NASTINFO, 32 (3), Autumn 2021: 56-76 ISSN: 2383-322X

Journal of National Studies on Librarianship and Information Organization (NASTINFO)

Research Article

Analysis of Citation-based Indicators to Determine the Relevance of Articles

M. Goltaji [1]

J. Abbaspour[2]

A. Jowkar[3]

S. M. Fakhrahmad[4]

A. Nikseresht[5]

1. PhD Candidate, Knowledge and Information Science Department, International Division, Shiraz University, Shiraz, Iran, marzieh.goltaji@gmail.com

2. PhD in Knowledge & Information Science, Assistant Professor, School of Education & Psychology, Shiraz University, Shiraz, Iran, (Corresponding Author), javad.abbaspour@gmail.com

3. PhD in Knowledge & Information Science, Professor, School of Education & Psychology, Shiraz University, Shiraz, Iran, ajowkar2003@yahoo.com

4. PhD in Computer Engineering, Associate Professor, Department of Computer Science and Engineering & IT, Shiraz University, Shiraz, Iran, fakhrahmad@shirazu.ac.ir

5. PhD in Computer Engineering, Assistant Professor, School of Education & Psychology, Shiraz University, Shiraz, Iran, nikseresht@gmail.com

Abstract

Purpose: The present study aimed to investigate the potential of citation-based indicators (Co-Citation, Bibliographic Coupling, Amsler, PageRank, HITS) to determine the relevance of articles.

Method: This is applied research with correlational approach. The population consisted of 26,262 articles in the PubMed Central open access subset of the CITREC, which had citation relationship with other articles based on all three traditional citation-based indicators (Co-Citation, Bibliographic coupling, Amsler). From among the citations in the research population, 30 were selected as basic ones, and the full-text of them were retrieved based on the mesh similarity. Then the similarities among the retrieved documents were extracted based on citation-based indicators. Each of the citation-based metrics was considered as independent variable and the mesh similarity as dependent variable. A MySQL database was created using WampServer simulation software and PHP My Admin. Then, using online demo of the CITREC test collection, an output was prepared. By entering the output into the MySQL database which contains the research data set, the main structure of its tables was created. Finally, by studying all the required codes from the CITREC source code package, we attempted to enter the required codes by applying necessary changes. The results were entered in the created MySQL database. By writing a query in SQL language, the set citation network was completely extracted and stored in a Comma-separated values (CSV) file. Then, a program was written in Python that could open and process this large file and calculate PageRank and HITS numbers (authority and Hub).

Findings: The results showed that all six measures studied had a significant and positive correlation with the relevance of articles. In other words, with increasing the values of each measure, the degree of relevance of the articles also increased. The highest correlation with the relevance of the articles belonged to the Amsler measure, followed by the Bibliographic Coupling. After Amsler and Bibliographic Coupling, the highest correlation was observed in the HITS(Authority) variable, and the PageRank variable was in the fourth place; Finally, the lowest correlation with the relevance of the articles was related to the Co-Citation and the HITS (Hub). Therefore, among the known Citation- based measure studied here, Amsler, Bibliographic Coupling, HITS(Authority) and PageRank metrics, respectively, had more potential to determine the relevance of articles rather than others.

Conclusion: Based on the findings, it can be concluded that the citation-based metrics studied are able to estimate the degree of relevance of articles. Therefore, they can be used in various information retrieval platforms, including search engines, citation- based databases, recommender systems, and even digital libraries to access articles, suggest similar articles, and rank retrieved results; Also, the Amsler measure as the less used in information retrieval systems than the two traditional Measure (Co- Citations and Bibliographic Coupling) needs to be considered more than ever. On the other hand, despite the fact that Co- Citations measure is used in some international information retrieval databases (such as Science Direct and CiteSeer) to retrieve relevant documents and suggest similar documents, it is less efficient than other metrics.

Keywords: Citation- Based Metrics, Relevance of Articles, Co-citation, Bibliographic Coupling, Amsler, PageRank, HITS

Recommended Citation

Goltaji, M., Abbaspour, J., Jowkar, A., Fakhrahmad, S., nikseresht, A. (2021). Analysis of Citation-based Indicators to Determine the Relevance of Articles. *Journal of National Studies on Librarianship and Information Organization (NASTINFO),* 32 (3): 56-76

10.30484/NASTINFO.2021.2942.2069

Received:06, Aug. 2021; accepted:18, Sep. 2021

فصلنامه علمی – پژوهشی
مطالعات ملی کتابداری و سازماندهی اطلاعات

مقاله پژوهشی

تحلیل سنجه‌های استنادمحور برای تعیین میزان ربط مقاله‌ها

مرضیه گل‌تاجی[۱]
جواد عباس‌پور[۲]
عبدالرسول جوکار[۳]
سیدمصطفی فخراحمد[۴]
علیرضا نیک‌سرشت[۵]

[۱]. دانشجوی دکتری علم اطلاعات و دانش‌شناسی، واحد بین الملل دانشگاه شیراز، شیراز، ایران mailto:University of Cambridge
[۲]. استادیار گروه علم اطلاعات و دانش‌شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه شیراز، شیراز، ایران (نویسنده مسئول)، javad.abbaspour@gmail.com
[۳]. استاد گروه علم اطلاعات و دانش‌شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه شیراز، شیراز، ایران jowkar2003@yahoo.com
[۴]. دانشیار گروه مهندسی و علوم کامپیوتر و فناوری اطلاعات، دانشکده مهندسی برق و کامپیوتر، دانشگاه شیراز، شیراز، ایران mfakhrahmad@gmail.com
[۵]. استادیار گروه علم اطلاعات و دانش‌شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه شیراز، شیراز، ایران nikseresht@gmail.com

چکیده

هدف: شناخت توانایی سنجه‌های استنادمحور (هم‌استنادی، زوج کتاب‌شناختی، امسلر، پیج‌رنک و هیتس(اعتبار و کانون)) برای تعیین میزان ربط مقاله‌ها با یکدیگر.

روش: پژوهش حاضر از نظر هدف، کاربردی و از لحاظ شیوه گردآوری داده‌ها، پژوهشی توصیفی از نوع همبستگی است. جامعه آماری، مجموعه مقالات موجود در زیرمجموعه دسترسی آزاد پاب‌مد سنترال مجموعه آزمون سایترک بود که بر اساس سه سنجه هم‌استنادی، زوج کتاب‌شناختی و امسلر با سایر مقالات رابطه استنادی داشتند. از میان ۲۶۲۶۲ مقاله، ۳۰ مقاله به‌عنوان مقالات پایه انتخاب شد و مقالات مرتبط با هر یک از آن‌ها بر اساس سنجه ربط مش بازیابی گردید؛ هر یک از سنجه‌های استنادمحور متغیر مستقل و سنجه ربط مش متغیر وابسته بود. با استفاده از نرم‌افزار شبیه‌ساز ومپ‌سرور و پی.اچ.پی.مای‌ادمین یک پایگاه مای.اس.کیو.ال ایجاد شد؛ سپس، با مطالعه کلیه کدهای مورد نیاز از بسته کد منبع سایترک، کدهای لازم با اعمال

تغییرات ضروری، اجرا و نتایج حاصل در پایگاه مای.اس.کیو.ال وارد شد. با نوشتن پرس‌وجو به زبان اس.کیو.ال، شبکه استنادی مجموعه به‌صورت کامل استخراج شد سپس با کدنویسی به زبان پایتون اعداد مربوط به پیج‌رنک و هیتس (اعتبار و کانون) به‌صورت جداگانه محاسبه گردید.

یافته‌ها: نتایج نشان داد تمامی شش سنجه در سطح یک صدم همبستگی معنادار و مثبت با میزان ربط مقاله‌ها داشت؛ به‌عبارت دیگر، با افزایش مقادیر هریک از سنجه‌ها، درجه ربط مقاله‌ها نیز افزایش یافت. بیشترین میزان همبستگی مربوط به سنجه امسلر و پس از آن، زوج کتاب‌شناختی بود. پس از سنجه‌های امسلر و زوج کتاب‌شناختی، بیشترین همبستگی میان متغیر هیتس(اعتبار) با ربط مقاله‌ها بود. متغیر پیج‌رنک در مرتبه چهارم قرار داشت؛ در نهایت، کم‌ترین میزان همبستگی با ربط مقاله‌ها، مربوط به سنجه‌های هم‌استنادی و هیتس(کانون) بود؛ بنابراین، از میان سنجه‌های استنادی بررسی‌شده در این پژوهش، سنجه‌های امسلر، زوج کتاب‌شناختی، هیتس(اعتبار) و پیج‌رنک بیش از سایر سنجه‌ها از پتانسیل لازم برای تعیین میزان ربط مقاله‌ها برخوردار بودند.

نتیجه‌گیری: بر اساس یافته‌های پژوهش می‌توان گفت سنجه‌های استنادمحور مطالعه‌شده قادرند درجه ربط مقاله‌ها را برآورد کنند و در بافتارهای مختلف بازیابی اطلاعات شامل موتورهای جست‌وجو، پایگاه‌های اطلاعاتی و استنادی، سامانه‌های پیشنهاددهنده و حتی کتابخانه‌های دیجیتالی برای دسترسی به مقالات مرتبط، پیشنهاد مقالات مشابه و رتبه‌بندی نتایج بازیابی کاربرد داشته باشند؛ همچنین، لازم است به سنجه امسلر که نسبت به دو سنجه سنتی هم‌استنادی و زوج کتاب‌شناختی، در سامانه‌های اطلاعاتی کمتر استفاده شده‌است، بیش از پیش توجه شود؛ از طرفی، علیرغم اینکه سنجه هم‌استنادی در برخی از پایگاه‌ها و سامانه‌های بازیابی اطلاعات بین المللی(مانند ساینس دایرکت و سایت سیر) برای بازیابی مدارک مرتبط و پیشنهاد مدارک مشابه استفاده می‌شود در مقایسه با سایر سنجه‌ها از کارایی کمتری برخودار است.

کلیدواژه‌ها: ربط مقاله‌ها، هم‌استنادی، زوج کتاب‌شناختی، امسلر، پیج‌رنک، هیتس، سنجه‌های استنادمحور

استناد به این مقاله:

گل‌تاجی، مرضیه، عباس پور، جواد، جوکار، عبدالرسول، فخراحمد، سید مصطفی، نیک سرشت، علیرضا. (۱٤۰۰). تحلیل سنجه‌های استنادمحور برای تعیین میزان ربط مقاله‌ها (۱٤۰۰). *فصلنامه مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۳۲ (۳): ۷٦-٥٦

10.30484/NASTINFO.2021.2942.2069

دریافت: ۱٤۰۰/۰٥/۱٥؛ پذیرش: ۱٤۰۰/۰٦/۲۷

مقدمه

سامانه‌های بازیابی اطلاعات شامل موتورهای کاوش، کتابخانه‌های دیجیتال و پایگاه‌های اطلاعاتی، از هر نوع و در هر سطح از پیچیدگی، هدف بنیادی مشترکی را دنبال می‌کنند و آن نمایش مدارک یا اطلاعات مرتبط با درخواست یا نیاز اطلاعاتی کاربران است (نشاط، ۱۳۸۲). در این سامانه‌ها برای تعیین میزان ربط پرس‌وجو[1] با مدرک یا مدارک از معیارهایی با عنوان "سنجه‌های شباهت" استفاده می‌شود و چنانچه مقدار شباهت از یک حد آستانه بیشتر باشد، مدارک به‌عنوان مدارک مرتبط بازیابی می‌شوند؛ همچنین، این سنجه‌ها شاخص‌های مناسبی برای رتبه‌بندی مدارک بازیابی‌شده‌اند و به سامانه‌ها اجازه می‌دهند تا با مرتب ساختن مدارک بر اساس میزان ربط، مدارک مرتبط‌تری را در صدر نتایج قرار دهند (Zhang and Korfhage, 1999)؛ همچنین سنجه‌های شباهت در بسترهای دیگری مانند طبقه‌بندی خودکار مدارک؛ تشکیل خوشه‌های مدارک مشابه؛ ایجاد سامانه‌های بازیابی گرافیکی و ترسیم نقشه‌های علوم نیز به‌کار می‌روند (Torres, Basnet, Sung, Mukkamala, & Ribeiro, 2009; Jouili, Tabbone, & Valveny,2010; Lin, Jiang, & Lee, 2013; Amer and Abdalla, 2020; Eminagaoglu, 2020).

علی‌رغم کاربرد وسیع و اهمیت سنجه‌های شباهت در بازیابی اطلاعات به‌طور گسترده و سامانه‌های بازیابی اطلاعات به‌صورت محدود، چالش اساسی این موضوع، مسئله ربط و رضایت کاربر است؛ نتایج پژوهش‌ها نشان می‌دهد کاربران از نتایج بازیابی‌شده توسط سامانه‌های بازیابی اطلاعات رضایت ندارند و بین نتایج بازیابی‌شده از طریق این سامانه‌ها با نظر کاربران قرابت چندانی وجود ندارد (سعدین‌خرم و عباس‌پور، ۱۳۹۸؛ حاجیان و چشمه‌سهرابی، ۱۳۹۹؛ Bar-Ilan, Keenoy, Yaari, & Levene, 2007; Hariri, 2011)؛ از طرفی، نتایج پژوهش‌های متعدد نشان داده است که کاربران تنها به چند نتیجه نخست توجه می‌کنند و سایر نتایج بازیابی‌شده را نادیده می‌گیرند (Bar-Ilan, Levene & Mat-Hassan, 2006; Lewandowski, 2008 ; Agrahri, Manickam, & Riedl, 2008; Bar-Ilan, Keenoy, Levene, & Yaari, 2009; Lewandowski, 2017). مهم‌ترین عواملی که می‌تواند بر این امر تأثیر مستقیم داشته باشد، کیفیت سنجه‌های ربط و شباهت است؛ به‌عبارت دیگر، هرچه سنجه‌های بهتری به کار گرفته شود، می‌توان انتظار نتایج بازیابی بهتری را داشت؛ بنابراین، سنجه‌های به‌کاررفته در سامانه‌ها برای شباهت‌یابی و ربط مدارک بایستی به‌حدی اثربخش باشد که نه‌تنها کلیه مدارک مرتبط با پرس‌وجو کاربر را بازیابی کند؛ بلکه مدارک را به نحوی رتبه‌بندی کند که با حداقل کوشش و زمان، نیاز اطلاعاتی کاربر پاسخ داده شود.

McGill, Koll, & Noreault (1979) ، ٦٧ سنجه ربط و شباهت مختلف به‌کاررفته در بازیابی اطلاعات را بررسی و مقایسه کردند (Quated in Zhang and Korfhage, 1999). بر اساس جست‌وجوهای انجام‌شده، در حال حاضر آمار دقیقی از تعداد سنجه‌های شباهت و ربط ابداع‌شده توسط متخصصان در دسترس نیست؛ ولی اکنون که در حدود چهار دهه از انتشار اثر آن‌ها می‌گذرد، به احتمال فراوان تعداد آن‌ها بسیار بیشتر از گذشته خواهد بود. سنجه‌های شباهت را بر اساس آنچه مبنای محاسبه شباهت قرار می‌گیرد، به سه گروه اصلی دسته‌بندی کرده‌اند: سنجه‌های متن‌محور، سنجه‌های استنادمحور و سنجه‌های ترکیبی (Boyack and Klavans, 2010). هرچند امروزه به‌دلیل گسترش و بهبود نظام‌های پردازش متن، سنجه‌های متن‌محور تحولات چشم‌گیری داشته‌اند؛ ولی سنجه‌های استنادمحور

[1]. Query

نیز دارای برخی از ویژگی‌ها و قابلیت‌های منحصربه‌فردند که آن‌ها را از برخی جهات از رویکردهای متن‌محور متمایز می‌سازد؛ ازجمله اینکه، سنجه‌های استنادمحور که در قالب پرس‌وجو توسط کاربران به نظام داده می‌شوند تحت‌تأثیر عبارت‌ها و کلمات گوناگونی قرار نمی‌گیرند؛ دسترسی به منابع و متون بین‌رشته‌ای را فراهم می‌سازند؛ و منابعی را که با یک موضوع خاص ارتباط دارند ولی با سامانه‌های بازیابی متنی ردگیری نمی‌شوند، نشان می‌دهند (Smith, 1981). علاوه بر این، داده‌های استنادی به‌راحتی قابل درک هستند و به‌سرعت هم گردآوری می‌شوند (Su and et al., 2011).

در برخی از سامانه‌های بازیابی اطلاعات، به‌ویژه پایگاه‌های استنادی و برخی موتورهای جست‌وجو، به‌منظور محاسبه میزان ربط مدارک با پرس‌وجو یا سایر مدارک، از پیوندهای استنادی میان مدارک و ساختار گراف آن‌ها استفاده می‌شود (Reyhani Hamedani, Lee, and Kim, 2013; Reyhani Hamedani and Kim, 2021). برخی از متخصصان معتقدند که استنادها اغلب برای اندازه‌گیری شباهت ضمنی میان مدارک علمی استفاده می‌شوند و بازیابی مدارک با استفاده از پیوندهای استنادی قادر است مدارک شناسایی‌نشده با پردازش متنی را بیابد (Eto, 2019)؛ در مقابل، عده‌ای هم عقیده دارند که این‌دسته از سنجه‌ها، تنها روابط استنادی میان مقالات علمی را درنظر می‌گیرند و از محتوای مقالات غافل می‌شوند (Reyhani Hamedani and et al., 2013).

از مهم‌ترین سنجه‌های سنتی مبتنی بر استناد می‌توان به سنجه‌های هم‌استنادی[1]، زوج کتاب‌شناختی[2]، و امسلر[3] اشاره کرد. هم‌استنادی پیوندی است که توسط نویسندگان جدید، میان مقالات پیشین برقرار می‌شود؛ بنابراین، هم استنادی، بسامد تعداد دفعاتی است که به دو مدرک همراه یکدیگر در مدارک بعدی استناد می‌شود (Small, 1973). در زوج کتاب‌شناختی، به مآخذ مشترک توجه می‌شود؛ به‌عبارتی، وجود یک مأخذ در دو مقاله، واحد اندازه‌گیری حد اشتراک آن دو مقاله محسوب می‌شود و هرچه دو مقاله در تعداد بیشتری از مآخذ خود مشترک باشند از لحاظ محتوایی به یکدیگر نزدیک‌ترند (Kessler, 1963). برای محاسبه شباهت مدارک با سنجه امسلر که شباهت پیوندی[4] نیز نامیده می‌شود، هم‌زمان هر دو سنجه زوج کتاب‌شناختی و هم‌استنادی لحاظ می‌شود (Bichteler and Eaton, 1980).

درکنار سه سنجه استنادمحور پیش‌گفته، که از قدمت بیشتری برخوردارند، دو سنجه هیتس[5] و پیج‌رنک[6] نیز وجود دارد که در اصل برای محیط وب و به‌ویژه رتبه‌بندی صفحات وب ایجاد شدند؛ اما پژوهشگران قابلیت به‌کارگیری آنان برای شناسایی مدارک علمی مهم و معتبر و رتبه‌بندی آن‌ها را بررسی و عملکرد آن‌ها را برای سنجش شباهت مدارک تأیید کردند (Zhuge and Zhang, 2010; Yin, Huang, & Li, 2011; Liu, Zhang, & Guo, 2012).

هر دو سنجه پیج‌رنک و هیتس، سنجه‌های بازگشتی و تکرارشونده مبتنی بر پیوند مدارک در محیط وب هستند؛ با این تفاوت که گرایش سنجه پیج‌رنک به‌سمت بازتاب مقالات قدیمی و سنجه هیتس به مقالات جدید است (Jiang, Sun, Yang, Zhuge, and Yao, 2016; Devi, Gupta and Dixit, 2014). در رتبه‌بندی صفحات وب با هیتس، ابتدا جست‌وجوی موضوعی با یک یا چند عبارت پرسشی انجام می‌گیرد، و تعدادی صفحه دارای بالاترین رتبه استخراج می‌شوند که

1. Cocitation
2. Bibliographic Coupling
3. Amsler

4. Linkage Similarity
5 .HITS
6 .PageRank

"مجموعه ریشه[1]" را شکل می‌دهند. این صفحات مرتبط‌ترین صفحات به پرس‌وجو مطرح‌شده فرض می‌شوند. سپس، صفحاتی که به وسیله پیوند، به مجموعه ریشه پیوند داده‌اند و یا مجموعه ریشه با آن‌ها پیوند برقرار کرده است به مجموعه افزوده می‌شوند و مجموعه‌ای بزرگ‌تر با نام "مجموعه پایه[2]" به وجود می‌آید که مبنای محاسبه الگوریتم هیتس است.. این الگوریتم به هر صفحه در این مجموعه یک امتیاز اعتبار[3] و یک امتیاز کانون[4] اختصاص می‌دهد (Kleinberg, 1999).

هیتس دو مقدار اعتبار و کانون را به ازای هر صفحه که به‌صورت بازگشتی متقابل روی هم اثر دارند محاسبه می‌کند؛ به‌عبارت دیگر، یک صفحه با امتیاز اعتبار بالا به‌وسیله تعدادی از صفحات با امتیاز کانون بالا (تعداد زیادی از صفحات با امتیاز کانون بالا به آن پیوند داده باشند) و یک صفحه با امتیاز کانون بالا به تعدادی صفحه با امتیاز اعتبار بالا اشاره می‌کند (صفحه‌ای است که به صفحات با امتیاز اعتبار بالا پیوند داده باشد) (Kleinberg, 1999; Kirsch, Gnasa, Won, and Cremers,2008).

عملکرد سنجه‌های استنادمحور در بازیابی از گذشته تا به امروز همواره مدنظر پژوهشگران بوده است. در پژوهش‌های اولیه، عملکرد سنجه‌های تعداد استنادات، زوج کتاب‌شناختی، هم‌استنادی، و امسلر در موضوع‌های مختلف بازیابی اطلاعات و همچنین در قیاس بایکدیگر آزموده شد. یافته‌های پژوهش Bichteler and Eaton (1980) بیانگر برتری مقیاس شباهت پیوندی امسلر (ترکیب زوج کتاب‌شناختی و هم‌استنادی) در مقایسه با سنجه زوج کتاب‌شناختی به‌تنهایی بود. در پژوهش دیگری، Shibata, Kajikawa, Takeda, and Matsushima (2009) سه سنجه هم‌استنادی، زوج کتاب‌شناختی و استناد مستقیم[5] را از نظر قابلیت شناسایی بهینه پیشگامان پژوهش و مدارک هسته با یکدیگر مقایسه‌کردند. یافته‌ها نشان داد استناد مستقیم در مجموع بهترین و هم‌استنادی بدترین عملکرد را نسبت به سایرین داشتند.

پس از ظهور شبکه جهانی وب- که اساس آن را فراپیوندها[6] تشکیل می‌دادند- سوالی که مطرح شد این بود که آیا از سنجه‌های پیشین، به‌ویژه سنجه‌های استنادمحور که قرابت نزدیکی با فراپیوندها داشتند، می‌توان در محیط وب هم استفاده کرد یا خیر؟ و در صورت مثبت‌بودن پاسخ این سوال، هر یک از این سنجه‌ها در چه بستری کارایی بهتری دارند؟ با توجه به اینکه استناد و سنجه‌های استنادمحور، اعتباری که مدارک به یکدیگر می‌دهند را در نظر نمی‌گیرند، توانایی سنجه‌های مبتنی بر گراف مانند الگوریتم‌های پیج‌رنک و هیتس در شباهت‌یابی مدارک مدنظر پژوهش‌ها قرار گرفت و قابلیت این سنجه‌های جدید استنادمحور در قیاس باهم، و در مقایسه با سنجه‌های پیشین آزموده شد.

Ma, Guan and Zhao (2008) با مقایسه دو سنجه پیج‌رنک و تعداد استنادات، تلاش کردند تا روش جدیدی برای اندازه‌گیری اهمیت مقالات علمی مبتنی بر پیج‌رنک گوگل عرضه کنند. آن‌ها نتایج حاصل از رتبه‌بندی مقالات علمی با پیج‌رنک را با تعداد استنادات به آن‌ها مقایسه و همبستگی زیادی بین استناد و پیج‌رنک مشاهده کردند. مقایسه این دو سنجه در پژوهش Herskovic and Bernstam (2005) تکرار شد و نتایج نشان داد عملکرد پیج‌رنک نسبت به تعداد استناد بهتر بود. در پژوهشی دیگر Lin (2008) دو سنجه جدید استنادمحور پیج‌رنک و هیتس را مقایسه کرد. وی قابلیت

1. Root Set
2. Base Set
3. Authority
4. Hub
5. Direct citation
6. Hyperlinks

پیج‌رنک و هیتس را در شبکه مقالات مرتبط پاب‌مد[1] برای بازیابی متون زیست‌پزشکی بررسی کرد. نتایج این پژوهش نشان داد که پیج‌رنک نسبت به هیتس برای تحلیل ساختار پیوند شبکه‌های مدارک مرتبط مؤثرتر است.

Yin and et al.,(2011) سه الگوریتم تحلیل پیوندی را که عبارت بودند از توزیع درجه[2]، پیج‌رنک و هیتس از نظر کارآمدی در بازیابی متون زیست پزشکی بررسی‌کردند. تحلیل داده‌ها نشان داد که هرچند هر سه الگوریتم مذکور منجر به بهبود بازیابی متون زیست‌پزشکی می‌شوند؛ اما الگوریتم توزیع درجه نسبت به سایر موارد عملکرد بهتری دارد و پیج‌رنک و هیتس به‌ترتیب در رتبه‌های بعدی قرار می‌گیرند؛ آن‌ها هم‌چنین یک چارچوب ترکیبی احتمالی[3] را پیشنهاد کردند که اطلاعات استنادی را با مدل وزن‌دهی احتمالی محتوامحور[4] ترکیب می‌کند. بر اساس بررسی‌های آنان این چارچوب باعث بهبود بازیابی متون زیست‌پزشکی می‌شود. Yoon, Kim & Park(2016) سنجه‌های رتبه شباهت[5]، شباهت معکوس (آر.وی.اس رتبه شباهت)[6]، رتبه نفوذ (رتبه پی)[7] و رتبه رابط (رتبه سی)[8] را معرفی کردند که به‌ترتیب نسخه بازگشتی هم‌استنادی، زوج کتاب‌شناختی و امسلرند. رتبه رابط نمره شباهت را بر اساس تعداد رابط‌ها در گراف غیرجهت‌دار محاسبه می‌کرد، این سنجه‌ها در مقایسه با نسخه‌های غیربازگشتی خود عملکرد بهتری داشتند و از میان همه آن‌ها رتبه رابط بهترین عملکرد را داشت.

Colavizza, Boyack, Van Eck, & Waltman (2018) در چهار نشریه از رشته‌های مختلف شباهت جفت مقالات هم‌استناد در سطوح مختلف را مطالعه‌کردند. نتایج نشان داد که با کاهش سطح هم‌استنادی شباهت مقالات افزایش می‌یابد. Jiang and et al. (2019) با توجه به نیاز پژوهشگران به دسترسی به مقالات زیست‌فناوری و همچنین ناتوانی سنجه‌های متن‌محور در بازیابی همه مدارک مرتبط، الگوریتم گراف چندتایی محتوا[9] را پیشنهاد دادند. نتایج آزمون این الگوریتم روی داده‌های موجود در پاب‌مد نشان داد که این الگوریتم به‌لحاظ بازیابی و رتبه‌بندی مقالات به‌طور معناداری از الگوریتم سنتی متن‌محور پاب‌مد بهتر عمل می‌کند.

Janssens, Gwinn, Brockman, Powell, & Goodman (2020) روش جست‌وجوی استنادمحور را گسترش دادند. این روش که برای کارایی بیشتر نسبت به کلیدواژه سنتی طراحی شده بود هم‌استنادها[10] نام داشت. نتایج نشان داد این روش، روندی مؤثر و درست برای یافتن مقالات مرتبط است.

با مرور پژوهش‌های انجام‌شده پیشین، می‌توان گفت همان‌گونه که Thompson, Panchev & Oakes, (2015) اشاره می‌کنند گرچه به سنجه‌های ربط و شباهت‌یابی در بسیاری از نظام‌های بازیابی و فرایندهای پردازش زبان طبیعی برای یافتن مرتبط‌ترین مدارک و مدارک مشابه با مدرک اصلی نیاز است و موفقیت این سامانه‌ها نیز تاحد زیادی به عملکرد صحیح این سنجه‌ها وابسته است؛ اما عملکرد این سنجه‌ها از سامانه‌ای به سامانه دیگر می‌تواند متفاوت باشد؛ سنجه‌ای که در یک سامانه اثربخش بود، در سامانه دیگر مؤثر نبود. همچنین در عمل نیز هر یک از این سنجه‌ها به‌صورت مجزا در پایگاه‌های اطلاعاتی، کتابخانه‌های دیجیتال و موتورهای جست‌وجو به کار رفته‌اند؛ برای مثال، ساینس‌دایرکت[11]، از سنجه هم‌استنادی برای پیشنهاد مدارک مشابه استفاده می‌کند (Eto, 2013; Eto, 2019)؛ در پایگاه‌های وب آو ساینس[12] و اسکوپوس[13] از سنجه زوج کتاب‌شناختی Burnham, 2006; Ahlgren and Jarneving,)

1. PubMed Related Articles
2. Degree Distribution
3. probabilistic combination framework
4. Content-based probabilistic weighting model
5. Simrank
6. Reverse Simrank(rvs-Simrank)
7. Penetrating (P-rank)
8. Connectors rank (C-rank)
9. Content Tuple Graph Algorithm (CTGA)
10. CoCites
11. ScienceDirect
12. Web of Scienc(WoS)
13. Scopus

2008; Nicolaisen and Frandsen, 2012; Char and Ajiferuke, 2013)، در کتابخانه دیجیتالی ای.سی.ام[1] و در پایگاه استنادی سایت سیر[2] از هم‌استنادی برای پیشنهاد مدارک مشابه و مرتبط به کاربر استفاده می‌شود (Wanjantuk and Keane, 2004; Lu, Janssen, Milios, Japkowicz, and Zhang, 2007; Gipp and Beel, 2009; Eto, 2013; Eto, 2019). موتور جست‌وجوی گوگل از پیج‌رنک و موتور جست‌وجوی اسک[3] از هیتس برای بازیابی و رتبه‌بندی اسناد مرتبط با پرس‌وجو کاربر استفاده می‌نمایند (Thelwall, 2003; Goswami, Mantri, & Bhattacharya, 2017).

در پژوهش‌های اشاره‌شده توانایی هر یک از این سنجه‌ها برای تعیین میزان ربط مدارک با پرس‌وجو و سایر مدارک اغلب به‌تنهایی آزمون شده است و اطلاعات چندانی از میزان اثربخشی آن‌ها بر روی یک مجموعه مدارک واحد وجود ندارد. با بررسی این سنجه‌ها بر روی یک مجموعه یکسان می‌توان درک دقیق‌تری از توانایی هر یک از آن‌ها به دست آورد. براین اساس پژوهش حاضر قصد دارد تا توانایی این سنجه‌ها را برای تعیین میزان ربط مقاله‌ها بررسی کند. با توجه به مسئله گفته‌شده این پژوهش در پی پاسخ‌گویی به این سوال است که میزان همبستگی سنجه‌های هم‌استنادی، زوج کتاب‌شناختی، امسلر، پیج‌رنک و هیتس با ربط مقاله‌ها چقدر است؟

روش پژوهش

جامعه آماری پژوهش، مجموعه مقالاتی بود که با نام "زیرمجموعه دسترسی آزاد پاب‌مد سنترال"[4] در مجموعه آزمون سایترک[5] آمده است. این مجموعه مشتمل بر ۲۵۵۳۳۹ عنوان مقاله است. سایترک کد دسترسی آزاد جاوا برای محاسبه ۳۵ سنجه شباهت مبتنی بر استناد و متن را عرضه می‌کند که زوج کتاب‌شناختی، هم‌استنادی و امسلر از جمله سنجه‌های شباهت مبتنی بر استناد در این مجموعه هستند. در این پژوهش، برای فراهم‌کردن امکان مقایسه شاخص‌ها با یکدیگر مقالاتی انتخاب شد که شاخص‌های استنادمحور بررسی‌شده را دارا باشند؛ به‌عبارت دیگر، بر اساس سه سنجه استنادمحور هم‌استنادی، زوج کتاب‌شناختی و امسلر با سایر مقالات رابطه استنادی داشته باشند (پیوست ۱)؛ در مرحله بعد، با استفاده از نرم‌افزار شبیه‌ساز و مپ‌سرور[6] و پی.اچ.پی.مای‌ادمین[7] یک پایگاه مای.اس.کیو.ال[8] ایجاد شد؛ سپس، با استفاده از نسخه نمایشی[9] برخط مجموعه آزمون سایترک، یک خروجی تهیه شد و با وارد کردن آن به پایگاه مای.اس.کیو.ال حاوی مجموعه داده‌های پژوهش، ساختار اصلی جدول‌های آن ایجاد شد؛ در نهایت، با مطالعه کدهای لازم از بسته کد منبع[10] سایترک، تلاش شد تا این کدها با اعمال تغییرات ضروری، اجرا و نتایج حاصل در پایگاه مای.اس.کیو.ال وارد شود.

در زیرمجموعه‌ی دسترسی آزاد پاب‌مد سنترال مجموعه آزمایشی سایترک، برای سنجش میزان شباهت مدارک با پرس‌وجوها، از سرعنوان‌های موضوعی پزشکی(مش) تخصیص‌یافته به هریک از مدارک استفاده می‌شود. سرعنوان‌های موضوعی پزشکی(مش)، اصطلاحنامه چند سلسله‌مراتبی از توصیفگرهای موضوعی است. متخصصان

1. ACM Digital Library
2. CiteSeer
3. ASK
4. PubMed Central Open Access Subset (PMC OAS)
5. CITREC (https://dke.uni-wuppertal.de/en/projects/citrec.html)
6. WampServer
7. phpMyAdmin
8. MySQL
9. Demo
10. Source Code

کتابخانه ملی پزشکی (ان.ال.ام) آمریکا[1] این اصطلاحنامه را ایجاد و به‌صورت دستی، مناسب‌ترین توصیفگرها را به مدارک موجود در مجموعه دیجیتال ان.ال.ام مدلاین[2] اختصاص می‌دهند. سایترک، مش را به‌عنوان استاندارد طلایی و قضاوت موضوعی متخصصان در نظر می‌گیرد. این قضاوت‌ها می‌توانند برای ایجاد یک استاندارد طلایی برای ربط موضوعی مناسب باشند. به این صورت که هر چه تعداد اشتراک توصیفگرهای اختصاص یافته به دو مدرک بیشتر باشد آن دو مقاله با یکدیگر شباهت بیشتری دارند؛ به‌عبارتی، ربط آن دو مقاله با هم بیشتر است. این استاندارد طلایی، پژوهشگران را قادر می‌سازد تا میزان انعکاس ربط موضوعی سنجه‌های شباهت مبتنی بر استناد را بسنجند؛ بنابراین، این سنجه به‌عنوان شاخص استاندارد پژوهش حاضر در نظر گرفته شد و متغیر ملاک پژوهش بود. از این تکنیک در پژوهش‌های متعدد دیگری نیز برای سنجش شباهت مدارک استفاده شده است (Batet et al. 2010, Eto, 2012, Lin and Wilbur, 2007; Zhu et al., 2009 quated in Gipp, Meuschke & Lipinski, 2015). در این مجموعه آزمون، سنجه شباهت مش(ربط متخصص) با اندازه‌گیری میزان شباهت میان هر زوج پرسش - مدرک، با استفاده از سنجه ژاکارد و با تقسیم اشتراک اصطلاحات مش بر اجتماع آنان محاسبه شده است.

برای تعیین ربط مدارک با یکدیگر، باید کدنویسی انجام و کد مش اینترسکشن[3] اجرا می‌شد. بدین‌منظور، یک جدول جدید در پایگاه مای.اس.کیو.ال ایجاد و هر مدرک با کل مدارک موجود در مجموعه مقایسه شد؛ سپس میزان ربط مدارک بر اساس میزان اشتراک توصیفگرهای مش آنان با یکدیگر به دست آمد که به‌صورت نزولی مرتب و در پایگاه مای.اس.کیو.ال با نام سیم مش اینترسکشن[4] ذخیره شد؛ در مرحله بعد، با نوشتن پرس‌وجو به زبان اس.کیو.ال[5]، شبکه استنادی مجموعه به‌صورت کامل استخراج شد و در قالب یک فایل مقادیر جداشده با کاما[6] ذخیره شد؛ به‌دلیل بزرگ بودن فایل، استفاده از نرم‌افزارهای موجود برای محاسبه پیج‌رنک و هیتس امکان‌پذیر نبود، بنابراین یک برنامه به زبان پایتون نوشته شد که قادر بود این فایل بزرگ را باز و پردازش کند و نیز سنجه‌ها را اجرا کند. بدین ترتیب، اعداد مربوط به پیج‌رنک و هیتس (اعتبار و کانون) به‌صورت جداگانه محاسبه و ذخیره گردید.

برای انتخاب نمونه، ابتدا به‌صورت تصادفی، ۳۰ مقاله از میان کلیه‌ی مقالات جامعه پژوهش، به‌عنوان مدارک پایه انتخاب شد و هر یک به‌منزله یک پرس‌وجو در نظر گرفته شد. با جست‌وجوی هر مقاله در جدول‌های مای.اس.کیو.ال ساخته‌شده در مرحله قبل، مدارک مرتبط با هر یک از ۳۰ مدرک پایه بر اساس سنجه شباهت مش شناسایی و بازیابی گردید؛ سپس، برای آن مقالات میزان ربط مقالات، سنجه‌های استنادمحور هم‌استنادی، زوج کتاب‌شناختی، امسلر، پیج‌رنک، هیتس (اعتبار و کانون) از جدول‌های مربوط استخراج گردید. برای سنجش توانایی شش سنجه زوج کتاب‌شناختی، هم‌استنادی، امسلر، پیج‌رنک، هیتس (اعتبار) و هیتس (کانون) برای تعیین میزان ربط مقاله‌ها با یکدیگر، این شش سنجه به‌عنوان متغیرهای مستقل و ربط مقاله‌ها براساس شاخص مش به‌عنوان متغیر وابسته وارد نرم افزار اس.پی.اس.اس نسخه ۲۳ شد. قبل از انجام آزمون همبستگی، فرض نرمال بودن

1. U.S. National Library of Medicine (NLM)
2. NLM's digital collection MEDLINE
3. Mesh-Intersections
4. Sim-Mesh-Intersections
5. SQL
6. Comma-separated values (CSV)

متغیرها از طریق آزمون کولموگروف اسمیرنوف بررسی شد(جدول۱).

جدول۱- نتایج آزمون کولموگروف اسمیرنوف سطح نرمال بودن متغیرها

متغیرها شاخص آماری	هم استنادی	امسلر	زوج کتابشناختی	پیج رنک	هیتس(اعتبار)	هیتس(کانون)	ربط
تعداد نمونه	۱۵۰۰۰	۱۵۰۰۰	۱۵۰۰۰	۱۵۰۰۰	۱۵۰۰۰	۱۵۰۰۰	۱۵۰۰۰
میانگین	۰/۱۲۵	۰/۷۹٤	۰/٦٥۷	۰/۰۰۰۰۰۰۰۳	۰/۰۰۰۱	۰/۰۰۰٤	۰/۲۳
انحراف استاندارد	۰/۷۷٦	۲/۲۳۸	۱/۷٥۹	۰/۰۰۰۰۰۰۰۳	۰/۰۰۰٤	۰/۰۰۰۸	۰/۰۷۸
سطح معناداری	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱	۰/۰۰۱

با توجه به این‌که برای تمامی متغیرها سطح معناداری از پنج صدم کمتر است، می‌توان گفت داده‌های پژوهش نرمال نیستند و بر این اساس از آزمون همبستگی اسپیرمن استفاده شد.

یافته‌های پژوهش

برای سنجش رابطه بین سنجه‌های استنادمحور با میزان ربط مقاله‌ها، با توجه به نرمال نبودن توزیع داده‌ها از آزمون ضریب همبستگی اسپیرمن استفاده شد. نتایج این آزمون نشان داد که تمامی شش سنجه در سطح یک صدم همبستگی معناداری با میزان ربط مقاله‌ها دارد. همان‌گونه که جدول ۲ نشان می‌دهد بیشترین میزان همبستگی مربوط به سنجه امسلر است. میزان این همبستگی ۰/۱۷٤ است و با اطمینان ۹۹ درصد می‌توان گفت که بین سنجه امسلر و ربط مقاله‌ها رابطه معناداری وجود دارد. با توجه به مثبت بودن جهت رابطه می‌توان گفت که با افزایش مقدار سنجه امسلر، میزان ربط مقاله‌ها نیز افزایش پیدا می‌کند($r=۰/۱۷٤$ ، $p \leq ۰/۰۱$).

جدول۲- نتایج آزمون همبستگی اسپیرمن برای سنجش رابطه بین سنجه‌های استنادمحور با میزان ربط مقاله‌ها

	متغیر	۱	۲	۳	٤	٥	٦	۷
۱	هیتس(اعتبار)	۱						
۲	هیتس(کانون)	۰/٤٥٦**	۱					
۳	پیج رنک	۰/۸۳٦**	۰/۲٤۱**	۱				
٤	هم استنادی	۰/۳٥٤**	۰/۲٤۳**	۰/۲٦۸**	۱			
٥	زوج کتاب‌شناختی	۰/۲۷۹**	۰/٥٦٤**	۰/۱٥۱**	۰/۳۰٥**	۱		
٦	امسلر	۰/۳۳۱**	۰/٥٦٦**	۰/۱۹٦**	۰/٤٥۹**	۰/۹٦٥**	۱	
۷	ربط	۰/۱٥۳**	۰/۰۸۹**	۰/۱۱۳**	۰/۰۸۹**	۰/۱٦۱**	۰/۱۷٤**	۱

**p ≤۰/۰۱

پس از سنجه امسلر، بیشترین میزان همبستگی، مربوط به سنجه زوج کتاب‌شناختی است. با توجه به ضریب همبستگی به‌دست‌آمده می‌توان گفت که بین دو سنجه زوج کتاب‌شناختی و ربط مقاله‌ها با اطمینان ۹۹ درصد رابطه معناداری وجود دارد؛ همچنین، با توجه به مثبت بودن جهت رابطه می‌توان گفت که با افزایش سنجه زوج کتاب‌شناختی، ربط مقاله‌ها نیز افزایش می‌یابد(p ≤۰/۰۱ ، r=۰/۱٦۱).

پس از سنجه‌های امسلر و زوج کتاب‌شناختی، بیشترین همبستگی بین متغیر هیتس(اعتبار) با ربط مقاله‌ها است. با توجه به مقدار ضریب همبستگی می‌توان گفت که با احتمال خطای کمتر از یک صدم درصد بین دو متغیر رابطه مثبت و معناداری وجود دارد. به‌عبارتی با افزایش هر واحد متغیر اعتبار هیتس، به میزان ۰/۱٥۳ متغیر ربط مقاله‌ها افزایش خواهد داشت (p ≤۰/۰۱ ، r=۰/۱٥۳).

در مرتبه چهارم میزان همبستگی ربط مقاله‌ها با سنجه‌های استنادی، متغیر پیج‌رنک قرار دارد که با توجه به مقدار ضریب همبستگی می‌توان گفت که با میزان خطای کمتر از یک صدم درصد بین دو متغیر رابطه مثبت و معناداری وجود دارد. به‌عبارتی با افزایش هر واحد متغیر پیج‌رنک، به میزان ۰/۱۱۳ متغیر ربط مقاله‌ها افزایش خواهد داشت (p ≤۰/۰۱ ، r=۰/۱۱۳).

با توجه به یافته‌های به‌دست‌آمده می‌توان گفت که کم‌ترین میزان همبستگی با ربط مقاله‌ها، مربوط به سنجه‌های هم‌استنادی و هیتس(کانون) است. براین اساس می‌توان گفت که سنجه امسلر که ترکیبی از سنجه‌های زوج کتاب‌شناختی و هم‌استنادی است، بهتر از سایر سنجه‌ها، می‌تواند میزان ربط مقاله‌ها را تعیین کند.

بحث و نتیجه‌گیری

در این پژوهش توانایی سنجه‌های استنادمحور هم‌استنادی، زوج کتاب‌شناختی، امسلر، پیج‌رنک و هیتس (اعتبار و کانون) برای تعیین میزان ربط مقاله‌ها تخمین زده شد. نتایج نشان داد که در بین سنجه‌های استناد محور، سنجه امسلر، زوج کتاب‌شناختی به‌ترتیب بیشترین توان را در تعیین میزان ربط مقاله‌ها داشتند. سنجه امسلر، هر دو سنجه هم‌استنادی و زوج کتاب‌شناختی را در خود لحاظ می‌کند. در این سنجه، به جای این‌که از هریک از دو شاخص پیشین به‌تنهایی استفاده شود، آن دو با یکدیگر ترکیب می‌شوند. به این ترتیب از تمامی علائمی که به‌نحوی نشانگر ارتباط موضوعی مدارک هستند استفاده می‌شود (Bichteler and Eaton, 1980). این یافته با نتایج به دست آمده از پژوهش Bichteler and Eaton(1980) همسو است. نتایج این پژوهش نیز نشان داد استفاده از این مقیاس نسبت به زوج

کتاب‌شناختی، منجر به بهبود بازیابی مدارک مرتبط می‌شود. گرچه به نظر می‌رسد از این سنجه نسبت به دو سنجه هم‌استنادی و زوج کتاب‌شناختی، در سامانه‌های اطلاعاتی کمتر استفاده شده است، بنابراین می‌توان این سنجه را بیش از پیش مدنظر قرار داد.

سنجه زوج کتاب‌شناختی، همان‌گونه که در مقدمه نیز بیان شد، علاوه بر این‌که عمدتاً در متون به‌عنوان یک سنجه کارآمد برای تعیین مدارک مرتبط یاد می‌شود؛ پایگاه‌های اطلاعاتی همانند وب آو ساینس و اسکوپوس نیز از آن برای بازیابی مدارک مرتبط استفاده می‌کنند (Burnham, 2006; Ahlgren and Jarneving, 2008; Nicolaisen and Frandsen, 2012; Char and Ajiferuke, 2013). براین اساس، همانند نتایج پژوهش‌های پیشین (برای مثال، Boyack and Klavans, 2010; Ahlgren, Chen, Colliander, & van Eck, 2020) که نشان داده بودند که عملکرد سنجه زوج کتاب‌شناختی نسبت به هم‌استنادی برای شناسایی شباهت میان مدارک بهتر است، در این پژوهش نیز یافته‌ها تأییدکننده این مهم بود.

در کنار دو سنجه امسلر و زوج کتاب‌شناختی، نتایج نشان داد دو سنجه هیتس (اعتبار) و پیج‌رنک نیز توانایی مناسبی برای تعیین میزان ربط مقاله‌ها دارند. این دو سنجه که از سنجه‌های جدیدتر استنادمحور هستند، با استفاده از ساختار پیوندی شبکه‌های مقالات، رویت‌پذیری و اعتبار را به‌طور هم‌زمان لحاظ می‌کنند. این یافته، یافته‌های پژوهش‌های پیشین مبنی بر قدرت الگوریتم‌های محیط وب از جمله پیج‌رنک و هیتس را در بازیابی و رتبه‌بندی مدارک مرتبط تأیید می‌کند (Liu and Lin, 2007; Chen and et al., 2007, Lin, 2008, Yin, Huang, Hu, & Li 2009; Zhuge and Zhang, 2010; Yin and et al.,2011; Liu, Zhang, & Guo, 2012). الگوریتم هیتس، علاوه بر ساختار گراف، از روش محتواکاوی نیز استفاده می‌کند و وابسته به پرس‌وجو است؛ بنابراین، در مقایسه با الگوریتم پیج‌رنک که صرفا ساختار گراف را در نظر گرفته و مستقل از پرس‌وجو است، در تعیین ربط مقاله‌ها موفق‌تر بوده است.

هم‌چنین نتایج نشان داد در بین سنجه‌های استنادی سنتی، سنجه هم‌استنادی کمترین توان را در تعیین میزان ربط مقاله‌ها داشت. عملکرد ضعیف‌تر سنجه هم‌استنادی نسبت به سایر سنجه‌ها در پژوهش‌های پیشین نیز به‌دست‌آمده است و قابل انتظار بود (Shibata and et al., 2009; Ahlgren and et al., 2020). این مسئله به این دلیل است که مقاله تازه انتشار یافته، هنوز نتوانسته است استناد قابل قبولی دریافت کند و بر این اساس نمی‌تواند با سایر مقالات هم‌استناد شود. البته باید به این نکته هم توجه کرد که مجموعه آزمون سایترک پویا نیست و قابلیت افزایش تعداد مقالات هم‌استناد نیز در آن وجود ندارد.

در این پژوهش گرچه تمامی سنجه‌های استنادمحور با ربط مقاله‌ها براساس سنجه مش، همبستگی در سطح یک صدم داشتند، اما ضرایب همبستگی چندان قوی نبود. در متون بر این نکته تأکید شده است که کاربران- به‌ویژه متخصصان موضوعی و پژوهشگران- به‌دلایل مختلف نیازمند سامانه‌های بازیابی کارآمدی هستند که تا حد امکان فقط مرتبط‌ترین و مهم‌ترین مدارک را در حداقل زمان برایشان بازیابی نمایند؛ برای مثال، بر اساس پژوهش‌ها کاربران تنها به چند نتیجه نخست توجه می‌کنند و سایر نتایج بازیابی‌شده را نادیده می‌گیرند (Bar-Ilan, et al. 2006; Lewandowski, 2008; Bar-Ilan, et al., 2009; Lewandowski, 2017). از جمله مهم‌ترین عواملی که بر تحقق این امر تأثیر مستقیم دارد، کیفیت سنجه‌هایی است که ربط بین مقاله‌ها با یکدیگر یا مدارک با پرس‌وجو را تعیین می‌کند؛ به‌عبارت دیگر، هرچه سنجه‌های بهتری به‌کار گرفته شود، می‌توان انتظار داشت

نتایج بازیابی بهتر شود. ضرایب همبستگی پایین سنجه‌های استنادمحور، می‌تواند نشان‌دهنده این باشد که استفاده از این سنجه‌ها به‌تنهایی برای یافتن مدارک مرتبط کفایت نمی‌کند و بایستی توانایی سایر سنجه‌ها مانند سنجه‌های متن‌محور و حتی ترکیب سنجه‌های متن‌محور و استنادمحور با یکدیگر مدنظر قرار گیرد. نتایج پژوهش‌های مختلف نیز نشان می‌دهد مدارک بازیابی‌شده با استفاده از سنجه‌های متن‌محور، با مدارکی که از طریق سنجه‌های استنادمحور بازیابی می‌شوند، متفاوت هستند (Mccain, 1989; Pao and Worthen, 1989, Ahlgren and Jarneving, 2008). توصیه به استفاده ترکیبی از دو گروه سنجه‌های متن‌محور و استنادمحور در تعدادی از پژوهش‌های پیشین تاکید شده است (Yin and et al., 2009; Boyack and Klavans, 2010; Reyhani Hamedani and et al., 2013; Reyhani Hamedani, Kim, & Kim, 2016; Jiang and et al., 2019; Ahlgren and et al., 2020; Boyack and Klavans, 2020).

سنجه‌های شباهت‌یابی مانند سنجه‌های مبتی بر استناد برای پیشنهاد مقالات مشابه در پایگاه‌های اطلاعاتی و در بافتارهای مختلف شامل موتورهای جست‌وجو، پایگاه‌های اطلاعاتی و استنادی، سامانه‌های پیشنهاددهنده و حتی کتابخانه‌های دیجیتالی کاربردهای بسیاری دارند و می‌توان از پتاسیل این سنجه‌ها، برای پیشنهاد مقالات مشابه و رتبه‌بندی نتایج بازیابی بهره برد.

در نهایت، باید به این امر اشاره کرد که با توجه به خلاءهای پژوهشی موجود در زمینه مقایسه سنجه‌های شباهت‌یابی، در این پژوهش، توانایی تعدادی از این سنجه‌ها برای تعیین میزان ربط مقاله‌ها بررسی شد؛ لازم است نتایج ذکرشده بر روی دیگر انواع منابع اطلاعاتی و دیگر مجموعه دادگان به دست آید تا بتوان با قطعیت بیشتری در مورد نتایج به‌دست‌آمده قضاوت کرد. الگوریتم‌ها و مدل‌های پیش‌بینی حاصل از ترکیب این سنجه‌ها در کنار یکدیگر می‌بایست در پژوهش‌های آتی مدنظر قرار گیرد.

References مآخذ

- Agrahri, A. K., Manickam, D. A. T., & Riedl, J. (2008, October). Can people collaborate to improve the relevance of search results?. *In Proceedings of the 2008 ACM conference on Recommender systems* (pp. 283-286). ACM. https://doi.org/10.1145/1454008.1454052
- Ahlgren, P., Chen, Y., Colliander, C., & van Eck, N. J. (2020). Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of PubMed publications. *Quantitative Science Studies*, *1*(2), 714-729. https://doi.org/10.1162/qss_a_00027
- Ahlgren, P., & Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping. *Scientometrics*, 76(2), 273-290. https://doi.org/10.1007/s11192-007-1935-1
- Amer, A. A., & Abdalla, H. I. (2020). A set theory based similarity measure for text clustering and classification. *Journal of Big Data*, *7*(1), 1-43. https://doi.org/10.1186/s40537-020-00344-3

- Bar-Ilan, J., Keenoy, K., Levene, M., & Yaari, E. (2009). Presentation bias is significant in determining user preference for search results—A user study. *Journal of the American Society for Information Science and Technology*, 60(1), 135-149. https://doi.org/10.1002/asi.20941
- Bar-Ilan, J., Keenoy, K., Yaari, E., & Levene, M. (2007). User rankings of search engine results. *Journal of the American Society for Information Science and Technology*, 58(9), 1254-1266. https://doi.org/10.1002/asi.20608
- Bar-Ilan, J., Levene, M., & Mat-Hassan, M. (2006). Methods for evaluating dynamic changes in search engine rankings: a case study. *Journal of Documentation*, 62(6), 708-729. https://doi.org/10.1108/00220410610714930
- Bichteler, J., & Eaton, E. A. (1980). The combined use of bibliographic coupling and cocitation for document retrieval. *Journal of the American Society for Information Science*, 31(4), 278-282. https://doi.org/10.1002/asi.4630310408
- Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12), 2389-2404. https://doi.org/10.1002/asi.21419
- Boyack, K. W., & Klavans, R. (2020). A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. *Quantitative Science Studies*, 1-16. https://doi.org/10.1162/qss_a_00085
- Burnham, J. F. (2006). Scopus database: A review. *Biomedical Digital Libraries*, 3(1), 1-8. https://doi.org/10.1186/1742-5581-3-1
- Char, D. C., & Ajiferuke, I. (2013, October). Comparison of the effectiveness of related functions in Web of Science and Scopus. *In Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI.* https://doi.org/10.29173/cais353
- Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm, *Journal of Informetrics*, 1(1), 8-15. https://doi.org/10.1016/j.joi.2006.06.001
- Colavizza, G., Boyack, K. W., Van Eck, N. J., & Waltman, L. (2018). The closer the better: Similarity of publication pairs at different cocitation levels. *Journal of the Association for Information Science and Technology*, *69*(4), 600-609. https://doi.org/10.1002/asi.23981
- Devi, P., Gupta, A., & Dixit, A. (2014). Comparative study of hits and pagerank link based ranking algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(2), 5749-5754.
- Eminagaoglu, M. (2020). A new similarity measure for vector space models in text classification and information retrieval. *Journal of Information Science*. https://doi.org/10.1177/0165551520968055
- Eto, M. (2013). Evaluations of context-based co-citation searching. *Scientometrics*, 94(2), 651-673. https://doi.org/10.1007/s11192-012-0756-z
- Eto, M. (2019). Extended co-citation search: Graph-based document retrieval on a co-citation network containing citation context information. *Information Processing & Management*, 56(6), 102046. https://doi.org/10.1016/j.ipm.2019.05.007
- Gipp, B., & Beel, J. (2009). Citation proximity analysis (CPA): a new approach for identifying related work based on co-citation analysis. In *ISSI'09: 12th International Conference on Scientometrics and Informetrics* (pp. 571-575). Retrieved June 20, 2020 from https://isg.beel.org/pubs/Citation%20Proximity%20Analysis%20(CPA)%20-%20A%20new%20approach%20for%20identifying%20related%20work%20based%20on%20Co-Citation%20Analysis%20--%20preprint.pdf.
- Gipp, B., Meuschke, N., & Lipinski, M. (2015). CITREC: An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central. In *iConference 2015*, Newport Beach, California. https://doi.org/10.5281/zenodo.3547372

- Goswami, P., Mantri, M., & Bhattacharya, M. (2017). Web Page Ranking Based On User Query. *International Journal of HIT Transaction on ECCN*, 3(2A), 21-33.
- Hajian, A., & CheshmehSohrabi, M. (2020). Ranking and Relevance in Noormags and RICEST Databases. *National Studies on Librarianship and Information Organization*, *31*(3), 72-92. DOI: 10.30484/nastinfo.2020.2472.1934 [In Persian]

[حاجیان، آزاده؛ چشمه‌سهرابی، مظفر (۱۳۹۹). رتبه‌بندی و ربط مقالات در پایگاه‌های اطلاعاتی نورمگز و رایسست. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۳۱(۳)، ۹۲-۷۲].

- Hariri, N. (2011). Relevance ranking on Google: Are top ranked results really considered more relevant by the users?. *Online Information Review*, 35(4), 598-610. https://doi.org/10.1108/14684521111161954
- Herskovic, J. R., & Bernstam, E. V. (2005). Using incomplete citation data for MEDLINE results ranking. In *AMIA Annual Symposium proceedings* (Vol. 2005, p. 316-320). American Medical Informatics Association.
- Janssens, A. C. J., Gwinn, M., Brockman, J. E., Powell, K., & Goodman, M. (2020). Novel citation-based search method for scientific literature: a validation study. *BMC medical research methodology*, *20*(1), 1-11. https://doi.org/10.1186/s12874-020-0907-5
- Jiang, X., Sun, X., Yang, Z., Zhuge, H., & Yao, J. (2016). Exploiting Heterogeneous Scientific Literature Networks to Combat Ranking Bias: Evidence From the Computational Linguistics Area. *Journal of the American Society for Information Science and Technology*, 67(7), 1679-1702. https://doi.org/10.1002/asi.23463
- Jiang, T., Zhang, Z., Zhao, T., Qin, B., Liu, T., Chawla, N. V., & Jiang, M. (2019, November). CTGA: Graph-based Biomedical Literature Search. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 395-400). IEEE. DOI: 10.1109/BIBM47256.2019.8983173
- Jouili S., Tabbone S., Valveny E. (2010) Comparing Graph Similarity Measures for Graphical Recognition. In: Ogier JM., Liu W., Lladós J. (eds) Graphics Recognition. *Achievements, Challenges, and Evolution. GREC 2009. Lecture Notes in Computer Science*, vol 6020. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13728-0_4
- Kirsch, S. M., Gnasa, M., Won, M., & Cremers, A. (2008). From PageRank to Social Rank: Authority-Based Retrieval in Social Information Spaces. In *Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectively* (pp. 134-154). IGI Global.
- Kessler, M.M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14(1), 10–25. https://doi.org/10.1002/asi.5090140103
- Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5), 604-632.
- Lewandowski, D. (2008). The retrieval effectiveness of web search engines: considering results descriptions. *Journal of documentation*, 64(6), 915- 937. Retrieved June 20, 2020 from https://arxiv.org/ftp/arxiv/papers/1511/1511.05800.pdf.
- Lewandowski D. (2017) Is Google Responsible for Providing Fair and Unbiased Results? In: Taddeo M., Floridi L. (eds) *The Responsibilities of Online Service Providers. Law, Governance and Technology Series*, vol 31, pp.61-77 Springer, Cham. https://doi.org/10.1007/978-3-319-47852-4_4
- Lin, J. (2008). PageRank without hyperlinks: Reranking with PubMed related article networks for biomedical text retrieval. *BMC Bioinformatics*, 9, 270. https://doi.org/10.1186/1471-2105-9-270
- Lin, Y. S., Jiang, J. Y., & Lee, S. J. (2013). A similarity measure for text classification and clustering. *IEEE transactions on knowledge and data engineering*, *26*(7), 1575-1590. doi: 10.1109/TKDE.2013.19.

- Liu, Y., & Lin, Y. (2007, October). Supervised HITS algorithm for MEDLINE citation ranking. In *Bioinformatics and Bioengineering (BIBE)*, 2007. *Proceedings of the 7th IEEE International Conference on* (pp. 1323-1327). IEEE.
- Liu, X., Zhang, J., & Guo, C. (2012, October). Full-text citation analysis: enhancing bibliometric and scientific publication ranking. In *Proceedings of the 21st ACM international conference on Informationand knowledge management* (pp. 1975-1979). ACM. https://doi.org/10.1145/2396761.2398555
- Lu, W., Janssen, J., Milios, E., Japkowicz, N., & Zhang, Y. (2007). Node similarity in the citation graph. *Knowledge and Information Systems*, 11(1), 105-129. https://doi.org/10.1007/s10115-006-0023-9
- Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing and Management*, 44:800-810. https://doi.org/10.1016/j.ipm.2007.06.006
- Mccain, W. (1989). Descriptor and Citation Retrieval in the Medical Behavioral Sciences Literature: Retrieval Overlaps and Novelty Distribution. *Journal of the American Society for Information Science*, 40(2), 110-114. https://doi.org/10.1002/(SICI)1097-4571(198903)40:2<110::AID-ASI5>3.0.CO;2-T
- McGill, M., Koll, M., & Noreault, T. (1979). An evaluation of factors affecting document ranking by information retrieval systems, Syracuse, NY: School of Information Studies, Syracuse University.
- Nicolaisen, J., & Frandsen, T. F. (2012). Consensus formation in science modeled by aggregated bibliographic coupling. *Journal of Informetrics*, *6*(2), 276-284. DOI: 10.1016/j.joi.2011.08.001
- Neshat, N. (2003). Hermeneutice and Information Retrieval. *Informology*, 2, 31-46. http://ensani.ir/file/download/article/20110108110840-0%20(1)

[نشاط، نرگس (۱۳۸۲). هرمنوتیک و بازیابی اطلاعات. *اطلاع‌شناسی*، ۲، ۳۱-۴۶].

- Pao, M. L., & Worthen, B. (1989). Retrieval Effectiveness by Semantic and Citation Searching. *Journal of the american society for information science*, 40(4), 226-235. https://doi.org/10.1002/(SICI)1097-4571(198907)40:4<226::AID-ASI2>3.0.CO;2-6
- Reyhani Hamedani, M., & Kim, S. W. (2021). On Investigating Both Effectiveness and Efficiency of Embedding Methods in Task of Similarity Computation of Nodes in Graphs. *Applied Sciences*, *11*(1), 162. https://doi.org/10.3390/app11010162
- Reyhani Hamedani, M., Kim, S. W., & Kim, D. J. (2016). SimCC: A novel method to consider both content and citations for computing similarity of scientific papers. *Information Sciences*, 334, 273-292. https://doi.org/10.1016/j.ins.2015.12.001
- Reyhani Hamedani, M.., Lee, S. C., & Kim, S. W. (2013, October). On combining text-based and link-based similarity measures for scientific papers. In *Proceedings of the 2013 Research in Adaptive and Convergent Systems* (pp. 111-115). ACM. DOI:10.1145/2513228.2513321
- Sadein Khorram, S., & Abbaspour, J. (2019). Article Ranking by Recommender Systems vs. Users' Perspectives. *National Studies on Librarianship and Information Organization*, *30*(3), 46-57. DOI: 10.30484/nastinfo.2019.2187.1838 http://nastinfo.nlai.ir/article_2346.html

[سعدین‌خرم، صبا؛ عباس‌پور، جواد (۱۳۹۸). سنجش رتبه‌بندی سامانه‌های پیشنهاددهندهٔ مقاله در تقابل با رتبه‌بندی کاربران. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۳۰(۳)، ۵۷-۴۶] .

- Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for information Science and Technology*, *60*(3), 571-580. https://doi.org/10.1002/asi.20994
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. https://doi.org/10.1002/asi.4630240406
- Smith, L. C. (1981). Citation Analysis. *Library Trends*, *30*(1), 83-106.

- Su, C., Pan, Y., Zhen, Y., Ma, Z., Yuan, J., Guo, H., ... & Wu, Y. (2011). PrestigeRank: A new evaluation method for papers and journals. *Journal of Informetrics*, *5*(1), 1-13. https://doi.org/10.1016/j.joi.2010.03.011
- Thelwall, M. (2003). Can Google's PageRank be used to find the most important academic Web pages?. *Journal of Documentation*, 59(2), 205-217. https://doi.org/10.1108/00220410310463491
- Thompson, V. U., Panchev, C., & Oakes, M. (2015, November). Performance evaluation of similarity measures on similar and dissimilar text retrieval. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)* (Vol. 1, pp. 577-584). IEEE.
- Torres, G. J., Basnet, R. B., Sung, A. H., Mukkamala, S., & Ribeiro, B. M. (2009). A similarity measure for clustering and its applications. *Int J Electr Comput Syst Eng*, *3*(3), 164-170.
- Wanjantuk, P., & Keane, J. A. (2004, October). Finding related documents via communities in the citation graph. In *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on* (Vol. 1, pp. 445-450). IEEE. doi: 10.1109/ISCIT.2004.1412885.
- Yin, X., Huang, X., Hu, Q., & Li, Z. (2009, April). Boosting biomedical information retrieval performance through citation graph: An empirical study. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 949-956). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-01307-2_100
- Yin, X., Huang, J. X., & Li, Z. (2011). Mining and modeling linkage information from citation context for improving biomedical literature retrieval. *Information processing & management*, 47(1), 53-67. https://doi.org/10.1016/j.ipm.2010.03.010
- Yoon, S. H., Kim, S. W., & Park, S. (2016). C-Rank: A link-based similarity measure for scientific literature databases. *Information Sciences*, 326, 25-40. https://doi.org/10.1016/j.ins.2015.07.036
- Zhang, J. & Korfhage, R. (1999).A Distance and Angle Similarity Measure Method. *Journal of the American Society for Information Science*, 50(9),772–778. https://doi.org/10.1002/(SICI)1097-4571(1999)50:9<772::AID-ASI5>3.0.CO;2-E
- Zhuge, H., & Zhang, J. (2010). Topological centrality and its e-Science applications. *Journal of the Association for Information Science and Technology*, 61(9), 1824-1841.
- https://doi.org/10.1002/asi.21353

پیوست ۱

شرایط انتخاب مقالات:

۱- مدرک انتخاب‌شده دارای شباهت امسلر با مدارک دیگر باشد(مجموعهA)

۲- مدرک انتخاب‌شده دارای شباهت زوج کتابشناختی با مدارک دیگر باشد (مجموعهB)

۳- مدرک انتخاب‌شده دارای شباهت هم‌استنادی با مدارک دیگر باشد(مجوعه C)

۴- مدرک انتخاب‌شده دارای شباهت مش با مدارک دیگر باشد(مجموعهM)

۵- بین مجموعه‌هایA ، B، Cو M اشتراک وجود داشته باشد.

پرسشی که با توجه به شرایط بالا برای انتخاب مدارک استفاده شد به شرح زیر است:

```
select distinct sim_amsler.document1,sim_amsler.document2
            from
        document,    sim_amsler,    sim_bibco,    sim_cocit,
        sim_mesh_intersections_27000
            where document.pmid = sim_amsler.document1
            and        ((sim_amsler.document1        =        sim_bibco.document1        and
 sim_amsler.document2 = sim_bibco.document2) )
            and ((sim_amsler.document1 = sim_cocit.document1 and sim_amsler.document2
 = sim_cocit.document2) )
            and (sim_amsler.document1 = sim_mesh_intersections_27000.document1 )
            and sim_bibco.value>0 and sim_cocit.value>0
            and document.id=@Random
```

هر دیتاست شامل ۳۰ مدرک بود و به ازای هر مدرک، بر اساس سنجه‌های شباهت یادشده، تعدادی مدرک دیگر وجود داشت که دارای میزانی از شباهت بودند. برای ایجاد شبکه استنادی برای استفاده در الگوریتم‌های پیج رنک و هیتس، از جدول refrence که حاوی مدارک استنادکننده و استنادشونده است استفاده شد و از بین آن‌ها مدارکی انتخاب شد که هم خود مدرک و هم استنادهای آن دارای سنجه شباهت مش باشند. شبکه استنادی به‌دست‌آمده گرافی با بیش از دو میلیون و سیصد هزار یال (ارتباط استنادی) بود. پرسشی که این شبکه استنادی را ایجاد کرد به‌شرح زیر است:

```
select distinct docpmid,          refpmid from
 (select distinct document as docpmid,refpmid from reference  where refpmid is not NULL
 and refpmid in
(select distinct doc1 from (
select distinct document2 as doc1 from citrec2.sim_mesh_intersections_27000
union
select distinct document1 as doc1 from citrec2.sim_mesh_intersections_27000
)t)
union
select distinct document as docpmid,refpmid from reference  where refpmid is not NULL  and
  document in
(select distinct doc1 from (
select distinct document2 as doc1 from citrec2.sim_mesh_intersections_27000
union
select distinct document1 as doc1 from citrec2.sim_mesh_intersections_27000
)t)  )tt
```