Scientometrics Research Journal

Scientific Bi-Quarterly of Shahed University
Vol. 8, No. 2, Autumn & Winter 2022-23 (Serial 16)

Research Article

DOI: 10.22070/RSCI.2021.13393.1449

Optimizing Confusion of Authors' Names in Persian Articles Using Random Forest Algorithm

Niloofar Mozafari [1*]

Narjes Vara [2]

1. Assistant Professor, Design and System Operations Department, Regional Information Center for Science and Technology, Shiraz, Iran. (Corresponding Author)

2. Faculty of Evaluation and Resource Development Department, Regional Information Center for Science and Technology & PhD Student in Library and Information Science, shiraz, Iran. Email: vara@ricest.ac.ir

Email:mozafari@ricest.ac.ir

Date of Reception: 04/05/2021

Date of Acceptation: 23/08/2021

Abstract

Purpose: Name is a key factor for distinguishing authors. In the academic databases that store information on papers, searching for the name of the article author is one of the most important elements in increasing visibility and the quantitative studies in the field of Scientology including the amount of citing works. The diversity of writings is one of the issues that lead to challenges in various scientific fields. In addition, the lack of writing standards in the Persian language and the lack of keyboards and standard codes, the habit of simply writing are among the factors that lead to the author's name disambiguation. Also, the spelling mistakes that occur by the writers in writing the name lead to the creation of different forms of writing for a single name. Considering the importance of solving the confusion of authors' names in Persian articles, this paper aims to propose a framework to solve the problem of confusion and dispersion of authors' names in Persian articles, which has led to a rupture and lack of comprehensiveness in information retrieval.

Methodology: The present research is an applied scientometrics method carried out by documentary procedure, and the required data is collected from the ISC database. The initial statistical population is 913 records during the period 2015 to 2017. The proposed framework consists of three stages: searching, matching, and grouping. In this regard, after initial pre-processing and feature extraction, the search operation is performed to find records that are potentially likely to be identical. Our method extracts two types of features including internal and external. The internal feature has been extracted from the author's information like first name, last name, affiliation, email, and co-authors. In addition, the external feature uses the scientific history of authors like articles and research interests. Next, in the search phase, the records that are potentially the same are identified. We propose a new method called Farsi-Soundex,

Scientometrics Research Journal

Scientific Bi-Quarterly of Shahed University
Vol. 8, No. 2, Autumn & Winter 2022-23 (Serial 16)

Research Article

DOI: 10.22070/RSCI.2021.13393.1449

Niloofar Mozafari [1*]

Narjes Vara [2]

Date of Reception: 04/05/2021

Date of Acceptation: 23/08/2021

which has been inspired by the well-known Soundex to categorize potential unique names. The same records are then found through further investigation in the adaptation phase, which is based on random forests. Therefore, the input of the matching stage is a group of records that have been detected the same based on the Farsi-Soundex algorithm. To specify whether these records are the same or not, a random forest algorithm has been applied to them. Finally, in the grouping stage, all the records that have been identified as the same using random forest are placed in one group by a hash-based algorithm.

Findings: The internal features of Email address, last name, and first name are the most significant features to optimize name-writing confusion. Also, the obtained results show the external features of the main subject and sub-subject provide the least effective features for solving the author name disambiguation problem in the academic database. In addition, using a random forest as a classifier in the matching phase, with an accuracy of over 99%, can solve the problem of confusion in writing the authors' names.

Conclusion: Results show the high efficiency of our framework in uniformity of names according to the criteria of accuracy, recall, and F value compared to the support vector machine, the nearest neighbor, and genetics. Our proposed method can be applied to scientific databases to standardize the names of the authors. In the future, we are investigating the efficiency of our proposed framework in a non-stationary environment in which the distribution of data may be changed over time.

Keywords: Name ambiguity, Article authors Persian articles, Random forest algorithm, Name Authority, Farsi-Soundex algorithm.

پژوهش‌نامه علم‌سنجی
دوفصلنامه علمی دانشگاه شاهد/ دوره ۸/ شماره ۲/ پاییز و زمستان ۱۴۰۱ (پیاپی ۱۶)

مقاله پژوهشی

DOI: 10.22070/RSCI.2021.13393.1449

بهینه‌سازی آشفتگی اسامی نویسندگان مقالات فارسی با استفاده از روش جنگل تصادفی

نیلوفر مظفری [1*]

نرجس ورع [2]

۱. استادیار، گروه پژوهشی طراحی و عملیات سیستم‌ها، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، شیراز، ایران. (نویسنده مسئول)

۲. عضو هیئت علمی گروه پژوهشی ارزیابی و توسعه منابع، مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری و دانشجوی دکتری علم اطلاعات و دانش شناسی. شیراز، ایران.

Email: vara@ricest.ac.ir

Email: mozafari@ricest.ac.ir

صفحه ۲۲۰-۲۰۳

دریافت: ۱۴۰۰/۰۲/۱۴

پذیرش: ۱۴۰۰/۰۶/۰۱

چکیده

هدف: ارائه چارچوبی جهت حل مشکل آشفتگی و پراکندگی اسامی نویسندگان در مقالات فارسی که منجر به گسیختگی و فقدان جامعیت در بازیابی اطلاعات شده است.

روش‌شناسی: پژوهش حاضر از نوع کاربردی علم‌سنجی است که به روش اسنادی انجام شده است. جامعه آماری را از ۹۱۳ رکورد از نام نویسندگان مقالات فارسی برگرفته از پایگاه استنادی علوم جهان اسلام، طی بازه زمانی ۱۳۹۵ تا ۱۳۹۷ تشکیل می‌دهد. چارچوب پیشنهادی از سه مرحله جستجو، تطابق و گروه‌بندی تشکیل شده است. در این راستا، بعد از پیش‌پردازش اولیه و استخراج ویژگی، عملیات جستجو با هدف یافتن رکوردهایی که بالقوه احتمال یکسان‌بودن آنها وجود دارد انجام شده و سپس رکوردهای یکسان از طریق بررسی‌های بیشتر در مرحله تطابق که مبتنی بر جنگل تصادفی است یافت می‌شود.

یافته‌ها: ویژگی‌های پست الکترونیک، نام خانوادگی و نام از مهم‌ترین ویژگی‌ها برای بهینه‌سازی آشفتگی نگارش اسامی هستند. استفاده از جنگل تصادفی به‌عنوان طبقه‌بند در مرحله تطابق، با دقت بالای ۹۹ درصد می‌تواند مشکل آشفتگی نگارش اسامی نویسندگان را برطرف نماید.

نتیجه‌گیری: نتایج نشان از کارایی بالای این روش در یکدست‌سازی اسامی با توجه به معیارهای دقت، بازیافت و مقدار اف نسبت به طبقه‌بندهای بردار پشتیبان، نزدیک‌ترین همسایه و ژنتیک دارد.

واژگان کلیدی: آشفتگی نگارش، جنگل تصادفی، نویسندگان مقالات فارسی، مستندسازی نام‌ها، الگوریتم ساندکس.

مقدمه و بیان مسئله

از میان حجم گسترده فعالیت‌های پژوهشی و تولیدات علمی، سهم عمده‌ای به مقالات اختصاص دارد که با انتشار یافته‌ها، نقش مؤثری در توسعه علمی و فنی کشور ایفا می‌کنند. سنجش و ارزیابی توانمندی علمی نیازمند بهره‌گیری از شاخص‌های استانداردی است؛ اما گسترش روزافزون تعداد مقالات و ورود بدون نظارت اطلاعات کتابشناختی آنها در پایگاه‌های علمی، دشواری‌های بازیابی اطلاعات را دوچندان کرده است. یکی از مهم‌ترین عناصر کتابشناختی مقالات، نام نویسنده است؛ به‌طوری‌که حتی یکی از روش‌های جستجوی منابع علمی روش جستجوی ستاره‌ها[1]ست. بدین معنی که کاربر به جای جستجوی موضوع یا کلیدواژه‌ای، نام مؤلف برجسته‌ای را در حوزه موضوعی خاص جستجو و به دنبال مقالات مرتبط با حوزه کاری خود در میان آثار وی می‌گردد (خسروی،۱۳۹۰). جستجوی نام نویسندگان نه‌تنها برای یافتن منابع علمی به قصد مطالعه و پژوهش بلکه برای مطالعات علم‌سنجی، جهت شناسایی نویسندگان برتر در سطح فردی نیز کاربرد دارد و امکان ردیابی حرکت و رشد علمی داخلی و بین‌المللی هر پژوهشگر را فراهم می‌آورد. همچنین به توسعه سیاست‌های علم و فناوری مبتنی بر منابع انسانی کمک می‌کند (Kawashima & Tomizawa, 2015). به عبارتی این مهم تنها به نام نویسندهٔ مقاله محدود نمی‌شود، بلکه در فرایند استناددهی نیز حائز اهمیت است.

در مباحث مربوط به پایگاه‌های اطلاعاتی، هر موردی که بخواهیم درباره آن اطلاعاتی را ذخیره کنیم یک موجودیت[2] نامیده می‌شود. بر این اساس نام نویسنده در یک پایگاه اطلاعات علمی یک موجودیت به شمار می‌رود و جستجو در آن پایگاه تنها در صورتی می‌تواند به بازیابی جامع و مانع منتهی شود که در درجه اول اصول نگارش اسامی و نشانی‌ها توسط نویسندگان رعایت و سپس ورود این اطلاعات کتابشناختی به‌صورت صحیح در پایگاه‌های اطلاعات علمی انجام گیرد؛ چراکه با نمایه‌سازی درست است که دستیابی دقیق به پیشینه علمی افراد، دانشگاه‌ها و نهادهای پژوهشی میسر می‌شود و رعایت‌نکردن آن منجر به گسیختگی و فقدان انسجام در بازیابی شده و مشکلاتی را ایجاد می‌کند (زلفی‌گل، شیری و کیانی بختیاری، ۱۳۸۶).

در زمینه نام نویسندگان، دو نوع آشفتگی وجود دارد. حالت نخست شامل نام یک نویسنده به شکل‌های نگارشی متنوع و حالت دوم دربردارنده یک نام، متعلق به نویسندگان مختلف است. این امر منجر به ترکیب مستندات شده و ممکن است اطلاعات یک نویسنده برای نویسنده دیگری نمایش داده شود (Kim & Kim, 2019). ازاین‌رو، باید بتوان با روش‌هایی، این موجودیت‌ها را از هم جدا کرد. تاکنون رویکردهای مختلفی برای بهبود این مشکل ارائه شده است، که در سه دسته کلی قرار دارند: روش‌هایی که فقط حالت نخست (On et al., 2006)، فقط حالت دوم (Fan et al., 2011)، (Kang et al., 2009) و یا هر دو حالت از مشکلات آشفتگی در اسامی را مورد توجه قرار داده‌اند (Cota et al., 2010)، (Han et al., 2004). برخی برای شناسایی نویسنده‌هایی که چند نام دارند؛ یا چند نویسنده با یک نام مشترک، از ویژگی روابط میان نویسندگان استفاده کرده‌اند (Shin et al., 2010)، (Fan et al., 2011)، (Zhang & Hasan, 2017)، (Kim & Kim, 2019). برخی دیگر صور مختلف نگارشی نام را به‌عنوان ویژگی مورد بررسی انتخاب کرده‌اند (مزروعی و همکاران، ۱۳۹۲؛ مرتضوی، ندیمی شهرکی، خانی مصطفی، ۱۳۹۶). اطلاعات محتوایی بخش‌های عنوان، چکیده و واژگان کلیدی نیز جهت تفکیک اشخاص دارای نام‌های مشابه مورد

1 . Stars Search
2 . Entity

استفاده قرار گرفته است (صادقی گورجی و همکاران، ۱۳۹۴).

همان‌گونه که اشاره شد، در هر یک از روش‌ها، از مجموعه‌ای ویژگی برای حل مشکل آشفتگی در نگارش نام‌ها استفاده شده که می‌تواند مبتنی بر روش‌های با نظارت یا بدون نظارت باشند. روش‌های بدون نظارت اغلب از توابع مشابهت برای بررسی شباهت بین ویژگی‌ها و گروه‌بندی رکوردهای مربوط به یک نویسنده استفاده می‌کنند (Cota et al., 2010)، (Bhattacharya & Getoor, 2006)، (Han et al., 2004). در مقابل روش‌های با نظارت از مجموعه آموزشی شامل نمونه‌های از قبل برچسب‌گذاری‌شده به‌منظور پیش‌بینی نویسنده یک رکورد یا تعیین دو رکورد متعلق به یک نویسنده یکسان استفاده می‌کنند (Torvik & Smalheiser, 2009)، (Ferreira et al., 2010).

در زبان فارسی به دلیل پیچیدگی و نبود یکدستی با چالش‌های نحوی، ریخت‌شناسی و معنایی واژگان مواجه هستیم (ستوده، هنرجویان، ۱۳۹۱). ازاین‌رو با توجه به اهمیت دقت بازیابی اطلاعات مربوط به پژوهشگران و نویسندگان مقالات، مسئله اصلی پژوهش از آنجا ناشی می‌شود که به دلیل تنوع فراوان نگارش در نام‌های نویسندگان مقالات فارسی، جامعیت بازیابی مدارک در پایگاه‌های استنادی و علمی، می‌تواند تحت تأثیر قرار گرفته و در هر جستجو، تعداد زیادی از مدارک مرتبط علی‌رغم وجود در پایگاه از دست برود. هرچند برخی پایگاه‌های اطلاعاتی، معیارهایی را برای جستجوی نگارش‌های مختلف نام در نظر گرفته‌اند؛ اما نمی‌توانند ارتباط میان آنها را تشخیص و یا برقرار نماید. به بیان دیگر، بین اشکال مختلف نام نویسندگان پیوندی وجود ندارد، تا اگر کاربر یک شکل از نام را جستجو کند، شکل‌های دیگر نام نیز در نتایج بازیابی لحاظ شود. همچنین نمی‌توان از کاربران انتظار داشت که تمامی صورت‌های احتمالی نگارش نام نویسنده را پیش‌بینی و جستجو نمایند؛ حتی در چنین حالتی نیز نتایج بازیابی نگارش‌های مختلف نام برخی از نویسندگان از نوسان چشمگیری برخوردار است. نظر به اینکه راهکارهای انسانی نیازمند مشارکت فعالانه نویسندگان متون و تایپیست‌هاست که روندی کند، بلندمدت و هزینه‌بر است؛ بنابراین ضروری است، راهکارهای خودکارسازی در فرایند پردازش، بیش از پیش مورد تأکید قرار گیرد. ازاین‌رو پژوهش حاضر در تلاش است بخشی از چالش‌هایی که به تنوع ریخت‌شناختی و آشفتگی نگارش نام‌های فارسی بازمی‌گردد و امکان بهبود آن به‌صورت ماشینی وجود دارد، را در راستای کنترل و یکدست‌سازی نام‌های نویسندگان در پایگاه‌های استنادی و اطلاعاتی به شکل خودکار ارائه دهد؛ به‌ترتیبی که بتوان هنجارسازی در الگوریتم جستجو را بهبود بخشید تا نام‌ها، صرف نظر از ریخت‌های مختلف، بازیابی شوند.

سؤال‌های پژوهش

در پژوهش حاضر، به سؤال‌های زیر پاسخ داده می‌شود:

۱. کدام‌یک از ویژگی‌های استفاده‌شده در تشخیص و بهبود آشفتگی نگارش نام‌های نویسندگان مقالات فارسی، نسبت به دیگر ویژگی‌ها از اهمیت بیشتری برخوردار هستند؟
۲. الگوریتم جنگل تصادفی به چه میزان می‌تواند در تشخیص و بهبود آشفتگی نگارش اسامی نویسندگان مقالات فارسی مؤثر واقع گردد؟
۳. استفاده از الگوریتم جنگل تصادفی به‌عنوان الگوریتم تطابق برای تشخیص و بهبود آشفتگی نگارش اسامی نویسندگان مقالات فارسی، به چه میزان موجب بهبود دقت در مقایسه با دیگر طبقه‌بندها می‌شود؟

چارچوب نظری

بهینه‌سازی آشفتگی اسامی نویسندگان به معنای یکسان‌سازی اسامی نویسندگان در انبار داده پایگاه‌های اطلاعات علمی و استنادی است. در این مطالعه اطلاعات کتابشناختی نویسندگان شامل نام و نام خانوادگی نویسنده، پست الکترونیکی، وابستگی سازمانی، عنوان مقاله و همچنین عنوان نشریه‌ای که نویسنده مقاله‌اش را در آن به چاپ می‌رساند، یک رکورد در نظر گرفته شده است.

به دلیل پیچیدگی مسئله و هزینه‌بربودن آن توسط نیروی انسانی، نیاز به روش‌های خودکار به‌منظور رفع آشفتگی اسامی نویسندگان ضروری به نظر می‌رسد. استفاده از روش‌های مبتنی بر هوش مصنوعی و یادگیری ماشین، راه‌حلی برای این چالش است. یادگیری ماشین به‌عنوان یکی از شاخه‌های وسیع و پرکاربرد هوش مصنوعی، به تنظیم و اکتشاف شیوه‌ها و الگوریتم‌هایی می‌پردازد که بر اساس آنها رایانه‌ها و سامانه‌ها، توانایی تعلّم و یادگیری پیدا می‌کنند. الگوریتم‌های یادگیری ماشین اغلب به‌عنوان با نظارت[1]، بدون نظارت[2] و تقویتی[3] دسته‌بندی می‌شوند (Pal et al., 2013). در این مقاله از طبقه‌بندی که رویکردی با نظارت است، به‌منظور بهینه‌سازی آشفتگی اسامی نویسندگان استفاده شده است. طبقه‌بندی[4] عملیاتی است که سازمان‌ها را قادر می‌سازد، در حل مسائل خاص مجموعه‌های بزرگ و پیچیده به کشف الگوهایی دست یابند. به عبارتی این فرایند، مجموعه داده‌ها را به قسمت‌های مشخص تقسیم می‌کند (Breiman, 2007).

طبقه‌بندهای زیادی تاکنون ارائه شده‌اند که از جمله می‌توان به ماشین بردار پشتیبان[5]، نزدیک‌ترین همسایگی[6]، درخت تصمیم[7] و جنگل تصادفی[8] اشاره کرد. ماشین بردار پشتیبان، به ازای هر نمونه، نقطه‌ای در فضای ویژگی‌ها ترسیم کرده و سعی در به دست‌آوردن یک صفحه جداکننده میان داده‌های آموزشی و آزمایشی می‌کند. این طبقه‌بند به جای استفاده از پارامترهای آماری از پارامترهای هندسی کلاس‌ها استفاده می‌نماید و درواقع یک صفحه بهینه در فضای ویژگی‌ها که بتواند بیشترین جداسازی نمونه‌ها را داشته باشد می‌یابد. درصورتی‌که داده‌ها به‌صورت خطی جداپذیر نباشد، با کرنلی غیرخطی به فضای با ابعاد بالاتر منتقل می‌شود و فراصفحه بهینه در آن فضا را تعیین می‌کند (Wu & Zhou, 2006). الگوریتم نزدیک‌ترین همسایگی یکی دیگر از طبقه‌بندهای معروف در یادگیری ماشین است. این الگوریتم از تشابه ویژگی‌ها برای تشخیص کلاس داده‌ها استفاده می‌کند. بدین صورت که به ازای هرکدام از داده‌های آزمایشی، همسایگان نزدیک به آن را در داده‌های آموزشی یافته و بر اساس برچسب آن داده‌ها، برای داده آزمایشی تصمیم می‌گیرد (Peterson, 2009).

درخت تصمیم، طبقه‌بندی است که تصمیم‌ها و پیامدهای هر تصمیم را در قالب شاخه‌هایی از درختان مدل می‌کند. این مدل از ساختار مشابه با فلوچارت استفاده می‌کند که هر گره داخلی یک تست روی یک ویژگی است و هر شاخه از این گره، نشان‌دهنده یکی از خروجی‌های این تصمیم است. درنهایت برگ‌های این درخت، هرکدام از خروجی‌های طبقه‌بند می‌باشد که حاصل از گذر از حالت‌های مختلف ویژگی‌هاست. مسیرها از ریشه به برگ، قوانین

1 . Supervised Learning
2 . Unsupervised Learning
3 . Reinforcement Learning
4 . Classification
5 . Support vector machine
6 . Nearest neighbour
7 . Decision tree
8 . Random forest

طبقه‌بند را مشخص می‌کند (Myles et al., 2004). در ساخت هر درخت تصمیم، از یک استراتژی پارتیشن‌بندی بازگشتی بالا به پایین استفاده می‌شود. یک درخت تصمیم، فضای ورودی را به مجموعه‌ای از نواحی مجزا تقسیم و یک مقدار پاسخ را به هر ناحیه اختصاص می‌دهد. اگرچه که روش‌های درختی از دیدگاه تفسیر نتایج ساده و موفق عمل می‌کنند؛ ولی محدودیت‌هایی نیز دارند. برای نمونه، میزان اندکی از آشفتگی در داده‌های آموزشی، منجر به ساخت درختی کاملاً متفاوت خواهد شد (Noori, 2011).

جنگل تصادفی با به‌کارگیری چندین درخت و سپس ترکیب نتایج، می‌تواند بر این مشکلات فائق آید. در این روش برای تشکیل هر درخت، دسته متفاوتی از الگوهای موجود، با درنظرگرفتن جایگزینی دوباره هر الگوی انتخاب‌شده ایجاد می‌شوند. اندازه این دسته نمونه‌برداری‌شده، برابر با تعداد کل الگوهای موجود خواهد بود. هر درخت بر اساس دسته الگوی انتخاب‌شده، تا ماکزیمم عمق از پیش‌تعیین‌شده رشد داده می‌شود. این عمق بر اساس حداقل تعداد الگوها در هر گره انتهایی تعیین می‌شود. بر اساس این الگوریتم، در مرحله رشد هر درخت، در هر گره، دسته‌ای از ویژگی‌ها که به‌صورت تصادفی انتخاب می‌شوند و بهترین انشعاب در میان دسته ویژگی انتخاب‌شده برای تشکیل گره‌های جدید بعدی هستند را در نظر می‌گیرد. این طبقه‌بند دقت بسیار بالایی نسبت به دیگر طبقه‌بندها دارد و مطالعات نشان‌دهندهٔ موفقیت این طبقه‌بند در کاربردهای مختلف است (Verikas et al., 2011). به همین دلیل در این پژوهش نیز از این طبقه‌بند استفاده شده است.

پیشینه پژوهش

مزروعی و همکاران (۱۳۹۲) روشی با نظارت به‌منظور دسته‌بندی مقالات با وجود ابهام در داده‌ها ارائه دادند. در این پژوهش یک الگوریتم طبقه‌بند دوکلاسه پیشنهاد شده که متعلق‌بودن و یا متعلق‌نبودن یک مقاله به یک نویسنده را با استفاده از جنگل تصادفی شبیه‌سازی می‌نماید. دهقان، محمودی و قاسم‌پور (۱۳۹۲) در مطالعه‌ای به مدارک نمایه‌شده محققان دانشگاه علوم پزشکی شیراز با آدرس وابستگی سازمانی غیراستاندارد در وب‌آوساینس و اسکوپوس پرداختند. نتایج نشان داد وجود تعداد قابل ملاحظه مدارک با آدرس‌های وابستگی سازمانی غیراستاندارد که منجر به عدم بازیابی مدارک مرتبط می‌گردد، نیاز به سیاست‌گذاری دقیق و صحیح برای یکپارچه‌سازی نام سازمانی دانشگاه و همچنین اطلاع‌رسانی دقیق به پژوهشگران برای پیشگیری از ادامه این روند را دارد؛ چراکه این امر مشکلاتی را در محاسبه تولیدات علمی، شاخص‌های علم‌سنجی و حتی رتبه‌بندی دانشگاه ایجاد می‌نماید. کیانی، داورپناه و فتاحی (۱۳۹۴) در پژوهشی به بررسی تأثیر خطاهای نظام‌مند موجود در طبقه‌بندی موضوعی آی‌اس‌آی[1] بر حجم تولیدات علمی و میزان رؤیت‌پذیری رشته‌ها پرداختند. نتایج نشان داد که نمایه‌شدن نادرست تولیدات علمی به‌نوعی انحراف و خطای نظام‌مند در نتایج حاصل از علم‌سنجی منجر می‌گردد. مرتضوی، ندیمی شهرکی، موسی‌خانی (۱۳۹۶) در پژوهشی به بهبود صحت ابهام‌زدایی نام نویسنده با استفاده از خوشه‌بندی تجمعی پرداخته‌اند. ازآنجاکه پایگاه‌های اطلاعاتی داده‌ها را از منابع مجزا و متعدد به دست می‌آورند؛ ازاین‌رو در ترتیب و کامل‌بودن ویژگی‌ها استانداردی وجود ندارد و همین مسئله منجر به ابهاماتی در این منابع می‌شود که در این میان ابهام نام از اهمیت ویژه برخوردار است. راهکار پیشنهادی در دو گام، عملیات ابهام‌زدایی را انجام می‌دهد. در گام نخست خوشه‌های اولیه با استفاده از "الگوریتم خوشه‌بندی سلسله‌مراتبی تجمعی با پارامترها و توابع اندازه‌گیری مشابهت مختلف" تولید می‌شوند. در گام دوم با بهره‌گیری

1 . ISI

از"الگوریتم خوشه‌بندی تجمعی" خوشه‌های تولیدشده به‌گونه‌ای ترکیب می‌شوند تا خوشه‌هایی با صحت بالاتر تولید شوند. مظفری (۱۴۰۰) روشی برای رفع ابهام نام نویسندگان نشریات انگلیسی با استفاده از الگوریتم ژنتیک ارائه داده است؛ که با استفاده از دو تابع برازش، میزان اهمیت ویژگی‌های استفاده‌شده را به دست می‌آورد. عبدی و نوروزی چاکلی (۱۴۰۰) در پژوهشی به ارزیابی تطبیقی تأثیر کنترل مستندات بر جایگاه بهره‌وری علمی پژوهشگران در پایگاه‌های گوگل‌اسکالر و ریسرچ‌گیت پرداختند. یافته‌ها نشان می‌دهد در پایگاه‌های گوگل‌اسکالر و ریسرچ‌گیت به‌جز چند مورد اصلی مانند نام کوچک، نام بزرگ، وابستگی سازمانی، ایمیل پژوهشگر، ابزار خاصی که بتواند برای احراز هویّت صحیح پژوهشگر راهگشا باشد یافت نشد. همچنین پایگاه‌های گوگل‌اسکالر و ریسرچ‌گیت، آسیب‌ها و خطاهای بسیاری در زمینهٔ کنترل مستندات اسامی متحمل می‌شوند و در زمینهٔ کنترل مستندات، پایگاه‌های معتبری مانند وب‌آوساینس و اسکوپوس نسبت به پایگاه‌های گوگل‌اسکالر و ریسرچ‌گیت از روش‌ها و ابزارهای کنترل گسترده‌تری استفاده می‌کنند. آسیب‌ها و خطاهای قابل توجه احراز هویت پژوهشگران، در دو پایگاه گوگل‌اسکالر و ریسرچ‌گیت نشان‌دهندهٔ عدم به‌کارگیری ابزار مستندسازی و سازمان‌دهی برای حل چالش‌ها و مسائل است که باعث آسیب‌های جدی و خطاهای پیش‌بینی‌نشدهٔ متعددی در زمینهٔ کنترل مستندات نویسندگان می‌شود.

پژوهش‌های مشابهی نیز در خارج از کشور روی رفع آشفتگی اسامی نویسندگان انجام شده است که از آن جمله می‌توان به این موارد، اشاره کرد: در پژوهشی که با استفاده از داده‌های پایگاه مدلاین[1] انجام شد، نشان داده شد که الگوریتم جنگل تصادفی بهتر از ماشین بردار پشتیبان[2]، عمل کرده و تنوعی از مشابهت‌ها را می‌توان در این الگوریتم لحاظ نمود (Treeratpituk & Giles, 2009). نتایج، دقت بالای ۹۰ درصد این الگوریتم در یافتن نویسندگان مشابه را نشان می‌دهد. در پژوهشی دیگر به رفع ابهام نام مخترعان بر اساس الگوریتم جنگل تصادفی پرداخته شد (Kim et al, 2016) که از یک تابع فاصله حاصل از طبقه‌بندی‌کننده جنگل تصادفی، برای خوشه‌بندی سوابق افراد استفاده شد. در آن پژوهش برای مقیاس‌پذیری بیشتر، خوشه‌بندی موازی نیز انجام شده است. نتایج، دقت و سرعت عملکرد این الگوریتم در ابهام‌زدایی اسامی را نشان داد. مدلی دیگر در سال ۲۰۱۷ (Silva & Silva) با استفاده از داده‌های پایگاه آتنتیکس[3] و الگوریتم جنگل تصادفی ارائه شد که داده‌ها را به دو دسته سوابق انتشارات نویسندگان و داده‌های آزمایش تقسیم کردند. اولین مقاله به‌عنوان پایگاه دانش انتشارات قبلی محققان عمل می‌کند و از دیگر مقالات نویسنده برای ارزیابی کیفیت مدل استفاده شده است. سپس، از تابع استخراج ویژگی برای ایجاد یک مجموعه داده با نویسندگان صحیح با عنوان "مطابقت" (مورد مثبت) استفاده شده است. بدین ترتیب برای هر نویسنده صحیح حداکثر ۴ نامزد اشتباه انتخاب شده و به آنها برچسب "بدون همخوانی" زده شده است (مورد منفی). برای انتخاب ۴ نامزد از کل لیست، از همبستگی بین ویژگی‌های هر یک از نامزدها و ویژگی‌های نویسنده واقعی با استفاده از ضریب همبستگی پیرسون استفاده شده است. نتایج، دقت بالای ۹۰ درصد برای نویسندگانی که دارای سابقه در پایگاه داده بودند و دقت بالای ۶۰ درصد را برای نویسندگان بدون سابقه علمی نشان می‌دهد. در سال ۲۰۲۰ ابهام‌زدایی نام نویسندگان در پایگاه پابمد[4] با استفاده از الگوریتم‌های طبقه‌بندی گروهی انجام گرفت (Jhawar et al., 2020). در آن پژوهش با بیان اینکه ابهام در مورد نام نویسنده یک مشکل رایج در کتابخانه‌های دیجیتالی است، مطالعه‌ای تجربی در زمینه

1 . Medline
2 . SVM
3 . Authenticus
4 . PubMed

انتشارات پژوهشی نمایه‌شده بر اساس نام نویسندگانی که به‌صورت عمومی از طریق پایگاه در دسترس بود انجام شد. همچنین دو الگوریتم جنگل تصادفی و درخت تصمیم‌گیری تقویت‌کننده گرادیان[1] برای رفع ابهام اسامی به کار گرفته شد. نتایج نشان داد الگوریتم جنگل تصادفی دقت، بازیافت و نمره اف[2] بالاتری تولید می‌کند، اما درخت تصمیم‌گیری تقویت‌کننده به‌صورت رقابتی عمل می‌کند.

نتایج پژوهش‌های انجام‌شده در خصوص نویسندگان مقالات به زبان انگلیسی، با استفاده از ویژگی‌های متفاوت و به روش‌های با نظارت و بدون نظارت، حاکی از دستیابی به درجات مختلفی از دقت در بهبود آشفتگی اسامی است. ولی یکدستی برابر انگلیسی اسامی، همچنان مشکل‌ساز است؛ به‌ویژه که همه مقاله‌های فارسی باید عنوان، نام انگلیسی نویسندگان و چکیده انگلیسی را نیز ارائه دهند. اما همان‌گونه که مشاهده گردید، پژوهشی که به شکل عملیاتی و کاربردی به بهبود مشکل تنوع و آشفتگی نگارش نام نویسندگان و کنترل مستند آن در مقالات به زبان فارسی و از نگاه علم‌سنجی پرداخته باشد، نتیجه‌ای دربرنداشت؛ غافل از اینکه نادیده‌انگاشتن اسامی نویسندگان مقالات و عدم یکدستی آنها منجر به کاهش دقت و عملکرد پایگاه‌های اطلاعاتی، سامانه‌های بازیابی اطلاعات و موتورهای جستجو و مطالعات حوزه علم‌سنجی خواهد بود.

روش‌شناسی پژوهش

پژوهش حاضر از نوع کاربردی علم‌سنجی است که به روش اسنادی انجام شده است. جامعه آماری اولیه بالغ بر ۱۰ هزار رکورد متشکل از نام نویسندگان مقالات فارسی برگرفته از پایگاه استنادی علوم جهان اسلام، طی بازه زمانی ۱۳۹۵ تا ۱۳۹۷ است. پس از بررسی اولیه و حذف داده‌های تکراری و مواردی که تنوع نگارشی و آشفتگی نداشتند، تعداد ۱۸۲۶ رکورد باقی ماند. سپس این داده‌ها نیز مورد پیش‌پردازش و پاک‌سازی قرار گرفت تا ناسازگاری میان آنها تا حد ممکن رفع گردد. پس از انجام این مرحله، داده‌های نمونه به ۹۱۳ رکورد تقلیل یافت. چارچوب پیشنهادی مبتنی بر الگوریتم‌های هوش مصنوعی و از سه مرحله جستجو[3]، تطابق[4] و گروه‌بندی[5] تشکیل شده؛ اما پیش از آن، پیش‌پردازش داده‌ها[6] و استخراج ویژگی[7] انجام گرفته است. بدین ترتیب که در مرحله پیش‌پردازش، تکنیک‌های پاک‌سازی داده انجام شد. سپس دو نوع ویژگی که شامل ویژگی داخلی که مستقیماً از اطلاعات خود نویسنده (نام، نام خانوادگی، وابستگی سازمانی، پست الکترونیکی نویسنده، نویسندگان همکار و همچنین میزان منحصربه‌فردبودن نام نویسنده[8]) و ویژگی خارجی که با استفاده از اطلاعات نشریه‌ای که نویسنده یا نویسندگان در آن نشریات مقاله(ها)یشان را به چاپ رسانده‌اند (عنوان نشریه، عنوان مقاله، موضوع اصلی و فرعی نشریه) استخراج گردید. در ادامه در مرحله جستجو، رکوردهایی که به‌صورت بالقوه احتمال یکی‌بودن آنها وجود دارد مشخص شدند. برای انجام این کار، با الهام از الگوریتم ساندکس، روشی جدید به نام ساندکس فارسی ارائه شده است که بر اساس نام خانوادگی نویسندگان، اسامی که بالقوه یکسان هستند در یک دسته قرار می‌گیرند. در مرحله تطابق با استفاده از الگوریتم جنگل

1 . gradient boosted decision trees
2 . F-measure
3 . search
4 . match
5 . grouping
6 . Data preprocessing
7 . Feature extraction
8 . Unique name

تصادفی[1] که طبقه‌بندی قوی در حوزه داده‌کاوی است، یکسان‌بودن رکوردها مورد بازبینی دقیق قرار می‌گیرد؛ بنابراین ورودی این مرحله، دسته‌ای از نویسندگان است که احتمال یکسان‌بودن آنها توسط الگوریتم ساندکس فارسی تشخیص داده شده است. یکسان‌بودن یا نبودن این نویسندگان به‌صورت دودویی توسط الگوریتم جنگل تصادفی مورد بررسی قرار گرفته و بنابراین خروجی این مرحله، به ازای هر جفت نویسنده، به‌صورت یک و صفر است. درنهایت، در مرحله گروه‌بندی، تمامی رکوردهایی که با استفاده از جنگل تصادفی یکسان تشخیص داده شده‌اند، در یک گروه قرار می‌گیرند که برای انجام این کار از یک الگوریتم مبتنی بر هش استفاده شده است.

چارچوب پیشنهادی در مراحل جستجو، تطابق و گروه‌بندی

- جستجو

در مرحله جستجو، گروهی از نویسندگان که احتمال یکی‌شدن آنها به‌صورت بالقوه وجود دارد، شناسایی و در یک گروه قرار گرفتند. جهت انجام این کار از الگوریتم ساندکس[2] که یک الگوریتم آوایی است، الهام گرفته شده است است (Lait, & Randell, 1996). این الگوریتم که به نام ساندکس فارسی پیشنهاد شده، قادر است روی داده‌های فارسی، کد ساندکس را تولید کند. ساندکس یک الگوریتم آوایی برای نمایه‌سازی و هش‌کردن حروف و کلمات با صدا به همان نحوی است که تلفظ می‌شود و از ترکیب یک حرف و عددی سه رقمی تشکیل شده است. این الگوریتم با هدف تفکیک آوایی کلمات همسان و همچنین دارای تفاوت املایی جزئی پایه‌ریزی شده و مهم‌تر اینکه با سرعت بسیار خوبی این کار را انجام می‌دهد. با توجه به کاربردهایی که کد ساندکس در بانک‌های اطلاعاتی دارد، این الگوریتم برای پیاده‌سازی مرحله جستجو انتخاب گردید.

کد ساندکس در زبان انگلیسی تعریف شده، ازاین‌رو به‌منظور اعمال این کد در زبان فارسی، باید الگویی ایجاد شود. بر این اساس ازآنجاکه حرف اول ساندکس برابر با همان حرف اول کلمه است، می‌بایست حرف اول آن را از فارسی به انگلیسی تعریف کرد. برای انجام این کار، حروف بر اساس نحوه نوشتار آنها دسته‌بندی شدند (جدول ۱).

جدول ۱. دسته‌بندی حروف فارسی

دسته	حرف	دسته	حرف
S	س، ث، ص، ش	F	ف
Z	ز، ژ، ظ، ذ، ض	H	ح، ه
T	ت، ط	B	ب، پ
Q	غ، ق	V	و
K	خ، ک، گ	N	ن
R	ر	Y	ی
D	د	A	ا، آ، ع
J	ج، چ	M	م
L	ل		

دسته اول شامل حروف ‘س’، ‘ث’، ‘ص’ و ‘ش’ هستند. حروف ‘س’، ‘ث’ و ‘ص’ هر سه با صدای S تلفظ

1 . Random Forest algorithm
2 . Soundex

می‌شوند. حرف ‘ش’ هم برای تبدیل از فارسی به انگلیسی به sh تبدیل می‌گردد و ازآنجاکه حرف اول مدّنظر است؛ بنابراین آن را هم به S نگاشت می‌کنیم.

حروف ‘ز’، ‘ژ’، ‘ظ’، ‘ذ’ و ‘ض’ همه به Z نگاشت شدند؛ چراکه همگی در نگارش از فارسی به انگلیسی به Z تبدیل می‌شوند. حرف ‘ژ’ نیز به Zh تبدیل می‌گردد که باز به دلیل اینکه حرف اول مهم است، Z را در نظر می‌گیریم.

حروف ‘ب’ و ‘پ’ هر دو را به B انتساب دادیم. بدین دلیل که در بعضی مواقع این دو حرف از نظر تلفظ به اشتباه ممکن است به جای یکدیگر استفاده شوند. علاوه‌براین، بعضی مواقع در بعضی سیستم‌ها حرف ‘پ’ به‌اشتباه ‘ب’ نوشته می‌شود.

طبق الگوریتم ساندکس انگلیسی، پس از تبدیل حرف اول، می‌بایست بقیه حروف به یک عدد نگاشت گردند. در این راستا شماره‌های ۱ تا ۹ برای نشان‌دادن دسته‌بندی‌های مختلف استفاده شد. پس از آن اگر عدد مربوطه کمتر از سه رقم بود، با قرارگیری اعداد صفر در قبل از آن، سه رقم مربوطه تکمیل گردید (جدول ۲).

جدول ۲. دسته‌بندی حروف فارسی به همراه کد اختصاص داده‌شده به هر گروه

کد	حرف
۱	س، ص، ث، ش، ز، ذ، ظ، ض
۲	د، ت، ط
۳	خ، غ، ق
۴	ر، ل
۵	ج، چ
۶	ف، و
۷	ح، ه
۸	ب، پ
۹	بقیه

درنهایت، حروفی که از نظر آوایی به هم نزدیک هستند، در یک دسته قرار می‌گیرند و با استفاده از یک کد سریع، اسامی که به یکدیگر شبیه و یا حتی با اشتباهات املایی کوچک هستند، کدهای شبیه به یکدیگر دریافت خواهند کرد. خروجی این مرحله (جستجو) یک جدول است که کد نویسنده به همراه کد سریع آن را نشان می‌دهد و بر اساس نام خانوادگی هر نویسنده به دست آمده است؛ بنابراین، با انجام فرایند فوق به هر نام یک کد اختصاص یافته که با استفاده از آن همان‌طور که پیش‌تر اشاره شد، اسامی که به یکدیگر شبیه و یا حتی دارای اشتباهات کوچک املایی هستند، کدهایی مشابه هم دریافت خواهند کرد.

- تطابق با استفاده از جنگل تصادفی و ارائه الگوریتم پیشنهادی یکدست‌سازی

ورودی این مرحله، دسته‌ای از رکوردهای نویسندگان است که احتمال یکسان‌بودن آنها به‌صورت بالقوه وجود دارد. در این مرحله، بررسی دقیق‌تری روی داده‌های موجود در هر دسته انجام می‌گیرد تا مشخص شود کدام جفت رکورد از نویسندگان درواقع متعلق به یک نویسنده است. بدین منظور از یک طبقه‌بند قوی در حوزه یادگیری ماشین به نام جنگل تصادفی استفاده شده است که با استفاده از داده‌های آموزشی یاد می‌گیرد و بر آن اساس، برای داده‌های

آزمایشی تصمیم‌گیری می‌کند. بدین منظور فضای مسئله بر اساس ویژگی‌ها به نواحی مختلفی تقسیم‌بندی می‌شود. به عبارت دیگر، هر بار زیرمجموعه‌ای از ویژگی‌ها انتخاب شده و روی هرکدام از این ویژگی‌ها یک درخت تصمیم با توجه به داده‌های آموزشی یاد داده می‌شود. درخت تصمیم ساخته‌شده روی هرکدام از این نواحی قادر به طبقه‌بندی داده‌ها با توجه به همان ویژگی‌هاست. سپس جنگل تصادفی از این درخت‌های تصمیم استفاده کرده و نتایج آنها را با یکدیگر ترکیب می‌کند. به عبارت دیگر، مدل پیش‌بینی‌کننده جنگل تصادفی بر اساس میانگین‌گیری از نتایج حاصل از تمامی درخت‌های تصمیم مربوطه استوار خواهد بود.

ازآنجاکه پست الکترونیکی یک ویژگی بسیار مؤثر در تشخیص یکسان و یا عدم یکسان‌بودن رکوردهای مختلف نویسندگان است، درصورتی‌که به ازای دو رکورد مختلف، این ویژگی موجود و کاملاً برابر بود، روش پیشنهادی، دو رکورد را یکسان درنظرمی‌گیرد. لازم به ذکر است که این ویژگی به ازای تمامی رکوردها، ممکن است موجود نباشد یا در صورت وجود کاملاً یکسان نباشند که در این صورت بر اساس تطابق مبتنی بر جنگل تصادفی توضیح داده شده، و در مورد رکوردها تصمیم‌گیری می‌شود.

- گروه‌بندی

در این مرحله، تمامی نویسندگان یکسان با استفاده از خروجی مرحله قبل در یک گروه قرار می‌گیرند. در مرحله قبل به ازای هر دو رکورد نویسنده، خروجی یک یا صفر به دست آمد که نشان‌دهنده یکسان‌بودن یا نبودن دو رکورد مورد نظر است. در مرحله گروه‌بندی، به ازای هر رکورد یک کد خوشه، به معنای خوشه نویسنده است، تخصیص داده می‌شود. بدین ترتیب که به ازای هر زوج نویسنده‌ای که یکسان هستند و یا برچسب یک دارند، درصورتی‌که کد نویسنده، قبلاً در مجموعه کدها موجود نباشد، کد خوشه جدید به آن کد نویسنده تعلق می‌گیرد. سپس کد نویسنده دوم بررسی و چنانچه این کد قبلاً در مجموعه کدها موجود باشد و کد خوشه متناظر با آن با کد خوشه نویسنده اول متفاوت باشد، کد خوشه‌ها به‌روزرسانی می‌شود. درنهایت روی تمامی رکوردهای نویسندگانِ موجود، این یکسان‌سازی انجام گیرد و تمامی رکوردهایی که کد خوشه آنها یکسان باشند، به این معناست که متعلق به یک نویسنده بوده و در پایگاه داده به دلایل مختلف به‌عنوان نویسندگان مجزا ذخیره شده بودند.

معیارهای ارزیابی

ارزیابی روش پیشنهادی، با استفاده از روش اعتبارسنجی ضربدری انجام شده است. این روش مستقل از داده‌های آموزشی است و تعیین می‌کند، نتایج یک تحلیل آماری روی مجموعه‌ای از داده‌ها تا چه اندازه قابل تعمیم است. اعتبارسنجی ضربدری شامل تقسیم داده‌ها به دو زیرمجموعه مکمل، انجام تحلیل روی مجموعه داده آموزشی و سپس اعتبارسنجی با استفاده از داده‌های تست است. به‌منظور کاهش پراکندگی، اعتبارسنجی چندین بار با بخش‌های مختلف انجام گرفته و سپس میانگین این نتایج به‌عنوان نتیجه نهایی گزارش می‌شود.

در روش اعتبارسنجی ضربدری k لایه[1]، ابتدا داده‌ها به k زیرمجموعه تقسیم می‌شود. سپس از این k زیرمجموعه، هر بار یکی برای اعتبارسنجی و k-1 مجموعه دیگر برای آموزش مورد استفاده قرار می‌گیرند. این روند k بار تکرار می‌شود تا بدین ترتیب همه داده‌ها شانسی برای عضویت در داده‌های آموزشی و تست پیدا کنند. ثابت شده است که مقدار k=10 بهترین مقداری است که می‌توان به نتایج دقیق و قابل اعتماد دست پیدا کرد (Breiman et al., 2010)؛

1 . K-fold

بنابراین در این مطالعه مجموعه‌ای از داده‌ها که برچسب آنها مشخص است، یعنی مشخص است که آیا متعلق به یک نویسنده هستند یا خیر، به 10 بخش تقسیم می‌شوند. روش پیشنهادی بر اساس ۹ بخش یاد می‌گیرد و سپس روی یک بخش باقی‌مانده به‌عنوان مجموعه تست، ارزیابی انجام می‌گیرد.

درنهایت، برای ارزیابی روش پیشنهادی روی داده‌های آموزشی و تست از دقت، بازیافت و مقدار اف استفاده می‌شود.

دقت، مشخص می‌کند که از میان تعداد جفت نویسندگانی که روش پیشنهادی آنها را یکسان در نظر گرفته است (نمونه‌های مثبت)، چند درصد واقعاً متعلق به یک نویسنده بوده‌اند (فرمول ۱).

فرمول ۱)
$$\text{دقت} = \frac{\text{مدارک بازیابی شده مربوط}}{\text{مدارک بازیابی شده}}$$

بازیافت، نشان می‌دهد که روش پیشنهادی تا چه اندازه توانسته است از میان تعداد نویسندگانی که واقعاً متعلق به یک موجودیت هستند، نویسندگان یکسان را پیدا کند (فرمول ۲).

فرمول ۲)
$$\text{بازیافت} = \frac{\text{مدارک مربوط بازیابی شده}}{\text{مدارک مربوط}}$$

درنهایت، مقدار اف به‌منظور ایجاد یک توازن میان این معیار به‌صورت زیر تعریف می‌شود (فرمول ۳).

فرمول ۳)
$$\text{مقدار اف} = \frac{2\times\text{دقت}\times\text{بازیافت}}{\text{دقت} + \text{بازیافت}}$$

یافته‌های پژوهش

پاسخ به سؤال اول پژوهش: کدام‌یک از ویژگی‌های استفاده‌شده در تشخیص و بهبود آشفتگی نگارش نام‌های نویسندگان مقالات فارسی، نسبت به دیگر ویژگی‌ها از اهمیت بیشتری برخوردار هستند؟

برای پاسخ به این سؤال که یکی از مهم‌ترین قدم‌ها در حوزه هوش مصنوعی و یادگیری ماشین است، ابتدا با توجه به داده‌ها، ویژگی‌های مختلفی استخراج شد. این ویژگی‌ها شامل دو نوع ویژگی داخلی و خارجی است. ویژگی داخلی مستقیماً از اطلاعات خود نویسنده استخراج می‌گردد که شامل نام، نام خانوادگی، وابستگی سازمانی، پست الکترونیکی نویسنده، نویسندگان همکار و همچنین میزان منحصربه‌فردبودن نام نویسنده است. ویژگی خارجی از اطلاعات نشریه‌ای که نویسنده مقاله(ها)یش را در آن چاپ کرده است، شامل عنوان نشریه، عنوان مقاله، موضوع اصلی و فرعی نشریه می‌باشد؛ بنابراین درنهایت ۱۰ ویژگی استخراج شد.

به‌منظور بررسی تأثیر ویژگی‌های مختلف در کارایی روش پیشنهادی، یک مجموعه در نظر گرفته و دقت روش پیشنهادی با ویژگی‌های موجود در این مجموعه بررسی گردید. ابتدا تک‌تک ویژگی‌ها، در این مجموعه قرار گرفته و دقت روش پیشنهادی را با توجه به آن محاسبه نمودیم. ویژگی برنده‌شده انتخاب می‌شود و تمام حالت‌های ممکن با دومین ویژگی، محاسبه شده و مجدداً با دقت محاسبه می‌گردد. این روند ادامه پیدا می‌کند و هر بار برترین مجموعه انتخاب می‌شود تا درنهایت تمامی ویژگی‌ها بررسی شوند.

در مرحله اول، بهترین ویژگی، پست الکترونیک نویسنده است. به عبارت دیگر این ویژگی از اهمیت بسیار زیادی برخوردار است؛ چراکه در صورت یکسان‌بودن این ویژگی برای دو رکورد، می‌توان با قطعیت اعلام کرد که دو رکورد

مختلف، واقعاً متعلق به یک نفر است؛ ولی درصورتی‌که این ویژگی یکسان نباشد، نمی‌توان در مورد یکسان‌بودن و یا نبودن دو رکورد از نویسنده تصمیم‌گیری کرد و باید از ویژگی‌های دیگر کمک گرفته شود. نتایج شبیه‌سازی روی داده‌های واقعی نیز مؤید این مسئله است.

ویژگی‌های بعدی که تأثیرگذاری بیشتری در تشخیص و رفع آشفتگی دارند، به‌ترتیب نام خانوادگی و نام نویسنده است؛ چراکه در مرحله جستجو، ابتدا آن دسته از نویسندگانی که با توجه به ویژگی نام خانوادگی تقریباً مشابه بودند، در یک دسته قرار گرفتند و ویژگی نام نیز خاصیت جداکنندگی بسیار زیادی را برای رکوردهای مختلف نویسندگان به وجود می‌آورد.

چهارمین ویژگی که میزان اهمیت آن با آزمایشات مشخص شد، وابستگی سازمانی است. به عبارت دیگر با استفاده از این ویژگی، می‌توان جداسازی بهتری از اسامی نویسندگان داشت. این امر نیز دور از انتظار نبود؛ زیرا با داشتن این ویژگی می‌توان محل تحصیل یا کار نویسنده را تعیین و تعداد بسیار زیادی از نویسندگان با این ویژگی قابل شناسایی خواهند شد. ویژگی‌های بعدی، به‌ترتیب درصد شباهت عناوین نشریات چاپ‌شده توسط دو نویسنده و همچنین عناوین مقالات آنهاست. به عبارت دیگر هر چقدر شباهت عناوین نشریات و مقالات چاپ‌شده توسط دو رکورد از نویسندگان بیشتر باشد، احتمال اینکه دو نویسنده واقعاً متعلق به یک موجودیت باشند بیشتر هست.

ویژگی بعدی، میزان منحصربه‌فردبودن نام است. هر چقدر نام یک نویسنده خاص‌تر باشد، راحت‌تر می‌توان یکسان‌سازی انجام داد. ازآنجاکه این نتایج روی داده‌های آموزشی مورد بررسی انجام گرفته است که درصد اسامی خاص در این لیست از نویسندگان کم بود؛ بنابراین ممکن است در مجموعه داده دیگری که اسامی خاص در آنها بیشتر باشد، این ویژگی اهمیت خود را پررنگ‌تر نشان دهد و قدرت تمییزکنندگی بیشتری را داشته باشد.

ویژگی بعدی که آزمایشات آن را بهتر از بقیه در نظر گرفته، نویسندگان مشترک با یک نویسنده است. درصد نویسندگان مشترک دو نویسنده می‌تواند به‌عنوان یک ویژگی مهم در شناسایی نویسندگان باشد. هر چقدر تعداد نویسندگان مشترک دو نویسنده بیشتر باشد، احتمال اینکه این دو نویسنده متعلق به یک موجودیت باشند، یا به عبارتی یک نفر باشند، بیشتر است.

دو ویژگی با درجه اهمیت کمتر از بقیه، درصد شباهت میان موضوع اصلی و فرعی نشریه نویسندگان است.

پاسخ به سؤال دوم پژوهش: الگوریتم جنگل تصادفی به چه میزان می‌تواند در تشخیص و بهبود آشفتگی نگارش اسامی نویسندگان مقالات فارسی مؤثر واقع گردد؟

نتیجه اجرای روش پیشنهادی با جنگل تصادفی روی داده‌های فارسی نشان می‌دهد؛ که این روش با دقت بیش از ۹۹ درصد در بهبود آشفتگی و به‌عبارتی یکسان‌سازی نام نویسندگان به زبان فارسی به شکل بسیار مؤثری عمل کرده است. این روش منجر به بهبود مشکلاتی از جمله کوتاه‌نویسی نام و نام خانوادگی، غلط تایپی، قراردادن یا ندادن خط فاصله بین دو بخش نام و نام خانوادگی و غیره شده؛ که در کدگذاری و بازیابی اطلاعات جامع مشکلاتی را ایجاد می‌کند (جدول ۳).

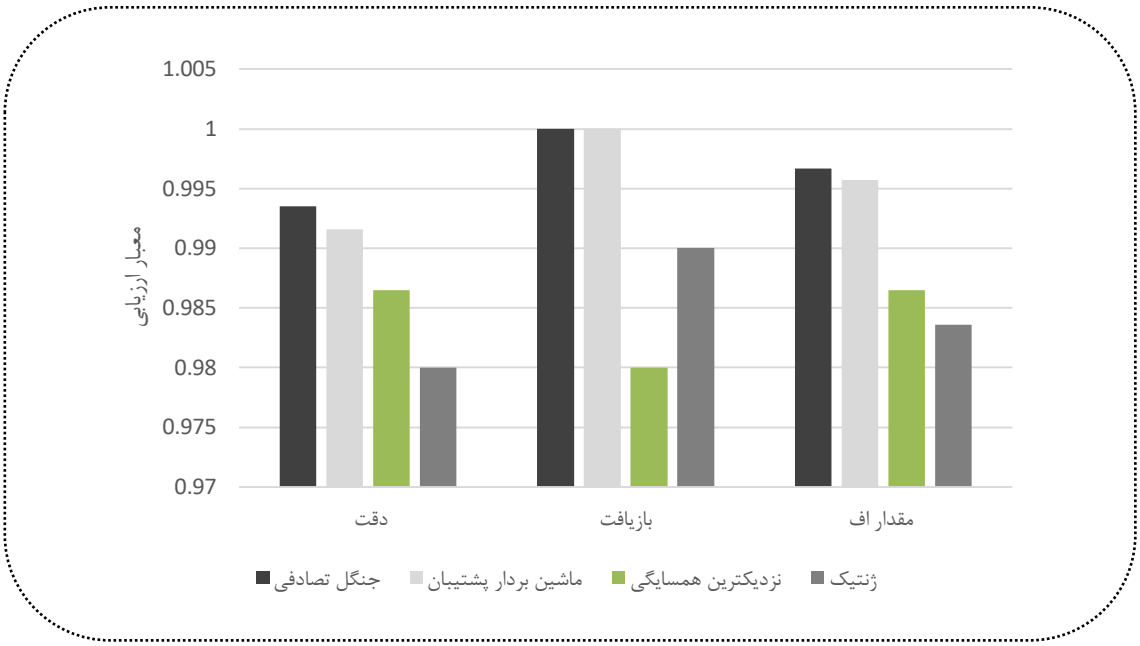
جدول ۳. نتیجه اجرای الگوریتم جنگل تصادفی روی داده‌های فارسی

	تعداد نمونه‌ها	دقت	بازیافت	مقدار اف
داده آموزشی	۸۲۲	۰.۹۹۳۴	۱	۰.۹۹۶۵
داده آزمایشی	۹۱	۰.۹۹۳۵	۱	۰.۹۹۶۷

همان‌گونه که پیش‌تر اشاره شد، برای ارزیابی به روش ۱۰ لایه، کل داده‌ها به ۱۰ قسمت تقسیم و هر بار یکی برای تست و بقیه برای آموزش مورد استفاده قرار گرفته است. بر این اساس تعداد ۹۱۳ رکورد بر ۱۰ تقسیم و هر بار ۹۱ رکورد برای تست و ۸۲۲ رکورد برای آموزش استفاده شدند. این روند ۳۰ بار انجام و درنهایت میانگین این ۳۰ بار گزارش شده است. بر این اساس، چارچوب روش پیشنهادی قادر است با دقت بسیار بالایی، وجود آشفتگی در اسامی نویسندگان مقالات فارسی را برطرف نماید.

پاسخ به سؤال سوم پژوهش: استفاده از الگوریتم جنگل تصادفی به‌عنوان الگوریتم تطابق برای تشخیص و بهبود آشفتگی نگارش اسامی نویسندگان مقالات فارسی، به چه میزان موجب بهبود دقت در مقایسه با دیگر طبقه‌بندها می‌شود؟

به‌منظور مقایسه عملکرد روش پیشنهادی با تطابق مبتنی بر جنگل تصادفی نسبت به دیگر طبقه‌بندهای معروف در حوزه یادگیری ماشین، آن را با دو طبقه‌بند ماشین بردار پشتیبان و نزدیک‌ترین همسایگی و همچنین روشی مبتنی بر ژنتیک (Mozafari, 2021) مقایسه کردیم (شکل ۱).



شکل ۱. مقایسه دقت، بازیافت و مقدار اف روش پیشنهادی با تطابق مبتنی بر جنگل تصادفی با طبقه‌بندهای دیگر

تمامی روش‌های مورد استفاده، از روندی یکسان برای بهینه‌سازی آشفتگی اسامی نویسندگان استفاده می‌کنند و تنها تفاوت آنها در مرحله تطابق است. ازآنجاکه تطابق مبتنی بر ماشین بردار پشتیبان، به ازای هر رکورد نویسنده، صفحه‌ای جداکننده در فضای ویژگی‌های داخلی و خارجی که بیشترین جداسازی را میان نمونه‌ها دارد رسم می‌کند؛ بنابراین دقت بسیار بالاتری نسبت به تطابق با ژنتیک و نزدیک‌ترین همسایگی دارد. ولی همان‌طور که این شکل نشان می‌دهد، تطابق مبتنی بر درخت تصادفی نسبت به دیگر روش‌ها از دقت، بازیافت و مقدار اف بالاتری برخوردار است که این امر نشان‌دهنده قدرت جنگل تصادفی نسبت به دیگر طبقه‌بندهای معروف می‌باشد. این توضیح لازم است که تمامی این آزمایش‌ها ۳۰ بار انجام گرفته و میانگین نتایج اعلام شده است.

بحث و نتیجه‌گیری

نام، یک ویژگی کلیدی برای تمایز بین افراد است. در پایگاه‌های اطلاعاتی جستجوی نام نویسنده/نویسندگان مقالات، یکی از مهم‌ترین عناصر در بازیابی منابع منتسب به یک فرد، افزایش رؤیت‌پذیری، مطالعات کمّی حوزه علم‌سنجی از جمله میزان استناد به آثار، ارزیابی فعالیت‌های علمی و پژوهشی یک نویسنده بر اساس تعداد تألیف‌ها و تعیین برون‌داد علمی، محاسبه برخی شاخص‌ها و غیره است. اما تنوع نگارشی و چگونگی ضبط آنها در نوشته‌ها، از مسائلی است که منجر به ایجاد چالش‌هایی در عرصه‌های مختلف علمی می‌شود. ریشه این چالش‌ها را می‌توان در مراحل مختلف چرخه حیات علمی یک مدرک، از مرحله تولید تا مرحله درون‌دهی عبارتِ جستجو یافت. هرچند واگرایی و عدم وحدت رویه در نگارش اسامی به دلیل ویژگی‌های زبانی، فرهنگی و اشتباهات تایپی جزء امور رایج است (کاواشیما و تامیزاوا، ۲۰۱۵)؛ اما نبود استاندارد نگارشی در زبان فارسی و رفتار سلیقه‌ای نویسندگان، نبود صفحه‌کلید و کدهای استاندارد، عادت به ساده‌نویسی و رعایت‌نکردن پیچیدگی‌های نگارشی از عوامل به‌وجودآمدن چنددستگی در نگارش اسامی است. همچنین اشتباهات املایی که توسط نویسندگان در نگارش نام رخ می‌دهد، نیز منجر به ایجاد صور مختلف نگارشی برای یک نام واحد می‌شود.

در پژوهش حاضر به روش جنگل تصادفی، با استفاده از یادگیری مبتنی بر نمونه برای رفع ابهام اسامی، الگوریتمی جهت بهبود آشفتگی اسامی پدیدآورندگان مقالات به زبان فارسی ارائه شده است. بر این اساس پس از شناسایی موارد ذکرشده تلاش بر این است، تا با استفاده از الگوریتم جنگل تصادفی این امر بهبود بخشیده شود. همچنین علاوه بر وجود جداول آوانگاری بین‌المللی، جدول دادگان و جدول دگرنویسی" کتابخانه کنگره" برای برگرداندن حروف به‌عنوان سه مرجع، اما در این پژوهش ساندکس فارسی پیشنهاد شده است که می‌تواند به‌عنوان الگویی جهت برگرداندن حروف در مرحله پیش‌پردازش و پاک‌سازی در پایگاه داده‌ها استفاده گردد.

نتایج نشان می‌دهد تطابق مبتنی بر درخت تصادفی نسبت به دیگر روش‌ها از دقت، بازیافت و مقدار اف بالاتری برخوردار است که این امر نشان‌دهنده قدرت جنگل تصادفی نسبت به دیگر طبقه‌بندهای معروف است. همچنین این روش برای داده‌های بسیار بزرگ قابلیت اجرا دارد؛ چراکه جنگل تصادفی از دسته روش‌هایی است که از چندین طبقه‌بند که در اینجا درخت تصمیم است، استفاده می‌کند. هرکدام از این درختان، فضای ورودی را به مجموعه‌ای از نواحی تقسیم می‌نماید و بر اساس هر ناحیه، یک تصمیم گرفته می‌شود. درنهایت جنگل تصادفی از میانگین این نتایج استفاده می‌کند. به همین دلیل می‌تواند نتایج به نسبت دقیق‌تری را ارائه نماید.

پیشنهادهای اجرایی پژوهش

- امکان ایجاد مستند اسامی مشاهیر و نویسندگان با استفاده از روش پیشنهادی پژوهش؛
- اعمال روش پیشنهادی پژوهش در بانک داده پایگاه‌های استنادی و علمی، جهت یکدست‌سازی اسامی نویسندگان.

پیشنهاد برای پژوهش‌های آتی

- بررسی شبکه همکاری نویسندگان به‌عنوان یک ویژگی در حل مشکل وجود آشفتگی در اسامی نویسندگان؛
- ارائه چارچوب پیشنهادی روش برای داده‌های برخط (مستلزم بررسی توزیع داده‌ها و وزن‌دهی به طبقه‌بندها در جنگل تصادفی با توجه به داده‌های جدید)؛

- توسعه ساندکس فارسی به‌عنوان مرجعی جهت برگرداندن اسامی فارسی به انگلیسی و ایجاد یکدستی در برگردان اسامی ایرانی در پایگاه‌های علمی.

فهرست منابع

خسروی، عبدالرسول (۱۳۸۳). ضرورت مستندسازی موضوع‌ها و نام‌های فارسی در محیط اینترنت. *پیام بهارستان*. ۴۱، ۸-۱۱.

خسروی، مریم (۱۳۹۰). آشفتگی نگارش نام پدیدآورندگان ایرانی در پایگاه اطلاعاتی آی.اس.آی. فصلنامه علمی پژوهشی پژوهشگاه علوم و فناوری اطلاعات ایران، ۴، ۴۵-۶۵.

دهقان، شیرین؛ محمودی، زلیخا؛ قاسم‌پور، محمد (۱۳۹۲). مدارک نمایه‌شدهٔ محققین دانشگاه علوم پزشکی شیراز با آدرس وابستگی سازمانی غیراستاندارد در Web of Science و Scopus. *مدیریت اطلاعات سلامت*. ۱۰ (۶): ۸۱۰-۸۱۸.

زلفی‌گل، محمدعلی؛ شیری، مرتضی و کیانی بختیاری، ابوالفضل (۱۳۸۶). اهمیت رعایت اصول نمایه‌سازی در مستندات علمی. *رهیافت*، ۳۹، ۳۷-۴۶.

صادقی گورجی، شهربانو؛ پوراحمد، علی‌اکبر؛ حاجی زین‌العابدینی، محسن و ضیایی، ثریا (۱۳۹۴). ارزیابی کارآمدی گوگل پژوهشگر در بازیابی اطلاعات نویسندگان دارای شکل‌های گوناگون نام: بررسی ضریب بازیافت و دقت. *پژوهشنامه کتابداری و اطلاع‌رسانی*، ۵ (۱)، ۲۰۵-۲۱۶.

عبدی، ساجده؛ نوروزی چاکلی عبدالرضا؛ اسدی سعید (۱۴۰۰). ارزیابی تطبیقی تأثیر کنترل مستندات بر جایگاه بهره‌وری علمی پژوهشگران در پایگاه‌های گوگل‌اسکالر و ریسرچ‌گیت. پژوهش‌نامه علم‌سنجی 7(13), 203-216.

کیانی، حمیدرضا؛ داورپناه، محمدرضا؛ فتاحی، رحمت‌الله (۱۳۹۴). بررسی تأثیر خطاهای نظام‌مند موجود در طبقه‌بندی موضوعی آی‌اس‌آی بر حجم تولیدات علمی و میزان رؤیت‌پذیری رشته‌ها. پژوهش‌نامه کتابداری و اطلاع‌رسانی. ۵ (۲): ۲۸۴-۲۶۳.

مرتضوی، سید محمد؛ ندیمی شهرکی، محمدحسین؛ موسی خانی، مصطفی (۱۳۹۶). بهبود صحت ابهام‌زدایی نام نویسنده با استفاده از خوشه‌بندی تجمّعی. *پردازش علائم و داده‌ها*، ۱۴ (۴)،۱۱۷-۱۲۸.

مزروعی سبدانی، نصیرالدین؛ ابراهیم‌پور کومله، حسین و نیک‌فرجام، علی‌محمد (۱۳۹۲). ارائه روش بانظارت به‌منظور دسته‌بندی مقالات با وجود ابهام در داده‌ها. دوازدهمین کنفرانس سیستم‌های هوشمند ایران، مجتمع آموزش عالی بم.

مظفری نیلوفر (۱۴۰۰). ارائه روشی مبتنی بر ژنتیک برای رفع ابهام نام نویسندگان مقالات. پژوهشنامه پردازش و مدیریت اطلاعات. ۳۶ (۳): ۸۱۶-۷۹۱.

Abdi, S., & Chakoli, A. N., Asadi, S. (2021). The comparative evaluation of authority control impact on the Iran researchers scientific productivity situation in the Google Scholar and ResearchGate. Scientometrics Research Journal. DOI: 4.4773.2019.rsci/22070.10 [In Persian]

Bhattacharya, I., & Getoor, L. (2006, April). A latent dirichlet model for unsupervised entity resolution. In *Proceedings of the 2006 SIAM International Conference on Data Mining* (pp. 47-58). Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611972764.5

Breiman, L., Friedman, J., Olsen, R., & Stone, C. (2010). Classification and Regression Trees (Wadsworth and Brooks/Cole, Monterey, CA, 1984).

Cota, R. G., Ferreira, A. A., Nascimento, C., Gonçalves, M. A., & Laender, A. H. (2010). An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations. *Journal of the American Society for Information Science and Technology*, *61*(9), 1853-1870. DOI: 10.1002/asi.21363

Dehghan, Sh., Mahmoodi, Z., Ghasempour, M. (2013). Indexed documents of researchers of Shiraz University of Medical Sciences with non-standard affiliation in Web of Science and Scopus, *Health Information Management*, 10(6):810-818. [In Persian]

Fan, X., Wang, J., Pu, X., Zhou, L., & Lv, B. (2011). On graph-based name disambiguation. *Journal of Data and Information Quality (JDIQ)*, *2*(2), 1-23. DOI: 10.1145/1891879.1891883

Ferreira, A. A., Veloso, A., Gonçalves, M. A., & Laender, A. H. (2010, June). Effective self-training author name disambiguation in scholarly digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp. 39-48). DOI: 10.1145/1816123.1816130

Jhawar, K., Sanyal, D. K., Chattopadhyay, S., Bhowmick, P. K., & Das, P. P. (2020, August). Author Name Disambiguation in PubMed using Ensemble-Based Classification Algorithms. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (pp. 469-470). DOI: 10.1145/3383583.3398568

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004, June). Two supervised learning approaches for name disambiguation in author citations. In *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries, 2004.* (pp. 296-305). IEEE. DOI: 10.1145/996350.996419

Huynh, T., Hoang, K., Do, T., & Huynh, D. (2013, March). Vietnamese author name disambiguation for integrating publications from heterogeneous sources. In *Asian Conference on Intelligent Information and Database Systems* (pp. 226-235). Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-642-36546-1_24

Kang, I. S., Na, S. H., Lee, S., Jung, H., Kim, P., Sung, W. K., & Lee, J. H. (2009). On co-authorship for author disambiguation. *Information Processing & Management*, *45*(1), 84-97. DOI: 10.1016/j.ipm.2008.06.006

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics*, *103*(3), 1061-1071. DOI: 10.1007/s11192-015-1580-z

Khosravi, A. (2004), The necessity of documenting Persian topics and names in the Internet environment, Payam Baharestan, 41:8-11.[In Persian]

Khosravi, M. (2011). The confusion of Iranian Author Names in ISI database. Scientific Research of Iran Research Institute of Science and Information Technology, 4:46-65. [In Persian]

Kiani, H., Davarpanah, M., Fattahi, R. (2015). Investigating the impact of systematic errors in the subject classification of ISI on the volume of scientific productions and the degree of visibility of fields. *Library and Information Science Research*, 5(2): 263-284. [In Persian]

Kim, K., Khabsa, M., & Giles, C. L. (2016). Random forest dbscan for uspto inventor name disambiguation. *arXiv preprint arXiv:1602.01792*. DOI: 10.48550/arXiv.1602.01792

Kim, J., & Kim, J. (2020). Effect of forename string on author name disambiguation. *Journal of the Association for Information Science and Technology*, 71(7), 839-855. DOI: 10.1002/asi.24298

Lait, A. J., & Randell, B. (1996). An assessment of name matching algorithms. *Technical Report Series-University of Newcastle Upon Tyne Computing Science*.

Mazroyi Sabadani, N., Ebrahimpour Komleh, H., Nikfarjam, A. (2013). A supervised approach for classification of papers with data ambiguation, 12th Iranian Conference on Intelligent Systems. Bam. [In Persian]

Mortazavi, S. M., Nadimi Shahraki, M. H., Mosakhani, M. (2017). Improving the accuracy of the author name disambiguation by using clustering ensemble. JSDP. 2018; 14 (4) :117-128. DOI: 10.29252/jsdp.14.4.117 [In Persian]

Mozafari, N. (2021). A Genetic-based Approach for Author Name Disambiguation Problem. *Iranian Journal of Information Processing Management*, *36*(3), 791-816. DOI: 10.52547/jipm.36.3.791. [In Persian]

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society, 18(6), 275-285. DOI: 10.1002/cem.873

Noori, A. (2011, July). On the relation between centrality measures and consensus algorithms. In *2011 International Conference on High Performance Computing & Simulation* (pp. 225-232). IEEE. DOI: 10.1109/HPCSim.2011.5999828

On, B. W., Elmacioglu, E., Lee, D., Kang, J., & Pei, J. (2006, December). Improving grouped-entity resolution using quasi-cliques. In *Sixth International Conference on Data Mining (ICDM'06)* (pp. 1008-1015). IEEE. DOI: 10.1109/ICDM.2006.85

Pal, A., R., A. Munshi, and D. Saha. (2013). An approach to speed-up the word sense disambiguation procedure through sense filtering. *International journal of Instrumentation and Control systems* (IJICS). 3(4), 29-41. DOI: 10.5121/ijics.2013.3403

Sadeghi Gouraji, Sh., Pourahman, A., Hajizeinolabedini, M., Zeiaei, S. (2015), Evaluation of the Effectiveness of Google Scholar in Authors' Information Retrieval Library and Information Science Research. 5(1): 205-2016. DOI: 10.22067/RIIS.V5I1.24674 [In Persian]

Shin, D., Kim, T., Jung, H., & Choi, J. (2010, April). Automatic method for author name disambiguation using social networks. In *2010 24th IEEE International Conference on Advanced Information Networking and Applications* (pp. 1263-1270). IEEE. DOI: 10.1109/AINA.2010.66

Silva, J. M., & Silva, F. (2017, April). Feature extraction for the author name disambiguation problem in a bibliographic database. In *Proceedings of the Symposium on Applied Computing* (pp. 783-789). DOI: 10.1145/3019612.3019663

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *3*(3), 1-29. DOI: 10.1145/1552303.1552304

Treeratpituk, P., & Giles, C. L. (2009, June). Disambiguating authors in academic publications using random forests. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries* (pp. 39-48). DOI: 10.1145/1555400.1555408

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern recognition*, 44(2), 330-349. DOI: 10.1016/j.patcog.2010.08.011

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. Expert systems with applications, 38(1), 223-230. DOI: 10.1016/j.eswa.2010.06.048

Zhang, B., & Al Hasan, M. (2017, November). Name disambiguation in anonymized graphs using network embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1239-1248). DOI: 10.1145/3132847.3132873

Zolfigol, M.A., Shiri, M., Kiani Bakhtiari, A. (2007). The importance of observing the principles of indexing in scientific documents, Rahyaft, 39:37-46. [In Persian]