Proposing a Method Based on Analysis of Bibliographic References **to Identify Related Scientific Articles** in Digital Libraries

Niloofar Mozafari
PhD in Artificial Intelligence; Assistant Professor; Regional Information Center for Science and Technology; Shiraz, Iran;
Email: mozafari@ricest.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 39 | No. 1 | pp. 333-356
Autumn 2023
https://doi.org/jipm.39.1

Received: 20, Jun. 2022 **Accepted: 14, Mar. 2023**

Abstract: The volume of scientific documents and articles has increased dramatically in the last decade. It makes too difficult to find the relevant documents based on the user's query. Information retrieval systems help researchers to find relevant scientific articles. One of the capabilities that help researchers to find the relevant papers is the feature of finding related scientific articles to an article. In other words, this feature allows the researcher to view other related articles by selecting one article. The purpose of this research is to present a method based on the analysis of bibliographic references to identify related scientific articles in digital libraries. The statistical population of this research is the articles published in the last 5 years in Persian and English publications indexed in the ISC in the field of computer science. The proposed method is able to find the articles that are most similar to the given article by analyzing the references of the articles and ranking them based on their similarity. To do that, after extracting the title and obtaining the similarity between the existing references among the articles, those articles that have the most similarity are identified and sorted based on their similarity. The proposed method has been compared with other methods, and the obtained results on both Persian and English data are promising.

Keywords: Bibliographic References, Information Retrieval, Similarity Measure, Precision

ارائه روشی مبتنی‌بر تحلیل مراجع کتابشناختی برای شناسایی مقالات علمی مرتبط در کتابخانه‌های دیجیتال

نیلوفر مظفری

دکتری هوش مصنوعی؛ استادیار؛
مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری؛
شیراز، ایران mozafari@ricest.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۱/۱۱/۰۹ **پذیرش:** ۱۴۰۲/۰۳/۰۹ مقاله برای اصلاح به مدت ۲۸ روز نزد پدیدآور بوده است.

چکیده: حجم مستندات و مقالات علمی در دهه اخیر افزایش بسیار زیادی یافته به‌گونه‌ای که شناسایی مقالات مرتبط در کتابخانه‌های دیجیتال را با چالش‌هایی روبه‌رو کرده است. هدف این پژوهش، ارائه روشی مبتنی بر تحلیل مراجع کتابشناختی برای شناسایی مقالات علمی مرتبط در کتابخانه‌های دیجیتال است. جامعه آماری این پژوهش مقالات منتشرشده در پنج سال اخیر در نشریات فارسی و انگلیسی نمایه‌شده در «پایگاه استنادی علوم جهان اسلام» در حوزه علوم کامپیوتر است. روش پیشنهادی قادر است با تحلیل مراجع کتابشناختی مقالات، مقالات مرتبط با یکدیگر را در کتابخانه‌های دیجیتال پیدا کرده و آن‌ها را بر اساس میزان شباهتشان مرتب نماید. در این راستا بعد از استخراج عنوان و به‌دست آوردن شباهت میان آن‌ها، آن دسته از مقالاتی که بیشترین شباهت را با یکدیگر دارند، شناسایی شده و بر اساس میزان شباهتشان مرتب می‌شوند. به‌منظور مقایسه، روش پیشنهادی با روش‌های دیگر مقایسه شد. نتایج به‌دست‌آمده روی داده‌های فارسی و انگلیسی نشان‌دهنده کارایی روش پیشنهادی در شناسایی مقالات مرتبط است.

کلیدواژه‌ها: تحلیل مراجع کتابشناختی، بازیابی اطلاعات، معیار شباهت، دقت

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در SCOPUS ،ISC ،LISTA و
jipm.irandoc.ac.ir
دورهٔ ۳۹ | شمارهٔ ۱ | ص‌ص ۳۳۳-۳۵۶
پاییز ۱۴۰۲
https://doi.org/jipm.39.1

۱. مقدمه

در سال‌های اخیر با ظهور وب و اینترنت به‌منزله یک سیستم اطلاع‌رسانی جهانی، توانایی بشر برای تولید و جمع‌آوری داده‌ها افزایش چشمگیری داشته است؛ به‌صورتی که بشر با حجم بسیار زیادی از داده و اطلاعات روبه‌رو شده است. از آنجا که درصد بسیار زیادی از این اطلاعات را مستندات و منابع متنی تشکیل می‌دهند، کشف دانش از این حجم نیاز به درک و شناسایی روابط میان آن‌هاست. از سوی دیگر، در عصر فناوری که در آن قرار داریم، اعتقاد بر این است که هر چیزی باید به‌صورت خودکار انجام گیرد. متن‌کاوی و یا کشف دانش از میان مستندات متنی، تلاشی برای نیل به این هدف است. اصلی‌ترین دلیلی که باعث شد متن‌کاوی در کانون توجهات قرار بگیرد، مسئله در دسترس بودن حجم وسیعی از داده‌های متنی و نیاز شدید به استخراج اطلاعات و دانش از این داده‌های متنی است (Feldman & Sanger 2007).

متن‌کاوی به معنای استخراج اطلاعات گران‌بها از حجم عظیم داده‌های متنی است. با توجه به نوع داده و همچنین حجم آن، مشخص نیست که چه اطلاعات گران‌بهایی در عمق این داده‌های متنی وجود دارد و تنها با کاوش در این داده‌هاست که می‌توان به این اطلاعات گران‌قدر دسترسی پیدا کرد. بنابراین، وظیفه اصلی متن‌کاوی، کاویدن و استخراج دانش از منابع عظیم داده متنی است تا اطلاعات گران‌بهایی که در حجم انبوهی از اطلاعات سطحی پنهان شده است، آشکار گردد. متن‌کاوی تلاش برای استخراج دانش از انبوه داده‌های متنی موجود است که به کمک مجموعه‌ای از روش‌های آماری و مدل‌سازی می‌تواند الگوها و روابط پنهان موجود در داده‌های متنی را تشخیص دهد. تحلیل متن[1] اصطلاحی است که گاهی به‌جای متن‌کاوی استفاده می‌شود که آن هم به فرایند تبدیل داده‌های متنی غیرساخت‌یافته به اطلاعات با معنا اطلاق می‌شود. برای تحلیل متن و یا به‌عبارتی متن‌کاوی، نیازمند الگوریتم‌های یادگیری ماشین هستیم (Hotho, Nürnberger & Paaß 2005).

یکی از کاربردهای بسیار مفید متن‌کاوی در موتور جست‌وجوست؛ جایی که کاربران نیاز به یافتن اطلاعات مرتبط به پرس‌وجو[2] دارند. موتور جست‌وجو در اصل هر برنامه کامپیوتری است که برای یافتن اطلاعات مورد نظر کاربر نوشته می‌شود و می‌تواند

1. text analysis
2. query

در هر حوزه‌ای مورد استفاده قرار گیرد. یک موتور جست‌وجو در واقع، یک سیستم پاسخ‌دهی است که اساساً از دو بخش اصلی تشکیل شده است: پایگاه داده اطلاعات و هسته موتور. اولین موتور جست‌وجو، «آرچی»[1] نام داشت که برای جست‌وجو میان عناوین مورد استفاده قرار می‌گرفت و توانایی نمایش محتوای وب را نداشت. موتورهای جست‌وجوی «ورونیکا»[2] و «جاگهد»[3] به‌دنبال پروژه «آرچی» با هدف نمایه[4]کردن متن ساده به‌وجود آمدند. به‌دنبال این موتورهای جست‌وجو، موتورهای جست‌وجوی دیگری به‌وجود آمدند که هر کدام، برای بهبود موتورهای قبلی تلاش می‌کردند تا اینکه حدود سال ۱۹۹۸ دامنه google.com ثبت گردید و از آن پس، «گوگل» به‌عنوان قوی‌ترین و پراستفاده‌ترین موتور جست‌وجو در تمامی پلتفرم‌ها معرفی شد و توانست محبوبیت بسیار زیادی میان کاربران کسب کند.

هر موتور جست‌وجو برای کشف، دسته‌بندی و رتبه‌بندی اسنادی که در اختیار دارد، نیازمند انجام سه فرایند کلی است که تحت عناوین خزیدن[5]، نمایه‌کردن[6] و رتبه‌بندی کردن[7] شناخته می‌شوند. با استفاده از فرایند خزیدن که به‌طور کلی توسط ربات‌های با عنوان خزنده[8] یا عنکبوت انجام می‌گیرد، داده‌های بسیار زیادی به پایگاه داده وارد می‌شوند. قبل از ذخیره داده‌ها در پایگاه داده، عملیات نمایه‌گذاری انجام می‌شود و در واقع، نحوه ذخیره‌کردن داده‌ها در پایگاه داده توسط این فرایند انجام می‌گیرد. لازم به ذکر است که هر موتور جست‌وجو، پایگاه داده مخصوص به خود را دارد و از یک فرایند نمایه‌گذاری با توجه به فیلدهای پایگاه داده بهره می‌برد. هنگام جست‌وجوی کاربر، موتور جست‌وجو با استفاده از الگوریتم‌های رتبه‌بندی، اسناد مرتبط با پرسش کاربر را یافته و آن‌ها را به‌صورت مرتب‌شده نمایش می‌دهد.

یکی از قابلیت‌هایی که در موتورهای جست‌وجو منجر به افزایش کاربرپسندی[9] و راحتی کاربر در استفاده از آن‌ها می‌شود، یافتن مقالات مرتبط با یک مقاله است. زمانی که کاربر یک عبارت را جست‌وجو می‌کند، موتور جست‌وجو مقالات مرتبط با آن را یافته و به وی نشان می‌دهد. از میان سیاهه‌ای که به کاربر نشان داده می‌شود، ممکن است کاربر یکی از آن مقالات را مرتبط با پرسش یافته و قصد دارد مقالات بیشتری که مرتبط با آن مقاله است، بیابد.

---

1. Archie
2. Veronica
3. Jughead
4. indext
5. crawling
6. indexing
7. ranking
8. crawler
9. user-friendly

از این رو، پژوهش حاضر تلاش دارد با ارائه روشی مبتنی بر تحلیل مراجع کتابشناختی، مقالات مرتبط با یک مقاله را بیابد؛ بدین صورت که وقتی موتور جست‌وجو مقاله‌ای را پیدا کرد، الگوریتم ارائه‌شده بتواند مقالات مرتبط با آن مقاله را پیدا کند.

در بخش بعدی مروری بر پیشینه پژوهش خواهیم داشت. سپس روش پیشنهادی برای شناسایی مقالات مرتبط ارائه خواهد شد. بخش ۴ به مرور یافته‌ها می‌پردازد و نتیجه‌گیری مقاله در بخش ۵ خواهد بود.

۲. پیشینه پژوهش

معیارهای شباهت متن، نقش بسیار مهمی در پژوهش‌های مرتبط با متن و یافتن مقالات مرتبط با یک مقاله ایفا می‌کنند و کاربردهای بسیار زیادی در زمینه‌های دیگر از جمله بازیابی متن[۱]، طبقه‌بندی متن[۲]، خوشه‌یابی اسناد[۳]، تشخیص موضوع[۴]، دنبال کردن موضوع[۵]، تولید سؤال[۶]، پاسخ به سؤال[۷]، نمره‌دهی به متون[۸]، ماشین ترجمه[۹]، خلاصه‌سازی متن[۱۰] و غیره دارند.

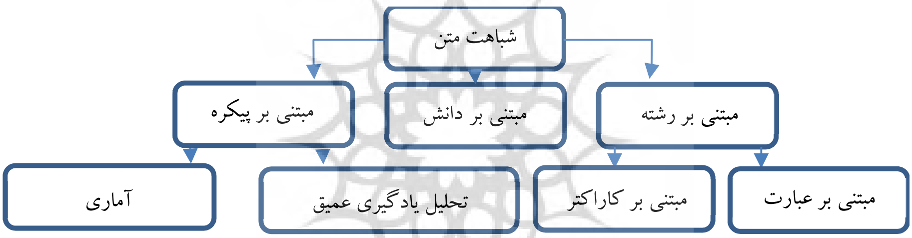
پیدا کردن شباهت میان کلمات یک بخش اساسی در یافتن شباهت متن است. در سطوح بعدی برای پیدا کردن شباهت میان جملات، پاراگراف‌ها و اسناد اهمیت دارند. کلمات می‌توانند به دو صورت لغوی[۱۱] و معنایی[۱۲] به یکدیگر شبیه باشند. کلمات به‌صورت لغوی با یکدیگر شبیه هستند در صورتی که دنباله کاراکترهای مشابهی داشته باشند. کلمات به‌صورت معنایی شبیه هستند اگر آن‌ها دارای یک مفهوم مشابه باشند؛ برای مثال، در یک بافت استفاده شوند و یا مثلاً یکی متضاد دیگری باشد.

تاکنون معیارهای مختلفی ارائه شده‌اند که به سه دسته کلی معیارهای مبتنی بر رشته[۱۳]، معیارهای مبتنی بر پیکره[۱۴] و معیارهای مبتنی بر دانش تقسیم می‌شوند (Gomaa & Fahmy 2013). شکل ۱، این دسته‌بندی را نشان می‌دهد (Farouk 2019). همان‌طور که در این شکل نشان داده شده، سه دسته اصلی برای به‌دست آوردن شباهت میان کلمات تاکنون ارائه شده‌اند. دسته اول، مبتنی بر پیکره هستند. به گفته دیگر، این گروه برای

1. information retrieval
2. text classification
3. document clustering
4. topic detection
5. topic tracking
6. question generation
7. question answering
8. essay scoring
9. machine translation
10. text summarization
11. lexically
12. semantically
13. string-based measures
14. corpus-based similarity

به‌دست آوردن شباهت میان کلمات از تحلیل کل پیکره استفاده می‌کنند و انتظار دارند که کلماتی که در کل پیکره به تعداد زیاد در کنار هم دیده می‌شوند، از نظر معنایی به یکدیگر شبیه باشند. روش‌های مبتنی بر یادگیری عمیق و روش‌های مبتنی بر تحلیل‌های آماری، دو دسته روش در این گروه هستند.

گروه دوم از روش‌های شباهت متنی، روش‌های مبتنی بر دانش هستند که عمدتاً از گراف ساخته‌شده توسط انسان برای به‌دست آوردن شباهت میان کلمات استفاده می‌کنند. در این شبکه عمدتاً ارتباط معنایی و روابط میان کلمات مشخص شده است. سرانجام، گروه سوم روش‌های مبتنی بر رشته هستند. این گروه از روش‌ها بدین صورت عمل می‌کنند که هر رشته را به‌صورت دنباله‌ای از کاراکترها در نظر گرفته و از شباهت میان کاراکترهای تشکیل‌دهنده برای به‌دست آوردن شباهت میان کلمات بهره می‌برند. در ادامه، هر کدام از این سه گروه اصلی توضیح داده می‌شود (Farouk 2019).



شکل ۱. روش‌های به‌دست آوردن شباهت متنی (Farouk 2019)

معیارهای مبتنی بر رشته به دو دسته معیارهای مبتنی بر کاراکتر[1] و مبتنی بر عبارت[2] تقسیم می‌شوند. یک معیار شباهت مبتنی بر کاراکتر، معیاری است که شباهت و یا عدم شباهت (فاصله) میان دو رشته متن را اندازه‌گیری می‌کند. الگوریتم طولانی‌ترین زیررشته مشترک[3] (یا به اختصار LCS)، شباهت میان دو رشته را بر اساس طول زنجیرهٔ پیوستهٔ کاراکترها که در هر دو رشته وجود دارد، در نظر می‌گیرد. الگوریتم «لوناشتاین-دامرا»[4] فاصله میان دو رشته را با شمردن کمترین تعداد عملیاتی که نیاز است یک رشته به رشته دیگری تبدیل شود، تعریف می‌کند. لازم به ذکر است که این عملیات به‌صورت اضافه، حذف و جایگذاری کاراکترهای رشته برای تبدیل یک رشته به دیگری است (Hall & Dowling1980; Peterson 1980).

1. character-based
2. term-based
3. longest common substring
4. Levenshtein-Damerau

معیارهـای شـباهت مبتنـی بـر عبـارت، شـباهت میـان دو رشـته را بـر اسـاس عبارت‌هـای تشـکیل‌دهنده آن به‌دسـت مـی‌آورد. بـه ایـن فاصلـه، فاصلـه بلاکـی[1]، فاصلـه منهتـن[2]، باکسـکار[3]، فاصلــه ارزش مطلــق[4]، فاصلــه L1[5]، فاصلــه بــلاک شــهر[6] هــم می‌گوینــد. ایــن فاصلــه بدیــن صـورت محاسـبه می‌شـود کـه مسـیر میـان دو عبـارت را به‌صـورت شـبکه‌ای در نظـر گرفتـه و تعـداد نقاطـی را کـه بـرای رسـیدن بـه مقصـد مـورد نیـاز اسـت، می‌شـمارد (Reynolds 1980). شـباهت کسینوسـی بـرای به‌دسـت آوردن شـباهت میـان دو رشـته کـه به‌صـورت یـک بـردار ذخیـره شـده‌اند، از کسـینوس زاویـه میـان آن‌هـا اسـتفاده می‌کنـد. ضریـب تـاس[7] به‌صـورت دو برابـر تعـداد عبارت‌هـای مشـترک در دو رشـته تقسـیم بـر تعـداد کل عبارت‌هـا در هـر دو رشـته محاسـبه می‌شـود (Dice 1945). معیـار جـاکارد[8] به‌صـورت تعـداد عبارت‌هـای مشـترک بـه کل تعـداد عبارت‌هـا در هـر دو رشـته محاسـبه می‌گـردد (Niwattanakul et al. 2013). فاصلـه اقلیدسـی[9] یـا فاصلـه L2 ریشـه مربـع مجمـوع فاصلـه میـان عناصـر دو بـردار تعریـف می‌شـود. از ایـن معیـار در پژوهش‌هـای دیگـر از جملـه (عباسـی و وزیـری ۱۳۹۴) و (سـلیمانی‌نژاد، سـلاجقه و طبیبی‌نیـا ۱۳۹۷) بـرای خوشـه‌بندی متـون اسـتفاده شـده اسـت. ضریـب تطابـق[10] یـک روش مبتنـی بـر بـردار بسـیار سـاده اسـت کـه به‌صـورت سـاده تعـداد عبارت‌هـای مشـابه غیرصفـر روی هـر دو بـردار را می‌شـمارد. ضریـب همپوشـانی[11] مشـابه ضریـب تـاس اسـت؛ امـا اگـر یکـی از رشـته‌ها زیرمجموعـه دیگـری باشـد، دو رشـته را به‌صـورت تطابـق کامـل در نظـر می‌گیـرد. از معیارهـای شـباهت کسینوسـی، جـاکارد و اقلیدسـی بـرای محاسـبه امتیـاز شـباهت اخبـار در زبان‌هـای انگلیسـی و هنـدی در Singh et al. (2021) اسـتفاده شـده اسـت.

معیـار شـباهت مبتنـی بـر پیکـره، یـک معیـار شـباهت معنایـی[12] اسـت کـه شـباهت میـان کلمـات را بـر اسـاس اطلاعـات به‌دسـت‌آمده از یـک پیکـره بـزرگ می‌یابـد. یـک پیکـره، یـک مجموعـه بـزرگ از متـون نوشـته شـده یـا صحبـت شـده اسـت کـه بـرای پژوهش‌هـای زبانـی اسـتفاده می‌شـود (Gomaa & Fahmy 2013).

تحلیـل معنایـی نهفتـه[13] (یـا بـه اختصـار LSA) یـک تکنیـک بسـیار محبـوب از شـباهت مبتنـی بـر پیکـره اسـت. LSA فـرض می‌کنـد کـه کلماتـی کـه از نظـر معنایـی بـه هـم نزدیـک هسـتند،

---

1. Block distance
2. Manhattan distance
3. Boxcar distance
4. absolute value distance
5. L1 distance
6. city block distance
7. Dice's coefficient
8. Jaccard similarity
9. Euclidean distance
10. matching coefficient
11. overlap coefficient
12. semantic similarity measures
13. latent semantic analysis

در بخش‌های مشابهی از متن رخ می‌دهند. این روش یک ماتریس کلمه به پاراگراف می‌سازد. در این ماتریس، ردیف‌ها کلمات منحصربه‌فرد موجود در متن هستند و ستون‌ها پاراگراف‌ها را نشان می‌دهند. سپس، یک تکنیک ریاضی تحت عنوان تجزیه مقدار منفرد[1] (یا به اختصار SVD) برای کاهش ابعاد استفاده می‌شود. سپس، کلمات به‌وسیلهٔ کسینوس زاویه میان دو بردار که هر بردار یک ردیف را نشان می‌دهد، مقایسه می‌شوند (Landauer & Dumais 1997).

بازیابی اطلاعات-اطلاعات مشترک نقطه‌ای[2] (یا به‌اختصار PMI-IR) روشی برای محاسبه شباهت میان جفت‌های کلمات است که در جست‌وجوی پیشرفته «آلتاویستا»[3] برای محاسبه احتمالات استفاده می‌شود. به‌طور عمده دو کلمه نزدیک به هم در یک صفحه وب[4] امتیاز شباهت PMI-IR بالایی نیز دارند (Turney 2001). اطلاعات مشترک نقطه‌ای هم‌وقوع مرتبه دوم[5] (یا به اختصار SCO-PMI) یک معیار شباهت معنایی با استفاده از اطلاعات مشترک نقطه‌ای است که لیستی از کلمات همسایه دو کلمه را از پیکره بزرگ بر اساس اهمیت آن‌ها مرتب می‌نماید (Islam & Inkpen 2008). مزایای استفاده از این معیار این است که قادر است شباهت میان کلماتی را که زیاد در کنار یکدیگر رخ نداده‌اند بر اساس کلمات همسایه آن‌ها محاسبه کند (Islam & Inkpen 2006).

فاصله گوگل نرمال‌شده[6] (NGD) یک معیار شباهت معنایی است که مجموعه‌ای از کلیدواژه‌ها را در نظر گرفته و با توجه به آن‌ها جست‌وجو انجام می‌دهد. سپس، از تعداد بازیابی‌های درست موتور جست‌وجوی گوگل برای به‌دست آوردن فاصله یا شباهت میان کلمات استفاده می‌کند. کلیدواژه‌ها با معنای مشابه یا یکسان در مفهوم زبان طبیعی متمایل به نزدیک در واحد فاصله گوگل هستند؛ در حالی که کلمات با معانی نامشابه گرایش به فاصله‌های دورتر با استفاده از این معیار فاصله دارند (Cilibrasi & Vitanyi 2007).

استخراج توزیعی کلمات مشابه با استفاده از هم‌وقوعی[7] (یا به‌اختصار DISCO) شباهت توزیعی میان کلمات است که فرض می‌کند که کلمات با معنای مشابه در بافت‌های مشابه در کنار یکدیگر دیده می‌شوند. مجموعه‌های متنی بزرگ‌تر به‌صورت

---

1. singular value decomposition
2. pointwise mutual information-information retrieval
3. AltaVista
4. Web page
5. second-order co-occurrence pointwise mutual information
6. normalized Google Distance
7. extracting distributionally similar words using Co-occurrences

آماری تحلیل می‌شوند تا شباهت توزیعی را به‌دست آورند. این معیار روشی است که شباهت توزیعی میان کلمات را با استفاده از پنجره بافت ساده با اندازه سه کلمه برای شمارش هم‌وقوعی محاسبه می‌نماید. زمانی که دو کلمه در معرض شباهت دقیق با توجه به DISCO قرار گرفتند، بردار کلماتشان از داده‌های نمایه‌شده بازیابی شده و سپس، شباهت میان آن‌ها بر اساس معیار «لین»[1] محاسبه می‌گردد (Lin 1998). اگر به کلمات مشابه بیشتری نیاز پیدا شد، DISCO بردار کلمه مرتبه دوم را برای کلمه داده‌شده برمی‌گرداند (Kolb 2009).

شباهت مبتنی بر دانش[2] یک معیار شباهت معنایی است که درجه شباهت میان کلمات را با استفاده از اطلاعات مستخرج از یک شبکه معنایی مشخص می‌کند (Gomaa & Fahmy 2013). به گفته دیگر، یکی از معیارهای شباهت معنایی است که بر اساس تعیین درجه شباهت میان کلمات با استفاده از اطلاعات مشتق‌شده از شبکه معنایی عمل می‌کند (Mihalcea, Corley & Strapparava 2006). «وردنت» معروف‌ترین شبکه معنایی در حوزه اندازه‌گیری شباهت مبتنی بر دانش میان کلمات است که پایگاه داده بزرگ لغوی غنی در حوزه زبان انگلیسی را جمع‌آوری کرده است (Kumar et al. 2018). اسامی، افعال، صفات، قیود درون مجموعه‌هایی از مترادف شناختی[3] قرار می‌گیرند که هر کدام یک مفهوم ویژه را نشان می‌دهد. این گروه‌ها با کمک مفاهیم معنایی و روابط لغوی به یکدیگر ارتباط پیدا می‌کنند (فتحیان ۱۴۰۰؛ حسینی بهشتی و رجبی ۱۴۰۰؛ فتحیان و همکاران ۱۳۹۹).

معیارهای شباهت مبتنی بر دانش می‌توانند به دو دسته معیارهای شباهت معنایی[4] و معیارهای ارتباط معنایی[5] تقسیم‌بندی شوند. مفاهیم مشابه معنایی بر اساس تشابهشان تشخیص داده می‌شوند. ارتباط معنایی از سوی دیگر، یک مفهوم کلی است که فقط به شکل یا فرم مفهوم گره نمی‌خورد. به گفته دیگر، شباهت معنایی نوعی از ارتباط میان دو کلمه است؛ در حالی که ارتباط معنایی محدوده وسیعی از ارتباطات از جمله «نوعی-از»[6]، «یک-مثال-خاص-از»[5]، «بخشی-از»[6]، «مخالف-از»[7] را دربرمی‌گیرد (Patwardhan, Banerjee & Pedersen 2003). «عسگریان» و همکاران (۱۳۸۶) از هسته آنتولوژی «وردنت» به‌عنوان دانش پس‌زمینه برای خوشه‌بندی متون استفاده کرده‌اند.

1. Lin measure
2. knowledge-based similarity
3. cognitive synonyms
4. measures of semantic similarity
5. measures of semantic relatedness
6. is-a-kind-of
5. is-a-specific-example-of
6. is-a-part-of
7. is-the-opposite-of

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

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

«حاجی غلامرضا» و همکاران (۱۴۰۱) شباهت معنایی جملات فارسی را با استفاده از تطبیق فضای برداری و یادگیری به‌دست آوردند. به‌منظور آموزش داده‌های آموزشی، در روش آن‌ها از ایده نگاشت بین زبانی استفاده می‌شود؛ بدین صورت که فضای برداری تعبیه کلمات انگلیسی را به فارسی نگاشت کرده و با کمک مدل آموزش داده‌شده در زبان انگلیسی، شباهت جملات فارسی را به‌دست می‌آورند.

دسته دیگر از معیارهای شباهت تحت عنوان معیارهای شباهت ترکیبی[1] شناخته می‌شوند که از آن دسته می‌توان به روش ترکیبی از معیارهای مبتنی بر پیکره و مبتنی بر دانش اشاره کرد (Mihalcea, Corley & Strapparava 2006). دو مورد از این معیارها، معیارهای مبتنی بر پیکره، و شش مورد دیگر مبتنی بر دانش هستند. ابتدا این هشت الگوریتم به‌صورت جداگانه مورد ارزیابی قرار گرفتند و سپس با یکدیگر ترکیب شدند. بهترین کارایی با استفاده از روشی که چندین معیار شباهت را درون یکی ترکیب می‌کند، به‌دست آمد.

1. hybrid similarity measures

روشی دیگر برای اندازه‌گیری شباهت معنایی میان جملات یا متون خیلی کوتاه بر اساس معنا و اطلاعات مبتنی بر ترتیب کلمه[1] در Li et al. (2006) مورد بررسی قرار گرفته است. ابتدا شباهت معنایی از یک پایگاه دانش لغوی[2] و پیکره استخراج گردید. سپس، روش پیشنهادی تأثیر ترتیب کلمات روی معنای جمله را بررسی می‌کند. معیارهای شباهت مستخرج از ترتیب کلمه، تعداد کلمات مختلف را به همراه تعداد جفت کلمات در ترتیب‌های مختلف اندازه‌گیری می‌کنند.

Buscaldi et al. (2012) روشی را ارائه کرده و نام آن را شباهت معنایی متن[3] (یا به اختصار STS) نامیدند. این روش شباهت دو متن را از ترکیب میان معنا و نحو اطلاعات تعیین می‌کند. آن‌ها دو تابع لازم (شباهت رشته[4] و شباهت معنایی کلمه[5]) و یک تابع اختیاری (شباهت ترتیب کلمه مشترک[6]) را در نظر گرفتند. روش STS به یک ضریب همبستگی پیرسون بسیار خوب رسیده و توانسته است نتایج Li et al. (2006) را بهبود دهد.

پژوهش‌هایی نیز هستند که معیار شباهت را برای یک حوزه خاص ارائه می‌کنند. Little et al. (2020) معیار شباهتی را برای توییت‌های سیاسی[7] که عمدتاً حاوی متون کوتاهی هستند، ارائه دادند. در پژوهش Qurashi, Holmes & Johnson (2020) تکنیک‌های مختلف برای اندازه‌گیری شباهت معنایی در اسنادی که برای سامانه‌های ایمنی مانند داده‌های راه‌آهن است، بررسی شده‌اند.

پژوهش‌های دیگری از جمله Atoum (2019) معیاری برای ارزیابی روش‌های شباهت متون ارائه دادند. آن‌ها ادعا کردند که معیار همبستگی پیرسون که برای تعیین کارایی روش‌های مختلف شباهت متون استفاده می‌شود، به داده‌های دورافتاده[8] وابسته است. بنابراین، نمی‌تواند در موقعیتی که داده‌ها بسیار شبیه و یا بسیار غیرمرتبط باشند، عمل کنند، و بنابراین با توسعه معیار همبستگی پیرسون به‌صورت مقیاس‌شده توانستند این مشکل را برطرف نمایند. لازم به ذکر است که همه پژوهش‌ها از معیار همبستگی پیرسون برای ارزیابی معیار شباهت ارائه‌شده استفاده نمی‌کنند. بعضی از پژوهش‌ها مانند Lakshmi & Baskar (2021) دو معیار شباهت مبتنی بر فاصله تکرار عبارت و وجود عبارت‌های مشترک در کنار یکدیگر ارائه کرده و کارایی معیارهای ارائه‌شده را با خوشه‌یابی نشان دادند.

مروری بر بیشینه‌ها نشان داد که با توجه به سه دسته کلی روش‌های به‌دست آوردن شباهت متون، می‌توان شباهت میان متون مختلف را به‌دست آورد. سپس، متونی که شباهت بسیار زیادی با یکدیگر دارند، به‌عنوان متون مرتبط در نظر گرفت. ویژگی‌ای که در یک مقاله علمی نسبت به یک متن ساده وجود دارد،

1. word order information
2. lexical knowledge base
3. semantic Text Similarity
4. string similarity
5. semantic word similarity
6. common-word order similarity
7. political tweets
8. outlier

وجود مراجع است که هدف این مقاله نیز یافتن روشی مبتنی بر تحلیل مراجع کتابشناختی برای یافتن مقالات علمی مرتبط است.

۳. روش پژوهش

در این بخش، ابتدا داده‌های پژوهشی توضیح داده می‌شوند. سپس الگوریتم پیشنهادی برای یافتن مقالات مرتبط با یک مقاله ارائه می گردد.

۳-۱. داده‌های پژوهشی

در این پژوهش از دو مجموعه داده فارسی و انگلیسی استفاده می‌شود. برای تهیه داده‌های فارسی، مقالات منتشرشده در ۵ سال اخیر موجود در ۸ نشریه که در جدول ۱، نشان داده شده‌اند، به تصادف استخراج گردیده است. این نشریات عبارت‌اند از: «پردازش علائم و داده‌ها»، «پژوهشنامه پردازش و مدیریت اطلاعات»، «رایانش نرم و فناوری اطلاعات»، «روش‌های عددی در مهندسی»، «علوم رایانش و فناوری اطلاعات»، «علوم رایانشی»، «محاسبات نرم»، و «مهندسی برق و مهندسی کامپیوتر ایران». همان‌طور که جدول ۱، نشان می‌دهد، موضوع سطح کلان این نشریات، علوم فیزیکی و موضوع سطح میانی، علوم کامپیوتر است. برای تهیه داده‌های انگلیسی نیز ۱۰ نشریه به تصادف از حوزه علوم کامپیوتر انتخاب شدند و مقالات نمایه‌شده در «پایگاه استنادی علوم جهان اسلام» در ۵ سال اخیر آن بازیابی گردیدند. جدول ۲، لیست نشریات انگلیسی این پژوهش را نشان می‌دهد که همه با موضوع سطح کلان علوم فیزیکی و سطح میانی علوم کامپیوتر هستند.

جدول ۱. لیست نشریات فارسی پژوهش

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

جدول ۲. لیست نشریات انگلیسی پژوهش

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

۳-۲. الگوریتم شناسایی مقالات مرتبط با تحلیل مراجع کتابشناختی

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

باشند. بنابراین، روش پیشنهادی از عنوان مراجع استفاده می‌کند و در صورتی که عنوان دو مرجع با یکدیگر شباهت بسیار زیاد داشته باشد، مرتبط در نظر گرفته می‌شوند.

به‌منظور استخراج عنوان مقالات، می‌بایست فرمت‌های مختلف مراجع بررسی شوند. فرمت‌های مختلفی برای تهیه مرجع وجود دارد که این فرمت‌ها با توجه به قوانین و دستورالعمل‌های موجود در هر نشریه مشخص می‌شود. در این پژوهش از فرمت‌های مختلف مرجع‌دهی که در scholar.google.com وجود دارد، استفاده شده و عبارت‌اند از: MLA، APA، Chicago، Harvard و Vancouver. در ادامه، فرمت‌های مختلف مرجع‌دهی با ذکر یک نمونه در هر فرمت توضیح داده می‌شود.

فرمت MLA

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

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

فرمت APA

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

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

فرمت Chicago

در این شیوه ارجاع‌دهی، ابتدا نام خانوادگی نویسندگان و سپس نام کوچک آن‌ها

به‌ترتیب مشارکتشان در مقاله مشخص می‌شود. در ادامه، عنوان مقاله در گیومه و نشریه نوشته می‌شوند. سری، شماره، سال انتشار و شماره صفحات نیز در انتهای این شیوه ارجاع‌دهی مشخص می‌گردند. نمونه‌ای از شیوه ارجاع با فرمت Chicago در ادامه آمده است:

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

فرمت Harvard

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

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

فرمت Vancouver

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

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

با توجه به اینکه داده‌های موجود در این پژوهش از نشریات مختلف گرفته شده و هر کدام از آن‌ها از یکی از این فرمت‌ها برای تهیه مراجع استفاده می‌نماید، می‌بایست تمامی حالات مختلف تهیه مرجع مد نظر قرار گیرد و با توجه به آن‌ها، عنوان موجود در هر مرجع را استخراج نمود. بدین منظور برنامه‌ای به زبان «پایتون» (نسخه ۳/۷) نوشته شد

که با در نظر گرفتن تمامی این حالات قرار گرفتن عنوان که ناشی از فرمت‌های مختلف مراجع است، عناوین موجود را استخراج نماید.

نمونه‌ای از اجرای برنامه با در نظر گرفتن مثال بالا در شکل ۲، ارائه شده است.

```
ng-performance tradeoff in cognitive radio enabled smart grid." IEEE Transactions on Smart Grid 4.1 (2013):
ognitive radio enabled smart grid.

 Zhang, Y., Maharjan, S., & Gjessing, S. (2013). Sensing-performance tradeoff in cognitive radio enabled sm
ognitive radio enabled smart grid

en, Xianghui Cao, Yan Zhang, Sabita Maharjan, and Stein Gjessing. "Sensing-performance tradeoff in cognitiv
ognitive radio enabled smart grid.

 X., Zhang, Y., Maharjan, S. and Gjessing, S., 2013. Sensing-performance tradeoff in cognitive radio enable
ognitive radio enabled smart grid

 Zhang Y, Maharjan S, Gjessing S. Sensing-performance tradeoff in cognitive radio enabled smart grid. IEEE
ognitive radio enabled smart grid
```

شکل ۲. نمونه‌ای از اجرای ماژول استخراج عنوان از فرمت‌های مختلف ارجاع‌دهی

سپس با استفاده از معیارهای شباهت، در صورتی که شباهت میان دو عنوان از یک مقدار آستانه بیشتر باشد، دو مقاله مرتبط در نظر گرفته شود. برای به‌دست آوردن شباهت میان دو عنوان مرجع، ابتدا عملیات پیش‌پردازش روی عنوان آن‌ها انجام گرفته و همچنین کلمات ایستا حذف می‌شوند. هر عنوان بر اساس کاراکتر فاصله[1] به برداری از عبارت‌ها تبدیل می‌شود و شباهت میان دو بردار بر اساس اشتراک عبارت‌های تشکیل‌دهنده دو بردار با توجه به معیار «جاکارد» به‌دست می‌آید. لازم به ذکر است که در صورتی که شباهت میان دو عبارت موجود در هر عنوان بر اساس الگوریتم «لونشتین-دامرا» از معیار آستانه که در این پژوهش ۰/۸ در نظر گرفته شده، بیشتر باشد، آن دو عبارت در معیار «جاکارد»، مشترک در نظر گرفته می‌شوند.

الگوریتم پیشنهادی برای یافتن مقالات مرتبط با مقاله داده‌شده با استفاده از تحلیل مراجع در شکل ۳، آمده است. ورودی این الگوریتم شناسه یک مقاله و همچنین تعداد مقالات بازیابی‌شده است. خروجی نیز لیست شناسه مقالات مرتبط با مقاله داده شده است. ردیف ۱۲ و ۱۳ تمامی شناسه‌های مقالات را بررسی می‌نماید. ردیف‌های ۱۴ تا ۱۶ الگوریتم تمامی مراجع هر کدام از شناسه‌ها را بررسی کرده و عنوان آن‌ها را استخراج می‌کند. ردیف ۱۷ در صورتی که شباهت عنوان استخراج‌شده با عنوان مورد پرسش

1. space

کاربر، از یک مقدار آستانه که در این پژوهش ۰/۸ در نظر گرفته شده، بیشتر باشد، یک مقدار به متغیر شمارش اضافه می‌کند. سرانجام، ردیف‌های ۲۳ تا ۲۵ مقالاتی را که بیشترین شباهت را با مقاله داده‌شده داشته باشند، بازیابی می‌کند.

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

شکل ۳. الگوریتم پیشنهادی برای یافتن مقالات مرتبط با یک مقاله

٤. تجزیه‌وتحلیل یافته‌ها

در این فصل به مطالعهٔ کارایی روش پیشنهادی می‌پردازیم. ابتدا معیار ارزیابی استفاده‌شده در این پژوهش معرفی می‌گردد و سرانجام، کارایی روش پیشنهادی را با توجه به معیار ارزیابی داده‌شده بررسی خواهیم کرد.

٤-١. معیار ارزیابی

در سامانه‌های بازیابی اطلاعات عمدتاً از معیارهای دقت[1] و بازیافت[2] برای سنجش

1. precision
2. recall

میزان کارایی روش‌های ارائه‌شده استفاده می‌شود. معیار بازیافت تعداد مقالات مرتبط بازیابی‌شده از میان کل مقالات مرتبط را در نظر می‌گیرد. بنابراین، این معیار نیاز به داشتن کل مقالات مرتبط به ازای هر پرسش کاربر است. از آنجا که کل مقالات مرتبط به ازای هر پرسش، به ازای داده‌های این پژوهش در دسترس نیست، نمی‌توان از این معیار برای ارزیابی روش پیشنهادی استفاده کرد.

معیار دقت بررسی می‌کند که از میان اسناد بازیابی‌شده، چند سند با پرسش کاربر مرتبط است. به‌عنوان مثال، در صورتی که تعداد اسناد بازیابی‌شده ۵ باشد و از این میان تنها ۲ سند مرتبط با مقاله مورد پرسش کاربر باشد، معیار دقت در این حالت ۰/۴ می‌شود. فرمول ۱، معیار دقت را نشان می‌دهد.

(۱) دقت =تعداد مقالات بازیابی‌شده مرتبط /تعداد مقالات بازیابی‌شده

٤-۲. یافته‌ها

به‌منظور ارزیابی دقیق‌تر روش پیشنهادی، به ازای داده‌های انگلیسی که از نشریات انگلیسی گرفته شده‌اند، ۳۰ مقاله به‌صورت تصادفی انتخاب شدند. همین رویه برای داده‌های فارسی نیز انجام گرفت و ۳۰ مقاله به‌صورت تصادفی از میان مجموعه مقالات انتخاب گردیدند.

روش پیشنهادی با چهار روش دیگر برای به‌دست آوردن مقالات مرتبط با یک مقاله مقایسه شده‌اند که این روش‌ها به‌ترتیب TF-IDF (Wang, Wang & Zhang 2010)، word2vec (Church 2017)، DOC2VEC (Le & Mikolov 2014) و BERT (Delving et al. 2018) است. جدول ۳، مقایسه میانگین دقت و انحراف معیار روش پیشنهادی را با روش‌های دیگر نشان می‌دهد. لازم به ذکر است که هر کدام از اعداد این جدول میانگین ۳۰ بار اجرای هر کدام از روش‌هاست.

جدول ۳. مقایسه دقت روش پیشنهادی با دیگر روش‌ها

	روش پیشنهادی	TF-IDF	Word2vec	DOC2VEC	BERT
داده‌های نشریات فارسی	۰/۷	۰/۴۲	۰/۲۷	۰/۴۱	۰/۵۳
داده‌های نشریات انگلیسی	۰/۶	۰/۴۲	۰/۴۷	۰/۳۹	۰/۶۰

به‌رغم روش پیشنهادی که مقالات مرتبط را بر اساس تحلیل مراجع بازیابی می‌کند، چهار روش

دیگر با تحلیل محتوای اطلاعات کتابشناختی مقالات مانند چکیده، عنوان و کلیدواژه عمل می کنند. بنابراین به‌منظور مقایسه، در قدم اول، چکیده، عنوان و کلیدواژه‌های مقالات مورد پیش‌پردازش قرار گرفت.

برای انجام این کار، ابتدا متن با توجه به کاراکترهای جداکننده[1] که شامل {، ؛ " ( ): . <>} هستند، به مجموعه‌ای از توکن‌ها تبدیل شد. سپس، عملیات ریشه‌یابی[2] روی آن‌ها انجام گرفت. هر واژه با کدی یکتا ذخیره شد. افزون بر واژه، تعداد رخداد آن در مجموعه مقالات نیز محاسبه و ذخیره گردید. همچنین ایست‌واژه[3]ها، علائم و اعداد حذف شدند. خروجی این مرحله، واژگان پردازش‌شده‌ای است که فرکانس تکرار آن‌ها در هر مقاله نیز مشخص شده است.

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

در روش مبتنی بر TF-IDF (Wang, Wang & Zhang 2010) هر مقاله که حاوی عنوان، کلیدواژه و چکیده است، به یک بردار تبدیل شد که هر کدام از اعضای آن بردار، TF-IDF کلمات آن مقاله است. بنابراین، کل داده‌ها به‌صورت یک بردار عددی تبدیل شدند و این رویه در فاز آفلاین انجام گرفت. حال زمانی که کاربر یک مقاله را انتخاب کرده و به‌دنبال مقالات مرتبط با آن مقاله می‌گردد، بردار آن مقاله با دیگر بردارهای مقالات موجود مقایسه شده و بردارهایی که نزدیک‌ترین فاصله را با آن دارند، بازیابی شدند. لازم به ذکر است که برای به‌دست آوردن فاصله میان دو بردار که هر کدام یک مقاله است، از معیار کسینوسی استفاده شده است. همان‌طور که جدول ۳، نشان می‌دهد، این روش توانسته است با دقت ۰/۴۲ مقالات مرتبط با یک مقاله را بازیابی نماید.

سه روش دیگری که در این پژوهش برای مقایسه روش پیشنهادی با آن‌ها استفاده شده از تعبیه کلمات[4] استفاده می کنند. تعبیه کلمات بردارهای عددی هستند که نمایانگر کلمات یک پیکره هستند و کاربردهای گسترده‌ای به‌خصوص در حوزه پردازش زبان طبیعی دارند. روش تعبیه کلمات اجازه می‌دهد که به‌طور غیرصریح، اطلاعاتی را از دنیای بیرون به مدل‌های زبانی اضافه کنید. در تعبیه کلمات، تمام کلمات استفاده‌شده در یک زبان، به‌وسیله مجموعه‌ای از اعداد اعشاری (در قالب یک بردار) نمایش داده می‌شود.

1. Delimitor characters
2. stemming
3. stop words
4. word embedding

در واقع، تعبیه کلمات، بردارهای nبُعدی هستند که تلاش می کنند معنای کلمات و محتوای آن‌ها را با مقادیر عددی خود ثبت و ضبط کنند.

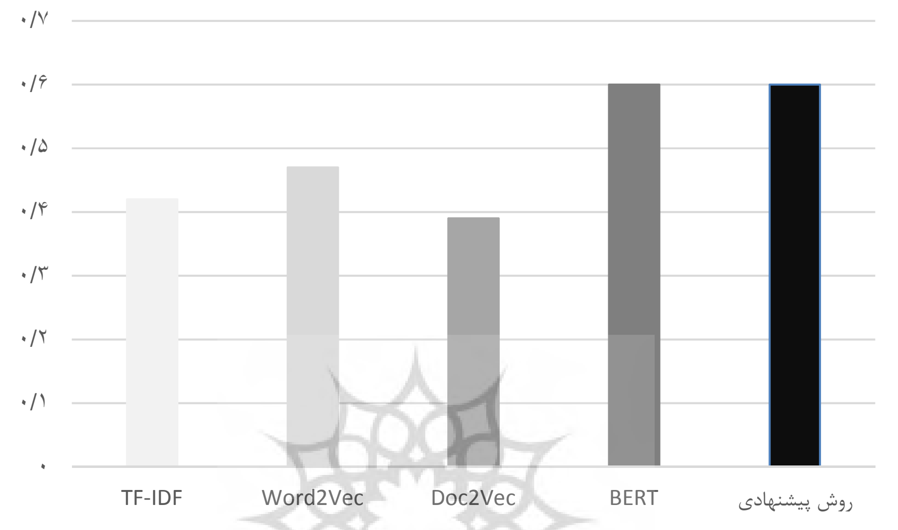
در روش‌های word2vec و DOC2VEC به کمک شبکه عصبی یک بردار با اندازه کوچک و ثابت برای نمایش تمام لغات و متون در نظر گرفته شده و با اعداد مناسب در فاز آموزش برای هر لغت این بردار محاسبه می‌شود. زمانی که کاربر یک مقاله را برای یافتن مقالات مرتبط انتخاب می کند، روش مبتنی بر word2vec ابتدا عملیات پیش‌پردازش را روی چکیده، عنوان و کلیدواژه مقاله انتخاب‌شده به‌کار گرفته و سپس آن را تبدیل به بردار می کند. سرانجام، میزان شباهت بردار مقاله داده‌شده با دیگر بردار مقالات با استفاده از معیار کسینوسی محاسبه می‌شود (Church 2017; Le & Mikolov 2014).

روش دیگری که در این پژوهش برای مقایسه با روش پیشنهادی به کار رفته، الگوریتم BERT است که توسط شرکت «گوگل» ارائه شده است و روی معماری ترنسفورمرها برای مدل‌سازی زبان‌ها عمل می کند (Delving et al. 2018). در اینجا هم بعد از پیش‌پردازش چکیده، کلیدواژه و عنوان مقالات، تمامی داده‌ها با استفاده از این مدل، آموزش می‌بینند و تمامی این عملیات در فاز آفلاین انجام می گیرد. در مرحله آنلاین، زمانی که کاربر یک مقاله را برای یافتن مقالات مرتبط انتخاب می‌نماید، عنوان، کلیدواژه و چکیده مقاله پیش‌پردازش شده و سپس تبدیل به بردار می گردد. سرانجام، شبیه‌ترین بردار به بردار پرسش کاربر با توجه به معیار کسینوسی بازیابی می‌شود.

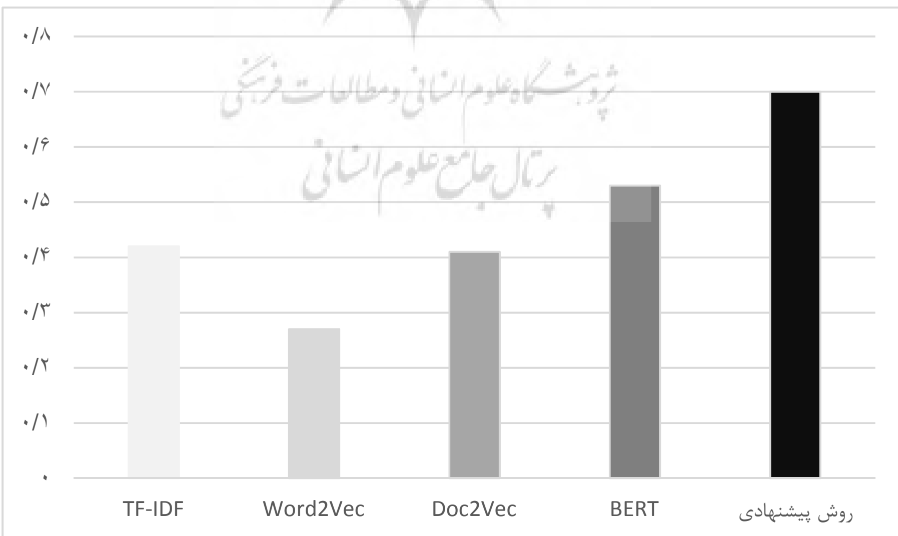
برای روشن‌تر شدن روند به‌دست آوردن نتایج، یک مثال توضیح داده می‌شود. یکی از مقالات موجود در داده‌ها به تصادف انتخاب گردید. سپس، عنوان این مقاله به همراه کلیدواژه و چکیده آن به روش‌های به‌دست آوردن مقاله مرتبط با تحلیل محتوا مانند TF-IDF، Word2vec، DOC2VEC و BERT داده شد. پارامتر T در الگوریتم ۱، که همان تعداد مقالات بازیابی‌شده است، ۵ در نظر گرفته شد. سپس به ازای هر کدام از روش‌ها پنج مقاله بازیابی‌شده مورد تحلیل و بررسی دقیق قرار گرفت تا مشخص شود کدام‌یک از آن‌ها با مقاله مورد پرسش ارتباط دارند. از آنجا که تمامی نشریاتی که برای داده‌های پژوهشی مورد استفاده قرار گرفتند از نظر موضوعی با تخصص پژوهشگر مطابقت داشت، بنابراین این تحلیل توسط پژوهشگر به‌عنوان عامل انسانی متخصص انجام گرفت. سپس تمامی منابع مقاله به همراه کل منابع تمامی مقالات به روش پیشنهادی داده شد و مقالات بازیابی‌شده مورد تحلیل قرار گرفت و مقالات مرتبط آن شناسایی گردید. همین روند برای تمامی ۳۰ مقاله فارسی و همچنین ۳۰ مقاله انگلیسی انجام گرفت. در انتها، میانگین و انحراف معیار در جدول ۳، گزارش گردید.

به‌منظور مقایسه شهودی الگوریتم پیشنهادی با دیگر روش‌ها نمودار میله‌ای آن‌ها رسم شد که شکل‌های ۴ و ۵ این مقایسه را روی داده‌های فارسی و انگلیسی نشان می‌دهد. همان‌طور که این شکل‌ها

نشان می‌دهند، روش پیشنهادی در هر دو مجموعه داده فارسی و انگلیسی توانسته است با دقت خیلی خوبی مقالات مرتبط با یک مقاله را بازیابی کند. از میان روش‌های مبتنی بر تحلیل محتوا، BERT توانسته است با دقت خیلی خوبی نسبت به دیگر روش‌ها مقالات مرتبط را بازیابی کند.



شکل ٤. مقایسه روش پیشنهادی با دیگر روش‌ها روی داده‌های انگلیسی



شکل ٥. مقایسه روش پیشنهادی با دیگر روش‌ها روی داده‌های فارسی

۵. بحث و نتیجه‌گیری

در سال‌های اخیر حجم مقالات علمی به‌طور فزاینده‌ای افزایش پیدا کرده است. از سوی دیگر، شناسایی مقالات مرتبط از اولین قدم‌های ضروری در هر پژوهشی است که می‌بایست توسط پژوهشگران انجام شود. هدف سامانه‌های بازیابی مقالات علمی، کمک به پژوهشگران در این راستاست. یکی از قابلیت‌هایی که در سامانه‌های بازیابی اطلاعات مقالات علمی به پژوهشگران کمک بسیار زیادی می‌کند، ویژگی یافتن مقالات علمی مرتبط با یک مقاله است؛ به گفته دیگر، ویژگی که به پژوهشگر اجازه می‌دهد با انتخاب یکی از مقالات بازیابی‌شده، دیگر مقالات مرتبط با آن را مشاهده نماید.

در این پژوهش، روشی مبتنی بر تحلیل مراجع کتابشناختی برای شناسایی مقالات مرتبط با یک مقاله ارائه شد. این روش بدین صورت عمل می‌کند که شباهت میان عنوان مراجع مقاله داده‌شده با عنوان مراجع دیگر مقالات محاسبه می‌شود و مقالاتی به‌عنوان مقاله مرتبط با یک مقاله بازیابی‌شده که عناوین آن‌ها بیشترین شباهت را با یکدیگر داشته باشند. بدین منظور از ترکیب معیار «جاکارد» و «لونشتین» برای به‌دست آوردن شباهت میان عناوین مقالات استفاده گردید.

به‌منظور ارزیابی روش پیشنهادی، آن را با روش‌های مبتنی بر تحلیل محتوا مقایسه نمودیم که یافته‌ها نشان‌دهنده کارایی روش پیشنهادی است. برای نشریات انگلیسی، BERT نسبت به دیگر روش‌های تحلیل محتوا از دقت بالاتری برخوردار است و در ادامه، روش‌های مبتنی بر word2vec، TF-IDF قرار دارند. روش مبتنی بر DOC2VEC کمترین دقت را داشته است. برای نشریات فارسی، BERT بیشترین دقت و word2vec کمترین دقت را دارد. در هر دو داده‌های فارسی و انگلیسی، روش پیشنهادی که در واقع، روشی مبتنی بر تحلیل استنادی است، بیشترین دقت را به همراه داشته است.

روش پیشنهادی در مواردی هیچ مقاله‌ای را بازیابی نکرد؛ زیرا از نظر استنادی با هیچ مقاله دیگری اشتراکی نداشت. البته، روش‌های دیگر نیز با اینکه مقالاتی را بازیابی کردند، ولی تنها یکی از آن‌ها با این مقاله تا اندازه‌ای مرتبط بود. در بعضی موارد نیز روش مبتنی بر تحلیل استنادی، تنها دو مقاله را مرتبط تشخیص داده و بازیابی نمود که آن دو نیز کاملاً مرتبط بودند؛ در صورتی که روش‌های مبتنی بر تحلیل محتوا نتوانستند مقالات مرتبط را بازیابی کنند.

به‌عنوان جمع‌بندی نهایی، نتایج اعمال روش پیشنهادی روی داده‌های فارسی و انگلیسی نشریات انتخاب‌شده در علوم کامپیوتر نشان‌دهنده کارایی آن در یافتن مقالات مرتبط با یک مقاله است. به گفته دیگر، در صورتی که پوشش جامعی از لیست مراجع مقالات وجود داشته باشد، روش پیشنهادی قادر خواهد بود با دقت بالایی مقالات مرتبط با یک مقاله را پیدا کند. بنابراین، الگوریتم ارائه‌شده می‌تواند در سامانه‌های بازیابی مقالات علمی که اطلاعات مراجع مقالات را داشته باشند، استفاده شود و یک

ویژگی ارزشمندی را به سامانه اضافه نماید که به کاربرپسندتر شدن سامانه بازیابی اطلاعات کمک شایانی می‌نماید. پژوهشگر به‌عنوان پژوهش‌های آتی این طرح، به‌دنبال ترکیب روش‌های تحلیل استنادی و تحلیل محتوا به‌منظور ارائه روشی قوی‌تر برای یافتن مقالات مرتبط با یک مقاله است.

فهرست منابع

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

حاجی غلامرضا، مینا، محمدرضا محمدزاده، سید محمدرضا محمدی، و محمدعلی کیوان‌راد. ۱۴۰۱. شباهت معنایی جملات فارسی با استفاده از تطبیق فضای برداری و یادگیری عمیق مقاله. *پدافند الکترونیکی و سایبری* ۲: ۴۳–۵۶.

حسینی بهشتی، ملوک‌السادات، و تقی رجبی. ۱۴۰۰. پیشنهاد طرح تدوین فرااصطلاح‌نامهٔ ایرانداک با تکیه بر الگو و ساختار نظام زبان واحد پزشکی (یو ام ال اس). *پردازش و مدیریت اطلاعات* ۱۰۵: ۲۲۹–۲۵۳.

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

عباسی، شیرین، و بابک وزیری. ۱۳۹۴. الگوریتم‌های خوشه‌بندی در داده‌های عظیم، کنفرانس بین‌المللی پژوهش‌های کاربردی در فناوری اطلاعات. کامپیوتر و مخابرات. دانشگاه آزاد اسلامی واحد تربت حیدریه.

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

فتحیان دستگردی، اکرم. ۱۴۰۰. انتشار معنایی: بازنمون معنایی انتشارات علمی مبتنی بر مجموعه هستی‌نگاری‌های اسپار. *مطالعات ملی کتابداری و سازماندهی اطلاعات*. ۳۲ (۳): ۲۳–۵۵.

ـــــــ، سید مهدی طاهری، اعظم صنعت‌جو، و محسن کاهانی. ۱۳۹۹. پیاده‌سازی روش داده‌های پیوندی در نظام کتابخانه‌ای: بررسی مؤلفه‌های مورد نیاز و ارائه یک الگو. *بازیابی دانش و نظام‌های معنایی* ۲۵: ۶۷–۹۵.

References

Atoum, I. 2019. Scaled Pearson’s correlation coefficient for evaluating text similarity measures. Infinite Study.?: Modern Applied Science.

Buscaldi, D., R. Tournier, N. Aussenac-Gilles, & J. Mothe. 2012. Irit: Textual similarity combining conceptual similarity with an n-gram comparison method. In * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012) (pp. 552-556). Montreal, Canada.

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

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

Devlin, J., M. W. Chang, K. Lee, & K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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

Farouk, M. 2019. Measuring sentences similarity: a survey. arXiv preprint arXiv:1910.03940.

Feldman, R., & J. Sanger. 2007. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge university press.

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

Hall, P. A., & G. R. Dowling. 1980. Approximate string matching. *ACM computing surveys (CSUR)* 12 (4): 381-402.

Hotho, A., A. Nürnberger, & G. Paaß. 2005. A brief survey of text mining. Journal for Language Technology and Computational Linguistics, 20 (1): 19-62.

Islam, A., & D. Inkpen. 2006. Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words. Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy (pp. 1033-1038).

_____. 2008. Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 2 (2): 1-25.

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

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

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

Landauer, T. K., & S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104 (2): 211.

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

Le, Q., & T. Mikolov. 2014. Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR. Beijing, China.

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

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

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

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

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

Mihalcea, R., C. Corley, & C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. Proceedings of the 21st national conference on Artificial intelligence, 1 (6): 775-780.

Niwattanakul, S., J. Singthongchai, E. Naenudorn, & S. Wanapu. 2013. Using of Jaccard Coefficient for keywords similarity. In Proceedings of the International Multiconference of Engineers and Computer Scientists 1 (6): 380-384).

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

Peterson, J. L. 1980. Computer programs for detecting and correcting spelling errors. *Communications of the ACM* 23 (12): 676-687.

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

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

Sadeghi, M., & J. Vegas, J. 2014. Automatic identification of light stop words for Persian information retrieval systems. *Journal of Information Science* 40 (4): 476-487.

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

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

Wang, N., P. Wang, & B. Zhang. 2010. An improved TF-IDF weights function based on information theory. In 2010 International Conference on Computer and Communication Technologies in Agriculture Engineering Chengdu, China. (Vol. 3, pp. 439-441).

نیلوفر مظفری

متولد ۱۳۶۴ دارای مدرک دکتری در رشته هوش مصنوعی از دانشگاه شیراز است. ایشان هم‌اکنون استادیار مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری است. یادگیری ماشین، یادگیری عمیق و پردازش زبان طبیعی از جمله علایق پژوهشی وی است.

تاکسونومی شناسایی مشتریان صنعت بانکی با به‌کارگیری یادگیری ماشین: مروری نظام‌مند با رویکرد فراترکیب

الهه باغانی
دانشجوی دکتری مدیریت فناوری اطلاعات؛ دانشکده مدیریت و اقتصاد؛ دانشگاه تربیت مدرس؛ تهران، ایران؛
Elaheh.Baghani@modares.ac.ir

شعبان الهی
دکتری مدیریت؛ استاد؛ گروه مدیریت؛
دانشکده علوم اداری و اقتصادی؛
دانشگاه ولی عصر (عج) رفسنجان؛ رفسنجان، ایران؛
پدیدآور رابط elahi@vru.ac.ir

علیرضا حسن‌زاده
دکتری مدیریت؛ استاد؛
مدیر گروه مدیریت فناوری اطلاعات؛ دانشکده مدیریت و اقتصاد؛ دانشگاه تربیت مدرس؛ تهران، ایران؛
ar_hassanzadeh@modares.ac.ir

علی رجب‌زاده
دکتری مدیریت؛ استاد؛ مدیر گروه مدیریت صنعتی؛
دانشکده مدیریت و اقتصاد؛ دانشگاه تربیت مدرس؛
تهران، ایران alirajabzadeh@gmail.com

دریافت: ۱۴۰۱/۰۷/۲۶ | **پذیرش:** ۱۴۰۱/۱۲/۰۸ | مقاله برای اصلاح به مدت ۱۸ روز نزد پدیدآوران بوده است.

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۲۲۵۱-۸۲۲۳
شاپا (الکترونیکی) ۲۲۵۱-۸۲۳۱
نمایه در SCOPUS ،ISC ،LISTA و
jipm.irandoc.ac.ir
دورهٔ ۳۹ | شمارهٔ ۱ | ص‌ص ۳۵۷-۳۹۲
پاییز ۱۴۰۲
https://doi.org/jipm.39.1

چکیده: امروزه، مشتری در هیچ صنعتی صرفاً به‌دنبال محصول نیست و دریافت سرویس شخصی‌سازی‌شدهٔ مبتنی بر نیازمندی‌های خود و خلق تجربه‌ای متفاوت از سازمان را انتظار دارد. از نگاه دیگر، طراحی سرویس‌های متناسب با نیاز مشتری مستلزم بررسی موشکافانه داده‌های مرتبط با مشتری در ابعاد مختلف است. بنابراین، شناخت مشتری نیازمند نگرشی نظام‌مند است تا اهداف، فاکتورهای تأثیرگذار، الگوریتم‌ها و روش‌های مناسب این حوزه شناسایی شود. پژوهش پیش رو با رویکرد فراترکیب، ابعاد حوزه شناخت مشتری صنعت بانکی و ملاحظات آن را با رویکرد داده‌محور و به‌کارگیری یادگیری ماشین تحلیل نموده است. از این رو، روش پژوهش بر حسب هدف کاربردی و بر حسب گردآوری اطلاعات فراترکیب است. برای انتخاب مقاله‌ها با جست‌وجو در پایگاه داده‌های معتبر «وب‌آوساینس» و «اسکوپوس»، ۴۳ سند که در فاصله زمانی ۲۰۱۶- ۲۰۲۲ منتشر شده،

به‌عنوان اسناد مرتبط و معتبر شناسایی و در ادامه نیز با رویکرد فراترکیب، بررسی و کدگذاری شدند. نتایج حاصل از فراترکیب منجر به شناسایی سه مقوله اصلی، ۱) اهداف شناسایی مشتری: درک بینش نسبت به مشتری، شناسایی ریسک مشتری، اهداف سازمانی، تعیین ارزش طول عمر مشتری و مدیریت محصول، ۲) فاکتورهای شناسایی مشتری: جمعیت‌شناختی، مالی و رفتاری، و ۳) الگوریتم‌های یادگیری ماشین: Regularization، Ensemble، Neural Networks، Probabilistic، Instanced Based، Dimensionality Reduction، Decision Tree، Bayesian، Regression، و Clustering گردید. بر اساس یافته‌های پژوهش جاری، متناسب با هدف شناسایی مشتری، داده‌های موجود و فاکتورهای انتخابی، الگوریتم‌های پایه و ترکیبی می‌توانند راهگشا باشد، اما نکته مهم پیش‌پردازش دقیق داده‌هاست. نکته دیگر اینکه تفکیکی در حوزه مشتری حقیقی و حقوقی صورت نپذیرفته است و اغلب مطالعات بر روی مشتری حقیقی تمرکز داشته‌اند که از دلایل آن می‌توان به پیچیدگی زنجیره تعاملات مالی مشتریان حقوقی اشاره نمود. همچنین با توجه به عدم تأکید بانک‌ها به شعب در خصوص لزوم تکمیل اطلاعات مندرج در فرم‌های افتتاح حساب یا عدم طراحی سرویس مناسب جهت تکمیل اطلاعات در بسترهای الکترونیک، شناخت دقیق‌تر مشتری مستلزم بازبینی در این فرایندهاست. شایان ذکر است که در هیچ‌یک از مطالعات انجام‌شده، شناخت مشتری تنها با استفاده از فاکتورهای جمعیت‌شناختی انجام نشده و بسته به هدف مطالعه، فاکتورها به‌صورت ترکیبی استفاده شده است.

کلیدواژه‌ها: شناخت مشتری، فاکتورهای شناسایی مشتری، یادگیری ماشین، فراترکیب، صنعت بانکی

۱. مقدمه

بررسی بازار خدمات مالی و بانکی نشان می‌دهد که با ایجاد بازارها و خدمات تلفیقی جدید، مرز بین خدمات مالی و صنایع مجاور نظیر صنایع خرده‌فروشی، مخابرات و فناوری از بین رفته است. از طرف دیگر، نفوذ و اختلال استارت‌آپ‌ها در حوزه‌های مختلف در صنعت بانکداری در حال گسترش است (World Economic Forum 2017). این است که ایجاد و حفظ وفاداری مشتری در بخش خدمات برای کسب‌وکار، چالش‌برانگیز و راهبردی است، همچنین هزینه‌های مرتبط با جذب مشتری جدید نسبت به نگهداری آن‌ها بسیار بیشتر است و مدیریت ارتباط با مشتری برای ادامه کار و موفقیت حیاتی است. مشکلی که اکنون بسیاری از بانک‌ها با آن روبه‌رو هستند، این است که مشتریان، خدمات مالی بسیار جذابی را از بازیگران جدید و غیربانکی دریافت می‌کنند. در حقیقت، بیش از ۵۰ درصد از نوآوری‌ها در بخش مالی توسط مؤسسات غیربانکی انجام می‌شود. برای پاسخگویی به این چالش‌ها، مؤسسات بانکی به‌وضوح باید در زمینه نوآوری بهتر عمل کنند. از طرف دیگر، طبق تحقیقات «گارتنر» در سال ۲۰۲۱، ۴۶ درصد از مشتریان نمیتوانند بین تجربیات دیجیتالی اکثر برندها تفاوت قائل شوند و تنها ۱۴ درصد از مشتریان اعلام

نموده‌اند که تجربه دیجیتال متفاوتی دریافت کرده‌اند (Gartner 2021). این آمار نشان می‌دهد که اغلب سازمان‌ها به‌رغم تلاش‌های صورت‌پذیرفته، در ایجاد تجربه جذاب برای مشتری شکست خورده یا در حال رفتن به مسیر اشتباه هستند. در واقع، در حال حاضر، خدمات بانکی تا حد زیادی به ارائه خدمت متمرکز است. به گفته دیگر، ذهنیت تولیدمحور[1]، بر عملکرد و فرایندهای بانک‌ها سایه افکنده و خدمات بانکی بدون درک کافی از نیازهای مشتری توسعه می‌یابند (Komulainen & Saraniemi, 2019).

تغییر رویکرد فوق و حرکت به سمت مشتری‌محوری[2] نیازمند نگاه موشکافانه‌تر به ویژگی‌های مشتری و رفتارهای مالی و رفتاری وی است. از طرفی، با توجه به حجم بسیار بالای داده‌های مشتری انتخاب ویژگی‌های تأثیرگذار و شناخت فناوری‌ها و روش‌هایی که فرایند شناخت را ضمن تسهیل، دقیق‌تر نمایند نیز از الزامات صنعت است. با عنایت به موارد فوق، نیاز به تحقیقات بیشتر و توسعهٔ محصولات شخصی‌سازی‌شده نظیر وام‌های خاص بسته به اطلاعات واقعی مشتریان (میزان درآمد، فرزند، پس‌انداز، نوع کسب‌وکار، تاریخچه تراکنش‌ها و غیره) از طریق فناوری‌های آنالیز داده و انواع روش‌های دسته‌بندی و ارزش‌گذاری مشتری ضروری است (Garcia-Mendez et al., 2020; Ladyzynski et al., 2019). در واقع، هرچه تحلیل اطلاعات و داده‌های مربوط به مشتری غنی‌تر باشد، مزیت رقابتی بانک بیشتر خواهد بود و حرکت سازمان به سمت تحول دیجیتال سریع‌تر خواهد شد (Behare et al., 2018).

در این راستا پژوهش‌های متعددی بر روی داده‌های دنیای واقعی، داده‌های جمع‌آوری‌شده از پرسشنامه، داده‌های شبیه‌سازی‌شده در پایگاه داده‌های آنلاین نظیر[3] و از طریق الگوریتم‌های مختلف یادگیری ماشین بر روی فاکتورهایی نظیر جمعیت‌شناختی، فاکتورهای مالی و غیره انجام شده است. لیکن این مطالعات بر پیاده‌سازی الگوریتم‌ها از طریق ابزارهای موجود و مقایسه دقت آن‌ها متمرکز است و مرور نظام‌مندی در این حوزه مورد توجه نبوده است. بنابراین، با توجه به شکاف نظری فوق، این پژوهش تلاش می‌کند که یک تاکسونومی در خصوص شناخت مشتری در صنعت بانکی ارائه نماید. همچنین، با توجه به اینکه تاکنون دسته‌بندی فاکتورهای شناختی مشتریان و روش‌ها و الگوریتم‌های یادگیری ماشین در صنعت بانکی ارائه نشده است، بنابراین خروجی پژوهش پیش رو نیاز صنعت بانکی در خصوص شناخت و انتخاب فاکتورها، روش‌ها و الگوریتم‌های یادگیری ماشین در پروژه‌های مرتبط، نظیر باشگاه مشتریان، طرح‌های وفاداری، طراحی سرویس‌های شخصی‌سازی‌شده و کمپین‌های تبلیغاتی را مرتفع خواهد کرد.

یکی دیگر از اهداف پژوهش جاری، ارتقای تجربهٔ مشتریان بانکی به‌عنوان یکی از مؤلفه‌های اصلی تحول دیجیتال و از طریق شناسایی مشتری و به‌تبع آن، ارائه سرویس‌های متناسب با نیاز وی است. در واقع،

1. product centric   2. customer centric   3. University of California Irvine (UCI)

خروجی این پژوهش می‌تواند به‌عنوان مرجعی برای پژوهشگران آتی و مدیران اجرایی این حوزه، جهت انتخاب الگوریتم‌های یادگیری ماشین و متغیرهای ورودی آن‌ها، متناسب با هدف شناسایی مشتری مورد استفاده قرار گیرد.

۲. مروری بر مبانی نظری و پیشینه پژوهش

در عصر تحول دیجیتال، تکنیک‌های یادگیری ماشین در هر صنعتی مفید واقع می‌شود. از یادگیری ماشین در حوزه‌های مختلف نظیر تجهیزات ورزشی (Pawełoszek, 2021)، صنعت برق (Razavi et al., 2019)، صنعت دیجیتال، رسانه‌های اجتماعی و بازار موبایل (Müller et al., 2018)، صنعت فین‌تک‌ها (A. Sheikha et al., 2019)، خرده‌فروشی آنلاین (Wu et al., 2020, Huseynov & Özkan Yıldırım; 2019)، صنعت بانکی (Bekamiri et al., 2020) و حل مشکلات پیچیده‌ای در صنایع دیگری نظیر بهداشت و درمان، سرمایه‌گذاری و بورس، و بیمه استفاده شده است. اما نقطه اشتراک استفاده از یادگیری ماشین در صنایع مختلف مدیریت روابط با مشتریان از طریق بخش‌بندی و شناخت دقیق مشتریان است.

در روابط با مشتری، جمع‌آوری و استفاده ماهرانه از اطلاعات مشتریان، ترجیحات و علایق آن‌ها بسیار مهم است. شناخت نیازها و انتظارات مشتریان و ایجاد پروفایل مشتری به سازمان‌هایی که هدف آن‌ها انطباق محصولات و خدمات با انتظارات مشتری به بهترین شکل ممکن است، کمک می‌کند (Jędrzejczyk, 2021). دانش بیشتر درباره مشتریان این امکان را به سازمان می‌دهد که فعالیت‌های بازاریابی و طراحی محصول خود را برای یک بخش مشتری خاص سفارشی کنند. دسته‌بندی صحیح و مؤثر مشتریان باعث بهبود سیاست‌های بازاریابی و استراتژی‌های سازمان می‌شود (Kovacs et al., 2021). به همین دلیل، این شناسایی در صنایع مختلف بر اساس داده‌ها، متغیرها و الگوریتم‌ها و روش‌های مختلف صورت پذیرفته است. یکی از مؤثرترین ابزارها برای درک انگیزه و رفتار مصرف‌کنندگان، تقسیم‌بندی مشتریان با توجه به فاکتورها یا ویژگی‌های رفتاری، جمعیتی و سایر فاکتورهای مؤثر در صنعت آن، خرده‌فروشی است (D Arli, 2017). بنابراین، در ادامه و با توجه به هدف و سؤالات پژوهش به بررسی پیشینه پژوهشی شناسایی مشتریان در صنعت بانکداری پرداخته شده است:

«کواکس، کو و عاصمی» در مقاله‌ای با استفاده از یک روش دو-مرحله‌ای و از طریق بررسی همزمان فاکتورهای دسته‌ای و عددی به شناسایی الگوهای سرمایه‌گذاری مشتریان

پرداخته‌اند. متغیرهای تأثیرگذار بر سرمایه‌گذاری افراد می‌توانند عوامل محیطی و عوامل شخصی باشند. از جمله عوامل محیطی می‌توان به ویژگی‌های اقتصادی، وضعیت جامعه و منابع در دسترس اشاره نمود و از جمله فاکتورهای شخصی می‌توان به میزان پس‌انداز افراد، ویژگی‌های رفتاری، دانش سرمایه‌گذاری و دارایی‌های مالی فرد نظیر میزان دارایی‌های ارزی، منزل مسکونی و غیره اشاره کرد (Kovacs et al., 2021). در پژوهشی دیگر «چن» داده‌های واقعی مشتریان بانک تایوان را بررسی و ریسک مشتری بر اساس الگوریتم SVM[1] دسته‌بندی نموده است. در این مقاله ۱۱ فلگ[2] مربوط به احتمال پولشویی یا حمایت تروریستی، ۷ فلگ سرویس و محصول، ۲ فلگ تراکنش‌های غیرمعمول مشخص می‌گردد. شایان ذکر است که در این پژوهش از درخت تصمیم برای نمایش دسته‌ها استفاده شده است (Chen, 2020b). در پژوهشی دیگر که با هدف شناسایی ریسک مشتری انجام شده، از تحلیل داده و یادگیری ماشین برای شناسایی ریسک اعتبار مشتری به‌صورت خودکار استفاده شده است. پژوهشگر برای کاهش ریسک ارائه وام به مشتری، از متدولوژی یک شبکه جدید سلسله‌مراتبی ژنتیکی عمیق از یادگیرندگان برای پیش‌بینی امتیازدهی اعتبار DGHNL[3] که شامل ۲۹ لایه، انواع یادگیرنده‌ها، انواع نرمال‌کننده داده، استخراج ویژگی، تابع «کرنل» و بهبود پارامترها بهره برده است. لازم به ذکر است که در این پژوهش از الگوریتم ژنتیک برای افزایش کارایی استفاده شده است (Plawiak et al., 2020). در پژوهشی که «فیرمن، سانتوسو و جاجادی» و همکارانش با هدف تقسیم‌بندی مشتری کارت اعتباری انجام داده‌اند، پژوهشگر با استفاده از الگوریتم mini batch kmeans و فاکتورهای جمعیت‌شناختی، جغرافیایی و رفتار مالی بر روی داده‌های مشتریان بانک اندونزی، به خوشه‌بندی مشتریان پرداخته است (Firman Pradana Rachman et al., 2021). در پژوهشی دیگر با هدف بهبود پروفایل مشتری و با استفاده از الگوریتم‌های k- means و Fuzzy C-means و improved k-means با ورودی ویژگی‌های جنسیت، وضعیت تأهل، سن و رفتارهای مالی به برچسب‌گذاری داده‌ها پرداخته شده است (Dawood et al., 2019).

سرانجام، اگرچه در مطالعاتی که تاکنون در این زمینه انجام شده، به برخی از جوانب پیاده‌سازی و مقایسه با مدل‌های پایه توجه شده است، اما تاکسونومی در

1. Support Vector Machine (SVM)
2. Flag
3. Deep Genetic Hierarchical Network of Learners (DGHNL)

خصوص الگوریتم‌های شناسایی مشتری و فاکتورهای اصلی در شناخت و ارزش‌گذاری مشتری بانکی وجود ندارد. بنابراین، این پژوهش در صدد است که با بررسی نکات کلیدی مطالعات منتخب، این شکاف پژوهشی را پوشش دهد. در این راستا و در مقایسه با مقالات عنوان‌شده در مرور مبانی نظری، نوآوری پژوهش جاری شامل موارد زیر است: ۱) دسته‌بندی ویژگی‌های مشتریان در ۳ طبقه رفتاری، جمعیت‌شناختی و مالی به‌صورت جامع و با جزئیات زیرمؤلفه‌های آن‌ها (۶۸ متغیر) و اشاره به تعدد ارجاعات در مقالات منتخب، ۲) شناسایی کلیه الگوریتم‌های استفاده‌شده در این حوزه (۳۸ الگوریتم) و دسته‌بندی الگوریتم‌ها در ۱۰ بخش با توجه به ویژگی‌های آن‌ها، و ۳) شناسایی حوزه‌هایی که به بانک در کسب سود و از طرفی به ارتقای تجربه مشتری کمک شایانی خواهد نمود.

۳. روش پژوهش

پژوهش پیش رو از منظر جمع‌آوری داده، فراترکیب بوده و از منظر هدف، در دسته مطالعات کاربردی قرار می‌گیرد. در روش فراترکیب پژوهشگر مطابق با هفت مرحله ارائه‌شده توسط «ساندلوسکی و باروسو»، ابتدا سؤال پژوهش را مشخص می‌کند و با بررسی ادبیات حوزه علمی و موضوع مدنظر به تحلیل یافته‌ها می‌پردازد. فراترکیب یکی از ۴ زیرمجموعه معرفی‌شده برای حوزه فرامطالعه است. فرامطالعه شامل ۴ روش فراترکیب[1]، فراتحلیل[2]، فراروش[3]، و فرانظریه[4] است. فراترکیب یک رویکرد هدفمند و منسجم برای تجزیه و تحلیل داده‌ها در مطالعات کیفی بوده و در واقع، فرایندی است که محققان را قادر می‌سازد که یک سؤال تحقیقی خاص را شناسایی کرده و سپس، شواهد کیفی را برای پاسخ به سؤال تحقیق جست‌وجو، انتخاب، ارزیابی، خلاصه‌سازی و ترکیب کنند (E Erwin et al., 2011). فراترکیب داده‌های کیفی را گردآوری می‌کند تا تفسیر جدیدی از زمینه تحقیق شکل دهد (Atkins et al., 2008). فراترکیب بر خلاف فراتحلیل که بر داده‌های کمی و رویکردهای آماری تأکید دارد، متمرکز بر مطالعات کیفی و تفسیر و تحلیل عمیق آن‌ها به جهت فهم عمیق‌تر است. روش پرکاربرد بعدی، فراتحلیل است که یک روش آماری است و برای ترکیب نتایج حاصل از مطالعات استفاده می‌شود (Møller & Myles,

1. meta-synthesis
2. meta-analysis
3. meta-study
4. meta-theory

2016). روش بعدی فراروش است که به بررسی روش پژوهش مطالعات قبلی می‌پردازد، و روش آخر فرانظریه است که نظریه‌های مطالعات گذشته را مورد تحلیل و بررسی قرار می‌دهد.

«ساندلوسکی و باروسو» برای انجام پژوهش فراترکیب روش هفت-مرحله‌ای را پیشنهاد دادند (Sandelowski & Barroso, 2006) در این پژوهش نیز با توجه به سؤالات مشخص‌شده از این فرایند استفاده شده است. محقق با بررسی ادبیات حوزه یادگیری ماشین و کاربرد آن در حوزه شناسایی و ارزش‌گذاری مشتری به سؤالات پژوهش پاسخ خواهد داد.

شکل ۱. فرایند هفت-مرحله‌ای فراترکیب

لازم به ذکر است که در راستای رعایت معیار وابستگی، در اجرای فرایند تحقیق سعی شده است با مستندسازی و تبیین دقیق مراحل، امکان بررسی و پیاده‌سازی مجدد فرایند برای سایر پژوهشگران مهیا گردد.

از آنجا که در مرور نظام‌مند، اسناد، اساس خروجی پژوهش خواهد بود، بنابراین، به‌منظور

تعیین روایی، از ابزار ارزیابی حیاتی[1] استفاده می‌شود. نحوه استفاده از این ابزار در بخش کنترل کیفیت اعلام خواهد شد. جهت بررسی پایایی نیز از روش توافق بین دو کدگذار استفاده خواهد شد؛ بدین صورت که افزون بر پژوهشگر، که اقدام به کدگذاری اولیه نموده، پژوهشگری دیگر همان متن را بدون اطلاع از کدهای اولیه و به‌صورت جداگانه کدگذاری خواهد کرد. در صورتی که کدهای این دو نفر تا درصد مشخصی اجماع داشته باشد، پایایی مورد تائید خواهد بود.

٤. مراحل انجام پژوهش

بر اساس مراحل ذکرشده در بخش قبل، پژوهش با طرح سؤال زیر آغاز خواهد شد.

٤-١. سؤال پژوهش

در روش فراترکیب به‌دلیل اینکه پژوهشگر به‌صورت اکتشافی عمل می‌کند، بنابراین، به‌دنبال سؤالات از جنس «چه چیزی؟» است. در پژوهش پیش رو به‌دنبال تاکسونومی شناسایی مشتری با رویکرد یادگیری ماشین هستیم. بنابراین، دو سؤال پژوهش به‌شرح زیر است:

- متغیرهای ورودی و الگوریتم‌های یادگیری ماشین جهت شناسایی مشتریان صنعت بانکی کدام است؟
- شناسایی مشتریان در صنعت بانکی با چه اهدافی انجام می‌گردد؟

٤-٢. مرور نظام‌مند مبانی نظری

در ادامه، توسعه کلیدواژه‌های پژوهش بر اساس سؤال انجام گرفت و بر همین اساس پایگاه داده‌ها به‌منظور شناسایی مقالات مرتبط با موضوع پژوهش و بر اساس معیارهای شمول مورد بررسی قرار گرفتند. بر اساس فرایند مرور نظام‌مند پایگاه‌های داده «وب‌آوساینس»[2] و «اسکوپوس»[3] با استفاده از تکنیک «جست‌وجوی پیشرفته» در بخش عنوان، چکیده و کلیدواژه مورد جست‌وجو قرار گرفتند. «وب‌آوساینس» پیشروترین پلتفرم جست‌وجوی استناد علمی و اطلاعات تحلیلی در جهان است (Li et al., 2018). این ابزار، هم به‌عنوان یک ابزار تحقیقاتی برای پشتیبانی از مجموعه وسیعی از وظایف علمی در حوزه‌های مختلف دانش و هم به‌عنوان مجموعه داده‌ای برای مطالعات داده در مقیاس

1. critical appraisals skills programme (CASP)
2. Web OF Science
3. Scopus

بـزرگ اسـتفاده می‌شـود و در هـزاران پژوهـش علمـی از آن اسـتفاده شـده اسـت. همچنیـن از «اسـکوپوس» به‌عنـوان مرجع تکمیلـی اسـتفاده می‌گـردد. در ادامـه و بـر اسـاس کلیـدواژه متنـوع مـورد جسـت‌وجو و بررسـی ترکیـب آن‌هـا کـه در جـدول ۱، اعـلام شـده، اسـناد اسـتخراج گردیـد. لازم بـه ذکـر اسـت کـه بـرای اسـتخراج مقـالات معتبـر و مرتبـط بـا موضـوع پژوهـش، جسـت‌وجو در بـازه زمانـی ۲۰۲۲-۲۰۱۶ و در دو پایـگاه داده مختلـف انجـام گردیـد.

جدول ۱. کلیدواژه‌های مطالعات مرتبط با حوزه یادگیری ماشین

فارسی	انگلیسی
یادگیری ماشین	Machine learning
الگوریتم‌های خوشه‌بندی	Cluster algorithms
الگوریتم‌های دسته‌بندی	Classification algorithms
ارزش مشتری	Customer value
تجربه مشتری	Customer experience
بخش‌بندی مشتری	Customer segmentation
مدیریت مشتری	Customer management
پروفایل مشتری	Customer profile
مشتری	Customer
بانک	Bank

۴-۳. غربالگری کیفی

در گام سـوم و به‌منظـور غربـال اسـناد کـه بـا کلیدواژه‌هـای اعلام‌شـده در جـدول ۱، به‌دسـت آمده‌انـد (۷۱۵ مقالـه)، ابتـدا عنـوان و کلیدواژه‌هـای اسـناد، منبـع انتشـار و همچنیـن سـاختار آن‌هـا بررسـی و در نهایـت، اسـناد غیرمعتبـر و غیرمرتبـط بـا حـوزه پژوهـش حـذف گردیدنـد و سـپس، مقـالات باقی‌مانـده بـا حـذف مقـالات تکـراری و بـا مطالعـه کامـل محتـوا و بـا اسـتفاده از روش CASP مـورد بررسـی اعتبـار کیفـی قـرار گرفتنـد. در جـدول زیـر معیارهـای شـمول مقـالات شـرح داده شـده اسـت:

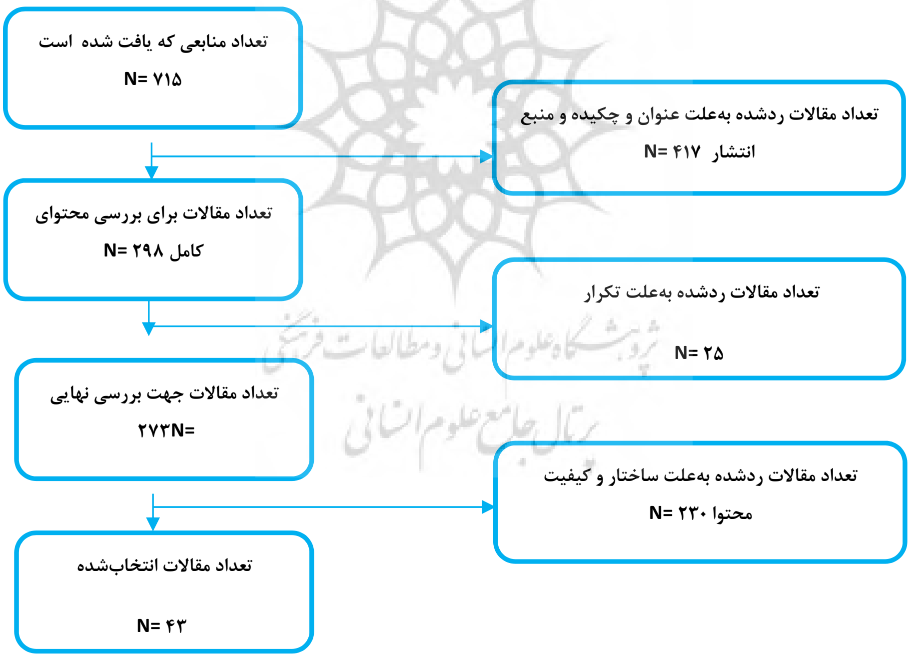
جدول ۲. معیارهای شمول و عدم شمول مقالات

معیارهای گزینش	توصیف و راهنماها	نتایج
معیار شمول	مقالاتی که: در جست‌وجوی پیشرفته پایگاه‌های اطلاعاتی «وب‌آوساینس» و «اسکوپوس» شناسایی شدند. موضوع: مقالات متمرکز بر شناسایی، ارزش‌گذاری و دسته‌بندی مشتریان بانکی با رویکرد یادگیری ماشین بود. زبان: مقالاتی که به زبان انگلیسی نوشته شده بودند. دوره زمانی: مقالاتی که در بازه زمانی ابتدای ۲۰۱۶ تا اواسط ۲۰۲۲ منتشر شده بودند. نوع: مقالات پژوهشی که در مجلات معتبر چاپ شده بودند. زمینه: سیستم‌های اطلاعاتی، مدیریت، علوم کامپیوتر، علوم مدیریت بازرگانی، هوش مصنوعی و یادگیری ماشین	
کلیدواژه‌ها	شامل: یادگیری ماشین، الگوریتم‌های خوشه‌بندی، الگوریتم‌های دسته‌بندی، ارزش مشتری، تجربه مشتری، بخش‌بندی مشتری، مدیریت مشتری، پروفایل مشتری، مشتری، بانک	
جست‌وجوی کلیدواژه‌ها	پایگاه‌های داده آنلاین با کلیدواژه‌های فوق جست‌وجو شد.	«وب‌آوساینس» و «اسکوپوس» ۷۱۵ مقاله
مقالات شناسایی‌شده	بررسی عنوان، چکیده و منبع انتشار	۲۹۸ مقاله
ترکیب	تعداد مقالات تکراری در «وب‌آوساینس» و «اسکوپوس»	۲۵ مقاله
نمونه انتخابی برای تحلیل		۲۷۳ مقاله
ارزیابی محتوا	ارزیابی مرحله دوم بر اساس ارزیابی کیفی مقالات	۴۳ مقاله
نمونه نهایی		۴۳ مقاله

ابزار CASP با استفاده از ۱۰ شرط، کیفیت مقالات را مورد ارزیابی قرار می‌دهد. به این ترتیب که هر یک از این شرایط، امتیازی بین ۱ تا ۵ می‌گیرد. مقالاتی که مجموع امتیاز آن‌ها ۲۵ یا بالاتر باشد، از لحاظ کیفی مورد تائید است. ۱۰ شرح عنوان‌شده در معیارهای برنامه مهارت ارزیابی حیاتی شامل تناسب اهداف مقاله مورد بررسی با اهداف پژوهش، به‌روز بودن پژوهشی مقاله، طرح مطرح‌شده در مقاله، روش نمونه‌گیری، روش و کیفیت جمع‌آوری داده‌ها، میزان انعکاس‌پذیری امکان بسط‌دادن نتایج و دستاوردها، میزان و نحوه رعایت نکات اخلاقی رایج در زمینه تدوین متون پژوهشی، میزان دقت در زمینه تجزیه و تحلیل داده‌ها، وضوح بیان در ارائه یافته‌ها، و ارزش کلی مقاله مورد بررسی است.

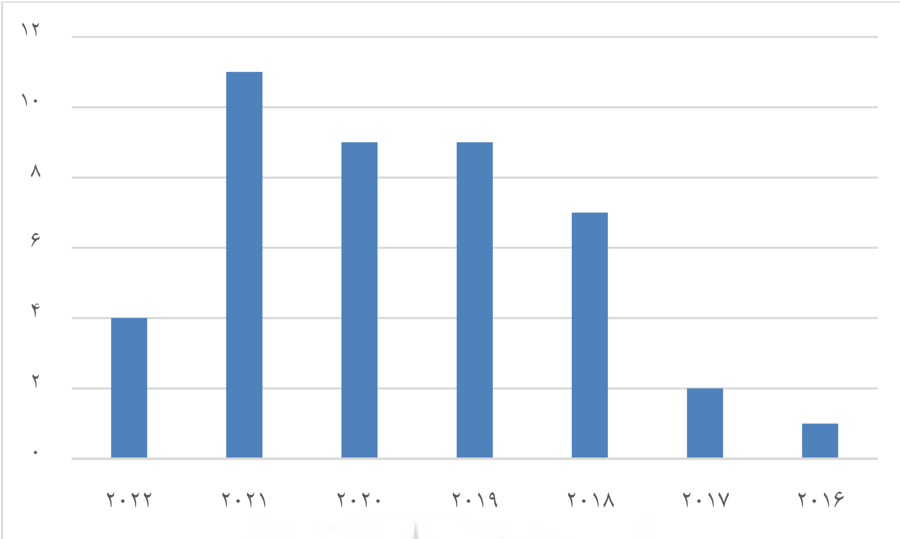
با توجه به وجود ۱۰ ویژگی که حداکثر امتیاز هر ویژگی ۵ است، بنابراین بیشترین نمره‌ای که هر مقاله بر اساس مقیاس CASP کسب می‌کند، ۵۰ است. دسته‌بندی مقالات بر اساس روش CASP شامل خیلی خوب (۵۰ -۴۱)، خوب (۴۰-۳۱)، متوسط (۳۰-۲۱)، ضعیف (۲۰-۱۱)، خیلی ضعیف (۱۰- ۰) است. فرایند انتخاب مقالات مرتبط در شکل ۲، مشخص شده است. شایان ذکر است که در این پژوهش مقالاتی که بین ۴۰ تا ۵۰ امتیاز کسب نمودند، استفاده شده است.

در نهایت، تعداد ۴۳ مقاله با کیفیت بالاتری تشخیص داده شد و با استفاده از روش کدگذاری و بهره‌گیری از نرم‌افزار MaxQda بررسی گردید. واژگان دارای بیشترین تکرار عبارت‌اند از: داده، مشتری، اعتبار، مدل‌ها، یادگیری، ویژگی‌ها، تکنیک، دسته‌بندی و خوشه‌بندی. در واقع، در حوزه مورد مطالعه تکرار این واژه‌ها به معنای اهمیت و تأثیر این موارد در یادگیری ماشین است.



شکل ۲. فرایند انتخاب مقالات

همچنین، در شکل ۳، روند انتشار مقالات ارائه شده است که نشان‌دهنده توجه جامعه پژوهش به شناخت مشتری مبتنی بر داده‌های واقعی است.



شکل ۳. تعداد مقالات به تفکیک سال

از لحاظ روش‌شناسی، اغلب پژوهش‌های انجام‌شده از روش کمی و عمدتاً با پیاده‌سازی الگوریتم‌های یادگیری ماشین با ابزارهای مختلف نظیر زبان برنامه‌نویسی «پایتون» و «متلب» همراه بوده است.

۴-۴. استخراج اطلاعات متون انتخابی

مشخصات مقالات به همراه اطلاعات کلیدی در جداول آتی درج گردیده است. این جداول می‌توانند راهنمای محققان آینده و همچنین مدیران فنی و اجرایی بانک‌ها، جهت انتخاب الگوریتم‌های یادگیری ماشین و فاکتورهای شناسایی مشتری مطابق با نیازمندی و هدف اجرایی ایشان باشند.

۴-۵. تجزیه و تحلیل و ترکیب یافته‌ها

پس از انتخاب مقالات با معیارهای مورد نظر، اسناد منتخب بررسی و کدهای مرتبط شناسایی و استخراج گردید. برای انجام کدگذاری فراترکیب، از رویکرد کدگذاری سه‌مرحله‌ای استفاده شد: کدگذاری باز، محوری، و انتخابی. کدهای دارای ماهیت مشابه، ذیل یک دسته قرار گرفتند و مفاهیم و مضامین خُرد را تشکیل دادند و در ادامه نیز مفاهیم مشابه یک مقوله را ایجاد کردند.

جدول ۳. مقوله‌ها و مفاهیم مرتبط

مقوله‌های اصلی	مفهوم ۱: اهداف شناسایی مشتری	مفهوم ۲: فاکتورهای مشتری	مفهوم ۳: الگوریتم‌های یادگیری ماشین
مفاهیم مرتبط	اهداف سازمانی بینش نسبت به مشتری تعیین ارزش و طول عمر مشتری مدیریت محصول	فاکتورهای رفتاری فاکتورهای مالی فاکتورهای جمعیت‌شناختی	Ensemble Probabilistic Neural Networks Rule System Regression Bayesian Decision Tree Dimensionality Reduction Instanced Based Clustering

٤-٦. کنترل کیفیت

بـا عنایـت بـه اینکـه کدگـذاری انجام‌شـده در پژوهش‌هـای فراترکیـب اسـاس مضامیـن و مقوله‌هـا و به‌تبـع آن پایـهٔ تحلیـل، مدل‌هـا یـا خروجـی پژوهـش اسـت، بنابرایـن، بـرای بررسـی پایایـی پژوهـش از ۱ نفـر درخواسـت گردیـد کـه اسـناد را به‌صـورت مجـزا کدگـذاری نمایـد. در صورتـی کـه نظـرات ایـن ۲ نفـر (نویسـنده و نفـر دوم) همگـرا باشـد، پایایـی پژوهـش تائیـد خواهـد شـد. بـرای ارزیابـی همگـرا بـودن کدهـای احصاشـده از ضریـب «کاپـا» اسـتفاده می‌شـود. زمانـی کـه ضریـب کمتـر از ۰/۲ بیانگـر توافـق ضعیـف، بیـن ۰/۲ و ۰/۴ متوسـط، بیـن ۰/۴ و ۰/۶ بـه نسـبت زیـاد، ۰/۶ و ۰/۸ زیـاد و بیشـتر از ۰/۸ باشـد، همگرایـی تقریبـاً کامـل اسـت. لازم بـه ذکـر اسـت کـه بـرای پژوهـش پیـش رو ضریـب «کاپـا» ۷۳/۶۸ درصـد محاسـبه شـده اسـت.

جدول ٤. کدهای احصاشده

مقوله	مفاهیم مرتبط	کدهای باز (مفاهیم و مضامین خرد)	منابع
اهداف شناسایی مشتری	اهداف سازمانی	مدیریت ارتباط با مشتری (۱)، تعیین احتمال جذب مشتری به سرویس (۲)، تعیین استراتژی وام‌دهی (۱)، شناسایی کلاهبرداری‌ها (۱)، تعیین استراتژی بازاریابی (۱)	Al-Rubaiee et al. (2018); Safarkhani & Moro (2021); A. Eltahir et al. (2022); Yanik & Elmorsy (2019); Chopra & Bhilare (2018)

مقوله	مفاهیم مرتبط	کدهای باز (مفاهیم و مضامین خرد)	منابع
	بینش نسبت به مشتری	پیش‌بینی رفتار مشتری (۱)، بهبود پروفایل رفتاری مشتری (۱)، پیش‌بینی رویگردانی مشتری (۴)، دسته‌بندی مشتری (۳)، شناخت بهتر مشتری (۳)، مدل‌سازی رفتاری کارت اعتباری مشتری (۱)، شناسایی الگوی سرمایه‌گذاری (۱)	Abbasimehr & Shabani (2021); Dawood et al. (2019); De Lima Lemos et al. (2022); Domingos et al. (2021); Bharathi et al. (2022); Long et al. (2019); K. Borna et al. (2019); Motevali et al. (2019); Xi Song et al. (2021); Al-Rubaiee et al. (2018); Firman Pradana Rachman et al. (2021); Alireza Sheikh et al. (2019); Ala'raj et al. (2021); Kovacs et al. (2021)
	تعیین ارزش و طول عمر مشتری	ارزیابی ارتباط استراتژی و طول عمر مشتری (۱)، ارائه مدل‌ها جامع (۱) CLV، تعیین ارزش مشتری (۳)	Plawiak et al. (2020); Hajipour & Esfahani (2019); Estrella-Ramon et al. (2017); Mosavi & Afsar (2018)
	مدیریت محصول	درک احساس مشتری در استفاده از سرویس (۱)، دسته‌بندی محصولات (۱)، پیش‌بینی استفاده از محصول (۱)، ترغیب به توصیه محصول (۱)	Calvo-Porral & Levy-Mangin (2020); Plawiak et al. (2020); Urkup et al. (2018)
	شناسایی ریسک مشتری	شناسایی، پیش‌بینی و جلوگیری (۷)، اجتناب از مشکلات اعطای اعتبار (۱)، پیش‌بینی بازپرداخت وام (۲)، اعتبارسنجی مشتری (۳)	Dayu Xu et al. (2018); Furio. Camillo & Liberati (2018); Ala'raj & Abbod (2016); Chen (2020a); Melo et al. (2020); Pandey et al. (2021); Mhlanga (2021); Çığşar & Ünal (2019); Mancisidor et al. (2021); Plawiak et al. (2020); Livieris et al. (2018)

مقوله	مفاهیم مرتبط	کدهای باز (مفاهیم و مضامین خرد)	منابع
فاکتورهای مشتری	فاکتورهای رفتاری	احساسات مشتری در زمان استفاده از سرویس (۱)، هدف افتتاح حساب (۱)، کشور با بیشترین معامله (۱)، معاملات برون‌مرزی (۱)، هدف از افتتاح حساب برای اتباع خارجی (۱)، ریسک مشتری (۱)، دارایی (۱)، رفتار خرید جانبی (۱)، مشتری فعال (۱)، وضعیت اعتبار مشتری (۳)، هدف وام (۴)، تکرار خرید (۱)، بستن حساب (۱)، عدم خرید (۱)، استفاده از سرویس‌های بانکی (وام) (۲)، حساب اعتباری (۴)، حساب پس‌انداز (۶)، رهن مسکن (۱)، وام‌های خرد (۳)، صندوق سرمایه‌گذاری (۱)، کارت نقدی (۱)، وام مسکن (۳)، وضعیت حساب بانکی (۲)، برنامه بازنشستگی (۱)، کارت اعتباری (۳))، تعداد محصولات مورد استفاده (۳)، استفاده از سرویس اعتباری اتومات (۱)، شکایت مشتری (۱)، درخواست انتقال اعتبار (۱)، استفاده از سرویس بیمه (۴)، دریافت رسید حقوق (۱)	Calvo-Porral & Levy-Mangin (2020); Dawood et al. (2019); Chen (2020a); Domingos et al. (2021); Motevali et al. (2019); Ashofteh & Bravo (2021); M. Torrens & A. Tabakovic (2022); Alam et al. (2020); Bharathi et al. (2022); De Lima Lemos et al. (2022); Seidlova et al. (2019)
	فاکتورهای مالی	هزینه تراکنش مشتری در کانال‌های مختلف (۱)، میانگین موجودی (۲)، میانگین دارایی (۱)، تعداد حساب (۱)، خریدهای مصرفی (لوازم تحریر) (۱)، غذا (۱)، بیمه (۱)، لباس (۱)، دکور و سلامت، خرید خرد، هزینه‌های اقامت، طلا، تحصیلات، میانگین تراکنش‌ها (۱)، نرخ سود دارایی‌ها (۱)، قبض (۴)، موجودی حساب (۳)، نرخ سود حساب (۱)، مدت‌زمان وام (۲)، ارزیابی وام (۱)، تعداد پرداخت‌ها در ماه (۱)، نرخ سود وام (۲)، تعداد وام‌ها (۲)، اعتبار در گردش (۱)، حجم پرداخت‌های قبلی (۲)، حجم خرید (۱)، سودآفرینی مشتری (۲)، سایر بدهی‌ها و تضامین (۴)، نرخ اقساط بر اساس درآمد قابل تصرف (۴)، سپرده پس‌انداز یا اوراق قرضه (۱)، میزان اعتبار (۶)، تاریخچه اعتبار (۲)، وضعیت حساب جاری (۱)، تراکنش‌های کارت اعتباری (۱)، وام‌های مسکن و ضروری (۲)، تعداد تراکنش (۳)، عدم بازپرداخت وام (۸)، RFM (۴)، LRFM(۱)	Chen (2020a); Plawiak et al. (2020); Estrella-Ramon et al., (2017); Motevali et al. (2019); Furio. Camillo & Liberati(2018); Ala'raj et al.(2021); Çığşar & Ünal, (2019); Dawood et al.(2019); Alam et al.(2020); Kovacs et al. (2021); Domingos et al.(2021); Ashofteh & Bravo (2021); Seidlova et al.(2019); Al-Qerem et al.(2020); Pandey et al.(2021); M. Torrens & A. Tabakovic(2022); De Lima Lemos et al.(2022); Firman Pradana Rachman et al.(2021); A. Eltahir et al.(2022); Hajipour & Esfahani(2019); Dayu Xu et al.(2018); Chopra & Bhilare (2018); Sivasankar et al.(2020); Safarkhani & Moro(2021)

مقوله	مفاهیم مرتبط	کدهای باز (مفاهیم و مضامین خرد)	منابع
	فاکتورهای جمعیت‌شناختی	صنعت (۱)، نوع شهر محل سکونت (۱)، کشور (۱)، وضعیت سلامت (۱)، تعداد سال اشتغال (۳)، تابعیت خارجی (۲)، تلفن (۳)، افراد تحت تکفل (۱)، وضعیت سکونت (۶)، آدرس (۷)، جنسیت (۱۱)، درآمد (۹)، تحصیلات (۱۰)، وضعیت تأهل (۱۰)، شغل (۱۰)، سن (۱۴)	A. Eltahir et al. (2022); M. Torrens & A. Tabakovic (2022); Bharathi et al. (2022); Firman Pradana Rachman et al. (2021); Pandey et al. (2021); Ala'raj et al. (2021); Ashofteh & Bravo (2021); Safarkhani & Moro (2021); Mancisidor et al. (2021); Chen (2020a); Plawiak et al. (2020); Al-Qerem et al. (2020); Dawood et al. (2019); Çığşar & Ünal (2019); Yanik & Elmorsy, (2019); Seidlova et al. (2019); Dayu Xu et al. (2018); Chopra & Bhilare (2018); Furio. Camillo & Liberati (2018); Estrella-Ramon et al. (2017)
انواع الگوریتم‌ها	Clustering	heirarchical clustering (1) -improved kmeans (1)- Fuzzy c-means (1)	Dawood et al. (2019); Kovacs et al. (2021)
	Instanced Based	KNN- (9) SOM (2)	Kovacs et al. (2021); De Lima Lemos et al. (2022); Bharathi et al. (2022); Domingos et al. (2021); Plawiak et al. (2020); Melo et al. (2020); Sivasankar et al. (2020); Alam et al. (2020); Yanik & Elmorsy (2019); Long et al. (2019); Livieris et al. (2018)
	Ensemble	Random Forest (18), Gradient Boosting (4), extra tree (3), AdaBoost (4), extreme gradient boosting (2), bagging (4), stacking (1), boosting (1), boosted trees (1), bagged neural network (1)	De Lima Lemos et al. (2022); M. Torrens & A. Tabakovic (2022); Bharathi et al. (2022); Pandey et al. (2021); Ala'raj et al. (2021); Ashofteh & Bravo (2021); Chen (2020a); Melo et al. (2020); Sivasankar et al. (2020); Al-Qerem et al. (2020); Alam et al. (2020); Çığşar & Ünal (2019); Long et al. (2019); Seidlova et al. (2019); Urkup et al. (2018); Dayu Xu et al. (2018); Lu et al. (2017)

مقوله	مفاهیم مرتبط	کدهای باز (مفاهیم و مضامین خرد)	منابع
	Dimenality Reduction	Linear discriminant analysis (1)	Furio. Camillo & Liberati (2018)
	Decision tree	Decision tree (12), C5.0 (3), J48 (1)	De Lima Lemos et al. (2022); M. Torrens & A. Tabakovic (2022); Bharathi et al. (2022); Ashofteh & Bravo (2021); Safarkhani & Moro (2021); Chen (2020a); Sivasankar et al. (2020); Al-Qerem et al. (2020); Long et al. (2019); Urkup et al. (2018); Dayu Xu et al. (2018); Chopra & Bhilare (2018)
	Baysian	Naive Bayes (9), Heirarchical Bayesian model (1), bernouliNB (1), GussianNB (1), BayesNet (1)	Bharathi et al. (2022); Ashofteh & Bravo (2021); Safarkhani & Moro (2021); Sivasankar et al. (2020); Al-Qerem et al. (2020); Çığşar & Ünal (2019); Long et al. (2019); Dayu Xu et al. (2018); Al-Rubaiee et al., (2018); Livieris et al., (2018); Estrella-Ramon et al., (2017)
	Regression	Logistic regression (16), Linear Regression (1)	A. Eltahir et al. (2022), De Lima Lemos et al. (2022), M. Torrens & A. Tabakovic (2022); Bharathi et al. (2022); Ala'raj et al. (2021); Ashofteh & Bravo (2021); Safarkhani & Moro (2021); Melo et al. (2020); Alam et al. (2020), Çığşar & Ünal (2019); Urkup et al. (2018); Furio. Camillo & Liberati (2018); Estrella-Ramon et al. (2017)
	Regularization	elastic net (1), Stochastic gradient descent classifier (SGD) (1)	Urkup et al. (2018); De Lima Lemos et al. (2022)
	Neural Network	RMkNN (1), RNN (1), MLP (3), LSTM (1), Neural network (3), Deep neural network (1), deep belief network (1), probabilistic neural networks (1), ant colony (1)	Ala'raj et al. (2021); Ashofteh & Bravo (2021); Domingos et al. (2021); Plawiak et al. (2020); Melo et al (2020); Dawood et al,. (2019), Çığşar & Ünal (2019); Seidlova et al. (2019), Dayu Xu et al. (2018); Livieris et al. (2018)

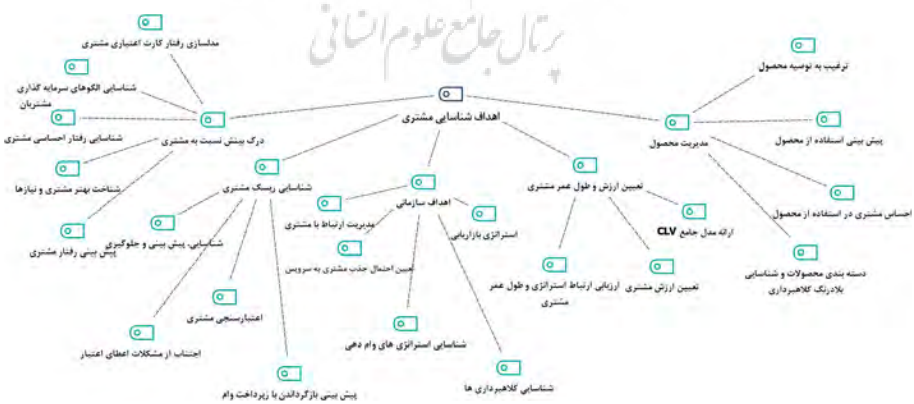
مقوله	مفاهیم مرتبط	کدهای باز (مفاهیم و مضامین خرد)	منابع
	probabilistic	SVM (16), Fuzzy system (1)	De Lima Lemos et al. (2022), Bharathi et al. (2022); Ala'raj et al. (2021); Ashofteh & Bravo (2021); Chen (2020a); Plawiak et al. (2020); Melo et al. (2020); Sivasankar et al. (2020); Long et al. (2019); Dayu Xu et al. (2018); Al-Rubaiee et al. (2018)

٤-٧. یافته‌ها

با عنایت به مراحل تشریح‌شده در بخش‌های پیشین پژوهش، کدهای استخراج‌شده در زیرمجموعه ۳ مقوله «اهداف شناسایی مشتری»، «فاکتورهای شناسایی مشتری»، و «الگوریتم‌های یادگیری ماشین» احصا شده است و هر یک در مدل نظری با ابزار MaxQda رسم و تشریح می‌شود:

٤-٧-١. مقوله اهداف شناسایی مشتری

نتایج حاصل از کدگذاری مقاله‌های منتخب نشان می‌دهد که در مقوله شناسایی مشتری، اغلب مقالات بر جنبه‌های شناسایی ریسک مشتری تمرکز دارند. همچنین اهداف دیگر، شامل کسب بینش نسبت به مشتری، تعیین ارزش و طول عمر مشتری، مدیریت محصول و اهداف سازمانی است. شکل ۴، که از نرم‌افزار MaxQda به‌دست آمده، مدل نظری کدهای مربوط به اهداف شناسایی مشتری را نشان می‌دهد.



شکل ٤. مدل نظری کدهای مربوط به مقوله اهداف شناسایی مشتری

۱) زمینـه شناسـایی ریسـک مشـتری: اهـداف شناسـایی ریسـک مشـتری شـامل شناسـایی، پیش‌بینی و جلوگیـری، پیش‌بینـی بازپرداخـت وام، اجتنـاب از مشـکلات اعطـای وام، و اعتبارسـنجی مشـتری اسـت. علـت تعـدد مطالعـات در زمینـهٔ ریسـک را می‌تـوان رشـد گسـترده دارایی‌هـای معـوق بانک‌هـا۱ دانسـت کـه بانک‌هـا را بـه سـمت کسـب مزیت‌هـای رقابتـی پایـدار، از جملـه تشـخیص، پیش‌بینـی و پیشـگیری از ریسـک‌های اعتبـاری سـوق داده اسـت (Al-Qerem et al., 2020). از جنبـه دیگـر، طراحـی مـدل بـرای پیش‌بینـی وضعیـت وام مشـتریان مهـم اسـت؛ چـرا کـه می‌توانـد از ضررهـای عظیـم جلوگیـری کنـد. بـه همیـن دلیـل، نه‌تنهـا توانایـی بازپرداخـت وام، بلکـه پرداخـت به‌موقـع آن بـرای بانک‌هـا از اهمیـت بالایـی برخـوردار اسـت. در حقیقـت شناسـایی و دنبال‌کـردن رفتـار مشـتری به‌صـورت فعالانـه در پیش‌بینـی مؤثـر رفتـار مشـتریانی کـه بازپرداخـت وام را متوقـف می‌کننـد، تأثیرگـذار اسـت. ارزیابـی رفتـار مشـتری و توانایـی بازپرداخـت وام‌هـا از امـکان بهره‌بـرداری وام توسـط افـراد بـا اعتبـار پایین‌تـر جلوگیـری می‌کنـد (Mhlanga, 2021).

در پژوهشـی بـا موضـوع کاهـش ریسـک اعتبـاری، ضمـن اشـاره بـه ضـرورت شناسـایی ریسـک مشـتری بـه گسـتردگی داده‌هـای بـدون سـاختار و نامتعـادل ایـن حـوزه پرداختـه شـده و الگوریتم‌هـای پایـه مختلـف را بـر روی داده‌هـای نمونـه بررسـی می‌کنـد (Pandey et al., 2021). بـرای شناسـایی، پیش‌بینـی و جلوگیـری از ریسـک مشـتری به‌طـور معمـول، فاکتورهایـی نظیـر حسـاب‌های مشـتری، مدت‌زمـان، تاریخچـه اعتبـار، هـدف، مبلـغ اعتبـار، حسـاب پس‌انـداز یـا اوراق قرضـه، اشـتغال جـاری، نـرخ اقسـاط بـر حسـب درصـد، درآمـد قابـل تصـرف، جنسـیت و سـایر بدهی‌هـا و وام‌هـای مشـتری مـورد بررسـی قـرار می‌گیرنـد.

۲) کسـب بینـش نسـبت بـه مشـتری: دومیـن حـوزهٔ پرتکـرار کـه مقـالات در مقولـه اهـداف شناسـایی مشـتری بـه آن پرداخته‌انـد، موضـوع کسـب بینـش مشـتری اسـت. ایـن حـوزه شـامل مدل‌سـازی رفتـار کارت اعتبـاری، شناسـایی الگـوی سـرمایه‌گذاری، شناسـایی رفتـار احساسـی، بهبـود پروفایـل رفتـاری، شـناخت دقیق‌تـر مشـتریان اسـت. علـت توجـه بـه ایـن حـوزه، لـزوم ارائـه تجربـه شخصی‌سازی‌شـده بـر اسـاس رفتـار و ویژگی‌هـای جمعیت‌شـناختی مشـتری اسـت. در واقـع، امـروزه، تمـام صنایـع از اهمیـت ارائـه تجربـه مشـتری عالـی و ایجـاد لحظـات نـاب بـر پایـه داده‌هـای مشـتری آگاه هسـتند. در پژوهشـی دیگـر، شناسـایی الگوهـای مواجهـه

1. non-performing assets (NPAs)

مشـتری بـا نقـاط تمـاس سـازمان بـر اسـاس دفعـات ارتبـاط مشـتری بـا نقـاط و درک رابطـه آن‌هـا بـا وفـاداری مشـتری مـورد بررسـی گرفتـه اسـت (Ieva & Ziliani, 2018). همچنیـن موضـوع پیش‌بینـی رویگردانـی مشـتری بـا توجـه بـه هزینه‌هـای جـذب مشـتری جدیـد و تأثیـر تبلیغـات دهـان بـه دهـان مشـتری ناراضـی، بیـش از پیـش مـورد توجـه پژوهشـگران حـوزه یادگیـری ماشـین و تجربـه مشـتری قـرار گرفتـه اسـت (Cheng et al., 2019; Coser et al., 2020; De Lima Lemos et al., 2022; Keramati et al., 2016; Milosevic et al., 2017; Shirazi & Mohammadi, 2019).

در یکـی از مقـالات بخش‌بنـدی یکـی از متداول‌تریـن و در عیـن حـال، مؤثرتریـن رویکردهـا جهـت ایجـاد پروفایـل مناسـب، شـناخت دقیق‌تـر مشـتری و مدیریـت مشـتریان معرفـی شـده اسـت (Alireza Sheikh et al., 2019). از طـرف دیگـر، بـا رشـد سـریع و عظیـم داده‌هـای مالـی، توسـعه مدل‌هـای امتیازدهـی اعتبـاری مؤثـر بسـیار حیاتـی اسـت (Dawood et al., 2019). توجـه گسـترده بـه ایـن حـوزه نیـز می‌توانـد بـه ایـن دلیـل باشـد کـه هـر سیسـتم بانکـی شـامل مجموعـه داده عظیمـی بـرای تراکنش‌هـای کارت‌هـای اعتبـاری مشـتریان اسـت و بنابرایـن، بانک‌هـا نیـاز بـه پروفایـل مشـتری دارنـد. پروفایـل مشـتریان بانـک از تصمیمـات صادرکننـده در مـورد اینکـه بـه چـه کسـی تسـهیلات بانکـی بدهـد و چـه حـد اعتبـار اعطـا کنـد، پشـتیبانی خواهـد کـرد. پروفایـل مشـتریان همچنیـن بـه بانک‌هـا کمـک می‌کنـد تـا درک بهتـری از مشـتریان بالقـوه و فعلـی خـود داشـته باشـند.

یکـی دیگـر از اهـداف اسـتفاده از الگوریتم‌هـای یادگیـری ماشـین در حـوزه درک بینـش مشـتری، اسـتخراج الگـوی سـرمایه‌گذاری اسـت. بـرای نمونـه، در پژوهشـی بـا اسـتفاده از روش خوشـه‌بندی دو-مرحلـه‌ای الگوهـای سـرمایه‌گذاری مشـتریان بالقـوه بانکـداری خـرد شناسـایی شـده اسـت. ایـن تحقیـق، ۱۵۴۲ پاسـخ دریافت‌شـده در نظرسـنجی سـرمایه‌گذاری آنلایـن را بـا تمرکـز بـر سـؤالاتی کـه بـه ترجیحـات سـرمایه‌گذاری پاسـخ‌دهندگان مرتبـط اسـت، بررسـی کـرده اسـت (Kovacs et al., 2021).

۳) تعییـن ارزش و طـول عمـر مشـتری: ارزش طـول عمـر مشـتری مفهـوم اصلـی بازاریابـی رابطه‌منـد اسـت و به‌طـور فزاینـده‌ای در مقـالات علمـی و کسـب‌وکار بـه آن پرداختـه می‌شـود (N. AbdolvandAmir & A. Albadvi, 2014). بـدون شـک یکـی از ابزارهـای اصلـی بـرای شناسـایی ارزش مشـتریان، معیـار ارزش طـول عمـر مشـتری CLV[1] اسـت. شـامل تمـام عناصـر سـودآوری

1. Customer life time value

مشتری (یعنی درآمدها و هزینه‌ها) و یک معیار آینده‌نگر (یعنی پیش‌بینی‌کننده) است. به‌طور کلی، CLV مبنای خوبی برای ارزیابی ارزش بازار یک شرکت فراهم می‌کند و ثابت شده است که تصمیمات بازاریابی بر اساس این معیار باعث بهبود عملکرد مالی شرکت‌ها می‌شود (Estrella-Ramon et al., 2017). بنابراین، در پژوهش‌هایی نظیر (Hajipour & Esfahani, 2019) به ارائه مدل‌های جدید برای تعریف ارزش‌های مشتریان بر اساس روش‌هایی نظیر(تازگی، تکرار و ارزش پولی) RFM[1] و تقسیم‌بندی مشتریان بانک با استفاده از الگوریتم یادگیری ماشین پرداخته شده است.

۴) اهداف سازمانی: گرچه شناخت مشتری به‌طور کلی باعث بهبود عملکرد سازمان در تمام فرایندها می‌شود، اما در برخی مقالات یادگیری ماشین به‌طور خاص و با تمرکز بر این هدف تدوین شده‌اند. به تیم ریسک توصیه می‌شود که از تکنیک‌های جدید برای دستیابی به دقت بهتر که منجر به استراتژی‌های وام مؤثر می‌شود، استفاده نمایند (Chopra & Bhilare, 2018).

طبق کدگذاری انجام‌شده بر روی مقالات منتخب، حوزه دیگر مورد استفاده در اهداف سازمان، بازاریابی است. در این پژوهش‌ها به ارائه بینشی در مورد تجزیه و تحلیل داده‌ها پرداخته شده و دسته‌ای از مشتریان را که به احتمال زیاد به کمپین‌های بازاریابی پاسخ مثبت می‌دهند، معرفی می‌شود (A. Eltahir et al., 2022; M. Torrens & A. Tabakovic, 2022; Marinakos & Daskalaki, 2017).

۵) مدیریت محصول: بانک‌ها به تقسیم‌بندی محصولات، درک احساس مشتری در زمان استفاده از محصول و ترغیب به توصیه محصول به مشتریان دیگر نیازمند هستند. با شناخت مشتریان از طریق الگوریتم‌های یادگیری ماشین، بانک‌ها می‌توانند به مشتریان خود خدمات بهتری ارائه دهند و اثربخشی را افزایش دهند. برای این منظور از روش‌های داده‌کاوی مختلفی استفاده می‌شود که امکان استخراج الگوها یا دانش‌های ضمنی و ناشناخته را از حجم عظیمی از داده‌ها فراهم می‌کند (Seidlova et al., 2019). یکی از حوزه‌های مورد استقبال مطالعات، بررسی رفتار مشتری (برای نمونه رفتار مالی و تحرک مکانی) و پیش‌بینی استفاده از محصولاتی مثل تسهیلات بانکی است. در حقیقت پیش‌بینی میزان استقبال مشتری از طرح‌ها و محصولات مختلف و درک احساس نسبت به محصول،

1. recency, frequency, monetary

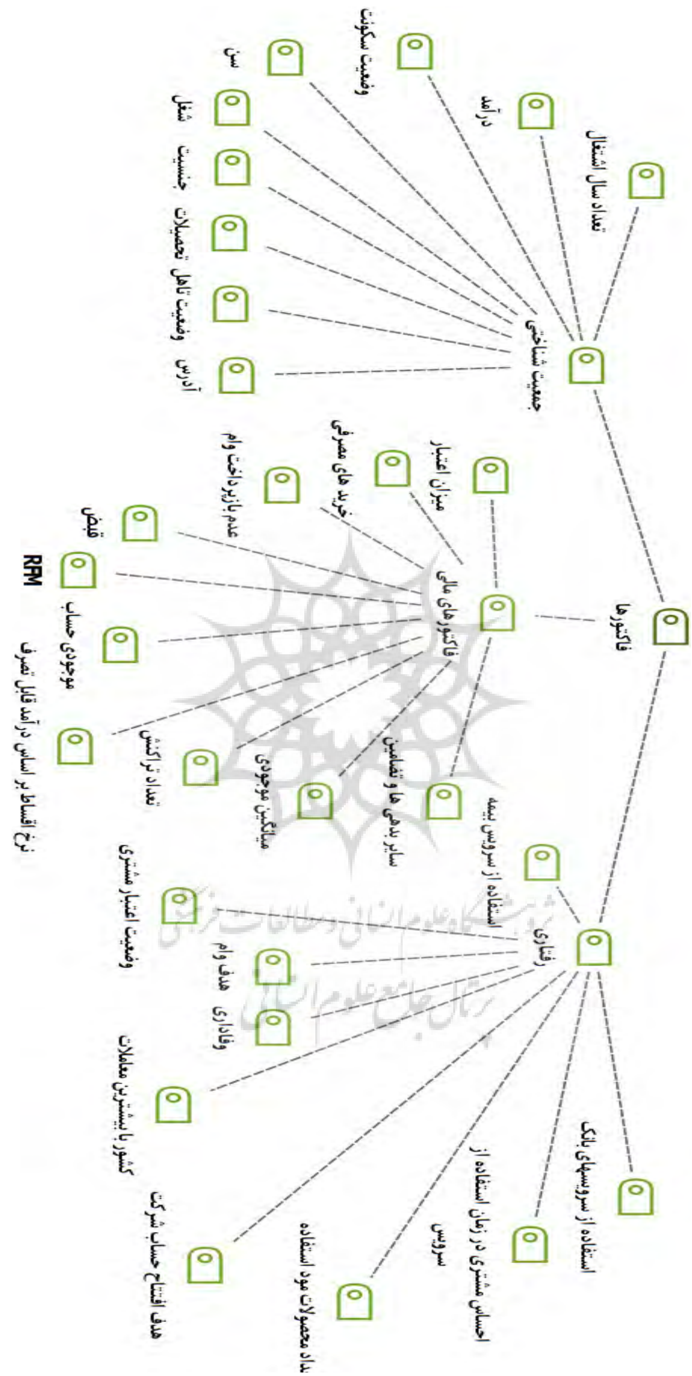
در بهبـود اسـتراتژی‌های طراحـی و تبلیـغ محصـول تأثیرگـذار اسـت (Urkup et al., 2018).

بـا عنایـت بـه بررسـی انجام‌شـده، بیشـترین تکـرار هـدف در مقـالات بررسی‌شـده، مربـوط بـه درک بینـش مشـتری و شناسـایی ریسـک مشـتری بـوده اسـت کـه نشـان‌دهندهٔ اهمیـت شـناخت مشـتری و درک مـدل رفتـاری وی جهـت اهـداف دیگـر نظیـر مدیریـت محصـول اسـت.

۴-۷-۲. مقوله کلان فاکتورهای شناسایی مشتری

نتایـج حاصـل از کدگـذاری مقاله‌هـای منتخـب نشـان می‌دهـد کـه فاکتورهـای مالـی، ویژگی‌هـای جمعیت‌شـناختی و رفتـار مشـتری در مقولـه فاکتورهـای انتخابـی اسـتفاده شـده اسـت. فاکتورهـا در اغلـب مقـالات، جهـت ورود بـه الگوریتم‌هـای یادگیـری ماشـین، به‌صـورت فاکتورهـای ترکیبـی[1] هسـتند. شـکل ۵، کـه از نرم‌افـزار MaxQda به‌دسـت آمـده، مـدل نظـری کدهـای مربـوط بـه فاکتورهـای انتخابـی را نشـان می‌دهـد.

۱. ترکیب فاکتورهای مالی، رفتاری و جمعیت‌شناختی



شکل ۵. مدل نظری کدهای مربوط به مقوله فاکتورها

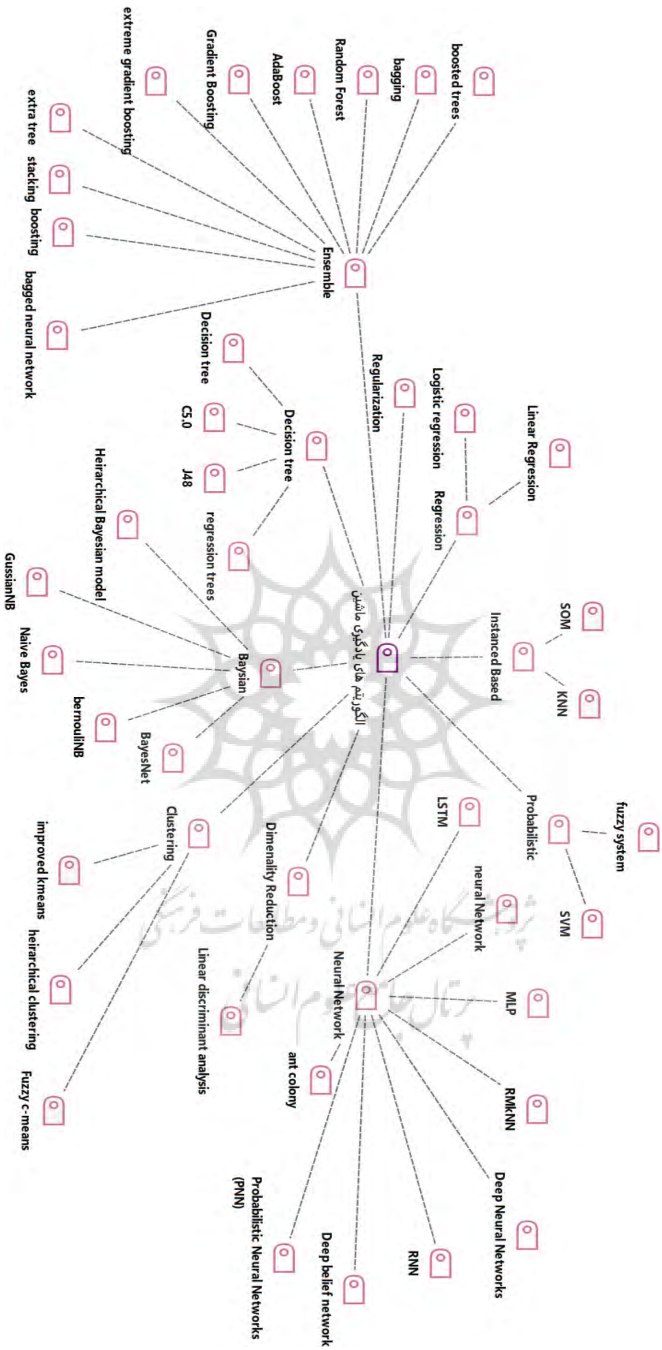
با عنایت به بررسی‌های انجام‌شده، بیشترین تکرار مربوط به ترکیب فاکتورهای مالی و جمعیت‌شناختی مشتریان است؛ چرا که هدف بسیاری از مقالات این حوزه، شناخت ریسک اعتباری مشتری و میزان سقف اعطای وام و تسهیلات است. مطابق با بررسی‌های انجام‌شده (جدول ۳)، فاکتور جمعیت‌شناختی به تنهایی عامل مناسبی برای شناخت مشتری نیست. همچنین، چنانکه از تعدد ارجاعات مشخص است، استفاده از روش RFM مورد استقبال خاصی در مقالات منتخب قرار نگرفته است. اما نکته‌ای که باید به آن توجه داشت، اینکه هر یک از فاکتورهای این مدل‌ها می‌توانند در بانک یا مؤسسه مالی مورد بررسی، شخصی‌سازی گردد. به‌عنوان نمونه، فاکتور F که تعدد مراجعات است، می‌تواند برای سرویس با تراکنش بالای بانک یا سرویس با ارزش افزودهٔ بالا برای بانک در نظر گرفته شود. فاکتورهای مالی دیگر نیز در مجموعهٔ مقالات بررسی‌شده، بر روی سهام مشتری و خرید و فروش وی تمرکز دارد؛ در حالی که دسترسی به این نوع اطلاعات در پایگاه داده‌های بانکی وجود ندارد و بنابراین، نیاز به بررسی امکان یا عدم امکان دسترسی از طریق سامانه‌های دیگر بانکی و غیربانکی و یا اکتساب این اطلاعات از طریق پرسشنامه است. برخی از فاکتورهای رفتاری نظیر احساس مشتری در زمان استفاده از سرویس و وفاداری نیازمند مصاحبه با مشتری است و قابل استخراج از پایگاه داده‌های بانکی نخواهد بود. نکته دیگر اینکه برخی از سرویس‌های عنوان‌شده در مقالات، نیازمند نگاشت با سرویس‌های بانک‌های ایرانی است.

۴-۷-۳. مقوله کلان الگوریتم‌های یادگیری ماشین

از عوامل تأثیرگذار در انتخاب الگوریتم‌های یادگیری ماشین، می‌توان به فاکتورهای مربوط به اطلاعات مشتری نظیر فاکتورهای جمعیت‌شناختی، مالی و رفتاری اشاره کرد. طبق بررسی‌های انجام‌شده الگوریتم‌های متنوعی جهت شناسایی و ارزش‌گذاری مشتری استفاده می‌شوند. پس از کدگذاری الگوریتم‌ها و تعیین دقت آن‌ها، الگوریتم‌های شناسایی‌شده در مقالات منتخب در ۱۰ دسته اصلی زیر دسته‌بندی گردیده و مدل نظری الگوریتم‌ها مطابق با شکل ۶، تدوین گردید:

Decision Tree
Dimensionality Reduction
Instanced Based
Clustering

Ensemble
Probabilistic
Neural Networks
Regularization
Regression
Bayesian



شکل ٦. مدل نظری کدهای مربوط به مقوله الگوریتم‌های یادگیری ماشین

هرچند عوامل مختلفی در انتخاب الگوریتم‌های یادگیری ماشین در حوزه شناسایی مشتری تأثیرگذار است، لیکن بر اساس کدگذاری انجام‌شده، الگوریتم‌های Random Forest (18)، SVM (19)، Logestic Regression (18)، Decission tree (14)، NaiveBayes (9)، KNN (11) و Kmeans (8) بیشترین ارجاعات را داشتند. با عنایت به بررسی دقیق مقوله کلان الگوریتم‌های یادگیری ماشین، نکات زیر می‌بایست مورد توجه قرار گیرند:

- اغلب پیاده‌سازی‌های انجام‌شده در مقالات، با ترکیب الگوریتم‌های پایه و از دسته‌های مختلف شکل ۶، طراحی شده است. به‌عنوان مثال، ترکیب SVM و Decission tree، طبق ارزیابی این الگوریتم‌ها دقت مدل‌های ترکیبی به مراتب بالاتر از الگوریتم‌های پایه است؛
- پیش از پیاده‌سازی الگوریتم‌های انتخابی از روش‌های پیش‌پردازش نظیر Dimension Reduction، feature Selection و Regularization استفاده می‌گردد. هرچه مرحله پیش‌پردازش داده‌ها نظیر حذف داده‌های دوره افتاده (با فاصله زیاد از سایر داده‌های تحقیق) و داده‌های بی‌مقدار دقیق‌تر انجام گردد، دقت مدل‌ها افزایش خواهد داشت. شایان ذکر است که دسته Regularization و Dimension Reduction به‌عنوان الگوریتم‌های پیش‌پردازش استفاده می‌شوند و با ترکیب الگوریتم‌های دیگر در شناسایی و ارزش‌گذاری مشتری مناسب خواهند بود؛
- هرچند روش‌های انتخاب ویژگی در دقت هرچه بیشتر خروجی الگوریتم‌ها کمک می‌کند، اما لازم است با توجه به تنوع و تعدد داده‌های مشتری ابتدا لیست فاکتورها توسط خبرگان بانکی محدودتر و داده‌های بی‌ربط حذف گردد؛
- حجم داده، هدف استفاده از الگوریتم، متغیرهای ورودی و میزان حساسیت زمینه در انتخاب الگوریتم‌های یادگیری ماشین تأثیرگذار است.

با عنایت به مقوله‌ها و مضامین شناسایی‌شده، تاکسونومی پیشنهادی به‌شرح زیر ارائه گردید:

شکل ۷. تاکسونومی حوزه شناسایی مشتریان با به‌کارگیری یادگیری ماشین

۵. بحث و نتیجه‌گیری

با عنایت به موارد فوق و اهمیت شناخت مشتری در صنعت بانکی، هدف اصلی پژوهش حاضر بر دسته‌بندی الگوریتم‌های یادگیری ماشین، فاکتورهای مؤثر در شناسایی و ارزش‌گذاری مشتری و شناسایی هدف‌های شناخت مشتری در صنعت بانکی قرار گرفت. به‌منظور بررسی اسناد از روش فراترکیب بهره گرفته شد. نتیجه حاصل از اجرای فراترکیب منجر به احصای ۳ مقوله کلان، ۱۸ زمینه و ۱۲۷ کد سطح ۱ گردید. پیش از این مقالات متعددی به حوزه مشتری و الگوریتم‌های یادگیری ماشین پرداخته بودند. با وجود این، این پژوهش‌ها به برخی ابعاد توجه نداشتند و هیچ‌کدام از آن‌ها به‌طور خاص بر روی مرور نظام‌مند این حوزه و در صنعت بانکی تمرکز ننموده بودند. همچنین در مقالات پیشین صرفاً به پیاده‌سازی الگوریتم‌ها پرداخته شده و دسته‌بندی و شناختی از

الگوریتم‌های پرکاربرد این حوزه و فاکتورهای ورودی ارائه نگردیده بود. بنابراین، در پژوهش جاری ضمن بررسی اهداف و کاربرد شناخت مشتری در صنعت بانکی، الگوریتم‌های استفاده‌شده دسته‌بندی گردید و متغیرهای ورودی الگوریتم‌ها نیز در ۳ دسته جمعیت‌شناختی، رفتاری، و مالی شناسایی شد. با عنایت به موارد فوق، تمایزات اصلی این پژوهش با مطالعات قبلی بدین ترتیب است: اول اینکه تاکنون مرور نظام‌مند در حوزه الگوریتم‌های یادگیری ماشین و کاربردهای آن در صنعت بانکی صورت نپذیرفته است. دوم، در پژوهش جاری سعی شده است در کنار بررسی جنبه‌های فنی موضوع و مباحث مربوط به پردازش داده، دسته‌بندی جامعی از فاکتورهای موجود جهت شناسایی و بخش‌بندی مشتری ارائه گردد که این مهم نیز در مطالعات قبلی صورت نپذیرفته بود. سوم، این پژوهش می‌تواند به‌عنوان دستورالعملی جهت آغاز پروژه‌های بخش‌بندی مشتریان بانکی مورد استفاده مدیران قرار گیرد. بنابراین، بر اساس یافته‌های پژوهش جاری و اهمیت داده‌ها در شناخت مشتری، پیشنهاد می‌گردد مدیران و تصمیم‌سازان حوزه داده نسبت به حفظ و جامعیت داده، طراحی پایگاه داده‌های متناسب با رشد داده‌های بانکی و نگهداری صحیح آن‌ها و همچنین، تکمیل اطلاعات مشتری در اولین تماس با بانک، تأکید و دقت نظر بیشتری داشته باشند. بر اساس یافته‌های پژوهش جاری، متناسب با هدف شناسایی مشتری، داده‌های موجود و فاکتورهای انتخابی، الگوریتم‌های پایه و ترکیبی می‌تواند راهگشا باشد. اما نکته مهم پیش‌پردازش دقیق داده‌هاست که این مهم با توجه به عدم تأکید بانک‌ها به شعب مبنی بر لزوم تکمیل اطلاعات مندرج در فرم‌ها توسط مشتری در زمان افتتاح حساب یا عدم طراحی سرویس مناسب جهت تکمیل اطلاعات در بسترهای الکترونیکی نیازمند بازبینی کامل است.

در کنار تمایزات اعلام‌شده، این پژوهش در چند زمینه با محدویت مواجه بوده که در تحقیقات آتی می‌تواند مدنظر قرار گیرد: بررسی‌ها بر روی مطالعات خارج از ایران انجام شده و منجر به این امر گردیده که برخی از فاکتورها به‌دلیل نبود سرویس‌های خاص در ایران، غیرکاربردی گردد. بنابراین، ضروری است خبرگان مؤسسات مالی ضمن تشکیل کارگروه متشکل از حوزه‌های مرتبط و در تعامل با مشتری نظیر بخش‌های بازاریابی و مرکز تماس نسبت به انتخاب فاکتورهای متناسب با بانک و نگاشت برخی فاکتورها به سرویس‌ها و شرایط داخلی کشور اقدام نمایند. این مهم می‌تواند در پژوهش‌های آتی و از طریق مصاحبه با خبرگان حوزه صورت پذیرد. به‌عنوان پیشنهاد برای پژوهش‌های آتی، می‌توان به استفاده از روش‌های مدل‌سازی نرم و رتبه‌بندی و سطح‌بندی فاکتورهای شناخت مشتری اشاره نمود.

نکته دیگری که خلأ آن در مقالات حوزه شناخت مشتری مشاهده شد، عدم تمرکز بر مشتریان حقوقی است. بنابراین، محققان آتی می‌توانند بر روی فاکتورهای شناسایی این دسته از مشتریان در مقالات که به‌طور قطع، پیچیدگی‌های خاص خود را خواهد داشت، مطالعه نمایند. پرداختن جامع به مقوله شناسایی مشتری افزون بر ارائه تاکسونومی که نقطه آغاز پژوهش در این زمینه است، نیازمند تبیین روابط بینابینی عوامل است، که هر یک از روابط می‌تواند در قالب تحقیقات آتی مدنظر قرار گیرد.

مقالات انتخابی با تمرکز بر داده‌های واقعی مشتری بوده است. بنابراین، داده‌های احساسی مشتری نظیر داده‌های متنی موجود در شبکه‌های اجتماعی به همراه شناخت الگوریتم‌های یادگیری ماشین در شاخه متن‌کاوی نیز می‌تواند در تحقیقات آتی مدنظر قرار گیرد. همچنین لازم به ذکر است که برخی از سؤالات و زمینه‌های حوزه شناسایی مشتری و به‌کارگیری یادگیری ماشین به‌دلیل محدودیت روش تحقیق مرور نظام‌مند مورد بررسی عملی قرار نگرفت. بنابراین، به‌عنوان شاخه دیگری از مطالعات آتی و به‌عنوان گام کاربردی‌سازی خروجی پژوهش، می‌توان خلأهای تجربی صنعت را نیز پوشش داد. به‌عنوان مثال، مطالعه پیش رو نشان داد که به‌رغم اینکه مطالعات متعددی در حوزه یادگیری ماشین و شناسایی مشتری انجام شده، تمرکز کافی در حوزه طراحی سرویس و ارائه سرویس شخصی‌سازی‌شده به مشتریان بانکی بر مبنای نتایج دسته‌بندی‌ها وجود نداشته است. بنابراین، در پژوهش‌های آتی استفاده از روش‌شناسی مورد مطالعه و علم طراحی، گام مهمی در صنعت خواهد بود. از طرف دیگر، فاکتورهای شناسایی‌شده در مقالات (متغیرهای مبنایی دسته‌بندی‌ها) نیازمند شخصی‌سازی در هر کشور متناسب با فرهنگ، قوانین بالادستی، بررسی وجود/عدم وجود برخی سرویس‌ها و متناسب با وجود داده‌هاست که می‌توان در مطالعات آتی از طریق مصاحبه نیمه‌ساختاریافته با سازمان‌های بالادستی نظیر بانک مرکزی و سایر نهادهای قانون‌گذار به این مهم دست یافت.

References

Abbasimehr, H. & M. Shabani. 2021. A new methodology for customer behavior analysis using time series clustering A case study on a bank’s customers. *Kybernetes* 50 (2): 221-242. https://doi.org/10.1108/k-09-2018-0506

AbdolvandAmir, N. & A. Albadvi. 2014. Customer Lifetime Value: Literature Scoping Map, and an Agenda for Future Research. *international Journal of Management Perspective* 1 (3):41-59.

Alam, T. M., K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, & M. Khushi. 2020. An investigation of credit card default prediction in the imbalanced datasets [Article]. *Ieee Access* 8: 201173-201198. https://doi.org/10.1109/ACCESS.2020.3033784

Al-A’raj, M. & M. F. Abbod. 2016. Classifiers consensus system approach for credit scoring [Article]. *Knowledge-Based Systems 104*: 89-105. https://doi.org/10.1016/j.knosys.2016.04.013

_____ & M. Majdalawieh. 2021. Modelling customers credit card behaviour using bidirectional LSTM neural networks. *Journal of Big Data* 8 (1): Article 69. https://doi.org/10.1186/s40537-021-00461-7

Al-Qerem, A., G. Al-Naymat, M. Alhasan, & M. Al-Debei. 2020. Default prediction model: The significant role of data engineering in the quality of outcomes [Article]. *International Arab Journal of Information Technology* 17 (4 Special Issue): 635-644. https://doi.org/10.34028/iajit/17/4A/8

Al-Rubaiee, H., K. Alomar, R. Qiu, & D. Li. 2018. Tuning of Customer Relationship Management (CRM) via Customer Experience Management (CEM) using sentiment analysis on aspects level [Article]. *International Journal of Advanced Computer Science and Applications* 9 (5): 300-312. https://doi.org/10.14569/IJACSA.2018.090540

Arli, D. 2017. “Investigating consumer ethics: a segmentation study. *Journal of Consumer Marketing,* 34 (7):636-645. https://doi.org/10.1108/JCM-08-2016-1908

Ashofteh, A., & J. M. Bravo. 2021. A conservative approach for online credit scoring* [Article]. *Expert Systems with Applications*, *176*, 16, Article 114835. https://doi.org/10.1016/j.eswa.2021.114835

Atkins, S., S. Lewin, H. Smith, M. Engel, A. Fretheim, & J. Volmink. 2008. Conducting a meta-ethnography of qualitative literature: Lessons learnt. *BMC Medical Research Methodology* 8 (1) 21. https://doi.org/10.1186/1471-2288-8-21

Behare, N., S. Waghulkar, & S. A. Shah. 2018. A Theoretical Perspective on Customer Experience (CX) in Digital Business Strategy. [2018 ieee international conference on research in intelligent and computing in engineering (rice iii)]. 3rd IEEE International Conference on Research in Intelligent and Computing in Engineering (RICE), Univ Don Bosco, San Salvador, EL SALVADOR.

Bekamiri, H., M. Mehraeen, A. Pooya & H. Sharif. 2020. A Stochastic Approach for Valuing Customers in Banking Industry: A Case Study. Industrial Engineering and Management Systems, 19 (4), 744-757. https://doi.org/10.7232/iems.2020.19.4.744

Bharathi, S. V., D. Pramod, & R. Raman. 2022. An Ensemble Model for Predicting Retail Banking Churn in the Youth Segment of Customers. *DATA*, *7* (5), Article 61. https://doi.org/10.3390/data7050061

Borna,K., Sh. Hoseini & M. Aghaei. 2019. Customer satisfaction prediction with Michigan-style learning classifier system. *SN Applied Sciences*. https://doi.org/https://doi.org/10.1007/s42452-019-1493-1

Calvo-Porral, C., & J. P. Levy-Mangin. 2020. An emotion-based segmentation of bank service customers. *International Journal of Bank Marketing*, *38* (7): 1441-1463. https://doi.org/10.1108/ijbm-05-2020-0285

Chen, T. H. 2020a. Do you know your customer? Bank risk assessment based on machine learning. *Applied Soft Computing*, *86*, Article 105779. https://doi.org/10.1016/j.asoc.2019.105779

Chen, T. H. 2020b. Do you know your customer? Bank risk assessment based on machine learning [Article]. *Applied Soft Computing Journal*, *86*, Article 105779. https://doi.org/10.1016/j.asoc.2019.105779

Cheng, L. C., C. C. Wu, & C. Y. Chen. 2019. Behavior Analysis of Customer Churn for a Customer Relationship System: An Empirical Case Study. *Journal of Global Information Management* 27 (1): 111-127. https://doi.org/10.4018/jgim.2019010106

Chopra, A., & P. Bhilare. 2018. Application of Ensemble Models in Credit Scoring Models [Article]. *Business Perspectives and Research* 6 (2): 129-141. https://doi.org/10.1177/2278533718765531

Çığşar, B., & D. Ünal. 2019. Comparison of Data Mining Classification Algorithms Determining the Default Risk. *Scientific Programming*, *2019*, 8706505. https://doi.org/10.1155/2019/8706505

Coser, A., A. Aldea, M. M. Maer-Matei & L. Besir. 2020. Propensity to Churn in Banking: what Makes Customers Close the Relationship with a Bank? *Economic Computation and Economic Cybernetics Studies and Research* 54 (2): 77-94. https://doi.org/10.24818/18423264/54.2.20.05

Dawood, E. A., E. Elfakhrany & D. A. Maghraby. 2019. Improve Profiling Bank Customer's Behavior Using Machine Learning. *Ieee Access* 7: 109320-109327. https://doi.org/10.1109/access.2019.2934644

Dayu Xu & Xuyao Zhang & Hailin Feng 2019. "Generalized fuzzy soft sets theory-based novel hybrid ensemble credit scoring model," *International Journal of Finance & Economics,* 24 (2): 903-921.

De Lima Lemos, R. A., T. C. Silva & B. M. Tabak. 2022. Propension to customer churn in a financial institution: a machine learning approach. *Neural Computing and Applications* 34 (14): 11751-11768. https://doi.org/10.1007/s00521-022-07067-x

Domingos, E., B. Ojeme & O. Daramola. 2021. Experimental Analysis of Hyperparameters for Deep Learning-Based Churn Prediction in the Banking Sector. *Computation* 9 (3): Article 34. https://doi.org/10.3390/computation9030034

Eltahir, A., T. Ahmed, A. Abdelmageed Mohammed, A. Hilal, & T, A. A. 2022. An Approach of Supervised and Unsupervised Machine Learning Model for E-CRM Bank's Marketing. *International Journal of Computer Science and Network Security* 22 (4): 625-636.

Erwin, E., M. J. Brotherson & J. A. Summers. 2011. Understanding Qualitative Metasynthesis: Issues and Opportunities in Early Childhood Intervention Research. *Journal of Early Intervention* 33 (3). https://doi.org/10.1177/1053815111425493

Estrella-Ramon, A., M. Sanchez-Perez, G. Swinnen & K. VanHoof. 2017. A model to improve management of banking customers. *Industrial Management & Data Systems* 117 (2): 250-266. https://doi.org/10.1108/imds-03-2016-0107

Firman Pradana Rachman, Handri Santoso & A. Djajadi. 2021. Machine Learning Mini Batch K-means and Business Intelligence Utilization for Credit Card Customer Segmentation. *International Journal of Advanced Computer Science and Applications* 12 (10): 218-227.

Furio. Camillo & C. Liberati. 2018. Personal values and credit scoring: new insights in the financial prediction. *Journal of the Operational Research Society* 69 (12) https://doi.org/10.1080/01605682.2017.1417684

Garcia-Mendez, S., M. Fernandez-Gavilanes, J. Juncal-Martinez, F. J. Gonzalez-Castano & O. B. Seara. 2020. Identifying Banking Transaction Descriptions via Support Vector Machine Short-Text Classification Based on a Specialized Labelled Corpus. *Ieee Access*, 8: 61642-61655. https://doi.org/10.1109/access.2020.2983584

Gartner. 2017. The customer experience marketing leader's first 100 days. https://www.gartner.com/en/marketing/insights/articles/the-customer-experience-marketing-leaders-first-100-days

Hajipour, B. & M. Esfahani. 2019. Delta model application for developing customer lifetime value [Article]. *Marketing Intelligence and Planning*, 37 (3): 298-309. https://doi.org/10.1108/MIP-06-2018-0190

Huseynov, F. & S. Özkan Yıldırım. 2019. Online Consumer Typologies and Their Sho pping Behaviors in B2C E-Commerce Platforms. *SAGE Open* 9 (2): 215824401985463. https://doi.org/10.1177/2158244019854639

Ieva, M., & C. Ziliani. 2018. Mapping touchpoint exposure in retailing: Implications for developing an omnichannel customer experience [Article]. *International Journal of Retail and Distribution Management* 46 (3): 304-322. https://doi.org/10.1108/IJRDM-04-2017-0097

Jędrzejczyk, W. 2021. Managing Customer Relations and Value in Organizations with the Use of IT Tools: Customer Segmentation on the Market of Eco-Innovative Services. Procedia Computer Science 192: 2816-2825. https://doi.org/10.1016/j.procs.2021.09.052

Keramati, A., H. Ghaneei & S. M. Mirmohammadi. 2016. Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation* 2 (1), Article 10. https://doi.org/10.1186/s40854-016-0029-6

Komulainen, H., & S. Saraniemi. 2019. Customer centricity in mobile banking: a customer experience perspective [Article]. *International Journal of Bank Marketing* 37 (5): 1082-1102. https://doi.

org/10.1108/ijbm-11-2017-0245

Kovacs, T., A. D. Ko & A. Asemi. 2021. Exploration of the investment patterns of potential retail banking customers using two-stage cluster analysis [Article]. *Journal of Big Data* 8 (1): 25, Article 141. https://doi.org/10.1186/s40537-021-00529-4

Ladyzynski, P., K. Zbikowski & P. Gawrysiak. 2019. Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications* 134: 28-35. https://doi.org/10.1016/j.eswa.2019.05.020

Li, K., J. Rollins & E. Yan. 2018. Web of Science use in published research and review papers 1997–2017: a selective, dynamic, cross-domain, content-based analysis. *Scientometrics* 115 (1): 1-20. https://doi.org/10.1007/s11192-017-2622-5

Livieris, I. E., N. Kiriakidou, A. Kanavos, V. Tampakas & P. Pintelas. 2018. On ensemble SSL algorithms for credit scoring problem [Article]. *Informatics* 5 (4), Article 40. https://doi.org/10.3390/informatics5040040

Long, H. V., L. H. Son, M. Khari, K. Arora, S. Chopra, R. Kumar, T. Le & S. W. Baik. 2019. A new approach for construction of geodemographic segmentation model and prediction analysis [Article]. *Computational Intelligence and Neuroscience*, *2019*, Article 9252837. https://doi.org/10.1155/2019/9252837

Lu, Q., Z. Cui, Y. Chen & X. Chen. 2017. Extracting optimal actionable plans from additive tree models [Article]. *Frontiers of Computer Science* 11 (1): 160-173. https://doi.org/10.1007/s11704-016-5273-4

Mancisidor, R. A., M. Kampffmeyer, K. Aas, & R. Jenssen. 2021. Learning latent representations of bank customers with the Variational Autoencoder. *Expert Systems with Applications*, *164*, Article 114020. https://doi.org/10.1016/j.eswa.2020.114020

Marinakos, G., & S. Daskalaki. 2017. Imbalanced customer classification for bank direct marketing [Article]. *Journal of Marketing Analytics* 5 (1): 14-30. https://doi.org/10.1057/s41270-017-0013-7

Melo, L., F. M. Nardini, C. Renso, R. Trani & J. A. Macedo. 2020. A novel approach to define the local region of dynamic selection techniques in imbalanced credit scoring problems [Article]. *Expert Systems with Applications* 152: 16, Article 113351. https://doi.org/10.1016/j.eswa.2020.113351

Mhlanga, D. 2021. Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment [Article]. *International Journal of Financial Studies* 9 (3) Article 39. https://doi.org/10.3390/ijfs9030039

Milosevic, M., N. Zivic & I. Andjelkovic. 2017. Early churn prediction with personalized targeting in mobile social games [Article]. *Expert Systems with Applications*, *83*, 326-332. https://doi.org/10.1016/j.eswa.2017.04.056

Møller, A. M., & P. S. Myles. 2016. What makes a good systematic review and meta-analysis? *BJA: British Journal of Anaesthesia* 117 (4): 428-430. https://doi.org/10.1093/bja/aew264

Mosavi, A. B., & A. Afsar. 2018. Customer Value Analysis in Banks Using Data Mining and Fuzzy Analytic Hierarchy Processes. *International Journal of Information Technology & Decision Making* 17 (3): 819-840. https://doi.org/10.1142/s0219622018500104

Motevali, M. M., A. M. Shanghooshabad, R. Z. Aram & H. Keshavarz. 2019. WHO: A New Evolutionary Algorithm Bio-Inspired by Wildebeests with a Case Study on Bank Customer Segmentation. *INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE* 33 (5), Article 1959017. https://doi.org/10.1142/S0218001419590171

Müller, J. M., B. Pommeranz, J. Weisser & K. I. Voigt. 2018. Digital, Social Media, and Mobile Marketing in industrial buying: Still in need of customer segmentation? Empirical evidence from Poland and Germany. *Industrial Marketing Management* 73: 70-83. https://doi.org/https://doi.org/10.1016/j.indmarman.2018.01.033

Pandey, M. K., M. Mittal & K. Subbiah. 2021. Optimal balancing & efficient feature ranking approach to minimize credit risk [Article]. *International Journal of Information Management Data Insights*, 1 (2), Article 100037. https://doi.org/10.1016/j.jjimei.2021.100037

Pawełoszek, I. 2021)/. Customer segmentation based on activity monitoring applications for the recommendation system. *Procedia Computer Science*, 192: 4751-4761. https://doi.org/https://doi.org/10.1016/j.procs.2021.09.253

Plawiak, P., M. Abdar, J. Plawiak, V. Makarenkov & U. R. Acharya. 2020. DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring [Article]. *Information Sciences*, 516: 401-418. https://doi.org/10.1016/j.ins.2019.12.045

Razavi, R., A. Gharipour, M. Fleury & I. J. Akpan. 2019. A practical feature-engineering framework for electricity theft detection in smart grids. *Applied Energy*, 238: 481-494. https://doi.org/https://doi.org/10.1016/j.apenergy.2019.01.076

Safarkhani, F., & S. Moro. 2021. Improving the Accuracy of Predicting Bank Depositor's Behavior Using a Decision Tree. *APPLIED SCIENCES-BASEL* (19), Article 9016. https://doi.org/10.3390/app11199016

Sandelowski, M., & J. Barroso. 2006. *Handbook for synthesizing qualitative research*. New york: Springer Publishing Company.

Seidlova, R., J. Poživil & J. Seidl. 2019. Marketing and business intelligence with help of ant colony algorithm [Article]. *Journal of Strategic Marketing* 27 (5): 451-463. https://doi.org/10.1080/0965254X.2018.1430058

Sheikha, A., T. Ghanbarpourb & D. Gholamiangonabadib. 2019. A Case Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. *Journal of Business-to-Business Marketing* 26 (2). https://doi.org/10.1080/1051712X.2019.1603420

Shirazi, F., & M. Mohammadi. 2019. A big data analytics model for customer churn prediction in the retiree segment. *International Journal of Information Management* 48: 238-253. https://doi.org/10.1016/j.ijinfomgt.2018.10.005

Sivasankar, E., C. Selvi & S. Mahalakshmi. 2020. Rough set-based feature selection for credit risk prediction using weight-adjusted boosting ensemble method [Article]. *Soft Computing* 24 (6): 3975-3988. https://doi.org/10.1007/s00500-019-04167-0

Torrens, M. & A. Tabakovic. 2022. A Banking Platform to Leverage Data Driven Marketing with Machine Learning. *Entrop* 24 (3). https://doi.org/10.3390/e24030347

Urkup, C., B. Bozkaya & F. S. Salman. 2018. Customer mobility signatures and financial indicators as predictors in product recommendation [Article]. *Plos One* 13 (7): 18, Article e0201197. https://doi.org/10.1371/journal.pone.0201197

World economic Forum. 2017. Shaping the Future of Retail for Consumer Industries. https://www3.weforum.org/

Wu, J., L. Shi, W. P. Lin, S. B. Tsai, Y. Li, L. Yang & G. Xu. 2020. An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and <i>K</i>-Means Algorithm. *Mathematical Problems in Engineering* 2020: 8884227. https://doi.org/10.1155/2020/8884227

Xi, Song, Matthew Tingchi Liu, Qianying Liu, & Ben NiuBen Niu. 2021. Hydrological cycling optimization-based multiobjective feature-selection method for customer segmentation. *International Journal of Intelligent Systems* 36 (16). https://doi.org/10.1002/int.22381

Yanik, S., & A. Elmorsy. 2019. SOM approach for clustering customers using credit card transactions. *International Journal of Intelligent Computing and Cybernetics* 12 (3): 372-388. https://doi.org/10.1108/ijicc-11-2018-0157

الهه باغانی

متولد سال ۱۳۶۷، دانشجوی دکتری مدیریت فناوری اطلاعات، گرایش کسب‌وکار هوشمند در دانشگاه تربیت‌مدرس است. ایشان هم‌اکنون معاون اداره شبکه و سخت‌افزار بانک سیناست.

مدیریت تجربه مشتری، طراحی سرویس دیجیتال، کسب‌وکار هوشمند و یادگیری ماشین از جمله علایق پژوهشی وی است.

شعبان الهی

عضو هیئت علمی نمونه کشور، پژوهشگر برتر استان و استاد سرآمد آموزشی وعضو هیئت علمی گروه مدیریت صنعتی دانشگاه ولی عصر رفسنجان می‌باشد. وی سابقه استادتمام مدیریت فناوری اطلاعات در دانشگاه تربیت مدرس، ریاست مرکز ملی مطالعات مدیریت، معاونت پژوهش و فناوری دانشکده مدیریت و اقتصاد دانشگاه تربیت مدرس و همچنین مدیرکل ارزیابی توسعه علوم معاونت علمی و فناوری ریاست جمهوری را در کارنامه خود دارد. دکتر الهی برای اولین بار در کشور رشته مدیریت فناوری اطلاعات را راه‌اندازی و رشته سیاست گذاری علم و فناوری را تأسیس و راه‌اندازی کرد و راه‌اندازی و سردبیری فصلنامه علمی و پژوهشی «پژوهش‌های مدیریت منابع سازمانی» را بر عهده داشت.

علایق تحقیقاتی ایشان شامل مدیریت فناوری اطلاعات، مدیریت صنعتی، مدیریت دانش و نوآوری، آینده‌پژوهی و آینده‌نگاری، کسب و کارهوشمند و دیجیتالی، سیستم خبره و هوش مصنوعی، حکمرانی و مدیریت تحول دیجیتالی، روش‌شناسی پژوهش، راهبر بودن و راهبری مؤثر است.

علیرضا حسن‌زاده

متولد سال ۱۳۴۴، دارای مدرک تحصیلی دکتری در رشته مدیریت سیستم‌ها از دانشگاه تهران است. ایشان هم‌اکنون استاد و مدیر گروه مدیریت فناوری اطلاعات دانشگاه تربیت‌مدرس است.

اینترنت اشیا، هوشمندی کسب‌وکار و تحلیل بیگ‌دیتا از جمله علایق پژوهشی وی است.

علی رجب‌زاده قطری

دارای مدرک تحصیلی دکتری مدیریت با گرایش مدیریت و تولید و عملیات از دانشگاه تربیت‌مدرس است. ایشان هم‌اکنون استاد گروه مدیریت صنعتی دانشکده مدیریت و اقتصاد دانشگاه تربیت‌مدرس است.

شبیه‌سازی سیستم‌های گسسته پیشامد، پویا عامل بنیان، تصمیم‌گیری، بازمهندسی فرایندهای سازمان‌های خدماتی و تولیدی، هوش مصنوعی در مدیریت از جمله علایق پژوهشی وی است.

پژوهش‌نامه
پردازش و
مدیریت
اطلاعات