بهبود صحت ابهام‌زدایی نام نویسنده با استفاده از خوشه‌بندی تجمّعی

سید محمد مرتضوی، محمد حسین ندیمی شهرکی* و مصطفی موسی‌خانی
دانشکده مهندسی کامپیوتر، واحد نجف آباد، دانشگاه آزاد اسلامی، نجف آباد، ایران
مرکز تحقیقات مِه داده، واحد نجف آباد، دانشگاه آزاد اسلامی، نجف آباد، ایران

چکیده

امروزه کتابخانه‌های دیجیتال از مهم‌ترین و سریع‌ترین منابع پژوهشی در جهان محسوب می‌شوند. از نقطه‌نظر مدیریت تجمیع دانش، توانایی جستجوی صحیح، دقیق و سریع مطالب علمی مد نظر کاربر، اهمیت زیادی دارد. پیچیدگی و وجود تشابه در بانک‌های اطلاعاتی موجب می‌شود این منابع در هنگام بهره‌برداری با چالش‌ها و ابهامات زیادی مواجه شوند و همین چالش‌ها دست‌مایه پژوهش‌های گسترده‌ای را در این حوزه شکل داده است. یکی از مهم‌ترین این چالش‌ها، وجود ابهام در نام نویسنده است. دراین‌خصوص روش‌های بسیاری با بهره‌گیری از روش‌های خوشه‌بندی نسبت به حل نام‌های مبهم مبادرت ورزیده‌اند. این روش‌ها تا حدودی توانسته‌اند مشکل را برطرف کنند؛ اما همچنان مسئله تکه‌تکه‌بودن خوشه‌ها و خطا در نتایج تولیدی، از معایب روش‌های موجود است. از سویی تجربه نشان داده که یک روش به‌تنهایی نتایجی با صحت بالا نمی‌تواند تولید کند. بدین‌منظور در این مقاله مدلی جهت حل مشکل ذکرشده ارائه شده 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 از "الگوریتم خوشه‌بندی تجمعی"، خوشه‌های تولیدشده به‌گونه‌ای ترکیب می‌شوند تا خوشه‌هایی غنی با درصد کمتری از تکه‌تکه‌بودن و صحت بالاتر تولید شوند. در ارزیابی الگوریتم پیشنهادی از "مجموعه دادگان DBLP، تحت معیار K" استفاده شده است. نتایج، بهبود قابل توجهی را در ترکیب خوشه‌های مذکور نشان می‌دهند.

واژگان کلیدی: کتابخانه‌های دیجیتال ، ابهام‌زدایی نام نویسنده، نام مبهم، خوشه‌بندی تجمعی

Improving the accuracy of the author name disambiguation by using clustering ensemble

Sayed Mohammad Mortazavi, Mohammad-Hossein Nadimi-Shahraki* & Mostafa Mosakhani

Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran

Big Data Research Center, Najafabad Branch, Islamic Azad University, Najafabad, Iran

* Corresponding author

* نویسندهٔ عهده‌دار مکاتبات

[DOI: 10.29252/jsdp.14.4.117]

Abstract

Today, digital libraries are important academic resources including millions of citations and bibliographic essential information such as titles, author's names and location of publications. From the view of knowledge accumulation management, the ability to search fast, accurate, desired contents, has a great importance. The complexity and similarity in these resources cause many challenges and ambiguities. One of the most of these challenges is the author name disambiguation which makes an extensive scope of research. Although many effective methods have been developed by using clustering techniques in disambiguation of the author's name, the accuracy of these methods is not acceptable and still there are some problems such as fragmentation and error in the produced results of these methods, since there is no uniform standard of citations, various combinations, and numerous, written, verbal patterns. In fact, experiences have shown that the use of a single method to disambiguate names does not provide results with a high accuracy despite concerns expressed above. In this paper, a new method is proposed to disambiguate author names in different formats and combinations with more accuracy. The proposed solution carries out the disambiguation in two steps; In the first step, agglomerative hierarchical clustering algorithm produces clusters using similar functions and different thresholds. In the second step, clusters produced by clustering ensemble technique in the previous stage are combined to provide more accurate clusters with less fragmentation. The proposed method is experimentally evaluated by conducted DBLP datasets with K criterion. The evaluation results show that the proposed method enhances the accuracy of disambiguation of author names in different formats.

Keywords: **Digital libraries, Author Name Disambiguation, Ambiguous name, Clustering Ensemble**

۱- مقدمه

کتابخانه‌ها محل ذخیرهٔ دانش هستند؛ به تعبیر دیگر آنها را مرکز فرهنگ و خرد می‌توان نامید.کتابخانه‌ها تکامل باورنکردنی را در طول تاریخ پشت‌سر گذاشته‌اند. این مؤسسات که در ابتدا تنها برای نگهداری و ذخیره مدارک ابداع شده بودند، تبدیل به مراکزی جهت تبادل اطلاعات شدند. کتابخانه‌های دیجیتال[1] بانک‌های اطلاعاتی پیچیده‌ای هستند که شامل مجموعه‌ای غنی از اشیای دیجیتال و فراداده‌ها[2] هستند. کتابخانه‌های دیجیتال شامل سرویس‌هایی جهت دریافت رکوردهای مرتبط با نویسنده‌ای خاص، جستجوهای مختلف، مرور، شخصی‌سازی و ساخت جوامع با زمینه‌های تخصصی خاص هستند. کتابخانه‌ها مانند DBLP[3]، CiteSeerx[4] و BDBCOMP[5] یکی از منابع مهم اطلاعاتی برای جوامع دانشگاهی شده است؛ زیرا آنها اجازه جستجو و کشف انتشارات مرتبط در یک روش متمرکز را فراهم می‌کنند [1]. این سامانه‌ها محتویاتشان را از منابع متعدد و مجزا به‌دست می‌آورند، بر این اساس استاندارد خاصی در ترتیب و کامل‌بودن ویژگی‌ها وجود ندارد که درنتیجه باعث وجود ابهامات[6] زیادی‌دراین منابع می‌شود. در این میان نام مبهم[7] از اهمیت بسیاری برخوردار بوده و مهم‌ترین مشکل در بین سایر ابهامات به‌وجود آمده در این منابع است. نام یک ویژگی کلیدی برای شناسایی افراد است؛ با توجه به اینکه به نام‌های افراد در دنیا شناسه یکتا تخصیص داده نشده است، اغلب در تشخیص افراد (نویسنده‌ها) با استفاده از نام آنها دچار اشتباه می‌شویم؛ زیرا ممکن است یک نویسنده چند نام داشته باشد یا اینکه چندین نویسنده یک نام یکسان را به اشتراک گذارند. مشکل نام‌های مبهم در منابع دیجیتال بر روی کارایی، در بازیابی اسناد تأثیر منفی می‌گذارد و همچنین کاربران را در شناسایی مقالات نویسندهٔ مورد نظر دچار اشتباه می‌کند. برای مثال ممکن است، اطلاعات بازیابی‌شده با نامی یکسان متعلق به نویسنده‌های مختلف در رکوردهای متفاوت باشند؛ یا نام نویسنده ممکن است با املای غلط به‌کار برده شده یا به‌صورت مخفف بیان شده باشد که کاربر (فرد جستجوگر) را دچار ابهام می‌کند. استانداردسازی اطلاعات به‌دست‌آمده از منابع گوناگون به‌دلیل وجود قالب‌ها و اشکال مختلف، کاری سخت و شاید غیرممکن باشد. برای حل چنین مسئله‌ای لازم است نام‌های موجود در بانک‌های اطلاعاتی کتابخانه‌های

[1] Digital libraries
[2] Metadata
[3] http://dblp.uni-trier.de
[4] http://citeseerx.ist.psu.edu/index
[5] http://www.lbd.dcc.ufmg.br/bdbcomp
[6] Ambiguities
[7] Ambiguous name

[DOI: 10.29252/jsdp.14.4.117]

دیجیتال و نام‌های واردشده به بانک، ابهام‌زدایی[1] شوند؛ یعنی هر رکورد را که شامل اطلاعات مقاله‌شناسی[2] از یک نویسنده با نام مبهم است در دسته‌ای قرار داد و نتایج جستجو، به‌صورت دسته‌بندی‌شده ارائه شود تا کاربر دچار ابهام در شناسایی نویسندهٔ مورد نظر نشود.

عنصر اصلی کتابخانه‌های دیجیتال، رکوردهایی حاوی سربرگ مقالات هستند. رکوردهای مقاله‌شناسی دارای ویژگی‌های مقاله‌شناسی و کتاب‌شناسی[3] بوده که شامل نام نویسنده‌ها، عنوان مقاله، سال انتشار، محل انتشار، تاریخ چاپ، وابستگی‌های نویسنده‌ها[4]، رایانامه و مواردی از این دست می‌باشند[2]. روش‌هایی ارائه شده است که با بهره‌گیری از این ویژگی‌ها و با استفاده از روش‌های گوناگون، سعی در حل نام‌های مبهم در کتابخانه‌های دیجیتال داشته‌اند؛ اما این مشکل تاکنون به‌طور قطع حل نشده است و هیچ‌کدام از روش‌های ارائه‌شده نتوانسته‌اند، تمامی نام‌های داخل یک فهرست را، صحیح ابهام‌زدایی کنند. به‌طور میانگین این روش‌ها، صحتی بین پنجاه الی نود درصد در نتایج نهایی به‌دست آورده‌اند.

مشکل به‌وجودآمده توسط نام‌های مبهم به دو گروه دسته‌بندی می‌شود [3]. نخستین گروه، ابهام‌زدایی مرجعی[5] است [4] که هدف آن، گروه‌بندی تغییرات مختلف نام یک موجودیت به یک خوشه است. شناسایی یکسان‌بودن چند نام نویسندهٔ به‌نسبه مشابه در مقالات مختلف برای خواننده، کار ساده‌ای نیست. با استفاده از روش‌های ابهام‌زدایی مرجعی، نام‌های مختلف یک شخص (نویسنده) دسته‌بندی می‌شود. نوع دوم که متفاوت از گروه نخست است، ابهام‌زدایی موجودیت نام[6] می‌باشد. ابهام‌زدایی موجودیت نام به‌عنوان شناسایی اشاره‌های گوناگون از موجودیت در اسناد متنی تعریف می‌شود [5]؛ اما با وجود ظهور چهار نمونه نام مبهم در منابع دیجیتال، نام‌های مبهم به‌طورکلی به دو شکل مختلف مشاهده می‌شوند:

نوع نخست: یک موجودیت نام به شکل‌های مختلف می‌تواند نوشته شود. به‌عبارت دیگر نام یک شخص ممکن است به چندین شکل به‌نسبه مشابه در مقالات و کتب مختلف نوشته شود (به‌علت سلیقهٔ نویسنده در نوشتن نام خود، خطا در هجی، تفاوت املا، تلفظ‌های گوناگون، نام‌های تغییر پیداکرده). در برخی موارد که کاربر یک نام را جستجو می‌کند، اگر آن شخص نام‌های مختلفی داشته باشد، ممکن است، همهٔ اسناد مرتبط با آن شخص را نتواند دریافت کند. برای مثال نام *محمدعلی حسینی‌یزدی* ممکن است در شکل‌های مختلف مانند محمد حسینی، محمدعلی حسینی و یا محمدعلی حسینی‌یزدی در مقالات نوشته شود. در جستجوهایی با استفاده از این نام‌ها در برخی از کتابخانه‌ها، این چند نام با نتایج مختلفی نمایش داده می‌شوند و تمام کارهای انجام‌شده توسط این نویسنده نمایش داده نمی‌شوند؛ درصورتی‌که همهٔ این نام‌ها تنها متعلق به یک شخص است. این نوع از نام مبهم را نام‌های به‌نسبه مشابه و یا رکوردهای جدا[7] می‌نامند.

نوع دوم: برعکس مورد قبل، هنگامی‌که دو موجودیت نام، املا و هجی یکسانی داشته باشند، مستندات آنها ممکن است ترکیب و باهم نمایش داده شوند که بدین ترتیب اطلاعات اشخاص دیگر برای یک نویسنده نمایش داده می‌شود. روش‌های ابهام‌زدایی در این نوع، نمونه‌های موجودیت را از هم جدا می‌کنند. برای مثال فرض کنید که نام محمد محمدی متعلق به دو شخص است، جستجویی صورت می‌پذیرد، یکی از این نام‌ها از دانشگاه تهران و دیگری از دانشگاه اصفهان است. اگر رکوردهای داخل کتابخانه ابهام‌زدایی و دسته‌بندی نشوند، در نتایج نمایش‌داده‌شده ممکن است رکوردهای هر دو نویسنده با هم نمایش داده شوند. این نوع از نام‌های مبهم را رکوردهای مخلوط[8] می‌نامند.

روش‌های ارائه‌شده برای حل نام‌های مبهم نویسنده‌ها در سه دسته کلی قرار دارند؛ روش‌هایی که فقط نوع نخست را مانند [6]، روش‌هایی که فقط نوع دوم را مانند [3]،[7] و گروهی از روش‌ها، که هر دو نوع، از نام‌های مبهم را ابهام‌زدایی می‌کنند[1]، [10]- [8]. به‌طورکلی ابهام‌زدایی موجودیت نام به‌عنوان مسئله خوشه‌بندی در نظر گرفته می‌شود که در آن هر خوشه به یک موجودیت معین اشاره می‌کند. کار اصلی روش‌های ارائه‌شده برای ابهام‌زدایی نام نویسنده، خوشه‌بندی‌کردن رکوردها است؛ به این شکل که هر خوشه به یک نویسنده خاص انتساب داده می‌شود. عملیات خوشه‌بندی با کمک ویژگی‌های موجود در مقالات

[1] Disambiguation
[2] Citation
[3] Bibliographic attributes
[4] Author affiliations
[5] Reference disambiguation
[6] Named Entity Disambiguation (NED)
[7] Split Citation (SC)
[8] Mixed Citation (MC)

[DOI: 10.29252/jsdp.14.4.117]

مانند نام نویسنده‌ها، عنوان مقاله، محل انتشار و مواردی از این دست انجام می‌شود.

روش‌های ارائه‌شده برای حل نام‌های مبهم از ترفندهای گوناگون، جهت حل این مشکل استفاده می‌کنند. برخی از روش‌های ارائه‌شده از روابط بین نویسندها بهره می‌برند که روابط نویسنده‌ها از طریق اطلاعات هم‌نویسندهای مقاله تشخیص داده می‌شوند؛ مانند کارهای انجام‌شده در مراجع [3]، [14]-[11]. روابط بین نویسنده‌ها اطلاعات مفیدی برای تشخیص نویسنده‌هایی که چند نام دارند یا چندین نویسنده که از یک نام مشترک استفاده می‌کنند، ارائه می‌دهند. برای مثال اگر دو نام مبهم از یک گروه، یک هم‌نویسندهٔ مشترک داشته باشند، به‌عنوان یک شخص در نظر گرفته می‌شوند. این فرضیه با درنظرگرفتن این مسئله صحیح است که در دنیای واقعی احتمال بسیار کمی وجود دارد دو نویسندهٔ مختلف با نام نسبتاً مشابه یک هم‌نویسندهٔ مشترک داشته باشند. علاوه‌براین، بیش‌تر پژوهش‌گران در زمینه‌های تخصصی خاصی فعالیت می‌کنند و به‌طورمعمول در آن زمینهٔ تخصصی مقاله می‌نویسند. به این دلایل مقاله‌های نوشته‌شده توسط نویسنده‌ها دربردارنده واژگان کلیدی هستند که بیان‌گر زمینهٔ تخصصی خاص هر نویسنده است [15]. در برخی روش‌ها، از اطلاعات معنایی، مانند عنوان، چکیده، واژگان کلیدی و غیره استفاده می‌کنند تا اشخاص با نام‌های مشابه را از یکدیگر تفکیک کنند.

چالش‌های بسیاری در این زمینه وجود دارد که باعث پیچیده‌شدن عملیات ابهام‌زدایی نام می‌شود. برای مثال نام‌های آسیایی به‌خاطر مختصربودن و هم‌نام‌بودن زیاد، عملیات ابهام‌زدایی را دچار خطا می‌کنند [1] و الگوریتم‌های ارائه‌شده در این نوع از نام‌ها جواب دقیقی به‌دست نمی‌آورند؛ یعنی به‌طورکلی به‌دلیل اشکال مختلف از نام‌های نویسنده‌های کل دنیا، یک الگوریتم دقیق وجود ندارد تا بتواند تمام نام‌های داخل یک فهرست، متعلق به کشورهای مختلف با اشکال گوناگون را حل کند. خوشه‌بندی از روش‌های کاربردی، مفید و مناسب جهت حل مسئلهٔ ابهام‌زدایی نام نویسنده‌ها، است. مشکل روش‌های مبتنی بر خوشه‌بندی، دست‌نیافتن به صحت کامل و تکه‌تکه[1]بودن نتایج برای یک نویسنده است. این تکه‌تکه‌بودن نتایج منجر به تولید خوشه‌های متعدد می‌شود که همگی تنها متعلق به یک نویسنده است. این مشکلات به‌دلیل عدم وجود شواهد[2] کافی جهت ترکیب و یکی‌کردن خوشه‌ها برای یک نویسنده ایجاد می‌شود.

در این مقاله روشی ارائه شده است که هدف آن استفاده از روش خوشه‌بندی تجمعی[3] در راستای کاهش تکه‌تکه‌بودن خوشه‌ها است. هدف از خوشه‌بندی تجمعی، ترکیب نتایج چند خوشه و به‌دست‌آوردن یک خوشه نهایی با صحت و درصد تکه‌تکه‌بودن پایین است. الگوریتم‌های خوشه‌بندی تجمعی، خوشه‌های بهتری تولید می‌کنند. روش پیشنهادی قادر به یکپارچه‌سازی نتایج حاصل از خوشه‌های توزیع‌شده یا تکه‌تکه‌شده از هر نویسنده است. این عمل، ابهام در نتایج نهایی روش‌های خوشه‌بندی را رفع می‌کند. در ادامه ابتدا به بررسی برخی از روش‌های پیشین در زمینهٔ ابهام‌زدایی نام پرداخته، سپس روش پیشنهادی و ارزیابی‌های آن ارائه می‌شود.

۲- پیشینهٔ پژوهش

روش‌های ابهام‌زدایی نام نویسنده‌ها در ادبیات به‌طورمعمول مبتنی بر روش‌های بانظارت[4] یا بدون نظارت[5] هستند. روش‌های بدون نظارت، برای گروه‌بندی‌کردن رکوردهای مربوط به یک نویسنده، از توابع مشابهت[6] جهت بررسی شباهت بین ویژگی‌ها استفاده می‌کنند. این توابع روی صفات موجود در رکوردهای مقالات یا بر روی داده‌های برگرفته از وب سایت‌ها تعریف شده‌اند. نمونه کارهای انجام‌شده از روش‌های بدون نظارت شامل [3]- [1]، [7] [9]، [20]- [16] می‌باشند.

در مقابل، روش‌های با نظارت از مجموعهٔ آموزشی که شامل نمونه‌های از قبل برچسب‌گذاری‌شده است، به‌منظور پیش‌بینی نویسنده یک رکورد یا تعیین اینکه آیا دو رکورد متعلق به یک نویسندهٔ یکسان هستند یا خیر، استفاده می‌کنند. کارهای انجام‌شده در [8]،[10]،[21]،[22] از روش‌های بانظارت جهت ابهام‌زدایی نام استفاده کرده‌اند.

در ادامه برخی از کارهای انجام‌شده جهت حل نام‌های مبهم با روش‌های خوشه‌بندی شرح داده می‌شود که در روش پیشنهادی ما مورد استفاده قرار گرفته‌اند.

[1] Fragmentation
[2] Evidences
[3] Clustering ensemble
[4] Supervised
[5] Unsupervised
[6] Similar functions

[DOI: 10.29252/jsdp.14.4.117]

در سال ۲۰۱۰ *کوتا* و همکاران [1] یک روش ابهام‌زدایی سلسله‌مراتبی اکتشافی[1] جهت حل هر دو مسئلهٔ رکوردهای جدا و رکوردهای مخلوط [23] ارائه کرده‌اند. این روش با استفاده از توابع مشابهت به‌همراه برخی از اکتشافات برای ابهام‌زدایی نام نویسنده‌ها ارائه شده است. این روش از ویژگی‌های هم‌نویسنده، عنوان و محل انتشار برای خوشه‌بندی استفاده می‌کند و پس از یک مرحله پیش‌پردازش، در دو گام، عملیات ابهام‌زدایی را انجام می‌دهد. مرحله پیش‌پردازش شامل حذف کلمات اضافی، کم‌اهمیت و گروه‌بندی‌کردن نویسنده‌های مبهم است. در هر گروه، نام‌های مشابه مبهم قرار می‌گیرند؛ که این عمل باعث کاهش پیچیدگی و افزایش صحت می‌شود. با درنظرگرفتن نامزدهایی برای خوشه‌بندی، در گام نخست این روش، تنها رکوردهایی که نام نویسندهٔ مبهم و یک هم‌نویسندهٔ مشابه دارند ترکیب می‌شوند. این فرضیه با درنظرگرفتن این مسئله است که در دنیای واقعی احتمال بسیار کمی وجود دارد دو نویسندهٔ مختلف با نام به‌نسبه مشابه، یک هم‌نویسندهٔ مشترک داشته باشند. در برخی موارد، رکوردهایی از گروه‌هایی که هیچ ویژگی مشترکی را به اشتراک نمی‌گذارند، با یکدیگر در یک خوشه قرار نمی‌گیرند؛ بنابراین شانس ترکیب انتشارات نویسنده‌های مختلف کاهش پیدا می‌کند. نکته قابل توجه برای حل هر دو زیر مسئلهٔ نام‌های مبهم این است که حل هم‌زمان آن دو، کار مشکلی است؛ زیرا این دو هدف در برخی از موارد تداخل دارند و یک راه‌حل خوب ایجاد تعادل صحیح بین آنهاست. دومین گام این روش، جهت کاهش تکه‌تکه‌بـودن خوشه‌های تولیدشده در مرحله قبل با استفاده از ویژگی‌های دیگر (ویژگی عنوان و محل انتشار مقاله) انجام می‌شود. در این مرحله، هرکدام از خوشه‌هایی که شامل عنوان و محل انتشار مشابه در رکوردهای خود باشند، با یکدیگر ترکیب می‌شوند. این مرحله بر اساس این فرضیه است که نویسنده‌ها بیشتر تمایل به انتشار مقالات در موضوع و محل انتشار یکسان، دارند. نتایج ارائه‌شده از این روش نشان می‌دهد که صحت آن تا حدود دوازده درصد از روش‌های بانظارت و بدون‌نظارت بیشتر است؛ اما این روش، در مواردی جهت ابهام‌زدایی رکوردهای خاص، دچار اشتباه شده است که به‌تبع در دیگر روش‌ها که از توابع مشابهت استفاده می‌کنند، دیده می‌شود. در شکست نخست این روش، دو نویسنده از گروه مبهم یکسان C.CHEN به‌عنوان یک نویسنده در نظر گرفته شده است. رکوردهای C1 و C2 شامل نام‌های مبهم مشابه (*chuen liang chen* و *c.chen*) و نام‌های هم‌نویسنده‌های مشابه (*biing feng wang* و *b.wang*) هستند. در این نمونه با استفاده از این روش، رکوردها به یک خوشه انتساب داده می‌شوند؛ درحالی‌که متعلق به دو نویسنده و دو خوشه مختلف هستند. در دومین نمونه شکست، رکوردها به‌اشتباه در گروه مبهم A.GUPTA قرار می‌گیرند. هر دو رکورد با یکدیگر در یک گروه قرار می‌گیرند؛ زیرا آنها نام نویسنده و محل انتشار مشابه دارند. درصورتی‌که رکورد نخست به *AMARNATH GUPTA* و رکورد دیگر به *AJAY GUPTA* اشاره می‌کند. سومین نمونه شکست، زمانی رخ می‌دهد که نام نویسنده‌ها در رکوردها مشابه‌اند و همچنین چندین کلمهٔ مشابه در عنوان مقاله دارند. در این نمونه هر رکورد شامل نام نویسندهٔ مشابه A.GUPTER و AJAY GUPTA و چهار عنوان مقاله (LOAD, BALANCE, ADAPTIV, METHOD) است. در نمونه شکست آخر دو رکورد از نویسنده‌ای یکسان (*AMAR GUPTA* پرفسور دانشکده مدیریت از دانشگاه ARIZONA) است. این رکوردها به گروه‌های مختلفی انتساب داده شده است؛ زیرا فهرست هم‌نویسنده‌ها، هیچ اشتراکی ندارند و در عنوان مقاله فقط یک واژه رایج وجود دارد و در محل انتشار مقالات نیز هیچ واژهٔ مشترکی وجود ندارد.

در سال ۲۰۱۰ *فراریا* و همکاران [24] از نتایج شبکه اجتماعی[2] برای ابهام‌زدایی نام استفاده کردند. آنها یک مطالعه از اهمیت اضافه‌کردن تحلیل شبکه اجتماعی به روش‌های سنتی در مسئلهٔ ابهام‌زدایی نام در کتابخانهٔ دیجیتال ارائه کردند. از طریق آزمایش‌ها با استفاده از زیرمجموعه‌ای از کتابخانهٔ واقعی، نشان داده شده که استفاده از تحلیل شبکه اجتماعی بهبود مهمی در کیفیت نتایج داشته است. آنها از تحلیل شبکه اجتماعی به‌عنوان مدرکی که دو نویسنده در دو مقاله مختلف در یک مجموعه داده با احتمال بیشتری نویسندهٔ یکسانی در دنیای واقعی هستند یا نه استفاده کرده‌اند. یک شبکه اجتماعی مجموعه‌ای از اشخاص یا بازیگرها است که هر بازیگر با زیرمجموعهٔ دیگر ارتباط برقرار می‌کند. در شبکه‌های همکارهای علمی، نویسنده‌ها اغلب بر روی حوزهٔ خاصی همکاری می‌کنند. همکاری بین دو نویسنده در وابستگی بین آنها تأثیر دارد.

[1] Heuristic based Hierarchical Clustering (HHC)

[2] Social network

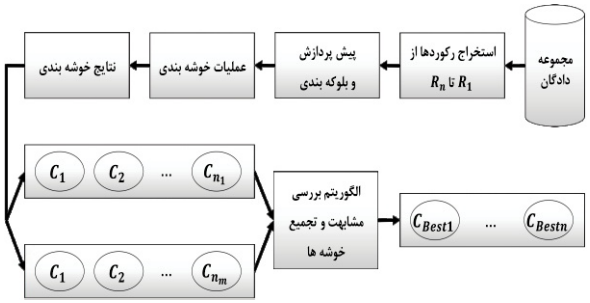
[DOI: 10.29252/jsdp.14.4.117]

ممکن است آنها به یک حوزهٔ یکسان علاقه داشته یا مربوط به سازمان یکسانی باشند. اگر فاصله بین دو نویسنده در شبکه کوچک باشد، شانس زیادی دارند که علایق یکسان داشته باشند و مربوط به سازمان یکسانی باشند. فاصله، اندازهٔ مهمی است؛ زیرا یک پدیده‌ای است که ابتدای کار شناخته می‌شود و به‌عنوان تأثیرات دنیای کوچک معرفی شده‌است؛ وحالتی است که هر دو شخص توسط یک مسیر شش الی هفت‌تـایی از اشخاص، به‌طور میانگین از یکدیگر جدا شده‌اند. برای مثال اهمیت مسیرهایی با فاصلهٔ بیش از سه نفر، در مسئلهٔ ابهام‌زدایی نام کاهش می‌یابد و ثابت شده است که مسیرهای برابر دو و کمتر بهترین میزان برای سنجش دو نام است. به‌علاوه اگر مسیرهای زیادی بین دو نویسنده وجود داشته باشد، نشان دهندهٔ این است که روابط قوی‌تری بین آنها وجود دارد؛ بنابراین اگر دو نویسنده از یک مجموعه داده، روابط قوی و درجه یکسانی از مشابهت معنایی داشتند، فرض می‌شود در دنیای واقعی با دقت بالایی تکرار شده‌اند. در ادامه روش پیشنهادی شرح داده می‌شود.

۳- روش پیشنهادی

روش‌های بسیاری برای حل نام‌های مبهم ارائه شده که در میان آنها به‌کارگیری روش خوشه‌بندی پرکاربرد بوده است. این روش‌ها دقتی بالاتر نسبت به روش‌های بانظارت داشته‌اند؛ اما دارای معایبی نیز هستند، که به‌تبع روش‌های دیگر ابهام‌زدایی نام نیز از این قاعده مستثنا نیستند. یکی از این معایب، درصد تکه‌تکه‌بودن بالا در نتایج خوشه‌بندی است که با این وجود باز هم شک و ابهام در شناسایی اطلاعات یک نویسنده موجود است. در این پژوهش، برای حل هر دو نوع از نام‌های مبهم، روشی ارائه شده است که با استفاده از الگوریتم‌های خوشه‌بندی تجمعی، صحت را افزایش و خوشه‌های تکه‌تکه‌شده از یک نویسنده را ترکیب می‌کند. با بررسی و کاوش در خوشه‌های تولیدشده و یافتن زوج اشتراکات در آنها، خوشه‌های تولیدشده را ترکیب و برای هر نویسنده یک خوشه منحصربه‌فرد می‌توان ایجاد کرد. نوآوری مدل پیشنهادی به‌کارگیری روش خوشه‌بندی تجمعی جهت ترکیب خوشه‌های متعدد (تکه‌تکه شده) یک نویسنده و ارائه خوشه‌ای منسجم و ترکیب‌شده از آنها است. در شکل (۱) فرآیند مدل پیشنهادی نشان داده شده است.

در مرحلهٔ نخست تمام رکوردهای متعلق به نویسنده‌ها با نام مشابه از پایگاه داده‌های کتابخانه‌های



(شکل-۱): مدل ارائه‌شده جهت ابهام‌زدایی نام نویسنده‌ها
(Figure-1): The proposed model for the author's name disambiguation

دیجیتال استخراج می‌شوند؛ سپس گام پیش‌پردازش که شامل حذف کلمات اضافی در عنوان و محل انتشار است و همچنین ماژول بلوکه‌بندی، انجام می‌شود. ماژول بلوکه‌بندی، نامزدهای مبهم را به طبقه‌های جداگانه گروه‌بندی می‌کند؛ سپس با به‌کارگیری یکی از روش‌های خوشه‌بندی مانند سلسله‌مراتبی تجمعی[1]، عملیات خوشه‌بندی گروه‌های مبهم انجام می‌شود. پس از خوشه‌بندی اولیه در این مرحله، خوشه‌های بسیاری برای هر نام ایجاد می‌شود. به‌طور میانگین برای هر نویسنده در هر اجرا، چهار الی بیست خوشه ایجاد می‌شود که این مسئله به‌عنوان یکی از چالش‌های اصلی در بحث خوشه‌بندی و ابهام‌زدایی نام است. در این پژوهش سعی می‌شود با استفاده از الگوریتم تشخیص مشابهت خوشه‌ها، محتویات مشترک، در خوشه‌هایی که متعلق به یک نویسنده است، تشخیص داده شود و درنهایت با تشخیص مشابهت میان خوشه‌ها، ترکیبات ممکن انجام و خوشه‌های غنی برای هر نویسنده ایجاد شود که علاوه‌بر داشتن صحت بالا، دارای اطلاعات و رکوردهای متعلق به یک نویسنده باشد. جهت تجمیع خوشه‌ها، ابتدا یکی از خوشه‌ها به‌عنوان مرجع در نظر گرفته و سپس بـــا استفاده از الگوریتم تشـــخیص مشـــابهت، هریک از خوشه‌ها با دیگر خوشه‌های مشابه آن ترکیب می‌شود.

الگوریتم (۱)، یک الگوریتم خوشه‌بندی سلسله‌مراتبی است که با عملکردی کلی و مشابه با دیگر روش‌های خوشه‌بندی، عملیات ابهام‌زدایی را انجام و یک فهرست اولیه از خوشه‌های ایجادشده از نویسنده‌ها تولید می‌کند. این الگوریتم با تغییرات جزئی برگرفته از روش پیشنهادی پیشین کار ما می‌باشد [25]. الگوریتم (۱)، یک فهرست از

[1] Agglomerative Hierarchical Clustering

[DOI: 10.29252/jsdp.14.4.117]

رکوردها A و یک فهرست از خوشه‌های رکوردها C_i را به‌عنوان ورودی دریافت می‌کند و یک فهرست جدید با هر رکورد a از A در برخی از خوشه‌های c از C_o تولید می‌کند. نام نویسنده در هر رکورد a با نام نویسنده از نخستین رکورد در داخل هر خوشهٔ c با استفاده از یک تابع مشابهت مقایسه می‌شود. اگر نام نویسندهٔ a با نام نویسنده از نخستین رکورد c مشابه و یک نام هم‌نویسنده از a مشابه با بعضی از نام‌های هم‌نویسنده‌ها در c باشد جواب مثبت را باز می‌گرداند، درنتیجه رکورد a در این خوشه c، درج می‌شود (خط۷)؛ در غیر این صورت یک خوشهٔ جدید با رکورد a ایجاد (خط۱۳) و به فهرست C_o اضافه می‌شود.

(الگوریتم-۱): خوشه‌بندی رکوردهای نویسنده‌های مبهم

(Algorithm -1): Ambiguous author records clustering

Input: List *A* of authorship records;
Output: List *Co* of authorship record clusters;
Begin
1 *Co* ← *Ci*;
2 **for each** *a* **in** *A* **do**
3 *inserted* ← **false**;
4 *c* ← first(*Co)*;
5 **while not** *inserted* **and** *c* ≠ **null do**
6 **if** the author name from *a* is similar to the author name from the first authorship record of *c* **and** there is a coauthor name in *a* that is similar to some coauthor name in *c* ***or*** there is a Title in *a* that is similar to some Title in *c* **or** there is a Venue title in *a* that is similar to some Venue title in *c* **then**
7 InsertAuthorshipRecord (*a*, *c)*;
8 *inserted* ← **true**;
9 **end if**
10 *c* ← next(*Co)*;
11 **end while**
12 **if** *inserted* =**false then**
13 *c* ← CreateNewCluster(*a)*;
14 Append(*Co*, *c)*;
15 **end if**
16 **end for**
End

پس از تولید فهرست خوشه‌های نام‌های مبهم، نوبت به بررسی مشابهت، جهت ترکیب و ادغام خوشه‌ها می‌رسد. در روش پیشنهادی از الگوریتم (۲) جهت ترکیب خوشه‌های تکه‌تکه‌شده از هر نویسنده استفاده می‌شود. در این الگوریتم فهرست خوشه‌های ایجادشده از مرحلهٔ قبل دریافت شده، سپس فهرست جدیدی که از ترکیب خوشه‌ها ایجاد شده است، ارائه می‌شود. جهت تشخیص خوشه‌های متعلق به یک نویسنده، فاصله بین هر جفت خوشه با استفاده از معیار مشابهت کوسینوس[1] محاسبه می‌شود؛ سپس جفت‌خوشه‌هایی که بیشترین مشابهت را با یکدیگر دارند (تا زمانی که برای هر خوشه یک متناظر پیدا شود)، به‌عنوان جفت خوشه‌های متناظر و مشابه انتخاب می‌شوند. این روند به‌صورت تکراری بین خوشه‌بندی مرجع و مرحلهٔ قبل با دیگر خوشه‌بندی‌ها انجام می‌شود.

(الگوریتم-۲): تشخیص مشابهت بین خوشه‌ها و ترکیب آنها

(Algorithm -2): Detect the similarity between clusters and their combination

Input: List *C* of clusters of authorship records;
Output: List C_{best} of clusters of authorship records;
Begin
1 *combination*←**true**;
2 **while** *combination* **do**
3 *combination* ←**false**;
4 **for each** *c*1 **in** *C* **do**
5 **for each** *c*2 **in** *C* **do**
6 *tn*1←Get authors name (*c*1);
7 *tn*2←Get authors name (*c*2);
8 *tt*1←GetWorkTitleTerms (*c*1);
9 *tt*2←GetWorkTitleTerms(*c*2);
10*tv*1←GetPublicationVenueTitleTerms(*c*1);
11*tv*2←GetPublicationVenueTitleTerms(*c*2);
12 **If** name Similarity**(***tn*1, *tn*2**)>** *name-threshold* **and** TitleSimilarity(*tt*1, *tt*2)>*title- threshold* **and** VenueSimilarity(*tv*1, *tv*2)>*venue-threshold* **then**
13 *c*1←*Combine*(*c*1, *c*2);
14 remove(*C*, *c*2);
15 remove duplicate citation records of each cluster c_1 in *C*.
16 *combination* ←**true**;
17 **end if**
18 **end if**
19 **end for**
20 **end for**
21 **end while**
End

پس از اعمال این الگوریتم، خوشه‌های متعلق به یک نویسنده که ابتدا به‌صورت خوشه‌های متعدد و تکه‌تکه‌شده بودند، ترکیب شده و در یک خوشه یا نزدیک به یک خوشه قرار می‌گیرند. الگوریتم (۲) در این روش، ابهامات موجود پس از اجرای روش‌های خوشه‌بندی را کاهش می‌دهد. با اضافه کـــــردن این الگوریتم به روش‌های دیگر، درصد تکه‌تکه‌بودن خوشه‌ها را به صفر می‌توان کاهش داد؛ به‌گونه‌ای که اطلاعات یک نویسنده در قالب یک خوشه

[1] Cosine similarity

[DOI: 10.29252/jsdp.14.4.117]

نمایش داده شود. در ادامهٔ ارزیابی‌ها و بررسی‌های انجام‌شده جهت تعیین کارایی و صحت روش پیشنهادی ذکر شده است.

۴- ارزیابی‌های روش پیشنهادی

آزمایش‌ها با هر گروه از نام‌های مبهم به‌صورت جداگانه انجام شده است که در آن نام نویسنده، فهرست نام هم‌نویسنده‌ها، عنوان مقاله و محل انتشار به‌عنوان صفات در نظر گرفته شده است. برای هر گروه نام مبهم ده مرتبه برنامه اجرا شده و نتایج هر مرتبه به ثبت رسیده تا میانگین آنها مشخص شود. در انتها میانگین تمام اجراها برای همهٔ گروه‌های نام مبهم به‌دست می‌آید. در این پژوهش پیاده‌سازی‌ها با زبان C# در محیط ویژوال استودیو نسخه ۲۰۱۲ انجام شده که بر روی رایانه‌ای با پردازنده i7 2/67 GHZ با 8GB RAM اجرا شده است.

مجموعه دادگان استفاده‌شده جهت بررسی صحت روش پیشنهادی از کتابخانه دیجیتال DBLP است. این مجموعه دربردارندهٔ چندین گروه نام مبهم است که تعداد رکوردهای استخراج‌شده از آن شامل ۴۲۸۷ رکورد و متعلق به ۲۲۰ نویسندهٔ مجزا است. همچنین در آزمایش‌ها از مقادیر ۰,۲ و ۰,۴ برای آستانه مشابهت[1] محل انتشار و عنوان مقاله در مجموعه دادگانDBLP استفاده شده است؛ زیرا این مقادیر بهترین نتایج را در آزمایش‌ها طبق اثبات صورت‌گرفته [1] تولید کرده است.

برای ارزیابی کیفیت خوشه‌های تولیدشده از معیار ارزیابی K [1] استفاده می‌شود. معیار K، از میانگین هندسی دو میانگین خالص‌بودن خوشه[2] (ACP) و خالص‌بودن نویسنده[3] (AAP) به‌دست می‌آید.

متغیر ACP خالص‌بودن خوشه‌های تولیدشده را با توجه به خوشه‌هایی که به‌صورت دستی تولید شده است، ارزیابی می‌کند. اگر خوشه‌های تولیدشده خالص باشند، جواب این متغیر یک می‌شود. فرمول ACP در رابطهٔ (۱) نشان داده شده است.

$$ACP = \frac{1}{N}\sum_{i=1}^{q}\sum_{j=1}^{R}\frac{n_{ij}^2}{n_i} \qquad (۱)$$

متغیرAAP تکه‌تکه‌بودن خوشه‌هایی را که به‌صورت خودکار تولید شده است، نسبت به خوشه‌های مرجع ارزیابی می‌کند. اگر تکه‌تکه‌بودن خوشه‌های تولیدشده کم باشند، نتیجه این متغیر نزدیک به یک است. مقادیر این متغیر بین صفر تا یک است. فرمول AAP در رابطهٔ (۲) نشان داده شده است.

$$AAP = \frac{1}{N}\sum_{j=1}^{R}\sum_{i=1}^{q}\frac{n_{ij}^2}{n_j} \qquad (۲)$$

در هر دو رابطهٔ (۱) و (۲)، متغیر R تعداد خوشه‌هایی که به‌صورت دستی تولید شده، می‌باشد (خوشه‌های مرجع). متغیر N تعداد کل رکوردها در گروه‌های نام مبهم است. متغیر q تعداد خوشه‌هایی است که توسط روش ابهام‌زدایی نام به‌صورت خودکار تولید شده است. متغیر n_{ij} تعداد کل عناصر خوشهٔ i است که به‌صورت خودکار تولید شده و متعلق به خوشهٔ j بوده که به‌صورت دستی انجام شده است. و متغیر n_i تعداد کل عناصر از خوشهٔ i که به‌صورت خودکار تولید شده، می‌باشد.

فرمول K طبق رابطهٔ (۳) محاسبه می‌شود.

$$K = \sqrt{ACP \times AAP} \qquad (۳)$$

در جدول (۱) میانگین نتایج پس از آزمایش‌های مکرر برای هر گروه نام مبهم در مجموعه دادگان DBLP، با روش پیشنهادی نشان داده شده است. در روش پیشنهادی هدف کاهش مقدار AAP است؛ البته حفظ و افزایش میزان صحت نیز حائز اهمیت است.

(جدول-۱): ارائه نتایج پس از ده بار اجرا برای تمام نام‌های مبهم مجموعه دادگان DBLP

(Table-1): The results after ten execution times for all ambiguous names of DBLP collections

نام‌های مبهم	ACP روش پیشنهادی	AAP روش پیشنهادی	K روش پیشنهادی
A.Gupta	0.87	0.79	0.83
A.Kumar	0.83	0.82	0.82
C.Chen	0.67	0.58	0.62
D.Johnson	0.92	0.69	0.80
J.Martin	0.91	0.79	0.85
J.Robinson	0.90	0.89	0.89
J.Smith	0.89	0.84	0.87
K.Tanaka	0.95	0.78	0.86
M.Brown	0.82	0.87	0.84
M.Jones	0.94	0.89	0.91
M.Miller	0.93	0.83	0.88
TOTAL	0.87	0.79	0.83

[1] Similarity threshold

[2] Average Cluster Purity (ACP)

[3] Average Author Purity (AAP)

[DOI: 10.29252/jsdp.14.4.117]

همان‌طور که مشاهده می‌شود، در جدول (۱) علاوه‌بر میزان صحت هر خوشه، مقدار تکه‌تکه‌بودن روش پیشنهادی نشان داده شده است. در برخی از گروه‌های نام‌های مبهم، پیچیدگی‌ها و مشکلاتی ذاتی وجود دارد که علت آن مشابهت زیاد نام‌های مبهم، نام هم‌نویسنده‌ها و حوزه‌های کاری است؛ اما با این‌حال، روش پیشنهادی با عملکرد تجمیع نتایج خوشه‌بندی توانسته است، درصد تکه‌تکه‌بودن خوشه‌ها را کاهش دهد که درنتیـــجه نهایتاً معیار K افزایش می‌یابد. به‌طورکلی باید میزان صحت و تکه‌تکه‌بودن بالا باشد تا بتوان اذعان کرد روش مورد نظر کارا و معقول بوده است.

در ارزیابی دیگری، مقایسه‌ای تحت شرایط یکسان بین نتایج روش پیشنهادی با روش‌های دیگر انجام شده است. جدول (۲)، میانگین نتایج نهایی تمامی گروه‌های نام مبهم را بر روی مجموعه دادگان DBLP با اجرای یکسان نشان می‌دهد.

(جدول-۲): میانگین نتایج روش پیشنهادی با روش‌های مقایسه‌ای

(Table-2): Average results of the proposed method with comparative methods

روش‌ها	ACP	AAP	K
روش پیشنهادی	**0.87**	**0.79**	**0.83**
HHC[1]	0.86	0.68	0.76
K-way[9]	0.75	0.47	0.59

همان‌طور که در جدول (۲) مشاهده می‌شود، روش پیشنهادی صحتی به‌نسبه بالاتر و درصد تکه‌تکه‌بودن بسیار کمتری نسبت به روش‌های مقایسه‌ای دارد و دارای بالاترین AAP یا کمترین میزان تکه‌تکه‌بودن است. درنتیجه با استفاده از تابع تجمیع و توابع مشابهت مناسب دیگر در این حوزه از پژوهش مشکل نام‌های مبهم تمامی افراد را با صحت بالاتر و تکه‌تکه‌بودن کمتری می‌توان حل کرد. دلیل این واقعیت، گستردگی اطلاعات هویت افراد و هم‌پوشانی‌های کم، در گذر زمان است. روش مقایسه‌ای نخست HHC توسط کوتا و همکاران [1] ارائه شده که میزان ۸۶ درصد صحت و ۶۸ درصد تکه‌تکه‌بودن را به‌دست آورده است که اغلب روش‌ها بـــا عملکرد مشابه روش پیشنهادی، در همین بازه بوده‌اند. روش مقایسه‌ای دوم روش خوشه‌بندی طیفی[1] [9] است. در این روش ابهام‌زدایی، هر رکورد توسط برداری از گراف بی‌جهت نمایش داده می‌شود که وزن لبه‌های بین دو رأس، بیانگر مشابهت بین رکوردها است که میزان ۷۵ درصد صحت و ۴۷ درصد تکه‌تکه‌بودن را به‌دست آورده است. درصد تکه‌تکه‌شدن در روش پیشنهادی نسبت به روش‌های پیشین بسیار کاهش یافته است که این بهبود با حفظ صحت و حتی افزایش آن همراه شده است. کاهش میزان تکه‌تکه‌بودن خوشه‌ها باعث افزایش مقدار K می‌شود لازم به ذکر است با اجرای الگوریتم پیشنهادی، صحت نیز بهبود دو الی پنج درصدی داشته و اطلاعات هر نویسنده در خوشه‌های صحیح خود قرار گرفته است.

[1] K-way spectral clustering method

۵- بحث و نتیجه‌گیری

در این مقاله یک روش جدید جهت کاهش تکه‌تکه‌بودن روش‌های خوشه‌بندی در مسئلهٔ ابهام‌زدایی نام نویسنده‌ها در کتابخانه‌های دیجیتال ارائه شده است. با استفاده از روش پیشنهادی در برخی موارد صحت ۹۵ درصد و به‌طور میانگین صحت ۸۳ درصد برای نام‌های مبهم مجموعه‌دادگان DBLP به‌دست آمده که نسبت به روش‌های پیشین بهبود داشته است. از دیگر نکات مثبت روش پیشنهادی به کاهش درصد تکه‌تکه‌بودن خوشه‌ها می‌توان اشاره کرد که درنتیجه باعث قرارگیری اطلاعات هر نویسنده در یک خوشه و کاهش ابهام در نتایج نهایی می‌شود.

۶- مراجع 6-References

[1] R. G. Cota, A. A. Ferreira, C. Nascimento, M. A. Gonçalves, and A. H. Laender, "An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations," *Journal of the American Society for Information Science and Technology,* vol. 61, pp. 1853-1870, 2010.

[2] B.-W. On and D. Lee, "Scalable Name Disambiguation using Multi-level Graph Partition," in *SDM*, 2007.

[3] X. Fan, J. Wang, X. Pu, L. Zhou, and B. Lv, "On graph-based name disambiguation," *Journal of Data and Information Quality (JDIQ),* vol. 2, p. 10, 2011.

[4] Z. Chen, D. V. Kalashnikov, and S. Mehrotra, "Adaptive graphical approach to entity resolution," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 2007, pp. 204-213.

[5] L. D. u Thu, "Named Entity Disambiguation in Digital Libraries," 2010.

[6] B.-W. On, E. Elmacioglu, D. Lee, J. Kang, and J. Pei, "Improving grouped-entity resolution using quasi-cliques," in *Data Mining, 2006. ICDM'06.*

[DOI: 10.29252/jsdp.14.4.117]

Sixth International Conference on, 2006, pp. 1008-1015.

[7] I.-S. Kang, S.-H. Na, S. Lee, H. Jung, P. Kim, W.-K. Sung*, et al.*, "On co-authorship for author disambiguation," *Information Processing & Management,* vol. 45, pp. 84-97, 2009.

[8] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, "Two supervised learning approaches for name disambiguation in author citations," in *Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on*, 2004, pp. 296-305.

[9] H. Han, H. Zha, and C. L. Giles, "Name disambiguation in author citations using a k-way spectral clustering method," in *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, 2005, pp. 334-343.

[10] J. Huang, S. Ertekin, and C. L. Giles, "Efficient name disambiguation for large-scale databases," in *Knowledge Discovery in Databases: PKDD 2006*, ed: Springer, 2006, pp. 536-544.

[11] B. Zhang and M. A. Hasan, "Name Entity Disambiguation in Anonymized Graphs using Link Analysis: A Network Embedding based Solution," *arXiv preprint arXiv:1702.02287,* 2017.

[12] S. Ressler, "Social network analysis as an approach to combat terrorism: Past, present, and future research," *Homeland Security Affairs,* vol. 2, pp. 1-10, 2006.

[13] F. H. Levin and C. A. Heuser, "Using Genetic Programming to Evaluate the Impact of Social Network Analysis in Author Name Disambiguation," in *AMW*, 2010.

[14] D. Shin, T. Kim, H. Jung, and J. Choi, "Automatic method for author name disambiguation using social networks," in *Advanced Information Networking and Applica-tions (AINA), 2010 24th IEEE Intern-ational Conference on*, 2010, pp. 1263-1270.

[15] Y. Ju, B. Adams, K. Janowicz, Y. Hu, B. Yan, and G. McKenzie, "Things and Strings: Improving Place Name Disambiguation from Short Texts by Combining Entity Co-Occurrence with Topic Modeling," in *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Procee-dings 20*, 2016, pp. 353-367.

[16] I. B. L. Getoor, "A Latent Dirichlet Model for Unsupervised Entity Resolution," in *Proceedings of the Sixth SIAM International Conference on Data Mining*, 2006, p. 47.

[17] I. Bhattacharya and L. Getoor, "Collective entity resolution in relational data," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 1, p. 5, 2007.

[18] Y. Song, J. Huang, I. G. Councill, J. Li, and C. L. Giles, "Generative models for name disambiguation," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 1163-1164.

[19] D. A. Pereira, B. Ribeiro-Neto, N. Ziviani, A. H. Laender, M. A. Gonçalves, and A. A. Ferreira, "Using web information for author name disambiguation," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, 2009, pp. 49-58.

[20] K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee, and J.-M. Ho, "Author name disambiguation for citations using topic and web correlation," in *Research and Advanced Technology for Digital Libraries*, ed: Springer, 2008, pp. 185-196.

[21] V. I. Torvik and N. R. Smalheiser, "Author name disambiguation in MEDLINE," *ACM Transactions on Knowledge Discovery from Data (TKDD),* vol. 3, p. 11, 2009.

[22] A. A. Ferreira, A. Veloso, M. A. Gonçalves, and A. H. Laender, "Effective self-training author name disambiguation in scholarly digital libraries," in *Proceedings of the 10th annual joint conference on Digital libraries*, 2010, pp. 39-48.

[23] W. W. Cohen, H. Kautz, and D. McAllester, "Hardening soft information sources," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 255-259.

[24] F. H. Levin and C. A. Heuser, "Evaluating the use of social networks in author name disambiguation in digital libraries," *Journal of Information and Data Management,* vol. 1, p. 183, 2010.

[25] M. H. Nadimi and M. Mosakhani, "A more Accurate Clustering Method by using Co-author Social Networks for Author Name Disambiguation," *Journal of Computing and Security,* vol. 1, 2015.

سید محمد مرتضوی مدرک کارشناسی ارشد خود را در رشته مهندسی کامپیوتر- هوش مصنوعی از دانشگاه آزاد اسلامی واحد نجف آباد در سال ۱۳۹۵ اخذ کرده‌انـد. ایشان در حال حاضر در زمینه داده‌کاوی مشغول پژوهش هستند.

نشانی رایانامه ایشان عبارت است از:

mortazavi.sm@sco.iaun.ac.ir

[DOI: 10.29252/jsdp.14.4.117]

محمد حسین ندیمی شهرکی مدرک دکترای خود را در رشته کامپیوتر-هوش مصنوعی از دانشگاه Putra مالزی در سال ۲۰۱۰ اخذ کرده‌اند. ایشان در حال حاضر استادیار دانشگاه آزاد اسلامی واحد نجف آباد و ریاست مرکز تحقیقات مِه داده (Big Data Research Center) را برعهده دارند و در زمینهٔ توسعه و تحلیل مِه داده‌ها و داده‌کـــاوی مشغول به تدریس و پژوهش هستند.

نشانی رایانامه ایشان عبارت است از:

nadimi@iaun.ac.ir

مصطفی موسی‌خـانی مدرک کارشناسی ارشد خود را در رشته کامپیوتر-نرم افزار از دانشگاه آزاد واحد نجف آباد در سال ۱۳۹۲ اخذ کرده اند. ایشان در حال حاضر مشغول به تحصیل در مقطع دکترا در واحد نجف‌آباد و در حال پژوهش در حوزه داده‌کاوی هستند.

نشانی رایانامه ایشان عبارت است از:

m_student1367@sco.iaun.ac.ir

[DOI: 10.29252/jsdp.14.4.117]