متن کاوی مقالات رشته مهندسی کامپیوتر بر اساس مدارک بازیابی شده از پایگاه Web of Science

علی سلطانی نژاد[1]، محمد احمدی نیا[2]

مقاله پژوهشی

چکیده

مقدمه: این پژوهش باهدف بررسی مستندات رشته مهندسی کامپیوتر بازیابی شده از پایگاه Web of Science به‌منظور انجام خوشه‌بندی و متن کاوی آن‌ها صورت گرفته است.

روش: روش این پژوهش از نوع توصیفی است و رویکرد متن کاوی را مورد نظر قرار داده است. جامعه پژوهش، مدارک حوزه مهندسی کامپیوتر نمایه شده در پایگاه Web of Science بود که در بازه زمانی ۲۰۰۴ تا ۲۰۱۴، ۶۱۸۶ رکورد گزارش شد. داده‌های جمع‌آوری شده، با استفاده از نرم‌افزارهای HistCite و Excel نسخه ۲۰۱۳ و همچنین نرم‌افزار RapidMiner نسخه ۷/۳ تجزیه و تحلیل شدند.

یافته‌ها: برای خوشه‌بندی پس از پیش‌پردازش داده‌ها و اجرای الگوریتم خوشه‌بندی k-means، ۸ خوشه اصلی با عناوین اینترنت و فناوری، امنیت سیستم‌های اطلاعات سلامت، انسان و تعامل با رایانه، وب پنهان، مدل‌های کامپیوتری، عملکرد سیستم‌های کامپیوتری، شبکه‌ها و پایگاه‌های اطلاعاتی، الگوریتم‌ها و روش‌های کشف دانش و خوشه‌ای نیز با عنوان سایر موضوعات تشکیل شد. به منظور ارزیابی خوشه‌ها از دو معیار دقت و بازیافت استفاده شد و برای هر دو معیار عدد ۰/۸۱ به دست آمد.

بحث و نتیجه‌گیری: با توجه به نتایج به دست آمده، موضوعات پژوهشی در رشته مهندسی کامپیوتر و خلاءهای پژوهشی این حوزه شناخته شده است و به پژوهشگران کمک می‌کند تا کارهای تحقیقاتی خود را بر این اساس انجام دهند.

واژه‌های کلیدی: متن کاوی، خوشه‌بندی، الگوریتم k-means، پایگاه Web of Science

ارجاع: سلطانی نژاد علی، احمدی نیا محمد. متن کاوی مقالات رشته مهندسی کامپیوتر بر اساس مدارک بازیابی شده از پایگاه Web of Science . مجله دانشکده مدیریت و اطلاع‌رسانی پزشکی کرمان ۱۳۹۶؛۳(۲):۲۰۹-۲۰۱

دریافت مقاله:۹۶/۵/۱۳ پذیرش مقاله:۹۶/۶/۱۳

۱. دانشجوی کارشناس ارشد، کامپیوتر-نرم افزار، دانشگاه آزاد واحد کرمان، کرمان، ایران
۲. استادیار، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد کرمان، کرمان، ایران
آدرس: دانشگاه آزاد اسلامی واحد کرمان، دانشکده علوم پایه، گروه مهندسی کامپیوتر، کرمان، ایران

تلفن:۰۳۴۳۲۲۳۴۹۵۴
Email: ali.soltanie@yahoo.com

مقدمه

با گسترش فناوری اطلاعات و ارتباطات در جهان و ورود سریع آن به زندگی روزمره مردم، مسائل و ضرورت‌های تازه‌ای به وجود آمده است. انسان توسعه‌یافته کسی است که به اطلاعات دسترسی داشته باشد و دسترسی به اطلاعات نه یک ضرورت که یک قدرت محسوب می‌شود. در نتیجه تلاش برای استخراج اطلاعات از داده‌ها، توجه بسیاری از افراد دخیل در صنعت اطلاعات و حوزه‌های وابسته را به خود جلب نموده است. حجم بالای داده‌ها دائماً در حال رشد در همه حوزه‌ها و نیز تنوع آن‌ها به شکل داده‌های متنی، اعداد، گرافیک‌ها، نقشه‌ها، عکس‌ها، تصاویر ماهواره‌ای و عکس‌های گرفته‌شده با اشعه ایکس نمایانگر پیچیدگی کار تبدیل داده‌ها به اطلاعات است (۱).

روش های متعددی وجود دارد که به کمک آن‌ها می‌توان اطلاعات موجود در متون پژوهشی را استخراج کرد؛ یکی از این روش‌ها، داده‌کاوی است. داده‌کاوی، هنر و علم آنالیز هوشمندانه داده‌ها است و هدف آن پیدا کردن بینش و دانش نسبت به داده‌های پژوهش می‌باشد. با توجه به افزایش روزافزون داده‌ها، داده‌کاوی در اکثر حوزه‌های پژوهش ازجمله مدیریت، مهندسی، علوم پزشکی و کتابداری استفاده شده است (۲). به عبارتی داده‌کاوی یا استخراج دانش از پایگاه داده‌ها، فرایند مهم شناسایی الگوهای معتبر، جدید و قابل‌فهم در میان انبوهی از داده‌هاست (۳). داده‌هایی که مورد استفاده برای داده‌کاوی هستند همیشه ساخت‌یافته نیستند و انجام داده‌کاوی بر روی آن‌ها کار مشکلی است (۳،۴). به همین منظور، از متن‌کاوی برای این کار استفاده می‌شود. متن‌کاوی به تحلیل هوشمند متن، داده‌کاوی متنی یا کشف دانش در متن نیز مشهور است. به‌طور کلی به فرایند استخراج دانش و اطلاعات مورد علاقه و مهم از مجموعه متنی غیر ساخت‌یافته اشاره دارد (۵). در این‌باره Salloum و همکاران نیز تأکید می‌کنند که تکنیک‌های متن کاوی نقش مهمی در تبدیل متن بدون ساختار به دانش اطلاعاتی بازی می‌کنند (۶).

به‌منظور انجام عمل متن کاوی، لازم است از شاخص‌های مرتبط در این فرایند استفاده شود که یکی از این شاخص‌ها، خوشه‌بندی می‌باشد. خوشه‌بندی روشی است که برای گروه‌بندی موجودیت‌های (مدارک) مشابه مورد استفاده قرار می‌گیرد. در این روش، مدارک در گروه‌هایی از پیش تعیین نشده به نام خوشه قرار می‌گیرند، به‌طوری که مدارک مشابه در کنار یکدیگر و مدارک نامشابه دور از یکدیگر خواهند بود (۸-۷). تاکنون شیوه‌های مختلفی از خوشه‌بندی پیشنهاد شده و مورد استفاده قرارگرفته است. یکی از روش‌هایی که در خوشه‌بندی مورد استفاده قرار می‌گیرد، الگوریتم خوشه‌بندی k-means می‌باشد. الگوریتم خوشه‌بندی k-means برای خوشه‌بندی داده‌های نیمه ساخت‌یافته یا غیر ساخت‌یافته استفاده می‌شود و به خاطر سادگی و توانایی کار با داده‌های متراکم، به یکی از روش‌های خوشه‌بندی رایج تبدیل‌شده است. از آنجا که این روش در بسیاری از مطالعات مانند مطالعه ابوالصدق (۹)، راد و همکاران (۱۰)، کدخدایی و شمس (۱۱) مورد استفاده و آزمون قرارگرفته است، به همین علت روش خوشه‌بندی منتخب این پژوهش نیز، الگوریتم خوشه‌بندی k-means بود.

پس از متن‌کاوی و خوشه‌بندی، لازم است روش به‌کار گرفته‌شده برای متن‌کاوی، مورد ارزیابی قرار گیرد. مهم‌ترین بخش متن‌کاوی، اعتبارسنجی به‌منظور بازیابی اطلاعات و داده‌های متنی است که می‌توان آن را هدف نهایی این کار دانست. هدف اصلی در بازیابی اطلاعات، جستجوی متون نزدیک به پرس و جوی کاربر است (۹).

برای دستیابی به چکیده مقالات رشته مهندسی کامپوتر، نیاز به بررسی نمایه‌نامه‌ها و چکیده نامه‌های بین‌المللی است که به صورت پایگاه‌های اطلاعاتی دربرگیرنده بخش عمده‌ای از اطلاعات علمی سراسر دنیا می‌باشند. یکی از این پایگاه‌ها، پایگاه Web of Science می‌باشد (۱۲،۱۳).

این پایگاه گرچه از نوع پایگاه داده ساخت یافته می‌باشد و اطلاعات ذخیره‌شده در این پایگاه، بر مبنای اطلاعات کتاب‌شناختی مقالات، ساختار یافته‌اند اما داده‌های مورد استفاده در این پژوهش، متن چکیده‌های مقالات می‌باشند که به زبان طبیعی نوشته شده‌اند و در نوشتن آن‌ها هیچ‌گونه اعمال نظری از طرف پایگاه، صورت نگرفته است. از آنجا که داده‌های بدون ساختار متنی مانند مقالات، اسلایدهای آموزشی، اجزای ایمیل‌ها و... شامل انبوهی از اطلاعات به هم پیوسته هستند که در قالب جداول پایگاه‌های اطلاعاتی رابطه‌ای قرار نمی‌گیرند (۱۴)، بنابراین بهترین روش برای کشف روابط میان متون آن‌ها، متن‌کاوی است.

به‌منظور شناسایی نرم‌افزار مناسب برای عمل متن‌کاوی، مطالعه‌ای به اعتبارسنجی خوشه‌بندی برای متن‌کاوی بر روی اخبار منتخب یک سایت خبری پرداخته است و این نتیجه به دست آمد که نرم‌افزار RapidMiner می‌تواند نرم‌افزار مناسب برای انجام عمل متن‌کاوی باشد (۱۵). همچنین پژوهشی با عنوان خوشه‌بندی متون فارسی به کمک الگوریتم k-means و باهدف ایجاد الگوریتمی برای استخراج کلمات فارسی انجام شد که داده‌های مورد استفاده در آن شامل متون فارسی برگرفته‌شده از سایت‌های خبری موجود در وب بود؛ و پس از انجام مراحل متن‌کاوی، روش الگوریتم خوشه‌بندی k-means، روشی کارا و بهینه معرفی شد (۱۱). به علاوه کاربرد متن‌کاوی در بررسی ادبیات مهندسی صنایع؛ در قالب یک پایان‌نامه کارشناسی ارشد مورد مطالعه قرار گرفت که طی آن پس از کاربرد فنون مختلف جهت آماده‌سازی داده‌های متنی، به کمک خوشه‌بندی k-means، خوشه‌هایی ایجاد شد که توانایی تخصیص برچسب‌های موضوعی مجموعه مقالات مهندسی صنایع را داشتند اعتبارسنجی خوشه‌ها نیز با استفاده از دو معیار دقت و بازیافت صورت گرفت که معیار دقت برای خوشه‌های به وجود آمده برابر با ۷۰٪ بود (۹). پژوهشی با هدف ارائه روشی برای استخراج کلمات کلیدی در زبان فارسی، با روش پیش‌پردازش متون، وزن دهی و رتبه‌بندی، ایجاد و تشکیل بردار جایگاه تکرار کلمه و محاسبه شباهت میان آن‌ها به‌منظور دستیابی به نتایج بهبودیافته در این حوزه انجام شد و ۱۰۰ مقاله با موضوع‌های مختلف از کنفرانس و مجلات معتبر فارسی به‌صورت تصادفی انتخاب شدند. نتایج نشان داد که خطاهای زبان فارسی همچون نشانه‌گذاری‌ها، شیوه‌های متفاوت نوشتاری کلمات و... بر انجام عمل متن کاوی تأثیرگذار هستند (۱۶). همچنین پژوهش دیگری به استخراج اصطلاحات اصلی و هسته ار متون احادیث معتبر که نقش اساسی در زبان عربی دارند، پرداخت. پس از استخراج کلمات با استفاده از فنون متن‌کاوی، دو معیار دقت و بازیافت برای ارزیابی کلمات استخراج‌شده محاسبه شد که معیار دقت با ۸۳٪ نشان‌دهنده مطلوبیت روش استخراج واژگان کلیدی زبان عربی در محدوده خاصِ انتخاب‌شده بود (۱۷).

بررسی مطالعات پیشین حاکی از اهمیت استفاده از روش‌های متن‌کاوی در پالایش حجم زیاد اطلاعاتی است که به صورت متن وجود دارند؛ و با اجرای این فرآیند می‌توان به داده‌هایی دست یافت که قابلیت بررسی از دیدگاه‌های متعددی همچون میزان هم‌آیی، دسته‌بندی و... را در پژوهش‌های مختلف دارند. در نهایت اعتبارسنجی باهدف بازیابی اطلاعات که از چگونگی آماده‌سازی، ذخیره‌سازی، استخراج اطلاعات، تجزیه‌وتحلیل متون و خوشه‌بندی آن‌ها تأثیر می‌پذیرد، بازده عملکرد الگوریتم‌های متن‌کاوی را نیز مورد بررسی قرار خواهد داد. هدف اصلی از انجام تمام این فرایندها، کمک به کاربر برای جلوگیری از اتلاف وقت و همچنین بازیابی اطلاعات متنی نزدیک با پرس و جوی کاربر خواهد بود.

ازاین‌رو به لحاظ اهمیت موضوع و با توجه به این‌که تاکنون در مقالات مهندسی کامپیوتر، در این زمینه تحقیقی صورت نگرفته است، پژوهش حاضر به استفاده از فنون متن‌کاوی به استخراج و خوشه‌بندی مقالات علمی آن‌که در پایگاه استنادی Web of Science می‌پردازد.

روش

این پژوهش با استفاده از الگوریتم‌های متن‌کاوی انجام شد. بستر مناسب برای گردآوری داده‌ها، پایگاه Web of Science بود. برای انجام پژوهش، مدارک مربوط به مجلات حوزه مهندسی کامپیوتر با محدودیت زمانی ۲۰۰۴ تا ۲۰۱۴ در نوامبر ۲۰۱۵ جستجو شد. سپس ۶۱۸۶ رکورد بازیابی شده به صورت «فول رکورد» و فرمت «متن ساده» در فایل‌های ۵۰۰ رکوردی ذخیره و برای تحلیل بیشتر وارد نرم‌افزارهای HistCite و Excel و سپس RapidMiner شدند.

روش پیشنهادی برای انجام عمل متن کاوی، شامل چند قسمت است:

۱. پیش‌پردازش داده‌ها

لازم است داده‌ها آماده سازی شوند. روش آماده سازی متون، چند مرحله‌ای است که به آن پیش‌پردازش داده‌ها اطلاق می‌شود.

- ✓ حذف کلمات بدون بار معنایی

کلماتی که بار معنایی خاصی ندارند و به صورت معمول در سیاهه بازدارنده قرار می‌گیرند مانند، for، so، this و... از میان واژگان، حذف می‌شوند.

- ✓ ریشه‌یابی کلمات

برای ریشه‌یابی کلمات، لازم است از الگوریتم‌های ریشه‌یابی استفاده شود که الگوریتم پیشنهادی، الگوریتم پورتر می‌باشد. در این الگوریتم آنچه اهمیت دارد این است که کلمات بر اساس میزان مشابهت‌های معنایی ریشه‌یابی می‌شوند تا مشکل ریشه‌یابی مترادف‌ها به حداقل برسد (۹).

۲. وزن دهی به کلمات متن

پس از پیش‌پردازش داده‌ها، لازم است کلمات، شاخص گذاری شوند. روش پیشنهادی، استفاده از روش TF-IDF است. در این روش از فرمول زیر برای محاسبه وزن هر واژه استفاده می‌شود:

رابطه (۱): $$W_{ij} = TF(w_i d_j) \times \log\left(\frac{D}{DF(w_i)}\right)$$

که در آن $TF(w_i d_j)$ تعداد تکرار کلمه w_i در متن d_j ، $DF(w_i)$ تعداد متونی که دربرگیرنده کلمه w_i بوده، D تعداد متون مورد بررسی و W_{ij} نیز وزن واژه w_i در متن j است (۱۸).

۳. تجزیه و تحلیل داده‌های متنی

پس از انجام مراحل آماده سازی، پیش‌پردازش و وزن دهی کلمات، پایگاه داده پالایش‌شده‌ای داشتیم که شامل ماتریسی با ابعاد ۵۹۷۹ رکورد (تعداد چکیده‌ها) و ۲۰۷ فیلد (ویژگی یا لغات) بود. پس از تبدیل داده‌ها به فرمت قابل‌قبول و استفاده از عملگر خوشه‌بندیk-means ، عملیات پردازش خوشه‌بندی انجام پذیرفت. مطلوب‌ترین تعداد برای تعیین خوشه‌ها، عدد ۱۲ بود. به این معنا که ۱۲ سرخوشه اصلی تشکیل شود.

۴. ارزیابی نتایج

بازیابی اطلاعات و داده‌های متنی، مهم‌ترین قسمت فرایند متن‌کاوی است و می‌توان آن را هدف نهایی از انجام همه مراحل قبلی دانست. هدف اصلی در بازیابی اطلاعات، جستجوی متن نزدیک به پرس‌وجوی کاربر است. دو مورد معیاری که مورد استفاده قرار می‌گیرند عبارت‌اند از معیار بازیافت و معیار دقت که با رابطه‌های زیر قابل اندازه‌گیری است (۹):

دقت = تعـداد متون بازیابی/ تعداد متون مرتبط بازیابی‌شده

بازیافت= تعداد متون مرتبط /تعداد متون مرتبط بازیابی‌شده موجود

در بخش تجزیه و تحلیل داده‌ها، از سه نرم‌افزار HistCite، Excel 2013 و RapidMiner 3/7 استـفاده شد.

یافته‌ها

با توجه به خوشه‌های تشکیل‌شده، یکی از خوشه‌های حاصله شامل بیش‌ترین میزان مشاهدات (شامل ۳۹۷۹ داده یعنی ۶۰ درصد کل داده‌ها) بود. به همین منظور در مورد این خوشه مجدداً تمامی تنظیمات نرم‌افزار مانند قبل انجام گرفت، غیر از تعداد خوشه‌هایی که k=۱۰ در نظر گرفته شد. دلیل انجام این کار این بود که بتوانیم موضوعات ازدست‌رفته را از میان این سطح خوشه‌بندی به دست بیاوریم.

نتیجه خوشه‌بندی نرم‌افزار با مقدار k=۱۲، ۱۲ خوشه غیر تهی بود و از این میان خوشه شماره ۱۱، شامل داده‌های پرت بود که کنار گذاشته شد. ازآنجایی‌که تابه حال در زمینه کشف خوشه‌های مهندسی کامپیوتر، کار خوشه‌بندی انجام‌نشده است، لذا برچسب‌گذاری خوشه‌ها، توسط محقق برای اولین بار انجام شد. به‌منظور برچسب‌گذاری مناسب خوشه‌ها، دقت و تمرکز نام‌گذاری خوشه‌ها را با استفاده از دو فاکتور اولویت لغات درون خوشه‌ها (لغات با بیش‌ترین فراوانی در خوشه) و مقایسه میزان فراوانی لغات مرتبط به یکدیگر در هر خوشه مورد سنجش قرار داده شد. ۱۱ خوشه به این صورت نام‌گذاری شد که در جدول ۱ نمایش داده‌شده است.

جدول ۱: عناوین موضوعات پژوهشی به دست آمده از خوشه‌بندی k-means

ردیف	نام خوشه
خوشه ۰	Internet & Technology
خوشه ۱	Classify & Cluster Algorithm
خوشه ۲	Information Security & Health System
خوشه ۳	Visual Technology
خوشه ۴	Knowledge & Information Concepts
خوشه ۵	Other Subjects
خوشه ۶	Semantic Web
خوشه ۷	Computer Models & Integrated Systems
خوشه ۸	System & Program Performance
خوشه ۹	Compute & Mining
خوشه ۱۰	Database & Network Subjects

به‌منظور ارزیابی عناوین منتخب هر خوشه، به نمونه‌گیری ۱۰ داده (چکیده) در هر خوشه پرداخته شد. در نمونه‌های گرفته‌شده از هر خوشه ۴ فیلد مهم در هر نمونه و ارزیابی این مورد که آیا عنوان منتخب برای هر خوشه با ۴ فیلد مورد نظر هر نمونه مطابقت دارد یا خیر، مورد بررسی قرار گرفت. ۴ فاکتور مورد بررسی عبارت بودند از: نام مقاله، منبع مقاله، لغات کلیدی نویسنده و لغات کلیدی نمایه.

بعد از خوشه‌بندی ۱۰۰ نمونه آزمایش (در هر خوشه ۱۰ نمونه) کارایی ۸ خوشه مطالعات مهندسی کامپیوتر به اثبات رسید و ۸ خوشه فوق مطابق جدول ۲ نام‌گذاری شدند. لازم به ذکر است که خوشه شماره ۵ با نام Other subject در جدول ۱، به دلیل گستردگی داده‌ها آن از این گروه کنار گذاشته شد. تعداد و درصد مشاهدات مربوط به هر خوشه، در جدول ۳ قابل مشاهده است.

جدول ۲: عناوین موضوعات نهایی حاصل از خوشه‌بندی

ردیف	نام خوشه
خوشه ۱	Internet & Technology
خوشه ۲	Health Information Systems Security
خوشه ۳	Human-Computer Interaction
خوشه ۴	Semantic Web
خوشه ۵	Computer Models
خوشه ۶	Computer Systems Performance
خوشه ۷	Networks & Databases
خوشه ۸	Knowledge Discovery Algorithms & Methods
خوشه ۹	Other Subjects

در بخش خوشه‌بندی ثانویه، خوشه‌بندی داده‌های موجود در خوشه Other Subject انجام شد، زیرا داده‌های موجود در این خوشه، فراوانی بسیار داشت. پس از بررسی نمونه‌های حاصل از خوشه‌بندی دوم، این نتیجه حاصل شد که در خوشه‌های به دست آمده، فراوانی داده‌ها، تقریباً مشابه و نزدیک با یکدیگر بود و امکان اختصاص عنوان خاصی برای خوشه‌ها یا ادغام در خوشه های مرحله اول نبود. به همین سبب، داده‌های این خوشه‌ها، با همان عنوان Other Subject در میان خوشه‌ها قرار گرفت. به‌منظور اهمیت لغات با بیش‌ترین فراوانی در هر خوشه که در مرحله اول تصمیم‌گیری در مورد عناوین موضوعی ذکرشده، ۱۰ لغت به ترتیب بیش‌ترین فراوانی در هر خوشه به صورت جدول ۴ ارائه شده است.

جدول ۳: تعداد و درصد مشاهدات در هر خوشه

ردیف	نام خوشه	تعداد مشاهدات در هر خوشه	درصد مشاهدات اختصاص‌یافته به هر خوشه
۱	Internet & Technology	۶۲۵	٪۱۰
۲	Health Information Systems Security	۲۵۹	٪۴
۳	Human-Computer Interaction	۸۸	٪۲
۴	Semantic Web	۹۳	٪۲
۵	Computer Models	۴۹۷	٪۸
۶	Computer Systems Performance	۸۲	٪۱
۷	Networks & Databases	۲۶۹	٪۵
۸	Knowledge Discovery Algorithms & Methods	۸۵	٪۱
۹	Other Subjects	۳۹۷۹	٪۶۷

جدول ۴: ده لغت با بیش‌ترین فراوانی در هر خوشه

ردیف	نام خوشه	۱۰ لغت با بیش‌ترین فراوانی
خوشه ۱	Internet & Technology	web, e-learning, online, science, internet, information, project, technology, electric, concept
خوشه ۲	Health Information & Systems Security	Information, security, health, data, human, product, medicine, cognitive, technology, brain
خوشه ۳	Human-Computer Interaction	image, camera, vision, algorithm, digital, human, recognition, information, data, stereo
خوشه ۴	Semantic Web	ontology, semantic, domain, knowledge, information, web, concept, automatic, data, support
خوشه ۵	Computer Models	integrate, virtual, product, model, plan, system, cad, information, decision, data
خوشه ۶	Computer Systems Performance	program, parallel, computer, performance, matrix, algorithm, memory, data, parameter, code
خوشه ۷	Networks & Databases	data, database, information, access, sensor, network, structure, system, algorithm, pattern
خوشه ۸	Knowledge Discovery Algorithms & Methods	cluster, fuzzy, algorithm, data, context, information, knowledge, concept, class, mine
خوشه ۹	Other Subjects	system, information, simulation, control, algorithm, network, science, data, optimize, support

به‌منظور ارزیابی روش خوشه‌بندی از دو معیار دقت و بازیابی استفاده شد که همان‌گونه که مشاهده می‌شود هر دو معیار دقت و بازیافت در حد مطلوب می‌باشد که مناسب بودن میزان این دو معیار نشان‌دهنده مطلوب بودن خوشه‌بندی ارائه‌شده متون می‌باشد.

تعداد متون بازیابی شده/ تعداد متون مرتبط بازیابی شده = دقت
=۹/۱۱=۰/۸۱

تعداد متون مرتبط / تعداد متون مرتبط بازیابی شده = بازیافت
موجود =۹/۱۱=۰/۸۱

بحث و نتیجه گیری

با توجه به نتایج حاصل از خوشه‌بندی در مرحله اول، با ارزیابی‌های صورت گرفته و تکنیک برچسب‌گذاری ۹ خوشه از ۱۲ خوشه از پیش تعیین‌شده در فرایند متن کاوی، اینترنت و فناوری، امنیت سیستم‌های اطلاعات سلامت، انسان و تعامل با رایانه، وب پنهان، مدل‌های کامپیوتری، عملکرد سیستم‌های کامپیوتری، شبکه‌ها و پایگاه‌های اطلاعاتی و الگوریتم‌ها و روش‌های کشف دانش و سایر موضوعات بودند. در تحقیق ابوالصدق (۹) با استفاده از همین تکنیک، خوشه‌های به دست آمده در حوزه مهندسی برچسب‌گذاری و نام‌گذاری شدند.

در مرحله دوم اعتبار خوشه‌بندی‌ها با دو معیار دقت و بازیافت مورد سنجش واقع گردید. تجزیه‌وتحلیل‌های مندرج گویای این مطلب‌اند که الگوریتم k-means توانایی بالقوه در به دست آوردن الگوهای مناسب از متون تخصصی را دارا است. در این رابطه نتایج تحقیق Chernyshova و همکاران (۱۵)، لحبیب و همکاران (۱۷) و همچنین کدخدایی و شمس (۱۱) صحت این عملکرد را تأیید نموده‌اند. بررسی خوشه‌های تشکیل‌شده از موضوعات مطرح‌شده در چکیده مقالات و تعیین موضوعات غالب خوشه‌ها، به تعیین خلأهای پژوهشی این حوزه کمک می‌کند و نشان می‌دهد پژوهشگران این حوزه در چه زمینه‌هایی بیشتر کارکرده‌اند و چه حوزه‌هایی مورد غفلت واقع‌شده است. با تجمیع دستاوردهای ذکرشده به محققین این امکان داده می‌شود تا قبل از تدوین کامل یک مقاله و تنها با تدوین یک چکیده، با ساختار مقالات مختلف مجلات معتبر

در سال‌های مختلف در زمینه کاری خود آشنا شوند و در صورت تمایل به ثبت مقاله خود، مجله‌های معتبر منتشرکننده مقالات در زمینه تخصصی به آن‌ها معرفی می‌شود. کارایی دیگر این عمل ترغیب محققین به استفاده از خوشه‌بندی k-means در جهت کشف خوشه‌های مختلف علمی در متون تخصصی می‌باشد. از طرفی استفاده از کلماتی که به عنوان کلمات کلیدی در خوشه‌بندی‌ها انتخاب‌شده‌اند، می‌تواند به کاربر در صرفه‌جویی در وقت و بازیابی اطلاعات مرتبط کمک کند.

تشکر و قدردانی

بدین وسیله از گروه مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد کرمان و خانم عاطفه ذوالفقارنسب دانشجوی کارشناسی ارشد کتابداری و اطلاع رسانی پزشکی، سپاسگزاری می‌گردد. این مقاله مستخرج از پایان‌نامه کارشناسی ارشد می‌باشد.

Reference:

1. Afsa A. Data mining using fractional particle extraction algorithm. 3rd Annual National Conference on Mechanical Engineering and Industrial Solutions;2015 May 20; Arg Scientific and Research Center, Mashhad; 2015. [In Persian].
2. Hood WW, Wilson CS. The literature of bibliometrics, scientometrics, and informetrics. Scientometrics 2001; 52(2):291-314.
3. Ramezani H, Alipur Hafezie M, Momeni A. Scientific drawings: techniques and methods. Journal of the Popularization of Science 2014; 5(6)53-84. [In Persian].
4. Berry MW. Survey of Text Mining: Clustering, Classification, and Retrieval. Berlin: Springer; 2006.
5. Esmaili M, Zahed A. An overview of text mining: concepts, techniques and challenges. 3rd International Conference on Applied Research in Computer and IT;2016 Feb 4; Malek-Ashtar University of Technology,Tehran;2016. [In Persian].
6. Salloum SA, Al-Emran M, Abdallah S, Shaalan K. Analyzing the arab gulf newspapers using text mining techniques; 2018. p. 396-405.
7. Yan W, Zhang B, Ma SH, Yang ZY. A novel regularized concept factorization for document clustering. Knowledge-Based Syst 2017; 135:147-58.
8. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. Boston: Addison-Wesley Longman Publishing; 2013.p.487-568.
9. Abolsedgh S. Application of text mining in the study of industrial engineering literature [dissertation]. Yazd: Yazd University; 2011. [In Persian].
10. Rad F, Parvin H, Dehbashi A, Minaie B. Improved clustering persian text based on keyword using linguistic and thesaurus knowledge. Signal and Data Processing 2016; 13(1):87-100. [In Persian]
11. Kadkhodaei P, Shams A. Clustering of persian texts using the algorithm. 2th Extending Industrial Applications of Information, Communication and Computations (EIAICC2013 Conference); 2013 Oct 30-31; Tabriz. [In Persian].
12. Okhovati M, Sadeghi H, Talebian A, Baneshi M. Citation analysis and mapping Library & Information Science in WOS citation database 1993-2011. Journal of Epistemology (Library and Information Science And Information Technology) 2013; 6(21):9-26. [In Persian].
13. Saboury AL. Iran science production in 2008. Rahyaft 2008; 43:21-31. [In Persian].
14. Shettar R, Shobha G. Survey on mining in semi-structured data. International Journal of Computer Science and Network Security 2007; 7(8):226-31.
15. Chernyshova G, Smorodin G, Ovchinnikov A. Technique of cluster validity for text mining. 6th International Conference Cloud System and Big Data Engineering (Confluence); 2016 Jun 14-15; IEEE, Noida, India; 2016.
16. Maadi M, Fouladi K. Providing a method for extracting key words in Persian language. 2nd International Conference and Third National Conference on the Application of New Technologies in Engineering Sciences; 2016 Feb 25. Mashhad, Iran: Civilica. [In Persian].
17. Lahbib W, Bounhas I, Slimani Y. Arabic terminology extraction and enrichment based on domain-specific text mining. 27th International Conference on Tools with Artificial Intelligence; 2015 Nov 9-11; IEEE, Vietri sul Mare, Italy; 2015.
18. Teimourpour B, Sepehri MM, Pezesh L. A new method for intelligent categorization of Scientific texts (case of iran's nanotechnology papers). Journal of Science & Technology Policy 2009; 2(2):1-14. [In Persian].

Text Mining of Computer Engineering Articles Based on the Documents Retrieved from the Web of Science Database

Soltani Nejad A[1], Ahmadiniya M[2]

Original Article

Abstract

Introduction: The aim of this study was to evaluate text mining and clustering of computer engineering documents retrieved from the Web of Science database.
Methods: This is a descriptive-analytical study which was conducted in a survey method using text mining approach. The research community was all computer engineering documents indexed in the Web of Science, among which 6016 cases were reported between 2004 and 2016. The collected data were analyzed by HistCite software, Excel version 2013 and RapidMiner version 7.3.
Results: In order to perform clustering, after preprocessing the data and running K-means (a clustering algorithm), 8 main clusters were established. The clusters were Internet and Technology, Security of Healthcare Information Systems, Human-Computer Interaction, Semantic Web, Computer Models, Computer Systems Performance, Networks & Databases, Knowledge Discovery Algorithms and Other Topics. To evaluate the clusters, two criteria of precision and recall were used and a value of 0.81 was obtained for both criteria.
Discussion and Conclusion: Using words selected as keywords in the clustering can help the user save time and retrieve the related information.
Key words: Text mining, Clustering, K-means algorithm, Web of Science Database

Citation: Soltani Nejad A, Ahmadiniya M. Text Mining of Computer Engineering Articles Based on Documents Retrieved From the Web of Science Database. J Manage Med Inform Sch 2017;3(2):211-209.

Received:2017/08/4 Accepted:2017/09/4

1. MSc Student of Computer-Software, Azad University of Kerman, Kerman, Iran
2. Assistant Professor, Department of Computer Engineering, Islamic Azad University of Kerman, Kerman, Iran
Address: Islamic Azad University, Kerman Branch, Department Computer Engineering, Faculty of Sciences, Kerman, Iran
Phone: 03432234954
Email: ali.soltanie@yahoo.com