سازماندهی اطلاعات در نظام‌های بازیابی اطلاعات

علی گزنی[1]

چکیده

هر نظام بازیابی اطلاعات (نرم‌افزار) دارای یک مبنای خاص برای تجزیه و تحلیل اطلاعات است ، که نظام براساس آن به تفسیر اطلاعات و مطابقت بین اقلام و درخواستهای اطلاعاتی پرداخته و بدین ترتیب بازیابی اطلاعات صورت می‌گیرد . این تجزیه و تحلیل سازماندهی اطلاعات نامیده می‌شود . بدون یک سازماندهی بهینه اطلاعات ، بازیابی اطلاعات به صورت کامل و دقیق صورت نخواهد گرفت . با توجّه به متفاوت بودن سیاست‌های بازیابی اطلاعات می‌بایست به صورت همزمان امکان استفاده از روشهای خودکار و نیمه خودکار فراهم آورده شود . پیش‌بینی سیاهه بازدارنده ، ایجاد انواع واژه نامه‌ها مانند واژه نامه ریشه لغات ، سیاهه پسوندها ، واژه نامه عبارات ، واژه‌نامه مفاهیم ، برقراری روابط سلسله مراتبی مفاهیم ، ریشه یابی واژگان ، محاسبه همبستگی و خوشه بندی اطلاعات همگی از امکاناتی هستند که می‌بایست در یک نظام بازیابی اطلاعات بهینه وجود داشته باشد . مقاله حاضر به بررسی این مفاهیم پرداخته است .

واژه‌های کلیدی : سازماندهی اطلاعات ، نظام های بازیابی اطلاعات ، فایل واژه‌نامه ، ریشه‌یابی واژگان، خوشه بندی اطلاعات .

مقدمه

بدون سازماندهی بهینه اطلاعات ، بازیابی اطلاعات به صورت کامل و دقیق صورت نخواهد گرفت . با توجه به متفاوت بودن سیاست‌های بازیابی اطلاعات می‌بایست به صورت همزمان امکان استفاده از روشهای خودکار و نیمه خودکار فراهم آورده شود . پیش‌بینی سیاهه بازدارنده ، ایجاد انواع واژه نامه‌ها مانند واژه‌نامه ریشه لغات ، سیاهه پسوندها ،

---

[1] ـ عضو هیات علمی کتابخانه منطقه‌ای علوم و تکنولوژی شیراز .

واژه‌نامه عبارات ، واژه‌نامه مفاهیم ، برقراری روابط سلسله مراتبی مفاهیم ، ریشه‌یابی واژگان ، محاسبه همبستگی و خوشه‌بندی اطلاعات همگی از امکاناتی هستند که می‌بایست در یک نظام بازیابی اطلاعات بهینه وجود داشته باشد . که در مقاله حاضر به بررسی این مفاهیم پرداخته شده است .

در انتهای جنگ جهانی دوّم یک سری مقالات علمی به رشته تحریر در آمدند که هرگز مورد استفاده عموم قرار نگرفت . در طول جنگ سرد (دهه ۱۹۵۰) گردآوری ، سازماندهی و اشاعه اطلاعات تخصصی (دبیزش و دکومانتاسیون) ، در زمینه‌های علمی و بخصوص تکنولوژیکی مورد توجه قرار گرفت . در طول سالهای ۱۹٤٥ تا دهه۱۹٦۰ نیاز به کنترل موضوعات یا نمایه سازی احساس شد . (Becker & Hayes, 1963; Bourne,1963; Herner, 1984; Weinberg, 1963) .

با گسترش منابع اطلاعاتی و ناهمگون بودن این منابع تاکید و توجه از نمایه‌سازی و شکلهای چاپی آن به سمت نظامهای بازیابی اطلاعات معطوف گشت ، زیرا صورتهای چاپی نمایه‌ای به تنهایی جوابگوی گسترش روزافزون اطلاعات نبودند (مانند چکیده‌نامه شیمی hemical Abstracts و نمایه پزشکی Index Mdeicus) . هر چند به صورت دقیق نمی‌توان تاریخ دقیقی را برای این تغییر جهت مشخص کرد، اما به عنوان یکی از اولین گامهایی که در این زمینه برداشته شده ، می‌توانیم از انتشار مطالبی در زمینه بازیابی اطلاعات توسط بوش[2] (۱۹٤٥) نام ببریم . با گذشت زمان رایانه‌های سریعتر و ارزانتری به بازار عرضه شد . این عامل خود باعث گردید تا چکیده‌های مدارک نیز با همان قالب زبان طبیعی و به عنوان قسمتی از رکوردهای کتابشناختی قابل ذخیره باشند . با افزایش قدرت نظامهای رایانه‌ای جستجو در چکیده مدارک نیز متداول گشت ، تا قبل از این ، نگهداری و ذخیره چکیده‌ها فقط با اهداف نمایشی صورت می‌گرفت . بتدریج جستجوی ریشه کلمات و کلمات متوالی مرسوم گشت و کم‌کم این امکان برای کاربران فراهم آورده شد تا به بازیابی اطلاعات مرتبط و مورد نیاز خود بپردازند . با هوشمندتر شدن نظامهای بازیابی اطلاعات بتدریج از اهمیت

---

2 . Bush

نمایه ها کاسته می شد، زیرا که این نظامها قادر به تشخیص عبارات و ترکیباتی بودند که به وسیله پدیدآورندگان متون اطلاعاتی بکار گرفته می‌شد . درحال حاضر و در بسیاری از نظام‌های بازیابی اطلاعات اختصاص توصیفگرها از طریق خود نظام صورت می‌گیرد . هر چند نمایه سازی هنوز هم از اهمیت برخوردار است ، اما پیشرفتهای صورت گرفته در این زمینه نمایه سازی (توسط انسان) را به صورت یک امر اختیاری در آورده است .

امروزه و با توجه به حجم انبوه اطلاعات ، دیگر ابزارها و شیوه های سنتی برای کنترل اطلاعات کافی نیست . استفاده از روش‌های غیر خودکار در این زمینه غیر اقتصادی بوده و مقرون به صرفه نمی‌باشد[٣]. هر چند که سازماندهی خودکار اطلاعات ، در نظام‌های بازیابی اطلاعات خارجی متداول می‌باشد ، اما در نظام‌هائی که در داخل کشور طراحی شده‌اند ، چیزی تحت عنوان سازماندهی اطلاعات وجود ندارد . در حالی که سازماندهی اطلاعات هسته مرکزی و ستون فقرات یک نظام بازیابی اطلاعات را تشکیل می‌دهد .

هر نظام اطلاعاتی دارای یک مبنای خاص برای تجزیه و تحلیل اطلاعات می‌باشد ، که نظام براساس آن به تفسیر اطلاعات و مطابقت بین مدارک و درخواستهای اطلاعاتی پرداخته و بدین ترتیب بازیابی اطلاعات صورت می‌گیرد . در کتابخانه‌ها معمولا این متخصصین رده‌بندی هستند که مناسب‌ترین طبقه را برای مدارک اطلاعاتی تعیین می‌کنند . در برخی مراکز در فرآیند نمایه سازی به هر مدرک توصیفگرهایی اختصاص داده می‌شود ، که بیانگر محتوای اطلاعاتی آنهاست و بازیابی اطلاعات براساس همین توصیفگرها صورت می‌گیرد . بنابراین برای انجام صحیح تجزیه و تحلیل اطلاعات به دستورالعملی دقیق و سنجیده نیاز داریم . این تجزیه و تحلیل سازماندهی اطلاعات نامیده می‌شود که در ادامه به بررسی جنبه‌های مختلف آن می‌پردازیم . در مقاله حاضر تاکید برروی تجزیه و تحلیل متون ، با قالب زبان طبیعی می‌باشد .

---

٣. به عنوان مثال سایت Google دارای ١/٣٦٨/٠٠٠/٠٠٠ صفحه اطلاعاتی می‌باشد . اگر قرار باشد سازماندهی اطلاعات در این سایت به صورت غیرخودکار صورت گیرد ، قاعدتاً متحمل هزینه بالائی خواهد بود . این سایت با آدرس HTTP://WWW.GOOGLE.COM در اینترنت قابل دسترسی می‌باشد .

تجزیه و تحلیل خودکار متون

یکی از شاخص‌های مهم در طبقه‌بندی نظام‌های بازیابی نحوه پردازش و تجزیه و تحلیل متون اطلاعاتی می‌باشد . به لحاظ اهمیت بحث تجزیه و تحلیل متون ، نظام‌های بازیابی اطلاعات به دو گروه اصلی هوشمند و غیرهوشمند تقسیم می‌شوند . در نظام‌های هوشمند پردازش اطلاعات به صورت خودکار صورت می پذیرد . در مقابل این نظام‌ها ، نظام‌های غیرهوشمند وجود دارند .

نظام‌های بازیابی اطلاعات هوشمند برای اولین بار بین سالهای ۱۹۶۲ تا ۱۹۶۵ در دانشگاه هاروارد طراحی و به کارگرفته شدند . در اینجا کاربرد واژه هوشمند برای نظام‌هائی است که در آنها تمام پردازشها ، برروی متن به صورت خودکار انجام می‌شود ، جستجو در آن صورت گرفته و مرتبط‌ترین اطلاعات متناسب با درخواست اطلاعاتی کاربر مورد بازیابی قرار می‌گیرد . در این نظام‌ها از رویه‌های متعددی برای طبقه‌بندی مدارک و درخواستهای اطلاعاتی استفاده می‌شود که این خود شامل الگوهای تطبیق واژه‌ها ، استفاده از واژه‌نامه‌ها برای گسترش دامنه مفاهیم و لغات ، استفاده از رویه‌های آماری و معنائی برای تعیین روابط موجود بین کلمات و مفاهیم ، رویه‌های ساخت عبارات و ... می‌باشد . بنابراین یک نظام باید قادر باشد ابزاری را فراهم آورد تا از طریق آن بتوان به گونه‌های مختلف به تحلیل محتوائی دست زد و در طی یک فرآیند تعاملی بین کاربر و نظام ، کاربر قادر خواهد بود ، تا نتایج دلخواه را بدست آورد . (Salton, 1964; Salton, 1965) .

ویژگیهای اصلی نظام‌های هوشمند

یک نظام اطلاعاتی هوشمند معمولاً دارای ویژگیهای زیر است :

۱ ـ سلسله عملیاتی که در طی آن اطلاعات مورد تجزیه و تحلیل قرار می‌گیرد باید به قدری واضح و عمیق باشد که بتوان مرتبط‌ترین اطلاعات ، مبتنی بر درخواستهای اطلاعاتی کاربران را در اختیار آنها قرار داد .

۲ ـ امکان بیان درخواست اطلاعاتی کاربر را به بهترین نحو ممکن فرآهم آورد .

۳ ـ یک مدرک با استفاده از فرمولهای مختلف قابل بازیابی باشد .

٤ ـ ضریب همبستگی بین کلمات براساس با هم رخدادن[4] کلمات و دفعات تکرار این با هم رخدادنها سنجیده می‌شود .

٥ ـ با استفاده از روشهای تجزیه و تحلیل نحوی ، عباراتی را برای شناسائی هر مدرک مشخص کنند و بین این عبارات نیز روابطی برقرار شود .

٦ ـ با استفاده از روشهای شناسائی آماری عبارات ، با استفاده از یک واژه‌نامه پیش ساخته ، همانند روش تجزیه و تحلیل نحوی ، عباراتی را برای شناسائی مدرک معرفی می‌کند با این تفاوت که میزان همبستگی بین ترکیبات سنجیده نمی‌شود .

٧ ـ رویه‌هایی وجود دارند که با استفاده از آنها درخواست کاربر مورد تجزیه و تحلیل قرار می‌گیرد و آنگاه با مدارک از قبل تجزیه و تحلیل شده مقایسه شده و بازیابی اطلاعات صورت می‌گیرد .

ضریب بازیابی و ضریب دقت به عنوان معیارهای ارزیابی در نظام‌های هوشمند مورد استفاده قرار می‌گیرند . ضریب بازیابی به درصد اطلاعات بازیابی شده نسبت به کل اطلاعاتی که احتمال بازیابی آنها وجود داشته است ، اطلاق می‌گردد و ضریب دقت به درصد اطلاعات مرتبط بازیابی شده نسبت به کل اطلاعات بازیابی شده ، اطلاق می‌گردد . در جواب این سئوال که آیا حدود مشخصی برای این ضرایب وجود دارد باید گفت که نسبت آنها با توجه به نیازهای اطلاعاتی کاربران ممکن است متفاوت باشد . بنابراین وقتی می‌توانیم یک نظام را مطلوب بدانیم که بتوان ضرایب بازیابی و دقت را به صورت دلخواه کنترل کرد .

مشکلات زبانشناختی در سازماندهی اطلاعات

هنگام پردازش متن‌هائی با قالب زبان طبیعی ، مسئله پیچیدگی زبان و بی‌قاعدگیهائی که در حوزه نحوی و معنائی وجود دارد خود را نشان خواهند داد . به عنوان مثال ، همانطوری که سالتون (١٩٦٤) نیز به آن اشاره دارد ، اگر برای اختصاص توصیفگرها به متن رویه‌ای نوشته شود ، احتمالاً با مشکلات زیر مواجه خواهیم بود :

---

4 . Co-occurrence

۱ ـ بعضی واژه‌ها به صورت مستقیم در واحدهای اطلاعاتی ظاهر نمی‌شوند ، اما به لحاظ عملکرد نحوی باید آنها را به حساب آورد (همانند استفاده از نشانه‌گذاریها در زبان انگلیسی) .

۲ ـ بعضی واژه‌ها برای رساندن معانی مشابه یا مرتبط بکار می‌روند (همانند مترادفها) .

۳ ـ بعضی از واژه‌ها به تناسب متنی که در آن آمده اند ، معانی متفاوتی دارند .

٤ ـ بعضی ترکیبات نحوی بیانگر یک ایده‌ء واحد و کلی می‌باشند .

٥ ـ در زبان طبیعی استفاده از ارجاعات غیرمستقیم متداول می‌باشد و گاهی اوقات پیدا کردن محلی که این ضمایر بدانجا اشاره دارند، مشکل می‌باشد .

٦ ـ معمولا بین واژه‌های داخل یک متن ارتباطاتی وجود دارد که فقط با خواندن متنهای قبلی قابل درک و استنتاج می‌باشند .

۷ ـ معانی بعضی واژه‌ها در طول زمان تغییر پیدا می‌کند و بدین لحاظ بین واژه‌های قدیم و جدید تداخل معنائی به وجود خواهد آمد .

واژه‌نامه‌ها

یک واژه‌نامه شامل گروهی از لغات یا ریشه لغات می‌باشد که در حوزه موضوعی خاصی طبقه‌بندی شده‌اند . هر نظام باید از روش خاصی برای متناسب سازی واژگان جهت بازیابی اطلاعات استفاده کند . یکی از راههای مؤثر در این زمینه استفاده از سیاهه واژگان است که آنرا واژه‌نامه یا تزاورس نیز می خوانند . در متن حاضر این سه واژه به یک مفهوم و به جای یکدیگر بکار می روند . واژه‌نامه‌ها قادر نیستند به صورت کامل ابهامات و پیچیدگی‌های موجود در زبان را برطرف کنند اما می‌توانند این بی‌قاعدگیها را کاهش دهند . واژه‌نامه‌ها می‌توانند از نقش مهمی در هدایت و واژه‌گزینی کاربران و بازیابی بهینه اطلاعات برخوردار باشند . در شکل (۱) نمونه‌ای از یک واژه‌نامه نمایش داده شده است .

واژه	شماره مفهوم	کدهای جملات
Block	58	070043040
Blueprint	324	070043
Bombard	346	070
Bond	105	043
Bookkeeping	001	070043
Boolean	28	00808080011
Borrow	28	070
Both	32178	070
Break	380	043
Brief	32232	070

شکل ١ : نمونه‌ای از یک واژه‌نامه ساخته شده در نظامهای بازیابی اطلاعات

این واژه‌ها با نظم الفبائی مرتب شده‌اند . شماره مربوط به مفهوم در وسط جدول دیده می‌شود . شماره‌های بالاتر از ٣٢٠٠٠ به لغات عمومی اختصاص داده شده است . ستون آخر حاوی کد جمله می‌باشد و در تحلیل نحوی به کار می‌رود .

انواع مهمتر واژه‌نامه‌هائی که در نظام‌های بازیابی اطلاعات به کار می‌روند عبارتند از :

١ ـ سیاهه بازدارنده[5] : شامل واژه‌هائی است که نباید مورد استفاده قرار گیرند .

٢ ـ واژه‌نامه ریشه لغات و سیاهه پسوندها : ریشه لغات و پسوندها میان ، به صورت جداگانه در هر مدخل نگهداری می‌شود .

٣ ـ واژه‌نامه عبارات : واژه‌ها براساس رخداد همزمان و میزان این رخداد به صورت یک عبارت در نظر گرفته می‌شوند . یک واژه‌نامه عبارات می‌تواند کیفیت و کارائی یک تحلیل محتوائی را بالا ببرد . با تعیین عبارات ، حالات مبهم معنائی در متن به مراتب کاهش خواهد یافت . به عنوان مثال واژه " برنامه نویسی" و " زبان" هر یک به تنهائی می‌توانند بیان کننده حالات مختلفی باشند ، اما قطعاً " زبان برنامه نویسی" یک معنای مشخص‌تری را القاء می‌کند .

5 . Stop Word List

٤ ـ واژه‌نامه مفاهیم (مترادفها) : در هر مدخل تعدادی از واژه‌های مترادف یا مفاهیم هم طبقه آورده شده است . گسترش دامنه مفاهیم و اصطلاحات در فرمول جستجو از طریق جایگزین کردن واژه‌های مبهم با واژه‌های هم طبقه از طریق این واژه‌نامه صورت می‌گیرد .

٥ ـ سلسله مراتب مفاهیم : مشابه طرحهای رده‌بندی کتابخانه‌ای می‌باشد ، با حرکت به بالای رده به قسمتهای اعم و با حرکت به قسمتهای پایینی به قسمتهای اخص خواهیم رسید .

مزایای اصلی واژه‌نامه مترادفها و عبارات ، با هدف تعیین محتوای مدارک را می‌توان در موارد زیر خلاصه کرد :

١ ـ امکان اختصاص دادن یک واحد اطلاعاتی به یک طبقه خاص را فراهم می‌آورد ، که در آنصورت می‌توان با جایگزینی و اضافه کردن واژه‌های مترادف در فرمول جستجو ، به نتایج بهتری در بازیابی اطلاعات دست یافت ، یا از طریق سنجش دفعات تکرار واژه‌های هم مفهوم (مترادف) در متن ، ضرایب مختلفی به هر مدرک اختصاص داد .

٢ ـ ابهامات زبانشناختی موجود در متون با استفاده از الگوی رخداد همزمان تا حدود زیادی کاهش می یابد .

٣ ـ برای تحلیل حوزه‌های مختلف موضوعی می‌توان از آنها استفاده کرد .

از ایرادات واژه‌نامه‌ها دشواری ساخت آنهاست ، بخصوص اگر در حوزه‌هائی باشد که دائما در حال تغییر و تحول باشند .

سازماندهی واژه‌ها

بهره‌گیری از واژه‌نامه‌ها به منظور ایجاد یکدستی معمولا دو گروه از سئوالات را بدنبال دارد ، که در ادامه به بررسی آنها می‌پردازیم :

الف ـ چه لغاتی باید به واژه‌نامه اضافه شود ؟

ب ـ کدام شیوه برای دسته‌بندی و طبقه‌بندی مفاهیم مناسب‌تر می‌باشد ؟

پاسخگوئی به سئوالات فوق وابسته به سیاستگذاری نظام در رابطه با ضرایب بازیابی و دقت می‌باشد . سه دسته عمده از لغات وجود دارند که باید در مورد شمول آنها در واژه‌نامه تصمیم‌گیری صورت گیرد ، این لغات عبارتند از :

١ ـ واژگان عمومی : لغاتی که دارای عملکرد عمومی بوده و در خارج از متن بیان کننده هیچ حالت موضوعی خاصی نمی‌باشند (مانند حروف ربط و اضافه) . افزودن این لغات موجب اشغال فضای اضافی و حذف آنها موجب ابهام در روابط معنائی ، خواهد شد .

٢ ـ واژگان با رخداد بالا[6] : تعدادی از واژه‌ها ، چه در حوزه‌های عمومی یا تخصصی (مانند واژه‌های نظام ، خودکار ، رایانه در علوم رایانه) از رخدادهای بالائی برخوردار هستند افزودن این گونه واژه‌ها موجب کاهش ضریب دقت خواهد شد و بازیابی ، اطلاعات نامرتبط زیادی را در پی خواهد داشت ، زیرا ، این واژه‌ها با توجه به دفعات تکرارشان از ضریب همبستگی بالائی با متن برخوردار هستند . در عین حال ، در صورت عدم درون‌داد آنها تعدادی از مدارک که می‌توانند در مقابل درخواست اطلاعاتی کاربر مورد بازیابی قرار گیرند بازیابی نخواهند شد .

٣ ـ واژگان با رخداد پایین[7] : گروهی از واژه‌ها از تعداد تکرار پایینی در متن‌ها برخوردار هستند . در صورتی که این گونه واژه‌ها نادیده گرفته شوند ، تعداد کمی از مدارک که در زمینه‌های مربوطه وجود دارند مورد بازیابی قرار نخواهند گرفت .

نوع طبقه‌بندی واژه‌ها و مفاهیم نیز از مسائلی هستند که در ضریب بازیابی و دقت تاثیر بسزائی خواهد داشت . بنابراین باید در مورد اعم یا اخص بودن طبقات تصمیم‌گیری صورت گیرد . در صورت اعم بودن طبقات شانس بازیابی مدارک غیرمرتبط بیشتر خواهد شد ، اما مدارک بیشتری مورد بازیابی قرار خواهند گرفت . در صورت اخص بودن طبقات ، بازیابی از ضریب دقت بالائی برخوردار خواهد بود ، اما احتمالاً تعدادی از مدارک مورد بازیابی قرار نخواهند گرفت . در هر صورت ، اگر واژه‌هائی که به لحاظ دفعات رخداد و خصوصیات دیگر

---

6 . *High Frequency*

7 . Low Frequency

مشابه می‌باشند در یک طبقه قرار گیرند ، و از لغات عمومی نیز برای ایجاد عبارات معنائی استفاده شود ، قاعدتاً از طبقات خاص‌تری بهره‌مند خواهیم بود .

اگر بخواهیم ضریب بازیابی را بالا ببریم ، باید تمام واژه‌هائی که می‌توانند بنحوی در بازیابی مدارک موثر باشند به واژه‌نامه افزوده شوند . هر چه واژه‌نامه عمومی‌تر باشد ، ضریب بازیابی افزایش و ضریب دقت کاهش خواهد یافت . در مقابل ، اگر بخواهیم ضریب دقت بالاتری داشته باشیم و اقلام مرتبط‌تری مورد جستجو قرار گیرند ، باید به سمت واژه‌نامه‌های تخصصی‌تر حرکت کنیم . در نهایت باید گفت ضرایب بازیابی و دقت می‌توانند برحسب محیط و نوع کاربران از درصدهای متفاوتی برخوردار باشند . برای مثال ، واژه‌نامه‌هائی که برای دو گروه دانش‌آموزان و دانشمندان طراحی می‌شود ، خیلی با هم متفاوت هستند .

در ادامه ، نکات مهمی که باید در ساخت واژه‌نامه در نظر گرفته شوند ، بیان می‌شود :

۱ ـ لازم است از درون داد مفاهیم نادر و کمیابی که احتمال درخواست برای آنها وجود ندارد، پرهیز شود.

۲ ـ واژه‌های خیلی عمومی که از تعداد رخداد بالائی نیز برخوردار هستند باید حذف شوند . به جای آن می‌توان از عباراتی ، که دارای معانی مشخص‌تری می‌باشند استفاده کرد به عنوان مثال ، واژه های "رایانه" و " کنترل " هر یک به تنهایی می توانند دامنه وسیعی را پوشش دهند . در حالی که عبارت " کنترل رایانه " مفهوم خاص و مشخص‌تری را بیان می‌کند .

۳ ـ کلماتی که دارای معانی مشخصی نیستند باید مورد مطالعه بیشتری قرار گیرند (مثلاً واژه " Hand " در تزاروس باید در زیر Biology قرار گیرد ولی تکرار زیاد آنها در متن می‌تواند به علت وجود عباراتی همچون " On The Other Hand " باشد) .

٤ ـ باید واژه‌های مبهم را با توجه به زمینه تخصصی که واژه در آن قرار دارد به مورد استفاده قرار داد و از تعمیم آن به سایر زمینه‌ها خودداری کرد .

٥ ـ در هر طبقه از مفاهیم ، واژگان باید از رخدادها و خصوصیات برابری برخوردار باشند .

استفاده از روشهای خودکار در ساخت واژه نامه‌ها

ایجاد واژه‌نامه‌های موضوعی نیاز به مهارت ، دانش ، تجربه و تلاش زیادی دارد . گروهی که مسئولیت ساخت واژه‌نامه را برعهده دارند ، علاوه بر داشتن تخصص موضوعی باید از مهارتهای کافی در زمینه زباشناسی و منطق برخوردار باشند . به دلیل گستردگی کار این کار به صورت گروهی و کمیته‌ای انجام می‌شود . این کمیته تعدادی سئوال را تعیین و مطرح می‌کند و در نهایت استانداردهای لازم را جهت تدوین واژه‌نامه پیشنهاد می‌کند (Dovel,1965) . باید توجه داشت که حتی پس از انجام کار ، و علیرغم تمام کارهای جدی که صورت پذیرفته ، ممکن است کسانی باشند که از واژه‌نامهٔ حاصل رضایت نداشته باشند .

در ساخت واژه‌نامه‌ها بهتر است از شیوه‌ای استاندارد استفاده شود . این امر مزایای زیر را در برخواهد داشت :

١ ـ علاوه بر کاهش هزینه‌ها و زمان تلف شده ، می‌توان به کنترل بازیابی اطلاعات پرداخت .

٢ ـ رویه‌های بازیابی می‌توانند در حوزه‌های مختلف موضوعی به کار برده شوند ، زیرا در اینجا با مشکلات ساخت واژه‌نامه‌ها مواجه نیستیم .

٣ ـ هر نوع تفاوت در حوزه‌های مختلف موضوعی را که ممکن است به دلیل کاربرد روشهای مختلف در ایجاد واژه‌نامه‌ها به وجود آیند و ممکن است در بازیابی موثر باشند را برطرف می‌کند .

٤ ـ امکان بررسی متغییرهائی که بر بازیابی تاثیر دارند را فراهم می آورد ، از جمله این متغییرها می‌توانیم از حجم واژه‌نامه ، تعداد طبقات مفاهیم ، و تعداد مفاهیم موجود در هر طبقه نام ببریم .

در ایجاد واژه‌نامه‌ها و تزاروسهای خودکار روش خاصی وجود ندارد . مهمترین مزیت این رویکرد این است که می‌توان به نحوی فعالیتها و کنشهای جستجوگر را کنترل کرد تا برای وارد کردن پرسش‌های مناسب به نظام بازیابی ، به دانش عمومی و تجربه زبانشناختی نیاز نداشته باشد . این مهم باعث همگانی شدن جستجو و در نتیجه بهره‌مندی از مزایای آن خواهد شد .

روشهای تمام خودکار

یکی از روشهای مورد استفاده در سازماندهی اطلاعات در نظام های بازیابی اطلاعات استفاده از روشهای تمام خودکار می‌باشد که تمام عملیات سازماندهی در این روش به صورت تمام خودکار صورت می‌گیرد .

در روش تمام خودکار تعدادی توصیفگر به هر مقاله اختصاص داده می‌شود . انتخاب این توصیفگرها براساس دفعات تکرار و خصوصیات ویژه‌ای می‌باشد که برای هر مجموعه مدرک در نظر گرفته می‌شود . در هر مورد مدارک نمونه ، به وسیله ماتریسی متشکل از مدارک نشان داده می‌شود . در این ماتریس ضرایب وزنی واژه‌ها در مدارک تعیین شده است . با استفاده از این ماتریس و روشهای آماری ، ضریب همبستگی بین واژه‌ها و مدارک به خوبی مشخص می‌گردد . میزان مشابهت و همبستگی بین مقالات موجود براساس رخداد همزمان واژه‌ها و مشخصه‌های واژه‌های مورد نظر می‌باشد .

اغلب روشهای مورد استفاده در ساخت واژه‌نامه‌ها و تزاروسها به صورت تمام خودکار ، به واژه‌های موجود در مجموعه مدارک وابسته می‌باشد .

Dennis, 1965; Dovel, 1965; Salton, 1966

روشهای نیمه خودکار

عملی بودن روش خودکار بیشتر از آنکه به یک شاخه موضوعی خاص وابسته باشد ، به مجموعه مدارک مورد پردازش بستگی دارد . بنابراین ، باید در رویه‌های پردازشی تغییراتی اعمال شود و از قضاوت انسانها نیز استفاده کنیم . روشهای مورد استفاده میتوانند براساس اهداف نظامها با هم فرق کند .

(Sparck-Jones, 1965; Levery, 1966; Abraham, 1965; Reisner, 1965)

در روش نیمه خودکار در ابتداء سیاهه‌ای از واژگان و تعداد رخداد آنها ایجاد می‌شود . این سیاهه مانند سیاهه‌ای است که در روش خودکار برای نشان دادن محتوای مدرک مورد

استفاده قرار می‌گرفت . معمولاً در این روش برای نمایش واژه‌ها از روش کوئیک[٨] استفاده می‌شود و واژه‌های مورد نظر در داخل جملات و در زیر یکدیگر مرتب می‌شوند .

مرحله بعدی طبقه‌بندی سیاهه واژگان می‌باشد . برای این منظور میتوان از روشهای مختلفی استفاده کرد :

١ ـ یک قضاوت غیر رسمی برای هر جفت واژه انجام گیرد تا مشخص شود که آیا آنها از لحاظ موضوعی با هم مترادف هستند .

٢ ـ یک مجموعه از قالبهای معنایی از قبل تعیین شده و واژه‌های متناسب با این قالب‌ها در یک گروه قرار می‌گیرند .

٣ ـ مجموعه‌ای از سئوالات ، برای گروه‌بندی واژه‌ها مطرح شده و مطابق با این سئوالات واژه‌ها در گروه‌ها قرار می‌گیرند .

در انتها ، برای هر واژه مشخصه‌های مختلفی تعیین می‌شود و سپس واژه‌نامه به همراه ماتریس خواص واژه‌ها قابل نمایش می‌باشد . سپس تفاوتهای معنایی بین واژه‌ها براساس مقایسه سطرهای ماتریس مشخص می‌شود . سطرهائی که با هم برابر هستند در یک گروه معنایی قرار می‌گیرند .

سیاهه بازدارنده

براساس سیاستهای و هدفهای نظام ، یک سیاهه بازدارنده تعریف می‌شود . این سیاهه شامل مجموعه علائم ، عبارات و کلماتی می‌باشد که هنگام سازماندهی اطلاعات از مجموعه علائم و کلمات استخراج شده حذف می گردند . به عنوان مثال حرف تعریف the ، یا حروف اضافه for,of می‌توانند جزء سیاهه بازدارنده تعریف شوند ، بنحوی که از مجموعه واژگان مدرک که مورد سازماندهی قرار می‌گیرند حذف گردند . استفاده از سیاهه بازدارنده از یک طرف می‌تواند حذف واژه‌های کم مایه که از رخداد بالائی نیز برخوردار هستند را به همراه

---

٨ ـ روش کوئیک ، اصولاً برای تولید نمایه‌ها براساس عناوین آثار کتابشناختی مورد استفاده قرار می‌گیرد . با استفاده از دستورالعمل‌های لازم حذف واژگان موجود در سیاهه بازدارنده ، رایانه می‌تواند از تمام واژه‌های نحوی مانند حروف تعریف و حروف اضافه چشم بپوشد ، اما تمام واژه‌های باقی مانده در عناوین آثار را به عنوان واژه‌های نمایه‌سازی انتخاب کند . نتیجه ، نمایه کلید واژه‌هایی است که به ترتیب الفبائی توام با عناوین مدارک چاپ می‌شوند .

داشته باشد و در همان ضمن می‌تواند از بین رفتن روابط معنایی را نیز بدنبال داشته باشد در بعضی نظام‌ها قیود ، ضمایر و حروف اضافه را جزو سیاهه بازدارنده قرار می‌دهند .

واژه نامه ریشه لغات و لیست پسوندها

ریشه‌های کلمات می‌توانند با اضافه شدن پسوندها و پیشوندها شکلهای مختلفی به خود بگیرند . در حقیقت بسیاری از کلمات موجود در متون دارای یک ریشه واحد بوده و به یک مفهوم اشاره می‌کنند . اگر ریشه‌یابی در مورد این واژه‌ها صورت نگیرد ، بازیابی اطلاعات به صورت کامل انجام نخواهد شد . ریشه‌یابی موجب نوعی یکدستی خواهد شد . به هر ریشه یک شماره اختصاص داده می‌شود . یک نمونه از واژه‌نامه ریشه در شکل (۲) نمایش داده شده است ، ریشه‌های موجود در این سیاهه برحسب دفعات رخداد آنها در متون مرتب شده‌اند .

شماره ترتیب	پسوند	ریشه	تعداد تکرار
2099	S	PLACE	11
2100		RESPONSE	11
2101		THICK	12
2102	ATION	TRUNC	15
2103	ICAL	ALPHABET	18
2014	ABLE	CAP	19

شکل ۲ : نمونه ای از یک واژه نامه ریشه

برای بدست آوردن ریشه کلمات در زبان انگلیسی روشهای مختلفی وجود دارد که در اینجا به بررسی مجموعه قواعد مطرح شده توسط لاوینز[9] می‌پردازیم :

۱ ـ اگر حرف انتهای کلمه یک صامت غیر از S باشد و بعدش S آمده باشد S حذف می‌شود .

۲ ـ اگر یک کلمه با ES ختم شود S که آخرین حرف می‌باشد برداشته می‌شود (این قاعده در مورد کلمات جمع یونانی الاصل مشکل ایجاد می‌کند) .

9 . Lovins

٣ ـ انتهای IEV به IEF و METR به METER تبدیل می‌شود .

٤ ـ اگر یک کلمه به ING ختم شود ING حذف می‌شود ، مگر اینکه بعد از حذف باقیمانده فقط یک حرف یا TH باقی مانده باشد .

٥ ـ اگر یک کلمه با ED ختم شود و قبل از آن یک حرف صامت وجود داشته باشد ED حذف می‌شود .

٦ ـ اگر انتهای لغت برداشته شد و آخر کلمه بدست آمده BB,DD,...,TT بود ، یکی از حروف مضاعف برداشته می‌شود (مثلاً BB می‌شود B) .

٧ ـ اگر انتهای واژه ION بود ، ION برداشته می‌شود مگر اینکه حروف باقیمانده دو یا یک حرف باشد . اگر حرف آخر ریشه صامت باشد و حرف قبل از آن صدادار باشد یک E اضافه می‌شود .

یک برنامه قابل استفاده برای ریشه‌یابی باید حدود ١٠ تا ٢٠ قاعده را در برداشته باشد و تعداد زیادی از موارد خاص را نیز بتواند پوشش دهد . در راستای ایجاد واژه‌نامه ریشه باید مسائل زیر را مد نظر داشته باشیم :

١ ـ آیا در واژه‌نامه باید شکل کامل لغت نگهداری گردد (ریشه و پسوند در یک مدخل ولی به صورت جداگانه) یا با حذف پسوند فقط ریشه اصلی نگهداری گردد ؟ تجربیات در این مورد نشان می‌دهد که باید از هر دو نوع واژه‌نامه یعنی شکل کامل و ریشه استفاده کرد به عنوان مثال در یک مجموعه از مدارک که واژگان موجود در چکیده های آنها بالغ بر ٥٠٠٠٠ عدد باشد ، اگر بخواهیم یک واژه‌نامه ریشه با ریشه کامل کلمات را داشته باشیم ، تعداد ٢٨٠٠ مدخل را خواهیم داشت و اگر ریشه را به صورت جزئی‌تر نگهداری کنیم ، تعداد مدخلها بالغ بر ٩٠٠ عدد خواهد شد (با این احتساب که هر مدخل دارای حداقل ٤ تکرار در چکیده بوده باشد) . به هر حال نگهداری ریشه کلمات نتایج بهتر و رضایت بخش‌تری را بدنبال خواهد داشت .

٢ ـ برای هر ریشه یک مدخل در نظر گرفته خواهد شد . واژه‌هائی که در یک مجموعه مدارک خاص یا حوزه موضوعی خاص دارای اهمیت نمی‌باشند یا باید با کد مخصوصی قابل تشخیص باشند و یا اینکه شامل نشوند .

۳ ـ آیا تمام واژه‌ها باید در واژه‌نامه وارد شوند ؟ یا از یک روش مبتنی بر تعداد تکرار واژگان استفاده شود (مثلاً از ۵ تکرار در یک مدرک کمتر نباشد و از ۱۰ تکرار هم بیشتر نباشد) .

٤ ـ یک سیاهه بازدارنده متشکل از واژه‌های عمومی برای جلوگیری از وارد شدن این واژه‌ها باید مورد استفاده قرار گیرد .

۵ ـ سیاست اصلی در نظام‌های هوشمند استفاده از کوتاهترین راه می‌باشد ، بنحوی که نیاز به انجام حداقل عملیات ممکن باشد .

ساختار واژه‌نامه‌های پسوندی شامل ریشه‌واژه ، شماره مربوط به ریشه و کد پسوند یا کد معنائی است . ریشه و پسوند باید به گونه‌ای آورده شوند که هیچ گونه ابهامی در ترکیب آندو با هم وجود نداشته باشد .

واژه‌نامه عبارات

در یک واژه‌نامه ریشه لغات ، هر مدخل براساس یک واژه یا ریشه واژه شکل می‌گیرد . امکان تحلیل و استنباط موضوعی از روی این واژه‌های منفرد کاری دشوار و همراه با خطا می‌باشد . در صورتی که ترکیب چند واژه معنای اخص‌تری را بیان می‌کند . این ترکیبات علاوه برآنکه امکان تحلیل موضوعی متن را فراهم می‌آورند ، در تعیین حوزه‌های موضوعی نیز مورد استفاده قرار می‌گیرند . به همین منظور در نظام های بازیابی اطلاعات از واژه‌نامه عبارات استفاده می‌شود . یک عبارت می تواند از دو یا سه یا چهار واژه یا مفهوم ایجاد شده باشد .

در ایجاد و ساخت واژه‌نامه عبارات از استراتژیهای مختلفی استفاده می‌گردد . به عنوان مثال این عبارات براساس دفعات رخداد بالای آنها در متون یا درخواستهای اطلاعاتی کاربر در نظر گرفته می‌شوند . البته امکان دارد قبل از تخصیص عبارات ، آنها را با یکی از واژه‌نامه‌های موجود نیز مطابقت دهند تا از صحت عبارات مطمئن شوند . در نظام‌های هوشمند به منظور ایجاد یکدستی از واژه نامه‌های عبارات موجود برای تعیین این عبارات استفاده می‌گردد .

به طور کلی برای ایجاد واژه‌نامه‌های عبارات از دو راهبرد استفاده می شود :

١ ـ واژه‌نامه‌های عباراتی که با روشهای آماری تهیه می‌شوند ، که براساس دفعات رخداد همزمان واژه‌ها در متن می‌باشد . در این روش هیچ گونه مطالعه‌ای برای درک روابط معنایی بین واژه‌های موجود در عبارت صورت نمی‌گیرد .

٢ ـ واژه‌نامه عبارات معنائی که در این روش علاوه بر تشخیص عبارات ، باید ارتباط معنائی بین عبارات نیز تشخیص داده شود ، تا به عنوان یک ترکیب قابل قبول پذیرفته شود .

تعیین سلسله مراتب مفاهیم

تعیین نظم سلسله مراتبی واژه‌ها از قسمتهای مهم در نظام‌های بازیابی اطلاعات می‌باشد استفاده از این سلسله مراتب‌ها امکان حرکت از مفاهیم اعم به اخص و از اخص به اعم را می‌سازد . اساس این روش نیز دفعات رخداد کلمات می‌باشد . کلمات با رخدادهای بالاتر در یک طبقه و کلمات با رخدادهای پایین‌تر در طبقات دیگری قرار می گیرند (Doyle, 1965) . پیوستگی سلسله مراتبی با استفاده از مشخصه دفعات رخداد بدست می آید . در اینجا باید این نکته را متذکر شویم که ایجاد یک نظم سلسله مراتبی بصورت تمام خودکار کاری بیهوده می‌باشد . زیرا ساختار حاصل بدون دخالت و قضاوت انسان ساختار متناسبی نخواهد داشت . در ادامه به بررسی مختصر دو روش ساخت این سلسله مراتبها می‌پردازیم :

در روش اول برای ایجاد سلسه مراتب ، اولین مرحله پردازش به صورت مستقیم و با طرح مجموعه سئوالاتی به منظور طبقه‌بندی اولیه واژه‌نامه صورت می‌گیرد . سئوالات بعدی در داخل هر یک از طبقات مطرح می‌گردد و این روند تا رسیدن به طبقات فرعی ادامه پیدا می‌کند .

در روش دوم در پردازش از دفعات رخداد کلمات استفاده می‌گردد و هیچ گروه از پیش تعیین شده‌ای وجود ندارد . به جای آنکه سلسله مراتب در ابتدا ایجاد و سپس واژه‌نامه براساس آن پایه‌ریزی شود ، برای شروع سیاهه‌ای از تکرار واژه‌ها و مشخصه‌های آن (ماتریس) تهیه می‌شود . در ابتدا واژه‌نامه به دو شاخه تقسیم می‌شود و در بالای آن واژه‌ای که بالاترین تعداد رخداد را دارد نگهداری می‌شود ، که آن را T_i می‌نامیم ، سپس تمام

واژه‌های مرتبط با T_i در یک شاخه ، و تمام واژه‌های نامرتبط با T_i در شاخه دیگری قرار می‌گیرند . گروهی از واژه‌ها که بالاترین رخداد را داشته است انتخاب می‌شود و از آن به عنوان معیاری برای شاخه شاخه کردن استفاده می‌شود و این رویه تا زمانی که واژه‌ها به گروههای کوچک تقسیم شوند ادامه پیدا می‌کند .

باید اضافه کرد که برای شاخه شاخه کردن باید تصمیمات زیر اتخاذ شود :

۱ ـ کلماتی که بالاترین تعداد رخداد را داشته اند ، به عنوان مرکز شاخه انتخاب می‌شوند و سایر واژه‌ها متناسب با میزان ارتباط آنها با واژه مرکزی تقسیم‌بندی و شاخه شاخه می‌شوند .

۲ ـ اگر یک واژه نتواند به صورت صحیح در یکی از دو طبقه قرار گیرد (چه مرتبط یا نامرتبط با واژه مرکزی) در جای خود به عنوان یک واژه اعم قرار می‌گیرد .

۳ ـ اگر چند واژه در یک گروه دارای دفعات رخداد بالائی باشند ، این واژه‌ها به یک مرحله بالاتر رانده شده و به عنوان یک تقسیم مستقل و کلی به حساب خواهند آمد .

آخرین مرحله پردازش برروی قالب سلسله مراتبی براساس واژه‌های مدارک یا ماتریس خواص صورت می‌گیرد . از مشخصه‌های اصلی در این میان ضریب وزنی (اهمیت) واژه‌ها می‌باشد . این مرحله به صورت تمام خودکار صورت می‌گیرد .

یک‌بُردار خواص با K بعد V_j,V_i که نمایانگر واژه های T_j,T_i میباشند ، ایجاد می شود و برای واژه‌های j,i میزان مشابهت و همبستگی از طریق فرمول زیر محاسبه می‌شود :

$$Cij = \frac{\sum_k MIN(V_k^i, V_k^j)}{\sum_k V_k^i}$$

در فرمول بالا C_{ij} بیانگر میزان همبستگی می‌باشد .

محاسبه همبستگی بین مدارک

همانطوریکه در مباحث قبلی نیز بیان شد ، دفعات رخدادهای همزمان واژه ها می‌تواند به عنوان معیاری برای سنجش میزان مشابهت و همبستگی بین دو متن بکار برود . میانگین

حاصل از این دفعات رخداد بنحو شایسته‌تری بیان کننده این همبستگی خواهد بود . اگر بخواهیم این همبستگی مقدار بالاتر و معنادارتری را نشان دهد باید این ارزیابی براساس لغات مشترک و غیرمشترک بین دو متن صورت گیرد . به منظور انجام این محاسبه فرمول‌های مختلفی پیشنهاد شده است که دراینجا به بررسی فرمول (McGill, Salton,1983) به قرار زیر می‌پردازیم :

$$SIM(j,k) = \frac{\sum_{i=1}^{n} t_{ij} . t_{ik}}{\sum_{i=1}^{n} (t_{ij})^2 + \sum_{i=1}^{n} (t_{ik})^2 - \sum_{i=1}^{n} t_{ij} . t_{ik}}$$

در این فرمول t_{ij} نشان دهنده ضریب وزنی (دفعات تکرار) واژه i در متن j و t_{ik} نشان دهند ضریب وزنی واژه i در متن k می باشد . n مجموع واژه‌های موجود در دو مدرک باشد . نکته مهم اینست که اگر واژه i در مدرک j موجود نباشد ، t_{ij} برابر صفر خواهد بود و اگر واژه i در هیچ یک از دو مدرک وجود نداشته باشد در آنصورت نیز $t_{ij}.t_{ik}$ برابر صفر خواهد بود .

خوشه بندی

هر رکورد اطلاعاتی از تعدادی مشخصه یا فیلد تشکیل شده است . هر رکورد اطلاعاتی در مقایسه با سایر رکوردهای اطلاعاتی از تعدادی مشخصه‌های منحصر بفرد و مشابه برخوردار است . با توجه به این مشابهت و نحوه توزیع آن ، می‌توان رکوردهای اطلاعاتی را در گروه‌های مشخصی دسته‌بندی کرد . به این عمل را خوشه‌بندی[10] می گویند . بعد از انجام خوشه‌بندی با درون داد هر مدرک جدید می توان تشخیص داد که این مدرک در کدام خوشه قرار می‌گیرد . با ایجاد این خوشه و با محاسبه درخواست اطلاعاتی کاربر می‌توان فهمید که کدام خوشه برای پاسخگوئی به پرسش جاری مناسب‌تر است . این نکته قابل توجه می‌باشد

10 . Clustering

که ابتدا خوشه‌بندی رکوردها مطابق با مشخصه‌هایشان صورت می‌گیرد . سپس رکوردها در داخل گروه‌های مختلفی قرار می‌گیرند . هیچ گروه از پیش تعیین شده‌ای در این رابطه وجود ندارد . به منظور ایجاد خوشه‌ها فرمول‌های مختلفی ارائه شده که در اینجا به بررسی یکی از ساده‌ترین فرمولهای موجود که توسط بُتر[11] ارائه شده است ، می‌پردازیم :

$$s_{ij} = \frac{T_{ij}}{T_{ii} + T_{jj} - T_{ij}}$$

در این فرمول T_{ij} بیانگر تعداد واژه‌های مشترکی می‌باشد که در دو رکورد i,j وجود دارد . T_{ji}, T_{ii} به ترتیب نشان دهنده کل واژه‌های موجود در رکوردهای j,i می‌باشند . اولین قدم در ایجاد یک خوشه تشکیل یک ماتریس از مشخصه‌های رکوردها می‌باشد . به هر رکورد به صورت مجزا یک سطر اختصاص داده می‌شود و هر واژه یا مشخصه به صورت یک ستون نمایش داده می‌شود .

در هر سلول ماتریس ، مقدار "۱" برای وجود و "۰" برای عدم وجود یک مشخصه در رکورد مورد نظر بکار برده می‌شود . به مثال شکل (۳) که برای مجموعه کوچکی از مدارک آورده شده است توجه کنید .

شماره رکورد	مشخصه یک	مشخصه دو	مشخصه سه	مشخصه چهار	مشخصه پنج	مشخصه شش
۱	۱	۰	۰	۱	۰	۰
۲	۱	۱	۰	۱	۰	۰
۳	۰	۰	۱	۱	۱	۱
۴	۰	۱	۱	۰	۰	۱
۵	۱	۰	۰	۱	۱	۰
۶	۰	۰	۱	۰	۱	۰
۷	۰	۱	۰	۱	۰	۱
۸	۱	۱	۱	۰	۰	۰

شکل ۳ : ماتریس مشخصه های رکوردهای اطلاعاتی

11 . Boner

اولین ماتریس مشابهت SM1 همانوریکه در شکل (٤) دیده می شود ، با محاسبه مقدار S برای رکوردها شکل می‌گیرد . به عنوان مثال ، در اولین رکورد ، مشخصه‌های ٤ و١ و در دومین رکورد مشخصه‌های ١و٢و٤ وجود دارد ، که در آنها دو مشخصه به صورت مشترک به کار رفته است ، بنابراین مقدار S برای رکوردهای ١و٢ برابر است با :

$$\frac{2}{2+3-2}=\frac{2}{3}=2/3$$

به منظور محاسبه SM2 مقدار SM1 مطابق با حد آستانه[١٢] در نظر گرفته شده و محاسبه می‌گردد . حد آستانه در مثال جاری برابر ٤٥/. می‌باشد . با تغییر حد آستانه تعداد ، اندازه و میزان همبستگی بین اعضای یک خوشه تغییر خواهد کرد . تعیین یک مقدار برای حد آستانه چیزیست که در عمل و تجربه مشخص می‌گردد و با تعیین مقادیر مختلف و خوشه‌بندی بهترین حد انتخاب می‌شود . محاسبه SM2 با حد آستانه ٤٥/. در شکل (٥) نشان داده شده است . پس از انجام این محاسبه ، خوشه‌های اولیه با مقدار ١ تشکیل می‌شوند . خوشه‌های بدست آمده در شکل (٦) نشان داده شده‌اند . خوشه‌ها براساس شماره رکورد مرتب شده‌اند . در ادامه مراحل زیر نیز باید صورت پذیرد :

١ ـ خوشه‌های تکراری را حذف کنید (رکورد ٥ تکرار رکورد ١ و رکورد ٦ تکرار رکورد ٣ بوده و حذف می‌شود . خوشه‌ها در حال حاضر عبارتند از : ٨،٧،٤،٣،٢،١).

٢ ـ خوشه‌هائی که در خوشه‌های دیگر وجود دارند را حذف کنید (خوشه ٢ در خوشه ١و٤ وجود دارد . خوشه در حال حاضر عبارتند از : ٨،٧،٤،٣،١) .

٣ ـ خوشه‌هائی که اعضاء آن در خوشه‌های دیگر تکرار شده‌اند را حذف کنید (خوشه ٤ در خوشه ٧ ، ٨ تکرار شده است و خوشه ٧ ، ٨ نیز تمام مقادیرشان در خوشه‌های ١ ، ٤ تکرار شده است . حذف خوشه‌های ٧ ، ٨ باعث کم شدن گروهها خواهد شد تا اینکه بخواهیم ٤ را حذف کنیم) .

---

[١٢] ـ مقدار حد آستانه می‌تواند از یک نظام به نظام دیگر متفاوت باشد و با تغییر آن و مشاهده نتایج می‌توان عدد صحیح را انتخاب کرد.

باقیمانده خوشه‌ها ۱، ۳ و ٤ می‌باشند که رکوردهای ۱ ، ۲ ، ۵ ؛ ۳ ، ٦ و ٤ ، ۷ ، ۸ را در بر دارند .

شماره رکورد	۱	۲	۳	٤	۵	٦	۷	۸
۱	۱	۲/۳	۱/۵	۰	۲/۳	۰	۱/۴	۱/۴
۲		۱	۱/۶	۱/۵	۲/۴	۰	۲/۴	۲/۴
۳			۱	۲/۵	۲/۵	۲/۴	۲/۵	۱/۶
۴				۱	۰	۱/۴	۲/۴	۲/۴
۵					۱	۱/۴	۱/۶	۱/۵
۶						۱	۰	۱/۴
۷							۱	۱/۵
۸								۱

شکل ۴ : ماتریس مشابهت رکوردهای اطلاعاتی

شماره رکورد	۱	۲	۳	٤	۵	٦	۷	۸
۱	۱	۱	۰	۰	۱	۰	۰	۰
۲	۱	۱	۰	۰	۱	۰	۱	۱
۳	۰	۰	۱	۰	۰	۱	۰	۰
٤	۰	۰	۰	۱	۰	۰	۱	۱
۵	۱	۱	۰	۰	۱	۰	۰	۰
٦	۰	۰	۱	۰	۰	۱	۰	۰
۷	۰	۱	۰	۱	۰	۰	۱	۰
۸	۰	۱	۰	۱	۰	۰	۰	۱

شکل ۵ : ماتریس مشابهت رکوردها با محاسبه حد آستانه

شماره خوشه‌های ایجاد شده	رکوردهای موجود در هر خوشه
١	٥،٢،١
٢	٨،٧،٥،٢،١
٣	٦،٣
٤	٨،٧،٤
٥	٥،٢،١
٦	٦،٣
٧	٧،٤،٢
٨	٨،٤،٢

شکل ۶: رکوردهای خوشه بندی شده

نتیجه گیری

امروزه با توجه به حجم انبوه اطلاعات و نگهداری اطلاعات تمام متن ، دیگر ابزارها و شیوه‌های سنتی برای کنترل اطلاعات کافی نمی‌باشند . استفاده از روش‌های غیرخودکار در این زمینه غیراقتصادی بوده و با توجّه به هوشمندتر شدن نظام‌ها مقرون به صرفه نمی‌باشد . هر چند که سازماندهی خودکار اطلاعات ، در نظام‌های بازیابی اطلاعات خارجی متداول می‌باشد ، امّا در نظام‌هائی که در داخل کشور طراحی شده‌اند ، چیزی تحت عنوان سازماندهی اطلاعات وجود ندارد . سازماندهی اطلاعات هسته مرکزی و ستون فقرات یک نظام بازیابی اطلاعات را تشکیل می‌دهد .

هر نظام اطلاعاتی دارای یک مبنای خاص برای تجزیه و تحلیل اطلاعات می‌باشد ، که نظام براساس آن به تفسیر اطلاعات و مطابقت بین اقلام و درخواستهای اطلاعاتی می‌پردازد. به لحاظ اهمیت سازماندهی اطلاعات نظام‌های بازیابی اطلاعات به دو دسته هوشمند و غیرهوشمند تقسیم می‌شوند .

هنگام پردازش متن‌هائی با زبان طبیعی ، مسئله پیچیدگی زبان و بی قاعدگیهائی که در حوزه نحوی و معنائی وجود دارد خود را نشان خواهند داد که باید این مشکلات را مد نظر قرار داد . در این میان واژه‌نامه‌ها قادر نیستند به صورت کامل ابهامات و پیچیدگیهای موجود در زبان را برطرف کنند اما می‌توانند این بی‌قاعدگیها را کاهش دهند . این واژه‌نامه‌ها

شامل سیاهه بازدارنده ، واژه‌نامه ریشه لغات و سیاهه پسوندها ، واژه‌نامه عبارات ، واژه‌نامه مفاهیم ، برقراری روابط سلسله مراتبی بین مفاهیم و ریشه‌یابی واژگان می‌شود .

به منظور سازماندهی واژه‌ها باید دامنه شمول واژه‌ها و نوع دسته‌بندی مفاهیم را مدنظر قرار داد . پاسخگوئی به مسائل فوق وابسته به سیاستگذاری نظام در رابطه با ضرایب بازیابی و دقت می‌باشد . اگر بخواهیم ضریب بازیابی را بالا ببریم باید تمام واژه‌هائی که می‌توانند بنحوی در بازیابی مدارک موثر باشند ، به واژه‌نامه اضافه شوند . هر چه واژه‌نامه عمومی‌تر باشد ، ضریب بازیابی افزایش و ضریب دقت کاهش خواهد یافت . در مقابل اگر بخواهیم ضریب دقت بالاتری داشته باشیم و اقلام مرتبط‌تری مورد جستجو قرار گیرند باید به سمت واژه‌نامه‌های تخصصی‌تر حرکت کنیم .

ساخت واژه‌نامه‌ها به صورت خودکار یا دو روش تمام خودکار و نیمه خودکار قابل اجرا می باشد ، به دلیل اینکه عملی بودن روش تمام خودکار بیشتر آنکه به یک شاخه موضوعی خاص وابسته باشد به مجموعه مدارک مورد پردازش بستگی دارد ، بنابراین در رویه‌های پردازشی تغییراتی اعمال شده و از قضاوت انسانها استفاده می‌شود .

به منظور بالا بردن کارائی نظام های بازیابی اطلاعات از روشهای محاسبه همبستگی و خوشه‌بندی اطلاعات نیز استفاده می‌گردد که محاسبه همبستگی به عنوان معیاری برای سنجش میزان مشابهت و همبستگی بین دو متن و خوشه‌بندی برای دسته‌بندی رکوردهای اطلاعاتی مشابه با هم مورد استفاده قرار می‌گیرد .

منـابع :

Abraham. C.T. 1965. Techniques for Thesaurus Organization and Evaluation. **Information Science** 4(4) :121-150.

Becker, J. Hayes, R. 1963. **Information Storage and Retrieval: Tools, Elements: Theories**. New York: John Wiley & Sons. Bedford. Massachusetts: C. Cleverdon.

Belkin, N. J. and Croft, W. B. 1987. Retrieval Techniques. **Annual Review of Information Science & Technology** (22):109-145.

Bonner, R.E. 1964. On Some Clustering Techniques. **IBM Journal of Research** (8): 22-32.

Bourne, C.P. 1963. **Methods of Information Handling**. New York: John Wiley & Sons.

Bush, V. 1945. "As We May Think". **Atlantic Monthly** 176 (1):101-108.

Dennis, S.F. 1965. **The Construction of a Thesaurus Automatically from a Sample of Text.** *Symposium on Statistical Association Methods for Mechanized Documentation.* Natl. U.S. Misc. Publ. 369 (December).

Dovel, J.A. and Heald J.H. 1965. **Project Lex Status**. Proceeding ADI Meeting. Santa Monica, Calif.

Doyle, L.B. 1965. "Expanding the Editing Function in Language Data Processing". **Commum**. ACM 8(4) : 126-138.

Herner, S. 1984. Brief History of Information Science. **Journal of American Society for Information Science** 35(5):157-163.

Levery, F. 1966. Organization et Consultation d 'un Thesaurus. **1965 FID Congress, Spartan Books.** Washington, D.C.

Lovins, J.B. 1968. Developing of Stemming Algorithm. **Mechanical Translation and Computational Linguistics** 11(1) :22-31 .

Resiner, P. 1965. Semantic, Diversity and a Growing Mac-Machine Thesaurus. **Information Science** 4(4):117-130.

Salton, G. 1964. **A Document Retrieval System For Mac-Machine Interaction**. Proceeding 19th ACM Natl. Philadelphia ,Pa.

Salton, G. 1965. Progress in Automatic Information Retrieval. **IEEE Spectrum** 2(8):90-103.

Salton, G. 1966. Data Manipulation and Programming Problems in Automatic Information Retrieval. **Commun**. ACM 9(3) 121-135.

Salton, G. and McGill, M. 1983. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill.

Sparck-Jones, K. 1965. Experiments in Semantic Classification. **Mechý. Transl** 8(4): 97-112.

Weinberg, A.M. and the President's Science Advisory Committee. 1963. **Science, Government, and Information: The Responsibilities of the Technical Community and the Government in the Transfer of Information**. Washington, D.C:U.S: Government Printing Office.