ارائهٔ روشی جدید برای شاخص‌گذاری خودکار و استخراج کلمات کلیدی برای بازیابی اطلاعات و خوشه‌بندی متون

فرهاد راد[1]، حمید پروین[2]، آتوسا دهباشی[3] و بهروز مینایی[4]

[1] دانشکدهٔ مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد یاسوج، یاسوج، ایران

[2و3] دانشکدهٔ مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد ممسنی، ممسنی، ایران

[2] باشگاه پژوهشگران جوان و نخبگان، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی، ایران

[4] دانشکدهٔ مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

چکیده

در زبان فارسی کلمات دارای صورت‌های نگارشی متنوعی هستند و پوشش کلیه حالات دستوری کلمات با به‌کارگیری یک سری قواعد معین ناممکن است؛ به همین دلیل استخراج کلمات کلیدی به‌طور خودکار از متون فارسی دشوار و پیچیده است. در این مقاله سعی شده است با استفاده از اطلاعات زبان‌شناختی و اصطلاح‌نامه، کلمات کلیدی بامعناتری استخراج شود. با استفاده از اصطلاح‌نامه که از نظامی ساختارمند برخوردار است، می‌توان شبکه کلمات کلیدی، شامل کلمات هم‌ارز، کلمات سلسله‌مراتبی و وابسته را تکمیل کرده و افزایش داد. بنابراین می‌توان توافق بین جستجوی کاربران و کلمات کلیدی متنی را بیشتر کرد و جامعیت جستجو را افزایش داد.

در گام نخست کلمات غیر مهم و عمومی حذف می‌شوند؛ سپس کلمات متن ریشه‌یابی و در ادامه برای مشخص‌شدن اهمیت نسبی کلمات با استفاده از روش‌های وزن‌دهی، یک وزن عددی به هر کلمه منسوب می‌شود که بیان‌گر میزان تأثیر کلمه در ارتباط با موضوع متن و در مقایسه با سایر کلمات به‌کار رفته در متن است. مجموعه عملیات بالا به‌خصوص استفاده از اصطلاح‌نامه باعث می‌شود که دسته‌بندی متون دقیق‌تر انجام گیرد و به‌نوعی رده علمی سلسله‌مراتبی متون در حوزهٔ بازیابی اطلاعات نیز مشخص می‌شود. نتایج آزمایش‌ها روی چندین متن در موضوعات مختلف نشان‌دهندهٔ دقت و توانایی روش پیشنهادی در استخراج کلمات کلیدی منطبق با خواست کاربر است و درنتیجه خوشه‌بندی دقیق‌تر متون است.

واژگان کلیدی: استخراج واژگان کلیدی، اصطلاح‌نامه، زبان‌شناختی، بازیابی اطلاعات.

۱- مقدّمه

ایده اصلی خوشه‌بندی اطلاعات، جداکردن نمونه‌ها از یکدیگر و قراردادن آنها در گروه‌های شبیه به هم است. به این معنی که نمونه‌های شبیه به هم باید در یک گروه قرار بگیرند و با نمونه‌های گروه‌های دیگر حداکثر تفاوت را داشته باشند (جین و همکاران، ۱۹۹۹؛ فاسلی و مارسیلیو، ۲۰۰۶). درواقع خوشه‌بندی داده‌ها یک ابزار ضروری برای یافتن گروه‌ها در داده‌های بدون برچسب است (استرل و گاش، ۲۰۰۲).

به‌صورت کلی روش‌های داده‌کاوی به دو گروه با ناظر و بدون ناظر تقسیم‌بندی می‌شوند. در روش‌های بدون ناظر متغیر هدفی تعریف نمی‌شود و الگوریتم داده‌کاوی همبستگی‌ها و ساختارهای بین تمام متغییرها را جستجو می‌کند. از مهم‌ترین روش‌های داده‌کاوی بدون ناظر، خوشه‌بندی را می‌توان نام برد.

در عصر کنونی، فناوری اطلاعات به‌یقین به‌عنوان یکی از مهم‌ترین مقوله‌های مورد بحث در میان متخصصان محسوب می‌شود. حجم عظیم و تاحدودی نامحدود اطلاعات موجود، باعث می‌شود تا استفاده از اطلاعات و درنتیجه مدیریت آن بسیار گسترده و درمواقعی پیچیده صورت پذیرد. با توجه به رشد و گسترش حجم اطلاعات و به موازات آن ضرورت به‌کارگیری و استفاده مؤثر از منابع اطلاعاتی، یکی از مهم‌ترین و اساسی‌ترین نیازهای موجود، قابلیت دست‌یابی به اطلاعات مورد نیاز در مدت زمان مناسب است. با وجود آن‌که اطلاعاتی که امروزه عرضه می‌شوند، صورت‌های

مختلفی از قبیل تصویر، صوت، پویانمایی و ... به خود گرفته‌اند، هنوز هم پراستفاده‌ترین و حجیم‌ترین اطلاعات موجود، متون غیر ساخت‌یافته هستند. یکی از کاربردهای خوشه‌بندی در همین مقوله است.

در این مقاله سعی شده است تا با استفاده از روابط موجود بین کلمات، به‌کمک اصطلاح‌نامه روش مناسبی برای ساخت خودکار شاخص در متون فارسی ارائه شود ضمن این‌که با به‌کارگیری روش خوشه‌بندی، متون موجود با دقت براساس محتوا و موضوع دسته‌بندی شوند.

سایر بخش‌های این مقاله به‌صورت زیر سازمان‌دهی شده است. در بخش دوم ادبیات موضوع شامل کارهای انجام‌شده در حوزه بازیابی متن، به‌ویژه در زبان فارسی مورد بررسی قرار می‌گیرد در این بخش همچنین مروری بر ساختار اصطلاح‌نامهٔ استفاده‌شده در این مقاله انجام می‌شود و مفاهیم مربوط به شاخص‌گذاری خودکار نظیر ریشه‌یابی و دیگر موارد بررسی می‌شوند. بخش سوم به ارائهٔ راه‌کار پیشنهادی اختصاص دارد. پیاده‌سازی راه‌کار پیشنهادی با تمرکز بر اصطلاح‌نامهٔ استفاده‌شده در بخش چهارم ارائه شده است. در بخش پنجم روش پیشنهادی روی مجموعه‌ای از متون فارسی مستخرج از روزنامه همشهری ارزیابی شده و با ارائه پرس‌وجوهای استاندارد کارایی روش بازیابی متن پیشنهادی مورد تحلیل قرار گرفته است و سرانجام بخش ششم به جمع‌بندی مطالب طرح‌شده اختصاص داده شده است.

۲- کارهای مرتبط

۲-۱ - تاریخچه شاخص‌گذاری خودکار

در سال ۱۹۹۹، برخی محققان که در زمینهٔ هوش مصنوعی کار می‌کردند با ارائه الگوریتم‌های پردازش ماشین، سعی به بالابردن کیفیت کلمات کلیدی استخراج‌شده کردند. (فرانک و همکاران، ۱۹۹۹). فرایند کلی برای استخراج کلمات کلیدی را برخی محققان در سال ۲۰۰۵ ارائه کردند که ابتدا کلمات کلیدی نامزد، تشخیص و به هر کلمه وزنی اختصاص داده شده و در نهایت کلمات کلیدی با بیش‌ترین وزن انتخاب می‌شدند (لیو و همکاران، ۲۰۰۵). در کاری دیگر در سال ۲۰۰۲، محققان تحلیل آماری و زبان‌شناختی را با یکدیگر ترکیب کردند (فرانتزی و همکاران، ۲۰۰۲). در سال ۲۰۰۵ محققان با توجه به مجموعه‌ای از سندهای آموزشی و کلمات کلیدی مشخص برای آن‌ها، فرایند استخراج کلمات کلیدی را به‌عنوان یک مسالهٔ رده‌بندی مدل کردند (فریتاس و همکاران، ۲۰۰۵). در سال ۲۰۰۶ از الگوریتم رده‌بندی درخت تصمیم در این زمینه استفاده شده است (ژانگ، ۲۰۰۶). در ادامه برخی پژوهش‌های خود را روی انتخاب کلمه نامزد متمرکز کردند و روشی بر مبنای مدل‌های چندکلمه‌ای را در سال ۲۰۰۸ ارائه دادند (لین و هوی، ۲۰۰۸). در همین راستا در سال ۲۰۰۳، محققی از روش‌های مربوط به زبان‌شناختی در سامانهٔ خود استفاده کرد (هالت، ۲۰۰۳). در سال ۲۰۰۳، سامانه‌ای را ارائه شد که براساس چهارکلمه‌ای‌ها و مدل فضای برداری کار می‌کرد و محاسبات آن ساده بود (رنز و همکاران، ۲۰۰۳). در سال ۲۰۰۴، محققان از دو منبع مختلف گنج‌واژه و دیگری روزنامه و صفحات خبری وب، برای استخراج کلمات کلیدی استفاده کردند و این نخستین سامانه‌ای بود که از گنج‌واژه استفاده می‌کرد (دیگان و همکاران، ۲۰۰۴). در سال ۲۰۰۶، محققی سعی کرد که از رابطهٔ سلسله‌مراتبی و طبقه‌بندی علوم در استخراج کلمات کلیدی استفاده کند که مزیت این روش گسترده، استخراج کلمات کلیدی و محدودیت آن درنظرگرفتن فقط به بخشی از خواص گنج‌واژه برای رده-بندی بود (هیون، ۲۰۰۶). در سال ۲۰۰۶، محققان سعی کردند که یک روش جدید پیشرفته برای استخراج کلمات کلیدی براساس اطلاعات معنایی پیشنهاد کنند (ویتن و مدلی، ۲۰۰۶). در همان سال، سعی کردند که سامانه‌ای را در زمینهٔ بررسی پرونده‌های محکومان به‌طور عملی و تجاری ارائه بدهند (کلین و همکاران، ۲۰۰۶). آن‌ها سامانه‌ای را برای استخراج کلمات کلیدی مبتنی بر گنج‌واژه ارائه دادند که برای مشخص‌کردن وضعیت پرونده‌های مجرمان استفاده می‌شد. اساس کار این سامانه مبتنی بر هستان‌شناسی بود که با توجه به درنظرگرفتن ارتباط معنایی بین کلمات قابل توجه است؛ ولی به هرحال به‌علت اعمال نشدن همهٔ موارد زبان شناختی شبیه به روابط اعم و اخص و سلسله‌مراتبی، نتایج این سامانه خالی از اشکال نبود (کلوین و استینبرگن، ۲۰۰۶). در سال ۲۰۰۷ روش کامل‌تری توسط محققان ارائه شد (رومرو و نینو، ۲۰۰۷). در سال ۲۰۰۸، در زمینهٔ استخراج کلمات کلیدی بر روی روزنامه کار شد (مارتینز و همکاران، ۲۰۰۸).

۲-۲- تعریف شاخص‌گذاری خودکار

براساس استاندارد شاخص‌گذاری (BS3700:1988)[1] شاخص‌ها مجموعه‌ای منظم از کلمات نشانه‌گذاری‌شده است

[1] British Indexing Standard

تا کاربران را قادر سازد اطلاعاتی را که محل آن‌ها در مدرک مشخص شده پیدا کنند.

جهت استفاده از روش شاخص‌گذاری خودکار باید داده‌ها به‌صورت ماشین‌خوان درآیند. این‌که واژگان از کدام محدوده از متن انتخاب شوند، بستگی به نرم‌افزار دارد. شرکت سافتک[1] نوعی برنامهٔ رایانه‌ای طراحی کرده که به ارائه خدمات شاخص می‌پردازد. این نرم‌افزار، اصولی برای آزمایش و بازیابی ساخته است. این نرم‌افزار از روش استفاده از واژه‌نامه بهره می‌برد. تغییرات واژگان متن با مطابقت‌دادن با واژه‌نامه‌های مختلف الکترونیکی اعمال می‌شود. همچنین این نرم‌افزار امکان تبدیل واژگان به ریشهٔ آن‌ها را جهت بازیابی بعدی فراهم می‌کند و علامت‌گذاری و محدودکردن واژه‌های ناخواسته را انجام می‌دهد. شکستن واژه‌های مرکب و ترجمه و انجام عمل ارجاع و مترادف‌سازی و ساخت عبارات را نیز انجام می دهد (سالتن و یانگ، ۱۹۸۳). درصورتی‌که شاخص‌ها به‌صورت خودکار از مدرک استخراج شده یا تولید گردند به آن شاخص‌گذاری، شاخص‌گذاری خودکار گفته می‌شود.

انجام صحیح شاخص گذاری به‌صورت دستی، به شخص بستگی دارد؛ چون تخصص بالایی را می‌طلبد؛ البته باید مدارک یک‌دستی در اختیار شخص قرار گیرد. به‌علاوه شخص باید به واژگان اصطلاح‌نامه نیز تسلط داشته باشد تا بتواند واژگان متعلق به مدارک را تجزیه و تحلیل کند.

اکنون به بررسی نقاط ضعف شاخص‌گذاری دستی پرداخته می‌شود که شامل موارد زیر است:

۱ - وقت زیادی را به خود اختصاص می‌دهد.

۲- تخصص بالایی را می‌طلبد.

۳- مخارج زیاد جهت استفاده از نیروی متخصص صرف می‌شود.

اما مزیت شاخص‌گذاری دستی، نسبت به شاخص‌گذاری خودکار این است که یک شخص شاخص‌گذار متخصص و مجرب به‌دلیل توانایی عقلانی‌اش، امکان درک صحیح محتوای مدرک را دارد و این باعث انتخاب صحیح شاخص‌ها خواهد شد.

۲-۳- تعریف اصطلاح‌نامه

اصطلاح‌نامه، مجموعه‌ای شامل واژه‌ها، اصطلاحات و اطلاعات مربوط به یک حوزهٔ خاص از معرفت بشری است. این مجموعه، واژگان زبان نمایه‌ای کنترل شده‌ای است که طوری سازمان یافته تا روابط پیشین میان مفاهیم (اعم و اخص و ...) را روشن کند (حری، ۱۳۸۳). واحد تشکیل‌دهنده اصطلاح‌نامه، واژه‌هایی است که تبلور اطلاعات و در برگیرندهٔ مسائل متن و مدرک مورد نظر است، که این‌ها را دراصطلاح واژه‌ها، یا نشانه‌های کلیدی و یا کلیدواژه می‌گویند. استخراج کلیدواژه از داخل متون و منابع به بازیابی اطلاعات متن کمک شایانی می‌کند.

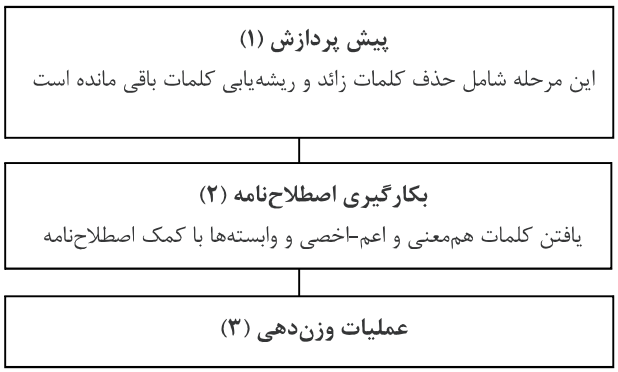
یکی از وظایف حتمی اصطلاح‌نامه این است که با نشان‌دادن روابط واژه‌ها با یکدیگر، روابط مفاهیمی را که این واژه‌ها بر آن‌ها دلالت دارند، نشان دهد. چنانچه این روابط، به‌طور دقیق ملاحظه شود و هر واژه‌ای در جایگاه معنایی خود قرار گیرد، نوعی تعریف ضمنی را نیز از هر اصطلاح به‌دست خواهد داد.

تاکنون، سه نوع رابطه، مورد توجه تدوین‌کنندگان اصطلاح‌نامه قرار گرفته است که عبارتند از (حری، ۱۳۸۳):

الف) رابطهٔ هم‌ارز[2] (رابطهٔ تعادل و ترادف): مانند رابطه حریق و آتش‌سوزی

ب) رابطهٔ سلسله‌مراتبی[3] (رابطه مرتبه‌ای یا رابطه کل و جزء): مانند رابطهٔ قرآن و آیه

ج) رابطهٔ همبسته[4]: مانند رابطهٔ قرآن و وحی



(شکل- ۱): معماری نمایه‌سازی پیشنهادی

۳- راه‌کارهای پیشنهادی

در این بخش راه‌کار پیشنهادی به‌منظور بهبود خوشه‌بندی متون فارسی به‌کمک اصطلاح‌نامه ارائه شده است. نمودار فرایند پیشنهادی در شکل (۱) نشان داده شده است. در ادامهٔ این فصل تک‌تک مراحل معرفی‌شده در نمودار به تفصیل مورد بررسی قرار می‌گیرد.

[1] http://www.softex.de

[2] Equivalence Relationship
[3] Hierarchical Relationship
[4] Associative Relationship

۳ـ۱ـ پیش‌پردازش متن

چنان‌که در شکل (۲) مشاهده می‌شود، در مرحلهٔ پیش‌پردازش، متون فارسی با استفاده از روش‌های فنی پیش‌پردازش پیشنهادی پالایش می‌شود تا کلمات اصلی متن استخراج شده و برای مرحلهٔ شاخص‌گذاری آماده شوند. این بخش شامل سه زیربخش است که در ادامه بحث خواهند شد.

(شکل- ۲): عملیات پیش‌پردازش روش پیشنهادی

۳ـ۱ـ۱ـ حذف کلمات Stop Word

واژه‌های عمومی زبان در شاخص‌گذاری، ارزش کمی دارند؛ به همین علت روند کلی شاخص‌گذاری با حذف واژه‌های عمومی آغاز می‌شود. از دلایلی که سبب می‌شود Stop Wordها را حذف کنیم، این است که این‌ها کمکی به بازیابی نمی‌کنند و دیگر این‌که حجم بالایی نیز دارند. این واژه‌ها حدود ۳۰ تا ۵۰ درصد متون را تشکیل می‌دهند. درواقع این واژه‌ها، واژه‌هایی هستند که کاربر مایل به جستجوی آنها نیست. این‌ها واژگانی هستند که در معنی جملات تاحدودی تأثیری ندارند؛ با وجود این در شکل‌گیری جملهٔ صحیح کمک می‌کنند؛ مانند حروف اضافه، بسیاری از قیدها، برخی از افعال (که خود ریشه‌اند)، حروف ربط و غیره. حذف آن‌ها به‌طورعمومی در مضمون کلی متن تأثیر منفی نمی‌گذارد؛ بلکه باعث خلاصه‌سازی متن می‌شود. اگرچه این همیشه درست نیست و به‌عنوان‌مثال واژه "تا" یک کلمه توقف است، اما گاهی حرف اضافه و گاهی نشان‌گر استنتاج است (مانند جمله "آمدم تا او را ببینم" که در این‌جا تا نشان‌گر علت آمدن است)؛ امّا با کمی چشم‌پوشی آن‌ها حذف می‌شوند. نکته قابل توجه این است که در این مرحله، تنها افعال ریشه از مستند حذف می‌شوند و حذف سایر افعال که ریشه نیستند، پس از مرحلهٔ ریشه‌یابی انجام می‌شود.

۳ـ۱ـ۲ـ ریشه‌یابی

اکنون به الگوریتم پیشنهادی، برای ریشه‌یابی پرداخته می‌شود. در این الگوریتم ابتدا واژه در فهرست واژگان که فقط اسامی جامد را در برمی‌گیرد و از قبل آماده شده است، جستجو می‌شود. در صورت وجود واژه در فهرست واژگان، عملیات ریشه‌یابی انجام نمی‌شود و خود واژه به‌عنوان ریشهٔ اصلی در نظر گرفته می‌شود. اما درصورتی‌که واژه در فهرست واژگان موجود نباشد، تا حد ممکن الگوریتم ریشه‌یابی روی آن واژه اجرا می‌شود. چنانچه واژهٔ حاصل‌شده در هر مرحله از الگوریتم، در فهرست واژگان موجود بود، عملیات مورد قبول واقع می‌شود و کار به پایان می‌رسد. در طراحی ریشه‌یاب سعی بر آن بوده است که ضمن حفظ انسجام، اجزا مستقل از یکدیگر باشند. بنابراین سامانه از چندین بخش تشکیل شده است. این بخش ها عبارتند از:

الف) مجموعه قوانین ریشه‌یابی (برای زدودن پسوند و پیشوند) که با استفاده از قوانین ساخت واژه در زبان فارسی استخراج شده‌اند.

ب) فهرست مصادر فارسی (اعم از افعال ساده، پیشوندی، مرکب و نیز عبارت‌های فعلی)

ج) فهرست واژگان جمع مکسر که برای هر واژهٔ جمع مکسر، مفرد آن را مشخص می‌کند.

د) فهرست واژگان جامد که کلیهٔ واژگان فارسی به غیر از افعال و مشتقات آن، جمع‌های مکسر و نیز مابقی واژگان مشتق را نگهداری می‌کند. به‌دلیل این‌که چنین فهرستی برای زبان آزمایشی بوده است، که اکنون حدود ۹۰۰۰ واژه را دربردارد، می‌توان به‌مرور آن را تکمیل‌تر کرد. بدیهی است هم‌نگاره‌های غیرواژگانی که در متون فارسی بسیار زیادند (مانند آسمانی که هم به معنی یک آسمان است و هم به معنی صفت آسمانی، ولی در متن به یک صورت نوشته می‌شوند)، ممکن است ریشه‌یاب به‌درستی کار نکند و هر دو را یک واژه ببیند؛ مگر آن‌که با روش‌های رایانه‌ای قاعدهٔ بنیاد آن‌ها را بازشناسی کرد.

برای آن‌که بتوان ریشهٔ واژگانی را که از زبان عربی وارد فارسی شده‌اند، استخراج کنیم، دو روش به نظر می‌رسد. یکی این‌که قواعد و ریشه‌های واژگان زبان عربی را نیز، هر چند محدود، در سامانه وارد کنیم، و یا اینکه با

استفاده از یک فرهنگ ترادف تمامی واژگان هم‌خانواده را فهرست کنیم.

۳ـ۱ـ۳ ـ حذف مجدد کلمات اضافی

در این مرحله به‌طورتقریبی تمام افعال و اسامی با ریشه‌ها جایگزین شده‌اند. حال طی یک مرحله، دوباره کلمات اضافی را براساس فایل کلمات اضافی که در گام نخست استفاده کرده‌ایم حذف می‌کنیم. به‌صورت‌فرضی برخی از افعال پس از حذف شناسه به افعال ریشه تبدیل می‌شوند که قابل حذف است. در پایان این مرحله بخش عظیمی از کلمات اضافی از متن حذف می‌شوند.

فهرست واژگانی که در جدول (۱) آمده توسط خود مؤلفان از روی کارهای بیجن‌خان و سمیعیان ایجاد شده که حدود ۸۰۰ لغت است که تنها بخشی از آن در جدول (۱) آمده است (بی‌جن‌خان ۱۳۸۳؛ سامییان، ۱۹۸۳).

(جدول- ۱): فهرست واژه‌های عمومی (stop word)

آن	ای	به	چندگانه	خودشان	روی	گرچه	میشوند	هر	همیشه
آنان	ایشان	بیشتر	چندین	خودم	زیرا	گرفته	میکند	هریک	همین
آنجا	این	بین	چنین	خودمان	سپس	لکن	میکنند	هست	هنوز
آنچه	اینجا	پس	چه	خویش	شامل	لیکن	نظر	هستند	هیچ
آنکه	اینست	تا	چون	داده	شاید	ما	نمیتوان	هستیم	هیچکدام
آنگاه	اینگونه	تایی	چیز	دارای	شد	مابین	نوع	هم	هیچگونه
آنها	اینها	تو	چیزی	دارد	شده	مانند	نیاز	همان	و
از	با	توسط	چیست	داشته	شما	مختلف	نیز	همانطور	وجود
است	باشد	چرا	حتی	در	شود	من	نیست	همانند	وگرنه
اگر	باید	چگونگی	خواهد	درباره	صورت	مورد	نیستند	همچنین	ولی
اگرچه	بدون	چگونه	خواهیم	دو	فقط	میباشد	ها	همچون	وی
الان	بر	چنان	خود	دیگر	کدام	میتوان	های	همدیگر	یا
اما	برای	چنانچه	خودت	دیگران	کرد	میتواند	هایی	همه	یک
انجام	بنابر	چند	خودتان	دیگری	کردن	میدهد	هر	همواره	یکدیگر
او	بنابراین	چندان	خودش	را	که	میشود	هرچه	همیشگی	یکی

۳ـ۲ـ به‌کارگیری اصطلاح‌نامه

در این مرحله با استفاده از اصطلاح‌نامه برای کلمات اصلی مستخرج از بخش پیش‌پردازشی، کلمات هم‌معنی و اجداد و وابسته‌ها شناسایی می‌شوند. به عبارت دیگر، برای تک‌تک کلمات اصلی متن کلمات هم‌معنی و پدر و وابسته استخراج و در جایی نگهداری می‌شود. از این کلمات بعدها برای وزن‌دهی استفاده خواهد شد. هدف آن است که در صورت دیدن کلمات هم‌معنی در متن، به جای آنکه به‌صورت مجزا برای هر یک وزنی در نظر گرفته شود، یک کلمه از میان آنها به‌عنوان نماینده انتخاب شده و مجموع وزن کلمات هم‌معنی و وابسته به نماینده در متن به وزن کلمهٔ نماینده اضافه شود. نحوهٔ به‌دست‌آوردن نماینده، به این شکل است که ابتدا تمام کلمات هم معنی و موجود در دایره ریشه کلمات موجود در گنج‌واژه یک گروه قرار می‌گیرند. یعنی به‌عنوان‌مثال برای کلمه "سگ"، مجموعه کلمات هم‌معنی و موجود در دایره ریشه کلمات موجود در گنج‌واژه، در یک گروه قرار می‌گیرند؛ سپس آنها براساس حروف الفبا مرتب می‌شوند. نخستین کلمه به‌عنوان ریشه کلمه انتخاب می‌شود.

برای مثال، کلمات موجود در متن شکل (۳) به هم وابسته هستند. با درنظرگرفتن یکی از سه کلمه مشخص در متن به عنوان نماینده، چنان‌که گفته شده است، وزن آن در متن، ۳ می‌شود.

.....وظیفه.....

.....خدمت.....

.....سربازی.....

.....

.....

(شکل- ۳): یک متن شامل سه واژهٔ مترادف

۳ـ۲ـ۱- کلمات هم‌ارز

روش استخراج کلمات کلیدی از متن از یک زاویه به دو صورت می‌باشد.

اگر متنی را در نظر بگیرید، اغلب اوقات به‌خاطر این‌که از تکرار یک کلمه جلوگیری شود، به‌طورمعمول کلماتی استفاده می‌شود که در آن متن به یک معنی به‌کار می‌روند (به الزام هم معنی نیستند؛ ولی در یک متن برای خواننده یک معنی دارند). مثل متنی که حاوی کلمات سربازی، اجباری و وظیفه است، درواقع این کلمات در متن شکل (۳) به‌دلیل ارتباط معنایی، همگی برای خواننده متن یک معنی دارند. این گونه کلمات را زنجیرهٔ واژگانی می‌گویند. وجود زنجیرهٔ واژگانی باعث پراکندگی می‌شود.

اشکال اصلی روش‌های موجود در بازیابی متن در زبان فارسی، عدم توجه به زنجیرهٔ واژگانی ا ست. اگر خودمان در متن مذکور بخواهیم کلمات کلیدی را وزن دهیم، با خواندن متن مذکور وزن بالایی به کلمهٔ سربازی می‌دهیم. این روش قادر است با دیدن کلمات هم‌معنی (زنجیرهٔ واژگان)، همه را به‌عنوان کلمه اصلی در نظر گرفته و بیش‌ترین وزن را به کلمه اصلی بدهد و بدین ترتیب روش پیشنهادی از توزیع وزن کلمات هم‌معنی که مانع از تشخیص دقیق رده متن مورد نظر می‌شود جلوگیری کند.

برای انجام این کار با استفاده از اصطلاح‌نامه برای کلمات اصلی موجود در متن کلمات هم‌معنی پیدا شده، و با دیدن هر یک از کلمات در متن، وزن کلمهٔ اصلی یک واحد اضافه می‌شود؛ یعنی یک کلمه به‌عنوان نماینده تعیین و هر بار به بسامد این کلمه اضافه می‌شود. برای تشریح به شکل (۴) دقت شود.

یادم می‌آید روزی که می‌خواستم به سربازی بروم خیلی ترسیده بودم زیرا از سربازی یک کابوس برایم ساخته بودند و به آن اجباری می‌گفتند. چاره‌ای نداشتم و باید به سربازی می‌رفتم زیرا اگر می‌خواستم جذب یک کار دولتی بشوم اولین سوالی که از من می‌پرسیدند این سوال هست "خدمت رفتی؟"، هنوز به سربازی نرفته دوستانم به من آشخور می‌گفتند، در حالی که سربازی یکی از اموری است که هر پسر سالم و بالغی باید آن را سپری کند.

(شکل- ۴): یک متن نمونه، حاوی واژه‌های مترادف

در متن شکل (۴) یک‌بار لغت سربازی، یک‌بار لغت اجباری، یک‌بار لغت خدمت و یک‌بار لغت آشخور تکرار شده است. بنابراین از متن شکل (۴) داده‌های جدول (۲) استخراج می‌شوند که نشان‌گر تکرار مفهوم سربازی در متن است. ذکر این نکته ضروری است که به جای هر چهار واژه مذکور، یک نماینده به نام سربازی در نظر گرفته شده است و فراوانی آن نیز ۴ شده است.

(جدول- ۲): چگونگی برخورد با واژه‌های مترادف در متن شکل (۴)

لغات	فراوانی زنجیره کلمات	
آشخور	۴	نماینده
خدمت	۴	
سربازی	۴	
وظیفه	۴	

در مثال جدول (۲)، اگر فقط فراوانی کلمات موجود در متن در نظر گرفته شوند، کلمات کلیدی مناسبی انتخاب نخواهند شد؛ درحالی‌که با رجوع به اصطلاح‌نامه و درنظرگرفتن زنجیره کلمات وابسته کلمات کلیدی مناسبی انتخاب می‌شود. به‌طور مثال با دیدن کلمه اجباری، مشخص می‌شود که این کلمه دارای ارتباط معنایی نزدیک به کلمه سربازی است و لذا به کلمه مرجح آن یعنی سربازی یک واحد اضافه می‌شود. همچنین در مورد کلمات خدمت و آشخور، هر کدام باعث می‌شوند یک واحد به فراوانی کلمهٔ سربازی اضافه شود. یعنی درنهایت کلمه سربازی در این متن دارای وزن انتخابی برابر چهار است، این باعث می‌شود که کلمه سربازی که یک کلمه کلیدی صحیح برای این قطعه متن است، به‌عنوان کلمه کلیدی انتخاب شود.

۳ـ۲ـ۲ـ رابطه اعم

اگرچه استفاده از کلمات هم‌معنی و وابسته با استفاده از اصطلاح‌نامه در وزن‌دهی کلمات موجود در متن کمک زیادی به رده‌بندی متن و فرایند بازیابی می‌کند، اما به‌الزام برای رده‌بندی صحیح متن‌ها کافی نیست. شکل (۵) را که در مورد آموزش و یادگیری است، در نظر بگیرید.

طرح «شناسنامه‌دارشدن مکان‌های آموزشی کشور» که از سال ۸۳ شروع شد و قرار بود تا شهریور ۸۴ انجام شود، با وجود گذشت ۵ سال هنوز هم ادامه دارد. در این طرح برای مدارس، آموزشگاه-ها، دانشگاه‌ها و پژوهشگاه‌ها شناسنامه تهیه می‌شود. لذا با انجام این طرح زمینه‌ی شناسایی مکان‌های مذکور فراهم می‌شود.

(شکل- ۵): یک متن نمونه حاوی واژه‌های اعم

در این متن کلمات مدارس، آموزشگاه، دانشگاه، پژوهشگاه و ... بیش از سایر کلمات تکرار شده‌اند؛ اما همان‌طور که ملاحظه می‌شود، این کلمات هم‌معنی نیستند و نمی‌توان یک کلمه را به‌عنوان نمایندهٔ همه آن‌ها در نظر گرفت؛ درحالی‌که در اصطلاح‌نامه مورد استفاده همه این کلمات در زیردهٔ آموزش و پرورش قرار می‌گیرد. در صورت عدم استفاده از اصطلاح‌نامه، ممکن است رده‌بندی صحیح متن مورد نظر به‌دلیل پراکندگی کلمات اصلی امکان‌پذیر نباشد؛ اما در روش پیشنهادی برای هر یک از این کلمات، اعم آن‌ها که همان آموزش است در نظر گرفته شده و به این ترتیب کلمه کلیدی آموزش امتیاز بالایی در متن پیدا می‌کند.

۳ـ۲ـ۳ـ رابطهٔ اخص

در این حالت از کلمات خیلی کلی‌تر به جزئی‌تر می‌رسیم. برای تشریح به مثال شکل (۶) دقت شود. در متن شکل (۶) یک‌بار لغت آنفولانزا، یک‌بار لغت بهداشت، یک‌بار لغت پزشکی، یک‌بار لغت سلامت و یک‌بار لغت درمان تکرار شده است. چنان‌که ملاحظه می‌شود، در این متن از کلمات دکتر، وزارت بهداشت، واکسن سلامتی و آنفولانزا ... استفاده شده است که این‌ها همگی کلمات کلی به‌شمار می‌آیند. از این کلمات چنین برمی‌آید که متن راجع به بهداشت و درمان است. زمانی‌که به زیرشاخه رجوع شود که به‌صورت جزئی به کدام رده اشاره شده است، کلمهٔ آنفولانزا یافت می‌شود. کلمه آنفولانزا دارای وزن معقولی در متن است؛ بنابراین نتیجه‌گیری می‌شود که به‌طور کلی متن در مورد آنفولانزا است (آنفولانزا کلمه اخص در این متن است). بدین ترتیب متن‌ها می‌تواند با دقت بالاتری رده‌بندی شوند و با همان دقت بالا پرس‌وجوی‌های کاربر پاسخ داده شوند؛ پس اگر کاربری به‌طور خاص راجع به متون آنفولانزا درخواست کند، در درجه نخست متن‌های مستقیم با آنفولانزا در اختیار او قرار خواهد گرفت و درجه‌های بعدی متون مربوط به بهداشت و درمان به وی برگشت داده خواهد شد.

۳ـ۳ـ مرحلهٔ وزن‌دهی

وزن‌دهی برای کلمات هم‌معنی (زنجیره واژگان) به این ترتیب انجام می‌شود که برای کلمات هم‌معنی موجود در متن به‌کمک اصطلاح‌نامه، وزنی مساوی در نظر گرفته می‌شود. اگر چند کلمهٔ هم‌معنی در متن وجود داشته باشند، یکی بر اساس اصطلاح‌نامه به‌عنوان نماینده انتخاب شده (کلمه مرجح) و در هر بار تکرار کلمات هم‌معنی یک واحد به وزن نماینده اضافه می‌شود. با توجه به این‌که نمی‌توان برای کلمه اعم و اخص یک کلمه موجود در متن، وزن مساوی با آن را اختصاص داد، لازم است روش ساخت‌یافته‌ای برای وزن‌دهی کلمات موجود در متن و وابسته‌های آن‌ها ارائه شود.

به گزارش و به نقل از پایگاه اطلاع‌رسانی دولت در دستورالعمل پیش‌گیری از آنفلوانزا در ادارات با برشمردن علائم و ویژگی‌های این بیماری آنچه که کارکنان برای سلامت و حفاظت خود در مقابل این بیماری باید انجام دهند به تفصیل بیان شده است. توصیه می‌شود شخص به یاد شده باید با مرکز بهداشتی درمانی که نزدیک محل کار است در ارتباط باشد. همچنین سیاست‌هایی جهت تسهیل دسترسی به خدمات بهداشتی درمانی، خدمات پزشکی، جهت ارتباط با پزشک و تزریق واکسن‌های لازم جهت پیش‌گیری از آنفولانزا فراهم گردد.

(شکل- ۶): یک متن نمونه حاوی واژه‌های اخص

۳-۳-۱- وزن‌دهی کلمات اعم و اخص

از آن‌جایی که اعم یک کلمهٔ اصلی در متن به‌الزام نمی‌تواند جایگزین آن کلمه شود و درصورت جایگزینی با کلمه اصلی مفهوم اصلی متن می‌تواند تغییر کند، نمی‌توان وزن مساوی برای کلمات اعم و اخص و کلمهٔ نماینده در نظر گرفت. در چنین مواردی بر اساس نوع کاربرد و میزان دقت مورد نیاز در رده‌بندی، وزنی کمتر از یک برای کلمات اعم یک کلمه اصلی در نظر گرفته می‌شود. برای مثال برای هر سطح در ساختار درختی رابطه، به‌ترتیب وزن‌های ۰/۵، ۰/۲۵ و ۰/۱۲۵ ... در نظر گرفته می‌شود. در مورد اخص‌ها هم‌چنین ترتیبی در نظر گرفته می‌شود (البته به این اعداد به‌صورت تجربی با آزمایش روی اعداد مختلف و بررسی نتایج دست یافته شده است).

در این مرحله به هر یک از کلمات، وزنی اختصاص می‌یابد، اگر کلمه جزء کلمات هم‌معنی باشد، وزن یک و اگر جزء کلمات اعم و اخص و همبسته باشد، وزن یک‌چهارم به آن اختصاص می‌یابد؛ سپس کلمات برحسب تعداد دفعات تکرارشان مرتب می‌شوند. در مرحلهٔ آخر پس از وزن‌دهی به کلمات دارای بیش‌ترین وزن به همراه پدر به‌عنوان کلمه کلیدی معرفی می‌شوند. متن شکل (۷) را در نظر بگیرید. دقت شود که در این متن با اعمال الگوریتم گفته‌شده به

جدول کلمات و وزن‌هایی به‌دست می‌آید که خروجی نهایی آن در جدول (۳) آورده شده است؛ که با توجه به این جدول کلمات کلیدی این متن عبارتند از: *آب‌های سطحی*، *رودخانه*، *دریاچه* و *تالاب*.

آب جاری در رودخانه‌ها، دریاچه‌ها و تالاب‌ها را آب سطحی می‌گویند. آب سطحی بطور طبیعی از طریق بارش (برف و باران) تامین می‌شود و با ورود به دریاها یا تبخیر و یا نفوذ عمقی به سفره‌های آب زیرزمینی از چرخه دسترسی خارج می‌شود. وقتی ما باتلاق‌ها، لجن‌زارها، تورب‌زارها، رودخانه‌ها و دریاچه‌ها، خورها و مناطق ساحلی اشاره می‌کنیم، متوجه می‌شویم که اکوسیستم‌های آبی در فراهم کردن آب پاک، حفظ تنوع زیستی، کاهش تاثیر تغییرات آب و هوایی، کنترل سیلاب و بسیاری منابع دیگر نقش بنیادی دارند.

(شکل- ۷): متن حاوی واژه های مختلف

۴ـ آزمایش‌ها

به‌منظور پیاده‌سازی روش پیشنهادی، از روش فنی خوشه‌بندی استفاده شده است. به این منظور یک فضای‌برداری شامل تمامی کلمات موجود در پایگاه داده‌ای متن ایجاد می‌شود. درنتیجه هر متن به‌صورت برداری از کلمات در فضای برداری نمایش داده شده است. به عبارت دیگر هر متن نقطه‌ای از این فضای‌برداری خواهد بود؛ با استفاده از روش خوشه‌بندی، متون مشابه و مربوط به یک موضوع، در یک خوشه قرار می‌گیرد. پرس‌وجوی کاربر به‌سادگی با مقایسه با مراکز خوشه‌ها پردازش شده و نزدیک‌ترین خوشه شامل شبیه‌ترین متن‌های موجود در خوشه به کاربر برگردانده می‌شود. همان‌گونه که در بخش قبل توضیح داده شد، این فرایند شامل چند گام اصلی است.

گام نخست پیش‌پردازش: همان‌گونه که بیان شد در این گام بایستی کلمات اضافی هر متن حذف، ریشه کلمات استخراج و کلمات اصلی متن وارد پایگاه داده متن شود. برای انجام این کار پرونده‌ای شامل کلمات اضافی متداول در نظر گرفته شده است، متن مورد نظر کلمه به کلمه خوانده شده و هر کلمه با کلمات موجود در فایل کلمات اضافی مقایسه می‌شود. درصورتی‌که کلمه مورد نظر در فایل کلمات اضافی وجود داشته باشد از متن حذف می‌شود، در غیر این صورت به رکورد متن مورد نظر در بانک اطلاعاتی پایگاه داده اضافه می‌شود. به این ترتیب با یک پویش ترتیبی، کلمات اضافه رایج در متن حذف می‌شوند. در گام دوم پیش‌پردازش با به‌کارگیری فرمول ساده‌ای برخی از پیشوندها و پسوندهای متداول از کلمات باقی‌مانده جدا شده و سعی می‌شود ریشه کلمات باقی بماند. برای مثال کلمه‌ای مانند درختان که "ان"، پسوند جمع برای آن به حساب می‌آید، از کلمه درخت حذف می‌شود. قابل توجه است که به سبب وجود استثناهای فراوان در کلمات فارسی، انجام مرحله حذف پیشوندها و پسوندها همواره و به‌سادگی امکان‌پذیر نیست. برای مثال برای کلمه مثل "باران" چون "ان" بخش اصلی کلمه است امکان حذف نیست. درنتیجه چنان که پیشتر گفته شد، با کمی چشم‌پوشی در این مرحله به حذف پیشوندها و پسوندهای کلمات که حرف اصلی آن بیش از سه حرف باشد اقدام شده است. مشاهدات و بررسی‌ها نشان داده است که در بسیاری از موارد درنظرگرفتن چنین معیاری منجر به حذف صحیح پیشوند و پسوند می‌شود؛ با این حال می‌توان از روش‌های فنی پیشرفته‌تری به‌منظور استخراج ریشه از کلمات اصلی استفاده شود. دقت ریشه‌یاب مورد استفاده بر روی یک متن که دارای ۳۸۹ کلمه بوده است، ۹۷/۹۴ درصد بوده است.

پس از انجام مراحل پیش‌پردازش متن، که طی آن کلمات اصلی متن استخراج شدند، نوبت به شمارش تعداد تکرار کلمات موجود در متن می‌رسد. در روش‌هایی که تا به حال در زبان فارسی به این منظور استفاده شده است، هم‌معنی بودن کلمات و وجود کلمات اعم و اخص در یک متن مورد توجه قرار نمی‌گرفته است. همان‌گونه که گفته شد از آنجایی که نویسندگان متن تمایل دارند به‌منظور تکراری‌نبودن کلمات، از کلمات هم‌معنی برای شکیل‌بودن متن استفاده کنند، عدم درنظرگرفتن مترادف‌ها باعث کاهش دقت فرایند طبقه‌بندی متن می‌شود. به عبارت دیگر چون یک کلمه به شکل‌های مختلفی در متن به‌کار رفته است، تعداد تکرار کلمه مورد نظر در متن توزیع می‌شود. درنتیجه شمارش ساده، معیار مناسبی برای تعیین رده یا موضوع متن نیست. با استفاده از یک اصطلاح‌نامه به‌سادگی می‌توان کلمات هم‌معنی را شناسایی کرد و برای همه این کلمات یک نماینده در متن در نظر گرفت و با دیدن کلمات مترادف در متن به فرا و انی‌کلمه نماینده اضافه شود. برای پیاده‌سازی این مرحله از یک اصطلاح‌نامه کامل متون فارسی (اصفا) استفاده شده است که در بخش قبل مورد بررسی قرار گرفته است .

۴ـ۱ـ بررسی ساختار اصطلاح‌نامه

این ساختار در یک فایل Access قابل نمایش است. ساختار اصطلاح‌نامه از ترکیبی از ساختار اصطلاح‌نامه اسلامی و اصفا

(خسروی، ۱۳۷۹) گرفته شده و داده‌های آن از پایگاه کتابخانه ملی استخراج شده و پس از تبدیل به فرمت لازم برگردانده شده است. لازم به ذکر است که تهیه این اصطلاح‌نامه بسیار زمان‌بر و طاقت‌فرسا بوده و قبل از به‌دست‌آوردن این اصطلاح‌نامه کارهای زیادی با اصطلاح‌نامه‌های متنوعی انجام شده است که به علت عدم کامل بودن اصطلاح‌نامه‌های پیشین، نتیجه خوبی عاید نشده است. این نخستین اصطلاح‌نامه‌ای می‌باشد که نتایج خوبی از آن منتج شده است.

۴ـ۲ـ نحوه انجام کار

از آن‌جایی که در گام نخست، هدف یافتن میزان تکرار کلمات مترادف در متن است، با شروع از متن کلمات اصلی، برای هر کلمه اصطلاح‌نامه مورد نظر مورد بررسی قرار می‌گیرد. درصورتی‌که کلمه خوانده‌شده، هم‌نام کلمه دیگری باشد که قبلاً در متن خوانده شده است، به جای اضافه‌کردن تکرار کلمه جدید خوانده شده، به شمارنده کلمه نخست، یعنی نماینده کلمه، یک واحد اضافه می‌شود. همان‌طور که در بخش قبل بیان شد، از آن جایی که نمی‌توان برای کلمه نماینده اعم و اخص، وزنی مساوی کلمات هم‌معنی در نظر گرفت، با دیدن اعم و اخص یک کلمه، به وزن کلمهٔ نماینده درصد مشخصی اضافه می‌شود. در این مقاله براساس پژوهش‌های انجام‌شده به این نتیجه رسیدیم که مقدار (۱/۲)[۲] است.

به این ترتیب با یک پویش ترتیبی کلمات اصلی مشخص می‌شود که آیا قبلاً هم‌نام آن کلمه در متن دیده شده بود، که در این صورت یک واحد به شمارنده اضافه می‌شود و اگر اعم یا اخص کلمه مورد نظر بود، مقدار ۱/۴ به شمارندهٔ کلمهٔ نماینده اضافه می‌شود. برای درک بهتر نحوهٔ پیاده‌سازی، متن کوتاه زیر را در نظر بگیرید که در آن مجموعه‌ای از کلمات مترادف و اعم و اخص قرار گرفته‌اند. بدین ترتیب کافیست برای هر کلمه خوانده شده جدید جستجویی در اصطلاح‌نامه صورت بگیرد. اگر در اصطلاح‌نامه برای کلمه مورد نظر، کلمات مترادف، اعم و اخص وجود داشت هر یک از این کلمات هم‌معنی و اعم و اخص در بانک داده‌های مربوط به آن متن جستجو و یافتن نماینده مورد نظر به شمارنده آن یک واحد اضافه می‌شود.

پس از مشخص شدن تعداد تکرارهای کلمات موجود در متن با در نظر گرفتن هم‌معنی‌ها، اعم‌ها و اخص‌های کلمات، نوبت به ایجاد بردار می‌رسد. بدین منظور برای هر متن یک بردار در نظر گرفته می‌شود که حاوی مقادیر صفات است. صفات در اینجا کلیه کلمات اصلی موجود در بانک داده‌ای متن‌ها هستند. بدین ترتیب هر عنصر از یک بردار نمایندهٔ یک کلمهٔ اصلی بوده و برای یک متن خاص درصورت دارا بودن کلمه مورد نظر، عنصر مربوطه در بردار، مقدار تکرار کلمه را گرفته و درصورت عدم حضور کلمه در متن مورد نظر، مقدار عنصر مورد نظر، صفر خواهد بود. نمونه‌ای از یک بردار برای متن ذکرشده در صفحهٔ قبل به شکل ۸ است.

V1	V2	V3	V4	V5	V6	…	V1256	V1257	…
0	2	0	1	8	3	…	2	4	…

شکل ۸- یک بردارمتن نمونه‌ای

با ساخت بردارهای مربوط به متن‌های مختلف با استفاده از روش kmeans متون موجود در بانک خوشه‌بندی می‌شود و برای هر خوشه یکی از بردارها به‌عنوان یک مرکز خوشه در نظر گرفته می‌شود تا در گام بعدی راه‌کار برای مقایسه پرس‌وجو و خوشه‌ها مورد استفاده قرار گیرد.

۵ـ نتایج و تفسیر

در این بخش، نتایج حاصل از ارزیابی روش پیشنهادی ارائه شده است. در بخش نخست، نتایج حاصل از انجام پیش‌پردازش متن مورد بررسی قرار گرفته است. در بخش دوم، خوشه‌بندی متون فارسی انتخابی به همراه کلمات کلیدی هر خوشه ارائه شده است.

۵ـ۱ـ نحوه انتخاب متون فارسی جهت آزمایش

به‌منظور آزمایش روش پیشنهادی مجموعه مقالات روزنامه همشهری در ۵ دسته مختلف از سایت همشهری[۱] گردآوری شده است. اطلاعات کلی این مقالات در شکل (۹) ارائه شده است.

مجموعه مقالات به‌گونه‌ای انتخاب شده‌اند که دایره وسیعی از کلمات مربوط به هر حوزه را پوشش دهند. به عبارت بهتر، از هر دسته مقالاتی با نویسندگان مختلف انتخاب شده است که از دایره‌ی لغت‌های مختلفی برای نوشتن مقالات استفاده شود. به این ترتیب، توانایی راه‌کار پیشنهادی در شناسایی متون مختلف با کلمات گوناگون اما در یک حوزهٔ مشخص بهتر نشان داده می‌شود.

[1] http://www.hamshahrionline.ir

(جدول- ۳): وزن‌دهی به کلمات متن شکل (۷)

شماره	کلمه	آدرس کلمه مرجح	وزن	مجموع
۱	آب های سطحی	۰	۱+۱+۰.۲۵+۰.۲۵+۰.۲۵+۱+۱	۴.۷۵
۲	آب های جاری (اخص)	۱	۰	۰
۳	آب گیرها	۱	۰	۰
۴	اقیانوس ها	۱	۰	۰
۵	باتلاق ها	۱	۰	۰
۶	تالاب ها	۱	۰	۰
۷	دریاچه ها	۱	۰	۰
۸	دریاها	۱	۰	۰
۹	رودخانه ها	۱	۰	۰
۱۰	سیلاب ها	۱	۰	۰
۱۱	مرداب ها	۱	۰	۰
۱۲	آب گیرها	۱	۰	۰
۱۳	اقیانوس ها	۱	۰	۰
۱۴	منابع آب (اعم)	۱	۰	۰
۱۵	آب شناسی	۱	۰	۰
۱۶	هیدرولوژی	۱	۰	۰
۱۷	رودخانه	۰	۱+۰.۲۵+۰.۲۵+۱	۲.۵
۱۸	آب های سطحی (۱/۴)	۱۷	۰	۰
۱۹	فرورفتگی های زمین	۱۷	۰	۰
۲۰	جویبارها(۱/۸)	۱۷	۰	۰
۲۱	جوی ها(۱/۸)	۱۷	۰	۰
۲۲	رودها (۱/۸)	۱۷	۰	۰
۲۳	نهرها (۱/۸)	۱۷	۰	۰
۲۴	شط ها (۱/۸)	۱۷	۰	۰
۲۵	دریاچه ها	۰	۱+۰.۲۵+۰.۲۵+۱	۲.۵
۲۶	آب های سطحی	۲۵	۰	۰
۲۷	فرورفتگی های زمین	۲۵	۰	۰
۲۸	تالاب ها	۰	۰.۲۵+۰.۲۵+۱+۱	۲.۵
۲۹	آب های سطحی	۲۸	۰	۰
۳۰	فرورفتگی های زمین	۲۸	۰	۰
۳۱	طبیعی	۰	۱	۱
۳۲	بارش	۰	۱	۱
۳۳	ریخت شناسی زمین	۳۲	۰	۰
۳۴	برف	۰	۱+۰.۲۵	۱.۲۵
۳۵	رطوبت	۳۴	۰	۰
۳۶	باران	۰	۱	۱
۳۷	رطوبت	۳۵	۰	۰
۳۸	تامین	۰	۱	۱
۳۹	مهندسی آب	۳۸	۰	۰
۴۰	ورود	۰	۱	۱
۴۱	تبخیر	۰	۱	۱
۴۲	نفوذ	۰	۱	۱
۴۳	عمقی	۰	۱	۱
۴۴	سفره	۰	۱	۱
۴۵	آب های زیرزمینی	۰	۱	۱
۴۶	چاه ها	۴۵	۰	۰
۴۷	آتشفشان ها	۴۵	۰	۰
۴۸	دریاچه های زیرزمینی	۴۵	۰	۰
۴۹	چشمه ها	۴۵	۰	۰
۵۰	آب های معدنی	۴۵	۰	۰
۵۱	لجنزارها	۰	۱	۱
۵۲	اکوسیستمها	۰	۱	۱
۵۳	آب	۰	۱	۱
۵۴	پاک	۰	۱	۱
۵۵	حفظ	۰	۱	۱
۵۶	تنوع زیستی	۰	۱	۱
۵۷	کاهش	۰	۱	۱
۵۸	تغییرات	۰	۱	۱
۵۹	آب و هوایی	۰	۱	۱
۶۰	کنترل	۰	۱	۱
۶۱	سیلاب	۰	۱	۱
۶۲	آب های سطحی	۶۱	۰	۰
۶۲	منابع	۰	۱	۱

ردیف	رده‌بندی موضوعی	تعداد مقالات	متوسط تعداد کلمات مقالات
۱	ورزشی	۱۴۶	۲۰۴
۲	اقتصادی	۱۵۴	۱۹۹
۳	شهری	۱۷۱	۱۲۳
۴	حوادث	۸۹	۱۶۰
۵	خارجی	۱۳۰	۱۷۷

(شکل- ۹): اطلاعات اولیه متون

۵ـ۲ـ نتایج حاصل از استخراج کلمات اصلی

همان‌گونه که بیان شد، روش پیشنهادی با دریافت یک متن فارسی، نخست اقدام به استخراج کلمات اصلی از طریق حذف کلمات اضافی و ریشه‌یابی می‌کند. در شکل (۱۰) برای هر دسته موضوعی نشان داده‌ایم که عملیات پیش‌پردازش متن به‌طور متوسط چه حجمی از کلمات را بیرون ریخته و چه درصدی را به‌عنوان کلمات اصلی به کاربر ارائه می‌کند.

ردیف	دسته بندی موضوعی	متوسط تعداد کلمات مقالات	متوسط تعداد کلمات پس از عملیات حذف و ریشه‌یابی
۱	ورزشی	۲۰۴	۱۴۹
۲	اقتصادی	۱۹۹	۱۳۵
۳	شهری	۱۲۳	۷۶
۴	حوادث	۱۶۰	۱۱۵
۵	خارجی	۱۷۷	۱۲۴

(شکل- ۱۰): فایل‌های پالایش شده

پس از استخراج کلمات اصلی نوبت به عملیات وزن‌دهی می‌رسد که در این مرحله وزن هر یک ازکلمات مشخص شده است. با آزمایش‌های انجام‌شده حد آستانه ۳ و ۴ برای انتخاب کلمات کلیدی از میان کلمات اصلی مناسب به نظر رسیده است. به این ترتیب کلمات اصلی موجود در متن که وزن آن‌ها بیش از حد آستانه باشد، به‌عنوان کلمه کلیدی هر متن برگردانده شده است. در این مرحله به‌طور متوسط بین ۵ تا ۱۰ کلمه از هر متن به‌عنوان کلمه کلیدی استخراج می‌شود. در برخی موارد این تعداد تنها یک کلمه و در مواردی بیش از ۲۰ کلمه بوده است.

به منظور آزمایش روش پیش‌پردازش پیشنهادی از هر دسته موضوعی در حدود دویست متن به‌طور تصادفی انتخاب و بررسی شد که روش پیشنهادی در چند درصد موارد کلمه اصلی را به‌اشتباه حذف کرده و در مواردی کلمهٔ اضافی به‌اشتباه به‌عنوان کلمه اصلی به کاربر ارائه شده است.

این بررسی به‌صورت دستی انجام شده است. نتایج حاصل از این آزمایش در شکل (۱۱) ارائه شده است.

ردیف	رده‌بندی موضوعی	درصد کلمات اضافی اشتباه	درصد کلمات اصلی اشتباها حذف شده
۱	ورزشی	٪۲۰	٪۹
۲	اقتصادی	٪۱۸	٪۱۱
۳	شهری	٪۲۲	٪۱۲
۴	حوادث	٪۱۶	٪۱۴
۵	خارجی	٪۱۵	٪۱۰

شکل ۱۱: نتایج بررسی کلمات کلیدی چندین متن

۵ـ۳ـ خوشه‌بندی

پس از پیش‌پردازش اولیه بر روی مستندات، به‌ازای هر سند یک بردار ویژگی استخراج می‌شود. به‌طور مثال فرض کنید m متن داریم. با فرض داشتن n کلمه نماینده متون، بردارهای ویژگی آن متون، می‌تواند به شکل (۱۲) باشد.

	کلمه اول	کلمه دوم	کلمه سوم	...	کلمه n-ام
متن اول	۱	۰	۱		۰
متن دوم	۰	۱	۱		۱
...					
متن m-ام	۱	۰	۰		۰

(شکل- ۱۲): بردارهای ویژگی تعدادی متون فرضی

چنان‌که پیش‌تر گفته شده، سطر نخست از این ماتریس یک متن را نمایش می‌دهد. در این متن کلمه نخست و سوم، از کلمات کلیدی بوده‌اند. این بردارهای ویژگی به یک الگوریتم خوشه‌بندی ارائه می‌شود که در آن تعداد خوشه‌ها پنج در نظر گرفته شده است. این الگوریتم خوشه‌بندی در این‌جا، الگوریتم k-means، در نظر گرفته شده است. ابتدا مرکز هر خوشه به‌صورت تصادفی انتخاب می‌شود. در هر تکرار از حلقه الگوریتم k-means مرکز خوشه با میانگین داده‌های متعلق به آن خوشه به‌روز می‌شود. معیار فاصله مورد استفاده در خوشه‌بندی، معیار فاصله اقلیدسی است. حلقه الگوریتم برای خوشه‌ها تا زمانی‌که مرکز برای خوش‌ها ثابت شود ادامه می‌یابد. جدول (۴) ماتریس تداخل را برای حالتی که از اصطلاح نامه و ریشه‌یابی استفاده نمی‌شود، نشان می‌دهد. جدول (۵) ماتریس تداخل برای حالتی را نشان می‌دهد که از اصلاح‌نامه استفاده نشود؛ ولی از ریشه‌یابی استفاده شود. جدول (۶) ماتریس تداخل را برای حالتی نشان می‌دهد که هم از اصطلاح‌نامه و هم از ریشه‌یابی استفاده شده است.

(جدول- ۴): ماتریس تداخل برای حالتی که از اصطلاح‌نامه و ریشه‌یابی استفاده نشود

خوشه	ورزشی	اقتصادی	شهری	حوادث	خارجی	Entropy	Purity
۱	۳۱	۱۶	۵۷	۱۹	۱۲	۱/۴۵	۴۲/۲۲
۲	۵	۴	۹	۹۷	۱۸	۰/۹۱	۷۲/۹۳
۳	۱۰۱	۲۷	۹	۵	۳	۰/۹۳	۶۹/۶۶
۴	۱۱	۲۳	۸	۱۱	۸۶	۱/۱۶	۶۱/۸۷
۵	۲۳	۸۴	۶	۱۴	۱۱	۱/۱۷	۶۰/۸۷
جمع	۱۷۱	۱۵۴	۸۹	۱۴۶	۱۳۰	۱/۱۲	۶۱/۵۹

(جدول- ۵): ماتریس تداخل برای حالتی که از اصطلاح‌نامه استفاده نشود ولی از ریشه‌یابی استفاده شود

خوشه	ورزشی	اقتصادی	شهری	حوادث	خارجی	Entropy	Purity
۱	۱۲	۲۱	۸	۹	۸۸	۱/۱۳	۶۳/۷۷
۲	۴	۲	۸	۱۰۲	۱۸	۰/۸۱	۷۶/۱۲
۳	۱۰۶	۲۷	۹	۳	۳	۰/۸۸	۷۱/۶۲
۴	۲۱	۹۱	۴	۱۳	۱۱	۱/۰۹	۶۵/۰۰
۵	۲۸	۱۳	۶۰	۱۹	۱۰	۱/۴۰	۴۶/۱۵
جمع	۱۷۱	۱۵۴	۸۹	۱۴۶	۱۳۰	۱/۰۶	۶۴/۷۸

(جدول- ۶): ماتریس تداخل برای حالتی که از اصطلاح‌نامه و ریشه‌یابی استفاده شود

خوشه	ورزشی	اقتصادی	شهری	حوادث	خارجی	Entropy	Purity
۱	۱۶	۱۰۷	۴	۸	۱۰	۰/۹۱	۷۳/۷۹
۲	۱۱	۱۵	۷	۶	۹۲	۱/۰۰	۷۰/۲۳
۳	۱۲۰	۲۰	۷	۳	۳	۰/۷۵	۷۸/۴۳
۴	۳	۱	۷	۱۱۸	۱۶	۰/۶۷	۸۱/۳۸
۵	۲۱	۱۱	۶۴	۱۱	۹	۱/۲۸	۵۵/۱۷
جمع	۱۷۱	۱۵۴	۸۹	۱۴۶	۱۳۰	۰/۹۱	۷۲/۶۱

در مرحله بعد برای محاسبه دقت هر خوشه از الگوریتم مجارستانی استفاده شده است.الگوریتم مجارستانی یک الگوریتم بهینه‌سازی است که به‌صورت یک ماتریس n * n است. به‌طور مثال اگر در این ماتریس درایه سطر i-ام و ستون j-ام، هزینه انجام j-امین کار توسط i-امین فرد باشد، در این‌صورت در این الگوریتم باید تقسیم وظایف بین

افراد را به‌گونه‌ای انجام داد که مجموع هزینه‌های افراد حداقل شود. در روش پیشنهادی این مقاله یک ردهٔ اصلی برای هر دسته در الگوریتم مجارستانی وجود دارد. رده‌های به‌دست آمده به این دسته اصلی ارائه می‌شوند؛ سپس محاسبه فراخوانی[1] و دقت[2] انجام می‌گیرد؛ و براساس این دو، معیار فیشر[3] محاسبه می‌شود.

$$Precision = \frac{TP}{TP + FP} \quad (۱)$$

$$Recall = \frac{TP}{TP + FN} \quad (۲)$$

$$F - Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (۳)$$

به‌منظور بررسی صحت عملکرد خوشه‌بندی از معیار فیشر و اطلاعات متقابل هنجارسازی‌شده[4] که معیارهای ترکیبی است استفاده می‌شود. این معیارها برای دو وضعیت استفاده از اصطلاح‌نامه و عدم استفاده از اصطلاح‌نامه در جدول (۷) ارائه شده است.

۶ـ جمع‌بندی و کارهای آینده

در این مقاله نکات زیر مورد توجه قرار گرفته است:
کارهای انجام‌شده در حوزهٔ پالایش متون فارسی مورد بررسی قرار گرفت. با استفاده از اصطلاح‌نامه وابسته در متون فارسی شامل هم‌معنی‌ها، کلمات اعم و اخص و وابسته‌ها شناسایی شدند. با استفاده از ارتباط بین کلمات موجود در متن با استفاده از اصطلاح‌نامه، وزن‌دهی دقیق انجام شده است و به این ترتیب کلمات کلیدی با دقت بیشتری استخراج شده‌اند. با تبدیل هر متن به یک فضای برداری که ابعاد آن کلمات کلیدی همه متون هستند، عمل خوشه‌بندی برای رده‌بندی متون بر اساس کلمات کلیدی استخراج‌شده انجام شد. نتایج حاصل از به‌کارگیری اصطلاح‌نامه در استخراج کلمات کلیدی متون فارسی بر خوشه‌بندی آن‌ها با حالت بدون اصطلاح‌نامه مقایسه شده است. نتایج بررسی‌شده نشان داده است که استفاده از اصطلاح‌نامه می‌تواند به رده‌بندی دقیق‌تر متون فارسی کمک کند.

(جدول- ۷): نتایج تجربی

روش	دقت	F-measure	NMI	Entropy	Purity
روش نزدیکترین همسایه وزن‌دار پیشنهادی یغمایی و تعبدی (۱۳۹۱)	%۶۷/۳۸	%۶۵/۰۶	%۱۶/۹۸	۱/۰۸	%۶۴/۸۷
روش k-means+SVD پیشنهادی علاقه‌بند و همکاران (۱۳۹۱)	%۶۷/۶۴	%۶۵/۴۵	%۱۷/۳۱	۱/۰۳	%۶۵/۰۹
روش آراسته و همکاران (Arasteh et. al, 2012)	%۶۴/۶۳	%۶۴/۹۳	%۱۵/۸۳	۱/۱۶	%۶۴/۳۳
عدم استفاده از اصطلاح‌نامه و ریشه‌یابی	%۶۱/۵۹	%۶۲/۰۱	%۱۴/۵۴	۱/۱۲	%۶۱/۵۹
استفاده از ریشه‌یابی عدم استفاده از اصطلاح‌نامه	%۶۴/۷۸	%۶۵/۱۴	%۱۶/۶۵	۱/۰۶	%۶۴/۷۸
استفاده از اصطلاح‌نامه و ریشه‌یابی	%۷۲/۶۱	%۷۲/۸۳	%۲۱/۳۸	۰/۹۱	%۷۲/۶۱

موارد زیر را می‌توان در کارهای آتی اعمال کرد: (الف) بهبود مرحله پیش‌پردازش، (ب) استفاده از روش خوشه‌بندی دیگر، و (ج) ابهام‌زدایی از کلمات.

۷- مراجع

بی‌جن‌خان، م. ، ۱۳۸۳. نقش پیکره‌های زبانی در نوشتن دستور زبان: معرفی یک نرم‌افزار رایانه‌ای، مجله زبان‌شناسی، سال ۱۹، شماره ۲.

حری، ع.، ۱۳۸۳. راهنمای تهیه و گسترش اصطلاح‌نامه یک زبانه، مرکز اسناد و مدارک علمی.

خسروی، ف.، ۱۳۷۹. اصطلاحنامه فرهنگی فارسی 'اصفا'، کتابخانه ملی جمهوری اسلامی ایران.

علاقه‌بند م.ر.، سعیدی محمدی، م.ر.، دزفولیان، م.ح.، ۱۳۹۱. خوشه‌بندی متون مبتنی بر مرکز دسته با استفاده از روش SVD و بهره‌گیری از نقاط همسایگی، نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، شهریور.

[1] Recall
[2] Precision
[3] F-Measure
[4] Normalized Mutual Information

یغمایی، ف.، تعبدی، س.، ۱۳۹۱. بهبود دسته بندی متون فارسی در روش همسایگی وزن دار، نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، شهریور.

Arasteh, A., Elahimanesh, M.H., Sharif, A., Minaei-Bidgoli, B., 2012. Semantically Clustering of Persian Words, The first international conference on Persian language processing, September.

Deegan, M., 2004. Keyword Extraction with Thesauri and Content Analysis, URL: http://www.rlg.org/en/page.php?Page_ID=17068

Faceli, K., Marcilio, C.P., 2006. *Multi-objective Clustering Ensemble*, Sixth International Conference on Hybrid Intelligent Systems (HIS'06), April.

Frank, E., 1999. Domain-Based Extraction of Technical Keyphrases, International Joint Conference on Artificial Intelligence, April.

Frantzi, K., Ananiadou, S., Mima, H., 2002. Automatic Recognition of Multi-word Terms: the C-value/NC-value Method, Digital Libraries, 3(2), pp. 115-130.

Freitas, N., Kaestner, A., 2005. Automatic text summarization using a machine learning approach, Brazilian Symposium on Artificial Intelligence (SBIA), April.

Hult, A., 2003. Improved automatic keyword extraction given more linguistic knowledge, Conference on Empirical Methods in Natural Language Processing (EMNLP), April.

Hyun, D., 2006. Automatic Keyword Extraction Using Category Correlation of Data, Heidelberg, pp. 224-230.

Jain, A. Murty, M.N., Flynn, P., 1999. *Data clustering: A review*, ACM Computing Surveys, 31(3), pp. 264–323.

Klein, M., Steenbergen, W.V., 2006. Thesaurus-based Retrieval of Case Law, International JURIX conference, April.

Liu, Y., Ciliax, B.J., Borges, K., Dasigi, V., Ram, A., Navathe S.B., Ingledine, R., 2005. Comparison of two schemes for automatic keyword extraction from MEDLINE for functional gene clustering, IEEE Computational Systems Bioinformatics Conference, April.

Maron, M.E., 1961. Automatic indexing: an experimental enquiry, Journal of the ACM, 8(1), pp. 404-417.

Martinez, J.L., 2008. Automatic Keyword Extraction for News Finder, Heidelberg, pp. 405-427.

Renz, I., 2003. Keyword Extraction for Text Characterization, Information Processing and Management, 31(5), pp. 226-237.

Romero, A., Nino, F., 2007. Keyword Extraction Using an Artificial Immune System, Information Retrieval, 5(2), pp. 216-231.

Salton, G., Hill, G., 1983. Introduction to Modern Information Retrival, MC Graw Hill, 1983.

Salton, G., Yang, C.S., 1973. On the specification of term values in automatic indexing, Journal of Documentation, 29, pp. 351-372.

Samiian, V., 1983. Origins of phrasal categories in Persian, an X-bar analysis, Ph.D dissertation, UCLA, 1983.

Strehl, A., Ghosh, J., 2002. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3, pp. 583–617.

Turney, P.D., 1999. Learning Algorithms for Keyphrase Extraction, Information Retrieval, 2(4), pp. 306-336.

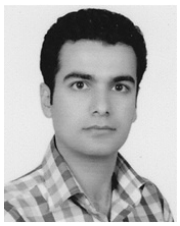
Witten, W., Medley, I.H., 2006. Thesaurus based autom-atic keyphrase indexing, ACM/IEEE-CS JCDL '06, April.

Zhang, Y., Heywood N.Z., Milios, E., 2006. World Wide Web Site Summarization Web Intelligence and Agent Systems, Technical Report, CS-2002-8.

حمید پروین کارشناسی خود را در سال ۱۳۸۵ از دانشگاه شهید چمران اهواز در زمینه مهندسی کامپیوتر گرایش نرم‌افزار اخذ کرد. وی کارشناسی ارشد و دکتری خود را در سال‌های ۱۳۸۷ و ۱۳۹۲ از دانشگاه علم و صنعت ایران در زمینه مهندسی کامپیوتر گرایش هوش مصنوعی اخذ کرد. تاکنون وی بیش از ۲۰ مقاله با نمایه SCIE (JCR) به چاپ رسانیده است. همچنین وی یک کتاب را ترجمه کرده است.

نشانی رایانامهٔ ایشان عبارت است از:

parvin@iust.ac.ir



فرهاد راد کارشناسی خود را در سال ۱۳۸۵ از دانشگاه شیراز در زمینه مهندسی کامپیوتر گرایش سخت‌افزار اخذ کرد. وی کارشناسی ارشد خود را در سال ۱۳۸۷ از دانشگاه علم و صنعت ایران در زمینه مهندسی کامپیوتر گرایش سخت‌افزار اخذ کرد. زمینه‌های پژوهشی وی بهینه‌سازی و شبکه بر روی تراشه است. وی هم اکنون دانشجوی دکترای واحد علوم و تحقیقات است.

نشانی رایانامهٔ ایشان عبارت است از:
f_rad@hotmail.com

آتوسا دهباشی کارشناسی خود را در سال ۱۳۸۵ از دانشگاه علم و صنعت ایران در زمینه مهندسی کامپیوتر گرایش سخت‌افزار و کارشناسی ارشد خود را در سال ۱۳۸۹ از دانشگاه علم و صنعت ایران در زمینه مهندسی کامپیوتر گرایش نرم‌افزار اخذ کرد. زمینه‌های پژوهشی وی متن‌کاوی است.

نشانی رایانامهٔ ایشان عبارت است از:
dahbashi_at@yahoo.com

بهروز مینایی بیدگلی دکترای خود را در رشته ی علوم و مهندسی کامپیوتر از دانشگاه ایالتی میشیگان آمریکا در سال ۱۳۸۴ گرفت. تخصص ایشان هوش مصنوعی و داده‌کاوی است و هم‌اکنون به‌عنوان عضو هیئت علمی دانشکدهٔ مهندسی کامپیوتر دانشگاه علم و صنعت به‌تدریس دروس هوش مصنوعی و نرم‌افزار مشغول هستند. ایشان سرپرستی گروه متن‌کاوی برای متون عربی و فارسی را در پژوهشکدهٔ داده‌کاوی نور نیز به عهده دارد. از سال ۱۳۸۶ ریاست بنیاد ملی بازی‌های رایانه ای بر عهدی ایشان است.

نشانی رایانامهٔ ایشان عبارت است از:
b_minaei@iust.ac.ir