A Persian Citation Parsing Method Using Support Vector Machine

Nasrollah Pakniat*
PhD in Mathematics; Assistant Professor; Iranian Research Institute for Information Science and Technology (IranDoc); Thran, Iran Email: pakniat@irandoc.ac.ir

Jalal A. Nasiri
PhD in Computer Engineering; Assistant Professor; Faculty of Mathematical Sciences; Ferdowsi University of Mashhad; Mashhad, Iran Email: jnasiri@um.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. 37 | No. 4 | pp. 1245-1268
Summer 2022
https://doi.org/10.35050/JIPM010.2022.009

Received: 28, Apr. 2021 **Accepted: 10, Oct. 2021**

Abstract: Human users can easily divide a bibliographic reference to its constructing fields such as authors, title, journal, year, etc. However, due to the variations in formats and errors made by the authors in citing documents, it is difficult to automate this task. There exist many solutions for this problem, known as citation parsing problem in the literature, however, none of them is compatible with Persian language. This is mainly due to high language-sensitivity of these solutions. Considering the important role of citation parsing in tasks such as autonomous citation indexing and information retrieval, in this paper, we propose an intelligent method for citation parsing in Persian language. The proposed method uses the support vector machine (SVM) classification method as its core. The results of testing the proposed method using a dataset designed for this task show 95% in average for precision, recall and F1 measures for extracting different fields from a bibliographic reference which is quite plausible.

Keywords: Citation Parsing, Classification, Multi-class Classification, Supports Vector Machine, Autonomous Citation Indexing

* Corresponding Author

تجزیهٔ متون استنادی در زبان فارسی با استفاده از ماشین بُردار پشتیبان

نصراله پاک‌نیت
دکتری ریاضی؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
پدیدآور رابط pakniat@irandoc.ac.ir

جلال‌الدین نصیری
دکتری مهندسی کامپیوتر؛ استادیار؛
دانشکدهٔ علوم ریاضی؛ دانشگاه فردوسی مشهد؛
مشهد، ایران jnasiri@um.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۰۲/۰۸ **پذیرش:** ۱۴۰۰/۰۷/۱۸ مقاله برای اصلاح به مدت ۴۵ روز 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 ۳۷ | شمارهٔ ٤ | ص‌ص ۱۲٤٥-۱۲٦۸
تابستان ۱٤۰۱
https://doi.org/10.35050/JIPM010.2022.009

چکیده: یک متن استنادی را می‌توان به‌عنوان مجموعه‌ای از مؤلفه‌ها مانند نام نویسندگان، عنوان، محل نشر، سال نشر، شمارهٔ صفحات و ... در نظر گرفت. در حالی که تجزیهٔ متون استنادی موجود در انتهای یک مدرک علمی توسط کاربر انسانی به‌راحتی انجام‌پذیر است، تنوع موجود در شیوه‌های استناددهی در کنار اشتباهات رخ‌داده توسط نویسندگان در نگارش این متون، خودکارسازی انجام این عملیات را دشوار نموده است. روش‌های زیادی برای خودکارسازی تجزیهٔ متون استنادی ارائه شده، اما این روش‌ها وابسته به زبان بوده و به‌کارگیری یک روش ارائه شده برای یک زبان در زبانی دیگر منجر به نتایجی اشتباه می‌شود. تحقیقات صورت‌گرفته بیانگر آن است که تاکنون هیچ روشی برای خودکارسازی تجزیهٔ متون استنادی در زبان فارسی ارائه نشده است. با توجه به این مهم و نقش گستردهٔ این مسئله در ساخت خودکار شبکه‌های استنادی مدارک علمی و فرایندهای بازیابی اطلاعات، در این مقاله به این مسئله پرداخته شده و با استفاده از روش یادگیری ماشین بُردار پشتیبان به‌عنوان یک دسته‌بند چنددسته‌ای، یک روش هوشمند برای مسئلهٔ تجزیهٔ متون استنادی در زبان فارسی ارائه شده است. با توجه به اهمیت انتخاب ویژگی‌های مناسب برای استفاده در دسته‌بند ماشین بُردار پشتیبان، در این پژوهش این مهم با توجه به ویژگی‌های استفاده‌شده در زبان انگلیسی و ویژگی‌های زبان فارسی و ارجاع‌دهی در این زبان انجام شده است. نتایج پیاده‌سازی و آزمایش روش پیشنهادی با استفاده از مجموعهٔ داده‌ای ایجادشده در این پژوهش نشانگر مقدار ۰/۹۵ برای پارامترهای دقت، فراخوانی و اف-۱ است.

کلیدواژه‌ها: تجزیه متون استنادی، دسته‌بندی، دسته‌بندی چنددسته‌ای، ماشین بُردار پشتیبان، ساخت خودکار شبکه‌های استنادی

۱. مقدمه

ارجاعات یا استنادها، که در مدارک علمی به‌وسیلهٔ متون استنادی به‌طور معمول، در انتهای مدارک آورده می‌شوند، نقش مهمی در کتابخانه‌های دیجیتال انتشار علمی مانند CiteSeer، arXiv e-Print، DBLP، Google Scholar و Scopus ایفا می‌کنند. افزون بر استفاده از ارجاعات به‌عنوان ابزاری برای یافتن اطلاعات مورد علاقه، از این داده‌ها به‌عنوان معیاری برای بررسی تأثیر و اهمیت یک مقالهٔ مشخص نیز استفاده می‌شود. در سطحی بالاتر، از میزان ارجاعات انجام‌شده به تحقیقات پژوهشگران به‌عنوان معیاری برای بررسی صلاحیت آن‌ها برای ارتقا، اعطای پژوهانه و پروژه‌های تحقیقاتی استفاده می‌شود. افزون بر موارد فوق، از ارجاعات به‌عنوان ابزاری کمکی در فرایندهای بازیابی اطلاعات مانند خوشه‌بندی خودکار مدارک و نمایه‌سازی نیز استفاده می‌شود. مسئلهٔ اصلی در تمام کاربردهای فوق، ارائهٔ روشی برای تجزیهٔ متون استنادی به اقلام اطلاعاتی اصلی تشکیل‌دهندهٔ آن‌ها مانند نام نویسندگان، عنوان، محل نشر، تاریخ و شمارهٔ صفحات است که مسئلهٔ "citation parsing" نامیده شده و در این مقاله از «تجزیهٔ متون استنادی» به‌عنوان معادل فارسی آن استفاده می‌شود. اگرچه تجزیهٔ متون استنادی توسط کاربر انسانی به‌سادگی انجام‌پذیر است، اما به‌دلایلی مانند اشتباهات صورت‌گرفته در وارد کردن داده‌ها، شیوه‌های مختلف به‌کاررفته در نگارش متون استنادی، اشکالات موجود در نرم‌افزارهای جمع‌آوری مراجع، به‌کار بردن اختصارات در نوشتن اسامی و محل‌های نشر، حذف برخی از قسمت‌های نام نویسندگان در اسامی چندبخشی و حجم زیاد داده‌ها، خودکارسازی این مسئله به‌سادگی امکان‌پذیر نیست. الگوریتم‌های زیادی برای تجزیهٔ یک متن استنادی در زبان انگلیسی ارائه شده که هر یک به شیوه‌ای قابل قبول قادر به تجزیهٔ یک متن استنادی ورودی هستند. با وجود این، روش‌های تجزیهٔ متون استنادی به زبان وابسته بوده و استفاده از یک روش ارائه‌شده برای یک زبان در زبانی دیگر به نتایجی با اشتباه بیشتر منجر می‌شود. در نتیجه، نمی‌توان از روش‌های موجود برای تجزیهٔ متون استنادی در سایر زبان‌ها در زبان فارسی استفاده نمود.

با توجه به اهمیت تجزیهٔ خودکار متون استنادی در کاربردهایی مانند نمایه‌سازی

خودکار و تحلیل استنادی، در این مقاله به مسئلهٔ تجزیهٔ خودکار متون استنادی در زبان فارسی پرداخته خواهد شد. در این راستا، ابتدا به بررسی روش‌های ارائه‌شده برای تجزیهٔ متون استنادی در زبان انگلیسی پرداخته و سپس، با در نظر گرفتن ویژگی‌های به‌کاررفته برای تشخیص اقلام اطلاعاتی مختلف تشکیل‌دهندهٔ متون استنادی در زبان انگلیسی و همچنین، ویژگی‌های زبان فارسی و مراجع در زبان فارسی، این ویژگی‌ها را بومی‌سازی کرده و با استفاده از روش یادگیری ماشین بُردار پشتیبان، روشی هوشمند برای تجزیهٔ خودکار متون استنادی نوشته‌شده به زبان فارسی و استخراج اقلام اطلاعاتی مختلف آن‌ها ارائه می‌کنیم. برای استفاده از روش‌های یادگیری ماشین در حل مسئلهٔ تجزیهٔ متون استنادی به مجموعه‌ای برچسب‌گذاری‌شده از متون استنادی به زبان فارسی نیاز داریم که با توجه به نبودِ چنین مجموعه‌ای، آن را ایجاد کرده و از آن در پیاده‌سازی و آزمایش روش پیشنهادی استفاده خواهیم کرد. لازم به ذکر است که در این مقاله تنها استناد به مقالات همایش‌ها و نشریه‌ها، کتاب‌ها، و پایان‌نامه‌ها و رساله‌های دانش‌آموختگان را در نظر گرفته و از دیگر انواع کمتر رایجِ استناد مانند منابع اینترنتی، فصلی از یک کتاب و یا گفت‌وگوهای شفاهی و یا فیلم و مصاحبه صرفنظر می‌کنیم که این امر با توجه به استفادهٔ بسیار کمتر از این نوع استنادها در مقایسه با چهار نوعِ در نظر گرفته‌شده قابل توجیه است.

۲. پیشینهٔ پژوهش

مسئلهٔ تجزیهٔ متون استنادی را می‌توان به‌عنوان قسمتی از مسئلهٔ کلی‌تر استخراج فراداده از متون علمی در نظر گرفت (Rizvi, Dengel and Ahmed 2020; Hashmi Afzal and ur Rehman 2020; Ahmed and Afzal 2020; Tkaczyk 2017 و). با توجه به ادبیات مسئله، روش‌های ارائه‌شده برای این مسئله را می‌توان به دو دستهٔ زیر تقسیم‌بندی کرد:

۱. روش‌های مبتنی بر قاعده[1]؛
۲. روش‌های مبتنی بر یادگیری ماشین[2].

تحقیقات صورت‌گرفته نشان‌دهندهٔ این است که استفاده از روش‌های مبتنی بر یادگیری ماشین در حالت کلی منجر به نتایج بسیار بهتری شده و روش‌های مبتنی بر

1. rule based methods
2. machine learning based methods

قاعده تنها در موقعیت‌هایی که بدانیم متون استنادی مورد آزمایش از مجموعه‌ای مشخص از قالب[1]ها یا شیوه‌های استناددهی تبعیت کرده و با استفاده از این قالب‌ها و شیوه‌ها ایجاد شده‌اند، کارا هستند (Nasar, Jaffry and Malik 2018 و Tkaczyk et al. 2018a). در ادامهٔ این بخش، روش‌های ارائه‌شده در هر یک از دو دستهٔ مورد نظر را بررسی خواهیم کرد.

۲-۱. روش‌های تجزیهٔ متون استنادی مبتنی بر قاعده

در این بخش، به‌طور خلاصه، به بررسی روش‌هایی که با استفاده از قواعد اکتشافی به تجزیهٔ متون استنادی می‌پردازند، خواهیم پرداخت. نویسندگانی همچون «لاورنس، گیلز و بولاکر» مسئلهٔ تطبیق خودکار متون استنادی را که کلی‌تر از مسئلهٔ تجزیهٔ متون استنادی است، بررسی کرده و بیان کرده‌اند که تجزیهٔ متون استنادی موجب بهبود نتایج در تطبیق متون استنادی خواهد شد (Lawrence, Giles and Bollacker 1999a, b). نویسندگانی هم با انجام تحلیل‌هایی ساختاری و نحوی، یک روش تجزیهٔ متون استنادی برای مقالات اسکن‌شده ارائه کرده‌اند (Besagni, Belaïd & Benet 2003). «دینگ، چاودوری و فو» با استفاده از مجموعه‌ای از قواعد به تجزیهٔ متون استنادی در مقالات چاپی و برخط پرداخته‌اند. نتایج به‌دست‌آمده نشانگر نتایج بهتر در مورد مقالات چاپی (به‌دلیل رعایت کردن بهتر فرمت‌های نگارش مراجع) بوده است (Ding, Chowdhury and Foo 1999). در پژوهش «هوآنگ» و همکاران، پایگاه داده‌ای شامل قالب‌های شناخته‌شده از متون استنادی ایجاد شده و سپس، تجزیهٔ متون استنادی با استفاده از این پایگاه داده و یکی از روش‌های همترازسازی توالی[2] به نام BLAST انجام می‌شود (Huang et al. 2004). «گوپتا» و همکاران در پژوهش خود از عبارات منظم و اطلاعاتی پایه‌ای در زمینهٔ مورد نظر (مانند لیستی از عناوین نشریات برای برچسب‌زنی به یک قسمت به‌عنوان نشریه) استفاده کرده و روشی دیگر برای تجزیهٔ متون استنادی ارائه داده‌اند (Gupta et al. 2009).

۲-۲. روش‌های تجزیهٔ متون استنادی مبتنی بر یادگیری ماشین

مسئلهٔ تجزیهٔ متون استنادی را می‌توان به‌عنوان یک مسئلهٔ برچسب‌گذاری توالی[3] یا یک مسئلهٔ دسته‌بندی چنددسته‌ای[4] در نظر گرفت. با توجه به کارایی مناسب روش‌های

1. template
2. sequence alignment
3. sequence labeling
4. multi-class classifier

یادگیری ماشین در حل مسائل برچسب‌گذاری توالی و دسته‌بندی چنددسته‌ای، در سالیان گذشته به‌طور گسترده از روش‌های یادگیری ماشین برای حل مسئلهٔ تجزیهٔ متون استنادی استفاده شده است. در پژوهش «هتزنر» از مدل مخفی «مارکوف» (HMM)[1] برای تجزیهٔ متون استنادی استفاده شده است (Hetzner 2008). HMM استفاده شده در این مقاله، فراوانی لغات و اطلاعات مکانی[2] لغات در اقلام اطلاعاتی را در نظر می‌گیرد. با این حال، این مدل ترتیب حضور لغات در اقلام اطلاعاتی را در نظر نمی‌گیرد. برای در نظر گرفتن این مهم و بهبود نتایج، در پژوهش «یین» و همکاران، از HMM دوتایی برای حل مسئلهٔ تجزیهٔ متون استنادی استفاده شده است که روابط دنباله‌های دوتایی لغات[3] را نیز در نظر می‌گیرد (Yin et al. 2004). «اجوکو، ژانگ و تانگ» در پژوهش خود از مدل HMM سه‌تایی[4] برای حل مسئلهٔ تجزیهٔ متون استنادی استفاده کرده‌اند (Ojokoh, Zhang and Tang 2011). در روش جدید از یک ماتریس انتقال سه‌بعدی استفاده شده که در نتیجهٔ آن، انتقال به یک حالت نه‌تنها به حالت قبل، بلکه به حالت قبل‌تر نیز وابسته است. در برخی دیگر از روش‌های موجود، مسئلهٔ تجزیهٔ متون استنادی به‌عنوان یک مسئلهٔ برچسب‌زنی توالی در نظر گرفته شده و از میدان‌های تصادفی شرطی[5] (CRF) برای حل آن استفاده شده است. با توجه به بررسی‌های صورت‌گرفته، CRF استفاده شده در کارهای انجام گرفته در این راستا مشابه بوده و تفاوت کارهای انجام گرفته در این زمینه در ویژگی‌های[6] به‌کاررفته در آن‌ها و همچنین، در زمینهٔ علمی که در آن مسئلهٔ تجزیهٔ متون استنادی مورد بررسی قرار گرفته، دیده می‌شود (Councill, Giles and Kan 2008; Lopez 2009; Zou, Le and Thoma 2010; Zhang et al. 2011; Kim et al. 2012; Peng and McCallum 2013; Tkaczyk et al. 2014; Tkaczyk et al. 2015; و Namikoshi et al. 2017).

«زو، لو و توما» مسئلهٔ تجزیهٔ متون استنادی را به‌عنوان یک مسئلهٔ دسته‌بند چنددسته‌ای در نظر گرفته و قلم اطلاعاتی متناظر با هر واحد موجود در متن استنادی را با استفاده از دسته‌بند چنددسته‌ای ماشین بردار پشتیبان[7] مشخص می‌کنند (Zou, Le and Thoma 2010). افزون بر این، نویسندگان توجه کرده‌اند که قواعدی اکتشافی وجود دارند که همیشه در مورد متون استنادی صحیح هستند. با توجه به این مهم، نویسندگان پس از دسته‌بندی هر

1. Hidden Markov Model (HMM)
2. position information
3. bigram sequential relation
4. trigram HMM
5. conditional random field (CRF)
6. feature
7. support vector machine (SVM)

واحد موجود در دنبالهٔ واحدهای یک متن استنادی مورد آزمایش، از این قواعد اکتشافی استفاده کرده و وجود تناقض با این قواعد را بررسی می‌کنند. در پژوهش آن‌ها برای طراحی دسته‌بند مورد نظر از ۱۵ ویژگی استفاده شده است: سه ویژگی برای بررسی وجود واحد در لغت‌نامه‌هایی از اسامی نویسندگان، عنوان مقالات و عنوان نشریات (ایجادشده از داده‌های ۱۰ سال مدلاین[1])، یک ویژگی برای بررسی مشابهت واحد با فرمت شمارهٔ صفحه، یک ویژگی برای بررسی شباهت واحد با فرمت نگارش نام یک نویسنده، (مانند N. و H.-N.)، یک ویژگی برای بررسی عددی ۴ رقمی و کمتر از سال فعلی بودن واحد، یک ویژگی برای بررسی بودن واحد به‌صورت‌های et، al. یا al.، یک ویژگی برای بررسی بودن واحد به‌صورت مشخص‌کنندهٔ شمارهٔ صفحه (p.، pp.، p یا pp)، یک ویژگی برای بررسی خاتمه‌یافتن واحد با "."، یک ویژگی برای بررسی بزرگ یا کوچک بودن حرف اول واحد، یک ویژگی برای بررسی وجود کاراکتری غیر از حرف در واحد، یک ویژگی برای بررسی وجود تنها رقم در واحد، یک ویژگی برای بررسی وجود همزمان عدد و حرف در واحد، یک ویژگی برای بررسی وجود تنها عدد و حرف در واحد و یک ویژگی برای نمایش موقعیت نرمال شدهٔ واحد در متن استنادی (نرمال شده با توجه به تعداد واحدهای موجود در متن استنادی). افزون بر این، از آنجا که در مسئلهٔ تجزیهٔ متون استنادی، واحدهای همسایه با احتمال بالاتری برچسب یکسانی خواهند داشت، در اینجا، در دسته‌بندی هر واحد، ویژگی‌های واحدهای قبل و بعد از آن نیز در نظر گرفته می‌شوند. در پژوهش «ژانگ» و همکاران بیان شده است که با توجه به ساختار منظم موجود در متون استنادی، مسئلهٔ تجزیهٔ مرجع را می‌توان به‌عنوان یک مسئلهٔ یادگیری توالی در نظر گرفت و در ادامه، از SVM ساختاری برای حل این مسئله استفاده کرد (Zhang et al., 2011). ویژگی‌های در نظر گرفته شده در این پژوهش نیز برای هر واحد همانند آنچه که در پژوهش Zou, Le and Thoma (2010) آمده، در نظر گرفته شده است.

بررسی‌های صورت‌گرفته در پژوهش Tkaczyk et al. (2018b, 2018c) نشان‌دهندهٔ این است که روشی وجود ندارد که همواره بهترین نتیجه را ارائه کند. با توجه به این فرض و همچنین با فرضِ در دسترس بودن چندین روش تجزیهٔ متون استنادی، در این پژوهش نویسندگان از روش‌های یادگیری ماشین برای شناسایی بهترین روش تجزیه از میان روش‌های موجود و به‌کارگیری آن استفاده کرده‌اند.

1. MEDLINE

در پژوهش Bhardwaj et al. (2017) فرض بر این بوده که اسناد علمی ورودی تصویر بوده و با اتخاذ رویکردی متفاوت به‌جای تبدیل تصویر به متن و سپس، تجزیهٔ متون استنادی، به‌طور مستقیم اقدام به این کار نموده‌اند. روش مورد استفاده در این پژوهش یادگیری عمیق بود. با این حال، با توجه به رویکرد اتخاذشده، نتایج این روش در مقایسه با نتایج دیگر روش‌های مبتنی بر یادگیری ماشین ضعیف‌تر است.

یادگیری عمیق از دیگر ابزارهایی است که از آن برای حل مسئلهٔ تجزیهٔ متون استنادی استفاده شده و نتایج مناسبی را ارائه کرده است (An et al. 2017 و Prasad, Kaur and Kan 2018). با این حال، روش‌های یادگیری عمیق به‌طور ذاتی نیازمند مجموعه‌ای عظیم از داده‌های آموزش بوده و برای به‌کارگیری در زبان‌هایی با منابع محدود مناسب نیستند.

۳. روش پژوهش

در این پژوهش با مطالعهٔ منابع علمی معتبر روشی پیشنهاد می‌شود که از طریق آن بتوان یک متن استنادی ورودی را به مؤلفه‌های سازندهٔ آن تجزیه کرد. برای جمع‌آوری اطلاعات پژوهش از روش کتابخانه‌ای استفاده شده و نتایج به‌کارگیری روش پیشنهادی روی مجموعهٔ داده‌ای طراحی‌شده برای این مسئله در این پژوهش، ارزیابی و تحلیل شده است. برای تحلیل نتایج و مقایسهٔ آن‌ها از شاخص‌های مختلفی نظیر دقت، فراخوانی و معیار اف-۱ استفاده شده است.

پیاده‌سازی روش ارائه‌شده با استفاده از زبان برنامه‌نویسی «پایتون» انجام شده و برای استفاده از دسته‌بند SVM چنددسته‌ای از کتابخانهٔ «سکلرن»[1] استفاده شده است. برای آموزش و بررسی کیفیت روش پیشنهادی، یک مجموعهٔ داده در این پژوهش ایجاد شده و با استفاده از روش اعتبارسنجی متقابل ۱۰ سطحی، روش پیشنهادی آموزش داده شده و مورد آزمایش قرار گرفته است.

٤. دسته‌بند ماشین بُردار پشتیبان (SVM)

همان‌طور که در بخش پیش مشخص گردید، استفاده از روش‌های مبتنی بر یادگیری ماشین در مقایسه با روش‌های مبتنی بر قاعده، نتایج بسیار بهتری برای حل مسئلهٔ تجزیهٔ متون استنادی ارائه می‌دهند. افزون بر این، بررسی پیشینه نشان می‌دهد که از میان

1. Sklearn

روش‌های یادگیری ماشین استفاده شده برای حل مسئلهٔ مورد نظر (HMM، CRF و SVM)، SVM و CRF نتایج بهتری ارائه کرده‌اند. با توجه به روند استفاده و برتری کلی SVM نسبت به CRF، در این مقاله روش SVM را انتخاب کرده و از آن برای حل مسئلهٔ تجزیهٔ متون استنادی در زبان فارسی استفاده خواهیم کرد. با توجه به این مهم، در این بخش جزئیات روش دسته‌بندی ماشین بُردار پشتیبان (SVM) ارائه خواهد شد. لازم به ذکر است که اغلب مطالب ذکرشده در این بخش با استفاده از پژوهش «کارگر» (۱۳۹۰) و «نصیری» (۱۳۹۴) نگاشته شده‌اند.

٤-۱. SVM برای دسته‌بندی داده‌های دودسته‌ای

ساده‌ترین نوع روش SVM برای دسته‌بندی داده‌های دودسته‌ای جدایی‌پذیر خطی، روش کلاسیک SVM است. این روش مبنای بسیاری دیگر از انواع روش‌های SVM نیز به شمار می‌رود (Vapnik 1995, 1998). فرض کنید تعداد M داده عددی به‌صورت بُردارهای ستونی m بعدی x_1، ...، x_M متعلق به دسته‌های ۱ یا ۲ در اختیار باشند. متناظر با هر x_i می‌توان یک اسکالر y_i در نظر گرفت به‌طوری که اگر x_i متعلق به دستهٔ ۱ باشد، y_i مقدار ۱ و چنانچه متعلق به دستهٔ ۲ باشد، مقدار ۱- دریافت کند. با توجه به این دسته‌بندی برای مجموعهٔ داده‌ها، می‌توان نمایشی به‌صورت $s = \{(x_i, y_i)\}_{i=1}^{M}$ برای آن‌ها در نظر گرفت. در ادامه، در روش SVM هدف این است که مجموعهٔ داده‌های $s = \{(x_i, y_i)\}_{i=1}^{M}$ به‌وسیلهٔ یک ابرصفحهٔ خطی $\bar{w}^T x + \bar{b} = 0$ که $\bar{w}^T \in R^M$ و $\bar{b} \in R$، به نحوی از یکدیگر جدا شوند که رابطهٔ زیر برقرار باشد:

$$\begin{cases} \bar{w}x_i + \bar{b} > 0 & y_i = 1 \\ \bar{w}x_i + \bar{b} < 0 & y_i = -1 \end{cases} \quad i = 1,2,,\dots,M. \qquad (۱)$$

رابطهٔ فوق را می‌توان به‌صورت زیر بازنویسی کرد:

$$\exists \varepsilon > 0 \; s.t.: \begin{cases} \bar{w}x_i + \bar{b} \geq \varepsilon & y_i = 1 \\ \bar{w}x_i + \bar{b} \leq \varepsilon & y_i = -1 \end{cases} \quad i = 1,2,,\dots,M \qquad (۲)$$

با ضرب طرفین نامعادلات موجود در رابطهٔ فوق در عدد $\frac{1}{\varepsilon} > 0$ و جای‌گذاری $w = \frac{1}{\varepsilon}\bar{w}$ و $b = \frac{1}{\varepsilon}\bar{b}$ خواهیم داشت:

$$y_i(w^T x_i + b) \geq 1 \qquad i = 1,2,,\dots,M \qquad (۳)$$

تعریف. مجموعهٔ داده‌های $s = \{(x_i, y_i)\}_{i=1}^{M}$ را جداپذیر خطی[1] گویند هرگاه $w^T \in R^m$ و $b \in R$ موجود باشند، به‌طوری که با استفاده از این پارامترها رابطهٔ فوق برای مجموعهٔ داده s برقرار باشد. در این صورت، ابرصفحهٔ $w^T x + b = 0$، ابرصفحهٔ جداکنندهٔ داده‌های s نامیده می‌شود. افزون بر این، دو ابرصفحهٔ $w^T x + b = 1$ و $w^T x + b = -1$ ابرصفحات حاشیه‌ای[2]، ناحیهٔ بین این دو ابرصفحهٔ حاشیهٔ سخت[3]، فاصلهٔ بین این دو ابرصفحهٔ پهنای حاشیه سخت و بُردار x_i که در یکی از ابرصفحات حاشیه‌ای صدق کند، بُردار پشتیبان[4] نامیده می‌شوند.

با توجه به این‌که ابرصفحات حاشیه‌ای با یکدیگر موازی هستند، بیشترین پهنای حاشیهٔ سخت مربوط به ابرصفحات حاشیه‌ای است که بیشترین فاصله را از یکدیگر داشته باشند. با در نظر گرفتن $\frac{2}{\|w\|}$ به‌عنوان فاصله بین دو ابرصفحهٔ حاشیه‌ای، هدف در استفاده از SVM عبارت است از کمینه‌کردن $\frac{1}{2}\|w\|^2$، به‌طوری که

$$y_i(w^T x_i + b) \geq 1 \qquad i = 1,2,,\dots,M, w \in R^m, b \in R \tag{۴}$$

تعریف. ابرصفحهٔ جداکننده با بیشترین حاشیهٔ سخت مربوط به مجموعهٔ داده‌های $s = \{(x_i, y_i)\}_{i=1}^{M}$ را ابرصفحهٔ جداکنندهٔ بهینه[5] یا به اختصار ابرصفحهٔ بهینه نامند. افزون بر این، تابع $D(x) = w^T x + b$ را تابع تصمیم[6] مجموعهٔ داده‌های $s = \{(x_i, y_i)\}_{i=1}^{M}$ گویند هرگاه $w^T x + b = 0$ ابرصفحهٔ بهینهٔ جداکنندهٔ این مجموعه باشند.

همان‌طور که پیش‌تر نیز بیان شد، هدف در استفاده از SVM عبارت است از استفاده از داده‌هایی با دستهٔ مشخص برای یافتن مدلی به‌منظور تعیین دستهٔ داده‌هایی که دستهٔ آن‌ها مشخص نیست. تابع تصمیم ذکرشده در تعریف فوق همان مدل مورد بحث است. تشخیص دستهٔ داده‌ها با استفاده از این تابع تصمیم به‌صورت زیر انجام می‌شود:

$$x \in \begin{cases} 1 \text{ دسته } D(x) > 0 \\ 2 \text{ دسته } D(x) < 0 \end{cases} \tag{۵}$$

اگر $D(x) = 0$ باشد، در این صورت x روی ابرصفحهٔ جداکنندهٔ بهینه قرار داشته و مدل حاصل قادر به تشخیص دستهٔ x نخواهد بود. در این حالت ممکن است بتوان با تغییر داده‌های اولیهٔ ساخت مدل، ابرصفحهٔ بهینهٔ دیگری به‌دست آورد تا تعیین‌کنندهٔ وضعیت دستهٔ x باشد.

---

1. linearly separable
2. marginal hyperplane
3. hard margin
4. support vector
5. optimal separating hyperplane
6. decision function

روش کلاسیک SVM بررسی شده برای داده‌هایی مناسب است که جداپذیر خطی باشند. داده‌هایی را که جداپذیر خطی نیستند، می‌توان با قبول درجه‌ای از خطا با استفاده از یک ابرصفحه جداسازی کرد. برای مجاز کردن خطا، رابطهٔ (۳) به‌صورت زیر تغییر داده می‌شود:

$$y_i(w^T x_i + b) \geq 1 - \varepsilon_i \qquad i = 1,2,,\dots,M \qquad (۶)$$

که ε_i یک متغیر کمبود نامنفی است.

حال فرض کنید مجموعهٔ داده‌های دودسته‌ای $s = \{(x_i, y_i)\}_{i=1}^{M}$ به کمک ابرصفحهٔ جداکننده به شیوه‌ای دسته‌بندی شوند که برخی از آن‌ها در دستهٔ اشتباه قرار گیرند. در این صورت، در اینجا برخی از داده‌ها در ناحیهٔ بین دو ابرصفحهٔ حاشیه‌ای قرار خواهند گرفت و ناحیهٔ بین دو ابرصفحهٔ حاشیه‌ای را حاشیهٔ نرم[1] گویند.

تعریف. فرض کنید دسته‌بندی مجموعهٔ داده‌های دودسته‌ای $s = \{(x_i, y_i)\}_{i=1}^{M}$ با استفاده از یک ابرصفحهٔ جداکننده به شیوه‌ای انجام شود که رابطهٔ (۶) برقرار باشد. در این صورت، $\|\varepsilon\|_p^p$ به‌عنوان خطای دسته‌بندی تعریف شده است که $p \in N$ و $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_M)$.

٤-۲. SVM برای دسته‌بندی داده‌های چنددسته‌ای

اغلب مسائل در دنیای واقعی، مانند مسئلهٔ مورد بررسی در این پژوهش، چنددسته‌ای بوده و استفاده از روش‌های دودسته‌ای به خودی خود برای حل این مسائل مفید نیستند. بنا بر فرمول‌بندی ارائه‌شده توسط Vapnik (1995)، تبدیل یک مسئلهٔ k دسته‌ای به k مسئلهٔ دودسته‌ای از متداول‌ترین روش‌های دسته‌بندی داده‌های چنددسته‌ای است. مجموعهٔ داده‌های $s = \{(x_i, y_i)\}_{i=1}^{M}$ را که $x_i \in R^m (i = 1,2,\dots,M)$، $\bar{y}_i \in \{1,2,\dots,k\}$، $k > 2$ تعداد دسته‌ها و M تعداد داده‌هاست، در نظر بگیرید. در این روش هر بار یکی از دسته‌ها از سایر دسته‌ها جداسازی می‌شود. برای این کار با فرض این‌که جداسازی دستهٔ j–ام از سایر دسته‌ها مد نظر باشد، داده‌ها به‌صورت مسئله‌ای دودسته‌ای است که در آن دستهٔ اول شامل داده‌های دستهٔ j–ام (متناظر با y_i=1) و دستهٔ دوم شامل داده‌های سایر دسته‌ها (متناظر با y_i=-1) در نظر گرفته می‌شوند. به‌عبارت دیگر:

1. soft margin

$$y_i = \begin{cases} 1 & \bar{y}_i = j \\ -1 & \bar{y}_i \neq j \end{cases} \qquad i = 1, \dots, M \qquad (۷)$$

در ادامه، جداسازی به‌وسیلهٔ روش‌های SVM دودسته‌ای انجام می‌شود. این عمل برای تمام دسته‌ها انجام می‌شود.

تعریف. مجموعهٔ داده‌های k دسته‌ای $s = \{(x_i, y_i)\}_{i=1}^{M}$ را در نظر بگیرید. فرض کنید برای j=1, ... k تابع تصمیم $D_j(x) = w_j^T x + b_j$ برای جداسازی داده‌های دستهٔ j–ام از داده‌های سایر دسته‌ها به‌دست آمده باشد. در این صورت، به شرط وجود j–ای یکتا که Dj(x) > 0، آنگاه دادهٔ x متعلق به دستهٔ j–ام است. در صورتی که رابطهٔ فوق به ازای چند j برقرار باشد و یا این رابطه به ازای هیچ مقداری از j برقرار نباشد، دادهٔ x قابل دسته‌بندی نیست.

۵. دسته‌بند پیشنهادی برای تجزیهٔ متون استنادی در زبان فارسی

در این بخش جزئیات روش پیشنهادی را مورد بررسی قرار خواهیم داد. در این راستا، ابتدا، جزئیات مجموعهٔ دادهٔ ایجادشده برای آموزش و آزمایش دسته‌بند پیشنهادی توصیف خواهد شد. در ادامه، مجموعهٔ ویژگی‌های در نظر گرفته‌شده در دسته‌بند نهایی بیان شده و سپس، نحوهٔ استفاده از دسته‌بند پیشنهادی ارائه خواهد شد. در نهایت، نتایج ارزیابی روش پیشنهادی ارائه می‌شود. لازم به ذکر است که مجموعهٔ اقلام اطلاعاتی (برچسب‌های) در نظر گرفته‌شده در این پژوهش برای تجزیهٔ متون استنادی در زبان فارسی عبارت‌اند از {نام نویسنده، عنوان، سال انتشار، عنوان نشریه، شماره/ دوره، شمارهٔ صفحه، نام انتشارات، آدرس، دانشگاه/ پژوهشگاه، نام همایش، ماه/ فصل، غیره).

۵-۱. ایجاد مجموعه‌ای برچسب‌گذاری‌شده از متون استنادی در زبان فارسی

وجود یک مجموعهٔ دادهٔ برچسب‌گذاری‌شده از ابزارهای ضروری در طراحی روش‌های تجزیهٔ متون استنادی است. افزون‌بر استفاده از این مجموعهٔ داده برای بررسی کیفیت تمام روش‌ها، در روش‌های مبتنی بر یادگیری ماشین از آن برای آموزش روش نیز استفاده می‌شود. با توجه به نبودِ چنین مجموعهٔ داده‌ای برای زبان فارسی و لزوم دسترسی به چنین داده‌ای، در این پژوهش مجموعهٔ داده‌ای از متون استنادی برچسب‌گذاری‌شده برای زبان فارسی ایجاد شده است. برای انجام این کار از مقالات نشریات علمی–پژوهشی نمایه‌شده در پایگاه «سیویلیکا» و پایان‌نامه‌ها و– رساله‌های ثبت شده در «پایگاه اطلاعات علمی ایران (گنج)» (در بازهٔ زمانی ۵ سال گذشته) استفاده شده و مجموعهٔ داده‌ای شامل

۱۰۰۰ متن استنادی ایجاد شده است. متون استنادی برچسب‌گذاری شده معادل هر یک از این متون استنادی استخراج‌شده نیز به‌صورت دستی توسط خبره ایجاد و ذخیره شده است. لازم به ذکر است که متون استنادی استخراج‌شده بدون هیچ تغییری در مجموعهٔ دادهٔ ایجادشده ذخیره شده و گاهی دارای اشتباهات نوشتاری نیز هستند. افزون بر این، با توجه به وجود منابع عربی در مراجع مقالات فارسی، مجموعهٔ دادهٔ ایجادشده شامل مراجعی به زبان عربی نیز است.

مجموعهٔ دادهٔ ایجادشده شامل منابعی از نوع مقالهٔ چاپ‌شده در نشریه، مقالهٔ چاپ‌شده در همایش، پایان‌نامه یا رسالهٔ دانشگاهی و کتاب است که تعداد موجود از هر دسته در مجموعهٔ داده برابر با ۲۵۰ عدد است.

۵-۲. استخراج ویژگی‌ها

در این قسمت مجموعهٔ ویژگی‌های در نظر گرفته شده در دسته‌بندِ پیشنهادی توصیف خواهد شد. ویژگی‌های در نظر گرفته‌شده با توجه به کارهای پیشینِ انجام‌شده در زمینهٔ تجزیهٔ متون استنادی در زبان انگلیسی، بررسی مجموعهٔ متون استنادی آماده‌شده در این پژوهش و ویژگی‌های زبان فارسی به‌دست آمده است. این ویژگی‌ها را با توجه به نوع می‌توان به انواع عددی، نقطه‌گذاری، مشخص‌کننده‌های مؤلفه‌ها (ویژگی‌های شمارهٔ ۱۲ تا ۲۸ در لیست زیر)، آماری به‌دست‌آمده با خزش (ویژگی‌های شمارهٔ ۲۹ تا ۳۲ در لیست زیر) و مرتبط با واحدهای قبل و بعد (ویژگی‌های شمارهٔ ۳۳ تا ۶۵ در لیست زیر) تقسیم‌بندی نمود که ویژگی‌های به‌کاررفته در هر نوع در جدول ۱، بیان شده‌اند.

جدول ۱. ویژگی‌های به‌کاررفته در دسته‌بند ارائه‌شده

نوع	ویژگی
ویژگی‌های عددی	۱) موقعیت واحد: واحد مورد نظر چندمین واحد در لیست واحدهای متن استنادی است. ۲) بررسی وجود تنها کاراکترهای عددی در واحد ۳) آیا واحد در مجموعهٔ اعداد ترتیبی نوشته‌شده به حروف، یعنی اول، اولین، یکمین، دوم، دومین و ... قرار دارد؟ بدین منظور از اعداد ترتیبی کمتر از صد استفاده شده است. ۴) تعداد واحدهای عددی قبل از واحد فعلی در لیست واحدهای متن استنادی ۵) آیا واحد نمایانگر سال است؟ عددی به‌عنوان سال در نظر گرفته می‌شود که بین ۱۰۰۰ تا ۲۰۲۰ باشد.

نوع	ویژگی
ویژگی‌های مربوط به علایم نقطه‌گذاری	۶) وجود کاراکتری غیر از کاراکترهای متناظر با حروف فارسی در واحد ۷) وجود همزمان عدد و حرف فارسی در واحد ۸) وجود تنها عدد و حرف فارسی در واحد ۹) وجود حداقل یک حرف انگلیسی در واحد ۱۰) وجود آخرین کاراکتر غیرفاصله قبل از واحد فعلی در متن استنادی در علایم نقطه‌گذاری {"،"، "."، "؛"، "(")، ")"، "«"، "»"، "؟"، "?"، ":"} ۱۱) وجود اولین کاراکتر غیرفاصله بعد از واحد فعلی در متن استنادی در علائم نقطه‌گذاری {"،"، "."، "؛"، "(")، ")"، "«"، "»"، "؟"، "?"، ":"}
ویژگی‌های متناظر با مشخص‌کننده‌های برچسب‌های مختلف	۱۲) وجود واحد در مجموعهٔ مشخص‌کننده‌های شمارهٔ صفحه: {"ص"، "صص"، "صفحات"، "صفحه"، "صفحه‌های"} ۱۳) بودن واحد در قالب شمارهٔ صفحه: در صورتی که واحد شامل "-" و تعدادی عدد باشد، به‌عنوان واحدی در قالب شمارهٔ صفحه در نظر گرفته می‌شود. ۱۴) وجود واحد در مجموعهٔ مشخص‌کننده‌های دوره و شماره: {"دوره"، "دوره‌ی"، "سال"، "س"، "د"، "شماره"، "شماره‌ی"، "ش"} ۱۵) وجود واحد در مجموعهٔ مشخص‌کننده‌های شمارهٔ چاپ: {"چاپ"، "چ"، "چ."} ۱۶) وجود واحد در مجموعهٔ مشخص‌کننده‌های شمارهٔ جلد: {"جلد"، "ج"، "ج."} ۱۷) نمایانگر مخفف نام یک نویسنده بودن واحد، یعنی آیا واحد به‌صورت "ا."، "آ."، "ب."، "پ."، ... است؟ ۱۸) مختوم بودن واحد به یکی از پسوندهای نام خانوادگی رایج در زبان فارسی. مجموعهٔ پسوندهای در نظر گرفته‌شده عبارت‌اند از {"پور"، "فرد"، "نیا"، "زاده"، "نژاد"، "راد"، "فر"، "ان"، "کیا"، "پناه"، "منش"}.[1] ۱۹) وجود واحد در مجموعهٔ نام ماه‌ها و فصل‌های سال ۲۰) وجود واحد در مجموعهٔ {"دیگران" و "همکاران"} ۲۱) وجود واحد در مجموعهٔ کلمات رایج در نام دانشگاه‌ها/پژوهشگاه‌ها: {"دانشگاه"، "پژوهشگاه"، "دانشکده"، "پژوهشکده"، "پیام"، "نور"، "آزاد"، "گروه"، "استان"، "ایران"، "شهید"} ۲۲) وجود واحد در مجموعهٔ {"ترجمه"، "ترجمه‌ی"، "مترجم"، "به کوشش"، "بامقدمه"، "با مقدمه‌ی"، "تصحیح"} ۲۳) وجود واحد در مجموعهٔ مشخص‌کننده‌های پایان‌نامه و رساله: {"استاد"، "راهنما"، "پایان‌نامه"، "رساله"، "کارشناسی"، "کارشناسی‌ارشد"، "دکتری"، "دکترا"، "رشته"، "رساله‌ی"، "راهنمایی"} ۲۴) وجود واحد در مجموعهٔ مشخص‌کننده‌های نام انتشارات: {"انتشارات"، "نشر"، "ناشر"، "الانتشار"، "چاپ"}

۱. لازم به ذکر است که فهرست ارائه‌شده با توجه به مجموعهٔ دادهٔ ایجادشده در این پژوهش تهیه شده است.

نوع	ویژگی
	۲۵) وجود واحد در مجموعهٔ اسامی شهرهای مهم کشور
	۲۶) وجود واحد در مجموعظ مشخص‌کننده‌های نام نشریه: {"فصلنامه"، "پژوهشنامه"، "مجله"، "دوفصلنامه"، "نشریه"، "نامه"، "دوماهنامه"، "ماهنامه"، "پژوهش"، "تحقیقات"، "مطالعات"}
	۲۷) وجود واحد در مجموعه مشخص‌کننده‌های نام همایش: {"همایش"، "کنفرانس"، "ایران"، "ملی"، "بین‌المللی"، "کنگره"، "کنگره‌ی"، "سمینار"، "انجمن"، "خلاصه"، "مقالات"}
	۲۸) وجود واحد در مجموعهٔ {"و"، "الی"، "تا"}
ویژگی‌های آماری به‌دست‌آمده با خزش روی پایگاه‌های اطلاعاتی	۲۹) وجود واحد در مجموعهٔ اسامی نویسندگان (استخراج شده با خزش از پایگاه سیویلیکا)
	۳۰) امتیاز اختصاص داده‌شده به واحد (بین یک تا ۱۰) با توجه به فراوانی حضور واحد در عنوان نشریات (به‌دست آمده با خزش در پایگاه‌های علمی)
	۳۱) امتیاز اختصاص داده‌شده به واحد (بین یک تا ۱۰) با توجه به فراوانی حضور واحد در عنوان مقالات چاپ‌شده تا به امروز (به‌دست آمده با خزش در پایگاه‌های علمی)
ویژگی‌های متناظر با واحدهای قبل و بعد از واحد فعلی	۳۲) برچسب اختصاص‌یافته به سومین واحد قبل از واحد فعلی در دنباله واحدهای متن استنادی
	۳۳) برچسب اختصاص یافته به دومین واحد قبل از واحد فعلی در دنبالهٔ واحدهای متن استنادی
	۳۴) برچسب اختصاص‌یافته به واحد قبل از واحد فعلی در دنبالهٔ واحدهای متن استنادی
	۳۵-۶۴) ویژگی‌های شماره ۱ تا ۳۰ برای واحد بعدی در در دنبالهٔ واحدهای متن استنادی

لازم به ذکر است که از بین ویژگی‌های ذکرشده در بالا، ویژگی‌های مشخص‌شده با خط کشیده در زیر آن‌ها برای اولین‌بار در این پژوهش معرفی شده و مورد استفاده قرار گرفته‌اند که به‌جز ویژگی دوم در این لیست، سایر ویژگی‌ها را می‌توان برای تجزیهٔ متون استنادی در هر زبانی مورد استفاده قرار داد.

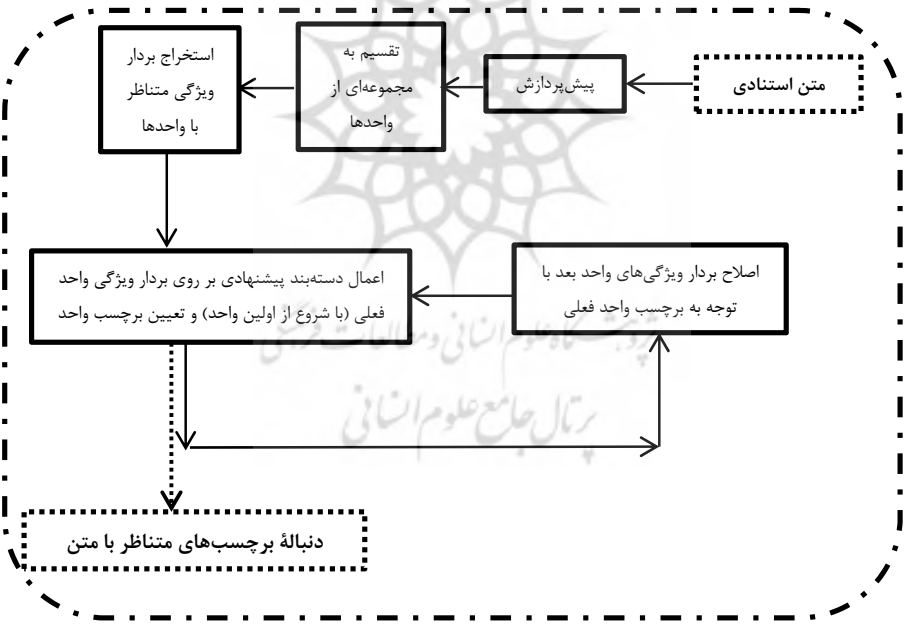
۵-۳. جزئیات روش پیشنهادی

در روش ارائه‌شده ابتدا، مجموعه‌ای از ویرایش‌ها روی متون استنادی منبع انجام می‌شود. مهم‌ترین ویرایش‌های انجام‌شده عبارت‌اند از:

- اصلاح نویسه‌های فارسی: با توجه به وجود چندین نویسه با ظاهر یکسان برای برخی از حروف فارسی، یکی از پیش‌پردازش‌های لازم برای متن‌کاوی در زبان فارسی، اصلاح و یکسان‌سازی نویسه‌هاست این مهم در روش ارائه‌شده صورت گرفته است.

- تبدیل برخی فاصله‌ها به نیم‌فاصله: از دیگر عوامل تأثیرگذار برای متن‌کاوی در زبان فارسی، فاصله‌ها و نیم فاصله‌ها هستند و استفادهٔ نادرست از آن‌ها منجر به کاهش کیفیت الگوریتم‌های متن‌کاوی خواهد شد. با توجه به این مهم، در روش پیشنهادی افزون‌بر اصلاح نویسه‌ها، فاصله و نیم‌فاصله‌ها را در کلمات تأثیرگذار نیز اصلاح می‌کنیم.

پس از ویرایش متون استنادی، هر متن استنادی به دنباله‌ای از واحدها تقسیم‌بندی می‌شود که در این پژوهش از هر لغت به‌عنوان یک واحد استفاده شده است. سپس، بُردار ویژگی‌های متناظر با هر واحد محاسبه می‌شود. در ادامه، پس از استخراج بُردار ویژگی متناظر با هر واحد هر متن استنادی موجود در مجموعهٔ دادهٔ آموزش، دسته‌بند SVM با استفاده از این داده‌ها آموزش داده می‌شود. در ادامه، برای تجزیهٔ یک متن استنادی جدید، همان‌طور که در شکل ۱، نمایش داده شده، به‌صورت زیر عمل می‌شود:



شکل ۱. شمای کلی روش پیشنهادی

برای هر واحد موجود در متن استنادی (به‌ترتیب از ابتدا تا انتها):

- همهٔ مؤلفه‌های بُردار ویژگی متناظر با آن به جز مؤلفه‌های ۳۲، ۳۳ و ۳۴ محاسبه می‌شود؛

- در صورتی که مقدار مؤلفهٔ اول بُردار ویژگی واحد مورد بررسی برابر با ۱ باشد، مقادیر مؤلفه‌های ۳۲، ۳۳ و ۳۴ آن برابر با null قرار داده می‌شود؛
- در غیر این صورت، مقدار مؤلفهٔ ۳۲ واحد فعلی برابر با مقدار مؤلفهٔ ۳۳ بُردار ویژگی واحد ما قبل، مقدار مؤلفهٔ ۳۳ واحد فعلی برابر با مقدار مؤلفهٔ ۳۴ بُردار ویژگی واحد ما قبل و مقدار مؤلفهٔ ۳۴ واحد برابر با برچسب اختصاص‌یافته به واحد ما قبل قرار می‌گیرد؛
- با استفاده از دسته‌بند آموزش‌دیده، برچسب واحد فعلی به‌دست آمده و افزون‌بر ذخیره، در بُردارهای ویژگی واحدهای بعد مورد محاسبه قرار می‌گیرد.

٦. ارزیابی روش پیشنهادی

برای ارزیابی روش پیشنهادی زبان برنامه‌نویسی «پایتون» پیاده‌سازی شده است. در پیاده‌سازی انجام‌شده، کتابخانهٔ sklearn برای استفاده از دسته‌بند SVM چنددسته‌ای خطی و با پارامترهای پیش‌فرض به‌کار رفته است. برای ارزیابی روش ارائه‌شده از مجموعهٔ دادهٔ طراحی‌شده در ابتدای این بخش و روش اعتبارسنجی متقابل ۱۰ سطحی استفاده شده است. در این روش، داده‌ها به‌طور تصادفی به ۱۰ بخش تقسیم‌بندی شده و هر بخش یک بار به‌عنوان دادهٔ آزمایش به دسته‌بند آموزش‌دیده با استفاده از داده‌های ۹ بخش دیگر داده شده و نتایج ارزیابی محاسبه می‌شود.

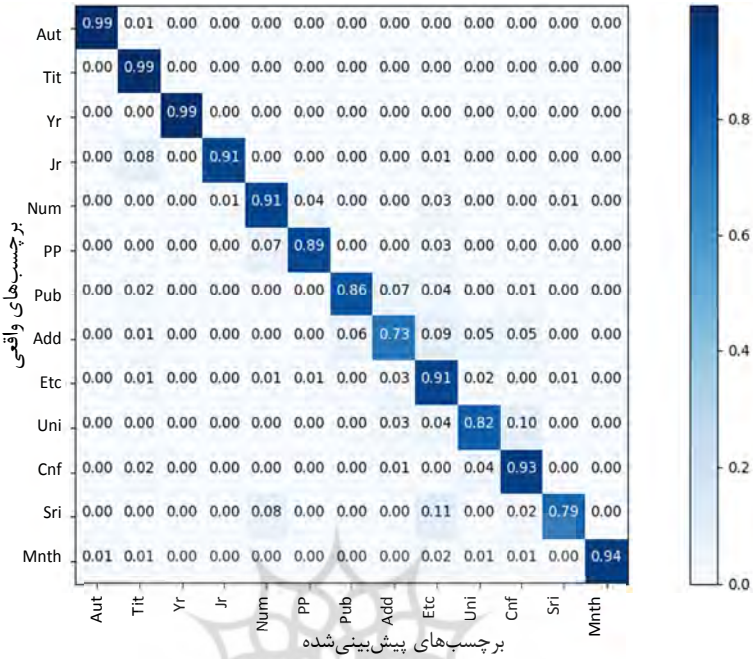
مجموعهٔ دادهٔ ایجادشده در این پژوهش شامل ۱۰۰۰ متن استنادی بوده که به هر واحد در هر رشته از این مجموعهٔ داده یک برچسب از مجموعهٔ برچسب‌های {نام نویسنده، عنوان، سال انتشار، نام نشریه، دوره/ شماره، شمارهٔ صفحه، نام انتشارات، آدرس، غیره، دانشگاه/ پژوهشگاه، همایش، ماه/ فصل} اختصاص یافته است.

نتایج پیاده‌سازی دسته‌بند پیشنهادی روی مجموعهٔ دادهٔ مورد نظر نشان می‌دهد که میانگین دقت، فراخوانی و اف-۱ در روش ارائه‌شده برابر با ۹۵ درصد است. برای بررسی بهتر، نتایج روش پیشنهادی در جدول ۲، و شکل ۲، نیز به‌صورت جزئی نمایش داده شده‌اند. با توجه به تعداد برچسب‌ها و برای میسرسازی نمایش آن‌ها، در شکل‌های زیر از Aut، Tit، Yr، Jr، Num، PP، Pub، Add، Etc، Uni، Cnf، Sri و Mnth به‌ترتیب برای نمایش برچسب‌های نام نویسنده، عنوان، سال انتشار، نام نشریه، شماره/ دوره، شمارهٔ صفحه، نام انتشارات، آدرس، غیره، دانشگاه/ پژوهشگاه، نام همایش و ماه/ فصل استفاده شده است.

جدول ۲. نتایج ارزیابی دسته‌بند پیشنهادی

برچسب	دقت	فراخوانی	F1	تعداد
نام نویسنده	۰/۹۹	۰/۹۹	۰/۹۹	۴۱۷۳
عنوان	۰/۹۸	۰/۹۹	۰/۹۹	۱۰۲۴۴
سال	۰/۹۹	۰/۹۹	۰/۹۹	۱۰۴۶
نشریه	۰/۹۷	۰/۹۱	۰/۹۴	۹۷۳
دوره/شماره	۰/۸۷	۰/۹۱	۰/۸۹	۴۲۳
شماره صفحه	۰/۸۹	۰/۸۸	۰/۸۹	۲۷۴
انتشارات	۰/۹۳	۰/۸۶	۰/۸۹	۸۶۶
آدرس	۰/۷۷	۰/۷۴	۰/۷۵	۷۹۳
غیره	۰/۹۲	۰/۹۱	۰/۹۱	۲۲۵۶
مؤسسه	۰/۸۰	۰/۸۱	۰/۸۱	۸۰۰
همایش	۰/۹۲	۰/۹۳	۰/۹۲	۱۹۰۱
شماره چاپ	۰/۷۴	۰/۷۵	۰/۷۵	۸۴
ماه/ فصل	۰/۹۸	۰/۹۶	۰/۹۷	۱۶۴
مجموع	۰/۹۵	۰/۹۵	۰/۹۵	۲۳۹۹۷

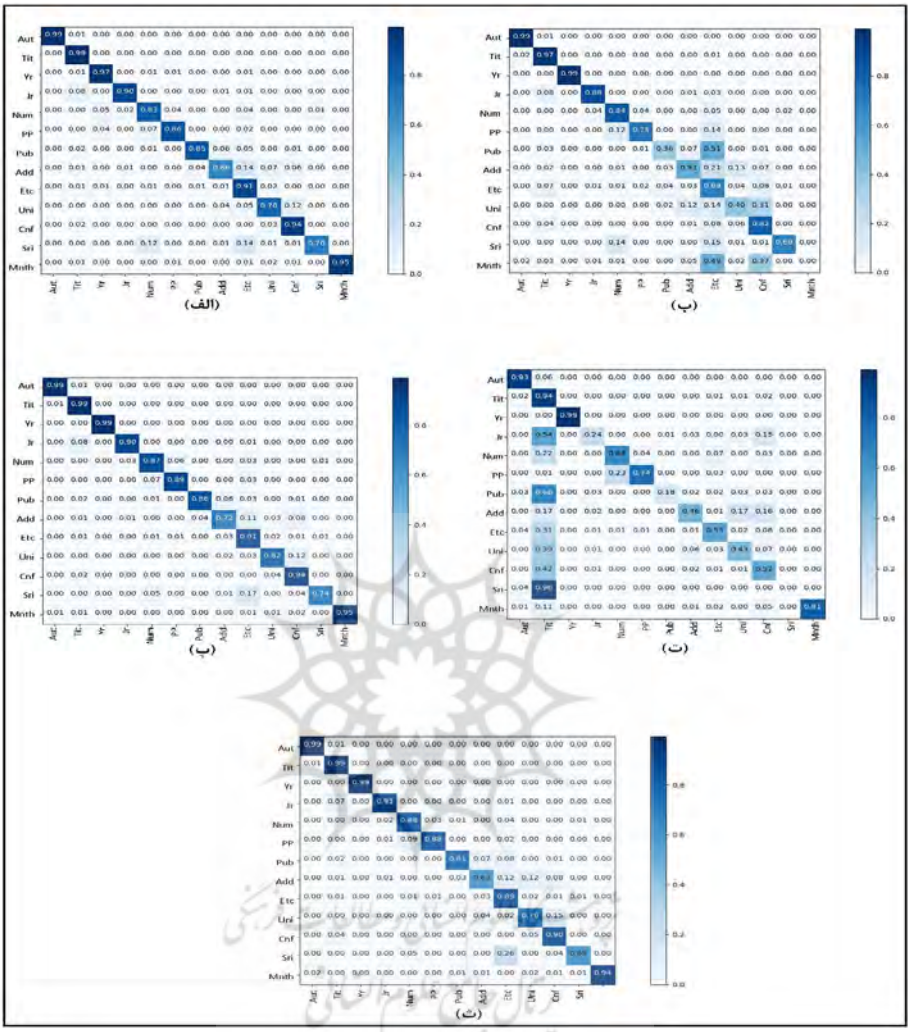
همان‌طور که در شکل ۲، مشخص است، عمدهٔ خطاهای موجود در نتایج روش پیشنهادی به‌صورت اختصاص برچسب قبل یا بعد، با توجه به‌ترتیب معمول برچسب‌ها در یک متن استنادی است. در پاراگراف‌های بعدی دلیل این مسئله را مشاهده خواهیم کرد.



شکل ۲. ماتریس درهم‌ریختگی روش پیشنهادی

برای بررسی میزان اثربخشی ویژگی‌های مختلف به‌کاررفته در روش پیشنهادی، در ادامه، با توجه به گروه‌بندی ارائه‌شده برای ویژگی‌ها، تأثیر آن‌ها را بر کیفیت روش پیشنهادی بررسی می‌کنیم. برای دستیابی به این هدف، به‌ترتیب، هر یک از گروه‌های موجود در گروه‌بندی بیان‌شده برای ویژگی‌ها را از لیست ویژگی‌ها حذف کرده و روش پیشنهادی را با توجه به این تغییر ارزیابی می‌کنیم. نتایج ارزیابی روش پیشنهادی با این تغییرات در شکل ۳، ارائه شده است. شکل ۳ (الف) نشان می‌دهد که حذف ویژگی‌های عددی منجر به کاهش کیفیت روش پیشنهادی در تشخیص برچسب واحدهای عددی یعنی سال انتشار، شماره/ دوره، شمارهٔ صفحه و شمارهٔ چاپ می‌شود. شکل ۳ (ب) نشان‌دهندهٔ این است که حذف ویژگی‌های مربوط به نقطه‌گذاری به‌طور محسوس منجر به کاهش کیفیت روش پیشنهادی در تشخیص برچسب واحدهایی که به‌طور معمول در انتهای یک متن استنادی بیان می‌شوند، شده است. با توجه به تغییرات زیاد در دنبالهٔ برچسب‌ها در انتهای متون استنادی و استفادهٔ بیشتر از علائم نقطه‌گذاری در این قسمت، این مسئله توجیه‌پذیر است. شکل ۳ (پ) نشان می‌دهد که مشخص‌کننده‌های مؤلفه‌ها دارای

نقشی اساسی در تعیین برچسب واحدهای یک متن استنادی هستند. همان‌طور که نتایج ارائه‌شده در این شکل نشان می‌دهد، حذف این ویژگی‌ها به کاهش چشمگیر کیفیت روش پیشنهادی در تشخیص برچسب واحدهایی که به‌طور معمول از این مشخص‌کننده‌ها در آن‌ها استفاده شده (مانند عنوان نشریه، ماه/ فصل، دانشگاه/ پژوهشگاه و ...)، منجر می‌شود. نتایج شکل ۳ (ت) بیانگر این است که تأثیر ویژگی‌های آماری به‌دست‌آمده با خزش بر روی پایگاه‌های اطلاعاتی بر روی نتایج به‌کارگیری روش پیشنهادی نامحسوس بوده و این ویژگی‌ها تنها اندکی باعث بهبود نتایج شده‌اند. شکل ۳ (ث)، نتایج روش پیشنهادی پس از حذف ویژگی‌های متناظر با واحدهای همسایه از لیست واحدهای یک واحد را نشان می‌دهد. همان‌طور که مشخص است حذف این ویژگی‌ها به‌طوری بسیار چشمگیر منجر به کاهش کیفیت روش پیشنهادی شده است. این نتیجه با توجه به این‌که مسئلهٔ تجزیهٔ متون استنادی یک مسئلهٔ برچسب‌گذاری توالی است، قابل توجیه است. این مسئله همچنین، می‌تواند توجیه‌کنندهٔ علت برچسب‌های اشتباه تشخیص داده‌شده توسط الگوریتم اصلی (در شکل ۲) نیز باشد؛ چرا که اهمیت برچسب واحدهای همسایه را نشان می‌دهد.



شکل ۳. ماتریس درهم‌ریختگی روش پیشنهادی پس از حذف الف) ویژگی‌های عددی، ب) ویژگی‌های مربوط به نقطه‌گذاری، پ) ویژگی‌های مربوط به مشخص‌کننده‌های برچسب‌های مؤلفه‌های مختلف، ت) ویژگی‌های آماری به‌دست‌آمده با خزش روی پایگاه‌های اطلاعاتی، و ث) ویژگی‌های متناظر با واحدهای همسایه

۷. نتیجه‌گیری و کارهای آینده

در این مقاله به مسئلهٔ تجزیهٔ متون استنادی پرداخته شد و با به‌کارگیری دسته‌بند SVM، روشی برای این مسئله در زبان فارسی ارائه گردید. در روش ارائه‌شده یک متن

استنادی ورودی به مجموعه‌ای از واحدها (کلمات) تقسیم‌بندی شده و در ادامه، از دسته‌بند SVM برای تعیین برچسب هر واحد، یا به بیانی کلی‌تر، برای تعیین مؤلفه‌های مختلف متن استنادی ورودی استفاده می‌شود. برای بررسی کیفیت روش ارائه‌شده، مجموعهٔ داده‌ای از متون استنادی برچسب‌گذاری‌شده در زبان فارسی تهیه شده و با استفاده از روش اعتبارسنجی متقابل ۱۰ سطحی، مجموعه آزمایشاتی بر روی روش پیشنهادی صورت گرفت. نتایج بررسی آزمایشات بر روی دسته‌بند پیشنهادی نشان‌دهندهٔ ۰/۹۵ برای هر سه پارامتر دقت، فراخوانی و اف-۱ است.

در طراحی دسته‌بند پیشنهادی از مجموعه‌ای از ویژگی‌ها استفاده شد که با توجه به ماهیت، می‌توان آن‌ها را به ویژگی‌های عددی، نقطه‌گذاری، مشخص‌کننده‌های مؤلفه‌ها، آماری و ویژگی‌های متناظر با واحدهای همسایه تقسیم‌بندی نمود. پس از بررسی کلی کیفیت روش پیشنهادی، مجموعه آزمایش‌هایی برای بررسی تأثیر انواع مختلف ویژگی‌های به‌کاررفته در دسته‌بند پیشنهادی صورت گرفت. نتایج به‌دست‌آمده نشان می‌دهد که ویژگی‌های عددی به‌کاررفته تأثیری محسوس در تشخیص برچسب واحدهای عددی (مانند سال انتشار، شماره/ دوره و ...) داشته؛ ویژگی‌های نقطه‌گذاری به‌طور محسوس در تشخیص برچسب واحدهایی که به‌طور معمول در انتهای یک متن استنادی بیان می‌شوند، مؤثر واقع شده، مشخص‌کننده‌های مؤلفه‌ها نقشی بسیار اساسی در تعیین برچسب واحدهای متن استنادی ورودی داشته و حذف این ویژگی‌ها منجر به کاهش چشمگیر کیفیت روش پیشنهادی شده، تأثیر ویژگی‌های آماری بر روی نتایج به‌کارگیری روش پیشنهادی نامحسوس بوده، و از آنجا که مسئلهٔ تجزیهٔ متون استنادی یک مسئلهٔ برچسب‌گذاری توالی است، ویژگی‌های متناظر با واحدهای همسایه نیز به‌طوری بسیار چشمگیر در کیفیت نتایج روش پیشنهادی مؤثر بوده‌اند.

از جمله کارهای قابل انجام برای ارتقای این پژوهش می‌توان به استفاده از دسته‌بندهای دیگر مانند ماشین بُردار پشتیبان دوقلو[1]، CRF و یادگیری عمیق برای حل این مسئله و مقایسهٔ نتایج با نتایج به‌دست‌آمده در این پژوهش اشاره کرد. افزون بر این، همان‌طور که ذکر شد، مسئلهٔ تجزیهٔ متون استنادی مسئله‌ای پایه‌ای برای بسیاری از مسائل دیگر است که از آن جمله می‌توان به تطبیق متون استنادی در یک سطح بالاتر و ساخت

1. twin SVM

خودکار شبکه‌های استنادی در حالت کلی اشاره کرد. با توجه به کیفیت مناسب روش ارائه‌شده می‌توان از آن به‌عنوان ابزاری پایه‌ای برای حل این مسائل استفاده کرد که در پژوهش‌های آتی به آن پرداخته خواهد شد.

فهرست منابع

نصیری، جلال‌الدین. ۱۳۹۴. بازشناسی اعمال انسان با رویکرد مقاوم‌سازی دسته‌بند تفکیکی. رسالهٔ دکتری دانشگاه تربیت مدرس.

کارگر، مرتضی. ۱۳۹۰. دسته‌بندی داده‌ها با استفاده از روش SVM. پایان‌نامهٔ کارشناسی دانشگاه شهید باهنر کرمان.

References

Ahmed, M. W., and M. T. Afzal. 2020. FLAG-PDFe: Features oriented metadata extraction framework for scientific publications. *IEEE Access* 8; 99458-99469.

An, D., L. Gao, Z. Jiang, R. Liu, and Z. Tang2017 .. Citation metadata extraction via deep neural network-based segment sequence labeling. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management; 1967-1970. Singapore, Singapore.

Besagni, D., A. Belaïd, & N. Benet. 2003. A segmentation method for bibliographic references by contextual tagging of fields. In Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on; 384-388. Edinburgh, Scotland.

Bhardwaj A., D. Mercier, A. Dengel, and S. Ahmed. 2017. DeepBIBX: Deep Learning for Image Based Bibliographic Data Extraction. In International Conference on Neural Information Processing; Cham, Switzerland. 286-293.

Councill, I. G., C. L. Giles, and M. Y. Kan. 2008. ParsCit: an Open-source CRF Reference String Parsing Package. In proceedings of the 6th edition of the Language Resources and Evaluation Conference (LREC). Morocco. 661-667.

Ding, Y., G. Chowdhury, and S. Foo. 1999. Template mining for the extraction of citation from digital documents. In Proceedings of the Second Asian Digital Library Conference. Taiwan. 47-62.

Gupta, D., B. Morris, T. Catapano, and G. Sautter. 2009. A new approach towards bibliographic reference identification, parsing and inline citation matching. In International Conference on Contemporary Computing; 93-102. Noida, India.

Hashmi, A. M., M. T. Afzal, and S. ur Rehman. 2020. Rule Based Approach to Extract Metadata from Scientific PDF Documents. In 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA); 1-4. Sydney, Australia.

Hetzner, E. 2008. A simple method for citation metadata extraction using hidden markov models. In Proceedings of the 8th ACM/ IEEE-CS joint conference on Digital libraries; Pittsburgh, Pennsylvania, USA. 280-284.

Huang, I. A., J. M. Ho, H. Y. Kao, and W. C. Lin. 2004. Extracting citation metadata from online publication lists using BLAST. In Pacific-Asia Conference on Knowledge Discovery and Data Mining; Sydney, Australia. 539-548.

Kim, Y. M., P. Bellot, J. Tavernier, E. Faath, and M. Dacos. 2012. Evaluation of BILBO reference parsing in digital humanities via a comparison of different tools». In Proceedings of the 2012 ACM symposium on Document engineering; Paris, France. 209-212.

Lawrence, S., C. Lee Giles, and K. Bollacker. 1999a. Digital libraries and autonomous citation indexing. *Computer* 32 (6): 67-71.

Lawrence, S., C. L. Giles, and K. D. Bollacker. 1999b. Autonomous citation matching». In Proceedings of the third annual conference on Autonomous Agents; 392-393. Seattle, Washington, USA.

Lopez, P. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In International Conference on Theory and Practice of Digital Libraries; 473-474. Glasgow, United Kingdom.

Namikoshi, D., M. Ohta, A. Takasu, and J. Adachi. 2017. CRF-based bibliography extraction from reference strings using a small amount of training data. Twelfth International Conference on Digital Information Management (ICDIM); 59-64.

Nasar, Z., S.W. Jaffry, and M.K. Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics* 117: 1931–1990.

Ojokoh, B., M. Zhang, and J. Tang. 2011. A trigram hidden Markov model for metadata extraction from heterogeneous references. *Information Sciences* 181 (9): 1538-1551.

Peng, F., and A. McCallum. 2013. Accurate information extraction from research papers using conditional random fields. https://aclanthology.org/N04-1042.pdf. (accessed April 13, 2013).

Prasad, A., M. Kaur, and M.Y. Kan. 2018. Neural ParsCit: a deep learning-based reference string parser. *International Journal of Digital Library* 19: 323–337.

Rizvi, S. T. R., A. Dengel, and S. Ahmed. 2020. A Hybrid Approach and Unified Framework for Bibliographic Reference Extraction. *IEEE Access*, 8; 217231-217245.

Tkaczyk, D. 2017. New Methods for Metadata Extraction from Scientific Literature. arXiv preprint arXiv:1710.10201.: http://arxiv.org/abs/1710.10201

_____, A. Collins, P. Sheridan, and J. Beel. 2018a. Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers. In Proceedings of the 18th ACM/IEEE on joint conference on digital libraries; 99-108. Fort Worth, Texas, USA.

Tkaczyk, D., R. Gupta, R. Cinti, and J. Beel. 2018b. Parsrec: A novel meta-learning approach to recommending bibliographic reference parsers. Dublin, Ireland. arXiv preprint arXiv: 1811. 10369.

Tkaczyk, D., P. Sheridan and J. Beel. 2018c. ParsRec: Meta-Learning Recommendations for Bibliographic Reference Parsing. In Proceedings of the Late-Breaking Results track part of the Twelfth ACM Conference on Recommender Systems (RecSys '18), Vancouver, BC, Canada, 2018.

Tkaczyk, D., P. Szostek, P. J. Dendek, M. Fedoryszak, and L. Bolikowski. 2014. Cermine--automatic extraction of metadata and references from scientific literature. In Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on; 217-221. Tours, France.

Tkaczyk, D., P. Szostek, M. Fedoryszak, P. J. Dendek, and L. Bolikowsk. 2015. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition* 18 (4): 317-335.

Vapnik, V. N. 1998. *Statistical Learning Theory*. NewYork: John Wiley & Sons.

_____. 1995. The Nature of Statistical Learning Theory. NewYork: Springer-Verlag.

Yin, P., M. Zhang, Z. Deng, and D. Yang. 2004. Metadata extraction from bibliographies using bigram HMM. In International Conference on Asian Digital Libraries; 310-319. Florida, USA.

Zhang, Q., Y. G. Cao, and H. Yu (2011). «Parsing citations in biomedical articles using conditional random fields». Computers in biology and medicine, 41 (4); 190-194.

Zhang, X., J. Zou, D. X. Le, and G. R. Thoma. 2011. A structural SVM approach for reference parsing. In 2010 Ninth International Conference on Machine Learning and Applications (pp. 479-484). IEEE. Washington DC, USA.

Zou, J., D. Le, and G. R. Thoma. 2010. Locating and parsing bibliographic references in HTML medical articles. *International Journal on Document Analysis and Recognition* 13 (2): 107-119.

نصراله پاک‌نیت

متولد سال ۱۳۶۵، دارای مدرک تحصیلی دکتری در رشتهٔ ریاضی از دانشگاه شهید بهشتی تهران است. ایشان هم‌اکنون استادیار پژوهشکدهٔ علوم اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

رمزنگاری، الگوریتم‌ها و متن‌کاوی از جمله علایق پژوهشی وی است.

جلال‌الدین نصیری

متولد سال ۱۳۶۲، دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی کامپیوتر گرایش نرم‌افزار از دانشگاه تربیت مدرس است. ایشان هم‌اکنون استادیار دانشگاه فردوسی مشهد است.

پردازش زبان‌های طبیعی و یادگیری ماشین از جمله علایق پژوهشی وی است.