مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری
Regional Information Center for Science and Technology RICeST

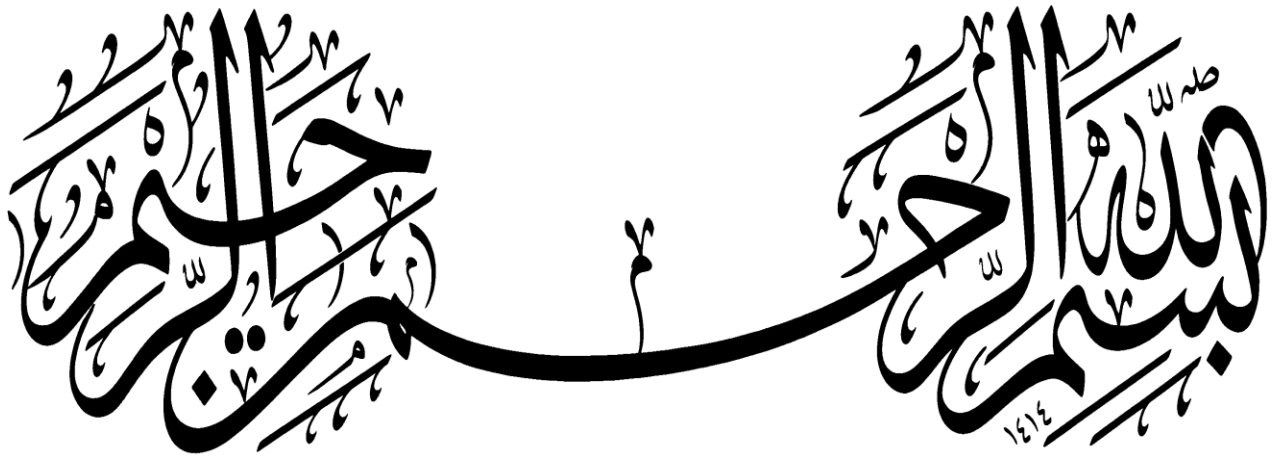
گزارش نهایی طرح

ارائه الگوریتمی برای یافتن مقالات علمی مرتبط با یک مقاله

توسط:

دکتر نیلوفر مظفری

شهریور ۱۴۰۱


بسم الله الرحمن الرحيم
١٤١٤

چکیده

زمینه و هدف: ظهور اینترنت موجب شده است که حجم مستندات و مقالات علمی در دهه اخیر افزایش بسیار زیادی پیدا کند که این امر، دستیابی به مقالات مورد نظر کاربر را با مشکلات زیادی روبرو کرده است. سامانه‌های بازیابی مقالات علمی تلاشی برای کمک به کاربر در جهت بازیابی مقالات و مستندات علمی مورد نظر کاربر هستند. یکی از قابلیت‌هایی که در سامانه‌های بازیابی اطلاعات مقالات علمی به کاربران و پژوهشگران کمک بسیار زیادی می‌کند، ویژگی یافتن مقالات علمی مرتبط با یک مقاله است. به عبارت دیگر، ویژگی که به پژوهشگر اجازه می‌دهد با انتخاب یکی از مقالات بازیابی شده، دیگر مقالات مرتبط با آن را مشاهده و بازیابی نماید.

روش‌شناسی: در این پژوهش، روشی مبتنی بر تحلیل مراجع برای یافتن مقالات مرتبط با یک مقاله ارائه شده است. جامعه پژوهش شامل داده‌های فارسی و انگلیسی مستخرج از نشریات حوزه علوم کامپیوتر می‌باشند. روش پیشنهادی بدین صورت عمل می‌کند که شباهت میان مراجع مقاله داده شده با مراجع دیگر مقالات محاسبه می‌کند و مقالاتی به عنوان مقاله مرتبط با یک مقاله بازیابی شده که عناوین آنها بیشترین شباهت را با یکدیگر داشته باشند. روش ارائه شده قادر است تا عناوین مقالات را در فرمت‌های مختلف APA، Vancouver، MLA، Chicago و Harvard استخراج کند. در نهایت، بر اساس شباهت میان عنوان مراجع، گراف شباهت مراجع در فاز آفلاین استخراج شده که در فاز آنلاین مقالات مرتبط بازیابی می‌گردد.

یافته‌ها: به منظور ارزیابی دقیق‌تر روش پیشنهادی، به ازای داده‌های انگلیسی که از نشریات انگلیسی گرفته شده‌اند، ۳۰ مقاله به صورت تصادف انتخاب شدند. همین رویه برای داده‌های فارسی نیز انجام گرفت و ۳۰ مقاله به صورت تصادفی از میان مجموعه مقالات انتخاب گردیدند. به ازای هر مقاله، اطلاعات کتابشناختی که حاوی عنوان، کلیدواژه و چکیده مقالات است به همراه لیست مراجع آن استخراج شد.

نتیجه‌گیری: نتایج اعمال روش پیشنهادی روی داده‌های فارسی و انگلیسی نشریات انتخاب شده در علوم کامپیوتر نشان‌دهنده کارایی آن در یافتن مقالات مرتبط با یک مقاله است. به عبارت دیگر، در صورتی که پوشش نسبتا جامعی از استنادهای مقالات علمی وجود داشته باشد، روش پیشنهادی قادر خواهد بود که با دقت بالایی مقالات مرتبط با یک مقاله را پیدا کند. الگوریتم ارائه شده می‌تواند در سامانه‌های بازیابی مقالات علمی که اطلاعات استنادی مقالات را داشته باشند، استفاده گردد و یک ویژگی ارزشمندی را به سامانه اضافه نماید که به کاربرپسندتر شدن سامانه بازیابی اطلاعات کمک شایانی می‌نماید.

واژگان کلیدی: تحلیل مراجع کتابشناختی، بازیابی اطلاعات، معیار شباهت، دقت.

فهرست مطالب

فهرست جداول

عنوان **صفحه**

فهرست اشکال

عنوان **صفحه**

اشکال فصل دوم

اشکال فصل سوم

اشکال فصل چهارم

فصل اول

مقدمه

۱. مقدمه

۱_۱ مقدمه

در سال‌های اخیر با ظهور وب و اینترنت به عنوان یک سیستم اطلاع‌رسانی جهانی، توانایی بشر برای تولید و جمع‌آوری داده‌ها افزایش چشمگیری داشته است؛ به صورتی که بشر با حجم بسیار زیادی از داده و اطلاعات روبرو شده است. از آنجا که درصد بسیار زیادی از این اطلاعات را مستندات و منابع متنی تشکیل می‌دهند، کشف دانش از این حجم نیاز به درک و شناسایی روابط میان آنها است. از طرف دیگر، در عصر تکنولوژی که در آن قرار داریم، اعتقاد بر این است که هر چیزی باید به صورت خودکار انجام گیرد. متن‌کاوی و یا کشف دانش از میان مستندات متنی، تلاشی برای نیل به این هدف است. اصلی‌ترین دلیلی که باعث شد متن‌کاوی کانون توجهات قرار بگیرد، مساله در دسترس بودن حجم وسیعی از داده‌های متنی و نیاز شدید به اینکه از این داده‌های متنی، اطلاعات و دانش سودمند بتوان استخراج کرد (Feldman et al., 1995).

متن‌کاوی به معنای استخراج اطلاعات گرانبها از حجم عظیم داده متنی است. با توجه به نوع داده و همچنین حجم آن، مشخص نیست که چه اطلاعات گرانبهایی در عمق این داده‌های متنی وجود دارد و تنها با کاوش در این داده‌هاست که می‌توان به این اطلاعات گرانقدر دسترسی پیدا کرد. بنابراین وظیفه اصلی متن‌کاوی، کاویدن و استخراج دانش از منابع عظیم داده متنی می‌باشد؛ تا اطلاعات گرانبهایی که در حجم انبوهی از اطلاعات سطحی پنهان شده است، آشکار گردد. متن‌کاوی تلاش برای استخراج دانش از انبوه داده‌های متنی موجود است که به کمک مجموعه‌ای از روش‌های آماری و مدلسازی می‌تواند الگوها و روابط پنهان موجود در داده‌های متنی را تشخیص دهد. تحلیل متن[۱] اصطلاحی است که گاها به جای متن‌کاوی استفاده می‌شود که آن هم به فرآیند

[۱] Text analysis

تبدیل داده‌های متنی غیرساخت‌یافته به اطلاعات با معنا اطلاق می‌شود. برای تحلیل متن و یا به عبارتی متن‌کاوی نیازمند الگوریتم‌های یادگیری ماشین می‌باشیم (Hotho et al., 2005).

یکی از کاربردهای بسیار مفید از متن‌کاوی در موتور جستجو[۱] انجام می‌گیرد؛ جایی که کاربران نیاز به یافتن اطلاعات مرتبط به پرس و جو[۲] دارند. موتور جستجو اساسا هر برنامه کامپیوتری است که برای یافتن اطلاعات مورد نظر کاربر نوشته می‌شود و می‌تواند در هر حوزه‌ای مورد استفاده قرار گیرد. یک موتور جستجو که در واقع یک سیستم پاسخ‌دهی است، اساسا از دو بخش اصلی تشکیل شده است: پایگاه داده اطلاعات و هسته موتور که از الگوریتم‌هایی تشکیل شده است. اولین موتور جستجو، آرچی[۳] نام داشت که برای جستجو میان عناوین مورد استفاده قرار می‌گرفت و توانایی نمایش محتوای وب را نداشت. موتورهای جستجوی ورونیکا[۴] و جاگ‌هد[۵] به دنبال پروژه آرچی با هدف نمایه[۶]کردن متن ساده بوجود آمدند. به دنبال این موتورهای جستجو، موتورهای جستجوی دیگری بوجود آمدند که هر کدام، برای بهبود قبلی تلاش می‌کردند تا اینکه حدود سال ۱۹۹۸ دامنه google.com ثبت گردید و از آن پس، گوگل به عنوان قوی‌ترین و پراستفاده‌ترین موتور جستجو در تمامی پلتفرم‌ها معرفی شد و توانست محبوبیت بسیار زیادی را میان کاربران کسب نماید.

هر موتور جستجو برای کشف، دسته‌بندی و رتبه‌بندی اسنادی که در اختیار دارد، نیاز به انجام سه فرآیند کلی دارد که تحت عناوین خزیدن[۷]، نمایه‌کردن[۸] و رتبه‌بندی کردن[۹] شناخته می‌شوند. با استفاده از فرآیند خزیدن که عمدتا توسط ربات‌های با عنوان خزنده[۱۰] یا عنکبوت انجام می‌گیرد، داده‌های بسیار زیادی به پایگاه داده وارد می‌شوند. قبل از ذخیره داده‌ها در پایگاه داده، عملیات نمایه‌گذاری انجام شده که در واقع نحوه ذخیره کردن داده‌ها در پایگاه داده توسط این فرآیند انجام می‌گیرد. لازم به ذکر است که هر موتور جستجو، پایگاه داده مخصوص به خود را دارد و از یک فرآیند نمایه‌گذاری با توجه به فیلدهای پایگاه داده بهره می‌برد. هنگام جستجوی کاربر، موتور جستجو با استفاده از الگوریتم‌های رتبه‌بندی، اسناد مرتبط با پرسش کاربر را یافته و آنها را به صورت مرتب‌شده نمایش می‌دهد.

---

۱ Search engine
۲ Query
۳ Archie
۴ Veronica
۵ Jughead
۶ indext
۷ crawling
۸ indexing
۹ ranking
۱۰ crawler

در ایران نیز چندین موتور جستجو طراحی و پیاده‌سازی شده است که هر کدام برای پاسخ به نیاز کاربر در یک حوزه است که یکی از آنها برای جستجوی مقالات علمی می‌باشد. این موتورهای جستجو قادرند تا از میان مقالات موجود در پایگاه‌های اطلاعاتی، مقالات مرتبط با پرسش کاربر را پیدا کنند. یکی از قابلیت‌هایی که در موتورهای جستجو منجر به افزایش کاربرپسندتر بودن[1] و راحتی کاربر در استفاده از آن می‌شود، یافتن مقالات مرتبط با یک مقاله است. زمانی که کاربر یک عبارت را جستجو می‌کند، موتور جستجو مقالات مرتبط با آن را یافته و به وی نشان می‌دهد. از میان لیستی که به کاربر نشان داده می‌شود، ممکن است کاربر یکی از آن مقالات را مرتبط با پرسش یافته و حال قصد دارد تا مقالات بیشتری که مرتبط با آن مقاله است، بیابد.

برای انجام این پژوهش، نیاز به پردازش زبان وجود دارد که این پردازش در هر سطح، نیازمند دانش، منابع و پیکره‌های مورد نیاز آن سطح و سطوح پایین‌تر است. در دسترس بودن منابع و دانش برای انجام تحقیق در حیطه‌ی پردازش زبان طبیعی از جمله چالش‌های پردازش زبان طبیعی است.

بنابراین این پژوهش تلاش دارد تا با ارائه تکنیکی، مقالات مرتبط با یک مقاله را بیابد؛ بدین صورت که وقتی موتور جستجو مقاله‌ای را پیدا کرد، الگوریتم ارائه بتواند مقالات مرتبط با آن مقاله را پیدا کند.

۱_۲ بیان مساله

قابلیت یافتن مقالات مرتبط با یک مقاله که به عنوان یک ویژگی ارزشمند در سامانه‌های بازیابی اطلاعات مقالات علمی قابل استفاده است، بدین صورت عمل می‌کند که با داشتن یک مقاله، به دنبال دیگر مقالات علمی مرتبط با آن مقاله می‌گردد و آنها را بازیابی می‌نماید.

فرض کنید که مجموعه‌ای از مقالات $P = \{p_1.p_2.p_3 \dots p_n\}$ وجود داشته باشد. به عبارت دیگر پایگاه داده حاوی n مقاله می‌باشد که به ازای هر مقاله، اطلاعات کتابشناختی عنوان، کلیدواژه و چکیده وجود دارد. همچنین لیست مراجع هر مقاله به صورت $Ref_{p_i} = \{R_1.R_2 \dots R_m\}$ وجود دارد. ویژگی یافتن مقالات مرتبط با یک مقاله بدین صورت عمل می‌کند که با داشتن مقاله p_i، T مقاله‌ای که در مجموعه P با p_i مرتبط هستند، بازیابی کند.

[1] User-friendly

۱_۳ ضرورت و اهمیت پژوهش

در هر سامانه بازیابی اطلاعات، طراحی و پیاده‌سازی روشی که بتواند به صورت کارآمد ارتباط میان مقالات را تعیین کند، همواره مورد توجه بسیاری از محققین در زمینه‌های هوش مصنوعی و بازیابی اطلاعات بوده است. تعیین ارتباط معنایی میان دو مقاله، معمولا از طریق محاسبات آماری و استفاده از معیارهای شباهت مختلف در سطوح مختلف، انجام می‌گیرد. این امر کاربردهای بسیار زیادی در حوزه پردازش زبان طبیعی دارد. در سال‌های اخیر، ظهور اینترنت و افزایش حجم داده‌های متنی از یک طرف و نیاز کاربر برای یافتن مقالات مورد نظر از طرف دیگر باعث شده است که این امر، مورد توجه محققین قرار بگیرد.

در مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، نیز با توجه به یکی از ماموریت‌های این سازمان مبنی بر "افزایش سهولت و پایداری در دستیابی به منابع اطلاعاتی تولید شده در ایران و منطقه از طریق بازطراحی و کاربرپسندی سیستم ذخیره و بازیابی اطلاعات"، می‌بایست به کاربران کمک کرد تا بتوانند به صورت موثرتری اطلاعات و مقاله‌های مورد نیازشان را بازیابی نمایند.

با توجه به چشم‌انداز مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری با دو ویژگی "سرآمد در ایران و منطقه در زمینه مدیریت داده، اطلاعات و دانش و فراهم کردن دسترسی به جامع‌ترین منابع اطلاعاتی و دانشی چندزبانی علمی و فنی" و همچنین "پیش‌رو در ارائه خدمات اطلاعاتی نوین و پایدار با استفاده از برنامه‌های نرم‌افزاری هوشمند برای پژوهشگران و دانشمندان ایران و منطقه"، طرح حاضر می‌تواند در نیل به این اهداف موثر واقع گردد. همچنین این پژوهش در راستای یکی از ماموریت‌های مرکز منطقه‌ای، تحت عنوان "افزایش سهولت و پایداری در دستیابی به منابع اطلاعاتی تولید شده در ایران و منطقه از طریق بازطراحی و کاربرپسندی سیستم ذخیره و بازیابی اطلاعات" می‌باشد. همچنین این پژوهش در راستای اهداف زیر، بر اساس سند راهبردی مرکز منطقه‌ای نیز است:

- بکارگیری فناوری‌های نوین ذخیره و بازیابی اطلاعات و دانش تولید شده (در هر قالب متنی و غیرمتنی) به منظور تسهیل ارائه خدمات سریع و دقیق به دانشمندان و پژوهشگران ایران و منطقه
- ترویج و حمایت از پژوهش‌های کاربردی در زمینه فناوری‌های مدیریت داده، اطلاعات و دانش در راستای بهینه‌سازی آنها.

۱_۴ مروری بر ساختار گزارش

در ادامه در فصل ۲، مبانی نظری و مروری بر روش‌های پیشین و همچنین بررسی سامانه‌های مشابه خواهیم داشت. فصل ۳ به روش روش‌شناسی پژوهش مورد بررسی قرار می‌گیرد. نتایج و یافته‌های بدست آمده در فصل ۴ بیان می‌شوند و در نهایت نتیجه‌گیری در فصل ۵ آمده است.

فصل دوم

مبانی نظری و پیشینه پژوهش

۲. مبانی نظری و پیشینه پژوهش

۲_۱ مقدمه

در این فصل، مبانی نظری پژوهش مورد بررسی قرار داده می‌شوند. در ادامه، مروری بر روش‌های گذشته خواهیم داشت. همچنین سامانه‌های مشابه برای بازیابی مقالات علمی را از دیدگاه قابلیت یافتن مقالات مرتبط با یک مقاله، بررسی خواهیم کرد.

۲_۲ مبانی نظری

در ادامه این فصل، تعاریف و مبانی نظری مورد استفاده در این پژوهش ارائه می‌گردد.

۲_۲_۱ متن‌کاوی

بشر با پیشرفت فناوری رایانه‌ای در ثبت و ذخیره‌سازی داده‌ها و پردازش آنها گامی بزرگ جهت کسب دانش برداشته است. در واقع داده نمایشی از واقعیت‌ها، معلومات، مفاهیم، رویدادها یا پدیده‌ها برای برقراری ارتباط، تفسیر یا پردازش توسط انسان یا ماشین است. از طرف دیگر واژه‌ی اطلاعات به معنی دانشی از طریق خواندن، مشاهده و آموزش بدست می‌آید، اطلاق می‌گردد و در حقیقت می‌توان گفت اطلاعات، داده‌هایی هستند که پس از جمع‌آوری پردازش شده و شکل مفهومی تولید کرده‌اند. اصلی‌ترین دلیلی که باعث شد متن‌کاوی کانون توجهات قرار بگیرد، مساله در دسترس بودن حجم وسیعی از داده‌های متنی و نیاز شدید به اینکه از این داده‌های متنی، اطلاعات و دانش سودمند بتوان استخراج کرد (Feldman et al., 1995).

بین داده‌ها و اطلاعات همانند خبر و اطلاع رابطه وجود دارد. خبری که دریافت می‌شود، پس از ارزیابی به اطلاع تبدیل می‌شود. داده‌ها نیز پردازش می‌شوند تا اطلاعات را پدید آورند. به بیان دیگر، اطلاع حاصل تکامل داده‌ها است. به این ترتیب بین داده‌ها و اطلاعات یک شکاف وجود دارد که اندازه این شکاف با حجم داده‌ها ارتباط مستقیم دارد. هر چه داده‌ها حجیم‌تر باشند این شکاف بیشتر خواهد بود و هر چه حجم داده‌ها کمتر و روش‌ها و ابزار پردازش داده‌ها کاراتر باشد، فاصله بین داده‌ها و اطلاعات کمتر است. امروزه افزایش سریع حجم داده‌ها به شکلی است که توانایی انسان برای درک این داده‌ها بدون ابزارهای پرقدرت میسر نمی‌باشد (عظیمی و شمس، ۱۳۹۴).

متن‌کاوی به معنای استخراج اطلاعات گرانبها از حجم عظیم داده متنی است. با توجه به نوع داده و همچنین حجم آن، مشخص نیست که چه اطلاعات گرانبهایی در عمق این داده‌های متنی وجود دارد و تنها با کاوش در این داده‌هاست که می‌توان به این اطلاعات گرانقدر دسترسی پیدا کرد. بنابراین وظیفه اصلی متن‌کاوی، کاویدن و استخراج دانش از منابع عظیم داده متنی می‌باشد؛ تا اطلاعات گرانبهایی که در حجم انبوهی از اطلاعات سطحی پنهان شده است، آشکار گردد. متن‌کاوی تلاش برای استخراج دانش از انبوه داده‌های متنی موجود است که به کمک مجموعه‌ای از روش‌های آماری و مدلسازی می‌تواند الگوها و روابط پنهان موجود در داده‌های متنی را تشخیص دهد. تحلیل متن[۱] اصطلاحی است که گاها به جای متن‌کاوی استفاده می‌شود که آن هم به فرآیند تبدیل داده‌های متنی غیرساخت‌یافته به اطلاعات با معنا اطلاق می‌شود. برای تحلیل متن و یا به عبارتی متن‌کاوی نیازمند الگوریتم‌های یادگیری ماشین می‌باشیم (Hotho et al., 2005).

۲_۲_۲ یادگیری ماشین[۲]

یادگیری ماشین به عنوان یکی از زیرشاخه‌های وسیع و پرکاربرد هوش مصنوعی[۳]، کامپیوتر را قادر به یادگیری با توجه به داده‌ها می‌کند. به عبارت دیگر، یادگیری ماشینی به تنظیم و اکتشاف الگوریتم‌هایی می‌پردازد که با توجه به آنها ماشین، می‌تواند یاد بگیرد. اگر یادگیری انسان را با وجود یک عامل و با تعامل با محیط بیرونی در نظر بگیریم، یادگیری ماشین با نوشتن برنامه، نمایش مثال‌های متعدد، تجربه‌ی محیط واقعی، مشاهده و بازخورد صورت می‌گیرد. زمانی الگوریتم‌های یادگیری ماشین، تاثیر خود را پررنگ‌تر نشان می‌دهند که مسئله توصیف‌ناپذیر باشد و همچنین در طول زمان تغییر کند. دسته‌بندی‌های مختلفی از الگوریتم‌های یادگیری ماشین وجود دارد که در این پژوهش، یادگیری بدون نظارت[۴] بررسی می‌شوند.

---

۱ Text analysis

۲ Machine learning

۳ Artificial intelligence

۴ Unsupervised learning

۲_۲_۳ یادگیری بدون نظارت

در این نوع شیوه یادگیری، هیچ داده آموزشی[۱] وجود ندارد. به عبارت دیگر، یادگیری بر روی داده‌های بدون برچسب[۲] به منظور یافتن الگوهای پنهان صورت می‌پذیرد. یکی از معروف‌ترین الگوریتم‌ها در این دسته، خوشه-یابی[۳] با هدف دسته‌بندی داده‌ها به گروه‌های مختلف است که به هر کدام از این دسته‌ها یک خوشه[۴] گفته می‌شود. مدل‌های موضوعی نیز از این شیوه یادگیری تبعیت می‌کنند.

۲_۲_۴ روش one-hot

ورودی اکثر روش‌های مبتنی بر یادگیری ماشین، بردار است. بنابراین در هر حوزه‌ای نیاز است که ابتدا داده‌ها را به بردار تبدیل شوند. یکی از روش‌های متداول برای تبدیل متن به بردار، روش نمایش one-hot است. در این شیوه نمایش، هر کلمه به یک بردار تبدیل می‌شود. فرض کنید که پیکره‌ای به طول N کلمه وجود دارد. حال برای نمایش هر کلمه، برداری به طول N کلمه در نظر گرفته می‌شود که هر خانه آن، متناظر با یک کلمه در پیکره است. با این پیش‌فرض، برای هر کلمه، یک بردار به طول N ایجاد می‌گردد که همه خانه‌های آن بجز خانه متناظر با آن کلمه، صفر خواهد بود و در خود ستون متناظر با کلمه مربوطه، عدد یک ذخیره می‌شود.

۲_۲_۵ TF-IDF

این روش، مبتنی بر روش one-hot است و تعداد تکرار یک کلمه در یک سند را در مقابل تعداد تکرار آن کلمه در مجموعه تمام اسناد در نظر می‌گیرد. به عبارت دیگر، این روش تعداد تکرار یک کلمه را به صورت نرمال‌شده محاسبه می‌نماید. سپس آن را در تعداد تکرار سندهایی که حاوی آن کلمه باشند، ضرب می‌کند. بنابراین کلماتی که در اسناد زیادی ظاهر شده باشند، وزن کمتری می‌گیرند. هرچقدر که تعداد تکرار یک کلمه در یک سند بیشتر باشد و آن کلمه در تعداد اسناد کمتری ظاهر شده باشد، وزن بیشتری می‌گیرد و از اهمیت بیشتری برخوردار می‌شود.

---

[۱] train

[۲] Unlabeled data

[۳] Clustering

[۴] Cluster

word2vec ۶_۲_۲

هدف این الگوریتم یافتن کلماتی است که برای پیش‌بینی کلمات همسایه در یک جمله یا سند مفید باشند. در word2vec تلاش بر این است که میانگین لگاریتم احتمال آمدن واژگان در اطراف یک کلمه (فرمول ۱) بیشینه گردد.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \quad (۱)$$

در این فرمول، c تعداد داده‌هایی که برای آموزش در دسترس است، نشان می‌دهد که طبیعتا با افزایش این مقدار، دقت و به طبع آن هزینه هم بالاتر می‌رود.

BERT ۸_۲_۲

یکی از الگوریتم‌های قوی در تعبیه کلمات، الگوریتمBERT است که مخفف نمایش رمزگذاری دوطرفه از ترانسفورمرها[۱] است و توسط گوگل در سال ۲۰۱۸ ارائه گردید. BERT از ترنسفورمر[۲] و مکانیزم توجه[۳] استفاده کرده و روابط بافتی را میان کلمات در یک متن یاد می‌گیرد. برای تولید مدل زبانی، BERT از یک رمزگذار بهره می‌برد. برخلاف مدل‌های جهت‌دار که کل متن را به صورت دنباله‌وار (از چپ به راست یا راست به چپ) می‌خوانند، رمزگذار ترنسفورمر کل دنباله کلمات را دوطرفه می‌خواند. این ویژگی به مدل اجازه می‌دهد تا مفهوم یک کلمه را بر اساس تمام اطرافیانش یاد بگیرد. بنابراین می‌توان گفت که این مدل، تعبیه کلمات مبتنی بر بافت است؛ برخلاف دیگر مدل‌های تعبیه کلمات که مستقل از بافت هستند.

۳_۲ پیشینه پژوهش

معیارهای شباهت متن، نقش بسیار مهمی را در پژوهش‌های مرتبط با متن و یافتن مقالات مرتبط با یک مقاله ایفا می‌کنند و کاربردهای بسیار زیادی در دیگر زمینه‌ها از جمله بازیابی متن[۴]، طبقه‌بندی متن[۵]، خوشه‌یابی

[1] Bidirectional Encoder Representations from Transformers
[۲] Transformer
[۳] Attention mechanism
[۴] Information retrieval
[۵] Text classification

اسناد[۱]، تشخیص موضوع[۲]، دنبال کردن موضوع[۳]، تولید سوال[۴]، پاسخ به سوال[۵]، نمره‌دهی به متون[۶]، ماشین ترجمه[۷]، خلاصه‌سازی متن[۸] و غیره دارد. پیدا کردن شباهت میان کلمات یک بخش اساسی در یافتن شباهت متن دارد که در سطوح بعدی برای پیدا کردن شباهت میان جملات، پارگراف‌ها و اسناد اهمیت دارد. کلمات می‌توانند به دو صورت لغوی[۹] و معنایی[۱۰] به یکدیگر شبیه باشند. کلمات به صورت لغوی با یکدیگر شبیه هستند، در صورتی که دنباله کارکترهای مشابهی داشته باشند. کلمات به صورت معنایی شبیه هستند، اگر آنها یک مفهوم مشابه را داشته باشند، مثلا در یک بافت استفاده شوند و یا مثلا یکی متضاد دیگری باشد. تاکنون معیارهای مختلفی ارائه شده‌اند که به سه دسته کلی معیارهای مبتنی بر رشته[۱۱]، مبتنی برپیکره[۱۲] و مبتنی بر دانش تقسیم می‌شوند (Gomaa & Fahmy, 2013). شکل ۱-۱ این دسته‌بندی را نشان می‌دهد (Farouk, 2019).

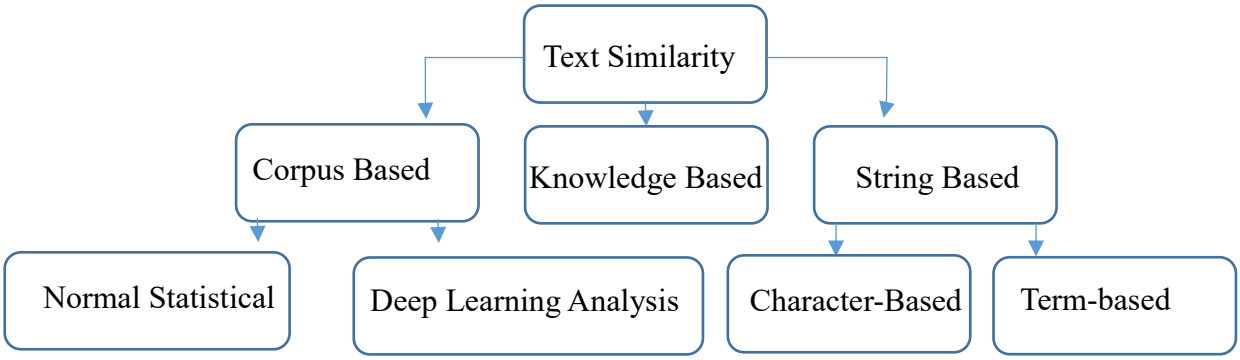
همانطور که این شکل نشان می‌دهد سه دسته اصلی برای بدست آوردن شباهت میان کلمات تاکنون ارائه شده‌اند. دسته اول مبتنی بر پیکره هستند. به عبارت دیگر، این گروه برای بدست آوردن شباهت میان کلمات از تحلیل کل پیکره استفاده می‌کنند و ایده آنها این است که کلماتی که در کل پیکره به تعداد زیاد در کنار هم دیده می‌شوند از نظر معنایی به یکدیگر شبیه هستند که از دو نوع روش برای تحلیل کل پیکره استفاده می‌کنند. یکی از آنها مبتنی بر روش‌های یادگیری عمیق است و دیگری مبتنی بر تحلیل‌های آماری. گروه دوم مبتنی بر دانش هستند که عمدتا از گراف ساخته شده توسط انسان برای بدست آوردن شباهت میان کلمات استفاده می‌کنند که در این شبکه عمدتا ارتباط معنایی و روابط میان کلمات مشخص شده است. دسته سوم مبتنی بر رشته هستند. بدین صورت که هر رشته را به صورت دنباله‌ای از کارکترها در نظر گرفته و از شباهت میان کارکترهای تشکیل-دهنده برای بدست آوردن شباهت میان کلمات بهره می‌برند. در ادامه هر کدام از این سه گروه اصلی معرفی می‌شوند (Farouk, 2019).

معیارهای مبتنی بر رشته به دو دسته معیارهای مبتنی بر کارکتر[۱۳] و مبتنی بر ترم[۱۴] تقسیم می‌شوند. یک معیار شباهت مبتنی بر کارکتر، معیاری است که شباهت و یا عدم شباهت (فاصله) میان دو رشته متن را اندازه‌گیری

---

۱ Document clustering
۲ Topic detection
۳ Topic tracking
۴ Question generation
۵ Question answering
۶ Essay scoring
۷ Machine translation
۸ Text summarization
۹ Lexically
۱۰ Semantically
۱۱ String-based measures
۱۲ Corpus-based similarity
۱۳ Character-based
۱۴ Term-based

می‌کند. الگوریتم طولانی‌ترین زیررشته مشترک[۱] و یا به اختصار LCS، شباهت میان دو رشته را بر اساس طول زنجیره پیوسته کارکترها که در هر دو رشته وجود دارد، در نظر می‌گیرد. الگوریتم لوشتین-دامرا[۲] فاصله میان دو رشته را با شمردن کمترین تعداد عملیاتی که نیاز است که یک رشته به دیگری تبدیل شود، تعریف می‌کند. لازم به ذکر است که این عملیات به صورت اضافه، حذف و جایگذاری کارکترهای رشته برای تبدیل یک رشته به دیگری است (Hall & Dowling, 1980; Peterson, 1980).



شکل ۲-۱: روش‌های بدست آوردن شباهت متنی، (Farouk, 2019)

فاصله جرو[۳] بر اساس تعداد و ترتیب کارکترهای مشترک میان دو رشته است که مشتقات املایی[۴] متداول را در نظر می‌گیرد و اساسا در حوزه پیوند رکورد[۵] کاربرد دارد (Jaro, 1995). معیار جرو-وینکلر[۶] برگرفته از معیار فاصله جرو است که از مقیاس پیشوند برای بدست آوردن شباهت میان دو رشته استفاده می‌کند (Winkler, 1990). الگوریتم نیدلمن-وونش[۷] یک نمونه از برنامه‌نویسی پویا است که می‌توان گفت اولین کاربرد برنامه‌نویسی پویا برای مقایسه دنباله‌های بیولوژیکی می‌باشد. این الگوریتم بدین صورت عمل می‌کند که یک همترازی کلی در دو رشته انجام می‌دهد تا بهترین همترازی را میان دو رشته داده شده پیدا کند. این روش زمانی مناسب است که دو رشته از نظر طول یکسان هستند و درجه‌ای از شباهت میان آنها وجود دارد (Needleman & Wunsch, 1970).

نمونه دیگری از برنامه‌نویسی پویا، الگوریتم اسمیت-واترمن است که در واقع یک همترازی محلی را روی دو رشته انجام می‌دهد. این الگوریتم برای دنباله‌های غیرمشابه که مشکوک به شباهت زیردنباله‌هایشان هستند، مناسب می‌باشد (Smith & Waterman, 1981). ان-گرم[۸] یک یک زیردنباله از n مورد از یک دنباله داده شده از متن

---

۱ Longest Common Substring
۲ Damerau-Levenshtein
۳ Jaro
۴ Spelling deviations
۵ Record linkage
۶ Jaro-Winkler
۷ Needlelan-Wunsch
۸ N-gram

است. الگوریتم‌های شباهت ان-گرم در واقع ان-گرم‌ها را از هر کارکتر یا کلمه در دو رشته مقایسه می‌کند. فاصله میان دو رشته داده شده با تقسیم تعداد ان-گرم‌های مشابه به ماکزیمم تعداد ان-گرم‌ها محاسبه می‌گردد (Barrón-Cedeno et al., 2010).

معیارهای شباهت مبتنی بر ترم، شباهت میان دو رشته را بر اساس ترم‌های تشکیل‌دهنده آن بدست می‌آورد. فاصله بلاکی[۱] تحت عنوان فاصله منهتن[۲]، باکسکار[۳]، فاصله ارزش مطلق[۴]، فاصله L1[۵]، فاصله بلاک شهر[۶] هم شناخته می‌شود. این فاصله بدین صورت محاسبه می‌شود که مسیر میان دو ترم را به صورت شبکه‌ای در نظر گرفته و تعداد نقاطی که برای رسیدن به مقصد مورد نیاز هست، می‌شمارد (Reynolds, 1980). شباهت کسینوسی[۷] برای بدست آوردن شباهت میان دو رشته که به صورت یک بردار ذخیره شده‌اند از کسینوس زاویه میان آنها استفاده می‌کند. ضریب تاس[۸] به صورت دوبرابر تعداد ترم‌های مشترک در دو رشته تقسیم بر تعداد کل ترم‌ها در هر دو رشته محاسبه می‌شود (Dice, 1945). معیار جاکارد[۹] به صورت تعداد ترم‌های مشترک به کل تعداد ترم‌ها در هر دو رشته محاسبه می‌گردد (Niwattanakul et al., 2013). فاصله اقلیدسی[۱۰] یا فاصله L2 ریشه مربع مجموع فاصله میان عناصر دو بردار تعریف می‌شود. از این معیار در پژوهش‌های دیگر از جمله (عباسی و وزیری، ۱۳۹۴) و (سلیمانی‌نژاد و همکاران، ۱۳۹۷) برای خوشه‌بندی متون استفاده شده است. ضریب تطابق[۱۱] یک روش مبتنی بر بردار بسیار ساده است که به صورت ساده تعداد ترم‌های مشابه غیرصفر روی هر دو بردار را می‌شمارد. ضریب همپوشانی[۱۲] مشابه ضریب تاس است اما اگر یکی از رشته‌ها زیرمجموعه دیگری باشد، دو رشته را به صورت تطابق کامل در نظر می‌گیرد. از معیارهای شباهت کسینوسی، جاکارد و اقلیدسی برای محاسبه امتیاز شباهت اخبار در زبان‌های انگلیسی و هندی در (Singh et al., 2021) استفاده شده است.

معیار شباهت مبتنی بر پیکره، یک معیار شباهت معنایی[۱۳] است که شباهت میان کلمات را بر اساس اطلاعات بدست آمده از یک پیکره بزرگ می‌یابد. یک پیکره، یک مجموعه بزرگ از متون نوشته شده یا صحبت شده است که برای پژوهش‌های زبانی استفاده می‌شود.

---

۱ Block distance
۲ Manhattan distance
۳ Boxcar distance
۴ Absolute value distance
۵ L1 distance
۶ City block distance
۷ Cosine similarity
۸ Dice’s coefficient
۹ Jaccard similarity
۱۰ Euclidean distance
۱۱ Matching coefficient
۱۲ Overlap coefficient
۱۳ Semantic similarity measures

الگوریتم HAL[1] یک فضای معنایی از هم‌وقوعی کلمات را تولید می‌کند. یک ماتریس کلمه در کلمه ایجاد می‌شود که هر عنصر این ماتریس، قدرت مشارکت میان کلماتی است که در هر ردیف و ستون نشان داده می‌شود. سپس کاربر الگوریتم این انتخاب را دارد که ستون‌هایی با آنتروپی[2] کم را از ماتریس حذف کند. در حالیکه متن مورد تحلیل و آنالیز قرار می‌گیرد، یک کلمه کانونی در آغاز یک پنجره با ۱۰ کلمه قرار می‌گیرد که هر کدام از این کلمات، مشخص می‌کند که کدامیک از همسایه‌های کلمه به عنوان هم‌وقوعی شمرده می‌شوند. مقادیر ماتریس به وسیله وزن معکوس هم‌وقوعی کلمات متناسب با فاصله از کلمه کانونی انباشه می‌شوند؛ کلمات همسایه نزدیکتر به کلمه کانونی از نظر معنایی به آن نزدیکتر هستند و بنابراین وزن بیشتری می‌گیرند. الگوریتم HAL همچنین اطلاعات مرتبط با ترتیب کلمات نیز رکورد می‌کند و این اطلاعات، بر اساس اینکه کلمات همسایه قبل یا بعد از کلمه کانونی قرار گرفته‌اند، محاسبه می‌شود (Lund & Burgess, 1996).

تحلیل معنایی نهفته[3] یا به اختصار LSA یک تکنیک بسیار محبوب از شباهت مبتنی بر پیکره است. LSA فرض می‌کند که کلماتی که از نظر معنایی به هم نزدیک هستند، در بخش‌های مشابهی از متن رخ می‌دهند. یک ماتریس که شامل کلمه به پارگراف است، ساخته می‌شود. در این ماتریس، ردیف‌ها کلمات منحصر به فرد موجود در متن هستند و ستون‌ها پارگراف‌ها را نشان می‌دهند. سپس یک تکنیک ریاضی تحت عنوان تجزیه مقدار منفرد۴ و یا به اختصار SVD برای کاهش ابعاد استفاده می‌شود. کلمات سپس به وسیله‌ی کسینوس زاویه میان دو بردار که هر ردیف را نشان می‌دهد، مقایسه می‌شوند (Landauer & Dumais, 1997).

آنالیز معنایی نهفته تعمیم‌داده شده[5] و یا به اختصار GLSA یک چارچوب برای محاسبه معنایی است که از بردارهای ترم و بردار انگیزه می‌گیرد. می‌توان گفت که GLSA یک تعمیمی از LSA می‌باشد که تمرکزش روی بردارهای ترم است و همچنین نیاز به معیار ارتباط معنایی میان ترم‌ها و یک روش کاهش ابعاد دارد. روش GLSA می‌تواند با هر معیار شباهت روی فضای ترم‌ها با روش مناسب کاهش ابعاد ترکیب شود (Matveeva et al., 2005).

تحلیل معنایی صریح[6] و یا به اختصار ESA معیاری است که ارتباط معنایی میان دو متن دلخواه را محاسبه می‌کند. تکنیک مبتنی بر ویکی‌پدیا ترم‌ها یا متون را به عنوان بردارها با ابعاد بالا نمایش می‌دهد؛ هر ورودی بردار وزن TF-IDF میان ترم و یک مقاله ویکی‌پدیا است. ارتباط معنایی میان دو ترم (متون) به وسیله معیار کسینوسی میان بردارهای متناظر بیان می‌شود (Gabrilovich & Markovitch, 2007).

---

[1] Hyperspace Analogue to Language

[2] Entropy

[3] Latent Semantic Analysis

[4] Singular value decomposition

[5] Generalized Latent Semantic Analysis

[6] Explicit Semantic Analysis

آنالیز معنایی صریح متقاطع[۱] و یا به اختصار CL-ESA یک تعمیم چندزبانه از ESA است. CL-ESA از یک مجموعه مرجع چندزبانه مثل ویکی‌پدیا بهره می‌بردتا یک سند را به عنوان یک بردار مفهومی مستقل از زبان نشان دهد. ارتباط دو سند در زبان‌های مختلف به وسیله‌ی معیار شباهت کسینوسی میان نمایش بردارهای مرتبط ارزیابی می‌گردد (Potthast et al., 2008).

بازیابی اطلاعات-اطلاعات مشترک نقطه‌ای[۲] و یا به اختصار PMI-IR روشی برای محاسبه شباهت میان جفت‌های کلمات است که در جستجوی پیشرفته آلتاویستا[۳] برای محاسبه احتمالات استفاده می‌شود. اغلب اوقات دو کلمه نزدیک به هم در یک صفحه وب[۴] امتیاز شباهت PMI-IR بالاتری نیز دارند (Turney, 2001). اطلاعات مشترک نقطه‌ای هم‌وقوع مرتبه دوم[۵] یا به اختصار SCO-PMI یک معیار شباهت معنایی با استفاده از اطلاعات مشترک نقطه‌ای است که لیستی از کلمات همسایه دو کلمه را از پیکره بزرگ بر اساس اهمیت آنها مرتب می‌نماید (Islam & Inkpen, 2008). مزایای استفاده از این معیار این است که می‌تواند شباهت میان دو کلمه که زیاد در کنار یکدیگر رخ نداده‌اند، محاسبه کند؛ زیرا آنها با کلمات همسایه مشترکی زیاد دیده شده‌اند (Islam & Inkpen, 2006).

فاصله گوگل نرمال شده[۶] (NGD) یک معیار شباهت معنایی گرفته شده از تعداد هیت[۷]های گرفته شده به وسیله‌ی موتور جستجوی گوگل برای مجموعه‌ای از کلیدواژه‌های داده شده است. کلیدواژه‌ها با معنی مشابه یا یکسان در مفهوم زبان طبیعی متمایل به نزدیک در واحد فاصله گوگل هستند؛ در حالیکه کلمات با معانی نامشابه گرایش به فاصله‌های دورتر با استفاده از این معیار فاصله می‌باشند (Cilibrasi & Vitanyi, 2007).

استخراج توزیعی کلمات مشابه با استفاده از هم‌وقوعی[۸] و یا به اختصار DISCO شباهت توزیعی میان کلمات است که فرض می‌کند که کلمات با معنی مشابه در بافت‌های مشابه هم در کنار یکدیگر دیده می‌شوند. مجموعه‌های متنی بزرگتر به صورت آماری تحلیل می‌شوند تا شباهت توزیعی را بدست آورند. این معیار روشی است که شباهت توزیعی میان کلمات با استفاده از پنجره بافت ساده با سایز سه کلمه برای شمارش هم‌وقوعی را محاسبه می‌نماید. زمانی که دو کلمه در معرض شباهت دقیق با توجه به DISCO قرار گرفتند، بردار کلماتشان از داده‌های نمایه شده بازیابی شده و سپس شباهت میان آنها بر اساس معیار لین[۹] محاسبه می‌گردد (Lin, 1998). اگر به کلمات مشابه بیشتری نیاز پیدا شد، DISCO بردار کلمه مرتبه دوم را برای کلمه داده شده، برمی‌گرداند. لازم به ذکر است که DISCO دو معیار شباهت اصلی DISCO1 و DISCO2 دارد که اولی شباهت مرتبه اول میان دو کلمه

---

[۱] Cross-Language Explicit Semantic Analysis

[۲] Pointwise Mutual Information-Information Retrieval

[۳] AltaVista

[۴] Web page

[۵] Second-order co-occurrence pointwise mutual information

[۶] Normalized Google Distance

[۷] hit

[۸] Extracting DIStributionally similar words using Co-occurrences

[۹] Lin measure

داده شده را بر اساس مجموعه‌های با همگذاری محاسبه کرده و DISCO2 شباهت مرتبه دوم میان دو کلمه ورودی را بر اساس کلمات مشابه به صورت توزیعی محاسبه می‌نماید (Kolb, 2009).

شباهت مبتنی بر دانش[۱] یک معیار شباهت معنایی است که درجه شباهت میان کلمات را با استفاده از اطلاعات مستخرج از یک شبکه معنایی، مشخص می‌کند (Gomaa & Fahmy, 2013). به عبارت دیگر، یکی از معیارهای شباهت معنایی است که بر اساس تعیین درجه شباهت میان کلمات با استفاده از اطلاعات مشتق‌شده از شبکه معنایی عمل می‌کند (Mihalcea et al., 2006). وردنت (Kumar et al., 2018) معروف‌ترین شبکه معنایی در حوزه اندازه‌گیری شباهت مبتنی بر دانش میان کلمات است که پایگاه داده بزرگ لغوی غنی در حوزه زبان انگلیسی را جمع‌آوری کرده است. اسامی، افعال، صفات، قیود درون مجموعه‌های از مترادف شناختی[۲] قرار می‌گیرند که هر کدام یک مفهوم ویژه را نشان می‌دهند. این گروه‌ها با کمک مفاهیم معنایی و روابط لغوی به یکدیگر ارتباط پیدا می‌کنند.

معیارهای شباهت مبتنی بر دانش می‌توانند به دو دسته معیارهای شباهت معنایی[۳] و معیارهای ارتباط معنایی[۴] تقسیم‌بندی شوند. مفاهیم مشابه معنایی بر اساس تشابهشان تشخیص داده می‌گردند. ارتباط معنایی از طرف دیگر یک مفهوم کلی است که فقط به شکل یا فرم مفهوم گره نمی‌خورد. به عبارت دیگر، شباهت معنایی نوعی از ارتباط میان دو کلمه است؛ در حالیکه ارتباط معنایی محدوده وسیعی از ارتباطات ار جمله "نوعی-از"[۵]، "یک-مثال-خاص-از"[۶]، "بخشی-از"[۷]، "مخالف-از"[۸] دربرمی‌گیرد (Patwardhan et al., 2003). عسگریان و همکاران (۱۳۸۶) از هسته آنتولویی وردنت به عنوان دانش پس‌زمینه برای خوشه‌بندی متون استفاده کردند.

شش معیار شباهت معنایی وجود دارد؛ سه مورد از آنها بر اساس محتوای اطلاعات هستند: رسنیک (res) (Bin et al., 2012)، لین (lin) (Lin, 1998) و ژیانگ و کنراث (jcn) (Jiang & Conrath, 1997). سه معیار دیگر بر اساس طول مسیر تعریف می‌شوند: لیکاک و چودورو (lch) (Leacock & Chodorow, 1998)، وو و پالمر (wup) (Wu & Palmer, 1994) و طول مسیر (path).

مقدار مرتبط در معیار res مساوی با محتوای اطلاعات (IC) کمترین زیرمجموعه (زیرمجموعه با اطلاعات بیشتر) است. محدوده این معیار بزرگتر یا مساوی با صفر می‌باشد. کران بالای این مقدار، بستگی به سایز پیکره استفاده شده برای بدست آوردن محتوای اطلاعات دارد. معیارهای lin و jcn هر دو محتوای اطلاعاتی مفاهیم را در نظر

---

[۱] Knowledge-based similarity

[۲] Cognitive synonyms

[۳] Measures of semantic similarity

[۴] Measures of semantic relatedness

[۵] is-a-kind-of

[۶] is-a-specific-example-of

[۷] is-a-part-of

[۸] is-the-opposite-of

می‌گیرند. معیار lin محتوای اطلاعاتی کمترین زیرمجموعه اطلاعاتی را با جمع‌کردن آنها اندازه‌گیری می‌کند؛ در حالیکه jcn اختلاف این مجموع را با محتوای اطلاعاتی کمترین زیرمجموعه مشترک در نظر می‌گیرد.

اسلامی‌نسب و جاویدان (۱۳۹۴)، به منظور بدست آوردن شباهت معنایی دو مقاله انگلیسی، ابتدا آنها را در سه بخش عنوان، کلمات کلیدی و چکیده تفکیک کرده و به هر قسمت یک وزن دادند. سپس شباهت هر کدام از قسمت‌ها به صورت دو به دو مقایسه شده و سپس نتاج نهایی بر اساس میانگین وزنی از قسمت‌های مختلف بدست آوردند. در پژوهشی دیگر از یک آنتولوژی برای بدست آوردن شباهت معنایی میان مقالات علمی بهره گرفته شده است که تمرکز آن روی شباهت معنایی در سطح سند است ((Liu et al., 2017 (a). در پژوهشی دیگر، یک پروفایل از مقاله علمی با توجه به قسمت‌های مختلف آن ساخته شده و شباهت پروفایل‌ها با تکنیک تعبیه کلمات بدست می‌آید ((Liue et al., 2017 (b).

بعضی از روش‌ها از تکنیک‌های یادگیری عمیق برای بدست آوردن شباهت متنی استفاده می‌کنند (Kenter & De Rijke, 2015). بدین منظور، یک پیکره بسیار بزرگ متنی برای آموزش مدل استفاده شده تا در نهایت مدل بتواند یک نمایش دیگری از کلمات در فضای برداری جدید را ارائه دهد. این مدل ساخته شده از کلمات به هم‌وقوعی کلمات در کل پیکره وابسته است. بدین صورت که از کل پیکره برای آموزش مدل استفاده می‌کند تا احتمال وقوع یک کلمه با توجه به کلمات دیگر را پیش‌بینی نماید (Church, 2017).

در سال‌های اخیر، معیارهای شباهتی ارائه شده‌اند که از ترکیب چند معیار استفاده می‌کنند که به آنها معیارهای شباهت ترکیبی[۱] می‌گویند. یکی از این معیارهای شباهت، ترکیبی از معیارهای مبتنی بر پیکره و مبتنی بر دانش است (Mihalcea et al., 2006). دو مورد از این معیارها معیارهای مبتنی بر پیکره هستند و شش مورد دیگر مبتنی بر دانش. ابتدا این هشت الگوریتم به صورت جداگانه مورد ارزیابی قرار گرفتند، سپس با یکدیگر ترکیب شدند. بهترین کارایی با استفاده از روشی که چندین معیارهای شباهت را درون یکی ترکیب می‌کند، بدست آمد.

روشی دیگر برای اندازه‌گیری شباهت معنایی میان جملات یا متون خیلی کوتاه بر اساس معنا و اطلاعات مبتنی بر ترتیب کلمه[۲] در (Li et al., 2006) مورد بررسی قرار گرفته است. ابتدا شباهت معنایی از یک پایگاه دانش لغوی[۳] و پیکره استخراج گردید. سپس، روش پیشنهادی تاثیر ترتیب کلمات روی معنی جمله را بررسی می‌کند. معیارهای شباهت مستخرج از ترتیب کلمه، تعداد کلمات مختلف را به همراه تعداد جفت کلمات در ترتیب‌های مختلف اندازه‌گیری می‌کنند.

بوسکالدی[۴] و همکاران (۲۰۱۲)، روشی را ارائه کرده و نام آن را شباهت معنایی متن[۵] و یا به اختصار STS نامیدند. این روش شباهت دو متن را از ترکیب میان معنا و نحو اطلاعات تعیین می‌کند. آنها دو تابع لازم (شباهت رشته[۶]

---

۱ Hybrid Similarity Measures

۲ Word order information

۳ Lexical knowledge base

۴ Buscaldi

۵ Semantic Text Similarity

۶ String similarity

و شباهت معنایی کلمه[۱]) و یک تابع اختیاری (شباهت ترتیب کلمه مشترک[۲]) را در نظر گرفتند. روش STS یک ضریب همبستگی پیرسون بسیار خوب را نتیجه می‌دهد و نتایج (Li et al., 2006) را بهبود داده است.
آگاروال[۳] و همکاران (۲۰۱۲) روشی را ارائه دادند که معیار ارتباط معنایی مبتنی بر پیکره[۴] را روی کل جمله به همراه امتیازهای شباهت معنایی مبتنی بر دانش را که برای کلماتی که تحت همان نقش‌های نحوی در هر دو جمله قرار می‌گیرند، ترکیب می‌کند. سپس همه این امتیازها به عنوان ویژگی[۵]هایی به مدل‌های یادگیری ماشین[۶] مانند رگرسیون خطی[۷] و مدل‌های بگینگ[۸] داده شده تا یک امتیاز که نشان‌دهنده شباهت میان جملات است، بدست آید. این روش نشان داده که ترکیب معیارهای شباهت مبتنی بر دانش و معیارهای شباهت مبتنی بر پیکره نسبت به معیارهای تنها مبتنی بر پیکره، یک بهبود قابل توجه را در محاسبه شباهت معنایی میان جملات، در پی داشته است.
یک همبستگی امیدوارکننده‌ای میان نتایج شباهت خودکار و دستی بوسیله ترکیب دو ماژول در (Buscaldi et al., 2012) ارائه گردید. اولین ماژول شباهت میان جملات را به وسیله‌ی شباهت مبتنی بر ان-گرم محاسبه می‌کند و دومین ماژول، شباهت میان مفاهیم در دو جمله را با استفاده از معیارهای شباهت مفهومی و وردنت حساب می‌کند.
یک سیستمی به نام UHP با نتایج همبستگی منطقی در (Bär et al., 2012) ارائه گردید که از یک مدل رگرسیون خطی ساده بر اساس داده آموزشی استفاده می‌کند تا چندین معیار شباهت را با یکدیگر ترکیب کند. این معیارها شباهت رشته‌ای، شباهت معنایی، مکانیزم‌های گسترش متن و معیارهای مرتبط به ساختار[۹] و سبک[۱۰] بودند.
پژوهش‌هایی نیز ارائه شدند که معیار شباهت را برای یک حوزه خاص ارائه می‌کنند. لیتل[۱۱] و همکاران (۲۰۲۰) معیار شباهتی را برای توییت‌های سیاسی[۱۲] که عمدتا حاوی متون کوتاهی هستند، ارائه دادند. پژوهشی دیگر (Qurashi et al., 2020) تکنیک‌های مختلف برای اندازه‌گیری شباهت معنایی در اسنادی که برای سامانه‌های ایمنی مانند داده‌های راه‌آهن است، بررسی کردند.

---

۱ Semantic word similarity
۲ Common-word order similarity
۳ Aggarwal
۴ Corpus-based semantic relatedness measure
۵ feature
۶ Machine learning
۷ Linear regression
۸ Bagging models
۹ Structure
۱۰ Style
۱۱ Little
۱۲ Political tweets

پژوهش‌های دیگری از جمله (Atoum, 2019) معیاری برای ارزیابی روش‌های شباهت متون ارائه دادند. آنها ادعا کردند که معیار همبستگی پیرسون که برای تعیین کارایی روش‌های مختلف شباهت متون استفاده می‌شود، به داده‌های دورافتاده وابسته است. بنابراین نمی‌تواند در موقعیتی که داده‌ها بسیار شبیه و یا بسیار غیرمرتبط باشند، عمل کنند. بنابراین با توسعه معیار همبستگی پیرسون به صورت مقیاس‌شده توانستند این مشکل را برطرف نمایند. لازم به ذکر است که همه پژوهش‌ها از معیار همبستگی پیرسون برای ارزیابی معیار شباهت ارائه شده، استفاده نمی‌کنند. بعضی از پژوهش‌ها مانند (Lakshmi & Baskar, 2021) دو معیار شباهت مبتنی بر فاصله تکرار ترم و وجود ترم‌های مشترک در کنار یکدیگر ارائه کرده و کارایی معیارهای ارائه شده را با خوشه‌یابی نشان دادند.

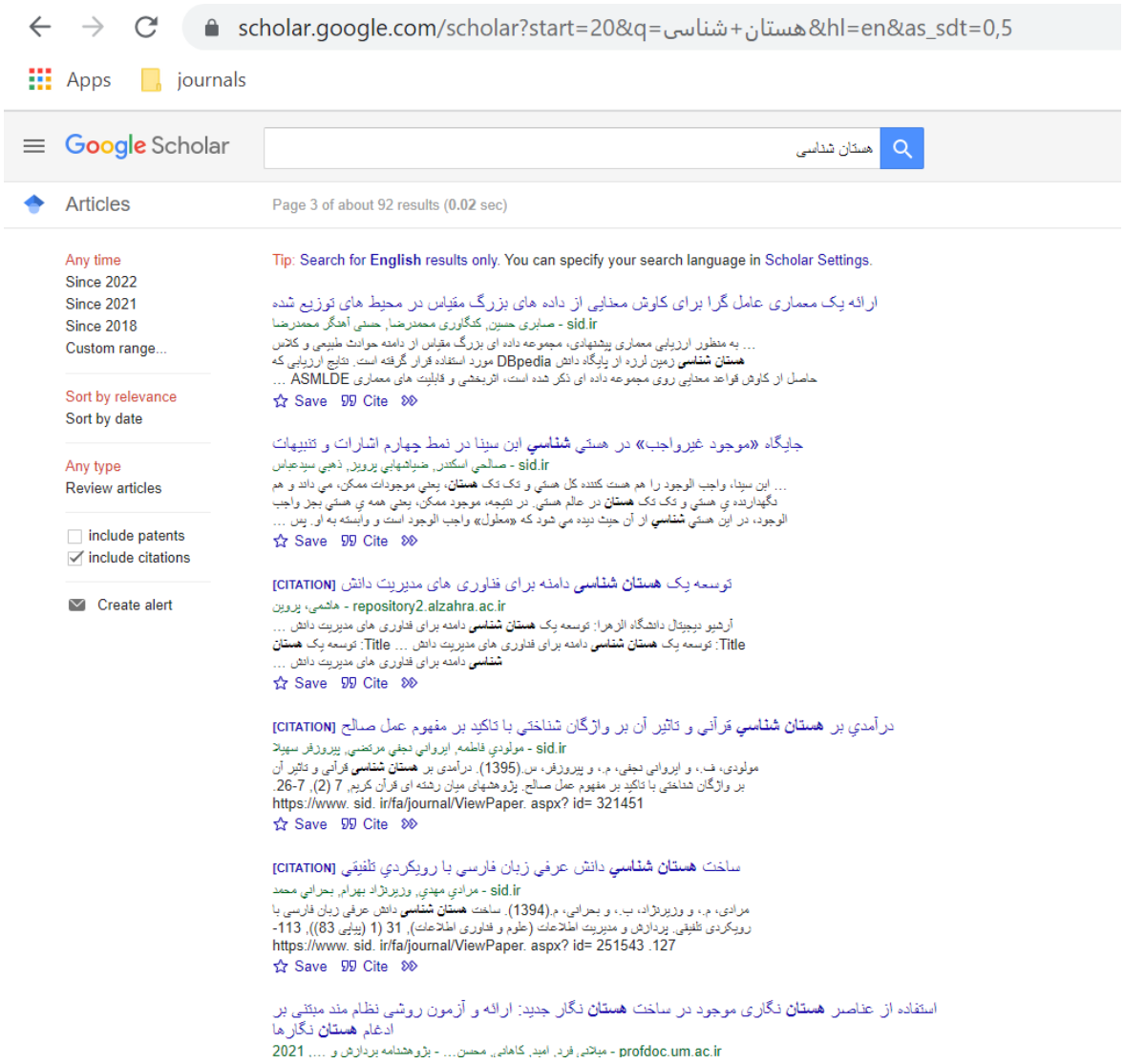
۴_۲ بررسی سامانه‌های مشابه

در این بخش، موتورهای جستجوی مختلفی از دو دیدگاه "وجود قابلیت یافتن مقالات مرتبط و همچنین "نحوه عملکرد" در صورت وجود همچین قابلیتی مورد بررسی قرار گرفتند. در ادامه، دو دسته موتورهای جستجو بررسی می‌شوند. دسته اول، موتور جستجوی گوگل به عنوان قوی‌ترین موتور جستجو در جهان و دسته دوم موتورهای جستجو مقالات فارسی در ایران است.

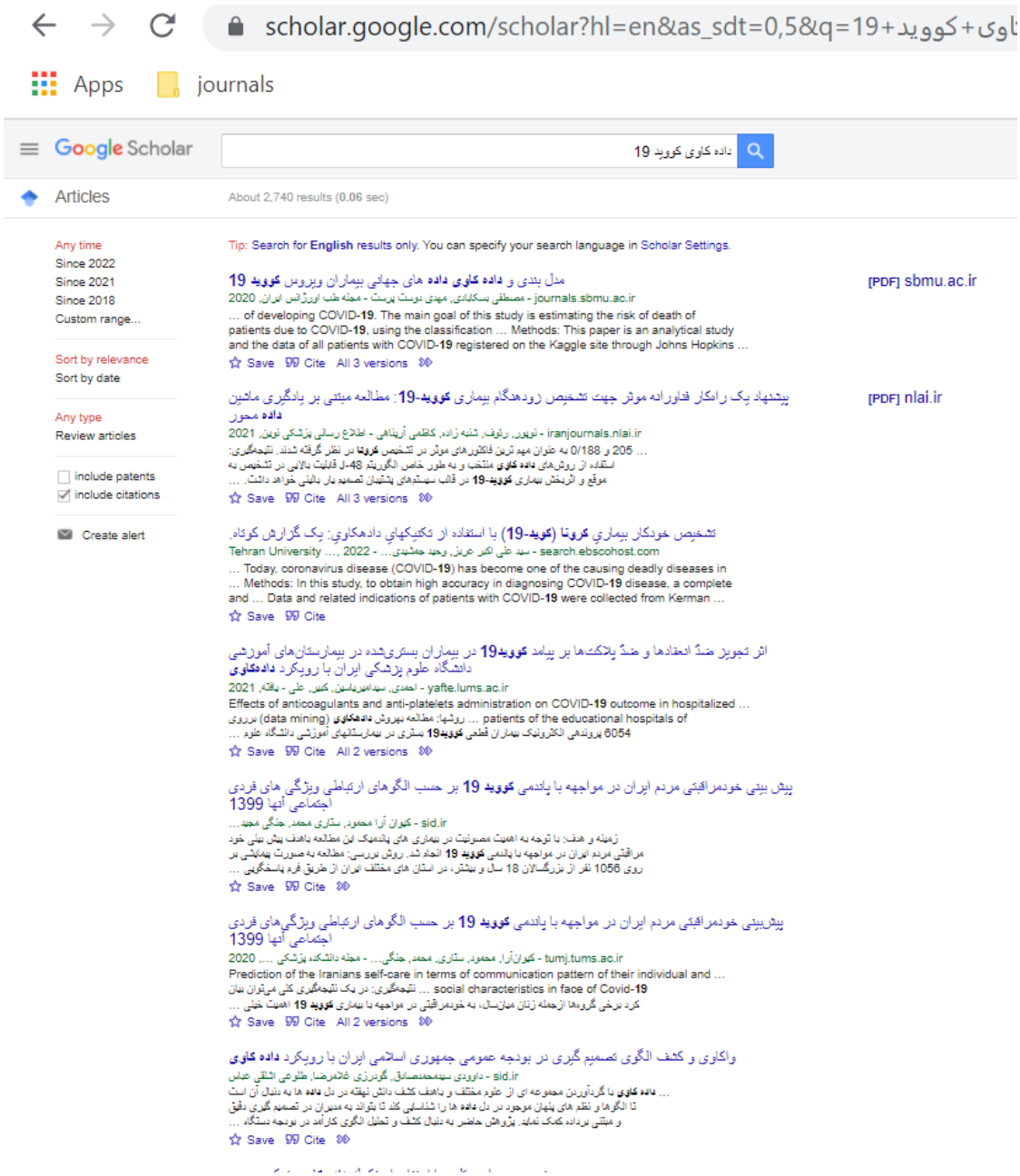
۴_۲_۱ موتور جستجوی گوگل[۱]

از ابتدای سال ۱۹۹۷ که موتور جستجوی گوگل معرفی گردید، بارها الگوریتم‌های آن مورد تغییرات گسترده و جزئی قرار گرفته و همچنین قابلیت‌های بسیار زیادی به آن اضافه شده است. یکی از قابلیت‌هایی که در موتور جستجوی گوگل برای مقالات، یافتن مقالات مرتبط و یا Related Article است. این قابلیت را برای جستجوی مقالات فارسی بررسی گردید. همانطور که شکل‌های ۲_۲ و ۲_۳ نشان می‌دهد، برای چندین کلیدواژه‌ای که مورد پرسش قرار گرفتند، این قابلیت اصلا وجود نداشت. لازم به ذکر است که این تعداد بسیار زیاد بود که فقط دو مورد به عنوان نمونه آورده شده است.

[۱] https://scholar.google.com/

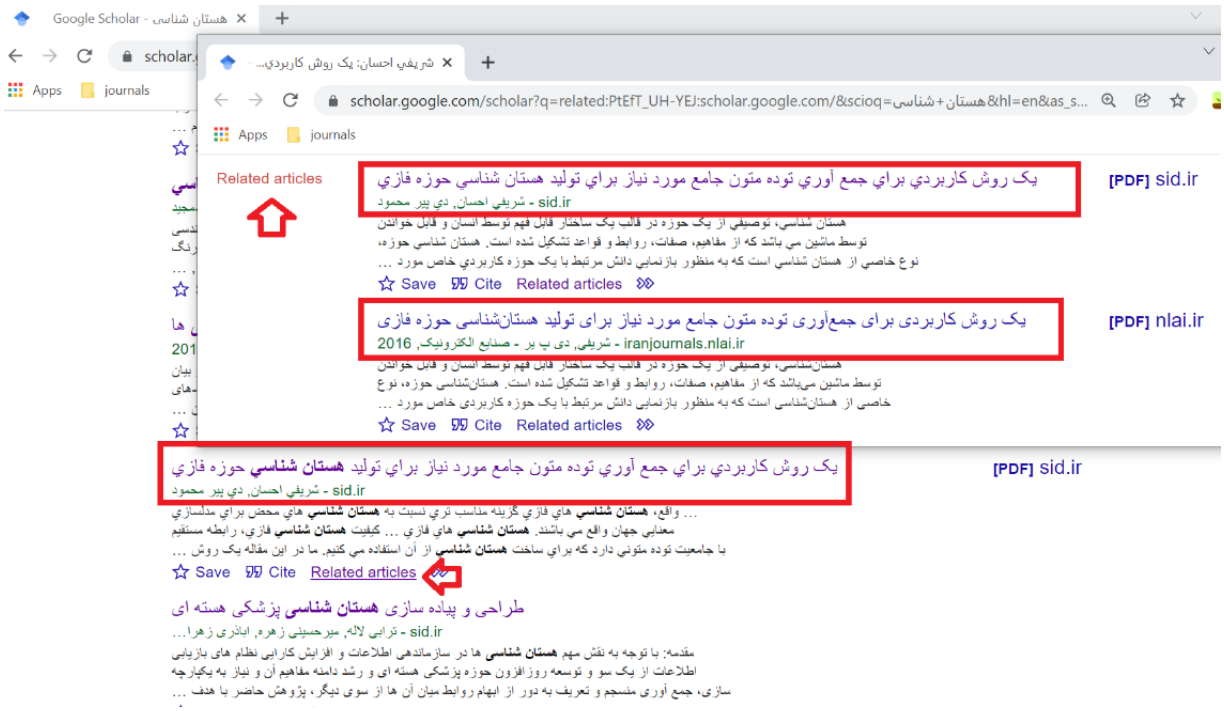


شکل ۲_۲: عدم وجود قابلیت یافتن مقالات مرتبط برای کلیدواژه "هستان‌شناسی"

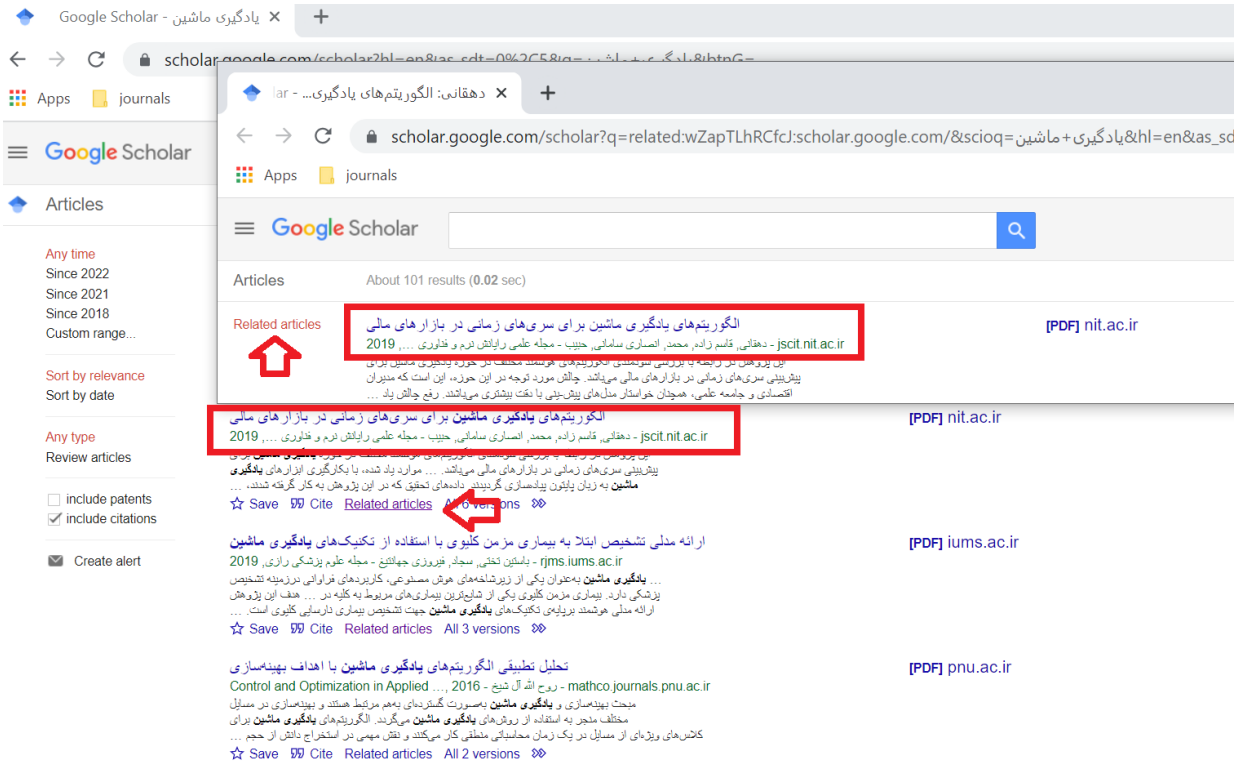


شکل ۲_۳: عدم وجود قابلیت یافتن مقالات مرتبط برای کلیدواژه "داده‌کاوی کووید۱۹"

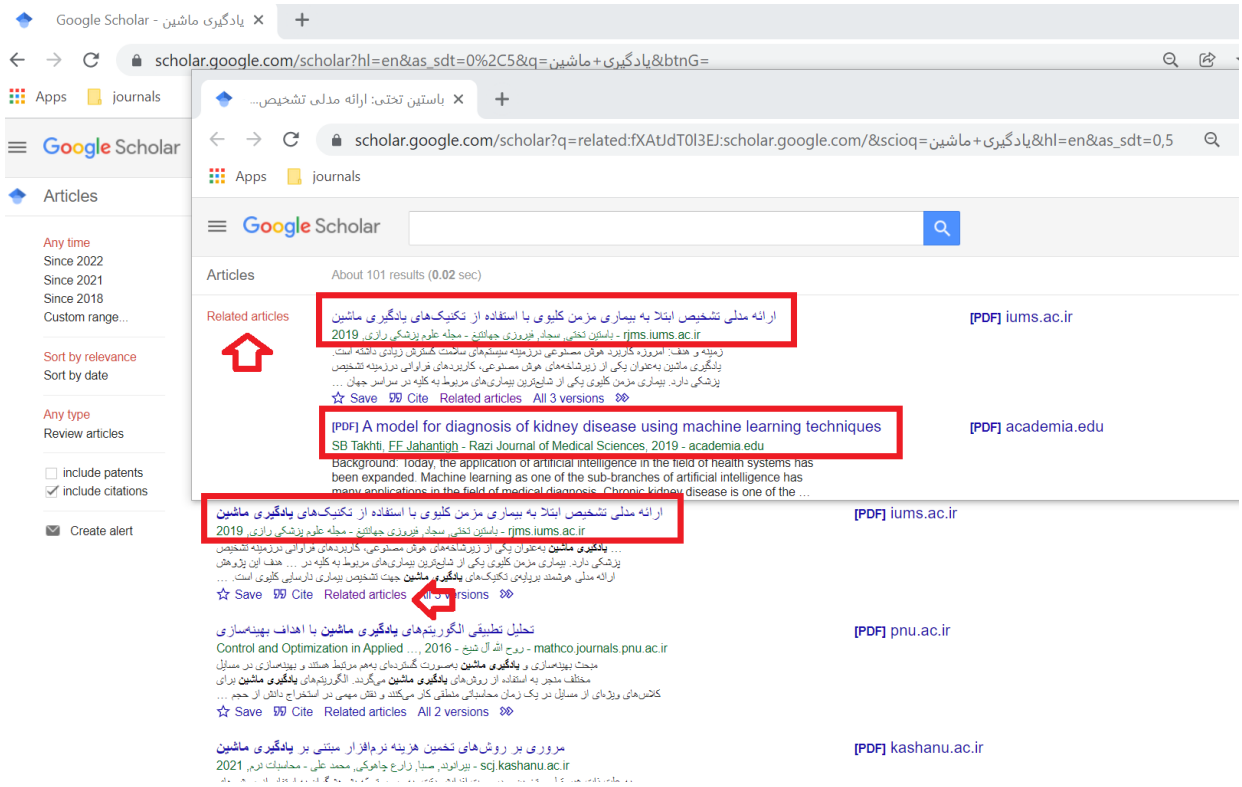
برای تعداد کمی از مقالات، این قابلیت related articles موجود بود ولی خود مقاله در نسخه‌های دیگر را نشان می‌داد. مانند موارد زیر:



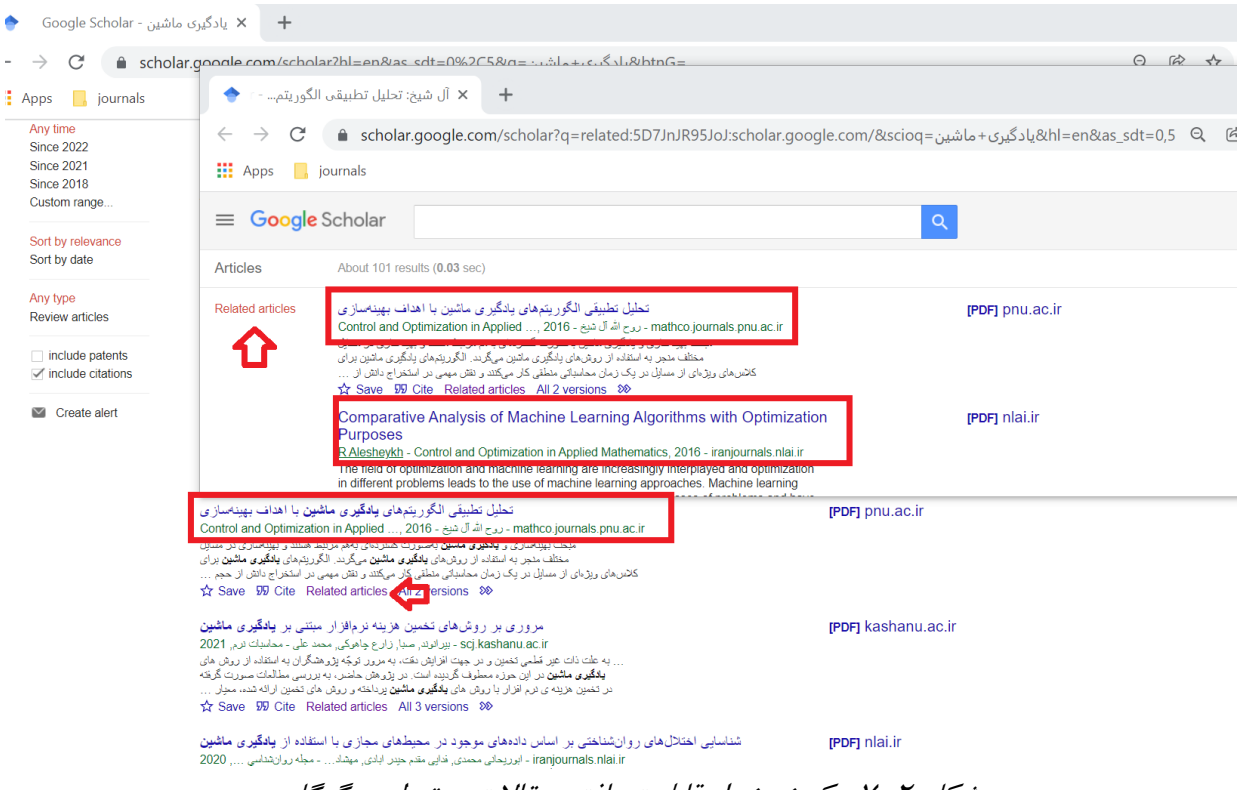
شکل ۲_۴: یک نمونه از قابلیت یافتن مقالات مرتبط در گوگل



شکل ۲_۵: یک نمونه از قابلیت یافتن مقالات مرتبط در گوگل

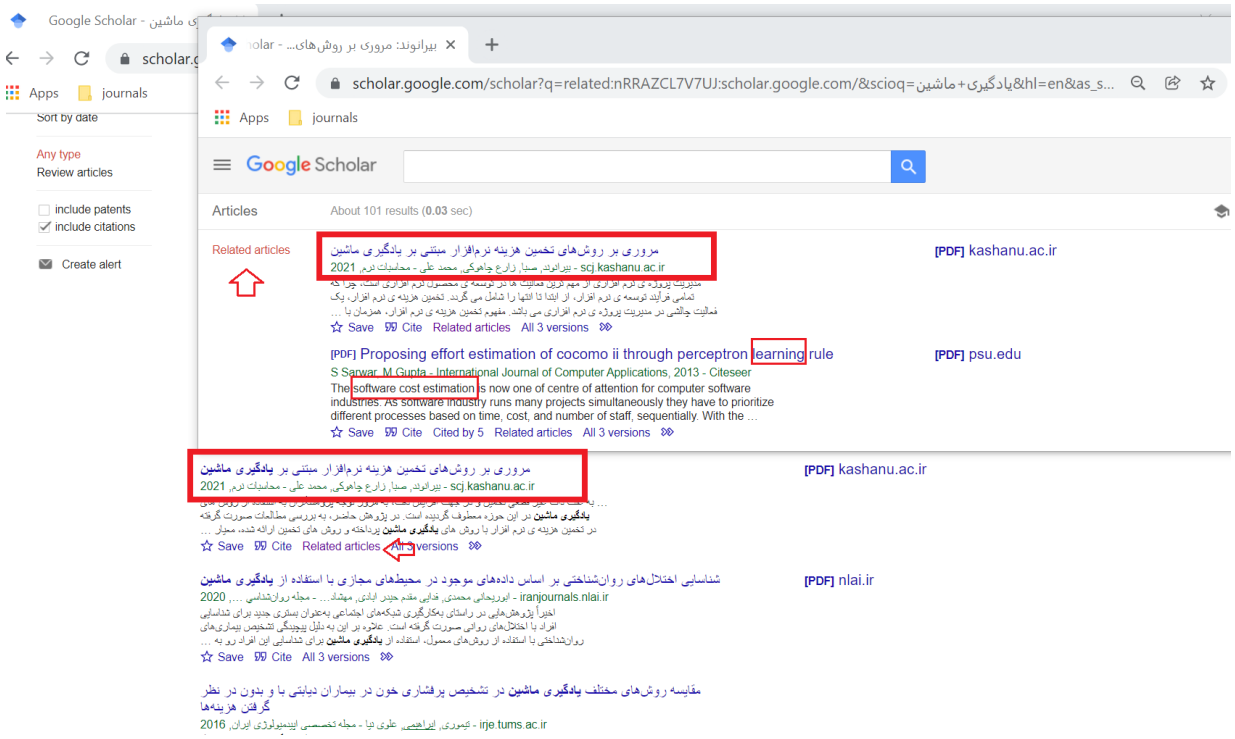


شکل ۲_۶: یک نمونه از قابلیت یافتن مقالات مرتبط در گوگل

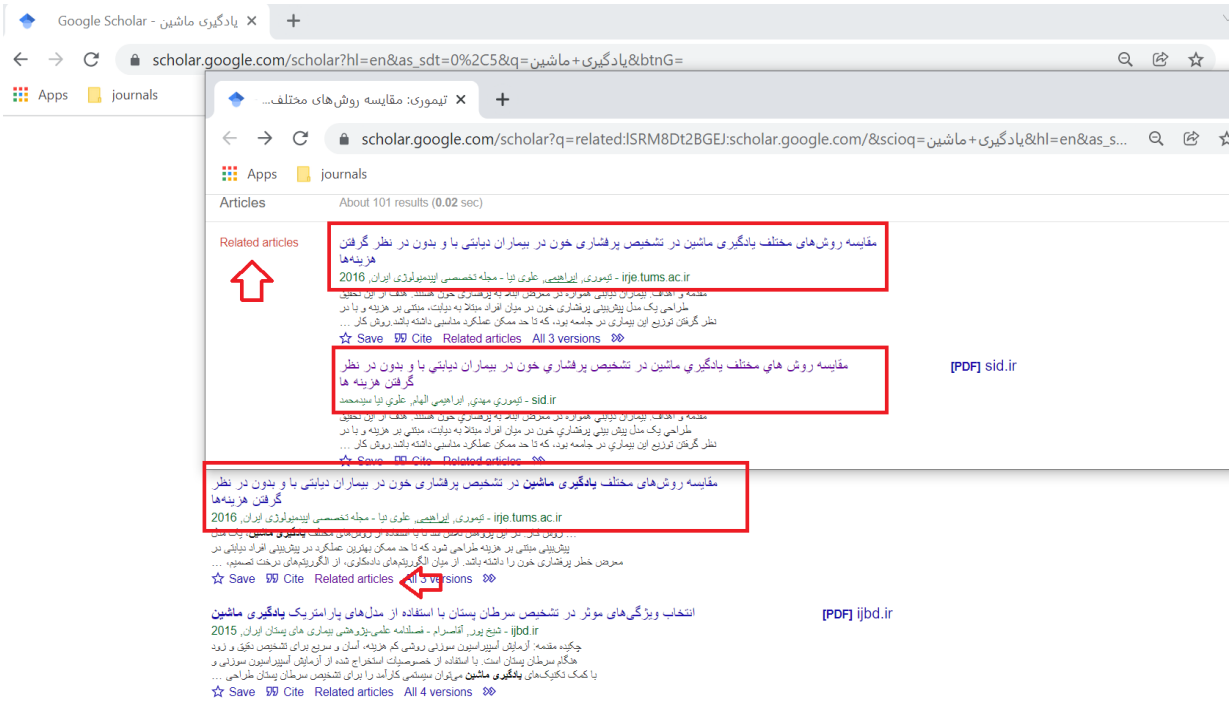


شکل ۲_۷: یک نمونه از قابلیت یافتن مقالات مرتبط در گوگل

یک سری موارد هم یافت شد که به نظر می‌رسد که فقط ترجمه کلمات کلیدی را در نظر گرفته و به دنبال مقالاتی می‌گردد که آن کلمه کلیدی در آنها تکرار شده باشد و آن مقالات را به عنوان related articles در نظر می‌گیرد. مانند شکل‌های ۲_۸ و ۲_۹.



شکل ۲_۸: یک نمونه از قابلیت یافتن مقالات مرتبط در گوگل



شکل ۲_۹: یک نمونه از قابلیت یافتن مقالات مرتبط در گوگل

نتیجه‌گیری از جستجوی گوگل:

۱) قابلیت پیدا کردن مقالات مرتبط یا همان Related Articles برای مقالات فارسی کم است و بسیاری از مقالات فارسی که توسط گوگل یافت می‌شوند، اصلا Related Articles برای آنها نشان داده نمی‌شود.

۲) آن دسته محدود از مقالات فارسی که برای آنها Related Articles موجود بود، در واقع نسخه‌های مختلف همان مقاله بود نه مقاله‌ای دیگر! در واقع همان مقاله بود که در پایگاه‌های مختلف نمایه شده بود و گوگل آنها را که در واقع نسخه‌های مختلف یک مقاله نمایه شده در پایگاه‌های مختلف را تحت عنوان Related Articles نشان داد.

۳) در صفحه Related Articles علاوه بر نسخه‌های دیگر مقاله در پایگاه‌های دیگر، به نظر می‌رسد که گوگل، کلمات کلیدی را از عنوان و کلیدواژه استخراج کرده و آنها را ترجمه می‌کند. سپس دنبال مقالات انگلیسی می‌گردد که در عنوان و کلیدواژه کلمات وجود داشته باشد.

۲_۲_۴ موتورهای جستجوی مقالات علمی فارسی

در این بخش، موتورهای جستجوی مقالات فارسی در ایران مورد بررسی قرار گرفتند و در مرحله اول، وجود قابلیت یافتن مقالات مرتبط در آنها واکاوی شد. جدول ۲_۱ لیست این موتورهای جستجوی علمی را به همراه آدرس و

وجود یا عدم وجود این قابلیت نشان می‌دهد. از میان موتورهای جستجوی مربوط به پایگاه‌های مگیران، نورمگز، پایگاه پژوهشکده اطلاعات و مدارک علمی ایران، پایگاه جهاد دانشگاهی، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پرتال جامع علوم انسانی و مرجع دانش، تنها موتورهای جستجوی پایگاه‌های نورمگز، جهاد دانشگاهی، مرجع دانش و پرتال جامع علوم انسانی این قابلیت را داشتند. در ادامه، این چهار موتور جستجو بیشتر مورد تحلیل قرار گرفته شد تا نحوه عملکرد آنها از منظر یافتن مقالات مرتبط بررسی گردد.

جدول ۲ _۱: پایگاه‌های داده فارسی برای جستجوی مقالات فارسی از دیدگاه داشتن ویژگی "یافتن مقالات مرتبط"

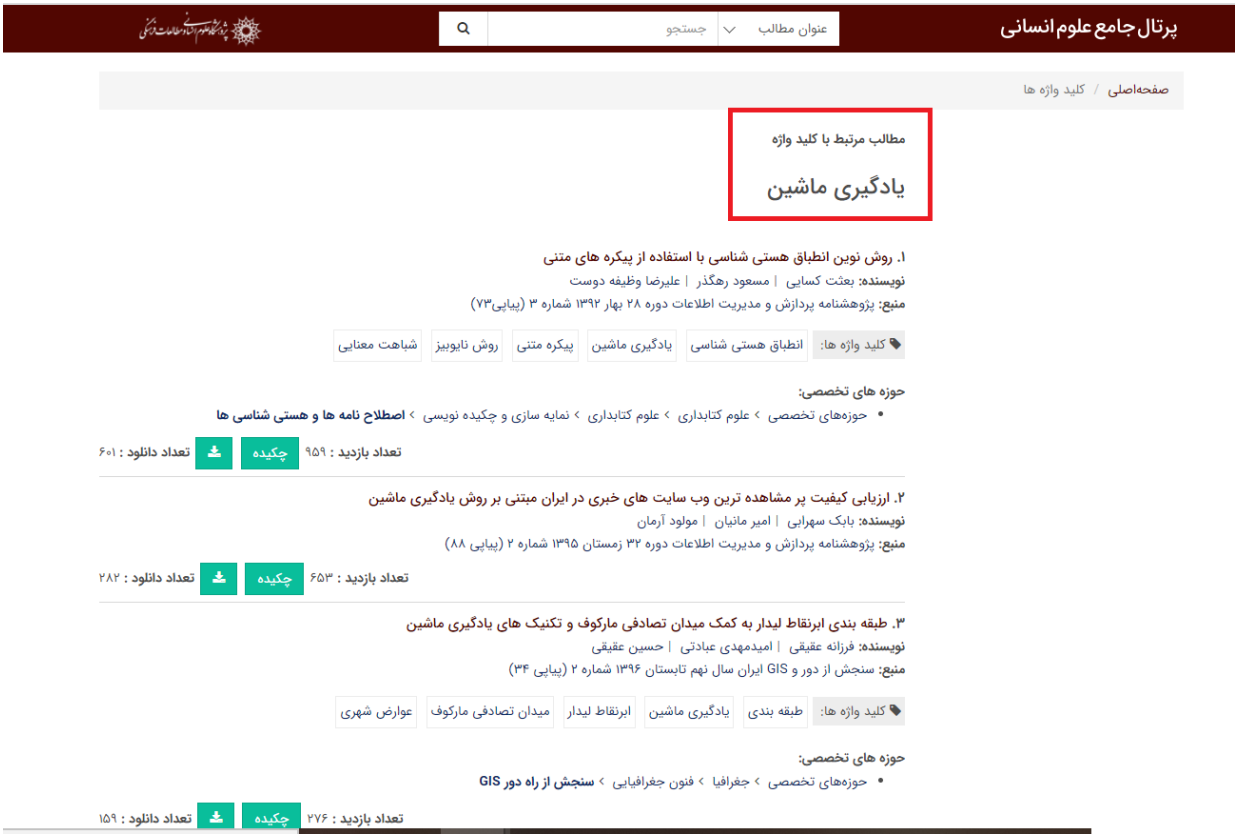
نام پایگاه	آدرس	قابلیت نمایش مقالات مرتبط
مگیران	https://www.magiran.com/	ندارد
نورمگز	www.noormags.ir	دارد
پژوهشکده اطلاعات و مدارک علمی ایران (گنج)	http://www.irandoc.ac.ir	ندارد
پایگاه بزرگ علمی - SIDجهاد دانشگاهی	www.sid.ir	دارد
مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری	www.ricest.ac.ir	ندارد
پرتال جامع علوم انسانی	www.ensani.ir	ندارد ولی می‌توان بر اساس کلمات کلیدی جستجو انجام داد.
مرجع دانش	https://civilica.com/	دارد

بررسی پرتال جامع علوم انسانی از منظر قابلیت یافتن مقالات مرتبط

پرتال جامع علوم انسانی که متعلق به پژوهشگاه علوم انسانی و مطالعات فرهنگی می‌باشد، قابلیت یافتن مقالات مرتبط را بدین صورت دارد که تنها می‌توان با استفاده از کلمات کلیدی جستجو انجام داد و مقالاتی که حاوی کلمات کلیدی انتخاب شده توسط کاربر هستند، نشان داده می‌شود.

شکل ۲_۱۰ نمونه‌ای از مقاله بازیابی شده توسط موتور جستجوی این پایگاه را نشان می‌دهد. در این مقاله، سه کلیدواژه "دشواری تکلیف"، "سالمندان" و "یادگیری ماشین" وجود دارد. با کلیک روی هر کدام از این کلیدواژه‌ها، مقالاتی که حاوی همان کلیدواژه هستند، نمایش داده می‌شود. به عنوان مثال، شکل ۲_۱۱ با کلیک روی کلیدواژه "یادگیری ماشین" بدست آمده است. همانطور که این شکل نشان می‌دهد، مقالاتی که بازیابی شده، ارتباطی با مقاله اولی ندارد و فقط کلمه یادگیری ماشین در آن وجود داد.

شکل ۲_۱۰: نمونه‌ای از عملکرد پرتال جامع علوم انسانی در یافتن مقالات مرتبط



شکل ۲_۱۱: مقالات بازیابی شده در پرتال جامع علوم انسانی با کلیک روی کلیدواژه "یادگیری ماشین"

بررسی پایگاه مرجع دانش از منظر قابلیت یافتن مقالات مرتبط

پایگاه مرجع دانش یا سیویلیکا یک پایگاه اینترنتی خصوصی است که به نمایه‌سازی و نشری مجموع مقالات همایش‌ها و کنفرانس‌های داخلی می‌پردازد. موتور جستجوی این پایگاه، قادر است مقالات مرتبط را نشان دهد؛ ولی میزان کارایی این سامانه در یافتن مقالات مرتبط باید مورد ارزیابی قرار گیرد. به عنوان مقال، نمونه‌ای از آن در شکل ۲_۱۲ در کادر قرمز مشخص شده است.

شکل ۲_۱۲: نمونه‌ای از عملکرد مرجع دانش در یافتن مقالات مرتبط

پایگاه جهاد دانشگاهی (SID) از منظر قابلیت یافتن مقالات مرتبط

موتور جستجوی پایگاه جهاد دانشگاهی، علاوه بر داشتن قابلیت یافتن مقالات مرتبط، قادر است تا میزان مرتبط بودن مقالات بازیابی شده را با رنگ‌بندی‌های مختلف نشان دهد. شکل ۲_۱۳ نمونه‌ای از عملکرد این پایگاه را در یافتن مقالات مرتبط با یک مقاله بازیابی شده نشان می‌دهد. البته میزان کارایی باید با دقت بیشتر بررسی گردد. این وب‌سایت ادعا کرده است که قادر است علاوه بر مقالات نشریات، مقالات همایشی مرتبط و همچنین مقالات بین‌الملل مرتبط نیز پیدا کند که البته میزان موفقیت این ادعا باید مورد بررسی قرار گیرد.


عنوان نشریه: بیماریهای پستان ایران

اطلاعات شماره: تابستان ۱۳۹۴ , دوره ۸ , شماره ۲ ; از صفحه ۱۵ تا صفحه ۲۳ .

عنوان مقاله: **انتخاب ویژگی های موثر در تشخیص سرطان پستان با استفاده از مدل های پارامتریک یادگیری ماشین**

نویسندگان: شیخ پور راضیه*, آقاصرام مهدی

آدرس: * یزد، خیابان چمران، خیابان رهبر، کوچه 17 رهبر، پلاک 91

چکیده: لطفا برای مشاهده چکیده به متن کامل (pdf) مراجعه فرمایید.

کلید واژه: سرطان پستان(Q2) | یادگیری ماشین(Q2) | انتخاب ویژگی(Q3) | روش های پارامتریک(Q2)

چارک 1 | چارک 2 | چارک 3 | چارک 4

موضوعات مرتبط: استخراج ویژگی | استوری | الگوی توانمندسازی خانواده محور | انتخاب مدل | انتخاب ویژگی | ترموگرافی پستان | تلفیق در سطح ویژگی | درخت تصمیم | درختان تصمیم گیری | ریشه دوم | سالمندان | سالمندی | سرطان سینه | سرطان | طیف نگاری رامان | کالوگیرو | کشف دانش | ماشین بردار پشتیبان | مدل آلماگرا | مدیریت رویگردانی مشتری

مقالات نشریه ای مرتبط:

بررسی و تجزیه و تحلیل مدل های بقا در سرطان پستان

انتخاب ویژگی مبتنی بر تئوری اطلاعات برای انتخاب ژن های مؤثر در تشخیص نوع سرطان با استفاده از داده های ریزآرایه

سرطان پستان مردان

ارزیابی ارزش در معرض ریسک شاخص سهام بر مبنای رویکردهای پارامتریک، شبه پارامتریک و غیرپارامتریک (مطالعه بورس اوراق بهادار تهران)

تشخیص سرطان پستان با استفاده از برآورد ناپارمتری چگالی احتمال مبتنی بر روش های هسته ای

مقالات همایشی مرتبط:

ندارد

مقالات بین المللی مرتبط:

.No item

ارتباط خیلی زیاد | ارتباط زیاد | مرتبط | ارتباط کمتر

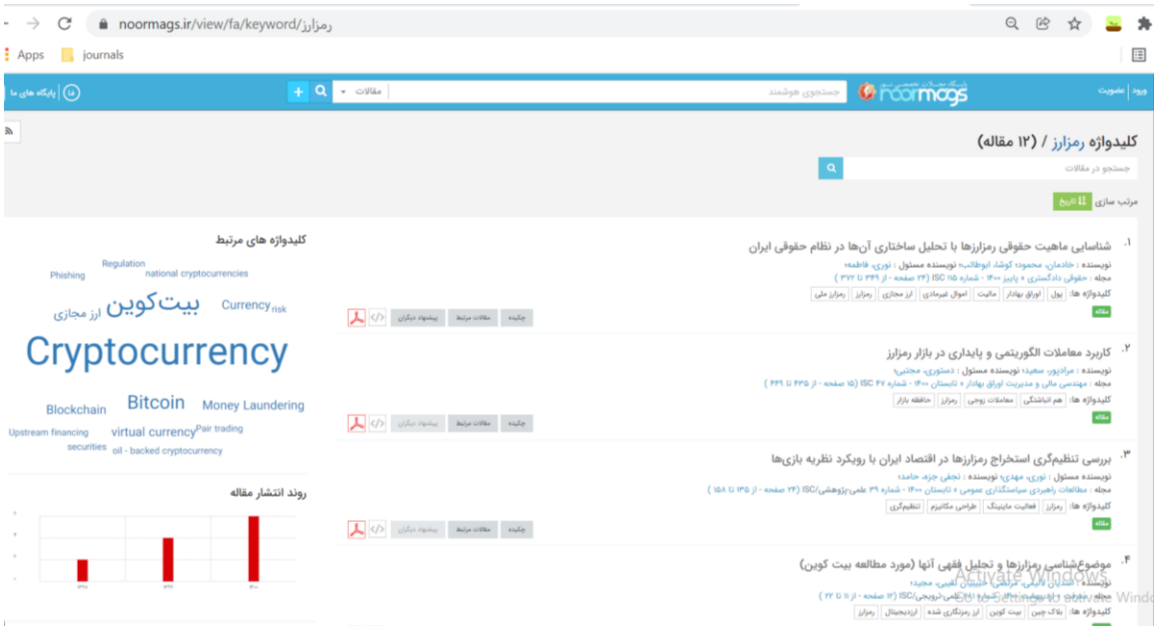
شکل ۲_۱۳: نمونه‌ای از عملکرد پایگاه جهاد دانشگاهی (SID) در یافتن مقالات مرتبط

بررسی پایگاه نورمگز از منظر قابلیت یافتن مقالات مرتبط

در این پایگاه مشابه با پرتال جامع علوم انسانی، می‌توان بر اساس کلیدواژه هم جستجو کرد. به عبارت دیگر روی هر کلیدواژه‌ای که کلیک کنیم، لیست مقالاتی که در کلیدواژه آنها، کلیدواژه مورد نظر وجود داشته باشد، نمایش داده می‌شود. به عنوان مثال، اگر یک مقاله حاوی کلیدواژه "رمزارز" باشد و روی این کلیدواژه کلیک کنیم، تمام مقالاتی که حاوی کلیدواژه "رمزارز" باشند، نمایش داده می‌شود (شکل ۲_۱۴).

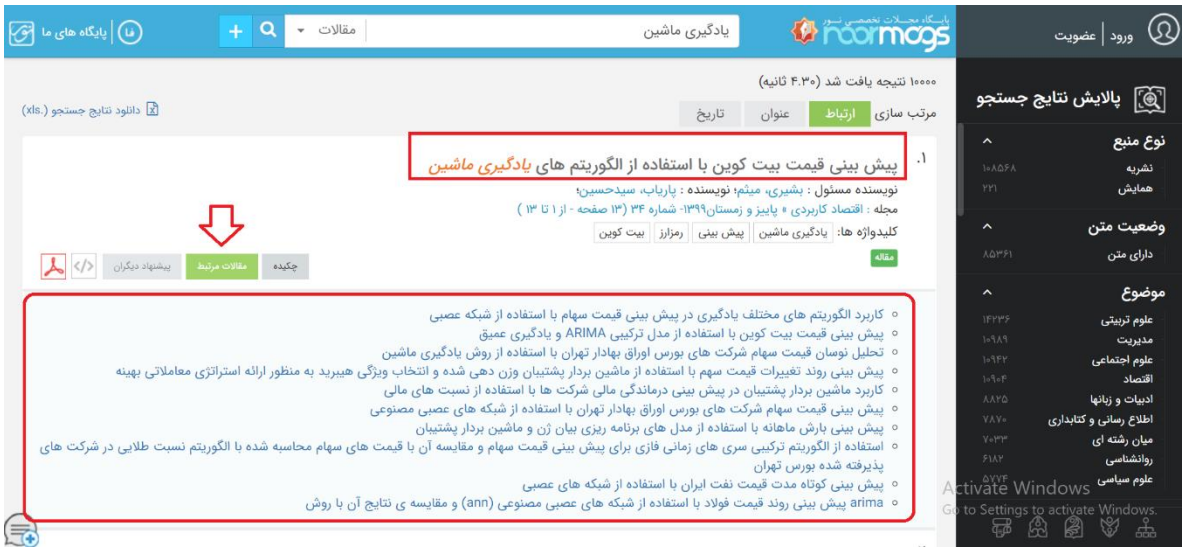
شکل ۲_۱۴: نمونه‌ای از عملکرد پایگاه نورمگز در یافتن مقالات مرتبط

شکل ۲_۱۵ نتایج کلیک روی کلیدواژه مرزارز در پایگاه نورمگز را نشان می‌دهد. در سمت چپ، روند انتشار مقاله، تعدادی مقالاتی که حاوی کلیدواژه رمزامرز هستند در سال‌های مختلف نشان می‌دهد. مثلا دو مقاله در سال ۱۳۹۸، چهار مقاله در سال ۱۳۹۹ و ۶ مقاله در سال ۱۴۰۰ که در کل ۱۲ مقاله با این کلیدواژه در پایگاه داده وجود دارند. علاوه بر این، لیست کلیدواژه‌های مرتبط با این کلیدواژه هم نشان داده شده است.



شکل ۲_۱۵: نمونه‌ای از عملکرد پایگاه نورمگز در یافتن مقالات مرتبط با استفاده از کلیدواژه "رمزارز"

پایگاه نورمگز، علاوه بر یافتن مقالات بر اساس کلیدواژه، می‌تواند مقالات مرتبط را نیز پیدا کند که در شکل ۲_۱۶ نمونه‌ای از آن، آورده شده است.



شکل ۲_۱۶: نمونه‌ای از عملکرد پایگاه نورمگز در یافتن مقالات مرتبط

کارایی نورمگز در قابلیت یافتن مقالات مرتبط در بعضی موارد با اشکال روبرو است. نمونه‌ای از آن در شکل ۲_۱۶ نشان داده شده است که به نظر هیچکدام از مقالات بازیابی شده به مقاله مورد نظر ارتباط ندارد. شکل ۲_۱۷ نمونه دیگری را نشان می‌دهد که فقط یکی (مشخص شده با رنگ سبز) را با اعماض می‌توان مرتبط در نظر گرفتم.

شکل ۲_۱۷: نمونه‌ای از عملکرد پایگاه نورمگز در یافتن مقالات مرتبط

از میان موتورهای جستجوی پایگاه‌های مقالات فارسی که در ایران وجود دارند، در زمان نگارش این گزارش، تنها نورمگز، مرجع دانش، جهاد دانشگاهی و پرتال جامع علوم انسانی قابلیت یافتن مقالات مرتبط را دارند و بقیه این قابلیت را ندارند. پرتال جامع علوم انسانی، تنها می‌تواند با انتخاب یک کلیدواژه، مقالاتی که حاوی آن کلیدواژه هستند، بیابد. در مورد کارایی سه پایگاه دیگر که این قابلیت را دارند، با اشکالاتی روبرو هستند. موتور جستجوی گوگل برای مقالات فارسی نیز در بعضی موارد اصلا نتوانسته مقالات مرتبط را پیدا کند و در بعضی موارد نیز نسخه‌های دیگر مقاله که در پایگاه‌های مختلفی ثبت شده‌اند، به عنوان مقاله مرتبط در نظر می‌گیرد. همچنین با ترجمه کلیدواژه‌های مقاله به انگلیسی، قادر است که مقالات انگلیسی آن کلیدواژه را به عنوان مقالات مرتبط با مقاله فارسی پیدا کند.

فصل سوم

روش‌شناسی پژوهش

۳. روش‌شناسی پژوهش

۳_۱ مقدمه

حجم مقالات علمی در سال‌های اخیر به شدت افزایش پیدا کرده است. از طرف دیگر، قدم اول در انجام یک پژوهش، مطالعه و بررسی مقالات و پژوهش‌های گذشته است. بنابراین یافتن مقالات مرتبط با یک مساله پژوهشگر از میان حجم بسیار زیاد مقالات یکی از چالش‌های موجود در انجام هر پژوهشی است. سامانه‌های بازیابی اطلاعات علمی کمک می‌کنند تا یک پژوهشگر بتواند مقالات مرتبط با پرسش خود را بازیابی نماید. یکی از قابلیت‌هایی که در بعضی از سامانه‌های بازیابی اطلاعات وجود دارد، ویژگی یافتن مقالات مرتبط با یک مقاله است. در این بخش الگوریتمی برای پیدا کردن مقالات مرتبط با یک مقاله بر اساس مراجع آن ارائه می‌گردد.

۳_۲ داده‌های پژوهش

در این پژوهش از دو مجموعه داده فارسی و انگلیسی استفاده می‌کند. برای تهیه داده‌های فارسی، مقالات منتشر شده در ۵ سال اخیر موجود در ۸ نشریه که در جدول ۳_۱ نشان داده شده‌اند، به تصادف استخراج گردیده است. این نشریات عبارتند از "پردازش علائم و داده‌ها"، "پژوهشنامه پردازش و مدیریت اطلاعات"، "رایانش نرم و فناوری اطلاعات"، "روش‌های عددی در مهندسی"، "علوم رایانش و فناوری اطلاعات"، "علوم رایانشی"، "محاسبات نرم"، "مهندسی برق و مهندسی کامپیوتر ایران". همانطور که جدول ۳_۱ نشان می‌دهد موضوع سطح کلان این نشریات، علوم فیزیکی و موضوع سطح میانی، علوم کامپیوتر است. از هر مقاله، اطلاعات کتابشناختی عنوان مقاله، چکیده کلیدواژه و مراجع انتخاب شدند. با تحلیلی که روی داده‌ها انجام گرفت، این نتیجه حاصل شد که تمامی مقاله‌ها حاوی عنوان و کلیدواژه و مراجع هستند ولی ۹۵٪ از داده‌ها چکیده ندارند.

جدول ۳_۱: لیست نشریات فارسی پژوهش

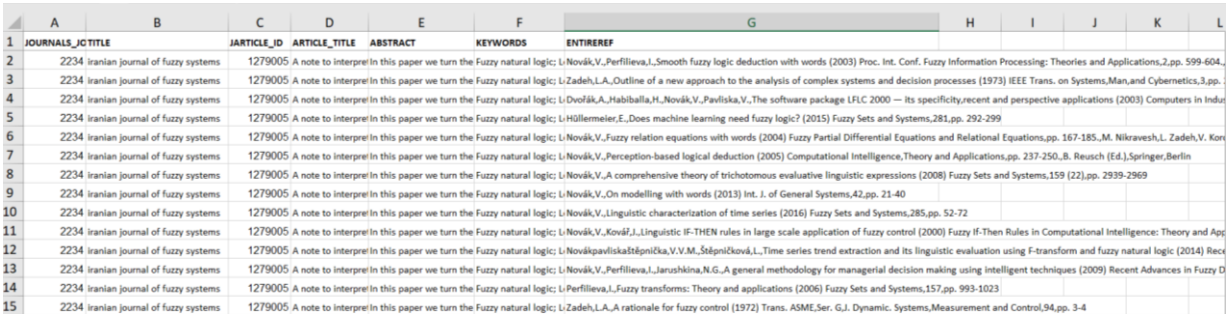
شماره	عنوان نشریه	آدرس	موضوع سطح کلان	موضوع سطح میانی
۱	پردازش علائم و داده‌ها	jsdp.rcisp.ac.ir	علوم فیزیکی	علوم کامپیوتر
۲	پژوهشنامه پردازش و مدیریت اطلاعات	jipm.irandoc.ac.ir	علوم فیزیکی	علوم کامپیوتر
۳	رایانش نرم و فناوری اطلاعات	jscit.nit.ac.ir	علوم فیزیکی	علوم کامپیوتر
۴	روش‌های عددی در مهندسی	www.jcme.iut.ac.ir/web/guest/homejcme.iut.ac.ir	علوم فیزیکی	علوم کامپیوتر
۵	علوم رایانش و فناوری اطلاعات	/jcsit.ir	علوم فیزیکی	علوم کامپیوتر
۶	علوم رایانشی	csj.isi.org.ir	علوم فیزیکی	علوم کامپیوتر
۷	محاسبات نرم	scj.kashanu.ac.ir	علوم فیزیکی	علوم کامپیوتر
۸	مهندسی برق و مهندسی کامپیوتر ایران	ijece.saminatech.ir	علوم فیزیکی	علوم کامپیوتر

برای تهیه داده‌های انگلیسی نیز ۱۰ نشریه در حوزه علوم کامپیوتر انتخاب شدند و مقالات نمایه شده در پایگاه استنادی علوم جهان اسلام در ۵ سال اخیر بازیابی گردیدند. جدول ۳_۲ لیست نشریات انگلیسی این پژوهش را نشان می‌دهد که همه این نشریات با موضوع سطح کلان علوم فیزیکی و سطح میانی علوم کامپیوتر می‌باشند. در نشریات انگلیسی نیز، اطلاعات عنوان، چکیده، مراجع و کلیدواژه مقالات مورد بررسی قرار گرفتند. تحلیل روی مقالات بازیابی شده از این نشریات نشان داد که عنوان و مراجع تمامی مقالات این نشریات موجود هست ولی کلیدواژه ۷٪ از مقالات دردسترس نیست و حدود ۳٪ از مقالات پژوهش فیلد چکیده آنها خالی است.

جدول ۳_۲: لیست نشریات انگلیسی پژوهش

شماره	عنوان نشریه	آدرس	موضوع سطح کلان	موضوع سطح میانی
۱	Journal Of Information Technology Management	/https://jitm.ut.ac.ir	علوم فیزیکی	علوم کامپیوتر
۲	Data Mining And Knowledge Discovery	https://www.springer.com/journal/10618	علوم فیزیکی	علوم کامپیوتر
۳	Journal Of Electrical And Computer Engineering Innovations	http://jecei.sru.ac.ir	علوم فیزیکی	علوم کامپیوتر
۴	Iranian Journal Of Fuzzy Systems	http://ijfs.usb.ac.ir	علوم فیزیکی	علوم کامپیوتر
۵	Journal Of Grid Computing	https://link.springer.com/journal/10723	علوم فیزیکی	علوم کامپیوتر
۶	Iranian Journal Of Mathematical Sciences And Informatics	http://www.ijmsi.ir	علوم فیزیکی	علوم کامپیوتر
۷	Journal Of Advances In Computer Engineering And Technology	http://jacet.srbiau.ac.ir	علوم فیزیکی	علوم کامپیوتر
۸	Journal Of Ai And Data Mining	http://jad.shahroodut.ac.ir	علوم فیزیکی	علوم کامپیوتر
۹	Iranian Journal Of Science And Technology, Transactions Of Electrical Engineering	http://ijste.shirazu.ac.ir	علوم فیزیکی	علوم کامپیوتر
۱۰	Journal Of Advances In Computer Engineering And Technology	http://jacet.srbiau.ac.ir	علوم فیزیکی	علوم کامپیوتر

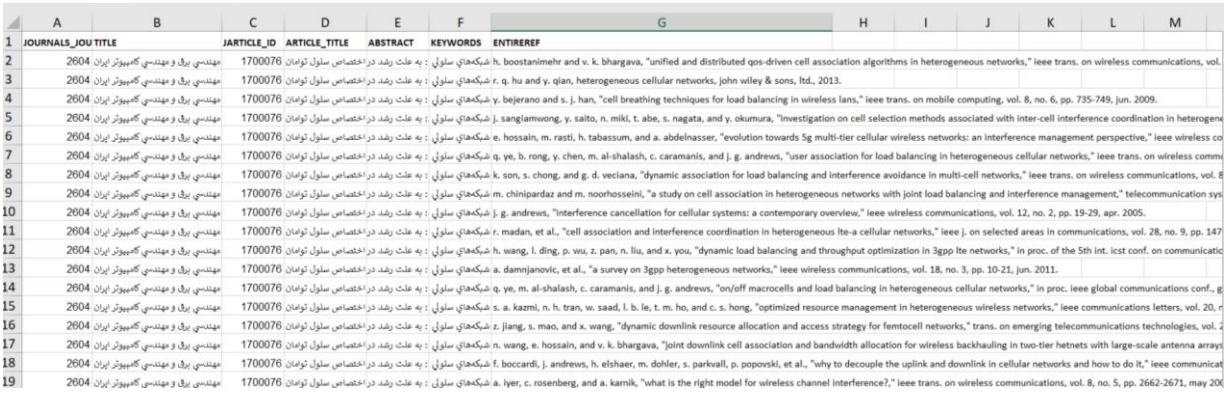
نمونه‌ای از داده‌ها در زبان فارسی و انگلیسی در شکل ۳_۱ و ۳_۲ نشان داده شده است.



	A	B	C	D	E	F	G
1	JOURNALS_JC	TITLE	ARTICLE_ID	ARTICLE_TITLE	ABSTRACT	KEYWORDS	ENTIREREF
2	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novák,V.,Perfilieva,I.,Smooth fuzzy logic deduction with words (2003) Proc. Int. Conf. Fuzzy Information Processing: Theories and Applications,2,pp. 599-604,
3	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Zadeh,L.A.,Outline of a new approach to the analysis of complex systems and decision processes (1973) IEEE Trans. on Systems,Man,and Cybernetics,3,pp.
4	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Dvořák,A.,Habiballa,H.,Novák,V.,Pavliska,V.,The software package LFLC 2000 — its specificity,recent and perspective applications (2003) Computers in Indu
5	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Hüllermeier,E.,Does machine learning need fuzzy logic? (2015) Fuzzy Sets and Systems,281,pp. 292-299
6	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novák,V.,Fuzzy relation equations with words (2004) Fuzzy Partial Differential Equations and Relational Equations,pp. 167-185.,M. Nikravesh,L. Zadeh,V. Kor
7	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novák,V.,Perception-based logical deduction (2005) Computational Intelligence,Theory and Applications,pp. 237-250.,B. Reusch (Ed.),Springer,Berlin
8	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novák,V.,A comprehensive theory of trichotomous evaluative linguistic expressions (2008) Fuzzy Sets and Systems,159 (22),pp. 2939-2969
9	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novák,V.,On modelling with words (2013) Int. J. of General Systems,42,pp. 21-40
10	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novák,V.,Linguistic characterization of time series (2016) Fuzzy Sets and Systems,285,pp. 52-72
11	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novák,V.,Kováľ,J.,Linguistic IF-THEN rules in large scale application of fuzzy control (2000) Fuzzy If-Then Rules in Computational Intelligence: Theory and Ap
12	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novákpavliskaštěpnička,V.V.M.,Štěpničková,L.,Time series trend extraction and its linguistic evaluation using F-transform and fuzzy natural logic (2014) Rec
13	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Novák,V.,Perfilieva,I.,Jarushkina,N.G.,A general methodology for managerial decision making using intelligent techniques (2009) Recent Advances in Fuzzy D
14	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Perfilieva,I.,Fuzzy transforms: Theory and applications (2006) Fuzzy Sets and Systems,157,pp. 993-1023
15	2234	iranian journal of fuzzy systems	1279005	A note to interpre	In this paper we turn the	Fuzzy natural logic; L	Zadeh,L.A.,A rationale for fuzzy control (1972) Trans. ASME,Ser. G,J. Dynamic. Systems,Measurement and Control,94,pp. 3-4

شکل ۳_۱: نمایی از داده‌های انگلیسی پژوهشی

ستون اول id نشریه و ستون بعدی عنوان نشریه است. ستون‌های بعدی اطلاعات مقاله را نشان می‌دهند که به ترتیب عبارتند از id مقاله، عنوان مقاله، چکیده و کلیدواژه‌ها. ستون آخر نیز مراجع هر مقاله را نشان می‌دهد. به عبارت دیگر هر ردیف این داده، یک مرجع از مراجع هر مقاله را مشخص می‌نماید. به عنوان مثال، در صورتی که مقاله ۲۰ مرجع داشته باشد، ۲۰ ردیف به ازای آن مقاله در این مجموعه داده وجود دارد که هر ردیف آن نشاندهنده یک مرجع از مراجع موجود در مقاله است.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	JOURNALS_JOU	TITLE	JARTICLE_ID	ARTICLE_TITLE	ABSTRACT	KEYWORDS	ENTIREREF						
2	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			h. boostanimehr and v. k. bhargava, "unified and distributed qos-driven cell association algorithms in heterogeneous networks," ieee trans. on wireless communications, vol.						
3	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			r. q. hu and y. qian, heterogeneous cellular networks, john wiley & sons, ltd., 2013.						
4	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			y. bejerano and s. j. han, "cell breathing techniques for load balancing in wireless lans," ieee trans. on mobile computing, vol. 8, no. 6, pp. 735-749, jun. 2009.						
5	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			j. sangiamwong, y. saito, n. miki, t. abe, s. nagata, and y. okumura, "investigation on cell selection methods associated with inter-cell interference coordination in heterogene						
6	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			e. hossain, m. rasti, h. tabassum, and a. abdelnasser, "evolution towards 5g multi-tier cellular wireless networks: an interference management perspective," ieee wireless co						
7	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			q. ye, b. rong, y. chen, m. al-shalash, c. caramanis, and j. g. andrews, "user association for load balancing in heterogeneous cellular networks," ieee trans. on wireless comm						
8	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			k. son, s. chong, and g. d. veciana, "dynamic association for load balancing and interference avoidance in multi-cell networks," ieee trans. on wireless communications, vol. 8						
9	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			m. chinipardaz and m. noorhosseini, "a study on cell association in heterogeneous networks with joint load balancing and interference management," telecommunication sys						
10	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			j. g. andrews, "interference cancellation for cellular systems: a contemporary overview," ieee wireless communications, vol. 12, no. 2, pp. 19-29, apr. 2005.						
11	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			r. madan, et al., "cell association and interference coordination in heterogeneous lte-a cellular networks," ieee j. on selected areas in communications, vol. 28, no. 9, pp. 147						
12	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			h. wang, l. ding, p. wu, z. pan, n. liu, and x. you, "dynamic load balancing and throughput optimization in 3gpp lte networks," in proc. of the 5th int. icst conf. on communicatic						
13	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			a. damnjanovic, et al., "a survey on 3gpp heterogeneous networks," ieee wireless communications, vol. 18, no. 3, pp. 10-21, jun. 2011.						
14	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			q. ye, m. al-shalash, c. caramanis, and j. g. andrews, "on/off macrocells and load balancing in heterogeneous cellular networks," in proc. ieee global communications conf., g						
15	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			s. a. kazmi, n. h. tran, w. saad, l. b. le, t. m. ho, and c. s. hong, "optimized resource management in heterogeneous wireless networks," ieee communications letters, vol. 20, r						
16	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			z. jiang, s. mao, and x. wang, "dynamic downlink resource allocation and access strategy for femtocell networks," trans. on emerging telecommunications technologies, vol. 2						
17	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			n. wang, e. hossain, and v. k. bhargava, "joint downlink cell association and bandwidth allocation for wireless backhauling in two-tier hetnets with large-scale antenna arrays						
18	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			f. boccardi, j. andrews, h. elshaer, m. dohler, s. parkvall, p. popovski, et al., "why to decouple the uplink and downlink in cellular networks and how to do it," ieee communicat						
19	2604	مهندسی برق و مهندسی کامپیوتر ایران	1700076	شبکه‌های سلولی : به علت رشد در اخصاص سلول توامان			a. iyer, c. rosenberg, and a. karnik, "what is the right model for wireless channel interference?," ieee trans. on wireless communications, vol. 8, no. 5, pp. 2662-2671, may 20(						

شکل ۳_۲: نمایی از داده‌های فارسی پژوهش

۳_۳ ارائه الگوریتم یافتن مقالات مرتبط با یک مقاله بر اساس تحلیل مراجع

در این پژوهش، روشی مبتنی بر تحلیل مراجع برای یافتن مقالات مرتبط با یک مقاله ارائه شده است. پژوهش‌های گذشته مقالاتی را مرتبط در نظر می‌گرفتند که مراجع یکسانی داشته باشند. در این پژوهش نه تنها مقالات با مراجع یکسان مرتبط در نظر گرفته می‌شوند؛ بلکه روش پیشنهادی مقالاتی که از نظر مراجع با هم شباهت دارند، به عنوان مقاله مرتبط در نظر می‌گیرد. به عبارت دیگر برای یافتن مقالات مرتبط با یک مقاله داده شده، مراجع آن مقاله با دیگر مقالات موجود مقایسه می‌گردد و مقالاتی بازیابی می‌گردد که بیشترین شباهت را از نظر مراجع با یکدیگر داشته باشند.

از آنجا که مقالات این پژوهش از نشریات مختلفی گردآوری شده‌اند و هر کدام الگوهای متفاوتی برای نگارش مقاله دارند، بنابراین مراجع داده‌های پژوهشی به طرق مختلفی می‌باشند که در ادامه، استانداردهای مختلف نگارش مراجع در مقالات معرفی می‌گردند. این فرمت‌ها که از scholar.google.com گرفته شده، به ترتیب عبارتند از MLA، APA، Chicago، Harvard و Vancouver.

- فرمت MLA

در این شیوه ارجاع‌دهی، ابتدا نام خانوادگی و سپس نام نویسندگان نوشته می‌شود. در ادامه، عنوان مقاله در گیومه و سپس عنوان نشریه به صورت مورب مشخص می‌شوند. در انتها نیز سری و شماره مقاله به همراه سال انتشار و شماره صفحات درج می‌گردد. نمونه‌ای از شیوه ارجاع‌دهی به یک مقاله با فرمت MLA در ادامه آمده است.

Deng, Ruilong, et al. "Sensing-performance tradeoff in cognitive radio enabled smart grid." *IEEE Transactions on Smart Grid* 4.1 (2013): 302-310.

- فرمت APA

این الگو که از محبوب‌ترین شیوه‌های ارجاع‌دهی در اکثر نشریات است، ابتدا نام نویسندگان را درج می‌نماید که برای این کار نام خانوادگی نویسنده اول، اولین حرف نام نویسنده اول را مشخص می‌کند که این رویه را برای تمامی نویسندگان به ترتیب انجام می‌دهد. در ادامه سال انتشار در پرانتز و سپس عنوان مقاله نوشته می‌شود. عنوان نشریه‌ای که مقاله در آن به چاپ رسیده‌است، سری و شماره و همچنین شماره صفحات در انتهای آن درج می‌شوند. نمونه‌ای از شیوه استناددهی با روش APA در زیر نشان داده شده است.

Deng, R., Chen, J., Cao, X., Zhang, Y., Maharjan, S., & Gjessing, S. (2013). Sensing-performance tradeoff in cognitive radio enabled smart grid. *IEEE Transactions on Smart Grid*, *4*(1), 302-310.

- فرمت Chicago

در این شیوه ارجاع‌دهی، ابتدا نام خانوادگی نویسندگان و سپس نام آنها به ترتیب مشارکتشان در مقاله مشخص می‌شود. در ادامه عنوان مقاله در گیومه و نشریه نوشته می‌شوند. سری، شماره، سال انتشار و شماره صفحات نیز در انتهای این شیوه ارجاع‌دهی مشخص می‌گردند. نمونه‌ای از شیوه ارجاع با فرمت Chicago در ادامه آمده است:

Deng, Ruilong, Jiming Chen, Xianghui Cao, Yan Zhang, Sabita Maharjan, and Stein Gjessing. "Sensing-performance tradeoff in cognitive radio enabled smart grid." *IEEE Transactions on Smart Grid* 4, no. 1 (2013): 302-310.

- فرمت Harvard

این شیوه ارجاع‌دهی، با نام خانوادگی و حرف اول نام نویسنده اول آغاز می‌گردد. نام نویسندگان به ترتیب مشارکت با این رویه از اول تا آخر نوشته می‌شود. سپس سال انتشار مقاله بدون پرانتز و عنوان مقاله و نشریه آن مشخص می‌شوند. در انتها نیز سری و شماره مقاله و شماره صفحات نگارش می‌شوند. نمونه‌ای از شیوه ارجاع‌دهی با این فرمت در ادامه آمده است:

Deng, R., Chen, J., Cao, X., Zhang, Y., Maharjan, S. and Gjessing, S., 2013. Sensing-performance tradeoff in cognitive radio enabled smart grid. *IEEE Transactions on Smart Grid*, *4*(1), pp.302-310.

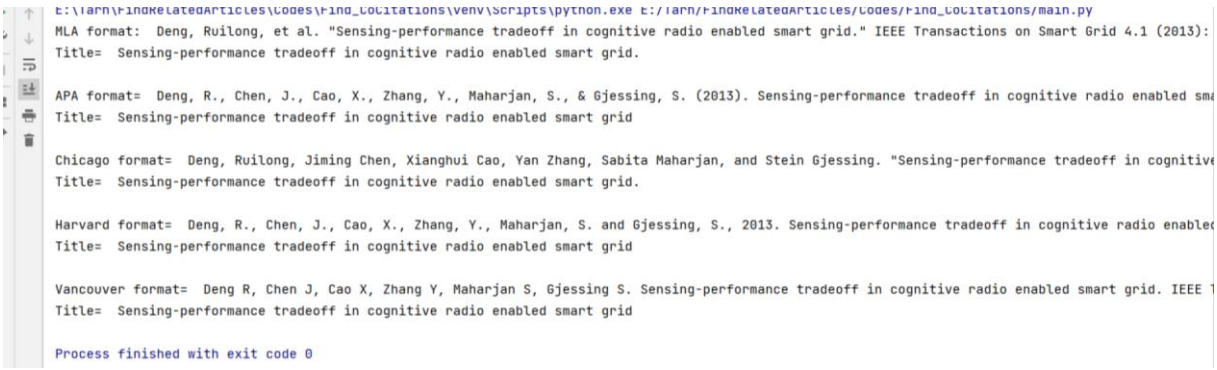
- فرمت Vancouver

فرمت Vancouver با نام خانوادگی نویسنده اول و سپس حرف اول نام نویسنده اول آغاز می‌گردد. سپس نام تمامی نویسندگان به همین ترتیب مشخص می‌گردند. در ادامه عنوان مقاله و سپس عنوان نشریه و سال انتشار نگارش می‌شوند. این فرمت با مشخص کردن سری، شماره و شماره صفحات مقاله پایان می‌یابد. نمونه‌ای از نحوه ارجاع‌دهی با این فرمت در ادامه مشخص شده است:

Deng R, Chen J, Cao X, Zhang Y, Maharjan S, Gjessing S. Sensing-performance tradeoff in cognitive radio enabled smart grid. IEEE Transactions on Smart Grid. 2013 Feb 18;4(1):302-10.

با توجه به اینکه داده‌های موجود در این پژوهش از نشریات مختلفی گرفته شده‌اند و هر کدام از آنها از یکی از این فرمت‌ها برای تهیه مراجع استفاده می‌نماید، می‌بایست تمامی حالات مختلف تهیه منبع مدنظر قرار بگیرد و با توجه به آنها، عنوان موجود در هر منبع را استخراج نماید. بدین منظور برنامه‌ای به زبان پایتون (نسخه ۳,۷) نوشته شد که با در نظر گرفتن تمامی این حالات قرار گرفتن عنوان که ناشی از فرمت‌های مختلف مراجع می‌باشد، عناوین موجود را استخراج نماید.

نمونه‌ای از اجرای برنامه با در نظر گرفتن مثال بالا در شکل ۳_۳ ارائه شده است.



شکل ۳_۳: نمونه‌ای از اجرای برنامه استخراج عنوان از فرمت‌های مختلف ارجاع‌دهی

به منظور نشان دادن کارایی روش پیشنهادی برای استخراج عنوان، الگوریتم روی چند داده دیگر از مجموعه داده‌های این پژوهش که به صورت تصادفی از مجموعه داده‌ها انتخاب شده‌اند، اجرا گردید که نتایج آن در شکل ۳_۴ آمده است.

شکل ۳_۴: نمونه‌هایی از اجرای کد استخراج عنوان روی چندین مرجع از مجموعه داده‌ها

الگوریتم پیشنهادی برای یافتن مقالات مرتبط با مقاله داده شده با استفاده از تحلیل مراجع در زیر نشان داده شده است. ورودی این الگوریتم شناسه یک مقاله و همچنین تعداد مقالات بازیابی شده است. خروجی نیز لیست شناسه مقالات مرتبط با مقاله داده شده می‌باشد. خط ۱۲ تا ۱۳ تمامی شناسه‌های مقالات را بررسی می‌نماید. خطوط ۱۴ تا ۱۶ الگوریتم تمامی مراجع هر کدام از شناسه‌ها را بررسی کرده و عنوان آنها را استخراج می‌کند. خط ۱۷ در صورتی که شباهت عنوان استخراج شده با عنوان مورد پرسش کاربر، از یک مقدار آستانه که در این پژوهش ۰,۸ در نظر گرفته شده، بیشتر باشد، یک مقدار به متغیر شمارش اضافه می‌کند. برای بدست آوردن

شباهت میان مراجع، ابتدا عنوان آنها استخراج می‌گردد. همانطور که قبلا توضیح داده شد مراجع ممکن است با فرمت‌های مختلف نگارش شده باشد که می‌بایست تمامی این تنوع فرمت‌ها لحاظ گردد. سپس عملیات پیش-پردازش روی عنوان انجام گرفته و همچنین کلمات ایستا حذف می‌شوند. هر عنوان بر اساس کارکتر فاصله[۱] به برداری از ترم‌ها تبدیل می‌شود و شباهت میان دو بردار بر اساس اشتراک ترم‌های تشکیل‌دهنده دو بردار با توجه به معیار جاکارد بدست می‌آید. لازم به ذکر است که در صورتی که شباهت میان دو ترم بر اساس الگوریتم لونشتین-دامرا از معیار آستانه که در این پژوهش ۰,۸ در نظر گرفته شده، بیشتر باشد، آن دو ترم در معیار جاکارد، مشترک در نظر گرفته می‌شوند. بدین ترتیب یک گراف از شباهت میان مراجع مقالات ساخته می‌شود. در نهایت خطوط ۲۳ تا ۲۵ مقالاتی را که بیشترین شباهت با مقاله داده شده داشته باشند، براساس گراف ساخته شده در مرحله قبل بازیابی می‌کند.

Algorithm

```
1   Input:
2       query_id: id of a given query
3       T: number of retrieved papers
4   Output:
5       R: list of related articles id
6   Definition:
7       id_list: list of all paper ids
8       title: title of query
9       ref_list: list of all references
10      ref_query: list of all references of the query
11  Begin:
12      for i=1,2…len(id_list) do
13          id_temp=id_list[i]
14          for j=1,2…len(ref_list[id_temp]) do
15              ref_temp= ref_list[id_temp][j]
16              title_temp=extract_title(ref_temp)
17              if similar(title,title_temp)>thr do
18                  count=count+1
19              end
20          end
21      end
22          co_ref.append(count)
23      for k=1,2,,,T do
24          return argmax(co_ref)
25      end
26  end
```

[۱] Space

فصل چهارم

یافته‌ها

۴. یافته‌ها

۴_۱ مقدمه

در این فصل به مطالعه‌ی کارایی روش پیشنهادی می‌پردازیم. ابتدا معیار ارزیابی استفاده شده در این پژوهش معرفی می‌گردند و در نهایت کارایی روش پیشنهادی را با توجه به معیار ارزیابی داده شده، بررسی خواهیم کرد.

۴_۲ معیار ارزیابی

در سامانه‌های بازیابی اطلاعات عمدتا از معیارهای دقت[۱] و بازیافت[۲] برای سنجش میزان کارایی روش‌های ارائه شده استفاده می‌شود. معیار بازیافت تعداد مقالات مرتبط بازیابی شده از میان کل مقالات مرتبط را در نظر می‌گیرد. بنابراین این معیار نیاز به داشتن کل مقالات مرتبط به ازای هر پرسش کاربر است. از آنجا که کل مقالات مرتبط به ازای هر پرسش به ازای داده‌های این پژوهش در دسترس نیست، نمی‌توان از این معیار برای ارزیابی روش پیشنهادی استفاده کرد.

معیار دقت بررسی می‌کند که از میان اسناد بازیابی شده، چند سند مرتبط با پرسش کاربر می‌باشد. به عنوان مثال در صورتی که تعداد اسناد بازیابی شده ۵ باشد و از این میان تنها ۲ سند مرتبط با مقاله مورد پرسش کاربر باشد، معیار دقت در این حالت ۰,۴ می‌شود. فرمول ۴_۱ معیار دقت را نشان می‌دهد.

دقت=تعداد مقالات بازیابی شده مرتبط / تعداد مقالات بازیابی شده (۴_۱)

[۱] precision

[۲] recall

۴_۳ یافته‌ها

به منظور ارزیابی دقیق‌تر روش پیشنهادی، به ازای داده‌های انگلیسی که از نشریات انگلیسی گرفته شده‌اند، ۳۰ مقاله به صورت تصادف انتخاب شدند. همین رویه برای داده‌های فارسی نیز انجام گرفت و ۳۰ مقاله به صورت تصادفی از میان مجموعه مقالات انتخاب گردیدند. به ازای هر مقاله، اطلاعات کتابشناختی که حاوی عنوان، کلیدواژه و چکیده مقالات است به همراه لیست مراجع آن استخراج شد. جدول ۱_۱ و ۱_۲ لیست عناوین فارسی و انگلیسی مقالاتی که به صورت تصادفی انتخاب شده‌اند، نشان می‌دهد.

جدول ۴ _۱: عناوین مقالات انتخاب شده در مجموعه داده فارسی به عنوان پرسش

شماره	عنوان
۱	روشی نوین برای پیش‌بینی ارتباط در شبکه‌های اجتماعی ناهمگن
۲	ارائه یک الگوریتم مبتنی بر رایانش مه جهت مسیریابی شبکه‌های حسگر بی‌سیم
۳	نظرکاوی افزایشی با استفاده از یادگیری فعال بر روی جریان متون
۴	پیشنهاد هشتگ در سیستم‌های میکروبلاگ توسط بردار موضوعی: مورد کاربرد توئیتر
۵	تاثیر الگوی موضوعی رفتار جستجوی کاربران نوجوان بر پیشنهاد پرس‌وجو
۶	پیشنهاددهنده تطبیقی منابع آموزشی بر اساس سبک یادگیری، بازخورد کاربر و الگوریتم اتوماتای یادگیر
۷	تشخیص جنسیت نویسنده مستقل از متن و زبان نوشتاری با استفاده از پالایش پویای نمادین مبتنی بر تبدیل رادن
۸	طبقه‌بندی و شناسایی وب سایت‌های فیشینگ به کمک مجموعه قوانین فازی و الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیب‌دار
۹	شناسایی افراد از طریق رگ‌های خونی انگشت دست در فضای رادون با به کارگیری الگوهای فضایی مشترک
۱۰	استخراج مفاهیم کلیدی با استفاده از شبکه قاب و زنجیره مفاهیم
۱۱	شناسایی پایدار فعالیت فیزیکی انسان بر اساس سنسورهای گوشی هوشمند
۱۲	ارزیابی کیفیت وب‌سایت روزنامه‌های سراسری ایران از نظر شاخص‌های بهینه‌سازی موتورهای کاوش(سئو)
۱۳	ارائه چارچوبی برای ارزیابی وبگاه‌ها از منظر معماری اطلاعات

۱۴	استخراج هوشمند مرز فراداده و متن در پایان نامه های فارسی با رویکردba_svm
۱۵	ارزیابی و پیش بینی عوامل کیفیت پاسخ ها در سیستم پرسش و پاسخ شبکه اجتماعی علمی ریسرچ گیت: مطالعه موردی قلمرو موضوعی مدیریت دانش
۱۶	ارائۀ یک الگوریتم ترکیبی با استفاده از الگوریتم کرم شب‌تاب، الگوریتم ژنتیک و جست‌وجوی محلی
۱۷	ارزش‌گذاری ارجاعات غیرمستقیم در شبکه‌های استنادی با استفاده از تلفیق داده‌ها
۱۸	الگوریتم چندمعیاره برای تعیین مسیر حرکت گره چاهک در شبکه‌های حسگر بی‌سیم
۱۹	مجموع فاصله‌های بین رئوس یک گراف
۲۰	آموزش شبکۀ عصبی مصنوعی با نسخۀ آشوب‌گونۀ الگوریتم جستجوی گرانشی و کاربرد آن در پیش‌بینی آلاینده‌های هوا: مطالعۀ قیاسی
۲۱	مروری بر روش‌های تخمین هزینۀ نرم‌افزار مبتنی بر یادگیری ماشین
۲۲	بررسی مولفه‌های سبکی نویسندگان پیام‌های الکترونیکی با تکیه بر پژوهش‌های انجام‌شده
۲۳	بهبود مسیریابی برای شبکه‌های موردی بین خودرویی (vanets) با استفاده از الگوریتم‌های الهام‌گرفته از طبیعت
۲۴	بازشناسایی فعالیت‌های انسان در ویدیو با استفاده از ویژگی‌های freakhog و ماشین بردار پشتیبان آبشاری
۲۵	مدل بومی ارزیابی کیفیت سایت‌های خبری(newsqual)
۲۶	تعیین و تشخیص ضربان قلب در سیگنال الکتریکی قلب برای کاربردهای پزشکی از راه دور
۲۷	بررسی تاثیر استفاده از روش‌های یادگیری ماشین تجمعی در شناسایی نظرهای هرز بر اساس ویژگی‌های رفتاری
۲۸	یک مدل محاسباتی چندوجهی از اعتماد و بی‌اعتمادی با آگاهی از زمینه در شبکه‌های اجتماعی برخط
۲۹	رتبه بندی ویژگی‌ها در تشخیص نظرات اسپم فارسی
۳۰	بهبود بازیابی تصاویر رنگی با استفاده از رنگ، بافت و شکل در روش کیسه کلمات بصری مبتنی بر امضاء

جدول ۴ _۲: عناوین مقالات انتخاب شده در مجموعه داده انگلیسی به عنوان پرسش

شماره	عنوان مقاله
۱	cloud computing application and its advantages and difficulties in the teaching process
۲	the impact of blockchain on accounting information systems
۳	the future of bitcoin as a tool for financial development
۴	effective learning to rank persian web content
۵	text analytics of customers on twitter: brand sentiments in customer support
۶	a grouping hotel recommender system based on deep learning and sentiment analysis
۷	iot future security challenges and recent solutions
۸	determining journal rank by applying particle swarm optimization-naive bayes classifier
۹	iot-based services in banking industry using a business continuity management approach
۱۰	continued usage of e-learning: a systematic literature review
۱۱	feature selection using a genetic algorithms and fuzzy logic in anti-human immunodeficiency virus prediction for drug discovery
۱۲	deep-learning-cnn for detecting covered faces with niqab
۱۳	graph-based extractive text summarization models: a systematic review
۱۴	the effects of social networking sites use on students’ academic performance at the university of taiz
۱۵	designing a mobile application for children: space science
۱۶	filter-based feature selection using information theory and binary cuckoo optimisation algorithm
۱۷	a decentralized polling system using ethereum technology
۱۸	stcp: a novel approach for congestion control in iot environment
۱۹	mapping grayscale images to colour space using deep learning
۲۰	implementing the blockchain technology in islamic financial industry: opportunities and challenges
۲۱	the mediating role of customer trust in affecting the relationship between online shopping factors and customer purchase decision
۲۲	the impact of covid-19 crisis upon the effectiveness of e-learning in higher education institution

the role of information technology on the muslim community in the era of globalization and digitalizatio	۲۳
the use of social media application as a factor influencing the students' decisions-making to enrol at private higher education institutions using smart pls	۲۴
the engineering of e-governance and technology in the management of secondary schools: case of the nouaceur delegation	۲۵
the effect of online marketing through social media platforms on saudi public libraries	۲۶
advertising strategy management in internet marketing	۲۷
A Constrained Optimization Approach to Integrated Active Fault Detection and Control	۲۸
an empirical study on the effectiveness of monkey testing for android applications	۲۹
a gender-aware deep neural network structure for speech recognition	۳۰

روش پیشنهادی با چهار روش دیگر برای بدست آوردن مقالات مرتبط با یک مقاله مقایسه شده‌اند که این روش‌ها به ترتیب TF-IDF (Zhang et al., 2010)، word2vec (Church, 2017)، DOC2VEC (Dai et al., 2015) و BERT (Delving et al., 2018) است. جدول ۴_۳ مقایسه میانگین دقت و انحراف معیار روش پیشنهادی را با روش‌های دیگر نشان می‌دهد. لازم به ذکر است که هر کدام از اعداد این جدول میانگین ۳۰ بار اجرای هر کدام از روش‌ها با توجه به پرسش‌های جداول ۴_۱ و ۴_۲ می‌باشد.

جدول ۴ _۳: مقایسه دقت روش پیشنهادی با دیگر روش‌ها

	روش پیشنهادی	**TF-IDF**	**Word2vec**	**DOC2VEC**	**BERT**
داده‌های نشریات فارسی	0.7±0.29	0.42±0.24	0.27±0.16	0.41±0.23	0.53±0.28
داده‌های نشریات انگلیسی	0.60±28	0.42±0.2	0.47±0.24	0.39±0.17	0.60±0.26

علیرغم روش پیشنهادی که مقالات مرتبط را بر اساس تحلیل مراجع بازیابی می‌کند، چهار روش دیگر با تحلیل محتوای اطلاعات کتابشناختی مقالات مانند چکیده، عنوان و کلیدواژه عمل می‌کنند. بنابراین در قدم اول، چکیده، عنوان و کلیدواژه مقالات مورد پیش‌پردازش قرار می‌گیرد.

برای انجام این کار، ابتدا متن با توجه به کارکترهای جداکننده[۱] که شامل { ، ؛ " ( ) : . < > } هستند، به مجموعه‌ای از توکن‌ها تبدیل می‌شود. سپس عملیات ریشه‌یابی[۲] روی آنها انجام می‌گیرد. هر واژه با کد منحصر به فردی ذخیره می‌شود. علاوه بر واژه، تعداد رخداد آن در مجموعه مقالات نیز محاسبه و ذخیره می‌گردد. همچنین ایست‌واژه[۳]ها، علائم و اعداد حذف می‌شوند. خروجی این مرحله، واژگان پردازش‌شده‌ای می‌باشد که فرکانس تکرار آنها در هر مقاله نیز مشخص شده است.

تشخیص واژه‌های ایستا یکی از مهمترین عملیات در متن‌کاوی است. واژه‌های ایستا معمولا خیلی زیاد در اسناد کل مجموعه رخ می‌دهند و عمدتا حاوی اطلاعات باارزشی در مورد متن و یا اسناد نیستند. بنابراین بهتر است که این واژه‌ها از کل مجموعه حذف گردند (Sadeghi & Vegas, 2014). در مرحله آخر از پیش‌پردازش، این واژه‌ها نیز حذف می‌گردند.

در روش مبتنی بر TF-IDF هر مقاله که حاوی عنوان، کلیدواژه و چکیده است به صورت یک بردار تبدیل می‌شود که هر کدام از اعضای آن بردار، TF-IDF کلمات آن مقاله می‌باشد. بنابراین کل داده‌ها به صورت یک بردار عددی تبدیل می‌شوند که این رویه در فاز آفلاین انجام می‌گیرد. حال زمانی که کاربر یک مقاله را انتخاب کرده و به دنبال مقالات مرتبط با آن مقاله می‌گردد، بردار آن مقاله با دیگر بردارهای مقالات موجود مقایسه شده و بردارهایی که نزدیکترین فاصله را با آن دارند، بازیابی می‌شوند. لازم به ذکر است که برای بدست آوردن فاصله میان دو بردار که هر کدام یک مقاله است، از معیار کسینوسی استفاده شده است. همانطور که جدول ۴_۱ نشان می‌دهد، این روش توانسته است با دقت ۰,۴۲ مقالات مرتبط با یک مقاله را بازیابی نماید.

سه روش دیگری که در این پژوهش برای مقایسه روش پیشنهادی با آنها استفاده شده است، از تعبیه کلمات[۴] استفاده می‌کنند. تعبیه کلمات بردارهای عددی هستند که نمایانگر کلمات یک پیکره هستند و کاربرهای گسترده-ای به خصوص در حوزه پردازش زبان طبیعی دارند. روش تعبیه کلمات، اجازه می‌دهد که به طور غیرصریح، اطلاعاتی را از دنیای بیرون به مدل‌های زبانی اضافه کنید. در تعبیه کلمات، تمام کلمات استفاده شده در یک زبان، به وسیله مجموعه‌ای از اعداد اعشاری (در قالب یک بردار) نمایش داده می‌گردد. در واقع تعبیه کلمات، بردارهای n بعدی هستند که تلاش می‌کنند معنای کلمات و محتوای آنها را با مقادیر عددی خود ثبت و ضبط کنند. لازم به ذکر است که برای آموزش در تمامی این الگوریتم‌ها از مدل پیش‌آموزش، استفاده شده است.

در روش‌های word2vec و DOC2VEC به کمک شبکه عصبی یک بردار با اندازه کوچک و ثابت برای نمایش تمام لغات و متون در نظرگرفته شده و با اعداد مناسب در فاز آموزش برای هر لغت این بردار محاسبه می شود. که البته فاز آموزش بعد از پیش‌پردازش داده‌ها انجام گرفته است. بنابراین زمانی که کاربر یک مقاله را برای یافتن

---

[۱] Delimitor characters

[۲] Stemming

[۳] Stop words

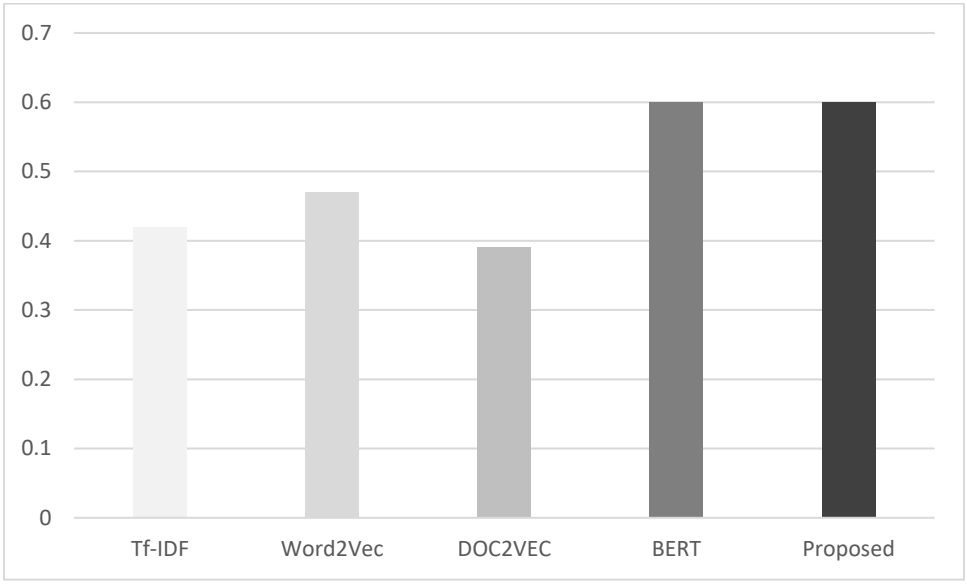
[۴] Word embedding

مقالات مرتبط انتخاب می‌کند، روش مبتنی بر word2vec ابتدا عملیات پیش‌پردازش را روی چکیده، عنوان و کلیدواژه مقاله انتخاب شده بکار گرفته و سپس آن را تبدیل به بردار می‌کند. در نهایت میزان شباهت بردار مقاله داده شده با دیگر بردار مقالات با استفاده از معیار کسینوسی محاسبه می‌گردد.

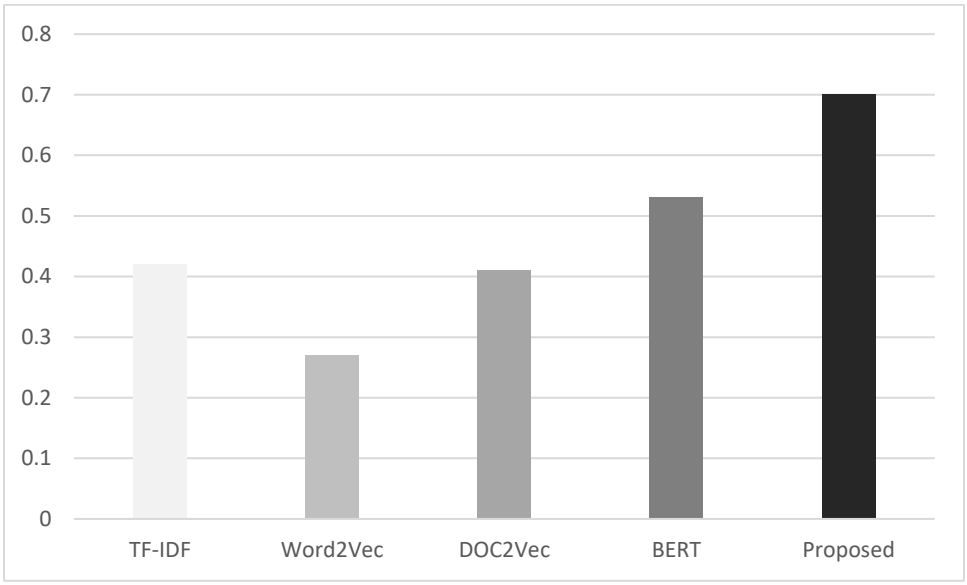
روش دیگری که در این پژوهش برای مقایسه با روش پیشنهادی استفاده شده است، الگوریتم برت است که توسط شرکت گوگل ارائه شده است و روی معماری ترنسفورمرها برای مدل‌سازی زبان‌ها عمل می‌کند. در اینجا هم بعد از پیش‌پردازش چکیده، کلیدواژه و عنوان مقالات، تمامی داده‌ها با استفاده از این مدل، آموزش می‌بینند که تمامی این عملیات در فاز آفلاین انجام می‌گیرد. در مرحله آنلاین، زمانی که کاربر یک مقاله را برای یافتن مقالات مرتبط انتخاب می‌نماید، عنوان، کلیدواژه و چکیده مقاله پیش‌پردازش شده و سپس تبدیل به بردار می‌گردد. در نهایت، شبیه‌ترین بردار به بردار پرسش کاربر با توجه به معیار کسینوسی بازیابی می‌گردد.

برای روشن‌تر شدن روند بدست آوردن نتایج، با یک مثال روند آن توضیح داده می‌شود. عنوان اولین پرسش در جدول ۴_۱ عبارت است از "روشی نوین برای پیش‌بینی ارتباط در شبکه‌های اجتماعی ناهمگن". عنوان این مقاله به همراه کلیدواژه و چکیده آن به روش‌های بدست آوردن مقاله مرتبط با تحلیل محتوا مانند TF-IDF، Word2vec، DOC2VEC و BERT داده شد. پارامتر T در الگوریتم ۱ که همان تعداد مقالات بازیابی شده است، پنج در نظر گرفته شد. سپس به ازای هر کدام از روش‌ها پنج مقاله بازیابی شده مورد تحلیل و بررسی دقیق قرار گرفت تا مشخص شود کدامیک از آنها با مقاله مورد پرسش ارتباط دارند. از آنجا که تمامی نشریاتی که برای داده‌های پژوهشی مورد استفاده قرار گرفتند، از نظر موضوعی با تخصص پژوهشگر مطابقت داشت، بنابراین این تحلیل توسط پژوهشگر به عنوان عامل انسانی متخصص انجام گرفت. سپس تمامی مراجع مقاله به همراه کل مراجع تمامی مقالات به روش پیشنهادی داده شد و مقالات بازیابی شده مورد تحلیل قرار گرفت و مقالات مرتبط آن شناسایی گردید. همین روند برای تمامی ۳۰ مقاله فارسی و همچنین ۳۰ مقاله انگلیسی انجام گرفت. در نهایت میانگین و انحراف معیار در جدول ۴_۳ گزارش گردید.

به منظور مقایسه شهودی الگوریتم پیشنهادی با دیگر روش‌ها نمودار میله‌ای آنها را رسم کرده که شکل‌های ۴_۱ و ۴_۲ این مقایسه را روی داده‌های فارسی و انگلیسی نشان می‌دهد. همانطور که این شکل‌ها نشان می‌دهد، روش پیشنهادی در هر دو مجموعه داده فارسی و انگلیسی توانسته است با دقت خیلی خوبی، مقالات مرتبط با یک مقاله را بازیابی کند. از میان روش‌های مبتنی بر تحلیل محتوا، BERT توانسته است با دقت خیلی خوبی نسبت به دیگر روش‌ها مقالات مرتبط را بازیابی کند.



شکل ۴_۱ : مقایسه روش پیشنهادی با دیگر روش‌ها روی داده‌های انگلیسی



شکل ۴_۲ : مقایسه روش پیشنهادی با دیگر روش‌ها روی داده‌های فارسی

فصل پنجم

بحث و نتیجه‌گیری

۵. بحث و نتیجه‌گیری

۵_۱ مقدمه

در سال‌های اخیر حجم مقالات علمی به طور فزاینده‌ای افزایش پیدا کرده است. از طرف دیگر، مطالعه روش‌های گذشته و یافتن مقالات مرتبط از اولین قدم‌های ضروری در هر پژوهشی است که می‌بایست توسط پژوهشگران انجام شود. هدف سامانه‌های بازیابی مقالات علمی، کمک به پژوهشگران در این راستاست. یکی از قابلیت‌هایی که در سامانه‌های بازیابی اطلاعات مقالات علمی به پژوهشگران کمک بسیار زیادی می‌کند، ویژگی یافتن مقالات علمی مرتبط با یک مقاله است. به عبارت دیگر، ویژگی که به پژوهشگر اجازه می‌دهد با انتخاب یکی از مقالات بازیابی شده، دیگر مقالات مرتبط با آن را مشاهده نماید.

در این پژوهش، روشی مبتنی بر تحلیل مراجع برای یافتن مقالات مرتبط با یک مقاله ارائه شد. این روش بدین صورت عمل می‌کند که شباهت میان مراجع مقاله داده شده با مراجع دیگر مقالات محاسبه می‌شود و مقالاتی به عنوان مقاله مرتبط با یک مقاله بازیابی شده که عناوین آنها بیشترین شباهت را با یکدیگر داشته باشند. بنابراین، برنامه‌ای به زبان پایتون برای استخراج عنوان از هر منبع توسعه داده گردید که این منبع می‌تواند با فرمت‌های مختلف APA، Vancouver، MLA، Chicago و Harvard باشد. در نهایت، روش پیشنهادی با تحلیل مراجع توانست مقالات مرتبط با یک مقاله را بازیابی نماید.

۵_۲ نتیجه‌گیری

به منظور ارزیابی روش پیشنهادی، آن را با روش‌های مبتنی بر تحلیل محتوا مقایسه نمودیم که یافته‌ها نشاندهنده کارایی روش پیشنهادی می‌باشد. این روش‌ها مبتنی بر TF-IDF، word2vec، DOC2VEC و BERT است. برای نشریات انگلیسی، BERT نسبت به دیگر روش‌های تحلیل محتوا از دقت بالاتری برخوردار است که در

ادامه آن، روش‌های مبتنی بر word2vec، TF-IDF قرار دارند. روش مبتنی بر DOC2VEC کمترین دقت را داشته است. برای نشریات فارسی، BERT بیشترین دقت و word2vec کمترین دقت را دارد. در هر دو داده‌های فارسی و انگلیسی، روش پیشنهادی که در واقع روشی مبتنی بر تحلیل مراجع است، بیشترین دقت را به همراه داشته است.

روش پیشنهادی در مواردی هیچ مقاله‌ای را بازیابی نکرد. به عنوان مثال، روش پیشنهادی نتوانست برای مقاله زیر، هیچ مقاله مرتبطی را پیدا کند؛ زیرا از نظر مراجع با هیچ مقاله دیگری اشتراکی نداشت. البته روش‌های دیگر نیز با اینکه مقالاتی را بازیابی کردند ولی تنها یکی از آنها با این مقاله تا اندازه‌ای مرتبط بود.

Title	designing a mobile application for children: space science
Keywords	stem education, earth and space science, mobile application, educational game
Abstract	the incorporation of stem education into the curriculum has been an aspiring objective for many nations around the world. most students choose not to pursue stem-discipline studies because they are losing interest slowly. moreover, the level of engagement required for stem education is limited due to inadequate interactive teaching and tools that facilitate effective learning in a classroom setting. the objective of this project is to assess how educational game applications can help incline students' interest in science, develop an educational game application, and conduct user experience testing. a mobile application on earth and space science has been developed for 10 – 11year-old school students. the project is based on the rapid application development methodology considering the short development timeframe. the application was created using the ionic framework, angular 5, c#.net and sqlexpress. the findings indicated that this experiment motivates students to be more inclined to science.

در بعضی موارد نیز روش مبتنی بر تحلیل مراجع، تنها دو مقاله را مرتبط تشخیص داده و بازیابی نمود که آن دو نیز کاملا مرتبط بودند. در صورتی که روش‌های مبتنی بر تحلیل محتوا نتوانستند مقالات مرتبط را بازیابی کنند. از آنجا که این پژوهش از شباهت میان مراجع که به صورت دودویی انجام می‌شود، استفاده می‌کند، پیچیدگی محاسباتی بالایی دارد؛ ولی تمام این پیچیدگی‌ها و محاسبات مربوط به فاز آفلاین است. به عبارت دیگر در زمان آفلاین، به صورت دودویی شباهت میان مراجع هر مقاله با دیگر مقالات محاسبه شده و در پایگاه داده ذخیره می‌شود. در فاز آنلاین، زمانی که کاربر یکی از مقالات موجود را برای یافتن مقالات مرتبط انتخاب نمود، الگوریتم پیشنهادی از محاسباتی که در فاز آفلاین انجام گرفته، استفاده کرده و از میان مقالات موجود در پایگاه، مقاله مرتبط را بازیابی می‌کند.

به عنوان جمع‌بندی آخر، نتایج اعمال روش پیشنهادی روی داده‌های فارسی و انگلیسی نشریات انتخاب شده در علوم کامپیوتر نشان‌دهنده کارایی آن در یافتن مقالات مرتبط با یک مقاله است. به عبارت دیگر، در صورتی که پوشش نسبتا جامعی از استنادهای مقالات وجود داشته باشد، روش پیشنهادی قادر خواهد بود که با دقت بالایی مقالات مرتبط با یک مقاله را پیدا کند.

۵_۳ پیشنهادهای اجرایی پژوهش

الگوریتم ارائه شده می‌تواند در سامانه‌های بازیابی مقالات علمی که اطلاعات استنادی مقالات را داشته باشند، استفاده گردد و یک ویژگی ارزشمندی را به سامانه اضافه نماید که به کاربرپسندتر شدن سامانه بازیابی اطلاعات کمک شایانی می‌نماید.

۵_۴ پیشنهاد برای پژوهش‌های آتی

به عنوان پژوهش‌های آتی این طرح، پژوهشگر به دنبال ترکیب روش‌های تحلیل مراجع و تحلیل محتوا به منظور ارائه روشی قویتر برای یافتن مقالات مرتبط با یک مقاله است. همچنین، ایجاد یک گراف از مقالات بر اساس تحلیل مراجع آنها به منظور استخراج اطلاعات بیشتر از مقالات، از دیگر پژوهش‌های آتی این پژوهش است.

مراجع

مراجع

اسلامی‌نسب، معصومه؛ جاویدان، رضا (۱۳۹۴). ارائه روشی بر اساس شباهت کسینوسی و شبکه واژگان جهت پیدا کردن میزان شباهت معنایی بین متون. هفتمین کنفرانس بین‌المللی اطلاعات و دانش، دانشگاه ارومیه.

حسنی آهنگر؛ محمدرضا؛ امیری جزه؛ علی (۱۴۰۰). بهبود دقت واژگان کلیدی استخراج شده از متن فارسی با استفاده از الگوریتم،Word2Vec. پردازش علائم و داده‌ها، شماره ۱ پیاپی ۴۷.

سلیمانی‌نژاد؛ سلاجقه، مژده، طبیبی‌نیا، الهام (۱۳۹۷). خوشه‌بندی مقالات علمی بر پایه الگوریتم k-means، مطالعه موردی: پایگاه پژوهشگاه علوم و فناوری. پژوهشنامه پردازش و مدیریت اطلاعات. دوره: ۳۴، شماره ۲: ص ۸۷۱-۸۹۶.

عباسی، شیرین؛ وزیری، بابک (۱۳۹۴). الگوریتم‌های خوشه‌بندی در داده‌های عظیم، کنفرانس بین‌المللی پژوهش-های کاربردی در فناوری اطلاعات، کامپوتر و مخابرات. دانشگاه آزاد اسلامی واحد تربت حیدریه.

عسگریان، احسان؛ حبیبی، جعفر؛ معاون، شهروز؛ معین‌زاده، حسین (۱۳۸۶). روشی جدید برای خوشه‌بندی مستندات متنی بر اساس آنتولوژی. سومین کنفرانس فناوری اطلاعات و دانش. دانشگاه فردوسی مشهد.

Aggarwal, N., Asooja, K., & Buitelaar, P. (2012). DERI&UPM: Pushing corpus based relatedness to similarity: Shared task system description. *In * SEM 2012: 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 2012) (pp. 643-647).

Atoum, I. (2019). Scaled Pearson's correlation coefficient for evaluating text similarity measures. *Infinite Study*.

Bär, D., Biemann, C., Gurevych, I., & Zesch, T. (2012). Ukp: Computing semantic textual similarity by combining multiple content similarity measures. *In * SEM 2012: 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 2012) (pp. 435-440).

Barrón-Cedeno, A., Rosso, P., Agirre, E., & Labaka, G. (2010, August). Plagiarism detection across distant language pairs. *In Proceedings of the 23rd International Conference on Computational Linguistics* (Coling 2010) (pp. 37-45).

Bin, Y., Xiao-Ran, L., Ning, L., & Yue-Song, Y. (2012, November). Using information content to evaluate semantic similarity on HowNet. *In 2012 Eighth International Conference on Computational Intelligence and Security* (pp. 142-145). IEEE.

Cilibrasi, R. L., & Vitanyi, P. M. (2007). The google similarity distance. *IEEE Transactions on knowledge and data engineering*, 19(3), 370-383.

Baker, K. (2005). Singular value decomposition tutorial. *The Ohio State University*, *24*.

Buscaldi, D., Tournier, R., Aussenac-Gilles, N., & Mothe, J. (2012). Irit: Textual similarity combining conceptual similarity with an n-gram comparison method. I*n * SEM 2012: 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 2012) (pp. 552-556).

Church, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155-162.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297-302.

Gabrilovich, E., & Markovitch, S. (2007, January). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *In IJcAI* (Vol. 7, pp. 1606-1611).

Gomaa, W. H., & Fahmy, A. A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, 68(13), 13-18.

Islam, A., & Inkpen, D. (2006, May). Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words. *In LREC* (pp. 1033-1038).

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data* (TKDD), 2(2), 1-25.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7), 491-498.

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.

Kenter, T., & De Rijke, M. (2015, October). Short text similarity with word embeddings. In *Proceedings of the 24th ACM international on conference on information and knowledge management* (pp. 1411-1420).

Kolb, P. (2009, May). Experiments on the difference between semantic similarity and relatedness. *In Proceedings of the 17th Nordic Conference of Computational Linguistics* (NODALIDA 2009) (pp. 81-88).

Kumar, D., Kumar, A., Singh, M., Patel, A., & Jain, S. (2018, November). Modern WordNet: An Affective Extension of WordNet. *In International Conference On Computational Vision and Bio Inspired Computing* (pp. 527-536). Springer, Cham.

Lakshmi, R., & Baskar, S. (2021). Efficient text document clustering with new similarity measures. *International Journal of Business Intelligence and Data Mining*, 18(1), 49-72.

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), 211.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. WordNet: An electronic lexical database, 49(2), 265-283.

Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE transactions on knowledge and data engineering*, 18(8), 1138-1150.

Lin, D. (1998, August). Extracting collocations from text corpora. *In First workshop on computational terminology* (pp. 57-63).

Little, C., Mclean, D., Crockett, K., & Edmonds, B. (2020). A semantic and syntactic similarity measure for political tweets. *IEEE Access*, 8, 154095-154113.

Liu, M., Lang, B., & Gu, Z. (2017) (a). Calculating semantic similarity between academic articles using topic event and ontology. *arXiv preprint arXiv*:1711.11508.

Liu, M., Lang, B., Gu, Z., & Zeeshan, A. (2017) (b). Measuring similarity of academic articles with semantic profile and joint word embedding. *Tsinghua Science and Technology*, 22(6), 619-632.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2), 203-208.

Martinez-Gil, J., & Pichler, M. (2014, September). Analysis of word co-occurrence in human literature for supporting semantic correspondence discovery. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business* (pp. 1-7).

Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005). Generalized latent semantic analysis for term representation. *In Proc. of RANLP* (p. 149).

Mihalcea, R., Corley, C., & Strapparava, C. (2006, July). Corpus-based and knowledge-based measures of text semantic similarity. *In Aaai* (Vol. 6, No. 2006, pp. 775-780).

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3), 443-453.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013, March). Using of Jaccard coefficient for keywords similarity. *In Proceedings of the international multiconference of engineers and computer scientists* (Vol. 1, No. 6, pp. 380-384).

Patwardhan, S., Banerjee, S., & Pedersen, T. (2003, February). Using measures of semantic relatedness for word sense disambiguation. *In International conference on intelligent text processing and computational linguistics* (pp. 241-257). Springer, Berlin, Heidelberg.

Potthast, M., Stein, B., & Anderka, M. (2008, March). A wikipedia-based multilingual retrieval model. *In European conference on information retrieval* (pp. 522-530). Springer, Berlin, Heidelberg.

Qurashi, A. W., Holmes, V., & Johnson, A. P. (2020, August). Document Processing: Methods for Semantic Text Similarity Analysis. *In 2020 International Conference on INnovations in Intelligent SysTems and Applications* (INISTA) (pp. 1-6). IEEE.

Reynolds, B. E. (1980). Taxicab geometry. *Pi Mu Epsilon Journal*, 7(2), 77-88.

Singh, R., & Singh, S. (2021). Text Similarity Measures in News Articles by Vector Space Model Using NLP. *Journal of The Institution of Engineers* (India): Series B, 102(2), 329-338.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1), 195-197.

Terra, E. L., & Clarke, C. L. (2003). Frequency estimates for statistical word similarity measures. In *Proceedings of the 2003 human language technology conference of the North American Chapter of the Association for Computational Linguistics* (pp. 244-251).

Turney, P. D. (2001, September). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. *In European conference on machine learning* (pp. 491-502). Springer, Berlin, Heidelberg.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.

Wu, Z., & Palmer, M. (1994). Verb semantics and lexical selection. arXiv preprint cmp-lg/9406033.

Proposing an Algorithm to Find Related Academic Articles to an Article

Niloofar Mozafari

Abstract

Background and Purpose
The emergence of the Internet has caused the volume of scientific documents and articles has increased dramatically in the last decade. It makes too difficult to find the relevant documents based on the user's query. The scientific information retrieval systems help researchers to find relevant scientific articles. One of the capabilities that help researchers to find the relevant papers is the feature of finding related scientific articles to an article. In other words, this feature allows the researcher to retrieve other related articles by selecting one article.
Methodology
In this research, an algorithm based on references analysis has been proposed to find scientific articles related to an article. The research population includes Persian and English data extracted from computer science publications. The proposed method works in such a way that it calculates the similarity between the references of the given article and the references of other articles. The presented method is capable of extracting the titles of articles in different formats APA, Vancouver, MLA, Chicago and Harvard formats. Finally, a similarity graph is extracted based on the similarity between the titles of the references.
Findings
In order to evaluate the proposed method more accurately, 30 articles were randomly selected for English data taken from English publications. The same procedure was done for Persian data and 30 articles were randomly selected from the collection of articles. For each article, bibliographic information containing the title, keyword and abstract of the articles along with the list of references have been extracted.
Conclusion
The results of applying the proposed method on Farsi and English data of selected publications in computer science show its effectiveness in finding articles related to an article. In other words, if there is a relatively comprehensive coverage of the citations of scientific articles, the proposed method will be able to find articles related to an article with high accuracy. The proposed algorithm can be used in scientific article retrieval systems that have the citation information of the articles, and it makes the information retrieval system more user-friendly.

Keywords: bibliographic references analysis, information retrieval, similarity measure, accuracy.

مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری

Regional Information Center for Science and Technology RICeST

Proposing an Algorithm to Find Related Academic Articles to an Article

By

Dr. Niloofar Mozafari

August 2021