Automatic Detection of the Boundary between Metadata and Body in Persian **Theses using BA_SVM**

Mohadese Rahnama
M.Sc. Student in Computer Engineering; Alzahra University; Tehran, Iran Email: m.rahnama@student.alzahra.ac.ir

Seyed Mohammad Hossein Hasheminejad*
PhD in Computer Engineering; Assistant Professor; Alzahra University; Tehran, Iran Email: smh.hasheminejad@alzahra.ac.ir

Jalal A Nasiri
PhD in Computer Engineering; Assistant Professor; Iranian Research Institute for Information Scienceand Technology (IranDoc); Tehran, Iran Email: j.nasiri@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 09, May 2020 **Accepted: 02, Nov. 2020**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 36 | No. 4 | pp. 1159-1180
Summer 2021

Abstract: Metadata extraction facilitates the process of indexing and improves information retrieval. Also automation of this process increases efficiency more than manual extraction. The example of the thesis metadata are names of students, professors, title, field, degree, abstract, keywords, etc. In this paper the aim is automatic boundary detection of metadata from the main body in Persian theses. Therefore, 250 theses collected from IRANDOC system. Features were extracted from paragraphs of each thesis then paragraphs were classified using support vector machine into 2 classes: metadata and body. In this study, Bat algorithm is used to set the parameter of SVM. The result reveals that the proposed method predicts type of paragraphs with 96.6 percent accuracy.

Keywords: Metadata Extraction, Information Extraction, Support Vector Machine (SVM), Metaheuristic Algorithm, Bat Algorithm (BA)

* Corresponding Author

استخراج هوشمند مرز فراداده و متن در پایان‌نامه‌های فارسی با رویکرد BA_SVM

محدثه رهنما
دانشجوی کارشناسی ارشد مهندسی کامپیوتر؛ دانشگاه الزهرا (س)؛ تهران، ایران؛
m.rahnama@student.alzahra.ac.ir

سید محمدحسین هاشمی‌نژاد
دکتری مهندسی کامپیوتر؛ استادیار؛ دانشگاه الزهرا (س)؛ تهران، ایران؛
پدیدآور رابط smh.hasheminejad@alzahra.ac.ir

جلال‌الدین نصیری
دکتری مهندسی کامپیوتر؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
j.nasiri@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۹/۰۲/۲۰ پذیرش: ۱۳۹۹/۰۸/۱۲ مقاله برای اصلاح به مدت ۳۵ 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 ۳۶ | شمارهٔ ٤ | ص‌ص ۱۱۵۹-۱۱۸۰
تابستان ۱٤۰۰

چکیده: استخراج فراداده باعث تسهیل در فرایند نمایه‌سازی و بهبود در بازیابی اطلاعات است. از سوی دیگر، خودکارسازی این فرایند سبب افزایش کارایی نسبت به استخراج دستی فراداده‌هاست. نام دانشجو، نام اساتید، عنوان، رشته و مقطع تحصیلی، چکیده، و کلمات کلیدی نمونه‌ای از فراداده‌های پایان‌نامه است. هدف در این مقاله شناسایی خودکار مرز فراداده و بدنهٔ اصلی در پایان‌نامه‌های فارسی است. بدین منظور، ۲۵۰ پایان‌نامهٔ ثبت‌شده در سامانهٔ «ایرانداک» جمع‌آوری شده است. ویژگی‌های مد نظر از هر پاراگراف استخراج شده و سپس، پاراگراف‌های پایان‌نامه با روش ماشین بردار پشتیبان به دو کلاس فراداده و بدنه طبقه‌بندی شد. در این پژوهش برای تنظیم پارامترهای الگوریتم ماشین بردار پشتیبان، الگوریتم فرامکاشفه‌ای خفاش به‌کار گرفته شده است. نتایج نشان می‌دهد که روش پیشنهادی با دقت ۹۶/۶ درصد نوع پاراگراف را تشخیص می‌دهد.

کلیدواژه‌ها: استخراج فراداده، استخراج اطلاعات، ماشین بردار پشتیبان، الگوریتم فرامکاشفه‌ای، الگوریتم خفاش

۱. مقدمه

با گسترش نوشتارهای علمی، دسترسی سریع و ساده به فراداده[1]های مقالات و پایان‌نامه‌ها بیش از پیش ضرورت می‌یابد. فرایند استخراج اطلاعات از متون علمی به‌صورت دستی و توسط انسان، بسیار زمان‌بر و پرهزینه خواهد بود. از این رو، خودکارسازی این فرایند می‌تواند در تسریع شاخص‌گذاری نوشتارهای علمی مؤثر باشد. در زمینهٔ استخراج فراداده از مقالات علمی و پایان‌نامه‌ها پژوهش‌هایی به‌ویژه در زبان انگلیسی صورت گرفته، اما در زبان فارسی تحقیقات کمتری انجام شده است.

فراداده‌ها از اطلاعات بسیار باارزش پایان‌نامه‌ها و رساله‌ها هستند. اولین و مهم‌ترین گام در استخراج خودکار فراداده‌ها تشخیص مرز شروع بدنهٔ پایان‌نامه است. ازاین رو، تشخیص خودکار مرز بین فراداده و متن اصلی پایان‌نامه در سامانه‌های ذخیره‌سازی موضوعی و بازیابی اطلاعات نقش مهمی ایفا می‌کند. در بسیاری از سامانه‌های ثبت و اشاعهٔ متون علمی ورود اطلاعات فراداده توسط عامل انسانی انجام می‌پذیرد.

هدف این مقاله تشخیص خودکار مرز فراداده‌ها و متن اصلی پایان‌نامه‌های فارسی است. بدین‌ترتیب، فرایند استخراج فراداده‌ها ساده‌تر صورت گرفته و نیازی به پردازش تمامی متن پایان‌نامه نیست. به‌عبارت دیگر، استخراج هوشمند پاراگرافی که ابتدای متن اصلی پایان‌نامه یا رساله باشد، هدف این پژوهش است. در شکل ۱، نمونه‌هایی از مرز فراداده و بدنه در پایان‌نامه‌های فارسی نشان داده شده است.

پرسش‌های اصلی این پژوهش به‌شرح زیر است که در انتهای پژوهش پاسخ داده می‌شود:

- ◇ آیا ترکیبی از ویژگی‌های مکاشفه، لغوی، هندسی و ویژگی‌های مرتبط با توالی پاراگراف‌ها می‌تواند دقت شناسایی هوشمند فراداده را افزایش دهد؟
- ◇ آیا استفاده از الگوریتم فراابتکاری خفاش[2] می‌تواند در دستیابی به پارامترهای بهینهٔ الگوریتم ماشین بردار پشتیبان مؤثر باشد؟

1. metadata
2. bat algorithm (BA)

(الف) (ب) (پ)

شکل ۱. نمونه‌هایی از مرز فراداده و بدنه در پایان‌نامه. در شکل‌های (الف)، (ب)، و (پ) به‌ترتیب عبارت‌های «فصل ۱:»، «فصل اول»، و «کلیات تحقیق» شروع متن اصلی و بدنهٔ پایان‌نامه هستند.

۲. پیشینهٔ پژوهش

استخراج اطلاعات[1]، فرایند تعیین و شناسایی اصطلاحات از پیش تعریف‌شده در حوزه‌ای خاص، بدون در نظر گرفتن اطلاعات غیرمرتبط است (Piskorski & Yangarber 2013). در این فرایند، موجودیت‌ها، خصیصه‌ها و ارتباط آن‌ها با یکدیگر به‌دست می‌آید. هدف از استخراج اطلاعات، دستیابی به اطلاعات ساخت‌یافته[2] از اسنادِ بدون ساختار[3] یا نیمه‌ساخت‌یافته[4] است (Adrian et al. 2017). فراداده، داده‌ای است که اطلاعاتی دربارهٔ دادهٔ دیگر می‌دهد و در انواع مختلفی دسته‌بندی می‌شوند؛ از جمله: توصیفی (مانند عنوان و نویسنده)، ساختاری (مثل تعداد صفحات) و اجرایی (مانند مالک) (Meng et al. 2018).

طبق مقالهٔ مروری «ناسار، جافری و مالیک»، به‌طور کلی، دو نوع اطلاعات از مقالات علمی قابل استخراج است. فراداده‌ها یکی از این اطلاعات است. برای نمونه، می‌توان به‌عنوان مقاله، نویسندگان، محل انتشار، تاریخ انتشار، کلمات کلیدی و ... اشاره کرد. تحلیل صفحات آغازین مقالات و تجزیهٔ رشته‌های مرجع[5] دو رویکرد متداول برای دستیابی به فراداده‌ها هستند. اطلاعات بالقوهٔ دیگری مانند موضوع پژوهش، مسئله، الگوریتم‌ها، نحوهٔ پاسخ به سؤالات پژوهش، مجموعه داده‌های به‌کاررفته، ابزارها، محدودیت‌های حوزهٔ مطالعه، روش ارزیابی، نتایج و پیشنهاد برای پژوهش‌های آینده نیز در مقالات علمی و دانشگاهی ظاهر می‌شوند که دستیابی به آن‌ها می‌تواند مفید باشد (Nasar, Jaffry, & Malik 2018).

1. information extraction (IE)
2. structured information
3. unstructured documents
4. semi-structured documents
5. references strings

بیشتر پژوهش‌های انجام‌شده در این زمینه بر استخراج فراداده‌ها و سایر اطلاعات ذکرشده از مقالات -به‌ویژه مقالات لاتین- تمرکز کرده‌اند. برای نمونه، هدف مقالهٔ «منگ و همکاران»، به‌دست آوردن اطلاعاتی دربارهٔ موضوع، روش‌های استفاده‌شده و داده‌های به‌دست‌آمده در مقالهٔ ورودی است (Meng et al. 2018). در پژوهشی دیگر، «صفدر و همکاران» به شناسایی شبه‌کدها و تحلیل منابع موجود دربارهٔ شبه‌کد و الگوریتم پیشنهادی در اسناد مقاله توجه کرده‌اند (Safder et al. 2020).

نمونه‌هایی از پژوهش‌های موجود در زمینهٔ استخراج فراداده‌ها از متون علمی عبارت‌اند از: «پنگ و مک‌کالوم» در مطالعهٔ خود سعی در تحلیل بخش‌های اول مقالات دارند و همچنین، رشتهٔ مراجع مقاله را به‌منظور استخراج فراداده‌های مورد نظر پردازش می‌کنند (Peng & McCallum 2006). استخراج فراداده‌های توصیفی از مقالات نشریه‌های علمی و تجزیهٔ فهرست منابع آن‌ها از اهداف مقالهٔ Tkaczyk et al. (2015) است. «بوخرز، آمهور و ستاب» بر استخراج[1] بخش منابع و دستیابی به اطلاعات آن متمرکز شده‌اند (Boukhers, Ambhore & Staab 2019). همچنین، هدف پژوهش Rizvi, Dengel & Ahmed (2019) تشخیص خودکار بخش فهرست منابع در مقالات پژوهشی است.

هدف پژوهش حاضر، استخراج فراداده‌های توصیفی صفحات آغازین پایان‌نامه است. برخی مطالعات انجام‌شده در این زمینه عبارت‌اند از:

«تن‌سازان و مهدوی» در پژوهش خود سعی دارند فراداده‌ها را از سرآیند[2] مقاله و بخش فهرست منابع فارسی و انگلیسیِ اسناد و مقالات علمی فارسی استخراج کنند. در این روش، اولین گام شناسایی بخش‌های آغازین و فهرست منابع متن مقالهٔ ورودی است. فراداده‌های هر مقاله از ابتدای متن آن شروع شده و به ابتدای مقدمه ختم می‌شود. بنابراین، برای شناسایی مرز شروع بدنهٔ اصلی مقاله، در ساده‌ترین حالت، یافتن کلمهٔ کلیدی «مقدمه» نشان‌دهندهٔ شروع بدنهٔ اصلی است (۱۳۹۶). البته، به‌دلیل تنوع قالب‌های ساختاری مقالات، این روش به‌تنهایی کفایت نمی‌کند. به‌عنوان نمونه ممکن است پیش از شروع بدنهٔ اصلی مقاله، چکیدهٔ انگلیسی قرار داشته باشد. در پژوهش حاضر این بخش جزو سرآیند مقاله در نظر گرفته نشده و راه‌حل‌هایی برای استثناها ارائه شده است. شناسایی فهرست منابع، تنها بر اساس ویژگی‌های متنی انجام شده است و استخراج

1. extracting
2. header

فراداده از سرآیند مقاله و فهرست منابع فارسی و انگلیسی با مدل آماری «میدان تصادفی شرطی»[1] انجام گرفته است.

در مطالعهٔ «پنگ و مک‌کالوم» مقاله‌ای با داده‌های متنی بررسی شده و فراداده‌های سرآیند و اطلاعات بخش منابع مورد توجه قرار گرفته است. به‌منظور استخراج اطلاعات از سرآیند مقالات، داده‌های استخراج‌شده از سرآیند شامل کلمات ابتدای پایان‌نامه تا اولین بخش مقاله (به‌طور معمول مقدمه) و یا کل کلمات صفحهٔ اول بوده‌اند. برای استخراج فراداده از بخش منابع نیز از رشتهٔ متنی منابع استفاده شده است. در این روش، افزون بر ویژگی‌های لغوی، ویژگی‌های مربوط به طرح‌بندی و هجی کلمات نیز در نظر گرفته شده و روش میدان تصادفی شرطی برای استخراج فراداده‌ها استفاده شده است (Peng & McCallum 2006).

هدف مقالهٔ «کرن» و همکاران استخراج فراداده از مقالات در قالب «پی‌دی‌اف»[2] است. پژوهشگران، مرحلهٔ پیش‌پردازش را در دو گام انجام داده‌اند: ۱) بلوک‌های متنی به‌دست آمده از هر سند را در گروه فراداده‌های مورد نظر طبقه‌بندی می‌کنند (بلوک‌های متنی که فراداده نیستند، در گروه «سایر» دسته‌بندی می‌شوند)؛ و ۲) هر بلوک متنی از چندین «توکن»[3] تشکیل شده است و در گام دوم طبقه‌بندی در سطح «توکن»ها انجام می‌شود. بدین‌ترتیب که هر یک از کلمات در چندین گروه طبقه‌بندی می‌شود. به‌عنوان نمونه، بلوک متنی که در گام اول برچسب «نویسنده» گرفته، در گام بعدی به هر کلمهٔ بلوک یکی از برچسب‌های نام، نام خانوادگی، ایمیل، وابستگی و ... تخصیص می‌یابد. در انتهای فرایند، فراداده‌های مورد نظر استخراج می‌شود (Kern et al. 2012).

«کان، لانگ و انگوین» و «کونگ» و همکاران در مطالعات خود با تحلیل ساختار منطقی سند[4]، فراداده‌های مورد نظر را استخراج کرده‌اند (Kan, Luong & Nguyen 2010; Cuong et al. 2015). روش ارائه‌شده در «کان، لانگ و نوین» به هر بخش از مقاله بر اساس نقش آن‌ها در سند، برچسبی از فراداده‌ها تخصیص می‌دهند (Kan, Luong & Nguyen 2010).

«دو» و همکاران در مطالعهٔ خود به نحوهٔ استخراج نام نویسندگان، راه‌های دسترسی به آن‌ها (مانند ایمیل، شمارهٔ تلفن، محل خدمت و ...) و تطبیق آن‌ها با نام هر نویسنده

1. conditional random field (CRF) 2. PDF 3. token
4. logical structure analysis

پرداخته‌انـد (Do et al. 2013). در پژوهـش حاضـر، پیـش از آغـاز فـاز شناسـایی فرادادههـا، نسـخهٔ «پـی‌دی‌اف»، بـا هـدف حفـظ ویژگی‌هـای مکانـی و هندسـی محتـوا بـه «ایکـس‌ام‌ال»[1] تبدیـل شـده و بـه هـر خـط بـر اسـاس نقـش آن در متـن مقالـه، برچسـبی بـه روش Kan, Luong & Nguyen (2010) اختصـاص دادهانـد. سـپس، خطوطـی کـه یکـی از دو برچسـب «نویسـنده» و یـا «وابسـتگی» را داشـته باشـند، وارد سیسـتم اصلـی می‌شـوند. سیسـتم شـامل دو مرحلـه اسـت: در مرحلـهٔ نخسـت، نـام نویسـنده (یـا نویسـندگان) و اطلاعـات وابسـتگی‌ها شناسـایی شـده، و در مرحلـهٔ بعـد، وابسـتگی‌های هـر نویسـنده بـا نـام او تطبیـق داده می‌شـود. بخـش شناسـایی بـا روش «یادگیـری ماشـین بـا نظـارت» و روش «میـدان تصادفـی شـرطی» انجـام شـده و بخـش تطبیـق و یافتـن ارتباطـات نیـز بـا «روش ماشـین بـردار پشـتیبان» صـورت می‌گیـرد.

«سـوزا، موریـرا و هیوسـر» در مطالعـهٔ خـود، بـا توجـه بـه اینکـه فرادادههـای یـک مقالـه در صفحـهٔ اول آن ظاهـر می‌شـود، فقـط بـه تحلیـل صفحـهٔ اول اکتفـا کردهانـد. پـردازش در دو مرحلـه تعریـف شـده اسـت: نخسـت، بخش‌هـای مختلـف صفحـهٔ اول، یعنـی سـرعنوان، عنـوان، اطلاعـات نویسـنده، بدنـه و پاورقـی مشـخص شـده، در مرحلـهٔ بعـد، فرادادههـای موجـود در هـر بخـش اسـتخراج می‌شـوند (Souza, Moreira & Heuser 2014).

Tkaczyk et al. (2015) بـرای اسـتخراج فـرادادهٔ از صفحـات ابتدایـی و فهرسـت منابـع مقالات علمـی، سیسـتمی بـا ورودی قالـب «پـی‌دی‌اف» ارائـه کردنـد[2]. از ایـن رو، نخسـت، محتـوای متنـی و ویژگی‌هـای تصویـری آن‌هـا اسـتخراج می‌شـود. ایـن سیسـتم خـود بـه سـه بخـش اسـتخراج سـاختار پایـه[3]، اسـتخراج فـرادادهٔ، و اسـتخراج فهرسـت مقـالات ارجاعـی[4] تقسـیم می‌شـود. نخسـت، سـاختار سلسـله‌مراتبی سـند به‌دسـت می‌آیـد. سـپس، نواحـی مقالـه بـه چهـار بخـش فـرادادهٔ، بدنـه، مراجـع و سـایر طبقه‌بنـدی می‌شـود؛ در ادامـه، ناحیه‌هایـی کـه برچسـب فـرادادهٔ گرفته‌انـد، بـه انـواع فرادادههـای از قبـل تعریف‌شـده طبقه‌بنـدی می‌شـود. هـر دو طبقه‌بنـدی بـا روش «ماشـین بـردار پشـتیبان» انجـام شـده اسـت (Cortes & Vapnik 1995). در بخـش اسـتخراج فهرسـت منابـع، رشـتهٔ متنـی مراجـع تجزیـه شـده و فرادادههـای منابـع مقـالات مشـخص می‌شـود (Tkaczyk et al. 2015).

تطبیـق گـراف سـاختار[5] سـند نیـز می‌توانـد در یافتـن موجودیت‌هـای مـورد نظـر مؤثـر

1. XML
2. http://cermine.ceon.pl/index.html
3. basic structure extraction
4. bibliography extraction
5. structure graph matching

باشد. برای نمونه، «کولی و بلید» در مطالعهٔ خود روشی ارائه کرده‌اند که افزون بر موجودیت‌های مقالات علمی، اسناد سازمانی و فاکتور خرید هم بررسی می‌شود (Kooli & Belaid 2016). در این روش، برچسب‌زنی موجودیت‌ها پس از استخراج متن از تصویر ورودی با استفاده از عبارات منظم و لغت‌نامه انجام می‌شود. در واقع، این برچسب‌زنی تنها بر اساس متن انجام شده و ویژگی‌های تصویری و ساختاری در مراحل بعدی مورد استفاده قرار می‌گیرد. در مرحلهٔ بعد، ساختارهای محلی در هر بخش از تصویر سند با یک گراف مدل می‌شود و با مدل‌هایی که سیستم از قبل یادگرفته (یادگیری به روش بدون نظارت یا خوشه‌بندی) تطبیق داده شده، نزدیک‌ترین مدل یافت می‌شود. از این مدل برای اصلاح خطاهای احتمالی و به‌روزرسانی استفاده می‌شود. در آخرین مرحله نیز موجودیت‌ها شناسایی می‌شوند.

«لیو» و همکاران در مقالهٔ خود با به‌کارگیری شبکه‌های عمیق[1] و ترکیب ویژگی‌های متنی و تصویری، فراداده‌ها را از سرآیند فایل «پی‌دی‌اف» مقالات استخراج کرده‌اند. در این روش با استفاده از شبکهٔ عصبی پیچشی[2]، ویژگی‌های متنی و تصویری به‌دست آمده و خروجی به شبکهٔ حافظهٔ طولانی کوتاه‌مدت[3] فرستاده می‌شود (Liu et al. 2017).

هدف مقاله «فریز» و همکاران، ارائهٔ ابزاری برای تبدیل فایل «پی‌دی‌اف» به «ایکس‌ام‌ال»، دستیابی به محتوای متنی و بخش‌های مختلف مقاله، مانند فراداده‌ها، بخش‌ها، زیربخش‌ها و مراجع است (Ferréset al. 2018). «کیو و ژو» مطالعهٔ دیگری در زمینهٔ استخراج فراداده‌ها از کتاب‌های دیجیتال انجام داده‌اند که در آن دسترسی به اطلاعاتی مانند عنوان کتاب، نویسنده، ناشر، تاریخ انتشار، شمارهٔ استاندارد بین‌المللی کتاب[4] (شابک) و ... بررسی شده است (Qiu & Zhou 2019). اگرچه مقالهٔ ذکرشده بر اسناد علمی تمرکز ندارد، اما از این جهت که به‌دنبال فراداده‌های توصیفی کتاب‌های دیجیتال است، با مسئله‌ای مشابه روبه‌روست.

در پژوهش حاضر، برای تنظیم پارامترهای ماشین بردار پشتیبان، از الگوریتم خفاش به‌عنوان یک الگوریتم فرامکاشفه‌ای[5]، بهره گرفته شده است؛

1. deep neural networks
2. convolutional neural network (CNN)
3. long short-term memory (LSTM)
4. ISBN
5. metahuristic

۲-۱. الگوریتم خفاش

مفهوم الگوریتم خفاش را نخستین‌بار Yang (2010) مطرح کرد. این الگوریتم با الهام از نحوهٔ جهت‌یابی خفاش با پژواک صدا طراحی شده و برای بهینه‌سازی مسایل پیوسته مناسب است. برای سادگی موضوع، قواعد زیر در نظر گرفته شده است:

- ◇ همهٔ خفاش‌ها از طریق ارسال صدا فاصله را درک کرده و تفاوت بین طعمه و موانع را تشخیص می‌دهند؛
- ◇ خفاش‌ها به‌صورت تصادفی و با سرعت v_i در مکان x_i پرواز می‌کنند و با فرکانس ثابت f_{min}، طول موج λ و بلندی صدای A_0 متغیر به‌دنبال طعمه می‌گردند. خفاش‌ها طول موج (یا فرکانس) پالس صدا و نرخ پالس انتشار $r \in [0,1]$ را بر اساس میزان نزدیکی به هدف خود به‌صورت خودکار تنظیم می‌کنند؛
- ◇ بلندی صدا نیز می‌تواند تغییر کند. در این الگوریتم فرض می‌شود که بلندی صدا از مقدار A_0، که عددی مثبت است، تا مقدار ثابت و کمینه A_{min} تغییر می‌کند؛
- ◇ همچنین، برای سادگی فرض می‌شود که از ردیابی صدا برای محاسبهٔ میزان تأخیر و نقشه‌برداری سه‌بعدی استفاده نمی‌شود.

به‌طور کلی، محدودهٔ فرکانس f در بازهٔ $[f_{min}, f_{max}]$ است و با طول موج در بازهٔ $[\lambda_{min}, \lambda_{max}]$ متناسب است.

از آن‌جا که حاصل ضرب λf مقداری ثابت و برابر سرعت است، در پیاده‌سازی الگوریتم فقط مقدار فرکانس تغییر می‌یابد.

در شبیه‌سازی، خفاش‌های مصنوعی هر کدام موقعیت و سرعتی دارند که در یک فضای dبعدی حرکت می‌کنند و در گام t موقعیت x_i^t و سرعت v_i^t آن‌ها به‌روزرسانی می‌شود (رابطهٔ ۱، رابطهٔ ۲، و رابطهٔ ۳).

رابطهٔ ۱ $$f_i = f_{min} + (f_{max} - f_{min})\beta$$

رابطهٔ ۲ $$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_*)f_i$$

رابطهٔ ۳ $$x_i^t = x_i^{t-1} + v_i^t$$

پارامتر $\beta \in [0,1]$ عددی تصادفی از توزیع یکنواخت است. x_* بهترین مکان سراسری است که از مقایسهٔ مکان همهٔ خفاش‌ها (در هر تکرار) به‌دست می‌آید.

هنگام جست‌وجوی محلی، پس از یافتن بهترین راه‌حل از راه‌حل‌های فعلی، راه‌حل

جدیدی به روش مسیر تصادفی برای هر خفاش تولید می‌شود (رابطهٔ ۴).

$$x_{new} = x_{old} + \varepsilon A^t$$ **رابطهٔ ٤**

پارامتر $\varepsilon \in [-1,1]$ عددی تصادفی است. $A^t =< A_i^t >$ میانگین بلندی صدای همهٔ خفاش‌ها در تکرار t است.

از طرفی، هنگامی که خفاش طعمه را پیدا می‌کند، بلندی صدای خود را کاهش و نرخ پالس را افزایش می‌دهد (رابطهٔ ۵).

$$A_i^{t+1} = \alpha A_i^t \ , \ r_i^{t+1} = r_i^0(1 - \exp(-\gamma t))$$ **رابطهٔ ٥**

α و γ دو عدد ثابت هستند. در واقع، α فاکتور خنک‌سازی است و برای هر $0 < \alpha < 1$ و $\gamma > 0$ اگر $t \to \infty$ داریم:

$$A_i^t \to 0 \ , \quad r_i^t \to r_i^0$$

برای سادگی پیاده‌سازی این الگوریتم، حالت $\alpha = \gamma = 0.9$ در نظر گرفته شده است.

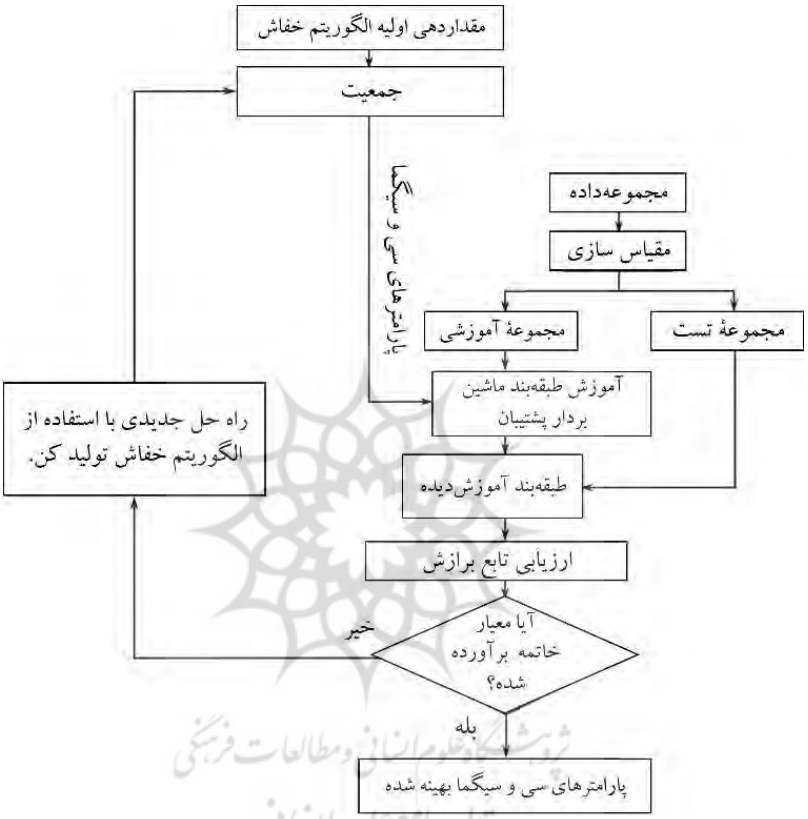
۲-۲. ترکیب ماشین بردار پشتیبان و الگوریتم خفاش

«ثروت، حسنیان و النقی» در مقالهٔ خود برای تنظیم پارامترهای ماشین بردار پشتیبان از الگوریتم خفاش بهره گرفته‌اند (Tharwat, Hassanien & Elnaghi 2017). این بهینه‌سازی به‌صورت دو آزمایش روی ۹ مجموعه دادهٔ دانشگاه کالیفرنیا[1] انجام شده است. در آزمایش اول، هدف، بهینه‌سازی پارامتر خطا C و پارامتر هستهٔ چندجمله‌ای[2] d بوده و در آزمایش دوم، پارامتر خطا C و پارامتر هستهٔ تابع پایهٔ شعاعی[3] σ بهینه‌سازی شده است. روش آن‌ها به این صورت است که ابتدا داده‌ها نرمال شده (رابطهٔ ۶) و به روش اعتبارسنجی متقابل k-fold و k=10 به دو دستهٔ داده‌های آموزشی و داده‌های آزمایشی تقسیم می‌شوند. سپس، الگوریتم خفاش اجرا شده و برای وضعیت هر خفاش، پارامترهای به‌دست آمده برای آموزش داده‌ها با ماشین بردار پشتیبان اعمال می‌شود. پس از محاسبهٔ نرخ خطا (رابطهٔ ۷) بهترین وضعیت (X^*) خفاش‌ها به‌دست می‌آید. در این جا، تابع برازش[4] همان نرخ خطاست، N_e تعداد نمونه‌هایی است که برچسب اشتباه خورده‌اند، و N تعداد کل داده‌های آزمایش است. این الگوریتم تا برآورده‌شدن معیارِ خاتمه ادامه می‌یابد (شکل ۲).

1. https://archive.ics.uci.edu/ml/datasets.php
2. polynomial kernel
3. radial basis function (RBF)
4. fitness function

$$v' = \frac{v - min}{max - min}$$ رابطهٔ ٦

$$Minimize: F = \frac{N_e}{N}$$ رابطهٔ ۷



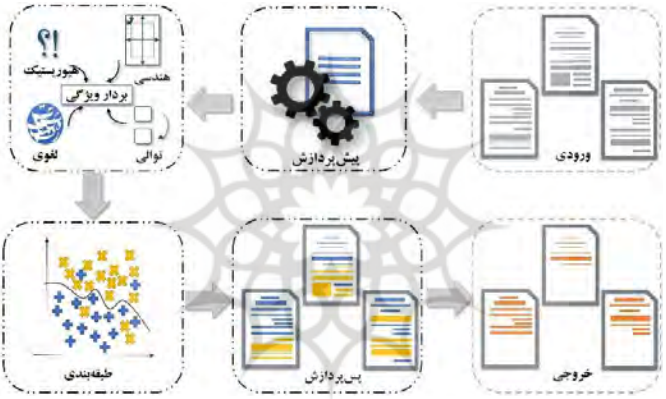
شکل ۲. نمودار گردش کار بهینه‌سازی ماشین بردار پشتیبان با الگوریتم خفاش
(Tharwat, Hassanien & Elnaghi 2017)

۳. روش پژوهش

در این بخش دربارهٔ روش پیشنهادی برای تشخیص مرز بدنهٔ اصلی در اسناد پایان‌نامه‌های فارسی توضیح داده می‌شود.

هر پایان‌نامه قبل از شروعِ فصل‌های اصلی از بخش‌هایی چون عنوان فارسی (مطابق با جلد فارسی)، اصالت و مالکیت، صورت‌جلسهٔ دفاع، تقدیم‌نامه، سپاسگزاری، چکیدهٔ فارسی و فهرست‌ها تشکیل می‌شود. هدف پژوهش حاضر، پیشنهاد روشی برای شناسایی

و جداسازی پاراگراف‌های مربوط به این بخش‌ها از بدنهٔ اصلی و شروع فصل‌های اصلی پایان‌نامه است. در این روش، متن پایان‌نامهٔ ورودی در پیش‌پردازش یکسان‌سازی[1] می‌شود. هدف این کار، یکسان‌سازی یونی‌کد[2] کاراکترهایی است که ممکن است چند یونی‌کد داشته باشند؛ مانند حرف «ی»، «و» و «ک». به‌دنبال آن ویژگی‌ها برای پاراگراف‌های فایل DOCX پایان‌نامه‌های ورودی محاسبه می‌شود. سپس، پاراگراف‌ها با کمک ماشین بردار پشتیبان در دو گروه بدنه و فراداده طبقه‌بندی می‌شوند. در نهایت، در فاز پس‌پردازش بررسی می‌شود که از کدام پاراگراف به بعد اکثریت پاراگراف‌ها برچسب بدنه دارند. به این ترتیب، بخش‌های آغازین پایان‌نامه به‌دست می‌آید و می‌توان فراداده‌های مورد نظر را از این بخش استخراج کرد (شکل ۳).



شکل ۳. مراحل روش پیشنهادی برای تشخیص مرز بدنهٔ پایان‌نامه‌های فارسی

یکی از بخش‌های سیستم ارائه‌شده در مقالهٔ «کازیک» و همکاران، استخراج محتوا با هدف تعیین نقش هر بخش از مقاله است (Tkaczyki et al. 2015). بدین‌ترتیب، بخش‌های مختلف مقالهٔ ورودی، با الگوریتم ماشین بردار پشتیبان در چهار گروه فراداده (عنوان، نویسندگان، چکیده، کلمات کلیدی و ...)، منابع، بدنه و سایر (تشکر و قدردانی، شمارهٔ صفحه و ...) طبقه‌بندی می‌شوند. بدین‌منظور، ویژگی‌هایی از بخش‌های مختلف مقاله استخراج می‌شود. در روش پیشنهادی نیز ویژگی‌های به‌دست‌آمده از هر پاراگراف، از برخی ویژگی‌های نام‌برده در سیستم ارائه شده در مقالهٔ «کازیک» اقتباس شده است.

1. normalization 2. unicode

ویژگی‌های استخراج‌شده به چهار دستهٔ کلی تقسیم می‌شوند: ویژگی‌های مکاشفه‌ای، ویژگی‌های هندسی، ویژگی‌های مربوط به توالی[1]، و ویژگی‌های لغوی. در مجموع، ۱۵ ویژگی برای هر پاراگراف محاسبه شده است (جدول ۱). ویژگی‌های ۱ تا ۱۲ از نوع ویژگی‌های مکاشفه‌ای هستند. ویژگی‌های ۱۳، ۱۴ و ۱۵ به‌ترتیب، ویژگی‌های مربوط به توالی پاراگراف‌ها، لغوی، و هندسی هستند. در ویژگی ۱۴ منظور از کلمات کلیدی رایج در پاراگراف‌های ابتدای بدنه، عباراتی مانند مقدمه، فصل اول، کلیات تحقیق و ... هستند.

تشخیص مرز پاراگراف‌های فراداده و بدنه با استفاده از الگوریتم ماشین بردار پشتیبان به یک مسئلهٔ طبقه‌بندی دو گروهی تبدیل شده است. فرایند آموزش و بهینه‌سازی مشابه مقالهٔ Tharwat, Hassanien & Elnaghi (2017) بوده و تابع پایهٔ شعاعی به‌عنوان هستهٔ الگوریتم ماشین بردار پشتیبان انتخاب شده است. نتایج به‌دست‌آمده در این مقاله نشان می‌دهد که ترکیب ماشین بردار پشتیان با الگوریتم خفاش، نرخ خطای کمتری نسبت به سایر روش‌ها، از جمله الگوریتم پرندگان[2] دارد؛ زیرا در الگوریتم پرندگان، تمامی پارامترها در ابتدای الگوریتم مقادیر ثابتی دارند، اما پارامترهای الگوریتم خفاش در طول تکرار می‌توانند تغییر یابند. به‌عبارت دیگر، الگوریتم پرندگان نوع خاصی از الگوریتم خفاش است. همچنین، نتایج آن‌ها نشان می‌دهد که بهینه‌سازی پارامترهای ماشین بردار پشتیبان با الگوریتم ژنتیک در مقایسه با الگوریتم خفاش، نرخ خطای بیشتری دارد. از این‌رو، در پژوهش حاضر، برای بهینه‌سازی پارامترها از الگوریتم خفاش به‌عنوان الگوریتم تکاملی استفاده شده است. پارامترهای اولیهٔ الگوریتم خفاش بر اساس مقالهٔ پایه مقداردهی شده (جدول ۲) و برای یافتن تعداد خفاش‌ها و تعداد تکرار، آزمایش‌های همین مقاله انجام شده است.

1. sequential
2. particle swarm optimization (PSO)

جدول ۱. ویژگی‌های استخراج‌شده از پاراگراف‌ها

ویژگی ها	
۱. تعداد تکرار کاراکترهای ’]‘ یا ’[‘	۹. تعداد تکرار کاراکترهای ’(‘ یا ’)‘ نسبت به کل کاراکترها
۲. تعداد تکرار کاراکترهای ’]‘ یا ’[‘ نسبت به کل کاراکترها	۱۰. تعداد تکرار علایم نگارشی نسبت به کل کاراکترها
۳. تعداد ارقام فارسی و انگلیسی	۱۱. تعداد کلمات پاراگراف
۴. تعداد ارقام فارسی و انگلیسی نسبت به کل کاراکترها	۱۲. آیا پاراگراف با عدد شروع شده است؟
۵. تعداد تکرار کاراکتر نقطه (.)	۱۳. شمارهٔ پاراگراف
۶. تعداد تکرار کاراکتر نقطه نسبت به کل کاراکترها	۱۴. تعداد کلمات کلیدی رایج در پاراگراف‌های ابتدای بدنه
۷. تعداد خطوط	۱۵. نسبت طول به عرض پاراگراف
۸. تعداد تکرار کاراکترهای ’(‘ یا ’)‘	

جدول ۲. مقداردهی اولیهٔ پارامترهای الگوریتم خفاش

پارامتر	مقدار
فرکانس (مقادیر کمینه و بیشینه)	$f_{max} = 2$ ، $f_{min} = 0$
نرخ پالس	r=0.5
بلندی صدا	A=0.5

پس از دسته‌بندی پاراگراف‌ها در دو گروه فراداده و بدنه، ضروری است پاراگراف شروع بدنهٔ اصلی پایان‌نامه شناسایی شود. به همین منظور، بررسی می‌شود که از کدام پاراگراف به بعد، اکثریت پاراگراف‌ها در گروه بدنه دسته‌بندی شده‌اند. بدین‌ترتیب، بخش‌های آغازین پایان‌نامهٔ ورودی از متن اصلی تفکیک می‌شود.

در این مقاله زبان برنامه‌نویسی «پایتون»[1] نسخهٔ ۳/۶ برای پیاده‌سازی فرایند یادگیری و بهینه‌سازی به‌کار گرفته شده است. الگوریتم‌ها روی سیستمی با پردازندهٔ «اینتل کور»[2]، حافظهٔ رم[3] ۳۲ گیگابات و سیستم عامل «اوبونتو»[4] ۱۶/۰۴ اجرا شده است.

1. Python 2. Intel core i7 6700K 3. RAM 4. Ubuntu 16.04

٤. تجزیه و تحلیل یافته‌ها

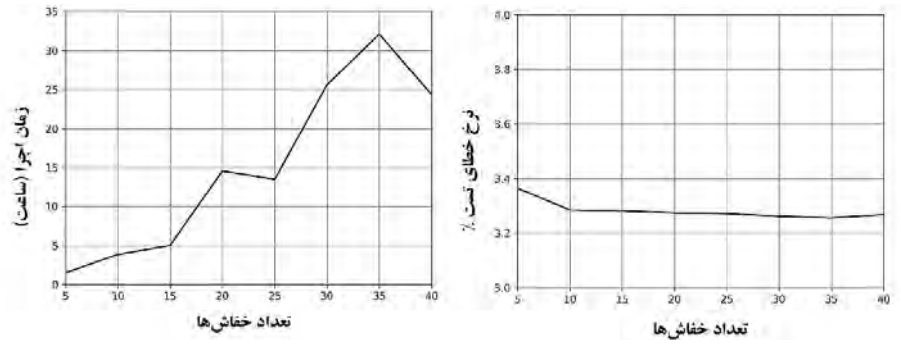
در این بخش دربارهٔ مجموعهٔ داده‌های به‌کار رفته، آزمایش‌های انجام‌شده برای یافتن پارامترها و نتایج به‌دست‌آمده از عملکرد طبقه‌بندی توضیح داده می‌شود.

شکل ٤. نمودار دسته‌بندی پایان‌نامه‌ها

مجموعهٔ داده‌ها شامل ۲۵۰ پایان‌نامهٔ ثبت‌شده در «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)[1] و در قالب DOCX هستند. داده‌ها به‌صورت تصادفی و از رشته‌های مختلف انتخاب شده‌اند. پایان‌نامه‌های انتخابی در پنج دستهٔ علوم انسانی، علوم پایه، فنی و مهندسی، هنر، و علوم پزشکی قرار می‌گیرند (شکل ۴). از فایل DOCX می‌توان اطلاعاتی چون محتوای پاراگراف‌ها، جداول، قالب و فونت، حاشیهٔ صفحات و ... را به‌دست آورد که در محاسبهٔ ویژگی‌ها مورد استفاده قرار می‌گیرند. از آنجا که جداکنندهٔ صفحات در فایل DOCX ظاهر نمی‌شود، در این پژوهش پاراگراف‌های پایان‌نامه‌ها مورد بررسی قرار گرفته‌اند. در مجموع، از ۲۵۰ پایان‌نامه، ۸۶۲۷۰ پاراگراف برای یادگیری به‌دست آمد. از بین آن‌ها تعداد ۴۰۴۱۸، پاراگراف‌های فراداده و تعداد ۴۵۸۵۲، پاراگراف‌های بدنهٔ اصلی هستند.

برای تعیین تعداد خفاش‌ها و تعداد تکرار آزمایش‌هایی انجام شد. در نخستین آزمایش میزان خطا به ازای مقادیر ۵ تا ۴۰ خفاش به‌دست آمد و زمان مورد نیاز برای اجرای الگوریتم محاسبه شد (نمودار ۱).

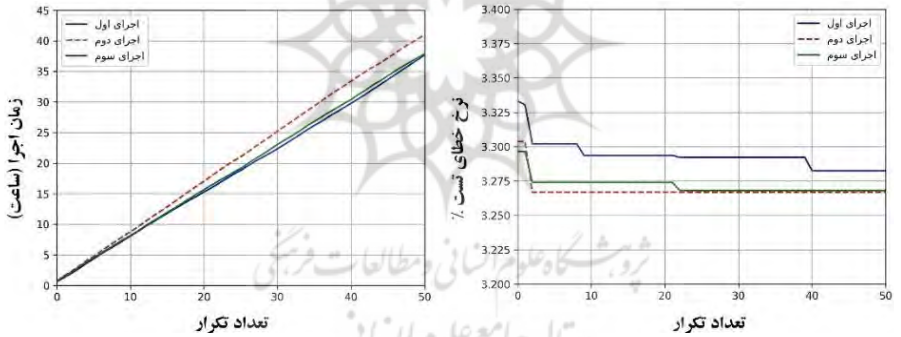
1. https://irandoc.ac.ir/



(ب) (الف)

نمودار ۱. آزمایش تأثیر تعداد خفاش بر عملکرد الگوریتم: (الف) نرخ خطای تست به ازای مقادیر مختلف خفاش‌ها؛ (ب) زمان اجرا به ازای مقادیر مختلف خفاش‌ها

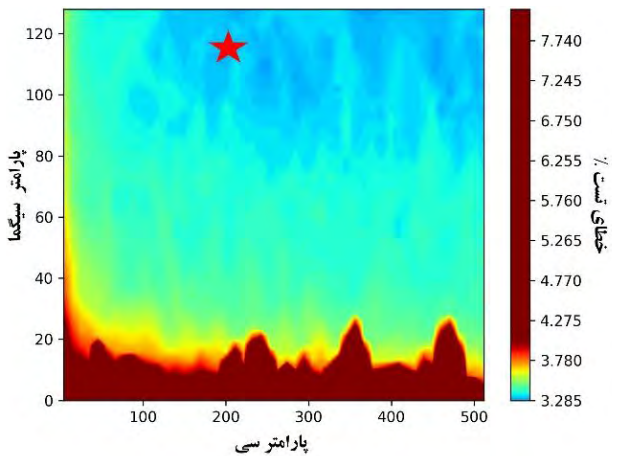
مطابق نمودار ۱، با افزایش تعداد خفاش‌ها، زمان اجرا نیز افزایش می‌یابد. همچنین، برای بررسی تأثیر تعداد تکرار بر عملکرد الگوریتم، سه آزمایش انجام شده است (نمودار ۲).



(ب) (الف)

نمودار ۲. آزمایش تأثیر تعداد خفاش بر عملکرد الگوریتم: (الف) نرخ خطای تست به‌تعداد تکرارهای مختلف؛ (ب) زمان اجرا به‌ازای افزایش تعداد تکرار

تابع نرخ خطا در این مسئله به ازای مقادیر مختلف پارامترهای C و σ رسم شده است (نمودار ۳). همان‌طور که مشاهده می‌شود، نرخ خطا دارای تعداد زیادی کمینهٔ محلی است.



نمودار ۳. نمودار تراز نرخ خطا به ازای پارامترهای خطا (c) و سیگما. نقطه‌ای که با ستاره مشخص شده نقطه‌ی کمینه است که به کمک الگوریتم خفاش به‌دست آمده است

نتایج به‌دست‌آمده از الگوریتم خفاش با روش جست‌وجوی شبکه‌ای[1] نیز مقایسه شده است (جدول ۳).

جدول ۳. مقایسهٔ عملکرد الگوریتم خفاش و جست‌وجوی شبکه‌ای

	Grid search-SVM	BA- SVM
درصد خطای تست	۳/۳۱۴	۳/۲۶۸

برای ارزیابی فرایند یادگیری، داده‌های آموزشی به روش اعتبارسنجی متقابل k-fold و k=10 جداسازی شده‌اند. با توجه به ویژگی‌های استخراج‌شده برای هر پاراگراف (جدول ۱)، ۸۰ درصد ویژگی‌های محاسبه‌شده از نوع ویژگی‌های مکاشفه‌ای هستند؛ از این رو، تأثیر سایر ویژگی‌ها و ترکیب آن‌ها با ویژگی‌های مکاشفه‌ای بر عملکرد طبقه‌بندی بررسی شده و نتایج در جدول ۴، خلاصه شده است. با توجه به یافته‌های به‌دست‌آمده، به نظر می‌رسد ویژگی هندسی تأثیری بر بهبود عملکرد نداشته و مفید نیست.

1. grid search

جدول ٤. تأثیر انواع ویژگی‌ها بر عملکرد طبقه‌بندی پاراگراف‌ها

نوع ویژگی/ معیارهای ارزیابی (میانگین)	دقت	صحت	بازخوانی	امتیاز اف ۱
مکاشفه‌ای	۷۶/۰	۸۷/۴	۷۶/۰	۷۹/۵
مکاشفه‌ای + هندسی	۷۶/۱	۸۷/۴	۷۶/۱	۷۹/۵
مکاشفه‌ای + لغوی	۸۷/۳	۸۳/۶	۸۳/۲	۸۳/۴
مکاشفه‌ای + توالی	۹۴/۴	۹۵/۵	۹۵/۲	۹۵/۳
مکاشفه‌ای + هندسی + لغوی	۸۳/۴	۸۳/۷	۸۳/۴	۸۳/۵
مکاشفه‌ای + توالی + لغوی	۹۶/۷	۹۵/۸	۹۵/۵	۹۵/۷
مکاشفه‌ای + هندسی + توالی	۹۵/۲	۹۵/۵	۹۵/۲	۹۵/۳
مکاشفه‌ای + هندسی + توالی+ لغوی	۹۶/۶	۹۵/۸	۹۵/۴	۹۵/۶

برای ارزیابی عملکرد طبقه‌بندی، معیارهای ارزیابی دقت، صحت، بازخوانی[1] و امتیاز اف ۱، برای گروه‌های فراداده و بدنه محاسبه شده است (جدول ۵). عملکرد روش پیشنهادی با بخش استخراج محتوا و طبقه‌بندی اولیهٔ قسمت‌های مختلف مقاله در سیستم «سرمین»[2] نیز مقایسه شده است (Tkaczyk 2015) (جدول ۶). سیستم «سرمین» در ارزیابی طبقه‌بندی الگوریتم ماشین بردار پشتیبان هسته‌های خطی، چندجمله‌ای، تابع پایهٔ شعاعی و سیگموید[3] را بررسی کرده و بهترین نتیجه در استفاده از هستهٔ تابع پایهٔ شعاعی به‌دست آمده است. در سیستم پیشنهادی نیز هستهٔ تابع پایهٔ شعاعی مورد استفاده قرارگرفته است.

جدول ۵. نتیجهٔ طبقه‌بندی پاراگراف‌های پایان‌نامه

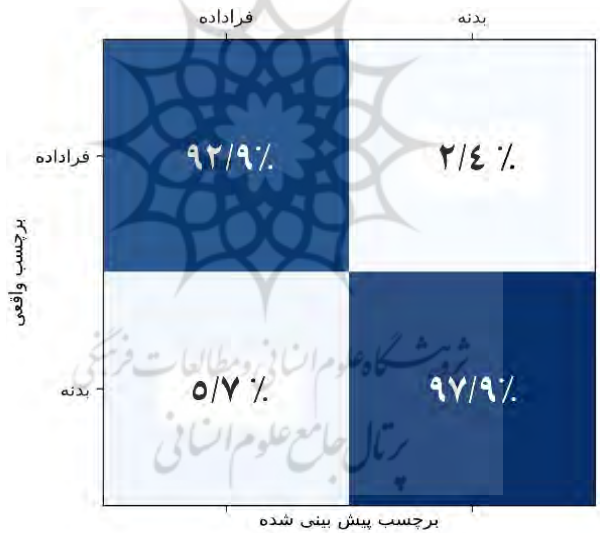
	دقت	صحت	بازخوانی	امتیاز اف ۱
پاراگراف‌های فراداده	۹۲/۹	۹۴/۱	۹۲/۹	۹۳/۵
پاراگراف‌های بدنه	۹۷/۹	۹۷/۵	۹۷/۹	۹۷/۷
میانگین	۹۶/۶	۹۵/۸	۹۵/۴	۹۵/۶

1. recall 2. CERMINE 3. Sigmoid

جدول ٦. مقایسهٔ عملکرد روش پیشنهادی و سیستم «سرمین» (Tkaczyk et al. 2015)

معیار/ روش	سرمین	روش پیشنهادی
میانگین امتیاز اف ۱ (%)	۹۳/۹	۹۵/۶

ماتریس درهم‌ریختگی دو گروه بدنه و فراداده نیز در شکل ۸، آمده است. همان‌طور که انتظار می‌رود، امکان رخداد دو خطا وجود دارد: نخست، تخصیص برچسب بدنه به پاراگراف‌های فراداده، و دوم تخصیص برچسب فراداده به پاراگراف‌های بدنه. خطای نوع اول به‌طور معمول، در پاراگراف‌های طولانی‌تر رخ می‌دهد. برای نمونه، پاراگراف‌های چکیده از دیگر بخش‌های آغازین پایان‌نامه دارای کاراکترهای بیشتری بود. از این رو، طولانی‌تر از سایر بخش‌هاست. در مقابل، خطای نوع دوم در تیترها یا پاراگراف‌های کوتاهِ بدنهٔ اصلی پایان‌نامه امکان رخداد دارد.



شکل ۸. ماتریس درهم‌ریختگی

به‌منظور مقایسهٔ عملکرد طبقه‌بندی در انواع پایان‌نامه‌ها، آزمایش دیگری انجام شده است (جدول ۷). همان‌طور که پیش‌تر اشاره شد، پایان‌نامه‌ها در پنج دستهٔ علوم پایه، علوم انسانی، فنی و مهندسی، هنر، و علوم پزشکی قرار می‌گیرند. طبق نتایج به‌دست‌آمده، به نظر می‌رسد که طبقه‌بندی پاراگراف‌ها در پایان‌نامه‌های علوم انسانی، عملکرد بهتری نسبت به سایر رشته‌ها دارد.

جدول ۷. مقایسهٔ عملکرد طبقه‌بند در انواع پایان‌نامه‌ها

دسته‌بندی	میانگین دقت (درصد)	میانگین صحت (درصد)	میانگین بازخوانی (درصد)
علوم پایه	۹۷/۰	۹۶/۶	۹۶/۰
علوم انسانی	۹۷/۵	۹۷/۰	۹۶/۵
فنی و مهندسی	۹۶/۵	۹۵/۰	۹۴/۰
هنر	۹۷/۴	۹۶/۹	۹۶/۴
علوم پزشکی	۹۶/۱	۹۵/۸	۹۵/۱

۵. نتیجه‌گیری

در پژوهش حاضر، جداسازی خودکار مرز فراداده‌ها از بدنهٔ پایان‌نامه‌های فارسی مورد بررسی قرارگرفته است. بدین منظور، مجموعهٔ داده‌ای متشکل از پایان‌نامه‌های علوم پایه، علوم انسانی، فنی و مهندسی، هنر، و علوم پزشکی جمع‌آوری شده است. برای پاراگراف‌های هر پایان‌نامه، ویژگی‌های مکاشفه‌ای، لغوی، توالی و هندسی محاسبه می‌شود. بر اساس یافته‌های پژوهش به نظر می‌رسد که ویژگی هندسی تأثیری در بهبود عملکرد طبقه‌بند نداشته است. سایر انواع ویژگی‌های تعریف‌شده برای حل این مسئله کفایت می‌کنند. برای طبقه‌بندی پاراگراف‌های پایان‌نامه‌ها، از الگوریتم ماشین بردار پشتیبان استفاده شده و پارامترهای آن با الگوریتم فرامکاشفه‌ای خفاش، بهینه‌سازی شده است. پاراگراف‌ها در دو گروه فراداده و بدنه طبقه‌بندی شده و نتایج نشان می‌دهد که مدل یادگیری به‌دست‌آمده با دقت ۹۶/۶ درصد می‌تواند نوع پاراگراف را تشخیص دهد. افزون بر این، طبق نتایج به‌دست‌آمده، عملکرد طبقه‌بند در پایان‌نامه‌های علوم انسانی بهتر از سایر رشته‌های دیگر بوده است. به نظر می‌رسد که تنوع در ساختار پایان‌نامه‌ها و تفاوت نوشتارها در رشته‌های گوناگون موجب تفاوت در نتایج به‌دست‌آمده شده است. با توجه به یافته‌های به‌دست‌آمده و عملکرد خوب روش پیشنهادی، برای پژوهش‌های آینده می‌توان پاراگراف‌های فراداده را با هدف دستیابی به فراداده‌های از پیش تعریف‌شدهٔ پایان‌نامه‌ها تحلیل کرد.

قدردانی

بدین‌وسیله از حمایت آزمایشگاه متن‌کاوی و یادگیری ماشین «پژوهشگاه علوم و فناوری اطلاعات ایران» در انجام این پژوهش تشکر و قدردانی می‌شود.

فهرست منابع

تن‌سازان، امیر، و محمدامین مهدوی. ۱۳۹۶. استخراج فراداده‌های متنی از مقاله‌های علمی به زبان فارسی با مدل آماری CRF *پژوهش‌های نظری و کاربردی در علم اطلاعات و دانش‌شناسی* ۷ (۱): ۳۲۱-۳۰۴.

References

Adrian, W., N. Leone, M. Manna, & C. Marte. 2017. Document Layout Analysis for Semantic Information Extraction. In: Esposito F., Basili R., Ferilli S., Lisi F. (eds) AI*IA 2017 Advances in Artificial Intelligence. AI*IA 2017. Lecture Notes in Computer Science, vol 10640. Springer, Cham. https://doi.org/10.1007/978-3-319-70169-1_20

Boukhers, Z., S. Ambhore, & S. Staab,. 2019. An end-to-end approach for extracting and segmenting high-variance references from pdf documents. ACM/IEEE Joint Conference on Digital Libraries (JCDL). Champaign, IL, USA.

Cortes, C., & V. Vapnik. 1995. Support-vector networks. *Machine Learning* volume. 20297–273 :.

Cuong, N., M. Kumar, M.-Y. Kan, & W. Lee. 2015. Scholarly document information extraction using extensible features for efficient higher order semi-crfs. JCDL '15*:* proceedings of the 15th acm/ieee-cs joint conference on digital libraries. Knoxville, Tennessee, USA.

Do, H., M. Chandrasekaran, P. Cho, & M.–Y. Kan. 2013. Extracting and matching authors and affiliations inscholarly documents. JCDL '13: proceedings of the 13th acm/ieee-cs joint conference on digital libraries. Indianapolis, Indiana, USA

Ferrés, D., H. Saggion, F. Ronzano, & À. Bravo. 2018. PDFdigest: an adaptable layout-aware pdf-to-xml textual content extractor for scientific articles.11 th Language Resources and Evaluation Conference (LREC). Miyazaki, Japan

Kan, M.-Y., M. Luong, & T. Nguyen. 2010. Logical structure recovery in scholarly articles with rich document features. *International Journal of Digital Library Systems* 1 (4): 1-23.

Kern, R., K. Jack, M. Hristakeva, & M. Granitzer. 2012. TeamBeam — meta-data extraction from scientific literature. *D-Lib Magazine*.

Kooli, N., & A. Belaid. 2016. Inexact graph matching for entity recognition in OCRed documents. 23rd International Conference on Pattern Recognition (ICPR). Cancun, Mexico. IEEE.

Liu, R., L. Gao, D. An, Z. Jiang, & Z. Tang. 2017. Automatic document metadata extraction based on deep networks. *Natural Language Processing and Chinese Computing*. Lecture Notes in Computer Science, 10619305-317 :

Meng, B., L. Hou, E. Yang, & J. Li. 2018. Metadata extraction for scientific papers. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. CCL 2018, NLP-NABD 2018. Lecture Notes in Computer Science, 11221: 111-122

Nasar, Z., S. Jaffry, & M. Malik. 2018. Information extraction from scientific articles: a survey. *Scientometrics* 1171990–1931 :(3) .

Peng, F., & A. McCallum. 2006. Information extraction from research papers using conditional random fields*. Information Processing & Management* 42963-979 :(4) .

Piskorski, J., & R. Yangarber. 2013. Information extraction: past, present and future. *In Multi-source, Multilingual Information Extraction and Summarization*. Berlin: Springer.

Qiu, S., & T. Zhou.2019 . A method of extracting metadata information in digital books. 10th International Conference on Information Technology in Medicine and Education (ITME). Qingdao, China.

Rizvi, S., A. Dengel, & S. Ahmed 2019. DeepBiRD: An automatic bibliographic reference detection approach.

Safder, I., S.-U. Hassan, A. Visvizi, T. Noraset, R. Nawaz, & S. Tuarob2020 .. Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. *Information Processing & Management* 57102269 :(6) .

Souza, A., V. Moreira, & C. Heuser2014 .. ARCTIC: metadata extraction from scientific papers in pdf using two-layer CRF. DocEng '14: Proceedings of the 2014 ACM symposium on Document engineering. New York, NY, USA.

Tharwat, A., A. Hassanien, & B. Elnaghi2017 .. A BA-based algorithm for parameter optimization of Support Vector Machine. *Pattern Recognition Letters* 9313-22 :.

Tkaczyk, D., P. Szostek, M. Fedoryszak, P. Dendek, & Ł. Bolikowski2015 .. CERMINE: automatic extraction of structured metadata from scientific literature. *International Journal on Document Analysis and Recognition (IJDAR)* 18: 317–335.

Yang, X.-S. 2010. A new metaheuristic bat-inspired algorithm. *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)* 284: 65-74.

محدثه رهنما

متولد سال ۱۳۷۴، دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی در دانشگاه الزهرا (س) است.

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

سید محمدحسین هاشمی‌نژاد

متولد سال ۱۳۶۴، دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی نرم‌افزار از دانشگاه تربیت مدرس است. ایشان هم‌اکنون استادیار دانشگاه الزهرا (س)، گروه مهندسی کامپیوتر است. موضوعات معماری نرم‌افزار، یادگیری ماشین و الگوریتم‌های تکاملی 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 گروه زبان‌شناسی رایانشی پژوهشکدهٔ علوم اطلاعات و مدیر آزمایشگاه متن‌کاوی و یادگیری ماشین در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

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