مقاله

جست‌وجوی اطلاعات و ساز و کارهای بهینه سازی آن

علی گزنی[1]

تاریخ دریافت: ۷۹/۳/۱۷

چکیده: تحقیق در زمینهٔ بازیابی اطلاعات راهگشای طراحی، کاربرد، و استفادهٔ مؤثر از ابزارهای کنترل اطلاعات بوده، و کنترل بر روی حجم گسترده‌ای از اطلاعات، محیط‌های اینترنت و بانک‌های صوتی ـ تصویری را به همراه دارد. مقالهٔ حاضر به صورت مختصر به تشریح یک نظام بازیابی اطلاعات (نرم‌افزار بازیابی اطلاعات) پرداخته و سپس بر روی فرمول جست‌وجو در این نظام‌ها متمرکز می‌شود، و آنگاه به بررسی امکانات و قابلیت‌های موجود در یک نظام بازیابی اطلاعات برای شکل‌گیری فرمول جست‌وجو می‌پردازد. امکانات و قابلیت‌های مورد بحث در این مقاله شامل عملگرهای منطقی، عملگرهای مجاورت، جست‌وجوی املائی، کوته‌سازی، تعیین محدوده در داده‌های کمی، گسترش دامنهٔ مفاهیم و اصطلاحات، ضریب وزنی، سیاههٔ واژه‌ها، مجموعه‌های بولی نمره‌گذاری شده، و اندازه‌گیری مشابهت‌هاست.

کلید واژه‌ها: بازیابی اطلاعات، نرم‌افزار، اندازه‌گیری، سیاههٔ واژگان

مقدمه

مسئله بازیابی اطلاعات از زمانی آغاز شد که بشر سعی کرد محیط پیرامون خود را کنترل یا حداقل از فشارهای خارجی که باعث نابودی او می‌شدند جلوگیری کند. بشر برای ایجاد محیطی مطلوب جهت ادامهٔ زندگی نیاز به اتخاذ تصمیمات سریع، صحیح، و دقیق داشت. کیفیت این تصمیمات به توانایی تصمیم گیرنده در حل مسائل وابسته بود ولی قبل از آن به میزان ارتباط و کیفیت اطلاعاتی وابسته بود که تصمیم گیرنده برای حل مشکل فراهم می‌آورد. به‌تدریج و در طول تاریخ، جمع‌آوری، سازماندهی و نگهداری اطلاعات امری متداول و مرسوم شد.

---

۱. کارشناس ارشد کتابداری و اطلاع‌رسانی کتابخانهٔ منطقه‌ای علوم و تکنولوژی شیراز

جنگ جهانی دوم را می‌توانیم نقطه تحرکی برای نظام‌های بازیابی اطلاعات[1] در نظر بگیریم، زیرا حجم انتشارات خصوصاً در زمینه‌های علمی، به صورت تصاعدی در حال افزایش بود. دامنه تحقیقات و پژوهش‌ها روزبه روز گسترده‌تر می‌گردید و این افزایش تدریجی، کنترل اطلاعات با استفاده از ابزارها و شیوه‌های قبلی را غیرممکن می‌ساخت.

واژه بازیابی اطلاعات در سال ۱۹۵۱ برای اولین بار توسط تئوریسینی بنام موئرز[2] به کار رفت. وی معتقد بود که یک نظام بازیابی اطلاعات باید امکان نمایه سازی و جست‌وجوی اطلاعات موجود در کتابخانه را فراهم آورد، او با استفاده از روش‌های نظری و بررسی ساختار تلگراف توانست برای کد گذاری اطلاعات از راهی مؤثر و کم خطا استفاده کند. اولین کتاب نظام‌های بازیابی اطلاعات در سال ۱۹۶۱ توسط ویکری[3] ارائه شد. در این کتاب نحوهٔ ایجاد نمایه‌ها و توصیف موضوعی مدارک به صورت گسترده‌تری مورد بررسی قرار گرفت. انگیزهٔ اصلی ویکری از تألیف این کتاب، رشد سریع انتشارات و بررسی روش‌های نمایه‌سازی در کتابداری بود.

(لنکستر[4]، ۱۹۷۹) بازیابی اطلاعات را عبارت از فرآیند جست‌وجو در میان مجموعه‌ای از مدارک با هدف تعیین آن دسته از مدارک که در حیطه موضوعی درخواست شده، باشند می‌داند.

(ساتکلیف[5]، ۱۹۹۲) مشخصه‌های اصلی نظام‌های بازیابی اطلاعات را به شرح زیر برمی‌شمارد:

۱. مجموعه‌ای از مدارک

۲. نوع سازماندهی و ساختار اطلاعات (شکل منطقی و فیزیکی اطلاعات ذخیره شده)

۳. روش بازیابی

۴. فرآیند بازیابی

۵. عملکرد بازیابی

به طور کلی نظام‌های بازیابی اطلاعات را می‌توانیم این‌گونه تعریف کنیم:

نظام‌هایی که به منظور پردازش و بازیابی داده‌های ساختار نیافته طراحی شده و به لحاظ نوع سازماندهی اطلاعات، ساختار پایگاه‌ها، روش، بازیابی، گروههای خدمت گیرنده و فرآیندی که در طی آن به درخواست‌های اطلاعاتی کاربران پاسخ می‌دهند، از دیگر نظام‌های موجود متمایز هستند.

مجموعهٔ داده‌ها

ارائهٔ یک شمای کلی از نظام‌های بازیابی اطلاعات به صورت‌های مختلف امکان‌پذیر است.

---

1. Information Retrieval System
2. Mooers
3. Vickery
4. Lancaster
5. Sutcliffe

یک راه آن است که این نظام‌ها را به عنوان مجموعه‌ای از رکوردها و فیلدهای اطلاعاتی در نظر بگیریم، بدین ترتیب که هر بانک اطلاعاتی دارای ساختار خاصی، متشکل از فیلدهای مختلف است و هر بانک اطلاعاتی می‌تواند بین یک تا N رکورد اطلاعاتی را در خود جای دهد است.

گردآوری و سازماندهی اطلاعات هر بانک اطلاعاتی براساس خط مشی تولید کننده بانک مربوط صورت می‌گیرد، بنابراین هر بانک اطلاعاتی دارای مشخصاتی مخصوص به خود است (به عنوان نمونه حوزه یا حوزه‌های موضوعی هر بانک). عناصر اطلاعاتی بانک‌های اطلاعاتی متفاوت بوده و شکل استانداری ندارد (مهراد و مقدسی، ۱۳۸۰). اطلاعات موجود در هر بانک مورد تجزیه و تحلیل قرار گرفته و به کوچکترین واحدهای اطلاعاتی معنادار تجزیه می‌شوند، آنگاه روابط لازم بین این واحدهای اطلاعاتی برقرار می‌شود(از جمله مشخص کردن روابط اعم و اخص، گروه‌بندی واحدهای اطلاعاتی و) و در نهایت واحدهای اطلاعاتی بدست آمده به همراه روابط میان آنها در یک فایل جدید ذخیره و نگهداری می‌شوند.

چرخهٔ دستورات

مطالعات انجام شده در زمینهٔ جست‌وجوی اطلاعات نشان دادند که، جست‌وجوگران بانک‌های متنی و کتابشناختی از یک چرخهٔ دستورات به صورت مشترک استفاده می‌کنند. در این مطالعات تأکید بر روی دستورات پراستفاده‌تر بود تا اینکه بخواهند سلسله مراتب استفاده از این دستورات را مورد توجه قرار دهند. این مطالعات نشان دادند که برمبنای دستورات انتخابی کاربران می‌توان پیش‌بینی کرد که جست‌وجوگران در چه مرحله‌ای از جست‌وجو قرار دارند و هدف آنها چیست؟ (پنای من[1]، ۱۹۷۸؛ چاپمن[2]، ۱۹۸۱).

هر دستور ممکن است با شاخص‌های مختلفی فراخوانی شود، برای مثال ترکیب فرمول جست‌وجو با روشی که برای بازبینی اطلاعات انتخاب می‌شود، از یک کاربر تا یک کاربر دیگر متفاوت است. در یک بررسی دیگر که توسط (چاپمن، ۱۹۸۱) صورت گرفت، وی توانست با گروه‌بندی دستورات، ترتیب استفاده از آنها و شاخص‌های استفاده شده به دسته‌بندی کاربران اطلاعاتی بپردازند.در ادامه به بررسی دستورات طبقه‌بندی شده و مورد استفاده در این چرخه می‌پردازیم:

۱. انتخاب یک بانک اطلاعاتی برای انجام جست‌وجو
۲. جست‌وجو برای واژه‌های مورد نظر در بانک واژگان
۳. ایجاد فرمول جست‌وجو و انجام جست‌وجو

---

1. Penniman 2. Chapman

۴. نمایش و بازبینی رکوردها
۵. سفارش مدارک
۶. درخواست برای اطلاعاتی دربارهٔ نظام بازیابی اطلاعات
۷. برقراری شاخص‌های نمایش و ارتباطی

ارائهٔ این الگوها و تکرار آنها خصوصاً برای اهدافی همانند طراحی نظام‌های بازیابی اطلاعات و تعلیم جست‌وجوگران بسیار حائز اهمیت است و براساس اطلاعات باز خوردهای حاصل از تکرار این الگوها می‌توان به طراحی نظام‌های بازیابی اطلاعات بهینه پرداخت.

انتخاب یک بانک اطلاعاتی برای انجام جست‌وجو

بعد از انتخاب بانک اطلاعاتی توسط کاربر، خلاصه‌ای از اطلاعات، همانند محدوده تاریخی رکوردها، تعداد رکوردها، قیمت و... را در اختیار کاربر قرار می‌دهد.

جست‌وجو برای واژه‌های مورد نظر در بانک واژگان

در نظام‌های بازیابی اطلاعات بعد از سازماندهی و تجزیه و تحلیل اطلاعات براساس داده‌های موجود در بانک‌های اطلاعاتی نظام، تعدادی فایل کمکی به منظور کمک به عملیات بازیابی اطلاعات ایجاد می‌شود. این فایل‌ها را می‌توان به عنوان سیاهه‌ای از کل واژه‌های موجود در بانک اطلاعاتی دانست که با نظم الفبایی در کنار هم قرار گرفته‌اند. این فایل‌ها از یک نظام به نظام دیگر می‌تواند ساختار بسیار ساده یا پیچیده‌ای داشته باشد. در شکل ساده، این فایل‌ها شامل ریشه واژه‌ها، پسوندها، پیشوندها، و تعداد تکرار آنها و در شکل پیچیده علاوه بر شکل ساده حاوی ارجاعات، سلسله مراتب‌ها، عبارات موجود در بانک اطلاعاتی و غیره است.

این بخش می‌تواند از اهمیت زیادی برخوردار باشد زیرا اطلاعاتی که بعداً توسط کاربر مورد بازیابی قرار می‌گیرد، بالقوه وابسته به اطلاعات صحیحی است که در این قسمت انتخاب می‌شود. حداقل استفاده‌ای که از بانک واژگان می‌شود این است که کاربر خواهد فهمید که آیا واژه‌های مورد نظر او در بانک اطلاعاتی وجود دارند؟ شکل صحیح آنها به چه صورت است؟ واژه در چند رکورد، در کدام فیلدها و به چه میزان تکرار شده است؟ اگر واژه درخواستی کاربر بد تایپ شده باشد یا در بانک واژگان وجود نداشته باشد، یک مقدار صفر برای آن نمایش داده می‌شود. کاربر با استفاده از این فایل‌ها می‌تواند به واژه‌های مرتبط با واژه‌های مورد نظر خود دست یابد و به اخص‌تر یا اعم‌تر کردن فرمول جست‌وجو بپردازد. یک نمونه بسیار ساده از این بانک‌ها در جدول ۱ قابل مشاهده می‌باشد.

جدول ۱. بانک واژگان براساس موجودی نظام بازیابی اطلاعات

واژه	تعداد تکرار
کتاب	۱۲۳۴
کتابت	۹۹۹
کتابخوانی	۶۵۷
کتابخانه	۳۰۰۱
کتابخانه‌ها	۱۰۰
کتابخانه‌های	۲۵۳۷
کتابدار	۱۲۰۰
کتابداری	۲۰۰
کتابشناختی	۱۲۹
کتابشناسی	۱۹۰۰

فرمول جست‌وجو

کاربر به وسیله یک عبارت یا مجموعه‌ای از عبارات، درخواست اطلاعاتی خود را به نظام اعلام می‌کند که به آن فرمول جست‌وجو یا به بیان دیگر پرسش گفته می‌شود. پرسش‌ها برحسب قابلیت نظام‌های بازیابی اطلاعات می‌توانند دارای گزینه‌هایی برای تعیین شکل، قالب خروجی و نتایج بازیابی نیز باشند. ارائهٔ پرسش‌ها به نظام‌های بازیابی اطلاعات عمدتاً به دو شیوه صورت می‌گیرد: زبان طبیعی و زبان رایانه. زبان طبیعی علاوه بر آنکه نیاز به یادگیری ندارد به لحاظ ماهیت از قابلیت بهتری برای بیان درخواست اطلاعاتی برخوردار است، اما باید گفت که تجزیه و تحلیل زبان طبیعی برای نظام مشکل‌تر بوده و نیاز به توانایی منطقی بالائی دارد. از سوی دیگر درخواست اطلاعات به زبان رایانه نیاز به یادگیری، کسب مهارت و تجربه دارد، اما تجزیه و تحلیل اطلاعاتی که با این شیوه طرح می‌شوند برای نظام بسیار آسان است.

ترکیب فرمول جست‌وجو با استفاده از عملگرهای موجود

بعد از آنکه واژه‌های مورد نظر برای بیان درخواست اطلاعات تعیین شد، با استفاده از امکانات، ابزار، و روشی که نظام در اختیار ما قرار می‌دهد، به گونه‌ای این واژه‌ها را در یک فرمول جست‌وجو با هم ترکیب می‌کنیم تا نتیجهٔ دلخواه از جست‌وجو حاصل گردد. در ادامه به توضیح این قابلیت‌ها، ابزارها و روش‌ها می‌پردازیم که شامل عملگرهای منطق بولی، عملگرهای مجاورت، جست‌وجوی املائی، کوته‌سازی، تعیین محدوده در داده‌های کمی، گسترش دامنهٔ مفاهیم و اصطلاحات، ضریب وزنی، سیاهه واژه‌ها، مجموعه بولی نمره‌گذاری شده، و اندازه‌گیری مشابهت‌هاست.

الف. عملگرهای بولی[1]

با استفاده از این عملگرها(AND,OR,NOT,XOR) کاربر می‌تواند میان تصورات و واژه‌هایی که برای بیان درخواست اطلاعاتی خود برگزیده است ارتباطی منطقی برقرار کند (داورپناه، ۱۳۷۶). عملگر منطقی نمادی است که برای مشخص کردن رابطهٔ منطقی بین دو مقدار یا مفهوم مورد استفاده قرار می‌گیرد. با استفاده از عملگر منطقی OR می‌توان محدودهٔ رکوردهای مورد نظر را افزایش و با استفاده از عملگر منطقی AND,NOT نیز محدوده آن را کاهش داد. در جدول ۲ این عملگرها مشاهده می‌شود.

جدول ۲. عملگرهای منطق بولی در نظام‌های بازیابی اطلاعات

عملگر	عملکرد	مثال	تشریح مثال
AND	در صورتی که به عنوان عملگر در کنار یک واحد اطلاعاتی ظاهر گردد، یک مجموعه فقط در صورتی که حاوی آن اطلاعات باشد در نتایج بازیابی ظاهر خواهد شد (اشتراک).	موضوع = "رایانه" و تاریخ = ۱۳۷۸	رکوردهای اطلاعاتی مورد بازیابی قرار می‌گیرد که فیلد موضوع آنها "رایانه" و همین‌طور فیلد تاریخ آنها برابر ۱۳۷۸ باشد.
OR	در صورتی که به همراه تعدادی از واحدهای اطلاعاتی به کار برود، در صورت وجود این واحدها در رکوردهای اطلاعاتی، آن رکوردها مورد بازیابی قرار خواهند گرفت (اجتماع).	موضوع = "کتابخانه" یا موضوع = "اطلاع‌رسانی"	رکوردهای اطلاعاتی مورد بازیابی قرار می‌گیرد که فیلد موضوع آنها "کتابخانه" یا "اطلاع‌رسانی" باشد.
NOT	در صورتی که به همراه یک تعداد واحد اطلاعاتی به کار برود، هر یک از واحدها در رکوردهای اطلاعاتی مشاهده شوند رکوردهای مربوط مورد بازیابی قرار نخواهد گرفت.	موضوع = "اطلاع‌رسانی" و "تاریخ نامساوی ۱۳۷۸	رکوردهای اطلاعاتی که موضوع آنها "اطلاع‌رسانی" باشد و تاریخ آنها برابر هر سالی غیر از سال ۱۳۷۸ باشد.
XOR	به دلیل درک دشوار آن توسط کاربران کمتر مورد استفاده قرار می‌گیرد ولی اگر به همراه دو واحد اطلاعاتی به کار برود به این مفهوم می‌باشد که هر یک از دو واحد اطلاعاتی که به صورت منفرد مورد مشاهده قرار گیرند مورد بازیابی قرار می‌گیرد ولی اگر با هم در یک مجموعه مشاهده شوند آن مجموعه مورد بازیابی قرار نمی‌گیرد.	موضوع = "رایانه" XOR موضوع = "اطلاع‌رسانی"	رکوردهای اطلاعاتی که موضوع آنها "رایانه" یا "اطلاع‌رسانی" باشد ولی نباید با هم در یک رکورد تکرار شده باشند.

ب. عملگرهای مجاورت[2]

واحدهای اطلاعاتی در موقعیت‌های مکانی مشخصی از رکوردهای اطلاعاتی قرار دارند. فواصل این واحدهای اطلاعاتی از یکدیگر می‌تواند نشان دهنده ارتباط یا عدم ارتباط معنائی

1. Boolean Operators 2. Adjacant Operators

این واحدها با یکدیگر باشد. برای تشخیص این ارتباط معنایی و بالا بردن ضریب دقت در بازیابی اطلاعات می‌توان از عملگرهای مجاورت استفاده کرد. به بیان دیگر با ترکیب عملگرهای منطق بولی و عملگرهای مجاورت می‌توانیم رکوردهای اطلاعاتی را مورد بازیابی قرار دهیم که حاوی واحدهای اطلاعاتی در فواصل مکانی معینی نسبت به هم باشند. به عنوان نمونه اگر دو واحد اطلاعاتی "کتابخانه" و "ساختمان" در یک بند اطلاعاتی[1] در فواصل نزدیک به هم قرار بگیرند احتمال اینکه در رابطه با "ساختمان کتابخانه" بحث شده باشد خیلی بالا است. شکل اصلی فرمول مجاورت به این صورت است:

واحد اطلاعاتی اول در M فاصله بند اطلاعاتی از واحد اطلاعاتی دوم قرار گرفته باشد.

مثال ۱: "ساختمان" در یک پاراگراف با "کتابخانه" باشند (ساختمان و کتابخانه باید در یک پاراگراف قرار داشته باشند)

مثال ۲: "ساختمان" در یک جمله با "کتابخانه" باشند (ساختمان و کتابخانه باید در یک جمله قرار داشته باشند)

یک نمونه دیگر از عملگرهای مجاورت، "ADJ" نامیده می‌شود و در صورتی که بین دو یا چند واحد اطلاعاتی به کار رود به شرطی که این واحدها با رعایت تقدم و بلافاصله بعد از هم قرار بگیرند، مورد بازیابی قرار خواهند گرفت، به عنوان مثال فرمول "استاندارد" ADJ "کتابخانه" رکوردهای اطلاعاتی را مورد بازیابی قرار خواهد داد که عبارت "استاندارد کتابخانه" را در خود جای داد باشند نه "کتابخانه استاندارد". نمونه‌ای از عملگرهای مجاورت در جدول ۳ مشاهده می‌شود.

جدول ۳. مهم‌ترین نمونه‌های عملگرهای مجاورت

عملگر	توصیف
W	با رعایت تقدم بین دو واژه تعداد صفر تا N واژه قرار خواهد گرفت.
N	بدون رعایت تقدم بین دو واژه تعداد صفر تا N واژه قرار خواهد گرفت.
P	دو واژه با هم در یک پاراگراف تکرار شده باشند.
S	دو واژه با هم در یک جمله یا یک فیلد فرعی تکرار شده باشند.
F	دو واژه با هم در یک فیلد قرار داشته باشد، که می‌توان نام فیلد را نیز به فرمول جست‌وجو اضافه کرد.

ج. جست‌وجوی املائی (هجایی)[2]

اشتباهات تایپی حاصل از غلط‌های املائی واردکنندگان اطلاعات، کاربران نظام و همچنین

۱. یک بند اطلاعاتی می‌تواند از یک واژه تا چند واژه، یک جمله، یک پاراگراف، یک فیلد یا یک فیلد فرعی تشکیل شده باشد.

2. Spelling Search

استفاده از روش‌های جدید برای درون داد اطلاعات همانند استفاده از پویشگر[1] و نرم‌افزار تشخیص نوری حروف، سبب شده تا از شیوهٔ جست‌وجوی املائی برای کاستن تأثیر این خطاها در بازیابی اطلاعات استفاده شود[2].

این روش هم در جست‌وجوی واژه‌ها و هم در تجزیه و تحلیل ابتدایی رکوردهای اطلاعاتی به منظور ایجاد فایل‌های کمکی در نظام کاربرد دارد. به عنوان مثال کاربر واژه COMPUTER را برای جست‌وجو وارد می‌کند و آنگاه واژه‌های CONPUTER, COMMPITER, COMPUTTER, COMPUTER نیز مورد جست‌وجو قرار خواهند گرفت. استفاده از این روش باعث افزایش ضریب بازیافت و کاهش ضریب دقت می‌شود.

د. کوته‌سازی[3]

ریشه هر واژه براساس پیشوند یا پسوندی که به آن اضافه می‌شود، اشکال مختلفی به خود می‌گیرد. بنابراین ریشه با ملحقاتی که به آن اضافه می‌شود به شکل‌های متفاوتی نوشته می‌شود. از سوی دیگر بعضی واحدهای اطلاعاتی با وجود یکسان بودن به صورت‌های مختلفی نوشته می‌شوند(همانند اسم پدید آورندگان). اهمیت کوته‌سازی بخصوص در نظام‌هایی که ریشه‌یابی واژه‌ها (در مراحله تجزیه و تحلیل داده‌ها) و یکدست سازی اطلاعات بخوبی صورت نمی‌گیرد بسیار پراهمیت است.

کوته‌سازی این امکان را فراهم می‌آورد که هنگام جست‌وجو بدون در نظر گرفتن حروف یا حرفهایی از یک واژه یا یک واحد اطلاعاتی به بازیابی بپردازیم. به این دلیل که هنگام تجزیه و تحلیل داده‌ها بعضی حروف و کاراکترهای خاص حذف می‌شوند. برای نمایش شکل کوته سازی در دستورات از این کاراکترها استفاده می‌شود(به عنوان نمونه#,?,*) دو الگوی رایج کوته سازی در جدول ۴ قابل مشاهده می‌باشند. کوته‌سازی باعث افزایش ضریب بازیافت و کاهش ضریب دقت می‌شود.

---

1. Scanner

۲. یکی از شیوه‌های درون داد اطلاعات در مراکز کتابخانه‌ای و اطلاع‌رسانی استفاده از اسکنر و نرم‌افزار تشخیص نوری حروف است. در این شیوه یک صفحه چاپی بدون نیاز به تایپ به نظام وارد می‌شود به نحوی که قابل ویرایش از نظر واژه‌پردازی باشد. در یک صفحه چاپی باکیفیت و وضوح خوب می‌توان بین ۹۰ تا ۹۹درصد بر روی صحت داده درون داد شده از جهت واژه‌پردازی حساب کرد.

3. Truncation

جدول ۴. نمونه‌هایی از عملگرهای کوته‌سازی

شرح	نمونه	نتایج حاصل از اجرای نمونه
کوته‌سازی برابر یک یا چند حرف با تعداد مشخص در مکان‌های مشخصی از واژه یا واحد اطلاعاتی	کتابخانه ؟؟	کتابخانه‌ای کتابخانه‌ها کتابخانه
کوته‌سازی برای یک مجموعه از حروف با تعداد نامشخص از واژه یا واحد اطلاعاتی	کتاب *	کتابخانه کتابخوانی کتابدار کتابشناسی کتابشناختی
کوته‌سازی به منظور نادیده گرفتن وجود یا عدم وجود یک یا چند حرف در واژه یا واحد اطلاعاتی	LABO#R	LABOUR LABOR

ه. تعیین محدوده در داده‌های کمی

داده‌های کمی می‌توانند بخشی از اطلاعات موجود در بانک‌های اطلاعاتی را تشکیل دهند.از نمونه‌های بارز این گونه داده‌ها می‌توانیم از مقادیر عددی و تاریخی نام ببریم. بنابر اهمیت این نوع داده‌ها نیاز خواهد بود که علاوه بر یک مورد خاص، در محدوده‌های خاصی نیز جست‌وجو صورت گیرد. عملگرهای مورد استفاده در تعیین محدوده‌ها عبارتند از: کوچکترین مساوی (=<)، بزرگتر مساوی (=>)، بزرگتر(>)، مساوی (=)، و نامساوی (=!). استفاده از این عملگرها باعث افزایش ضریب دقت و کاهش ضریب بازیافت می‌شوند.

و. گسترش مفاهیم و اصطلاحات

نظام‌های بازیابی اطلاعات با استفاده از ابزارها و شیوه‌های مختلف به گسترش مفاهیم موجود در درخواست اطلاعاتی کاربر می‌پردازند و برای این منظور با استفاده از اصطلاحنامه‌ها و واژه‌نامه‌های موجود یا ایجاد آنها بر پایه اطلاعات موجود در نظام می‌پردازند. ایجاد این اصطلاحنامه‌ها و واژه‌نامه‌ها براساس داده‌های موجود در بانک اطلاعاتی مبحث گسترده‌ای است که در مقاله حاضر به بحث دربارهٔ آن نمی‌پردازیم.

ز. کاربرد ضریب وزنی [1]

اگر بخواهیم هر عضو فرمول جست‌وجو در بازیابی از دقت بالاتری برخوردار باشد، می‌توانیم از روش ضریب وزنی استفاده کنیم. منظور از ضریب وزنی این است که کاربر به هر

1. Weight Search

عضو دل‌خواه از فرمول جست‌وجو یک عدد صحیح که مشخص کننده اهمیت عضو در میان اعضا است اختصاص می‌دهد. برای مثال به پرسش زیر توجه کنید:

(TENNIS(.8)OR GOLF (.4)AND CHAMPION(.6)

در چنین حالتی، نظام بازیابی اطلاعات بدون در نظر گرفتن ضرایب وزنی تعیین شده به جست‌وجوی رکوردها می‌پردازد. سپس با استفاده از ضرایب وزنی تعیین شده توسط کاربر به نمره‌گذاری رکوردها می‌پردازد. جدول ۵. بیانگر موضوع است.

جدول ۵. محاسبه ضرایب وزنی

محتوای رکورد	نمره‌ای که به رکورد اختصاص داده می‌شود
اگر یک رکورد حاوی TENNIS و CHAMPION باشد	.8+.6=1.4
اگر یک رکورد حاوی GOLF و CHAMPION باشد	.4+.6=1.0
اگر یک رکورد حاوی هر سه واژه باشد	.8+.4+.6=1.8

به ازای هر بار تکرار هر یک از واژه‌های عضو بیش از یک بار در رکوردهای اطلاعاتی، ضریب وزنی واژه مربوط، به کل نمره رکورد بازیابی شده اضافه می‌گردد. بعد از انجام جست‌وجو و انجام نمره‌گذاری توسط نظام، رکوردها برحسب بالاترین نمرات مرتب می‌شوند و کاربر می‌تواند تا N رکورد بازیابی شده را که دارای نمره بالاتری هستند، انتخاب کند.

با وجود مزایای این روش کاربران باید یاد بگیرند که چگونه باید ضرایب وزنی را به عضوهای فرمول اختصاص دهند و در کنار آن باید متذکر شد که نتایج بدون ضریب وزنی از ارزش بالایی برخوردار نخواهند بود.

ح. سیاهه واژه‌ها

برای آنکه یک روش آسان‌تر نسبت به روش تعیین ضرایب وزنی برای کاربران تدارک ببینیم می‌توانیم از روش فرمول نویسی "سیاهه واژه‌ها" استفاده کنیم. این شیوه همان کار ضرایب وزنی را با یک روش دیگر انجام می‌دهد. به فرمول زیر توجه کنید:

TENNIS,GOLF,CHAMPION,CUT,TOURNAMENT,WINNER

در سیاهه فوق بین هر جفت واژه عملگر OR فرض می‌شود و هر کدام از آنها دارای ضرایب وزنی یکسان هستند. معادل این فرمول با استفاده از عملگرهای بولی به صورت زیر خواهد بود:

TENNIS.OR.GOLF.OR.CHAMPION.OR.CUT.OR.TOURNAMENT.OR.WINNER

در نهایت، رکوردها براساس بالاترین تعداد واژه‌های فوق که در رکوردها ظاهر می‌شوند و همچنین براساس تعداد تکرار آنها در رکوردها نمره گذاری می‌شوند و براساس همین نمره گذاری

مرتب می‌شوند. در این شیوه از عملگر AND نمی‌توان استفاده کرد و به همین دلیل ضریب دقت پایین است. در این روش می‌توان برای هر واژه یک ضریب وزنی نیز مشخص کرد.

ت. مجموعه‌های بولی نمره‌گذاری شده

همانند دو روش قبل در این روش نیز براساس تعداد تکرار اعضای فرمول جست‌وجو در رکوردهای بازیابی شده، به هر کدام از رکوردها یک نمره اختصاص می‌یابد و در نهایت رکوردها بر حسب همین نمره در خروجی مرتب می‌شوند.این نمره نشان دهنده درصد ارتباط رکورد بازیابی شده حاضر با در خواست اطلاعاتی کاربر است. به عنوان مثال بعد از اجرای فرمول (Information And Retrieval And System) در سایت Infoseek قسمتی از نتایج جست‌وجو به صورت زیر نمره‌گذاری شده و بازگردانده شد. این نمره‌گذاری در جلو عبارت Relevance به صورت درصد بیان شده است.

نتایج جست‌وجو و نمره‌گذاری شده سایت Infoseek

1. Food Safety Information from NC State

<http://www.ces.ncsu.edu/depts/foodsci/aqentifo/>

Food Safety resources for consumers, educators, researchers and absolutelyanlone that wants to know more aboutfood satety prevention and issues. Developed by The NC Cooperative Extension Service and...

Relevance: 100% Date: 11 May 1999, Size 9.Ok,

http: /www.ces.ncsu.edu/depts/foodsci/agentinfo/ find similar paqes

2. CIIR Demo Paqe <http://ciir.cu.umass,edu/>

Sorry, this site is geared for browsers that support frames.

Relevance: 98% Date: 28 Sep 1999, Size 1.9K, http://ciir.cs.umass.edu/find similar paqes

3. California Highway conditions <http://www.amdahl.com/internet/qeneral/travel/ ca_hiqhway. html>

Access highway condition information byregion of highway type.

Relevance: 85% Date: 8 Oct 1999, Size 25.gk,

http://www.amdahl.com/internet/general/travel/ca_ hiqhway.html find similar paqes

4. CAIRSS<http://imr.utsa.edu/CAIRSS.html>

Institute for Music Research ACIRSS for Music (Computer_ Assisted Information Retrieval Service System) Search CAIRSS CAIRSS is a bibliographic database of music research literature in music education, music psychology, music...

Relevance: 81% Date: 28 Aug 1998, Size 3.7K, http://imr.utsa.edu/CAIRSS.html Find similar pages.

ی. اندازه‌گیری مشابهت

دسترسی آسان کاربر به حداقل یک رکورد که بالاترین میزان ارتباط را با درخواست اطلاعاتی او داشته باشد بسیار حائز اهمیت است. اولین رکوردی که به این صورت توسط نظام مورد بازیابی قرار گیرد می‌تواند اساسی برای جست‌وجوهای بعدی کاربر باشد. اگر کاربر چنین رکوردها را بازیابی، و به جهت مرتبط بودن با درخواست اطلاعاتیش آن را تائید کرد، آنگاه نظام به تجزیه و تحلیل رکورد می‌پردازد و با معیارهای ارزشیابی متن و دادن ضریب وزنی به هر یک از واژه‌ها اقدام به انجام یک جست‌وجوی دیگر می‌کند و رکوردهای مرتبط را مورد بازیابی قرار می‌دهد. در این روش کاربر نیاز نخواهد داشت که به بازنویسی دوباره فرمول بپردازد.

انجام جست‌وجو

پس از آنکه درخواست اطلاعاتی توسط کاربر به نظام وارد شد، نظام درخواست اطلاعاتی را مورد تجزیه و تحلیل قرار داده و آنگاه با انجام عملیات جست‌وجو مجموعه‌هایی را که با آن منطبق هستند، به صورت خروجی به کاربر نشان می‌دهد که آن را نتیجه جست‌وجو می‌نامند. منظور از تجزیه و تحلیل، ترجمه و تفسیر درخواست اطلاعاتی مطابق با زبان رایانه است. با بررسی نتیجهٔ جست‌وجو کاربر در می‌یابد که آیا جست‌وجو رضایت‌بخش بوده است یا باید از فرمول جدیدی استفاده کند، و یا اینکه باید بعد از انجام تغییرات لازم دوباره فرمول جست‌وجو به اجرا بگذارد. در مجموعهٔ رکوردهای اطلاعاتی بازیابی شده رکوردهایی وجود دارند که با درخواست اطلاعاتی کاربر منطبق نیستند که آنها را ریزش کاذب اطلاعاتی می‌نامند. ریزش کاذب اطلاعات می‌تواند از عوامل مختلفی ناشی شود، از جمله عدم سازماندهی بهینه اطلاعات، محدودیت‌های زبان نرم‌افزاری برای بیان پرسش، بازخورد ضعیف نظام در تعامل با کاربر، تشریح بد یاناقص درخواست اطلاعاتی.

نمایش و بازبینی رکوردها

باز خورد خوب نظام در این قسمت می‌تواند نقش مهمی در هدایت کاربر برای رسیدن به

اطلاعات مورد نظرش داشته باشد. در این قسمت کاربر باید بتواند به سئوالات زیر پاسخ دهد:

چه کارهایی با نیاز اطلاعاتی کاربر مطابقت دارند؟

آیا اطلاعاتی وجود دارد که بتوان واژه‌هایی که حوزه‌های موضوعی وسیعی را پوشش داده‌اند تشخیص داد؟

آیا اطلاعاتی وجود دارد تا بتوان واژه‌هایی که می‌تواند موجب اخص‌تر شدن دامنهٔ جست‌وجو شوند را تشخیص داد؟

آیا مجموعهٔ رکوردهای بازیابی شده با نیازهای اطلاعاتی کاربر مطابقت دارد؟

آیا اطلاعاتی وجود دارد که از طریق آن کاربر بتواند به تصحیح منطق بولی فرمول جست‌وجو بپردازد؟

سفارش مدارک

در بعضی از نظام‌ها فقط اطلاعات کتابشناختی نگهداری می‌شود، در اینگونه نظام‌ها میان زمان جست‌وجو و دریافت اطلاعات مدت زمانی تلف خواهد شد. اما در نظام‌های تمام متن کاربر قادر خواهد بود، مستقیماً مدارک را مورد مطالعه قرار داده و در مورد انتخاب یا عدم انتخاب آنها تصمیم‌گیری کند، در این نظام‌ها کاربر مستقیماً مدارک را دریافت می‌کند و نیازی به سفارش آنها نیست.

درخواست برای اطلاعاتی درباره نظام بازیابی اطلاعات

بیشتر نظام‌های بازیابی اطلاعات از راهنمای پیوسته برخوردارند. و اگر کاربر در مورد دستورات یا دیگر ترکیبات نظام نیاز به راهنمایی داشت، اطلاعات لازم در اختیار او قرار می‌گیرد. نظام‌های بازیابی اطلاعات حرفه‌ای همیشه سعی دارند به سئوالات کاربران نهایی پاسخ دهند، همانند (در اینجا چه کاری در حال انجام است؟ در اینجا چگونه باید عمل کنم؟) برای پاسخ دادن به بعضی از این سئوالات نظام باید از اهداف کاربر آگاه باشد و یک چارچوب برای ارزیابی کاربر در اختیار داشته باشد. پاسخ دادن به بسیاری از سئوالات دیگر به راحتی و با تفسیر دستورات یا گزینه‌های انتخاب شده توسط کاربر صورت می‌گیرد. کاربرانی که از تجربه کمتری برخوردار هستند نیاز به راهنمایی بیشتری در این زمینه خواهند داشت؟

برقراری شاخص‌های نمایشی و ارتباطی

شاید یک از علل مهم ناموفق بودن جست‌وجو این باشد که کاربران نمی‌توانند با خدمات

اطلاعاتی میزبان ارتباط برقرار کنند، زیرا کلمات عبور را اشتباه وارد می‌کنند یا شماره تلفن مربوط به شبکه ارتباطی را درست وارد نمی‌کنند. کاربران در این گونه مواقع نمی‌دانند که چگونه باید از صحت کار خود مطمئن شوند. در این گونه مواقع هنوز ارتباط با نظام بازیابی اطلاعات برقرار نشده است که نظام بتواند از خود عکس‌العمل نشان دهد.

نتیجه‌گیری

در مجموعه چرخهٔ دستورات باید چهار مقوله زیر را به عنوان مهم‌ترین اهداف خود مدنظر قرار دهیم:

- بهینه سازی انتخاب واژگان جست‌وجو، توسط کاربر؛
- بهینه سازی فرمول جست‌وجوی کاربر؛
- بهینه سازی و تنظیم تعداد رکوردهای بازیابی شده؛
- بهینه سازی ضریب دقت و بازیابی یا بهینه سازی کل نظام بازیابی اطلاعات

بعد از انجام این مراحل در صورت رضایت جست‌وجوگر، جست‌وجو به مرحله بعد منتقل خواهد شد در غیر اینصورت جست‌وجو دوباره تکرار خواهد شد. معنی تکرار انجام دوباره یک کار است. در جست‌وجو تکرار به معنی تکرار پردازش فرمول جست‌وجو و اصلاح آن تا رسیدن به نتیجه دل‌خواه است. هدف ممکن است رسیدن به یک اندازه دل‌خواه یا اطلاعاتی خاص یا رکوردهای خاصی باشد، که دارای بالاترین ارزش اطلاعاتی هستند. اطلاعات به‌دست آمده امکان تصمیم‌گیری در مورد این پرسش که آیا جست‌وجو باید دوباره انجام شود را در اختیار ما قرار می‌دهد، و این قضاوتی است که از باز خورد حاصل از نتایج جست‌وجو بدست می‌آید. در این مورد بعضی عوامل به صورت مرسوم و متعارف چنانچه قبلاً در چرخه دستورات به آن اشاره شد وجود دارند، ولی قالب‌های باز خورد به صورت کامل در جست‌وجو مشخص نشده‌اند.

بعضی از کاربران ترجیح می‌دهند که با توجه به محتوای رکوردهای بازیابی شده، درباره نحوه و چگونگی ایجاد و اجرای دوبارهٔ فرمول جست‌وجو تصمیم‌گیری کنند، نه براساس تعداد رکوردهای بازیابی شده، اما در مواقعی که تعداد رکوردها از حد معینی بالاتر باشد، انجام این بررسی مشکل به نظر می‌رسد. حالت معمول این قضیه به این صورت است که اگر بعد از اجرای فرمول جست‌وجو تعداد رکوردهای زیادی مورد بازیابی قرار نگرفته، یک فرمول دیگر ایجاد و اجرا می‌شود، به عنوان مثال اگر نتیجه یک بازیابی اطلاعات ۳۵۰۰ رکورد باشد، حتی اگر رکوردهای مورد نظر کاربر در میان این رکوردها وجود داشته باشند، بررسی آنها بسیار مشکل است و باید فرمول جست‌وجو اصلاح شود تا تعداد رکوردهای کمتری مورد بازیابی قرار بگیرند.

از سوی دیگر فرض کنید برای ایجاد یک فرمول جست‌وجو تعداد رکوردهای بازیابی شده ۲۳ عدد باشد این اشتباه است اگر از این تعداد خوشحال شده و فکر کنیم جست‌وجوی خوبی انجام شده است یا برعکس فکر کنیم جست‌وجوی خوبی صورت نگرفته است، زیرا هر چند تعداد این رکوردها کم است ولی احتمالاً اطلاعات مفیدی از این تعداد رکورد بدست خواهد آمد. در این حالت تعداد رکوردها به آن اندازه کوچک است که کاربر می‌تواند آن را مورد بازبینی قرار داده و درباره فرمول جست‌وجوی بعدی تصمیم‌گیری کند. بازخورد به اصلاح فرمول جست‌وجو کمک می‌کند، و این خود باعث اجرای دوباره جست‌وجو و بازیابی تعداد زیاد یا کمتری از رکوردها می‌شود. در این هنگام مسئله قابلیت و توانایی‌های به تصویر کشیدن اطلاعات نمایان می‌شود.

جست‌وجوگران با تجربه و متخصصان بازیابی باید به این نکته توجه داشته باشند که انجام بازیابی مستلزم صرف وقت و هزینه است و در این مسیر باید با اصلاح و بازنویسی و تکرار چرخه دستورات به نتایج دلخواه خود دست یابیم.

مآخذ

۱. داورپناه، محمدرضا (۱۳۷۶). "قواعد استنتاج و فرمولبندی پرسش در جست‌وجوی کامپیوتری". **پیام کتابخانه**، دبیرخانه هیأت امنای کتابخانه‌های عمومی کشور، دوره ۷ (شماره ۲).

۲. مهراد، جعفر؛ مقدس، جلیل (۱۳۷۸). "جست‌وجوی رایانه‌ای پایگاه‌های اطلاعاتی علم شیمی". **پیام کتابخانه**، دبیرخانه هیأت امنای کتابخانه‌های عمومی کشور، دوره ۹ (شماره ۳).

3. Chapman, J.L.(1980). "A State Transition Analysis of On- line Information - Seeking Behavior", *Journal of the American Society for Information Science* 32(2): 107-116.

4. Lancaster, F.W.(1979). *Information Retrieval Systems: Characteristics, Testing, and Evaluation.* 2nd ed. New York: Wiley.

5. Penniman, W.D. *A Stochastic Process Analysis of On- line Behavior,* Proceedings of the American Society for Information Science 38th Annual Meeting.

6. Tague - Sutcliffe, J.(1992). "The Pragmatics of Information Retrieval Experimentation, Revisited," *Information Processing & Management,* 28: 467-490.