وزارت علوم، تحقیقات، و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)

طرح پژوهشی

ارائه روشی هوشمند برای استخراج کلیدواژه از مستندات علمی زبان فارسی براساس سیستم‌های پیشنهاددهنده

مجری

آزاده محبی

همکار

عمار جلالی منش

اسفند ۱۳۹۷


بسم الله الرحمن الرحیم

به نام خدا

وزارت علوم، تحقیقات و فناوری
پژوهشگاه علوم و فناوری اطلاعات ایران
ایرانداک

شماره: ۳۳/۱۵۰۳۲
تاریخ: ۱۳۹۷/۱۲/۲۱
پیوست: ندارد

پژوهش، مدیریت دانش، آموزش

گواهی

انجام طرح پژوهشی:

- **عنوان:** ارائه روشی هوشمند برای استخراج کلیدواژه از مستندات علمی زبان فارسی براساس سیستم‌های پیشنهاددهنده
- **مجری:** دکتر آزاده محبی
- **همکار:** دکتر عمار جلالی‌منش
- **گروه تحقیقاتی:** امیر بادامچی، فرنوش بیات ماکو
- **شمارهٔ قرارداد:** ۳۳/۳۰۴۲
- **تاریخ قرارداد:** ۱۳۹۶/۳/۱۰
- **تاریخ شروع:** ۱۳۹۶/۳/۱۰
- **تاریخ پایان:** ۱۳۹۷/۱۲/۱۸
- **ناظر:** دکتر جلال‌الدین نصیری

در پژوهشگاه علوم و فناوری اطلاعات ایران گواهی می‌شود. گزارش پایانی این طرح، بر پایهٔ داوری در نشست ۳۲۳ (تاریخ ۹۷/۱۲/۲۱) شورای پژوهشی پژوهشگاه به تصویب رسیده است. لازم می‌دانم مراتب قدردانی خود را به مجری، همکار، و ناظر این طرح تقدیم دارم.

با آرزوی بهروزی
پرویز شهریاری
معاون پژوهش و آموزش

تهران . خیابان انقلاب اسلامی . چهارراه فلسطین . شمارهٔ ۱۰۹۰ صندوق پستی ۱۳۷۱-۱۳۱۸۵
تلفن: ۶۶۴۹۴۹۸۰ - ۶۶۹۵۱۲۳۰ دورنگار: ۶۶۴۶۲۲۵۴ تلفن گویا: ۶۶۴۹۴۹۵۴
irandoc.ac.ir
info@irandoc.ac.ir

ارائه روشی هوشمند برای استخراج کلیدواژه از مستندات علمی زبان فارسی براساس سیستم‌های پیشنهاددهنده

مدیر طرح پژوهشی: آزاده محبی، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
نشانی: تهران، خیابان انقلاب، چهارراه فلسطین، شماره ۱۰۹۰، طبقه ۴، اتاق ۴۰۴
صندوق پستی: ۱۳۱۵۷۷۳۳۱۴ **تلفن:** ۶۶۴۹۴۹۸۰-۰۲۱(داخلی ۴۷۷)
وبگاه: https://www.irandoc.ac.ir
رایانامه: *mohebi@irandoc.ac.ir*

همکاراصلی طرح پژوهشی: عمار جلالی‌منش، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
گروه تحقیقاتی: امیر بادامچی، فرنوش بیات ماکو

چکیده

استخراج کلیدواژه یکی از مهمترین قدم‌های نمایه‌سازی مستندات محسوب می‌شود. کلیدواژه‌های یک سند، توصیفگرهای مفهومی هستند که می‌توانند در جستجو و بازیابی اطلاعات و نیز اشاعه آنها بکارگرفته شوند. در پایگاه‌های دربردارنده اسناد علمی مانند پایگاه علمی گنج پژوهشگاه علوم و فناوری اطلاعات ایران، کلیدواژه‌ها نقش مهمتری دارند، و تخصیص کلیدواژه‌های تخصصی نیز چالش‌برانگیزتر خواهد بود، زیرا در این پایگاه‌ها اسناد تخصصی با حوزه‌های علمی مختلفی وجود دارند. با توجه به افزایش حجم تولید و ثبت مستندات علمی، نیاز است که فرایند نمایه‌ساز و تخصیص کلیدواژه با سرعت بیشتری صورت گیرد و از روش‌های ماشینی هوشمند برای پیشنهاد و تخصیص کلیدواژه استفاده گردد. در بسیاری از پایگاه‌های اطلاعات علمی دنیا از روش‌های ماشینی و خودکار در کلیه فعالیت‌های فرایند نمایه‌سازی یا بخشی از آنها استفاده می‌شود. تعدادی از این روش‌ها بر مبنای تحلیل آماری متون و استفاده از روش‌های یادگیری ماشین هستند، تعدادی بر مبنای تحلیل معنایی متون به واسطه اصطلاح‌نامه‌های تخصصی و هستان‌شناسی، و در تعدادی دیگر از این روش‌ها از تلفیق هر دو استفاده می‌شود. بر همین اساس، در این طرح پژوهشی روشی برای پیشنهاد کلیدواژه به مستندات علمی فارسی ارائه شده که بر مبنای روش‌های هوشمند پردازش متن و یادگیری ماشین عمل می‌کند. روش پیشنهادی بر مبنای سیستم‌های پیشنهاددهنده و استدلال نمونه‌محور طراحی شده که براساس آن، مجموعه‌ای از کلیدواژه‌های مرتبط با یک سند به نمایه‌ساز پیشنهاد شود تا نمایه‌ساز سریعتر بتواند از بین آنها، کلیدواژه‌های مناسب را انتخاب کند. روش پیشنهادی براساس استدلال نمونه محور عمل می‌کند که در آن فرض بر این است که اسناد مشابه می‌توانند کلیدواژه‌های مشابه داشته باشند. بر همین اساس، ابتدا اسناد مشابه با یک سند جدید براساس روش‌های TFIDFو روش‌های بازنمایی کلمه-به-بردار، بازیابی می‌شوند. سپس کلیدواژه‌های کاندید از بین اسناد مشابه درنظر گرفته می‌شوند و در نهایت براساس یک تابع رتبه‌بندی، کلیدواژه‌های مناسب از بین آنها انتخاب می‌شوند. روش پیشنهادی بر روی مجموعه‌ای از اسناد پایگاه گنج در سه حوزه فنی و مهندسی، هنر و ادبیات، و علوم انسانی، پیاده‌سازی شده و نتایج آن با معیارهایی نظیر دقت، فراخوانی و نظرات متخصصین ارزیابی شده است.

کلیدواژه‌ها: استخراج کلیدواژه؛ سیستم‌های پیشنهاددهنده؛ استدلال نمونه محور؛ روش بازنمایی کلمه=به=بردار؛ بازیابی اطلاعات ؛ یادگیری ماشین.

فهرست مطالب

عنوان **شماره صفحه**

فهرست شکل‌ها

عنوان **شماره صفحه**

فهرست جدول‌ها

عنوان **شماره صفحه**

سپاسگزاری

در ابتدا خداوند منان را شاکرم که فرصت خدمت را برای بنده فراهم کرد.

از ناظر محترم، جناب آقای دکتر نصیری برای ارائه نظرات ارزشمند که در بهبود نتایج طرح موثر بوده، سپاسگزارم.

برخود لازم می‌دانم که از سرکار خانم دهسرایی، مدیر واحد سازماندهی و تحلیل اطلاعات ایرانداک، تشکر کنم که از ابتدای تعریف این طرح پژوهشی و در مرحله ارزیابی نتایج طرح، به عنوان متخصص حوزه نمایه‌سازی اسناد علمی، همکاری صمیمانه‌ای داشتند. علاوه بر آن، از همکاری صمیمانه خانم‌ها پیروزکیا، صادقیان، محمودیان وموسیوند، کارشناسان واحد سازماندهی و تحلیل اطلاعات برای ارائه نظرات تخصصی در ارزیابی نتایج طرح، سپاسگزارم.

همچنین از اعضای آزمایشگاه تعامل انسان و ماشین روبوداک، به خصوص خانم بیات ماکو، خانم جبل‌عاملی و آقای سلیمی برای مشارکت و همفکری در موضوعات فنی طراحی و برنامه‌نویسی تشکر می‌کنم.

فصل اول:

مقدمه و مساله پژوهش

۱. مقدمه و مساله پژوهش

نمایه‌سازی مستندات علمی فارسی به خصوص پایان نامه‌ها، در پژوهشگاه علوم و فناوری اطلاعات ایران مورد بررسی قرار گرفته است و در حال حاضر این فرایند توسط نیروی انسانی متخصص حاضر در پژوهشگاه انجام می‌شود. با توجه به افزایش چشمگیر تولید مستندات علمی، لازم است که این فرایند با دقت و سرعت بیشتری انجام پذیرد. در حال حاضر به طور متوسط هر نمایه‌ساز ۱۶/۱ سند را در روز نمایه می‌کند و با توجه به آنکه در حال حاضر ۱۳ نمایه‌ساز در ایرانداک فعالیت می‌کنند، به طور متوسط ۲۰۹/۳ سند در روز نمایه می‌شود. این رقم در حالیست که روزانه به طور متوسط ۲۰۰ عنوان پایان‌نامه در روز به ثبت می‌رسد. با توجه به افزایش مستمر نرخ تولید اسناد علمی در طول زمان، انتظار می‌رود که در چند سال آینده نرخ نمایه‌سازی در روز کمتر از نرخ اسناد ورودی باشد. علاوه بر ورود روزانه پایان‌نامه‌ها، هم اکنون حدود ۱۴۰ هزار عنوان پایان‌نامه از پیش باقی مانده که هنوز نمایه نشده‌اند. بنابراین با توجه به افزایش حجم مستندات علمی، لازم است که راهکاری برای افزایش سرعت نمایه‌سازی ارائه گردد.

امروزه الگوریتم‌ها و روش‌های متنوعی برای نمایه‌سازی خودکار مستندات علمی استفاده می‌شود. این روش‌ها بر مبنای ترکیبی از الگوریتم‌های یادگیری ماشین و پردازش معنایی متون، با استفاده از هستان شناسی و اصطلاحنامه‌های تخصصی هستند. یکی از بخش‌های اصلی فرایند نمایه‌سازی، که در پژوهشگاه انجام می‌شود، اختصاص کلیدواژه مناسب به یک سند است. این کلیدواژه‌ها می‌توانند از بین مجموعه‌ای از کلیدواژه‌های موجود (واژگان کنترل شده) انتخاب شوند، یا حتی خود کلیدواژه جدیدی باشند که براساس تخصص و تجربه نمایه‌ساز و اصطلاحنامه‌های تخصصی، تخصیص داده می‌شوند. با توجه به حجم بسیار وسیع کلیدواژه‌های کنترل شده در پایگاه گنج، انتخاب کلید واژه مناسب توسط نمایه‌ساز بسیار زمانبر خواهد بود. در این طرح پژوهشی، روشی پیشنهاد می‌شود که طی آن برای یک سند جدید، مجموعه‌ای بسیار محدودتر از واژه‌های کاندید برای نمایه‌سازی پیشنهاد می‌گردد تا نمایه‌ساز از بین این مجموعه محدود، کلیدواژه‌های مناسب را انتخاب کند. این مجموعه براساس داده‌های تاریخی نمایه شده

در گنج ایجاد خواهد شد. رویکرد اصلی در این طرح استفاده از الگوریتم‌های سیستم‌های پیشنهاددهنده برای پیشنهاد کلیدواژه مناسب به نمایه‌ساز است.

۱-۱. سیستم‌های پیشنهاددهنده و نمایه‌سازی متون

سیستم‌های پیشنهاددهنده جزو سیستم‌های پشتیبانی تصمیم هستند که از ابزارهای ماشینی و روش‌های خودکار استفاده می‌کنند تا اقلامی جهت استفاده به کاربر براساس مطلوبیت وی پیشنهاد دهند. «قلم» یا آیتم[۱] به چیزی گفته می‌شود که از طریق سیستم به کاربر پیشنهاد می‌گردد که بسته به کاربرد سیستم می‌تواند در قالب کالا یا خدمت باشد. سیستم‌های پیشنهاددهنده معمولا در حوزه‌هایی به کار گرفته می‌شوند که کاربر به دلیل عدم وجود دانش یا مهارت کافی، یا حتی زمان کافی، امکان پیدا کردن اقلام مطلوب خود را از یک مجموعه ندارد (Indurkhya & Damerau 2010; Ricci et al. 2011).

به عنوان مثال، در سیستم‌های پیشنهاددهنده براساس رویکرد پالایش همکارانه[۲]، که یکی از پرکاربردترین انواع سیستم‌های پیشنهاددهنده است، هرکاربر، پروفایلی دارد که بیانگر اقلامی است که در گذشته از آنها استفاده کرده است. برای پیشنهاد اقلام جدید به یک کاربر، لازم است که تاریخچه علاقه‌مندی‌های آن کاربر براساس مواردی که در گذشته از آنها استفاده کرده، با تاریخچه علاقه‌مندی‌های کاربران دیگر مقایسه شود. براساس این مقایسه، کاربرانی که علاقه‌مندی‌های مشابه داشتند استخراج می‌شوند که به آنها شرکاء[۳] گفته می‌شود. سپس بررسی می‌شود که چه اقلامی در بین علاقه‌مندی‌های شرکاء موجود است که کاربر مورد نظر قبلا از آنها استفاده نکرده است، و این اقلام براساس میزان تشابه، به کاربر پیشنهاد می‌شود. نمود کاربرد این سیستم‌ها را می‌توانیم در وبسایت‌های فروش آنلاین محصولات مشاهده کنیم. در رویکردهای همکارانه معمولا از ماهیت و ویژگی‌های یک قلم برای پیشنهاد آن به کاربر استفاده نمی‌شود، بلکه براساس میزان شباهت علاقه‌مندی‌های کاربر با سایر کاربران، آیتم‌های جدید پیشنهاد می‌شود (Bridge et al. 2005).

بخشی از پیشینه پژوهش در زمینه بکارگیری سیستم‌های پیشنهاددهنده در فرایند نمایه‌سازی معطوف به استفاده از این سیستم‌ها در برچسب‌زنی[۴] متون است. البته در نمایه‌سازی معمولا براساس محتوی یک سند، کلیدواژه‌ها به آن اختصاص می‌یابد، حال آنکه در برچسب‌زنی، برچسب یک سند می‌تواند علاوه بر محتوی، توصیف ویژگی‌های غیرمحتویی آن مانند اندازه سند، فرمت یا ساختار آن هم باشد (Hedden

---

[۱] Item

[۲] Collaborative filtering

[۳] Partners

[۴] Tagging

(2014. لیکن در ادامه پژوهش‌هایی قید می‌شود که روی برچسب‌زنی براساس محتوی، یعنی همان نمایه‌سازی براساس کلیدواژه‌ها، متمرکز هستند.

برخی پژوهش‌ها در این خصوص روی روش‌های پیشنهاددهی برچسب که برای یک فرد شخصی‌سازی شده‌اند، متمرکز است. در اینگونه روش‌ها براساس علایق فرد و حتی ارتباطات اجتماعی وی در یک شبکه اجتماعی، برچسب‌ها برای یک متن پیشنهاد می‌شود. این روش‌ها در زمینه‌هایی نظیر برچسب‌زنی محتویات وبلاگ‌ها یا پیام‌های شبکه‌های اجتماعی استفاده می‌شود (Tuarob et al. 2013; Mishne 2006).

بخش دیگری از تحقیقات روی تحلیل محتوی یک سند برای کشف موضوعات پنهان داخل یک سند برای برچسب‌زنی آن متمرکز است. در این نوع از تحقیقات، با اتخاذ رویکرد «مدلسازی موضوع»[۱] از روش‌های «تحلیل معنایی نهفته»[۲] برای کشف موضوعات پنهان در یک متن استفاده می‌شود و برچسب‌ها براساس این روش‌ها پیشنهاد می‌گردند(Krestel et al. 2009).

علاوه بر رویکردهای فوق، توآروب و همکاران (Tuarob et al. 2015) روشی را برای پیشنهاد برچسب در نمایه‌سازی خودکار مستندات پیشنهاد دادند. وجه مشترک کار ایشان با پژوهش حاضر اینست که در هر دو فرایند نمایه‌سازی براساس اطلاعات فراداده‌ها انجام می‌شود و در هر دو از اطلاعات مستندات نمایه شده برای برچسب‌زنی یک سند جدید استفاده می‌شود. لیکن در پژوهش توآروب و همکاران، یک روش کاملا خودکار برای نمایه‌سازی پیشنهاد شده و بخش پیشنهاد برچسب به عنوان بخشی از آن فرایند درنظر گرفته شده است.

براساس اطلاعات نویسنده، در حوزه نمایه‌سازی مستندات فارسی با استفاده از سیستم‌های پیشنهاددهنده تاکنون پژوهشی صورت نگرفته است. لیکن در این گزارش تحقیقاتی که تاحدی با این پژوهش مرتبط است، ارائه می‌گردد. جلالی منش و همکاران (جلالی منش و همکاران، ۱۳۹۳)، یک مدل مفهومی را برای فرایند نمایه‌سازی ماشینی مستندات علمی ایرانداک ارائه نمودند که در آن زیرسیستم‌های مختلفی در نظر گرفته شده است. لیکن با توجه به اینکه تمرکز ایشان روی ارائه یک مفهومی براساس شناسایی زیرسیستم‌های لازم برای این منظور بوده، به روش‌ها و الگوریتم‌های استخراج کلیدواژه‌ها و چگونگی اختصاص آنها به اسناد اشاره‌ای نشده است.

در اکثر پژوهش‌هایی که در زمینه نمایه‌سازی خودکار مستندات فارسی صورت گرفته، معمولا یک فرایند پیش پردازش متن دیده می‌شود که از طریق آن واژه‌های کاندید برای نمایه آن سند، طی عملیات

---

[۱] Topic modeling

[۲] Latent Semantic Analysis

ریشه‌یابی و حذف ایست واژه[1] ، استخراج می‌گردند. سپس براساس ویژگی‌هایی نظیر فراوانی واژه‌ها یا فراوانی اسنادی که آن واژه را در بردارند، کلیدواژه‌ها از بین آنها تعیین می‌گردند (بشیری و همکاران،۱۳۸۴؛ تشکری و میبدی، ۱۳۸۲).

۲-۱. هدف و گام‌های پژوهش

هدف غایی از انجام این طرح افزایش سرعت در فرایند نمایه‌سازی مستندات علمی فارسی (پایان‌نامه‌ها و رساله‌های فارسی) است. این هدف با استفاده از اهداف فرعی زیر حاصل می‌شود:

- ارائه روشی برای محاسبه شباهت بین یک سند نمایه نشده با مجموعه‌ای از اسناد علمی نمایه شده، برای یافتن کلیدواژه‌های کاندید
- ارائه روشی مبتنی بر متن کاوی برای پیشنهاد کلیدواژه‌ها به نمایه سازان، به صورت خودکار برای نمایه‌سازی مستندات علمی

در این پژوهش از روش کتابخانه‌ای برای استخراج مطالعات و پژوهش‌های مرتبط استفاده می‌شود. در واقع ابتدا روش‌های موجود در حوزه سیستم‌های پیشنهاددهنده و رویکردهای معمول در این حوزه مطالعه می‌گردد. سپس روش‌های بازیابی متون مشابه یک متن و معیارهای مختلف سنجش میزان تشابه بررسی می‌شود. با مطالعه این روش‌ها، روش مناسب برای مساله موجود طراحی یا انتخاب می‌گردد. پس از آن با بررسی روش‌های موجود استخراج کلیدواژه، روش مناسب برای استخراج و پیشنهاد کلیدواژه از بین متون مشابه ارائه می‌گردد.

لازم به ذکر است که از آنجاییکه در روش پیشنهادی مجموعه‌ای از کلیدواژه‌ها پیشنهادی می‌شود، این روش به تنهایی نمی‌تواند جایگزین فرایند نمایه‌سازی شود، بلکه نتایج حاصل از این روش می‌تواند در کنار دانش نمایه‌ساز و یا سایر روش‌های معمول استخراج کلیدواژه که براساس استخراج کلیدواژه از متن کامل سند عمل می‌کنند، قرار گیرد.

روش پیشنهادی روی مجموعه‌ای محدود از داده‌های نمایه شده آزمایش و ارزیابی می‌گردد و نتایج آنها براساس معیارهایی نظیر دقت و بازخوانی گزارش می‌شود. بنابراین یکی از بخش‌های مهم این پژوهش آماده‌سازی مجموعه‌ای از داده‌های منتخب برای آزمایش و ارزیابی روش پیشنهادی است.

این پژوهش از دو گام اصلی زیرتشکیل شده است (شکل ۱-۱):

۱. مطالعه و بررسی ادبیات موضوع در زمینه سیستم‌های پیشنهاددهنده، شامل:

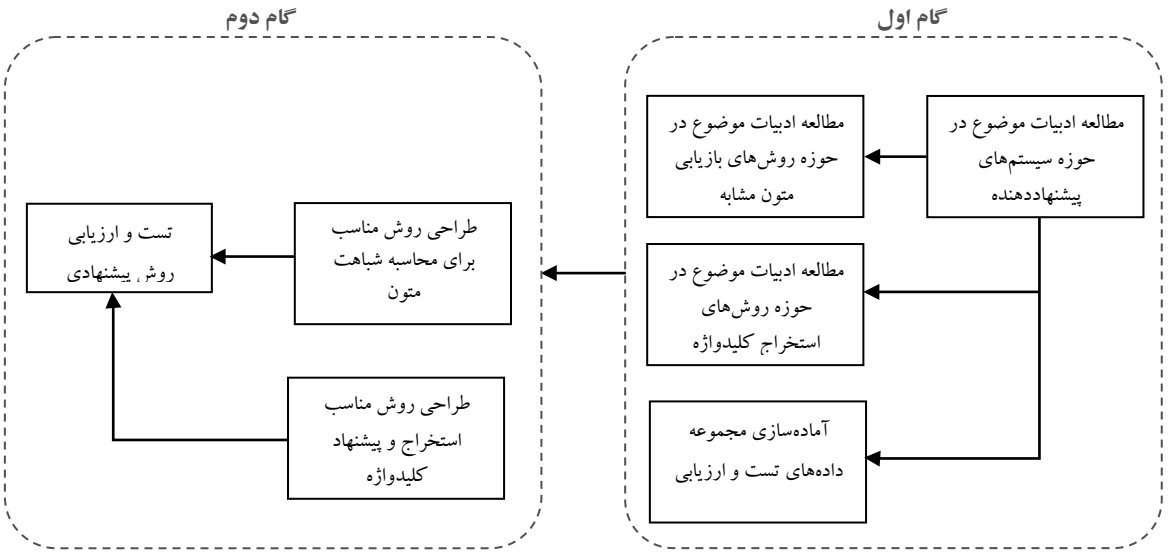
- شناسایی روش‌های بازیابی متون مشابه در یک پایگاه متنی، با تاکید بر روش‌های محاسبه شباهت متون

[1] Stop words

- مطالعه روش‌های استخراج کلیدواژه از بین متون بازیابی شده
- آماده سازی مجموعه داده‌های آزمایشی برای آزمایش روش پیشنهادی

۲. طراحی و اجرای روش پیشنهادی روی داده‌های آزمایشی

- طراحی روش مناسب برای محاسبه شباهت بین متون
- طراحی روش مناسب برای استخراج کلیدواژه‌های کاندید جهت پیشنهاد به نمایه‌ساز
- آزمایش و ارزیابی روش پیشنهادی روی مجموعه داده‌های آزمایشی



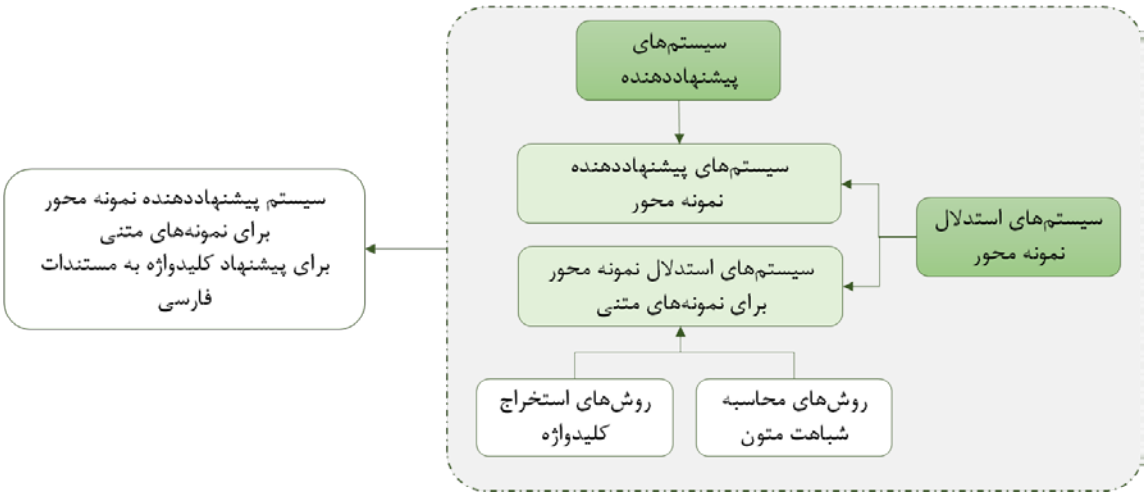
شکل ۱-۱- گام‌های اصلی پژوهش

فصل دوم:

مفاهیم و ادبیات موضوع

۲. مفاهیم و ادبیات موضوع

از آنجاییکه هدف نهایی این طرح ارائه روشی برای پیشنهاد کلیدواژه به مستندات علمی زبان فارسی است، حوزه‌های پژوهشی اصلی که در این فصل بررسی شده، در شکل ۲-۱ به صورت شماتیک آمده است.



شکل ۲-۱- حوزه‌های پژوهشی اصلی مطالعه شده در این طرح برای تهیه ادبیات موضوع

در مطالعه صورت گرفته در ابتدا مفاهیم سیستم‌های پیشنهاددهنده و انواع آنها معرفی شده‌اند. سیستم پیشنهاددهنده مورد بحث در این طرح پژوهشی، لازم است که از بین کلیدواژه‌های موجود در پایگاه داده اسناد برای پیشنهاد کلیدواژه جدید به یک سند جدید استفاده کند، از استدلال نمونه محور و سیستم‌های پیشنهاددهنده مبتنی بر این استدلال[۱] استفاده می‌شود. بنابراین در ادامه، مفاهیم مربوط به استدلال نمونه محور و سیستم‌های پیشنهاددهنده در این حوزه مرور شده است. علاوه بر آن، از آنجاییکه داده‌های سیستم پیشنهاددهنده در این طرح، داده‌های متنی هستند، برخی از مهمترین پژوهش‌ها در زمینه سیستم‌های پیشنهاددهنده متنی[۲] و سیستم‌های استدلال نمونه محور با نمونه‌های متنی[۱] نیز معرفی شده‌اند.

[۱] Case-Based Recommener Systems

[۲] Textual Recommender Systems

با توجه به آنکه روش‌های شباهت متون و اسناد، و روش‌های استخراج واژه از متون در سیستم‌های پیشنهاددهنده و استدلال نمونه محور با نمونه‌های متنی، نقش کلیدی را ایفا می‌کنند، در ادامه نیز پژوهش‌های اصلی مرور شده‌اند. همچنین انواع رویکردهای استخراج کلیدواژه از مستندات و برخی از مهمترین آنها تشریح شده است. در انتها نیز، رویکرهای بازنمایی برداری کلمات برای محاسبه شباهت بین متون توضیح داده شده است.

۱-۲. سیستم‌های پیشنهاددهنده

سیستم‌های پیشنهاددهنده، ابزارها و روش‌های نرم‌افزاری هستند که با ارائه اقلام[۲] مناسب (کالا، داده، اطلاعات و غیره)، کاربران خود را در فرآیند تصمیم‌گیری در حوزه‌های گوناگون حمایت می‌کنند(Ricci et al. 2011). این سیستم‌ها ترکیبی از فرایندهایی نظیر بازیابی اطلاعات، فیلترسازی، مدل‌سازی کاربر، یادگیری ماشین و تعامل انسان و کامپیوترهستند (Bridge et al. 2005) که می‌توانند راه‌حل‌های موثری را برای مقابله با افزونگی اطلاعاتی ارائه دهند. سیستم‌های پیشنهاددهنده در مقایسه با تحقیقات صورت گرفته در مورد سایر ابزار و روش‌های سیستم‌های اطلاعاتی سنتی (به عنوان مثال، پایگاه‌های داده یا موتورهای جستجو) موضوعات جدیدی هستند که اواسط دهه ۱۹۹۰ به عنوان یک حوزه تحقیقاتی مستقل به وجود آمدند و از محبوبیت فراوانی در حوزه تجارت الکترونیک برخوردار هستند. ارائه‌دهندگان خدمات به دلایل مختلفی، نظیر افزایش مقدار فروش، فروش اقلام متنوع، افزایش رضایت کاربر و درک بهتر نیازهای کاربران خود، از این سیستم‌ها استفاده می‌کنند. همچنین این سیستم‌ها در حوزه‌های گوناگون به منظور حمایت موثر از وظایف و اهداف کاربر مورد استفاده قرار می‌گیرند.هرلاکر و همکاران (Herlocker et al. 2004) در یک مقاله که مرجع کلاسیک در این زمینه شناخته شده است به معرفی یازده کاربرد اصلی سیستم‌های پیشنهاددهنده پرداخته‌اند که عبارتند از: حاشیه نویسی[۳] در متن، یافتن مجموعه‌ای از اقلام مناسب، یافتن تمام اقلام مناسب، توصیه یک دنباله معنی‌دار از اقلام، پیشنهاد یک بسته یکپارچه و مرتبط از اقلام، پشتیبانی در مرورو جستجوی اقلام، یافتن یک پیشنهاددهنده قابل اطمینان، ارتقا اطلاعات تخصصی کاربران، ارائه جمع‌آوری و ارائه نظرات افراد، استفاده از سیستم‌های پیشنهاددهنده برای کمک به سایرین، و اثرگذاری روی افراد. از بین کاربردهای ذکر شده برخی مستقیما برای پیشنهاد اقلام مناسب کاربر است، و برخی دیگر دیدگاه «فرصت طلبانه»[۴] را

---

[۱] Textual Case-Based Resoning Systems

[۲] Items

[۳] Annotation

[۴] Opportunistic

دربر دارد به این مفهوم که از داده‌های سیستم‌های پیشنهاددهنده و فواید جانبی آن بهره گرفته می‌شود (Ricci, Francesco;Rokach, Lior;Shapira, Bracha; Kantor 2011).

اکثر سیستم‌های پیشنهاددهنده دارای ویژگی های عملیاتی مشترک هستند. براساس این ویژگی‌ها می‌توان معیارهایی را برای ارزیابی عملکرد سیستم‌ها نیز تعریف نمود. این ویژگی‌ها عبارتند از (Bobadilla et al. 2013):

۱. مرتبط بودن[۱]: واضح‌ترین هدف عملی سیستم پیشنهاددهنده، ارائه مرتبط‌ترین اقلام با توجه به نیاز و ویژگی‌های کاربر است. اگر چه مرتبط بودن پیشنهادات مهمترین هدف اینگونه سیستم‌ها است اما نمی‌تواند به تنهایی کافی باشد.
۲. نوآوری[۲] : سیستم‌های پیشنهاددهنده می‌توانند با توصیه اقلامی جدید که قبلا توسط کاربر استفاده نشده مفید واقع شوند هستند.
۳. سرنتیپیتی[۳] : در این حالت سیستم اقلام غیرمنتظره‌ای را پیشنهاد می‌کند. تفاوت سرنتیپیتی و نوآوری در این است که اقلام پیشنهادی برای کاربر غیر منتظره هستند.
۴. تنوع[۴] : سیستم‌های توصیه شده به طور معمول مجموعه‌ای از بهترین اقلام را پیشنهاد می‌دهند و اگر همه این موارد توصیه شده خیلی شبیه باشند، احتمال اینکه هیچ یک از آنها مفید واقع نشود افزایش می‌یابد. از سوی دیگر، هنگامی که اقلام توصیه شده متنوع باشند، احتمال بیشتری وجود دارد که حداقل یکی از این موارد مطلوب باشد. تنوع برای اطمینان از این است که توصیه‌های مکرر اقلام مشابه به کاربر ارائه نگردد.

۱-۱-۲. منابع دانشی در سیستم‌های پیشنهاددهنده

به طور کلی چهار منبع اصلی برای برای تولید دانش در سیتسم‌های پیشنهاددهنده وجود دارد که پیشنهادها براساس آنها پیشنهاد می‌شود(Felfernig & Burke 2008):

۱. کاربر
۲. سایر کاربران همسان[۵]
۳. داده‌های موجود درباره اقلام

---

[۱] Relevance

[۲] Novelty

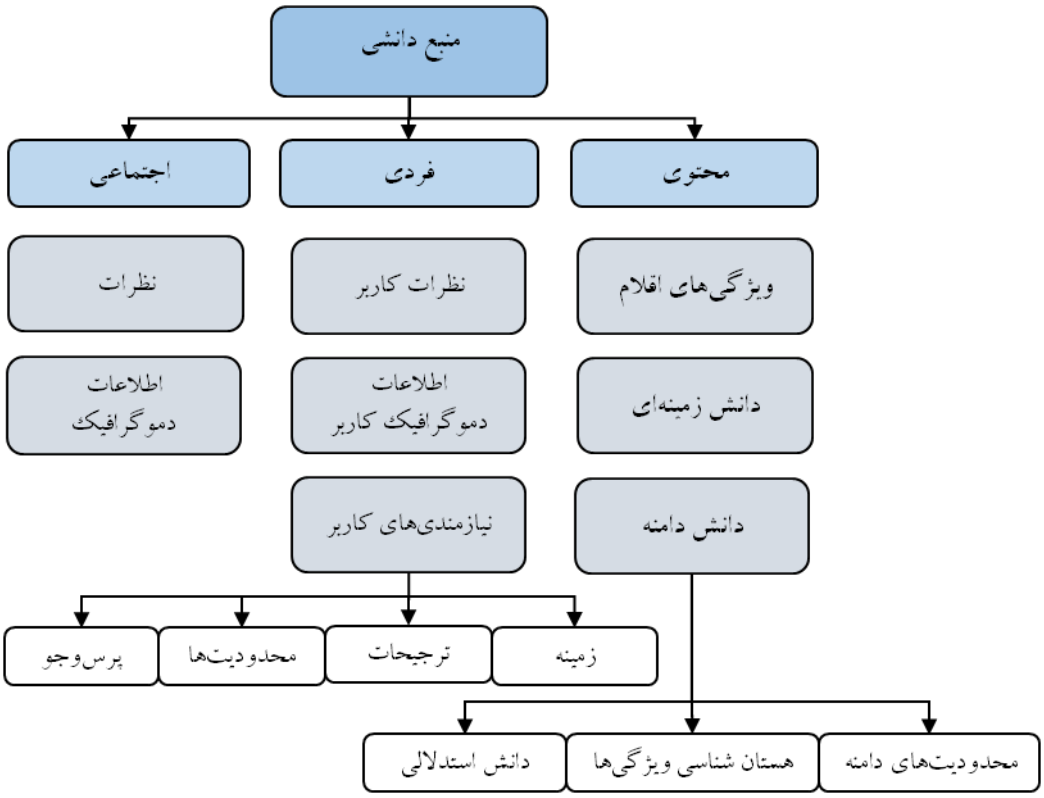
[۳] Serendipity

[۴] Diversity

[۵] Peer users

۴. دامنه و دانش زمینه‌ای و موضوعی سیستم

شمایی از این دسته‌بندی و زیرمجموعه‌های هر دسته درشکل ۲-۲ آمده است.



شکل ۲-۲- انواع منابع دانشی در سیستم‌های پیشنهاددهنده (Felfernig & Burke 2008)

در این دسته‌بندی دانش اجتماعی[۱]، همان دانش درباره کاربران سیستم است، مانند رتبه‌بندی‌های عددی موجود در پروفایل کاربران یا داده‌های دموگرافیک [۲] که به منظور نمایش شباهت میان کاربران و تکمیل رتبه‌بندی‌ها استفاده می‌شود. همچنین به منظور شخصی سازی پیشنهادها، اطلاعات فردی نیز درباره کاربر ضروری است. در واقع دسترسی به ترجیحات کاربر به منظور ارائه یک سری پیشنهادها می تواند کافی باشد اما در شرایطی که کاربر با یک درخواست و پیش زمینه خاص از سیستم استفاده می‌کند سیستم باید بتواند به آن پاسخگو باشد. در این حالت اطلاعات مورد نیاز در مورد خواسته کاربر از طریق پرسش[۳] و در نظر گرفتن یکسری محدودیت ها روی گزینه‌ها انجام خواهد شد. نهایتا، سیستم به منظور تطبیق دادن

[۱] Collaborative / social knowledge

[۲] Demographic data

[۳] Query

نیازهای کاربر و اقلام موجود، نیازمند دانش محتوا[۱] است که می‌تواند شامل مواردی نظیر ویژگی اقلام[۲]، دانش زمینه‌ای[۳] و دانش دامنه کاربرد سیستم باشد. در ادامه توضیحات بیشتری درباره هر یک ار منابع دانشی ارائه می‌گردد.

۲-۱-۱-۱. دانش کاربر

دانش مربوط به هر یک از کاربران در یک سیستم پیشنهاددهنده ضروری است. ارتباط بین کاربر و دانش مشارکتی می‌تواند کاملا دوسویه باشد، به این مفهوم که *نظر یک کاربر* یا *اطلاعات دموگرافیک* وی هنگامیکه قرار است پیشنهادی به وی داده شود، یک دانش فردی محسوب می‌شود، لیکن هنگامیکه از این اطلاعات برای پیشنهاد به کاربر دیگری استفاده می‌شود، آنگاه این اطلاعات از نوع اجتماعی است. اطلاعات تاریخی مربوط به ترجیحات کاربر در برخی از سیستم‌های پیشنهاددهنده ممکن است کافی باشد، لیکن در بسیاری از سیستم‌ها لازم است که از اطلاعات دیگری نیز استفاده شود تا *نیازمندی‌های کاربر* مشخص گردد. اطلاعات زیر می‌تواند علاوه بر ترجیحات کاربر مورد استفاده قرار گیرد(Felfernig & Burke 2008):

- پرس‌وجو: کاربر می‌تواند نیاز خود را در قالب یک عبارت پرس‌وجو به سیستم ارائه کند.
- محدودیت‌ها: در برخی از سیستم‌های پیشنهاددهنده، کاربر می‌تواند نیاز خود را در قالب مجموعه‌ای از محدودیت‌ها بیان کند. به عبارت دیگر مشخص کند که اقلام مطلوب وی باید حداقل دارای چه شرایطی باشند.
- ترجیحات: منظور از ترجیحات مواردی است که کاربر وجود آن ویژگی‌ها را رد اقلام پیشنهادی مطلوب می‌داند، نه اینکه الزامی برای رعایت آنها در اقفلام پیشنهادی باشد.
- زمینه: زمینه کاربر به معنی شرایط بیرونی مربوط به سیستم پیشنهاددهنده یا کاربر است، به عنوان مثال موقعیت مکانی کاربر در یک سیستم پیشنهاددهنده رستوران نقش کلیدی را ایفا می‌کند.

۲-۱-۱-۲. دانش اجتماعی

دانش اجتماعی اطلاعاتی که درباره سایر کاربران، از منظر یک کاربر که قرار است به وی پیشنهاد اقلام صورت بگیرد، وجود دارد. این اطلاعات عبارتند از رتبه‌بندی‌ها (نظرات)، و داده‌های دموگرافیک برای مشخص کردن شباهت بین کاربران. البته در پژوهش‌های کنونی، از اطلاعات موجود در شبکه‌های اجتماعی نظیر نظرات نوشتاری کاربران و برچسب‌زنی اجتماعی نیز استفاده می‌شود.

---

[۱] Content

[۲] Item attributes

[۳] Contextual knowledge

۲-۱-۱-۳. دانش محتوی

دانش محتوی طیف وسیعی را شامل می‌شود که در دسته‌بندی ارائه شده در شکل ۲-۲، چهار دسته اصلی برای آن درنظر گرفته شده است(Felfernig & Burke 2008):

- ویژگی‌های اقلام: منظور از ویژگی‌های اقلام، اطلاعاتی که ماهیت آنها را توصیف می‌کند.
- دانش زمینه‌ای: استفاده از دانش زمینه‌ای، نوع جدیدی از تحقیقات را در شسیستم‌های پیشنهاددهنده بوجود آورده که به آنها سیستم‌های پیشنهاددهنده زمینه-آگاه[1] می‌گویند(Adomavicius & Tuzhilin 2011). در این نوع از سیستم‌ها، دانش زمینه‌ای تعامل کاربر با سیستم نظیر موقعیت مکانی وی، زمان تعامل، شرایط محیطی،... به عنوان منبع دانشی متفاوت در ارائه پیشنهادها در نظر گرفته می‌شود.
- دانش دامنه: دانش دامنه اطلاعاتی را درباره نوع اقلام پیشنهادی، نوع کاربرانی که با سیستم تعامل می‌کنند، در بر دارد. این اطلاعات به روش‌ها و فرمت‌های مختلف زیر می‌توانند در سیستم وجود داشته باشند:
  - دانش استدلالی: دانش استدلالی این امکان را فراهم می‌کند که سیستم‌ها پینشهادهایی را ارائه دهد که از نظر منطقی با دامنه کاربرد سیستم سازگار باشد.
  - هستان شناسی ویژگی‌ها: اینگونه اطلاعات می‌تواند برای تعیین ارتباط معنایی بین ویژگی‌های اقلام و تعریف آنها مفید باشد.
  - محدودیت‌های دامنه: براساس دامنه کاربرد سیستم، ممکن است محدودیت‌هایی در ارائه برخی از اقلام به برخی کاربران وجود داشته باشد، به عنوان مثال در یک سیستم پیشنهاددهنده داروهای تقویتی، افرادی که دیابت دارند، نمی‌توانند از برخی از محصولات استفاده کنند.

۲-۱-۱-٤. انواع سیستم‌های پیشنهاددهنده

سیستم‌های پیشنهاددهنده را معمولا براساس روشی که برای پیشنهاد اقلام در آنها استفاده می‌شود، دسته‌بندی می‌کنند. روش‌های پیشنهاد اقلام نیز بر مبنای منبع دانش و اطلاعاتی که در سیستم‌ها برای پیشنهاد اقلام جدید وجود دارد، بکار گرفته می‌شوند. براساس یک دسته بندی کلاسیک، سیستم‌های پیشنهاد دهنده به پنج دسته اصلی تقسیم‌بندی می‌شوند (Burke 2007; Ricci, Francesco;Rokach, Lior;Shapira, Bracha; Kantor 2011):

[1] Context-aware

۱. مبتنی بر پالایش همکارانه: فرضیه اصلی رویکرد پالایش همکارانه این است که اگر شخص «الف» علایق مشابه شخص «ب» در یک مسئله داشته باشد، در یک مسئله جدید احتمال اینکه شخص «الف» نظری مشابه با «ب» داشته باشد نسبت به یک فرد تصادفی بیشتر است. بنابراین در این نوع از سیستم‌ها، اقلام تنها براساس رتبه‌بندی که سایر کاربران به آنها داده‌اند، پیشنهاد می‌شوند. بنابراین در این سیستم‌ها، برای پیشنهاد اقلام جدیدی به یک کاربر، تاریخچه رتبه‌بندی وی از اقلام با تاریخچه سایر کاربران در سیستم مقایسه می‌شود، و کاربرانی که رتبه‌بندی مشابهی با کاربر مفروض ارائه داده‌اند شناسایی می‌شوند. در نهایت پیشنهاد اقلام جدید براساس اقلام کاربران مشابه صورت می‌گیرد. روش‌های زیر معمولا برای شناسایی و پیشنهاد اقلام جدید استفاده می‌شود:

- روش‌های مبتنی بر حافظه [1]: این روش‌ها به عنوان مدل‌های پالایش همکارانه مبتنی بر همسایگی [2] نیز شناخته می‌شوند که در آن رتبه بندی ترکیبات کاربر و اقلام بر اساس همسایگی آنها پیش بینی می‌شود. از مزایای روش‌های مبتنی بر حافظه، کاربرد ساده آنها است. از مشکلات این روش‌ها این است که با ماتریس‌های رتبه بندی پراکنده[3] به خوبی کار نمی‌کنند. مدل‌های مبتنی بر همسایگی خود به دو دسته پالایش همکارانه مبتنی بر کاربر [4] و پالایش همکارانه مبتنی بر اقلام[5] تقسیم می‌شوند (Aggarwal 2016b).
- روش‌های مبتنی بر مدل[6]: در روش‌های مبتنی بر مدل، روش‌های یادگیری ماشین[7] و داده کاوی[8] به منظور پیش بینی مورد استفاده قرار می‌گیرند. برخی از نمونه‌های روش‌های مبتنی بر مدل عبارتند از درخت تصمیم‌گیری[9]، مدل های مبتنی بر قانون[10]، روش های بیزی[11] و مدل‌های فاکتور پنهان[12]. بسیاری از این مدل‌ها مانند فاکتور پنهان دارای پوشش بسیار خوب حتی برای ماتریس های رتبه بندی پراکنده هستند (Aggarwal 2016a).

---

[1] Memory-based methods
[2] Neighborhood based collaborative filtering
[3] Sparse Ratings Matrices
[4] User-based collaborative filtering
[5] Item-based collaborative filtering
[6] Model-based methods
[7] Machine Learning
[8] Data Mining
[9] Decision Trees
[10] Rule-based Models
[11] Bayesian Methods
[12] Latent Factor Models

۲. محتوا محور[1]: در این سیستم‌ها از ویژگی‌های توصیفی اقلام برای تهیه پیشنهادها استفاده می شود و واژه محتوا در واقع به همان توصیف‌ها اشاره دارد. در روش‌های مبتنی بر محتوا، رتبه بندی‌ها و رفتار کاربران با اطلاعات محتوایی اقلام ترکیب می‌شود. در این روش توصیف اقلامی که با رتبه‌بندی برچسب گذاری شده‌اند به عنوان داده آموزشی استفاده می‌شود. سیستم از دو نوع اطلاعات برای پیشنهاد استفاده می کند: ویژگی‌های اقلام و رتبه‌بندی‌هایی که کاربر به آن اقلام داده است. این سیستم‌ها، مساله پیشنهاددهی را در قالب ک مساله دسته‌بندی[2] در نظر می گیرند به گونه‌ای که این دسته‌بندی، می تواند اقلام مطلوب و نامطلوب (دسته‌بندی دو کلاسه) را برای کاربر براساس ویژگی‌های آنها مشخص نماید(Burke 2007). سیستم‌های مبتنی بر محتوا هنگامی که مقدار کافی داده‌های رتبه بندی در دسترس نباشند، بسیار کارا هستند. اگر چه سیستم‌های پیشنهاددهنده محتوا محور در پیشنهاد اقلام جدید نیز بسیار کارا هستند اما بازدهی خوبی در مواجه با کاربر جدید ندارند، زیرا در این حالت مدل یادگیری برای کاربر هدف نیاز به سابقه کاربر و رتبه‌بندی وی دارد. به همین دلیل دسترسی به حجم زیاد اطلاعات رتبه بندی در اینگونه روش‌های روی دقت و عملکرد سیستم تاثیر زیادی دارد. همچنین در اینگونه از سیستم‌ها، به دلیل استفاده از کلیدواژه‌ها و محتوا، در بسیاری از موارد پیشنهادها شبیه هم هستند و تنوع در پیشنهادها کم است. (Jure Leskovec, Jure, Rajaraman & Ullman 2014).

۳. دموگرافیک: در این سیستم‌ها، اطلاعات مربوط به ویژگی‌های کاربر[3]، منبع اطلاعاتی اصلی جهت ارائه پیشنهادهاست. ویژگی‌های کاربر معمولا در قالب پرسشنامه و هنگام ثبت نام کاربر دریافت می‌شود. در اینگونه از سیستم‌ها، شباهت بین کاربر کنونی و سایر کاربران براساس ویژگی‌های آنها، محاسبه می‌شود. سپس برای هر کاربر مجموعه از کاربران همسایه که بیشترین شباهت را دارند مشخص می‌شوند. در نهایت به کلیه اقلامی که کاربران همسایه آنها را رتبه‌بندی کرده‌اند، رتبه‌ای اختصاص می‌یابد تا براساس آن رتبه، به کاربر کنونی اقلام مطلوب پیشنهاد شود(Al-Shamri 2016).

۴. دانش محور[4] : این سیستم‌ها زمانی استفاده می‌شوند که در ابتدای کار، اطلاعات و اقلام کافی در سیستم موجود نیست. بیشتر سیستم‌های یادگیری از جمله سیستم‌های پیشنهاددهنده با چالش

---

[1] Content-Based
[2] Classification
[3] User profile
[4] Knowledge -Based recommender Systems

«شروع سرد»[1] مواجه هستند. شروع سرد به این معناست که در ابتدای شروع کار سیستم که ممکن است تعداد کاربران و رتبه‌بندی آنها از اقلام محدود باشد، سیستم‌های اطلاعات کافی را برای یافتن اقلام مناسب برای یک کاربر جدید را ندارد. برای برطرف نمودن این چالش، سیستم‌های پیشنهاددهنده دانش محور این اجازه را به کاربر می‌دهند که کاربر خواسته خود را به طور مستقیم بیان کند یا از منابع دانشی دیگری استفاده شود. دو دسته مهم این سیستم‌ها عبارتند از(Burke 2007):

- سیستم‌های پیشنهاددهنده محدودیت محور[2] : در این حالت کاربر محدودیت و درخواست های خود را در مورد ویژگی‌های اقلام (به عنوان مثال بیشترین و کمترین مقدار) به سیستم ارائه می‌دهد.
- سیستم‌های پیشنهاددهنده نمونه محور[3] : در این سیستم‌ها معیارهای شباهت براساس ویژگی اقلام به منظور بازیابی اقلام مشابه استفاده می‌شود. بنابراین در این نوع سیستم‌ها، تعریف معیار شباهت در وهله اول اهمیت زیادی دارد که برای تعیین این معیار معمولا از دانش زمنیه‌ای حاصل از تعامل و بازخورد از کاربر نیز استفاده می‌شود(Felfernig & Burke 2008).

۵. ترکیبی[4] : سیستم‌های مبتنی بر پالایش همکارانه بر رتبه‌بندی‌های جمعی کاربران متکی هستند، سیستم‌های محتوا محور براساس توصیف‌های متنی و رتبه‌بندی‌های کاربر هدف پیشنهاد ارائه می‌دهد و سیستم‌های دانش محور براساس تعامل با کاربر کار می‌کند. این سیستم‌ها هر یک از منابع دانشی مختلفی استفاده می‌کنند و دارای نقاط قوت و ضعفی هستند. برخی از سیستم‌های پیشنهاددهنده، مانند سیستم‌های مبتنی بر دانش، در شروع سرد که در آن مقدار قابل توجهی از اطلاعات در دسترس نیست، موثرتر است. سایر سیستم‌های پیشنهاددهنده مانند روش‌های مبتنی بر پالایش همکارانه، هنگامی که مقدار زیادی از اطلاعات در دسترس است، موثرترند. استفاده از سیستم‌های ترکیبی این امکان را می‌دهد که جنبه‌های مختلف از انواع سیستم‌ها برای به دست آوردن بهترین خروجی ترکیب شود(Burke 2007).

---

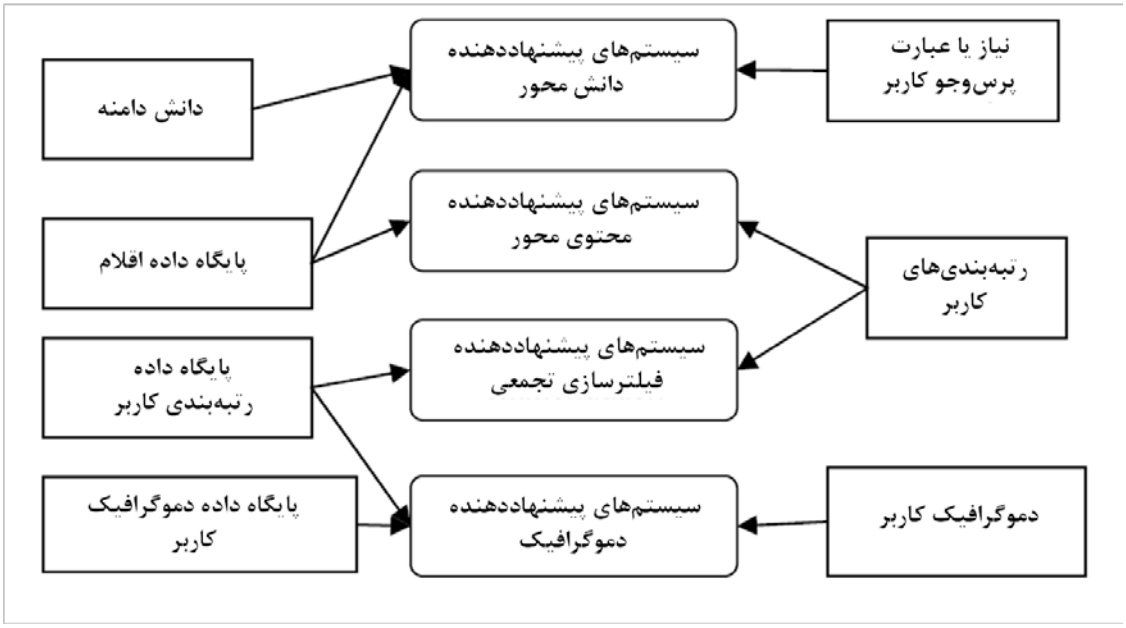
[1] Cold -start

[2] Constraint-based recommender systems

[3] Case-based recommender systems

[4] Hybrid Recommender Systems

در شکل ۲-۳ انواع سیستم‌های پیشنهاد دهنده به همراه نوع منبع دانشی که در آنها استفاده می‌شود، آمده است. سیستم‌های پیشنهاددهنده ترکیبی، بسته به اینکه از ترکیب چه نوع سیستمی حاصل شده‌اند، از منابع دانشی متفاوتی استفاده می‌کنند.



شکل ۲-۳- انواع سیستم‌های پیشنهاددهنده و نوع منبع دانش مورد استفاده در هر یک (Burke 2007)

۲-۱-۲. سیستم‌های پیشنهاددهنده متنی[1]

در سال های اخیر تحلیل داده‌های متنی و استفاده از ابزارهای متن کاوی[2] در سیستم‌های پیشنهاددهنده توجه زیادی را در بخش‌های مختلف تحقیقاتی به خود جلب کرده است.

۲-۱-۲-۱. نمونه‌هایی از پژوهش‌ها در زمینه سیستم‌های پیشنهاددهنده متنی

اولین بار در سال ۲۰۰۶، آشوین و همکارانش با بکارگیری ابزارهای متن کاوی به معرفی سیستم پیشنهاددهنده پرداختند که به کاربر در هنگام سفارش آنلاین یک محصول کمک می‌کرد. سیستم پیشنهاددهنده، این امکان را به کاربر میدهد که علایق و خواسته های خود را در غالب متن ارائه دهد و سیستم با تحلیل متن دریافتی بهترین پیشنهاد را به کاربر خود پیشنهاد میدهد. به این منظور، ورودی ساختارنیافته [3] مشتری به شکل ساختاریافته [4] تبدیل می‌شود و تمامی اصطلاحات نامربوط، علائم نشانه‌گذاری و کلمات غیرکلیدی حذف می‌شود. همچنین ریشه‌سابی و یکسان سازی کلات با ریشه‌های

[1] Textual Recommender systems
[2] Text mining
[3] Unstructured
[4] Structured

یکسان صورت می‌گیرد. در نهایت، توصیه ها با اجرای قوانین که شرایط آنها مطابق با شرایط ورودی مشتریان است و از طریق امتیازدهی به کالاهای موجود ساخته می‌شود (Ittoo et al. 2006).

لی و همکارانش (Li et al. 2010) یک سیستم پیشنهاددهنده با در نظر گرفتن اطلاعات زمینه‌ای[۱] ارائه نمودند که در آن با استفاده از ابزارهای متن کاوی (پردازش زبان های طبیعی[۲]) و استخراج اطلاعات از میان نظرات متنی کاربران و ترکیب آن با رتبه دهی‌های آنها، مشکل شروع سرد در سیستم‌های پیشنهاددهنده را تا حدی مرتفع نمودند.

آشیار (Aciar 2014) سیستم پیشنهاد دهنده ای در حوزه آموزش های الکترونیکی معرفی کرده است که با بکارگیری ابزارهای متن کاوی و تحلیل متن های موجود در گروه‌های تبادل نظر[۳] و استخراج دانش، کاربر را در یافتن مناسب‌ترین فردی که دانش کافی در زمینه مورد نظر کاربر داشته باشد و بتواند پاسخگو باشد یاری می‌کند. در این سیستم بعد از دریافت سوال کاربر و استخراج لغات کلیدی، مناسب‌ترین انجمن ها و افراد مشخص می‌شود. سپس با استفاده از اطلاعات شخصی کاربران انجمن، تحلیل متون نوشته شده توسط آنها و امتیازدهی دیگر کاربران بهترین فرد را که پاسخگوی کاربر باشد معرفی می‌کند.

کفالاس و مانولوپولوس (Kefalas & Manolopoulos 2017) سیستم پیشنهاددهنده‌ای را ارائه کردند که در آن اطلاعات مربوط به مکان، زمان و محتوی پیغام‌های نظرات کاربران درباره اقلام مختلف نیز درنظر گرفته می‌شود. برای استخراج اطلاعات از متن نظرات کاربران از روش‌های متن‌کاوی استفاده کردند. یکی از معیارهایی که برای شباهت دو کاربر در این سیستم درنظر گرفته شده، میزان شباهت متن نظرات آنهاست که براساس واژگان مشترک بین دو متن محاسبه شده است.

سان و همکاران (Sun et al. 2010) با معرفی روش SimRank به مقایسه شباهت متون وبلاگ‌ها، ایجاد گراف میان وبلاگ نویسان و رتبه بندی آنها، مناسب ترین وبلاگ را به کاربران خود پیشنهاد می‌دهند. کاربران ترجیح می دهند مشترک وبلاگ نویسان اصلی در حوزه علاقه‌مندی‌شان باشند. اما از آنجایی که وبلاگ‌ها معمولا به هم مرتبط هستند و از یکدیگر کپی‌برداری می‌کنند، یافتن وبلاگ مناسب معمولا کار ساده‌ای نیست. در این مقاله، با استفاده از روش تحلیل مولفه اصلی[۴] برای کاهش بُعد پست‌های یک وبلاگ استفاده شده است. سپس با استفاده از روش مدل فضای برداری، شباهت پست‌های

---

[۱] Contextual

[2] Natural Language Processing

[3] Forums

[۴] Principal Component Analysis

وبلاگ‌ها با استفاده از مقدار TF-IDF برای هر کلمه غیر تکراری در تمام پستهای وبلاگ محاسبه شده و در نهایت از طریق الگوریتم PageRank اهمیت یک وبلاگ محاسبه مشخص می‌شود.

حریری و همکاران(Hariri et al. 2011) یک سیستم پیشنهاد دهنده زمینه‌ای در حوزه گردشگری را معرفی کردند که اطلاعات زمنیه‌ای را با تحلیل متن نظرات کاربران بدست می‌آورد. در تحلیل متن نظرات کابران از روش L-LDA [۱]استفاده می‌شود که یک روش مدلسازی موضوع[۲] به شمار می‌آید. در واقع نظرات کاربران متونی درنظر گرفته می‌شوند که حاوی موضوعاتی هستند و موضوعات همان انواع سفرهای ممکن در سیستم به شمار می‌آیند که قبلا در سیستم تعریف شده استو با استفاده از روش‌های متن کاوی و روش‌های یادگیری با ناظر لازم است موضوعات نظرات کاربران مشخص شود و این موضوعات در قالب اطلاعات زمینه‌ای در سیستم پیشنهاددهنده درنظر گرفته شوند.

۲-۲. سیستم‌های استدلال نمونه محور

سیستم‌های «استدلال نمونه محور» مبتنی بر این فرضیه هستند که مشکلات جدید معمولا شبیه مشکلاتی هستند که قبلا با آنها مواجه شده ایم. بنابراین یک راه حل موفق در گذشته می‌تواند برای حل مسائل جدید نیز مفید واقع شود. این سیستم‌ها دارای پایگاهی از نمونه‌ها[۳] هستند که شامل مجموعه‌ای از مشکلات حل شده و راه حل‌های آنهاست و مشکلات جدید با اقتباس از راه حل‌هایی که برای مسائل مشابه در گذشته مورد استفاده قرار گرفته، حل می‌شود (Bridge et al. 2005).

۱-۲-۲. فرایندهای استدلال نمونه محور

نحوه استنتاج در سیستم‌های استدلال نمونه محور شامل چهار فرایند اصلی است (Humphreys et al. 2003):

۱. بازیابی[۴]: در این مرحله هنگامی که سیستم یک مساله جدید را دریافت می‌کند به جستجو در میان پایگاه نمونه‌ها و امتیازدهی هر نمونه براساس میزان شباهت به نمونه جدید می‌پردازد تا نمونه‌های موجودی که به نمونه جدید بیشترین شباهت را دارند، پیدا کند.
۲. استفاده مجدد[۵]:در این مرحله سیستم، راه حل موجود برای نمونه‌های مشابه را با مساله جدید تطبیق می‌دهد و بهترین راه حل را برای نمونه جدید، پیشنهاد می‌کند.

---

[۱] Labeled-Latent Drichlet Allocation

[۲] Topic Modeling

[۳] Cases

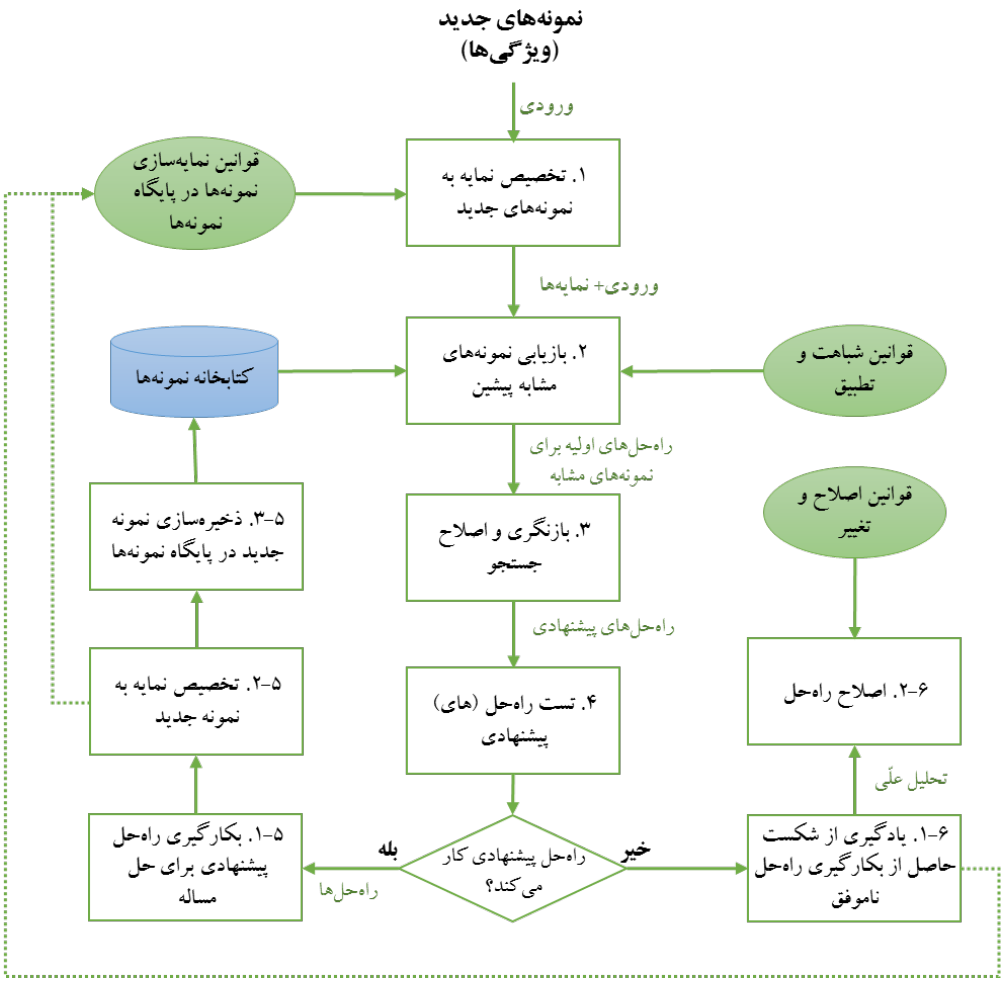
[۴] Retrieve

[۵] Reuse

۳. بازنگری[۱]: سیستم راه‌حل را روی نمونه جدید آزمایش می‌کند و بازنگری های لازم را انجام می‌دهد.

۴. نگهداری[۲]: پس از اینکه راه حل با موفقیت با نمونه جدید، تطبیق داده شده، نتیجه به عنوان یک نمونه جدید در پایگاه نمونه‌ها ذخیره می‌شود.

در شکل ۲-۴ جزییات فرایندهای فوق به صورت شماتیک نمایش داده شده است.



شکل ۲-٤- نحوه عملکرد استدلال نمونه محور (Turban et al. 2010)

---

[۱] Revise

[۲] Retain

استدلال نمونه محور دو جنبه یادگیری و استنتاج را در بردارد. در حوزه هوش مصنوعی وقتی درباره یادگیری صحبت می‌کنیم، معمولا تعمیم‌پذیری یا استدلال و استقرا به ذهن می‌آید. حال آنکه در استدلال نمونه محور، یادگیری با تعمیم یا استقرا انجام نمی‌شود.

یادگیری در استدلال نمونه محور از دو جنبه رخ می‌دهد:

۱) استفاده از نمونه‌های قبلی و تطبیق راه‌حل آنها برای حل یک مساله جدید، به جای تولید کل راه حل از ابتدا،

۲) یادگیری از طریق تکامل سیستم، با افزایش تعداد نمونه‌ها، که منجر به ایجاد جواب‌های بهتر خواهد شد (Kolodner 1992).

یادگیری در استدلال نمونه محور یک فرایند مستمر است به گونه‌ای که با افزایش تعداد نمونه‌ها به پایگاه نمونه‌ها، یادگیری تقویت می‌شود. در واقع یادگیری یک نتیجه جانبی از فرایند استدلال است به گونه‌ای که هنگامیکه یک مساله با موفقیت حل می‌شود، تجربه حاصل از حل آن مساله (مساله و راه‌حل آن به همراه بازخوردهای دریافت شده از کاربر حین ارائه راه‌حل) در سیستم ذخیره می‌شود. همچنین زمانی که یک راه‌حل با شکست مواجه می‌شود، یعنی سیستم راه‌حل ارائه شده به نتیجه نرسیده و مساله حل نشده، اطلاعات شکست (مساله، راه‌حل ارائه شده و دلایل شکست آن) در سیستم ذخیره می‌شود تا در آینده اشتباه مشابه تکرار نشود (Aamodt & Plaza 1994).

۲-۲-۲. سیستم‌های پیشنهاددهنده نمونه محور

همانگونه که عنوان شد، سیستم‌های پیشنهاددهنده محتوی محور، بسته به نوع مساله ودانش موجود درآن، می‌توانند براساس استدلال نمونه محور طراحی شوند. در بسیاری از این سیستم‌ها، نقد کاربر و بازخورد وی از پیشنهاد ارائه شده نقش مهمی را در اصلاح پیشنهاد (یا همان راه‌حل در ادبیات استدلال نمونه محور) ایفا می‌نماید (Felfernig & Burke 2008).

در سیستم‌های پیشنهاددهنده نمونه محور توصیه‌هایی براساس شباهت بین درخواست‌های هدف [۱] و نمونه‌های موجود ارائه می‌شود. به عبارت دیگر، ترجیح سیستم یافتن اقلامی درون پایگاه نمونه‌هاست که دارای بیشترین شباهت به درخواست ارائه شده به سیستم را داشته باشد (Bridge et al. 2005).

نکته قابل توجه این است که این سیستم‌ها در مواردی به خوبی کار می‌کنند که ویژگی‌های نمونه‌ها و درخواست کاربر به خوبی تعریف شده باشند (Smyth 2007).

سیستم‌های پیشنهاددهنده نمونه محورخود به سه دسته اصلی تقسیم می‌شوند:

---

[۱] Target query

- سیستم‌های مکالمه‌ای[۱]: در این حالت ترجیحات کاربر به شکل تکرار حلقه‌های دریافت بازخورد از جانب کاربر تعریف می‌شود (Vasudevan & Chakraborti 2014).
- سیستم‌های مبتنی بر جستجو[۲] : ترجیحات کاربر با استفاده از یک سری سوالات متوالی از پیش تعیین شده، مشخص می‌شود.
- سیستم‌های مبتنی بر رهیابی[۳] : کاربر یکسری تغییرات را برای اقلام پیشنهاد شده درخواست می‌کند تا مورد مطلوب شناسایی شود.

۲-۲-۲-۱. نمونه‌هایی از سیستم‌های پیشنهاددهنده نمونه محور

دانگ و همکاران (Dong et al. 2014) در پژوهش خود یک سیستم پیشنهاددهنده نمونه محور ارائه نمودند که در آن به جای آنکه فقط از نظرات کاربران برای پیشنهاد نمونه‌های مشابه استفاده شود، دو منبع دانشی دیگر نیز برای توصیف نمونه‌ها بکار گرفته می‌شود: ۱) تکمیل ویژگی‌های توصیفی کالا با استناد به اطلاعات فراداده‌ای مربوط به هر کالا که آن کالا را توصیف می‌کند، ۲) استخراج ویژگی‌های کالا از نظرات کاربران با استفاده از روش‌های متن‌کاوی. سپس یک معیار جدید شباهت برای یافتن نمونه‌های مشابه پیشنهاد کردند که بر مبنای شباهت کسینوسی و شباهت معنایی تعریف شده است.

گاتزیورا و سانچه-ماری (Gatzioura & Sànchez-Marrè 2015) در پژوهش خود، برخلاف سایر پژوهش‌ها در این حوزه که هر محصول را در قالب یک نمونه (Case) درنظر می‌گیرند، یک تراکنش را به عنوان یک نمونه در نظر گرفتند. بنابراین در این حالت، ویژگی‌های هر تراکنش (یا هر نمونه)، با اقلام (محصولات) آن تراکنش تعیین می‌شود. هر نمونه، شامل یک مساله یعنی اقلامی که انتخاب شده‌اند، و راه‌حل آن مساله یعنی اقلامی که باید انتخاب شوند تا آن تراکنش کامل شود، است. بنابراین شباهت بین نمونه‌ها، براساس شباهت بین اقلام آنها محاسبه می‌شود. شباهت بین دو قلم نیز، براساس یک ساسله مراتب معنایی بین اقلام محاسبه می‌شود.

اوشی و همکاران (O'Shea et al. 2014) سیستم پیشنهاددهنده‌ای را برای تغییر رفتار کاربران با عنوان NudgeAlong ارائه کردندکه بر مبنای استدلال نمونه محور عمل می‌کند و پیغام‌هایی را برای کاربران برای این منظور ارسال می‌نماید. تغییر رفتار کاربران در مواردی که نیاز است که کاربران به خرید یک کالا یا خدمت تشویق شوند و نظر آنها درباره استفاده از آن کالا یا خدمت تغییر کند، مناسب است؛ مثلا

---

[۱] Conversational systems

[۲] Search-based systems

[۳] Navigation-based systems

افراد را تشویق کنند که از دستگاه‌های گرم‌کننده‌ای استفاده کنند که انرژی کمتری مصرف می‌کند. برای این منظور، این سیستم ابتدا از هر کاربر براساس اطلاعات استفاده وی از آن محصول یا خدمت، یا اطلاعات شخصی او، یک پروفایل ایجاد می‌کند. برای ارسال پیام برای یک کاربر، لازم است که کاربران مشابه و پیام‌هایی که برای آنها ارسال شده، از سیستم بازیابی شود. سپس، براساس این اطلاعات بازیابی شده، پیام مناسب برای کاربر تولید شود. همچنین اطلاعات مربوط به اثربخش بودن پیام در تغییر رفتار کاربر نیز در سیستم ذخیره و استفاده می‌شود.

بوسباهی و چورفی (Bousbahi & Chorfi 2015) یک سیستم پیشنهاددهنده بر مبنای استدلال نمونه محور ارائه کردند که برای پیشنهاد دروس مناسب به کاربران در یک سیستم آموزش الکترونیکی بکار می‌رود. این سیستم در بستر MOOC[1] طراحی شده، که یک پورتال دروس انلاین به شمار می‌رود. در این سیستم، کاربر نیاز درسی خود را در قالب ویژگی‌هایی نظیر عنوان درس، مکان، زبان، و هزینه آن ارائه می‌کند، و سیستم با توجه به این ویژگی‌ها، دروس مشابه را جستجو می‌کند.

به طور کلی، سیستم‌های پیشنهاددهنده نمونه محور در حوزه های مختلف به خصوص در خرید محصولات از فروشگاه‌های الکترونیکی و خدمات مسافرتی کاربرد دارند(Lorenzi & Ricci 2005). همچنین این سیستم‌ها زمانی که اطلاعات زمینه‌ای[2] و یا اطلاعات مربوط به کاربر خاص مربوط به زمان انتخاب باید گنجانده شود، بسیار مفید هستند (Gatzioura & Sànchez-Marrè 2015).

۳-۲. روش‌های بازیابی و محاسبه شباهت متون در سیستم‌های نمونه‌محور[3]

روش‌های شباهت متون، نقش مهمی را در سیستم‌های نمونه محور که بر مبنای داده‌های متنی کار می‌کنند، ایفا می‌کنند. در بسیاری از این سیستم‌ها، شرح نمونه‌ها به صورت متن نمایش داده می‌شوند و برای بازیابی نمونه‌های مشابه، لازم است که یک معیار مناسب برای محاسبه میزان شباهت نمونه جدید با سایر نمونه‌های موجود در پایگاه نمونه‌ها استفاده شود. بنابراین در این بخش خلاصه‌ای از مهمترین روش‌های محاسبه میزان شباهت متون ارائه می‌گردد.

در استدلال نمونه محور برای نمونه‌های متنی[4]، هر یک از چهار فرایند اصلی بازیابی، استفاده مجدد، بازنگری، و نگهداری تحت تاثیر قرار می‌گیرد. در فرایند بازیابی، ابتدا لازم است که ویژگی‌های نمونه متنی جدید استخراج شود تا بتوان با استفاده از آن ویژگی‌ها، نمونه جدید را با نمونه‌های دیگر در پایگاه نمونه‌ها مقایسه نمود. استخراج ویژگی به این دلیل است که معمولا نمونه‌های متنی به صورت ساختارنیافته

---

[1] Massive Open Online Courses

[2] Contextual

[3] Text similarity

[4] Textual case based reasoning

هستند، که با استخراج ویژگی، این نمونه‌ها در قالب ساختاریافته بازتعریف می‌شوند. پس از آن لازم است که نمونه ساختاریافته جدید با نمونه‌های موجود در پایگاه نمونه‌ها مقایسه شود. در اینجا لازم است روش‌های محاسبه شباهت بین متون بکارگرفته شود.

در فرایند استفاده مجدد، لازم است که براساس مسئله و ساختار راه‌حل‌های آن، راه‌حل مناسب برای نمونه جدید ایجاد گردد و پس از بازنگری راه‌حل (در صورت نیاز)، نمونه جدید نیز به همراه راه‌حل مربوط به آن در پایگاه نمونه‌ها نمایه شود(Weber et al. 2005).

بنابراین براساس نتایج پژوهش ویبر و همکاران (Weber et al. 2005)، در سیستم‌های استدلال نمونه محور برای نمونه‌های متنی، چهار چالش اصلی مطرح است:

- شباهت بین نمونه‌ها
- بازنمایی نمونه‌ها در قالب داده‌های ساختاریافته
- استفاده مجدد از نمونه‌ها برای طراحی راه‌حل برای نمونه متنی جدید
- بهبود و توسعه روش‌های خودکارسازی بازنمایی و بازیابی نمونه‌ها

در ادامه برخی از پژوهش‌ها در حوزه شباهت متون معرفی می‌شوند.

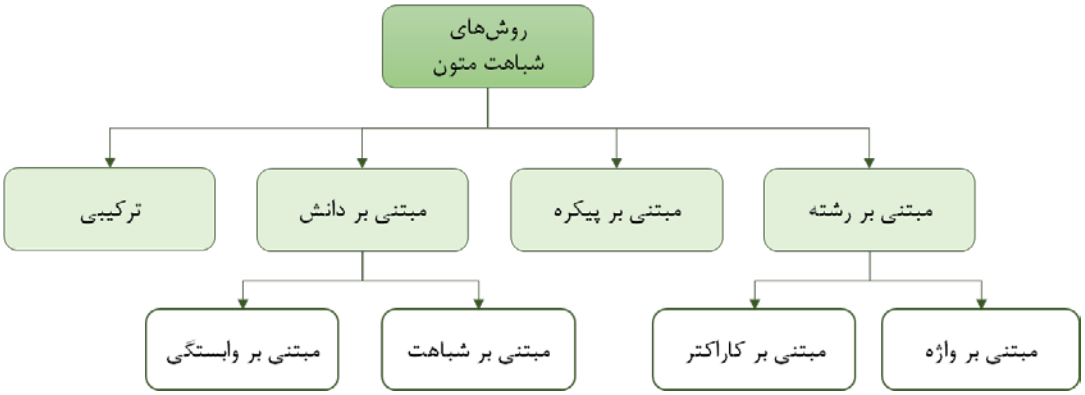
هنگامیکه از تشابه بین متون صحبت می‌کنیم بسته به هدف ما از محاسبه شباهت دو متن، روش‌های متنوعی می‌تواند استفاده شود. به عنوان مثال، هنگامیکه بخواهیم متن مشکوک به سرقت علمی را با یک متن دیگری مقایسه کنیم، روش‌هایی استفاده می‌شوند که امکان مقایسه در سطح کاراکترهای دو متن را نیز ارائه دهند. لیکن هنگامیکه می‌خواهیم دو متن را از نظر معنایی مقایسه کنیم، آنگاه روش‌های معنایی کاربرد بیشتری دارند.

براساس مطالعه گوما و فاهمی(Gomaa & Fahmy 2013)، انواع روش‌های شباهت متون را می‌توان در چهار دسته اصلی قرار داد (شکل ۲-۵):

- روش‌های مبتنی بر رشته کلمات: در این روش‌ها، هر متن در قالب رشته‌ای از کلمات و هر کلمه در قالب رشته‌ای از حروف در نظر گرفته می‌شود. در این‌گونه از روش‌ها، معیار شباهت برمبنای کاراکترها و واژگان است، به‌گونه‌ای که میزان شباهت واژگان هر دو متن حتی در سطح کاراکترهای هر دو واژه نیز ارائه می‌شود.
- روش‌های مبتنی بر پیکره: در این دسته از روش‌ها، یک معیار معنایی درنظر گرفته می‌شود که میزان شباهت بین دو متن را براساس میزان شباهت واژگان آنها برمبنای رفتار آن واژگان در یک پیکره بزرگ (به عنوان مثال هم‌رخدادی واژگان در آن پیکره) محاسبه می‌کند.

- روش‌های مبتنی بر دانش: در این نوع از روش‌ها یک شبکه معنایی واژگانی تعریف می‌شود، و شباهت بین واژگان دو متن و در نهایت شباهت بین دو متن از طریق ارتباطات این شبکه تعیین می‌گردد.
- روش‌های ترکیبی: این روش‌های با ترکیب دو یا چند روش فوق حاصل می‌شوند.

در ادامه هر یک از روش‌های فوق و برخی از مهمترین پژوهش‌ها در این زمینه تشریح می‌شوند.



شکل ۲-۵- دسته‌بندی روش‌های شباهت متون (Gomaa & Fahmy 2013)

۲-۳-۱-۱. روش‌های مبتنی بر رشته

این روش‌ها خود به دو دسته مبتنی بر کاراکتر و مبتنی بر کلمه یا واژه تقسیم‌بندی می‌شوند. روش‌های مبتنی بر کاراکتر معمولا برای مقایسه کاراکترهای دو متن و نه محتوی معنایی آنها استفاده می‌شوند. از بین روش‌های موجود در این زمینه، روش «بزرگترین زیررشته مشترک[۱]» که بنسون و همکاران (Benson et al. 2013) نیز یک نمونه تعمیم یافته آن را معرفی کرده‌اند، کاربرد دارد. همچنین از فاصله دامرو-لونشتاین[۲] نیز استفاده می‌شود که این فاصله براساس میزان تغییراتی که لازم است تا یک کلمه به کلمه دیگری با بکارگیری عملیاتی نظیر درج، جایگزینی یا حذف حروف تبدیل شود.

روش‌های مبتنی بر واژه، به جای مقایسه رشته‌ها براساس کارکترها، از واژگان تشکیل‌دهنده آنها استفاده می‌شود. در این روش‌های فاصله بین دو رشته و در نهایت دو متن برمبنای واژه‌های مشترک آنهاست. مهمترین معیارها و توابع فاصله که در این نوع از روش‌ها استفاده می‌شوند، عبارتند از: فاصله بلاک یا منهتن[۳]، فاصله کسینوسی، فاصله براساس ضریب دایس[۴]، فاصله اقلیدسی[۱]، و فاصله جاکارد[۲]. در اینگونه

[۱] Longest Common SubString (LCS)

[۲] Damerau-Levenshtien

[۳] Block or Manhatan Distnace

[۴] Dice Coefficient

از روش‌ها، معمولا از یک مدل فضای برداری[۳] استفاده می‌شود که در آن هر متن به عنوان برداری از واژگان در نظر گرفته می‌شود، و مقادیر مولفه‌های این بردار براساس حضور یا عدم خضور (یا میزان تکرار) هر واژه در متن تعیین می‌گردد. در نهایت برای سنجش تشابه بین دو متن، فاصله بردارهای متناظر با آن دو براساس یکی از معیارهای فاصله که پیش‌تر ذکر شد، در نظر گرفته می‌شود(Manning & Raghavan 2009).

۲-۳-۱-۲. روش‌های مبتنی بر پیکره

در این روش‌ها مقایسه بین دو متن فراتر از مقایسه بین واژگان آنهاست. در واقع در مقایسه بین واژگان، شباهت معنایی نیز درنظر گرفته می‌شود. این شباهت معنایی براساس پیکره‌ای از واژگان با استفاده از معیارهایی نظیر هم‌رخدادی دو واژه می‌تواند محاسبه شود. یکی از روش‌ها در این حوزه، روش HAL[۴] است که در آن از مفهوم هم‌رخدادی کلمات در یک پنجره از متن استفاده می‌شود و شباهت معنایی بین دو کلمه براساس حضور همزمان دو کلمه در این پنجره سنجیده می‌شود(Lund & Burgess 1996). روش «تحلیل معنای پنهان» (LSA)[۵] نیز از جمله روش‌هایی است که فراتر از مقایسه واژگان عمل می‌کند. در این روش، مانندمدل فضای برداری ماتریس کلمه–سند در نظر گرفته می‌شود لیکن با انجام عملیات «تجزیه مقدار یگانه» (SVD)[۶]، روی این ماتریس، ارتباط معنایی پنهان بین واژگان و اسناد مشخص می‌شود. در واقع در این روش، ماتریس کلمه–سند به فضایی با بُعد پایین‌تری نگاشته می‌شود که در آن فضا، شباهت بین اسناد ، فراتر واژگان مشترک بین آنها محاسبه خواهد شد(Landauer & Dumais 1997; Manning & Raghavan 2009). تغییرات متنوعی روی این روش در تحقیقات پس از آن ارائه شده مانند روش تحلیل معنای پنهان احتمالی[۷]، یا روش تحلیل معنای پنهان تعمیم یافته[۸]. برخی از روش‌های نیز براساس «اطلاعات متقابل»[۹] است که مهمترین آنها روش «اطلاعات متقابل نقطه‌ای»[۱۰] است.

---

۱ Euclidean Distance

۲ Jaccard Distance

۳ Vector Space Model

۴ Hyperspave Analogue to Language

۵ Latent Semantic Analysis

۶ Singular Value Decomposition

۷ Probabilistic Latent Semantic Analysis

۸ Genralized Latent Semantic Analysis

۹ Mutual Information

۱۰ Pointwise Mutual Information

در این روش با استفاده از هم‌رخدادی دو کلمه، میزان اطلاعات متقابل بین دو کلمه که بیانگر ارتباط معنایی آن دو است، تعریف می‌شود (Mihalcea et al. 2006; Turney 2001).

یکی دیگری از روش‌های مبتنی بر پیکره روش بازنمایی کلمات[۱] کلمه-به-بردار (Word2Vec) است که توسط میکولوف و همکاران (Mikolov, Chen, et al. 2013) پیشنهاد شد. در این روش با آموزش یک شبکه عصبی سه لایه روی مجموعه عظیمی از داده‌های متنی، یک بردار چندبعدی برای هر کلمه تعیین می‌شود. شیوه عملکرد این شبکه عصبی بر اساس هم‌رخدادی واژگان در یک پنجره ثابت از کلمات است. بردار مشخص شده برای هر کلمه می‌تواند معیاری برای مقایسه معنایی کلمات با هم و در نهایت مقایسه متون باشد. در باره این روش و شیوه عملکرد آن در بخش‌های بعدی این گزارش توضیحات بیشتری ارائه خواهد شد.

۲-۳-۱-۳. روش‌های مبتنی بر دانش

دسته سوم شباهت‌یابی براساس دانش است که میزان شباهت میان کلمات را از طریق اطلاعات استخراج شده از شبکه‌ای معناشناختی محاسبه می‌کند. شبکه WordNet یک نمونه پراستفاده در زبان انگلیسی از این شبکه واژگان است. در این روش‌ها، شباهت معنایی دو واژه با استناد به شبکه واژگانی معناشناختی، براساس معنای مشترک دو واژه تعریف می‌شود. در برخی از این روش‌ها، محتوی اطلاعاتی بین واژگان اندازه‌گیری می‌شود و در برخی دیگر شباهت، براساس طول ارتباطات شبکه‌ای آنها در شبکه معناشناختی تعرفی می‌شود. به عنوان مثال وو و پالمر معیاری را براساس طول ارتباطات با اندازه‌گیری عمق دو واژه در شبکه معناشناختی و «رده‌بندی‌کننده کوچکترین مقدار مشترک» (LCS)[۲] آن دو واژه تعریف کردند(Gomaa & Fahmy 2013).

۲-۳-۱-٤. روش‌های ترکیبی

روش‌های ترکیبی در بیشتر موارد نتایج بهتری را در محاسبه شباهت متون ارائه کرده‌اند. میهالسی و همکاران (Mihalcea et al. 2006) انواع روش‌های مبتنی بر دانش و مبتنی بر پیکره را پیاده‌سازی کردند، و چندین روش از هر نوع را با هم ترکیب نمودند. نتایج بدست آمده حاکی از این بود که روش‌های ترکیبی عموما نتایج بهتری را ارائه می‌دهند. در ادامه برخی از مهمترین روش‌های ترکیبی در حوزه شباهت متون معرفی می‌شوند.

ایسلام و اینکپن (Islam & Inkpen 2009)، یک معیار مبتنی بر پیکره را با روش بزرگترین زیررشته مشترک ترکیب کردند و از این معیار جدید برای محاسبه شباهت بین دو متن کوتاه استفاده نمودند. معیار

---

[۱] Word embedding

[۲] Least Common Subsumer

مبتنی بر پیکره براساس روش «اطلاعات متقابل نقطه‌ای» که در آن از هم‌رخدادی مرتبه دوم استفاده می‌شود (SOC-PMI)[۱] است و در بکارگیری روش بزرگترین زیر رشته، یک معیار نرمال‌شده را بر مبنی این معیار تعریف نموده‌اند که چالش شباهت جملات بزرگ با کوچک نیز برطرف گردد.

آگاروال و همکاران (Aggarwal et al. 2012) روش ترکیبی برای شباهت دو متن پیشنهاد کردند. در این روش ترکیبی، شباهت بین دو واژه با ترکیب دو معیار تعریف می‌شود: معیار ارتباط معنایی مبتنی بر پیکره در یک جمله و معیار شباهت بین نقش دستوری دو کلمه در جملات مختلف. برای تعریف هر یک از این معیارها، ویژگی‌هایی تعریف شده و در نهایت از روش‌های یادگیری ماشین مانند رگرسیون خطی برای تعیین شباهت استفاده شده است.

معیار شباهتی که بوسکالدی و همکاران (Buscaldi et al. 2012)، ارائه کردند از ترکیب دو ماژول حاصل شده است: ماژول محاسبه شباهت بین جملات براساس مدل‌های n-گرم، و ماژول محاسبه شباهت بین مفاهیم[۲] براساس معیار شباهت مفاهیم و شبکه WordNet. ماژول n-گرم براساس مدل CKPD[۳] تعریف شده که در ابتدا در حوزه سیستم‌های پرسش و پاسخ استفاده می‌شد. ماژول شباهت بین مفاهیم براساس معیار شباهت بین مفاهیم در شبکه معنایی WordNet که توسط وو و پالمر (Wu & Palmer 1994) پیشنهاد شد، تعریف شده است. در این معیار شباهت بین دو مفهوم براساس عمق آنها در شبکه واژگان تعریف می‌شود.

٤-۲. استخراج خودکار کلیدواژه

رشد روزافزون داده‌های متنی در قالب مستندات الکترونیکی، کاربران را با چالش افزونگی اطلاعات مواجهه ساخته است. جهت مواجهه با این چالش و استفاده بهتر از اطلاعات موجود، یکی از راهکارهای پیشنهادی استخراج کلیدواژه‌ها از متون است. استخراج خودکار کلیدواژه به فرایند خودکار استخراج کلمات مهم از متن گفته می‌شود که فهمیدن مفاهیم و مفهوم اصلی متن را هموارتر می‌سازد (Jiang 2012). علاوه بر این، درک بهتری از مرتبط بودن متن با پرس‌وجو[۴] کاربر را در زمان کمتری فراهم می‌آورد. کلیدواژه‌ها می‌توانند به صورت تک‌کلمه[۵] یا مجموعه‌ای از کلمات[۶] باشند که توصیفی از محتوای متن را فراهم می‌آورند.

---

[۱] Second Order Co-occurance PMI

[۲] Concepts

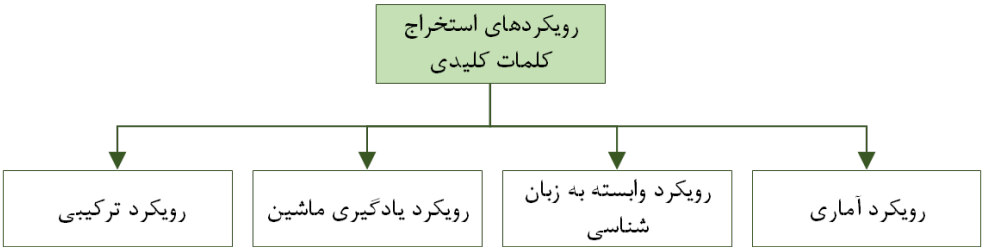
[۳] Clustered Keywords Positional Distance

[4] Query

[۵] Keyword

[۶] Keyphrase

رویکردهای متنوعی برای استخراج خودکار کلیدواژه‌ها پیشنهاد و استفاده شده است (Hasan & Ng 2014; Zhang et al. 2008) که مطابق شکل ۲-۶ به چهار دسته اصلی تقسیم‌بندی می‌شوند.



شکل ۲-۶- دسته‌بندی انواع رویکردهای استخراج خودکار کلیدواژه‌ها

- رویکرد آماری[۱]: در این رویکرد (Siddiqi & Sharan 2016) بدون استفاده از داده‌های آموزشی[۲] امکان تشخیص کلمات وجود دارد. این رویکرد براساس ویژگی‌های آماری نظیر رخداد یا عدم رخداد هر کلمه در سند، موقعیت مکانی کلمه در سند، فراوانی کلمه، فراوانی معکوس کلمه در سند و هم‌رخدادی آن[۳] با سایر کلمات کلید عمل می‌کند.
- رویکرد زبان‌شناختی: در این رویکرد به منظور استخراج کلیدواژه‌ها از ویژگی‌های زبان‌شناسی استفاده می‌شود و عموما بر اساس مدل‌های n-گرم[۴]، تحلیل واژگانی، تحلیل نحوی وتحلیل گفتمان است (Siddiqi & Sharan 2015; Hulth 2003; Ercan & Cicekli 2007).
- رویکرد یادگیری ماشین: روش‌های استفاده شده در رویکرد یادگیری ماشین عموما به دو دسته یادگیری با ناظر و یادگیری بدون ناظر تفکیک می شوند. در یادگیری با ناظر با استفاده از مجموعه‌ای از داده‌های آموزشی که متشکل از اسنادی است که کلیدواژه‌ها آنها مشخص شده، مدلی برای تعیین کلیدواژه‌ها در یک سند مشخص می‌شود. سپس با استفاده از این مدل، برای هر کلمه کاندید در یک سند جدید، مشخص می‌شود که آیا آن کلمه می‌تواند یک کلمه کلیدی باشد یا خیر. روش‌هایی که بیشترین استفاده را در استخراج کلیدواژه‌ها بر مبنای یادگیری با ناظر دارند، عبارتند از: دسته بندی بیزی ساده[۵]، ماشین بردار پشتیبان[۶] و شبکه‌های عصبی مصنوعی. در روش‌های یادگیری بدون ناظر، استخراج کلیدواژه‌ها بدون مدل صورت می‌گیرد که در آنها از

---

[1] Statistical Approaches
[۲] Training Data
[۳] Word Co-occurrence
[4] N-Gram
[۵] Naïve Bayes
[۶] Support Vector Machine

روش‌های مبتنی بر گراف و خوشه‌بندی استفاده می‌شود (Hulth 2004; K. Zhang, H. Xu, J. Tang 2006; Ercan & Cicekli 2007; Hasan & Ng 2014).

- رویکرد ترکیبی: رویکرد ترکیبی، با تلفیق دو یا چند رویکرد فوق، حاصل می‌شود. به عنوان مثال در برخی از روش‌های یادگیری با ناظر از ویژگی‌های زبان‌شناختی نیز به عنوان ویژگی کلمات استفاده می‌شود(Hulth 2004; Ercan & Cicekli 2007).

۱-٤-۲. مروری بر پژوهش‌های مهم در استخراج خودکار کلیدواژه‌ها

یکی از ساده‌ترین روش‌های موجود برای استخراج کلیدواژه‌ها فراوانی کلمه[1] در سند است (Luhn 1957). در این نوع از روش‌ها، میزان تکرار کلمه، معیار اصلی است که براساس آن یک کلکه به عنموان کلمه کلیدی شناخته می‌شود. اما یکی از مشکلات این روش ارائه جواب‌های نامطلوب[2] است.

اورتونو و همکاران (Ortuno et al. 2002) روش را برای استخراج کلیدواژه‌ها از یک سند پیشنهاد کردند که در آن دسته بندی کلمات در یک سند بر اساس انحرافات استاندارد فاصله بین دو کلمه هم‌رخداد متوالی انجام می‌شود. با توجه به نتایج تحقیق ایشان، کلیدواژه‌ها در یک سند معمولا انحرافات استاندارد بزرگتری دارند.

از جمله روش‌های مهم برای استخراج کلیدواژه‌ها و مواجهه با مشکلات استخراج کلمه صرفا بر اساس فراوانی کلمه، استفاده از فراوانی کلمه-فراوانی معکوس سند[3] (TF-IDF) است(Salton et al. 1975) . این روش میزان تکرار یک کلمه در یک سند را در مقابل تعداد تکرار آن در مجموعه کلیه مستندات در نظر می‌گیرد. به عبارت دیگر این روش مبتنی بر چند سند است که علاوه بر میزان تکرار یک کلمه در سند، به تعداد اسنادی که آن کلمه در آنها تکرار شده است هم توجه می‌کند تا اثر کلمات عمومی را در استخراج کلیدواژه‌ها کاهش دهد. البته در این روش گاهی اوقات کلمات پرتکرار که کلمات تخصصی و کلیدی نیز هستند، ممکن است حذف شوند یا به اشتباه برخی کلمات دیگر به عنوان کلیدواژه در نظر گرفته شود.

یکی دیگر از روش‌های مهم برای استخراج کلیدواژه‌ها استفاده از الگوریتم رتبه‌بندی براساس گراف یعنی TextRank (Mihalcea Rada 2004) است. در این روش با استفاده از گراف هم‌رخدادی پس از شناسایی برچسب واژگان باتوجه به ارتباط این واژگان در گراف، کلمات مهم سند استخراج می‌شود. طبق آزمایش‌های انجام شده (Rose et al. 2010) روش RAKE در مقایسه با TextRank که از اندازه

---

[1] Term frequency

[2] Unrefined

[3] Term Frequency-Inverse Document Frequency

پنجره با مقدار مشخص برای در نظر گرفتن هم‌رخدادی کلمات استفاده می‌کند، از نظر محاسباتی پیچیدگی کمتری دارد.

علاوه بر مطالعات انجام شده، مطالعات دیگری نیز از رویکردهای یادگیری با ماشین برای استخراج کلیدواژه‌ها از یک سند استفاده کرده اند. به عنوان مثال نجفی و درونه (Najafi & Darooneh 2015) برای یافتن مهمترین کلمات یک سند از خوشه بندی و تجزیه وتحلیل ابعاد فراکتال[۱] استفاده کرده اند. در این رویکرد، فاصله بین دو رخداد کلمات متوالی به عنوان یک مجموعه فراکتال در نظر گرفته می‌شود و نتایج حاکی از آن است که کلیدواژه‌ها ابعاد فرکتال بزرگتری دارند، در حالی‌که ابعاد فراکتال کلمات بی‌اهمیت کم است.

در رویکرد یادگیری ماشین معمولا با استفاده از داده‌های آموزشی یک مدل آموزش داده می‌شود و سپس براساس مدل ساخته شده، کلمات و عبارات کلیدی از اسناد جدید استخراج می‌گردد. از جمله مهم‌ترین روش‌ها در این رویکرد می‌توان به پژوهش ویتن و همکاران (Witten et al. 1999) و هولت (Hulth 2003) اشاره کرد که برای ساختن مدل آموزشی از بعضی ویژگی‌هایی همچون فراوانی کلمه-فراوانی معکوس سند، اولین رخداد کلمه و برچسب واژگانی کلمات استفاده شده است.

۲-٤-۱-۱. روش استخراج سریع و خودکار کلیدواژه‌ها

یکی از روش‌های مهم برای مواجهه با حذف کلیدواژه‌های تخصصی حوزه خاص که در مستندات بیشتری، به‌عبارتی در کل پیکره‌ی متنی تکرار شده‌اند، استفاده از روش «استخراج خودکار سریع کلیدواژه‌ها» (RAKE)[۲] (Rose et al. 2010) است.

روند استخراج کلیدواژه‌ها براساس این روش به این صورت است که در ابتدا ایست واژه حذف می‌شود (به عنوان مثال «برای»، «هستند»، «وجود دارد»، و غیره). سپس، برداری برای سایر کلمات باقی‌مانده در متن که کلیدواژه‌های کاندید هستند، ایجاد می‌شود. پس از آن، یک گراف هم‌رخدادی جفت برای کلیدواژه‌های کاندید بدون استفاده از اندازه پنجره[۳] ایجاد می‌شود و در نهایت با استفاده از فراوانی کلمه در سند و درجه کلمه[۴] - که مربوط به تعداد دفعاتی است که کلمه کلیدی کاندید شده با سایر کلیدواژه‌های کاندید هم‌رخداد شده است-کلیدواژه‌های نهایی استخراج می‌شود.

در این روش برای هر کلیدواژه کاندید، یک امتیاز محاسبه می‌شود که براساس دو شاخص فراوانی کلمه، $freq(k_j, d_i)$ ، و درجه کلمه، $\deg(k_j, d_i)$، به صورت زیر محاسبه می‌شود:

---

۱ Fractal

Rapid Automatic Keyword Extraction(RAKE)۲

۳ Window

4 Word degree

$$score(k_j, d_i) = \frac{\deg(k_j, d_i)}{freq(k_j, d_i)} \quad (۱-۲)$$

کلیدواژه‌ها براساس امتیاز محاسبه‌شده ، مرتب می شوند. معمولا یک سوم کلمات کلیدی به عنوان کلمات کلیدی نهایی برای هر مستند علمی انتخاب می شود.

۲-٤-۱-۲. استخراج خودکار کلیدواژه‌ها از متون فارسی

همانند متون لاتین برای متون فارسی نیز روش‌های مختلفی برای استخراج کلیدواژه‌ها به‌کار گرفته شده که البته پژوهش‌ها در زبان فارسی بسیار محدودتر است. همانند سایر زبان‌ها در زبان فارسی، به‌منظور استخراج کلیدواژه‌ها ابتدا پیش‌پردازش‌هایی[۱] برای اسناد انجام می‌شود. یکی از چالش‌های پیش پردازش زبان فارسی وجود فضای خالی بین کلمات است که ممکن است به اشتباه منجر به تفکیک کلمات چندبخشی شود و یک کلمه با پیشوند یا پسوند، به صورت چند کلمه متمایز به‌نظر برسد، که در اینصورت کلمات کلیدی به طور صحیح استخراج نمی‌شوند(Khozani & Bayat 2011). به‌منظور مواجهه با این چالش روش‌های متعددی ارائه شده است تا ریشه‌های کلمات فارسی مشخص شود (Mokhtaripour & Jahanpour 2006; Taghva et al. 2005). باتوجه به اهمیت پیش پردازش در استخراج کلمات علی رغم روش‌های موجود، ریشه یابی کلمات فارسی همچنان یکی از چالش‌هاست.

از جمله رویکردهای مهم برای اسناد فارسی می توان به روش صیادی و همکاران (Sayyadiharikandeh et al. 2012) اشاره کرد که با استفاده از الگوریتم PostRank بُعد اسناد متنی موجود در یک وبلاگ را کاهش می‌دهد تا تم اصلی وبلاگ را مشخص کند. به عبارت دیگر در این روش، کلماتی که به نحو بهتری نمایش‌دهنده مفهوم یک وبلاگ هستند، مشخص می‌شود. PostRank از یک نمایش‌دهنده گراف نحوی با در نظر گرفتن برخی از ویژگی‌های ساختاری وبلاگ برای استخراج کلیدواژه‌ها استفاده می کند.

رویکرد دیگری برای متن فارسی توسط کیان و زاهدی (Kian & Zahedi 2011) پیشنهاد شده که با استفاده از استخراج کلیدواژه‌ها، امکان دسترسی به محتوای صفحات وب را بهبود می‌دهد. برای استخراج کلیدواژه‌ها از دو تابع امتیازدهی استفاده می‌شود که اولین تابع امتیازدهی سعی در یافتن اهمیت کلمات بر اساس ویژگی های آماری دارد و دومین تابع نتایج حاصل از امتیازدهی تابع اول را با استفاده از نتایج موتورهای جستجوی پراستفاده مانند گوگل، یاهو و ام. اس.ان[۲] ارزیابی می کند، به این ترتیب که نتیجه

---

[۱] Preprocessing

[۲] Google, Yahoo, MSN

صفحه‌ی اول را به عنوان بازخورد دریافت می‌کند وسپس اولین تابع را با استفاده از الگوریتم ژنتیک بهینه می‌کند.

علاوه بر این، خوزانی و بیات، (Khozani & Bayat 2011) رویکرد جدیدی با استفاده از رویکردهای آماری برای استخراج کلیدواژه‌ها ارائه کرده‌اند که نه تنها قادر به استخراج کلیدواژه‌های تک کلمه‌ای است بلکه قابلیت استخراج کلیدواژه‌های دو کلمه‌ای را نیز برای متن فارسی دارد. در این رویکرد از دو معیار فراوانی کلمه و تعداد اسناد دربردارنده کلمه[۱] برای استخراج کلیدواژه‌های تک کلمه‌ای استفاده شده است و سپس با استفاده از ماتریس ساخته شده برای کلیدواژه‌های تک کلمه‌ای، کلیدواژه‌های دوکلمه‌ای نیز استخراج شده‌اند.

۲-۵. روش‌های بازنمایی کلمات براساس مدل‌های برداری

یکی از رویکردها برای پردازش متون در مسائل پردازش زبان طبیعی، اینست که واژگان به صورت بردارهای عددی نمایش داده شوند. این رویکرد در روش‌های استخراج کلیدواژه که براساس الگوریتم‌های یادگیری ماشین عمل می‌کنند بکارگرفته می‌شود، زیرا این نمایش برداری می‌تواند بیانگر بردار ویژگی‌ها باشد.

برخی از این روش‌ها براساس فراوانی عمل می‌کنند مانند روش کدگذاری تک-مولفه‌ای[۲]، و روش‌های هم-رخدادی و تحلیل معنای پنهان[۳] و برخی دیگر نیز برمبنای مدل‌های پیش‌بینی عملی می‌کنند مانند روش کلمه-به-بردار[۴]. در ادامه مهمترین این روش‌ها در این حوزه تشریح می‌شوند.

۱-۵-۲. روش کدگذاری تک-مولفه‌ای

یک روش رایج برای نمایش برداری کلمات روش کدگذاری تک-مولفه‌ای است که طول هر بردار برابر با تعداد کل کلمات منحصر به فرد در متون است. اساسا هر بردار متناظر با یک کلمه از متن موردنظر است که فقط برای خود خانه متناظر با کلمه در صورت موردنظر عدد یک و برای همه خانه ها بجز خانه متناظر با آن کلمه عدد صفر در نظر می‌گیرد، این روش به نوعی کیسه کلمات صفر و یکی[۵] محسوب می‌شود.

---

[۱] Document frequency

[۲] One-Hot encoding

[۳] Latent Semantic Analysis (LSA)

[۴] Word2Vec

[۵] Boolean Bag of Word

روشی دیگر برای نمایش کلمات، نمایش برداری به صورت کیسه کلمات[۱] است. این روش کیسه‌ای از کلمات و فراوانی وقوع آن‌ها را در یک متن (سند) توصیف می کند.

بنابراین اگر پیکره متنی $\boldsymbol{D}$ که شامل اسناد $\{d_1, \dots, d_M\}$ و N کلمه متمایز باشد، آنگاه می‌توان یک ماتریس ماتریس کلمه–سند[۲] با اندازه $M \times N$ در نظر گرفت که هر سطر ماتریس نشان‌دهنده فراوانی هر کلمه در سند d_i است.

روشی دیگر برای بردار کلمات براساس فراوانی، روش «فراوانی کلمه–فراوانی معکوس سند»(TFIDF)[۳] است (Salton 1989). تفاوت این روش با روش کیسه کلمات، در نظر گرفتن اهمیت کلمه در سند مربوطه و در کل پیکره‌ی متنی است. به‌نوعی این روش فراوانی یک کلمه در یک سند را در مقابل فراوانی آن در کل پیکره متنی می‌سنجد. از آنجاییکه در یک پیکره متنی مشخص، کلماتی که در همه اسناد تکرار شده‌اند، بار معنایی برای تمایز یک سند از سند را نخواهند داشت، فراوانی هر کلمه براساس تعداد اسنادی که آن کلمه در آنها تکرار شده است، تعدیل می‌شود، که به صورت زیر محاسبه می گردد:

$$\boldsymbol{TF}\ (\boldsymbol{t}, \boldsymbol{d_i}) = \textbf{frequency of term } \boldsymbol{t} \textbf{ in document } \boldsymbol{d_i} \qquad (۲\text{-}۲)$$

$$IDF\ (t) = \frac{N}{|\{d \in \boldsymbol{D}\ | t \in d\}|} \qquad (۳\text{-}۲)$$

$$TFIDF = TF\ (t, \boldsymbol{d_i}).IDF\ (t) \qquad (۴\text{-}۲)$$

فرمول (۲-۲) برای محاسبه فراوانی کلمه است. فرمول (۲–۳)، برای محاسبه فراوانی معکوس سند، که صورت کسر، نشانده تمام اسناد موجود در پیکره متنی و مخرج کسر، تعداد اسنادی است که کلمه موردنظر درآنها آمده است. فرمول (۲–۴)، حاصل‌ضرب فراوانی کلمه در یک سند در فراوانی معکوس کلمه برای محاسبه‌ی امتیاز یک کلمه باتوجه به کل پیکره‌ی متنی است. مطابق فرمول، این روش امتیاز کمتری به کلماتی‌که فراوانی بیشتری در کل مجموعه مستندات دارند قائل می‌شود. در نهایت هر درایه از ماتریس کلمه–سند، بیانگر وزن کلمه براساس فرمول (۲–۴) در یک سند است(Manning & Raghavan 2009).

اگر چه روش نمایش برداری بر اساس فراوانی کلمات بسیار ساده است، اما با معایب متعددی همراه دارد. یکی از این معایب آن، در نظر نگرفتن روابط معنایی بین دو کلمه و ترتیب آنها درمتن است. علاوه بر این، به دلیل تعداد بیشمار صفرها در بردارها نیاز به فضای ذخیره سازی بیشتری خواهیم داشت.

---

[۱] Bag-of-Word

[۲] Term-document matrix

[۳] Term Frequency-Inverse Document Frequency

۲-۵-۲. روش ماتریس هم‌رخدادی و تحلیل معنایی پنهان[۱]

ماتریس هم‌رخدادی، کنار هم قرارگیری دوبه‌دوی کلمات را در یک پنجره مشخص، در یک پیکره متنی بیان می‌کند. به عبارت دیگر، هر درایه از ماتریس هم‌رخدادی نشاندهنده تعداد دفعاتی است که دو کلمه باهم (در یک پنجره‌ای مشخص از کلمات) در یک پیکره متنی رخداده‌اند. طبق روش هم‌رخدادی، کلماتی که متعلق به یک حوزه خاص هستند و از نظر معنایی شبیه‌اند، باهم در یک متن رخ می‌دهند. برتری این روش به روش فراوانی کلمه-فراوانی معکوس سند، در نظر گرفتن روابط معنایی کلمات در یک سطح اولیه است.

ماتریس هم-رخدادی می‌تواند در تحلیل‌های بعدی نظیر خوشه‌بندی اسناد، یا بازیابی و دسته‌بندی آنها به عنوان نمایش برداری از کلمات در نظر گرفته شود. لیکن یکی از نمونه‌های پرکاربرد آن، استفاده از این ماتریس، به عنوان ماتریس ویژگی های اولیه در روش تحلیل معنایی پنهان (LSA) است، که برای کشف روابط معنایی کلمات و مستندات بکار گرفته می‌شود. طبق این روش، کلماتی که در یک سند رخ‌داده‌اند، از نظر معنایی به هم مرتبط هستند و مستنداتی که دارای موضوع مشابهی هستند نیز کلمات مشابهی دارند(Deerwester et al. 1990).

برای استخراج معنای نهفته هرسند و کلمات به کاررفته در آن از ماتریس هم-رخدادی یا ماتریس کلمه-سند برای نمایش این داده‌ها استفاده می‌شود. روش آنالیز معنایی پنهان برای کاهش ابعاد این ماتریس از روشی برای کاهش بُعد فضای ویژگی‌ها، به‌نام تجزیه مقادیر منفرد (SVD)[۲] استفاده می‌کند. با تصویر کردن کلمات و متون در زیرفضای با بُعدکمتر، هر کلمه یک بازنمایی برداری با ابعاد بسیار کمتر از ابعاد ماتریس اولیه خواهد داشت. سپس با محاسبه‌ی شباهت کسینوسی بین دو بردار که سطرهای ماتریس هستند، می‌توان ارتباط معنایی بین کلمات را یافت. مقادیر نزدیک به یک کلماتی هستند که از نظر معنایی شباهت بیشتری دارند و مقادیر نزدیک به صفر کلماتی هستند که شباهت معنایی کمتری دارند(Manning & Raghavan 2009).

ماتریس کلمه-سند، تنها فراوانی رخداد هر کلمه در سند مربوطه را نمایش می‌دهد، اما ترتیب کلمات در سند را در نظر نمی‌گیرد، و به‌نوعی مشابه روش کیسه‌کلمات است. برای در نظر گرفتن ترتیب کلمات می‌توان به‌جای ماتریس کلمه-سند از ماتریس هم‌رخدادی استفاده کرد.

---

[۱] Latent Sematic Analysis (LSA)

[۲] Singular value Decompision

۳-۵-۲. روش کلمه-به-بردار

یکی از مهمترین الگوریتم‌های بازنمایی کلمه، روش کلمه-به-بردار که در سال ۲۰۱۳ توسط میکولوف، محقق شرکت گوگل، برای نمایش برداری کلمات در فضای ابعادی کم پیشنهاد شد (Mikolov, Chen, et al. 2013). روش کلمه-به-بردار، براساس یک شبکه عصبی با یک لایه پنهان[۱] طراحی شده که کلمات را به عنوان بردارهایی در یک فضای برداری چند بعدی نمایش می‌دهد. درواقع این روش، یک مدل پیش‌بینی را با آموزش[۲] شبکه عصبی روی مجموعه‌ای عظیم از متون می‌سازد که براساس آن می‌تواند پیش‌بینی کند، هر کلمه چه میزان به کلمه دیگری ارتباط دارد. لیکن پس از آموزش شبکه عصبی، تنها از وزن‌های مشخص شده در لایه پنهان به عنوان بردار بازنمایی هر کلمه استفاده می‌شود.

مدل آموزش داده شده، روابط معنایی بین کلمات را با توجه به نحوه قرارگیری آن‌ها در پیکره‌ی متنی استخراج می‌کند. دقت مدل به تعداد دفعات هم‌رخدادی کلمات درون یک متن درپیکره متنی وابسته است. بنابراین هر چه حجم داده‌های پیکره بیشتر باشد، دقت این روش افزایش خواهد یافت.

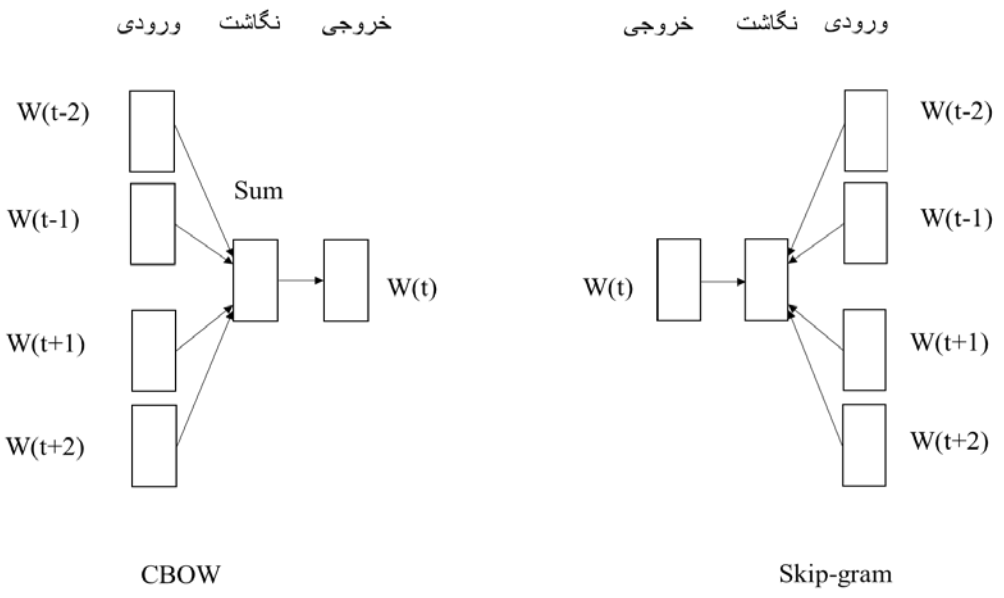
برای آموزش شبکه عصبی از دو مدل استفاده می‌شود: مدل کیسه لغات پیوسته (CBOW)[۳] و مدل اسکیپ گرام (Skip-gram). در ادامه هر یک از این دو مدل و نحوه کارکرد روش بردار-به کلمه تشریح می‌شود. در شکل ۲-۷، ساختار دو مدل CBOW و Skip-gram نمایش داده شده است، که $w(t)$ کلمه هدف، N اندازه پنجره و کلمات همسایگی از $w(t-N)$ تا $w(t+N)$ هستند.

---

[۱] Hidden Layer

[۲] Training

[۳] Continuous Bag of Words



شکل ۲-۷–ساختار شماتیک مدل CBOW وSkip-gram (Mikolov, Chen, et al. 2013)

در مدل CBOW شبکه عصبی آموزش داده می‌شود که با یک اندازه پنجره مشخص، با استفاده از کلمه‌های همسایگی (Context)،کلمه هدف[۱] را پیش بینی می‌کند. بنابراین در این مدل کلمات برای آموزش شبکه، همسایگی یک کلمه هدف به عنوان ورودی به شبکه عصبی داده می‌شود و وزن‌ها براساس پیکره متنی، تعیین می‌شوند. در مدل Skip-gram، شبکه عصبی آموزش داده می‌شود که با یک اندازه پنجره مشخص، با استفاده از یک کلمه هدف، کلمه‌های همسایگی را پیش بینی کند.

در CBOW، قبل از پیش بینی کلمه هدف، از بردارهای کلمات زمینه محلی میانگین گرفته می‌شود. اما درSkip-gram عمل میانگیری برای بردارهای کلمات انجام نمی‌شود. به نظر می‌رسد این روش در مواقعی که کلمات خاص و کم تکرار در پیکره زیاد است، بهتر عمل می‌کند زیرا از بردار کلمه با بردار سایر کلمات زمینه‌ی محلی برای پیش بینی، میانگین گرفته نمی‌شود.

در هر دو مدل، آموزش با توجه به زمینه محلی[۲] خود، که در اصل توسط یک پنجره از کلمات همسایگی تعریف شده، انجام می‌شود. هنگامی که آموزش همگرا شد، بردار خروجی نادیده گرفته می‌شود و تنها وزن موجود بین لایه پنهان و لایه خروجی به عنوان نمایشگر برداری کلمات مورد استفاده قرار می‌گیرد.

---

[۱] Target Word

[۲] Local Context

در CBOW، برای پیش بینی کلمه هدف، بردارهای کلمات زمینه محلی تجمیع و مورد استفاده قرار می‌گیرد. بنابراین، اگر پیکره متنی شامل کلمات $w_1, w_2, \ldots, w_T$ و اندازه‌ی پنجره برابر با c باشد. تابع هدف برابر است با:

$$\frac{1}{T}\sum_{t=1}^{T}\log p(w_t|\sum_{-c\le j\le c, j\neq 0} w_{t+j}) \quad (۵-۲)$$

در واقع در این تابع احتمال رخداد کلمه w_t را با توجه به کلمات همسایگی آن با پنجره به طول c در نظر می‌گیرد.

برای مدل Skip-gram، کلمات زمینه محلی به طور مستقل برای کلمه هدف، پیش بینی می‌شوند. بنابراین تابع هدف برابر است با:

$$\frac{1}{T}\sum_{t=1}^{T}\log p(w_{t+1}|\sum_{-c\le j\le c, j\neq 0} w_t) \quad (۶-۲)$$

برای محاسبه‌ی احتمال $p(w_{t+j}|w_t)$ میتوان از تابع سافت مکس[۱] استفاده کرد:

$$p(w_o|w_I) = \frac{\exp({v'_{w_o}}^T v_{w_I})}{\sum_{w=1}^{W}\exp({v'_w}^T v_{w_I})} \quad (۷-۲)$$

که v_w و v'_w بازنمایی برداری کلمه w هستند و W نیز تعداد کل کلمات است. البته این فرمول از نظر محاسباتی بسیار سنگین است زیرا تعدا کلمات یک پیکره بزرگ که معمولا برای آموزش مدل استفاده می‌شود، بسیار زیاد است و محاسبات آن نیز زمانبر خواهد بود. به همین منظور دو استراتژی پیشنهاد شده است که برآورد مناسبی را برای تابع سافت مکس ارائه می‌کند: «سافت مکس سلسله مراتبی»[۲] و «نمونه‌برداری منفی»[۳] که شرح آنها در مقاله اصلی میکولوف آمده است(Mikolov, Sutskever, et al. 2013).

برای الگوریتم کلمه-به-بردار، پارامترهای متعددی جهت آموزش مدل وجود دارد.

- اندازه پنجره: اندازه پنجره، تعداد کلمات همسایگی کلمه هدف، زمینه محلی، برای آموزش مدل را تعیین می‌کند. افزایش اندازه پنجره ممکن است شامل کلماتی شود که با کلمه هدف مرتبط نیستند و در نتیجه زمان آموزش مدل را طولانی‌تر و کاهش اندازه پنجره منجربه حضور ایست واژه می‌شود. بنابراین انتخاب صحیح اندازه پنجره تاثیر بسزایی در عملکرد مدل دارد.

---

[۱] Softmax

[۲] Hierarchical Softmax

[۳] Negative Sampling

- ابعاد بردار[۱]: تعداد ابعاد بردار، اندازه لایه پروجکشن را نشان می دهد.
- نمونه‌برداری[۲]: این پارامتر به‌منظور حذف کلماتی که فراوانی بیشتری از حد آستانه مشخص دارند، می‌باشد. به‌عبارت‌دیگر، این کلمات پرتکرار همانند ایست واژه بار معنایی برای استخراج روابط معنایی بین کلمات ندارند. احتمال این کلمات به صورت زیر محاسبه می‌شود.

$$P(w_t) = 1 - \sqrt{\frac{t}{f(w_t)}} \qquad (۸-۲)$$

که در آن صورت کسر t ، حدآستانه، و مخرج کسر $f(w_t)$ ، فراوانی کلمات در کل پیکره متنی است.

- کمترین فراوانی[۳]: این پارامتر به حذف کلماتی که کمتر از مقدار خاصی در کل پیکره متنی ظاهر شده است، می‌پردازد.
- نمونه‌برداری منفی: هدف از نمونه‌برداری منفی، افزایش سرعت زمان آموزش است. به‌عبارت‌دیگر، در هرتکرار و بروزرسانی وزن‌های شبکه، به‌جای به‌روزرسانی بردارهای تمام کلمات، تنها بخشی ازکلمات به روزرسانی می‌شود. به‌صورت تصادفی چند نمونه از کلمات، انتخاب می‌شوند، احتمال اینکه این کلمات نمونه با کلمه‌هدف، w_t، شباهتی داشته باشند بسیار کم است. نمونه گیری منفی با هدف به حداکثر رساندن شباهت بین کلماتی که با کلمه هدف w_t در متن ظاهر می شود و به حداقل رساندن شباهت بین w_t و کلمات نمونه است.

۲-۶. جمع‌بندی

در این بخش مروری بر مفاهیم و مهمترین پژوهش‌ها در حوزه‌های زیر ارائه شده است:

- سیستم‌های پیشنهاددهنده
- سیستم‌های پیشنهاددهنده متنی
- استدلال نمونه‌محور
- سیستم‌های پیشنهاددهنده براساس استدلال نمونه محور
- روش‌های بازیابی و شباهت متون
- روش‌های استخراج کلیدواژه

---

[۱] Vector Dimensions

[۲] Subsampling

[۳] Min-count

همانطور که عنوان شد، هدف از انجام این پژوهش، ارائه روشی است که بتواند برای یک سند علمی فارسی (پایان‌نامه و رساله)، کلیدواژه‌های مناسب را به نمایه‌ساز پیشنهاد دهد. این کلیدواژه‌ها از بین کلیدواژه‌های استفاده شده توسط نمایه‌سازان که در مجموعه داده‌ها موجود است استخراج می‌شود.

با توجه به دامنه کار این طرح و ویژگی‌های آن، از بین انواع سیستم‌های پیشنهاددهنده، تمرکز این طرح بر استفاده از سیستم پیشنهاددهنده محتوی محور است و از بین انواع استدلال‌ها در سیستم‌های پیشنهاددهنده محتوی محور، استدلال نمونه محور استفاده خواهد شد. علت این امر اینست که در این طرح، برای استخراج کلیدواژه برای یک سند علمی تنها به اطلاعات آماری آن سند اکتفا نخواهد شد بلکه قرار است با اتکا به دانش ضمنی موجود در مجموعه عظیم نمونه داده‌های نمایه‌سازی شده، کلیدواژه‌های جدید پیشنهاد شود. بنابراین لازم است که از روش‌ها تعیین شباهت بین دو متن (شباهت بین سند جدید با سایر سندهای نمایه‌شده در پایگاه داده) استفاده شود.

یک سند جدید علاوه بر متنی که محتوی آن را توصیف می‌کند، یعنی عنوان و چکیده آن، دارای اطلاعاتی نظیر موضوع سند و کلیدواژه‌های اختصاص داده شده به آن (توسط نویسنده) نیز هست. بنابراین برای مقایسه بین دو سند می‌توان علاوه بر مقایسه محتوی آنها، از این اطلاعات نیز استفاده نمود.

در فصل بعدی، با بهره‌گیری از روش‌های بازنمایی کلمات و نیز استدلال نمونه‌محور، روشی برای پیشنهاد کلیدواژه برای یک سند پیشنهاد می‌شود.

فصل سوم:

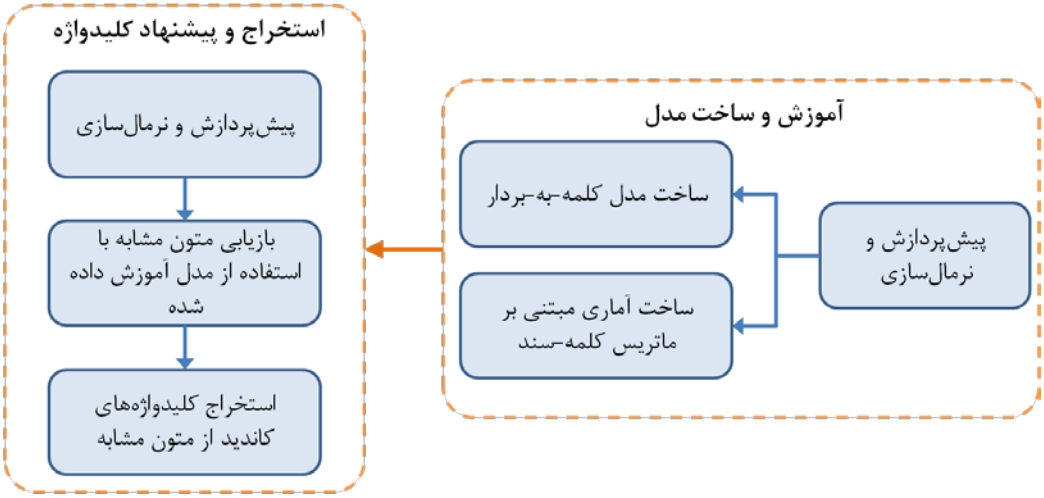
روش پیشنهادی

۳. روش پیشنهادی

روش پیشنهادی برای استخراج کلیدواژه برای یک متن، براساس سیستم‌های پیشنهاددهنده عمل می‌کند و شیوه استنتاج آن نیز براساس استدلال نمونه محور است که پیشتر درباره آن در فصل ۲ توضیح داده شد. در واقع در روش پیشنهادی، کلیدواژه‌های پیشنهادی برای یک سند، از متن سند استخراج نمی‌شود، بلکه براساس میزان شباهت سند با اسناد قبلی، از دانش موجود در اسنادی که کلیدواژه به آنها اختصای داده شده، کلیدواژه‌های مناسب استخراج می‌شوند. در ادامه این بخش روش پیشنهادی و بخش‌های مختلف آن به تفصیل توضیح داده می‌شود.

۳-۱. اجزای اصلی روش پیشنهادی

روش پیشنهادی براساس معماری سیستم‌های پیشنهاددهنده عمل می‌کند. روش پیشنهادی دو بخش اصلی دارد: ۱) آموزش و ساخت مدل، و ۲) پیشنهاد کلیدواژه برای یک سند جدید، که در شکل ۳-۱ نمایش داده شده است.



شکل ۳-۱- ساختار روش پیشنهادی

در بخش اول، با توجه به مجموعه پیکره اسناد نمایه‌شده (یعنی اسنادی که به آنها کلیدواژه اختصاص داده شده است)، مدل بازنمایی کلمات، براساس مدل بردار-کلمه (Word2Vec) ساخته می‌شود. از این مدل برای تعیین شباهت بین اسناد مختلف در مرحله بازیابی و نیز استخراج و پیشنهاد کلیدواژه‌های مناسب استفاده خواهد است. خروجی این مدل، بازنمایی برداری چندبُعدی برای هر کلمه در پیکره است. مسلما پیش از آموزش مدل، نیاز است که پیش‌پردازش‌هایی برای نرمال‌سازی متن روی متون پیکره انجام شود.
بخش دوم، که بیانگر کارکرد اصلی روش پیشنهادی است، خود از چندین قسمت تشکیل شده است و در آن برای یک سند جدید، مجموعه‌ای از کلیدواژه‌ها پیشنهاد می‌شود. شیوه استخراج و پیشنهاد کلیدواژه در این بخش، براساس استدلال نمونه محور انجام می‌شود. در این بخش، پس از نرمال‌سازی و پیش‌پردازش سند جدید، اسناد مشابه با آن با استفاده از مدل آموزش داده شده در بخش اول، بازیابی می‌شوند. سپس مجموعه‌ای از کلیدواژه‌ها از بین کلیدواژه‌های اسناد بازیابی شده، براساس میزان شباهتشان با سند جدید، انتخاب و پیشنهاد می شوند. بنابراین بخش دوم خود شامل قسمت‌های زیر است:

- پیش‌پردازش و نرمال‌سازی شامل تعیین موضوع سند جدید (برای مشخص نمودن مدل مناسب برای بازیابی متون مشابه)، و پیش‌پردازش متنی سند جدید
- بازیابی اسناد مشابه و تشکیل مجموعه کلیدواژه‌های کاندید از بین اسناد مشابه
- محاسبه امتیاز کلیدواژه‌های کاندید و میزان شباهت هر یک با سند جدید
- استخراج و پیشنهاد کلیدواژه‌های مناسب برای سند جدید از بین کلیدواژه‌های کاندید

۲-۳. آموزش و ساخت مدل

در این بخش، دو مدل ساخته می‌شود: مدل آماری مبتنی بر ماتریس کلمه-سند براساس TFIDF، و مدل برداری کلمه-به-بردار برای بازنمایی برداری کلمات ایجاد می‌شود تا از طریق آن بتوان شباهت بین متون و کلمات برای بازیابی اسناد مشابه، و نیز شباهت بین یک کلیدواژه کاندید با سند جدید را محاسبه نمود.
با توجه به آنکه اسناد مطرح برای پیشنهاد کلیدواژه، اسناد تخصصی هستند، نیاز است که ابتدا مجموعه اسناد از نظر موضوعی به دسته‌های مختلف تقسیم‌بندی شوند و فرایند آموزش برای هر دسته از اسناد به صورت جداگانه انجام شود. در نهایت برای هر دسته از اسناد، مدل خاص مربوط به موضوع آن دسته حاصل می‌شود. در مسئله مربوط به اسناد تخصصی علمی مانند اسناد پایان‌نامه‌هاو رساله‌ها، معمولا هر سند یک برچسب موضوعی کلان دارد که بر همان اساس می‌توان اسناد را پیش از انجام فرایند آموزش به دسته‌های مختلف دسته‌بندی نمود. در صورتیکه چنین برچسب موضوعی فراهم نباشد، نیاز است که با استفاده از روش‌های دسته‌بندی خودکار، اسناد دسته‌بندی شوند. در این پژوهش براساس تنوع داده‌های موجود، چهار دسته فنی-مهندسی، علوم انسانی، هنر و ادبیات، و علوم پایه در نظر گرفته شده است.

پس از تشکیل دسته‌های موضوعی از اسناد، نیاز است که برای هر دسته، به صورت تصادفی، سه مجموعه داده مشخص شود: مجموعه داده آموزش، مجموعه داده اعتبارسنجی، و مجموعه داده آزمایش. در واقع در این مرحله تنها از مجموعه داده آموزش برای آموزش مدل استفاده می‌شود و مجموعه داده اعتبارسنجی برای تنظیم پارامترهای مدل و اعتبارسنجی آن بکار گرفته خواهد شد و مجموعه داده آزمایش نیز برای آزمایش نهایی روش پیشنهاد کلیدواژه و اعلام دقت روش مورد استفاده قرار می‌گیرد.
همانطور که در فصل دوم اشاره شد، مدل بردار-به-کلمه از پارامترهای مختلفی تشکیل شده است که تعیین مقادیر آنها برای آموزش مدل نیازمند انجام آزمایش‌های مکرر است که شیوه تعیین آنها در فصل پیاده‌سازی و ارزیابی تشریح می‌شود.
پیش از آموزش مدل، نیاز است که عملیات پیش‌پردازش روی داده‌ها انجام گیرد. طی این عملیات ابتدا ایست واژه حذف می‌گردد، و سپس با استفاده از یک تابع lemmatizer عملیات زیر روی متون اعمال می‌گردد:

- اصلاح نویسه‌ها
- اصلاح نشانه‌گذاری
- رعایت فاصله و نیم‌فاصله
- حذف اعراب و تشدید
- حذف علایم نگارشی (غیر از نقطه)
- استخراج جملات
- ریشه‌یابی واژه‌ها

یک نمونه از عملکرد پیش‌پردازش در جدول ۳-۱ آمده است. در این نمونه، پس از حذف ایست واژه از lemmatizer کتابخانه ابزار hazm که برای پردازش زبان طبیعی فارسی کاربرد دارد، استفاده شده است.

جدول ۳-۱- نمونه متن پیش‌پردازش شده

متن اولیه	متن پس از پیش پردازش
مقایسه سبک های تفکر، رفتار سازمانی و عملکرد مدیران انتخابی و انتصابی هدف از پژوهش حاضر، مقایسه سبک های تفکر، رفتارسازمانی و عملکرد مدیران انتخابی و انتصابی در دانشگاه سیستان و بلوچستان است .روش پژوهش حاضر توصیفی _ پیمایشی و از نوع مقایسه ای می‌باشد که جامعه ی آماری شامل کلیه ی مدیران گروههای آموزشی)انتخابی (و مدیران و معاونین دانشکده های دانشگاه سیستان وبلوچستان)انتصابی (در سال تحصیلی 92-93به تعداد50 نفر می باشد.یافته‌های پژوهش حاکی از آن است که بین سبک های تفکر و رفتار شهروندی سازمانی، بین سبک های تفکر و عملکرد مدیران، بین رفتار شهروندی و عملکرد مدیران، تفاوت معناداری وجود دارد .بررسی های انجام شده حاکی از آن است که با مدیریت موثر و تقسیم کار مناسب که ویژگی های فردی و شخصیتی افراد را در نظر گرفته باشد می توان شرایطی ایجاد کرد که افراد نقش خود را با کیفیت و کمیتی فراتر از آنچه که تعریف شده و انتظار می رفته ایفا نموده و به عنوان شهروند سازمانی متخصص و متعهد به تمامی ارزش ها و آرمان سازمان پایبند و در راستای تأمین انتظارات ذینفعان گام بردارند.	مقایسه سبک تفکر رفتار سازمان عملکرد مدیر انتخاب انتصاب هدف پژوهش مقایسه سبک تفکر رفتارسازمانی عملکرد مدیر انتخاب انتصاب دانشگاه سیستان بلوچستان روش پژوهش توصیف _ پیمایش مقایسه جامعه آمار شامل کلیه مدیر گروه آموزش انتخاب مدیر معاونین دانشکده دانشگاه سیستان وبلوچستان انتصاب تحصیل ۹۳ - ۹۲ تعداد ۵۰ نفر یافته پژوهش حاکی سبک تفکر رفتار شهروند سازمان سبک تفکر عملکرد مدیر رفتار شهروند عملکرد مدیر تفاوت معنادار بررسی حاکی مدیریت موثر تقسیم کار مناسب ویژگی فرد شخصیت افراد نظر گرفت شرایط ایجاد افراد نقش کیفیت کمیت فراتر تعریف انتظار می‌رفته ایفا نموده شهروند سازمان متخصص متعهد تمام ارزش آرمان سازمان پایبند راستا تأمین انتظارات ذینفع

گرچه در برخی از مدل‌های کلمه-به-بردار، ایست واژه حذف نمی‌شود، لیکن علت حذف ایست واژه پیش از آموزش مدل اینست که این کلمات بار معنایی برای مشخص نمودن محتوی یک سند ندارند و اضافه کردن آنها تنها نویز مدل وتعداد واژگان آموزش دیده شده در مدل را افزایش می‌دهد. در بسیاری از پژوهش‌ها نیز بسته به ماهیت مسئله، ایست واژه پیش از آموزش مدل حذف شده است.

این نکته قابل ذکر است که هرگونه عملیاتی که پیش از آموزش مدل روی داده‌ها انجام می‌شود، لازم است در بخش بعدی نیز عینا برای یک سند جدید، پیش از هر کاری اعمال شود.

۳-۳. استخراج و پیشنهاد کلیدواژه

پس از آموزش مدل، در بخش قبل، برای هر سند جدید، پنج کار اصلی انجام می‌شود، تا در نهایت کلیدواژه‌های مناسب پیشنهاد گردد. در ادامه هر یک از این مراحل تشریح می‌گردد.

۳-۳-۱. پیش‌پردازش و نرمال‌سازی

همانگونه که در شکل ۳-۱ نیز نمایش داده شده، برای هر سند جدید ابتدا موضوع آن مشخص می‌شود، تا معلوم شود که در نهایت از چه مدل و مجموعه داده‌ای برای بازیابی کلمات کاندید استفاده گردد. سپس عملیات پیش‌پردازش روی سند انجام می‌گیرد که این عملیات عیناً همان فرایندی است که در مرحله آموزش، پیش از آموزش مدل، روی داده‌ها صورت گرفته است.

۳-۳-۲. بازیابی متون مشابه

برای بازیابی متون مشابه از دو مدل مختلف استفاده شده است، که یکی براساس ارتباط معنایی عمل می‌کند و دیگری براساس مدل‌های آماری مبتنی بر ماتریس کلمه–سند برمبنای TF-IDF عمل می‌کند:

- مدل برداری معنایی، مبتنی بر مدل کلمه–به–بردار
- مدل برداری آماری، مبتنی بر TF-IDF

در مدل معنایی برداری، شباهت بین دو متن، براساس فاصله بین نمایش برداری دو متن که مطابق با مدل کلمه–به–بردار بدست آمده، محاسبه می‌شود. بنابراین اگر سند d_i پس از پیش‌پردازش به کلمات $W_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$ تبدیل شده باشد، آنگاه برای هر کلمه یک بازنمایی برداری براساس مدل کلمه–به–بردار وجود دارد. بنابراین بردار متن d_i به صورت زیر محاسبه می‌شود:

$$V_{avg}(d_i) = \frac{1}{n}\sum_{k=1}^{n} V_{W2V}(w_{ik}) \qquad (۱-۳)$$

که $V_{W2V}(w_{ik})$ همان بردار متناظر با کلمه w_{ik} در مدل کلمه–به–بردار است. در واقع در فرمول بالا، بردار هر متن براساس میانگین بردار کلمات تشکیل دهنده آن متن محاسبه می‌شود.

در نهایت شباهت بین دو متن با استفاده از شباهت کسینوسی بین دو بردار آنها محاسبه می‌شود:

$$Sim_{W2V}(d_i, d_j) = \cos\big(V_{avg}(d_i), V_{avg}(d_j)\big) \qquad (۲-۳)$$

مدل برداری آماری، براساس ماتریس کلمه–سند عمل می‌کند که هر درایه آن بیانگر مقدار TFIDF هر کلمه در هر سند است و براساس فرمول (۲-۴) محاسبه می‌گردد. در نهایت شباهت بین دو سند، براساس شباهت کسینوسی بین بردارهای آنها در ماتریس کلمه–سند، بدست می‌آید.

با تجمیع شباهت بین دو سند در هر روش، در نهایت داریم:

$$Sim_{Final}(d_i, d_j) = (1-\propto)\ Sim_{W2V}(d_i, d_j) + \propto\ Sim_{TFIDF}(d_i, d_j) \qquad (۳-۳)$$

که $0 \leq \propto \leq 1$ مشخص می‌کند که تا چه میزان در محاسبه شباهت نهایی به شباهت معنایی اهمیت می‌دهیم. هر چه $\propto$ بزرگتر باشد، وزن روش معنایی که براساس مدل کلمه-به-بردار عمل می‌کند، بیشتر خواهد بود.

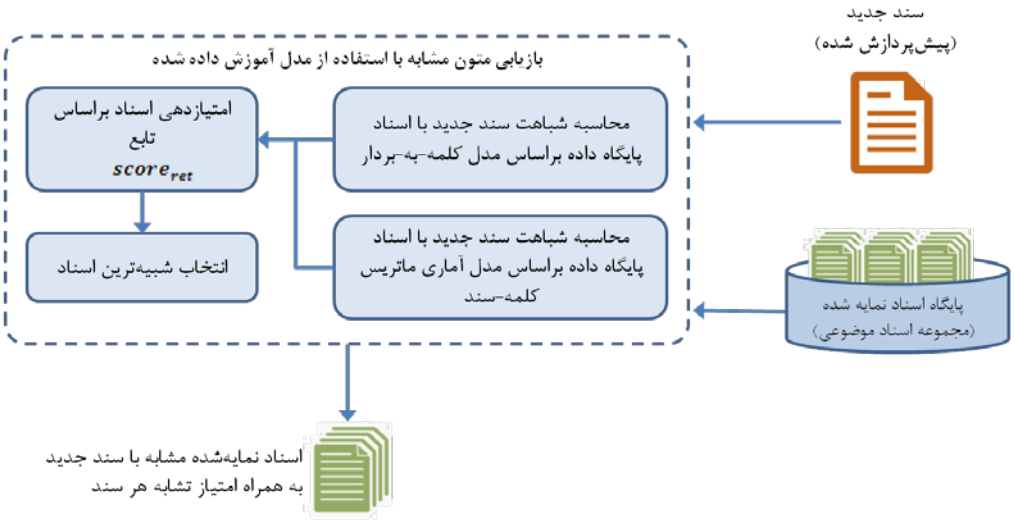
در نهایت برای هر سند جدید d_{new} با توجه به میزان شباهت آن با سایر اسناد، مجموعه ای از M سندی که بیشترین تشابه را دارند به همراه میزان تشابه هر یک، حاصل می‌شود:

$$\boldsymbol{\mathcal{M}_{new} = \{\langle d_1^*, score(d_1^*)\rangle, \langle d_2^*, score(d_2^*)\rangle, \dots, \langle d_M^*, score(d_M^*)\rangle\}} \qquad (۴-۳)$$

که اسناد براساس میزان شباهتشان از ۱ تا M مرتب شده اند و $score_{ret}(d_m^*)$ بیانگر میزان شباهت بین سند جدید با سند d_m^* در فرایند بازیابی است:

$$\boldsymbol{score_{ret}(d_m^*, d_{new}) = Sim_{Final}(d_m^*, d_{new})} \qquad (۵-۳)$$

در شکل شکل ۳-۲ فرایند و ورودی و خروجی‌های این بخش نمایش داده شده است.

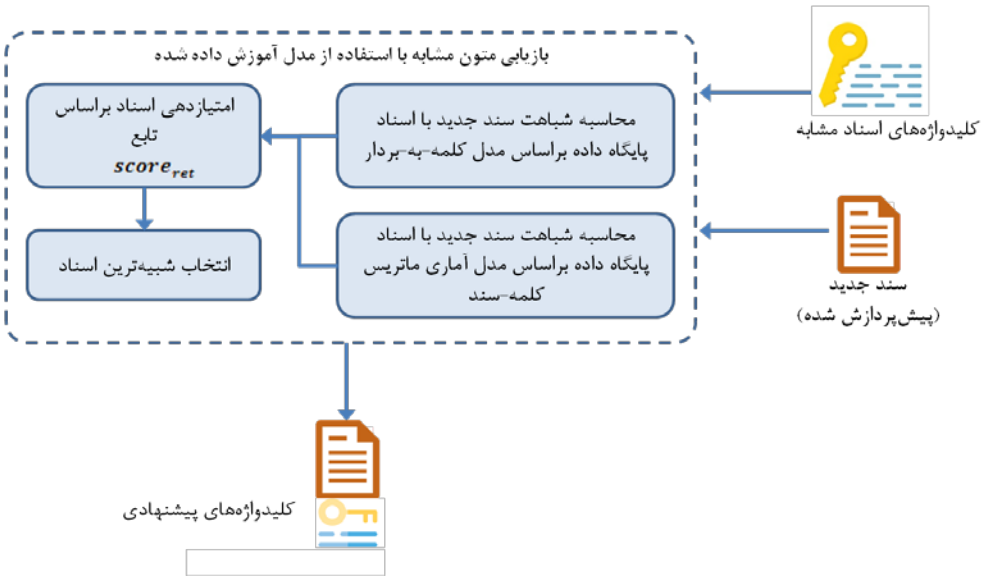


شکل ۳-۲- فرایند بازیابی اسناد مشابه

لازم به ذکر است که در این مرحله تنها عنوان و چکیده هر سند در نظر گرفته می‌شود و کلیدواژه‌ها در فرایند بازیابی دخیل نمی‌شوند.

۳-۳-۳. استخراج کلیدواژه‌های کاندید

هر یک از اسناد مشابه بازیابی شده در مرحله قبل، کلیدواژه‌هایی دارند. در این مرحله، کلیدواژه‌های همه اسناد مشابه، استخراج شده و براساس دو روش امتیازدهی، رتبه‌بندی می‌شوند که در ادامه تشریح می‌گردد. در شکل شکل ۳-۳ فرایند و ورودی و خروجی‌های این بخش نمایش داده شده است.



شکل ۳-۳- فرایند امتیازدهی و پیشنهاد کلیدواژه‌ها

۳-۳-۳-۱. روش امتیازدهی براساس امتیاز بازیابی

در این روش امتیاز هر کلیدواژه براساس میزان شباهت اسناد دربرگیرنده آن کلیدواژه محاسبه می‌شود. بنابراین برای هر کلیدواژه یک بردار، $V_{key}(k_l)$ ، M تایی بر مبنای مدل تک-مولفه‌ای باینری در نظر گرفته می‌شود که هر مولفه از آن می‌تواند مقدار ۰ یا ۱ را اتخاذ کند. هر مولفه از این بردار متناظر با یک سند از اسناد $\mathcal{M}_{new}$ است و مقدار صفر یا یک برای آن بیانگر اینست که کلیدواژه متعلق به سند هست یا خیر، بنابراین:

$$V_{key}(k_l) = (v_{1l}, \dots, v_{Ml}) \qquad (۶-۳)$$

$$v_{mk} = \begin{cases} 1 & if\ k_l\ belongs\ to\ d_m^* \\ 0 & otherwise \end{cases}$$

از طرفی هر سند نیز دارای یک امتیاز است که مطابق با فرمول (۳-۵) محاسبه شده است. بنابراین می‌توان یک بردار M تایی از امتیازها برای سند جدید در نظر گرفت:

$$\boldsymbol{V_{ret}} = (\boldsymbol{score}(\boldsymbol{d_1^*}), \dots, \boldsymbol{score}(\boldsymbol{d_M^*})) \quad (۷-۳)$$

حال می‌توان یک تابع امتیاز براساس ضرب نقطه‌ای بین دو بردار $\boldsymbol{V_{ret}}$ و $\boldsymbol{V_{key}}(\boldsymbol{k_l})$ در نظر گرفت که بیانگر امتیاز هر کلمه کلیدی باشد. براین اساس، کلماتی که در اسناد شبیه‌تر هستند، امتیاز بیشتری خواهند داشت. از طرفی، کلیدواژه‌ایی که در تعداد اسناد بیشتری از بین اسناد بازیابی شده هستند، نیز نیاز است که وزن بیشتری را نسبت به سایر کلیدواژه‌ها داشته باشند. بنابراین در محاسبه امتیاز نهایی می‌توان ضریبی را برای این منظور لحاظ نمود. بنابراین تابع امتیاز کلیدواژه ه$\boldsymbol{k_l}$ براساس امتیاز بازیابی به صورت زیر است:

$$\boldsymbol{score_1}(\boldsymbol{k_l}, \boldsymbol{d_{new}}) = \frac{\boldsymbol{V_{ret}} \cdot \boldsymbol{V_{key}}(\boldsymbol{k_l})}{|\boldsymbol{V_{ret}}|} \ \log(\boldsymbol{f}(\boldsymbol{k_l}) + \frac{\boldsymbol{1}}{\boldsymbol{4}}\boldsymbol{M}) \quad (۸-۳)$$

که $f(k_l)$ بیانگر تعداد اسناد از بین M سند بازیابی شده است، که k_l کلیدواژه آنها بوده است. . در واقع با ضرب عبارت فوق در لگاریتم، می‌خواهیم وزن کلیدواژه‌هایی را که در بیش از یک چهارم اسناد بازیابی شده آمده‌اند، بیشتر در نظر بگیریم. همانطور که از شیوه محاسبه مشخص است، در این روش، امتیاز یک کلیدواژه تنها براساس امتیاز شباهت اسناد دربردارنده آن کلیدواژه محاسبه می‌شود و ماهیت معنایی آن درنظر گرفته نمی‌شود.

۳-۳-۳-۲. روش امتیازدهی براساس شباهت با متن جدید

در روش دوم، مفهوم کلیدواژه و ارتباط معنایی آن با سند جدید به عنوان معیاری برای محاسبه امتیاز کلمه درنظر گرفته می‌شود. در این روش، هر کلیدواژه از نظر معنایی با تک تک واژگان سند جدید از نظر معنایی، براساس مدل برداری کلمه-به-بردار مقایسه می‌شود.

برای این منظور برای هر کلیدواژه یک بردار معنایی براساس مدل آموزش داده شده کلمه-به-بردار، درنظر گرفته می‌شود.

هر کلیدواژه ممکن است از چندین کلمه تشکیل شده باشد. بنابراین نیاز است که ابتدا عملیات پیش‌پردازش، همانند آنچه در بخش پیش‌پردازش و نرمال‌سازی گفته شد، روی آن انجام شود.

در صورتیکه مدل کلمه-به-بردار آموزش دیده، کلیدواژه را نداشته باشد، آنگاه نمی‌توان امتیازی براساس این روش برای آن حساب نمود و مقدار NAN برای آن لحاظ می‌شود.

در صورتیکه کلیدواژه، پس از پیش‌پردازش بیش از یک کلمه داشته باشد، آنگاه برای هر یک از کلمات آن (در صورت موجود بودن) براساس مدل کلمه-به-بردار، یک نمایش برداری وجود خواهد داشت و نمایش برداری هر کلیدواژه در نهاست میانگین بردار کلمات تشکیل دهنده آن خواهد بود.

بنابراین بردار نمایش معنایی برای هر کلیدواژه $\boldsymbol{k_l}$ به صورت زیر محاسبه می‌شود:

$$\boldsymbol{V_{sem}(k_l)} = \begin{cases} \boldsymbol{V_{W2V}(k_l)}, & \boldsymbol{if\ k_l\ is\ a\ single\ word\ and\ W2V\ model\ exists} \\ \boldsymbol{\frac{1}{I}\sum_{x_i\ in\ k_l} V_{W2V}(x_i)}, & \boldsymbol{if\ k_l\ is\ a\ multi\ word} \\ \boldsymbol{NAN}, & \boldsymbol{otherwise.} \end{cases} \quad (۹-۳)$$

که $\boldsymbol{I}$ بیانگر تعداد کلمات تشکیل دهنده یک کلیدواژه پس از انجام عملیات پیش‌پردازش است.
اگر بردار کلید واژه NAN نباشد، آنگاه می‌توان آن را بردار با سند جدید d_{new} مقایسه نمود. هر یک از واژگان سند جدید، پس از پیش‌پردازش، یک نمایش برداری براساس مدل آموزش دیده کلمه–به–بردار دارند. واژگانی که در مدل موجود نباشند، در نظر گرفته نمی‌شوند.
کلیدواژه‌هایی که بردار آنها NAN است، در واقع حکم داده‌های گم شده[۱] را دارند که مدل برای آنها آموزش داده نشده و اطلاعاتی درباره آنها موجود نیست. برای لحاظ کردن این نوع از داده‌ها در مسائل یادگیری ماشین رویکردهای متنوعی وجود دارد. یکی از این رویکردها براساس توزیع آماری داده است (García-Laencina et al. 2010). در همین راستا برای تعیین مقادیر این بردار، توزیع این مقدار برای مجموعه داده‌های آموزشی بررسی شده، و به صورت تصادفی از آن توزیع نمونه‌برداری شده است.
اگر واژگان d_{new} را که مدل کلمه–به–بردار هم برای آنها موجود است، داخل مجموعه $\mathcal{R}_{\text{new}}$ قرار گیرد، آنگاه می توان شباهت معنایی هر واژه با کلیدواژه k_l، براساس مدل برداری آنها محاسبه کرد و در نهایت امتیاز هر کلیدواژه براساس میانگین این شباهت‌ها حساب شود. بنابراین:

$$\boldsymbol{score_2(k_l, d_{new}) = \frac{1}{|\mathcal{R}_{new}|} \sum_{\mathbf{r}\in\mathcal{R}_{\text{new}}} \cos(\, V_{sem}(k_l),\, V_{W2V}(r)\,)} \quad (۱۰-۳)$$

٤-۳-۳. پیشنهاد کلیدواژه براساس امتیاز نهایی

دو امتیاز بدست آمده برای هر کلیدواژه در نهایت در هم ضرب می‌شود و امتیاز نهایی حاصل می‌شود:

$$\boldsymbol{score_{final}(k_l) = score_1(k_l, d_{new}) \times score_2(k_l, d_{new})} \quad (۱۱-۳)$$

[۱] Missing data

در نهایت لازم است که کلیدواژه‌های اصلی از بین کلیدواژه‌های کاندید به توجه به امتیاز نهایی آنها، انتخاب شوند. یکی از روش‌های انتخاب می‌تواند براساس تعداد باشد، به این مفهوم که از بین موارد کاندید، تعداد مشخصی از کلیدواژه‌ها که بیشترین امتیاز را دارند، انتخاب شوند.

روش دیگر اینست که براساس مقدار امتیاز نهایی کلیدواژه‌ها عمل شود و کلیدواژه‌هایی که امتیازشان از یک حد آستانه بیشتر است، انتخاب شوند.

برای تعیین آستانه امتیاز یا تعداد مشخص کلیدواژه‌ها نیاز است که آزمایش‌های متعددی براساس مجموعه داده‌های اعتبارسنجی انجام شود تا مقادیر این دست از پارامترها مشخص شود.

فصل چهارم:

آماده‌سازی و پیش‌پردازش داده‌ها

٤. آماده‌سازی و پیش‌پردازش داده‌ها

برای پیاده‌سازی و ارزیابی روش پیشنهاد کلیدواژه برای مستندات علمی فارسی، از مجموعه‌ای از پایان‌نامه‌های کارشناسی ارشد و رساله‌های دکتری (پارسا) کشور استفاده شده است. این مجموعه شامل ۳۱۸۴۰ پایان‌نامه و رساله است. هر رکورد داده‌ای در این مجموعه شامل اطلاعات فراداده‌ای مربوط به یک پایان‌نامه یا رساله است که از بین آنها، اقلام اطلاعاتی زیر در این پژوهش مورد استفاده قرار می‌گیرد:

- عنوان پارسا
- موضوع پارسا (موضوع اصلی و موضوع فرعی)
- کلیدواژه‌های پارسا (کلیدواژه‌هایی که نمایه‌ساز تخصیص داده و کلیدواژه‌های نویسنده)
- خلاصه پارسا

مدل داده‌ای این مجموعه داده همان مدل داده‌ای پایگاه گنج است که در محیط SQL قابل دسترسی است. هر رکورد اطلاعاتی می‌تواند تعدادی نامحدودی کلیدواژه داشته باشد و این تعداد برای پارساهای مختلف، متفاوت است. هر کلیدواژه یکی از انواع زیر است:

- کلیدواژه نویسنده: منظور کلیدواژه‌ای است که توسط نویسنده پارسا هنگام ثبت آن سند در گنج ثبت شده است.
- کلیدواژه کنترل‌شده: کلیدواژه‌ای که توسط نمایه‌ساز برای سند تخصیص یافته است.
- کلیدواژه اصطلاح‌نامه: کلیدواژه‌ای که از اصطلاح‌نامه‌های تخصصی استخراج شده و توسط نمایه‌ساز برای سند ثبت شده است.

در جدول ۴-۱ اطلاعات آماری مربوط به این مجموعه داده آمده است.

جدول ٤-١-مشخصات آماری مجموعه داده‌ها

تعداد کل رکوردها (پارساها)	۳۱۸۴۰
تعداد کلیدواژه‌ها	۴۶۶۰۳
تعداد موضوعات اصلی	۴
تعداد موضوعات فرعی	۱۵۲
تعداد کل رکوردهای با حداقل یک کلیدواژه کنترل‌شده یا کلیدواژه اصطلاحنامه	۲۵۷۲۶

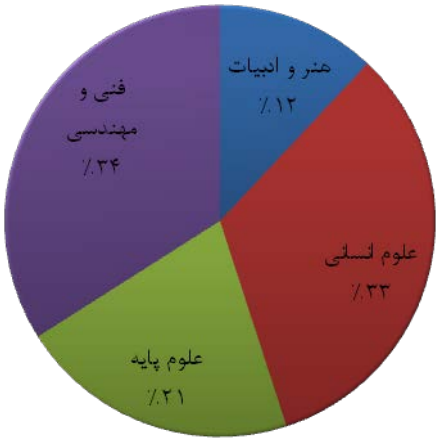
برای آماده‌سازی داده‌ها دو عملیات «مرتب‌سازی موضوعات» و «پاکسازی و استانداردسازی داده‌ها» انجام شده است. مرتب‌سازی موضوعات برای تشکیل سلسه‌مراتب موضوعی انجام شده و پاکسازی و استانداردسازی داده‌ها نیز برای نرمال کردن متون واستانداردسازی کاراکترهای فارسی صورت گرفته است. توضیحات مربوط به هر عملیات در بخش جداگانه در ادامه آمده است.

٤-١. مرتب‌سازی موضوعات داده‌ها

برای بهبود عملیات بازیابی و تعیین شباهت متون، لازم است که یک دسته‌بندی موضوعی از داده‌ها وجود داشته باشد که البته اکثر داده‌های گنج دارای این دسته‌بندی موضوعی هستند. برای یکسان‌سازی توزیع داده‌ها در هر موضوع، دسته‌بندی موضوعی جدیدی روی داده‌ها انجام شده که شامل دو سطح است:

- سطح اول یا موضوعات اصلی شامل: فنی و مهندسی، علوم پایه، علوم انسانی، و هنر و ادبیات
- سطح دوم یا موضوعات فرعی: هر موضوع اصلی شامل مجموعه‌ای از موضوعات فرعی است که در جدول ۴-۲ نمایش داده شده است.

تعیین این دسته‌بندی می‌تواند در دقت بازیابی نمونه‌ها در سیستم استدلال نمونه محور و تخصیص کلیدواژه‌های مناسب بسیار مفید باشد. توزیع پارساها در هر موضوع اصلی در شکل ۴-۱ آمده است.



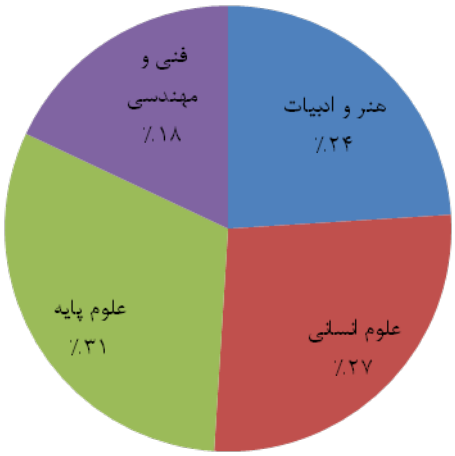
شکل ٤-١- توزیع پارساها در موضوعات اصلی

جدول ٤-۲- دسته‌بندی موضوعی، سطح اول و سطح دوم

سطح اول (موضوع اصلی)	سطح دوم (موضوع فرعی)		
فنـــــــی و مهندسی (۲۷ موضـــوع فرعی)	• کشاورزی، عمومی و میان رشته‌ای • مهندسی برق و الکترونیک • مهندسی مکانیک • علوم کامپیوتر،سخت افزار و معماری • زمین شناسی مهندسی • مهندسی عمومی و میان رشته‌ای • مهندسی بافت و سلول • علوم کامپیوتر،مهندسی نرم افزار • رباتیک و کنترل	• مهندسی شیمی • مهندسی صنایع • مهندسی متالوژی • مکانیک • مهندسی دریایی • مهندسی نفت • فناوری اطلاعات • مهندسی عمران • مهندسی مصنوعات • تحقیق در عملیات	• علوم کامپیوتر، هوش مصنوعی • علوم کامپیوتر، سیستمهای اطلاعاتی • مهندسی محیط زیست • مخابرات، ارتباطات راه دور • مهندسی هوا و فضا • مهندسی زیست پزشکی • کشاورزی، مهندسی و ماشین آلات • مهندسی صنایع
علوم پایه (٤۷ موضـــوع فرعی)	• فیزیک عمومی و میان رشته ای • شیمی معدنی و هسته ای • فیزیک اتمی، مولکولی و شیمیایی • بیوشیمی و زیست شناسی ملکولی • علوم کامپیوتر،کاربردهای میان رشته ای • زیست شناسی دریایی و آب شیرین • علوم مواد و مواد زیستی • زیست شناسی ریاضی و کامپیوتری • علوم مواد. مواد ترکیبی • علوم مواد. میان رشته ای • علوم انسانی (زیست پزشکی) • علوم مواد. منسوجات • علوم مواد، نمونه سازی و قطعات	• ریاضیات • شیمی آلی • زیست شناسی • زمین شناسی • شیمی تجزیه • ریاضیات کاربردی • شیمی فیزیک • آمار و احتمالات • فیزیک مواد چگال • فیزیک ذرات و میدانها • فیزیک هسته ای • شیمی کاربردی • شیمی عمومی و میان رشته‌ای • فیزیک کاربردی • زیست شناسی سلولی • علوم و فناوری نانو • علوم اعصاب	• علوم مواد. کاغذ و چوب • فیزیک سیالات و پلاسما • ژیوشیمی و ژیوفیزیک • زیست فناوری و میکروبشناسی کاربردی • زیست شناسی تکوینی • میکروبشناسی • علوم زمین (میان رشته ای) • بیوفیزیک • علوم پلیمر • علوم ، میان رشته ای • علوم آزمایشگاهی • علوم و فناوری هسته ای • علوم کامپیوتر، نظریه‌ها و روشها • الکتروشیمی • علوم و فناوریهای حمل و نقل • فیزیک ریاضی • شیمی

سطح اول (موضوع اصلی)	سطح دوم (موضوع فرعی)		
علوم انسانی (۴۲ موضوع فرعی)	• روانشناسی • اقتصاد کشاورزی، سیاستگذاری کشاورزی • علوم رفتاری • مدیریت • اقتصاد • حقوق • جغرافیا • آموزش و پرورش و تحقیقات تربیتی • روانشناسی تربیتی • روانشناسی عمومی و میان رشته‌ای • علوم سیاسی • تاریخ • جامعه شناسی • ارگونومیک • سوء مصرف مواد (خدمات اجتماعی)	• روانشناسی بالینی • علوم انسانی (میان رشته ای) • علوم اطلاعات ، کتابداری و دانش شناسی • ارتباط شناسی (علوم اجتماعی)، علوم ارتباطات • اوقات فراغت، سرگرمی، ورزش، گردشگری و هتلداری • جمعیت شناسی • روانشناسی کاربردی • روانشناسی ، روانکاوی • مطالعات خانواده • خدمات اجتماعی	• علوم اجتماعی (روشهای ریاضی) • حمل و نقل • مطالعات ناحیه ای • قوم شناسی • سالمندی (خدمات اجتماعی) • روانشناسی بیولوژیک • روانشناسی بالینی • جرم شناسی، بزهکاری • مردم شناسی • کسب و کار ، بازرگانی • روابط بین الملل • روانشناسی اجتماعی • مطالعات زنان • برنامه ریزی و توسعه • مطالعات فرهنگی • تاریخ علوم اجتماعی • مسایل اجتماعی
هنر و ادبیات (۳۶ موضوع فرعی)	• ادبیات فارسی • ادبیات انگلیسی • معماری • ادبیات عربی • شهرسازی • هنرهای نمایشی • نقاشی • باستانشناسی • ادبیات فرانسه • فلسفه هنر • ادبیات آلمانی • طراحی • گرافیک و چاپ	• طراحی پارچه • ادبیات کودکان و نوجوانان • سینما (فیلمنامه نویسی) • ادبیات اسپانیایی • خوشنویسی • ادبیات تطبیقی، ادبیات جهان • سینما (تدوین) • ادبیات ترکی • هنرهای نمایشی (طراحی صحنه) • مرمت و احیای بافتهای تاریخی • پژوهش در هنر	• ادبیات نمایشی • نقاشی ایرانی ، نگارگری • موسیقی ایرانی • عکاسی و فیلمبرداری • هنرهای تزیینی • موسیقی شناسی، موسیقی‌های قومی • سینما (کارگردانی) • ادبیات روسی • موسیقی کلاسیک • فرش و صنایع دستباف • هنرهای نمایشی (بازیگری) • صنایع دستی و هنرهای سنتی

در شکل ۴-۲ توزیع تعداد موضوعات سطح دوم (فرعی) در موضوعات سطح اول نمایش داده شده است.



شکل ٤-۲- توزیع تعداد موضوعات سطح دوم (فرعی) در موضوعات سطح اول (اصلی)

۲-٤. پاکسازی و استانداردسازی داده‌ها

مجموعه داده‌های استفاده شده از نظر نگارشی اشکالات فراوانی دارند که برخی از آنها عبارتند از:

- استفاده از کاراکترهای غیر معتبر در بخش‌های مختلف داده‌ها
- فیلدهای خالی موجود در جداول پایگاه داده
- غلط‌های املایی در عنوان و کلیدواژه‌ها
- استفاده از کیبوردهای غیر استاندارد برای وارد کردن اطلاعات توسط کاربران
- استفاده از عبارت‌ها به صورت چند زبانه در یک خانه از جدول پایگاه داده
- وجود یک کلمه به شکل‌های مختلف یا با کاراکترهای متفاوت در قسمت‌های مختلف
- استفاده از کدهای کاراکتری مختلف برای کاراکترهایی مانند «ی» و «ک»

برای برطرف نمودن این مشکلات یک فرایند پاکسازی اولیه روی مجموعه داده انجام شده است. بخشی از این فرایند به صورت دستی و بخشی از براساس مجموع‌ای از قوانین به صورت خودکار و بخش دیگر به صورت خودکار لحاظ شده است.

- فرایند دستی پاکسازی داده‌های کلیدواژه‌ها: فرایند دستی پاکسازی داده‌ها تنها بر روی فیلد کلیدواژه‌ها اعمال شده است. زیرا این فیلد در بخش نمایش مصور اطلاعات (که در بخش‌های بعدی توضیح داده خواهد شد استفاده خواهد شد) برای انجام این فرایند، سامانه‌ای طراحی شده

که به طور همزمان چندین کاربربه صورت دستی می‌توانند صحت داده‌ها را بررسی نمایند. براساس این فرایند، هر کاربر، برای هر کلمه کلیدی یکی از سه برچسب صحیح، نادرست و مبهم را انتخاب می‌کند. برچسب صحیح به این معنی است که نگارش کلمه صحیح است، نادرست به این معنی است که نگارش کلمه با استفاده از مجموعه‌ای از قوانین پاکسازی به صورت خودکار می‌تواند اصلاح شود و برچسب مبهم به این مفهوم است به راحتی نمی‌توان نگارش کلمه را با استفاده از قوانین خودکار اصلاح نمود. در نهایت کلمات نادرست با استفاده از قوانین خودکار اصلاح شده‌اند. استفاده از قوانین خودکار برای اصلاح کلیدواژه‌های مبهم خود نیازمند پردازش‌های جداگانه و استفاده از مدل‌های صرفی و نحو زبانی است که از دامنه پژوهشی این طرح خارج است. بنابراین برای کاهش تعداد رکوردهایی که کلیدواژه‌های مبهم دارند، دو راهکار درنظر گرفته شد:

- یک بررسی روی توزیع کلیدواژه‌های مبهم روی رکوردها انجام شد. تعداد بسیار زیادی از کلیدواژه‌های مبهم تنها یکبار یا فقط در یک رکورد آمده بودند، با حذف آنها و رکوردهای متناظر با آنها، این کلمات از پایگاه داده حذف شدند.
- از طرف دیگر موضوعات سطح دوم نیز بررسی شد، برخی از موضوعات تنها تعداد محدودی رکورد اطلاعاتی را داشتند در حالیکه کلیدواژه‌های مبهم بسیار را دربرداشتند. این موضوعات نیز به طور کامل از پایگاه داده حذف شدند.

در نهایت تعداد موضوعات سطح دوم به ۱۵۲ موضوع کاهش یافت در حالیکه تعداد رکوردها از ۳۹۹۹۷ رکورد به ۳۱۸۴۰ و تعداد کلیدواژه از ۱۴۳۲۹۲ به تعداد ۴۶۶۰۳، یعنی بیش از یک سوم حالت اولیه کاهش یافت.

- فرایندخودکار پاکسازی داده‌ها: در فرایند خودکار پاکسازی داده‌ها، با استفاده از مجموعه‌ای از قوانین، تا حد امکان، داده‌ها به یک فرمت نگارشی استاندارد تبدیل می‌شوند. این فرایند برای کلیه فیلدهای اطلاعاتی یک رکورد انجام می‌شود. مجموعه قوانینی که در این فرایند برای استانداردسازی داده‌ها بکار گرفته شده عبارتند از:
  - □ تبدیل انواع «ی» به یک نویسه با کدینگ یکسان
  - □ تبدیل انواع «ک» به یک نویسه با کدینگ یکسان

- □ حذف نمودن همزه آخر از کلمات مانند حذف همزه از املاء. همانطور که گفته شد چون این حذف هم بر منابع و هم از پرس‌وجوی کاربر اجرا می‌شود ایجاد مشکل نخواهد نمود.
- □ حذف تشدید و تنوین از همه کلمات
- □ تبدیل چند (دو یا بیشتر) فاصله به یک فاصله
- □ تبدیل فاصله انگلیسی به فاصله فارسی
- □ تبدیل أ به ا
- □ تبدیل ئ به ی
- □ تبدیل ؤ به و
- □ حذف تمامی اعراب گذاری‌ها
- □ تبدیل نیم فاصله‌ها به فرم یکسان (Shift + Ctrl + 2)
- □ حذف تمامی «ـ»(Shift+J) از عبارات به دلیل بی معنا بودن آن. برای مثال «تـــــــــا» به «تا » تبدیل می‌گردد.
- □ وارد کردن نیم فاصله در موارد خاص مثل «ها» در آخر کلمات و «می» در ابتدای افعال
- □ خذف «ۀ» از آخر کلمات
- □ تبدیل همه اعداد انگلیسی به اعداد فارسی

همانطور که ذکر شد این مراحلی بر روی همه فیلدهای منابع شامل چکیده، کلیدواژه‌ها و عنوان انجام شده است و بر روی هر سند ورودی نیز انجام می‌شود.

نکته دیگری که باید به آن توجه کرد این است که با توجه به این که پیش پردازش متن می‌تواند متن اولیه را دست خوش تغییر بسیار کرده و حتی در مواردی باعث شود که معنای متن تغییر کند لازم است که برای هر فیلد اطلاعاتی یکی فیلد اطلاعاتی دیگر اضافه شود و متن پیش پردازش شده در آن ذخیره گردد. لذا الگوریتم‌های جستجو بر روی فیلدهای پیش پردازش شده اعمال خواهند شد و الگوریتم‌های نمایش و بازیابی بر روی فیلدهای اصلی عمل خواهند کرد. این کار در اکثر موتورهای جستجو انجام می‌شود.

۳-٤. پیش‌پردازش و نرمال‌سازی داده‌ها

به منظور آماده‌سازی داده‌های پاکسازی شده جهت استفاده در الگوریتم‌های پردازش متن، عملیاتی برای واژه‌برداری[۱]، حذف ایست واژه، و بن‌واژه‌سازی[۲] انجام می‌شود.

بنابراین متن مورد نظر به واژگان قابل پردازش تقطیع خواهد شد. در جدول ۴-۳، متن یک سند پیش و پس از انجام عملیات فوق نمایش داده شده است. لازم به ذکر است که این عملیات بر روش همه داده‌های مجموعه داده و نیز هر سند و رکورد جدید اعمال می‌شود.

جدول ٤-۳- متن سند پیش و پس از آماده‌سازی، استانداردسازی، پیش‌پردازش و نرمال‌سازی

متن اولیه	متن آماده‌سازی شده و پیش‌پردازش شده
بررسی رابطه علّیت میان نوآوری و رشد اقتصادی در کشورهای منتخب منا)طی دوره‌ی(1995-2011 بر اساس نظریات رشد درون‌زا، دانش، نوآوری و تکنولوژی از مهم‌ترین عوامل تأثیر گذار بر رشد اقتصادی محسوب می‌شوند. از طرفی این دیدگاه وجود دارد که با افزایش رشد اقتصادی امکانات و منابع مالی بیشتری در اختیار کارآفرینان قرار می‌گیرد و این می‌تواند به نوبه‌ی خود ابداعات و نوآوری را گسترش دهد. در واقع، ممکن است یک جریان دایره وار بین نوآوری و رشد اقتصادی وجود داشته‌باشد؛ به‌طوری‌که نوآوری، رشد اقتصادی را افزایش داده و رشد اقتصادی خود می‌تواند به افزایش نوآوری منجر شود. علی‌رغم اهمیت عامل نوآوری در فرآیند رشد اقتصادی، به لحاظ تجربی مطالعات اندکی در این زمینه خصوصاً در مورد کشورهای در حال توسعه صورت گرفته‌است. از این‌رو، هدف طراحی شده برای این پژوهش، بررسی رابطه‌ی علّی میان نوآوری و رشد اقتصادی در گروهی از کشورهای منتخب منا می‌باشد. بدین منظور، پس از مرور مبانی نظری و مطالعات انجام‌شده در این زمینه، با استفاده از اقتصاد سنجی داده‌های تابلویی و آزمون‌های علّیت، رابطه‌ی علّیت میان نوآوری و رشد اقتصادی و هم‌چنین نحوه‌ی تأثیرگذاری نوآوری بر رشد اقتصادی در 22کشور منتخب منا طی سال‌های 1995تا 2011مورد بحث و بررسی قرار گرفته	بررسی رابطه علیت نوآور رشد اقتصاد کشور منتخب منا دوره ۱۹۹۵ -۲۰۱۱ نظریات رشد درون‌زا دانش نوآور تکنولوژی مهم عوامل تأثیر رشد اقتصاد محسوب شد طرف دیدگاه افزایش رشد اقتصاد امکانات منابع مالی بیشتر اختیار کارآفرین قرار گرفت توانست نوبه ابداعات نوآور گسترش دایره نوآور رشد اقتصاد داشته‌باشد به‌طوری‌که نوآور رشد اقتصاد افزایش رشد اقتصاد توانست افزایش نوآور منجر اهمیت عامل نوآور فرآیند رشد اقتصاد لحاظ تجربی مطالعات اندک خصوصا کشور توسعه گرفته‌است این‌رو هدف طراح پژوهش بررسی رابطه نوآور رشد اقتصاد گروهی کشور منتخب منا بدین منظور مرور مبانی نظر مطالعات انجام‌شده اقتصاد سنج داده تابلو آزمون علیت رابطه علیت نوآور رشد اقتصاد هم‌چنین نحوه تأثیرگذاری نوآور رشد اقتصاد ۲۲ کشور منتخب منا سال ۱۹۹۵ ۲۰۱۱ بحث بررسی قرار گرفته است یافته داد رابطه علیت طرفه نوآور رشد اقتصاد کوتاه‌مدت بلندمدت هم‌چنین کشور مطالعه ارتباط مثبت معنادار نوآور رشد اقتصاد حال ارتباط صادرات

[۱] Tokenization

[۲] Lemmatization

است. یافته‌ها نشان می‌دهد که یک رابطه‌ی علّیت یک طرفه از سمت نوآوری به رشد اقتصادی هم در کوتاه‌مدت و هم در بلند مدت وجود دارد. هم‌چنین، در گروه کشورهای مورد مطالعه، ارتباط مثبت و معناداری میان نوآوری و رشد اقتصادی وجود دارد. در حالی که، هیچ گونه ارتباطی میان صادرات کالاهای با فناوری برتر با رشد اقتصادی وجود ندارد. هم‌چنین، متغیرهای سرمایه‌گذاری مستقیم خارجی و تشکیل سرمایه ناخالص بر خلاف متغیر مخارج دولت، رابطه‌ی مثبت و معنی‌داری با رشد اقتصادی دارند.	کالا فناوری برتر رشد اقتصاد هم‌چنین متغیر سرمایه گذار مستقیم خارج تشکیل سرمایه ناخالص خلاف متغیر مخارج دولت رابطه مثبت معنی‌دار رشد اقتصاد

فصل پنجم:

پیاده‌سازی و ارزیابی

۵. پیاده‌سازی و ارزیابی

در ابن بخش ابتدا چگونگی پیاده‌سازی روش پیشنهادی و ابزارهای مورد استفاده برای آن تشریح می‌گردد. سپس روش پیشنهادی بر روی مجموعه داده‌ای که در فصل چهارم توضیح داده شد، ارزیابی می‌گردد.

۵-۱. پیاده‌سازی

برای پیاده‌سازی روش پیشنهادی از کتابخانه متن‌باز جنسیم[۱](Rehurek & Sojka 2010) و هضم[۲] در پایتون استفاده شده است. کتابخانه جنسیم، یک کتابخانه برای انجام عملیات پردازش متن نظیر نمایه‌سازی متون، بازیابی اطلاعات و مدل‌های برداری نمایش متون است. کتابخانه هضم نیز شامل مجموعه‌ای از توابع برای پردازش زبان طبیعی است که توابعی نظیر واژه–واژه کردن[۳]، ریشه‌یابی، و برچسب‌زنی اجزای سخن، را برای زبان فارسی ارائه می‌دهد. کتابخانه هضم یک کتابخانه پردازش زبان طبیعی فارسی در محیط پایتون است که به صورت منبع‌باز و رایگان در دسترس است.

برای پیاده‌سازی روش پیشنهادی از زبان برنامه‌نویسی پایتون ۳ استفاده شده است. هر یک از اجزای روش پیشنهادی که در فصل ۳ توصیف شده، به صورت یک تابع مجزا در پایتون کدنویسی و پیاده‌سازی شده است.

مدل کلمه–به–بردار برای دسته‌های موضوعی از اسناد به صوت جداگانه آموزش داده شده است. همانطور که در فصل پیش عنوان شد، چهار دسته موضوعی برای داده‌ها درنظر گرفته شده است که روش پیشنهادی برای سه دسته پیاده‌سازی و ارزیابی شده است، یعنی سه دسته: هنر و ادبیات، فنی و مهندس، و علوم انسانی. دسته علوم پایه به دلیل گستردگی و تنوع موضوعی بالا، در این مطالعه در نظر گرفته نشده است.

---

[۱] Gensim

[۲] Hazm

[۳] Tokenize

پس از آموزش مد کلمه-به-بردار برای هر دسته موضوعی از اسناد، برای هر کلمه در مجموعه داده، یک بازنمایی برداری عددی حاصل می‌شود که بر همین اساس می‌توان برای یک کلمه خاص، براساس فاصله کسینوسی بین دو بردار، کلمات مشابه را بازیابی نمود. نمونه‌ای از نحوه عملکرد مدل کلمه-به-بردار برای کلمات مختلف در حوزه‌های موضوعی مختلف، در جدول ۵-۱ آمده است.

جدول ۵-۱- نمونه‌ای از عملکرد مدل آموزش داده شده کلمه-به-بردار برای بازیابی کلمات مشابه

موضوع	کلمه	کلمات مشابه بازیابی شده (به ترتیب)
هنر و ادبیات	مولوی	مثنوی، عطار، خاقانی، سنایی، سعدی، حافظ، غزلیات، سروده، عاشقانه، مولانا، منظومه، فروغ، تمثیل، غزل، اسلوب، بدیع، احادیث، بیدل، قصاید، ناصرخسرو، قصیده، آیات
	هنر	هنرمند، سنت، تجسم، مدرن، نقاشی، عکاسی، هنرپروری، انتزاعی، معناگرا، ایرانی، باستان، اسلام، آیین، تئاتر، تحول، کهن، اصیل، هخامنشی، ارزشمند، سینما
	قاجار	صفوی، صفویه، ایلخان، ساسان، نگارگری، دوران، قاجاریه، سلجوقی، تیمور، راس، عصر، هخامنشی، آرامگاه، دوره، باستانی، مکتب، سکه، مانده، سفالینه، منبر، پهلوی
	ادبیات	عرب، معاصر، عامیانه، نویسندگان، ادب، کهن، عامه، هرسین، نوجوان، کلاسیک، برجسته، جامعه‌شناسی، سینما، غنا، بازشناخت، شاخه، نثر، حوزه، غنی، مکاتب، نقد، نو، مشروطیت
فنی و مهندسی	عصبی	بیزی، ANFIST، نروفازی، بیز، سیناپس، پرسپترون، مصنوعی، LVQ ، ANN، فازی-عصبی، شبکه عصبی، سخت‌افزاری، بولتزمن
	سازه	لرزه، ساختمان، میراگرها، زلزله، قاب، دیوار، بهسازی، نامنظم، بتنی، پی، شمع، ساختگاه، بنا، روسازه، سکو، مهاربند، طبقات، جداگر، بادبند
	ترمودینامیک	ترمواکونومیکی، سینتیک، اگزرژی، اگزرژتیک، اکسرژی، مونود، ریاضی، تعادل، ترموفیزیکی، سینیتیکی، ریفرمر، بخار-مایع، هیدرودینامیک، هالدن
علوم انسانی	خلاقیت	انگیزش، یادگیری، خلاق، مولفه، خودتنظیمی، خودکارآمدی، تفکر، فرصت‌یابی، شایستگی، مهارت، خودانگیزی، آمادگی، یادگیرنده، اثربخش
	صنعت	صنایع، صادرات، خودرو، سیمان، پتروشیمی، نساجی، داروسازی، کاشی، سرامیک، محصولات، هتلداری، بنگاه، لاستیک، کشاورزی، کارخانجات، کارخانه، فرآورده، فولاد، معدن، صادرکنندگان، فلز
	سرمایه‌گذاری	نقدینگی، دارایی، بدهی، مازاد، اهرم، سودآوری، صندوق، جاری، ریسک، بازده، پرتفوی، پورتفوی، نقدشوندگی، خالص، بنگاه، آتی، مبادله، معاملات، سهام، چرخه، نگهداشت
	استراتژیک	راهبردی، استراتژی، چابک، نوآور، هوشمندی، پیاده‌سازی، کسب و کار، پشتیبان،

		فرایند، فناورانه، همسو، یکپارچه، مزیت، رقابت، بازاریابی، پروژه، پیشبرد، برنامه‌ریزی

پس از آموزش مدل، شیوه عملکرد روش پیشنهادی به این ترتیب است که برای هر سند جدید، ابتدا حوزه موضوعی از بین چهار دسته فوق شناسایی می‌شود، سپس مدل آموزش داده شده متناسب با آن برای پیشنهاد کلیدواژه بکار گرفته می‌شود.در جدول ۵-۲ تا ۵-۳، نمونه‌ای از اسناد به همراه کلیدواژه‌های اصلی آنها و کلیدواژه‌های پیشنهادی در سه موضوع مهندسی، هنر و ادبیات، و علوم انسانی آمده است.

جدول ۵-۲- نمونه کلیدواژه‌های پیشنهادی براساس رویکرد پیشنهادی در موضوع هنر و ادبیات (چند متن اول و کلیدواژه اول نمایش داده شده است)

عنوان متن جدید	نگاهی نو به استاپ موشن در گرافیک دیجیتال
عنوان متون مشابه بازیابی شده	مطالعه و تحلیل تیزرها و انیمیشنهای تبلیغاتی موفق جهان (تکنیک استاپ موشن) قابلیت‌های استاپ موشن در بازنمود فضاهای گروتسک، ذهنی و وهمی تاثیر گرافیک بر سیر تحول تکنیک های عکاسی دیجیتال
کلیدواژه‌های پیشنهادی	پویانمایی، تیزر، گرافیک، سوررئالیسم، ارتباطات، بازنمایی، تکامل، تکنولوژی
کلیدواژه‌های اصلی	تکنیک ایست-حرکت، تیزر، تکنیک استاپ-موشن، تیزر تبلیغاتی، پویانمایی، تبلیغات، خلاقیت
عنوان متن جدید	آسیب ویژه نحوی زبان: مطالعه ای در یادگیری زبان دوم
عنوان متون مشابه بازیابی شده	ساختار لغوی-معنایی زبان آموزان چندتوانشی زبان دوم (زبان انگلیسی) در زبان اول (زبان فارسی) بررسی فراگیری سوالات استفهامی ساده زبان انگلیسی توسط زبان آموزان تک زبانه فارس و دو زبانه کرد-فارس در چارچوب دستور زبان زایشی مطالعه مقایسه‌ای جهت مجهول در زبان فرانسه و فارسی و مشکلات کاربرد آن نزد زبان‌آموزان ایرانی
کلیدواژه‌های پیشنهادی	زبان‌آموز، یادگیری زبان دوم، مجهول، تک‌زبانی، دوزبانگی، یادگیری، زبان فارسی، فراگیری زبان
کلیدواژه‌های اصلی	نحو، آسیب ویژه نحوی، کمینه‌گرایی، مجهول، زبان‌آموز، آسیب ویژه زبانی، یادگیری زبان دوم، زبان‌شناسی، رمزگذاری
عنوان متن جدید	جستاری در نقاشی معاصر ایران با تاکید بر موانع و چالش‌ها
عنوان متون مشابه بازیابی شده	بررسی واقع‌گرایی در نقاشی معاصر ایران (دهه‌ی ۷۰ و ۸۰ شمسی) نمود زیبایی شناسانه هنر اسلامی در نقاشی معاصر ایران نوشتار به مثابه تصویر در نقاشی معاصر ایران
کلیدواژه‌های پیشنهادی	فرانوگرایی، ایران، نقاشی معاصر، واقع‌گرایی، هنر جدید، زیبایی‌شناسی، نقاشی، کلاژ، معماری
کلیدواژه‌های اصلی	هنر معاصر، نقاشی معاصر، نوگرایی، پسانوگرایی، کثرت گرایی، ایران، مدرنیت، مدرنیته، مدرنیسم، پست مدرنیسم، نقاشی معاصر ایران

جدول ۵-۳- نمونه کلیدواژه‌های پیشنهادی براساس رویکرد پیشنهادی در موضوع علوم انسانی (چند متن و کلیدواژه اول نمایش داده شده است)

عنوان متن جدید	رابطه سبک های فرزندپروری ادراک شده با احساس تنهایی و ترس از ارزیابی منفی در نوجوانان
عنوان متون مشابه بازیابی شده	بررسی نقش واسطه ای ادراک از سبک‌های فرزندپروری والدین در رابطه بین سبک‌های تفکر پدران با سبک‌های تفکر دانش‌آموزان
	بررسی رابطه سبک‌های فرزندپروری ادراک شده، طرحواره های ناسازگار اولیه با احساس تنهایی در نوجوانان
	بررسی رابطه بین سبک های فرزندپروری با خلاقیت دانش آموزان مقطع متوسطه شهرستان قزوین
کلیدواژه‌های پیشنهادی	والدین، دانش‌آموز، ادراک، تنهایی، نوجوانان، خلاقیت، شیوه فرزندپروری، اقتدارگرایی
کلیدواژه‌های اصلی	شیوه پرورش کودک، احتمال خطر ادراکی، ترس، تنهایی، رابطه والد-کودک، ادراک، نوجوان، سبک‌های فرزندپروری ادراک شده
عنوان متن جدید	بررسی تأثیر توانمندسازی کارکنان بر وفاداری مشتریان با تأکید بر نقش میانجی کیفیت خدمات و رضایت مشتریان در صنعت بانکداری
عنوان متون مشابه بازیابی شده	بررسی تأثیر کیفیت خدمات بر وفاداری مشتریان در صنعت بانکداری شهرستان فردیس کرج
	بررسی تاثیر کیفیت ارتباط کارکنان بر وفاداری مشتریان در بانک کشاورزی (با تاکید بر استان گلستان)
	بررسی مسیر تاثیر عوامل وفاداری کارکنان، کیفیت خدمات، رضایت و وفاداری مشتری بر عملکرد شرکت‌های صنعت نرم‌افزاری ایران
کلیدواژه‌های پیشنهادی	کیفیت خدمات، تعهد، وفاداری، اعتماد، بازاریابی، بانکداری، وفاداری مشتری، ایران، رضایت مشتری، عملکرد، بانک ملی، کارکنان
کلیدواژه‌های اصلی	کیفیت خدمات، وفاداری مشتری، صنعت بانکداری، بانک ملی ایران، توانمندسازی، کارکنان، رضایت مشتری، روش کمترین مربعات، رضایت مشتری، کمترین مربعات جزئی
عنوان متن جدید	سیاست جنایی قضائی ایران در قبال جرایم رایانه‌ای
عنوان متون مشابه بازیابی شده	تاثیر پیشگیرانه قانون گذاریهای نظام تقنینی ایران در حوزه جرایم (IT)
	بررسی جرم جعل رایانه‌ای در قلمرو حقوق کیفری ایران
	نقش محیط سایبر در وقوع جرایم
کلیدواژه‌های پیشنهادی	جرم، پیشگیری، جرم‌شناسی، اقتصاد، مدرنیت، اینترنت، امنیت، محرومیت، تمامیت داده، بزه‌کاری، فضای سایبر، فضای مجازی
کلیدواژه‌های اصلی	سیاست کیفری، حقوق جزا، جرم سایبری، رسیدگی قضایی، جرم سایبری یا رایانه‌ای، جرم رایانه‌ای، امنیت داده‌ها، فضای مجازی، رویه قضایی، جرم‌شناسی
عنوان متن جدید	مقایسه سبک های تفکر، رفتار سازمانی و عملکرد مدیران انتخابی و انتصابی
عنوان متون مشابه بازیابی شده	بررسی اثر رفتار شهروندی سازمانی مدیران بر عملکرد کارکنان آموزش و پرورش ناحیه ۳ کرج
	رابطه بین سبک تفکر مدیران مدارس ابتدایی با سبک رهبری تحولی آنان بر اساس مدل بس و اولیو
	رابطه سبک‌های تفکر و خلاقیت با عملکرد شغلی مدیران مدارس متوسطه نواحی یک و دو شیراز
کلیدواژه‌های پیشنهادی	رفتار شهروندی سازمانی، تفکر، عملکرد، مدیران، شیراز، جو سازمانی، رفتار شهروندی، آموزش و پرورش، دانشگاه، رفتار سازمانی، آموزش، سازمانی، شخصیت، ارزیابی عملکرد
کلیدواژه‌های اصلی	رفتار شهروندی سازمانی، انتصاب، سبک تفکر، رفتار سازمانی، عملکرد مدیران، رفتار سازمانی، مدیران، تفکر، عملکرد

جدول ٥-٤- نمونه کلیدواژه‌های پیشنهادی براساس رویکرد پیشنهادی در موضوع فنی و مهندسی (چند متن اول و چند کلیدواژه اول نمایش داده شده است)

عنوان متن جدید	ارائه یک مدل مفهومی دیفرانسیلی به منظور شبیه‌سازی عملکرد حرارتی و بهینه‌سازی موتور استرلینگ
عنوان متون مشابه بازیابی شده	مدلسازی دینامیکی و تحلیل عملکرد حالت گذرای موتور هیسترزیس دور بالا
	مدل سازی ترموهیدرلیکی بازیاب موتور استرلینگ
	طراحی و ساخت موتوراسترلینگ نوع آلفا و مقایسه نتایج آزمایشگاهی با پیش‌بینی‌های مدل
کلیدواژه‌های پیشنهادی	بهینه‌سازی، شبیه‌سازی، جریان متلاطم، افت فشار، ترمودینامیک، اصطلاک، بازده، عملکرد، ترموهیدرولیک، موتور استرلینگ، الگوس جریان، سیلندر موتور، نشت، همرفت، جدایش
کلیدواژه‌های اصلی	موتور استرلینگ، تحلیل ترمودینامیکی، تحلیل حرارتی، عملکرد حرارتی، اثر شاتل، تحلیل پلی تروپیک، ترمودینامیک سرعت محدود، ویژگی ترمودینامیکی، شبیه‌سازی، بهینه‌سازی، افت فشار، اصطکاک، جریان نشتی
عنوان متن جدید	تأثیر متغیرهای جوشکاری همزن اصطکاکی بر خواص اتصال غیر همجنس آلیاژهای آلومینیم ۲۰۲۴ به ۷۰۷۵
عنوان متون مشابه بازیابی شده	بررسی خواص مکانیکی و ریزساختار جوشکاری آلیاژ آلومینیوم ۲۰۲۴-T3 به روش جوشکاری اصطکاکی اغتشاشی
	جوشکاری همزن اصطکاکی آلیاژ منیزیم AZ31
	بررسی تاثیر پارامترهای رئوکست روی ریزساختار و خواص مکانیکی آلیاژ Al-6Si-2Mg
کلیدواژه‌های پیشنهادی	جوشکاری اصطکاکی اغتشاشی، خواص مکانیکی، استحکام کششی، ریزساختار، آلیاژ آلومینیم، جوشکاری اصطکاکی، جوشکاری همزن اصطکاکی، خمش(شکل‌دهی)، تحلیل عددی، ساختار گلبولی، آلیاژ سیلیسیم، ویژگی‌های تجهیزات
کلیدواژه‌های اصلی	جوشکاری اصطکاکی اغتشاشی، خواص مکانیکی، آلیاژ آلومینیوم، جوشکاری اصطکاکی، سرعت چرخشی، ریزساختار، خمش(شکل‌دهی)، استحکام کششی
عنوان متن جدید	تثبیت خاک‌های ریز دانه با استفاده از مواد پلیمری
عنوان متون مشابه بازیابی شده	بررسی کاربرد پودر لاستیک به عنوان ماده تثبیت‌کننده ارزان در خاک رس نرم
	تاثیر میکروسیلیس بر مقاومت و تورم‌پذیری خاک تثبیت شده با آهک
	مطالعه رفتار مکانیکی خاک رس اصلاح شده با سیمان و الیاف نخ تایر
کلیدواژه‌های پیشنهادی	خاک رسی، تثبیت خاک، آهک، تثبیت، خاک، زهکشی، مکانیک خاک، نانوذره، نانوسیلیس، سیمان، خاک رس، تقویت، تورم، مقاومت، مقاومت برشی، افزودنی، تراکم، شبکه زمین، آماس
کلیدواژه‌های اصلی	تثبیت خاک، ریزدانه، خاک رسی، بسپار، آهک، مقاومت فشاری، ظرفیت تحمل

۲-۵. ارزیابی و تنظیم پارامترها

برای ارزیابی روش پیشنهادی مجموعه داده‌ها به سه دسته آموزش، اعتبارسنجی و آزمایش تقسیم‌بندی شده است، که ۸۰ درصد داده‌ها برای آموزش، ۱۰ درصد برای آزمایش و ۱۰ درصد برای اعتبارسنجی درنظر گرفته شده است.

روش پیشنهادی از پارامترهای مختلفی تشکیل شده است که برخی از آنها مربوط به بخش آموزش، و برخی دیگر مربوط به بخش بازیابی و امتیازدهی کلیدواژه‌های کاندید است.

پارامترهای بخش آموزش، مربوط به مدل کلمه-به-بردار است که مقادیر آنها عبارتند از:

- اندازه پنجره: اندازه پنجره، معمولا بین ۵ تا ۸ در نظر گرفته می‌شود، که در این آزمایش مقدار ۸ برای آن درنظر گرفته شده است.
- ابعاد بردار: ابعاد بردار عدد ۱۰۰ درنظر گرفته شده است.
- نمونه‌برداری: مقداری که به صورت معمول در نظر گرفته می‌شود عدد ۲۵ است.
- کمترین فراوانی: به صورت معمول عددی بین ۲ تا ۵ درنظر گرفته می‌شود که در اینجا عدد ۵ درنظر گرفته شده است.
- نمونه‌برداری منفی: در صورتیکه بخواهیم از نمونه‌برداری منفی استفاده کنیم، مقدار این پارامتر ۱ در نظر گرفته می‌شود.

هر یک از پارامترهای فوق نیز می‌تواند براساس مجموعه داده‌های اعتبارسنجی و با طراحی آزمایش‌های مختلف تنظیم شود، لیکن برای جلوگیری از پیچیدگی محاسباتی در این طرح مقادی معمول برای این پارامترها درنظر گرفته شده است.

برای بخش بازیابی اسناد مشابه و امتیازدهی، ۴ پارامتر مختلف وجود دارد:

- تعداد اسناد مشابه انتخاب شده (M)
- وزن روش بازیابی براساس مدل برداری ماتریس سند-کلمه (α)
- تعداد کلیدواژه‌های انتخابی از بین کلیدواژه‌های کاندید (K)

برای تنظیم سه پارامتر دیگر براساس مجموعه داده اعتبارسنجی، آزمایش‌های متعددی براساس معیارهای ارزیابی انجام شده است. برای ارزیابی و تنظیم پارامترها در روش پیشنهادی به راحتی نمی‌توان معیار دقت[۱] وبازخوانی[۱] و معیار F[۲] که به صورت معمول در این گونه از مسائل استفاده می‌شود، بکار گرفت، زیرا:

---

[۱] Precision

- برخی از کلیدواژه‌های پیشنهادی در فهرست کلیدواژه‌های اصلی نیستند لیکن با موضوع ارتباط دارند. علت این امر اینست که گاهی نمایه‌ساز که به صورت دستی کلیدواژه‌ها انتخاب نموده، دقت کافی در انتخاب آنها نداشته است.
- در روش پیشنهادی فهرست حداکثری از کلیدواژه‌ها پیشنهاد می‌شود تا از بین آنها انتخاب شود. این فهرست می‌تواند در بازیابی اطلاعات مورد استفاده قرار گیرد. بنابراین تعداد کلیدواژه‌های پینشهادی معمولا سه الی چهار برابر بیش از تعداد کلیدواژه‌های اصلی است که بر همین اساس به راحتی نمی‌توان دقت را محاسبه نمود. زیرا دقت نسبت تعداد اقلام درست به کل اقلام است که چون تعداد کل اقلام معمولا سه الی چهار برابر تعداد کل اقلام درست است، این مقدار، معنادار نخواهد بود. در اینگونه از موارد، برای محاسبه دقت، معیارهای دیگری مانند «میانگین حد متوسط دقت» (MAP)[۳] که معمولا در سیستم‌های پیشنهاددهنده از آنها نیز استفاده می‌شود، بکار گرفته می‌شوند. این معیار در ادامه تشریح می‌گردد.

به همین دلیل برای بررسی عملکرد روش پیشنهادی می‌توان از سه روش استفاده نمود:

- بازخوانی: مقدار فراخوانی همچنان می‌تواند معیار مناسبی برای سنجش روش پیشنهادی باشد. که مسلما این معیار با افزایش تعداد کلیدواژه‌های پیشنهادی (K) می‌تواند افزایش یابد.
- معیار MAP: استفاده از معیار «میانگین حد متوسط دقت» برای محاسبه میزان دقت روش پیشنهادی. این معیار معمولا در سیستم‌های پیشنهاددهنده و بازیابی اطلاعات که تعداد اقلام زیادی به کاربر پیشنهاد می‌شود و ترتیب اقلام مرتبط نیز اهمیت دارد، استفاده می‌گردد. در این معیار لزوما با افزایش تعداد کلیدواژه‌های پیشنهادی، مقدار دقت کاهش نمی‌یابد، بلکه آنچه که اهمیت دارد اینست که از کدامیک از موارد پیشنهاد شده مرتبط در بالای فهرست پیشنهادی قرار دارند و امتیاز بالاتری دارند.
- تشابه معنایی: میزان شباهت کلیدواژه‌های پیشنهادی با سند مورد نظر و نیز میزان شباهت کلیدواژه‌های اصلی آن سند با هم مقایشه شود تا معلوم شود که چقدر کلیدواژه‌های پیشنهاد شده، با سند مذکور مشابه هستند. برای این منظور می‌توان از مدل کلمه-به-بردار استفاده نمود.
- فرد متخصص: افراد متخصص حوزه‌های موضوعی برای تعیین ارتباط یا عدم ارتباط کلیدواژه‌های پیشنهادی بکار گرفته شوند و نظرات آنها در نهایت ملاک ارزیابی قرار گیرد.

---

[۱] Recall

[۲] F-measure

[۳] Mean Average Precision

در این طرح، از هر چهار روش برای برررسی عملکرد روش پیشنهادی استفاده شده است.

۱-۲-۵. ارزیابی براساس بازخوانی

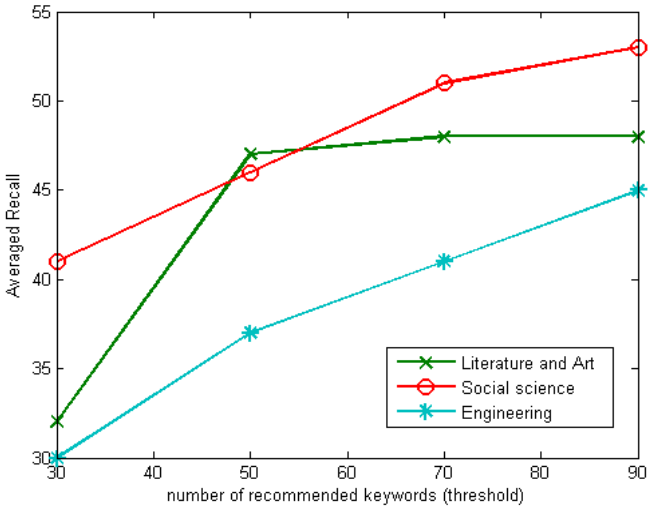
بازخوانی براساس رابطه زیر حساب می‌شود:

$$\boldsymbol{Recall} = \frac{\#\{\boldsymbol{recommended\ keywords}\} \cap \{\boldsymbol{original\ keywords}\}}{\#\{\boldsymbol{original\ keywords}\}} \quad (۱-۵)$$

برای تنظیم پارامترهای مسئله و بررسی عملکرد روش پیشنهادی بر روی تغییرات پارامترها، آزمایش‌های مختلفی انجام شده که نتایج آنها در جدول ۵-۵ و شکل ۵-۱ آمده است.

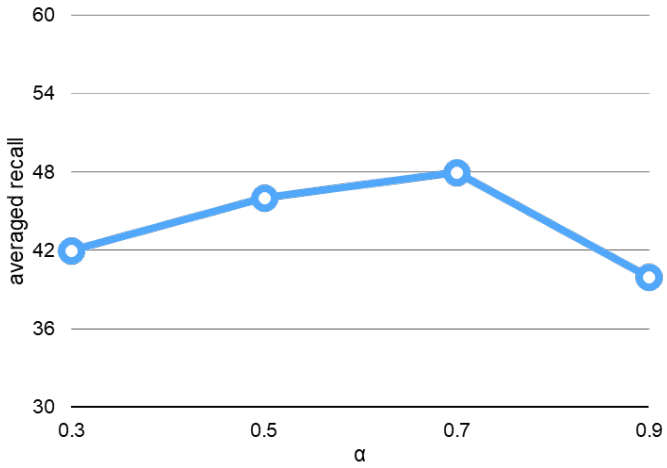
جدول ۵-۵- بررسی عملکرد روش پیشنهادی با تغییر تعداد کلیدواژه‌های پیشنهادی ($\alpha = 0.7, M = 20$)

موضوع	تعداد کلیدواژه‌های پیشنهادی (K)	بازخوانی (درصد)
هنر و ادبیات	۳۰	۳۲
	۵۰	۴۲
	۷۰	۴۸
	۹۰	۴۸
علوم انسانی	۳۰	۴۱
	۵۰	۵۰
	۷۰	۵۱
	۹۰	۵۲
فنی و مهندسی	۳۰	۳۰
	۵۰	۳۷
	۷۰	۴۱
	۹۰	۴۴



شکل ۵-۱- عملکرد روش پیشنهادی با تغییر پارامتر تعداد کلیدواژه‌های پیشنهادی($\alpha = 0.7, M = 20$)

همچنین عملکرد روش پیشنهادی با تغییر مقدار α برای داده‌های هنر و ادبیات در شکل ۵-۲ آمده است.



شکل ۵-۲- عملکرد روش پیشنهادی برای مقادیر مختلف α

۵-۲-۲. ارزیابی براساس معیار «میانگین حد متوسط دقت» (MAP)

در سیستم‌هایی که ترتیب موارد بازیابی شده اهمیت دارد و مواردی که بازیابی می‌شوند براساس رتبه مرتب می‌شوند، معیارها « میانگین حد متوسط دقت» برای سنجش عملکرد سیستم، مناسب‌تر است. به طور خاص در سیستم‌های پیشنهاددهنده از این معیار هم برای سنجش میزان دقت نیز استفاده می‌شوند، زیرا معمولا در این دست از سیستم‌ها فهرستی از موارد، بیش از حد نیاز کاربر، به وی پیشنهاد می‌شود تا از بین آنها انتخاب نماید. بنابراین پیشنهاد موارد زیاد، که ممکن است تعدادی از آنها هم کاملا مرتبط نباشند، لزوما یک امتیاز منفی محسوب نمی‌شود، بلکه در این گونه از موارد، سیستمی مناسب خواهد بود که

بتوانـد مـوارد مـرتبط را در N پیشـنهاد اول قـرار دهـد(Manning & Raghavan 2009; Shani & Gunawardana 2011). در روش‌های کلاسیک اندازه‌گیری دقت، به عنوان مثال اگر سیستمی ۲۰ نمونـه پیشنهاد دهد و از بین این ۲۰ نمونه ۱۰ نمونه مرتبط باشند، آنگاه، فارق از اینکه ایـن ۱۰ نمونـه در کجـای فهرست پینشهادی قرار دارند، دقت سیستم ۵۰ درصد است. در صورتیکه در روش «میانگین حد متوسـط دقت»، رتبه نمونه‌های مرتبط نیز اهمیت دارد. به عنوان مثال اگر سیستمی بتواند ۱۰ نمونه مرتبط را در اول فهرست پیشنهادها قرار دهد، دقت بالاتری خواهد داشت نسبت به سیستمی که این ۱۰ نمونه را انتها قـرار دهد.

معیار «میانگین حد متوسط دقت» (MAP) براساس پیشنهاد K کلیدواژه، میانگین $AP@K$ بـرای M سـند مختلف است. مقدار $AP@K$ (حد متوسط دقت درK) برای یک سند d_i به صورت زیر محاسبه می‌شود:

$$AP@N(d_i) = \frac{1}{m}\sum_{k=1}^{K} P(k).rel(k) \qquad (۲-۵)$$

که $rel(k)$ مساوی مقدار یک است در صورتیکه کلیدواژه k ام مرتبط با d_i باشد و صفر را می‌گیرد در صورتیکه کلیدواژه مرتبط نباشد و m تعداد کل کلیدواژه‌های مـرتبط درکـل فضـای مسـئله اسـت. $P(k)$ دقت را برای k کلیدواژه اول نشان می‌دهد که به صورت زیر محاسبه می‌شود:

$$P(k) = \frac{\#\{recommended\ keywords\} \cap \{original\ keywords\}}{k} \qquad (۳-۵)$$

در نهایت $MAP@N$ یعنی «میانگین حد متوسط دقت»، به صورت زیر محاسبه می‌شود:

$$MAP@K = \frac{1}{N}\sum_{i=1}^{N} AP@N(d_i) \qquad (۴-۵)$$

که N در آن تعداد اسناد مجموعه آزمایش اسـت. در واقـع $MAP@N$ میـانگین مقـدار $AP@K$ بـرای N سند مختلف است. بر همین اساس برای داده‌های آزمایشی، این مقـدار محاسـبه شـده کـه در جـدول ۵-۶ نمایش داده شده است.

جدول ۵-٦- مقدار میانگین حد متوسط دقت ($MAP@K$) برای $K = 50$

موضوع	میانگین حد متوسط دقت ($MAP@K$)
هنر و ادبیات	۲۷/۳
علوم انسانی	۲۷/۵

فنی و مهندسی	۲۲/۲

۳-۲-۵. ارزیابی براساس شباهت معنایی

همانطور که در فصل سوم نیز عنوان شد، تابع امتیازدهی و رتبه‌بندی کلیدواژه‌های کاندید از دو منبع اطلاعاتی برای تعیین امتیاز کلیدواژه‌ها استفاده می‌کند:

۱. امتیازدهی براساس شباهت با متن جدید
۲. امتیازدهی براساس امتیاز اسناد بازیابی شده

ترکیب این دو منبع اطلاعاتی می‌تواند به روش‌های مختلفی انجام شود، که در فصل سوم روش عنوان شده در فرمول (۳-۱۱) به عنوان روش اصلی در نظر گرفته شده است، لیکن چند حالت مختلف نیز بررسی شده است.

در همین راستا، چهار حالت مختلف ترکیب این دو منبع اطلاعاتی به صورت زیر بررسی شده است:

۱. حالت مجموع مربعات (sum of squared): در این حالت، تابع امتیاز نهایی به صورت زیر تعریف می‌شود:

$$score_{final}(k_l) = score_1^*(k_l, d_{new})^2 + score_2(k_l, d_{new})^2 \qquad (۵-۵)$$

که در آن تابع امتیاز $\boldsymbol{score_1^*(k_l, d_{new})}$ نیز به صورت زیر در نظر گرفته می‌شود:

$$score_1^*(k_l, d_{new}) = \frac{V_{ret} \cdot V_{key}(k_l)}{|V_{ret}|} \qquad (۶-۵)$$

در واقع در این روش، فراوانی اسنادی که یک کلیدواژه کاندید در آنها آمده، در نظر گرفته نمی‌شود.

۲. حالت مجموع مربعات به همراه اثر نمایی فراوانی اسناد (sum of squared*exp.): در این حالت، تابع امتیاز نهایی همانند رابطه (۵-۲) تعریف می‌شود، لیکن تابع امتیاز $\boldsymbol{score_1^*(k_l, d_{new})}$ به صورت زیر تعریف می‌شود:

$$score_1^*(k_l, d_{new}) = \frac{V_{ret} \cdot V_{key}(k_l)}{|V_{ret}|} \cdot e^{(2(f(k_l)-1)/M)} \qquad (۷-۵)$$

در این حالت، فراوانی اسنادی که یک کلیدواژه کاندید را دارند به صورت نمایی در امتیاز نهایی اثر خواهد بود.

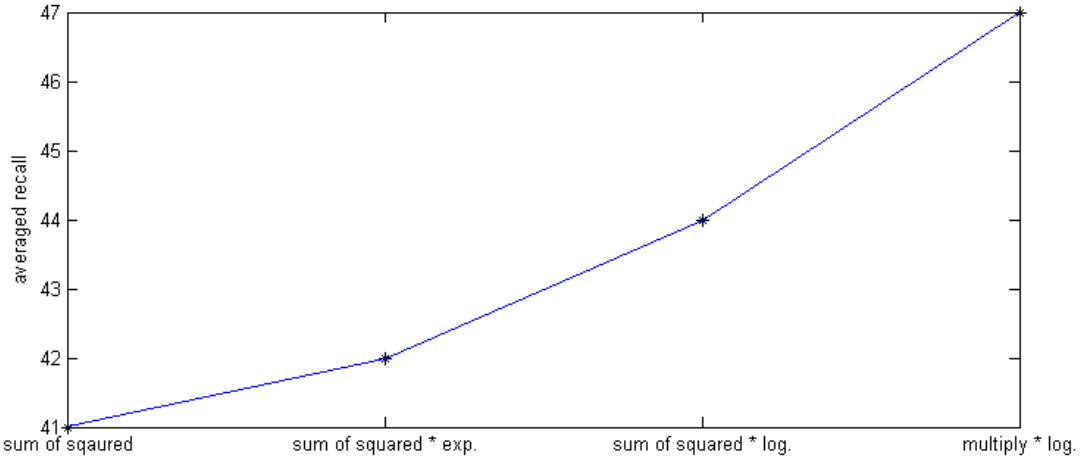
۳. حالت مجموع مربعات به همراه اثر لگاریتمی فراوانی اسناد (sum of squared*log.): در این حالت، تابع امتیاز نهایی همانند رابطه (۵-۲) تعریف می‌شود، و تابع امتیاز $score_1(k_l, d_{new})$ نیز همانند آنچه در فصل سوم تعریف شد، در نظر گرفته می‌شود، یعنی :

$$score_1(k_l, d_{new}) = \frac{V_{ret} \cdot V_{key}(k_l)}{|V_{ret}|} \cdot \log(f(k_l) + \frac{1}{4}M) \qquad (۸-۵)$$

در این حالت، فراوانی اسنادی که یک کلیدواژه کاندید را دارند به صورت لگاریتمی در امتیاز نهایی اثر خواهد بود.

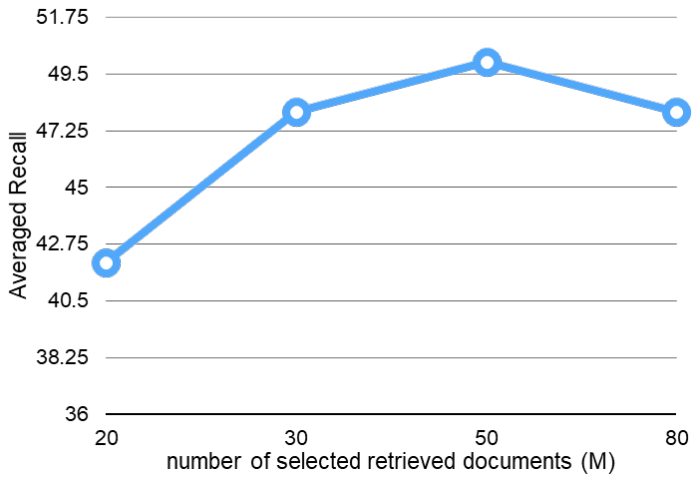
۴. حالت حاصلضرب امتیازها به همراه اثر لگاریتمی فراوانی سند (multiply * log): این حالت همانند شیوه‌ای که در فصل سوم به آن اشاره شد و از آنجاییکه منجر به عملکرد بهتر در روش پیشنهادی می‌شود، در آزمایش‌های متعدد، این روش مبنا درنظر گرفته شده است.

عملکرد روش پیشنهادی براساس ۴ روش امتیازدهی فوق، در شکل ۵-۳ برای مجموعه داده هنر و ادبیات آمده است. همانطور در شکل مشخص است، روش چهارم، عملکرد بهتری را نسبت به سایر روش‌ها ارائه می کند.



شکل ۵-۳- عملکرد روش پیشنهادی براساس روش‌های امتیازدهی مختلف برای کلیدواژه‌های کاندید

همچنین عملکرد روش پیشنهادی با تغییر تعداد اسناد بازیابی شده منتخب یعنی M بررسی شده است که نتایج آن برای $\boldsymbol{\alpha = 0.7}$ و $Threshold = 50$ برای مجموعه داده هنر و ادبیات در شکل ۵-۴ نمایش داده شده است.



شکل ٥-٤- عملکرد روش پیشنهادی براساس پارامتر تعداد اسناد بازیابی منتخب

لازم به ذکر است که یکی از دلایل اصلی کم بودن مقدار بازخوانی اینست که در بسیاری از موارد کلیدواژه‌های اصلی در مجموعه داده‌های آموزش وجود ندارد. بنابراین تحت هیچ شرایطی این کلیدواژه‌ها در نتایج پیشنهادی وجود نخواهند داشت. در صورتیکه این کلیدواژه‌ها در محاسبه مقدار افزایش یابد.

از این رو، برای ارزیابی بهتر روش پیشنهادی، از معیار دیگری برمبنای تشابه معنایی کلیدواژه‌های پیشنهادی با سند، استفاده شده است. بر اساس این معیار، میزان تشابه معنایی کلیدواژه‌های اصلی سند با خود سند، و با میزان تشابه معنایی کلیدواژه‌های پیشنهادی با سند، با هم مقایسه می‌شود. بنابراین اگر $\{k_1^*, \dots k_{n_k}^*\}$ کلیدواژه‌های اصلی سند $\boldsymbol{d_k}$، و $\{k_1, \dots k_{m_k}\}$ کلیدواژه‌های پیشنهادی باشند، آنگاه شباهت بین هر کلیدواژه را با سند می‌توان براساس مدل کلمه-به-بردار براساس رابطه (۳-۹) که در فصل سوم تعریف شد، محاسبه نمود و میانگین این شباهت را در نظر گرفت:

$$\boldsymbol{x_k} = \frac{\mathbf{1}}{\boldsymbol{m_k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m_k}} \boldsymbol{score_2(k_i, d_k)} \qquad (۹-٥)$$

$$\boldsymbol{y_k} = \frac{\mathbf{1}}{\boldsymbol{m_k}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n_k}} \boldsymbol{score_2(k_i^*, d_k)} \qquad (۱۰-٥)$$

در نهایت می‌توان میانگین و انحراف معیار خطای مربعات را برای دو نمونه $\boldsymbol{x_k}$ و $\boldsymbol{y_k}$ محاسبه نمود که نتایج آن برای داده‌های هر حوزه در جدول زیر آمده است.

جدول ۵-۷- میانگین خطای مربعات بین شباهت کلیدواژه‌های اصلی با سند و شباهت کلیدواژه‌های پیشنهادی با سند

حوزه موضوعی	میانگین خطای مربعات	انحراف معیار خطای مربعات	تعداد نمونه‌های آزمایش
هنر و ادبیات	۰٫۰۱۱۹	۰٫۰۱۷۳	۱۸۷
علوم انسانی	۰٫۰۱۹۴	۰٫۰۲۵۸	۵۲۰
فنی و مهندسی	۰٫۰۱۵۶	۰٫۰۲۴۰	۵۴۳

٤-۲-۵. ارزیابی براساس نظر متخصص

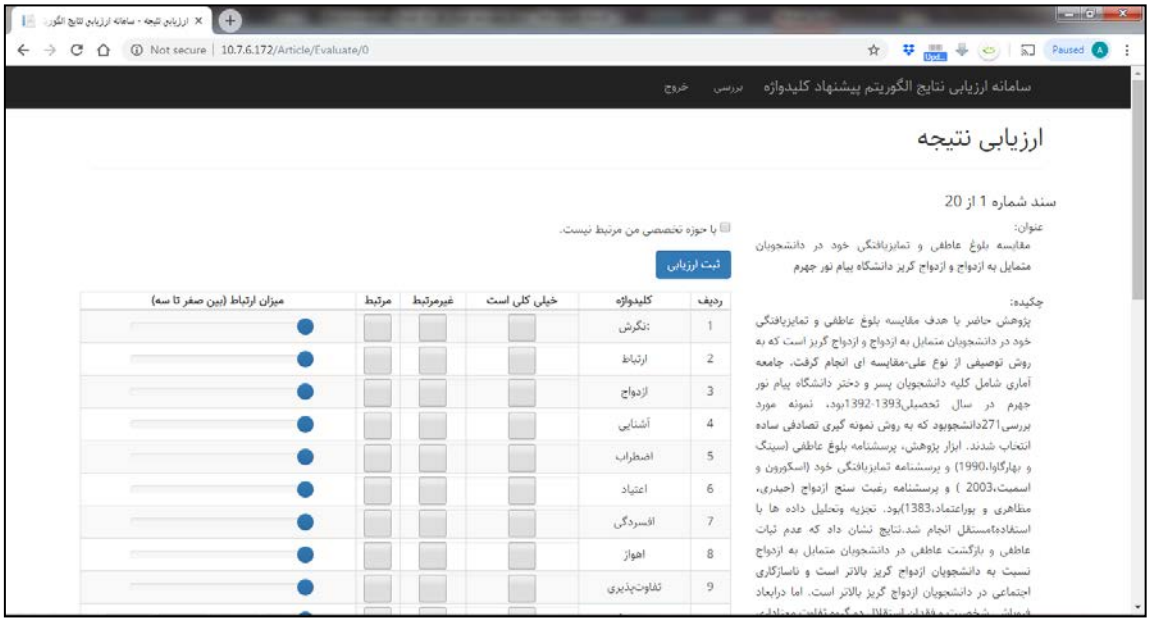
به منظور ارزیابی کلیدواژه‌های پیشنهادی براساس نظر متخصص، از متخصصین نمایه‌ساز پژوهشگاه علوم و فناوری اطلاعات ایران استفاده شده است. در هر یک از سه حوزه هنر و ادبیات، فنی و مهندسی، و علوم انسانی، نظر ۴ متخصص درباره ۲۵ سند دریافت شده (روی هم برای هر سه حوزه ۷۵ سند) است. برای دریافت نظرات متخصصین، سامانه ارزیابی مبتنی بر وب طراحی شده تا متخصصین بتوانند نظراتشان را در آن وارد نمایند. در این سامانه هر کاربر پس از ثبت‌نام، حوزه موضوعی مرتبط را انتخاب می‌کند (شکل ۵-۵). سپس هر سند در یک صفحه مجزا به همراه کلیدواژه‌های پیشنهادی برای آن نمایش داده می‌شود (شکل ۵-۶، (ب)). کاربر براساس ارتباط معنایی بین کلیدواژه و سند، گزینه‌های «خیلی کلی است»، «مرتبط» و «غیرمرتبط» را انتخاب می‌کند. در صورتیکه کاربر گزینه مرتبط را انتخاب کند، نیاز است که براساس میزان ارتباط یکی از اعداد ۱، ۲ یا ۳ (عدد ۱ بیانگر کمترین ارتباط است) را در نوار عددی متناظر با هر کلیدواژه انتخاب کند (شکل ۵-۷).

الف) صفحه اول سامانه ارزیابی

ب) صفحه ثبت‌نام/ورود کاربر

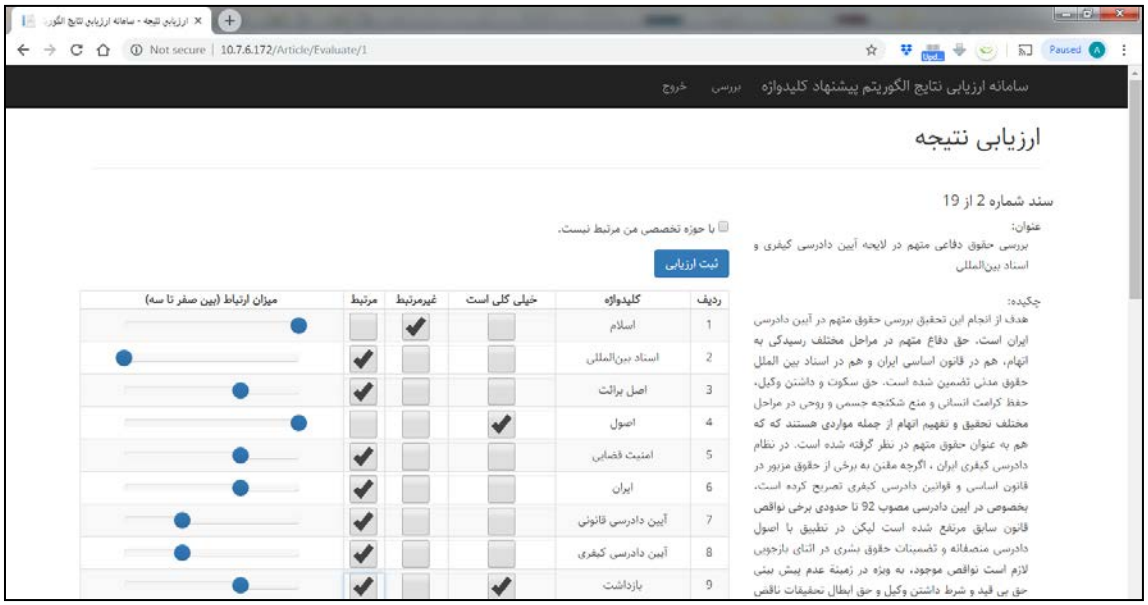
شکل ۵-۵- صفحات اول و ورود به سامانه ارزیابی کلیدواژه‌ها براساس نظرات متخصص

الف) انتخاب حوزه موضوعی



ب) ارزیابی کلیدواژه‌ها برای هر سند

شکل ۵-۶- صفحات انتخاب حوزه موضوعی و ارزیابی کلیدواژه‌ها



شکل ۵-۷- امتیازدهی به کلیدواژه‌ها

نتایج بدست آمده براساس این سامانه در جدول زیر نمایش داده شده است.

جدول ۵-۸- نتایج سامانه ارزیابی

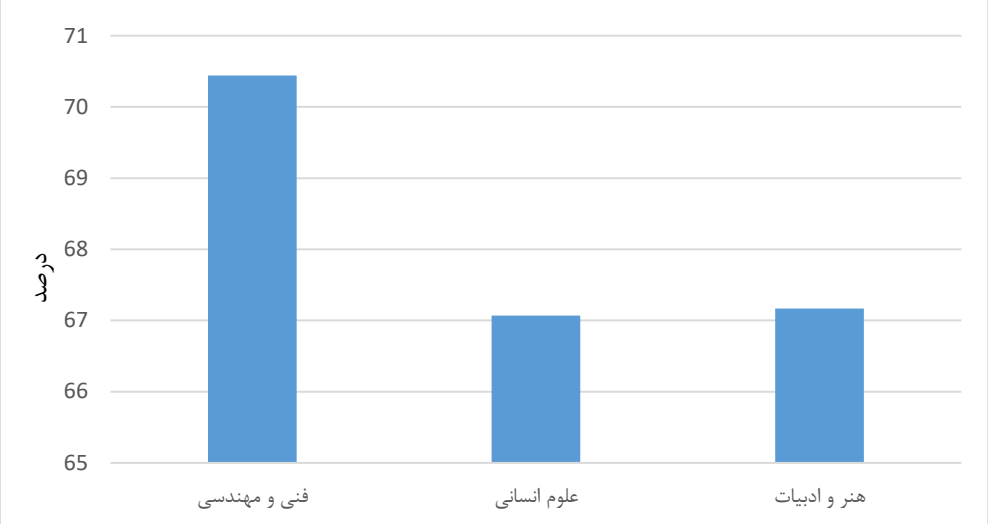
معیار ارزیابی / حوزه	متوسط تعداد کلیدواژه‌های خیلی کلی	متوسط تعداد کلیدواژه‌های مرتبط	متوسط مجموع امتیاز کلیدواژه‌ها	متوسط حداکثر امتیاز ممکن برای کلیدواژه‌های مرتبط	متوسط میزان ارتباط کلیدواژه‌های پیشنهادی (درصد)
فنی و مهندسی	۴/۶۹	۱۰/۸۴	۲۲/۸۷	۳۲/۵۲	۷۰/۴۴
علوم انسانی	۱۰/۰۲	۲۱/۰۹	۴۲/۴۶	۶۳/۲۷	۶۷/۰۷
هنر و ادبیات	۷/۶۵	۱۳/۶۶	۲۷/۳۵	۴۰/۹۸	۶۷/۱۷

براساس جدول فوق، به طور متوسط ۱۵ کلیدواژه از بین کلیدواژه‌های پیشنهاد شده مرتبط هستند. با مقایسه این عدد با متوسط تعداد کلیدواژه‌های اسناد که ۷ است، می‌توان گفت که الگوریتم پیشنهادی در شناسایی کلیدواژه‌های مرتبط از نظر تعداد، توانسته حداقل به اندازه کلیدواژه‌های اصلی سند، کلیدواژه مرتبط بیابد.

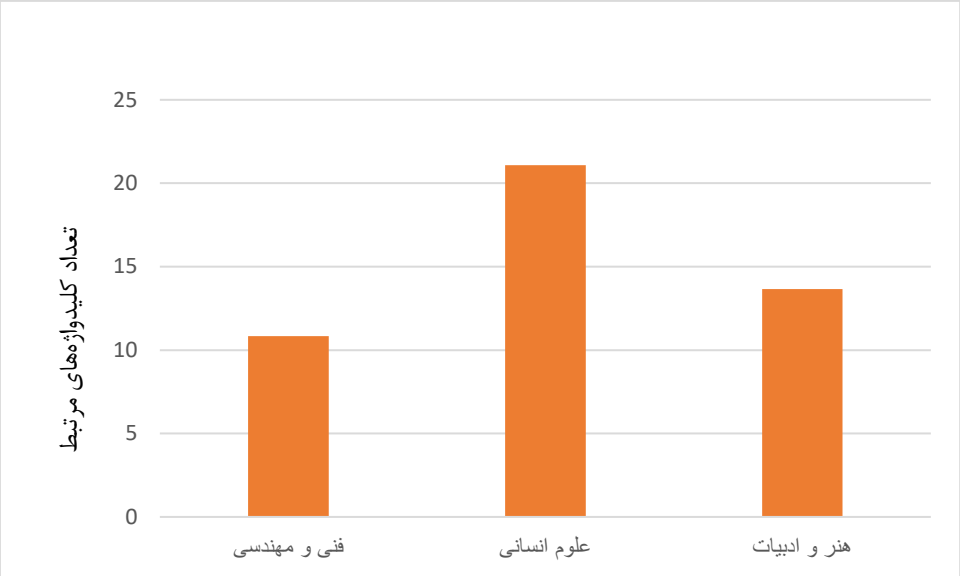
از طرف دیگر با بررسی امتیاز کلیدواژه‌های مرتبط، می‌توان دریافت که تا چه حد این کلیدواژه‌ها مرتبط هستند و از نظر معنایی و مفهومی با سند مورد نظر همخوانی دارند. برای این منظور «متوسط مجموع امتیاز کلیدواژه‌های هر سند» نیز محاسبه شده است. سپس این مقدار با «متوسط حداکثر مقدار امتیاز ممکن» مقایسه شده است. به عنوان مثال برای سندی که ۱۰ کلیدواژه مرتبط شناسایی شده، حداکثر امتیاز

کلیدواژه‌ها مقدار ۳۰ خواهد بود، زیرا هر کلیدواژه حداکثر ۳ امتیاز از لحاظ میزان ارتباط می‌تواند داشته باشد. برای محاسبه «متوسط میزان ارتباط کلیدواژه‌های پیشنهادی (درصد)» برای این سند، مجموع امتیاز کلیدواژه‌های مرتبط بر مقدار ۳۰ تقسیم شده است و در ۱۰۰ ضرب شده است. هر چه این مقدار بیشتر باشد، نشان می‌دهد که کلیدواژه‌های مرتبط شناسایی شده، از نظر معنایی و مفهومی با سند مرتبط‌تر هستند.

نتایج جدول فوق در شکل ۵-۸ و شکل ۵-۹ به صورت مشخص‌تری برای مقایسه سه حوزه نمایش داده شده است. براساس این شکل می‌توان گفت که علی رغم اینکه در حوزه فنی و مهندسی کلیدواژه‌های کمتری مرتبط شناسایی شده، لیکن میزان ارتباط کلیدواژه‌های مرتبط بیشتر است، به بیان دیگر در حوزه فنی و مهندسی کلیدواژهای تخصصی بیشتری شناسایی شده است. لیکن در دو حوزه دیگر تعداد کلیدواژه‌های مرتبط بیشتر است، در حالیکه میزان ارتباط آنها به نسبت حوزه فنی و مهندسی کمتر است.



شکل ۵-۸- مقایسه میزان ارتباط کلیدواژه‌های مرتبط در حوزه‌های مختلف



شکل ۵-۹- مقایسه تعداد کلیدواژه‌های مرتبط در حوزه‌های مختلف

فصل ششم:

نتیجه‌گیری و پیشنهادات آتی

٦. نتیجه‌گیری و پیشنهادات آتی

در این پژوهش، روشی برای پیشنهاد خودکارکلیدواژه برای مستندات علمی فارسی پیشنهاد شده است. تعداد کلیدواژه‌های پیشنهادی در این روش بسیار بیشتر از تعداد کلیدواژه‌های معمول برای یک سند علمی است زیرا کلیدواژه‌های می‌تواند درعملیات بازیابی و جستجو و تحلیل محتوی سند مفید باشد.

روش پیشنهادی بر مبنای عملکرد سیستم‌های پیشنهاددهنده محتوی محور عمل می‌کند وشیوه عملکرد آن برای پیشنهاد کلیدواژه براساس منطق استدلال نمونه‌محور است. دو بخش اصلی در روش پیشنهادی وجود دارد: بخش آموزش برای ساخت مدل و بخش بازیابی و پیشنهاد کلیدواژه.

روش پیشنهادی برای تعدادی از پایان‌نامه‌های ارشد و دکتری تعدادی از دانشگاه‌های ایران پیاده‌سازی شده است. این پایان‌نامه‌ها از پایگاه اطلاعات گنج متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) استخراج شده است.

برای بهبود عملکرد روش پیشنهادی، مجموعه داده‌ها به سه حوزه موضوعی هنر و ادبیات، علوم انسانی، و فنی-مهندسی تقسیم‌بندی شده و برای هر دسته موضوعی به صورت جداگانه پیاده‌سازی شده است.

برای بررسی اثرگذاری پارامترهای مختلف در روش پیشنهادی، آزمایش‌های مختلفی انجام شده است و عملکرد آن براساس شاخص بازخوانی و شباهت بین کلیدواژه‌های پیشنهادی و سند، بررسی شده است.

براساس آزمایش‌های انجام شده، می‌توان گفت که آموزش مدل‌های جداگانه برای هر حوزه موضوعی بر روی مجموعه داده‌های موضوعی، می‌تواند نتایج بهتری را ارائه دهد. همچنین از آنجاییکه نتایج روش پیشنهادی وابسته به تعداد اسناد مشابه بازیابی شده است، بنابراین الگوریتم بازیابی نقش مهمی را در عملکرد روش پیشنهادی ایفا می‌نماید.

بررسی شباهت معنایی بین کلیدواژه‌های پیشنهادی با سند و شباهت معنایی بین کلیدواژه‌های اصلی با سند، نشان می‌دهد که اختلاف معناداری بین شباهت‌ها وجود ندارد و بر همین اساس می‌توان گفت که از نظر معنایی کلیدواژه‌های پیشنهادی با سند از نظر معنایی مرتبط هستند.

همانطور که در فصل اول و اهداف طرح نیز مطرح شد، کلیدواژه‌های پیشنهاد شده در کنار دانش نمایه‌ساز یا سایر روش‌های استخراج کلیدواژه که بر اساس استخراج واژگان مهم از متن کامل سند عمل

می‌کنند، می‌تواند بکار گرفته شود و قرار نیست که به طور کامل جایگزین فرایند نمایه‌سازی شود. روش پیشنهادی می‌تواند به عنوان یکی از منابع دانشی در فرایند نمایه‌سازی درنظر گرفته شود.

بنابراین یکی از پیشنهادها برای تحقیقات آتی اینست که روش‌های استخراج کلیدواژه از متن که بر مبنای اطلاعات آماری واژگان کاندید درون متن عمل می‌کنند در کنار روش پیشنهادی در نظر گرفته شود.

علاوه بر آن، از آنجاییکه عملکرد روش‌های مدلسازی مانند کلمه-به-بردار وابسته به حجم داده‌های آموزش است، برای بهبود عملکرد روش پیشنهادی می‌توان از مجموعه داده بزرگ تری با داده‌های موضوعی متنوع تر استفاده نمود.

یکی دیگر از پیشنهادات اینست که با افزایش مجموعه داده‌ها، حوزه‌های موضوعی به صورت تخصصی‌تر و با تعداد بیشتر مشخص گردد و در فرایند بازیابی اسناد مشابه، به پایگاه داده موضوعی تخصصی‌تر مراجعه شود.

منابع

۷. منابع

جلالی‌منش، عمار؛ علیدوستی، سیروس؛ خسروجردی، محمود. نمایه‌ساز ماشینی منابع فارسی: مدلی یک‌پارچه برای پژوهشگاه علوم و فناوری اطلاعات ایران .پژوهشنامه پردازش و مدیریت اطلاعات. ۱۳۹۲؛ ۲۹ (۲) :۴۲۵-۴۵۱.

بشیری، حسن؛ کربلایی، فاطمه؛ موسوی، شیرین.طراحی و ارزیابی نمایه‌ساز خودکار متون فارسی. یازدهمین کنفرانس بین‌المللی کامپیوتر انجمن کامپیوتر ایران. تهران. ۱۳۸۴.

تشکری، مسعود؛ میبدی، محمدرضا. ساخت یک نمایه ساز خودکار برای متون فارسی. یازدهمین کنفرانس مهندسی برق. ۱۳۸۲.

Aamodt, A. & Plaza, E., 1994. Case-based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Community*, 7(1), pp.39–59. Available at: http://dl.acm.org/citation.cfm?id=196108.196115.

Aciar, S., 2014. Peer Recommendation Based on Text Mining Algorithm. In *2014 13th Mexican International Conference on Artificial Intelligence*. pp. 37–42.

Adomavicius, G. & Tuzhilin, A., 2011. Context-Aware Recommender Systems. In F. Ricci et al., eds. *Recommender Systems Handbook*. Boston, MA: Springer US, pp. 217–253. Available at: https://doi.org/10.1007/978-0-387-85820-3_7.

Aggarwal, C.C., 2016a. Model-Based Collaborative Filtering. In *Recommender Systems: The Textbook*. Cham: Springer International Publishing, pp. 71–138. Available at: https://doi.org/10.1007/978-3-319-29659-3_3.

Aggarwal, C.C., 2016b. Neighborhood-Based Collaborative Filtering. In *Recommender Systems: The Textbook*. Cham: Springer International Publishing, pp. 29–70. Available at: https://doi.org/10.1007/978-3-319-29659-3_2.

Aggarwal, N., Asooja, K. & Buitelaar, P., 2012. DERI&UPM: Pushing Corpus Based Relatedness to Similarity: Shared Task System Description. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. SemEval '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 643–647. Available at: http://dl.acm.org/citation.cfm?id=2387636.2387745.

Al-Shamri, M.Y.H., 2016. User profiling approaches for demographic recommender systems. *Knowledge-Based Systems*, 100, pp.175–187. Available at: http://dx.doi.org/10.1016/j.knosys.2016.03.006.

Benson, G., Levy, A. & Shalom, B.R., 2013. Longest Common Subsequence in k Length Substrings. In N. Brisaboa, O. Pedreira, & P. Zezula, eds. *Similarity Search and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 257–265.

Bobadilla, J. et al., 2013. Recommender systems survey. *Knowledge-Based Systems*, 46, pp.109–132. Available at: http://dx.doi.org/10.1016/j.knosys.2013.03.012.

Bousbahi, F. & Chorfi, H., 2015. MOOC-Rec: A Case Based Recommender System for MOOCs. *Procedia - Social and Behavioral Sciences*, 195, pp.1813–1822. Available at: http://www.sciencedirect.com/science/article/pii/S1877042815038744.

Bridge, D. et al., 2005. Case-based recommender systems. *The Knowledge Engineering Review*, 20(3), p.315. Available at: http://www.journals.cambridge.org/abstract_S0269888906000567.

Burke, R., 2007. Hybrid Web Recommender Systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl, eds. *The Adaptive Web: Methods and Strategies of Web Personalization*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 377–408. Available at: https://doi.org/10.1007/978-3-540-72079-9_12.

Buscaldi, D. et al., 2012. IRIT: Textual Similarity Combining Conceptual Similarity with an N-gram Comparison Method. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*. SemEval '12. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 552–556. Available at: http://dl.acm.org/citation.cfm?id=2387636.2387729.

Deerwester, S. et al., 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), pp.391–407.

Dong, R., O'Mahony, M.P. & Smyth, B., 2014. Further Experiments in Opinionated Product Recommendation. In L. Lamontagne & E. Plaza, eds. *Case-Based Reasoning Research and Development*. Cham: Springer International Publishing, pp. 110–124.

Ercan, G. & Cicekli, I., 2007. Using lexical chains for keyword extraction. *Information Processing and Management*, 43(6), pp.1705–1714.

Felfernig, A. & Burke, R., 2008. Constraint-based recommender systems: technologies and research issues. In *Proceedings of the 10th international conference on Electronic commerce ICEC '08*. ACM, pp. 1–10. Available at: http://portal.acm.org/citation.cfm?id=1409544.

García-Laencina, P.J., Sancho-Gómez, J.-L. & Figueiras-Vidal, A.R., 2010. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2), pp.263–282. Available at: https://doi.org/10.1007/s00521-009-0295-6.

Gatzioura, A. & Sànchez-Marrè, M., 2015. A Case-Based Recommendation Approach for Market Basket Data. *IEEE Intelligent Systems*, 30(1), pp.20–27.

Gomaa, W.H. & Fahmy, A.A., 2013. A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13).

Hariri, N. et al., 2011. Context-aware recommendation based on review mining. In *Proceedings of the 9th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP 2011)*. p. 30.

Hasan, K.S. & Ng, V., 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.1262–1273.

Hedden, H., 2014. Tagging versus indexing. *The Indexer*, 32(2), pp.81–82.

Herlocker, J.L. et al., 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1), pp.5–53. Available at: http://portal.acm.org/citation.cfm?doid=963770.963772.

Hulth, A., 2004. *Combining machine learning and natural language processing for automatic keyword extraction*, Department of Computer and Systems Sciences [Institutionen för Data-och systemvetenskap], Univ.

Hulth, A., 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*. pp. 216–223.

Humphreys, P., Mcivor, R. & Chan, F., 2003. Using case-based reasoning to evaluate supplier environmental management performance. *Expert Systems with Applications*, 25, pp.141–153.

Indurkhya, N. & Damerau, F.J., 2010. *Handbook of natural language processing*, CRC Press.

Islam, A. & Inkpen, D., 2009. Semantic similarity of short texts. *Recent Advances in Natural Language Processing V*, 309, pp.227–236.

Ittoo, A.R., Zhang, Y. & Jiao, J., 2006. A Text Mining-based Recommendation System for Customer Decision Making in Online Product Customization. In *2006 IEEE International Conference on Management of Innovation and Technology*. pp. 473–477.

Jiang, J., 2012. Information Extraction from Text. In C. C. Aggarwal & C. Zhai, eds. *Mining Text Data*. Boston, MA: Springer US, pp. 11–41. Available at: http://dx.doi.org/10.1007/978-1-4614-3223-4_2.

Jure Leskovec, Jure, Rajaraman, A. & Ullman, J.D., 2014. Recommendation Systems. In *Mining of Massive Data sets*. Cambridge University Press, pp. 307–341.

K. Zhang, H. Xu, J. Tang, J.Z.L., 2006. Keyword Extraction Using Support Vector Machine. *Proceedings of the Seventh International Conference on Web-Age Information Management (WAIM2006), Hong Kong, China*, pp.85–96.

Kefalas, P. & Manolopoulos, Y., 2017. A time-aware spatio-textual recommender system. *Expert Systems with Applications*, 78, pp.396–406.

Khozani, S.M.H. & Bayat, H., 2011. Specialization of keyword extraction approach to Persian texts. *Proceedings of the 2011 International Conference of Soft Computing and Pattern Recognition, SoCPaR 2011*, pp.112–116.

Kian, H. & Zahedi, M., 2011. an Efficient Approach for Keyword Selection; Improving Accessibility of Web Contents By General Search Engines. *International Journal of Web & …*, 2(4), pp.81–90.

Kolodner, J., 1992. An introduction to case-based Reasoning. *Artificial Intelligence Review*, 6, pp.3–34.

Krestel, R., Fankhauser, P. & Nejdl, W., 2009. Latent Dirichlet Allocation for Tag Recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*. RecSys '09. New York, NY, USA: ACM, pp. 61–68. Available at: http://doi.acm.org/10.1145/1639714.1639726.

Landauer, T.K. & Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic

analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), p.211.

Li, Y. et al., 2010. Contextual recommendation based on text mining. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. pp. 692–700.

Lorenzi, F. & Ricci, F., 2005. Case-Based Recommender Systems: A Unifying View. In B. Mobasher & S. S. Anand, eds. *Intelligent Techniques for Web Personalization*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 89–113.

Luhn, H.P., 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1(4), pp.309–317.

Lund, K. & Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), pp.203–208.

Manning, C.D. & Raghavan, P., 2009. An Introduction to Information Retrieval. *Online*, 1, p.1.

Mihalcea, R. et al., 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*. pp. 775–780.

Mihalcea Rada, T.P., 2004. TextRank: Bringing Order into Texts. In *Conference on Empirical Methods in Natural Language Processing*.

Mikolov, T., Sutskever, I., et al., 2013. Distributed Representations of Words and Phrases and their Compositionality. , pp.1–9.

Mikolov, T., Chen, K., et al., 2013. Efficient Estimation of Word Representations in Vector Space. , pp.1–12.

Mishne, G., 2006. AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts. In *Proceedings of the 15th International Conference on World Wide Web*. WWW '06. New York, NY, USA: ACM, pp. 953–954. Available at: http://doi.acm.org/10.1145/1135777.1135961.

Mokhtaripour, A. & Jahanpour, S., 2006. Introduction to a new Farsi stemmer. *Proceedings of the 15th ACM international*, (January 2006), pp.826–827.

Najafi, E. & Darooneh, A.H., 2015. The fractal patterns of words in a text: a method for automatic keyword extraction. *PloS one*, 10(6), p.e0130617.

O'Shea, E. et al., 2014. NudgeAlong: A Case Based Approach to Changing User Behaviour. In L. Lamontagne & E. Plaza, eds. *Case-Based Reasoning Research and Development*. Cham: Springer International Publishing, pp. 345–359.

Ortu??o, M. et al., 2002. Keyword detection in natural languages and DNA. *Europhysics Letters*, 57(5), pp.759–764.

Rehurek, R. & Sojka, P., 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50.

Ricci, Francesco;Rokach, Lior;Shapira, Bracha; Kantor, P.B. ed., 2011. *Recommender Systems Handbook*,

Ricci, F., Rokach, L. & Shapira, B., 2011. Introduction to Recommender Systems Handbook. In F. Ricci et al., eds. *Recommender Systems Handbook*. Boston, MA: Springer US, pp. 1–35. Available at: http://dx.doi.org/10.1007/978-0-387-85820-3_1.

Rose, S. et al., 2010. CO RI Automatic keyword extraction. *Text Mining: Applications and Theory*, pp.1--277.

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of

Information by Computer. *Analysis and Retrieval of Information by Computer*, p.530.

Salton, G., Yang, C.S. & Yu, C.T., 1975. A theory of terms importance in automatic text analysis. *Journal of American Society for Information Sciences*, 26(1), pp.33–44.

Sayyadiharikandeh, M., Ghodsi, M. & Naghibi, M., 2012. PostRank: A New Algorithm for Incremental Finding of Persian Blog Representative Words. In *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics*. WIMS '12. New York, NY, USA: ACM, p. 17:1--17:6.

Shani, G. & Gunawardana, A., 2011. Evaluating recommendation systems. In *Recommender systems handbook*. Springer, pp. 257–297.

Siddiqi, S. & Sharan, A., 2015. Keyword and Keyphrase Extraction Techniques: A Literature Review. *International Journal of Computer Applications*, 109(2), pp.975–8887.

Siddiqi, S. & Sharan, A., 2016. Keyword extraction from single documents using mean word intermediate distance. *International Journal of Advanced Computer Research*, 625(25), pp.138–145.

Smyth, B., 2007. Case-based recommendation. *The adaptive web*, pp.342–376.

Sun, C. et al., 2010. SimRank: A link analysis based blogger recommendation algorithm using text similarity. In *2010 International Conference on Machine Learning and Cybernetics*. pp. 3368–3373.

Taghva, K., Beckley, R. & Sadeh, M., 2005. A stemming algorithm for the Farsi language. *International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume II*, p.158–162 Vol. 1.

Tuarob, S. et al., 2015. A generalized topic modeling approach for automatic document annotation. *International Journal on Digital Libraries*, 16(2), pp.111–128. Available at: http://dx.doi.org/10.1007/s00799-015-0146-2.

Tuarob, S., Pouchard, L.C. & Giles, C.L., 2013. Automatic tag recommendation for metadata annotation using probabilistic topic modeling. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*, p.239. Available at: http://dl.acm.org/citation.cfm?id=2467696.2467706.

Turban, E., Sharda, R. & Delen, D., 2010. *Decision Support and Business Intelligence Systems* 9th ed., Upper Saddle River, NJ, USA: Prentice Hall Press.

Turney, P., 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *Machine Learning: ECML 2001*, pp.491–502.

Vasudevan, S.R. & Chakraborti, S., 2014. Enriching Case Descriptions Using Trails in Conversational Recommenders. In L. Lamontagne & E. Plaza, eds. *Case-Based Reasoning Research and Development: 22nd International Conference, ICCBR 2014, Cork, Ireland, September 29, 2014 - October 1, 2014. Proceedings*. Cham: Springer International Publishing, pp. 480–494. Available at: https://doi.org/10.1007/978-3-319-11209-1_34.

Weber, R.O., Ashley, K.D. & Brüninghaus, S., 2005. Textual case-based reasoning. *The Knowledge Engineering Review*, 20(3), pp.255–260.

Witten, I.H. et al., 1999. KEA: Practical Automatic Keyphrase Extraction. *In Proceedings of the fourth ACM conference on Digital libraries, ACM*, pp.254–261.

Wu, Z. & Palmer, M., 1994. Verbs Semantics and Lexical Selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*. ACL '94. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 133–138. Available at: https://doi.org/10.3115/981732.981751.

Zhang, C. et al., 2008. Automatic Keyword Extraction from Documents Using Conditional Random Fields. *Journal of Computational Information Systems*, 4(3), pp.1169–1180.

پیوست:
بخش‌های اصلی کد

۸. پیوست: بخش‌های اصلی کد روش پیشنهادی

تابع آموزش مدل کلمه-به-بردار

```
def trainWord2Vec(subject, datasetType):
    #load training data
    trainingFileName = 'corpus-' + subject + '-'+ datasetType +'.json'
    trainingArticles = json.load(codecs.open(trainingFileName, 'r', 'utf-8-sig'))

    # create word2vec model and save it
    datasetAllWords = []
    word2vecInput = []
    for article in trainingArticles:
        articleWords = []
        normalizedTitle = normalize(article['title'])
        wordList = word_tokenize(normalizedTitle)
        articleWords.append(wordList)
        normalizedAbstract = normalize(article['abstract'])
        sentenceList = sent_tokenize(normalizedAbstract)
        for sentence in sentenceList:
            sentenceWords=[]
            wordList = word_tokenize(sentence)
            for word in wordList:
                sentenceWords.append(word)
            articleWords.append(sentenceWords)
        flatList = [item for sublist in articleWords for item in sublist]
        word2vecInput.append(flatList)
    logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.INFO)
    model = gensim.models.Word2Vec(word2vecInput)
    trainingModelFileName = 'word2vec-' + subject + '-' + datasetType +  '-model.wv'
    model.wv.save_word2vec_format(trainingModelFileName)
    newModelName = subject + '-word2vec.model'
    model.save(newModelName)
    return model
```

تابع نرمال‌سازی و پیش‌پردازش داده‌ها

```
def normalize(text):
    # Normalize
    normalizedText = hazmNormalizer.normalize(text)

    # Replace some characters
    normalizedText = normalizedText.replace('؟','').replace(".", "").replace("
,",
    "").replace(":","").replace("/", "").replace("!", "").replace("?", "").rep
lace("‘", "").replace("؛",
    "").replace(";", "").replace("(", "").replace(")", "").replace("»", "").re
place("«", "")
    # Remove stop words
    wordList = word_tokenize(normalizedText)
    newWordList =[]
    for word in wordList:
        if word not in stopwordList:
            newWordList.append(word)
    # Stemming
    resultWordList =[]
    for word in newWordList:
        #word = hazmStemmer.stem(word)
        word = hazmLemmatizer.lemmatize(word)
        if (word.find('#') != -1):
            word = word.split('#')[0]
        resultWordList.append(word)
    newText = " ".join(str(x) for x in resultWordList)
    return newText
```

تابع اصلی برای بازیابی و پیشنهاد کلیدواژه

```
def findSimilarArticle(keywordNumberThreshold, beta, alpha, subject, word2vecM
odel, trainingArticles,
    testArticles, articleAvgVectors, resultFileName, similarArticleCount):

    f = codecs.open(resultFileName, "w", "utf-8-sig")
    f2 = codecs.open('pValueResult.txt', "w", "utf-8-sig")
    articleIndex = 0
    precisionSum = 0
    recallSum = 0
    articleCount = len(testArticles)
    print('number of test articles: ' + str(articleCount))

    tfidfScores = getTFIDFScore(trainingArticles, testArticles)
    CurrentArticleCounter = 0
    for i in range(len(testArticles)):
        article = testArticles[i]
        currentTestArticleId =  article['id']
        articleIndex+=1
        print('article ' + str(articleIndex) + ' of ' + str(len(testArticles))
)
        f.write('test article:' + os.linesep)
        f.write(article['title'] + os.linesep)
        f.write(article['abstract'] + os.linesep)
        f.write('similar articles:' + os.linesep)
        f.write('article keywords:' + os.linesep)
        keywords = getArticleKeywords(currentTestArticleId)

        currentArticleOriginalKeywords = keywords
        print('test article keywords:')
        print(keywords)
        currentArticleKeywordCount = len(currentArticleOriginalKeywords)

        for keyword in keywords:
            f.write(keyword + ', ')
        f.write(os.linesep + 'similar keywords:' + os.linesep)

        articleDistance ={}

        # extract top similar articles
        articleAvgVector = calculateVector(word2vecModel, testArticles, curren
tTestArticleId)
        currentIndex = 0

        for currentTrainingArticleId in articleAvgVectors:
            word2vecDist = distance.cosine(articleAvgVectors[currentTrainingAr
ticleId], articleAvgVector)
            tfidfScore = tfidfScores[int(currentTrainingArticleId), int(curren
tTestArticleId)]
            if tfidfScore >1:
                tfidfDist = 0
                print('tfidf >1')
            else:
                    tfidfDist = 1 - tfidfScore
            currentIndex += 1

            dist = (1-alpha) * word2vecDist + alpha * tfidfDist
            articleDistance[currentTrainingArticleId] = dist
        print('retrieval finished**********************************')
```

```
54.         sortedDistance = sorted(articleDistance.items(), key=operator.itemgett
    er(1))
55.
56.         for i in range(similarArticleCount):
57.             articleId = int(list(sortedDistance)[i][0])
58.             articleDis = float(list(sortedDistance)[i][1])
59.             similarArticle = getArticle(trainingArticles, articleId)
60.             f.write('similar item: '+ str(i) + os.linesep)
61.             f.write('article Id: '+ str(articleId) + os.linesep)
62.             f.write('article Score: '+ str(articleDis) + os.linesep)
63.             f.write(similarArticle['title'] + os.linesep)
64.             f.write(similarArticle['abstract'] + os.linesep)
65.         f.write(os.linesep)
66.
67.         allKeywords = []
68.         recommendedKeywords = []
69.         articleKeywords = {}
70.         firstScore = {}
71.         Xj = {}
72.         Y = {}
73.         secondScore = {}
74.         sumScore = {}
75.         # extract keywords
76.         for i in range(similarArticleCount):
77.             articleId = int(list(sortedDistance)[i][0])
78.             articleDis = float(list(sortedDistance)[i][1])
79.             keywords = getArticleKeywords(articleId)
80.             articleKeywords[articleId] = keywords
81.             for keyword in keywords:
82.                 if keyword not in allKeywords:
83.                     allKeywords.append(keyword)
84.         allKeywordsList=allKeywords
85.         allKeywordsCount = len(allKeywordsList)
86.         allSum = 0
87.         selectedSum = 0
88.         for i in range(allKeywordsCount):
89.             print('keyword ' + str(i) + ' of ' + str(allKeywordsCount))
90.
91.             # first sorting method
92.             currentKeyword = allKeywordsList[i]
93.             for j in range(similarArticleCount):
94.                 articleId = int(list(sortedDistance)[j][0])
95.                 ##retrieval score of each document
96.                 score_retrieval = calculateScore(word2vecModel, testArticles,
    articleAvgVectors, articleId, currentTestArticleId)
97.                 keywords = getArticleKeywords(articleId)
98.                 if currentKeyword in keywords:
99.                     selectedSum += score_retrieval
100.                        else:
101.                            allSum += score_retrieval
102.                    currentScore = selectedSum / allSum
103.                    firstScore[currentKeyword] = currentScore
104.
105.                    # second sorting method
106.                    Y[currentKeyword] = calculateY(currentKeyword, word2vecMode
    l, testArticles, currentTestArticleId)
107.                    secondScore[currentKeyword] = Y[currentKeyword]
108.                    if (secondScore[currentKeyword]>0):
109.                        sumScore[currentKeyword] = (firstScore[currentKeyword]+
     beta * secondScore[currentKeyword]) / 2
110.                    else:
111.                        sumScore[currentKeyword] = firstScore[currentKeyword]
```

```
112.                sortedSumScore = sorted(sumScore.items(), key=operator.itemgett
    er(1), reverse=True)
113.                sortedSumScore2 = sorted(secondScore.items(), key=operator.item
    getter(1), reverse=True)
114.                sortedSumScore1 = sorted(firstScore.items(), key=operator.itemg
    etter(1), reverse=True)
115.                for i in range(len(sortedSumScore)):
116.                    keyword = list(sortedSumScore)[i][0]
117.                    keywordScore = float(list(sortedSumScore)[i][1])
118.                    currentFirstScore = firstScore[keyword]
119.                    currentSecondScore = secondScore[keyword]
120.                    f.write('similar item: '+ str(i) + os.linesep)
121.                    f.write('keyword: '+ keyword + ', ' + 'FirstScore: '+ str(c
    urrentFirstScore) + ', SecondScore: '+ str(currentSecondScore) + ', Score: '+
    str(keywordScore) + os.linesep)
122.
123.                    # save recommended keywords
124.                    if (keyword not in recommendedKeywords):
125.                        recommendedKeywords.append(keyword)
126.
127.                for i in range(len(sortedSumScore2)):
128.                    keyword = list(sortedSumScore2)[i][0]
129.                    keywordScore2 = float(list(sortedSumScore2)[i][1])
130.                    currentSecondScore = secondScore[keyword]
131.                    f.write('similar item: '+ str(i) + os.linesep)
132.                    f.write('keyword: '+ keyword + ', ' + 'SecondScore: '+ str(
    currentSecondScore) + os.linesep)
133.
134.
135.                for i in range(len(sortedSumScore1)):
136.                    keyword = list(sortedSumScore1)[i][0]
137.                    keywordScore1 = float(list(sortedSumScore1)[i][1])
138.                    currentFirstScore = firstScore[keyword]
139.                    f.write('similar item: '+ str(i) + os.linesep)
140.                    f.write('keyword: '+ keyword + ', ' + 'FirstScore: '+ str(c
    urrentFirstScore) + os.linesep)
141.
142.                #calculate similarity between original keywords and the article

143.                currentArticleKeywordCount = len(currentArticleOriginalKeywords
    )
144.                similarityOriginalKeyword = 0
145.                avgSimilarityOriginalKeywords = 0
146.                originalKeywordW2VSim = 0
147.                similarityOriginalKeywordsVector = []
148.                for j in range(currentArticleKeywordCount):
149.                    currentOriginalKeyword = currentArticleOriginalKeywords[j]

150.                    print(currentOriginalKeyword)
151.                    originalKeywordW2VSim = calculateY(currentOriginalKeyword,
    word2vecModel, testArticles, currentTestArticleId)
152.                    similarityOriginalKeywordsVector.append(originalKeywordW2VS
    im)
153.                    print(similarityOriginalKeywordsVector)
154.                    #similarityOriginalKeywordsVector[j] = originalKeywordW2VSi
    m
155.                    similarityOriginalKeyword += originalKeywordW2VSim
156.                avgSimilarityOriginalKeywords = similarityOriginalKeyword / cur
    rentArticleKeywordCount
157.
158.                # calculate recall
159.                commonKeywordCount = 0
160.                recall = 0
```

```
161.                recommendedKeywordsCount = len(recommendedKeywords)
162.                if recommendedKeywordsCount > keywordNumberThreshold:
163.                    recommendedKeywordsCount = keywordNumberThreshold
164.                similarityRecommendKeywords = 0
165.                similarItemsCount = 0
166.                similarityRecommendKeywordsVector = []
167.                avgSimilarityRecommendKeywords = 0
168.
169.                for i in range(recommendedKeywordsCount):
170.                    #calculate averaged semantic similarity between recomm keyw
    ords and est document
171.                    currentRecommendedKeyword = recommendedKeywords[i]
172.                    secondScoreRecommenKeyword = secondScore[currentRecommended
    Keyword]
173.                    if secondScoreRecommenKeyword > 0:
174.                        similarityRecommendKeywords += secondScoreRecommenKeywo
    rd
175.                        similarityRecommendKeywordsVector.append(secondScoreRec
    ommenKeyword)
176.                        similarItemsCount += 1
177.
178.                    if (currentRecommendedKeyword in currentArticleOriginalKeyw
    ords):
179.                        commonKeywordCount += 1
180.
181.                avgSimilarityRecommendKeywords = similarityRecommendKeywords /
     similarItemsCount
182.
183.                ## Kolmogorov Smirnov Test
184.                ## H0= same distribution,     H1=not the same distribution
185.                pValue = 0
186.                [KSstatistics, pValue] = stats.ks_2samp(similarityOriginalKeywo
    rdsVector, similarityRecommendKeywordsVector)
187.                f2.write('p-
    value ' + str(CurrentArticleCounter) + ' = ' + str(pValue) + os.linesep)
188.                CurrentArticleCounter += 1
189.
190.                 if (currentArticleKeywordCount>0):
191.                    recall = commonKeywordCount / currentArticleKeywordCount
192.                 f.write('recall: '+ str(recall) + os.linesep)
193.                f2.write('recall: '+ str(recall) + os.linesep)
194.                f.write('Original Keywords Similarity: ' + str(avgSimilarityOri
    ginalKeywords) + os.linesep)
195.                f.write('Recomm Keywords Semantic Similarity: ' + str(avgSimila
    rityRecommendKeywords) + os.linesep)
196.                recallSum += recall
197.            recallAll = recallSum / articleCount
198.             f.write('recall all: '+ str(recallAll) + os.linesep)
199.            f2.write('recall all: '+ str(recallAll) + os.linesep)
200.            f.write(os.linesep)
201.            f2.write(os.linesep)
202.            f.close()
203.            f2.close()
```

تابع محاسبه شباهت معنایی براساس مدل بردار-به-کلمه بین کلیدواژه و سند

```
## calculte the average distnace between a keyword and an article,
    ## based on the distance between the keyword and every term in the article

def calculateY(currentKeyword, word2vecModel, testArticles, currentTestArticleId):

        result = 0
        currentKeywordVector = [0] * 100
        currentKeywordVector = calculateXj(word2vecModel , currentKeyword)
        currentKeywordExist = sum(calculateXj(word2vecModel , currentKeyword))

 ##             currentKeywordVector = word2vecModel.wv.word_vec(currentKeyword)
        if currentKeywordExist == 0:
            pass

        else:
                result = 0
                sumSimilarity = 0

                currentTestArticle=getArticle(testArticles, currentTestArticleId)
                articleWords = []
                normalizedTitle = normalize(currentTestArticle['title'])
                wordList = word_tokenize(normalizedTitle)
                articleWords.append(wordList)
                normalizedAbstract = normalize(currentTestArticle['abstract'])

                sentenceList = sent_tokenize(normalizedAbstract)
                for sentence in sentenceList:
                        sentenceWords=[]
                        wordList = word_tokenize(sentence)
                        for word in wordList:
                                sentenceWords.append(word)
                        articleWords.append(sentenceWords)
                articleWordsList = [item for sublist in articleWords for item in sublist]
                similarWordCount = 0
                for word in articleWordsList:
                        try:
                                wordVector = word2vecModel.wv.word_vec(word)
                                similarity = spatial.distance.cosine(wordVector, currentKeywordVector)
                                sumSimilarity += similarity
                                similarWordCount +=1
                        except:
                                pass
                wordCount = len(articleWordsList)  ## number of all terms in the new article
                result = sumSimilarity / similarWordCount
        return result

```

ABSTRACT

Keyword Extraction is known as a key step in document indexing. Keywords are semantic and contenet-based descriptors of a document, which can be used in document retrieval and representation. In databases containing scientific documents, such as Ganj in Irannian Research Institue for Information Science and Technology (IranDoc), it is even more critical to assign meaningful keywords for documents, since the documents are from different academic disciplines and contain technical terms.

As the number of scientific documents grows exponentially, having an automatic and intelligent keyword extraction technique is getting more critical. There are various keyword extraction techniques which are either based on statistical features of the text or machine learning approaches, and sometimes a combination of both. In this research, we are proposing a new keyword extraction method for Persian scientific documents based on recommender systems and case-based reasoning. The proposed method is designed based case-based reasoning in which the main assumption is that similar documents share similar keywords. There are two main steps in the proposed approach: first, similar documents to a given new document are retrieved based on TFIDF and word2vec model, second, the candidate keywords are extracted from retrieved documents and ranked based on a new scoring scheme, and a set of keyword are selected from the candidate keywords based on their score. The proposed method is tested and avaluated on a set of documents of Ganj database in three different subject areas (Art, Humanities and Engineering), based on precision, recall and expert panel.

Keywords: keyword extraction; recommender systems; case-based reasoning; word2vec embedding; information retrieval; machine learning.

Iranian Research Institute
for Information Science and Technology
(I r a n D o c)

Research Project

A New Intelligent Approach For Keyword Extraction From Persian Scientific Documents Based On Recommender Systems

By:

Azadeh Mohebi

Ammar Jalalimanesh

Winter 2019