-- Journal of Knowledge Retrieval and Semantic Systems---
Vol. 9, No. 32, Autumn 2022, PP.191-219.
jks.atu.ac.ir
DOI: 10.22054/jks.2021.59944.1425

An Overview of Automatic Text Classification

Reza Dehkhodaie Ph.D Student in Library and Information Science, Knowledge Management, Tarbiat Modares University, Tehran, Iran.

Atefeh Sharif* Assistant Professor, Library and Information Science Dept, Faculty of Management, Tarbiat Modares University, Tehran, Iran.

Abstract

Nowadays, various online resources are growing and disseminating rapidly. In order to organize these resources, attempts have been made to use automatic classification, which often uses statistical algorithms and machine learning. Recently, attention has been drawn to the use of library classifications. The main challenge here is that classification is an abstract, thought-provoking process, and machine techniques and artificial intelligence have not yet been able to completely replace the human mind. In this paper, we provide an overview of the importance of automatic classification, machine learning, and practical algorithms and techniques of clustering and classification like K-nearest neighbor, Bayesian models, artificial neural networks, deep learning, and hybrid classifications. Also, the steps of automatic classification of web pages and the techniques used in each step were mentioned. Achieving a clearer understanding of automatic classification will enable LIS experts to communicate with experts in the field of artificial intelligence and computers. This could pave the way for interdisciplinary research.

Keywords: Automatic Text Classification, Classification, Library Classification, Machine Learning, Web Page Classification.

* Corresponding Author: atefehsharif@gmail.com
How to Cite: Dehkhodaie, Reza. (2022). An Overview of Automatic Text Classification. *Knowledge Retrieval and Semantic Systems*, *9*(32), 191-219.

eISSN: 2783-1795 ISSN:2980-8243 Accepted: *September 16, 2022* Received: *April 6, 2022* Original Research

-- فصلنامه علمی بازیابی دانش و نظام‌های معنایی---------
دوره ۹، شماره ۳۲، پاییز، ۱۴۰۱، ص ۱۹۱ تا ۲۱۹
jks.atu.ac.ir
DOI: 10.22054/JKS.2021.59944.1425

مقاله پژوهشی
تاریخ دریافت: ۱۴۰۱/۰۱/۱۷؛ تاریخ پذیرش: ۱۴۰۱/۰۶/۲۵
ISSN:2980-8243
eISSN:2783-1795

مروری بر رده‌بندی خودکار متن

رضا دهخدایی
دانشجوی دکتری علم اطلاعات و دانش‌شناسی، گرایش مدیریت دانش، دانشگاه تربیت مدرس، تهران، ایران

عاطفه شریف *
استادیار، گروه علم اطلاعات و دانش‌شناسی، دانشگاه تربیت مدرس، تهران، ایران

چکیده

منابع با سرعت بسیار زیادی در حال رشد و انتشار هستند و در این میان سهم منابع دیجیتال و وبی بسیار مشهود است. به منظور سازماندهی این منابع، تلاش‌هایی برای رده‌بندی خودکار صورت گرفته که غالباً از الگوریتم‌های آماری و یادگیری ماشینی استفاده می‌کنند. همچنین در برخی منابع، استفاده از رده‌بندی‌های کتابخانه‌ای نیز توصیه شده است. اصلی‌ترین چالشی که در این زمینه وجود دارد آن است که رده‌بندی، فرآیندی انتزاعی و نیازمند تفکر است و روش‌های ماشینی و هوش مصنوعی هنوز نتوانسته‌اند به طور کامل جایگزین ذهن انسان شوند. در این مقاله ضمن بیان اهمیت رده‌بندی خودکار به مفاهیم یادگیری ماشینی و روش‌ها و الگوریتم‌های پرکاربرد در خوشه‌بندی و رده‌بندی مانند کا- نزدیکترین همسایه، مدل بیز، شبکه‌های عصبی مصنوعی، یادگیری عمیق، و طبقه‌بندی‌های ترکیبی پرداخته شد. همچنین مراحل رده‌بندی خودکار صفحات وب و روش‌های مورد استفاده در هر مرحله مورد اشاره قرار گرفت. رسیدن به درک روشن‌تری از موضوع رده‌بندی خودکار، امکان هم‌زبانی با متخصصان حوزه هوش مصنوعی و کامپیوتر را فراهم آورده و زمینه‌ساز پژوهش‌های میان‌رشته‌ای خواهد بود.

کلیدواژه‌ها: رده‌بندی، رده‌بندی خودکار متن، رده‌بندی صفحات وب، رده‌بندی کتابخانه‌ای، یادگیری ماشین.

* نویسنده مسئول: atefehsharif@gmail.com

مقدمه

امروزه به سبب رشد منابع اطلاعاتی در محیط وب، با انبوهی از اطلاعات و منابع روبه‌رو هستیم. اگرچه موتورهای کاوش در بازیابی اطلاعات عملکرد مطلوبی دارند، اما در سازماندهی با محدودیت‌هایی مواجه هستند (Choi & Yao, 2005). توجه به این امر می‌تواند کاربران را در جست‌وجویی موثر یاری نماید؛ زیرا کاربران با محدود کردن نتایج جست‌وجو نیز با انبوهی از نتایج روبه‌رو هستند که معمولاً فقط چند نتیجه اول را موردتوجه قرار می‌دهند.

علاوه بر این، بازیابی مؤثر تنها با جست‌وجوی کلیدواژه‌ای از طریق موتورهای کاوش امکان‌پذیر نیست. یکی از راه‌حل‌های این مشکل، رده‌بندی صفحات وب است. صفحات وب به وسیله پیوندشان با دیگر اسناد و نتایج جست‌وجو و همچنین قالبی نیمه ساختاریافته در فرمت اچ تی ام ال شناخته می‌شوند (Lassri et al, 2019). در این موارد، استفاده از یک رده‌بندی موضوعی مطابق طرح‌های استاندارد می‌تواند به بهبود بازیابی کمک کند و کاربران را به سمت مرور موضوعی منابع هدایت نماید. تفاوت جست‌وجو و مرور در آن است که زمانی که کاربر آدرس مشخصی دارد از طریق آن به جست‌وجو می‌پردازد. اما در مواقعی که کاربر تعریف مشخصی از نیاز خود ندارد، در بین منابع موضوعی به مرور منابع می‌پردازد. اما انجام این امر با توجه به حجم تولید منابع به صورت روزافزون امری دشوار و هزینه‌بر است. در چنین شرایطی استفاده از طرح رده‌بندی خودکار برای اختصاص خودکار اسناد متنی به گروه‌های از پیش تعیین شده می‌تواند نقش مفیدی را ایفا نماید.

نکته حائز اهمیت این است که رده‌بندی منابع دیجیتال امری حیاتی است، به این دلیل که به علت نبود نسخه فیزیکی، گم شدن منابع دیجیتال به معنای از بین رفتن آن‌ها است. از این‌رو، وجود رده‌بندی سلسله مراتبی با فراهم‌آوردن امکان مشاهده موضوعات در سلسله مراتب خود و در کنار یکدیگر بازیابی اطلاعات و مدارک را تسهیل و سرعت می‌بخشد (اسماعیل‌پور، ۱۳۸۶). با این حال صفحات وب بسیار متنوع هستند به نحوی که شاید هیچ نظام رده‌بندی نتواند امید داشته باشد که تمام موضوعات را با جزئیات پوشش دهد Choi (& Yao, 2005). از طرفی به سبب حجم زیاد اطلاعات و همچنین به سبب رشد روزافزون

منابع و وقت‌گیر بودن این فعالیت و هزینه‌بر بودن آن، پژوهش‌ها به سمت استفاده از رده‌بندی خودکار سوق پیدا کرده است. اما باید توجه داشت که روش‌های هوش مصنوعی، یادگیری ماشینی و رده‌بندی خودکار، هنوز به طور کامل نتوانسته جایگزین ذهن انسان شود.

با این وجود، برای مدتی، پروژه‌های مختلفی برای ساخت راهنماهای وبی با هدف مکان‌یابی اطلاعات و به عنوان جایگزینی برای موتورهای جست‌وجوی وب اجرا شد. این راهنماها، صفحات وب را در یک سلسله مراتب موضوعی قرار می‌دادند تا کاربر بتواند با مرور سلسله مراتبی، اطلاعات موردنیاز خود را پیدا کند. علیرغم سادگی در جست‌وجوی مطالب توسط راهنماها، ویرایش، نگهداری و به‌روزرسانی آن‌ها با توجه به تولید روزافزون منابع و اطلاعات وبی امری دشوار است؛ زیرا اختصاص صفحات وب به عنوان فهرست راهنما منحصراً نیازمند وجود ویراستاران انسانی است (Stamou et al, 2006). راهنماهای گوگل و یاهو و همچنین دیموز[1] نمونه‌هایی از این پروژه‌ها بود که اکنون متوقف شده‌اند. اگر عنصر انسانی حذف شود و این امر به صورت خودکار انجام پذیرد، امکان تعلق گرفتن اشتباه یک سند یا صفحه به دسته‌ای اشتباه وجود دارد. اگر هم به صورت دستی انجام شود، در بهترین حالت بسیار زمان‌بر و هزینه‌بر است و غیر از آن بحث سلیقه و یا ناهم‌خوانی کلیدواژه مورد جست‌وجوی کاربر با کلیدواژه‌های تعیین شده برای هر دسته وجود دارد.

به‌طور کلی چنین برداشت می‌شود که پروژه‌های رده‌بندی خودکار به ویژه برای منابع وبی، در عمل با مشکلاتی روبه‌رو بوده و بعضاً شکست خورده و یا متوقف شده‌اند. اما در این میان این پرسش مطرح می‌شود که آیا باید این موضوع را پذیرفت و از امر رده‌بندی منابع وب دست کشید؟ تلاش‌های صورت گرفته در ارائه روش‌ها و بهینه‌سازی روش‌های رده‌بندی خودکار، رشد روزافزون منابع در محیط وب و نیازی که برای سازماندهی مؤثر منابع جهت دستیابی و بازیابی مؤثر اطلاعات توسط کاربران ایجاد می‌شود، نشان از پاسخ منفی به این پرسش است. رسالت مهم جامعه کتابداری و علم اطلاعات و به صورت خاص‌تر حوزه مدیریت اطلاعات و دانش، توجه به نیازهای کاربران است تا در مدیریت اطلاعات،

1. DMOZ

منابع مناسب را در زمان مناسب در اختیار کاربر مناسب قرار دهند. مشاهده می‌شود که در سال‌های اخیر، توجه صاحب‌نظران و متخصصان رده‌بندی به جایگاه رده‌بندی‌های کتابخانه‌ای و اهمیت آن‌ها در کاربردشان برای رده‌بندی خودکار جلب شده، اگر چه در میان منابع فارسی کمتر به موضوعات این چنینی پرداخته شده است.

سباستینی (۱۹۹۹)[1] در مقاله خود به بررسی یادگیری ماشینی در رده‌بندی خودکار متن و سه مشکل شامل ارائه اسناد، ساختار رده‌بندی و ارزیابی رده‌بندی پرداخته است. ایکونوماکیس و همکاران (۲۰۰۵)[2] با بررسی مباحث نظری در حوزه رده‌بندی خودکار به بررسی روند رده‌بندی متن با استفاده از روش‌های یادگیری ماشینی پرداختند. اسماعیل‌پور (۱۳۸۶) در نوشته خود مروری بر رده‌بندی خودکار، رویکردها و چالش‌های آن داشت و همچنین اشاره‌ای به پروژه‌های انجام شده، روندها و روش‌های مورد استفاده در رده‌بندی خودکار کرده است.

جورابچی و مهدی (۲۰۱۱)[3] رویکردی بر رده‌بندی خودکار منابع آرشیوی در کتابخانه‌ها و مخازن دیجیتال بر مبنای یک طرح رده‌بندی کتابخانه‌ای استاندارد یعنی دهدهی دیویی ارائه دادند و با استفاده از CiteSeer به ارزیابی آن پرداختند. دالال و زاوری (۲۰۱۱)[4] به شرح استراتژی‌های عمومی در رده‌بندی خودکار متن پرداختند و همچنین راه‌حل‌هایی برای مسائل عمده این حوزه یعنی پرداختن به متن بدون ساختار، مدیریت ویژگی‌های زیاد و انتخاب روش یادگیری مناسب مطابق با طرح رده‌بندی ارائه دادند. لاسری، بنلاهمار و تراگها (۲۰۱۹)[5] به بیان ویژگی‌های رده‌بندی صفحات وب و مروری بر الگوریتم‌های یادگیری ماشینی با هدف رده‌بندی صفحات وب پرداختند.

ماو، بالاکریشنان، رانا و راوانا(۲۰۲۰)[6] در مقاله خود با هدف ارائه یک دید از روندها و شکاف‌های پژوهش‌های مربوط به رده‌بندی متن با استفاده از ساختار و الگوهای تحلیل

1. Sebastiani, F
2. Ikonomakis, M., Kotsiantis, S., & Tampakas, V
3. Joorabchi, A., & Mahdi, A. E
4. Dalal, M. K., & Zaveri, M. A
5. Lassri, S, Benlahmar, H, Tragha, A
6. Maw, M., Balakrishnan, V., Rana, O., & Ravana, S. D

پژوهش‌ها به بررسی مشکلات و روش‌های حل مشکلات در این حوزه پرداختند. آن‌ها با استفاده از رهنمودهای پیترسن و همکاران (۲۰۰۷) به مطالعه نظام‌مند ۹۶ پژوهش از پنج پایگاه داده الکترونیکی از سال ۲۰۰۶ تا ۲۰۱۷ پرداختند و ۹ مشکل اصلی در پژوهش‌های این حوزه را شناسایی کردند. ایشان همچنین، تکامل پژوهش‌های حوزه رده‌بندی متن را طی ۱۲ سال مورد بررسی قرار دادند.

جمع‌بندی پیشینه‌ها نمایانگر آن است که در پژوهش‌های مختلف چه به صورت مروری و یا پژوهشی، به موضوع رده‌بندی خودکار و روش‌های آن از جنبه‌های مختلف پرداخته شده است. با این حال این موضوع در بین منابع داخلی، به میزان کافی مورد اشاره قرار نگرفته است. از این جهت، در این نوشتار تلاش می‌شود ضمن مرور موضوع رده‌بندی خودکار، انواع روش‌های خودکار رده‌بندی با تأکید بر صفحات وب بیان شوند تا درک روشن‌تری از روش‌های رده‌بندی خودکار در میان متخصصان علم اطلاعات و دانش‌شناسی حاصل شود.

رده‌بندی خودکار متن[1] و طرح‌های رده‌بندی کتابخانه‌ای

با گسترش کتابخانه‌ها و مخازن دیجیتالی در جوامع دانشگاهی به منظور تسهیل انتشار مؤثر برون‌دادهای علمی در قالب‌های مختلف و به صورت الکترونیک، استفاده از فراداده‌ها برای منابع داده‌ای بدون ساختار می‌تواند به توصیف، مکان‌یابی، بازیابی و دسترسی کارآمد به منابع دیجیتال بیانجامد. استفاده از داده‌کاوی و فرآیندهای کشف دانش با هدف شناسایی محتواها طبق رده‌بندی‌های استاندارد و تاکسونومی در کتابخانه‌ها و محیط دیجیتال امری ضروری به نظر می‌رسد (Joorabchi & Mahdi, 2011). همچنین رشد چشمگیر متون دیجیتال در محیط وب، محققان را به سمت تمرکز بر فعالیت‌هایی چون دسته‌بندی، خوشه‌بندی و رده‌بندی متن‌ها به صورت دستی و خودکار و در کلاس‌های خاص سوق داده است.

اما این موضوع با چالش‌هایی نظیر پیچیدگی زبان‌های طبیعی و قالب متن بدون ساختار روبه رو است و رده‌بندی دستی بر روی میلیاردها سند متنی به دلیل زمان‌بر و هزینه‌بر بودن

1. Automatic Text Classification (ATC)

غیرممکن است (Maw et al, 2020). رده‌بندی خودکار متن از زمان ایجاد منابع و اسناد دیجیتال همیشه امری مهم، کاربردی و تحقیقاتی بوده و امروزه به دلیل مقادیر بسیار زیادی از اسناد متنی، یک ضرورت است (Ikonomakis et al, 2005). همچنین رده‌بندی در عمل با چالش‌هایی مواجه است. این چالش‌ها به قرار زیر است: (Dalal & Zaveri, 2011).

رده‌بندی متون بدون ساختار. برخی از اسناد مانند مقالات علمی پژوهشی در قالب‌های از پیش تعیین شده‌ای نوشته می‌شوند که رده‌بندی آن‌ها را آسان می‌کند. اما بیشتر اسناد به روشی بدون ساختار نوشته می‌شوند، لذا رده‌بندی باید براساس ویژگی‌هایی مانند وجود یا نبود کلمات کلیدی و تعداد دفعات وقوع آن‌ها انجام شود.

مدیریت تعداد زیادی ویژگی. انتخاب ویژگی‌های مفید با استفاده از روش‌های پیش‌پردازش آماری و معنایی شامل کلمات ساده، کلمات کلیدی مشخص شده یا استخراج شده توسط کاربر یا فراداده‌ها هستند.

بازیابی فراداده‌های مفید برای رده‌بندی. اطلاعات مربوط به فراداده در رده‌بندی مفید است. کلمات کلیدی، اسامی نظیر نام افراد و مکان‌ها، عنوان سند، نام نویسنده و غیره با استفاده از فراداده‌ها بیان می‌شوند که این امر در رده‌بندی کمک‌کننده است.

مدل‌سازی: انتخاب روش مناسب. برای مدل‌سازی و رده‌بندی، روش‌های مختلفی وجود دارد که آشنایی با آن‌ها و انتخاب مناسب‌ترین روش ضرورت دارد.

از اوایل دهه ۹۰ با پیشرفت حوزه یادگیری ماشینی، این حوزه با حوزه رده‌بندی خودکار متون مرتبط شد. درواقع الگوریتم‌های این حوزه با استفاده از رده‌بندی‌های دستی تابعی ایجاد می‌کنند که برای پیش‌بینی رده اسنادی که فاقد برچسب هستند، استفاده می‌شود. با این حال سیستم مذکور خالی از اشکال نیست و زمانی‌که تعداد طرح‌ها افزایش پیدا می‌کند، دقت آن‌ها کاهش می‌یابد. بعضی از این مشکلات نظیر سلسه مراتب‌های عمیق و بالا، توزیع‌های کج و به اصطلاح چوله[1] و داده‌های پراکنده هستند. درنتیجه تلاش‌های اخیر به سمت توسعه آمار و احتمالات مبتنی بر الگوریتم‌های یادگیری ماشینی جهت یافته است Joorabchi & (Mahdi, 2011).

1. skewed data distribution

توجه متخصصان به جایگاه طرح‌های کتابخانه‌ای نظیر دیویی و کنگره، درخصوص منابع وب نیز کاربرد یافت و در این زمینه پروژه‌های مختلفی نظیر Noradic WAIS/ World Wide Web با استفاده از رده‌بندی دهدهی جهانی، GERHARD با استفاده از رده‌بندی دهدهی جهانی، DESIRE با استفاده از اصطلاحنامه مهندسی و رده‌بندی دهدهی جهانی انجام شدند. مهم‌ترین آن‌ها را می‌توان پروژه اسکورپیون[1] دانست که در سال ۱۹۹۸ توسط او سی ال سی آغاز شد. در این پروژه ابتدا از نسخه الکترونیکی دیویی استفاده شد و در پروژه جدید، رده‌بندی کتابخانه کنگره استفاده شد (اسماعیل‌پور، ۱۳۸۶). اسکورپیون از روش خوشه‌بندی مبتنی بر فراوانی برای یافتن مرتبط‌ترین اسناد برای رده‌بندی استفاده می‌کند (Joorabchi & Mahdi, 2011). برای تورق بهتر در منابع اینترنت، روش‌هایی نظیر خوشه‌بندی و تکنیک‌های مبتنی بر محتوا یا مبتنی بر استناد وجود دارد. این ویژگی‌ها را در شبکه‌های عصبی مصنوعی می‌توان یافت که گره‌ها در آن نمایانگر عناصر بازیابی اطلاعات نظیر کلیدواژه، نویسنده و غیره هستند، و پیوندهای موجود برای انتقال ورودی از لایه‌ای به لایه دیگر استفاده شده و درنهایت، خروجی شبکه، منجر به بازیابی مدرک می‌شود (اسماعیل‌پور، ۱۳۸۶).

رده‌بندی متن بخش مهمی از متن‌کاوی و از حوزه‌های حیاتی پژوهش در پردازش زبان طبیعی[2] است و به عنوان فرایندی تعریف می‌شود که یک سند را براساس محتوای متنی و ویژگی‌های استخراج شده آن به یک یا مجموعه‌ای از دسته‌های از پیش تعریف شده اختصاص می‌دهند و این فرآیند شامل چهار مرحله است: مرحله پیش پردازش/ نمایش اسناد، استخراج ویژگی، انتخاب ویژگی/ تبدیل ویژگی و درنهایت مرحله آموزش و یادگیری / رده‌بندی (Maw et al, 2020). رده‌بندی خودکار، فرآیند اختصاص یک سند یا متن به مجموعه‌ای از کلاس‌های از پیش تعریف‌شده به صورت خودکار و با استفاده از روش‌هایی نظیر یادگیری ماشین است. رده‌بندی متن بر اساس کلمات یا ویژگی‌های مهم متن اسناد انجام

1. Scorpion
2. Natural Language Processing (NLP)

می‌شود و چون کلاس‌ها و رده‌ها از قبل تعریف شده‌اند، این فرآیند از نوع یادگیری ماشینی تحت نظارت است (Dalal & Zaveri, 2011).

روش‌های خاصی با بررسی عملکرد، سرعت و قابلیت استفاده، در نظر گرفته شده که متداول‌ترین آن‌ها شامل نمایش بر اساس فراوانی یک کلمه یا کلمات و داده‌های آموزش با مدل‌های مختلف نظیر ماشین بردار پشتیبانی[۱]، قضیه بیز[۲]، درخت تصمیم[۳] و غیره است. اما مشکلات اساسی که در روش‌های سنتی یادگیری ماشینی وجود داشت که منجر به ظهور روش‌های جدیدتر و موثرتر شد. به عنوان مثال، یک چالش در پردازش متن، پرداختن به ساختارهای معنایی و نحوی زبان‌های انسانی است. لذا رویکردهای یادگیری عمیق مانند شبکه‌های عصبی تکرار شونده[۴] و شبکه‌های عصبی حلقوی[۵] پدید آمدند که نتایج دقیق‌تری را بدون نیاز به دانش معنایی قبلی زبان‌های خاص ارائه می‌دهند (Maw et al, 2020).

مدتی است توجه دانشمندان به اهمیت و قابلیت طرح‌های رده‌بندی کتابخانه‌ای در استفاده برای رده‌بندی خودکار متون جلب شده است. به عقیده گولوب[۶] (۲۰۰۶)، یی[۷] (۲۰۰۷) و مارکی (۲۰۰۶)[۸]، رویکردی ضعیف‌تر برای رده‌بندی خودکار متن[۹] وجود دارد که به جامعه علوم کتابداری نسبت داده می‌شود، کمتر به الگوریتم‌ها و بیشتر به استفاده از واژگان کنترل شده جامع، نظیر طرح‌های رده‌بندی کتابخانه‌ای و اصطلاحات کنترل شده[۱۰] در رده‌بندی دستی منابع کتابخانه‌ای متمرکز است. دو نمونه رایج در کتابخانه‌ها استفاده از رده‌بندی دهدهی دیویی و رده‌بندی کتابخانه کنگره است (Joorabchi & Mahdi, 2011). هر نوع نظم‌دهی به مدارک نوعی رده‌بندی محسوب می‌شود. اما در معنایی خاص‌تر، سازماندهی نظام‌مند مدارک براساس موضوع، رده‌بندی کتابخانه‌ای محسوب می‌گردد (برومند، ۱۳۸۱).

1. Vector Machine (SVM)
2. Naïve Bayes (NB)
3. decision trees
4. Recurrent Neural Networks (RNN)
5. Convolutional Neural Networks (CNN)
6. Golub, K
7. Yi, K.
8. Markey, K.
9. Automatic Text Classification (ATC)
10. controlled vocabularies

استفاده از طرح‌های رده‌بندی کتابخانه‌ای می‌تواند تمام دانش بشری را پوشش دهد. این رویکرد در سیستم‌های رده‌بندی خودکار متن دو دسته را شامل می‌شود: ۱) سیستم‌های مبتنی بر انطباق رشته‌ای: این سیستم‌ها مبتنی بر الگوریتم‌های یادگیری ماشینی[1] نیستند، بلکه در عوض از روش تطبیق رشته به رشته بین کلمات موجود در سیاهه اصطلاحات اصطلاحنامه و طرح‌های رده‌بندی و کلمات موجود در متن رده‌بندی استفاده می‌کنند. در اینجا، سند ورودی بدون برچسب، به عنوان یک درخواست جست‌وجو در برابر طرح‌های رده‌بندی کتابخانه و اصطلاح‌نامه قرار می‌گیرند و نتیجه این جست‌وجو، رده موردنظر را برای آن مدرک نشان می‌دهد. نمونه بارز این طرح پروژه اسکورپیون است. ۲) سیستم‌های مبتنی بر یادگیری ماشینی: این سیستم‌ها از الگوریتم‌های یادگیری ماشینی برای رده‌بندی مطابق با طرح‌های رده‌بندی کتابخانه‌ای نظیر دهدهی دیویی و رده‌بندی کتابخانه کنگره استفاده می‌کنند. نتایج پژوهش گلوب و همکاران[2] نشان داد که روش دوم موثرتر عمل می‌کند (Joorabchi, & Mahdi, 2011). از این‌رو، به دلیل کاربرد گسترده و اساسی یادگیری ماشینی در بحث رده‌بندی خودکار، در ادامه به موضوع یادگیری ماشینی اشاره می‌شود.

یادگیری ماشینی

یادگیری ماشینی نوعی الگوسازی از روی داده‌ها است که برای پیش‌بینی و یا استخراج دانش از داده‌ها به کار می‌رود. تکنولوژی‌های یادگیری ماشینی در بینایی ماشین، پردازش صوت، پردازش زبان طبیعی، علوم عصبی، سلامتی و همچنین در حوزه اینترنت اشیا کاربرد چشم‌گیری یافته است. مراحل یادگیری ماشین شامل پیش‌پردازش، یادگیری و ارزیابی است. پیش‌پردازش شامل پاک‌سازی، استخراج، تغییر و ترکیب داده‌ها است تا ناسازگاری‌ها و نویزهای داده‌های خام از بین رفته و داده‌ها در شکلی مناسب جهت یادگیری به صورت ورودی اعمال شوند. در بخش یادگیری، الگوریتم‌هایی به کار گرفته می‌شود تا مدل‌سازی و یادگیری صورت گیرد (باغبانی، ۱۳۹۶).

1. machine learning
2. Golub et al

یادگیری ماشینی سه نوع الگوریتم دارد؛ شامل یادگیری نظارت‌شده، یادگیری نظارت‌نشده و یادگیری تقویتی. در یادگیری نظارت‌شده که اغلب سیستم‌ها از این نوع استفاده می‌کنند، سیستم سعی می‌کند تا از نمونه‌های قبلی یادگیری کند. یعنی الگوها را براساس مثال‌های داده شده فرا می‌گیرد. این روش شامل دسته‌بندی و رگرسیون است. در دسته‌بندی، متغیر به یک دسته از پیش‌تعیین‌شده تعلق می‌گیرد و برای مثال ایمیل به هرزنامه و غیرهرزنامه تعلق می‌گیرد. در رگرسیون، یک متغیر دارای یک مقدار حقیقی نظیر قد است. متغیرها در دسته‌بندی، گسسته و درگرسیون، پیوسته هستند (حصارکی، ۱۳۹۹).

در یادگیری نظارت‌نشده الگوریتم‌ها خود به دنبال ساختارهای موجود در داده‌ها هستند. این مورد شامل قوانین انجمنی و خوشه‌بندی است. در قوانین انجمنی هدف، کشف قواعدی است که بخش بزرگی از داده‌ها را توصیف کند. در ادامه مقاله، موضوع خوشه‌بندی به صورت مفصل‌تری بحث شده است. نهایتاً در یادگیری تقویتی یک برنامه که با محیطی پویا در ارتباط است باید به هدف خاصی دست یابد. سپس بازخوردهایی با عنوان پاداش و تنبیه فراهم می‌کند و فضای حل مسئله را به دنبال آن هدایت می‌کند. درواقع ماشین می‌آموزد که تصمیمات مشخصی را که در معرض آزمون و خطا و در محیطی دائمی قرار دارد، اتخاذ نماید (حصارکی، ۱۳۹۹).

یکی از حوزه‌هایی که در بحث رده‌بندی پیوند نزدیکی با یادگیری ماشینی دارد و از الگوریتم‌های مورداستفاده در آن در بحث رده‌بندی به وسیله یادگیری ماشینی نیز استفاده می‌شود، داده‌کاوی است. عمده راهکارهای داده‌کاوی که منجر به مدل‌سازی داده‌ها و یادگیری می‌شود و می‌توان آن را در مباحث مربوط به رده‌بندی خودکار نیز یافت، شامل خوشه‌بندی[1] و رده‌بندی است.

خوشه‌بندی

خوشه‌بندی فرآیند گروه‌بندی مجموعه‌ای از داده‌ها است تا آن‌ها را در طبقاتی از نمونه‌های مشابه قرار دهند، به این صورت که داده‌های موجود در درون هر خوشه مشابه و داده‌های

1. clustering

موجود با دیگر خوشه‌ها متفاوت باشند. خوشه‌بندی و تحلیل آن کاربردهایی در تحلیل داده‌ها، پردازش تصاویر، تشخیص الگو و تحلیل‌های تجاری دارد. موضوع اصلی در خوشه‌بندی شناسایی تشابه و عدم تشابه نمونه‌ها است تا نمونه‌های مشابه در یک گروه قرار و نمونه‌های غیرمشابه در خوشه‌های متفاوت قرار گیرند (اسماعیلی، ۱۳۹۱).

رده‌بندی

رده‌بندی و تخمین، دو شکل از تحلیل داده‌ها هستند که می‌توان با کمک آن‌ها به استخراج مدلی جهت توصیف داده‌ها و استفاده از آن‌ها در فرآیندهای بعدی پرداخت. این روش که جزو روش‌های یادگیری با نظارت است در پی آن است تا با برقراری ارتباطی میان داده‌ها به ارائه مدلی از این ارتباطات بپردازد، تا بعد از آن هر مفهومی که در چارچوب این ارتباطات قرار گرفت را در یک گروه قرار دهد. این بخش را از آن جهت یادگیری با نظارت می‌نامند که یک فرآیند دو مرحله‌ای است. در مرحله اول، داده‌های آموزش ایجاد می‌شوند که در آن‌ها برچسب کلاس تمام نمونه‌ها مشخص است و براساس آن مدل ساخته می‌شود. یادگیری در این مرحله انجام می‌شود. در مرحله دوم با استفاده از داده‌های آزمایش که در آن برچسب کلاس مشخص نیست به ارزیابی مدل می‌پردازند (اسماعیلی، ۱۳۹۱).

مدل‌سازی داده‌ها برای رده‌بندی خودکار چه به صورت رده‌بندی انجام شود و چه خوشه‌بندی دارای روش‌ها و الگوریتم‌های ویژه‌ای است.

درنهایت در بخش ارزیابی که مرحله سوم یادگیری ماشینی است، کارایی مدل در بازده آن با توجه به یادگیری شده‌ها، سنجیده می‌شود (باغبانی، ۱۳۹۶). به عنوان مثال در رده‌بندی سنجیده می‌شود که مدل ایجاد شده چقدر در انجام رده‌بندی موفق خواهد بود. آشنایی با روش‌های به کار رفته در خوشه‌بندی و رده‌بندی به درک بهتری از این فرایندها می‌انجامد، لذا در ادامه مورد بحث قرار گرفته است.

روش‌ها و الگوریتم‌های خوشه‌بندی و رده‌بندی

به صورت کلی الگوریتم‌های خوشه‌بندی را در پنج گروه می‌توان قرار داد. این گروه‌ها شامل خوشه‌بندی مبتنی بر فراوانی، خوشه‌بندی مبتنی بر افراز[1]، خوشه‌بندی سلسله مراتبی[2]، خوشه‌بندی مبتنی بر تراکم یا چگالی داده‌ها[3]، خوشه‌بندی مبتنی بر شبکه‌های شطرنجی گرید[4] و خوشه‌بندی مبتنی بر مدل[5] است (اسماعیلی، ۱۳۹۱). لازم به ذکر است که هر یک از این گروه‌ها می‌توانند مشتمل بر چندین الگوریتم باشند و موارد مذکور جزو اهم موارد است.

اما برای رده‌بندی، الگوریتم‌ها و مدل‌های مختلفی نظیر کا- نزدیکترین همسایه، مدل بیز، ماشین‌های بردار پشتیبان و غیره وجود دارد که در پروژه‌های مختلف انواع متفاوت آن استفاده شده است. به عنوان مثال چونگ و نو[6] در پروژه خود در حوزه اقتصاد، صفحات وب را در ۷۵۷ زیرگروه از طرح رده‌بندی دیویی قرار دادند و از الگوریتم کا- نزدیکترین همسایه استفاده کردند. پونگ و دیگران[7] سیستم رده‌بندی خودکار متن را مبتنی بر رده‌بندی کتابخانه کنگره توسعه دادند و از الگوریتم‌های بیز و کا- نزدیکترین همسایه بهره بردند. فرانک و پاینتر[8] از الگوریتم خطی ماشین‌های بردار پشتیبان برای رده‌بندی بیش از ۲۰ هزار منبع پژوهشی اینترنت مبتنی بر طرح کتابخانه کنگره استفاده نمودند. وانگ[9] از الگوریتم‌های بیز و ماشین‌های بردار پشتیبان برای رده‌بندی مجموعه داده‌ها بر مبنای رده‌بندی دهدهی دیویی استفاده نمود. البته روش‌های دیگری نیز برای رده‌بندی نظیر رده‌بندی مبتنی بر شروط، الگوریتم‌های ژنتیک، مجموعه‌های فازی و یا رگرسیون وجود دارد و همچنین هرکدام از این روش‌ها شامل انواع متفاوتی می‌توانند باشند؛ به‌طور مثال برای الگوریتم درخت تصمیم،

1. partitioning method
2. hierarchical
3. density-based
4. grid-based
5. model-based
6. Chung & Noh
7. Pong et al
8. Frank & Paynter
9. Wang

الگوریتم‌های مختلفی نظیر ID3, C4.5, CART و CHAID وجود دارد (اسماعیلی، ۱۳۹۱). در ادامه، تعدادی از الگوریتم‌ها و روش‌های مذکور که در پژوهش‌های مربوط به رده‌بندی خودکار بیشتر مورد استفاده قرار گرفته‌اند، مورد اشاره قرار می‌گیرند.

الف) کا- نزدیکترین همسایه: تمرکز این روش بر رده‌بندی نمونه‌ها بر مبنای همسایگی آن‌ها است و نمونه‌های مشابه بر مبنای برچسب هدف، برچسب‌گذاری می‌شوند. بنابراین رده‌بندی نمونه‌ها وابسته به برچسب هدف نقاط همسایه است. اگر برای یک نمونه بیش از یک همسایه وجود داشته باشد، از رأی‌گیری استفاده می‌شود و برچسب مربوط به همسایه دارای اکثریت آرا به عنوان برچسب نمونه جدید انتخاب می‌شود. برای اینکه مشخص شود دو نمونه چقدر به هم نزدیک هستند و کدام نمونه در نزدیک‌ترین همسایگی با نمونه مورد نظر وجود دارد، لازم است که میزان شباهت محاسبه شود. برای اینکار از توابع فاصله نظیر فاصله اقلیدسی، منهتن[1] و همینگ[2] استفاده می‌شود (Lassri et al, 2019). در این روش ابتدا مثال‌هایی به حافظه سپرده می‌شود و بعد از آن مشاهدات با آن مثال‌هایی که در حافظه موجودند مقایسه می‌شوند و بر مبنای نزدیک‌ترین همسایگی دسته‌بندی می‌شوند (باغبانی، ۱۳۹۶).

ب) ماشین‌های بردار پشتیبان: از روش‌های رده‌بندی تفکیکی است و تمرکز آن بر یافتن فرضیه‌ای برای تضمین کمترین خطای واقعی است. در این روش به مجموعه‌های آموزشی مثبت و منفی نیاز هست تا تصمیم‌هایی را جست‌وجو کند که داده‌های مثبت و منفی را در فضای چند سطحی جدا و همچنین اسنادی را با ورودی‌های چند سطحی کنترل نماید و ویژگی‌های نامربوط را از بین برد. با این حال، اشکال عمده این روش در پیچیدگی و صرف زمان و فضای زیاد در هنگام آموزش است (Lassri et al, 2019). اینگونه الگوریتم‌ها مدل‌های دسته‌بندی ایجاد می‌کنند. این الگوریتم به دنبال یک صفحه مافوق[3] و با نمایش داده‌ها در یک فوق فضا[4] به جداسازی بهتر کلاس‌ها کمک می‌کنند (باغبانی، ۱۳۹۶).

1. manhattan distance
2. hemming distance
3. hyper plane
4. hyper space

ج) قضیه بیز: این روش از یک رویکرد احتمالی برای رده‌بندی استفاده می‌کند. کلاسی که دارای بیشترین میزان احتمال باشد، برچسب نمونه را تعیین می‌کند. این قضیه این‌طور فرض می‌کند که ویژگی‌های ورودی از نظر آماری مستقل از هم هستند و بر این اساس هیچ ورودی بر ورودی دیگر تأثیر نمی‌گذارد. اما همین فرض استقلال به این علت که همیشه صادق نیست، از اشکالات این روش محسوب می‌شود (Lassri et al, 2019). بسته به اینکه ویژگی‌ها مستقل یا وابسته باشند از دو روش Naive یا Network استفاده می‌شود (باغبانی، ۱۳۹۶).

د) شبکه عصبی مصنوعی[1]: براساس شبکه‌های عصبی بیولوژیکی عمل می‌کنند که شامل گروه‌های به هم پیوسته نورون‌ها هستند و برای پردازش اطلاعات از رویکردی اتصال گرایانه برای محاسبات استفاده می‌کند. این روش یک سیستم انطباقی است و براساس اطلاعات ورودی یا خروجی شبکه در طول مرحله یادگیری، ساختار خود را تغییر می‌دهد Lassri et al, 2019). درواقع از ساختار نورون‌ها برای یادگیری جهت دسته‌بندی، پیش‌بینی و برچسب زدن داده‌ها با متصل کردن نورون‌های مصنوعی با توابع فعال‌ساز و تجمیع‌کننده (پرسپترون‌ها) در لایه و اتصال لایه به شبکه استفاده می‌کنند (باغبانی، ۱۳۹۶). این سیستم شامل مجموعه‌ای از گره‌ها و اتصالات میان آن‌ها است که شبکه‌ای را تشکیل می‌دهند و هر گره واحدی محاسباتی از شبکه است و می‌تواند بر روی ورودی‌ها پردازش انجام دهد (اسماعیلی، ۱۳۹۱).

ه) یادگیری عمیق[2]: برای رفع چالش اصلی هوش مصنوعی در شبیه‌سازی مغز انسان، یادگیری عمیق ویژگی‌ها را هنگام عبور از چندین لایه پنهان یاد می‌گیرد. شبکه‌های عمیق به دلیل انتقال اطلاعات آموخته شده از لایه‌های قبلی به لایه‌های آینده، ویژگی‌های انتزاعی سطح بالا را به تدریج یاد می‌گیرند. تفاوت یادگیری عمیق با شبکه عصبی معمولی در همین برقراری ارتباط با لایه‌های پیشین به دلیل عدم استقلال از آن‌هاست و لذا یادگیری بهتر انجام می‌شود.

1. Artificial Neural Network (ANN)
2. deep learning

و) درخت‌های تصمیم: برای تقسیم داده‌ها به مناطق خالص یعنی مناطقی که فقط یک کلاس نمونه دارند، مورداستفاده قرار گرفتند، زیرا با داده‌های واقعی، داشتن زیرمجموعه خالص امکان‌پذیر نیست. درواقع هدف این است که هر زیر مجموعه دارای نمونه‌ای از یک کلاس واحد باشد. مرزهایی که موجب جدایی این زیرمجموعه‌ها می‌شوند را مرز تصمیم‌گیری می‌نامند (Lassri et al, 2019). درخت تصمیم از تعدادی گره و شاخه تشکیل شده است که درآن هر گره یا برگ بیانگر کلاس‌ها هستند و در هریک از گره‌های غیر برگ، تصمیم‌گیری براساس یک یا چند صفت خاصه صورت می‌گیرد (اسماعیلی، ۱۳۹۱).

ز) طبقه‌بندی‌های ترکیبی[1]: برای بهبود عملکرد سیستم از چندین روش یادگیری ماشینی استفاده می‌گردد. این رویکردها معمولاً از دو جزء عملکردی استفاده می‌کنند. در مرحله اول داده‌های خام به عنوان ورودی گرفته می‌شود و نتایج متوسطی ایجاد می‌کند. سپس در مرحله دوم این نتایج متوسط به عنوان ورودی درنظر گرفته می‌شود و نتایج نهایی ارائه می‌گردد. این روش‌ها مبتنی بر رده‌بندی‌های آبشاری[2] مختلف است. همچنین می‌توان از خوشه‌بندی برای پردازش ورودی‌ها استفاده نمود تا نمونه‌های غیرنماینده را از هر کلاس حذف شود و یا برای طراحی رده‌بندی از نتایج خوشه‌بندی به عنوان نمونه‌های آموزشی استفاده گردد. لذا در سطح اول می‌توان براساس روش‌های یادگیری تحت نظارت یا بدون نظارت عمل کرد. همچنین این رده‌بندی‌ها می‌توانند دو روش مختلف را برای بهینه‌سازی در جهت یادگیری و همچنین پیش‌بینی استفاده نمایند (Lassri et al, 2019).

ح) رده‌بندی کننده‌های کلی[3]: این اصطلاح به ترکیب چندین الگوریتم یادگیری ضعیف یا یادگیرندگان ضعیف اشاره داد و برای بهبود عملکرد رده‌بندی کننده‌های منفرد ایجاد شدند. یادگیرندگان ضعیف در گروه‌های آموزشی تعلیم دیده‌اند تا در عملکرد کلی، مؤثرتر عمل نمایند. یکی از رایج‌ترین موارد برای این بخش روش اکثریت آرا است (Lassri et al, 2019).

1. hybrid classifiers
2. cascading different classifiers
3. ensemble classifiers

در پایان باید متذکر شد که هیچ روش واحدی برای انواع رده‌بندی برتر از سایر روش‌ها نیست. مثلاً روش بیز براساس فرض استقلال مشروط در میان صفات است. این روش اگر تعداد کافی نمونه آزمایش از هر گروه وجود داشته باشد، رده‌بندی احتمالی از یک سند متنی را ارائه می‌دهد. با این حال اجرای آن ساده و زمان یادگیری کمتر است، اما عملکرد آن برای دسته‌هایی که با ویژگی‌های بسیار کمی تعریف شده‌اند مناسب نیست Dalal, & Zaveri, 2011). مزیت الگوریتم شبکه عصبی در آن است که به جای دیکته کردن دستورات به یک سیستم از طریق برنامه‌نویسی، خود سیستم برای ارائه پاسخ مناسب به اتفاقات آموزش می‌یابد و البته این روش بیشتر در حوزه زیست‌شناسی کاربرد دارد (تیم پژوهش راهبرد، ۱۴۰۰).

اما درخت تصمیم برخلاف روش بیز در بین ویژگی‌های خود استقلال را درنظر نمی‌گیرد. در نمایش درخت تصمیم رابطه بین ویژگی‌ها به عنوان پیوند ذخیره می‌شود و هنگامی که تعداد نسبتاً کمتری از ویژگی وجود دارد، درخت تصمیم به عنوان روشی برای رده‌بندی متن مورد استفاده قرار می‌گیرد، با این وجود مدیریت آن برای تعداد زیادی از ویژگی‌ها دشوار می‌شود. محققان با ترکیب بعضی روش‌ها، دقت رده‌بندی را بهبود بخشیده‌اند (Dalal & Zaveri, 2011). درخت‌های تصمیم به صورت خاص در تحلیل تصمیمات و در جهت انتخاب مناسب‌ترین استراتژی کاربرد دارند و از آنجا که جنبه بصری دارند، مطالعه آن‌ها راحت‌تر است. ماشین‌های بردار پشتیبان بهترین دسته‌بندی و تفکیک خطی میان داده‌ها را مشخص می‌کنند و یک رده‌بندی قوی برای تفکیک داده‌ای پیچیده محسوب می‌شوند (تیم پژوهش راهبرد، ۱۴۰۰). درنهایت مزیت روش کا- نزدیکترین همسایه در آن است که باعث می‌شود اطلاعات و یا صفحات جدید به نزدیک‌ترین دسته شبیه به خود تعلق گیرند و با این حال این چالش وجود دارد که انطباق کامل وجود نداشته باشد.

با معرفی روش‌های مذکور باید متذکر شد که انتخاب از بین آن‌ها تابع سیاست‌ها، حوزه موضوعی و مباحث میدانی است و همان‌طور که در ابتدا ذکر شد نمی‌توان گفت که روشی

بر دیگری برتری دارد. اکنون با معرفی روش‌ها و الگوریت‌های مذکور، با توجه به اینکه رده‌بندی خودکار برای متون و خصوصاً صفحات وب کاربرد مهمی دارد، در ادامه به این موضوع پرداخته می‌شود.

رده‌بندی خودکار صفحات وب

رده‌بندی خودکار به ویژه برای صفحات وب به چند دلیل ضرورت دارد؛ ۱) اطلاعات زیاد موجود رده‌بندی دستی را برای متخصصان انسانی دشوار می‌کند، ۲) به تخصص‌های زیادی نیاز دارد، ۳) ماهیت پویا و ناپایدار منابع دانش به‌خصوص صفحات وب و ۴) زمان و هزینه بیشتر موردنیاز (Selvakuberan et al, 2008).

خوشه‌بندی امکان بازیابی اسناد مربوطه را افزایش می‌دهد. با این کار اسناد مشابه در گروه‌هایی قرار می‌گیرند که توسط یک رأس نمایندگی می‌شوند. در این روش، پرسش‌های کاربران با گروه‌ها مقایسه شده و مشابه‌ترین گروه به عنوان مرتبط‌ترین نتیجه انتخاب می‌شود. مزیت رده‌بندی خودکار و خوشه‌بندی این است که اسناد یک گروه ممکن است لزوماً حاوی اصطلاحات مورداستفاده در پرس‌وجوی کاربر نباشند، اما درواقع موردنیاز کاربر هستند. همچنین با این روش‌ها می‌توان از مشکلات ناشی از مترادف یا هم‌آوائی جلوگیری کرد (Eito-Bru, 2014).

رده‌بندی صفحات وب به دنبال تشخیص این موضوع است که یک صفحه وب به چه دسته یا دسته‌هایی تعلق دارد. در مواردی که یادگیری از طریق صفحات وب مطرح می‌شود، این موضوع در زمینه یادگیری ماشینی قرار می‌گیرد. در این زمینه باید در نظر داشت که صفحات وب، اسناد نیمه ساختاریافته‌ای هستند که در قالب اچ تی ام ال نوشته می‌شوند، سپس از طریق پیوند در ارتباط قرار می‌گیرند، اغلب کوتاه هستند و استفاده از متن آن‌ها برای تجزیه‌وتحلیل‌ها مناسب نیست و منابع آن‌ها متعدد، ناهمگن و به طور پویا درحال تغییر است (Choi, & Yao, 2005). از آنجا که صفحات وب در قالب اچ‌تی‌ام‌ال و به صورت نیمه ساختاری و حاوی برچسب، قالب و غیره هستند، لازم است قبل از استفاده از روش‌های متن کاوی در جهت رده‌بندی، مقداری پیش پردازش در آن‌ها انجام شود (Selvakuberan et al, 2008).

به طور کلی، مرور منابعی که در ادامه به آن‌ها اشاره خواهد شد، نمایانگر این است که رده‌بندی صفحات وب طی ۶ مرحله انجام می‌شود:

۱. مشخص شدن موضوع و کلیدواژه‌ها؛

۲. تشکیل بردار ویژگی؛

۳. پیش‌پردازش؛

۴. کاهش بعد؛

۵. تشکیل بردار سند برای استفاده به عنوان گروه آموزش برای رده‌بندی؛ و

۶. اختصاص کد رده‌بندی به مدرک

در مرحله اول باید موضوع و کلیدواژه‌های متن مشخص شوند. در این مرحله معمولاً برای تجزیه‌وتحلیل متن از روش‌ها و فنون نمایه‌سازی خودکار استفاده می‌شود. این روش‌ها سه نوع هستند.

الف) فایل‌های امضایی[1]: که یک نوع کددهی به مدارک هستند و الگوریتم‌های خاص خود را دارند و به وسیله کوتاه‌سازی لغات با استفاده از سیاهه‌های بازدارنده و ریشه‌یابی این کار را انجام می‌دهند.

ب) نمایه معکوس[2]: در یک پایگاه اطلاعاتی، همه رکوردها با کد خود در فایل اصلی ذخیره شده‌اند. به دلیل وقت‌گیر بودن جست‌وجو در این فایل به‌خصوص هنگامی که تعداد رکوردها زیاد باشد، برنامه‌نویسان، فایلی را ایجاد کرده‌اند که فیلدها را با شماره رکوردهای آن‌ها به صورت معکوس، دربر دارد. لذا برنامه به جای بررسی تک تک فایل‌ها، در فایل معکوس رکوردهایی را که شماره آن‌ها در برابر این کلمه قرار گرفته‌اند، بازیابی می‌کند (اسماعیل‌پور، ۱۳۸۶).

ج) البته در بعضی موارد از روش‌های ریشه‌یابی هم استفاده می‌شود، بدین معنا که ریشه کلمات جدا می‌شوند (Ikonomakis et al, 2005; Dalal & Zaveri, 2011). سپس بررسی می‌شود که کلمات انتخاب شده از کدام قسمت متن بازیابی شده‌اند. کلماتی که در برخی

1. signature files
2. inverted files

بخش‌های خاص نظیر عنوان، مقدمه، چکیده، عناوین فصول و بحث و نتیجه‌گیری یافت شوند، نسبت به دیگر قسمت‌های متن درجه اعتبار بیشتری دارند (اسماعیل‌پور، ۱۳۸۶).

مرحله دوم تشکیل بردار ویژگی است. ویژگی‌ها یا صفات شامل کلمات مهم یا عباراتی است که رخداد بالایی در متن دارند و نماینده متن هستند (Dalal & Zaveri, 2011). در این مرحله، صفحات وب که معمولاً از رشته‌های کاراکترها، فراپیوندها، تصاویر و برچسب‌های اچ‌تی‌ام‌ال تشکیل شده‌اند، به یک بردار ویژگی برای حذف اطلاعات کم اهمیت‌تر و استخراج ویژگی‌های برجسته تبدیل می‌شوند. دو رویکرد برای رده‌بندی صفحات وب شامل رویکرد موضوع‌محور[1] و ژانرمحور[2] ارائه شده است (Choi & Yao, 2005)

الف) در رویکرد موضوع‌محور می‌توان سلسه مراتب موضوعی برای صفحات وب و سپس جست‌وجوی مبتنی بر موضوع ایجاد کرد. برای مثال در یاهو، موضوعاتی که ذیل موضوع علم قرار می‌گیرند، می‌توانند شامل مواردی از قبیل کشاورزی، نجوم، شیمی، زیست‌شناسی و غیره باشد.

ب) رویکرد دیگر، مبتنی بر ژانر است (Choi & Yao, 2005). در رده‌بندی مبتنی بر ژانر یا رده‌بندی مبتنی بر ساختار، رده‌بندی بر اساس عوامل عملکردی و نوع متن انجام می‌شود؛ برای مثال انواع ژانرهایی که در این نوع وجود دارند، شامل فهرست محصولات، خریدهای آنلاین، تبلیغات، فراخوان مقاله و غیره هستند. این مورد برای یافتن علایق فوری به کاربران کمک می‌کند. ژانر، با درنظرگرفتن محتوا، فرم و هدف، نوعی کنش ارتباطی و با یک هدف ارتباطی جمعی و جنبه‌های مشترک فرمی تعریف شده است. دو ویژگی این نوع رده‌بندی شامل ویژگی‌های ارائه که نحوه ارائه وب را مشخص می‌کند و ویژگی‌های نحوی شامل ویژگی‌های پیچیدگی متن، ویژگی‌های سطح کاراکتر[3]، ویژگی‌های طرح و ویژگی‌های واژگانی است که به تجزیه‌وتحلیل کاربرد کلمات در قالب کلمات بازدارنده و کلمات کلیدی می‌پردازد (Choi & Yao, 2005). اسناد این نوع دارای قالب‌های مختلف، واژگان

1. subject-based
2. genre-based
3. character level features

متفاوت و اغلب سبک نوشتاری متفاوتی حتی برای اسناد داخل یک ژانر هستند و این به آن معناست که داده‌ها ناهمگن هستند (Ikonomakis et al, 2005).

در مرحله سوم پیش پردازش انجام می‌شود. در روش مبتنی بر موضوع پیش‌فرض این است که متن صفحات نمایانگر محتوا هستند و لذا برای بازیابی ویژگی‌های مهم متنی، ابتدا صفحات باید پیش پردازش شوند (Choi & Yao, 2005). قبل از رده‌بندی صفحات وب، ۸۰ درصد پیش پردازش نیاز است. پیش پردازش برای صفحات وب شامل مراحل زیر است:

الف) در ابتدا تگ‌های اچ تی ام ال وزن‌دهی و سپس حذف می‌شوند.

ب) بعد از آن واژگان بازدارنده حذف می‌شوند.

ج) سپس مطابق قانون لون[1]، کلماتی که رخداد آن‌ها در متن پایین‌تر از آستانه مشخص شده است. به دلیل اینکه تأثیر زیادی در متن ندارند حذف می‌گردند. د) درنهایت ریشه کلمات باقی مانده استخراج می‌شوند. یکی از موارد کاربردی در این زمینه الگوریتم پورتر[2] است.

پس از پیش‌پردازش در مرحله چهارم کاهش بعد انجام می‌شود. یک ویژگی رده‌بندی متن این است که تعداد ویژگی‌ها و یا کلمات یا عبارات منحصر به فرد می‌تواند بسیار باشد که این مورد، موانع بزرگی در استفاده از بسیاری از الگوریتم‌های پیشرفته یادگیری برای رده‌بندی متن ایجاد می‌کند، لذا برای رفع این مشکل از روش‌های کاهش بعد استفاده می‌شود (Ikonomakis et al, 2005). تا تحلیل‌ها را در فضای کوچک‌تری انجام دهند و کیفیت افزایش یابد. این‌گونه توابع دارای دو حالت هستند: ویژگی‌های انتخابی و استخراجی (Sebastiani, 1999). در روش انتخابی، درواقع ویژگی‌ها از مجموعه اصلی انتخاب می‌گردند. اما در روش استخراجی، انتخاب از مجموعه اصلی نیست و به نوعی واژه‌ها و ویژگی‌ها از جایی دیگر اختصاص داده می‌شوند. به صورت کلی، معیارهای انتخاب ویژگی را می‌توان به دو مجموعه تقسیم کرد: یک مجموعه فقط میزان رخداد یک ویژگی را نشان

1. Lune
2. Porter Stemmer

می‌دهد. مانند انتخاب ویژگی با استفاده از فراوانی سند[1]، اطلاعات متقابل[2]، آنتروپی متقابل[3] و نسبت شانس[4]. مجموعه دیگر تمام مقادیر ممکن یک ویژگی را شامل می‌شود از جمله انتخاب ویژگی با استفاده از کسب اطلاعات[5] و مربع آماری (Choi & Yao, 2005). به دلیل مشکلات موجود در ویژگی‌های ترکیبی نظیر هم‌سانی، هم‌آوائی و مترادفات، که موجب عدم بهینگی برای نمایش محتوای سند می‌شود، می‌توان از روش‌های استخراج ویژگی با ایجاد ویژگی‌های مصنوعی که فاقد مشکلات مذکور هستند، استفاده نمود.

برای استخراج ویژگی دو رویکرد وجود دارد: نمایه‌سازی معنایی نهفته[6] و خوشه‌بندی کلمات[7](Sebastiani, F, 1999) رویکرد نمایه‌سازی معنایی نهفته برای غلبه بر یک مشکل اساسی در روش‌های بازیابی به وجود آمده است، زیرا در این روش‌ها سعی بر آن است تا کلمات جست‌وجو را با کلمات اسناد مطابقت دهند. زمانی که موضوع بازیابی مفهومی توسط کاربران مطرح شود، یا اصطلاحات تحت اللفظی در پرسش کاربر با اصطلاحات یک سند مرتبط مطابقت نداشته باشد، یا زمانی که کلمات چندین معنی داشته باشند؛ مشکلاتی در روش‌های بازیابی معمول رخ می‌دهد. در روش مذکور، فرض بر این است که برخی ساختارهای معنایی نهفته اساسی در داده‌ها وجود دارد که با تصادفی بودن انتخاب کلمه با توجه به بازیابی، تا حدی پنهان می‌ماند. برای حدس این ساختارها و رفع نویزها از فنون آماری استفاده می‌شود. در نمایه‌سازی معنایی نهفته، یک ماتریس از ارتباطات اسناد و اصطلاحات ایجاد و یک فضای معنایی ساخته می‌شود که در آن اصطلاحات و اسناد مشابه در نزدیکی یکدیگر قرار می‌گیرند. از این‌رو، بازیابی با استفاده از اصطلاحات موجود در یک پرس‌وجو در یک نقطه از فضا شناسایی و اسناد موجود در آن بخش به کاربر نمایش داده می‌شود (Deerwester et al, 1990). در رویکرد خوشه‌بندی اصطلاحات و یا کلمات،

1. document frequency
2. mutual information
3. cross entropy
4. odd ratio
5. information gain
6. latent semantic indexing
7. word clustering

درواقع کلمات با درجه بالایی از رابطه معنایی زوجی با خوشه‌ها را مورد هدف قرار می‌دهد که در این روش ممکن است از خوشه‌ها (یا مرکز آن‌ها) به جای اصطلاحات به عنوان ابعاد فضای بردار استفاده شود (Sebastiani, 1999).

در مرحله پنجم پس از انتخاب ویژگی، متن سند به عنوان برداری برای سند نشان داده می‌شود و با استفاده از الگوریتم یادگیری ماشین گروه آموزش برای رده‌بندی ایجاد می‌گردد، سپس این گروه مورد آزمایش قرار می‌گیرند و اگر مشخص شود که دقت گروه آموزش دیده برای مجموعه آزمون شده قابل قبول است، این مدل برای رده‌بندی نمونه‌های جدید اسناد متنی استفاده می‌شود (Dalal & Zaveri, 2011).

در مرحله آخر، پس از مشخص شدن موضوع و کلیدواژه‌های متن، باید کد رده‌بندی به مدرک اختصاص یابد که این کار با استفاده از روش‌های آماری و برحسب اولویت انجام می‌گیرد. برای مثال مارسلا و ملتبای[1] (۲۰۰۰) بیان داشتند که اسکورپیون بر اساس یک پایگاه اطلاعاتی قابل جست‌وجو در فایل رده‌بندی دیویی ساخته شده است. از این رو، مدرکی که آماده دریافت کد رده‌بندی است، همانند یک پرسش در برابر سیستم رده‌بندی دیویی قرار می گیرد و کدهای رده‌بندی را به عنوان نتیجه جست‌وجو دریافت و برای بازیابی از نرم‌افزار اسمارت[2] استفاده می‌کند که برای این امر استفاده از اصطلاحنامه‌های وب‌محور در سازماندهی و بازیابی بهتر، می‌تواند مؤثر باشد (اسماعیل‌پور، ۱۳۸۶).

در یک رویکرد، رده‌بندی مبتنی بر پروفایل[3] انجام می‌شود. در این روش پس از انتخاب صفحات وب آموزشی، از الگوریتم‌های یادگیری ماشینی و روش‌های رده‌بندی برای رده‌بندی صفحات وب استفاده می‌شود. این مورد در دو مرحله آموزش و سپس آزمایش انجام می‌گیرد. در برخی موارد از یک مرحله اعتبارسنجی نیز استفاده می‌شود. برای این کار یک پروفایل و یا یک سری ویژگی‌ها را به عنوان نمونه‌هایی از دسته، از پیش تعریف شده و برای هر دسته از مجموعه‌ای از صفحات وب آموزشی را استخراج می‌کنند. پس از آموزش

1. Marcella & Maltby
2. SMART
3. profile based classifiers

همه دسته‌ها، صفحات وب جدید را رده‌بندی می‌نمایند. پس از آن، بردار ویژگی با مشخصات تمام دسته‌ها مقایسه و امتیازدهی می‌شود. اگر صفحه وب جدید با بیش از یک امتیاز در یک دسته از صفحات وب به بیش از یک گروه اختصاص یابد، صفحه جدید به دسته‌ای تعلق می‌گیرد که دارای بالاترین امتیاز است. نمونه‌هایی از رده‌بندی‌ها با استفاده از این روش، رده‌بندی روکچیو[1]، ماشین بردار پشتیبانی، رده‌بندی شبکه‌های عصبی و رده‌بندی لاینر لست اسکوار فیت[2] است (Choi & Yao, 2005). این نوع رده‌بندی نمایانگر مقوله‌ای صریح برای تصمیم‌گیری است. پس از ایجاد مجموعه آموزش، یادگیری با استخراج ویژگی‌ها از این مجموعه ایجاد می‌شود (Sebastiani, 1999). به عنوان مثال برای دسته‌های مختلف، پروفایلی تشکیل می‌شود و ویژگی‌هایی به آن تعلق می‌گیرد. برای مثال تعیین می‌گردد که افرادی که دارای یک جنسیت مشخص، محدوده‌ای از وزن و اندازه قد و ویژگی‌هایی از این قبیل هستند، در دسته مشخص شده قرار گیرند تا در کلاس خاصی قرار داده شوند.

در روشی دیگر برای رده‌بندی، رده‌بندی بر مبنای یادگیری نقش‌ها و رده‌بندی مبتنی بر یادگیری قاعده[3] انجام می‌شود که از بارزترین آن‌ها قوانین اگر- آنگاه[4] است. برای این نوع رده‌بندی، از صفحات وب آموزشی برای یک گروه از قوانین و برای توصیف این گروه استفاده می‌شود. قوانین منطبق بر اساس نتایج این قانون، به پیش‌بینی کلاس صفحات وب می‌پردازند. سه مثال برای این نوع، قانون نرمال انقطاعی[5]، قاعده انجمن[6] و درخت تصمیم[7] است (Choi & Yao, 2005). به عنوان مثال قواعدی برای دسته‌ها در نظر گرفته می‌شود به این صورت که برای ویژگی‌های انسانی، اگر قد بالای ۱۹۰ سانتی متر بود، آنگاه در رده بلند قدها قرار گیرد.

1. Rocchio
2. Linear Least Square Fit
3. rule learning based classifiers
4. if-then
5. disjunctive normal form rule
6. association rule
7. decision tree

نوع دیگر رده‌بندی، رده‌بندی مبتنی بر مثال مستقیم[1] است. این روش نمایشی صریح و واضح از گروه‌های مورد مطالعه ایجاد نمی‌کند و قضاوت در آن مبتنی بر دیگر گروه‌های مشابه است (Sebastiani,1999). برای این نوع رده‌بندی، یک صفحه وب رده‌بندی شده به عنوان پرسشی در برابر مجموعه‌ای از مثال‌های دسته‌بندی شده قرار می‌گیرد و سپس به دسته‌ای تعلق می‌گیرد که دارای بیشترین شباهت با مجموعه‌های نمونه هستند (Choi & Yao, 2005). به این نوع، سیستم یادگیری تنبل گفته می‌شود. زیرا مرحله آموزش در آن وجود ندارد که روش کا- نزدیکترین همسایگان[2] یک نمونه از آن است (Sebastiani,1999). درواقع مواردی از قبل دسته‌بندی می‌شوند و برای هرکدام یک مورد به عنوان نمونه در نظر گرفته می‌شود. سپس صفحات جدید با این نمونه‌ها مطابقت داده می‌شوند و با هر کدام که دارای بیشترین شباهت بود، به دسته مربوط به آن نمونه تعلق می‌گیرد.

روش دیگر، رده‌بندی مبتنی بر پارامترها[3] است که در آن برای تخمین پارامترهای یک توزیع احتمالی از نمونه‌های آموزشی استفاده می‌گردد. یک مثال آن قضیه بیز[4] است. این‌گونه موارد کمک می‌کند تا به صورت خودکار مدلی برای رده‌بندی ایجاد شود و صفحات دیگر بر مبنای آن رده‌بندی گردد (Choi & Yao, 2005).

نهایتاً در رویکردی متفاوت برای رده‌بندی خودکار با عنوان رده‌بندی خودکار متن مبتنی بر کتابشناسی بر اساس کاوش و استفاده از شبکه‌های استنادی میان اسناد علمی ارائه شده است. در این روش این‌طور فرض شده است که منابع مورد استفاده در یک سند می‌تواند در دسته‌بندی مشابه قرار گیرد. سپس سه مرحله برای آن در نظر گرفته شده است: ۱) شناسایی و استخراج منابع از اسناد، ۲) جست‌وجوی فرادادههای رده‌بندی موضوعی منابع مورد استناد از اوپک‌ها و ۳) اختصاص کلاس به فرادادههای اسناد با کمک مکانیزم‌های وزن‌دهی.

1. direct example based classifiers
2. K-Nearest Neighbors (KNN)
3. Parameter Based Classifiers
4. Naïve Bayes

رده‌بندی کتابخانه‌ای مورد استفاده، رده‌بندی دیویی بوده است (Joorabchi & Mahdi, 2011).

نتیجه‌گیری

رده‌بندی منابع یکی از موضوعات کلیدی در رشته علم اطلاعات و دانش‌شناسی و در فضای سازماندهی و بازیابی اطلاعات است. منابعی که به طور گسترده در حال تولید هستند، اگر به صورت مناسب سازماندهی نشوند، نمی‌توانند به نحو صحیح به دست جامعه هدف برسند و مورد استفاده قرار گیرند. اهمیت رده‌بندی در میان انواع مختلف منابع و مدارک مشهود است. در این زمینه، منابع دیجیتال به دلیل برخی از ویژگی‌ها دارای چالش‌های جدی هستند. اهمیت رده‌بندی در منابع دیجیتال از آن جهت است که گم شدن اینگونه منابع به معنای نابودی آن‌هاست؛ چرا که فاقد معادل فیزیکی هستند. از آنجا که حجم منابع دیجیتالی با سرعت بسیاری در حال افزایش است و هیچ‌گونه کنترلی بر آن وجود ندارد، لذا رده‌بندی آن‌ها به صورت دستی عملاً از لحاظ هزینه و زمان غیرممکن است. از این‌رو، رویکرد رده‌بندی خودکار مطرح شد. در رده‌بندی خودکار از روش‌های آماری مختلف و روش‌های یادگیری ماشینی برای رده‌بندی منابع به صورت خودکار استفاده می‌گردد. با این وجود، این مبحث با چالش‌هایی رو به رو است. اصلی‌ترین چالشی که در این زمینه وجود دارد آن است که اصولاً رده‌بندی را فرآیندی انتزاعی می‌دانند و این فرایند نیازمند تفکر انسانی است و روش‌های ماشینی و هوش مصنوعی هنوز نتوانسته‌اند به‌طور کامل جایگزین ذهن انسان شوند. از طرفی ممکن است در حین رده‌بندی خطاهایی رخ دهد و یک منبع به اشتباه به دسته‌ای غیرمرتبط تعلق گیرد. با تمام این تفاسیر، به دلیل رشد منابع وبی و تمایل کاربران در به‌کارگیری محیط دیجیتال و وب، نباید از این مهم دست کشید و باید به دنبال راهی برای رفع چالش‌های رده‌بندی خودکار بود.

در زبان انگلیسی پژوهش‌های قابل توجهی در موضوع رده‌بندی خودکار به انتشار رسیده و غالباً شامل گزارش کاربرد هر یک از روش‌ها و الگوریتم‌ها و نتایج به دست آمده می‌باشد، اما این موضوع کمتر در مقالات فارسی مورد توجه بوده است. در این مقاله تلاش شد با

شناسایی مقالات پژوهشی و مروری منتشر شده به زبان انگلیسی که به ویژه بر رده‌بندی خودکار وب تأکید داشتند، چارچوبی از اهمیت رده‌بندی خودکار، روش‌ها و الگوریتم‌های مورد استفاده در خوشه‌بندی و رده‌بندی و همچنین مراحل رده‌بندی خودکار وب ترسیم گردد. استفاده از طرح‌های رده‌بندی کتابخانه‌ای به عنوان بخشی از فرایند رده‌بندی خودکار متن، امکان استفاده از ابزارهای بازنمایی دانش را که در رشته علم اطلاعات و دانش‌شناسی به توسعه و ساخت آن‌ها پرداخته می‌شود، در کنار سایر روش‌ها نوید می‌دهد. رسیدن به درک روشن‌تری از موضوع رده‌بندی خودکار، امکان هم‌زبانی با متخصصان حوزه هوش مصنوعی و کامپیوتر را فراهم آورده و زمینه‌ساز پژوهش‌های میان‌رشته‌ای خواهد بود.

ORCID

Reza Dehkhodaie https://orcid.org/0000-0001-7296-9564

Atefeh Sharif https://orcid.org/0000-0003-4761-6761

منابع

اسماعیل‌پور، رضیه. (۱۳۸۶). رویکردها و چالش‌های رده‌بندی خودکار منابع اطلاعاتی در محیط جدید. *کتابداری و اطلاع‌رسانی*، ۱۰(۲)، ۹۱-۱۰۶.

اسماعیلی، مهدی. (۱۳۹۱). مفاهیم و تکنیک‌های دادهکاوی. بازیابی شده در ۱۳۹۹/۰۹/۲۰ از https://aghazeh.com/.

باغبانی، شهناز. (۱۳۹۶). *تکنیک‌ها و روش‌های یادگیری ماشین روی کلان داده*. کنفرانس ملی فناوری‌های نوین در مهندسی برق و کامپیوتر، اصفهان.

برومند، فیروزه. (۱۳۸۱). *رده‌بندی کتابخانه‌ای*. دایره‌المعارف کتابداری و اطلاع‌رسانی. تهران: کتابخانه ملی جمهوری اسلامی ایران.

تیم پژوهش راهبرد. (۱۴۰۰). طبقه‌بندی در داده‌کاوی. بازیابی شده در ۱۴۰۰/۰۳/۲۴ از https://raahbord.com/classification-in-data-mining/.

حصارکی، الهه. (۱۳۹۹). یادگیری ماشینی (Machine Learning) چیست؟ بازیابی شده در ۲۴/ ۱۴۰۰/۰۳ از https://b.fdrs.ir/pd.

هان، ژیاوی؛ کمبر، میشلین و پی، ژان. (۱۳۹۱). *مفاهیم و تکنیک‌های داده‌کاوی*، ترجمه مهدی اسماعیلی. تهران: نیاز دانش.

References

Choi, B., & Yao, Z. (2005). Web page classification. In *Foundations and Advances in Data Mining*, 221-274. Springer, Berlin, Heidelberg.

Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: a technical review. *International Journal of Computer Applications*, *28*(2), 37-40.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407.

Eito-Brun, R. (2014). Knowledge dissemination patterns in the information retrieval industry: A case study for automatic classification techniques. *World Patent Information*, *39*, 50-57.

Golub, K., Hagelbäck, J., & Ardö, A. (2018). *Automatic classification using DDC on the Swedish Union Catalogue*. In 18th European Networked Knowledge Organization Systems Workshop (NKOS 2018), *Porto, Portugal, September 13, 2018*, 4-16. CEUR-WS. org.

Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS transactions on computers*, *4*(8), 966-974.

Joorabchi, A., & Mahdi, A. E. (2011). An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. *Journal of Information Science*, *37*(5), 499-514.

Lassri, S, Benlahmar, H, Tragha, A. (2019). Machine Learning for Web Page Classification: A Survey. *International Journal of Information Science and Technology*, *3*(5), 38-50.

Maw, M., Balakrishnan, V., Rana, O., & Ravana, S. D. (2020). TRENDS AND PATTERNS OF TEXT CLASSIFICATION TECHNIQUES: A SYSTEMATIC MAPPING STUDY. *Malaysian Journal of Computer Science*, *33*(2), 102-117.

Qi, X., & Davison, B. D. (2009). Web page classification: Features and algorithms. *ACM computing surveys (CSUR)*, *41*(2), 1-31.

Sebastiani, F. (1999). A tutorial on automated text categorisation. In *Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence*, 7-35. Buenos Aires, AR.

Selvakuberan, K., Indradevi, M., & Rajaram, R. (2008). Combined Feature Selection and classification–A novel approach for the categorization of web pages. *Journal of Information and Computing Science*, *3*(2), 83-89.

Stamou, S., Ntoulas, A., Krikos, V., Kokosis, P., & Christodoulakis, D. (2006). January). Classifying web data in directory structures. In *Asia-Pacific Web Conference*, 238-249. Springer, Berlin, Heidelberg.

Baghbani, Shahnaz. (2016). *Techniques and methods of machine learning on big data*, National Conference of New Technologies in Electrical and Computer Engineering, Isfahan. [In Persian].

Boroumand, Firouzeh. Library Classification. (2002).. *Encyclopedia of librarianship and information.* Tehran: National Library of the Islamic Republic of Iran. [In Persian].

Esmaili, Mehdi. (2011). *Concepts and techniques of data mining.* Retrieved on 20/09/2013 from **https://aghazeh.com//** [In Persian].

Esmailpour, Razieh. (2007). Approaches and challenges of automatic classification of information sources in the new environment. *Library and Information Science, 10*(2), 91-106. [In Persian].

Han, Jiawei; Kember, Micheline and Pi, Jean. (2011). *Concepts and techniques of data mining, translated by Mehdi Esmaili.* Tehran: Niaz-e Danesh. [In Persian].

Hesaraki, Elaheh. (2019). What is machine learning? Retrieved on 24/03/1400 from **https://b.fdrs.ir/pd**. [In Persian].

Rahbord Research Team. (2021). Classification in data mining. Retrieved 03/24/1400 from **https://raahbord.com/classification-in-data-mining/**. [In Persian].

استناد به این مقاله: دهخدایی، رضا. (۱۴۰۱). مروری بر رده‌بندی خودکار متن. *بازیابی دانش و نظام‌های معنایی*، ۹(۳۲)، ۱۹۱-۲۱۹.