استخراج خودکار عبارتهای کلیدی از متون مقاله‌های فارسی

علی گزنی[۱]

چکیده

در پژوهش حاضر، عبارتهای کلیدی از متون مقاله‌های فارسی به صورت خودکار جداسازی گردیده است. استخراج عبارتها مبتنی بر روشهای آماری، نحوهٔ توزیع واژگان، مجاورت و ... صورت پذیرفته است. سیستمی که بر پایه پژوهش حاضر طراحی گردیده، با توجه به بازخوردهای کاربر از قابلیت یادگیری برخوردار است، با توجه به بازخوردهای کاربر از قابلیت یادگیری برخوردار است، به گونه‌ای که در طول زمان مرتباً به کارایی آن افزوده می‌شود. استخراج عبارتهای کلیدی می‌تواند در پهنه گسترده‌ای از مسائل از جمله در طراحی سیستمهای بازیابی اطلاعات، کاربر داشته باشد.

کلیدواژه‌ها: کلیدواژه، عبارتهای کلیدی، مقاله‌های فنی

هدف از استخراج عبارتهای کلیدی[۲] متون مقاله‌ها[۳]، تسریع در تعیین حوزه موضوعی مقاله‌های چاپ شده می‌باشد. این کار فکری است و به آشنایی کلی با موضوع مورد نظر، مهارت و تجربه نیاز دارد. بنابراین، انرژی‌ای که می‌تواند در راههای دیگری صرف گردد، باید در راه تسهیل دسترسی دیگران به اطلاعات مصرف گردد. افزایش

---

۱. عضو هیئت علمی کتابخانه منطقه‌ای علوم و تکنولوژی شیراز

۲. Key Phrase

۳. در رشته‌های فنی (علوم و تکنولوژی) نویسندگان اغلب مجبورند از مجموعه واژگان خاصی برای گسترش مقاله خود استفاده کنند. استخراج عبارتهای کلیدی با استفاده از روش حاضر برای این‌گونه مقاله‌ها، پاسخ بهتری را در بر خواهد داشت.

روزافزون تعداد مقاله‌های فنی و حجم انبوه اطلاعات از جمله مواردی هستند که بر اهمیت این مسئله می‌افزایند. به علاوه، استخراج عبارتهای کلیدی در اغلب اوقات تحت تأثیر پیش‌زمینه‌های قبلی، عقاید شخصی یا تمایل فرد به انجام سریعتر کار قرار می‌گیرد. بنابراین، کیفیت کار در بین افراد مختلف و حتی افراد یکسان در زمانهای مختلف، متفاوت است (مهراد، ۱۳۷۳).

مقاله حاضر به تشریح پژوهش انجام شده در زمینه استخراج عبارتهای کلیدی متون مقاله‌های فنی می‌پردازد. در روش مورد نظر، متن مقاله در قالب ماشین‌خوان به برنامه رایانه‌ای طراحی شده تحویل می‌گردد. برنامه در روشی قابل مقایسه با آنچه توسط انسان خوانده می‌شود، اطلاعات را مورد پردازش قرار می‌دهد و از میان کلیه واژگان موجود در مقاله، عبارتهای کلیدی را که نشان‌دهنده مرتبط‌ترین عبارتها و اطلاعات مقاله می‌باشند، استخراج می‌کند. این عبارتها می‌توانند به عنوان راهنمایی برای قضاوت در مورد متن مقاله مورد استفاده قرار گیرند. بنابراین، عبارتهای کلیدی مستقیماً از میان نوشته‌های نویسنده انتخاب می‌گردند.

عبارتهای کلیدی

عبارتهای کلیدی متن نشان‌دهندهٔ مفاهیم و موضوع مقاله بوده، می‌توانند در موارد زیر مورد استفاده قرار گیرند:

۱. استخراج خودکار عبارتهای کلیدی، یک متن بلند را به خلاصه‌ای کوتاه تبدیل می‌کند. به عنوان مثال، می‌توان از این ویژگی در مرورگرهای وب[۱] استفاده کرد؛ بدین ترتیب که کاربر با فشار دادن یک دکمه، عبارتهای کلیدی متن را مشاهده و در نتیجه به حوزهٔ موضوعی متن مورد نظر پی می‌برد. برای مثال، شکل۱ نتیجه یک جستجو از موتور کاوش Google را نشان می‌دهد. پیوند Key Phrases در این صفحه اضافه شده است. با کلیک کردن بر روی این گزینه عبارات کلیدی متن نمایش داده می‌شوند.

---

۱. Web Browser

Information Technology Association of America Website
Information Technology Association of America Headlines, … April ۲۴, ۲۰۰۳-Organizing for Results: Information Technology Structures and Staffing WEBCAST …
Description: Trade association representing the broad spectrum of the world-leading US IT industry.
That's why …
Category: Business> Information Technology > Associations
www.itaa.org/-۵۵k-Cached-Similar Pages Key Phrases

شکل ۱. عبارتهای کلیدی در نمایش جستجو

۲. عبارتهای کلیدی می‌توانند به عنوان قسمتی از نتایج جستجو همراه با سایر مشخصه‌های متن بازیابی شده (همانند عنوان، قسمتهایی از متن، URL و ...) یا به جای آنها نمایش داده شوند. در شکل ۱ می‌توان تصور کرد که به جای پیوند Key Phrases عبارتهای کلیدی همراه با سایر قسمتهای جستجو شده، نمایش داده شوند.

۳. در مواردی که به مشخصه‌هایی بیش از نامگذاری صرف به منظور درک سریعتر متن نیاز داریم، عبارتهای کلیدی می‌توانند مفید باشند. به عنوان مثال، اگر نام یک فایل یا نامه الکترونیکی[۱] به عنوان برچسب[۲] با عبارتهای کلیدی ادغام گردند، حالت بهتری را ایجاد می‌کنند. در این حالت، مشاهده عبارتهای کلیدی همراه با عنوان، به فهم محتوای نامه کمک بیشتری می‌کند.

٤. برجسته‌کردن[۳] عبارتهای کلیدی در متون الکترونیکی می‌تواند به مرور سریع و اجمالی متن کمک کند.

۵. کمک به نویسنده یا ویراستار در تخصیص عبارتهای کلیدی به متن. انجام این کار به صورت خودکار می‌تواند به عنوان یک استاندارد، نوعی یکدستی و مطابقت نوشته با کارکرد سیستم بازیابی اطلاعات و در نتیجه اطلاع‌رسانی صحیح‌تر را به همراه داشته باشد.

---

۱. Email
۲. Label
۳. Highligth

٦. در مواردی که با مشکل پهنای خط یا مطابق با اصول نمایش گرافیکی اطلاعات با محدودیت فضای نمایشی[۱] مواجه هستیم، نمایش عبارتهای کلیدی بسیار مفید است. اصولاً در کشورهای جهان سوم که خطوط از سرعت و پهنای خط پایینی برخوردارند و در مکانهایی که محدودیت فیزیکی وجود دارد، همانند صفحات نمایش رایانه (اندازه ثابت)، حالت مطلوبتری را ایجاد می‌کند.

۷. استخراج خودکار عبارتهای نمایه‌ای متون نشریات و صفحات وب، خواندن و جستجوی اطلاعات نشریات را برای خوانندگان تسهیل می‌کند.

۸. حضور عبارتهای کلیدی در نتایج جستجو می‌تواند به اصلاح و تعریف مجدد فرمول جستجو و حتی تغییر دیدگاه کاربران از ساختار موجود در یک زمینه خاص کمک کند؛ یعنی کاربران می‌توانند با افزودن، حذف واژگان دامنه جستجو را محدودتر کرده، ضریب دقت را بالاتر ببرند. در نتیجه، بالابردن ضریب دقت[۲] یا با گسترده‌ترکردن دامنهٔ جستجو و در نتیجه به بالابردن ضریب بازیابی[۳] کمک می‌کند. بنابراین می‌توان عبارتهای کلیدی را به عنوان جزئی لازم برای سیستمهای بازیابی اطلاعات معرفی کرد.

۹. در مفاهیم سازماندهی اطلاعات در سیستم‌های بازیابی اطلاعات[٤] (۱) می‌توان به گونه‌ای مؤثر از عبارتهای کلیدی در خوشه‌بندی[٥] و طبقه‌بندی مدارک استفاده کرد.

---

۱. تعریف علمی نمایش گرافیکی اطلاعات عبارت است از محاسبه و انتقال علایم به اشکال هندسی به صورتی قابل درک و مشاهده توسط انسان، به منظور فهم و کشف روابط پنهان موجود بین عناصر مختلف داده‌ها (۲). نمایش گرافیکی اطلاعات به عنوان یک شیوه علمی و زیرشاخه‌ای از مبحث تعامل انسان و رایانه و با استفاده از تواناییهای گرافیکی رایانه‌ها اهداف زیر را دنبال می‌کند:

۱) بالابردن سرعت فهم و پردازش اطلاعات توسط انسان در طی فرایند ادراک و کم کردن درگیریهای ذهنی او

۲) ایجاد و برقراری ارتباط بین اجزای مختلف اطلاعات

۳) انجام عملیاتهای پیچیده با اعمالی بسیار ساده

۲. Precision

۳. Recall

٤. هر نظام بازاریابی اطلاعات (نرم‌افزار) دارای یک مبنای خاص برای تجزیه و تحلیل اطلاعات است، که نظام بر اساس آن به تفسیر اطلاعات و مطابقت بین اقلام و درخواستهای اطلاعاتی پرداخته و بدین ترتیب بازیابی اطلاعات صورت می‌گیرد. این تجزیه و تحلیل «ساماندهی اطلاعات» نامیده می‌شود.

۵. دسته‌بندی رکوردهای اطلاعاتی در گروههای مختلف در کل سیستم بازیابی اطلاعات با توجه به مشخصه‌های مشابه و نحوه توزیع آنها به منظور طبقه‌بندی آنها را «خوشه‌بندی اطلاعات» (Clustring) می‌گویند.

تعیین اهمیت واژگان

برای تعیین عبارتهایی که می‌توانند به عنوان عبارتهای کلیدی متن مورد استفاده قرار گیرند، به یک معیار برای مقایسه و نمره‌گذاری محتوای اطلاعاتی مقاله نیاز داریم. عاملی که رتبه اهمیت هر عبارت به وسیله آن تعیین می‌گردد، تجزیه و تحلیل کلمات موجود در جملات می‌باشد. آنچه در مقاله حاضر برای اندازه‌گیری رتبه اهمیت یک کلمه مناسب تشخیص داده شده و پیشنهاد می‌شود تعداد رخداد کلمه، مجاورت مکانی این واژگان با یکدیگر و موقعیت مکانی آنها در مقاله نسبت به هم می‌باشد. نکته‌ای که در اینجا باید به آن اشاره شود اینکه برای رایانه، واژگان موجود در متون همانند یک سری اشیای فیزیکی می‌باشند. ماشین می‌تواند تشخیص دهد که آیا بعضی اشیا با هم مشابه اند یا نه، ماشین می‌تواند این قبیل یافته‌ها را به خاطر داشته باشد و می‌تواند بر روی آنهایی که قابل شمارش هستند، محاسبه انجام دهد. ماشین تمام این کارها را با استفاده از یک رویه از قبل برنامه‌ریزی شده انجام می‌دهد. در اینجا از هوش انسان فقط برای تهیه این برنامه‌ها استفاده می‌گردد.

دلیل استفاده از تعداد رخداد برای اندازه‌گیری رتبه اهمیت، بر این باور استوار است که نویسنده معمولاً از واژگان معینی برای پیشبرد، بحث یا تشریح دقیق جنبه‌های مختلف موضوع موردنظر استفاده و آنها را تکرار می‌کند. تعداد رخداد هر واژه می‌تواند به عنوان عامل تعیین درجه اهمیت واژگان مورد استفاده قرار گیرد. در غالب اوقات، واژه‌های معینی وجود دارند که با یکدیگر یک گروه را تشکیل می‌دهند، باید به این واژه‌ها رتبه اهمیت بالاتری اختصاص داد. در این میان، بعضی کلمات برای نشان دادن میزان ارتباط واژه‌ها با یکدیگر و گروه‌بندی آنها به کار می‌روند. به این واژه‌های رابط، نمره‌ای اختصاص داده نمی‌شود. این قبیل کلمات عمومی را می‌توان با تعریف یک سیاهه بازدارنده و تکمیل این سیاهه در طول زمان، حذف کرد (دیانی، ۱۳۸۱). به همین منظور، علاوه بر اینکه می‌توان حد بالایی[۱] را برای رخداد واژگان در نظر گرفت، یک سیاهه بازدارنده که قابلیت افزایش

---

۱. Threshold

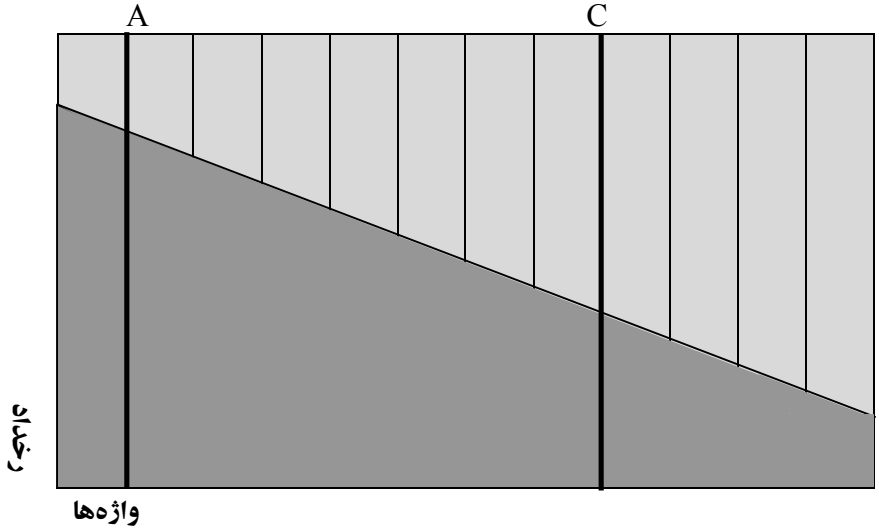
و کاهش آن توسط کاربر وجود دارد، در سیستم گنجانده می‌شود که از این طریق اثر بعضی از واژگان را خنثی و آنها را نادیده گرفت. تعیین حد بالا بر این حقیقت استوار است که واژگان عمومی نظیر و، به، با و ... از رخداد بسیار بالایی در مقایسه با سایر کلمات موجود در مقاله برخوردارند.

در تعیین رتبه اهمیت، از مسائل زبان‌شناختی همانند گرامر استفاده نمی‌شود. در یک نگاه کلی باید گفت، در روش حاضر حتی بین شکلهای مختلف کلمات نیز تفاوتی گذاشته نشده است. بنابراین، حالتهای مختلف کلمات جستجو، جستجوهای، جستجوها با یکدیگر یکسان است. در روش حاضر، به ارتباطهای منطقی و معنایی مورد نظر نویسنده توجه‌ای نشده است. به بیان دیگر، پس از بررسی متن، فهرستی از کلمات متن با یک نظم نزولی، بر حسب تعداد رخدادشان ایجاد و مرتب‌سازی می‌شوند.

رویّه‌ای که در روش حاضر مورد استفاده قرار می‌گیرد، بسیار ساده بوده و از نظر اقتصادی مقرون به صرفه است. این در حالی است که هر چه روش پیچیده‌تر باشد، ماشین باید سلسله عملیات بیشتری را متحمل گردد، که این خود باعث افزایش هزینه پردازشها خواهد شد. دلیل انتخاب یک روش ساده برای کار بر روی مقاله‌های فنی این است که با توجه به ماهیت مقاله‌های فنی، احتمال بسیار کمی وجود دارد که یک واژه برای نشان دادن بیش از یک مفهوم به کار رفته باشد یا نویسنده از کلمات متفاوتی برای نشان دادن بیش از یک مفهوم استفاده کند. حتی اگر یک نویسنده به دلایل نگارشی به انتخاب واژه‌های مترادف بپردازد، به زودی از این کار خسته شده و دوباره به استفاده از کلمه‌ای که اولین بار برای بیان مفهوم خود از آن استفاده کرده است، می‌پردازد. فهرستی از واژه‌های به دست آمده مطابق روش حاضر در نمودار شکل (۱) قابل مشاهده می‌باشد. چنانکه قبلاً نیز به آن اشاره کردیم، کلمات عمومی رخداد بالایی دارند که این خود موجب اختلال در سیستم می‌گردد. امکان کاهش تأثیر این اختلال، با ذخیره یک سیاهه از واژگان عمومی به صورت جداگانه، مقایسه این واژگان با واژگان متن و حذف واژگان عمومی از متن، وجود دارد. یک روش ساده‌تر این است که برای حصول اطمینان با استفاده از روشهای آماری، حدی را برای بالاترین رخداد تعیین کنیم. اگر خط A در شکل (۱) نشان‌دهنده این حد باشد، آنگاه تنها واژگانی که در سمت راست این خط می‌باشند، با اهمیت در نظر گرفته

می‌شوند. به دلیل اینکه میزان رخداد به عنوان یک معیار برای تعیین اهمیت واژگان تعیین شده است، باید حدّ پایینی نیز در این رابطه در نظر گرفته شود. در این قسمت، خط C نشان‌دهنده این حد باشد. تعیین یک محل مناسب برای این دو خط تجربی بوده و با توجه به بررسی نمونه‌های مقاله‌های چاپ شده در سطح وسیع، قابل تعیین می‌باشد. این امکان در سیستم حاضر وجود دارد که این محل (میزان حد بالا و پایین رخداد کلمات) برای تغییر خصوصیات خروجی حاصل، تغییر داده شود.

گهگاه دیده شده واژگانی غیر عمومی نیز، در سمت چپ خط A ظاهر شده‌اند. اگر برنامه به خوبی فرمول‌بندی شده باشد، محل واژه‌ها در نمودار می‌توانند بر از دست رفتن حد تمایز دلالت کند. تعیین یک مقدار برای حد، در یک رشته می‌تواند برای شاخه‌های خاصّ آن رشته یا حتی در رشته‌های دیگر نیز کاربرد داشته باشد. می‌توان به منظور افزایش کیفیت، حد تمایز را افزایش داد. در بعضی حالات تعدادی از واژه‌های عمومی در سمت راست خط A قرار می‌گیرند. در این حالت، تعداد این واژگان کم بوده و به علاوه با کمک سیاهه بازدارنده، تأثیر آنها به حداقل خواهد رسید.



شکل ۱. نمودار رخداد ـ کلمات بر روی محور افقی تک تک واژه‌ها به ترتیب تعداد رخداد آنها نمایش داده شده‌اند.

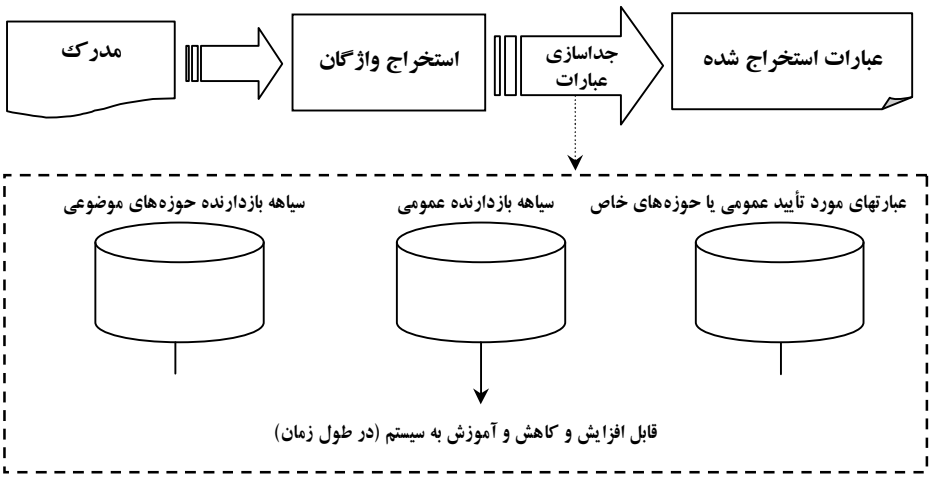
استخراج عبارتهای کلیدی

در روش حاضر با معانی واژه‌ها کاری نداریم و پیوند و ترکیب کلمات با یکدیگر به صورت قوی مورد بحث قرار نمی‌گیرد. البته، از این خصوصیت که فاصلهٔ کمتر واژه‌ها از یکدیگر بر جنبه خاصی از یک موضوع دلالت می‌کند، استفاده شده است. بنابراین، رخداد بالای کلمات مختلف در مجاورت یکدیگر، نشان دهندهٔ احتمال مرتبط بودن این واژه‌ها با محتوای مقاله است. رتبه اهمیت مجاورت می‌تواند بر اساس خصوصیات زبانی نوشته‌ها متفاوت باشد. به صورت فیزیکی، واژه‌هایی که برای بیان تصورهای متجانس ذهنی به کار می‌روند از لحاظ مکانی در موقعیت نزدیکتری نسبت به هم قرار دارند. تقسیم متن نوشته به جملات، پاراگرافها، فصلها و ... از راههای دیگری است که در آن درجه همبستگی تصورات با یکدیگر مشخص‌تر می‌شود.

مرحله بعدی ترکیب واژگانی است که حروف آغازین آنها با یکدیگر مشابه است؛ همانند جستجو، جستجوها، جستجوهای. این کار با یک تحلیل ساده آماری با مقایسه هر جفت واژه به صورت حرف به حرف در فهرست الفبایی واژه‌ها صورت می‌گیرد. اگر تعداد حروف مشابه آغازین برابر عدد چهار بود، دو واژه با یکدیگر مشابه فرض می‌شوند. متناسب با ماهیت زبان فارسی، این عدد می‌تواند بین ۳ تا ۵ قابل تغییر باشد. هرچند در این روش تطبیق، امکان خطا وجود دارد، اما به نظر نمی‌رسد خطاها بیش از ۵٪ باشد. بنابراین، در نتیجه نهایی تأثیری نخواهد داشت. در اینجا باید این نکته را خاطر نشان کرد که این مقدار در برنامه قابل تنظیم و تغییر می‌باشد. در مرحله بعدی ماشین، تعداد رخداد واژه‌های مشابه را مورد محاسبه قرار می‌دهد. مطابق با حد پایینی تعیین شده برای رخداد کلمات، واژه‌هایی که رخداد آنها از این میزان کمتر باشد، حذف می‌شوند و واژه‌های باقیمانده دوباره مرتب‌سازی می‌شوند. واژه‌های باقیمانده، وضعیت کلمات با اهمیت را نشان می‌دهند.

در نهایت، عبارتهای کلیدی از میان کلمات با رخداد بالایی که در کنار یکدیگر در سطح مقاله تکرار شده‌اند مشخص و با توجه به تکرارشان به هر کدام نمره‌ای اختصاص

داده می‌شود و کاربر می‌تواند سیاهه این عبارتها همراه با نمره‌های آنها را مشاهده کند. می‌توان مجموعه عبارتهایی را که تعداد تکرار آنها کمتر از حدّ معینی می‌باشد، حذف کرد. انجام این کار با توجه به سیاست سیستم در افزایش ضریب دقت یا بازیابی صورت می‌گیرد. تعیین تعداد واژه‌هایی که می‌توانند در میان واژگان موجود در عبارتها تکرار شوند و همچنین تعریف سیاهه‌ای از واژه‌های مجاز بین عبارت نیز در سیستم امکان‌پذیر می‌باشد. موقعیت مکانی واژگان در کلّ مقاله (مثلاً در عنوان) نیز در تعیین عبارتهای کلیدی مورد توجه قرار می‌گیرد که در مقاله دیگری توسط نویسنده به صورت جزئی مورد بررسی قرار می‌گیرد.



شکل ۲. مدل سیستم استخراج عبارتهای کلیدی

فرمول حاضر بر روی ۵۰ مقاله ۲۹۰ تا ۵۰۰۰ واژه‌ای مورد آزمایش قرار گرفت و مبتنی بر نتایج این جرأت حاصل گردید که عبارتهای استخراج شده به منظور ارزیابی در اختیار ۸ نفر از متخصصان نمایه‌سازی قرار گیرد.

یک مثال برای استخراج عبارتهای کلیدی در شکل (۳) از نشریه فصلنامه کتابخانه مرکزی آستان قدس رضوی وجود دارد و عبارتهای استخراج شده توسط سیستم در جدول (۱) قابل مشاهده است.

سازماندهی اطلاعات در نظامهای بازیابی اطلاعات

علی گزنی[۱]

چکیده

هر نظام بازیابی اطلاعات (نرم‌افزار) دارای یک مبنای خاص برای تجزیه و تحلیل اطلاعات است، که نظام بر اساس 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 توجه به متفاوت بودن سیاستهای بازیابی اطلاعات باید به صورت همزمان امکان استفاده از روشهای خودکار و نیمه‌خودکار فراهم آورده شود. پیش‌بینی سیاهه بازدارنده، ایجاد انواع واژه‌نامه‌ها مانند واژه‌نامه ریشه لغات، سیاهه پسوندها، واژه‌نامه عبارات، واژه‌نامه مفاهیم، برقراری روابط سلسله مراتبی مفاهیم، ریشه‌یابی واژگان، محاسبه همبستگی و خوشه‌بندی اطلاعات همگی از امکاناتی هستند که باید در یک نظام بازیابی اطلاعات بهینه وجود داشته باشد. مقاله حاضر، به بررسی این مفاهیم پرداخته است.

واژه‌های کلیدی: سازماندهی اطلاعات، نظامهای بازیابی اطلاعات، فایل واژه‌نامه، ریشه‌یابی واژگان، خوشه‌بندی اطلاعات.

مقدمه

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

۱. عضو هیئت علمی کتابخانه منطقه‌ای علوم و تکنولوژی شیراز

شکل۳. مقاله سازماندهی اطلاعات در سیستمهای بازیابی اطلاعات

جدول ۱. عبارتهای استخراج شده از مقاله سازماندهی اطلاعات در نظامهای بازیابی اطلاعات

عبارتهای استخراج شده	
واژه‌نامه	واژه‌نامه ریشه‌یابی
بازیابی اطلاعات	سازماندهی خودکار اطلاعات
نظامهای بازیابی اطلاعات	روش خودکار
سازماندهی اطلاعات	سازماندهی واژه‌ها
واژه‌نامه ریشه	واژه‌نامه مفاهیم
رخدادهای بالاتر	سازماندهی واژه‌ها
رکورد اطلاعاتی	ریشه واژه
بازیابی مدارک	خوشه‌بندی اطلاعات
واژه‌نامه عبارتهای	نظام خودکار
نظام اطلاعاتی	

آموزش سیستم

همان‌گونه که در شکل (۲) قابل مشاهده است، سیستم استخراج عبارتهای کلیدی می‌تواند یک سیر تکاملی را طی کند؛ بدین نحو که در طول زمان با توجه به آموزشهای کاربر، سیستم عملکرد خود را مطابق با نیاز و آموزشها تغییر می‌دهد. در سیستم حاضر، این آموزش می‌تواند در ٤ مقوله مورد توجه قرار گیرد که عبارتند از:

الف) سیاهه بازدارنده عمومی

این بانک شامل واژگان عمومی مشترک میان کلیهٔ حوزه‌های موضوعی می‌گردد. واژگان عمومی نظیر: به، با، و، که و ... در این مقوله قرار می‌گیرند. این قبیل واژگان که معمولاً تعداد رخداد بالایی دارند به حوزه خاصی وابسته نبوده و می‌توانند به عنوان یک بانک عمومی در کلیه شاخه‌ها تعریف شوند.

ب) سیاهه بازدارنده حوزه‌های موضوعی

با نمایش نتایج عبارتهای استخراج شده، کاربر با علامتگذاری، عبارتهای مرتبط را به خروجی ارسال می‌کند. در این مرحله، سیستم سیاهه‌ای از واژگان تأیید نشده را در بانکی تحت عنوان سیاهه بازدارنده حوزه‌های موضوعی نگهداری می‌کند. اطلاعات این بانک در پردازشهای بعدی مورد توجه قرار گرفته و عبارتهای موجود در این بانک به صورت خودکار از فهرست نتایج حذف می‌گردند.

ج) عبارتهای مورد تأیید عمومی یا حوزه‌های خاص

همانند مورد (ب)، سیستم سیاهه‌ای از عبارتهای تأیید شده کاربر در حوزه‌های

مختلف را نگهداری و بر اساس آن در پردازشهای بعدی، این عبارتهای به عنوان عبارتهای تأیید شده مورد توجه قرار می‌گیرند.

د) سیاهه‌ای از واژگان مترادف در حوزه‌های موضوعی

کاربر با معرفی واژگان مترادف، سیستم را قادر می‌سازد تا در پردازشها آنها را تشخیص و به عنوان یک واحد مورد محاسبه و پردازش قرار دهد.

نتیجه‌گیری

نتایج حاصل بر روی مقاله‌های فنی نشان داد که انتخاب خودکار عبارتهای کلیدی به نحوی که بیانگر موضوع کلی مقاله باشند عملی است و این عبارتها تا حدود زیادی شبیه عبارتهایی هستند که توسط انسان از میان نوشته انتخاب می‌شوند. مزیت اصلی این روش، یکدستی و یکنواختی آنهاست. به دلیل دخالت نداشتن تواناییها و تمایلات انسانها و استخراج عبارتهای با تحلیل آماری کلمات به کار رفته توسط نویسنده، عبارتهای استخراج شده از شایستگی، یکدستی و پویایی برخوردارند. هنگامی که عبارتهای کلیدی در سطح وسیع در اختیار کاربران قرار گرفت، کاربران یاد خواهند گرفت که چگونه آنها را درک کنند و چگونه مفاهیم مورد نظر خود را تشخیص دهند. البته، این احتمال نیز وجود دارد که با توجه به سبک نگارش نویسنده در گسترش مطالب، عبارتهای درجه دو انتخاب شوند یا نتایج حاصل نامناسب باشند، اما می‌توان حالتهای استثنایی را به سیستم آموزش داد؛ بدین صورت که سیستم را به شاخه‌های مختلف در حوزه‌های مختلف دانش تقسیم کرد و آنگاه سیستم رفتار خود را مطابق با آموزشهای کاربر و همچنین کارهای قبلی تغییر داد.

منابع

دیانی، محمدحسین (۱۳۸۱)، سازماندهی اطلاعات، جی تیلور، انتشارات کتابخانه رایانه‌ای، ص ۲۱۲ـ۱۶۳.

مهراد، جعفر (۱۳۷۳)، سیستمها و خدمات اطلاع‌رسانی کامپیوتر، هرمان و ویسمان، انتشارات نوید، ص۱۲٤ـ۹۷.

S. Rose (۱۹۹۹) The Sunflower Visual Metaphor: A New Paradigm for Dimensional Compression. EEE Proceedings of the Symposium on Information Visualization, PP. ۱۳۱-۱۳۴.