Journal of Information and Communication Technology

Volume 16, Issue 59-60, Spring and Summer 2024, pp. 240-263

Survey on the Applications of the Graph Theory in the Information Retrieval

Maryam Piroozmand[1], Amir Hosein Keyhanipour[2*], Ali Moeini[3]

[1] Faculty of Engineering Science, School of Engineering, University of Tehran, Tehran, Iran
[2] Computer Engineering Department, Faculty of Engineering, College of Farabi, University of Tehran, Tehran, Iran
[3] Faculty of Engineering Science, School of Engineering, University of Tehran, Tehran, Iran

Received: 05 June 2023, Revised: 12 August 2023, Accepted: 05 September 2023
Paper type: Review

Abstract

Due to its power in modeling complex relations between entities, graph theory has been widely used in dealing with real-world problems. On the other hand, information retrieval has emerged as one of the major problems in the area of algorithms and computation. As graph-based information retrieval algorithms have shown to be efficient and effective, this paper aims to provide an analytical review of these algorithms and propose a categorization of them. Briefly speaking, graph-based information retrieval algorithms might be divided into three major classes: the first category includes those algorithms which use a graph representation of the corresponding dataset within the information retrieval process. The second category contains semantic retrieval algorithms which utilize the graph theory. The third category is associated with the application of the graph theory in the learning to rank problem. The set of reviewed research works is analyzed based on both the frequency as well as the publication time. As an interesting finding of this review is that the third category is a relatively hot research topic in which a limited number of recent research works are conducted.

Keywords: Graph Theory, Information Retrieval, Learning to Rank, Knowledge Graph, Graph-based Dataset Representation.

* Corresponding Author's email: keyhanipour@ut.ac.ir

دوفصلنامه
فناوری اطلاعات و ارتباطات ایران

سال شانزدهم، شماره های ۵۹ و ۶۰، بهار و تابستان ۱۴۰۳، صفحه ۲۴۰ الی ۲۶۳

بررسی کاربردهای نظریه گراف در بازیابی اطلاعات

مریم پیروزمند[۱]، امیرحسین کیهانی پور[۲*]، علی معینی[۳]

[۱] گروه الگوریتم و محاسبات، دانشکده علوم مهندسی، دانشکدگان فنی، دانشگاه تهران، تهران، ایران

[۲] گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشکدگان فارابی، دانشگاه تهران، تهران، ایران

[۳] گروه الگوریتم و محاسبات، دانشکده علوم مهندسی، دانشکدگان فنی، دانشگاه تهران، تهران، ایران

تاریخ دریافت: ۱۴۰۲/۰۳/۱۵ تاریخ بازبینی: ۱۴۰۲/۰۵/۲۱ تاریخ پذیرش: ۱۴۰۲/۰۶/۱۴

نوع مقاله: مروری

چکیده

نظریه گراف بواسطه توانمندی در مدلسازی روابط پیچیده بین عناصر در مسائل مختلف، بصورت گسترده مورد استفاده قرار گرفته است. از سوی دیگر، بازیابی اطلاعات یعنی استخراج اطلاعات مورد نیاز کاربر، به عنوان یکی از مسائل مهم در دنیای الگوریتم و محاسبات مطرح است. با توجه به کارآمدی راهکارهای مبتنی بر گراف در بازیابی اطلاعات، این مقاله، به بررسی تحلیلی و دسته‌بندی کاربردهای نظریه گراف در بازیابی اطلاعات، می‌پردازد. این راهکارها در سه دسته کلی، قابل تفکیک هستند؛ دسته نخست، شامل الگوریتم‌هایی می‌باشد که در آنها از بازنمایی گرافی دادگان در فرآیند بازیابی اطلاعات، استفاده می‌شود. دسته دوم پژوهش‌ها، به حل مسئله بازیابی معنایی اطلاعات با استفاده از نظریه گراف می‌پردازند و نهایتا دسته سوم، مربوط به یادگیری رتبه‌بندی با استفاده از نظریه گراف است. این سه دسته بصورت جزئی‌تر در هشت زیردسته، دسته‌بندی شده‌اند. همچنین از منظر آماری، پژوهش‌های صورت گرفته در هر دسته بر اساس تعداد و سال انتشار، بررسی شده‌اند. از جمله یافته‌های این مطالعه، این است که دسته سوم، هم از نظر تعداد پژوهش‌ها و نیز سال انتشار آنها، شاخه نوظهوری محسوب می‌شود و می‌تواند حوزه تحقیقاتی جالب توجهی برای محققان محسوب شود.

کلیدواژگان: نظریه گراف، بازیابی اطلاعات، یادگیری رتبه‌بندی، گراف دانش، بازنمایی گرافی دادگان.

* رایانامهٔ نویسنده مسؤول: keyhanipour@ut.ac.ir

۱- مقدمه

بازیابی اطلاعات به عنوان یکی از بنیادی‌ترین مسائل دنیای فناوری اطلاعات، فرآیندی است که هدف آن فراهم نمودن امکان دستیابی کاربر به اطلاعات مورد نیاز، می‌باشد. با این وجود، فرآیند بازیابی اطلاعات، با چالشهای زیادی مواجه است. بخشی از این معضلات، بواسطه حجم بسیار زیاد اطلاعات مورد جستجو، بروز می‌کند. از سوی دیگر غیر ساخت‌یافته بودن اطلاعات و مسائل مربوط به تشخیص دقیق نیازمندی‌های کاربران، شناسایی اطلاعات مرتبط با نیاز آنان را امری دشوار ساخته است. این چالشها در خصوص سامانه‌های بازیابی اطلاعات وب، به ویژه جویشگرهای وب و نیز سامانه‌های توصیه‌گر، بسیار جدی‌تر است. از این رو، بهره‌گیری از قابلیت‌های دیگر ابزارهای نظری و عملیاتی به منظور ارتقا کیفیت بازیابی اطلاعات، همواره یکی از علاقه‌مندی‌های پژوهشگران و صاحب‌نظران بوده است. بصورت مشخص، نظریه گراف، به عنوان رویکردی برای مدلسازی روابط میان عناصر درگیر در مسائل مختلف که به لحاظ نظری، پیشینه و غنای بالایی دارد، به عنوان یکی از رویکردهای ارتقای بازیابی اطلاعات و ارائه مدل‌های کارآمد در این زمینه، همواره مورد توجه جامعه پژوهشگران بوده است.

در این نوشتار، بصورت مشخص، به مطالعه، بررسی و دسته‌بندی تحقیقات صورت گرفته در زمینه بهره‌گیری از نظریه گراف در فرآیند بازیابی اطلاعات، خواهیم پرداخت. نویسندگان این مقاله بر این باور هستند که این کار می‌تواند ضمن تبیین کارآمدی استفاده از نظریه گراف در بازیابی اطلاعات، دیدگاه مناسبی در خصوص انواع کاربرد این نظریه در حل چالش‌های فرآیند بازیابی اطلاعات، در اختیار پژوهشگران قرار دهد و در عین حال، با دسته‌بندی و تحلیل نقاط قوت و ضعف این روش‌ها، شرایط را به منظور ارائه روش‌های کارآمدتر، فراهم آورد. برای نیل به این مقصود، در این نوشتار، بیش از پنجاه پژوهش شاخص در این زمینه، مورد مطالعه و بررسی قرار گرفته است و کوشش شده است تا با دسته‌بندی و ارزیابی تحلیلی آنها، شناخت جامعی در زمینه پژوهش‌های صورت گرفته در رابطه با بکارگیری نظریه گراف در حوزه بازیابی اطلاعات، فراهم آید.

در ادامه، در بخش ۲، دسته‌بندی جامعی از پژوهش‌های انجام شده در زمینه بازیابی اطلاعات بر پایه نظریه گراف، در قالب سه دسته کلی و هشت زیر دسته، ارائه خواهد شد. پس از آن، در بخش ۳، به اختصار نسبت به معرفی شاخص‌های ارزیابی مورد استفاده در ارزیابی عملکرد الگوریتم‌های بازیابی اطلاعات مبتنی بر گراف، پراخته خواهد شد و سپس در بخش‌های ۴، ۵ و ۶، پژوهش‌های مرتبط با هر یک از سه دسته فوق‌الذکر، بصورت تحلیلی، بررسی خواهد شد. نهایتا بخش ۷ نیز به جمع‌بندی و نتیجه‌گیری حاصل از این پژوهش اختصاص یافته است.

۲- دسته‌بندی پژوهش‌های صورت گرفته

با توجه به کاربردی بودن نظریه گراف در مدل‌سازی کارآمد عناصر مسئله بازیابی اطلاعات، پژوهش‌های گسترده‌ای در زمینه بکارگیری نظریه گراف در فرآیند بازیابی اطلاعات، صورت گرفته است. این پژوهش‌ها چه از منظر رویکرد و نحوه حل مسئله و چه از نظر شاخص‌های ارزیابی مورد استفاده در ارزیابی عملکرد الگوریتم‌های ارائه شده و نیز مجموعه دادگان مورد استفاده در فرآیند ارزیابی، بسیار متنوع و متفاوت هستند. از این رو به منظور تسهیل بررسی این پژوهش‌ها، ابتدا بایستی به نحوی، اقدام به دسته‌بندی آنها نمود. برای نیل به این مقصود، در این نوشتار با عنایت به نوع مسئله مورد مطالعه در پژوهش‌های مورد مطالعه و همچنین نحوه بهره‌گیری از نظریه گراف در فرآیند بازیابی اطلاعات در آنها، کوشش شده است تا دسته‌بندی کلانی از این پژوهش‌ها، ارائه گردد و سپس به صورت تفصیلی، به بررسی روش‌های ارائه شده در هر دسته نیز پرداخته شود.

بطور کلی، می‌توان حوزه بازیابی اطلاعات را از نظر نوع داده‌های مورد بازیابی، به دو دسته عمده بازیابی اطلاعات متنی و بازیابی اطلاعات وب، تقسیم‌بندی نمود [۱]. از سوی دیگر، این تفکیک می‌تواند از منظر نحوه انجام فرآیند بازیابی اطلاعات نیز صورت گیرد [۲]. در این صورت، روش‌های متداول تحت عنوان سبد کلمات (BOW) و روش‌های بازیابی معنایی، می‌توانند مورد استفاده، واقع شوند. با توجه به این ملاحظات، می‌توان گفت که بطور کلی، چهار دسته کلان از الگوریتم‌ها در حوزه بازیابی اطلاعات، مطرح هستند. از سوی دیگر، در بررسی‌های انجام شده در حوزه کاربرد نظریه گراف در بازیابی اطلاعات، روش بازیابی اطلاعات متنی عمدتا با استفاده از مدل سبد کلمات، صورت گرفته است. لذا، کاربردهای نظریه گراف در فرآیند بازیابی اطلاعات، در سه دسته کلی، به شرح ذیل، قابل تفکیک و دسته‌بندی است:

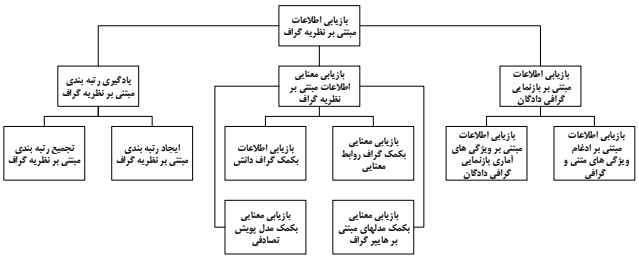
- الف) بازیابی اطلاعات بر مبنای ارائه بازنمایی گرافی از دادگان بازیابی اطلاعات که در دو دسته کلی زیر، قابل تفکیک می‌باشند:
  - بازنمایی گرافی دادگان جهت استخراج ویژگی‌های مبتنی بر گراف و ترکیب این ویژگی‌ها با ویژگی‌های پایه دادگان مسئله به منظور ارائه راهکارهای نوین در حوزه

بازیابی اطلاعات.

- استخراج ویژگی‌های آماری دادگان و ارائه بازنمایی گرافی این ویژگی‌ها

- ب) بازیابی معنایی اطلاعات با استفاده از نظریه گراف که در چهار دسته زیر، قابل دسته‌بندی می‌باشند [۳]:
  - بازیابی اطلاعات با استفاده از گراف دانش
  - ایجاد گراف موجودیت بر اساس دادگان مسئله بازیابی اطلاعات
  - مدل‌های مبتنی بر هایپر گراف
  - مدل‌های مبتنی بر گراف پویش تصادفی
- ج) ارائه بازنمایی گرافی از مسائل حوزه یادگیری رتبه‌بندی که در دو دسته اصلی ایجاد رتبه‌بندی و تجمیع رتبه‌بندی، تقسیم‌بندی می‌شوند.

در شکل ۱ نمایی از این دسته‌بندی ارائه شده است.



شکل ۱. دسته‌بندی پیشنهادی از روش‌های مطرح شده در زمینه بکارگیری نظریه گراف در بازیابی اطلاعات

در ادامه این بخش، به بیان برخی از پژوهش‌های اصلی صورت گرفته در هر یک از سه دسته فوق‌الذکر، پرداخته خواهد شد. ضمن اینکه بایستی خاطر نشان نمود، در انتخاب پژوهش‌های مرتبط با هر دسته، عمدتا تاکید بر انتخاب مقالات سال‌های اخیر بوده است. علاوه بر آن، دسته‌بندی فوق‌الذکر، بر مبنای پژوهش‌های صورت گرفته است. بر این اساس، ممکن است حوزه‌های پژوهشی بالقوه‌ای در ذیل هر دسته مطرح باشد که تا کنون به آنها پرداخته نشده باشد. شناسایی این شکاف‌های تحقیقاتی و عرضه آنها به عنوان حوزه‌های تحقیقاتی بکر، یکی از اهداف اصلی این نوشتار می‌باشد.

۳- شاخص‌های ارزیابی مورد استفاده در ارزیابی عملکرد الگوریتم‌های بازیابی اطلاعات مبتنی بر گراف

در ارزیابی الگوریتم‌های ارائه شده در حوزه بازیابی اطلاعات بر پایه نظریه گراف، مشخص شد، عمده پژوهش‌ها بر اساس شاخص‌های استاندارد بازیابی اطلاعات نظیر P@n، MAP، MRR، ERR@n و NDCG، نسبت به بررسی عملکرد خود در قبال روش‌های پایه، اقدام نموده‌اند. از این رو، در این بخش، به اختصار نسبت به معرفی این شاخص‌ها خواهیم پرداخت.

- شاخص MRR[1]: عکس موقعیت نخستین پاسخ مرتبط در بین فهرست نتایج جستجو را نشان می‌دهد [۱]:

$$MRR = \frac{1}{n}$$

که n برابر با موقعیت اولین پاسخ مرتبط با پرس‌وجوی کاربر در بین فهرست نتایج جتسجو می‌باشد.

- دقت[2] در نقطه k ($P@k$): اگر قضاوت انسانی در خصوص میزان مرتبط بودن اسناد به پرس‌وجوهای کاربران، به‌صورت دودویی در اختیار باشد، مثلاً برچسب اسناد مرتبط، یک و برچسب اسناد نامرتبط، صفر باشد، در این صورت، $P@k$ به‌صورت زیر تعریف می‌شود [۱]:

$$P@k(\pi, l) = \frac{\sum_{t \le k} I_{\{l_{\pi^{-1}(t)}=1\}}}{k}$$

که $I_{\{.\}}$ تابع شاخص است و $l_{\pi^{-1}(j)}$ به برچسب سند قرار گرفته در موقعیت j لیست مرتب نتایج، π، اشاره می‌کند.

- متوسط میانگین دقت[3] (MAP): بر اساس تعریف ارائه شده از دقت در موقعیت k، میانگین دقت[4] (AP) عبارتست از [۱]:

$$AP(\pi, l) = \frac{\sum_{k=1}^{m} P@k.I_{\{l_{\pi^{-1}(t)}=1\}}}{m_1}$$

در رابطه فوق، m تعداد کل اسناد متناظر پرس‌وجوی q است و m_1 تعداد زیر مجموعه‌ای از این اسناد است که میزان مرتبط بودن آنها برابر با یک باشد. مقدار میانگین AP روی همه پرس‌وجوهای آزمون، متوسط میانگین دقت (MAP) نامیده می‌شود.

- بهره انباشته کاهشی هنجار شده[5] (NDCG) در نقطه k ($NDCG@k$): این شاخص بر پایه میانگین‌گیری از شاخص

[1] Mean Reciprocal Rank
[2] Precision
[3] Mean Average Precision
[4] Average Precision
[5] Normalized Discounted Cumulative Gain

بهره انباشته کاهشی[1] (DCG)، محاسبه می‌شود. شاخص بهره انباشته کاهشی (DCG) می‌تواند قضاوت‌های میزان مرتبط بودن را که به‌صورت طبقه‌بندی‌های مرتب چندگانه ارائه شده باشند، در ارزیابی مورد استفاده قرار دهد؛ ضمن آنکه در تعریف آن، یک ضریب کاهشی بر اساس موقعیت اسناد، در نظر گرفته شده است [۱]. برای بیان این موضوع به زبان ریاضی، فرض کنید که لیست مرتبی از نتایج نظیر π به ازای پرس‌وجوی q وجود داشته باشد؛ در این صورت، مقدار شاخص DCG در موقعیت k به‌صورت زیر، تعریف می‌شود:

$$DCG@k(\pi, l) = \sum_{j=1}^{k} G\left(l_{\pi^{-1}(j)}\right)\eta(j)$$

که در آن، $G(.)$ رتبه یک سند است که عموماً به صورت $G(z) = (2^z - 1)$ بیان می‌شود. همچنین $\eta(j)$ یک ضریب کاهشی بر اساس موقعیت سند است، که غالباً به‌صورت $\eta(j) = 1/log(j+1)$ تعریف می‌شود. با هنجار سازی مقادیر $DCG@k$ با یک مقدار بیشینه (مثلاً Z_k) شاخص بهره انباشته کاهشی هنجار شده ($NDCG$) حاصل می‌گردد که عبارتست از:

$$NDCG@k(\pi, l) = \frac{1}{Z_k}\sum_{j=1}^{k} G\left(l_{\pi^{-1}(j)}\right)\eta(j)$$

- رتبه متقابل مورد انتظار[2] (ERR) در نقطه k ($ERR@k$): این شاخص احتمال برآورده شدن نیاز کاربر پس از مشاهده k سند نخست موجود در فهرست نتایج جستجو را نشان می‌دهد. اگر احتمال تامین نیاز کاربر در سند k برابر با $p(q, d_k)$ باشد و طول فهرست پاسخ‌های به پرس‌وجوی کاربر، K سند باشد، آنگاه [۱]:

$$ERR@k = \sum_{k=1}^{K}\frac{1}{k}p(q, d_k)\prod_{i=1}^{k-1}(1 - p(q, d_i))$$

۴- بازیابی اطلاعات بر مبنای ارائه بازنمایی گرافی از دادگان بازیابی اطلاعات

ایده کلی این روش‌ها، بهره‌گیری از قابلیت نظریه گراف در مدل‌سازی روابط پیچیده بین عناصر مسئله، به منظور ارائه یک مدل مبتنی بر گراف از دادگان مسئله بازیابی اطلاعات است. البته رویکرد انجام اینکار در هر زیر دسته متفاوت است. به عنوان مثال، الگوریتم‌های دسته نخست، مستقیما با اعمال نظریه گراف روی دادگان مورد استفاده، یک بازنمایی گرافی از آن، ایجاد نموده و سپس با استخراج ویژگی‌های مبتنی بر این گراف و ترکیب ویژگی‌های بدست آمده با ویژگی‌های پایه دادگان مسئله، به ارائه الگوریتم‌های بازیابی اطلاعات متناسب با مسئله مورد بررسی می‌پردازند. در مقابل، از ویژگی‌های آماری دادگان مسئله، اقدام به تولید بازنمایی گرافی متناظر با این ویژگی‌ها می‌کنند.

۴-۱- بازنمایی گرافی دادگان جهت استخراج ویژگی‌های مبتنی بر گراف و ترکیب این ویژگی‌ها با ویژگی‌های پایه دادگان مسئله

از جمله پژوهش‌های صورت گرفته در دسته اول می‌توان به کار تحقیقاتی [۴] اشاره نمود که در آن، روشی برای جستجوی افراد مرتبط با حوزه علمی مورد نظر کاربر، پیشنهاد شده است. در روش پیشنهادی، برای تخمین میزان تخصص پژوهشگران مختلف، از دو رویکرد مختلف استفاده شده است. روش نخست، با استفاده از اطلاعات متنی از جمله داده‌های مندرج در پروفایل محققان مختلف، تخمینی از میزان خبرگی آنان را بدست می‌دهد و در مقابل، روش دوم، بر اساس گراف ارجاعات بین مقالات مختلف، میزان تخصص هر پژوهشگر را تعیین می‌نماید. سپس با تلفیق این دو تخمین با کمک روش‌های یادگیری رتبه‌بندی تحت نظارت و نیز الگوریتم‌های تجمیع رتبه‌بندی و نیز الگوریتم‌های ترکیب اطلاعات، رتبه‌بندی نهایی، تعیین می‌گردد. در آزمایش‌های صورت گرفته به منظور ارزیابی روش پیشنهادی، از مجموعه داده DBLP استفاده شده است. چند نمونه از ویژگی‌های گراف ارجاعات که در این تحقیق مورد استفاده واقع شده است، عبارتند از: شاخص Hirsch که به ازای یک پژوهشگر خاص، ترکیبی از میزان کلی تولید علم توسط وی و نیز میزان اعتبار موسسه تحقیقاتی متبوع او را نشان می‌دهد، شاخص Hirsch مبتنی بر پرس‌وجوی کاربر که میزان معتبر بودن یک پژوهشگر در یک حوزه علمی خاص را نشان می‌دهد، شاخص Hirsch مبتنی بر زمان، شاخص آلفا که میزان تاثیر گذاری مهمترین مقالات یک فرد یا موسسه متبوع وی را نشان می‌دهد و نیز شاخص e که میزان مازاد ارجاعات یک محقق را که در محاسبه شاخص h لحاظ نمی‌شوند، را تعیین می‌کند. نتایج ارزیابی نشان می‌دهد که ترکیب ویژگی‌های متنی مستخرج از پروفایل و نیز ویژگی‌های برگرفته از گراف ارجاعات، توانسته است مطابق شاخص‌های ارزیابی استاندارد نظیر P@n، NDCG@n و MAP، نسبت به الگوریتم‌های مرجع، به عملکرد مناسب‌تری دست یابد. همچنین در [۵] برای

[1] Discounted Cumulative Gain

[2] Expected Reciprocal Rank

رتبه‌بندی محصولات مورد درخواست کاربران در سیستم‌های درخواست دهنده، خصوصیات ساختار گراف و مدل شبکه عصبی را ادغام کرده است. این کار دو مزیت دارد: اول اینکه با استفاده از ساختار گراف می‌توان مشکل کم تعداد بودن تعداد درخواست‌ها به ازای برخی ازکالاها در سیستم‌های توصیه‌گر را حل کرد (مقابله با پراکندگی محصولات و درخواست‌ها). همچنین می‌توان از ساختار گراف دوبخشی، اطلاعاتی در خصوص ارتباط گره‌های بین هر بخش گراف بدست آورد (ترکیب اطلاعات خارجی ناهمگن برای کمک به بهبود رتبه‌بندی در شبکه عصبی). در این مقاله بر روی دادگان CIKM Cup 2016 Track 2 مدل رتبه‌بندی شامل گراف برای جستجوی محصولات ارائه داده است و با استفاده از مقایسه شاخص‌های MRR، MAP و NDCG نشان داده است که مدل پیشنهادی، عملکرد بهتری نسبت به مدل‌های ارائه شده قبلی دارد. مزیت این الگوریتم استفاده از اطلاعات خاصیت تعدی در پیش بینی ارتباطات است که با استفاده از خواص گراف و از ارتباطات بین آیتم‌ها حاصل می‌شود. همچنین در مقاله [۶]، به موضوع تولید یک روش حاشیه نگاری[1] خودکار برای داده‌های چند رسانه‌ای پرداخته شده است، به نحوی که به ازای هر قلم داده چند رسانه‌ای، متن مستخرج از این روش، عملاً معادل متن توصیف آن داده خواهد بود و بر اساس آن، می‌توان رتبه‌بندی داده‌های چند رسانه‌ای را بر اساس نیاز اطلاعاتی کاربر، به صورت یک الگوریتم یادگیری رتبه‌بندی، انجام داد. در پژوهش‌های قبلی که بر روی حاشیه نگاری خودکار ارایه شده، معمولا از ابرداده[2] استفاده شده است که در اغلب موارد در دسترس نیست. در روش پیشنهادی، بر اساس تعاملات کاربر در دسترسی به هر یک از اقلام داده چند رسانه‌ای، گرافی تحت عنوان گراف تعاملات[3] ایجاد می‌گردد. با توجه به این واقعیت که تعداد این داده‌ها بسیار زیاد است، به منظور کاهش پیچیدگی زمانی الگوریتم، به صورت تدریجی و بازگشتی، در هر بازه زمانی مشخص، تعدادی از داده‌های شاخص‌تر و حائز اهمیت بیشتر، در تولید زیر گراف تعاملات، مورد استفاده قرار می‌گیرند. نکته قابل توجه در ایده پیشنهادی، استفاده از دو دسته از ویژگی‌های مختلف در توصیف مشخصات گراف تعاملات است. یکی از این دو دسته، ویژگی‌های ایستا[4] است که عمدتاً بر پایه نتایج آماری حاصل از ترجیح بعضی از اقلام داده چند رسانه‌ای نسبت به بقیه توسط کاربران، استخراج می‌شوند و وابسته به ساختار گراف ترجیحات نیستند. در مقابل، دسته دوم، عمدتا شامل ویژگی‌های توپولوژیک و ساختاری[5] گراف مانند ضریب خوشه‌بندی است. این دو دسته ویژگی، به عنوان ورودی به الگوریتم‌های پایه رتبه‌بندی نظیر RankBoost و AdaRank، عرضه می‌شوند و امکان ایجاد اولویت‌بندی بین اقلام داده‌ای ممکن را فراهم می‌سازند. برای ارزیابی عملکرد روش پیشنهادی، از مجموعه سوابق اطلاعات وبسایت Douban طی یک بازه شش ساله و در دو بخش کتاب و فیلم، استفاده شده است. همچنین از الگوریتم‌های شاخص یادگیری رتبه‌بندی نظیر AdaRank و AdaBoost، در ارزیابی نتایج روش پیشنهادی، بهره گرفته شده است. ضمنا شاخص‌های ارزیابی مورد استفاده عبارتند از: P@n، MAP و NDCG. نکته جالب توجه در خصوص نتایج آزمایش‌های به عمل آمده، این است که طی آنها سعی شده است تا نقش و تاثیر هر یک از این دو دسته ویژگی‌های ایستا و توپولوژیک در فرآیند یادگیری رتبه‌بندی مطابق شاخص‌های ارزیابی رتبه‌بندی، تعیین شود. در این خصوص ذکر این نکته نیز حائز اهمیت است که در برخی شرایط، بکارگیری صرف ویژگی‌های توپولوژیک، منجر به بهبود بیشتر رتبه‌بندی در قیاس با ویژگی‌های ایستا شده است. ضمن آنکه همواره ترکیب این دو دسته ویژگی‌ها نسبت به هر دسته، منجر به عملکرد بهتری می‌شود. با توجه به ضعف عملکرد این الگوریتم در شرایط محدود بودن محتوای متنی، توسعه الگوریتم به منظور جبران این نقیصه، حائز اهمیت است. همچنین نظر به هزینه‌بر بودن اجرای متوالی گام‌های استخراج ویژگی‌ها و رتبه‌بندی زیر گراف‌ها، می‌توان به ادغام این دو گام در قالب یک گام واحد نیز پرداخت. دستیابی به یک شرط توقف خودکار که مصالحه مابین هزینه و دقت عملکرد را ممکن سازد، نیز جالب توجه می‌باشد. از سوی دیگر، نویسندگان [۷]، بر مبنای نظریه گراف‌ها، الگوریتم یادگیری رتبه‌بندی جدیدی را برای بهبود عملکرد سامانه‌های توصیه‌گر ارائه نموده‌اند. ایده اصلی آنها، محاسبه دسته‌ای از ویژگی‌ها بر اساس هایپرگراف[6] ارتباط بین کاربران و کالاهای عرضه شده توسط سامانه است. مسئله مورد نظر در این مقاله، بهبود عملکرد سامانه‌های توصیه‌گر با عرضه گونه جدیدی از ویژگی‌ها موسوم به ویژگی‌های طیفی[7] است. در واقع ویژگی‌های بصری و متنی که در الگوریتم‌های پایه رتبه‌بندی در سامانه‌های توصیه‌گر، مورد استفاده قرار می‌گیرند، ممکن است به ازای همه اقلام کالایی و همه کاربران، موجود نباشد. در این شرایط، ویژگی‌های طیفی همچنان قابل محاسبه و استفاده خواهند بود. بر این اساس، ایده اصلی این مقاله، ایجاد یک هایپرگراف[8] روی کاربران و اقلام است. یک هایپرگراف،

[1] Annotation
[2] Metadata
[3] Interaction Graph
[4] Static Features
[5] Topologic Features
[6] Hypergraph
[7] Spectral Features
[8] Hypergraph

تعمیمی از گراف ساده است که در آن یک هایپریال[1] تعدادی از رئوس (و نه لزوماً فقط دو راس) را به هم متصل می‌سازد. در ادامه، مجموعه‌ای از ویژگی‌ها موسوم به ویژگی‌های طیفی با استفاده از ماتریس لاپلاسین، استخراج می‌شوند. برای این منظور از ماتریس لاپلاسین، مقادیر ویژه استخراج می‌شود و سپس با استفاده از k مقدار ویژه بیشینه، ماتریس ویژگی‌های طیفی[2]، استخراج می‌گردد. ویژگی‌های طیفی، میزان مشابهت کاربران و اقلام کالایی را در نمایش گرافی سامانه‌های توصیه‌گر نشان می‌دهند. به کمک این ویژگی‌ها، یک الگوریتم یادگیری رتبه‌بندی جفتی[3] ایجاد می‌گردد که قادر است اولویت نسبی کالاها به ازای یک کاربر خاص را تعیین نماید. به منظور ارزیابی عملکرد روش پیشنهاد شده، از مجموعه دادگان وبسایت آمازون در دو بخش البسه و جواهرات، استفاده شده است. نتایج ارزیابی عملکرد الگوریتم ارائه شده نسبت به روش‌های مرجع نظیر فاکتورسازی ماتریس احتمالاتی[4]، رتبه‌بندی شخصی‌شده بیزین[5]، رتبه‌بندی شخصی‌شده بیزین بر مبنای ترجیح گروهی[6] و نیز رتبه‌بندی شخصی‌شده بیزین بصری[7]، حاکی از تفوق این الگوریتم بر اساس شاخص‌های F1 و NDCG@n می‌باشد. از آنجا که روش پیشنهادی در شرایطی قادر به عملکرد مناسب است که بازخورد صریح کاربران به ازای کالاهای مورد نظر در اختیار باشد، لذا توسعه این الگوریتم در حالتی که بازخورد صریح در اختیار نباشد، حائز اهمیت خواهد بود. همچنین در مقاله [۸] به منظور یافتن میزان مشابهت بین عناصر داده‌ای مختلف نظیر اسناد و پرس‌وجوها و بهبود نتایج رتبه‌بندی، الگوریتم جدیدی ارائه شده است. ابتدا بایستی توجه داشت که در کارهای قبلی عمدتاً از دو دسته از روش‌های رتبه‌بندی یعنی «روش‌های رتبه‌بندی مبتنی بر ویژگی‌ها» و «روش‌های مبتنی بر گراف بین اسناد و پرس‌وجوها»، استفاده شده است. بر این اساس، در این مقاله، نقاط قوت دو دسته روش فوق‌الذکر، ترکیب شده است. در روش ارائه شده، مسئله یادگیری مشابهت سند و پرس‌وجو را با استفاده از یک گراف دوبخشیِ بدون جهت متناظر با اطلاعات کلیک‌های کاربران، مدلسازی شده است که متشکل از اسناد و پرس‌وجوها است. در این گراف، ابرداده‌ی[8] گره‌ها، شامل انواع مختلفی از ویژگی‌ها است و وزن یال‌ها نیز نشان دهنده تعداد کلیک‌های کاربران می‌باشد. علاوه بر آن، فرض می‌شود که ابرداده غنی به ازای هر یک از گره‌های این گراف، موجود باشد. این ابرداده، شامل اطلاعات ویژگی‌های مختلف به ازای اسناد یا پرس‌وجوها است. بر اساس این مدل، در این پژوهش، الگوریتم یادگیری رتبه‌بندی جدیدی، پیشنهاد شده است که بر مبنای الگوریتم M-PLS عمل می‌کند. روش M-PLS، یک توسعه چند وجهی از فرآیند آماری PLS[9] است. بطور خلاصه در الگوریتم PLS، برای مدلسازی رابطه بین دو یا چند مجموعه داده، آنها را به یک فضای نهفته[10]، نگاشت می‌کنند. بدین ترتیب، بررسی هم‌خطی بودن[11] این مجموعه‌های داده، ممکن خواهد شد. برای ارزیابی الگوریتم پیشنهاد شده، عملکرد آن روی مجموعه داده غیر عمومی موتور جستجوی Bing متعلق به شرکت مایکروسافت بررسی شده است. این ارزیابی‌ها که بر اساس شاخص‌های ارزیابی MAP و NDCG@n صورت گرفته است، حاکی از برتری روش فوق نسبت به الگوریتم‌های پایه رتبه‌بندی نظیر BM25 است. همچنین در این مقاله، بلحاظ نظری، ثابت شده است که الگوریتم ارائه شده، عملکرد بهینه سراسری دارد. به عنوان کارهای تحقیقاتی آتی می‌توان موازی سازی الگوریتم پیشنهادی جهت استفاده از مجموعه‌های داده حجیم‌تر و در نتیجه بهبود عملکرد و نتایج را بررسی کرد. در [۹] نیز الگوریتمی برای حل مسئله رتبه‌بندی رئوس گراف‌های دوبخشی بر اساس یک شاخص معین، ارائه شده است. گراف‌های دو بخشی را می‌توان به عنوان مدل پایه در تحلیل بسیاری از سامانه‌های مهم بازیابی اطلاعات نظیر سامانه‌های بازاریابی برخط، سامانه‌های توصیه‌گر، سامانه‌های پرسش و پاسخ و نظایر آنها در نظر گرفت. بنابراین رتبه‌بندی رئوس در این گراف‌ها حائز اهمیت بالایی می‌باشد. برای این منظور، در روش پیشنهاد شده، از اطلاعات ساختاری گراف دوبخشی به عنوان اطلاعات پایه استفاده شده است و در ادامه به منظور بهبود عملکرد، از ترکیب آن با ابرداده مربوط به رئوس گراف، استفاده شده است. بصورت مشخص، فرض می‌شود که هر راس از یک بخش گراف، به صورت بالقوه با تمامی رئوس از بخش مقابل، اتصال دارد که میزان این ارتباط با یک مقدار عددی تحت عنوان وزن این یال، مشخص می‌گردد (در صورت عدم وجود اتصال، وزن مربوطه، صفر در نظر گرفته می‌شود). سپس، مشابه الگوریتم PageRank، به هر گره، یک درجه اهمیت تخصیص می‌یابد. این وزن، با استفاده از ایده الگوریتم HITS، بصورت تقویتی و طی یک فرآیند بازگشتی، از ترکیب خطی درجه اهمیت و وزن رئوس مرتبط از بخش مقابل، محاسبه می‌گردد. وجه تمایز الگوریتم ارائه شده موسوم به BiRank نسبت به روش‌های کلاسیک رتبه‌بندیِ مبتنی

[1] Hyperedge
[2] Spectral Feature Matrix
[3] Pairwise Learning to Rank
[4] Probabilistic Matrix Factorization
[5] Bayesian Personalized Ranking
[6] Group Preference-based Bayesian Personalized Ranking
[7] Visual Bayesian Personalized Ranking
[8] Metadata
[9] Partial Least Squares
[10] Latent Space
[11] Collinearity

بر پویش تصادفی[1]، این است که این الگوریتم، طی فرآیند محاسبات مکرّر، یک تابع منظم سازی را بهینه می‌کند که وظیفه هموار سازی گراف با استفاده از اطلاعات بردار پرس‌وجو را به عهده دارد. ارزیابی عملکرد الگوریتم پیشنهادی در دو مسئله متفاوت یعنی پیش‌بینی میزان محبوبیت کالاها در آینده و نیز توصیه کالاهای مورد علاقه به کاربران، بر مبنای شاخص NDCG@n انجام شده است. سطح ارزیابی نخست بر اساس گراف‌های مصنوعی با توزیع درجه یکنواخت و power-law انجام شده است و علاوه بر آن، در مسئله نخست، از سه نوع داده واقعی مختلف یعنی YouTube، Flicker و Last.fm استفاده شده است و به همین ترتیب، در ارزیابی مسئله دوم دادگان Yelp و Amazon، بکار گرفته شده است. نتایج این ارزیابی‌ها حاکی از عملکرد مطلوب روش پیشنهادی در قیاس با الگوریتم PageRank و یکی از توسعه‌های موسوم به TagRW می‌باشد. به منظور توسعه این الگوریتم، سه ایده بکارگیری ویژگی‌های بیشتری از اسناد و پرس‌وجوها، توسعه روش پیشنهادی با استفاده از چارچوب بیزین و نیز یادگیری ادغام پارامترها به صورت خودکار، بیان شده است. علاوه بر موارد فوق، در [۱۰] روشی برای پیشنهاد مخاطب در شبکه‌های اجتماعی، ارائه شده است که در آن، شبکه اجتماعی به صورت یک گراف جهت‌دار، مدل‌سازی شده است. در این گراف، گره‌ها معادل افراد عضو شبکه اجتماعی مورد نظر هستند و انواع ارتباط بین آنها (نظیر دوستی و تعاملات مختلف)، معادل یال‌های این گراف هستند. به ازای هر گره، سه دسته همسایه، شامل همسایه‌های ورودی (افرادی که پیوندهای به گره مورد بحث دارند)، همسایه‌های خروجی (افرادی که گره مورد نظر، پیوندهایی به آنها دارد) و نیز اجتماع هر دو دسته فوق، می‌توان در نظر گرفت. به این ترتیب به سادگی می‌توان مسئله پیشنهاد مخاطب را به مسئله رتبه‌بندی، نگاشت نمود. بر این اساس، مسئله پیشنهاد مخاطب به یک کاربر مشخص، بصورت اولویت‌بندی همسایگان کاربر مورد نظر بر مبنای میزان مشابهت آنان به وی، قابل مدلسازی خواهد بود. در پیاده‌سازی این الگوریتم، از روش رتبه‌بندی پایه BM25 استفاده شده است. همچنین به منظور ارزیابی روش ارائه شده، از مجموعه دادگان Twitter و Facebook و نیز شاخص‌های ارزیابی MAP و NDCG@n استفاده شده است. این بررسی‌ها حاکی از توفیق روش پیشنهادی نسبت به الگوریتم‌های شناخته شده رتبه‌بندی نظیر VSM و PageRank می‌باشد. همچنین در [۱۱] بر مبنای بکارگیری یادگیری رتبه‌بندی، روشی برای اولویت‌بندی فرمولهای درج شده در یک مقاله علمی، ارائه شده است. در روش پیشنهادی، دو دسته ویژگی‌های ثانویه به ازای هر فرمول، محاسبه می‌شود که یک دسته، بر اساس تحلیل ارجاعات بین مقالات و دسته دوم بر اساس تحلیل معنایی درون مقاله مورد نظر، بدست می‌آید. ابتدا با استفاده از یک پیکره بزرگ، گراف ارجاعات به فرمول‌ها، تولید و تحلیل می‌شود تا بکمک آن، اهمیت نسبی هر یک از فرمول‌ها تعیین شود. برای این منظور، در گراف ایجاد شده، بر اساس تحلیل توپولوژیک پیوندهای مقالات، ویژگی‌های بین مقاله‌ای فرمول‌ها، استخراج می‌شود. سپس مدل تعبیه کلمات[2] به منظور استخراج ویژگی‌های درونی مقاله، مورد استفاده قرار می‌گیرد تا رابطه معنایی بین فرمول مورد نظر و آن مقاله تعیین شود. بر این اساس، جمعاً یازده ویژگی مختلف در قالب سه دسته ویژگی به شرح زیر استخراج می‌شود: سه ویژگی مبتنی بر خصوصیات فرمول، پنج ویژگی درون مقاله‌ای و نیز سه ویژگی بین مقاله‌ای. نهایتاً با اِعمال الگوریتم یادگیری رتبه‌بندی ListNet روی ویژگی‌های استخراج شده، رتبه‌بندی فرمول‌ها، صورت می‌گیرد. برای ارزیابی روش ارائه شده از زیر مجموعه دادگان ویکی‌پدیا در تاریخ ۲۲-۷-۲۰۱۶ شامل ۳۳۲ مقاله استفاده شده است. بررسی صورت گرفته نشان دهنده برتری عملکرد روش پیشنهادی نسبت به روش‌های پایه بر مبنای شاخص‌های P@n و NDCG@n است. از سوی دیگر، مراجعه به جویشگرهای کدهای برنامه‌نویسی، یک روش رایج بین برنامه‌نویسان به منظور یافتن راهکارهای پیاده‌سازی در حل مسائل پیش رو به شمار می‌رود. از این رو، بکارگیری روش‌های یادگیری رتبه‌بندی به منظور اولویت‌بندی پاسخ‌های پیشنهادی، حائز اهمیت است. در پژوهش [۱۲]، روشی برای رتبه‌بندی نمونه‌های کد پیاده‌سازی شده بر اساس سوال کاربر، عرضه شده است. در این الگوریتم، بر اساس رابطه فراخوانی بین قطعات کد مختلف، ابتدا گرافی موسوم به گراف فراخوانی بین آنها ایجاد می‌شود و سپس میزان PageRank هر قطعه کد در این گراف، محاسبه می‌شود. علاوه بر این ویژگی، یازده ویژگی دیگر که برخی از آنها وابسته به سوال کاربر و برخی مربوط به خصوصیات ذاتی قطعات کد هستند، نیز استخراج می‌شوند. دو نمونه از این ویژگی‌ها عبارتند از: مشابهت متن پرس‌وجوی کاربر و محتوای هر قطعه کد، میزان معروفیت کد که بر اساس میزان مشابهت بین الگوی پیاده‌سازی استفاده شده در آن کد و الگوی غالب در پیاده‌سازی قطعات کد موجود در مخزن مورد بررسی، تعیین می‌شود. این دوازده ویژگی، به الگوریتم رتبه‌بندی RankBoost عرضه می‌شود تا فرآیند یادگیری رتبه‌بندی، انجام گردد. داده محک مورد استفاده برای ارزیابی روش پیشنهادی، بکمک خزش داده‌های موجود در GoogleCode با برچسب Android بدست آمده است و شامل ۳۶۰۰۰۰ قطعه کد مربوط به ۵۸۶ پروژه پیاده‌سازی شده با

[1] Random Walk

[2] Word Embedding

زبان اندروید می‌باشد. ارزیابی صورت گرفته نشان می‌دهد الگوریتم ارائه شده بر اساس شاخص‌های ارزیابی NDCG و ERR، عملکرد مناسبی داشته است. در [۱۳] نیز الگوریتم جدیدی برای یادگیری رتبه‌بندی ارائه شده است که از ویژگی‌های گراف دو بخشی متناظر با کلیک‌های کاربران روی اسناد وب، در فرآیند رتبه‌بندی استفاده می‌کند. در این گراف، وزن متناظر با یک زوج سوال و سند، برابر با تعداد کلیک‌های صورت گرفته بین این زوج است. اسناد و پرس و جوهای کاربران با استفاده از مدل فضای برداری، نمایش داده می‌شوند. در مرحله بعدی که یک فرآیند تکرار شونده است، فرآیند انتشار بردار بین اسناد و پرس وجوهای هم کلیک، انجام می‌شود. بدین معنی که مثلا اگر از سمت پرس و جوها شروع کنیم، به ازای هر سند، جمع وزندار بردارهای متناظر با پرس و جوهای هم کلیک، بردار متناظر با آن سند را بدست می‌دهد. در تکرار بعدی، بردار متناظر یک پرس‌وجو با استفاده از جمع وزندار بردارهای متناظر با اسناد هم کلیک متناظر، محاسبه می‌شود. این توالی تا زمان همگرایی این بردارها تداوم خواهد داشت. در ارزیابی کاربرد این الگوریتم از سوابق تراکنش‌های کاربران جویشگر Yahoo شامل حدود هشت میلیارد پرس‌وجو و نیز حدود سه میلیارد سند استفاده شده است. ارزیابی‌های بعمل آمده بر اساس شاخص NDCG@n حاکی از بهبود رتبه‌بندی با استفاده از روش پیشنهادی است. از سوی دیگر، پژوهشگران در [۱۴] الگوریتم یادگیری رتبه‌بندی جدیدی برای سامانه‌های توصیه‌گر ارائه کرده‌اند که بر مبنای رتبه‌بندی محلی و ادغام سراسری عمل می‌کند. استفاده از این روش، امکان آن را فراهم می‌آورد که از چندین فضای ویژگی‌های ریز دانه در قالب یادگیری رتبه‌بندی استفاده شود. در این الگوریتم، ایجاد گراف‌های وزن دار کاربران و کالاها مبنای مراحل بعدی است. در ایجاد گراف کاربران، چنانچه دو کاربر به کالای واحدی ارتباط داشته باشند، بین آنها یال ایجاد می‌شود. وزن متناظر با این یال، بر اساس میزان رای دهی این دو کاربر به کالاهای مشترک، تعیین می‌گردد. همین روند در تولید گراف کالاها نیز استفاده می‌شود. سپس در گام نخست الگوریتم ارائه شده، فرآیند گام زدن تصادفی دو مرحله ای روی گراف کاربران و اقلام کالاها، اعمال می‌شود تا کاربران و کالاهای با حداکثر میزان ارتباط، در قالب گروه‌های محلی محدود، شناسایی شود. در مرحله بعد، با استفاده از الگوریتم رتبه‌بندی بیزین زوج-محور روی این گروه‌های محلی، لیست‌های رتبه‌بندی محلی در هر گروه، حاصل می‌شود. نهایتا یک روش ادغام سراسری روی این لیست‌های مرتب محلی، اجرا می‌شود تا برای هر یک از کاربران، برترین اقلام کالایی، شناسایی شوند. به منظور ارزیابی روش پیشنهادی، از مجموعه دادگان Movielens-100K، Movielens-1M و نیز Yahoo!R3 استفاده شده است. مطابق نتایج حاصل از این آزمایش‌ها، روش پیشنهاد شده توانسته است بنا به شاخص‌های ارزیابی MAP و NDCG، عملکرد بهتری نسبت به روش‌های پایه رتبه‌بندی در حوزه سامانه‌های توصیه‌گر نظیر PersonalRank، CLLORMA و RWLMA داشته باشد. یک رویکرد تحقیقاتی دیگر در این حوزه، موضوع رتبه‌بندی عادلانه نتایج جستجو می‌باشد که با هدف به حداقل رسانیدن نابرابری در عرضه اطلاعات، انجام می‌شود. در پژوهش [۱۵]، هدف از این مقوله، ارائه نتایج تحقیقات گروه‌های مختلف محققان در جستجوی محتوای علمی است. برای این منظور، در این مقاله روش‌های متفاوتی ارائه شده است. در روش نخست از ویژگی‌های مختلف مقالات نظیر عنوان، نام نویسندگان، محل انتشار، تعداد مراجع و نیز تعداد ارجاعات به آنها، به منظور رتبه‌بندی آنها بر اساس الگوریتم یادگیری رتبه‌بندی پیشنهاد شده در [۱۶]، استفاده شده است. در ادامه برای یکسان کردن شانس اسناد با میزان مرتبط بودن همسان به منظور رتبه‌بندی همانند، یک عنصر تصادفی با توزیع یکنواخت، به میزان درجه مرتبط بودن مقالات، افزوده می‌شود. هدف روش دوم، تخصیص شانس یکسان جهت عرضه تحقیقات انجام شده توسط گروه‌های مختلف محققین است. برای این منظور، از گراف مشارکت محققین در مقالات منتشر شده، استفاده شده است. سپس گروه‌های محققین بکمک روش [۱۷] و بر مبنای فاصله آنها در گراف مذکور، از این گراف شناسایی می‌شوند. بدین ترتیب هر پژوهشگر در گروهی قرار می‌گیرد که یا با دیگر اعضای آن، مقاله مشترکی داشته باشد یا اینکه با آنها بواسطه پژوهشگر مشترک دیگری، مرتبط شده باشد. به ازای یک پرس‌وجوی مشخص، از هر گروه پژوهشگران، مقالات مرتبط با استفاده از ایده روش نخست، شناسایی و رتبه‌بندی می‌شوند. ارزیابی روش‌های پیشنهادی بر اساس شاخص‌های MAP و NDCG@n با استفاده از داده Semantic Scholar (S2) Open Corpus که شامل اطلاعات مربوط به حدود ششصد پرس‌وجوی متفاوت است، نشان می‌دهد که روش ساده‌تر اول، عملکرد بهتری داشته است. ضمن اینکه روش نخست، این مزیت را نیز دارا است که در فرآیند بازیابی و جستجو، نیازی به اطلاعات مازاد مندرج در مقالات در رابطه با محققین ندارد. یک رویکرد دیگر مربوط به فرآیند تست نرم‌افزار است که بخش مهمی از مهندسی نرم‌افزار را در بر می‌گیرد که به منظور رفع نواقص موجود و اصلاح نرم‌افزار، انجام می‌شود. با این حال، در فرآیند تست سامانه‌های نرم‌افزاری بزرگ و پیچیده، غالباً تعداد زیادی ایراد، کشف می‌شود که پس از اصلاح نرم‌افزار، موارد تست متناظر با آنها تدوین می‌شود. در [۱۸]، دو دسته راه‌حل برای اولویت‌بندی موارد تست و نیز کلاس‌های مورد تست در فرآیند

مهندسی نرم‌افزار، ارائه شده است که بر اساس اهداف تست (مثلا میزان تاثیر خطاهای نرم‌افزاری) عمل می‌کند. یک مجموعه از روش‌های پیشنهادی، الگوریتم یادگیری رتبه‌بندی هستند که از مجموعه‌ای از ویژگی‌های متناظر با کد در حال تست و نیز تعدادی از خصوصیات فرآیند تست نرم‌افزار نظیر سوابق هر تست، استفاده می‌کنند. در مقابل، روش‌های پیشنهاد شده در دسته دوم، الگوریتم‌های یادگیری تقویتی هستند که عامل یادگیرنده، طی هر گام از تعامل با محیط پیرامونی، از فرآیند رتبه‌بندی به منظور بهبود عملکرد خود، بهره می‌گیرد. بدین ترتیب این امکان وجود خواهد داشت که با عنایت به تغییرات مداوم نتایج تست در طی فاز تحویل نرم‌افزار، فرآیند یکپارچه سازی پیوسته نرم‌افزار، انعطاف و پویایی بیشتری داشته باشد. فرآیند پیشنهادی، شامل دو مرحله انتخاب مجموعه تست‌ها در مرحله فعلی و سپس اولویت‌بندی آنها به منظور بررسی می‌باشد. در گام نخست، بر مبنای تحلیل ایستای وابستگی بین اجزا نرم‌افزار، در سطح کلاس‌ها، به صورت دقیق و کم هزینه می‌توان زیر مجموعه‌ای از کل تست‌ها به منظور بررسی در وهله بعدی تحویل نرم‌افزار، انتخاب نمود. برای این منظور، گراف وابستگی بین کلاس‌ها ایجاد می‌شود. از بررسی این گراف، می‌توانیم کلاس‌های تحت تاثیر ناشی از اصلاحات تست در دیگر کلاس‌های نرم‌افزاری را شناسایی کنیم و مجموعه موارد تست مرحله بعد را تعیین کنیم و سپس بر اساس دو شاخص میزان تاثیر آنها در کشف خطاهای نرم‌افزار و نیز مدت زمان مورد نیاز جهت انجام، موارد تست منتخب را اولویت‌دهی و رتبه‌بندی کنیم. در ارزیابی عملکرد روش‌های پیشنهادی از اطلاعات مربوط به حدود هزار پروژه موجود در GitHub استفاده شده است. این بررسی‌ها بر اساس شاخص پایه FPA[1] حاکی از عملکرد مناسب روش‌های پیشنهاد شده نسبت به الگوریتم‌های کلاسیک است. همچنین در [۱۹] روشی برای رتبه‌بندی پژوهشگران شاخص در حوزه‌های مختلف علوم ارائه شده است که بر مبنای شاخص‌های مستخرج از گراف نویسندگان مشترک مقالات علمی، اقدام به شناسایی محققان برتر می‌کند. روش سنتی انجام اینکار بر مبنای محاسبه شاخص‌های علم سنجی نظیر تعداد مقالات منتشر شده، میزان ارجاعات به این مقالات و نیز h-index عمل می‌کنند که از میان آنها میزان ارجاعات به مقالات محقق مورد نظر، نقش پر رنگ تری دارد. عموما مدتی پس از چاپ یک مقاله، ارجاع دیگر محققان به آن شروع می‌شود و میزان ارجاعات به یک مقاله طی یک بازه زمانی چندین ساله، مبنای دقیقی برای ارزیابی اهمیت آن مقاله محسوب می‌شود. از این رو، استفاده از این قبیل شاخص‌ها برای رتبه‌بندی محققان جوان، چندان دقیق نخواهد بود. به منظور حل این مسئله، در این پژوهش، روشی برای اولویت‌بندی پژوهشگران بر مبنای شاخص‌های میزان مرکزیت گره‌ها در گراف نویسندگان مشترک، ارائه شده است. در ارزیابی روش پیشنهادی از اطلاعات مربوط به ۶۷۱ کاندیدای مربوط به ۲۴ جایزه بین‌المللی معتبر در حوزه ریاضیات استفاده شده است. آزمایش‌های صورت گرفته نشان داده است که مطابق معیار P@n، رتبه‌بندی بر مبنای شاخص‌های شبکه گراف نویسندگان مشترک، در مقایسه با شاخص‌های علم سنجی، در شناسایی محققین برتر، تقریبا عملکرد مشابهی داشته است. با این حال، با استفاده از روش پیشنهادی، می‌توان محققان جوان و شاخص در هر حوزه را شناسایی نمود.

۴-۲- استخراج ویژگی‌های آماری دادگان و ارائه بازنمایی گرافی این ویژگی‌ها

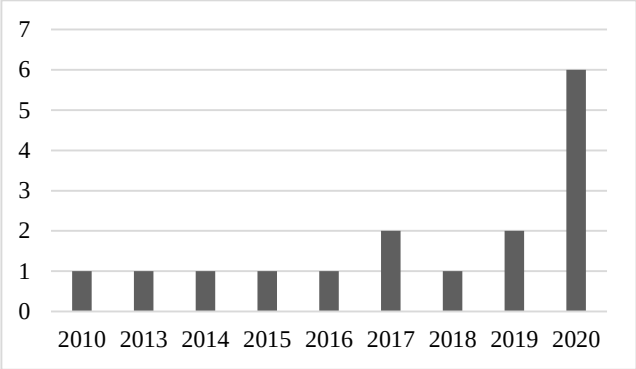
به عنوان نمونه تحقیقات دسته دوم، می‌توان به [۲۰] اشاره نمود که در آن، یک مدل رتبه‌بندی بر اساس الگوریتم تعمیم یافته PageRank معرفی شده است. در این روش، با استفاده از تحلیل بیزین نتایج بازی‌های انجام شده، گرافی از تیم‌ها و ارتباطات بین آنها ساخته شده است. پارامترهای این مدل با هدف حداقل کردن فاصله تابع زیان بین رتبه‌بندی واقعی تیم‌ها و پیش بینی مدل، تخمین زده می‌شوند. برای ارزیابی روش پیشنهادی از داده‌های ده سال مسابقات بسکتبال که از سایت NBA استخراج شده، استفاده شده است. مزیت استفاده از این مدل در مقایسه با PageRank این است که می‌توان دانش اولیه را در مدل گنجاند و بدین ترتیب می‌توان مدل را با سایر مدل‌های رتبه‌بندی ترکیب کرد. همچنین مقادیر پارامترها در مدل را می‌توان بر اساس بازی‌های جدید اصلاح و بروزرسانی کرد. علاوه بر آن، تابع مربوط به ارتباطات گره‌ها یک بردار چند بعدی است که اطلاعات بیشتری نسبت به امتیاز بازی‌ها یا برنده شدن تیمها دارد. در آزمایش‌های به عمل آمده، عملکرد الگوریتم پیشنهادی نسبت به الگوریتم‌های پایه رتبه‌بندی، از لحاظ همبستگی رتبه‌بندی نتایج کندال، دقت پیش بینی نتایج در هر فصل و نیز دقت نتایج بازی‌های Playoff بررسی شده است. این بررسی‌ها نشان می‌دهد که مدل ارائه شده در بیشتر فصول، عملکرد بهتری نسبت به سایر مدل‌های رتبه‌بندی داشته است. روش پیشنهادی، تاثیر حاشیه‌ای هر یک از بازی‌های انجام شده توسط یک تیم را در فرآیند ارزیابی قدرت آن تیم در نظر می‌گیرد. با این حال، در این فرآیند، عامل زمان در نظر گرفته نشده است، یعنی اثر

[1] Fault Percentile Average

یک بازی ممکن است در طول زمان تضعیف شود. به منظور توسعه روش پیشنهادی، می‌توان به نتیجه هر بازی، وزنی متناسب با فاصله زمانی آن نسبت به زمان حال را اختصاص داد. علاوه بر این، بررسی گونه‌های دیگر توابع پیوند و نیز توابع زیان، می‌تواند جالب توجه باشد. ضمنا عملکرد روش پیشنهادی، به منظور رتبه‌بندی تیم‌ها یا ورزشکاران دیگر رشته‌های ورزشی نیز می‌تواند مورد بررسی قرار گیرد. همچنین در [۲۱]، دسته‌ای از الگوریتم‌های یادگیری رتبه‌بندی روی گراف‌ها ارائه شده است که در آنها، هدف کلی، رتبه‌بندی اشیایی است که به هم مرتبط هستند. الگوریتم‌های استفاده شده بر مبنای ایده graph regularization طراحی شده‌اند که قبلا در زمینه حل مسائل یادگیری گراف‌ها ارائه شده بود. در این تحقیق، مسئله رتبه‌بندی به صورت گرافی وزن‌دار بازنمایی شده است که وزن بین دو راس، نشان دهنده ترتیب نسبی بین آن دو است. بر این اساس، به ازای یادگیری رتبه‌بندی صحیح رئوس این گراف، تابع جریمه‌ای تعریف می‌شود که به ازای یک روش رتبه‌بندی خاص، بر اساس وزن یال‌های گراف فوق، محاسبه می‌شود. بصورت خلاصه، در روش پیشنهاد شده، از طریق بازتولید فضای هسته هیلبرت، تابع رتبه‌بندی مستخرج از گراف، فرا گرفته می‌شود. این امر باعث می‌شود، الگوریتم‌های بدست آمده، حائز ویژگی‌های جالب توجهی نظیر پایداری و تعمیم‌پذیری باشند. پیاده‌سازی این ایده به دو روش یعنی با استفاده از رگرسیون SVM و نیز توسعه الگوریتم پایه RankSVM با تابع جریمه فوق‌الذکر، انجام شده است. به منظور ارزیابی عملکرد روش‌های پیشنهادی، کارآیی آنها در دو مسئله پایه، مورد بررسی و ارزیابی قرار گرفته است که عبارتند از: تشخیص میزان کیناس در شبکه تعامل پروتئین‌ها و نیز تعیین میزان مشارکت ساختارهای شیمیایی پایه در تولید داروهای مورد نظر. نتایج حاصل از آزمایش‌های صورت گرفته، حاکی از برتری شکل دوم پیاده‌سازی روش پیشنهادی نسبت به گونه نخست پیاده‌سازی این الگوریتم بر اساس شاخص‌های P@n، MAP و NDCG@n است. به عنوان توسعه‌های آتی، دو مسیر پیشنهاد شده است که یکی از آنها، استفاده از توابع جریمه دیگر بجای Hinge Loss Function است که در این مقاله استفاده شده است. همچنین، تلاش برای بهبود عملکرد روی شاخص‌های دیگر بازیابی نظیر MAP و MeanNDCG، مسیر توسعه پیشنهاد شده بعدی است.

در جدول ۱ خلاصه پژوهش‌های بررسی شده در دسته الف (بازیابی اطلاعات بر مبنای ارائه بازنمایی گرافی از دادگان بازیابی اطلاعات) آمده است.

از سوی دیگر، فراوانی تعداد مقالات مرتبط بررسی شده در دسته نخست در شکل ۲ آمده است. بر اساس این نمودار، اگر چه از سال‌های گذشته، پژوهش‌هایی در این زمینه انجام شده است، اما عمده این تحقیقات مربوط به سالهای اخیر است.



شکل ۲. فراوانی تعداد مقالات مرتبط بررسی شده در دسته نخست (بازیابی اطلاعات با استفاده از بازنمایی گرافی دادگان)

۵- بازیابی معنایی اطلاعات با استفاده از نظریه گراف

وجه مشترک الگوریتم‌های ذیل این بخش، در پرداختن آنها به مقوله بازیابی معنایی اطلاعات با بهره‌گیری از نظریه گراف می‌باشد که بنا به اهمیت و گستره کاربرد وسیع روش‌های بازیابی معنایی اطلاعات، به صورت مستقل در این بخش، مورد مطالعه قرار گرفته‌اند. البته این دسته از روش‌ها، بر اساس نحوه بکارگیری نظریه گراف در فرآیند بازیابی معنایی اطلاعات، خود به چهار زیر دسته کلی، تفکیک شده‌اند:زیر دسته اول، پژوهش‌هایی را در بر می‌گیرد که از تکنیک گراف دانش در فرآیند بازیابی معنایی اطلاعات، استفاده می‌کنند. در مقابل، زیر دسته دوم از این پژوهش‌ها، با ایجاد گراف روابط معنایی بین عناصر داده‌ای مرتبط با مسئله بازیابی اطلاعات، به انجام بازیابی معنایی اطلاعات می‌پردازند. از سوی دیگر، در زیر دسته سوم پژوهش‌های بررسی شده، از مدل‌های مبتنی بر هایپر گراف در بازیابی معنایی استفاده شده است؛ و نهایتا زیر دسته چهارم، شامل پژوهش‌هایی می‌باشد که با بکارگیری مدل پویش تصادفی، عملیات بازیابی معنایی اطلاعات را انجام می‌دهند.

جدول ۱. خلاصه مقالات ارزیابی شده در دسته الف (بازیابی اطلاعات بر مبنای ارائه بازنمایی گرافی از دادگان بازیابی اطلاعات)

مقاله	دسته‌بندی روش	سال چاپ	ایده اصلی	نقاط قوت	نقاط ضعف
[4]	الف-۱	۲۰۱۵	یافتن پژوهشگران بر اساس ادغام اطلاعات پروفایل و گراف ارجاعات به کمک عملگرهای ترکیب اطلاعات	شناسایی پژوهشگران با دقت مطلوب	وجود پارامترهای متعدد جهت پیکربندی الگوریتم
[5]	الف-۱	۲۰۱۹	بهبود توصیه‌گرها با ایجاد گراف دوبخشی بین خریداران و کالاها و استخراج ویژگی‌های گرافی و ادغام آنها با مشخصات کالاها توسط یادگیری ژرف	دقت مطلوب به ازای محصولات با تعداد فروش محدود	هزینه محاسباتی بالا
[6]	الف-۱	۲۰۱۴	مدل‌سازی تعامل کاربر با محتوا توسط نظریه گراف جهت حاشیه‌نگاری خودکار در شبکه‌های اجتماعی برخط	توجه به سوابق تعامل کاربران با داده‌ها و نیز استفاده از ابرداده متناظر با این محتوا	هزینه محاسباتی بالا و عملکرد ضعیف در شرایط محدود بودن داده
[7]	الف-۱	۲۰۱۹	ارائه مفهوم ویژگی‌های طیفی از هایپرگراف روابط بین کاربران و کالاها در سامانه توصیه‌گر	امکان استفاده در شرایط فقدان ویژگی‌های بصری و متنی برای همه کالاها یا کاربران	نیاز به وجود بازخورد صریح کاربران در قبال کالاها
[8]	الف-۱	۲۰۱۳	مدل‌سازی رفتار کاربران با استفاده از گراف دوبخشی تعاملات کاربران با داده‌ها	استفاده از اطلاعات سوابق کاربران در ارتقای فرآیند بازیابی اطلاعات	عدم مقیاس‌پذیری محاسبات در قبال کلان داده‌ها
[9]	الف-۱	۲۰۱۷	رتبه‌بندی رئوس گراف‌های دوبخشی با ترکیب مشخصات ساختاری گراف و ابرداده رئوس	قابلیت کاربرد در سامانه‌های مختلف بازیابی اطلاعات	هزینه محاسباتی بالا و عدم استفاده از شاخص‌های استاندارد ارزیابی بازیابی اطلاعات
[10]	الف-۱	۲۰۲۰	نگاشت مسئله پیشنهاد مخاطب در شبکه‌های اجتماعی به مسئله رتبه‌بندی	امکان پیاده‌سازی توسط روش‌های کم هزینه رتبه‌بندی نظیر BM25	عدم توجه به تنوع گرایش‌ها و خصوصیات افراد پیشنهادی
[11]	الف-۱	۲۰۱۸	رتبه‌بندی فرمول‌های مقالات علمی بر اساس ترکیب ویژگی‌های متنی و گراف ارجاعات درون و برون مقاله‌ای	دقت قابل قبول به دلیل ترکیب ویژگی‌های متنی و گراف	پیچیدگی زمانی زیاد
[12]	الف-۱	۲۰۱۷	رتبه‌بندی قطعات کد بر اساس ادغام مشخصات پرس‌وجوها و ویژگی‌های قطعات کد	عملکرد مطلوب در رتبه‌بندی	ارزیابی محدود روش پیشنهادی
[13]	الف-۱	۲۰۱۶	توسعه مدل فضای برداری با ایجاد گراف دوبخشی بین اسناد وب و کلیک‌های کاربران	عملکرد مطلوب در شرایط نویز و محدود بودن تعاملات کاربران	هزینه محاسباتی بالا
[14]	الف-۱	۲۰۲۰	ایجاد گراف‌های کاربران و کالاها و رتبه‌بندی محلی گره‌ها در زیرگروه‌ها و سپس ادغام رتبه‌بندی‌های محلی	امکان موازی‌سازی پردازش داده‌ها به منظور کاهش هزینه محاسبه	محدودیت عملکرد به وجود اطلاعات امتیازدهی کالاها توسط کاربران
[15]	الف-۱	۲۰۲۰	رتبه‌بندی عادلانه نتایج جستجوهای علمی با ترکیب ویژگی‌های گراف مشارکت محققین و نیز ویژگی‌های متناظر با متن مقالات	کارآمدی مناسب در رتبه‌بندی	هزینه محاسباتی بالا
[18]	الف-۱	۲۰۲۰	رتبه‌بندی موارد تست نرم‌افزار بر اساس گراف وابستگی بین اجزای نرم‌افزاری و سوابق تست	قابلیت بکارگیری در مدیریت فرآیند مهندسی نرم‌افزار در پروژه‌های کلان نرم‌افزاری	عدم امکان استفاده در نرم‌افزارهای با مقیاس کم یا متوسط
[19]	الف-۱	۲۰۲۰	رتبه‌بندی محققان شاخص توسط گراف نویسندگان مشترک	کارآمدی مناسب به خصوص شناسایی محققان جوان‌تر	عدم استفاده از همه اطلاعات گراف نویسندگان
[20]	الف-۲	۲۰۲۰	رتبه‌بندی تیم‌های ورزشی بر اساس سابقه بازی‌های انجام شده با استفاده از گراف بازی‌ها	قابلیت بکارگیری دانش اولیه و نیز امکان یادگیری رتبه‌بندی پویا	عدم توجه به عامل زمان در مدل گراف بازی‌ها
[21]	الف-۲	۲۰۱۰	توسعه روش‌های یادگیری رتبه‌بندی با مدل‌سازی مسئله رتبه‌بندی بر اساس نظریه گراف	قابلیت بکارگیری در مسائل مختلف بازیابی اطلاعات	هزینه محاسباتی بالا و عملکرد نه چندان مطلوب

۵-۱- بازیابی اطلاعات با استفاده از گراف دانش

از جمله پژوهش‌های گروه نخست، می‌توان به [۲۲ و ۲۳] اشاره نمود که در آنها ضمن مرور قابلیت‌های داده ساختار گراف‌های دانش به بیان طیف وسیعی از کاربردهای آنها در سامانه‌های پرسش و پاسخ، سامانه‌های توصیه‌گر و نیز سامانه‌های بازیابی اطلاعات می‌پردازد. در [۲۴] نشان داده شده است که جستجوی معنایی بر پایه هستان شناسی [1] می‌تواند بر اساس شاخص P@n، منجر به بهبود روش

[1] Ontology

کلاسیک جستجوی مبتنی بر کلید واژه‌های کاربران شود. برای این منظور، آنها معماری نوینی بر سامانه‌های پرسش و پاسخ ارائه کردند که با ایجاد نمایه مفهوم-محور و ادغام آن با نمایه مبتنی بر اسناد، اقدام به بازیابی معنایی اسناد نماید. بصورت مشخص، در این تحقیق از سامانه PowerAqua [۲۵] برای نگاشت کلید واژه‌های موجود در پرس‌وجوهای مطرح شده به زبان طبیعی به عناصر هستان شناسی استفاده شده است. در [۲۶] روشی برای نمایش یکپارچه پایگاه‌های داده ترکیبی ارائه شده است که قادر به ادغام داده‌های ساخت یافته و نیز داده‌های بدون ساختار است. این کار از طریق سه‌تایی RDF[1] شامل فاعل، مسند و مفعول و ایجاد نمایش گرافی متناظر با آن، صورت گرفته است. بدین ترتیب، انواع داده‌ها شامل داده‌های ساخت یافته موجود در پایگاه‌های داده رابطه ای، داده‌های غیر ساخت یافته مربوط به موجودیت‌ها و نیز روابط استخراج شده از متون غیر ساخت یافته، قابل یکپارچه سازی خواهند بود. ارزیابی‌های صورت گرفته حاکی از بهبود عملکرد این روش نسبت به روش‌های پایه مطابق شاخص‌های دقت[2] و صحت[3] است. در [۲۷] سامانه‌ای تحت عنوان SaHaRa برای جستجوی موجودیت-محور[4] ارائه شده است که قادر به جستجوی اطلاعات مورد نیاز کاربر در مخزن اخبار است. بازیابی اطلاعات با استفاده از مدل‌های زبانی صورت می‌گیرد و امکان بازیابی متون و موجودیت‌ها وجود دارد. در این سامانه دو نگاه سند-محور و موجودیت-محور به صورت توامان استفاده شده است. بر مبنای این دو دیدگاه، در SaHaRa علاوه بر متن اخبار، پیوندهای اخبار مرتبط با آنها و نیز موجودیت‌های متناظر نیز نمایش داده می‌شود. در [۲۸] چندین نوع از روابط بین موجودیت‌ها مورد بررسی قرار گرفته است تا بهترین حضور توامان آنها به منظور تولید متون داستانی، شناسایی و مورد استفاده قرار گیرد. در انجام این تحقیق، از مخزن بزرگی از دادگان TREC در رابطه با بازیابی سوالات پیچیده کاربران، استفاده شده است و کارآمدی ۱۲ ویژگی مختلف مربوط به موجودیت‌ها و حضور توامان آنها و نیز ترکیب آنها بکمک روش‌های یادگیری رتبه‌بندی، بر اساس شاخص‌های دقت و MAP، مورد ارزیابی قرار گرفته است. همچنین در [۲۹] روشی تحت عنوان MIRRE[5] برای بازیابی داده‌های چند رسانه‌ای ارائه شده که قادر است بصورت غیرخطی و چند وجهی، اسنادِ شامل داده‌های چند رسانه‌ای یکپارچه‌سازی شده را جستجو نماید. در روش پیشنهاد شده، وجوه مختلف داده‌های چند رسانه‌ای و نیز عناصر رسانه‌ای موجود در آنها (اعم از متن، صوت، تصویر، ویدیو و نظایر آن) را در فرآیند جستجو مورد استفاده قرار دهد. برای نیل به این منظور، از مدل داده‌ای گراف استفاده شده است که در آن، گره‌ها، اسناد چند رسانه‌ای هستند و یال‌ها، معادل شباهت چند وجهی و روابط معنایی بین گره‌ها می‌باشند. این شباهت‌سنجی بر مبنای مجموع مشابهت مابین عناصر داده‌ای موجود در اسناد، محاسبه می‌گردد. علاوه بر آن، در این پژوهش، واسط کاربری جدیدی نیز برای جستجوی اطلاعات چند رسانه‌ای، طراحی و پیاده‌سازی شده است. در این سامانه، داده‌های چند رسانه‌ای بصورت نمایه معکوس[6]، نمایه‌گذاری می‌شوند و پرس‌وجوهای کاربران نیز به صورت معمول، شامل تعدادی کلید واژه متنی هستند. با انجام شباهت سنجی بین پرس‌وجوهای کاربران و اجزاء چند وجهی هر داده چند رسانه‌ای، میزان ارتباط هر یک از این اجزا با پرس‌وجوی کاربر مشخص می‌شود و سپس بر اساس مدل گرافی تهیه شده، عناصر مشابه به سوال کاربر در گراف متناظر، تعیین و بازیابی می‌شوند. برای سنجش عملکرد الگوریتم ارائه شده، از دادگان I-Search استفاده شده است. این دادگان شامل بیش از ده هزار داده چند رسانه‌ای است که توصیف آنها در قالب اسناد XML نیز موجود است. ارزیابی‌های صورت گرفته بر اساس شاخص‌های تشخیص میزان کاربر پسندی سامانه، نشان می‌دهد که میزان کاربردی بودن و رضایتمندی کاربران در نتیجه استفاده از واسط کاربری تهیه شده برای ارائه نتایج عملکرد روش پیشنهادی، مطلوب بوده است. محققان این پژوهش، در نظر دارند روش‌های خوشه‌یابی و داده‌کاوی را به منظور بهبود عملکرد روش ارائه شده، در فرآیند محاسبات، دخیل کنند و در عین حال، ارتقای عملکرد واسط کاربری تهیه شده نیز مد نظر آنان است. در [۳۰] الگوریتم رتبه‌بندی جدیدی برای سامانه‌های پرسش و پاسخ[7] ارائه شده است که بر اساس خوشه‌بندی گراف‌های دانش با مقیاس بزرگ و اولویت‌بندی زیرگراف‌های ایجاد شده، عمل می‌کند به نحوی که زیرگراف‌های با اولویت بالاتر به منظور تهیه پاسخ به ازای سوالات کاربران، مورد استفاده قرار می‌گیرند. بدین ترتیب، ضمن کاهش هزینه محاسباتی کار با گراف‌های بسیار بزرگ، صرفا بخش‌هایی از گراف دانش در فرآیند پاسخ‌دهی، استفاده می‌شوند که حاوی اطلاعات با میزان حداکثر مرتبط بودن به سوالات کاربران باشند. در عین حال، بر اساس شاخص‌های ارزیابی دقت، صحت و F1، نسبت به الگوریتم‌های پایه نیز بهبود، حاصل شده است. از سوی دیگر، در [۳۱] روشی برای پیوند دهی ضمنی موجودیت‌های اطلاعاتی موجود در گراف‌های دانش ارائه شده است که در آن، موجودیت‌های

[1] Resource Description Framework
[2] Precision
[3] Recall
[4] Entity-oriented Search
[5] Multimedia Information Retrieval Results Exploration
[6] Inverted Index
[7] Question Answering

اطلاعاتی بر مبنای توییت‌های کاربران، اولویت‌بندی و مرتبط می‌شوند. بر اساس نتایج ارزیابی این روش طبق شاخص‌های MRR و P@n، روش پیشنهادی، عملکرد مناسب داشته است. همچنین در [۳۲] روشی برای برای پیشنهاد دروس به دانشجویان بر اساس رتبه‌بندی دروس بر اساس سوابق تحصیلی آنها ارائه شده است که از گراف دانش برای اولویت‌بندی دروس، استفاده می‌کند. در [۳۳] کاربرد روش‌های مبتنی بر شبکه‌های عصبی برای رتبه‌بندی گراف پرس‌وجوی کاربر به منظور حل مسئله پاسخ‌دهی به پرسش‌های پیچیده با استفاده از گراف دانش، مورد بررسی قرار گرفته است. علاوه بر آن، با توجه به ساختار گراف‌های متناظر با پرسش‌های کاربران، الگوریتم جدیدی برای اولویت‌بندی گراف‌های پرسش‌های کاربران بر مبنای گراف دانش در اختیار، ارائه شده است. به منظور سنجش عملکرد این روش نسبت به الگوریتم‌های پایه، از شاخص‌های ارزیابی استانداردی نظیر دقت، صحت، MRR و F1 استفاده شده است. علاوه بر این، در [۳۴] روشی برای انتساب خطاهای گزارش شده سامانه‌های نرم‌افزاری به مولفه‌های تشکیل دهنده آنها، ارائه شده است که قادر است بر اساس اطلاعات موجود در سوابق انتساب‌های درست یا نادرست قبلی، عملکرد مطلوبی داشته باشد. برای این منظور، بر اساس وظایف مولفه‌های مختلف سامانه نرم‌افزاری مورد نظر و نیز روابط منطقی بین آنها، یک گراف دانش، ایجاد می‌شود و طی فرآیند یادگیری، انتساب صحیح خطاها به مولفه‌های نرم‌افزاری متناظر، صورت می‌گیرد. در مقایسه عملکرد این روش نسبت به دیگر روش‌های مشابه، شاخص‌های P@n و NDCG@n، بهره گرفته شده است. همچنین در [۳۵] روشی برای بازیابی موجودیت‌های اطلاعاتی متناظر با نیازهای کاربران از گراف‌های دانش بسیار بزرگ، ارائه شده است که در آن، ابتدا زیرگراف‌های مرتبط با پرس‌وجوهای کاربران، شناسایی و هرس می‌شوند و سپس با استفاده از شبکه‌های عصبی عمیق، اقدام به اولویت‌بندی موجودیت‌های مرتبط در این زیرگراف‌ها می‌کند. همچنین در ارزیابی عملکرد الگوریتم ارائه شده نسبت به دیگر روش‌های شاخص از معیارهای ارزیابی دقت و MAP استفاده شده است.

۵-۲- ایجاد گراف موجودیت بر اساس دادگان مسئله بازیابی اطلاعات

دسته بعدی از پژوهش‌ها مربوط به ایجاد گراف روابط معنایی بر اساس دادگان مسئله بازیابی اطلاعات می‌باشد. از جمله این تحقیقات می‌توان به [۱] اشاره نمود که در آن مروری بر الگوریتم‌های ارائه شده در حوزه جستجوی موجودیت-محور ارائه شده است. بر اساس این مطالعه، فاصله معناداری بین روش‌های بازیابی متن-محور و بازیابی دانش-محور وجود دارد. الگوریتم‌های بازیابی متن-محور، قابلیت لحاظ نمودن روابط معنایی پیچیده را ندارند، در حالیکه روش‌های بازیابی دانش-محور، فاقد امکان پردازش پرس‌وجوهای مبتنی بر کلید واژه‌های کاربران و نیز بازیابی بر اساس رتبه‌بندی نتایج، می‌باشند. از سوی دیگر در [۳۶] روشی برای بازنمایی اسناد بکمک گراف مفهوم[1] ارائه شده است تا از این طریق بتوان شباهت معنایی اسناد را تعیین نمود. در این تحقیق از ابزار TAGME[2] به منظور اتصال اسناد به مفاهیم Wikipedia استفاده شده است. سپس گرافی از مفاهیم به عنوان رئوس و روابط بین آنها به عنوان یال‌ها تشکیل می‌شود. در این گراف، وزن یال‌ها متناظر با نوع رابطه متناظر با مفاهیم دو سوی آن یال، تعیین می‌شود. این روابط می‌توانند از سه گونه زمینه، رده و ساختار باشند. ضمنا وزن متناظر با رئوس هم معادل میزان مرکزیت نزدیکی[3] آنها در نظر گرفته شده است. بر اساس این وزن دهی، میزان شباهت مفاهیم بصورت تابعی از شباهت رئوس متناظر در گراف مفهوم و نیز وزن یال‌های متصل کننده آنها تعیین می‌شود. مقایسه این روش با الگوریتم‌های مشابه، مطابق شاخص همبستگی پیرسن[4]، صورت گرفته است. از سوی دیگر در [۳۷] با استفاده از فاکتورسازی تنسور[5]، تنسور[5]، معناشناسی نهفته در روابط موجودیت‌ها استخراج شده است. در این الگوریتم، به منظور تشریح روابط بین موجودیت‌ها بر اساس مسندهای مختلف، یک تنسور تعریف شده است که بصورت چندین ماتریس مجاورت می‌باشد و هر ماتریس به ازای هر مسند روی بعد سوم تنسور، در نظر گرفته شده است. همچنین برای رتبه‌بندی نتایج، از روش یادگیری رتبه‌بندی جفتی استفاده شده که عملکرد آن با استفاده از درخت‌های رگرسیون ارتقا یافته، به ازای شاخص NDCG بهینه شده است. در این کار از ویژگی‌های واژه-محور و نیز ویژگی‌های ساختاری، استفاده شده است. ضمنا بایستی اشاره نمود که ارزیابی عملکرد این روش، بر اساس معیارهای P@n، MAP و NDCG@n انجام شده است. همچنین در [۳۸] روشی برای حل مسئله پیوند زمینه‌ای[6] ارائه شده است. در این مسئله کوشش می‌شود هر خبر ارائه شده، جامع باشد به نحوی که شامل پیشینه موضوع آن خبر و نیز اطلاعات زمینه‌ای مرتبط باشد تا بتواند

[1] Concept Graph
[2] https://tagme.d4science.org/tagme/
[3] Closeness centrality
[4] Pearson Correlation
[5] Tensor Factorization
[6] Background Linking

درک صحیحی را در ذهن مخاطب ایجاد نماید. روش پیشنهاد شده که بر مبنای الگوریتم رتبه‌بندی LambdaMART می‌باشد، برای رتبه‌بندی هر خبر، ترکیبی از ویژگی‌های استاندارد متناظر با متن آن خبر و نیز ویژگی‌های مربوط به گراف موجودیت‌ها را ایجاد می‌کند. ابتدا به ازای هر خبر، موجودیت‌های بالقوه، شناسایی و استخراج می‌شوند و سپس گراف روابط بین موجودیت‌ها، بر مبنای ارتباطات مابین آنها تشکیل می‌شود. برای اینکار از مخزن دانش ویکی‌پدیا، به عنوان گراف کلی روابط معنایی بین موجودیت‌ها، استفاده می‌شود. در این گراف، که غیر جهت‌دار و وزن‌دار می‌باشد، رئوس، همان موجودیت‌ها هستند و وزن یال‌ها، معادل میزان ارتباط معنایی بین موجودیت‌های متناظر است. در ادامه، گراف بدست آمده هرس می‌شود بنحوی که تنها شامل بزرگترین مولفه همبند باشد. سپس با بکارگیری الگوریتم تشخیص انجمن Girvan-Neman، از مولفه همبند فوق، بزرگترین انجمن موجود، شناسایی و استخراج می‌شود. در مرحله بعدی، استخراج ویژگی‌های متناظر با اسناد صورت می‌گیرد. این ویژگی‌ها که خود یا مرتبط با اسناد یا پرس‌وجوهای کاربران هستند، می‌توانند از متن اسناد و نیز گراف معنایی روابط مابین موجودیت‌ها استخراج شوند. در گام بعدی، با استفاده از الگوریتم یادگیری رتبه‌بندی LambdaMART، مدل‌های رتبه‌بندی مختلفی تولید می‌شود که در هر یک از آنها، از زیرمجموعه محدودی از کل ویژگی‌های تولید شده طی مراحل قبلی، استفاده شده است. نهایتاً در مرحله ترکیب، تعداد مشخصی از مدل‌های مرحله قبلی با الگوریتم ساده CombSUM، با هم ادغام می‌شوند تا خروجی نهایی، بدست آید. الگوریتم ارائه شده در حالات مختلف ترکیب با تعداد اجراهای متفاوت با الگوریتم پایه BM25 با شاخص‌های NDCG، P@n و Reciprocal Rank مقایسه می‌شود. برای ارزیابی روش پیشنهادی، از Washington Post Corpus vs2 و Annotated Corpus New York Times، استفاده شده است. با این کار مجموعه آموزشی غنی‌تر شده است اما این کار به دلیل عدم هماهنگی بین درجات مرتبط بودن در این دو مجموعه داده، باعث محدودیت‌های ناشی از یادگیری انتقالی[1] شده است. نتایج نشان می‌دهد که تعداد اجراهای بیشتر ترکیب منجر به بهبود نتیجه رتبه‌بندی نمی‌شود. با این حال، مطابق نتایج ارزیابی طبق شاخص‌های P@n و NDCG@n، ایده ترکیب نتایج، منجر به بهبود عملکرد می‌شود. به علاوه هر چه اسناد بیشتری برای رتبه‌بندی بکار گرفته شوند، عملکرد الگوریتم پیشنهادی نسبت به الگوریتم BM25 پایین می‌آید. دلایل ضعیف بودن روش پیشنهادی آن است که اولا روش یادگیری انتقالی به کارگرفته شده است، که از رسیدن بعضی موضوعات به بالاترین درجه اثربخشی جلوگیری می‌کند. دوم اینکه ممکن است از اسناد، گراف‌های موجودیت خوش فرمی ایجاد نشود و این امر باعث می‌شود الگوریتم یادگیرنده صرفا از ویژگی‌های متن برای امتیازدهی استفاده کند.

۵-۳- مدل‌های مبتنی بر هایپر گراف

از جمله روش‌های دسته سوم که مبتنی بر بکارگیری هایپر گراف‌ها می‌باشند، می‌توان به [۳۹] اشاره نمود که در آن هایپر گراف پرس‌وجوهای کاربران، پیشنهاد و طراحی شده است. در این هایپر گراف، گره‌ها متناظر با مفاهیم متناظر با پرس وجو هستند و یال‌ها، وابستگی‌های بین زیر مجموعه‌هایی از این گره‌ها و اسناد را نمایش می‌دهند. بر این اساس، هایپر گراف پرس وجوها قادر به مدلسازی سطوح بالاتر وابستگی بین واژه‌ها می‌باشد. به منظور تخصیص امتیاز به زوج‌های اسناد و پرس وجوها، از نمایش گراف فاکتور به ازای هایپر گراف پرس وجوها استفاده شده است. برای مقایسه عملکرد این روش با الگوریتم‌های پایه از شاخص‌های MAP و ERR@n، استفاده شده است. از سوی دیگر در [۴۰] مفهومی تحت عنوان هایپرگراف معنایی[2] معرفی شده است که راهکاری برای بازنمایی دانش استخراج شده از مخازن داده بر اساس هایپرگراف‌های مرتب معکوس می‌باشد. در این ساختار، گره‌ها که معادل واژه‌ها هستند، می‌توانند دارای ترتیب نسبی در هایپر یال متناظر باشند و موجودیت‌ها را به صورت توالی از کلمات، توصیف نمایند. همچنین، با توجه به ویژگی بازگشتی بودن بودن این هایپرگراف، می‌توان هر هایپر یال را بر اساس دیگر هایپریال‌ها بیان نمود. محصول این پژوهش، ارائه یک کتابخانه نرم‌افزاری در زبان پایتون موسوم به GraphBrain[3] است که امکان پردازش هایپر گراف‌های معنایی را در کاربردهای مرتبط با پردازش زبان طبیعی و جستجو و استنتاج دانش، فراهم می‌سازد. همچنین در [۴۱] الگوریتمی تحت عنوان ENT Rank، یک روش رتبه‌بندی موجودیت‌ها بر اساس بکارگیری هایپر گراف‌ها ارائه شده است، که در آن از اطلاعات متنی به منظور بهبود بازیابی موجودیت‌ها استفاده شده است. در این الگوریتم، هایپر گراف مورد استفاده به گراف چندگانه متناظر با حضور توامان موجودیت‌ها، تبدیل می‌شود و در آن از ویژگی‌های متعددی نظیر ویژگی‌های همسایگان و نیز ویژگی‌های مرتبط بودن زمینه، به منظور ایجاد مدل یادگیری رتبه‌بندی، استفاده می‌شود. در واقع مدل پیشنهادی، گونه‌ای از الگوریتم پویش تصادفی با راه‌اندازی مجدد

[1] Transfer Learning
[2] Semantic Hypergraph
[3] https://graphbrain.net

تصادفی، محسوب می‌شود. در ارزیابی عملکرد این روش، از شاخص‌های MAP و NDCG@n استفاده شده است.

۵-۴- مدل‌های مبتنی بر گراف پویش تصادفی

دسته چهارم از کاربردهای نظریه گراف در بازیابی معنایی اطلاعات مربوط به مدل‌های بازیابی معنایی مبتنی بر پویش تصادفی می‌باشد. در این دسته می‌توان از الگوریتم‌های نظیر ObjectRank و HopRank نام برد که روی گراف‌های RDF قابل اعمال هستند. همچنین روش‌های DING و PageRank معنایی نیز مطرح هستند که قابلیت ترکیب گراف وب و گراف موجودیت‌ها را دارند. الگوریتم ObjectRank [۴۲] بر پایه الگوریتم PageRank برای جستجوی کلید واژه‌ها در پایگاه‌های داده‌ای است که به صورت گراف‌های برچسب‌دار، مدلسازی شده‌اند. در واقع در این الگوریتم، مشابه الگوریتم HITS از یک مجموعه پایه از اسناد به منظور ایجاد گراف موضوعی، استفاده می‌شود. در این الگوریتم، دو شاخص ObjectRank سراسری و نیز ObjectRank به ازای همه گراف‌های مبتنی بر کلید واژه‌ها، محاسبه و با هم ادغام می‌شوند. مقایسه این الگوریتم نسبت به روش‌های مشابه بر اساس پیچیدگی زمان و فضای محاسباتی مورد نیاز، انجام شده است. الگوریتم HubRank توسعه‌ای از الگوریتم ObjectRank می‌باشد که منجر به ارتقای کارآمدی آن می‌شود [۴۳]. در سنجش عملکرد روش پیشنهادی، علاوه بر شاخص‌های پیچیدگی زمان و فضای محاسباتی مورد نیاز، از شاخص صحت نیز استفاده شده است. الگوریتم HopRank به منظور مدلسازی نحوه تعامل انسانی با شبکه‌های معنایی طراحی شده است [۴۴]. بر اساس نتایج حاصل از تحلیل رفتار کاربران انسانی در وبسایت BioPortal[1] که مخزنی از هستان‌شناسی‌های زیست پزشکی است، مشخص شده است که کاربران بجای حرکت تصادفی روی گره‌های آنتولوژی مورد بررسی، غالباً گرایش به پرش به گره‌هایی در فاصله مشخص k از گره فعلی را دارا می‌باشند. احتمال متناظر با این پرش که تحت عنوان k-hop نامیده می‌شود، افرازی روی گره‌ها تحت عنوان HopPortation را به وجود می‌آورد.

گروه دوم از روش‌های مطرح شده در دسته چهارم از الگوریتم‌های بازیابی مبتنی بر نظریه گراف، شامل الگوریتم‌های نظیر DING و PageRank معنایی می‌باشد. بطور مشخص، در الگوریتم DING، یک روش تحلیل سلسله مراتبی از پیوندها ارائه شده است که بر اساس یکی از گسترش‌های PageRank تحت عنوان DatasetRank عمل می‌کند [۴۵] و روی یک مدل دولایه‌ای از وب معنایی، اعمال می‌شود. ضمنا ایده اصلی الگوریتم DatasetRank، ادغام رتبه محلی موجودیت‌ها و نیز احتمال پرش به دیگر مجموعه‌های داده، می‌باشد. در ارزیابی عملکرد این روش، از معیارهای ارزیابی استاندارد حوزه بازیابی اطلاعات نظیر P@n، MAP و NDCG@n استفاده شده است. همچنین در الگوریتم PageRank معنایی، توسعه‌ای از الگوریتم پایه PageRank به منظور پیشنهاد پیوند بر اساس گراف علایق کاربران، مطرح شده است [۴۶]. در این پژوهش از دو نوع بازنمایی گرافی از علایق کاربران، استفاده شده است که یکی از آنها بصورت دوتایی‌هایی از کاربر-داده است و دومی شامل سه‌تایی‌هایی از کاربر-داده-منبع می‌باشد. مطابق ارزیابی‌های صورت گرفته طبق شاخص‌های MAP و NDCG، بهترین عملکرد این الگوریتم مربوط به شرایطی است که در آن، گراف‌های متناظر با علایق کاربران، تراکم کمتری داشته باشند.

در جدول ۲، خلاصه مقالات ارزیابی شده ذیل دسته ب (بازیابی معنایی اطلاعات با استفاده از نظریه گراف)، بیان شده است.

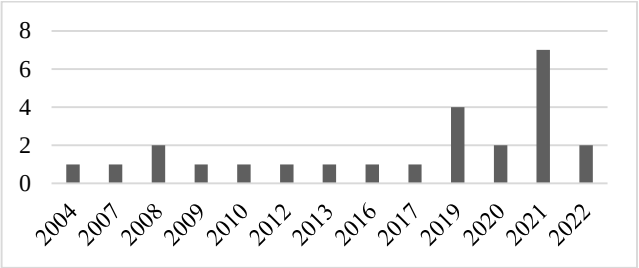
به لحاظ آماری، تعداد مقالات مرتبط بررسی شده در دسته دوم در شکل ۳ آمده است. مطابق این آمار، کاربرد نظریه گراف در فرآیند بازیابی معنایی اطلاعات، در سال‌های اخیر مورد توجه بیشتری قرار گرفته است.

جدول ۲. خلاصه مقالات ارزیابی شده در دسته ب (بازیابی معنایی اطلاعات با استفاده از نظریه گراف)

مقاله	دسته‌بندی روش	سال چاپ	ایده اصلی	نقاط قوت	نقاط ضعف
[3]	ب-۲	۲۰۲۱	مرور مدل‌های مبتنی بر گراف برای جستجوی موجودیت-محور	بررسی و دسته‌بندی مدل‌های متعدد جستجوی موجودیت-محور مبتنی بر گراف	محدود بودن به جستجوی موجودیت-محور
[22]	ب-۱	۲۰۲۰	بررسی کاربردهای گراف دانش در سامانه‌های پرسش و پاسخ و سامانه‌های توصیه‌گر	شناسایی و دسته‌بندی تحقیقات صورت گرفته در زمینه کاربرد گراف دانش و چالش‌های آن	محدود بودن مقاله به معرفی روش‌های پایه و عدم بررسی توسعه‌های روش‌های مطالعه شده
[23]	ب-۱	۲۰۲۲	بررسی روش‌های استخراج گراف دانش و به ویژه گراف‌های دانش زمان-محور و بررسی برخی از	ارائه دسته‌بندی سلسله مراتبی از پژوهش‌های انجام شده و نیز نرم‌افزارهای	عدم توجه به نحوه کاربری روش‌های بررسی شده در دسته‌بندی ارائه شده

[1] https://bioportal.bioontology.org

		کاربردهای گراف دانش در سامانه‌های بازیابی اطلاعات		مرتبط	
[24]	ب-۱	توسعه یک روش جستجوی معنایی بر پایه هستان شناسی	۲۰۰۸	ایجاد نمایه مفهوم-محور و ادغام آن با نمایه مبتنی بر اسناد به منظور بازیابی معنایی اسناد	محدود بودن اطلاعات قابل استخراج از گراف دانش مبتنی بر وب معنایی
[26]	ب-۱	ارائه روشی برای نمایش یکپارچه پایگاه‌های داده ترکیبی بر اساس گراف‌های RDF	۲۰۰۸	امکان ادغام داده‌های ساخت یافته و نیز داده‌های بدون ساختار	عدم استفاده از روش‌های پردازش زبان طبیعی در ایجاد گراف
[27]	ب-۱	جستجوی موجودیت-محور با استفاده از مدل‌های زبانی	۲۰۰۹	ترکیب توامان دو نگاه سند-محور و موجودیت-محور در انجام جستجو	محدود بودن کاربری به جستجوی اخبار برخط
[28]	ب-۱	استخراج ویژگی‌های وابسته به روابط بین موجودیت‌ها جهت پاسخ‌دهی به سوالات پیچیده کاربران	۲۰۲۱	بهره‌گیری از روش‌های یادگیری رتبه‌بندی برای ادغام ویژگی‌های استخراج شده	عملکرد ضعیف در شرایط غنی نبودن پایگاه دانش
[29]	ب-۱	بازنمایی رابطه معنایی بین اسناد چند رسانه‌ای با استفاده از مدل گراف	۲۰۲۱	امکان تعمیم روش پیشنهادی به جستجو در انواع مختلف داده	عدم استفاده از روش‌های یادگیری ماشین
[30]	ب-۱	رتبه‌بندی نتایج سامانه‌های پرسش و پاسخ بر اساس خوشه‌بندی گراف‌های دانش حجیم	۲۰۲۱	کاهش هزینه محاسباتی ناشی از پردازش گراف‌های بسیار بزرگ	بزرگی نسبی زیرگراف‌های تولید شده
[31]	ب-۱	پیوند دهی ضمنی موجودیت‌های اطلاعاتی موجود در گراف‌های دانش	۲۰۲۱	اولویت‌بندی موجودیت‌های اطلاعاتی طبق توییت‌های کاربران	عدم ترکیب اطلاعات ویژگی‌های صریح و ضمنی با یکدیگر
[32]	ب-۱	پیشنهاد دروس به دانشجویان بر اساس سوابق تحصیلی آنها	۲۰۲۰	استفاده از گراف دانش برای اولویت‌بندی دروس	نیاز به حجم نسبتا زیاد داده برای آموزش مدل تصمیم‌گیری
[33]	ب-۱	اولویت‌بندی گراف‌های پرسش‌های کاربران بر مبنای گراف دانش موجود	۲۰۱۹	قابلیت پاسخ‌دهی به پرسش‌های پیچیده با استفاده از گراف دانش	عدم امکان انتقال دانش از مدل‌های زبانی تولید شده قبلی
[34]	ب-۱	انتساب خطاهای گزارش شده سامانه‌های نرم‌افزاری به مولفه‌های تشکیل دهنده آنها	۲۰۲۱	ایجاد گراف دانش بر اساس وظایف مولفه‌های مختلف سامانه نرم‌افزاری مورد نظر و نیز روابط بین آنها	عدم استفاده از اطلاعات مربوط به موضوع و رده خطا
[35]	ب-۱	بازیابی موجودیت‌های اطلاعاتی طبق نیاز کاربر از گراف‌های دانش بسیار بزرگ	۲۰۲۲	استخراج روابط زمینه‌ای بین موجودیت‌ها برای فرآیند بازیابی	نیاز به حجم نسبتا زیاد داده آموزشی
[36]	ب-۲	محاسبه شباهت معنایی اسناد به کمک گراف مفهوم	۲۰۱۶	استفاده از مفاهیم ویکیپدیا در مدل‌سازی	دقت نسبتا پایین مولفه تشخیص هدف اسناد
[37]	ب-۲	استخراج معناشناسی نهفته در روابط موجودیت‌ها	۲۰۱۳	مقیاس پذیری مناسب روی گراف‌های حجیم	نیاز به حجم زیاد داده برای آموزش سیستم
[38]	ب-۲	حل مسئله پیوند زمینه‌ای با ترکیب ویژگی‌های متناظر با متن خبر و نیز ویژگی‌های مربوط به گراف موجودیت‌ها	۲۰۲۱	بکارگیری خصوصیات متناظر با گراف موجودیت‌ها و تلفیق آن با ویژگی‌های متنی	عدم امکان استفاده از الگوریتم پیشنهادی به ازای خبرهای با میزان محتوای متنی محدود
[39]	ب-۳	استفاده از هایپرگراف پرس وجوها به منظور مدلسازی سطوح بالاتر وابستگی بین واژه‌ها	۲۰۱۲	تخصیص امتیاز به زوج‌های اسناد و پرس وجوها توسط نمایش گرافی	هزینه محاسباتی نسبتا بالا
[40]	ب-۳	بازنمایی دانش استخراج شده از مخازن داده بر اساس هایپرگراف‌های مرتب معکوس	۲۰۱۹	ارائه یک کتابخانه نرم‌افزاری در زبان پایتون	محدود بودن دادگان مورد استفاده در تولید مدل
[41]	ب-۳	رتبه‌بندی موجودیت‌ها بر اساس بکارگیری هایپر گراف‌ها	۲۰۱۹	تبدیل هایپرگراف مورد استفاده به گراف چندگانه متناظر با حضور توامان موجودیت‌ها	عدم استفاده از داده‌های متنی در تولید هایپر گراف
[42]	ب-۴	توسعه الگوریتم PageRank برای جستجو در پایگاه‌های داده مدلسازی به صورت گراف‌های برچسب دار	۲۰۰۴	استفاده از ایده روش HITS در ایجاد گراف موضوعی از اسناد موجود	تعداد زیاد پارامترها در روش پیشنهادی
[43]	ب-۴	توسعه روش [42]	۲۰۰۷	استفاده از روش جستجوی مبتنی بر مجاورت در گراف روابط معنایی بین موجودیت‌ها	هزینه محاسباتی بالا بدلیل عدم نمایه سازی و خوشه‌بندی گراف مورد پردازش
[44]	ب-۴	توسعه روش [43]	۲۰۱۹	مدلسازی نحوه تعامل انسانی با شبکه‌های معنایی	نیاز به حجم نسبتا زیاد سوابق تعاملات انسانی
[45]	ب-۴	توسعه الگوریتم PageRank با استفاده از مدل دو لایه‌ای از وب معنایی	۲۰۱۰	ادغام رتبه محلی موجودیت‌ها	رتبه‌بندی مستقل از پرس‌وجوی کاربر
[46]	ب-۴	توسعه الگوریتم PageRank با کمک تکنیک‌های بازیابی معنایی	۲۰۱۷	بازنمایی گرافی از علایق کاربران	محدود بودن به منابع مستقیما مرتبط با علایق کاربران



شکل ۳. روش‌های وب کاوی

۶- بکارگیری نظریه گراف در یادگیری رتبه‌بندی

رویکرد یادگیری رتبه‌بندی به عنوان یک رویکرد نوین و در عین حال کارآمد در حل مسئله رتبه‌بندی، مورد اقبال گسترده جامعه پژوهشگران قرار گرفته است و از این رو، تلاش بر ارتقای عملکرد روش‌های موجود و یا ارائه راهکارهای نوین، طی حدود یک دهه اخیر، همواره مورد توجه جدی، بوده است. بر این اساس، در این بخش، به صورت خاص به بررسی نحوه بکارگیری نظریه گراف در حل مسئله یادگیری رتبه‌بندی، پرداخته شده است. با توجه به تقسیم‌بندی کلی روش‌های یادگیری رتبه‌بندی به دسته عمده تجمیع رتبه‌بندی و ایجاد رتبه‌بندی [۴۷]، الگوریتم‌های مورد بررسی در این بخش، ذیل دو زیر دسته فوق، تقسیم‌بندی شده‌اند.

۶-۱- مدل‌های مبتنی بر تجمیع رتبه‌بندی

از جمله این پژوهش‌ها می‌توان به [۴۸] اشاره نمود که در آن، روشی برای حل مسئله ادغام رتبه‌بندی[1] بر اساس بکارگیری نظریه گراف‌ها ارائه شده است.در این پژوهش، از یادگیری بدون نظارت استفاده می‌شود. این روش قادر است نسبت به ترکیب لیست‌های رتبه‌بندی داده که از رتبه‌بندهای مختلف دریافت می‌شود، اقدام نماید. روش پیشنهادی بر اساس مفهومی تحت عنوان گراف ادغام[2] بنا شده است. بر این اساس، روش ارائه شده، شامل یک فرآیند دو مرحله‌ای است که مرحله برون‌خط، وظیفه بازنمایی مجموعه پاسخ بصورت گراف‌های ادغام را بر عهده دارد. در مقابل، بخش برخط، مسئولیت پردازش پرس‌وجوی دریافتی از کاربر و تهیه رتبه‌بندی نهایی را دارا می‌باشد. در هر دو مرحله، گام استخراج گراف ادغام وجود دارد. طی این گام، بر اساس رتبه‌بندی دریافت شده از هر یک از رتبه‌بندی کننده‌های پایه، یکپارچه‌سازی رتبه‌بندی‌های متناظر با پرس‌وجوی مورد نظر انجام می‌شود. بدین ترتیب، فرآیند جستجو به مسئله یافتن زیرگراف‌های کمینه مشترک[3]، تبدیل می‌شود. بررسی عملکرد روش پیشنهادی روی دادگان مختلفی صورت گرفته است که مربوط به داده‌های متنی و تصاویر و نیز داده‌های چند رسانه‌ای می‌باشند. این دادگان عبارتند از: Ohsumed و Brodatz برای اطلاعات متنی، MPEG-7 و Soccer برای اطلاعات تصاویر و نیز دادگان UW و UKBench برای دادگان شامل ترکیب متون و داده‌های چند رسانه‌ای. بررسی‌های صورت گرفته روی دادگان فوق، مطابق شاخص NDCG@n، نشان داده است که روش ارائه شده، عملکرد مناسب‌تری نسبت به روش‌های پایه در مسئله ادغام رتبه‌بندی نظیر CombSUM، CombMNZ، CombMIN، CombMAX و BordaCount داشته است. مقایسه روش فوق با روش‌های یادگیری تحت نظارت و نیز ارتقای الگوریتم فوق برای مقیاس‌پذیری در شرایط واقعی، از جمله مسیرهای توسعه آتی این روش است. همچنین در [۴۹] روشی برای یکپارچه‌سازی فهرست پیشنهادات در سامانه‌های توصیه‌گر ارائه شده است که بر مبنای الگوریتم فرا ابتکاری[4] موسوم به تکامل تفاضلی[5]، عمل می‌کند. در واقع، روش پیشنهادی یک الگوریتم یادگیری تحت نظارت است که در آن، با بکارگیری الگوریتم تکامل تفاضلی، تلاش شده است تا عملکرد روش پیشنهادی بر مبنای شاخص متوسط دقت[6] (AP)، بهینه شود. روش ارائه شده، این قابلیت را دارد که بر مبنای علایق کاربر، تنظیم شود. در پیاده‌سازی روش پیشنهادی، از چهار الگوریتم پایه SVD، BPR، WMF و WARP استفاده شده است . هر چهار الگوریتم فوق، بر مبنای روش فاکتور سازی ماتریسی[7] طراحی شده‌اند و ویژگی‌های پایه هر فرد جمعیت تکاملی را به ازای هر کالا و هر کاربر، بر اساس ماتریس کالا-کاربر، محاسبه می‌کنند. این ویژگی‌ها را اصطلاحاً ویژگی‌های نهفته[8] می‌نامند. به منظور بررسی عملکرد روش فوق، از دادگان MovieLens 100k استفاده شده است. ارزیابی‌های صورت گرفته نشان می‌دهد که الگوریتم ارائه شده، بر مبنای شاخص‌های ارزیابی نظیر P@n و MAP، قادر است نسبت به الگوریتم‌های پایه ترکیب اطلاعات نظیر Borda Count و Majority Voting، عملکرد مناسب‌تری داشته باشد. به منظور توسعه روش پیشنهادی، می‌توان تکنیکی برای ارزیابی اولیه توصیه‌کننده‌های پایه، طراحی نمود به نحوی که پیش از یکپارچه‌سازی فهرست‌های پیشنهادی، پیشنهاد

[1] Rank Aggregation
[2] Fusion Graph
[3] Minimum Common Subgraphs
[4] Metaheuristic
[5] Differential Evolution
[6] Average Precision
[7] Matrix Factorization
[8] Latent Features

دهنده‌های ضعیف، شناسایی و حذف شوند. از سوی دیگر، تصمیم‌گیری برای محل تحصیل، یکی از دغدغه‌های اصلی دانشجویان است. وجود رتبه‌بندی‌های مختلف، اتخاذ تصمیم مناسب را با مشکلاتی مواجه می‌کند. بر این اساس، در پژوهش [۵۰]، روشی برای رتبه‌بندی دانشگاه‌های جهان، ارائه شده است که قادر است از ترکیب رتبه‌بندی‌های مختلف انجام شده، فهرست یکپارچه‌ای را تولید کند. روش پیشنهادی، یک الگوریتم مبتنی بر گراف است که بر پایه ایجاد و پردازش گراف رقابت بین دانشگاه‌های مختلف، طراحی شده است. در این گراف جهت دار و وزن‌دار، هر گره، معادل یک دانشگاه است و هر یال از یک دانشگاه به دانشگاه دیگر، به صورت تابعی از میزان برتری دانشگاه اول نسبت به دانشگاه دوم در رتبه‌بندی‌های پایه است. بر مبنای این برتری، شاخصی تحت عنوان کیفیت برای هر دانشگاه تعریف می‌شود که معادل نسبت درجه خروجی به ورودی به ازای راس متناظر در گراف رقابت است. رتبه‌بندی نهایی دانشگاه‌ها بر اساس شاخص فوق، صورت می‌گیرد. به منظور بررسی و تحلیل روش پیشنهادی، از مجموعه پنج دادگان مختلف شامل USNEWS، ARWU، QS، THE و URAP استفاده شده است. آزمایش‌های انجام شده طبق شاخص‌های دقت و صحت، نشان می‌دهد روش پیشنهادی، قادر است رتبه‌بندی جامعی از دانشگاه‌ها را ارائه دهد و در عین حال، نویز موجود در لیست‌های رتبه‌بندی دریافتی را نیز حذف کند. پژوهشگران این تحقیق در نظر دارند کاربرد الگوریتم پیشنهادی را در حوزه‌های دیگری نظیر رتبه‌بندی برندها، رتبه‌بندی نتایج جستجو در وب و رتبه‌بندی تیم‌های ورزشی، بررسی کنند. علاوه بر آن، در [۵۱] روشی مبتنی بر گراف برای ترکیب رتبه‌بندی در حوزه بازیابی تصاویر، ارائه شده است. ایده اصلی در الگوریتم، انتخاب ویژگی‌های موثر در فرآیند رتبه‌بندی است. برای این منظور، گرافی ایجاد می‌شود که گره‌های آن، ویژگی‌های مختلف تصاویر نظیر رنگ، بافت و نظایر آنها هستند و وزن یال‌های بین گره‌ها، میزان مکمل بودن نسبی آنها را نشان می‌دهد. این گراف، بصورت تدریجی، بر مبنای آستانه‌های مختلف میزان موثر بودن ویژگی‌ها، بروز رسانی می‌شود تا بر این اساس، امکان شناسایی موثرترین ویژگی‌ها حاصل شود. سپس یال‌های این گراف بر مبنای آستانه‌های همبستگی رتبه‌بندی ویژگی‌های متناظر، بروز می‌شود. بدین ترتیب، یال‌ها شامل اطلاعات میزان مکمل هم بودن ویژگی‌ها خواهند بود. گراف تولید شده، قادر است تخمینی از میزان موثر بودن هر ویژگی و مکمل آن نسبت به بقیه ویژگی‌ها را بدست دهد. ویژگی‌های منتخب نهایی بر مبنای مولفه‌های همبند گراف فوق، تعیین می‌شوند. ارزیابی عملکرد این الگوریتم روی دادگان مختلفی نظیر Flowers، Holidays، Corel5k و UKBench، طبق شاخص MAP نشان می‌دهد که روش پیشنهادی قادر است ترکیبات کارآمدی از ویژگی‌های کلیدی در فرآیند رتبه‌بندی را شناسایی و مورد استفاده قرار دهد. بکارگیری روش‌های مختلف ترکیب اطلاعات و ترکیب گونه‌های مختلف الگوریتم ارائه شده، از جمله مسیرهای تحقیقاتی پیشنهای در این پژوهش است.

۶-۲- مدل‌های مبتنی بر ایجاد رتبه‌بندی

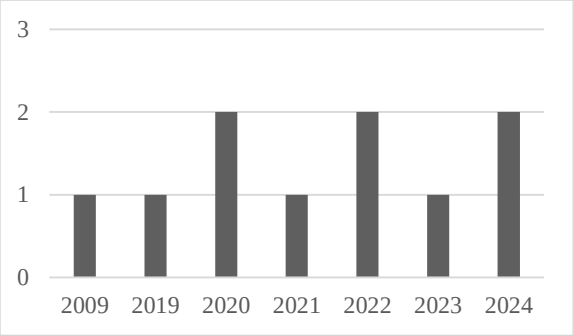
از نمونه‌های کاربرد نظریه گراف در ایجاد رتبه‌بندی می‌توان به [۵۲ و ۵۳] اشاره نمود که در آنها الگوریتم یادگیری رتبه‌بندی جدیدی ارائه شده است که طی آن، ابتدا بر مبنای محاسبه شاخص Kendall's tau بین هر زوج از ویژگی‌های موجود در دادگان پایه مورد استفاده در یادگیری رتبه‌بندی، گراف شباهت بین ویژگی‌های دادگان یادگیری رتبه‌بندی، ایجاد می‌گردد. در ادامه، بر مبنای یک آستانه حداقل شباهت، این گراف، هرس می‌شود. سپس از الگوریتم خوشه‌یابی طیفی[1] به منظور اِفراز گراف فوق، استفاده می‌شود تا ویژگی‌های با میزان مشابهت قابل توجه، در خوشه‌های یکسانی قرار گیرند. در مرحله بعد و بر مبنای خوشه‌یابی بعمل آمده، دو گام موازی، طی می‌شود. در یکی از این دو گام، از هر خوشه، زیر مجموعه محدودی از ویژگی‌های مهم‌تر، انتخاب می‌شود و نهایتاً از ترکیب آنها، الگوریتم یادگیری رتبه‌بندی جدیدی، حاصل می‌شود. در گام موازی دوم نیز از گراف شباهت ویژگی‌ها، برای محاسبه میزان اهمیت و مرتبط بودن نسبی ویژگی‌های پایه به کمک الگوریتم Biased PageRank، استفاده می‌شود. برای ارزیابی عملکرد الگوریتم ارائه شده، از مجموعه دادگان LETOR، استفاده شده است. نتایج حاصل از آزمایش‌های انجام شده، طبق شاخص‌های MAP و NDCG@n نشان می‌دهد که الگوریتم پیشنهاد شده، عملکرد بهتری نسبت به روش‌های مطرح در یادگیری رتبه‌بندی نظیر RankSVM دارد. به منظور توسعه این روش، می‌توان از روش‌های ترکیب اطلاعات بر مبنای وزن‌دهی نسبی ویژگی‌های منتخب در هر خوشه استفاده نمود. همچنین استفاده از ویژگی‌های منتخب در توسعه روش‌های رتبه‌بندی پایه، جالب توجه می‌باشد. علاوه بر آن بررسی استفاده از شاخص‌های آماری دیگری نظیر Spearman's rho جهت ایجاد گراف مشابهت بین ویژگی‌ها، حائز اهمیت می‌باشد. از سوی دیگر در [۵۴]، یک الگوریتم یادگیری رتبه‌بندی تصاویر، ارائه شده است که قادر است بر اساس ادغام خصوصیات متن صفحات و نیز ویژگی‌های متناظر با تصاویر، اقدام به اولویت‌بندی

[1] Spectral clustering

تصاویر نماید. در این الگوریتم با استفاده از ویژگی‌های تصویری سطح پایین متناظر با تصاویر، گراف مشابهت بین تصاویر مختلف ایجاد می‌شود و بکمک روش یادگیری خروجی حاشیه‌ای ساخت‌یافته[1] [۵۵] و ترکیب آنها با ویژگی‌های متنی تصاویر، اقدام به اولویت‌بندی آنها می‌کند. ضمناً در سنجش عملکرد این روش از شاخص NDCG@n استفاده شده است. در بررسی صورت گرفته در [۵۶]، بیان شده است که به دلیل وجود تعداد کم پیوندهای بین پرس و جوها و صفحات وب در گراف دوبخشی بین این دو و نیز عدم تعادل بین تعداد صفحات(تریلیون) و تعداد پرس و جوهای کاربران (بیلیون) و همچنین در تعداد حاشیه نویسی‌ها، استفاده و افزایش مقیاس شبکه‌های عصبی مبتنی بر گرافهای پیچشی، دشوار است. لذا در این مقاله ابتدا دو زیرگراف تک بخشی Q و W با حفظ ارتباطات معرفی شده اند و سپس یک ساختار LTRGCN پیشنهاد شده است که یک کانال ارتباطی یادگیری رتبه‌بندی است. این ساختار از این دو زیرگراف برای نمونه‌برداری استفاده می‌کند و در نهایت امتیازات رتبه‌بندی را پیش بینی می‌کند. برای ارزیابی LtrGCN از دو مجموعه داده‌ی دنیای واقعی و آزمایش‌های برخط بر اساس آزمون A/B در یک موتور جستجوی مقیاس بزرگ استفاده شده است. نتایج برون‌خط نشان می‌دهد که LtrGCN می‌تواند شاخص NDCG را بین ۲,۸۹ تا ۳,۹۷ درصد نسبت به روشهای پایه افزایش دهد. الگوریتم LtrGCN با ترافیک واقعی در یک موتور جستجوی مقیاس بزرگ نیز استفاده شده است و پیشرفتهای چشمگیری را در حالت برخط نیز نشان می‌هد. همچنین در [۵۷]، با استفاده از یک روش مبتنی بر نظریه گراف، بازنمایی دادگان مختلف یادگیری رتبه‌بندی و مقایسه تطبیقی آنها صورت گرفته است. در روش پیشنهادی، به ازای هر مجموعه داده یادگیری رتبه‌بندی، گرافی تحت عنوان گراف مشابهت ویژگی‌ها ایجاد می‌شود که گره‌های آن، ویژگی‌های ارائه شده در مجموعه داده مورد بررسی می‌باشد و وزن هر یال، متناظر با میزان شاخص کندال به ازای زوج ویژگی‌های دو سر آن یال می‌باشد. از گراف حاصل و نیز مولفه غول‌پیکر آن، مجموعه‌ای از خصوصیات ساختاری و نیز تعدادی از خصوصیات متناظر با ویژگی‌ها استخراج می‌شود. این خصوصیات، امکان مقایسه دادگان محک در حوزه یادگیری رتبه‌بندی را فراهم می‌سازد. در [۵۸] نیز الگوریتمی برای رتبه‌بندی گره‌ها در گراف متناظر با راه‌های مواصلاتی ارائه شده است. این الگوریتم، شامل یک مولفه جاسازی است که از شبکه عمیق رمزگذار برای یادگیری بازنمایی نهفته به ازای هر قطعه از جاده استفاده می‌کند. از سوی دیگر با بکارگیری روش ادغام گراف‌های چندگانه، امکان نگاشت بخش‌های مختلف شبکه راه‌ها را فراهم می‌آورد. بدین ترتیب، یک روش یادگیری رتبه‌بندی به منظور اولویت‌بندی بخش مختلف شبکه مواصلاتی بر اساس مشخصات مختلف آنها نظیر ویژگی‌های ترافیکی، تعداد باندهای جاده و نیز متوسط سرعت در هر بخش، بدست می‌آید.

در جدول ۳، خلاصه مقالات ارزیابی شده ذیل دسته ج (بازنمایی گرافی از مسائل حوزه یادگیری رتبه‌بندی)، آمده است.

به لحاظ آماری، تعداد مقالات مرتبط بررسی شده در دسته سوم در شکل ۴ آمده است. بر اساس داده‌های این نمودار، رویکرد استفاده از نظریه گراف در یادگیری رتبه‌بندی، رویکرد نوینی است و عمده تحقیقات در این حوزه، مربوط به چند سال اخیر است.



شکل ۴. فراوانی تعداد مقالات مرتبط بررسی شده در دسته سوم (بکارگیری نظریه گراف در یادگیری رتبه‌بندی)

۷- جمع‌بندی و نتیجه‌گیری

بازیابی اطلاعات که با هدف تهیه اطلاعات مورد نیاز کاربر، انجام می‌شود، فرآیندی دشوار می‌باشد که همواره نیازمند ارتقا و بهبود می‌باشد. از سوی دیگر غنای نظریه گراف و کاربردهای گسترده آن در حل مسائل روزمره، پژوهشگران را به استفاده از این نظریه به منظور طراحی الگوریتم‌های کارآمد بازیابی اطلاعات ترغیب نموده است. با عنایت به موفقیت قابل توجه این الگوریتم‌ها، در این پژوهش، به بررسی کاربرد نظریه گراف در فرآیند بازیابی اطلاعات پرداخته شد و در این خصوص، پنجاه و دو مورد از تحقیقات مرتبط، مورد مطالعه و دسته‌بندی قرار گرفت.

[1] Margin structured output learning

جدول ۳. خلاصه مقالات ارزیابی شده در دسته ج (بازنمایی گرافی از مسائل حوزه یادگیری رتبه‌بندی)

مقاله	دسته‌بندی روش	سال چاپ	ایده اصلی	نقاط قوت	نقاط ضعف
[48]	ج-۱	۲۰۱۹	بازنمایی پرس‌وجوها و فهرست رتبه‌بندی دریافتی از هر رتبه‌بندی کننده گراف ادغام	امکان استفاده برای رتبه‌بندی گونه‌های مختلف داده‌های متنی و چند رسانه‌ای	هزینه محاسباتی زیاد
[49]	ج-۱	۲۰۲۲	استخراج ویژگی‌های نهفته به ازای پیشنهاد دهنده‌های پایه و استفاده از روش فرا ابتکاری تکامل تفاضلی	عملکرد مطلوب در بازیابی اطلاعات	عملکرد ضعیف در صورت ضعف پیشنهاد دهنده‌های پایه
[50]	ج-۱	۲۰۲۱	ایجاد گرافی جهت دار و وزن‌دار موسوم به گراف رقابت از روی رتبه‌بندی‌های پایه	حذف تاثیر نویز موجود در لیست‌های رتبه‌بندی پایه	هزینه محاسباتی زیاد
[51]	ج-۱	۲۰۲۰	ارائه روشی برای ترکیب رتبه‌بندی در حوزه بازیابی تصاویر که در آن، با تشکیل گراف مکمل بین ویژگی‌های مختلف تصاویر، ویژگی‌های موثر در فرآیند رتبه‌بندی، انتخاب و ترکیب می‌شوند.	کارآمدی و موثر بودن روش ارائه شده بواسطه استفاده از تعداد محدود ویژگی‌ها بواسطه انتخاب ویژگی‌های مکمل	عدم استفاده از عملگرهای مطرح در حوزه ترکیب اطلاعات به منظور ادغام ویژگی‌های منتخب
[52]	ج-۲	۲۰۲۲	ایجاد گراف مشابهت ویژگی‌های دادگان یادگیری رتبه‌بندی و استفاده از زیرمجموعه محدودی از ویژگی‌های پایه	هزینه محاسباتی اندک	بررسی محدود عملکرد روش پیشنهادی
[53]	ج-۲	۲۰۲۰	ایجاد گراف مشابهت بین ویژگی‌ها و انتخاب ویژگی‌ها با حداقل افزونگی	استفاده صرف از ویژگی‌های مهم در فرآیند یادگیری رتبه‌بندی	بررسی محدود عملکرد روش پیشنهادی
[54]	ج-۲	۲۰۰۹	یادگیری رتبه‌بندی تصاویر بر اساس ادغام خصوصیات متن و نیز ویژگی‌های تصاویر	استفاده از ویژگی‌های تصویری سطح پایین تصاویر در تولید مدل	نیاز به حجم زیاد دادگان آموزشی
[56]	ج-۲	2023	استفاده از شبکه گراف پیچشی به منظور ارائه یک روش یادگیری رتبه‌بندی	امکان بکارگیری در شرایط تعداد کم بودن پیوندهای بین پرس و جوها و صفحات وب در گراف دوبخشی بین این دو و نیز عدم تعادل بین تعداد صفحات و تعداد پرس و جوهای کاربران و همچنین در تعداد حاشیه نویسی‌ها	بهبود محدود نسبت به روش‌های پایه علیرغم هزینه محاسباتی بالا
[57]	ج-۲	2024	ارائه یک روش مبتنی بر نظریه گراف برای بازنمایی دادگان مختلف یادگیری رتبه‌بندی و مقایسه تطبیقی آنها	امکان مقایسه دادگان با مختصات و ویژگی‌های متفاوت	عدم ترکیب این روش بازنمایی دادگان با روش‌های آماری کلاسیک
[58]	ج-۲	2024	اولویت‌بندی قطعات شبکه راه‌ها با استفاده از خصوصیات مختلف مسیرها با بکارگیری روش ادغام گراف‌های چندگانه	در نظر گرفتن مشخصات مختلف مسیرهای مواصلاتی نظیر ویژگی‌های ترافیکی، تعداد باندهای جاده و نیز متوسط سرعت در هر بخش از جاده	نیاز به حجم زیاد داده آموزش

بر این اساس، مشخص شد، این تحقیقات، در سه دسته کلی، قابل تفکیک می‌باشند. دسته نخست، شامل تحقیقاتی است که در آنها از بازنمایی گرافی از دادگان در فرآیند بازیابی اطلاعات، استفاده شده است. در مقابل، دسته دوم شامل کاربرد نظریه گراف در بازیابی معنایی اطلاعات است و نهایتاً دسته سوم مربوط به بکارگیری نظریه گراف در فرآیند یادگیری رتبه‌بندی است. از سوی دیگر، در یک بررسی دقیق‌تر مشخص شد که تحقیقات دسته نخست، در دو زیر دسته، قابل گروه‌بندی است: در زیر دسته اول، از بازنمایی گرافی دادگان جهت استخراج ویژگی‌های مبتنی بر گراف و ترکیب این ویژگی‌ها با ویژگی‌های پایه دادگان، استفاده شده است؛ در حالی که زیر دسته دوم شامل روش‌هایی است که در آنها ویژگی‌های آماری دادگان استخراج می‌شود و سپس، یک بازنمایی گرافی از این ویژگی‌ها، به منظور ارتقای فرآیند بازیابی اطلاعات، ارائه می‌شود. در مقابل، دسته دوم، شامل پژوهش‌هایی است که در آنها از نظریه گراف به منظور بازیابی معنایی اطلاعات، استفاده شده است. در این خصوص، پژوهش‌های بررسی شده، از چهار رویکرد متفاوت، شامل: استفاده از گراف دانش، ایجاد گراف موجودیت، بکارگیری مدل‌های مبتنی بر هایپر گراف و نیز مدل پویش تصادفی، استفاده کرده‌اند. نهایتاً دسته سوم مربوط به بهره‌گیری از نظریه گراف در فرآیند یادگیری رتبه‌بندی است که خود در دو زیر دسته اصلی، یعنی ایجاد رتبه‌بندی و تجمیع رتبه‌بندی، تقسیم‌بندی می‌شوند.

از سوی دیگر، به لحاظ آماری، در این نوشتار، از دسته نخست، یعنی بکارگیری بازنمایی گرافی از دادگان در فرآیند بازیابی اطلاعات،

تعداد ۱۶ پژوهش شاخص، شناسایی و بررسی شد. در مقابل، از دسته دوم یعنی بکارگیری نظریه گراف در فرآیند بازیابی معنایی اطلاعات، تعداد ۲۵ تحقیق، شناسایی و ارزیابی شد و نهایتاً از دسته سوم تحقیقات مربوط به بکارگیری نظریه گراف در یادگیری رتبه‌بندی، تعداد ۷ پژوهش، مورد مطالعه و بررسی، قرار گرفت. بر این اساس، بنظر می‌رسد در دو دسته اول و دوم، تعداد نسبتا زیادی از تحقیقات، صورت گرفته است، در حالی که بکارگیری نظریه گراف در یادگیری رتبه‌بندی، کماکان می‌تواند یک موضوع تحقیقاتی جذاب در حوزه بازیابی اطلاعات باشد.

در عین حال، به عنوان پیشنهادهایی به منظور توسعه آتی هر دسته از روش‌های بررسی شده، می‌توان به موارد ذیل اشاره نمود. در دسته نخست که به بازیابی اطلاعات بر مبنای ارائه بازنمایی گرافی از دادگان بازیابی اطلاعات، پرداخته است، مواردی نظیر بکارگیری نظریه ترکیب اطلاعات و استفاده از توان محاسباتی عملگرهای ترکیب اطلاعات به منظور ادغام ویژگی‌های گرافی و ویژگی‌های پایه به طور خاص با توجه به نایقینی موجود در ویژگی‌های پایه و ویژگی‌های گرافی و بکارگیری نظریه استدلال شهودی و انتگرال‌های فازی، از جمله زمینه‌های توسعه روشهای بازیابی اطلاعات در این دسته به نظر می‌رسند. در خصوص دسته دوم کاربردهای گراف در بازیابی اطلاعات که مربوط به بازیابی معنایی اطلاعات با استفاده از نظریه گراف می‌باشد، با عنایت به تعدد و تنوع عناصر داده‌ای متناظر با اسناد و پرس و جوهای کاربران و نیز پیچیدگی‌های ارتباطات معنایی مابین آنها، بکارگیری مدل‌های یادگیری ژرف به منظور شناسایی الگوهای کارآمد در بازیابی اطلاعات یک ضرورت جدی، محسوب می‌شود. ضمنا وجود انواع موجودیت‌های داده‌ای و نیز روابط توام با درجاتی از نایقینی مابین آنها چه در متن اسناد و چه در واژگان بکار رفته در پرس و جوی کاربر، نیاز مبرم به بکارگیری نظریه‌های توانمند در مدلسازی و مواجه با نایقینی نظیر استدلال شهودی و انتگرال‌های فازی را بوجود آورده است. نهایتا در دسته سوم که مربوط به بهره‌گیری از نظریه گراف در یادگیری رتبه‌بندی است، استفاده از مدل‌های یادگیری ماشین قدرتمندی نظیر یادگیری ژرف امکان ارائه مدل‌های پیچیده‌تر و کامل‌تر یادگیری رتبه‌بندی را فراهم می‌آورد. در عین حال توجه به این نکته که سامانه‌های بازیابی اطلاعات سامانه‌های کاربر محوری هستند، بکارگیری نظریه گراف به منظور مدلسازی رفتار تعاملی کاربر با این سامانه‌ها بسیار مفید خواهد بود. در کنار همه این موارد بهره‌گیری از رویکرد ترکیب اطلاعات به منظور تلفیق کارآمد عناصر اطلاعاتی بسیار حائز اهمیت خواهد بود. ضمنا توجه به ارتباط معنایی اجزای اسناد و پرس و جوهای کاربر و نیز ویژگی‌های متناظر با آنها و بکارگیری مدل‌های بازیابی معنایی اطلاعات بخش مهمی از معضلات مدل‌های پایه فعلی مورد استفاده در سامانه‌های بازیابی اطلاعات را که عمدتا مبتنی بر مدل سبد کلمات هستند، تقلیل خواهد داد.

مراجع

[1] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi: 10.1017/cbo9780511809071.

[2] R. Baeza-Yates, *Modern Information Retrieval: The Concepts and Technology Behind Search*, Second Eit. Addison-Wesley Professional, 2011.

[3] J. Devezas and S. Nunes, "A Review of Graph-Based Models for Entity-Oriented Search," *SN Computer Science*, vol. 2, no. 6. Springer, pp. 1–36, Nov. 01, 2021. doi: 10.1007/s42979-021-00828-w.

[4] C. Moreira, P. Calado, and B. Martins, "Learning to rank academic experts in the DBLP dataset," *Expert Syst.*, vol. 32, no. 4, pp. 477–493, Aug. 2015, doi: 10.1111/exsy.12062.

[5] Y. Zhang, D. Wang, and Y. Zhang, "Neural IR meets graph embedding: A ranking model for product search," in *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, May 2019, pp. 2390–2400. doi: 10.1145/3308558.3313468.

[6] T. Xu, H. Zhu, E. Chen, B. Huai, H. Xiong, and J. Tian, "Learning to annotate via social interaction analytics," *Knowl. Inf. Syst.*, vol. 41, no. 2, pp. 251–276, Oct. 2014, doi: 10.1007/s10115-013-0717-8.

[7] W. Yu and Z. Qin, "Spectrum-enhanced pairwise learning to rank," in *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, May 2019, pp. 2247–2257. doi: 10.1145/3308558.3313478.

[8] W. Wu, H. Li, and J. Xu, "Learning query and document similarities from click-through bipartite graph with metadata," in *WSDM 2013 - Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, 2013, pp. 687–696. doi: 10.1145/2433396.2433481.

[9] X. He, M. Gao, M. Y. Kan, and D. Wang, "BiRank: Towards Ranking on Bipartite Graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 57–71, Jan. 2017, doi: 10.1109/TKDE.2016.2611584.

[10] J. Sanz-Cruzado, P. Castells, C. Macdonald, and I. Ounis, "Effective contact recommendation in social networks by adaptation of information retrieval models," *Inf. Process. Manag.*, vol. 57, no. 5, p. 102285, Sep. 2020, doi: 10.1016/j.ipm.2020.102285.

[11] K. Yuan, L. Gao, Z. Jiang, and Z. Tang, "Formula Ranking within an Article," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, May 2018, pp. 123–126. doi: 10.1145/3197026.3197061.

[12] H. Niu, I. Keivanloo, and Y. Zou, "Learning to rank code examples for code search engines," *Empir. Softw. Eng.*, vol. 22, no. 1, pp. 259–291, Feb. 2017, doi: 10.1007/s10664-015-9421-5.

[13] S. Jiang *et al.*, "Learning query and document relevance from a web-scale click graph," in *SIGIR 2016 - Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2016, pp. 185–194. doi: 10.1145/2911451.2911531.

[14] X. Yang and B. Wang, "Local ranking and global fusion for personalized recommendation," *Appl. Soft Comput. J.*, vol. 96, p. 106636, Nov. 2020, doi: 10.1016/j.asoc.2020.106636.

[15] A. Ferraro, L. Porcaro, and X. Serra, "Balancing Exposure and Relevance in Academic Search," 2020. Accessed: Nov. 14, 2022.

Sciences, Social-Informatics and Telecommunications Engineering, *LNICST*, 2020, vol. 331, pp. 19–30. doi: 10.1007/978-3-030-62205-3_2.

[33] G. Maheshwari, P. Trivedi, D. Lukovnikov, N. Chakraborty, A. Fischer, and J. Lehmann, "Learning to Rank Query Graphs for Complex Question Answering over Knowledge Graphs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11778 LNCS, pp. 487–504. doi: 10.1007/978-3-030-30793-6_28.

[34] Y. Su *et al.*, "Reducing Bug Triaging Confusion by Learning from Mistakes with a Bug Tossing Knowledge Graph," in *Proceedings - 2021 36th IEEE/ACM International Conference on Automated Software Engineering, ASE 2021*, 2021, pp. 191–202. doi: 10.1109/ASE51524.2021.9678574.

[35] P. Jafarzadeh, Z. Amirmahani, and F. Ensan, "Learning to Rank Knowledge Subgraph Nodes for Entity Retrieval," in *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2022, pp. 2519–2523. doi: 10.1145/3477495.3531888.

[36] Y. Ni *et al.*, "Semantic documents relatedness using concept graph representation," in *WSDM 2016 - Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, Feb. 2016, pp. 635–644. doi: 10.1145/2835776.2835801.

[37] N. Zhiltsov and E. Agichtein, "Improving entity search over linked data by modeling latent semantics," in *International Conference on Information and Knowledge Management, Proceedings*, 2013, pp. 1253–1256. doi: 10.1145/2505515.2507868.

[38] O. Irrera and G. Silvello, "Background Linking: Joining Entity Linking with Learning to Rank Models.," *ceur-ws.org*, vol. 2816, 2021, Accessed: Aug. 24, 2022. [Online]. Available: http://ceur-ws.org/Vol-2816/paper6.pdf

[39] M. Bendersky and W. B. Croft, "Modeling higher-order term dependencies in information retrieval using query hypergraphs," in *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2012, pp. 941–950. doi: 10.1145/2348283.2348408.

[40] T. Menezes and C. Roth, "Semantic Hypergraphs," Aug. 2019, Accessed: Apr. 07, 2023. [Online]. Available: http://arxiv.org/abs/1908.10784

[41] L. Dietz, "ENT rank: Retrieving entities for topical information needs through entity-neighbor-text relations," in *SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2019, pp. 215–224. doi: 10.1145/3331184.3331257.

[42] B. A., "ObjectRank : Authority-Based Keyword Search in Databases," *VLDB 2004*, 2004.

[43] S. Chakrabarti, "Dynamic personalized pagerank in entity-relation graphs," in *16th International World Wide Web Conference, WWW2007*, May 2007, pp. 571–580. doi: 10.1145/1242572.1242650.

[44] L. Espín-Noboa, F. Lemmerich, S. Walk, M. Strohmaier, and M. Musen, "HopRank: How semantic structure influences teleportation in PageRank (a case study on bioPortal)," in *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, May 2019, pp. 2708–2714. doi: 10.1145/3308558.3313487.

[45] R. Delbru, N. Toupikov, M. Catasta, G. Tummarello, and S. Decker, "Hierarchical link analysis for ranking web data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6089 LNCS, no. PART 2, pp. 225–239. doi: 10.1007/978-3-642-13489-0_16.

[46] C. Musto, G. Semeraro, M. de Gemmis, and P. Lops, "Tuning personalized pagerank for semantics-aware recommendations based on linked open data," in *Lecture Notes in Computer Science* [Online]. Available: https://fair-trec.github.io/2020/doc/guidelines-2020.pdf

[16] C. J. C. Burges, "From ranknet to lambdarank to lambdamart: An overview," 2010. Accessed: Nov. 14, 2022. [Online]. Available: http://www.ccs.neu.edu/home/vip/teach/MLcourse/4_boosting/materials/msr-tr-2010-82.pdf

[17] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, Oct. 2008, doi: 10.1088/1742-5468/2008/10/P10008.

[18] A. Bertolino, A. Guerriero, B. Miranda, R. Pietrantuono, and S. Russo, "Learning-to-rank vs ranking-to-learn: Strategies for regression testing in continuous integration," in *Proceedings - International Conference on Software Engineering*, Jun. 2020, pp. 1261–1272. doi: 10.1145/3377811.3380369.

[19] S. Maqsood, M. A. Islam, M. T. Afzal, and N. Masood, "A comprehensive author ranking evaluation of network and bibliographic indices," *Malaysian J. Libr. Inf. Sci.*, vol. 25, no. 1, pp. 31–45, Apr. 2020, doi: 10.22452/mjlis.vol25no1.2.

[20] J. Shi and X.-Y. Tian, "Learning to Rank Sports Teams on a Graph," *Appl. Sci.*, vol. 10, no. 17, p. 5833, Aug. 2020, doi: 10.3390/app10175833.

[21] S. Agarwal, "Learning to rank on graphs," *Mach. Learn.*, vol. 81, no. 3, pp. 333–357, Dec. 2010, doi: 10.1007/s10994-010-5185-8.

[22] X. Zou, "A Survey on Application of Knowledge Graph," in *Journal of Physics: Conference Series*, Apr. 2020, vol. 1487, no. 1, p. 012016. doi: 10.1088/1742-6596/1487/1/012016.

[23] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "A Survey on Knowledge Graphs: Representation, Acquisition, and Applications," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 33, no. 2, pp. 494–514, Feb. 2022, doi: 10.1109/TNNLS.2021.3070843.

[24] M. Fernandez *et al.*, "Semantic search meets the Web," in *Proceedings - IEEE International Conference on Semantic Computing 2008, ICSC 2008*, 2008, pp. 253–260. doi: 10.1109/ICSC.2008.52.

[25] V. Lopez, M. Sabou, and E. Motta, "PowerMap; Mapping the real semantic web on the fly," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2006, vol. 4273 LNCS, pp. 414–427. doi: 10.1007/11926078_30.

[26] K. Byrne, "Populating the Semantic Web-Combining Text and Relational Databases as RDF Graphs," The University of Edinburgh, 2008.

[27] K. Balog *et al.*, "SaHaRa: Discovering Entity-Topic Associations in Online News," 2009.

[28] P. Oza, L. D.-C. workshop Proceedings, and U. 2021, "Which entities are relevant for the story?," in *Text2Story'21 Workshop*, 2021, pp. 41–48. Accessed: Apr. 03, 2023. [Online]. Available: https://par.nsf.gov/biblio/10300275

[29] U. Rashid, K. Saleem, and A. Ahmed, "MIRRE approach: nonlinear and multimodal exploration of MIR aggregated search results," *Multimed. Tools Appl.*, vol. 80, no. 13, pp. 20217–20253, May 2021, doi: 10.1007/s11042-021-10603-x.

[30] H. Gao, L. Wu, P. Hu, Z. Wei, F. Xu, and B. Long, "Graph-augmented Learning to Rank for Querying Large-scale Knowledge Graph," Nov. 2021, Accessed: Mar. 17, 2023. [Online]. Available: http://arxiv.org/abs/2111.10541

[31] H. Hosseini and E. Bagheri, "Learning to rank implicit entities on Twitter," *Inf. Process. Manag.*, vol. 58, no. 3, p. 102503, May 2021, doi: 10.1016/j.ipm.2021.102503.

[32] H. Wu and F. J. Meng, "Research on the Application of Personalized Course Recommendation of Learn to Rank Based on Knowledge Graph," in *Lecture Notes of the Institute for Computer*

[53] J. Y. Yeh and C. J. Tsai, “Graph-based Feature Selection Method for Learning to Rank,” in *ACM International Conference Proceeding Series*, Nov. 2020, pp. 70–73. doi: 10.1145/3442555.3442567.

[54] B. Geng, L. Yang, and X.-S. Hua, “Learning to Rank with Graph Consistency,” 2009. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2009/08/MSRA-TR-CAR.pdf

[55] J. Fan, H. Luo, Y. Gao, and R. Jain, “Incorporating concept ontology for hierarchical video classification, annotation, and visualization,” *IEEE Trans. Multimed.*, vol. 9, no. 5, pp. 939–957, Aug. 2007, doi: 10.1109/TMM.2007.900143.

[56] Y. Li et al., “LtrGCN: Large-Scale Graph Convolutional Networks-Based Learning to Rank for Web Search,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 14174 LNAI, pp. 635–651, 2023, doi: 10.1007/978-3-031-43427-3_38.

[57] A. H. Keyhanipour, “Graph-based comparative analysis of learning to rank datasets,” Int. J. Data Sci. Anal., vol. 17, no. 2, pp. 165–187, Mar. 2024, doi: 10.1007/S41060-023-00406-8/METRICS.

[58] M. Xu and J. Zhang, “MGL2Rank: Learning to rank the importance of nodes in road networks based on multi-graph fusion,” Inf. Sci. (Ny)., vol. 667, p. 120472, May 2024, doi: 10.1016/J.INS.2024.120472.

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, vol. 10249 LNCS, pp. 169–183. doi: 10.1007/978-3-319-58068-5_11.

[47] H. Li, “Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition,” *Synth. Lect. Hum. Lang. Technol.*, vol. 7, no. 3, pp. 1–123, Oct. 2015, doi: 10.2200/S00607ED2V01Y201410HLT026.

[48] I. C. Dourado, D. C. G. Pedronette, and R. da S. Torres, “Unsupervised graph-based rank aggregation for improved retrieval,” *Inf. Process. Manag.*, vol. 56, no. 4, pp. 1260–1279, Jul. 2019, doi: 10.1016/j.ipm.2019.03.008.

[49] M. Bałchanowski and U. Boryczka, “Aggregation of Rankings Using Metaheuristics in Recommendation Systems,” *Electronics*, vol. 11, no. 3, p. 369, Jan. 2022, doi: 10.3390/electronics11030369.

[50] Y. Zhang, Y. Xiao, J. Wu, and X. Lu, “Comprehensive world university ranking based on ranking aggregation,” *Comput. Stat.*, vol. 36, no. 2, pp. 1139–1152, Jun. 2021, doi: 10.1007/s00180-020-01033-8.

[51] L. P. Valem and D. C. G. Pedronette, “Graph-based selective rank fusion for unsupervised image retrieval,” *Pattern Recognit. Lett.*, vol. 135, pp. 82–89, Jul. 2020, doi: 10.1016/j.patrec.2020.03.032.

[52] J. Y. Yeh and C. J. Tsai, “A Graph-based Feature Selection Method for Learning to Rank Using Spectral Clustering for Redundancy Minimization and Biased PageRank for Relevance Analysis,” *Comput. Sci. Inf. Syst.*, vol. 19, no. 1, pp. 141–164, Jan. 2022, doi: 10.2298/CSIS201220042Y.