استفاده از الگوریتم‌های یادگیری ماشین در استخراج مشابهت علمی کشورها

سیده فاطمه نورانی[*1]
استادیار، گروه کامپیوتر، دانشکده فنی و مهندسی دانشگاه پیام نور، تهران، ایران

رعنا نقدی
کارشناسی ارشد، گروه کامپیوتر، دانشکده فنی و مهندسی، دانشگاه پیام نور، تهران، ایران

مدیریت اطلاعات
دوره ۹، شماره ۲
پاییز و زمستان ۱۴۰۲

چکیده: امروزه، تولید علم در تمام کشورها، اولویت مهمی شناخته شده است؛ زیرا توسعهٔ علمی پایه‌ای برای توسعهٔ فناوری است و توسعهٔ فناوری نیز، اساس رشد اقتصادی و رفاه اجتماعی است. به همین دلیل، سنجش سطح کمی و کیفی تولیدات علمی جوامع، بسیار اهمیت دارد. علم‌سنجی و کتاب‌سنجی ابزارهایی هستند که برای اندازه‌گیری و ارزیابی تولیدات علمی در جوامع استفاده می‌شوند. این نوع مطالعات و بررسی‌ها، در زمینه‌های مختلف آموزشی و پژوهشی یا به‌منظور تصمیم‌گیری، سیاست‌گذاری و آینده‌نگری در مؤسسه‌ها و سازمان‌ها کاربردهای وسیعی دارند. در این زمینه، یکی از ابزارهای کاربردی پایگاه اطلاعاتی سایمگو است که داده‌های ارزشمندی، از جمله عملکرد علمی کشورهای دنیا را در حوزه‌های علمی مختلف فراهم می‌کند و به‌عنوان منبع اطلاعاتی مناسب برای انجام چنین تحقیقاتی استفاده می‌شود. در این مقاله، شباهت‌های علمی کشورها و حوزه‌های علمی آن‌ها با ایران، در بازهٔ زمانی مشخص و برمبنای دو شاخص کتاب‌سنجی، یعنی تعداد مستندات و شاخص هرش شناسایی شده است. در ادامه با استفاده از شباهت به‌دست‌آمده و به‌کارگیری الگوریتم‌های تشخیص جوامع لووین و لیدن، به خوشه‌بندی و در نتیجه ارائهٔ میزان مشابهت علمی کشورها و حوزه‌های علمی پرداخته شده است. میزان مشابهت گزارش‌شده در این مقاله نشان می‌دهد که کشورهایی با ضریب هم‌بستگی بیش از ۰/۹ با ایران، در روند تولید علمی (از نظر تعداد مستندات و شاخص هرش) شباهت بسیار زیادی با این کشور دارند. در این پژوهش، در الگوریتم لیدن مقدار سیلوئت بهتر نشد؛ اما با اختلاف کمی، تغییری در بحث ماژولاریتی به‌وجود آمد. گفتنی است تغییر ایجاد شده به‌دلیل ماهیت این الگوریتم است که براساس ماژولاریتی کار می‌کند و زمان اجرا، الگوریتم لیدن به‌طور محسوسی بهتر از الگوریتم لووین است.

کلیدواژه‌ها: یادگیری ماشین، مشابهت علمی، داده‌کاوی، خوشه‌بندی

---

۱. نویسنده مسئول: sf.noorani@pnu.ac.ir

مقدمه

علم‌سنجی[1] یکی از حوزه‌های مطالعاتی است که به بررسی و تحلیل عملکرد علمی و پژوهشی افراد، سازمان‌ها، کشورها و مجامع علمی می‌پردازد. در این حوزه، از شاخص‌های متنوعی مانند تعداد استنادها[2]، ضریب تأثیر[3]، شاخص هرش[4] و شاخص SJR[5] برای ارزیابی و نمایش اثربخشی فعالیت‌های علمی دانشمندان در زمینه‌های مختلف استفاده می‌شود (Khokhlov, 2020; Roldan-Valadez, Salazar-Ruiz, Ibarra-Contreras & Rios, 2019).

در عصر حاضر، علم‌سنجی معیاری برای توسعه‌یافتگی شناخته می‌شود؛ چراکه تولید علم نمایانگر سطح پیشرفت و توسعهٔ جامعه است. این شاخص‌ها، علاوه‌بر نمایش دستاوردهای علمی، تصویری از توانایی‌های یک جامعه در پیشبرد علم و دانش ارائه می‌دهند و در تعیین جایگاه کشورها در جهان نقشی کلیدی ایفا می‌کنند.

در هر کشور و جامعه‌ای شناخت و ارزیابی پژوهش‌های انجام‌شده، نه‌تنها موضوع مورد توجه محققان هر رشته است؛ بلکه امری مهم و ضروری برای برنامه‌ریزان و سیاست‌گذاران پژوهشی آن کشور به‌شمار می‌آید (فرزین یزدی و رضایی شریف‌آبادی، ۱۳۹۶). از این رو سازمان‌ها و مؤسسه‌ها با آگاهی از موقعیت علمی و پژوهشی خود و رقبا، ضمن تشخیص نقاط قوت و ضعف، می‌توانند جهت ارتقا و توسعه برنامه‌ریزی کنند؛ بنابراین با رشد روزافزون دانش و افزایش رقابت، سنجش محصولات علمی با استفاده از شیوه‌های علم‌سنجی، به موضوعی مهم و پراهمیت تبدیل شده است (احقاقی و فتحیان، ۱۴۰۰). شاخص‌های علم‌سنجی ابزارهای مؤثری برای سیاست‌گذاران فراهم می‌کنند تا تصمیمات آگاهانه‌تری اتخاذ کنند. از طریق این شاخص‌ها، امکان دستیابی به پیشرفت علمی و توسعهٔ اجتماعی ـ اقتصادی پایدار تسهیل می‌شود (Dikusar & Cujba, 2024). این شاخص‌ها نه‌تنها به پژوهشگران، بلکه به سازمان‌های علمی و نهادهای تأمین بودجه کمک می‌کنند تا تصمیمات بهتری در تخصیص منابع و اولویت‌بندی‌های پژوهشی اتخاذ کنند (Wani & Zainab, 2017).

علم‌سنجی همچنین ابزاری کلیدی برای شناسایی اولویت‌های پژوهشی و پیش‌بینی روندهای آینده، به‌ویژه در زمینه همکاری‌های بین‌المللی است (Sallam et al., 2024).

یکی از روش‌های نوین علم‌سنجی، استفاده از ابزارهای داده‌کاوی[6] است. داده‌کاوی به‌معنای استخراج اطلاعات و الگوهای مفید از حجم وسیعی از داده‌ها و کشف روندها و پیش‌بینی‌هاست. در علم‌سنجی، این روش می‌تواند برای تحلیل داده‌های علمی مانند تعداد مقالات، ارجاعات، همکاری‌های بین‌المللی و دیگر شاخص‌های علمی استفاده شود (Wang, Long, Zeng, Chen & Yishan, 2024).

1. Bibliometric
2. Citation
3. Impact Factor
4. H-Index
5. https://Scimagojr.com
6. Data Mining

در این خصوص، پژوهش‌هایی با استفاده از ابزارهای داده‌کاوی، انجام شده است. برای مثال، در پژوهش نوروزی چالکی، نوروزی چالکی و چهره‌نگار[1] (۲۰۲۳) رابطه میان شاخص‌های فرهنگی، اقتصادی و تحقیق و توسعه با جایگاه و تأثیر تولیدات علمی کشورهای هند، ترکیه، ایران، عربستان سعودی و پاکستان در آسیای مرکزی و غربی بررسی شده است. دورهٔ زمانی پژوهش از ۲۰۰۱ تا ۲۰۲۰ بوده و داده‌های لازم از منابعی چون یونسکو، سایمگو و JCR[2] جمع‌آوری شده است. تحلیل داده‌ها با استفاده از نرم‌افزارهای متلب، اکسل و الگوریتم شبکه عصبی پیش‌خور[3] انجام شده است. نتایج نشان می‌دهد که وضعیت آموزشی، اقتصادی و تحقیق و توسعه، بر جایگاه علمی این کشورها در منطقه تأثیر زیادی دارد.

هدف پژوهش حاضر، بهره‌گیری از تکنیک‌های تشخیص جامعه[4] و خوشه‌بندی[5] در حوزهٔ داده‌کاوی با هدف یافتن مشابهت‌های علمی کشورها، برحسب معیارهای مختلف در داده‌های پایگاه اطلاعاتی سایمگو و تحلیل این روابط است. خوشه‌بندی **و** تشخیص جامعه، از جمله تکنیک‌های داده‌کاوی برای تحلیل ساختارهای پیچیده داده‌ها هستند. الگوریتم‌های تشخیص جامعه، به عنوان ابزاری مؤثر در تحلیل شبکه‌های اجتماعی و علمی، می‌توانند نقش مهمی در شناسایی ساختارهای ارتباطی و همکاری‌های علمی ایفا کنند. این الگوریتم‌ها با تجزیه‌وتحلیل داده‌های علمی و شبکه‌های ارتباطی، به شناسایی جوامع علمی و الگوهای همکاری بین محققان و مؤسسه‌های مختلف کمک می‌کنند. استفاده از این الگوریتم‌ها در تحقیقات علمی ایران می‌تواند به کشف ارتباطات پنهان و بهینه‌سازی همکاری‌های علمی با دیگر کشورها منجر شود.

اهمیت این پژوهش این است که با بهره‌گیری از تکنیک‌های مدرن تحلیل داده، می‌توان نقشه‌ای جامع از وضعیت ارتباطات علمی ایران با دیگر کشورها ترسیم کرد. این به نهادهای علمی و پژوهشگران کمک می‌کند تا نقاط قوت و ضعف همکاری‌های علمی را شناسایی و بهبودهای لازم را اعمال کنند. در واقع این پژوهش می‌تواند پایه‌ای برای تحقیقات آینده و توسعهٔ استراتژی‌های علمی و فناوری در ایران باشد.

در این پژوهش، به‌صورت خاص، از الگوریتم‌های تشخیص جامعه لووین[6] و لیدن[7] برای علم‌سنجی استفاده شده است. این روش‌ها با شناسایی جوامع و گروه‌های مرتبط در شبکه‌های علمی، به تحلیل دقیق‌تر روابط و ساختارهای دانش کمک می‌کنند. دیتاست به‌کار گرفته شده در مقاله حاضر، از سایت سایمگو[8] استخراج شده است. این سایت پایگاهی است که از مقالات علمی، ارجاعات، همکاری‌های بین‌المللی و رتبه‌بندی دانشگاه‌ها، اطلاعات گسترده‌ای ارائه می‌دهد و با پوشش متنوع موضوعات علمی، ابزار ارزشمندی برای تحلیل دقیق روندهای علمی و ساختارهای دانش محسوب می‌شود. اطلاعات سایمگو، سهم ارزشمندی در رتبه‌بندی کیفی مجلات دارد و براساس داده‌های اسکوپوس توسعه یافته

1. Noroozi Chakoli, Noroozi Chakoli & Chehrenegar
2. Journal Citation Reports
3. Feed Forward Neural Network
4. Community Detection
5. Clustering
6. Louvain
7. Leiden
8. https://www.scimago.com

است. سایمگو رتبه‌بندی را در دو سطح رتبه‌بندی نشریه‌ها و نیز رتبه‌بندی کشورها در ۲۷ حوزۀ موضوعی و ۳۰۹ زیرشاخه موضوعی ارائه می‌دهد.

در این مقاله، منظور از مشابهت علمی کشورها، میزان شباهت آن‌ها بر اساس تعداد اسناد تولیدی (شامل مقاله‌ها و کتاب‌ها) و شاخص هرش است. داده‌های استفاده شده، اطلاعات علمی ۲۴۰ کشور از سال ۱۹۹۶ تا سال ۲۰۲۱ است که از پایگاه اطلاعاتی سایمگو استخراج شده است.

در ادامه، ابتدا پیشینه پژوهش و سپس روش پژوهش بررسی شده است. در ادامه به ارزشیابی و تحلیل روش پیشنهادی پرداخته می‌شود و پس از آن نتیجه‌گیری و پیشنهادها ارائه خواهد شد.

پیشینه پژوهش

عرفان‌منش، جهرمی، حسینی و غلامحسین‌زاده[1] (۲۰۱۳) در پژوهشی با هدف ارزیابی بهره‌وری و تولیدات علمی پژوهشی ۱۰ کشور برتر آسیا (چین، ژاپن، هند، کره جنوبی، تایوان، هنگ‌کنگ، سنگاپور، ایران، تایلند و مالزی) در سال‌های ۱۹۹۶ تا ۲۰۱۰، به جست‌وجو در پایگاه اطلاعاتی اسکوپوس[2] پرداختند. برای این منظور، از روش آمار توصیفی استفاده شده است. داده‌ها با نرم‌افزار اکسل تجزیه‌وتحلیل شدند. همچنین، از نرم‌افزار یوسی‌آی نت[3] برای تجسم شبکه همکاری کشورها استفاده شد.

ریس[4] (۲۰۱۴) به پژوهش در مورد رشد مجلات نمایه شده کشورهای امریکا لاتین و کارائیب طی سال‌های ۲۰۰۶ تا ۲۰۰۹ پرداخت و نشان داد که رشد مجلات نمایه شده کشورهای مطرح‌شده، قبل از آنکه تحت تأثیر تحولات جامعه علمی باشد، از سیاست‌های نمایه‌سازی پایگاه داده ISI تأثیر پذیرفته ا ست. در این پژوهش، از دو روش آماری برازش‌های خطی و ضریب هم‌بستگی برای تجزیه‌وتحلیل داده‌ها استفاده شده است.

محمد اسماعیل، ریاحی و صحبتی‌ها (۱۳۹۳) با بررسی پارامترهای کمی و کیفی مجلات علمی ایران در پایگاه داده اسکوپوس، به این نتیجه رسیدند که بیشترین رشد مجلات مربوطه به حوزه پزشکی بوده و پارامتر کیفیت نسبت به کمیت، وضعیت قابل تعریفی نداشته است. ابازری، ریاحی، صحبتی‌ها، صیامیان و یمین فیروز (۱۳۹۹) نشان دادند که در حوزۀ پزشکی، ایران در قیاس با سایر کشورهای منطقه‌ای مدیترانه شرقی، مجلات عملی نمایه شده بیشتری را انتشار داده است؛ اما به‌لحاظ کیفیت، به غیر از ایران و امارات مجلات سایر کشورها، کیفیت مطلوبی نداشته‌اند، در این پژوهش، از نرم‌افزار آماری اکسل برای تجزیه‌وتحلیل اطلاعات و از نرم‌افزار نود ایکس‌ال[5] برای ترسیم گراف‌ها استفاده شده است.

در خصوص کیفیت مجلات در حوزه کشاورزی نیز پژوهش‌هایی انجام گرفته است. از آن جمله پژوهش وینارکو، آبریزه و طاهره[6] (۲۰۱۶) که نشان می‌دهد شاخص انتشار مجلات در پایگاه‌های

1. Erfanmanesh, Jahromi, Hosseini & Gholamhosseinzadeh
2. Scopus
3. UCINet
4. Reyes
5. NodeXL
6. Winarko, Abrizah & Tahira

اسکوپوس به زبان انگلیسی در مقایسه با سایر زبان‌های دیگر موفق‌تر بوده است، در این پژوهش روش برازش و تحلیل هم‌بستگی و رگرسیون به کار برده شد.

احمدیان دیوکتی، رازقی و آقاجانی (۱۳۹۹) نشان دادند که رشد تولیدات علمی کشور ایران با وجود محدودیت‌های دو دهه اخیر، بازهم چشمگیر می‌باشد. همچنین در پژوهش گزارش شده در سطح منطقه داشته است بازهم چشمگیر می‌باشد، با استفاده از نرم‌افزار ایویوز و مدل آریما به پیش‌بینی روند تولیدات علمی کشور تا سال ۲۰۳۰ پرداخته شد.

جنوی، مرادی و پاکزاد (۱۳۹۹)، نشان داده شده که ایران در شاخص‌های کمی مانند «تعداد مقالات در هر یک‌میلیون نفر از جمعیت»، «نسبت فارغ‌التحصیلان آکادمیک و حوزه مقالات نمایه‌سازی شده بین‌المللی» و همچنین «رابطه مقالات نمایه‌سازی شده بین‌المللی به تعداد اعضای هیئت‌علمی»، دارای رشد صعودی بوده و امکان پژوهش این شاخص‌ها با توجه به کمیت مطلوبشان برای سال ۱۴۰۴دور از انتظار نیست. ولی در مورد شاخص‌های کیفی مانند «میزان استنادات در واحد انتشارات»، با وجود سیر صعودی در سالیان اخیر، مقدار به‌دست‌آمده با میزان هدف‌گذاری برای این شاخص در سال ۱۴۰۴فاصله زیادی دارد، در این پژوهش از روش توصیفی ـ تحلیلی و از آزمون‌های شاپیرو ـ ویلک[1]، کروسکال ـ والیس[2] و من ـ ویتنی[3] استفاده شد.

در این راستا در پژوهش جنوی و شاهمرادی (۱۳۹۹) گزارش شده است که به تعیین جایگاه رقابت‌پذیری ایران در منطقه و شناسایی و رتبه‌بندی حوزه‌های علمی با استفاده از شاخص پیچیدگی علمی انجام‌شده، می‌پردازد. در این پژوهش نشان داده شده است که ایران رتبهٔ چهلم جهانی و رتبهٔ هشتم منطقه‌ای را به‌لحاظ شاخص پیچیدگی علمی به خود اختصاص داده است و ترکیه مهم‌ترین رقیب تولیدات علمی بعدازآن کشورهای عربستان، پاکستان، مصر و اردن به‌ترتیب دیگر رقبای علمی ایران در منطقه قلمداد می‌شوند، این مقاله با روش برازش و الگوریتم تکرارشونده انجام شده است.

آزادی احمدآبادی (۱۴۰۱) وضعیت موجود تولید علم تحلیل و سیاست‌گذاری‌های مناسب جهت ارتقای سطح کمّی و کیفی به‌منظور افزایش سطح تولیدات علمی کشور با استفاده از نرم‌افزار اکسل و اس‌پی‌اس‌اس بیان شده است.

درادکه، ابوعلیگه، عطالله و منصور[4] (۲۰۲۲) از شبکهٔ عصبی کانولوشنال، برای تجزیه‌وتحلیل علم‌سنجی و طبقه‌بندی تحقیقات استفاده کردند. نتایج مقاله نشان می‌دهد که شبکهٔ عصبی نسبت به KNN، NBM و SVM برای این کار دقت بیشتری دارد. در این پژوهش، از روش پایتون و NLP استفاده شده است.

هدف مقالهٔ حاضر، یافتن مشابهت علمی کشورها و حوزه‌های علمی در بازه زمانی مشخص برمبنای دو شاخص کتاب‌سنجی، یعنی تعداد مستندات و شاخص هرش است؛ همچنین با استفاده از شباهت

1. Shapiro - Wilk
2. Kruskal-Wallis
3. Mann-Whitney
4. Daradkeh, Abualigah, Atalla & Mansoor

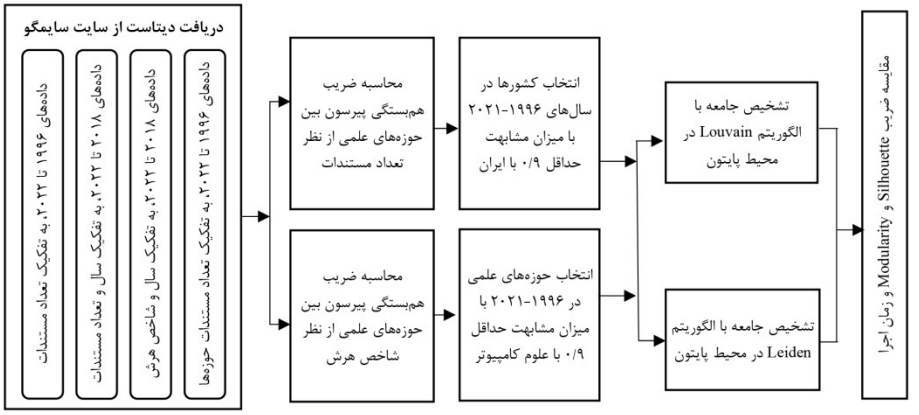
به‌دست‌آمده و با به‌کارگیری الگوریتم‌های تشخیص جوامع لووین و لیدن، به خوشه‌بندی آن‌ها پرداخته خواهد شد. نوآوری پژوهش حاضر را می‌توان در موارد زیر خلاصه کرد:

- در مقایسه با مقاله درادکه و همکاران (۲۰۲۲)، نوآوری این پژوهش در استفاده از الگوریتم‌های تشخیص جامعه است که جزء دسته الگوریتم‌های خوشه‌بندی است و به برچسب کلاس‌های از قبل تعیین‌شده نیاز ندارد. در مقاله مذکور از تکنیک‌های طبقه‌بندی[1] استفاده شده است. در طبقه‌بندی برچسب کلاس‌ها باید از قبل مشخص باشند.
- در مقایسه با مقالهٔ جنوی و شاهمرادی (۱۳۹۹) به‌کارگیری الگوریتم لیدن نوآوری پژوهش حاضر محسوب می‌شود. آن‌ها به‌منظور پیدا کردن مشابهت علمی، از روش برازش و الگوریتم تکرارشونده بر مبنای پایگاه اطلاعاتی سایمگو استفاده کرده‌اند.
- مقالهٔ آزادی احمدآبادی (۱۴۰۱) تنها رشد کیفی و کمی برون‌داده‌های علمی جمهوری اسلامی ایران را مبنای پژوهش خود قرار داده است؛ ولی در پژوهش حاضر، تمامی کشورهای پایگاه اطلاعاتی سایمگو، به‌منظور آزمون فرضیه‌های مطرح‌شده بررسی شده است. همچنین در پژوهش آزادی احمدآبادی (۱۴۰۱) از نرم‌افزارهای اکسل و اس‌پی‌اس‌اس استفاده شده است؛ اما در پژوهش حاضر از الگوریتم لیدن در محیط پایتون استفاده خواهد شد.
- در پژوهش خرمی (۱۳۹۶) داده‌های پایگاه سایمگو بررسی شده و از الگوریتم لووین به‌عنوان ابزار و از نرم‌افزار گفی[2] برای تجزیه‌وتحلیل استفاده شده است. در پژوهش حاضر، نظر به قدرت پایتون، از این زبان و الگوریتم لیدن استفاده خواهد شد.

روش‌شناسی پژوهش

در این مقاله، با استفاده از تکنیک داده‌کاوی، شناسایی و استخراج مشابهت علمی بین کشور ایران و سایر کشورها انجام شده است. برای این منظور، همان طور که در شکل ۱ نمایش داده شده است، ابتدا، داده‌های مورد نیاز از پایگاه سایمگو استخراج و پس از آن، ضریب هم‌بستگی پیرسون[3] بین ایران و سایر کشورها از نظر تعداد مستندات و شاخص هرش محاسبه می‌شود. این کار در محیط اس‌پی‌اس‌اس انجام شده است. در ادامه کشورهایی که با ایران مشابهت حداقل ۰/۹ دارند، برای خوشه‌بندی و تشخیص جامعه انتخاب و با استفاده از محیط پایتون و الگوریتم‌های لووین و لیدن تشخیص جامعه انجام می‌شود. در نهایت جوامع (خوشه‌های مشابه) با استفاده از ضریب سیلوئت و محاسبهٔ پیمانگی و زمان اجرا ارزیابی می‌شوند.

1. Classification
2. Gephi
3. Pearson Correlation Coefficient



شکل ۱. مراحل اجرای طرح پیشنهادی در مقاله

دیتاست

مجموعه داده‌های استفاده شده در این پژوهش، از پایگاه اطلاعاتی سایمگو استخراج شده است. سایمگو، شامل اطلاعات علمی ۲۴۰ کشور در ۲۷ حوزهٔ علمی مانند کشاورزی، علوم کامپیوتر، دام‌پزشکی و... است. این اطلاعات برمبنای شاخص تعداد سند و شاخص هرش موجود است. شاخص هرش به تعداد مقاله‌های[1] کشوری اشاره دارد که هر یک، حداقل تعداد بار مشخصی (h) از آن‌ها نقل‌قول دریافت کرده‌اند. درواقع، به تعداد مقاله‌هایی اشاره دارد که در یک کشور خاص منتشر شده و توسط دیگران، حداقل همان تعداد بار نقل‌قول شده است. با استفاده از این شاخص، می‌توان بررسی کرد که در کشورهای مختلف، چند مقاله با تعداد نقل‌قول‌های حداقلی وجود دارد و این امر می‌تواند به‌عنوان معیار برای اندازه‌گیری تأثیر و اعتبار پژوهشگران و نویسندگان در آن کشورها استفاده شود.

پایگاه اطلاعاتی سایمگو، علاوه‌بر رتبه‌بندی نشریه‌ها، رتبه‌بندی کشورها را نیز برحسب مستندات و شاخص هرش نمایش می‌دهد.

در پژوهش حاضر، داده‌ها از بخش رتبه‌بندی کشورها جمع‌آوری شده است. همان طور که در شکل ۱ نشان داده شده است، در این پژوهش، چهار مجموعه داده متفاوت استخراج شده است.

۱. داده‌های سال‌های ۱۹۹۶ تا ۲۰۲۱ بر اساس تعداد مستندات علمی کشورها (شامل ۲۹۴۰۳ رکورد): با تحلیل این داده‌ها، مشابهت علمی ۲۴۰ کشور در ۲۵ سال، بر پایه مستندات علمی آن‌ها شناسایی می‌شود.

۲. داده‌های سال‌های ۱۹۹۶ تا ۲۰۲۱ بر اساس تعداد مستندات، به‌تفکیک حوزه‌های مختلف علمی (شامل ۷۲۹ رکورد): با استفاده از این سری از داده‌ها، مشابهت علمی کشورها در ۲۵ سال و به‌تفکیک حوزه‌های مختلف علمی بررسی می‌شود.

1. Article

۳. داده‌های سال‌های ۲۰۱۸ تا ۲۰۲۲ به‌تفکیک سال، بر اساس تعداد مستندات علمی (شامل ۳۲۴۰رکورد): با این سری از داده‌ها، میزان شباهت ۳۰ کشور اولِ سایمگو، با ایران بر اساس تعداد مستندات تولیدشده و به تفکیک سال بررسی می‌شود.

۴. داده‌های سال‌های ۲۰۱۸ تا ۲۰۲۲ به تفکیک سال، بر اساس شاخص هرش (شامل ۸۳۷ رکورد): با استفاده از این داده‌ها، میزان شباهت ۳۰ کشور اولِ سایمگو با ایران براساس شاخص هرش و به‌تفکیک سال مشخص می‌شود.

ضریب هم‌بستگی پیرسون

به‌منظور پیدا کردن مشابهت علمی کشورها، پیش از خوشه‌بندی، از ضریب هم‌بستگی پیرسون استفاده شده است. این ضریب مقداری بین ۱ تا ۱- است که مقدار ۱ به‌معنای هم‌بستگی مثبت، مقدار ۰ به‌معنای عدم هم‌بستگی و مقدار ۱- به‌معنای هم‌بستگی منفی است. هم‌بستگی نشان می‌دهد که چقدر تغییرات یک متغیر (مثل تعداد مستندات علمی ایران) با تغییرات متغیر دیگر (مثل تعداد مستندات علمی کشور دیگری) هماهنگ است (حسن‌زاده و مداح، ۱۴۰۲) برای مثال، اگر هر سال تعداد مستندات علمی ایران و کشور X هم‌زمان افزایش یابد، هم‌بستگی مثبت است، اگر یکی افزایش و دیگری کاهش یابد، هم‌بستگی منفی است و اگر هیچ الگویی بین تغییراتشان نباشد، هم‌بستگی صفر است.

رابطهٔ ۱ نحوهٔ محاسبه ضریب هم‌بستگی را نشان می‌دهد. در این رابطه $cov(X.Y)$cov، کواریانس بین دو متغیر x, y است و نشان‌دهندهٔ رابطهٔ خطی بین دو متغیر است. $\sigma_{X\mathrm{x}}$ و $\sigma_{Y\mathrm{y}}$ به‌ترتیب انحراف معیار متغیر x و y است. $\mu_{X\mathrm{x}}$ و μ_{Y}y میانگین دو متغیر و E امید ریاضی است.

$$P_{X.Y} = \frac{cov(X.Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$ رابطهٔ ۱)

در این مقاله پیش از استفاده از الگوریتم‌های تشخیص جامعه، ضریب هم‌بستگی بین شاخص تعداد مستندات و شاخص هرش کشورها محاسبه می‌شود. شکل ۲ نمونه‌ای از این مقادیر را نشان می‌دهد.

	A	B	C	D	E	F	G	H
1		چین	ایالات متحده	هند	انگلستان	آلمان	ایتالیا	ژاپن
2	چین	1	0.673534	0.970815	0.681098	0.849916	0.793677	0.863684
3	ایالات متحده	0.673534	1	0.65127	0.991048	0.951519	0.975706	0.922232
4	هند	0.970815	0.65127	1	0.66886	0.819501	0.773539	0.822984
5	انگلستان	0.681098	0.991048	0.66886	1	0.949033	0.968617	0.903159
6	آلمان	0.849916	0.951519	0.819501	0.949033	1	0.987899	0.98254
7	ایتالیا	0.793677	0.975706	0.773539	0.968617	0.987899	1	0.969661
8	ژاپن	0.863684	0.922232	0.822984	0.903159	0.98254	0.969661	1
9	کانادا	0.7221	0.992691	0.705016	0.991253	0.961248	0.982134	0.927048
10	استرالیا	0.671491	0.984918	0.654032	0.991364	0.930584	0.955742	0.887654
11	فرانسه	0.84197	0.941767	0.810717	0.934482	0.993694	0.985962	0.982028
12	اسپانیا	0.723017	0.974418	0.701783	0.988298	0.956142	0.966709	0.907022
13	فدراسیون روسیه	0.888829	0.65533	0.852704	0.69503	0.830895	0.755711	0.805455
14	کره جنوبی	0.967293	0.809609	0.936053	0.808626	0.928382	0.890993	0.941984
15	برزیل	0.710365	0.980064	0.675838	0.974221	0.947118	0.972041	0.925663
16	ایران	0.915834	0.878942	0.872882	0.872651	0.949768	0.935651	0.960463
17	هلند	0.634156	0.996183	0.614806	0.991576	0.934798	0.96466	0.897744
18	ترکیه	0.812454	0.961991	0.794984	0.960401	0.970987	0.978326	0.960765
19	ربستان صعودی	0.96971	0.704324	0.946901	0.705708	0.86611	0.814135	0.878613

شکل ۲. بخشی از داده‌های سال‌های ۲۰۱۸ تا ۲۰۲۲ به‌تفکیک سال و بر اساس تعداد مستندات علمی

الگوریتم‌های تشخیص جامعه

خوشه‌بندی یکی از تکنیک‌های مهم داده‌کاوی است که به تقسیم داده‌ها به گروه‌های مشابه یا خوشه‌ها بر اساس شباهت و ویژگی مشترک آن‌ها می‌پردازد. هدف اصلی خوشه‌بندی، ایجاد گروه‌هایی است که داده‌های داخل هر خوشه به یکدیگر شباهت بیشتر و با داده‌های خارج از خوشه‌ها شباهت کمتری داشته باشند (نصرتی و رحمانی، ۱۴۰۱)

تشخیص جامعه، مدلی از خوشه‌بندی است که با داده‌های از نوع گراف کار می‌کند. تشخیص جامعه در شبکه‌های اجتماعی به‌معنای شناسایی و تحلیل گروه‌ها، شبکه‌ها و الگوهای ارتباطی درون شبکه است (Bedi & Sharma, 2016). در تشخیص جامعه، گروه‌های مشابه، در مجموعه ای از گره‌های یک ساختار گرافیکی، تشخیص داده می‌شوند (اسمعیلی آبدر و جهانشاهی، ۱۳۹۹). دو الگوریتم مهم در تشخیص جامعه، لووین و لیدن هستند (Anuar et al., 2024).

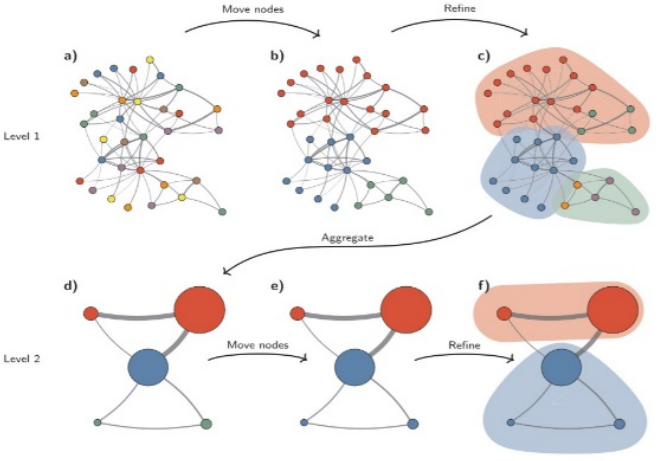
در سال ۲۰۰۸ بلاندل و همکارانش، الگوریتم لووین را ارائه دادند که به تشخیص جامعه بر پایه پیمانگی می‌پردازد (Blondel, Guillaume, Lambiotte & Lefebvre, 2008). این الگوریتم در ابتدا با بهینه‌سازی محلی، به‌دنبال جوامع کوچک می‌گردد و سپس با ادغام جوامع کوچکی که توانایی ایجاد جوامع بزرگ‌تر را دارند، خوشه‌بندی را ادامه می‌دهد. فرض کنید یک شبکه وزن‌دار با n رأس موجود است. این الگوریتم ابتدا هر رأس را یک جامعه در نظر می‌گیرد، سپس الگوریتم برای هر رأس i، انجمن همسایه j را به نحوی می‌یابد که به ازای حذف i از انجمن خودش و ملحق کردن آن به انجمن j، شاخص پیمانگی بیشتر شود. در غیر این صورت، رأس i در جامعه خودش باقی می‌ماند. این عمل به‌طور مکرر برای تمامی رئوس تکرار می‌شود تا زمانی که دیگر تغییری اعمال نشود. این مرحله اول الگوریتم است که در یک نقطهٔ بهینهٔ محلی متوقف می‌شود. نقطه‌ای که هم‌پیمانگی بیشتر با تغییر انجمن هیچ رأسی به‌دست نمی‌آید. سپس در مرحله دوم، الگوریتم با ادغام جوامع کوچک که توانایی ایجاد جوامع بزرگ‌تر را دارند، ادامه می‌دهد. این دو مرحله تا جایی ادامه می‌یابند که تغییری در جوامع ایجاد نشود و شاخص پیمانگی نیز به حالت بیشینه خود دست یافته باشد (Dollmann, 2023; Anuar et al., 2021 & Gilad & Sharan, 2023)

در الگوریتم لووین مشکلی وجود دارد. همان طور که بیان شد، گره‌ها بهترین جامعه محلی را بر اساس تابع هدفی که شباهت جامعه را بیشینه می‌کند، انتخاب می‌کنند و به آن جامعه متصل می‌شوند. اما وقتی یک گره به جامعه دیگری منتقل می‌شود، ممکن است این گره در جامعه قبلی به‌عنوان پلی عمل کند که ارتباط بین اجزا مختلف آن جامعه را برقرار می‌کند. با حذف این گره از جامعه قبلی، ارتباط داخلی جامعه قبلی قطع می‌شود و جامعه گسسته می‌شود. این مشکل در الگوریتم لووین به‌طور متداول در عمل رخ می‌دهد و تکرار الگوریتم ممکن است مشکل را تشدید کند (Traag, Waltman & Van Eck., 2019).

در سال ۲۰۱۹، تراگ و همکارانش الگوریتم لیدن را برای حل مشکل لووین ارائه دادند. الگوریتم لیدن، علاوه‌بر سرعت بالا، تضمین می‌کند که جوامعی را استخراج کند که بیشترین و بهترین اتصال بین اعضای آن وجود داشته باشد (Anuar et al., 2021; Traag et al., 2019).

این الگوریتم پیچیده‌تر و در عین حال دقیق‌تر و سریع‌تر از لووین است. الگوریتم لیدن در سه مرحله اجرا می‌شود. مرحلهٔ اول، بهینه‌سازی پیمانگی، مرحلهٔ دوم اصلاح تقسیم‌بندی و مرحلهٔ سوم تجمیع جامعه است. این الگورتیم در شبکه‌های کوچک، متوسط و بزرگ عملکرد خوبی دارد.

مرحلهٔ دوم، یعنی اصلاح تقسیم‌بندی، به‌معنای بهبود تقسیم‌بندی گره‌ها در جوامع است. این بهبود می‌تواند با جابه‌جایی یا ادغام گره‌ها از جوامع مختلف باهدف افزایش پیمانگی صورت گیرد. شکل ۳ مراحل کلی این الگوریتم را نشان می‌دهد.



شکل ۳. تشخیص انجمن به روش الگوریتم لیدن

منبع: (Traag,et al, 2019)

ارزیابی الگوریتم‌های تشخیص جامعه با شاخص پیمانگی

منظور از شاخص پیمانگی[1]، مقداری بین ۱- تا ۱ است که با رابطهٔ ۲ به دست می‌آید. حداکثر میزان پیمانگی زمانی است که تمام رئوس هر خوشه به هم وصل باشند و یالی خوشه‌های متفاوت را به هم متصل نکند. در این حالت بهترین حالت تشخیص جامعه اتفاق افتاده است. رابطهٔ ۲ نحوه محاسبه پیمانگی را نشان می‌دهد.

$$Q = \frac{1}{2m}\sum_{i.j=1}^{n}(A_{ij} - \frac{k_i\, k_j}{2m})\delta(g_i \,.\, g_j)$$ رابطهٔ ۲)

1. Modularity

در این رابطه، Q نمایانگر ارزش ماژولاریتی است؛ m تعداد کل یال‌ها در شبکه است؛ A_{ij} مقدار ماتریس مجاورت است که نشان می‌دهد آیا گره i و گره j باهم یال دارند یا خیر؛ k_i و k_j درجه گره‌هاست و تعداد یال‌هایی را نشان می‌دهد که به هر گره وارد می‌شود. g_i و g_j برچسب جوامع متناظر با گره‌ها هستند. $\delta(g_i . g_j)$ تابع دلتا است که اگر g_i و g_j برابر باشند، ۱ و در غیر این صورت برابر صفر است.

ارزیابی الگوریتم‌های تشخیص جامعه با شاخص سیلوئت

یکی دیگر از معیارهای مهم ارزیابی خوشه‌بندی، شاخص سیلوئت است. این شاخص براساس میزان نزدیکی اعضای درون هر خوشه به یکدیگر و تفاوت اعضای خوشه‌های متفاوت ساخته می‌شود. میانگین این شاخص می‌تواند مقداری در بازه [۱+،۱-] اختیار کند. اگر میانگین شاخص نزدیک به عدد ۱ باشد، آنگاه مدل خوشه‌بندی رضایت‌بخش تلقی می‌شود. مقادیر منفی و نزدیک به صفر این شاخص، حاکی از نامناسب بودن مدل و عملکرد ضعیف الگوریتم خوشه‌بندی در ایجاد خوشه‌ها است. محاسبه این شاخص برای یک نمونه داده مانند X_i در سه‌گام انجام می‌شود:

گام ۱: محاسبه متوسط فاصله داده X_i از تمام داده‌های دیگر در خوشه خودش (a_i).

گام ۲: محاسبه متوسط فاصله داده X_i از تمام داده‌های دیگر در خوشه دیگر (b_i).

گام ۳: محاسبه ضریب سیلوئت با استفاده از رابطهٔ ۳.

رابطهٔ ۳)
$$S_i = \frac{b_i - a_i}{\max(b_i . a_i)}$$

در این رابطه، a_i و b_i، به‌ترتیب بیانگر میانگین فاصله بین مشاهده i با سایر مشاهدات در یک خوشه مشابه و میانگین فاصله مشاهده i به تمام مشاهدات در خوشه‌های دیگر است. به‌منظور بررسی مناسب بودن یک روش خوشه‌بندی، متوسط S_i برای تمام داده‌ها محاسبه می‌شود (مجرد، پروین، نجاتیان، رضایی و باقری فرد، ۱۳۹۹).

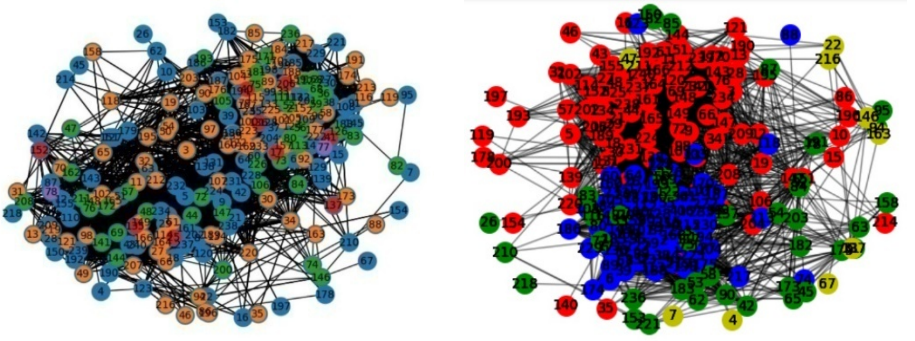
یافته‌های پژوهش

همان طور که در قسمت روش پژوهش بیان شد، چهار دسته داده از سایت سایگو استخراج شد، سپس برای هر سری از داده‌ها، ابتدا با استفاده از ضریب هم‌بستگی پیرسون، شباهت محاسبه شد. در مرحله بعد، از الگوریتم‌های لووین و لیدن جهت تشخیص جوامع علمی متفاوت استفاده شد. در ادامه این بخش، نتایج گزارش می‌شوند.

تشخیص جوامع علمیِ کشورها، بر اساس تولید اسناد علمی در بازهٔ ۱۹۹۶ تا ۲۰۲۱

در این مرحله، جدولی به ابعاد ۲۴۰×۲۴۰ (به تعداد کشورها) ایجاد و تعداد اسناد تولید شدهٔ کشورها در بازه زمانی ۱۹۹۶ تا ۲۰۲۱ در آن قرار گرفت. در مرحله بعد، ضریب هم‌بستگی پیرسون به‌صورت دوبه‌دو، برای کشورها محاسبه شد. این ضریب میزان شباهت علمی در موضوع مدنظر را نشان می‌دهد. در مرحله بعد، گرافی ایجاد می‌شود که کشورها رئوس آن‌ها و ضرایب پیرسون وزن یال‌های بین گره‌ها را نشان

می‌دهند. این گراف، در مرحله بعد به الگوریتم لووین و لیدن جهت تشیخص جامعه داده می‌شود. نتایج در شکل ۴ نمایش داده شده است. در بازه ۱۹۹۶ تا ۲۰۲۱، الگوریتم لووین کشورها را براساس شباهت در تولید اسناد علمی به چهار خوشه (جامعه) متفاوت تقسیم می‌کند که با رنگ‌های قرمز، آبی، زرد و سبز متمایز شده‌اند. نتایج تشخیص جامعه روی این داده‌ها، توسط الگوریتم لیدن، پنج خوشه است که با رنگ‌های نارنجی، آبی، سبز، قرمز، بنفش متمایز شده‌اند.



الف) خوشه‌بندی بر اساس الگوریتم لووین

ب) خوشه‌بندی با استفاده از الگوریتم لیدن

شکل ۴. خوشه‌بندی کشورها با توجه به شباهت در تعداد اسناد تولید شده

بر اساس نتایج تحلیل خوشه‌بندی، در الگوریتم لووین، کشورهایی مانند ایران، ایالات متحده، انگلستان، کره جنوبی، آلمان، ایتالیا و چین در خوشهٔ قرمز، کشورهای اتیوپی، افغانستان، سنگال، سومالی، پاراگوئه، اریتره در خوشهٔ آبی، کشورهای بلیز، پاناما، قرقیزستان، مغولستان، فیجی، مالدیو در خوشهٔ سبز و کشورهای آذربایجان، بلاروس، ازبکستان، تاجیکستان، ارمنستان در خوشهٔ زرد قرار گرفته‌اند.

در الگوریتم لیدن، توزیع کشورها متفاوت است: ایران همراه با انگلستان و آلمان، افغانستان، سنگال، سومالی، پاراگوئه، اریتره، پاناما، قرقیزستان، مغولستان، فیجی، ازبکستان، تاجیکستان، ارمنستان در خوشهٔ آبی، کشورهای چین، کره جنوبی، بلاروس در خوشهٔ سبز، کشور ایالات متحده در خوشهٔ قرمز و کشورهای ایتالیا، اتیوپی، بلیز، مالدیو، آذربایجان در خوشهٔ نارنجی جای گرفته‌اند.

از آنجا که هر دو الگوریتم برای اجرا، به شروع اولیه نیاز دارند و این شروع اولیه می‌تواند تصادفی انتخاب شود، نتایج اجراها با یکدیگر کمی متفاوت‌اند. از این رو، هر دو الگوریتم ۳۰ بار اجرا و میانگین نتایج آن‌ها در نظر گرفته شد. مقادیر سیلوئت[1] و ماژولاریتی[2] و نیز، زمان اجرای این ۳۰ اجرا در جدول ۱ نشان داده شده است.

1. Silhouette
2. Modularity

جدول ۱. نتایج ارزیابی خوشه‌بندی الگوریتم‌های لووین و لیدن

	لووین			لیدن		
ردیف	سیلوئت	ماژولاریتی	زمان	سیلوئت	ماژولاریتی	زمان
۱	۰/۰۳۳۶۴۵۱۵۸	۰/۳۶۴۲۵۶۶۹۱	۰/۱۳۹۸۹۰۶۷۱	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۰۱۳۸۱۴۲۱۱
۲	۰/۰۲۸۹۰۵۶۶۳	۰/۳۶۱۱۹۰۸۵۳	۰/۰۸۲۰۰۰۲۵۶	۰/۰۵۵۰۲۶۰۶۷	۰/۳۶۶۹۳۵۰۷۵	۰/۰۱۴۵۹۵۵۰۹
۳	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۶۰۵۲۴۹۴	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۰۲۹۲۸۸۲۹۲
۴	۰/۰۲۱۳۶۵۷۲	۰/۳۶۳۰۲۲۶۵۸	۰/۰۶۸۸۰۳۳۱	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۱۲۵۹۴۴۶۱
۵	۰/۰۵۲۱۴۵۶۸۲	۰/۳۶۵۶۳۷۷۴۱	۰/۰۶۴۱۴۵۵۶۵	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۱۳۷۰۲۱۵۴
۶	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۰۸۱۳۸۰۸۴۴	۰/۰۵۵۵۰۸۹۸۳	۰/۳۶۶۹۴۱۹۷۲	۰/۰۱۴۶۴۰۸۰۸
۷	۰/۰۰۰۳۹۰۱۹۵	۰/۳۶۲۳۱۹۲۲۲	۰/۰۸۷۳۹۶۸۶	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۲۳۹۸۶۳۴
۸	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۱۲۸۸۸۷۴۱۵	۰/۰۲۹۲۱۷۵۹۸	۰/۳۶۶۹۴۴۵۵۴	۰/۰۱۴۰۲۱۳۹۷
۹	۰/۰۵۵۵۰۸۹۸۳	۰/۳۶۶۹۴۱۹۷۲	۰/۰۸۸۸۹۰۳۱۴	۰/۰۲۹۲۱۷۵۹۸	۰/۳۶۶۹۴۴۵۵۴	۰/۰۱۷۵۹۵۲۹۱
۱۰	۰/۰۵۵۵۰۸۹۸۳	۰/۳۶۶۹۴۱۹۷۲	۰/۰۹۵۶۴۱۶۱۳	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۱۴۰۷۸۱۴
۱۱	۰/۰۵۵۵۰۸۹۸۳	۰/۳۶۶۹۴۱۹۷۲	۰/۰۸۰۳۶۶۳۷۳	۰/۰۲۹۲۱۷۵۹۸	۰/۳۶۶۹۴۴۵۵۴	۰/۰۱۸۳۷۳۰۱۳
۱۲	۰/۰۵۲۱۴۵۶۸۲	۰/۳۶۵۶۳۷۷۴۱	۰/۱۶۶۶۲۵۹۷۷	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۰۱۴۰۸۰۵۲۴
۱۳	۰/۰۳۳۶۴۵۱۵۸	۰/۳۶۴۲۵۶۶۹۱	۰/۱۴۸۰۶۷۴۷۴	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۱۴۶۴۰۸۰۸
۱۴	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۱۲۰۵۳۹۶۶۵	۰/۰۵۵۰۲۶۰۶۷	۰/۳۶۶۹۳۵۰۷۵	۰/۰۲۵۸۸۳۶۷۵
۱۵	۰/۰۳۳۶۴۵۱۵۸	۰/۳۶۴۲۵۶۶۹۱	۰/۰۶۸۷۱۵۰۹۶	۰/۰۳۰۰۴۴۲۵۶	۰/۳۶۵۹۹۵۳۸۴	۰/۰۱۳۹۴۴۱۴۹
۱۶	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۹۱۷۹۹۹۷۴	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۲۳۲۱۳۳۸۷
۱۷	۰/۰۰۷۷۲۸۸۶	۰/۳۶۵۹۰۶۳۹	۰/۱۳۳۵۹۳۷۹۸	۰/۰۲۹۲۱۷۵۹۸	۰/۳۶۶۹۴۴۵۵۴	۰/۰۱۳۷۶۶۵۲۷
۱۸	۰/۰۵۵۵۰۸۹۸۳	۰/۳۶۶۹۴۱۹۷۲	۰/۱۱۲۷۱۴۰۵۲	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۰۲۳۷۴۵۷۷۵
۱۹	۰/۰۲۱۳۶۵۷۲	۰/۳۶۳۰۲۲۶۵۸	۰/۱۲۵۸۳۷۰۸۸	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۱۵۰۵۳۰۳۴
۲۰	۰/۰۴۹۳۷۸۵۴	۰/۳۶۴۰۱۲۲۶۵	۰/۰۸۶۹۲۵۷۴۵	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۰۱۷۸۹۱۴۰۷
۲۱	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۰۸۳۸۰۱۰۳۱	۰/۰۲۹۲۱۷۵۹۸	۰/۳۶۶۹۴۴۵۵۴	۰/۰۱۳۲۵۳۴۵
۲۲	۰/۰۵۲۱۴۵۶۸۲	۰/۳۶۵۶۳۷۷۴۱	۰/۰۷۵۳۸۵۰۹۴	۰/۰۳۳۶۴۵۱۵۸	۰/۳۶۴۲۵۶۶۹۱	۰/۰۱۳۹۶۶۵۶
۲۳	۰/۰۲۲۹۹۷۷۰۱	۰/۳۶۳۵۰۶۵۲۷	۰/۰۷۱۳۱۶۷۱۹	۰/۰۲۸۸۰۷۴۱۹	۰/۳۶۶۱۲۸۶۵۴	۰/۰۲۳۲۲۲۲۰۸
۲۴	۰/۰۰۹۷۶۶۸۴۲	۰/۳۶۵۸۹۳۶۵۷	۰/۱۹۴۵۴۵۷۴۶	۰/۰۳۳۶۴۵۱۵۸	۰/۳۶۴۲۵۶۶۹۱	۰/۰۱۳۴۸۸۵۳۱
۲۵	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۷۹۹۷۳۹۳۶	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۲۴۶۱۳۶۱۹
۲۶	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۹۵۴۱۷۹۷۶	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۱۳۲۷۷۰۵۴
۲۷	۰/۰۵۴۷۰۸۴۲	۰/۳۶۵۶۱۸۹۱۲	۰/۰۶۲۲۹۳۰۵۳	۰/۰۲۹۷۴۰۲۷۹	۰/۳۶۶۹۴۶۹۹	۰/۰۱۲۵۶۶۸۰۵
۲۸	۰/۱۳۳۵۶۵۲۰۴	۰/۳۶۱۷۹۰۰۹۶	۰/۰۸۵۵۷۳۶۷۳	۰/۰۳۰۹۷۷۱۰۲	۰/۳۶۶۸۱۳۷۱۹	۰/۰۱۴۵۲۸۰۳۶
۲۹	۰/۰۲۱۳۶۵۷۲	۰/۳۶۳۰۲۲۶۵۸	۰/۰۸۶۱۵۸۵۱۴	۰/۱۷۲۶۸۵۵۲۱	۰/۳۶۳۲۱۶۰۹۱	۰/۰۱۴۶۸۰۶۲۴
۳۰	۰/۰۴۷۲۱۰۸۱۸	۰/۳۶۲۹۹۵۴۴۵	۰/۱۴۶۵۵۴۲۳۲	۰/۰۵۵۰۲۶۰۶۷	۰/۳۶۶۹۳۵۰۷۵	۰/۰۱۴۰۳۵۴۶۳
میانگین	۰/۰۳۷۷۴۴۹۵	۰/۳۶۵۰۶۸۵۰	۰/۱۰۰۴۵۸	۰/۰۳۷۹۳۰۶۵	۰/۳۶۶۴۴۱۶۷۹	۰/۰۱۶۸۸۴۷۰۸

با توجه به نتایج، مقدار سیلوئت در دو الگوریتم تفاوت قابل توجهی ندارند؛ اما شاخص ماژولاریتی با اختلاف کمی در لیدن، بهتر از لووین است؛ یعنی الگوریتم لیدن نسبت به لووین تشخیص جامعهٔ بهتری داشته است. همچنین، زمان اجرای الگوریتم لیدن کمتر است؛ یعنی الگوریتم لیدن سریع‌تر اجرا می‌شود. ردیف آخر این جدول، میانگین مقادیر هر دو الگوریتم را نشان می‌دهد. این مقدار نیز موید سرعت بیشتر و ماژولاریتی بیشتر الگوریتم لیدن در مقایسه با لووین، در خوشه‌بندی است.

تشخیص جوامع علمیِ کشورها، بر اساس حوزه‌های علمی در بازهٔ ۱۹۹۶ تا ۲۰۲۱

از آنجا که ۲۷ حوزه علمی متفاوت در پایگاه اطلاعاتی سایمگو در نظر گرفته شده است، جدولی به ابعاد ۲۷×۲۷ تهیه شد. این جدول مربوط به شاخص تعداد مستندات در بازهٔ زمانی ۱۹۹۶تا ۲۰۲۱ به‌تفکیک حوزه‌های علمی است. با محاسبهٔ ضریب هم‌بستگی، میزان مشابهت حوزه‌های علمی متفاوت استخراج می‌شود. برای نمونه، در جدول ۲، مشابهت بین حوزهٔ علوم کامپیوتر با برخی از حوزه‌ها نشان داده شده است.

جدول ۲. شش مورد از مشابه‌ترین حوزه‌های علمی با علوم کامپیوتر

۲۰۲۱-۱۹۹۶			
ریاضیات	۰/۹۹	پرستاری	۰/۷۷
علم تصمیم‌گیری	۰/۹۹	هنر و علوم انسانی	۰/۷۶
مهندسی	۰/۹۹	روان‌شناسی	۰/۷۵

همان طور که ردیف‌های جدول نشان می‌دهد، بیشترین هم‌بستگی به‌لحاظ تعداد مستندات بین «علوم کامپیوتر» و «ریاضیات» و «علم تصمیم‌گیری» و «حوزه مهندسی» و کمترین میزان هم‌بستگی، بین رشتهٔ علوم کامپیوتر و پرستاری و هنر و علوم انسانی و روان‌شناسی است؛ یعنی برای مثال، روند رشد مستندات مرتبط با علوم کامپیوتر و ریاضیات به‌طور قابل توجهی شبیه هم هستند. این امر ممکن است به‌دلیل وابستگی مفاهیم و اصول مشترکی در این حوزه‌ها باشد و میزان هم‌بستگی کمتر نشان می‌دهد که مستندات مرتبط با علوم کامپیوتر کمترین شباهت و ارتباط را با حوزه‌های پرستاری، هنر و علوم انسانی و روان‌شناسی دارند. این ممکن است به‌دلیل تفاوت‌های بزرگ در موضوعات و محتواهای این حوزه‌ها باشد.

میزان مشابهت علمی ۳۰ کشور برتر با ایران بر اساس تعداد مستندات

با محاسبهٔ ضریب هم‌بستگی در بازه سال‌های ۲۰۱۸ تا ۲۰۲۲، کشورهایی که از نظر تعداد مستندات بیشترین ضریب هم‌بستگی با ایران را داشتند، جدا کردیم. ۳۰ کشور برتر در تولید اسناد، در جدول ۳ مشاهده می‌شود. در این جدول ضریب هم‌بستگی بین ایران و این کشورها نیز، به تفکیک سال نشان داده شده است.

جدول ۳. میزان شباهت ۳۰ کشور برتر، با ایران بر اساس تعداد مستندات تولیدشده به تفکیک سال

ردیف	۲۰۱۸		۲۰۱۹		۲۰۲۰		۲۰۲۱		۲۰۲۲	
۱	آمریکا	۰/۸۴	آمریکا	۰/۸۵	چین	۰/۹۲	چین	۰/۹۱	چین	۰/۹۱
۲	چین	۰/۹۳	چین	۰/۹۲	آمریکا	۰/۸۶	آمریکا	۰/۸۷	آمریکا	۰/۸۷
۳	انگلستان	۰/۸۱	انگلستان	۰/۸۳	انگلستان	۰/۸۴	انگلستان	۰/۸۶	هند	۰/۸۷
۴	آلمان	۰/۹۴	هند	۰/۸۵	هند	۰/۸۹	هند	۰/۸۹	انگلستان	۰/۸۷
۵	هند	۰/۸۹	آلمان	۰/۹۴	آلمان	۰/۹۴	آلمان	۰/۹۵	آلمان	۰/۹۴
۶	ژاپن	۰/۹۶	ژاپن	۰/۹۶	ایتالیا	۰/۹۱	ایتالیا	۰/۹۲	ایتالیا	۰/۹۳
۷	فرانسه	۰/۹۲	ایتالیا	۰/۹۱	ژاپن	۰/۹۶	ژاپن	۰/۹۶	ژاپن	۰/۹۶
۸	ایتالیا	۰/۹۰	فرانسه	۰/۹۳	فدراسیون	۰/۷۷	کانادا	۰/۸۹	کانادا	۰/۹۰
۹	کانادا	۰/۸۴	فدراسیون	۰/۷۸	فرانسه	۰/۹۳	فرانسه	۰/۹۴	استرالیا	۰/۸۷
۱۰	استرالیا	۰/۸۱	کانادا	۰/۸۶	کانادا	۰/۸۸	فدراسیون	۰/۷۳	فرانسه	۰/۹۴
۱۱	فدراسیون	۰/۷۸	استرالیا	۰/۸۳	استرالیا	۰/۸۶	استرالیا	۰/۸۷	اسپانیا	۰/۸۸
۱۲	اسپانیا	۰/۸۶	اسپانیا	۰/۸۶	اسپانیا	۰/۸۷	اسپانیا	۰/۸۸	فدراسیون	۰/۸۱
۱۳	کره	۰/۹۸	کره	۰/۹۷	برزیل	۰/۸۸	کره	۰/۹۶	کره	۰/۹۶
۱۴	برزیل	۰/۸۶	برزیل	۰/۸۹	کره	۰/۹۶	برزیل	۰/۹۰	برزیل	۰/۹۱
۱۵	هلند	۰/۷۸	هلند	۰/۷۹	ایران	۱	ایران	۱	ایران	۱
۱۶	ایران	۱	ایران	۱	هلند	۰/۸۲	هلند	۰/۸۴	هلند	۰/۸۵
۱۷	لهستان	۰/۹۶	لهستان	۰/۹۵	لهستان	۰/۹۴	ترکیه	۰/۹۵	ترکیه	۰/۹۶
۱۸	سوئیس	۰/۸۵	ترکیه	۰/۹۳	ترکیه	۰/۹۳	لهستان	۰/۹۴	عربستان	۰/۹۱
۱۹	ترکیه	۰/۹۲	سوئیس	۰/۸۶	سوئیس	۰/۸۶	سوئیس	۰/۸۸	لهستان	۰/۹۵
۲۰	سوئد	۰/۸۷	اندونزی	۰/۶۰	اندونزی	۰/۵۳	اندونزی	۰/۴۱	سوئیس	۰/۸۸
۲۱	تایوان	۰/۹۵	سوئد	۰/۸۷	سوئد	۰/۸۹	سوئد	۰/۹۰	سوئد	۰/۹۰
۲۲	بلژیک	۰/۸۷	تایوان	۰/۹۳	تایوان	۰/۹۴	عربستان	۰/۹۳	تایوان	۰/۹۶
۲۳	اندونزی	۰/۶۳	مالزی	۰/۷۹	مالزی	۰/۸۴	تایوان	۰/۹۵	مالزی	۰/۸۵
۲۴	مالزی	۰/۸۲	بلژیک	۰/۸۸	بلژیک	۰/۸۹	مالزی	۰/۸۴	مصر	۰/۹۷
۲۵	دانمارک	۰/۸۲	دانمارک	۰/۸۵	عربستان	۰/۹۴	بلژیک	۰/۹۰	اندونزی	۰/۶۱
۲۶	اتریش	۰/۹۰	پرتغال	۰/۹۱	دانمارک	۰/۸۷	مصر	۰/۹۸	پاکستان	۰/۹۶
۲۷	پرتغال	۰/۹۱	آفریقا	۰/۷۶	پرتغال	۰/۹۱	پاکستان	۰/۹۵	بلژیک	۰/۹۱
۲۸	آفریقای	۰/۷۵	اتریش	۰/۹۱	مصر	۰/۹۸	پرتغال	۰/۹۱	پرتغال	۰/۹۲
۲۹	مکزیک	۰/۹۳	مکزیک	۰/۹۳	آفریقا	۰/۷۷	دانمارک	۰/۸۸	دانمارک	۰/۸۹
۳۰	چک	۰/۹۵	عربستان	۰/۹۵	مکزیک	۰/۹۴	مکزیک	۰/۹۴	آفریقا	۰/۸۰

با توجه به جدول ۳، بیشترین میزان هم‌بستگی ایران از نظر تعداد اسناد تولیدی، در سال ۲۰۱۸ به‌ترتیب با کشورهای کره جنوبی، ژاپن، لهستان، تایوان و چک بوده است. ایران در سال ۲۰۱۹، به‌ترتیب با کشورهای کره جنوبی، ژاپن، لهستان و همچنین عربستان سعودی، بیشترین میزان هم‌بستگی را دارد.

در سال ۲۰۲۰، کره جنوبی جای خود را به مصر داده و سپس کره جنوبی و ژاپن بیشترین میزان شباهت داشته است. در سال ۲۰۲۱، همچنان مصر جایگاه خود را حفظ کرده است و پس از مصر، کشورهای ژاپن و کره جنوبی، آلمان، ترکیه، تایوان، پاکستان در بالاترین میزان شباهت با ایران قرار گرفته‌اند. در سال ۲۰۲۲ نیز کشور مصر، ژاپن، کره جنوبی و سپس ترکیه، تایوان، پاکستان دارای مشابهت بالای ۰/۹۵ ازنظر تعداد تولید سند با ایران هستند که در سال ۲۰۲۲، میزان مشابهتشان افزایش یافته است.

همچنین نتایج جدول ۲ نشان می‌دهد که در سال‌های ۲۰۱۸ و ۲۰۱۹، از جهت تعداد مستندات تولید شده، ایران در جایگاه ۱۶ قرار دارد. این جایگاه در سال‌های ۲۰۲۰، ۲۰۲۱ و ۲۰۲۲ با یک پله ارتقا، به جایگاه ۱۵ رسیده است.

میزان مشابهت ایران با ۳۰ کشور برتر به‌لحاظ شاخص هرش

در این قسمت، میزان مشابهت علمی کشورها با ایران، براساس شاخص هرش بررسی شد. جدول ۴ میزان شباهت ۳۰ کشور اولِ سایمگو با ایران را که خود چهل‌ویکمین کشور بر اساس شاخص هرش است، در بازه زمانی ۲۰۱۸ تا ۲۰۲۲ و به‌تفکیک سال نشان می‌دهد. نکتهٔ مهم این است که مقدار شاخص هرش در پنج سال اخیر تغییراتی نداشته است؛ بنابراین ما در این مقاله، به محاسبهٔ ضریب هم‌بستگی شاخص هرش در سال ۲۰۲۲ کفایت کردیم.

جدول ۴. میزان شباهت ۳۰ کشور برتر در شاخص هرش با ایران در سال ۲۰۲۲

ردیف	کشور	میزان هم‌بستگی هرش	ردیف	کشور	میزان هم‌بستگی هرش
۱	ایالات متحده	۰/۵۴	۱۶	اسرائیل	۰/۶۲
۲	انگلستان	۰/۵۵	۱۷	کره جنوبی	۰/۸۷
۳	آلمان	۰/۶۶	۱۸	اتریش	۰/۶۳
۴	کانادا	۰/۵۶	۱۹	فنلاند	۰/۶۳
۵	فرانسه	۰/۶۷	۲۰	هند	۰/۹۵
۶	هلند	۰/۵۶	۲۱	نروژ	۰/۵۶
۷	استرالیا	۰/۶۶	۲۲	سنگاپور	۰/۸۹
۸	ایتالیا	۰/۶۷	۲۳	برزیل	۰/۷۲
۹	ژاپن	۰/۶۹	۲۴	هنگ‌کنگ	۰/۸۴
۱۰	سوئیس	۰/۶۴	۲۵	فدراسیون روسیه	۰/۷۱
۱۱	چین	۰/۸۹	۲۶	لهستان	۰/۷۹
۱۲	اسپانیا	۰/۷۷	۲۷	نیوزلند	۰/۵۵
۱۳	سوئد	۰/۶۱	۲۸	ایرلند	۰/۷۲
۱۴	بلژیک	۰/۶۶	۲۹	تایوان	۰/۸۹
۱۵	دانمارک	۰/۶۵	۳۰	یونان	۰/۸۵

با توجه به جدول ۳، کشور چین در سه سال اخیر در مبحث تعداد مستندات در ردیف اول قرار دارد؛ اما با توجه به جدول ۴، از نظر شاخص هرش، ایالات متحده رتبهٔ برتر را دارد و چین در رتبهٔ یازدهم قرار گرفته است؛ این بدین معناست که مستندات علمی ایالات متحده توسط سایر پژوهشگران بیشتر استناد شده‌اند. همچنین، با توجه به جدول ۴ مشاهده می‌شود، بر اساس شاخص هرش، ایران بیشترین میزان شباهت را به‌ترتیب با کشورهای هند، چین، سنگاپور و تایوان دارد؛ یعنی به‌لحاظ استناد به مقالات پژوهشگران ایرانی، ایران در ردیف کشورهای مذکور است.

نتیجه‌گیری و پیشنهادها

در این مقاله از الگوریتم لیدن و لووین به‌منظور خوشه‌بندی کشورها براساس تعداد مستندات و شاخص هرس استفاده شد. از نظر فنی، نتایج این پژوهش نشان داد که الگوریتم لیدن نسبت به الگوریتم لووین می‌تواند جوامعی را استخراج کند که بیشترین و بهترین اتصال بین اعضا را تشخیص دهد. مقدار ماژولاریتی با اختلاف کمی بهتر از الگوریتم لیدن بهبود یافت؛ ولی زمان اجرای الگوریتم لیدن، به‌طور محسوسی بهتر از الگوریتم لووین شد.

با بررسی تولید اسناد علمی کشورهای مختلف از سال ۱۹۹۶ تا ۲۰۲۱ و تحلیل شباهت‌های علمی در بین آن‌ها، الگوریتم لووین، کشورها را به چهار خوشه (با رنگ‌های قرمز، آبی، زرد و سبز) و الگوریتم لیدن به پنج خوشه (با رنگ‌های نارنجی، آبی، سبز، قرمز و بنفش) تقسیم‌بندی کرد. این تقسیم‌بندی‌ها نمایانگر ساختارهای علمی و همکاری‌های بین‌المللی در زمینه‌های مختلف پژوهشی هستند.

علاوه‌براین، با تهیه یک جدول برای ۲۷ حوزهٔ علمی متفاوت و محاسبهٔ مشابهت‌ها بین آن‌ها، مشخص شد که حوزه‌های علمی مختلف نیز به میزان متفاوتی با یکدیگر مرتبط هستند. به‌ویژه، حوزه‌های «علوم کامپیوتر»، «ریاضیات»، «علم تصمیم‌گیری» و «مهندسی»، بیشترین هم‌بستگی را در تولید مستندات دارند، در حالی که کمترین میزان هم‌بستگی مربوط به «علوم کامپیوتر» و حوزه‌های «پرستاری»، «هنر» و «علوم انسانی و روان‌شناسی» است.

این نتایج نه تنها به درک بهتر از شباهت‌ها و تفاوت‌های علمی بین کشورهای مختلف کمک می‌کند، بلکه می‌تواند به‌عنوان ابزاری برای شناسایی فرصت‌های همکاری و تحقیقات مشترک در سطح بین‌المللی عمل کند. از این رو، پیشنهاد می‌شود که محققان و سیاست‌گذاران از این تحلیل‌ها برای تقویت پیوندهای علمی و پژوهشی در عرصه‌های خاص استفاده کنند.

در تحلیلی دیگر در این مقاله، با تمرکز بر کشور ایران، میزان شباهت علمی ایران با ۳۰ کشور برتر در تولید اسناد و همچنین شاخص هرش در بازه زمانی ۲۰۱۸ تا ۲۰۲۲ پرداخته شد. نتایج نشان می‌دهند که ایران در سال‌های مختلف به طور مداوم در حال ارتقای خود در تولید مستندات علمی بوده و در سال‌های اخیر، به جایگاه ۱۵ در بین کشورهای برتر جهان از این منظر دست‌یافته است.

تحلیل هم‌بستگی‌ها با کشورهای مختلف نشان می‌دهد که در سال ۲۰۱۸، کره جنوبی، ژاپن، لهستان، تایوان و چک بیشترین شباهت را از نظر تولید اسناد با ایران داشتند. این روند در سال‌های بعدی

به‌ترتیبی دچار تغییر شد که نشان‌دهندهٔ الگوی همکاری و تبادل علمی پایدار بین ایران و این کشورهاست. به‌ویژه، کشور مصر در سال‌های اخیر جایگاه خود را در صدر این فهرست حفظ کرده است و ارتباط مستحکمی با ایران برقرار کرده است.

از سوی دیگر، بررسی شاخص هرش نیز نشان می‌دهد که در حالی که چین به‌لحاظ تعداد مستندات در صدر قرار دارد، ایالات متحده از نظر استناد به مقالات برتر است. این یافته‌ها به‌وضوح نشان‌دهندهٔ تفاوت‌های موجود بین تولید مستندات و تأثیرگذاری آن‌ها بر جامعه علمی جهانی است. همچنین، ایران با کشورهایی چون هند، چین، سنگاپور و تایوان بیشترین مشابهت را در استناد به پژوهشگران خود دارد که بر ارتباطات علمی قوی این کشورها با ایران دلالت دارد.

این نتایج می‌تواند به سیاست‌گذاران و محققان کمک کند تا بیشتر روی تقویت همکاری‌های علمی بین‌المللی تمرکز کنند و از این طریق، به رشد علمی و پژوهشی کشور کمک کنند. در نهایت، این مطالعه اهمیت پیگیری و بهینه‌سازی روابط علمی را برای گسترش سواد علمی و حضور بیشتر در عرصه‌های جهانی پررنگ می‌کند.

از جمله محدودیت‌های این پژوهش، می‌توان به بازه‌های زمانی پایگاه سایمگو اشاره کرد. بازه‌های زمانی به‌دقت یک سال تقسیم شده است و این امکان وجود ندارد که اطلاعات را برای یک بازه زمانی معین برای مثال از سال ... تا سال ... مشاهده شود.

نتایج حاصل از پژوهش حاضر می‌تواند با بررسی نتایج الگوریتم‌های دیگر تشخیص جامعه و استفاده از سایر ضرایب هم‌بستگی توسعه پیدا کند. همچنین از آنجا که تمرکز مقاله حاضر بر تعداد اسناد و شاخص هرش بود، می‌توان پارامترهای دیگری مانند مستندات قابل استناد (دوره سه ساله)[1]، استناد[2]، خود استنادی[3]، میانگین استناد به هر مدرک[4]، به‌عنوان ملاکی برای پیدا کردن هم‌بستگی بین کشورها مورد استفاده قرار گیرند.

فهرست منابع

آزادی احمدآبادی، قاسم (۱۴۰۱). تحلیل و ارزیابی رشد کمّی و کیفی برون‌دادهای علمی جمهوری اسلامی ایران، *پژوهش‌نامه علم‌سنجی*، ۸ (۲)، ۲۶۵- ۲۸۶.

ابازری، زهرا؛ ریاحی، عارف؛ صحبتی‌ها، فریبا؛ صیامیان، حسن؛ یمین فیروز، موسی (۱۳۹۹). بررسی تطبیقی رشد مجلات و مقالات حوزه پزشکی در کشورهای عضو منطقه‌ای مدیترانه شرقی در پایگاه اطلاعاتی اسکوپوس (۲۰۱۲-۲۰۰۲). *مجله دانشکده پیراپزشکی دانشگاه علوم پزشکی تهران (پیاورد سلامت)*، ۹(۳)، ۲۳۵- ۲۴۸.

1. Citable Documents (3 years)
2. Citations
3. Self-Citations
4. Citations per Document

احقاقی الهام؛ فتحیان محمد (۱۴۰۰). علم‌سنجی و فراترکیب وضعیت موجود پژوهش‌های حوزه شبکه‌های همکاری بین سازمانی. *فصلنامه توسعه مدیریت فناوری*، ۹(۱)، ۱۱- ۳۹.

احمدیان دیوکتی، محمدمهدی؛ رازقی، نادر؛ آقاجانی، حسنعلی (۱۳۹۹). آینده‌پژوهی تولیدات علمی ایران تا سال ۲۰۳۰ با استفاده از مدل ARIMA. *مطالعات کتابداری و علم اطلاعات*، ۱۲(۱)، ۱۵۳- ۱۷۳.

اسمعیلی آبدر، سمیه؛ جهانشاهی، محسن (۱۳۹۹). استفاده از الگوریتم GSO برای تشخیص جوامع در شبکه‌های اجتماعی پویا. *نشریه فناوری اطلاعات و ارتباطات انتظامی*، ۱(۲)، ۴۳-۵۳.

جنوی، المیرا؛ شاهمرادی، بهروز (۱۳۹۹). سنجش جایگاه رقابت‌پذیری علمی ایران در منطقه با استفاده از شاخص پیچیدگی علمی. *پژوهش‌نامه علم‌سنجی*، ۵(۱)، ۶۷- ۸۴.

جنوی، المیرا؛ مرادی، شیما؛ پاکزاد، مهدی (۱۳۹۹). ارزیابی وضعیت انتشارات علمی ایران بر مبنای نقشه جامع علمی کشور. *پژوهش‌نامه علم‌سنجی*، ۶(۱)، ۲۱۳- ۲۳۶.

حسن‌زاده، رمضان؛ مداح، محمد تقی (۱۴۰۲). *روش‌های آماری در علوم رفتاری*. انتشارات روان.

خرمی، محیا (۱۳۹۶). *تحلیل داده‌های پایگاه سایمگو به‌منظور استخراج روابط شباهتی جغرافیایی و موضوعی*، پایان‌نامه کارشناسی ارشد رشته مهندسی فناوری اطلاعات گرایش تجارت الکترونیک.

فرزین یزدی، محبوبه؛ رضایی شریف آبادی، سعید (۱۳۹۶). بررسی تولیدات علمی حوزه موضوعی هوش مصنوعی در کشورهای خاورمیانه طی سال‌های ۱۹۹۶ تا ۲۰۱۴. *پژوهش‌نامه علم‌سنجی*، ۳(۲)، ۹۷- ۱۱۴.

مجرد، موسی؛ پروین، حمید؛ نجاتیان، صمد؛ رضایی، وحیده؛ باقری فرد، کرم‌اله (۱۳۹۹). خوشه‌بندی مبتنی بر گراف با استفاده از آزمون ویلکاکسون جهت استخراج ارتباطات بیولوژیکی سلول‌ها و بافت‌ها. *مجله مهندسی برق دانشگاه تبریز*، ۵۰(۳)، ۱۳۷۳- ۱۳۸۲.

محمد اسماعیل، صدیقه؛ ریاحی، عارف و صحبتی‌ها، فریبا (۱۳۹۳). ارزیابی کمی و کیفی مجلات ایران در پایگاه استنادی اسکوپوس طی سال‌های ۲۰۱۲-۲۰۰۰. *مجله علم‌سنجی کاسپین*، ۱(۱)، ۳۳-۳۹.

نصرتی، وحید؛ رحمانی، محسن (۱۴۰۱). ارائه روش انتخاب ویژگی مبتنی بر خوشه‌بندی در مسئله تشخیص هرزنامه. *مدیریت اطلاعات*، ۸(۱)، ۲۰۲- ۲۲۴.

Anuar, S. H. H., Abas, Z. A., Yunos, N. M., Zaki, N. H. M., Hashim, N. A., Mokhtar, M. F., ... & Nizam, A. F. (2021, December). Comparison between Louvain and Leiden algorithm for network structure: a review. *In Journal of Physics: Conference Series*, 2129 (1). 012028. IOP Publishing.

Bedi, P. & Sharma, C. (2016). *Community detection in social networks*. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 6(3), 115-135.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10),10008.

Daradkeh, M. Abualigah, L. Atalla, S. & Mansoor, W. (2022). Scientometric analysis and classification of research using convolutional neural networks: A case study in data science and analytics. *Electronics*, 11(13),2066.

Dikusar A. & Cujba, R. (2024). Scientometric Approach in Determining the Role of Science in Socioeconomic Development of Society. *Journal of Social Sciences*, *7*(2), 159–169.

Dollmann, M. M. (2023). Graph Clustering: A Comparison of Louvain and Leiden. *Conf. Ser.* 2129 012028.

Erfanmanesh, M., Jahromi, R. B. Hosseini, E. & Gholamhosseinzadeh, Z. (2013). Scientific productivity, impact and collaboration of the top Asian countries in Scopus during 1996-2010. *Collnet Journal of Scientometrics and Information Management*, 7(1), 97-110.

Gilad, G. & Sharan, R. (2023). From Leiden to Tel-Aviv University (TAU): exploring clustering solutions via a genetic algorithm. *PNAS nexus*, 2(6), pgad180.

Khokhlov, A. N. (2020). How scientometrics became the most important science for researchers of all specialties. *Moscow University Biological Sciences Bulletin*, *75*(4), 159-163.

Noroozi Chakoli, A., Noroozi Chakoli, S. & Chehrenegar, L. (2023). Is there relationship between cultural-economic indicators and the scientific status of countries? Analysis of Western and Central Asian countries using a neural network algorithm. *27th International Conference on Science, Technology and Innovation Indicators* (STI 2023).

Reyes, C. F. (2014). Growth of the number of indexed journals of Latin America and the Caribbean: The effect on the impact of each country. *Scientometrics*, 98,197-209.

Roldan-Valadez, E., Salazar-Ruiz, S. Y., Ibarra-Contreras, R. & Rios, C. (2019). Current concepts on bibliometrics: a brief review about impact factor, Eigenfactor score, CiteScore, SCImago Journal Rank, Source-Normalised Impact per Paper, H-index, and alternative metrics. *Irish Journal of Medical Science* (1971-), (188), 939-951.

Sallam, M., Mohammadi, M., Sainsbury, F., Nguyen, N. T., Kimizuka, N., Muyldermans, S. & Benešova-Schäfer, M. (2024). Bibliometric and scientometric analysis of PSMA-targeted radiotheranostics: knowledge mapping and global standing. *Frontiers in oncology*, *14*, 1397790.

Traag, V.A., Waltman, L. & Van Eck, N.J. (2019). *From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports*, 9(1), 5233.

Wang, X., Long, S., Zeng, L., Chen, C. & Yishan, L. (2024, June). Mapping the Evolution and Future Trajectories of Network Mining: A Scientometric Analysis (2004–2023).

In *2024 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM)* (pp. 468-473). IEEE.

Wani, Z. A. & Zainab, T. (2017). A review of eminence of scientometric indicators in scientific research productivity', *COLLNET Journal of Scientometrics and Information Management*, 11(2), 273–285.

Winarko, B., Abrizah, A. & Tahira, M. (2016). An assessment of quality, trustworthiness and usability of Indonesian agricultural science journals: stated preference versus revealed preference study. *Scientometrics*, 108, 289-304.

Using Machine Learning in Extracting the Scientific Similarity of Countries

Seyede Fatemeh Noorani [*1]
Assistant Prof., Department of Computer Science, Faculty of Engineering and Technology, Payam Noor University, Tehran, Iran

Raana Nagdi
MSc., Department of Computer Science, Faculty of Engineering and Technology, Payam Noor University, Tehran, Iran

Abstract
Today, the production of science is recognized as an important priority in all countries, because scientific development is the basis for the development of technology, and the development of technology is also the basis of economic growth and social welfare. For this reason, measuring the quantitative and qualitative level of scientific production of societies is very important. Scientometrics and bibliometrics are tools used to measure and evaluate scientific productions in societies. These types of studies and reviews have wide applications in various educational and research fields or for decision-making, policy-making and foresight in institutions and organizations. In this context, one of the useful tools is the Symgo database, which provides valuable data such as the scientific performance of the countries of the world in various scientific fields, and can be used as a suitable source of information for conducting such research. This database provides valuable information and data related to the scientific performance of different countries in various scientific fields and can be used as a scientific database for conducting such research. The purpose of this article is to find the scientific similarity of countries and scientific fields in a certain period of time based on two bibliometric indicators, namely the number of documents and the H-index. Then we will cluster using the similarity obtained by applying Louvain and Leiden community detection algorithms, based on which we will bring analysis. In this research, although the Silhouette value did not improve in the Leiden algorithm, we had a change in the Modularity discussion with a slight difference, and that is because of the nature of this algorithm, which works based on Modularity, and the execution time of the Leiden algorithm was significantly better than the Louvain algorithm.

Keywords: Machine learning, Scientific similarity, Data mining, Clustering.

1.. Corresponding Author: sf.noorani@pnu.ac.ir