Improved Keyword Extraction for Persian Academic Texts Using RAKE Algorithm;
Case Study: Persian Theses and Dissertations

Elaheh Mehrabi
B.Sc Student in Industrial Engineering; Amirkabir University of Technology; Tehran, Iran Email: Elaheh.mehrabi77@aut.ac.ir

Azadeh Mohebi*
Assistant Professor; Faculty of Information Technology; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: mohebi@irandoc.ac.ir

Abbas Ahmadi
Associate Professor; Department of Industrial Engineering and Management Systems; Amirkabir University of Technology; Tehran, Iran Email: abbas.ahmadi@aut.ac.ir

Iranian Journal of Information Processing and Management

Received: 26, Jul. 2020 **Accepted: 15, Nov. 2020**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 37 | No. 1 | pp. 197-228
Autumn 2021

Abstract: Keywords and key phrases are subsets of most relevant words or phrases that summarize contents of a document while they play a critical role in information and document retrieval. Keyword extraction from scientific text is challenging and time-consuming due to the technical and multi-subject nature of the text, while the number of documents requiring keywords is increasing. There are various algorithms and methods developed for automatic keyword extraction. Rapid Automatic Keyword Extraction (RAKE) is a popular algorithm in this domain. RAKE's decisions are based on the observation that keywords generally contain multiple words and they rarely include stopwords and words with minimum lexical meanings. Candidate keywords are a set of single-word or multi-word sequences selected based on the scores assigned to them by some scoring criteria in RAKE.

In this research, a new modified version of RAKE algorithm is proposed in which candidate keyword scoring scheme is improved to increase precision and recall in the keyword extraction process. The proposed algorithm is to cover some of the main weaknesses of RAKE algorithm, especially in Persian scientific documents. To study the weaknesses of RAKE algorithm and evaluating the proposed modified version of RAKE, a set of metadata of Persian theses and dissertations are used. The result of test and evaluation of the proposed algorithm confirm improvement in precision, recall and F-measure.

* Corresponding Author

We study effectiveness of RAKE in extracting keywords from Persian texts. We find that RAKE algorithm often extracts long phrases with redundant words on Persian texts, leading to low accuracy. In this paper, we study sources of scoring inefficiency of RAKE algorithm and propose an improved version of RAKE algorithm with a novel scoring mechanism. Our scoring mechanism overcomes some of the weaknesses in RAKE's original scoring for Persian texts and yields better results. Our evaluations on Persian corpus demonstrate that our improved RAKE algorithm outperforms original RAKE algorithm by extracting more accurate keyword. Our results show that improved RAKE achieves more than 20% higher precision and recall on average compared to original RAKE.

Keywords: Keyword Extraction, RAKE Algorithm, Part of Speech Tagging, Natural Language Processing, Persian Scientific Document

بهبود الگوریتم RAKE برای استخراج کلیدواژه از متون علمی فارسی؛ مطالعهٔ موردی: پایان‌نامه‌ها و رساله‌های فارسی

الهه محرابی
دانشجوی کارشناسی مهندسی صنایع؛ دانشگاه صنعتی امیرکبیر؛ تهران، ایران؛
Elaheh.mehrabi77@aut.ac.ir

آزاده محبی
دکتری مهندسی طراحی سیستم‌ها؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
پدیدآور رابط mohebi@irandoc.ac.ir

عباس احمدی
دکتری مهندسی صنایع؛ دانشیار؛ دانشگاه صنعتی امیرکبیر؛ تهران، ایران abbas.ahmadi@aut.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۹/۰۵/۰۵ **پذیرش:** ۱۳۹۹/۰۸/۲۵ مقاله برای اصلاح به مدت ۲۸ 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 ۱ | ص‌ص ۱۹۷-۲۲۸
پاییز ۱۴۰۰

چکیده: کلمات کلیدی زیرمجموعه‌ای از کلمات یا عبارات یک سند هستند که می‌توانند معنای سند را توصیف کنند و در فرایند بازیابی اطلاعات نقش مهمی ایفا کنند. از آنجا که عملیات استخراج کلیدواژه یا عبارات کلیدی از متون تخصصی و علمی کاری تخصصی و زمان‌بر بوده و حجم اسناد علمی که نیاز به کلیدواژه دارند روزافزون است، الگوریتم‌های مختلفی برای استخراج تخصصی و خودکار کلیدواژه و عبارات کلیدی به اسناد طراحی و پیاده‌سازی شده‌اند. RAKE یک الگوریتم پرکاربرد برای استخراج کلمات کلیدی از متون است. اساس کار الگوریتم RAKE، کلمات کلیدی و عموماً حاوی چندین کلمه (یعنی عبارت کلیدی) هستند، ولی علائم نگارشی یا کلمات بی‌معنا یا ایست‌واژه‌ها را شامل نمی‌شوند. در این الگوریتم از برچسب‌گذاری دستوری کلمات به‌عنوان ابزاری برای تعیین ضریب اهمیت آن‌ها در جملات استفاده می‌شود. کلیدواژه‌ها مجموعه‌ای از توالی‌های چندکلمه‌ای یا تک‌کلمه‌ای هستند که طبق معیارهای خاصی امتیازدهی می‌شوند. در این پژوهش، یک نسخهٔ بهبودیافته از الگوریتم استخراج خودکار کلیدواژه (RAKE) ارائه شده است. در نسخهٔ بهبودیافته سعی شده با ایجاد تغییراتی در معیارهای امتیازدهی عبارات کاندید، دقت

و بازخوانی عبارات کلیدی استخراج‌شده افزایش یابد. راهکار ارائه‌شده برای بهبود الگوریتم RAKE با در نظر گرفتن ضعف‌های موجود در رویکردهای وزن‌دهی دراین الگوریتم به‌ویژه برای زبان فارسی و مستندات علمی پیشنهاد شده است. برای بررسی نقاط ضعف الگوریتم RAKE و ارائهٔ راهکار پیشنهادی از مجموعه‌ای از فراداده‌های پایان‌نامه و رساله‌های فارسی استفاده شده است. راهکار پیشنهادی روی این داده‌ها آزمایش و ارزیابی شده و باعث افزایش دقت، بازخوانی و معیار F شده است.

کلیدواژه‌ها: استخراج کلیدواژه، الگوریتم RAKE، برچسب‌گذاری دستوری، پردازش زبان طبیعی، مستندات علمی فارسی

۱. مقدمه

استخراج کلیدواژه از متون، یکی از حوزه‌های اصلی در پژوهش‌های مربوط به پردازش زبان طبیعی است. پردازش زبان طبیعی حوزه‌ای تخصصی درعلوم رایانه و هوش مصنوعی محسوب می‌شود که به بهره‌گیری از ابزارها، روش‌ها و الگوریتم‌ها برای پردازش و درک داده‌های طبیعی مبتنی بر زبان مربوط است و به‌طور معمول، در قالب‌های ساخت‌نیافته‌ای وجود دارند. تمام متون و اسناد برای ارائهٔ یک نظریه یا انتقال اطلاعات نوشته می‌شوند و هر یک، ایدهٔ منحصربه‌فرد خود را دنبال می‌کنند. برای انتقال مفهوم و ایدهٔ اصلی هر متن می‌توان آن را در مجموعه‌ای از لغات یا عبارات خلاصه نمود. این لغات، کلیدواژه نام دارند. کلیدواژه‌های یک سند مهم‌ترین کلمات یک سند هستند و ایدهٔ اصلی و بنیادی یک متن را در قالب عباراتی کوتاه، خلاصه و سودمند بیان می‌کنند و بنابراین، می‌توانند در جست‌وجو و بازیابی سند و طبقه‌بندی موضوعی اسناد به‌کار روند. عبارات کلیدی به‌طوری گسترده در زمینه‌های مختلفی همچون بازیابی اطلاعات، پردازش زبان‌های طبیعی، موتورهای جست‌وجو و ... به‌کار می‌روند. در سیستم‌های بازیابی اطلاعات، از واژگان کلیدی برای توصیف اسناد به‌دست‌آمده از یک تحقیق یا پرسشنامه استفاده می‌شود. نمایه‌سازان در نمایه‌سازی اسناد و خوانندگان در خواندن داده‌های متنی بزرگ موجود در کتابخانه‌های دیجیتال در یافتن بخش‌های مورد نظر خود با مشکل روبه‌رو می‌شوند. در خوشه‌بندی اسناد، کاربرد کلیدواژه‌ها به‌عنوان ویژگی‌های سند مزایای مختلفی دارد. کاربرد کلیدواژه‌ها به‌طور مثال، می‌تواند کیفیت خوشه‌ها را بهبود بخشد، و در ساخت توضیحات مختصر و دقیق (برچسب‌ها) برای خوشه‌های تولیدشده نیز سودمند باشد. کلیدواژه‌ها همچنین، در موتورهای جست‌وجو برای استخراج اطلاعات

مورد نیاز مفید هستند. کلیدواژه‌های موضوعی به‌طور خلاصه محتوای اصلی موضوعات را نشان می‌دهند. این عبارات به کاربران کمک می‌کنند تا محتوای اصلی یک موضوع را درک کنند و تصمیم بگیرند که خواندن سند را ادامه دهند یا نه (Alami Merrouni, Frikh & Ouhbi 2020). با افزایش حجم اسناد و متون و ذخیره‌سازی آن‌ها به‌صورت الکترونیک و نیاز به جست‌وجو و بازیابی آن‌ها در بسیاری از موارد فرایند اختصاص کلیدواژه به اسناد به‌صورت خودکار انجام می‌شود. تاکنون روش‌های متعددی برای استخراج و اختصاص خودکار کلیدواژه پیشنهاد و به‌کار گرفته شده است. بعضی از این روش‌ها بر پایهٔ شاخص‌های آماری عمل می‌کنند و مبنای انتخاب کلیدواژه‌ها را فراوانی آن‌ها در متن در نظر می‌گیرند و از شاخص‌هایی نظیر TF-IDF (فراوانی کلمه–فراوانی محتوای معکوس)[1] استفاده می‌کنند (Chen, Chen & Liang 2016). برای استخراج و اختصاص خودکار کلیدواژه‌ها از روش‌های مبتنی بر یادگیری ماشین نیز بهره‌گیری می‌شود. بعضی از این روش‌ها خود به دو دستهٔ روش‌های باناظر و روش‌های بدون ناظر تقسیم‌بندی می‌شوند (Alami Merrouni, Frikh & Ouhbi 2020; Papagiannopoulou & Tsoumakas 2019). افزون بر روش‌های پیشنهادی، انواع روش‌ها را می‌توان بر پایهٔ زمینهٔ کاربرد آن‌ها نیز دسته‌بندی نمود. به‌عنوان نمونه، در برخی از روش‌ها برای استخراج کلیدواژه‌ها از مجموعه‌ای از مستندات استفاده شده است (Bayatmakou, Ahmadi & Mohebi 2017؛ محبی و جلالی‌منش ۱۳۹۸) و برخی نیز بر استخراج کلیدواژه از یک سند متمرکز بوده‌اند. از بین روش‌های استخراج کلیدواژه از یک سند، برخی روش‌ها برای اسناد کوتاه طراحی شده‌اند (Rose et al. 2010; Timonen et al. 2013). در این پژوهش، تمرکز روی طراحی و پیاده‌سازی روشی برای استخراج کلیدواژه از یک سند علمی کوتاه است. منظور از سند علمی کوتاه سندی است که چکیدهٔ آن موجود است و لازم است کلیدواژه‌ها تنها با استفاده از چکیده و یا عنوان استخراج شوند. برای این منظور، در این پژوهش تنها چکیده، عنوان و کلیدواژه‌های بخشی از پایان‌نامه‌ها و رساله‌های ثبت‌شده در پایگاه اطلاعات علمی ایران موسوم به «گنج»، به‌عنوان مطالعهٔ موردی برای پیاده‌سازی روش پیشنهادی در نظر گرفته شده است. در پایگاه «گنج» تاکنون نزدیک به ۶۰۰ هزار پایان‌نامه و رساله ثبت شده است. هم‌اکنون فرایند اختصاص کلیدواژه به مدارک ثبت‌شده و نیز نمایه‌سازی آن‌ها توسط متخصص

1. term frequency- inverse document frequency (TF-IDF)

نمایه‌ساز انجام می‌گیرد، ولی به‌دلیل حجم بالای این مدارک و درخواست روزافزون بر ثبت پایان‌نامه‌ها و رساله‌ها، وجود روشی خودکار برای تخصیص کلیدواژه به این اسناد ضروری است. استفاده از روش خودکار نمایه‌سازی برای نمایه‌سازی خودکار مستندات پایگاه «گنج» می‌تواند سرعت نمایه‌سازی را افزایش داده و امکان اختصاص کلیدواژه‌های بیشتر به اسناد را فراهم آوَرَد. کلیدواژه‌های یک سند در «گنج»، همانند هر پایگاه اطلاعاتی تخصصی، نقش مهمی در بازیابی اطلاعات از آن پایگاه ایفا می‌کند. از طرف دیگر، استفاده از روش‌های ماشینی مانند آنچه که در این پژوهش به آن پرداخته شده، در نمایه‌سازی می‌تواند امکان تخصیص کلیدواژه‌های بیشتری را فراهم آورد و حتی در مواردی با در دست داشتن امتیاز کلیدواژه‌های پیشنهادی توسط الگوریتم، در امتیازدهی و رتبه‌بندی اسناد در بازیابی اطلاعات مؤثر باشد.

استخراج کلیدواژه از مستندات علمی را می‌توان از متن کامل مقاله انجام داد و یا برای این کار می‌توان تنها از چکیده و عنوان سند استفاده کرد. «خطیر و گنجه‌فر» در پژوهشی توزیع کلیدواژه‌ها را در عنوان و چکیدهٔ اسناد «گنج» بررسی کردند. نتایج پژوهشی آن‌ها نشان داد که استفاده از عنوان و چکیده به‌تنهایی در داده‌های «گنج» نیز می‌تواند منجر به استخراج کلیدواژه‌های مناسب شود (۱۳۹۷). در پژوهش حاضر نیز از چکیده، عنوان و کلیدواژه‌های بخشی از پایان‌نامه‌ها و رساله‌های پایگاه «گنج» استفاده می‌شود. بنابراین، یکی از چالش‌هایی که وجود دارد، کوتاه‌بودن متن هر سندی است که کلیدواژه از آن استخراج می‌شود. روش‌هایی برای استخراج کلیدواژه از متون کوتاه پیشنهاد شده که اکثر آن‌ها برای متون انگلیسی به‌کار رفته‌اند. اگرچه بیشتر این روش‌ها در ظاهر وابستگی زبانی ندارند و می‌توان از آن‌ها برای متون به هر زبان دیگری نیز به‌کار برد، لیکن معمولاً عملیات پیش‌پردازش متن در این روش‌ها که وابسته به عناصر زبانی است، نقش مهمی در کیفیت نتیجهٔ نهایی ایفا می‌کند. بنابراین، به‌سادگی نمی‌توان روش‌های موجود برای زبان انگلیسی را برای متون دیگر زبان‌ها به‌کار برد. در این پژوهش از یکی از این روش‌ها به نام RAKE (الگوریتم سریع و خودکار استخراج کلیدواژه)[1] استفاده می‌شود. روش RAKE الگوریتمی است که با بهره‌گیری از فهرستی از ایست‌واژه‌ها[2] و برچسب‌گذاری دستوری

1. rapid automatic keyphrase extraction (RAKE)
2. stop words

کلمـات[1] و ...، کلیدواژه‌هـای کاندیـد را از متـن اسـتخراج می‌کنـد. در ایـن پژوهـش به‌طـور ویـژه، افـزون بـر مـرور چالش‌هـای به‌کارگیـری ایـن روش روی متـون علمـی فارسـی موجـود در پایـگاه «گنـج»، راهـکاری نویـن بـرای بهبـود وزن‌دهـی کلیدواژه‌هـا پیشـنهاد می‌شـود.

پژوهش‌های پیشین

در یـک دسـته‌بندی کلـی می‌تـوان روش‌هـای اسـتخراج کلیـدواژه از متـون را در دو گـروه «باناظـر» و «بـدون ناظـر» طبقه‌بنـدی کـرد (Aggarwal & Zhai 2012). البتـه، اخیـراً روش‌هـای مبتنـی بـر یادگیـری عمیـق نیـز بـه ایـن دسـته‌بندی اضافـه شـده اسـت (Alami Merrouni, Frikh & Ouhbi 2020). در روش باناظـر **بـا اسـتفاده** از مجموعـهٔ متـون و در دسـت داشـتن کلیدواژه‌هـای هـر یـک از آن‌هـا، مـدل مناسـب تعییـن می‌شـود. سـپس، بـرای هـر متـن بـا اسـتفاده از مـدل آمـوزش داده‌شـده، بـر اسـاس مجموعـه‌ای از متـون، کلیـدواژهٔ مناسـب تخصیـص داده می‌شـود؛ در حالـی کـه در روش بـدون ناظـر، کلیدواژه‌هـا بـدون اسـتفاده از متـون و عمدتـاً براسـاس محتـوی خـود سـند اسـتخراج می‌شـوند (Sathya & Abraham 2013). روش‌هـای بـدون ناظـر بـه جهـت عـدم نیـاز بـه دامنـه و متـون مرجـع و داده‌هـای برچسب‌گذاری‌شـده مـورد توجـه هسـتند. از سـوی دیگـر، روش‌هـای باناظـر توانایـی بالاتـری در مدل‌سـازی دارنـد و از دقـت بالاتـری برخوردارنـد (Meng et al. 2017). از جملـه روش‌هـای اسـتخراج کلیـدواژه بـدون ناظـر می‌تـوان بـه روش‌هـای آمـاری، روش‌هـای مبتنـی بـر گـراف، و روش‌هـای مبتنـی بـر مـدل زبانـی اشـاره‌کرد. همچنیـن، از روش‌هـای اسـتخراج کلیـدواژه باناظـر می‌تـوان از روش‌هـای مبتنـی بـر یادگیـری عمیـق[2] نـام بـرد (Alami Merrouni, Frikh & Ouhbi 2020; Papagiannopoulou & Tsoumakas 2019). از طـرف دیگـر، روش‌هـای اسـتخراج کلیـدواژه را بـر پایـهٔ اینکـه از چندیـن سـند کلیـدواژه اسـتخراج می‌کننـد یـا از یـک سـند، می‌تـوان بـه دو دسـته تقسـیم‌بندی کـرد. برخـی از ایـن روش‌هـا نیـز بـرای اسـتخراج کلیـدواژه از اسـناد کوتـاه کاربـرد دارنـد و برخـی بـرای اسـتخراج کلیـدواژه از اسـناد بلنـد.

بسـیاری از الگوریتم‌هـای اسـتخراج کلیـدواژه از متـون بـر مبنـای معیـار فراوانـی کلمـات در متـن هسـتند. از آن جملـه می‌تـوان بـه الگوریتم‌هـای TF-IDF، TextRank و RAKE اشـاره‌کرد. یکـی از قدیمی‌تریـن روش‌هـای وزن‌دهـی بـه کلمـات در متـون روش TF-IDF اسـت. ایـن

1. part of speech tagging (POS)
2. deep learning

روش، روشی آسان و پرکاربرد برای پیدا کردن چگالی کلیدواژه‌ها در بازیابی اطلاعات است که بر پایهٔ فراوانی کلمات تعریف می‌شود. خروجی‌های این الگوریتم بیشتر تک‌کلمه هستند، نه عبارات کلیدی. این در حالی است که در بسیاری از متون، عبارات کلیدی که شامل چند کلمه باشند، بسیار گویاتر و کارآمدتر از تک‌کلمات هستند. از طرف دیگر، در بسیاری از موارد کلمه‌ای که بیشترین فراوانی را دارد، لزوماً نمی‌تواند بیانگر مفهوم اصلی متن باشد.

«میهالسی رادا» در مقالهٔ خود الگوریتم TextRank را معرفی کرده است. این الگوریتم یک گراف از کلمات و روابط بین آن‌ها در یک سند متنی ایجاد می‌کند. سپس، مهم‌ترین رأس‌های گراف (کلمات) را بر پایهٔ نمراتشان شناسایی می‌کند. از مشکلات این الگوریتم می‌توان به‌سرعت پایین‌تر آن نسبت به الگوریتم‌های دیگر اشاره کرد (Mihalcea Rada 2004). «اندرید و والنسیا» (۱۹۹۸) تحقیقات خود را بر پایهٔ مقایسهٔ تعداد کلمات در یک سند و یک مجموعه از اسناد و نوشته‌ها انجام داده‌اند (Andrade & Valencia 1998). «رز» و همکاران برای اولین‌بار «الگوریتم سریع و خودکار استخراج کلیدواژه» (RAKE) را معرفی کردند (Rose et al. 2010). این الگوریتم مستقل از مجموعهٔ کلمات و مجموعه‌ای از اسناد[1] و تا حدودی مستقل از زبان متن، کلمات و عبارات کلیدی را استخراج می‌کند و حتی برای متون کوتاه هم کاربرد دارد.

در مورد مستندات فارسی نیز پژوهش‌های متعددی برای استخراج کلیدواژه انجام شده است. «کلانتری» و همکاران در پژوهش اخیر خود انواع روش‌های استخراج کلیدواژه برای مستندات فارسی را مرور کردند. در این پژوهش، نویسندگان ۲۰ مقاله در حوزهٔ استخراج کلیدواژه را بررسی کردند که از بین آن‌ها ۷ مقاله برای متون علمی روش استخراج کلیدواژه را پیاده‌سازی و ارزیابی کرده بودند (کلانتری و همکاران ۱۳۹۹). مجموعهٔ داده‌های دو پژوهش «بشیری» و همکاران (۱۳۸۲) و «تشکری» و همکاران (۱۳۸۴) از بین ۷ پژوهش، داده‌هایی متشکل از پایان‌نامه‌ها و مقالات فارسی داشتند و در این دو پژوهش از روش‌های آماری استفاده شده است. همچنین، «کلانتری» و همکاران در مقالهٔ مروری خود، سه روش از بین ۲۰ پژوهش انجام‌شده را به‌عنوان روش‌هایی پیشنهاد کردند که انسجام نسبی داشتند و امکان تعمیم آن‌ها روی داده‌های دیگر نیز تا حدی وجود

1. corpus

دارد. یکی از این سه روش، پژوهش «حسینی‌خواه و احمدی» (۱۳۹۲) است که در مقالهٔ خود از شبکهٔ رقابتی LVQ[1] و شبکهٔ عصبی MLP[2] برای استخراج کلیدواژه استفاده کردند و مجموعه‌ای ۸۰ تایی از اخبار فارسی را برای آموزش شبکه به‌کار گرفتند. روش دیگر پژوهش «ویسی و افلاکی» (۱۳۹۴) است که در آن از روش‌های آماری بر روی داده‌های خبری استفاده کردند و پژوهش سوم، پژوهش «باسره، درهمی و ظریف‌زاده» (۱۳۹۶) است. این پژوهش نیز بر روش‌های آماری بر روی متون خبری استوار است (۱۳۹۹). در نهایت، می‌توان گفت که در پژوهش‌های متون فارسی تاکنون از روش RAKE برای متون فارسی علمی استفاده نشده و در عین حال، به‌دلیل در دسترس نبودن مجموعهٔ دادهٔ استاندارد در زمینهٔ مستندات علمی فارسی، اکثر پژوهش‌ها روی مجموعهٔ داده‌های پیکرهٔ همشهری صورت گرفته است. روش‌های پیشنهادشده بر روی مستندات علمی فارسی نیز بیشتر بر روی مقالات بوده و بیشتر بر روش‌های با ناظر و روش‌های آماری مبتنی بوده‌اند.

از آنجاکه در پژوهش حاضر، روش جدیدی برای مبنای الگوریتم RAKE برای استخراج کلمات و عبارات کلیدی از متون کوتاه فارسی‌زبان پیشنهاد می‌شود، در ادامه، پژوهش‌هایی که بر روی این الگوریتم انجام شده، بررسی می‌شوند.

«توشارا، مونیکا و مانگامورو» عملکرد سه الگوریتم بدون نظارت RAKE، TextRank و PositionRank را با یکدیگر مقایسه کردند. طبق نتایج این تحقیق، الگوریتم PostionRank به‌دلیل توجه‌کردن به مکان کلمه در متن دقت بیشتری دارد. از جمله مشکلات ذکرشده برای الگوریتم RAKE لزوم در دست داشتن مجموعهٔ کاملی از ایست‌واژه‌ها و عدم عملکرد دقیق در صورت نداشتن این مجموعه است (Thushara, Mownika & Mangamuru 2019).

«هاک» الگوریتم RAKE را جهت پیاده‌سازی روی زبان بنگالی مورد بررسی قرار داد. از جمله مشکلات این الگوریتم در زبان بنگالی که «هاک» به آن اشاره می‌کند، عدم تفاوت امتیاز دو عبارت با کلمات یکسان، اما با ترتیب متفاوت است. همچنین، در صورت ادغام ایست‌واژه‌ها با یک کلمه، این کلمات توسط الگوریتم قابل تشخیص نیستند. او جهت رفع کاستی‌های عملکرد این الگوریتم برای زبان بنگالی، الگوریتم پیشنهادی RAKE را معرفی کرد. تفاوت این الگوریتم با الگوریتم اصلی در شیوهٔ تعیین امتیاز کلمات است (Haque 2019).

1. learning vector quantization 2. multi- layer perceptrons

«سیدیکی و شاران» در پژوهش خود در مورد مستندات هندی روشی را برای بهبود الگوریتم RAKE پیشنهاد کردند. از آنجا که در زبان هندی فهرست منسجمی از ایست‌واژه‌ها وجود ندارد، آن‌ها نخست، روشی را برای ایجاد این فهرست ارائه کردند. سپس، توابع جدیدی برای امتیازدهی به کلیدواژه‌های کاندید پیشنهاد کرده و نتایج را روی متون ادبی بلند آزمایش نمودند (Siddiqi & Sharan 2018).

در زمینهٔ مستندات علمی فارسی تاکنون پژوهشی برای به‌کارگیری و بهبود الگوریتم RAKE انجام نشده است. تنها پژوهشی که ساختار آن تا حدی با این الگوریتم شباهت دارد، پژوهشی است که توسط «توکلی‌زاده راوری» انجام شده است. وی در پژوهش خود مدلی را ارائه کرد که مشابه الگوریتم RAKE عمل می‌کرد. این الگوریتم بدین‌صورت بود که با تعیین و حذف کلماتی که از نظر ساختار جمله و نقش آن‌ها در جمله نمی‌توانند کلیدواژه باشند، (مانند افعال، قیدها، حروف اضافه، ...)، جملات را به عباراتی متشکل از مجموعه‌ای از اسم و صفت تبدیل کرد که عبارات کاندید برای کلیدواژه بودند. سپس، بر اساس مجموعه‌ای از شاخص‌های آماری و وزن‌دهی عبارات کاندید، کلیدواژه‌ها را استخراج نمود (۱۳۹۴).

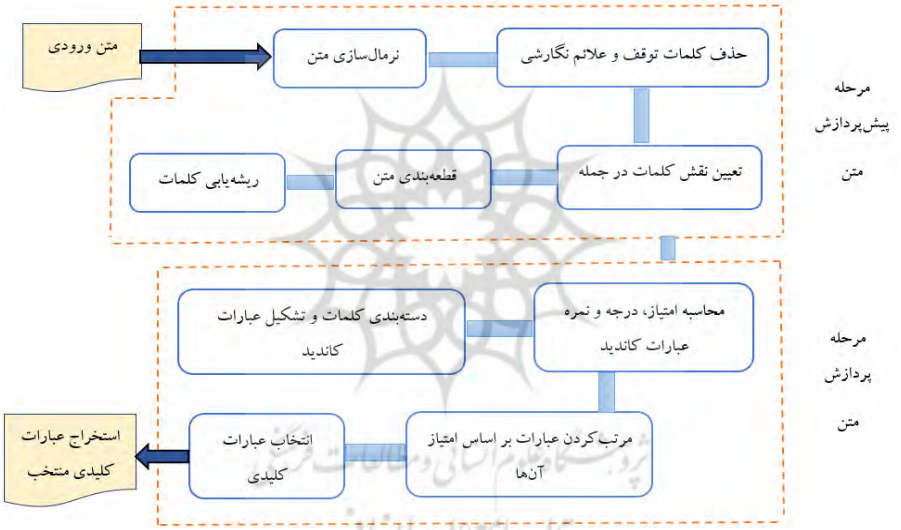
در پژوهش حاضر الگوریتم RAKE را برای پایان‌نامه‌های فارسی (چکیده و خلاصه) انتخاب کرده‌ایم. این الگوریتم از روش‌های یادگیری ماشین بدون نظارت بهره می‌گیرد و مستقل از منبع اسناد و زبان متون است و می‌توان از آن برای متون فارسی هم استفاده کرد. همچنین، بر خلاف برخی از الگوریتم‌های پیشین، خروجی‌های این الگوریتم اغلب عبارات کلیدی (دربرگیرندهٔ دو کلمه یا بیشتر) هستند که با توجه به ساختار و قواعد خاص زبان فارسی در اغلب موارد عبارات کلیدی در مقایسه با تک‌کلمات، بهتر می‌توانند مفهوم اصلی متن را بیان کنند. اگرچه این الگوریتم، همانند دیگر الگوریتم‌ها در زبان انگلیسی، به ظاهر وابستگی زبانی ندارد، لیکن پیش‌پردازش‌هایی که باید روی متن در زمان به‌کارگیری الگوریتم انجام شود، روی نتیجه تأثیر زیادی خواهد گذاشت. این پیش‌پردازش‌ها عموماً به ساختار زبانی وابسته هستند. این الگوریتم پیش از این برای زبان فارسی پیاده‌سازی نشده است. بنابراین، این الگوریتم در آغاز، برای زبان فارسی پیاده‌سازی می‌شود، سپس، مشکلات و کاستی‌های آن را در متن فارسی بررسی می‌کنیم و جهت بهبود عملکرد آن تغییراتی پیشنهاد می‌شود.

شرح الگوریتم RAKE

«رز» و همکاران الگوریتم RAKE را جهت استخراج خودکار کلیدواژه در زبان انگلیسی معرفی کردند (Rose et al. 2010). این الگوریتم تنها روی یک سند اجرا می‌شود و نیازی به استفاده از مجموعهٔ داده مجزا در آن نیست. در ادامه، شیوهٔ عملکرد این الگوریتم توضیح داده می‌شود.

مراحل الگوریتم

الگوریتم RAKE از دو مرحلهٔ اصلی پیش‌پردازش متن و پردازش تشکیل شده که در شکل ۱، فرایندهای هر مرحله نمایش داده شده است.



شکل ۱. مراحل اصلی الگوریتم RAKE

پیش‌پردازش متن

گاهی اوقات حروفی مثل «ک» یا «ی» در زبان فارسی در رایانه با کدهای مختلف نوشته می‌شوند. برای پردازش، نخست، باید همهٔ این حروف یکسان شوند و متن نرمال گردد. یکی از مهم‌ترین کارها در مرحلهٔ پیش‌پردازش، حذف ایست‌واژه‌ها و علائم نگارشی است. ایست‌واژه‌ها، کلماتی هستند که چه از لحاظ معنایی و چه از دید قواعد دستور زبان ارزش انتخاب شدن به‌عنوان کلیدواژه را ندارند و باید از متن حذف شوند.

به‌طور معمول، افعال و حروف ربط و اضافه فاقد ارزش معنایی هستند و در فهرست ایست‌واژه‌ها قرار می‌گیرند. همچنین، علائم نگارشی و اعداد نیز در این فهرست جای دارند. برچسب‌گذاری دستوری کلمات[1] نیز که برای تعیین ضریب اهمیت کلمات و تعیین نقش کلمات در جمله است، می‌توانند برای حذف ایست‌واژه‌ها به‌کار رود. منظور از برچسب‌گذاری کلمات، نسبت‌دادن برچسب‌های دستوری همچون اسم، صفت، حرف اضافه و ... به آن‌هاست. از آنجا که فعل‌ها و حروف اضافه به‌طور معمول، نقش تعیین‌کننده‌ای در کلیدواژه‌ها ندارند، در الگوریتم‌های استخراج کلیدواژه عموماً واژگانی که این برچسب‌ها را در متن دارند از فهرست اولیهٔ کلمات کاندید برای کلیدواژه حذف می‌شوند. بنابراین، در بسیاری از الگوریتم‌های استخراج کلیدواژه، در مراحل پیش‌پردازش، ایست‌واژه‌ها و کلماتی که بر پایهٔ برچسب نقش آن‌ها در جمله بار معنایی ندارند، حذف می‌شوند.

برای تهیهٔ فهرست ایست‌واژه‌ها از یکی از پروژه‌های گیت‌هاب[2] استفاده شده و برای غنی‌سازی آن، نتایج پژوهش «سمائی و رسولی» (۱۳۹۹) نیز در آن لحاظ شده است. در پژوهش حاضر از مدل برچسب‌گذار موجود در کتابخانهٔ «هضم»[3] استفاده شده است. «هضم» یکی از بسته‌های زبان «پایتون» است. چگونگی برچسب‌گذاری دستوری توسط مدل برچسب‌زن کتابخانهٔ «هضم» روی یک نمونهٔ متن کوتاه در شکل ۲، نمایش داده شده است. در این شکل، برچسب دستوری کلمات با حروف انگلیسی نشان داده شده، که N و Ne بیانگر اسامی جمع و مفرد، AJ صفت، P حرف اضافه، V فعل، NUM عدد و PUNC علائم نگارشی را نشان می‌دهند.

1. part of speech (POS) tagging
2. https://github.com/kharazi/persian-stopwords
3. https://www.sobhe.ir/hazm/

متن ورودی :

«هوش مصنوعی شاخه‌ای از علوم کامپیوتر است؛ که توسط دستگاه‌ها ارائه می‌گردد. درعلم کامپیوتر دستگاه هوشمند یک عامل عقلانی انعطاف‌پذیر است.»

نتیجه‌ی برچسب‌گذاری دستوری کلمات توسط مدل فوق:

[('هوش', 'Ne'), ('مصنوعی', 'AJ'), ('شاخه\u200cای', 'N'), ('از', 'P'), ('علوم', 'N'), ('کامپیوتر', 'N'), ('است', 'V'), ('؛', 'PUNC'), ('که', 'CONJ'), ('توسط', 'Pe'), ('دستگاه\u200cها', 'N'), ('ارائه\u200cمی\u200cگردد', 'V'), ('.', 'PUNC'), ('در', 'P'), ('علم', 'Ne'), ('کامپیوتر', 'Ne'), ('دستگاه', 'Ne'), ('هوشمند', 'AJ'), ('یک', 'NUM'), ('عامل', 'Ne'), ('عقلانی', 'AJe'), ('انعطاف\u200cپذیر', 'AJ'), ('است', 'V'), ('.', 'PUNC')]

شکل ۲. نمونهٔ برچسب‌گذاری دستوری کلمات

به‌طور کلی، عملیات پیش‌پردازش متن را می‌توان به‌صورت زیر خلاصه نمود:

- ◇ نرمال‌سازی متن؛
- ◇ حذف ایست‌واژه‌ها و علائم نگارشی؛
- ◇ برچسب‌زنی دستوری کلمات؛
- ◇ قطعه‌بندی متن[1]؛
- ◇ ریشه‌یابی کلمات[2].

چالش‌های مرحلهٔ پیش‌پردازش متن

ویژگی‌های خاص زبان فارسی، به‌ویژه در بحث نگارش، مرحلهٔ پیش‌پردازش را با چالش‌هایی مواجه می‌کند (توکلی‌زاده راوری ۱۳۹۴). برخی از این ویژگی‌ها عبارت‌اند از:

- ◇ تنوع نوشتاری در برخی از کلمات مانند «تهران» و «طهران»؛
- ◇ عدم رعایت درست نیم‌فاصله‌ها و فاصله‌ها؛
- ◇ صورت‌های مختلف نوشتاری به‌صورت جدا و سرهم مانند «آن‌ها» و «آنها»؛
- ◇ استفاده از اصل کلمات انگلیسی با نوشتار فارسی، به‌طور نمونه وب‌سایت (به‌جای website)؛

1. tokenizing 2. stemming

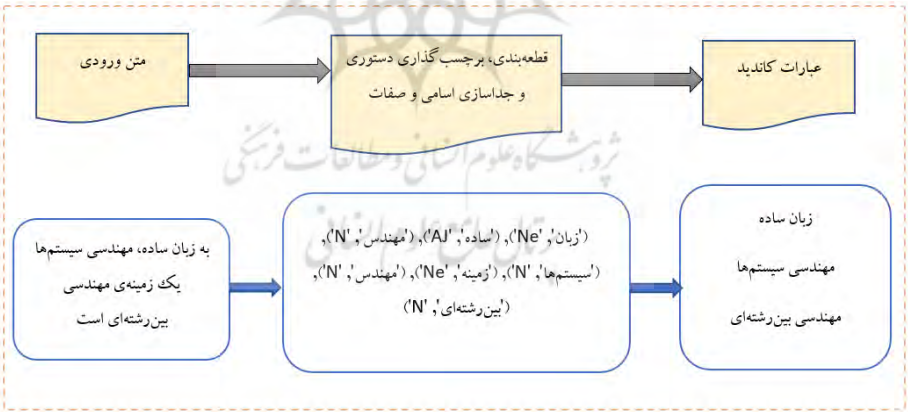
◇ تنوع نوشتاری در حروفی مانند ک یا ی که در صفحه‌کلیدهای مختلف ممکن است کدهای مختلفی داشته باشند و به شیوه‌های مختلف نوشته شوند (Hosseinikhah, Ahmadi & Mohebi 2018).

پردازش متن

در این مرحله با دسته‌بندی کلمات، عبارات کلیدی کاندید تشکیل می‌شوند. سپس، معیارهای وزن‌دهی شامل درجه[1]، فراوانی[2]، و امتیاز[3] عبارات کلیدی کاندید محاسبه می‌شود. در انتها، پس از مرتب‌سازی عبارات بر پایهٔ امتیازهایشان، عبارات برگزیده استخراج می‌گردد. در ادامه، گام‌های مرحلهٔ پردازش متن تشریح می‌شوند.

◇ دسته‌بندی کلمات و تشکیل عبارات کاندید

در مرحلهٔ پردازش متن، نخست عبارات کاندید برای کلیدواژه مشخص می‌شوند. برای این منظور، در هر جمله کلماتی که قبل و بعد از آن‌ها ایست‌واژه وجود داشته باشد، در یک دسته به‌عنوان عبارت کاندید قرار می‌گیرد. به این ترتیب، دسته‌بندی جدیدی بر اساس عبارات به‌وجود می‌آید که نمونه‌ای از این دسته‌بندی در شکل ۳، نمایش داده شده است.



شکل ۳. تشکیل عبارات کاندید

1. degree 2. frequency 3. score

◇ محاسبهٔ امتیاز و وزن عبارات کاندید

برای هر کلمهٔ w_j که در عبارات مختلفی دیده شده، درجهٔ کلمه به‌صورت زیر تعریف می‌شود:

$$deg_w(w_j) = \sum_{p_i \in \mathrm{P}_{w_j}} length(p_i.w_j) \quad (۱)$$

که در آن P_{w_j} مجموعهٔ تمام عباراتی است که شامل کلمهٔ w_j است:

$$\mathrm{P}_{w_j} = \{p_i | w_j \in p_i\} \quad (۲)$$

و p_iها عبارت‌هایی هستند که کلمهٔ w_j درآن‌ها وجود دارد و $length(p_i.w_j)$ تعداد کلمات هر عبارت p_i است. اگر $freq(w_j)$ بیانگر فراوانی کلمهٔ w_j باشد، آنگاه، امتیاز هر کلمه این‌گونه تعریف می‌شود:

$$score_w(w_j) = \frac{deg_w(w_j)}{freq(w_j)} \quad (۳)$$

در پایان، برای هر عبارت کاندید p_i، امتیاز آن برابر است با مجموع امتیازهای کلمات تشکیل‌دهندهٔ آن. بنابراین:

$$score_{RAKE}(p_i) = \sum_{j=1}^{n} score_w(w_j) \quad (۴)$$

که در آن فرض بر این است که عبارت p_i دارای n کلمه $w_1.w_2.\dots.w_n$ است. شبه‌کد الگوریتم RAKE در شکل ۴، آمده است.

Algorithm: RAKE

```
1: procedure start
2:      Text=[] ← Input text
3:      preprocessing section
            Def Clean()
            Text ← Clean[Text]
            Text ← word_tokenize[text]
            Pos_tag[text]
            Adjective_tags = ["AJ","AJe"]
        for word in pos_tag[text]:
            if word in adjective_tags:
                lemmatized_text.append(word)
4:      processing section
            Wanted_pos=["N","AJ","AJe","Ne"]
            Partitioned_phrases=[] ← partitioned lemmatized text using stopwords
        for phrase in phrases:
            for word in phrase:
                frequency[word]+=1
                degree[word]+=len (phrase)
                word_score[word]=degree[word]/frequency[word]
                phrase_score+=word_score[word]
5:      print phrase_scores
6: End procedure
```

شکل ٤. شبه‌کد الگوریتم RAKE

چالش‌های اصلی پیاده‌سازی الگوریتم RAKE

با توجه به اینکه الگوریتم RAKE نخست، جهت استخراج عبارات کلیدی در زبان انگلیسی طراحی شده، با پیاده‌سازی الگوریتم RAKE روی متون استاندارد فارسی چالش‌هایی به‌دنبال دارد. برخی از این چالش‌ها عبارت‌اند از:

- برخی از عبارات کلیدی که توسط RAKE استخراج شده، طولانی هستند که می‌تواند برآمده از حذف‌نشدن یک حرف ربط، وجود توضیحات اضافی و غیرضروری، و نیز رشته‌ای از کلمات باشد که به‌صورت صفت یا مضاف و مضاف‌الیه به‌دنبال هم قرار گرفته‌اند. در نظر گرفتن حرف ربط «و» گاهی می‌تواند مشکل ایجاد کند. به‌عنوان نمونه، اگر «و» را در فهرست ایست‌واژه‌ها قرار ندهیم، عبارات طولانی مانند «منابع توزیع پراکنده و منابع تجدیدپذیر انرژی» حاصل می‌شود. در عین حال، اگر حرف «و» را در فهرست ایست‌واژه‌ها در نظر بگیریم، عباراتی نظیر «آموزش و پرورش» نیز به‌صورت دو کلمه جدا در نظر گرفته می‌شوند. این چالش گاهی در زبان انگلیسی هم دیده می‌شود.

◇ در الگوریتم RAKE به عبارات کلیدی طولانی‌تر، بدون توجه به ارزش کلمات تشکیل‌دهندهٔ آن‌ها وزن بیشتری نسبت به عبارات کوتاه‌تر اختصاص داده می‌شود. به‌طور مثال، عبارت «رشتهٔ تحصیلی مهندسی صنایع» نسبت به «مهندسی صنایع» وزن بیشتری دارد؛ در حالی که عبارت دوم برای انتخاب شدن به‌عنوان کلیدواژه مناسب‌تر است. دلیل ایجاد این مشکل روش وزن‌دهی عبارت کاندید در الگوریتم RAKE است.

برای برطرف کردن این دست از چالش‌ها، در ادامه، مدل RAKE بهبودیافته پیشنهاد می‌شود. در مدل پیشنهادی، سه راهکار برای تغییر روش محاسبهٔ امتیاز عبارات کاندید معرفی می‌شوند و در پایان مدل پیشنهادی با ترکیب این سه راهکار عمل می‌کند.

مدل RAKE بهبودیافته

در مدل پیشنهادی سه راهکار برای بهبود عملکرد الگوریتم RAKE پیشنهاد می‌شود که در ادامه شرح داده خواهد شد.

نرمال‌سازی درجهٔ عبارات نسبت به طول آن‌ها

در بین عبارات مختلف به‌طور معمول، عبارات طولانی‌تر نسبت به عبارات کوتاه‌تر اطلاعات بیشتری را در اختیار ما می‌گذارند. بر این پایه، در الگوریتم RAKE هم، به عبارات طولانی‌تر وزن‌های بیشتری اختصاص داده می‌شود و در نتیجه، عبارات طولانی‌تر نسبت به عبارات کوتاه‌تر ترجیح داده می‌شوند. این در حالی است که حالات بسیاری وجود دارد که در آن عبارات کوتاه‌تر جهت انتخاب شدن به‌عنوان عبارات کلیدی مناسب‌تر هستند. به‌طور مثال، در مقایسه بین دو عبارت «امر آموزش و پرورش» و «آموزش و پرورش»، عبارت دوم عبارت مناسب‌تری برای کلیدواژه است؛ زیرا عبارت «امر آموزش و پرورش» توضیح اضافی دارد و حضور کلمهٔ «امر» در این عبارت ارزش معنایی مهم و خاصی ندارد و مفهوم جدیدی تولید نمی‌کند. بنابراین، طولانی کردن عبارت کلیدی در این مورد ضرورتی ندارد. این در حالی است که در الگوریتم RAKE این موضوع در نظر گرفته نشده و همواره به عبارات طولانی‌تر وزن بیشتری اختصاص داده می‌شود.

برای حل این مشکل پیشنهاد می‌شود که معیار طول عبارات در رابطه‌های وزن‌دهی در نظر گرفته شود. به بیان دیگر، رابطه‌ها از لحاظ طول عبارت نرمال‌سازی شود تا امتیاز عبارات بدون تأثیرگذاری طول عبارت‌ها به‌دست آید. در این حالت است که

تنها طولانی بودن عبارات به‌عنوان معیار برتری آن‌ها محسوب نخواهد شد، بلکه امتیاز و اهمیت کلمات تشکیل‌دهندهٔ عبارت، بدون توجه به طول عبارت نیز در نظر گرفته می‌شود. بنابراین، برای کلمهٔ w_j و عبارت p_i رابطه‌های جدید به‌صورت زیر تعریف می‌شوند:

$$deg'_w(w_j) = \frac{deg_w(w_j)}{length(p_i.w_j)} \quad (۵)$$

$$deg'_w(p_j) = \sum_{\substack{j=1 \\ p_i \epsilon p_w}}^{n} \frac{deg_w(w_j)}{length(p_i.w_j)} \quad (۶)$$

$$score'_w(w_j) = \frac{deg'_w(w_j)}{freq(w_j)} \quad (۷)$$

که در آن $deg'_w(w_j)$ درجهٔ جدید برای کلمهٔ w_j، و $score'_w(w_j)$ امتیاز جدید برای کلمهٔ w_j است. در پایان، امتیاز هر عبارت بر پایهٔ درجهٔ جدید طبق رابطهٔ زیر محاسبه می‌شود:

$$score'_{RAKE}(p_i) = \sum_{j=1}^{n} score'_w(w_j) \quad (۸)$$

وزن‌دهی به عبارات تکراری

فرض کنیم سه عبارت «آموزش و پرورش»، «امر آموزش و پرورش» و «امر آموزش و پرورش کودکان» در متن مورد نظرمان وجود دارند. بنا بر راهکار پیشنهادشده در قسمت قبل، توضیحات اضافی در عبارات کلیدی برای ما خوشایند نیست. بنابراین، لزوماً نباید به عبارات طولانی‌تر وزن بیشتری اختصاص داده شود. بنا بر رابطهٔ (۸)، عبارات طولانی لزوماً امتیاز بالایی نخواهند داشت. اکنون حالتی را می‌توان در نظر گرفت که این عبارات به دفعات در متن تکرار شده باشند. به‌طور مثال، عبارت «آموزش و پرورش»، ۳ بار، عبارت «امر آموزش و پرورش»، ۲ بار و عبارت «امر آموزش و پرورش کودکان»، ۱۰ بار در متن استفاده شده باشد. در این شرایط به نظر می‌رسد که وجود کلمهٔ «کودکان» از نظر نویسنده در این عبارت اهمیت داشته و مفهوم کامل‌تر و متفاوتی را منتقل می‌کند. بنابراین، می‌توان بر پایهٔ دفعات تکرار، وزنی برای هر عبارت نیز در نظر گرفت تا ارزش عبارات به‌طور درست‌تری تعیین شود. در مدل پیشنهادی، به‌جای محاسبهٔ امتیاز عبارت بر پایهٔ

الگوریتم RAKE، پیشنهاد می‌شود که فراوانی عبارت نیز در نظر گرفته شود. بنابراین:

$$score_{new}(p_i) = score'_{RAKE}(p_i) \times freq(p_i) \quad (۹)$$

استفاده از مفهوم انحراف استاندارد

شاخص انحراف استاندارد مهم‌ترین شاخص پراکندگی است. در اصل، انحراف استاندارد اندازه‌ای از میزان انحرافی است که یک توزیع از اعداد از میانگین خود دارند. منطق انحراف استاندارد نشان دادن متوسط میزان فاصلهٔ هر مورد از میانگین است. تفسیر انحراف استاندارد به این صورت است که هرچه مقدار انحراف استاندارد بیشتر باشد، پراکندگی اعداد از میانگین‌شان هم بیشتر است. برای نمونه، N عضو به‌صورت $x_1x_2x_{3\dots}x_N$ با میانگین μ، **انحراف استاندارد هر عضو از میانگین برابر است با:**

$$SD_i(x) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \quad (۱۰)$$

با توجه به توضیحات فوق دربارهٔ مفهوم انحراف استاندارد، در ادامه، در محاسبات وزن‌دهی عبارات کاندید در مدل پیشنهادی از این معیار استفاده شده است. اگر یک عبارت به‌صورت ترکیبی از کلمات $w_1w_2w_{3\dots}w_N$ باشد و این کلمات بر پایهٔ معیار امتیازدهی الگوریتم RAKE امتیازهایی به‌صورت $s_1.s_2.s_3 \dots .s_N$ داشته باشند، در مدل پیشنهادی، به‌جای قرار دادن امتیاز کلمات، که حاصل تقسیم تعداد کلمات بر درجهٔ آن‌هاست، انحراف استاندارد کلمات از میانگین امتیازهایشان قرار داده می‌شود. یعنی امتیازات عبارت فوق به شکل $SD_1.SD_2.SD_3 \dots .SD_N$ تغییر می‌کند و در نتیجه، امتیاز کل عبارت برابر با مجموع انحراف استانداردهای کلمات تشکیل‌دهندهٔ عبارت یعنی $\sum_1^N SD_i$ خواهد بود.

این راهکار برگرفته از راهکار پیشنهادی است که Siddiqi & Sharan (2018) در مطالعهٔ خود نیز آن را پیشنهاد کردند. علت اصلی این انتخاب این است که عباراتی که در کلیدواژه هستند، به‌طور معمول، رفتاری متفاوت از دیگر کلمات دارند و امتیازشان هم متفاوت از امتیاز دیگر عبارات است. بنابراین، از مفهوم انحراف استاندارد برای این منظور استفاده می‌شود.

امتیاز نهایی هر عبارت بر پایهٔ مدل پیشنهادی

با ترکیب سه راهکار پیشنهادی، نحوهٔ امتیازدهی به عبارات در مدل RAKE بهبودیافته

برای زبان فارسی به‌صورت زیر است:

$$Score_{proposed}(p_i) = \sum_{j=1}^{N} SD_i\ (score'_{RAKE}(w_j)) \times freq(p_i) \quad (۱۲)$$

که در آن فرض بر این است که عبارت p_i دارای کلمات $w_1 w_2 w_3 \ldots w_N$ است. در رابطهٔ پیشنهادی برای محاسبهٔ امتیاز هر عبارت، سه راهکار پیش‌گفته در نظر گرفته شده است:

- ◇ راهکار ۱: نرمال‌سازی طول عبارات با استفاده از امتیاز جدید برای کلمات $(score'_{RAKE}(w_j))$
- ◇ راهکار ۲: در نظر گرفتن فروانی عبارت $(freq(p_i))$
- ◇ راهکار ۳: در نظر گرفتن انحراف استاندارد امتیاز کلمات $(\sum_{j=1}^{N} SD_i\ (score'_{RAKE}(w_j)))$

پیاده‌سازی و ارزیابی الگوریتم RAKE بهبودیافته

در این پژوهش برای پیاده‌سازی الگوریتم از زبان «پایتون» استفاده شده است. این زبان به‌دلیل داشتن بسته‌های قدرتمند و کاربردی، به‌ویژه برای پردازش زبان بسیار مناسب است. در ادامه، مجموعهٔ داده‌هایی که برای پیاده‌سازی الگوریتم استفاده شده، معرفی می‌شوند. پس از آن، نتایجِ پیاده‌سازی الگوریتم پیشنهادی از نظر دقت، بازخوانی و معیار F، بررسی می‌شوند.

مجموعهٔ داده‌ها

مجموعهٔ دادهٔ مورد استفاده در این پژوهش بخشی از فراداده‌های متون علمی است که در پایگاه اطلاعات علمی ایران (گنج)[1] قابل دسترسی است. این متون، پایان‌نامه‌ها و رساله‌های فارسی است که در بخشی از فراداده‌های آن‌ها از عنوان، چکیده، و کلیدواژه‌ها استفاده شده است. مجموعهٔ داده‌های استفاده شده شامل ۵۰۰ متن علمی است که دو نمونه از آن‌ها در جدول ۱، آمده است. کلیدواژه‌های هر سند، مجموعه‌ای از کلیدواژه‌هایی است که نویسندهٔ سند و متخصص نمایه‌سازی به آن اختصاص داده است. هر پایان‌نامه و رساله‌ای که در پایگاه «گنج» ثبت می‌شود، توسط متخصص موضوعی و نمایه‌ساز کنترل

1. http://ganj.irandoc.ac.ir

و ویرایش می‌شود و در صورت نیاز، کلیدواژه‌های بیشتر و تخصصی‌تری به آن اختصاص داده می‌شود. در این پژوهش همهٔ این کلیدواژه‌ها به‌عنوان کلیدواژه‌های سند در نظر گرفته می‌شود. هر سند به‌طور متوسط و بدون در نظر گرفتن عنوان از ۲۵۰ تا ۳۵۰ کلمه تشکیل شده است. عنوان هر یک از اسناد هم به‌طور متوسط دربرگیرندهٔ ۱۰ تا ۲۰ کلمه است.

جدول ۱. نمونه‌ای از داده‌های استفاده شده

عنوان	چکیده	تعداد کلمات (چکیده و عنوان)	واژه‌ها و عبارات کلیدی
بررسی تطبیقی نقش شخص ثالث در عدم اجرای قرارداد و آثار آن در حقوق ایران و اسناد بین‌المللی	یکی از مباحث مهمی که در حقوق قراردادها مطرح می‌شود، بحث نقش شخص ثالث در عدم اجرای قرارداد است. با توجه به تخصصی‌شدن کارها، امروزه در عرصهٔ تجاری در بسیاری از موارد قرارداد توسط طرفین بالمباشره انجام نمی‌شود، بلکه تمام یا بخشی از تعهدات طرفین توسط اشخاص ثالث اجرا می‌شوند. بدیهی است در تعهداتی که اجرای آن‌ها قایم به شخص نباشد، ...	۲۸۶	نظام حقوقی، مسئولیت (حقوقی)، ایران، معافیت، عدم امکان اجرای قرارداد، شخص ثالث، اسناد بین‌المللی
تحلیلی بر ظرفیت‌پذیری گردشگری در مقاصد شهری؛ مطالعهٔ موردی: شهر مراغه	کم‌وکیف محصولات مقاصد گردشگری، نقش بسیار مهمی در توسعه و ارتقای محصول و مقصد گردشگری دارند و از سوی دیگر، برنامه‌ریزان گردشگری به این نتیجه واقف گفته‌اند که رقابت در گردشگری بدون توسعهٔ پایدار امری مضر است، چرا که جاذبه‌های گردشگری ...	۳۴۸	مراغه، ظرفیت‌پذیری حمل‌ونقل، گردشگری، جاذبه جهانگردی، ظرفیت‌سازی، عوامل اقتصادی، بخش کشاورزی

پیاده‌سازی

برای پیاده‌سازی الگوریتم پیشنهادی از زبان برنامه‌نویسی «پایتون»، «کتابخانهٔ جنسیم»[1] و «کتابخانهٔ هضم» استفاده شده است. برای ارزیابی الگوریتم پیشنهادی و مقایسهٔ عملکرد آن، الگوریتم RAKE در چند حالت مختلف دیگر نیز پیاده‌سازی شده است. در جدول ۲، چند نمونه از نتایج الگوریتم پیشنهادی و الگوریتم اصلی RAKE آمده است.

1. Gensim

جدول ۲. نتایج الگوریتم RAKE بهبودیافته و الگوریتم RAKE اصلی

عنوان	کلیدواژه‌های اصلی	کلیدواژه‌های الگوریتم RAKE بهبودیافته	کلیدواژه‌های الگوریتم RAKE اصلی
بررسی تأثیر اجتناب مالیاتی بر هزینهٔ سرمایهٔ حقوق مالکانه شرکت	اجتناب از مالیات، هزینهٔ سرمایه، حقوق صاحبان سهام، سهامدار نهادی، بورس اوراق بهادار تهران، صورت مالی	بورس اوراق بهادار تهران، هستهٔ مرکزی نظام مالیات، حقوق صاحبان سهام، توسعهٔ سیستم اطلاعات مالیات، مالیات شهر تهران، اجتناب از مالیات	شرایط ابهام ذاتی اجتناب مالیات سوء، هزینهٔ حقوق صاحبان سهام، مصرف نادرست منابع مالک، اجتناب مالیات انتقال منابع، برنامهٔ اجتناب مالیات، بررسی تأثیر اجتناب مالیات
تحلیلی بر ظرفیت‌پذیری گردشگری در مقاصد شهری؛ مطالعهٔ موردی: شهر مراغه	ظرفیت‌سازی، گردشگری، عوامل اقتصادی، بخش کشاورزی، مراغه، ظرفیت‌پذیری حمل‌ونقل گردشگری، جاذبهٔ جهانگردی	گردشگری، شهر مراغه، ظرفیت پذیرش کافی، برآورد ظرفیت پذیرش منطقه، بخش کشاورزی، جاذبه جهانگردی	مقصد گردشگری شهر مراغه، ظرفیت پذیرش کافی، جامعهٔ میزبان تفاوت معنادار، برآورد ظرفیت پذیرش منطقه، مقصد گردشگری نقش، توجه جامعه علم

با مقایسهٔ نتایج حاصل از الگوریتم RAKE اصلی و RAKE پیشنهادی بهبودیافته روی مجموعهٔ داده‌ها می‌توان به نتایج زیر رسید:

- ◇ برخی از عبارات کلیدی که از پیاده‌سازی الگوریتم RAKE اصلی به‌دست آمده‌اند، بی‌معنا هستند؛ به‌عنوان مثال، عبارت «فقیر مناطق محسوب» را در نظر بگیرید. این عبارت که از پردازش جملهٔ «... از لحاظ بعد مسکن منطقه ۴، فقیرترین مناطق محسوب می‌شوند ...» حاصل شده است، فاقد ارزش معنایی است. از طرف دیگر، به‌طور کلی این عبارت محتوای کلیدی برای این متن ندارد. همچنین، کلمهٔ «محسوب» بخشی از فعل «محسوب می‌شود» بوده و به‌صورتی بی‌نظم و به‌هم‌خورده در این عبارت کلیدی نمایان شده است. این در حالی است که این دست از مشکلات در الگوریتم بهبودیافته کمتر مشاهده می‌شود.
- ◇ در الگوریتم اصلی RAKE عبارات طولانی‌تر به عبارات کوتاه‌تر، به‌طور مطلق ترجیح داده شده‌اند، در حالی که با توجه به تغییراتی که در الگوریتم بهبودیافته ایجاد کردیم، این مشکلات کمتر وجود دارد.
- ◇ در نتایج حاصل از پیاده‌سازی الگوریتم اصلی RAKE روی متون، گاهی با عبارات کلیدی تکراری مواجه هستیم. به‌طور مثال، در یکی از متون مشاهده می‌کنیم که

دو عبارت «خلق سیستم‌های پیچیده» و «سیستم‌های پیچیده» هر دو در عبارات کلیدی وجود دارند، در صورتی که به وجود هیچ‌یک از آن دو نیازی نیست. این در حالی است که با توجه به تغییرات اعمال‌شده در الگوریتم بهبودیافته، با این مشکل کمتر مواجه هستیم.

ارزیابی و مقایسه

روش پیشنهادی را از دو جنبه ارزیابی می‌کنیم:

- ارزیابی روش پیشنهادی برای بهبود الگوریتم RAKE، با مقایسهٔ نتایج الگوریتم بهبودیافته با نتایج الگوریتم RAKE اصلی و بررسی نحوهٔ اثرگذاری هر یک از راهکارهای پیشنهادی؛
- ارزیابی الگوریتم RAKE بهبودیافته و مقایسهٔ الگوریتم پیشنهادی با تعدادی از روش‌های اصلی استخراج کلیدواژه.

در ادامه، نتایج این ارزیابی آمده است.

ارزیابی الگوریتم RAKE پیشنهادی و مقایسهٔ آن با الگوریتم اصلی

برای ارزیابی و مقایسهٔ عملکرد الگوریتم پیشنهادی و بررسی تأثیر تغییراتی که در الگوریتم بهبودیافته پیشنهاد شده، از نظر معیار دقت، بازخوانی و معیار F، آزمایش‌های مختلف انجام شده است.

دقت برابر است با نسبت تعداد عبارات کلیدی درست استخراج‌شده به کل عبارات کلیدی استخراج‌شده. بازخوانی عبارت است از نسبت تعداد عبارات کلیدی درست استخراج‌شده به عبارات کلیدی اصلی متن. بنابراین، اگر X_A کلمات کلیدی استخراج‌شده توسط الگوریتم مورد نظر و X_B کلمات کلیدی استاندارد متن مورد نظر (کلیدواژه‌های استاندارد پایگاه «گنج») باشد، آنگاه دقت، بازخوانی و معیار F به‌صورت زیر تعریف می‌شود:

$$Precision = \frac{X_A \cap X_B}{X_B} \quad (۱۳)$$

$$Recall = \frac{X_A \cap X_B}{X_A} \quad (۱۴)$$

$$F - measure = 2 \times \frac{precison \times Recall}{precision + Recall} \quad (۱۵)$$

دقت و بازخوانی محاسبه‌شده بر اساس مقایسهٔ دقیق بین کلیدواژه‌ها نیست، زیرا اگر مبنا را مقایسهٔ دقیق بگذاریم، در برخی از موارد ممکن است با وجود یک فاصله یا نیم‌فاصله و حتی «ی» بین کلیدواژه‌های استخراج‌شده و کلیدواژه‌های اصلی اختلاف ایجاد کند و بنا بر عدم تطابق آن‌ها گذاشته شود. این مبنای مقایسه در برخی از پژوهش‌های حوزهٔ استخراج کلیدواژه نیز به‌کار گرفته شده است (Lau & Baldwin 2016; Siddiqi & Sharan 2018). بنابراین، در این مقایسه اگر فاصله بین دو کلیدواژه از یک مقداری کمتر باشد، یکسان در نظر گرفته می‌شوند. در ادامه، نتایج پیاده‌سازی الگوریتم را در پنج حالت زیر به تفکیک بررسی می‌کنیم:

- ◇ الگوریتم RAKE اصلی؛
- ◇ الگوریتم RAKE تغییریافته پس از اعمال راهکار ۱ (نرمال‌سازی درجهٔ عبارات نسبت به طول آن‌ها)؛
- ◇ الگوریتم RAKE تغییریافته پس از اعمال راهکار ۲ (وزن‌دهی به‌عبارات تکراری)؛
- ◇ الگوریتم RAKE تغییریافته پس از اعمال راهکار ۳ (استفاده از مفهوم انحراف استاندارد)؛
- ◇ الگوریتم RAKE بهبودیافته (پس از اعمال هر سه راهکار ۱ و ۲ و ۳، روش پیشنهادی).

جدول ۳. میانگین مقادیر دقت، بازخوانی و معیار F در حالت‌های مختلف

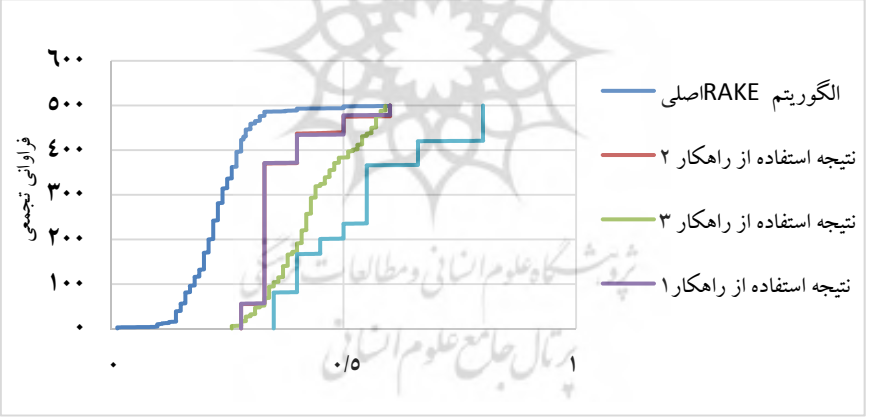
الگوریتم	دقت	بازخوانی	معیار F
الگوریتم RAKE اصلی	۰/۲۳	۰/۲۲	۰/۲۱
الگوریتم RAKE و راهکار ۱	۰/۳۶	۰/۲۲	۰/۲۸
الگوریتم RAKE و راهکار ۲	۰/۳۶	۰/۲۲	۰/۲۸
الگوریتم RAKE و راهکار ۳	۰/۴۲	۰/۲۹	۰/۳۵
الگوریتم RAKE بهبودیافته	۰/۵۳	۰/۳۷	۰/۴۴

در شکل ۵، بررسی نمودار فراوانی تجمعی مقادیر دقت در ۵ حالت به‌تفکیک نشان می‌دهد که از بین روش‌های نشان داده‌شده، نمودار فراوانی نتایج حاصل از پیاده‌سازی الگوریتم نهایی بهبودیافته توسط هر سه راهکار، در مقادیر دقت بالاتری نسبت به دیگر روش‌ها متمرکز است. به بیان دیگر، با پیاده‌سازی این روش روی مجموعهٔ متون، تعداد

بیشتری دقت با مقادیر بالاتر به‌دست می‌آید. به این ترتیب، روش‌های به‌کاررفته را بر حسب میزان دقت آن‌ها به‌صورت زیر رتبه‌بندی می‌کنیم:

۱. الگوریتم RAKE بهبودیافته (پس از اعمال هر سه راهکار ۱ و ۲ و ۳ با هم)؛
۲. الگوریتم RAKE تغییریافته پس از اعمال راهکار ۳ (استفاده از مفهوم انحراف استاندارد)؛
۳. الگوریتم RAKE تغییریافته پس از اعمال راهکار ۱ و الگوریتم RAKE تغییریافته پس از اعمال راهکار ۲ (عملکرد این دو راهکار به‌تنهایی تقریباً مشابه یکدیگر است)؛
۴. الگوریتم RAKE اصلی.

در شکل ۶، با بررسی نمودار فراوانی تجمعی مقادیر بازخوانی در ۵ حالت به‌تفکیک می‌توان دریافت که از بین روش‌های نشان داده‌شده، نمودار فراوانی نتایج حاصل از پیاده‌سازی الگوریتم نهایی بهبودیافته توسط هر سه راهکار، در مقادیر بازخوانی بالاتری نسبت به دیگر روش‌ها متمرکز است. به بیان دیگر، با پیاده‌سازی این روش روی مجموعهٔ متون، تعداد بیشتری بازخوانی با مقادیر بالاتر به‌دست می‌آید.



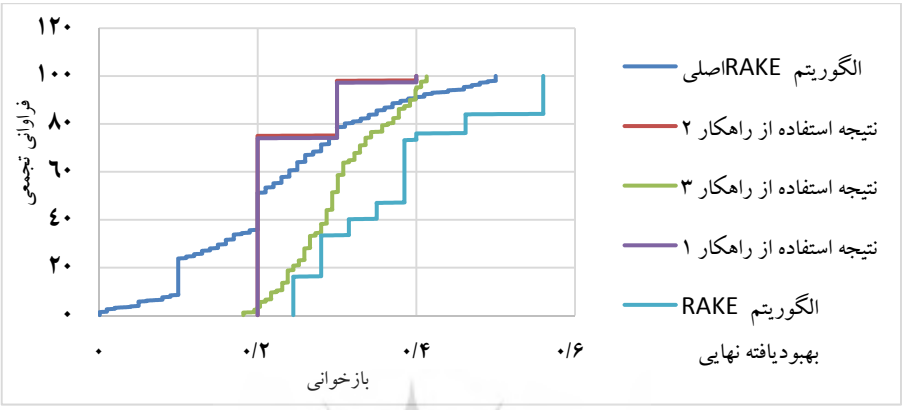
شکل ۵. نمودار فراوانی تجمعی دقت در ۵ حالت مورد بررسی

به این ترتیب، روش‌های به‌کاررفته را بر حسب میزان بازخوانی آن‌ها به‌صورت زیر رتبه‌بندی می‌کنیم:

۱. الگوریتم RAKE بهبودیافته (پس از اعمال هر سه راهکار ۱ و ۲ و ۳ با هم)؛
۲. الگوریتم RAKE تغییریافته پس از اعمال راهکار ۳ (استفاده از مفهوم انحراف استاندارد)؛
۳. الگوریتم RAKE تغییریافته پس از اعمال راهکار ۱ و الگوریتم RAKE تغییریافته پس از

اعمال راهکار ۲ (عملکرد این دو راهکار به‌تنهایی تقریباً مشابه یکدیگر است)؛

۴. الگوریتم RAKE اصلی.



شکل ٦. نمودار فراوانی تجمعی بازخوانی در ٥ حالت مورد بررسی

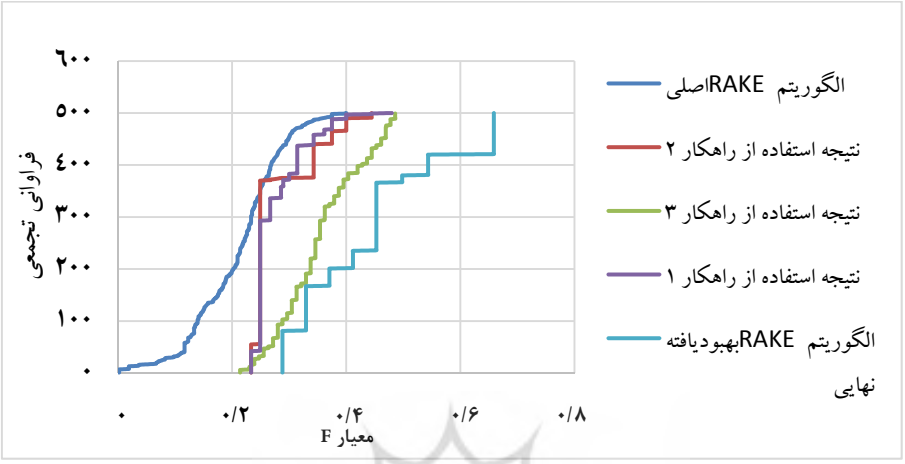
در شکل ۷، بررسی نمودار فراوانی تجمعی مقادیر معیار F در ۵ حالت به‌تفکیک نشان می‌دهد که از بین روش‌های نشان داده‌شده، نمودار فراوانی نتایج حاصل از پیاده‌سازی الگوریتم نهایی بهبودیافته توسط هر سه راهکار، در مقادیر معیار F بالاتری نسبت به دیگر روش‌ها متمرکز است. به بیان دیگر، با پیاده‌سازی این روش روی مجموعه متون، تعداد بیشتری بازخوانی با مقادیر بالاتر به‌دست می‌آید. به این ترتیب، روش‌های به‌کاررفته را بر حسب مقادیر معیار F آن‌ها به‌صورت زیر رتبه‌بندی می‌کنیم:

۱. الگوریتم RAKE بهبودیافته (پس از اعمال هر سه راهکار ۱ و ۲ و ۳)؛
۲. الگوریتم RAKE تغییریافته پس از اعمال راهکار ۳ (استفاده از مفهوم انحراف استاندارد)؛
۳. الگوریتم RAKE تغییریافته پس از اعمال راهکار ۱ و الگوریتم RAKE تغییریافته پس از اعمال راهکار ۲ (عملکرد این دو راهکار به‌تنهایی تقریباً مشابه یکدیگر است)؛
۴. الگوریتم RAKE اصلی.

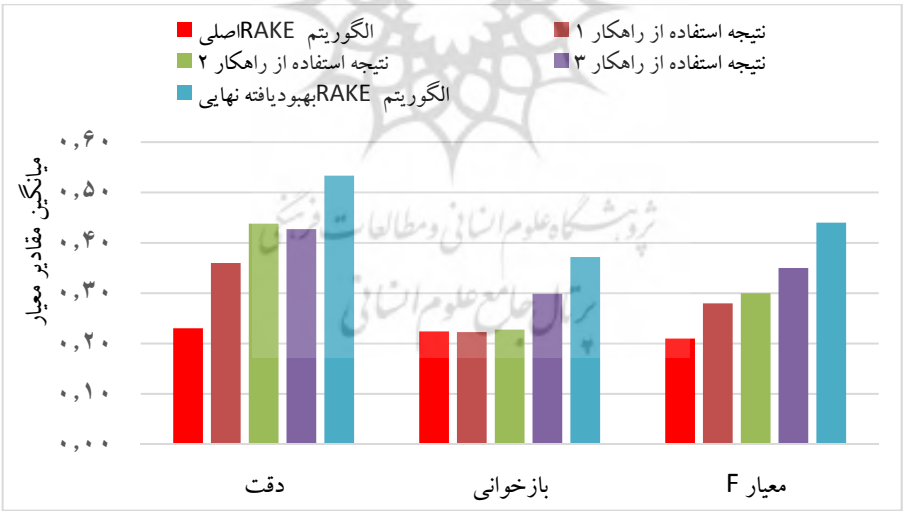
میانگین سه شاخص «دقت»، «بازخوانی» و«معیار F» در ۵ حالت ذکرشده در شکل ۸ آمده است.

طبق نتایج به‌دست‌آمده میانگین هر سه معیار در الگوریتم بهبودیافتهٔ نهایی نسبت به دیگر روش‌ها بیشتر است. با استفاده از تغییرات اعمال‌شده دقت حاصل از پیاده‌سازی

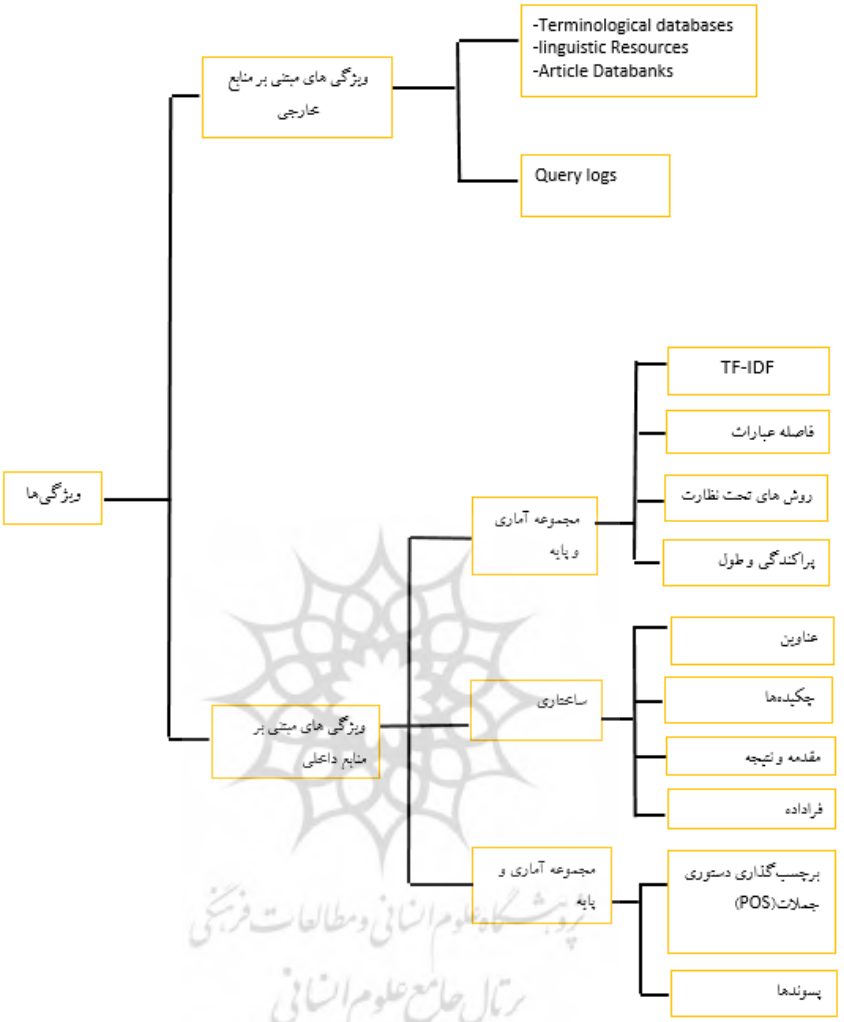
الگوریتم بهبودیافته روی متون فارسی در حدود ۳۰ درصد، بازخوانی در حدود ۱۵ درصد و معیار F در حدود ۲۰ درصد نسبت به الگوریتم اصلی افزایش می‌یابد.



شکل ۷. نمودار فراوانی تجمعی معیار F در ۵ حالت مورد بررسی



شکل ۸. هیستوگرام میانگین دقت، بازخوانی و معیار F در ۵ حالت مورد بررسی



شکل ۹. ویژگی‌های کلی تشکیل‌دهندهٔ یک عبارت کلیدی (Alami Merrouni, Frikh & Ouhbi 2020)

مقایسهٔ عملکرد روش پیشنهادی با سایر روش‌های استخراج کلیدواژه

روش‌های موجود برای استخراج کلیدواژه بسیار متنوع هستند. از دیدگاهی، روش‌های استخراج کلیدواژه در سه دستهٔ کلی قرار می‌گیرند: روش‌های باناظر، روش‌های بدون ناظر و روش‌های یادگیری عمیق. هر یک از این روش‌ها مزایا و معایبی دارند. از دیدگاه‌های دیگری نیز می‌توان روش‌های استخراج کلیدواژه را دسته‌بندی کرد؛ به‌عنوان مثال، از نظر ویژگی‌های مورد استفاده برای استخراج کلیدواژه؛ از نظر استخراج از متون کوتاه یا

بلند؛ و یا استخراج از چند سند یا از یک سند. در شکل ۹، یک نمای کلی از دسته‌بندی ویژگی‌های پایه که یک عبارت کلیدی را مشخص می‌کند، آورده شده است (Alami Merrouni, Frikh & Ouhbi 2020). در این دسته‌بندی، در مرتبهٔ اول، روش‌های استخراج کلیدواژه را از این نظر که آیا در آن‌ها از منابعی غیر از خود سند استفاده شده یا خیر، مورد توجه است. پس از آن بسته به اینکه از چه ویژگی‌هایی برای استخراج کلیدواژه استفاده می‌شود، می‌توان روش‌ها را در دسته‌های مختلف قرار داد.

برای انتخاب روش مناسب جهت مقایسه با الگوریتم پیشنهادی بر مبنای این دسته‌بندی و نیز بر پایهٔ مطالعهٔ مروری که Alami Merrouni, Frikh & Ouhbi (2020) انجام داده‌اند، عمل می‌کنیم. از بین انواع روش‌هایی که در آن مطالعهٔ مروری آمده، روش‌هایی را در نظر می‌گیریم که بر مبنای استخراج کلیدواژه از یک سند عمل می‌کنند. افزون بر آن، روش‌هایی برای مقایسه مناسب هستند که به‌طور معمول برای متون کوتاه نتایج بهتری را ارائه می‌دهند. همچنین، به‌دلیل اینکه روش پیشنهادی ما یک روش بدون ناظر است، آن‌ها را با روش‌های بدون ناظر مقایسه می‌کنیم.

از بین انواع روش‌های موجود، روش آماری TF-IDF (فراوانی کلمه-فراوانی محتوای معکوس) یک روش پایه‌ای در حوزهٔ استخراج کلیدواژه است که بر اساس فراوانی کلمات عمل می‌کند. همچنین، در بین روش‌های موجود، روش TextRank یک روش بدون ناظر مبتنی بر گراف است که به‌طور معمول، برای متون کوتاه نیز به‌کار می‌رود (Mihalcea Rada 2004). افزون بر آن، روش YAKE یک روش جدید برای استخراج کلیدواژه است که در بین روش‌های بدون ناظر نتایج خوبی ارائه کرده است (Campos et al. 2020). بنابراین، در ادامه، نتایج الگوریتم پیشنهادی را روی داده‌های این پژوهش، برای سه الگوریتم TF-IDF، الگوریتم TextRank، و الگوریتم YAKE! بررسی می‌کنیم.

برای پیاده‌سازی روش TF-IDF از پژوهش Tripathi (2018) استفاده شده است. در ادامه، در جدول ۴، میانگین مقادیر سه معیار ارزیابی دقت، بازخوانی و معیار F این الگوریتم‌ها در زبان فارسی برای داده‌های این پژوهش آمده است.

جدول ٤. میانگین مقادیر دقت و بازخوانی و معیار F در الگوریتم‌های مختلف

نام الگوریتم	دقت	بازخوانی	معیار F
الگوریتم TF-IDF	۰/۰۷	۰/۱۰	۰/۰۸
الگوریتم TextRank	۰/۲۰	۰/۰۹	۰/۱۲
الگوریتم YAKE!	۰/۲۶	۰/۱۲	۰/۱۷
الگوریتم RAKE	۰/۲۳	۰/۲۲	۰/۲۱
الگوریتم پیشنهادی (RAKE بهبودیافته)	۰/۵۳	۰/۳۷	۰/۴۴

الگوریتم‌های TextRank، TF-IDF و RAKE بر پایهٔ فراوانی کلمات در متن عمل می‌کنند. TF-IDF روشی آسان و پرکاربرد برای پیدا کردن چگالی کلیدواژه‌ها در بازیابی اطلاعات است. خروجی‌های این الگوریتم اغلب تک‌کلمه هستند، نه عبارات کلیدی. این در حالی است که در بسیاری از متون، عبارات کلیدی که شامل چند کلمه باشند، بسیار گویاتر و کارآمدتر از تک‌کلمات هستند. همچنین، در بسیاری از موارد لزوماً کلمه‌ای که بیشترین فراوانی را دارد، نمی‌تواند بیانگر مفهوم اصلی متن باشد.

الگوریتم TextRank یک گراف از کلمات و روابط بین آن‌ها در یک سند متنی ایجاد می‌کند. سپس، مهم‌ترین رأس‌های گراف (کلمات) را بر پایهٔ نمراتشان شناسایی می‌کند. از مشکلات این الگوریتم می‌توان به‌سرعت پایین‌تر آن نسبت به الگوریتم‌های دیگر اشاره کرد (Mihalcea Rada 2004).

بنابراین، با توجه به توضیحات فوق و میانگین مقادیر دقت، بازخوانی و معیار F در این سه الگوریتم در جدول ۴، الگوریتم RAKE بهبودیافته عملکرد بهتری بر روی اسناد فارسی دارد. لازم به ذکر است که به‌طور معمول، روش‌های استخراج کلیدواژه دقت و بازخوانی بالایی را به‌نسبت سایر روش‌های پردازش زبان طبیعی ندارند. به‌عنوان مثال، در پژوهش مروری Alami Merrouni, Frikh & Ouhbi. (2020)، دقت و بازخوانی روش‌های استخراج کلیدواژه بررسی شده است. این مقادیر به‌ندرت بیش از ۵۰ درصد گزارش شده‌اند و بیشتر مقادیری بین ۲۰ تا ۳۰ درصد دارند. یکی از مهم‌ترین دلایل پایین بودن دقت و بازخوانی، چالش مقایسهٔ کلیدواژه‌های استخراج‌شده با کلیدواژه‌های اصلی است. به‌طور معمول، برای سنجش این موضوع از مقایسهٔ دقیق یا شبه‌دقیق بین کلیدواژه‌ها

استفاده می‌کنند. این در حالی است که ممکن است با این روش سنجش دو کلیدواژه با هم یکسان نباشند، لیکن از نظر معنایی با هم یکسان باشند. یکی دیگر از دلایل پایین بودن این مقادیر، محدود بودن تعداد کلیدواژه‌های اصلی است که به‌طور معمول، بیش از ۸ عدد نیستند. برای رفع این دسته از چالش‌ها، روش‌های معنایی و مقایسهٔ غیردقیق نیز پیشنهاد شده (Siddiqi & Sharan 2018) که می‌توان آن‌ها را در پژوهش‌های آتی مورد توجه قرار داد.

نتیجه‌گیری و مطالعات آتی

در این مقاله روشی برای بهبود استخراج کلمات کلیدی از متن فارسی بر پایهٔ الگوریتم RAKE ارائه شده است. برای این منظور، از ترکیب روش‌های پردازش زبان طبیعی مانند پیش‌پردازش متن، برچسب‌گذاری دستوری کلمات، ریشه‌یابی، حذف ایست‌واژه‌ها و ... با مفاهیم آماری چون انحراف استاندارد استفاده شد. وجه تمایز این پژوهش در ایجاد تغییراتی در روش‌های وزن‌دهی به عبارات کلیدی مانند استفاده از معیار طول عبارات در رابطه‌ها و بهره‌گیری از معیار انحراف استاندارد است. این تغییرات باعث شد که در وهلهٔ اول، عبارات صرفاً به‌دلیل طولانی بودن وزن بیشتری دریافت نکنند، بلکه ارزش کلمات مستقل از طول عبارت در نظر گرفته شود. این امر از وجود توضیحات اضافی در کلیدواژه‌های خروجی مدل جلوگیری می‌کند. افزون بر این، با تغییرات اعمال‌شده در مدل، از عبارات کلیدی تکراری جلوگیری شده است. همچنین، مشاهده کردیم که دو معیار ارزیابی دقت و بازخوانی نیز در مدل جدید ارائه‌شده به‌طور میانگین ۳۰ و ۱۵ درصد افزایش یافت. الگوریتم ارائه‌شده در این پژوهش می‌تواند در فرایند نمایه‌سازی اسناد پایگاه «گنج» به‌کار رود و نمایه‌ساز می‌تواند از نتایج این الگوریتم برای تسریع عملیات نمایه‌سازی و افزایش دقت آن استفاده کند.

با توجه به اهمیت زیاد دستیابی به کلیدواژه‌های غنی در متون فارسی، تلاش جهت افزایش هرچه بیشتر دقت و بهبود عملکرد الگوریتم‌های استخراج کلیدواژه دارای اهمیت است. به این منظور، راهکارهایی از جمله تهیهٔ لیست‌هایی جامع‌تر از ایست‌واژه‌ها و دسته‌بندی موضوعی ایست‌واژه‌ها (علمی، تاریخی، جغرافیایی و ...) در بهبود فرایند استخراج کلیدواژه‌ها تأثیرگذار خواهد بود. همچنین، استفاده از مدل‌های برچسب‌گذار دستوری دقیق‌تر و با درصد خطای کمتر منجر به افزایش دقت نتایج خواهد شد.

سپاسگزاری

این پژوهش با حمایت آزمایشگاه تعامل انسان و ماشین «ربوداک» در پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) به انجام رسیده است. از پژوهشگران آزمایشگاه که در این پژوهش ما را یاری کردند، سپاسگزاریم.

فهرست منابع

باسره، مریم، ولی درهمی، و سجاد ظریف‌زاده. ۱۳۹۶. ارائهٔ روشی برای استخراج خودکار عبارات کلیدی از اخبار وب. *مجله مهندسی برق دانشگاه تبریز* ۴۷ (۸۱): ۸۵۷-۸۶۶.

بشیری، حسن، فاطمه کربلائی، و شیرین موسوی. ۱۳۸۴. طراحی و ارزیابی نمایه‌ساز خودکار متون فارسی. *یازدهمین کنفرانس بین‌المللی کامپیوتر تهران*: انجمن کامپیوتر ایران، پژوهشگاه دانش‌های بنیادی.

تشکری، مسعود، و محمدرضا میبدی. ۱۳۸۲. ساخت یک نمایه‌ساز خودکار برای متون فارسی. *یازدهمین کنفرانس مهندسی برق شیراز*: دانشگاه شیراز.

توکلی‌زاده راوری، محمد. ۱۳۹۴. مدل دومرحله‌ای شکاف-گلچین برای نمایه‌سازی خودکار متون فارسی. *تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی* ۲۱ (۱): ۱۳-۴۰.

خطیر، اشکان، و سهیل گنجه‌فر. ۱۳۹۷. تحلیل توزیع و تمرکز کلیدواژه‌های پارساها: میزان تطابق با توصیفگرها، عنوان، و چکیده. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۴ (۱): ۴۱۱-۴۲۸.

سمائی، سید مهدی، و بهروز رسولی. ۱۳۹۹. *شناسایی ویژگی‌های زبان علم در مدارک علمی فارسی*. تهران: پژوهشگاه علوم و فناوری اطلاعات ایران.

کلانتری، عاطفه، عبدالرسول جوکار، سید مصطفی فخراحمد، جواد عباسپور، مسعود مرتضوی،امیر جوادی، زهرا پوربهمن. ۱۳۹۹. استخراج کلمات و عبارات کلیدی از متون فارسی: مروری بر پژوهش‌های صورت گرفته. *پژوهشنامه پردازش و مدیریت اطلاعات* http://Jipm.irandoc.ac.ir(دسترسی در ۱۳۹۹/۷/۲۰)

محبی، آزاده، و عمار جلالی‌منش. ۱۳۹۸. *ارائهٔ روشی هوشمند برای استخراج کلیدواژه از مستندات علمی زبان فارسی بر اساس سیستم‌های پیشنهاددهنده*. تهران: پژوهشگاه علوم و فناوری اطلاعات ایران.

ویسی، هادی، و نیلوفر افلاکی. ۱۳۹۴. استخراج کلمات کلیدی متن فارسی با استفاده از آنالیز آماری. *کنفرانس بین‌المللی مهندسی و علوم کاربرد*. دبی.

References

Aggarwal, C. C., & C. Zhai. 2012. Mining Text Data. In *Springer*. https://doi.org/10.1111/j.1751-1097.1972.tb06217.x (accessed July 18, 2020)

Alami Merrouni, Z., B. Frikh, & B. Ouhbi. 2020. Automatic keyphrase extraction: a survey and trends. *Journal of Intelligent Information Systems* 54 (2): 391–424. https://doi.org/10.1007/s10844-019-00558-9

Andrade, M. A., & A. Valencia. 1998. Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. *Bioinformatics* 14 (7): 600–607. https://doi.org/10.1093/bioinformatics/14.7.600

Bayatmakou, F., A. Ahmadi, & A. Mohebi. 2017. Automatic query-based keyword and keyphrase extraction. *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, 325–330. Shiraz, Iran.

Campos, R., V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, & A. Jatowt. 2020. YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences* 509: 257–289. https://doi.org/10.1016/j.ins.2019.09.013

Chen, J., C. Chen, & Y. Liang. 2016. *Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word 133*: 114–117. https://doi.org/10.2991/aiie-16.2016.28

Haque, M. 2019. Automatic Keyword Extraction from Bengali Text Using Improved RAKE Approach. *2018 21st International Conference of Computer and Information Technology, ICCIT 2018*, 1–6. https://doi.org/10.1109/ICCITECHN.2018.8631917 Bangladesh.

Hosseinikhah, T., A. Ahmadi, & A. Mohebi. 2018. A new Persian text summarization approach based on natural language processing and graph similarity. *Iranian Journal of Information Processing Management* 33 (2): 885–914.

Lau, J. H., & T. Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. *ArXiv Preprint ArXiv:1607.05368*.

Meng, R., S. Zhao, S. Han, D. He, P. Brusilovsky, & Y. Chi. 2017. Deep keyphrase generation. *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *1*, 582–592. https://doi.org/10.18653/v1/P17-1054 Vancouver, Canada.

Mihalcea Rada, T. P. 2004. TextRank: Bringing Order into Texts. *Conference on Empirical Methods in Natural Language Processing*. Waikiki, Honolulu.

Papagiannopoulou, E., & G. Tsoumakas. 2019. A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10* (September 2018), 1–45. https://doi.org/10.1002/widm.1339 (accessed July 18, 2020)

Rose, S., D. Engel, N. Cramer, & W. Cowley. 2010. Automatic Keyword Extraction from Individual Documents. *Text Mining: Applications and Theory*, *October 2017*, 1–20. https://doi.org/10.1002/9780470689646.ch1(accessed July 18, 2020)

Sathya, R., & Z. Abraham. 2013. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence* 2 (2): 34–38. https://doi.org/10.14569/ijarai.2013.020206

Siddiqi, S., & A. Sharan. 2018. *Improved RAKE Models to Extract Keywords from Hindi Documents BT - Information Systems Design and Intelligent Applications* (V. Bhateja, B. Le Nguyen, N. G. Nguyen, S. C. Satapathy, & D.-N. Le (eds.); pp. 472–483). Singapore: Springer.

Thushara, M. G., T. Mownika, & R. Mangamuru. 2019. A comparative study on different keyword extraction algorithms. *Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019*, *Iccmc*, 969–973. https://doi.org/10.1109/ICCMC.2019.8819630 Erode, India.

Timonen, M., T. Toivanen, M. Kasari, Y. Teng, C. Cheng, & L. He. 2013. Keyword Extraction from Short Documents Using Three Levels of Word Evaluation. *Communications in Computer and Information Science* 415: 130–146. https://doi.org/10.1007/978-3-642-54105-6_9

Tripathi, M. 2018. *How to process textual data using TF-IDF in Python*. Free Code Camp. https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3/ (accessed July 18, 2020)

Zhang, C., H. Wang, Y. Liu, D. Wu, Y. Liao, & B. Wang. 2008. Automatic Keyword Extraction from Documents Using Conditional Random Fields. *Journal of Computational Information Systems* 4 (3): 1169–1180.

الهه محرابی

متولد سال ۱۳۷۷ و دانشجوی کارشناسی در رشتهٔ مهندسی صنایع در دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) است. حوزه‌های هوش مصنوعی، یادگیری ماشین، پردازش زبان‌های طبیعی و تعامل انسان و یارانه از جمله علایق پژوهشی وی است.

آزاده محبی

دارای مدرک دکتری در رشتهٔ مهندسی طراحی سیستم‌ها از دانشگاه واترلو کاناداست. ایشان هم‌اکنون استادیار پژوهشکدهٔ فناوری اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

تعامل انسان و کامپیوتر، داده‌کاوی، سیستم‌های هوشمند، بازشناسی الگو، متن‌کاوی و بازیابی اطلاعات از جمله علایق پژوهشی وی است.

عباس احمدی

دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی طراحی سیستم‌ها از دانشگاه واترلو کاناداست. ایشان هم‌اکنون دانشیار دانشکدهٔ مهندسی صنایع و سیستم‌های مدیریت دانشگاه صنعتی امیرکبیر (پلی‌تکنیک تهران) است. وی دارای مقالات متعدد در مجلات و کنفرانس‌های معتبر داخلی و خارجی است.

تحلیل و هوشمندی کسب‌وکار، سیستم‌های هوشمند، داده‌کاوی، بازیابی اطلاعات و دانش و سیستم‌های سلامت از جمله علایق پژوهشی وی است.