A Genetic-based Approach for Author Name Disambiguation Problem

Niloofar Mozafari
PhD in Artificial Intelligence; Assistant Professor; Regional Information Center for Science and Technology; Islamic World Science Citation Center; 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. 36 | No. 3 | pp. 791-816
Spring 2021

Received: 19, Apr. 2020 **Accepted: 07, Oct. 2020**

Abstract: In the recent years, with the increasing volume of articles and the use of Internet and search engine services, the author name disambiguation problem has received a lot of attention. Name disambiguation can occur when one is seeking a list of publications of an author who has used different name variations and also when there are multiple other authors with the same name. So far, various methods have been proposed to solve this problem, each of which has its own advantages and disadvantages. Despite years of research, the name disambiguation problem remains largely unresolved. In this study, we propose an algorithm to identify several records that belong to one author. For this purpose, a new criterion has been proposed to determine the similarity between the two records. Since this study addresses the approximate matching of authors' records, the importance of the fields in each record is determined by the coefficients. In order to get the optimal coefficients, we propose a genetic algorithm to learn from the available samples. The proposed method has been evaluated with two fitness functions on experimental data and the results are promising.

Keywords: Name Disambiguation Problem, Levenshtein Distance, Genetic Algorithm, Fitness Function

ارائهٔ روشی مبتنی بر ژنتیک برای رفع ابهام نام نویسندگان مقالات

نیلوفر مظفری

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

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۹/۰۱/۳۱ **پذیرش:** ۱۳۹۹/۰۷/۱۶ مقاله برای اصلاح به مدت ۳ ماه و نیم نزد پدیدآوران بوده است.

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۶ | شمارهٔ ۳ | ص‌ص ۷۹۱-۸۱۶
بهار ۱۴۰۰

چکیده: امروزه، با افزایش روزافزون حجم مقالات از یک طرف و استفاده از اینترنت و خدمات موتورهای جست‌وجو از طرف دیگر، روش‌های ابهام‌زدایی از اسامی پژوهشگران بسیار مورد توجه قرار گرفته است. تاکنون روش‌های مختلفی برای حل این مشکل ارائه شده که هر یک مزایا و معایب خاص خود را دارند. هدف این مقاله ارائهٔ راهکاری جهت شناسایی رکوردهای متعددی است که به یک نویسنده تعلق دارند. بدین‌منظور بعد از استخراج ویژگی‌های داخلی و خارجی نویسندگان، یک معیار جدید جهت مشخص کردن میزان مشابهت میان دو رکورد ارائه شده است. اهمیت هر یک از ویژگی‌های ارائه‌شده با استفاده از الگوریتمی مبتنی بر ژنتیک با دو تابع برازش مختلف تعیین می‌شود تا از طریق یادگیری نمونه‌های موجود بهینه‌ترین ضرایب به‌دست آید. روش پیشنهادی با دو تابع برازش روی داده‌های آزمایشی مورد ارزیابی و مقایسه قرار گرفته و نتایج حاصل نشان‌دهندهٔ افزایش دقت در روش پیشنهادی با هر دو تابع برازش نسبت به روش قبلی است.

کلیدواژه‌ها: ابهام نام نویسندگان، فاصله لونشتین، الگوریتم ژنتیک، تابع برازش

۱. مقدمه

امروزه، عناوین در کتابخانه‌های دیجیتال به نهادهای پیچیده‌ای تبدیل گشته‌اند که قادر به انجام فعالیت‌های گسترده از جمله بازیابی اطلاعات و همچنین توانایی مدیریت و ارزیابی هستند. کتابخانه‌های دیجیتال قادر به ارائهٔ خدمات بسیار زیادی به کاربران در جهت جست‌وجو و بازیابی اسناد مرتبط با پرسش کاربر هستند به‌صورتی که کاربر قادر است با مکانیزم‌های جست‌وجوی مختلف موجود در کتابخانه‌ها، اسناد مرتبط با پرسش خود را بازیابی نماید. همچنین، کاربر می‌تواند پژوهش‌های نویسندهٔ دلخواهش را جست‌وجو کند و این امر نشان‌دهندهٔ قدرت بالای کتابخانه‌های دیجیتال در انتشار دانش میان پژوهشگران است. این است که در سال‌های اخیر پژوهش‌های زیادی روی این حوزه انجام گرفته است (قاسمی الوری و چشمه‌سهرابی ۱۳۹۹).

کتابخانه‌های دیجیتال با بهره‌گیری از فناوری‌های جدید در تلاش هستند تا کیفیت خدمات خود را بهبود بخشند. در عصر اطلاعاتی رقابتی امروز، برای راضی نگه‌داشتن کاربران می‌بایست از فناوری‌های جدید بیش از پیش استفاده کرد (رزمی شندی، نوروزی و علیپور حافظی ۱۳۹۹). از طرف دیگر، امروزه رشد انفجارگونهٔ داده‌ها و استفاده از آن‌ها یکی از چالش‌های کتابخانه‌ها و مراکز اطلاع‌رسانی است. افزایش حجم دستاوردهای علمی نویسندگان مختلف موجب شده است که سازماندهی اطلاعات با دشواری‌هایی همراه شود. یکی از این چالش‌ها تحت عنوان ابهام نام نویسندگان[1] شناخته می‌شود. ابهام نام نویسندگان به معنای وجود ابهامات مختلف در نگارش اسامی نویسندگان است که می‌تواند ناشی از عوامل مختلف باشد.

ابهام نام نویسندگان به‌طور عمده خود را به دو صورت نشان می‌دهد که ما آن را هم‌مرجعی و چندمرجعی می‌نامیم. در واقع، هم‌مرجعی به معنای وجود یک نگاشت یک به چند میان نویسندگان و اسامی است (یک نویسنده با اسامی مختلف). به‌عبارت دیگر، گاهی اوقات نام یک نویسنده به دو یا چند شکل مختلف در مقالات مختلف دیده می‌شود که می‌تواند ناشی از وجود اسامی مختلف برای یک نویسنده باشد. به‌عنوان مثال، نویسنده‌ای با نام Mohammad-Ali Hasani Shirazi ممکن است در مقاله‌های مختلف به شکل‌های Mohammad Hasani، M. Hasani Shirazi، Mohammad-Ali Shirazi، Hasani Shirazi و غیره وجود داشته

1. author name disambiguation

باشد. در تمامی این حالت‌ها باید بتوان تمامی این اسامی را به یک نویسنده نسبت داد. افزون بر این، مشکل مورد نظر می‌تواند ناشی از خطای تایپی مانند Mohamad-Ali Hasni Shirazi نیز باشد.

مشکل بعدی و یا چندمرجعی به معنای نگاشت چند به یک میان نویسندگان و اسامی آن‌هاست (چند نویسنده با نام یکسان). فرض کنید دو نویسندهٔ مختلف با نام یکسان Mohammad-Ali Hasani Shirazi وجود داشته باشد که یکی از آن‌ها در حوزهٔ علوم کامپیوتر و دیگری در حوزهٔ علم اطلاعات و دانش‌شناسی فعالیت دارد. این نام به‌رغم یکسان‌بودن، متعلق به دو شخص متفاوت است و سیستم باید بتواند میان این دو تمایز قائل شود.

وجود ابهام در اسامی نویسندگان به‌طور مستقیم روی کارایی و کاربرپسندبودن[1] کتابخانه‌های دیجیتال تأثیر منفی می‌گذارد. به‌عبارت دیگر، کارایی کتابخانه‌های دیجیتال را به‌دلیل کاهش بازیابی و همچنین، میزان کاربرپسندبودن را به‌دلیل عدم شناسایی مقالات نویسندهٔ خاصِ مورد جست‌وجوی کاربر نهایی کاهش می‌دهد. یکسان‌سازی نام نویسندگان شامل کلیهٔ راهکارهایی است که کمک می‌کند تمامی حالاتی که به بروز مشکل در درج نام نویسنده منجر می‌شود، کاهش یابد.

پژوهش‌هایی که در این حوزه انجام می‌شود، می‌تواند راهگشای این مشکلات باشد که عمدتاً به دو دستهٔ کلی تکنیک‌های مبتنی یادگیری ماشینی[2] و تکنیک‌های مبتنی بر روش‌های غیریادگیری ماشینی[3] تقسیم‌بندی می‌شوند (Hussain & Asghar 2017). تکنیک‌های مبتنی بر یادگیری ماشینی بر اساس مشاهدات قبلی یک مدل را ساخته و سپس، به پیش‌بینی داده‌های جدید می‌پردازند که خود به سه دسته روش‌های یادگیری با نظارت[4]، یادگیری بدون نظارت[5] و یادگیری نیمه‌نظارتی[6] تقسیم می‌شوند (Wang et al. 2011؛ Imran, Syed Gillani and Marchese 2013؛ Huynh et al. 2013؛ Han et al. 2015؛ Seol et al. 2016 و مرتضوی، ندیمی شهرکی و موسی‌خانی ۱۳۹۶). پژوهش‌هایی که از روش‌های غیر یادگیری ماشین برای یکسان‌سازی نام استفاده می‌کنند، به دو دستهٔ تکنیک‌های مبتنی بر گراف[7] و روش‌های اکتشافی[8] طبقه‌بندی می‌شوند (Fan et al. 2011؛ Wang et al. 2011؛ Tang et al. 2011 و Shin et al. 2014).

1. user-friendly
2. machine learning based techniques
3. non-machine learning based techniques
4. supervised learning
5. unsupervised learning
6. semi-supervised learning
7. graph-based techniques
8. heuristic

در مطالعهٔ حاضر، یک روش بانظارت جهت رفع ابهام اسامی نویسندگان ارائه شده است که روی هر دو مشکل هم‌مرجعی و چندمرجعی قابل اعمال است. افزون بر این، به‌دلیل استفاده از روش‌های یادگیری در تخمین پارامترها، روش پیشنهادی نیازی به تنظیم پارامترها توسط انسان ندارد و به‌صورت خودکار این پارامترها را از خود داده‌ها یاد می‌گیرد. همچنین، نیازی به دانستن تعداد نویسندگان مبهم در داده‌ها وجود ندارد.

به‌طور خلاصه، در این مقاله روشی برای حل مسئلهٔ ابهام نام نویسندگان ارائه شده است. ابهام در نام نویسندگان به معنای وجود شکل‌های مختلف اسامی یک نویسنده و یا وجود چند نویسنده با یک نام است. ابتدا، اطلاعات مختلف نویسندگان و همچنین مقالات منتشرشده توسط آن‌ها وارد سامانه می‌شود. روش پیشنهادی بعد از پردازش این اطلاعات و استخراج ویژگی‌ها، عملیات جست‌وجو را انجام می‌دهد. خروجی این فرایند گروه‌هایی از نویسندگان است؛ به‌طوری که هر گروه، اسامی نویسندگانی را شامل می‌شود که بیشترین شباهت را با یکدیگر دارند. به‌عبارت دیگر، خروجی این مرحله آن اسامی را می‌یابد که احتمال یکی‌بودن آن‌ها به‌صورت بالقوه وجود دارد. سپس، تطبیق[1] در هر گروه، یکی‌بودن دو نام را بررسی می‌کند و در نهایت، در هر گروه نام نویسندگانی که متعلق به یک موجودیت هستند، بازیابی می‌شود. به‌منظور به‌دست آوردن اهمیت ویژگی‌های استفاده‌شده در این مقاله، الگوریتمی مبتنی بر ژنتیک ارائه گردیده است.

در این مقاله به‌دنبال پاسخ به سؤالات زیر هستیم:

- ◇ میزان اهمیت ویژگی‌های لازم برای رفع ابهام اسامی نویسندگان چگونه است؟
- ◇ میزان بهبود روش پیشنهادی نسبت به روش‌های قبلی برای حل مشکل ابهام نام نویسندگان چگونه است؟

۲. پیشینهٔ پژوهش

امروزه، عناوین در پژوهش‌هایی که از تکنیک‌های یادگیری بانظارت برای یکسان‌سازی اسامی استفاده می‌کنند، طبقه‌بند[2] را بر اساس داده‌های آموزشی آموزش داده و سپس، کارایی مدل را بر اساس داده‌های تست مورد ارزیابی قرار می‌دهند. یکی از این پژوهش‌ها از طبقه‌بند درختی تقویت‌شده[3] برای یکسان‌سازی نام نویسندگان استفاده می‌کند که از چهار مرحله تشکیل

1. match
2. classifier
3. boosted tree classifier

می‌شود (Wang et al. 2012). در مرحلهٔ اول، یک پیش‌پردازش روی نام و آدرس دانشگاهی نویسنده بر اساس تطابق مستقیم نام و نام خانوادگی و آدرس دانشگاهی انجام می‌گیرد. در مرحلهٔ بعد، با تعریف یک معیار شباهت، میزان شباهت نویسنده‌ها مشخص می‌شود. مرحلهٔ سوم یک غربالگری با استفاده از نرخ اشتباه انجام می‌دهد، و در مرحلهٔ آخر طبقه‌بند درختی تقویت‌شده روی داده‌ها اعمال می‌شود. روش پیشنهادی مبتنی بر طبقه‌بند درختی تقویت‌شده نمی‌تواند اسامی نویسندگان با نرخ بالای اشتباه را طبقه‌بندی کرده و به بررسی دستی توسط انسان نیاز دارد.

پژوهشی دیگر مبتنی بر شبکهٔ عصبی عمیق[1] در سال ۲۰۱۴ ارائه گردید که ویژگی‌ها را به‌صورت خودکار یاد می‌گیرد و ابهام اسامی را برطرف می‌نماید (Tran, Huynh and Do 2014)، بدین‌ترتیب که در مرحلهٔ اول، داده‌ها به‌عنوان ورودی به الگوریتم داده شده و نحوهٔ نمایش داده‌ها محاسبه می‌گردد، و در مرحلهٔ دوم، ویژگی‌های پایه را گرفته و ویژگی‌ها را در لایه‌های پنهان یاد می‌دهد تا ابهام اسامی را برطرف نماید. آخرین لایهٔ شبکهٔ عصبی عمیق احتمال یکسان‌بودن دو نمونه از اسامی را محاسبه می‌کند. یکی از مشکلات این روش به‌دست آوردن تعداد بهینهٔ لایه‌های پنهان است.

از جملهٔ دیگر پژوهش‌هایی که از روش‌های یادگیری با نظارت استفاده می‌کنند، می‌توان به پژوهشی که در سال ۲۰۱۵ انجام شده، اشاره کرد (Han et al. 2015). آن‌ها به ازای هر یک از نام‌های موجود در پایگاه داده یک طبقه‌بند ایجاد کرده و با توجه به اسامی موجود در یک مقاله، از مقاله‌های دارای نویسندگان با اسامی مبهم رفع ابهام می‌کنند. آن‌ها همچنین، به‌منظور کاهش ابعاد ویژگی‌ها از آنالیز مؤلفهٔ اصلی[2] استفاده می‌کنند.

پنج تکنیک یادگیری ماشین با نظارت که شامل جنگل تصادفی[3]، ماشین بردار پشتیبان[4]، k همسایه نزدیک[5]، درخت تصمیم‌گیری[6] و نایوبیز[7] است، برای حل مسئلهٔ ابهام نام نویسندگان در سال ۲۰۱۳ ارائه گردید (Huynh et al. 2013). آن‌ها مجموعه‌ای از ویژگی‌ها را از دادهٔ نشریات استخراج کرده و بعد از آموزش طبقه‌بندها با کمک داده‌های آموزشی، کارایی روش‌ها را با کمک داده‌های تست مورد بررسی قرار دادند که البته، کارایی روش آن‌ها به تمیز[8] بودن داده‌ها وابسته است.

1. deep neural network
2. principal component analysis
3. random forest
4. support vector machine
5. K-nearest neighbours
6. decision tree
7. Naïve bayes
8. clean

یک روش با نظارت به‌منظور دسته‌بندی مقالات با وجود ابهام در داده‌ها توسط «مزروعی سبدانی، ابراهیم‌پور کومله، و نیک‌فرجام» ارائه گردید که دارای دو مرحلهٔ اصلی پیش‌پردازش و دسته‌بندی است. در واقع، یک الگوریتم طبقه‌بند دوکلاسه پیشنهاد شده که متعلق‌بودن و یا متعلق‌نبودن یک مقاله به یک نویسنده را با استفاده از جنگل تصادفی شبیه‌سازی می‌نماید (۱۳۹۲).

پژوهش‌های بیشتر در این حوزه روی ارائهٔ تابع پیوند، به‌ویژه دقت آن متمرکز شده‌اند (Tejada, Knoblock and Minton 2001؛ Han et al. 2004؛ Torvik et al. 2005؛ Bekkerman & Mccallum 2005). «تجادا، نوبلوک و مینتون» از یک درخت تصمیم به‌منظور یادگیری قوانین نگاشت بر اساس شباهت صفات میان رکوردها استفاده کردند (Tejada, Knoblock and Minton 2001). «هان» و همکاران دو طبقه‌بند نایوبیز ترکیبی و ماشین بردار پشتیبان را به‌منظور رفع ابهام نام نویسندگان در پایگاه داده DBLP ارائه کردند (Han et al. 2004). در پژوهشی دیگر «هوآنگ، ارتکین و جیلز» از خوشه‌یابی مبتنی بر چگالی به‌منظور دسته‌بندی کردن نویسندگان CiteSeerX بر اساس فردادهایی همچون وابستگی، پست الکترونیک، آدرس، دگرگونی‌های اسامی و URLهای مقاله‌ها استفاده کردند. لازم به ذکر است که در پژوهش مذکور از تابع فاصلهٔ مبتنی بر ماشین بردار پشتیبان به‌منظور تابع شباهت دوتایی استفاده شده است (Huang, Ertekin and Giles 2006). «سونگ» و همکاران یک الگوریتم رفع ابهام مبتنی بر موضوع بر اساس PLSA و LDA ارائه کردند تا ابهام اسامی نویسندگان را بر اساس محتوای مقاله‌ها بیابند (Song et al. 2007).

در پژوهش‌های پیشین به‌منظور یکسان‌سازی نام نویسندگان، نیاز به استفاده از یک معیار شباهت است. این روش‌ها عمدتاً بر پایهٔ کدگذاری آوایی[1] و تطابق الگو[2] هستند. تکنیک‌های مختلفی تاکنون برای هر کدام از راه‌حل‌ها ارائه شده است. همچنین، روش‌هایی پیشنهاد شده است که این دو دسته از راه‌حل‌ها را با هم ترکیب کرده و کیفیت نتیجه را بهبود بخشد.

روش‌های کدگذاری آوایی تلاش می‌کنند یک رشته نام را عمدتاً بر اساس نحوهٔ تلفظش به یک کد تبدیل کنند. بدیهی است که این روش‌ها وابسته به زبان[3] هستند. یکی از قدیمی‌ترین و معروف‌ترین شیوه‌های کدگذاری آوایی، «ساندکس»[4] (Lait &

1. phonetic encoding 2. pattern matching 3. language dependent 4. Soundex

Randell 1996) است که اولین‌بار در سال ۱۹۹۶ توسط Zobel & Dart مطرح شد. این روش، اولین کاراکتر را ذخیره و بقیهٔ کاراکترها را به یک عدد تبدیل می‌کند. کدگذاری‌های دیگری نیز برای بهبود کارایی «ساندکس» ارائه گردیدند. از آن جمله می‌توان به «فانکس»[1] (Holmes & McCabe 2002) و «فانیکس»[2] (Hodge & Austin 2003) اشاره کرد. «فانکس» قبل از کدگذاری، یک مجموعهٔ پیش‌پردازش را روی نام انجام می‌دهد، اما در «فانیکس» با تعریف بیش از یک‌صد قوانین تبدیل حروف، پیش‌پردازش بیشتری نسبت به «فانکس» انجام می‌گیرد. این مجموعه قوانین باعث می‌شود که دقت آن نسبت به روش‌های قبلی بهتر شود، ولی از طرفی به‌دلیل تعریف این مجموعه نسبتاً زیاد از قوانین، پیش‌پردازش طولانی خواهد شد و بنابراین، نسبت به روش‌های قبلی کندتر عمل می‌کند.

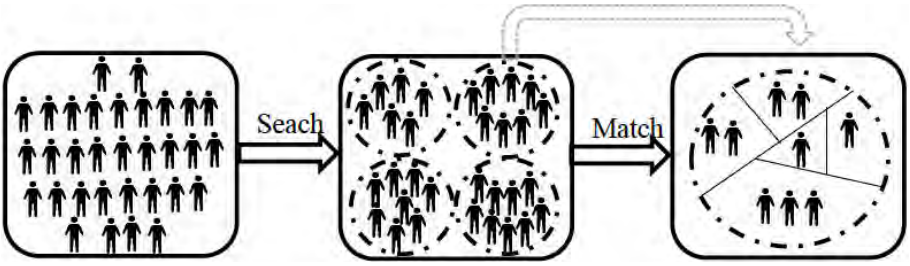
یکی دیگر از شیوه‌های کدگذاری «نایسیس»[3] (Sayers 2018) است که مشابه «فانکس» و «فانیکس»، بر اساس تبدیل حروف به اعداد بر اساس یک مجموعه از قوانین کار می‌کند؛ اما کدی که تولید می‌کند فقط بر اساس حروف است. «متافون» جفتی[4] (Philips 2000) الگوریتم کدگذاری دیگری است که تلاش می‌کند روی کلمه‌های غیر انگلیسی نیز کار کند. مشابه «نایسیس»، کد ساخته‌شده فقط بر اساس حروف بوده و علاوه بر این، مشابه «فانیکس»، قوانین بسیار زیادی دارد که موقعیت هر حرف را در کلمه به همراه حروفی که بعد از آن حرف آمده است، در نظر می‌گیرد. بر خلاف دیگر تکنیک‌های کدگذاری، در مواردی خاص به ازای بعضی از کلمات دو کد به‌عنوان خروج تولید می‌شود.

تکنیک‌های تطابق الگو به‌صورت متداول برای تخمین تطابق رشته به‌کار می‌روند که کاربردهای بسیار زیادی در حوزه‌های مختلف دارند. فاصلهٔ «لونشتین»[5] (Navarro 2001) یکی از تکنیک‌های متداول در این دسته است که به‌عنوان کمترین تعداد عملیاتی است که یک رشته را به دیگری تبدیل می‌کند. این عملیات می‌توانند اضافه[6]، حذف[7] و تعویض[8] باشند. این معیار هر یک از عملیات را با هزینهٔ یک در نظر می‌گیرد. این فاصله می‌تواند به یک معیار شباهت در فاصله بین صفر و یک تبدیل گردد. در صورتی که دو رشته کاملاً یکسان باشند، فاصله لونشتین آن‌ها، عدد یک و در صورتی که هیچ ارتباطی با یکدیگر نداشته باشند، این فاصله صفر است.

1. Phonex
2. Phonix
3. NYSIIS
4. double metaphone
5. Leveshtein
6. insertion
7. deletion
8. substitutions

۳. روش پژوهش

روش پیشنهادی از سه مرحلهٔ انتخاب ویژگی، جست‌وجو و تطابق تشکیل شده است. شکل ۱، چارچوب کلی روش پیشنهادی را نشان می‌دهد. در ادامه، هر کدام از این مراحل با جزئیات شرح داده می‌شود.



شکل ۱. چارچوب کلی روش پیشنهادی متشکل از دو فاز جست‌وجو و تطابق

۳-۱. انتخاب ویژگی

انتخاب ویژگی یکی از گام‌های اصلی در داده‌کاوی است که به جرأت می‌توان گفت نقش بسیار مهمی در موفقیت و یا شکست عملیات داده‌کاوی دارد. در این مطالعه از دو نوع ویژگی استفاده شده است که آن‌ها را ویژگی‌های داخلی[1] و خارجی[2] می‌نامیم. دلیل این نامگذاری این است که ویژگی‌های داخلی به‌طور مستقیم از اطلاعات خود نویسنده استفاده می‌کنند که شامل نام[3]، نام خانوادگی[4]، وابستگی دانشگاهی[5]، پست الکترونیکی[6] نویسنده، نویسندگان همکار[7] وی و همچنین، میزان منحصربه‌فرد بودن نام نویسنده[8] است. در بخش ویژگی‌های خارجی از اطلاعات نشریه‌ای که نویسنده یا نویسندگان در آن مقاله(ها) یشان را به چاپ رسانده‌اند، استفاده می‌شود. این ویژگی‌ها شامل گروه علمی و زیرگروه نشریه برای نشریات وزارت علوم[9] و موضوع کلی و موضوعات اختصاصی برای نشریات وزارت بهداشت[10] است. در ادامه، توضیحاتی در مورد هر یک از ویژگی‌ها ارائه می‌شود.

نام و نام خانوادگی نویسنده از مهم‌ترین ویژگی‌هایی هستند که معمولاً جست‌وجو

1. internal features
2. external features
3. first name
4. last name
5. affiliation
6. email address
7. co-author
8. unique name
9. https://journals.msrt.ir
10. https://journals.research.ac.ir/

بر اساس آن‌ها صورت می‌گیرد. این نام‌ها که می‌توانند به‌صورت نگاشت یک به چند میان نام و نویسنده باشد (بدین معنا که هر نام می‌تواند متعلق به چندین نویسنده باشد)، عمدتاً باعث ابهام در شناسایی نویسندگان می‌شود. پست الکترونیک هر نویسنده ویژگی است که با کمک آن می‌توان شناسایی نویسندگان را با دقت بالایی انجام داد. فقط مشکل موجود در داده‌ها این است که این ویژگی به ازای تمامی نویسندگان تعریف نشده و تنها برای تعداد کمی از نویسندگان در پایگاه داده وجود دارد. همچنین، پست الکترونیک نویسنده ممکن است در گذار زمان تغییر کند. هر نویسنده بسته به وابستگی که به یکی از دانشگاه‌ها و یا سازمان‌ها دارد، دارای یک وابستگی دانشگاهی است که البته، ممکن است در گذر زمان نیز این وابستگی دانشگاهی و حتی پست الکترونیک وی تغییر کند. بنابراین، این دو ویژگی به‌صورت یک بردار تعریف می‌شود که هر عضو آن نشان‌دهندهٔ یک وابستگی دانشگاهی و پست الکترونیک در یک برههٔ زمانی است.

برای به‌دست آوردن ویژگی نام و نام خانوادگی به ازای دو نویسنده، از فرمول ۱، استفاده می‌شود:

$$f_n = 1 - \left({}^{lev(x_1, x_3)}/_{max_{len(x_1,x_2)}}\right) \quad (۱)$$

در این فرمول، f_n در واقع، ویژگی است که با توجه به نام و نام خانوادگی به‌دست می‌آید و $lev(x_1, x_3)$ فاصلهٔ «لونشتین»[1] (Navarro 2001) میان دو رشته است.

ویژگی بعدی که این مقاله آن را در نظر گرفته، میزان شباهت وابستگی سازمانی دو نویسنده است. یک نویسنده ممکن است چندین وابستگی سازمانی در زمان‌های مختلف داشته باشد (به‌عنوان مثال، دانشجوی دکتری که در یک دانشگاه فارغ‌التحصیل شده و در یک دانشگاه دیگر به‌عنوان هیئت عملی مشغول به کار است). بنابراین، وابستگی سازمانی هر نویسنده به‌عنوان یک متغیر وابسته به زمان تعریف می‌شود که وابستگی سازمانی نویسنده را در زمان‌های مختلف نشان می‌دهد.

برای به‌دست آوردن شباهت میان دو متغیر از وابستگی سازمانی می‌بایست شباهت تک‌تک وابستگی‌های سازمانی متغیر اول و دوم در زمان‌های مختلف را با یکدیگر مقایسه نماییم و ماکزیمم این شباهت را گزارش کنیم. به‌منظور به‌دست آوردن شباهت میان

1. Levenshtein

یک وابستگی سازمانی با دیگری، ابتدا عملیات نشانه‌گذاری انجام می‌شود. سپس، یک مجموعه از کلمات را که در تمامی وابستگی‌های سازمانی وجود دارد، از مجموعهٔ کلمات حذف می‌کنیم. این لیست از کلمات به‌صورت مجموعه {university, department, and, of, city, the, school, faculty, for, a, an} تعریف می‌شوند. لازم به ذکر است که این کلمات به‌صورت حروف بزرگ و کوچک در نظر گرفته می‌شوند. سپس، فاصلهٔ جاکارد[1] (Niwattanakul et al., 2013) میان دو وابستگی سازمانی را به‌عنوان شباهت میان دو وابستگی سازمانی گزارش می‌کنیم.

یکی دیگر از ویژگی‌هایی که با کمک آن می‌توان یک نویسنده را شناسایی کرد، نویسندگان همکار وی 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 و یا به‌عبارتی، یک نفر باشند، بیشتر است. برای به‌دست آوردن لیست نویسندگان مشترک میان دو نویسندهٔ مورد نظر، نویسندگان همکار نویسندهٔ اول در تمامی مقالات را به‌دست آورده و اشتراک این مجموعه را با نویسندگان همکار نویسندهٔ دوم تهیه می‌کنیم و در نهایت، بر مینیمم طول این دو لیست تقسیم کرده و به‌عنوان ویژگی‌های همکار دو نویسنده گزارش می‌کنیم.

آخرین ویژگی داخلی که این پژوهش از آن استفاده کرده، میزان منحصربودن نام یک نویسنده است. هر چقدر نام یک نویسنده خاص‌تر باشد، یکسان‌سازی را راحت‌تر می‌توان انجام داد. این ویژگی به‌صورت زیر تعریف می‌شود.

$$uniq = log_{N/2}^{N/x} \qquad (۲)$$

که در آن N تعداد کل مقالات است و x تعداد مقالاتی است که یکی از نویسندگان آن با این نام باشد. هر چقدر این مقدار بیشتر باشد، آن نام منحصربه‌فردتر است و احتمال این‌که مورد ابهام قرار بگیرد، کمتر.

این پژوهش از چهار ویژگی خارجی استفاده می‌کند که بر اساس اطلاعات نشریه به‌دست می‌آید. از آنجا که هر نویسنده ممکن است مقالات مختلف را در چندین نشریه

1. Jaccard

با موضوعات مختلف به چاپ رسانده باشد، بنابراین، تمام ویژگی‌های خارجی هر نویسنده به‌صورت یک بردار تعریف می‌شود. به‌عنوان مثال، ویژگی زیرگروه با یک بردار تعریف می‌شود که هر عضو این بردار مربوط به یکی از نشریاتی است که نویسنده مقاله‌اش را در آن به چاپ رسانیده است.

هر نویسنده مقاله‌ای را در نشریه‌ای به چاپ می‌رساند. بنابراین، از حوزهٔ تحقیقاتی نشریه که آن را موضوع اصلی و فرعی می‌نامیم، می‌توان استفاده کرد. این دو ویژگی در نشریات وزارت علوم با گروه علمی و زیرگروه و برای نشریات وزارت بهداشت با موضوع کلی و موضوعات اختصاصی مشخص شده‌اند. دو ویژگی خارجی دیگری که در این پژوهش از آن استفاده شده، درصد شباهت عناوین نشریه و همچنین، درصد شباهت عناوین مقالات دو نویسنده است. برای به‌دست آوردن این ویژگی‌ها از معیار «جاکارد» استفاده می‌کنیم که در واقع، نسبت اشتراک به اجتماع اعضای دو لیست است.

۳-۲. جست‌وجو

در مرحلهٔ جست‌وجو گروهی از نویسندگان که به‌طور بالقوه احتمال یکی‌شدن آن‌ها وجود دارد، شناسایی شده و در یک گروه قرار می‌گیرند. در این مرحله از الگوریتم «ساندکس» که یک الگوریتم آوایی است، استفاده شده است. «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 یک عدد سه رقمی تشکیل شده است. این الگوریتم در واقع، با هدف تفکیک آوایی کلمات همسان و دارای تفاوت املایی جزئی پایه‌ریزی شده و کاربردهایی در بانک‌های اطلاعاتی مرتبط با پرونده‌های سرشماری، ثبت احوال، پاسخگویی اطلاعات تلفن و غیره دارد. با توجه به این‌که کد «ساندکس» اشتباهات آوایی را در نظر می‌گیرد و همچنین، اشتباهات املایی بسیار جزئی را هم می‌تواند پوشش دهد و مهم‌تر این‌که با سرعت بسیار خوبی این کار را انجام می‌دهد، برای پیاده‌سازی مرحلهٔ جست‌وجو انتخاب گردید.

حرف اول در «ساندکس» همیشه همان حرف اول کلمه است. شماره‌های باقی‌مانده از ۱ تا ۶ نشان‌دهندهٔ دسته‌بندی‌های مختلف است و اگر عدد مربوط کمتر از سه رقم باشد، با قرارگرفتن اعداد صفر قبل از آن، سه رقم مربوط تکمیل می‌شود. همچنین، در صورتی که عدد مربوط بیشتر از سه رقم باشد، اعداد اضافی از سمت راست حذف می‌شوند.

جدول ۱، لیست حروف به همراه کدی را که به هر گروه از حروف انتساب داده می‌شود، نشان می‌دهد. از این پس کد را به نام کد سریع[1] می‌نامیم.

جدول ۱. دسته‌بندی حروف به همراه کد اختصاص داده‌شده به هر دسته

کد	۱	۲	۳	٤	٥	٦
حرف	B, F, P, V	C, G, J, K, Q, S, X, Z	D, T	L	M, N	R

در این مرحله یک جدول ایجاد می‌گردد که حاوی کد نویسنده و کد سریع آن نویسنده است که با توجه به نام خانوادگی وی به‌دست می‌آید. بنابراین، خروجی مرحلهٔ جست‌وجو دسته‌بندی نویسندگان بر اساس کد سریع آن‌هاست، به‌صورتی که تمامی نویسندگانی که کد سریع آن‌ها یکی باشد، در یک گروه یا دسته قرار گیرد.

۳-۳. تطابق

در این مرحله، تمامی نویسندگان موجود در یک گروه یا دسته بررسی شده و شباهت آن‌ها با استفاده از فرمول ۳ محاسبه می‌گردد. در صورتی که این شباهت از یک مقدار آستانه بیشتر باشد، می‌توان این نتیجه را گرفت که این دو نامِ نویسنده متعلق به یک موجودیت هستند.

$$sim(R_i.R_j) = \sum_{k=1}^{M}[1 - \frac{d(R_{i.k}.R_{j.k})}{\max\{l_{R_{i.k}}.l_{R_{j.k}}\}}]W_k \qquad (۳)$$

در این فرمول R_i و R_j رکوردهای iام و jام از پایگاه داده هستند. M تعداد ویژگی‌ها را نشان می‌دهد. $l_{R_{i.k}}$ و $l_{R_{j.k}}$ طول ویژگی kام از رکورد iام و jام است. W_k بردار وزن است؛ به‌صورتی که هر عنصر آن میزان اهمیت ویژگی kام را نشان می‌دهد. $d(R_{i.k}.R_{j.k})$ فاصلهٔ دو ویژگی از یک رکورد را مشخص می‌نماید. الگوریتم ۱ روند کار مرحلهٔ تطابق روش پیشنهادی مقاله را نشان می‌دهد.

مقادیر W_k در الگوریتم ۱، میزان اهمیت هر کدام از ویژگی‌ها و θ مقدار آستانه را نشان می‌دهد. به‌منظور به‌دست آوردن این مقادیر، روش پیشنهادی از الگوریتم‌های تکاملی الهام گرفته است تا از طریق یادگیری از نمونه‌های موجود، بهینه‌ترین ضرایب محاسبه گردد.

1. FCODE

Algorithm1

Input: Records in each cluster (R)

Output: label

1: $label = \emptyset$

2: $s = \emptyset$

3: for i=1 to len(R) do:

4: for j=1 to len(R) do:

5: if emailAdrrs(Ri)== emailAdrrs(Rj)

6: label=1

7: return label

8: end if

9: for k=1 to M do:

10: $s = s + [1 - \frac{d(R_{i.k}.R_{j.k})}{max\{l_{R_{i.k}}.l_{R_{j.k}}\}}]W_k$

11: $end\ for$

12: $if\ s \geq \theta\ do$:

13: label=1

14: $end\ if$

15: return label

16: $end\ for$

17: $end\ for$

الگوریتم‌های ژنتیک به‌عنوان یکی از الگوریتم‌های تکاملی، با الهام‌گرفتن از اصول انتخاب طبیعی داروین تلاش می‌کنند تا از تکامل ژنتیکی به‌عنوان یک الگوی حل مسئله استفاده کنند. بدین‌صورت که مسئله به‌صورت ژن و کروموزم الگوبرداری کرده و آن‌ها را توسط یک تابع برازش[1] که با توجه به مسئله تعیین می‌گردد، مورد بررسی قرار می‌دهند. بهترین‌های هر نسل به نسل بعد منتقل شده و بدین‌صورت روند تکامل را پیاده‌سازی

1. fitness function

می‌نمایند. البته، با درصدی که در مسئله مشخص می‌شود، ژن‌ها می‌توانند کاملاً تغییر کنند و بدین‌صورت اجازهٔ ورود فضای جدید به مسئله و یا به‌عبارتی تنوع را می‌دهند. بر اساس انتخاب داروین، از میان جمعیت، تنها موجوداتی باقی می‌مانند که بیشترین تطابق را با شرایط محیطی خود داشته باشند.

به‌صورت کلی، اساس کار الگوریتم‌های ژنتیکی به‌صورت زیر است: ابتدا با توجه به نوع مسئله، کروموزم‌ها تعریف می‌شوند. سپس، جمعیتی از کروموزم‌ها وارد مسئله می‌شوند و بر اساس تابع برازش برای بقا تلاش می‌کنند. بنابراین، قدم بعدی انتخاب تابع برازش مناسب با توجه به مسئله است. کروموزم‌ها با یکدیگر ترکیب می‌شوند و از میان کروموزم‌های والد و فرزندان، بهترین‌ها که با توجه به تابع برازش مشخص می‌شوند، به نسل بعد می‌روند. بر این اساس، کروموزم‌هایی باقی خواهند ماند که تطابق بیشتری با شرایط محیطی خود داشته باشند. البته، در تولید هر نسل، عامل‌های تصادفی نیز که موجب تولید ژن‌های جدید هستند، دخیل خواهند بود.

الگوریتم پیشنهادی برای یادگیری وزن‌ها بدین‌صورت عمل می‌کند که ابتدا جمعیتی از کروموزم‌ها تولید می‌شوند. هر کروموزم یک مجموعهٔ ده‌تایی است که ژن‌های آن به‌ترتیب نشان‌دهندهٔ نام، نام خانوادگی، وابستگی دانشگاهی، موضوع اصلی، موضوع فرعی، نویسندگان همکار، نام منحصربه‌فرد، عنوان نشریه، عنوان مقاله و مقدار آستانه هستند (جدول ۲).

جدول ۲. کروموزم روش پیشنهادی

مقدار آستانه	عنوان مقاله	عنوان نشریه	نام منحصر به‌فرد	نویسندگان همکار	موضوع فرعی	موضوع اصلی	وابستگی دانشگاهی	نام خانوادگی	نام
(۰-۱)	(۰-۱)	(۰-۱)	(۰-۱)	(۰-۱)	(۰-۱)	(۰-۱)	(۰-۱)	(۰-۱)	(۰-۱)

همان‌طور که در جدول ۲، مشاهده می‌شود، این کروموزم از ۱۰ ژن تشکیل شده که مقدار هر یک از ژن‌ها عددی میان ۰ و ۱ است. بر اساس الگوریتم ژنتیک، بهترین‌ها که با تابع برازش به‌دست می‌آیند، به نسل بعد منتقل می‌شوند. به‌عبارت دیگر، کیفیت هر کروموزم توسط تابع برازش به‌دست می‌آید.

در این مطالعه دو تابع برازش ارائه شده است. اولین تابع برازش بر اساس نسبت تعداد جفت نویسندگانی که یکسان‌بودن آن‌ها درست تشخیص داده شده (hit) به تعداد

جفت نویسندگانی که یکسان‌بودن آن‌ها درست تشخیص داده نشده است (mis) است. فرمول ۴ اولین تابع برازش ارائه‌شده در مقاله را نشان می‌دهد.

$$f_{HM}(R) = \frac{hit}{mis} \quad (۴)$$

در این فرمول hit و mis به‌صورت زیر تعریف می‌شوند:

$$hit = TP + TN \quad (۵)$$

$$mis = FN + FP$$

TP تعداد جفت نویسندگان یکسانی است که روش پیشنهادی آن‌ها را یکسان در نظر گرفته است. FP تعداد جفت نویسندگانی را نشان می‌دهد که روش پیشنهادی آن‌ها را یکسان در نظر گرفته، ولی در حقیقت متعلق به دو نویسندهٔ مختلف هستند. TN تعداد جفت نویسندگانی که به‌رغم شباهت ظاهری متعلق به دو نویسندهٔ متفاوت هستند و روش پیشنهادی این تفاوت را به‌درستی درک کرده است. FN تعداد جفت نویسندگانی را نشان می‌دهد که در حقیقت یک نویسنده را نشان می‌دهد، ولی روش پیشنهادی به اشتباه آن‌ها را متعلق به دو نویسنده در نظر گرفته است.

تابع برازش دوم از مقدار F1 الهام گرفته‌شده و به‌صورت زیر تعریف می‌شود.

$$F_{NT}(R) = \frac{2TP}{N + TP - TN} \quad (۶)$$

٤. تجزیه و تحلیل یافته‌ها

٤-١. داده‌ها

برای جمع‌آوری داده‌های این پژوهش، ابتدا بر اساس تقسیم موضوعی نشریات وزارت علوم و وزارت بهداشت، تعداد ۲۰ نشریه را با موضوعات تخصصی مختلف با روش نمونه‌گیری تصادفی انتخاب کردیم. سپس، اطلاعات مقالات مختلف از پایگاه‌های اطلاعات علمی مرتبط با این نشریات بررسی و آن‌ها که بیش از یک رکورد به ازای یک نام موجود بود، جمع‌آوری گردید. این اطلاعات شامل نام نویسند(گان) مقاله، وابستگی دانشگاهی نویسند(گان)، پست الکترونیکی آن‌ها و همچنین، اطلاعات مقاله است. سپس،

1. f-measure

تمامی این اطلاعات با یکدیگر ترکیب شد که منجر به ایجاد اسامی مختلف گردید. در مرحلهٔ بعد، با کمک متخصصان نمایه‌سازی در پایگاه استنادی علوم جهان اسلام، نام‌هایی که متعلق به یک نویسنده بودند، مشخص گردید. در نهایت، بعد از پردازش‌های نهایی و حذف بعضی از رکوردها، تعداد ۱۷۰۴ نام ایجاد گردید که ۸۰۴ نام همسان هستند. البته، لازم به ذکر است که از این تعداد بعضی ۲ نام ممکن است متعلق به یک نویسنده باشد و بعضی دیگر ۳، ۴ یا حتی بیشتر.

جدول ۳، نمونه‌ای از داده‌های مورد استفاده در این مقاله را نشان می‌دهد. در این جدول، ستون اول، نام و نام خانوادگی نویسنده، ستون دوم، وابستگی سازمانی، ستون سوم و چهارم، عنوان نشریه و عنوان مقالهٔ وی است. در این جدول، تمامی سطرها متعلق به یک نویسنده است که با مقادیر متفاوتی ذخیره شده است.

جدول ۳. نمونه‌ای از داده‌ها

عنوان مقاله	عنوان نشریه	وابستگی سازمانی	نام و نام خانوادگی
Who is Responsible?	International Journal of Cancer Management	Cancer Research Center, Shohada Hospital, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran	Mohammad Esmaeil Akbari
Ethics of Palliative Surgery in Esophageal Cancer	Iranian Journal of Cancer Prevention	Cancer Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran	ME Akbari
Burden of the central nervous system malignancies in Iran 2003	Iranian Journal of Cancer Prevention	Professor of Surgery- Cancer Research Center, Shahid Beheshti Medical University, Tehran, Iran	ME. Akbari
The Role of Patient in Patient Management	International Journal of Cancer Management	Cancer Research Center, Shohada Hospital, Shahid Beheshti University of Medical Sciences, Tehran, IR Iran	Mohammad Esmaeil Akbari
Five and Ten Years Survival in Breast Cancer Patients Mastectomies vs. Breast Conserving Surgeries Personal Experience	Iranian Journal of Cancer Prevention	Professor of Surgical oncology, Cancer Research Center, Shahid Beheshti University (MC), Tehran, Iran	ME Akbari

عنوان مقاله	عنوان نشریه	وابستگی سازمانی	نام و نام خانوادگی
Ten year breast cancer screening and follow up in 52200 women in Shahre-Kord, Iran (1997-2006)	Iranian Journal of Cancer Prevention	Professor of surgical Oncology, Cancer Research Center, Shahid Beheshti Medical University (MC), Iran	ME Akbari
OCT-4, an Embryonic Stem Cell Marker Expressed in Breast, Brain and Thyroid Carcinomas Compared to Testicular Carcinoma	Iranian Journal of Cancer Prevention	Cancer Research Centre, Shahid Beheshti University of Medical Sciences, Tehran, Iran	ME Akbari
Psychosocial Care for Breast Cancer: Physicians' Perspective	Iranian Journal of Cancer Prevention	Cancer Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran	ME Akbari
Burden of the central nervous system malignancies in Iran 2003	Iranian Journal of Cancer Prevention	Professor of Surgery- Cancer Research Center, Shahid Beheshti Medical University, Tehran, Iran	ME. Akbari
The need for palliative care services in Iran; an introductory commentary	Iranian Journal of Cancer Prevention	Professor of Cancer Surgery, Cancer Research Centre, Shahid Beheshti University of Medical Sciences, Iran	ME. Akbari

٤-٢. معیارهای ارزیابی

به‌منظور ارزیابی روش پیشنهادی از روش اعتبارسنجی ضربدری[1] استفاده شد. این روش، مستقل از داده‌های آموزشی بوده و مشخص می‌کند که نتایج یک تحلیل آماری روی مجموعه‌ای از داده‌ها تا چه اندازه قابل تعمیم است. اساس کار این روش، تقسیم داده‌ها به دو مجموعهٔ جدا از هم، تحلیل روی یک مجموعه و سپس، اعتبارسنجی روی مجموعهٔ دیگر است. به‌منظور رسیدن به نتایج دقیق‌تر، این تقسیم‌بندی چندین‌بار صورت گرفته و در نهایت، میانگین این نتایج به‌عنوان نتیجهٔ نهایی گزارش می‌شود.

روش اعتبارسنجی ضربدری k-fold بدین‌صورت عمل می‌کند که داده‌ها به k زیرمجموعه تقسیم می‌شود. از این k زیرمجموعه، هر بار یکی برای اعتبارسنجی و k-1

1. cross-validation

مجموعهٔ دیگر برای آموزش مورد استفاده قرار می‌گیرد. این روند k بار تکرار می‌شود و بدین‌صورت، همهٔ داده‌ها شانسی برای عضویت در داده‌های آموزشی و تست پیدا می‌کنند. ثابت شده است که مقدار k=10 بهترین مقداری است که می‌توان به نتایج دقیق و قابل اعتماد دست پیدا کرد (Breiman 2017). بنابراین، مجموعه‌ای از داده‌ها که برچسب آن‌ها مشخص است، یعنی مشخص است که آیا متعلق به یک نویسنده هستند یا خیر، به ۱۰ بخش تقسیم می‌گردد. روش پیشنهادی با هر یک از توابع برازش بر اساس ۹ مجموعه از ۱۰ زیرمجموعهٔ تولیدشده، آموزش داده شده و وزن‌ها به همراه مقدار آستانه تعیین می‌گردد و سپس، روی یک مجموعه باقی‌مانده به‌عنوان مجموعه تست، ارزیابی انجام می‌گیرد. برای ارزیابی روش پیشنهادی روی داده‌های آموزشی و تست از دقت[1]، بازیافت[2] و مقدار F[3] استفاده می‌شود.

دقت (فرمول ۷) مشخص می‌کند که از میان تعداد جفت نویسندگانی که روش پیشنهادی آن را یکسان در نظر گرفته است (نمونه‌های مثبت)، چند درصد واقعاً متعلق به یک نویسنده بوده است.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (۷)$$

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.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (۸)$$

و در نهایت، مقدار F به‌منظور ایجاد یک توازن میان این معیار به‌صورت زیر تعریف می‌شود:

$$\text{f} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (۹)$$

٤-۳. نتایج

نتایج اجرای الگوریتم پیشنهادی با توجه به تابع برازش اول (FHM) به ازای یک‌بار اجرای روش ضربدری 10-fold به‌عنوان نمونه در جدول ۴، نشان داده شده است. به‌عبارت

1. precision 2. recall 3. f-measure

دیگر، داده‌ها به ۱۰ قسمت تقسیم شده‌اند و هر بار ۹ قسمت برای آموزش و یکی برای تست استفاده شده است. بنابراین، ۱۰ مجموعه از دادهٔ اولیه ایجاد می‌گردد که با مجموعهٔ ۱ تا ۱۰ مشخص شده‌اند. مقادیر R، P و F در این جدول، به‌ترتیب، نشان‌دهندهٔ دقت، یادآوری و مقدار F در داده‌های آموزشی و تست در هر یک از مجموعهٔ داده که با set1 تا set10 مشخص شده، هستند. میانگین نتایج روی ۱۰ مجموعه به‌عنوان نتیجهٔ اجرای الگوریتم از یک‌بار اجرای الگوریتم با روش ضربدری 10-fold است (ردیف آخر این جدول). لازم به ذکر است که برای رسیدن به نتایج دقیق‌تر، این آزمایش ۳۰ بار انجام گرفته و میانگین همهٔ این نتایج به‌عنوان نتیجهٔ نهایی گزارش شده است.

جدول ٤. نمونه‌ای از اجرای الگوریتم با تابع برازش FHM روی داده آموزشی و آزمایشی

	داده‌های آموزشی										داده‌های آزمایشی			
	TP	FP	TN	FN	P	R	F	TP	FP	TN	FN	P	R	F
مجموعه ۱	۳۴۱	۴۹	۳۵۵	۲۰	۰/۸۷	۰/۹۴	۰/۹۱	۳۸	۷	۳۸	۳	۰/۸۴	۰/۹۳	۰/۸۸
مجموعه ۲	۳۴۳	۵۲	۳۵۰	۲۰	۰/۸۷	۰/۹۴	۰/۹۱	۳۶	۴	۴۳	۳	۰/۰۹	۰/۹۲	۰/۹۱
مجموعه ۳	۳۳۹	۴۸	۳۵۷	۲۱	۰/۸۸	۰/۹۴	۰/۹۱	۴۰	۸	۳۶	۲	۰/۸۳	۰/۹۵	۰/۸۹
مجموعه ۴	۳۴۶	۵۱	۳۴۷	۲۱	۰/۸۷	۰/۹۴	۰/۹۱	۳۳	۵	۴۶	۲	۰/۸۷	۰/۹۴	۰/۹۰
مجموعه ۵	۳۳۶	۵۲	۳۵۵	۲۲	۰/۸۷	۰/۹۴	۰/۹۰	۴۳	۴	۳۸	۱	۰/۹۱	۰/۹۸	۰/۹۵
مجموعه ۶	۳۳۸	۵۳	۳۵۴	۲۰	۰/۸۶	۰/۹۴	۰/۹۰	۴۱	۳	۳۹	۳	۰/۹۳	۰/۹۳	۰/۹۳
مجموعه ۷	۳۴۵	۵۲	۳۴۷	۲۱	۰/۸۷	۰/۹۴	۰/۹۰	۳۴	۴	۴۶	۲	۰/۸۹	۰/۹۴	۰/۹۲
مجموعه ۸	۳۴۶	۵۰	۳۴۶	۲۳	۰/۸۷	۰/۹۴	۰/۹۰	۳۳	۶	۴۷	۰	۰/۸۵	۱	۰/۹۲
مجموعه ۹	۳۳۸	۴۸	۳۵۶	۲۳	۰/۸۸	۰/۹۴	۰/۹۰	۴۱	۸	۳۷	۰	۰/۸۴	۱	۰/۹۱
مجموعه ۱۰	۳۳۷	۵۱	۳۵۴	۲۳	۰/۸۷	۰/۹۴	۰/۹۰	۴۲	۵	۳۹	۰	۰/۸۹	۱	۰/۹۴
میانگین	۳۴۰/۹	۵۰/۶	۳۵۲/۱	۲۱/۴	۰/۸۷	۰/۹۴	۰/۹۰	۳۸/۱	۵/۴	۴۰/۹	۱/۶	۰/۸۸	۰/۹۶	۰/۹۲

همان‌گونه که اشاره شد، به ازای هر نویسنده، ویژگی‌های داخلی و خارجی استخراج و به‌صورت یک بردار تعریف می‌شود. سپس، به ازای هر نویسنده یک کد با توجه به نام خانوادگی به‌دست آمده است. تمامی نویسندگانی که کد آن‌ها یکسان است، به معنای شباهت بسیار زیاد در نام خانوادگی آن‌ها و بالتبع یکسان‌بودن بالقوهٔ آن‌ها است. نویسندگانی که در یک گروه قرار گرفته‌اند، به‌صورت دوبه‌دو بررسی می‌شوند و در

صورت یکسان‌بودن برچسب ۱ و در غیر این‌صورت برچسب ۰ به آن‌ها تعلق می‌گیرد. لازم به ذکر است که صحت این برچسب‌گذاری دوبه‌دویی نویسندگان توسط کارشناسان خبرهٔ موضوعی نیز تأیید شده است.

الگوریتم پیشنهادی مبتنی بر تابع برازش اول (FHM) توانسته با دقت بالایی نویسندگان یکسان را در هر دو گروه آموزشی و تست پیدا کند. در صورتی که دو نویسنده متعلق به دو موجودیت متفاوت باشند، روش پیشنهادی توانسته با دقت بسیار خوبی آن‌ها را تشخیص دهد. مقادیر بالای TP و TN مؤید این امر است. مقدار یادآوری بزرگ الگوریتم پیشنهادی روی هر دو مجموعهٔ آموزشی و تست (در بعضی از مجموعه‌ها مقدار ۱۰۰ درصد) نشان‌دهندهٔ کارایی بالای تابع برازش است.

نتایج اجرای الگوریتم پیشنهادی با توجه به تابع برازش دوم (FNT) به ازای یک‌بار اجرای روش ضربدری 10-fold به‌عنوان نمونه در جدول ۵، نشان داده شده است. همان‌طور که مشاهده می‌شود، روش پیشنهادی با تابع برازش دوم نیز توانسته به‌خوبی عملیات رفع ابهام نویسندگان را انجام دهد؛ به‌عبارت دیگر، جفت نویسندگانی را که متعلق به یک موجودیت بودند، مقدار ۱ و آن‌ها را که متعلق به موجودیت‌های مختلف هستند، مقدار ۰ نسبت دهد. مقادیر نسبتاً بالای دقت، بازیافت و مقدار F مؤید این امر است.

همان‌طور که قبلاً نیز عنوان گردید، جداول ۴ و ۵ نتیجهٔ اجرای یک‌بار الگوریتم با دو تابع برازش با روش ضربدری 10-fold است و در نهایت، میانگین روی این ۱۰ مجموعه به‌عنوان نتیجه گزارش می‌شود. به‌منظور رسیدن به نتیجهٔ دقیق‌تر، این آزمایش ۳۰ بار انجام گرفته و میانگین این نتایج به‌عنوان نتیجهٔ نهایی در نظر گرفته می‌شود. به‌عبارت دیگر، آزمایشی مشابه جداول ۴ و ۵، ۳۰ بار انجام گرفته و هر بار میانگین روی ۱۰ مجموعه به‌دست آمده است. سپس، میانگین این نتایج در ۳۰ بار گزارش می‌شود. جدول ۶، وزن نهایی برای هر یک از ویژگی‌ها را با توجه به هر دو تابع برازش نشان می‌دهد (پاسخ به پرسش ۱).

جدول ۵. نمونه‌ای از اجرای الگوریتم با تابع برازش FNT روی دادهٔ آموزشی و تست

داده‌های آزمایشی					داده‌های آموزشی								
F	R	P	FN	TN	FP	TP	F	R	P	FN	TN	FP	TP	
۰/۹۰	۰/۹۷	۰/۸۵	۱	۴۶	۶	۳۳	۰/۹۱	۰/۹۵	۰/۸۷	۲۰	۳۴۵	۵۲	۳۴۸	مجموعه ۱
۰/۹۰	۰/۹۸	۰/۸۳	۱	۳۱	۹	۴۵	۰/۹۱	۰/۹۴	۰/۸۷	۲۰	۳۶۰	۴۹	۳۳۶	مجموعه ۲
۰/۹۰	۰/۹۱	۰/۸۹	۴	۳۵	۵	۴۲	۰/۹۱	۰/۹۵	۰/۸۶	۱۷	۳۵۶	۵۳	۳۳۹	مجموعه ۳
۰/۹۶	۰/۹۶	۰/۹۶	۲	۳۹	۲	۴۳	۰/۹۰	۰/۹۵	۰/۸۶	۱۹	۳۵۲	۵۶	۳۳۸	مجموعه ۴
۰/۹۰	۰/۹۵	۰/۸۶	۲	۴۰	۶	۳۸	۰/۹۱	۰/۹۵	۰/۸۷	۱۹	۳۵۱	۵۲	۳۴۳	مجموعه ۵
۰/۹۳	۰/۹۳	۰/۹۳	۳	۴۱	۳	۳۹	۰/۹۰	۰/۹۵	۰/۸۶	۱۸	۳۵۰	۵۵	۳۴۲	مجموعه ۶
۰/۹۳	۱	۰/۸۷	۰	۴۰	۶	۴۰	۰/۹۰	۰/۹۴	۰/۸۷	۲۱	۳۵۱	۵۲	۳۴۱	مجموعه ۷
۹۱۰	۰/۹۵	۰/۸۸	۲	۴۴	۵	۳۵	۰/۹۱	۰/۹۵	۰/۸۷	۱۹	۳۴۷	۵۳	۳۴۶	مجموعه ۸
۰/۸۸	۰/۹۴	۰/۸۳	۲	۴۴	۷	۳۳	۰/۹۱	۰/۹۵	۰/۸۷	۱۹	۳۴۷	۵۱	۳۴۸	مجموعه ۹
۰/۹۳	۰/۹۳	۰/۹۳	۳	۴۰	۳	۴۰	۰/۹۰	۰/۹۵	۰/۸۶	۱۸	۳۵۱	۵۵	۳۴۱	مجموعه ۱۰
۰/۹۱	۰/۹۵	۰/۸۸	۲	۴۰	۵/۲	۳۸/۸	۰/۹۱	۰/۹۵	۰/۸۷	۱۹	۳۵۱	۵۲/۸	۳۴۲	میانگین

جدول ٦. وزن نهایی هر یک از ویژگی‌ها در روش پیشنهادی با دو تابع برازش

مقدار آستانه	عنوان مقاله	عنوان نشریه	نام منحصربه‌فرد	نویسندگان همکار	موضوع فرعی	موضوع اصلی	وابستگی دانشگاهی	نام خانوادگی	نام	
۰/۳	۰/۹۷	۰/۱	۰/۵۸	۰/۹۸	۰/۱	۰/۱	۰/۵۶	۰/۱	۰/۹۸	وزن‌های نهایی با تابع برازش FHM
۰/۳۴	۰/۸۵	۰/۱۱	۰/۴۶	۰/۸۷	۰/۱	۰/۱	۰/۶۲	۰/۲۱	۰/۹۸	وزن‌های نهایی با تابع برازش FNT

لازم به ذکر است که یکی از ویژگی‌های داخلی که در این جدول مشخص نشده، پست الکترونیکی نویسنده است. از آنجا که این ویژگی به ازای تمامی داده‌های این پژوهش موجود نبود، در کروموزوم الگوریتم ژنتیک پیشنهادی لحاظ نگردید. به ازای هر جفت نام نویسنده، در صورتی که پست الکترونیک هر دو موجود باشد و با یکدیگر برابر باشد، به معنای این است که این دو نام متعلق به یک نویسنده است. بنابراین، از این ویژگی به‌عنوان یک ویژگی نهایی لحاظ شده است؛ یعنی همان‌طور که در الگوریتم ۱، نیز مشخص شده، اگر پست الکترونیک دو نویسنده موجود و برابر باشد، آن دو نام را متعلق به یک نویسنده اعلام می‌کنیم. اگر پست الکترونیک موجود و نابرابر بود، نمی‌توان نظر

نهایی داد؛ چرا که ممکن است واقعاً متعلق به دو نویسندهٔ متفاوت باشد و یا یک نویسنده با دو آدرس پست الکترونیک.

همان‌طور که جدول ۶، نشان می‌دهد، ویژگی‌های نام از نام خانوادگی وزن بالاتری نسبت به دیگر ویژگی‌ها دارند. دلیل آن این است که این روش در مرحلهٔ جست‌وجو، تمامی اسامی را که از نظر نام خانوادگی به یکدیگر شبیه هستند، در یک گروه قرار می‌دهد و سپس، در هر گروه، عملیات تطابق را انجام می‌دهد. بنابراین، طبیعی است که اسامی که در هر گروه قرار می‌گیرند، از نظر نام خانوادگی به یکدیگر شبیه هستند و در نتیجه، این ویژگی وزن کمتری به خود می‌گیرد.

عنوان مقاله و نویسندگان همکار نیز وزن‌های بالایی توسط هر دو روش گرفته‌اند که نشان می‌دهد در داده‌های این پژوهش، این دو ویژگی خاصیت تمیزکنندهٔ بیشتری نسبت به دیگر ویژگی‌ها دارند. لازم به ذکر است که این وزن‌ها با توجه به داده‌های این پژوهش به‌دست آمده است که طبیعتاً مقادیر آن‌ها در داده‌های دیگر ممکن است متفاوت باشد و این امر ناشی از نحوهٔ توزیع داده‌ها در ویژگی‌های مختلف و قدرت تمیزکنندهٔ ویژگی‌ها در داده‌هاست.

به‌منظور بررسی پرسش ۲، روش پیشنهادی با دو تابع برازش با یکدیگر و همچنین، با ویژگی‌های روش «مزروعی سبدانی، ابراهیم‌پور کوملـه و نیک‌فرجام» (۱۳۹۲) مقایسه شد. نتایج در جداول ۷ و ۸ نشان داده شده است. همان‌طور که این جدول نشان می‌دهد، روش پیشنهادی با هر دو تابع برازش توانسته تعداد بسیار زیادی از زوج نویسندگانی را که در واقع، متعلق به یک موجودیت بوده‌اند، به‌درستی بیابد. در مجموعه‌های آموزشی و تست، مقادیر بالای TP مؤید این امر است. علاوه بر این، مقادیر TN نیز برای هر دو مجموعهٔ آموزشی و تست به ازای هر دو تابع برازش پیشنهادی بالاست. به‌عبارت دیگر، روش پیشنهادی تعداد بسیار زیادی از جفت نویسندگانی را که در واقع، متعلق به یک موجودیت بودند و یا به‌رغم شباهت ظاهری در نام، متعلق به دو نویسندهٔ متفاوت هستند، یافته است. مقادیر FN نیز به ازای هر دو تابع برازش و روی هر دو مجموعهٔ آموزشی و تست نیز پایین است که نشان‌دهندهٔ این است که تعداد کمی از جفت نویسندگانی که متعلق به یک موجودیت بودند، روش پیشنهادی آن‌ها را مجزا تشخیص داده است. در نهایت، مقادیر بالای دقت، بازیافت و مقدار F در هر دو تابع برازش روش پیشنهادی روی مجموعهٔ دادهٔ آموزشی و تست نشان‌دهندهٔ کارایی بالای روش پیشنهادی است. علاوه بر

این، نتایج نشان می‌دهد که روش پیشنهادی با دو تابع برازش تفاوت بسیار زیادی با یکدیگر در میانگین تعداد اجراهای بالا ندارد.

جدول ۷. مقایسهٔ روش پیشنهادی با روش‌های پیشین از منظر معیارهای TN، FP، TP و FN

آزمایشی				آموزشی				
FN	TN	FP	TP	FN	TN	FP	TP	
۱/۹۸	۳۹/۷	۶/۰۲	۸۳/۳	۱۹/۰۲	۳۵۱/۳	۵۱/۹۷	۳۴۲/۷۰	روش پیشنهادی با تابع برازش FNT
۲/۱۷	۳۹/۹۲	۵/۸	۳۸/۱۱	۲۰/۸۳	۳۵۳/۰۸	۵۰/۲	۳۴۰/۸۹	روش پیشنهادی با تابع برازش FHM
۵/۱۷	۳۹/۳۵	۶/۳۸	۳۵/۱۱	۴۷/۸۳	۳۴۸/۶۵	۵۴/۶۲	۳۱۳/۸۹	روش مزروعی

جدول ۸. مقایسهٔ روش پیشنهادی با روش‌های پیشین از منظر معیارهای دقت، بازیافت و مقدار F

آزمایشی			آموزشی			
مقدار F	بازیافت	دقت	مقدار F	بازیافت	دقت	
۰/۹۰	۰/۹۵	۰/۸۶	۰/۹۱	۰/۹۵	۰/۸۷	روش پیشنهادی با تابع برازش FNT
۰/۹۰	۰/۹۵	۰/۸۶	۰/۹۱	۰/۹۴	۰/۸۷	روش پیشنهادی با تابع برازش FHM
۰/۸۶	۰/۸۷	۰/۸۵	۰/۸۶	۰/۸۷	۰/۸۵	روش مزروعی

۵. نتیجه‌گیری

در این مطالعه، روشی برای رفع ابهام نام نویسندگان ارائه گردید. این روش از دو مرحلهٔ جست‌وجو و تطابق تشکیل شده است. در مرحلهٔ جست‌وجو با دادن یک کد به هر نویسنده، نویسندگانی که بالقوه احتمال یکسان‌بودن آن‌ها هست، در یک گروه قرار می‌گیرند. مرحلهٔ تطابق، تمامی نویسندگانی را که در یک گروه قرار دارند، مورد بررسی قرار می‌دهد و با ارائهٔ یک معیار شباهت، میزان شباهت دو نویسنده را تعیین کرده و در صورتی که میزان شباهت آن‌ها از یک مقدار آستانه بیشتر باشد، آن‌ها را یکسان و در غیر این‌صورت متعلق به دو موجودیت مختلف گزارش می‌دهد. روش پیشنهادی از ویژگی‌های داخلی و خارجی نویسندگان استفاده کرده و به هر یک از ویژگی‌ها یک وزن می‌دهد که در واقع، نشان‌دهندهٔ میزان اهمیت هر کدام از ویژگی‌ها در مسئله است.

به‌منظور تعیین این وزن‌ها و همچنین، مقدار آستانه و یا به‌عبارتی پاسخ به پرسش ۱ مقاله، یک الگوریتم ژنتیک با دو تابع برازش مختلف ارائه شده است تا بتواند وزن‌ها را

از داده‌ها یاد بگیرد. الگوریتم پیشنهادی با دو تابع برازش روی داده‌های آموزشی اجرا شده و با توجه به این داده‌ها که در آن‌ها نام‌های متعلق به یک نویسنده توسط متخصصان نمایه‌سازی مشخص گردیده، میزان اهمیت هر یک از ویژگی‌ها و یا به‌عبارتی، وزن‌ها را به‌دست می‌آورد. لازم به ذکر است که وزن‌های به‌دست‌آمده با توجه به داده‌های آموزشی این پژوهش به‌دست آمده است و طبیعتاً در هر نوع داده‌ای با توجه به توزیع داده در ویژگی‌های مختلف ممکن است این وزن‌ها تغییر کند. علاوه بر این، ویژگی پست الکترونیک در صورت موجود بودن در پایگاه دادهٔ مورد بررسی از اهمیت بسیار زیادی برخوردار است. اگر پست الکترونیک دو نویسنده موجود و برابر باشد، آن دو نام را متعلق به یک نویسنده اعلام می‌کنیم. اگر پست الکترونیک موجود و نابرابر بود، نمی‌توان نظر نهایی داد؛ چرا که ممکن است واقعاً به دو نویسندهٔ مختلف و یا یک نویسنده با دو آدرس پست الکترونیک متعلق باشد. به‌عبارت دیگر، در صورت موجود بودن و برابر بودن پست الکترونیک به ازای دو داده، آن‌ها متعلق به یک نویسنده هستند و باید یکسان‌سازی روی آن‌ها انجام گیرد. در غیر این‌صورت باید یکسان‌بودن یا نبودن آن دو داده را با توجه به الگوریتم تطابق به‌دست آورد.

به‌منظور پاسخ به پرسش ۲ مقاله، آزمایشات را ۳۰ بار تکرار کرده و میانگین آن‌ها را به‌عنوان نتیجهٔ نهایی در نظر گرفتیم و همچنین، با روش پیشین در این راستا مقایسه نمودیم. نتایج شبیه‌سازی روی داده‌های آزمایشی نشان‌دهندهٔ کارایی بالای هر دو تابع برازش و برتری آن نسبت به روش دیگر است. همچنین، نتایج نشان می‌دهد که روش پیشنهادی با دو تابع برازش تفاوت بسیار زیادی با یکدیگر در میانگین تعداد اجراهای بالا ندارند.

این الگوریتم را می‌توان در سامانهٔ یکسان‌ساز اسامی نویسندگان به‌کار برد. اگر یک مورد مثبت -به معنای یکسان‌بودن دو نویسنده- توسط الگوریتم تشخیص داده شد، می‌توان یک پیام به مدیر سامانه و یا به پست الکترونیکی نویسنده ارسال کرد و در صورت تأیید نویسنده و یا مدیر سامانه، عملیات یکسان‌سازی نام نویسنده انجام شود. بنابراین، مقدار FN اهمیت بیشتری نسبت به FP دارد؛ چرا که اگر مورد اشتباهی نیز توسط الگوریتم تشخیص داده شد، این تغییرات به‌دلیل عدم تأیید نویسنده و یا ادمین سامانه در پایگاه اعمال نمی‌شود.

فهرست منابع

رزمی شندی، مسعود، یعقوب نوروزی، و مهدی علیپور حافظی. ۱۳۹۹. ارائهٔ الگوی مفهومی به‌کارگیری

اینترنت اشیا در خدمات نوین کتابخانه‌های دیجیتال ایران. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۵ (۳): ۶۹۳-۷۲۸.

قاسمی الوری، مینا، و مظفر چشمه‌سهرابی. ۱۳۹۹. تحلیل کمی و انتقادی پژوهش‌های حوزهٔ کتابخانه‌های دیجیتالی در ایران، *پژوهشنامه پردازش و مدیریت اطلاعات* ۴ (۳۵): ۹۲۱-۹۵۲.

مرتضوی، سید محمد، محمدحسین ندیمی شهرکی، و مصطفی موسی‌خانی. ۱۳۹۶. بهبود صحت ابهام‌زدایی نام نویسنده با استفاده از خوشه‌بندی تجمعی. *پردازش علائم و داده‌ها* ۳۴ (۴): ۱۱۷-۱۲۷.

مزروعی سبدانی، نصیرالدین، حسین ابراهیم‌پور کوملـه، و علی‌محمد نیک‌فرجام. ۱۳۹۲. ارائهٔ روش با نظارت به‌منظور دسته‌بندی مقالات با وجود ابهام در داده‌ها. دوازدهمین کنفرانس سیستم‌های هوشمند ایران، مجتمع آموزش عالی بم.

References

Bekkerman, R., & A. McCallum. 2005. Disambiguating web appearances of people in a social network. In Proceedings of the 14th international conference on World Wide Web, pp. 463-470. Chiba, Japan.

Breiman, L. 2017. *Classification and regression trees*. Routledge.

Fan, Xiaoming, Jianyong Wang, Xu Pu, Lizhu Zhou, and Bing Lv. 2011. On graph-based name disambiguation. *Journal of Data and Information Quality (JDIQ)* 2 (2): 1-23.

Giles, C. Lee, Hongyuan Zha, and Hui Han. 2005. Name disambiguation in author citations using a k-way spectral clustering method. In Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries (JCDL'05), pp. 334-343. IEEE. Denver, CO USA.

Han, Donghong, Siqi Liu, Yachao Hu, Bin Wang, and Yongjiao Sun. 2015. ELM-based name disambiguation in bibliography. *World Wide Web* 18 (2): 253-263.

Han, Hui, Lee Giles, Hongyuan Zha, Cheng Li, and Kostas Tsioutsiouliklis. 2004. Two supervised learning approaches for name disambiguation in author citations. In Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries. pp. 296-305. IEEE. Tuscon AZ USA.

Hodge, Victoria J., and Jim Austin. 2003. A comparison of standard spell checking algorithms and a novel binary neural approach. *IEEE transactions on knowledge and data engineering* 15 (5): 1073-1081.

Holmes, David, and M. Catherine McCabe. 2002. Improving precision and recall for soundex retrieval. In Proceedings. International Conference on Information Technology: Coding and Computing, pp. 22-26. IEEE. Las Vegas, Nevada.

Huang, Jian, Seyda Ertekin, and C. Lee Giles. 2006. Efficient name disambiguation for large-scale databases. In European conference on principles of data mining and knowledge discovery, pp. 536-544. Berlin, Heidelberg: Springer.

Hussain, Ijaz, and Sohail Asghar. 2017. A survey of author name disambiguation techniques: 2010–2016. *The Knowledge Engineering Review* 32: e22 DOI:https://doi.org/10.1017/S0269888917000182.

Huynh, Tin, Kiem Hoang, Tien Do, and Duc Huynh. 2013. Vietnamese author name disambiguation for integrating publications from heterogeneous sources." In Asian Conference on Intelligent Information and Database Systems, pp. 226-235. Berlin, Heidelberg: Springer.

Imran, Muhammad, Syed Gillani, and Maurizio Marchese. 2013. A real-time heuristic-based unsupervised method for name disambiguation in digital libraries. *D-Lib Magazine* 19 (9):1.

Lait, Andrew J., and Brian Randell. 1996. An assessment of name matching algorithms. Technical Report Series.University of Newcastle upon Tyne Computing Science.

Navarro, Gonzalo. 2001. A guided tour to approximate string matching. *ACM computing surveys (CSUR)* 33 (1): 31-88.

Niwattanakul, S., J. Singthongchai, E. Naenudorn, and S. Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In Proceedings of the international multi-conference of engineers and computer scientists 1 (6): 380-384.

Philips, Lawrence. 2000. The double metaphone search algorithm. *C/C++ Users Journal* 18 (6): 38-43.

Sayers, Adrian. 2014. *NYSIIS: Stata module to calculate nysiis codes from string variables. Statistical Software Components* S457936, Boston: College Department of Economics. Revised 21 Jul 2018.

Seol, Jae-Wook, Seok-Hyoung Lee, and Kwang-Young Kim. 2016. Author disambiguation using co-author network and supervised learning approach in scholarly data. *International Journal of Software Engineering and Its Applications* 10 (4): 73-82.

Shin, Dongwook, Taehwan Kim, Joongmin Choi, and Jungsun Kim. 2014. 2014. Author name disambiguation using a graph model with node splitting and merging based on bibliographic information. *Scientometrics* 100 (1): 15-50.

Song, Yang, Jian Huang, Isaac G. Councill, Jia Li, and C. Lee Giles. 2007. Efficient topic-based unsupervised name disambiguation. In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, pp. 342-351. Vancouver BC Canada.

Tang, Jie, Alvis CM Fong, Bo Wang, and Jing Zhang. 2011. A unified probabilistic framework for name disambiguation in digital library. *IEEE Transactions on Knowledge and Data Engineering* 24 (6): 975-987.

Tejada, Sheila, Craig A. Knoblock, and Steven Minton. 2001. Learning object identification rules for information integration. *Information Systems* 26 (8): 607-633.

Torvik, Vetle I., Marc Weeber, Don R. Swanson, and Neil R. Smalheiser. 2005. A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for information science and technology* 56 (2): 140-158.

Tran, Hung Nghiep, Tin Huynh, and Tien Do. 2014. Author name disambiguation by using deep neural network. In Asian Conference on Intelligent Information and Database Systems, pp. 123-132. Cham: Springer.

Wang, Xuezhi, Jie Tang, Hong Cheng, and S. Yu Philip. 2011. Adana: Active name disambiguation. In 2011 IEEE 11th international conference on data mining, pp. 794-803. IEEE. Vancouver, British Columbia, Canada.

Wang, Jian, Kaspars Berzins, Diana Hicks, Julia Melkers, Fang Xiao, and Diogo Pinheiro. 2012. A boosted-trees method for name disambiguation. *Scientometrics* 93 (2): 391-411.

Zobel, Justin, and Philip Dart. 1996. Phonetic string matching: Lessons from information retrieval. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 166-172. Zurich Switzerland.

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.

داده‌کاوی، یادگیری ماشین، تحلیل شبکه‌های اجتماعی و پردازش زبان‌های طبیعی از جمله علایق پژوهشی وی است.