A Case-Based Recommender System for Persian Scientific Document Indexing

Azadeh Mohebi*
PhD in Systems Design Engineering; Assistant professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: mohebi@irandoc.ac.ir

Azadeh Fakhrzdaeh
PhD in computer image processing; Assistant professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: Fakhrzadeh@mail.irandoc.ac.ir

Marzieh Zarinbal
PhD in Industrial engineering; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran;
Email: zarinbal@irandoc.ac.ir

Iranian Journal of Information Processing and Management

Received: 09, Jan. 2023 **Accepted: 10, Apr. 2023**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 39 | No. 2 | pp. 599-626
Winter 2024
https://doi.org/10.22034/jipm.2023.704737

Abstract: Keyword extraction is a key step in document indexing. Keywords are semantic and content-based descriptors of a document, which can be used in document retrieval and representation. In databases containing scientific documents, such as Ganj in Irannian Research Institue for Information Science and Technology (IranDoc), it is even more critical to assign meaningful keywords for documents, since the documents are from different academic disciplines and contain technical terms.

As the number of scientific documents grows exponentially, having an automatic and intelligent keyword extraction technique is getting more critical. There are various keyword extraction techniques that are either based on statistical features of the text or machine learning approaches, and sometimes a combination of both. In this research, we propose a new keyword extraction method for Persian scientific documents based on recommender systems and case-based reasoning. The proposed method is designed based on case-based reasoning in which the main assumption is that similar documents share similar keywords. There are two main steps in the proposed approach: first, similar documents to a given new document are retrieved based on TFIDF and word2vec model, and second, the candidate keywords are extracted from retrieved documents and ranked based on a new scoring scheme, and a set of keywords are selected from the candidate keywords based on their score. The proposed method is tested and evaluated on a set

* Corresponding Author

of documents of Ganj database in three different subject areas (Art, Humanities and Engineering), based on precision, recall and expert panel.

Keywords: Keyword Extraction, Recommender Systems, Case-Based Reasoning, Word2Vec Word Embedding, Information Retrieval, Machin Learning, Indexing

توسعه سیستم پیشنهاددهنده بر مبنای استدلال نمونه‌محور برای نمایه‌سازی مستندات علمی فارسی

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

آزاده فخرزاده
دکتری پردازش تصویر؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
Fakhrzadeh@mail.irandoc.ac.ir

مرضیه زرین‌بال
دکتری مهندسی صنایع؛ استادیار؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛ تهران، ایران؛
zarinbal@irandoc.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۱/۱۱/۰۹ | **پذیرش:** ۱۴۰۲/۰۱/۲۱ | مقاله برای اصلاح به مدت ۹ روز نزد پدیدآوران بوده است.

چکیده: استخراج کلیدواژه یکی از مهم‌ترین قدم‌های نمایه‌سازی مستندات محسوب می‌شود. کلیدواژه‌های هر سند، توصیفگرهای مفهومی هستند که می‌توانند در جست‌وجو و بازیابی اطلاعات و نیز اشاعه آن‌ها به‌کار گرفته شوند. در پایگاه‌های دربردارنده اسناد علمی مانند پایگاه علمی «گنج» متعلق به «پژوهشگاه علوم و فناوری اطلاعات ایران»، کلیدواژه‌ها نقش مهم‌تری دارند، و بنابراین، تخصیص کلیدواژه‌های تخصصی نیز چالش‌برانگیزتر خواهد بود، زیرا در این پایگاه اسناد تخصصی با حوزه‌های علمی مختلفی وجود دارند. با توجه به افزایش حجم تولید و ثبت مستندات علمی، نیاز است که فرایند نمایه‌سازی و تخصیص کلیدواژه با سرعت بیشتری صورت گیرد و از روش‌های ماشینی هوشمند برای پیشنهاد و تخصیص کلیدواژه استفاده گردد. در بسیاری از پایگاه‌های اطلاعات علمی دنیا از روش‌های ماشینی و خودکار در کلیه فعالیت‌های فرایند نمایه‌سازی یا بخشی از آن‌ها استفاده می‌شود. بعضی از این روش‌ها بر مبنای تحلیل آماری متون و استفاده از روش‌های یادگیری ماشین هستند، تعدادی بر مبنای تحلیل معنایی متون به‌واسطه اصطلاحنامه‌های تخصصی و هستان‌شناسی، و در تعدادی دیگر از این روش‌ها از تلفیق هر دو استفاده

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 ۳۹ | شمارهٔ ۲ | ص‌ص ۵۹۹-۶۲۶
زمستان ۱۴۰۲
https://doi.org/10.22034/jipm.2023.704737

می‌شود. بر همین اساس، در این طرح پژوهشی روشی برای پیشنهاد کلیدواژه به مستندات علمی فارسی ارائه شده که بر مبنای روش‌های هوشمند پردازش متن و یادگیری ماشین عمل می‌کند. روش پیشنهادی، بر مبنای سیستم‌های پیشنهاددهنده و استدلال نمونه‌محور طراحی شده که بر اساس آن، مجموعه‌ای از کلیدواژه‌های مرتبط با یک سند به نمایه‌ساز پیشنهاد شود تا نمایه‌ساز سریع‌تر بتواند کلیدواژه‌های مناسب از بین آن‌ها را انتخاب کند. روش پیشنهادی بر اساس استدلال نمونه‌محور عمل می‌کند که در آن فرض بر این است که اسناد مشابه می‌توانند کلیدواژه‌های مشابه داشته باشند. بر همین اساس، ابتدا اسناد مشابه با یک سند جدید بر اساس روش‌های TFIDF و روش‌های بازنمایی کلمه-به-بردار بازیابی می‌شوند. سپس، کلیدواژه‌های کاندید از بین اسناد مشابه در نظر گرفته می‌شوند و سرانجام، بر اساس یک تابع رتبه‌بندی، کلیدواژه‌های مناسب از بین آن‌ها انتخاب می‌شوند. روش پیشنهادی بر روی مجموعه‌ای از اسناد «پایگاه گنج» در سه حوزه فنی و مهندسی، هنر و ادبیات، و علوم انسانی پیاده‌سازی، و نتایج آن با معیارهایی نظیر دقت، فراخوانی و نظرات متخصصان ارزیابی شده است..

کلیدواژه‌ها: کیفیت داده، استخراج کلیدواژه، سیستم‌های پیشنهاددهنده، استدلال نمونه‌محور، روش بازنمایی کلمه-به-بردار، بازیابی اطلاعات، یادگیری ماشین

۱. مقدمه

عناوین نمایه‌های یک سند، مجموعه‌ای از واژگان یا مفاهیمی است که بیانگر محتوای آن هستند (Frické 2012a)، و هدف از نمایه‌سازی، بهبود جست‌وجو، بازیابی و سازماندهی اسناد و اطلاعات است. بنابراین، نمایه‌ها باید به‌گونه‌ای انتخاب شوند که باعث افزایش دقت و فراخوانی در فرایند بازیابی اطلاعات شوند. بر این اساس، نمایه‌سازی مستندات علمی با توجه به افزایش حجم و تنوع آن‌ها، یکی از چالش‌های مهم در بازیابی اطلاعات در پایگاه‌های تخصصی و وب به‌شمار می‌رود. پیش‌تر، از روش‌های دستی (نمایه‌ساز انسانی) برای نمایه‌سازی مستندات استفاده می‌شد، لیکن با پیشرفت فناوری‌های بازیابی اطلاعات و روش‌های هوشمند تحلیل خودکار متن، روش‌های ماشینی جدیدی برای این منظور توسعه یافتند. در طراحی و توسعه و ارزیابی عملکرد هر نظام یا روشی، دو موضوع از اهمیت ویژه‌ای برخوردار است: قابلیت اطمینان[1] و اعتبار[2] (Frické 2012b). قابلیت اطمینان به این مفهوم است که در یک سیستم یا روش (اعم از اینکه انسانی یا ماشینی باشد) تا چه حد ورودی یکسان منجر به خروجی یکسان می‌شود. اعتبار به این مفهوم است که تا چه میزان خروجی‌های یک نظام یا سیستم، معتبر و صحیح است. نمایه‌ساز انسانی لزوماً قابل

1. reliability 2. validity

اطمینان نیست و مانند سایر فرایندهای انسانی، عوامل انسانی متعددی روی نتیجه آن اثر می‌گذارد و نتیجه‌ای که قابل اطمینان نباشد به‌طور مسلم از اعتبار کمی برخوردار است. نمایه‌سازی ماشینی قابلیت اطمینان بالایی دارد، اما ممکن است معتبر نباشد. به‌رغم اینکه ممکن است دقت روش‌های ماشینی در برخی از موارد کمتر از روش‌های دستی باشد، با توجه به حجم بالای اطلاعات و سرعت زیاد نمایه‌سازی ماشینی، و قابل اطمینان بودن آن، این روش‌ها برای نمایه‌سازی وب و مستندات تخصصی استفاده می‌شوند (Bayatmakou, Ahmadi & Mohebi 2017; Firoozeh et al. 2020; Rose et al. 2010; Thushara, Mownika & Mangamuru 2019).

مهم‌ترین قدم در نمایه‌سازی یک سند استخراج و تخصیص کلیدواژه است. بیشتر روش‌های توسعه‌یافته برای استخراج خودکار کلیدواژه، تنها با اکتفا به متن سند، کلیدواژه‌هایی را استخراج می‌کنند (Alami Merrouni, Frikh & Ouhbi 2020; Hasan & Ng 2014)؛ اما روش‌هایی هم هستند که از سایر منابع مانند اصطلاحنامه‌ها یا پیکره‌های تخصصی برای این منظور استفاده می‌کنند (Hasan & Ng 2014; Pay & Lucci 2017; Sharma, Jain & Aggarwal 2018). در بسیاری از روش‌های استخراج کلیدواژه، از ویژگی‌های مختلف زبانی و آماری نظیر فراوانی، هم‌رخدادی کلمات، محل رخداد، نقش کلمه در جمله، و نحوه نگارش کلمه در متن سند (برای مثال، به‌صورت بولد یا ایتالیک یا با حروف بزرگ) استفاده می‌شود (Firoozeh et al. 2020). لیکن استخراج قوانینی که بتوانند کلیه این ویژگی‌ها را پوشش دهند و هنگام استخراج و تخصیص کلیدواژه به‌نوعی بیانگر رویّه استدلال و منطق ذهن نمایه‌ساز باشند، چالش‌برانگیز است و تنظیم این قوانین به‌راحتی امکان‌پذیر نیست (Frické 2012a). در طراحی روش‌های هوشمند برای این دست از مسائل که به‌راحتی در قالب قوانین مشخص و معینی نمی‌گنجند، یک رویکرد، بهره‌گیری از استدلال نمونه‌محور[1] است. مبنای روش‌های استدلال نمونه‌محور این است که مسائل مشابه، راه‌حل‌های مشابه دارند. بنابراین، برای مسئله استخراج کلیدواژه برای یک سند که راه‌حل آن، کلیدواژه‌های استخراج‌شده برای سند است، فرض بر این است که می‌توان از کلیدواژه‌های نمونه‌های مشابهی (اسناد مشابهی) که پیش‌تر نمایه شده‌اند، استفاده کرد (Kolodner 1992; Weber, Ashley & Brüninghaus 2005).

1. case-based reasoning

در این پژوهش، از استدلال نمونه‌محور برای نمایه‌سازی مستندات علمی (استخراج کلیدواژه از اسناد علمی) فارسی استفاده شده است. در واقع، برای نمایه‌سازی یک سند، از اسنادی که پیش‌تر نمایه شده‌اند و کلیدواژه‌های مناسب دارند به‌عنوان یک منبع اطلاعاتی، برای استخراج کلیدواژه استفاده می‌شود. بر همین اساس، یک سیستم پیشنهاد کلیدواژهٔ طراحی‌شده که با استفاده از استدلال نمونه‌محور، برای یک سند، کلیدواژه مناسب پیشنهاد می‌دهد. این سیستم می‌تواند به‌صورت کاملاً خودکار برای نمایه‌سازی مستندات علمی فارسی به‌کار گرفته شود، یا در کنار سایر روش‌های استخراج کلیدواژه به نمایه‌ساز کمک کند.

برای طراحی سیستم پیشنهادی که بر مبنای استدلال نمونه‌محور عمل کند، دو سؤال اصلی وجود دارد که در این پژوهش بررسی می‌شوند: برای تعیین کلیدواژه‌های یک سند (۱) چگونه اسناد مشابه یک سند شناسایی می‌شوند؟، و(۲) چگونه از کلیدواژه‌های اسناد مشابه و اطلاعات سند مورد نظر، برای تعیین کلیدواژه‌های سند استفاده می‌شود؟ به‌منظور پاسخگویی به این دو سؤال، سیستم پیشنهاددهنده دو بخش اصلی دارد: در بخش اول، کلیدواژه‌های کاندید برای سند جدید بر اساس اسناد مشابه انتخاب می‌شوند و در بخش دوم، کلیدواژه‌های کاندید بر اساس معیارهای مختلفی از جمله میزان ارتباط آن‌ها با سند رتبه‌بندی خواهند شد.

در این پژوهش برای آزمایش و ارزیابی روش پیشنهادی، از بخشی از داده‌های پایگاه اطلاعاتی «گنج»[1] متعلق به پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) که حاوی پایان‌نامه‌ها و رساله‌های فارسی کشور است، استفاده شده است. برای این منظور ۳۱۸۴۰ پایان‌نامه و رساله در موضوعات مختلف به‌صورت تصادفی استخراج شده و تنها عنوان، موضوع، کلیدواژه‌ها و خلاصهٔ هر سند در نظر گرفته شده است.

ساختار این پژوهش به شرح زیر است: به‌دلیل اهمیت سیستم‌های پیشنهاددهنده مبتنی بر استدلال نمونه‌محور در این پژوهش، در بخش ۲، شرح مختصری در این باره ارائه شده است. در بخش ۳، پیشینه پژوهش آمده و در بخش ۴، سیستم پیشنهادی و اجزای مختلف آن تشریح شده است. ارزیابی و پیاده‌سازی در بخش ۵، بررسی شده و در بخش۶، مهم‌ترین یافته‌ها از پیاده‌سازی سیستم پیشنهادی بیان شده است. سرانجام، در بخش ۷، نتیجه‌گیری و پیشنهاد برای پژوهش‌های آتی آمده است.

1. www.ganj.irandoc.ac.ir

۲. سیستم‌های پیشنهاددهنده و استدلال نمونه‌محور

سیستم‌های پیشنهاددهنده را به‌طور معمول بر اساس روشی که برای پیشنهاد اقلام در آن‌ها استفاده می‌شود، دسته‌بندی می‌کنند. روش‌های پیشنهاد اقلام نیز بر مبنای منبع دانش و اطلاعاتی که در سیستم‌ها برای پیشنهاد اقلام جدید وجود دارد، به‌کار گرفته می‌شوند. بر اساس یک دسته‌بندی کلاسیک، سیستم‌های پیشنهاددهنده به پنج دسته اصلی تقسیم‌بندی می‌شوند: روش‌های مبتنی بر پالایش همکارانه، روش‌های محتوامحور، روش‌های دموگرافیک، روش‌های دانش‌محور، و روش‌های ترکیبی (Burke 2007; Recommender Systems Handbook 2011). سیستم‌های پیشنهاددهنده محتوامحور، یکی از روش‌های مهم و پرکاربرد در این حوزه است (Adomavicius & Tuzhilin 2005; Baltrunas 2008). در این سیستم‌ها از ویژگی‌های توصیفی اقلام برای تهیه پیشنهادها استفاده می‌شود و واژه محتوا در واقع، به همان توصیف‌ها اشاره دارد. سیستم‌های محتوامحور بسته به نوع مسئله و دانش موجود درآن می‌توانند بر اساس استدلال نمونه‌محور طراحی شوند (Smyth 2007). در سیستم‌های پیشنهاددهندهٔ نمونه‌محور توصیه‌هایی بر اساس شباهت بین درخواست‌های هدف[1] و نمونه‌های موجود ارائه می‌شود. به گفته دیگر، ترجیح سیستم، یافتن اقلامی درون پایگاه نمونه‌هاست که بیشترین شباهت به درخواست ارائه‌شده به سیستم را داشته باشد (Bridge et al. 2005). فرایند استدلال نمونه‌محور در این سیستم‌ها شامل چهار گام اصلی است (Humphreys et al. 2003):

- ◇ بازیابی[2]: سیستم برای یک مسئله جدید، برای یافتن نمونه‌های مشابه با مسئله جدید، در میان پایگاه نمونه‌ها جست‌وجو می‌کند و به هر نمونه بر اساس میزان شباهت آن امتیاز می‌دهد؛
- ◇ استفاده مجدد[3]: راه‌حل موجود برای نمونه‌های مشابه با مسئله جدید تطبیق می‌یابد و بهترین راه‌حل پیشنهاد می‌شود؛
- ◇ بازنگری[4]: راه‌حل روی مسئله جدید آزمایش می‌شود و بازنگری‌های لازم انجام می‌گیرد؛
- ◇ نگهداری[5]: پس از اینکه راه‌حل با موفقیت با مسئله جدید تطبیق داده شد، نتیجه به‌عنوان یک نمونه جدید در پایگاه نمونه‌ها ذخیره می‌شود.

1. target query 2. retrieve 3. reuse 4. revise 5. retain

در استدلال نمونه‌محور، یادگیری با تعمیم یا استقرا انجام نمی‌شود، بلکه با تکیه بر دانش نهفته در نمونه‌های قبلی و تطبیق راه‌حل آن‌ها برای حل مسئله جدید و نیز تکامل سیستم، با افزایش تعداد نمونه‌ها صورت می‌گیرد (Kolodner 1992).

۳. پیشینهٔ پژوهش

در این بخش مهم‌ترین پژوهش‌ها در زمینه به‌کارگیری سیستم‌های پیشنهاددهنده در حوزه اسناد علمی، و نیز روش‌های استخراج کلیدواژه برای مستندات لاتین که در آن‌ها از اصطلاحنامه‌ها یا پیکره‌های تخصصی استفاده شده، بررسی می‌شوند. همچنین، برخی از روش‌های استخراج کلیدواژه برای مستندات فارسی معرفی می‌شوند.

سیستم‌های پیشنهاددهنده در حوزه اسناد علمی، اغلب نمونه‌محور هستند و کاربردهای متفاوتی دارند: سیستم‌های پیشنهاددهنده برای اختصاص داور به مقالات مجلات و کنفرانس‌ها (Wang, Shi & Chen 2010)، خلاصه‌سازی خودکار مقالات علمی (Roul & Arora 2019)، پیشنهاد رویدادهای علمی (Klamma, Cuong & Cao 2009)، پیشنهاد ارجاعات برای اختراعات (Oh et al. 2013)، کمک به برنامه‌های آموزشی دانش‌آموزان (Ahn et al. 2005)، تشخیص سرقت علمی (Bohra & Barwar 2022)، و اختصاص کلیدواژه به یک سند (Huang, Névéol & Lu 2011).

برای استخراج کلیدواژه از مستندات لاتین با استفاده از اصطلاحنامه‌ها یا پیکره‌های تخصصی، در مقالات زیادی از «مش»[1]، اصطلاحنامه کنترل‌شده حوزه پزشکی استفاده شده است. «هوانگ، نوول و لو» برای تخصیص کلیدواژه‌ها به هر سند جدید، ابتدا k همسایه نزدیک[2] به آن سند را بر اساس شباهت محتوایی بازیابی کرده و سپس تمام کلیدواژه‌هایی را که از «مش» به این مقالات مرتبط شده‌اند، با استفاده از یک الگوریتم رتبه‌بندی شباهت به مقاله جدید اختصاص دادند (Huang, Névéol & Lu 2011). «مائو و لو» روشی به اسم «مش ناو»[3] را معرفی کردند که در آن با استفاده از سه روش مختلف یک لیست از کلمات مرتبط با سند مورد نظر را از «مش» استخراج و سپس با یک روش یادگیری رتبه‌بندی، کلمات را بر اساس ارتباط آن‌ها با سند مورد نظر رتبه‌بندی کردند (Mao & Lu 2017).

1. MeSH
2. k-nearest Neighborhoods
3. MeSH Now

از آنجا که زبان فارسی دارای چالش‌های بی‌شمار در مقایسه با زبان انگلیسی است، بسیاری از الگوریتم‌های استخراج کلیدواژه که روی داده‌های انگلیسی نتایج به‌نسبت خوبی ارائه می‌کنند، برای زبان فارسی کارآمد نیستند. برای متون فارسی، «کیان و زاهدی» رویکردی را پیشنهاد کردند که با استفاده از استخراج کلیدواژه‌ها، دسترسی به محتوای صفحات وب بهبود می‌یابد. برای این منظور از دو تابع امتیازدهی استفاده می‌شود که با اولین تابع سعی در یافتن اهمیت کلمات بر اساس ویژگی‌های آماری دارد و دومین تابع نتایج حاصل از امتیازدهی تابع اول را با استفاده از نتایج موتورهای جست‌وجوی پُراستفاده مانند «گوگل»، «یاهو» و «ام‌اس‌ان»[1] ارزیابی می‌کند؛ به این ترتیب که نتیجه صفحهٔ اول را به‌عنوان بازخورد دریافت، و سپس اولین تابع را با استفاده از الگوریتم ژنتیک بهینه می‌کند (Kian & Zahedi 2011).

افزون بر این، «خوزانی و بیات» رویکرد جدیدی با استفاده از رویکردهای آماری برای استخراج کلیدواژه‌ها ارائه کرده‌اند که نه‌تنها قادر به استخراج کلیدواژه‌های تک-کلمه‌ای است، بلکه قابلیت استخراج کلیدواژه‌های دو-کلمه‌ای را نیز در متن فارسی دارد. در این رویکرد از دو معیار فراوانی کلمه و تعداد اسناد دربردارنده کلمه[2] برای استخراج کلیدواژه‌های تک-کلمه‌ای استفاده شده، سپس با استفاده از ماتریس ساخته‌شده برای کلیدواژه‌های تک-کلمه‌ای، کلیدواژه‌های دو-کلمه‌ای نیز استخراج می‌شود (Khozani & Bayat 2011).

از جمله رویکردهای مهم برای استخراج کلیدواژه از اسناد فارسی می‌توان به روش «صیادی، قدسی و نقیبی» اشاره کرد که با استفاده از الگوریتم PostRank بُعد اسناد متنی موجود در یک وبلاگ را کاهش می‌دهند تا موضوع اصلی وبلاگ را مشخص کنند. به گفته دیگر، در این روش، کلماتی که به نحو بهتری نمایش‌دهنده مفهوم یک وبلاگ هستند، مشخص می‌شود (Sayyadiharikandeh, Ghodsi & Naghibi 2012). PostRank از گراف نحوی با در نظر گرفتن برخی از ویژگی‌های ساختاری وبلاگ برای استخراج کلیدواژه‌ها استفاده می‌کند.

«محرابی، محبی و احمدی» روشی را بر مبنای الگوریتم RAKE (Rose et al. 2010) پیشنهاد دادند که به‌طور خاص برای مستندات علمی فارسی طراحی شده است. در پژوهش ایشان، برخی از ویژگی‌های امتیازدهی به عبارات کاندید در الگوریتم RAKE،

1. Google, Yahoo, MSN
2. document frequency

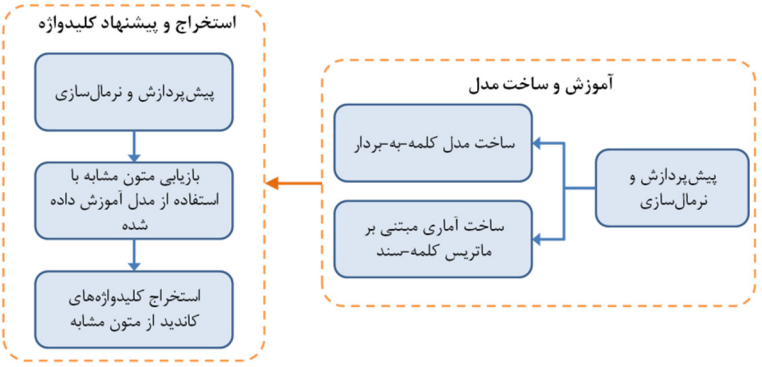
متناسب با ویژگی‌های زبان فارسی تغییر یافته تا سرانجام، نتایج بهتری حاصل شود (Mehrabi, Mohebi & Ahmadi 2021).

در پژوهش‌های انجام‌شده در زبان انگلیسی، تعداد محدودی از اطلاعات اسناد مشابه برای استخراج کلیدواژه استفاده کرده‌اند که در حوزه نمایه‌سازی مستندات علمی پزشکی هستند و بر مبنای اصطلاحنامه‌های تخصصی موجود در آن حوزه عمل می‌کنند. در زبان فارسی نیز پژوهش‌های انجام‌شده روی استفاده از ویژگی‌های آماری و ساختاری در محتوای سند، متمرکز بوده و پژوهشی بر مبنای رویکرد استدلال نمونه‌محور و سیستم‌های پیشنهاددهنده که بر مبنای اطلاعات اسناد مشابه کار کند، انجام نشده است.

٤. روش پیشنهادی

روش پیشنهادی برای استخراج کلیدواژه در واقع، یک سیستم پیشنهاددهنده است که بر مبنای استدلال نمونه‌محور عمل می‌کند. ابتدا، مجموعه‌ای از کلیدواژه‌های کاندید استخراج می‌شود، سپس، بر اساس یک تابع امتیازدهی، این کلیدواژه‌ها رتبه‌بندی شده و کلیدواژه‌ها با بهترین امتیاز پیشنهاد می‌شوند. مجموعه کلیدواژه‌های کاندید برای یک سند از متن سند استخراج نمی‌شود، بلکه بر اساس میزان شباهت سند با اسناد پیشین و با کمک دانش موجود در این اسناد دارای کلیدواژه است.

در شکل ۱، نحوه استخراج کلیدواژه‌های کاندید برای یک سند، در سیستم پیشنهادی نمایش داده شده که دارای دو مرحله اصلی است: آموزش و ساخت مدل، و استخراج و پیشنهاد کلیدواژه.



شکل ۱. روش پیشنهادی برای استخراج کلیدواژه‌های کاندید

در مرحله نخست، با توجه به اینکه اسناد مورد استفاده برای پیشنهاد کلیدواژه، اسناد تخصصی هستند، نیاز است ابتدا مجموعه اسناد از نظر موضوعی به دسته‌های مختلف تقسیم شده و فرایند آموزش برای هر دسته از اسناد به‌صورت جداگانه انجام شود. در نهایت، برای هر دسته، مدل کلمه-به-بردار[1] و ماتریس کلمه-سند مربوط به موضوع آن دسته حاصل می‌شود. در مسئله مربوط به اسناد تخصصی علمی مانند اسناد پایان‌نامه‌ها و رساله‌ها، معمولاً هر سند یک برچسب موضوعی کلان دارد که بر همان اساس می‌توان اسناد را پیش از انجام فرایند آموزش به دسته‌های مختلف دسته‌بندی نمود. در این پژوهش بر اساس تنوع داده‌های موجود، چهار دسته فنی-مهندسی، علوم انسانی، هنر و ادبیات، و علوم پایه در نظر گرفته شده است.

در مرحله دوم، پس از تشکیل دسته‌های موضوعی از اسناد، نیاز است برای هر دسته به‌صورت تصادفی، سه مجموعه داده مشخص شود: مجموعه داده آموزش، مجموعه داده اعتبارسنجی، و مجموعه داده آزمایش. از مجموعه داده آموزش برای آموزش مدل استفاده می‌شود و مجموعه داده اعتبارسنجی برای تنظیم پارامترهای مدل و اعتبارسنجی آن به‌کار گرفته خواهد شد و مجموعه داده آزمایش نیز برای آزمایش نهایی روش پیشنهادی و اعلام دقت روش استفاده می‌شود.

٤-١. مرحله نخست: آموزش و ساخت مدل

٤-١-١. پیش پردازش و نرمال‌سازی

پیش از آموزش مدل، نیاز است که عملیات پیش‌پردازش روی داده‌ها انجام گیرد. در مرحلهٔ اول، نرمال‌سازی یا یکسان‌سازی انجام می‌شود. هدف از نرمال‌سازی یکسان‌کردن کلمات موجود در اسناد از نظر کدهای کاراکترهای فارسی است. پس از آن ایست‌واژه‌ها[2] حذف شده و سپس با استفاده از یک تابع برای یافتن ریشه کلمات[3] عملیاتی نظیر اصلاح نویسه‌ها، اصلاح نشانه‌گذاری، رعایت فاصله و نیم‌فاصله، حذف اعراب و تشدید، حذف علایم نگارشی (غیر از نقطه)، استخراج جملات، و ریشه‌یابی واژه‌ها انجام می‌شود. برای این منظور از ابزار هضم[4] استفاده شده که نظیر واژه-واژه کردن[5]، ریشه‌یابی، و برچسب‌زنی اجزای سخن[6] را برای زبان فارسی ارائه می‌دهد.

1. word2vec
2. stop words
3. iemmatizer
4. Hazm
5. tokenize
6. part of speech tagging (POS)

٤-١-٢. آموزش و ساخت مدل بازیابی اسناد مشابه

برای بازیابی اسناد مشابه از دو مدل استفاده شده است: مدل آماری مبتنی بر فراوانی و مدل ارتباط معنایی.

روش بازیابی مبتنی بر فراوانی[1]

روش TFIDF، فراوانیِ کلمه در یک سند را در مقابل فراوانی آن در کل پیکره می‌سنجد:

$$TF\ (t, d_i) = \text{frequency of term } t \text{ in document } d_i \qquad IDF\ (t) = \frac{N}{|\{d \in \boldsymbol{D}\ | t \in d\}|} \tag{۱}$$

$$TFIDF\ (t, \boldsymbol{d_i}) = TF\ (t, \boldsymbol{d_i}).IDF\ (t) \tag{۲}$$

در رابطه (۱)، فراوانی کلمه t در سند $\boldsymbol{d_i}$ ($TF\ (t, \boldsymbol{d_i})$) و IDF که فراوانی معکوس سند است، محاسبه می‌شود، که N تعداد تمام اسناد موجود در پیکره متنی، و D مجموعه اسنادی است که کلمه t در آن‌ها آمده است. در رابطه (۲)، TFIDF برای هر کلمه محاسبه می‌شود که در نهایت، هر درایه از ماتریس کلمه-سند، بیانگر وزن کلمه بر اساس رابطه (۲) در یک سند است (Manning & Raghavan 2009).

روش بازیابی معنایی مبتنی بر مدل کلمه-به-بردار

روش کلمه-به-بردار (Mikolov et al. 2013)، بر اساس مدل آموزش دیده‌شده روی یک شبکه عصبی با یک لایه پنهان[2] طراحی شده که کلمات را به‌عنوان بردارهایی در یک فضای برداری چندبعدی نمایش می‌دهد. با استفاده از این مدل، می‌توان روابط معنایی بین کلمات را با توجه به نحوه قرارگیری آن‌ها در پیکرهٔ متنی استخراج کرد. دقت مدل به تعداد دفعات هم‌رخدادی[3] کلمات درون یک سند در پیکره متنی وابسته است. بنابراین، هرچه حجم داده‌های پیکره بیشتر باشد، دقت این روش افزایش می‌یابد.

برای آموزش شبکه عصبی از دو مدل استفاده می‌شود: مدل کیسه لغات پیوسته[4] و مدل اسکیپ-گرام[5]. در CBOW، قبل از پیش‌بینی کلمه هدف، از بردارهای کلمات زمینه محلی[6] میانگین گرفته می‌شود. اما در Skip-gram عمل میانگین‌گیری برای بردارهای کلمات انجام نمی‌شود. این روش در مواقعی که کلمات خاص و کم‌تکرار در پیکره زیاد باشد، بهتر عمل می‌کند، زیرا از بردار کلمه-به-بردار سایر کلمات زمینه محلی برای پیش‌بینی، میانگین گرفته نمی‌شود.

---

1. TFIDF-Term Frequency Inverse Document Frequency
2. hidden layer
3. co-occurence
4. continuous bag of words (CBOW)
5. Skip-gram
6. local context

پس از آموزش مدل کلمه-به-بردار برای هر دسته موضوعی از اسناد، برای هر کلمه در مجموعه داده، یک بازنمایی برداری عددی محاسبه می‌شود. با کمک این بردار می‌توان برای یک کلمه خاص کلمات مشابه را بازیابی نمود. شباهت نیز بر اساس فاصله کسینوسی بین دو بردار محاسبه می‌شود. نمونه‌ای از نحوه عملکرد مدل کلمه-به-بردار برای کلمات مختلف در حوزه‌های موضوعی مختلف، در جدول ۱۱، آمده است.

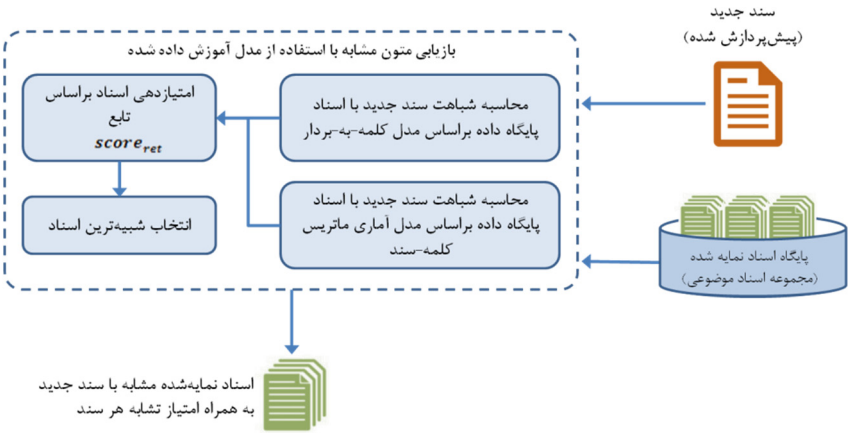
جدول ۱۱. نمونه‌ای از عملکرد مدل آموزش داده‌شده کلمه-به-بردار برای بازیابی کلمات مشابه

موضوع	کلمه	کلمات مشابه بازیابی‌شده (به‌ترتیب)
هنر و ادبیات	مولوی	مثنوی، عطار، خاقانی، سنایی، سعدی، حافظ، غزلیات، سروده، عاشقانه، مولانا، منظومه، فروغ، تمثیل، غزل، اسلوب، بدیع، احادیث، بیدل، قصاید، ناصرخسرو، قصیده، آیات
	ادبیات	عرب، معاصر، عامیانه، نویسندگان، ادب، کهن، عامه، هرسین، نوجوان، کلاسیک، برجسته، جامعه‌شناسی، سینما، غنا، بازشناخت، شاخه، نثر، حوزه، غنی، مکاتب، نقد، نو، مشروطیت
فنی و مهندسی	عصبی	بیزی، ANFIST، نروفازی، بیز، سیناپس، پرسپترون، مصنوعی، LVQ، ANN، فازی-عصبی، شبکه عصبی، سخت‌افزاری، بولتزمن
	سازه	لرزه، ساختمان، میراگرها، زلزله، قاب، دیوار، بهسازی، نامنظم، بتنی، پی، شمع، ساختگاه، بنا، روسازه، سکو، مهاربند، طبقات، جداگر، بادبند
علوم انسانی	خلاقیت	انگیزش، یادگیری، خلاق، مؤلفه، خودتنظیمی، خودکارآمدی، تفکر، فرصت‌یابی، شایستگی، مهارت، خودانگیزی، آمادگی، یادگیرنده، اثربخش
	سرمایه‌گذاری	نقدینگی، دارایی، بدهی، مازاد، اهرم، سودآوری، صندوق، جاری، ریسک، بازده، پرتفوی، پورتفوی، نقدشوندگی، خالص، بنگاه، آتی، مبادله، معاملات، سهام، چرخه، نگه‌داشت

٤-٢. مرحله دوم، استخراج و پیشنهاد کلیدواژه‌ها

٤-٢-١. بازیابی اسناد مشابه

برای بازیابی اسناد مشابه از دو مدل مختلف آموزش‌دیده استفاده شده است (شکل ۲).



شکل ۲. بازیابی اسناد مشابه

در مدل معنایی، اگر سند d_i پس از پیش‌پردازش، به کلمات $W_i = \{w_{i1}, w_{i2}, \dots, w_{in}\}$ تبدیل شده باشد (n بیانگر تعداد کلمات سند d_i)، آنگاه برای هر کلمه یک بازنمایی برداری بر اساس مدل کلمه-به-بردار وجود دارد. بر این اساس، بردار سند d_i به‌صورت زیر محاسبه می‌شود:

$$V_{avg}(d_i) = \frac{1}{n}\sum_{k=1}^{n} V_{W2V}(w_{ik}) \qquad (۳)$$

که $V_{W2V}(w_{ik})$ همان بردار متناظر با کلمه w_{ik} در مدل کلمه-به-بردار است. در واقع، در رابطه (۳)، بردار هر سند بر اساس میانگین بردار کلمات تشکیل‌دهنده آن سند محاسبه می‌شود.

سرانجام، شباهت بین دو سند با استفاده از شباهت کسینوسی بین دو بردار آن‌ها محاسبه می‌شود:

$$Sim_{W2V}(d_i, d_j) = cos\ \left(V_{avg}(d_i), V_{avg}(d_j)\right) \qquad (۴)$$

مدل مبتنی بر TFIDF، بر اساس ماتریس کلمه-سند عمل می‌کند که هر درایه بیانگر مقدار TFIDF هر کلمه در هر سند است و بر اساس رابطه (۲) محاسبه می‌گردد. در نهایت، شباهت بین دو سند، بر اساس شباهت کسینوسی بین بردارهای آن‌ها در ماتریس کلمه-سند، به‌دست می‌آید.

با تجمیع شباهت بین دو سند در هر دو مدل، سرانجام:

$$Sim_{Final}(d_i, d_j) = (1-\propto)\, Sim_{W2V}(d_i, d_j) + \propto\ Sim_{TFIDF}(d_i, d_j) \qquad (۵)$$

که $0 \leq \propto \leq 1$ مشخص می‌کند که تا چه‌میزان در محاسبه شباهت نهایی به شباهت

معنایی اهمیت می‌دهیم. هرچه $\propto$ بزرگ‌تر باشد، وزن روش معنایی که بر اساس مدل کلمه-به-بردار عمل می‌کند، بیشتر خواهد بود. پارامتر $\propto$ در فرایند آموزش و اعتبارسنج تعیین می‌شود.

در نهایت، برای هر سند جدید d_{new} با توجه به میزان شباهت آن با سایر اسناد، مجموعه $\mathcal{M}_{\text{new}}$ که متشکل از M سندی است که بیشترین تشابه را دارند، به همراه میزان تشابه هر یک حاصل می‌شود:

$$\boldsymbol{\mathcal{M}_{\text{new}} = \{\langle \mathbf{d}_1^*, \mathbf{score}(\mathbf{d}_1^*)\rangle, \langle \mathbf{d}_2^*, \mathbf{score}(\mathbf{d}_2^*)\rangle, \dots, \langle \mathbf{d}_M^*, \mathbf{score}(\mathbf{d}_M^*)\rangle\}} \quad (۶)$$

در مجموعه فوق، اسناد بر اساس میزان شباهتشان از ۱ تا M مرتب شده‌اند و $\text{score}_{\text{ret}}(\text{d}_\text{m}^*)$ بیانگر میزان شباهت بین سند جدید با سند d_m^* در فرایند بازیابی است:

$$\boldsymbol{score_{ret}(d_m^*, d_{new}) = Sim_{Final}(d_m^*, d_{new})} \quad (۷)$$

۴-۲-۲. استخراج کلیدواژه‌های کاندید و پیشنهاد کلیدواژه مناسب

در این مرحله کلیدواژه‌های اسناد مشابه، بر اساس دو روش امتیازدهی رتبه‌بندی می‌شوند. در روش امتیازدهی بر اساس امتیاز بازیابی، برای هر کلیدواژه یک بردار، $V_{key}(k_l)$، Mتایی بر مبنای مدل تک-مؤلفه‌ای باینری در نظر گرفته می‌شود که هر مؤلفه از آن می‌تواند مقدار ۰ یا ۱ را اتخاذ کند. هر مؤلفه از این بردار متناظر با یک سند از مجموعه اسناد مشابه $(\mathcal{M}_{\text{new}})$ است و مقدار صفر یا یک برای آن بیانگر تعلق کلیدواژه به سند است:

$$\boldsymbol{V_{key}(k_l) = (v_{1l}, \dots, v_{Ml})}$$
$$\boldsymbol{v_{mk} = \begin{cases} 1 & if\ k_l\ belongs\ to\ d_m^* \\ 0 & otherwise \end{cases}} \quad (۸)$$

از طرفی، هر سند نیز دارای یک امتیاز است که بر اساس رابطه (۷) محاسبه شده است. بنابراین، می‌توان یک بردار $\boldsymbol{M}$تایی از امتیازها برای سند جدید در نظر گرفت:

$$\boldsymbol{V_{ret} = (score(d_1^*), \dots, score(d_M^*))} \quad (۹)$$

حال، می‌توان یک تابع امتیاز بر اساس ضرب نقطه‌ای بین دو بردار $\boldsymbol{V_{ret}}$ و $\boldsymbol{V_{key}(k_l)}$ محاسبه کرد که بیانگر امتیاز هر کلمه کلیدی باشد. بر این اساس، کلماتی که در اسناد شبیه‌تر هستند، امتیاز بیشتری خواهند داشت. همچنین، کلیدواژه‌ای که در تعداد اسناد بیشتری از بین اسناد بازیابی‌شده هستند، نیاز است که وزن بیشتری را نسبت به سایر کلیدواژه‌ها داشته باشند. بنابراین، در محاسبه امتیاز نهایی می‌توان ضریبی را برای این

منظور لحاظ نمود. بنابراین، تابع امتیاز کلیدواژه k_l بر اساس امتیاز بازیابی به‌صورت زیر محاسبه می‌شود:

$$score_1(k_l, d_{new}) = \frac{V_{ret} \cdot V_{key}(k_l)}{|V_{ret}|} \log(f(k_l) + \frac{1}{4}M) \qquad (۱۰)$$

که $f(k_l)$ بیانگر تعداد اسناد از بین M سند بازیابی‌شده‌ای است، که k_l کلیدواژه آن‌ها بوده است. در واقع، با ضرب عبارت فوق در عبارت لگاریتمی، وزن بیشتری برای کلیدواژه‌هایی که در بیش از یک چهارم اسناد بازیابی‌شده آمده‌اند، در نظر گرفته‌ایم. در این روش، امتیاز کلیدواژه تنها بر اساس امتیاز شباهت اسناد دربردارنده آن کلیدواژه محاسبه می‌شود و ماهیت معنایی آن در نظر گرفته نمی‌شود.

در روش امتیازدهی بر اساس شباهت با سند جدید، مفهوم کلیدواژه و ارتباط معنایی آن با سند جدید به‌عنوان معیاری برای محاسبه امتیاز کلمه کاندید در نظر گرفته می‌شود. در این روش، هر کلیدواژه از نظر معنایی با تک‌تک واژگان سند جدید از نظر معنایی، بر اساس مدل برداری کلمه-به-بردار مقایسه می‌شود. برای این منظور برای هر کلیدواژه یک بردار معنایی بر اساس مدل آموزش داده‌شده کلمه-به-بردار، در نظر گرفته می‌شود.

هر کلیدواژه ممکن است از چندین کلمه تشکیل شده باشد. بنابراین، نیاز است که ابتدا عملیات پیش‌پردازش، همانند آنچه که در بخش پیش‌پردازش و نرمال‌سازی گفته شد، روی آن انجام شود. در صورتی که مدل کلمه-به-بردار برای کلیدواژه، برداری نداشته باشد، آنگاه نمی‌توان امتیازی بر اساس این روش برای آن حساب نمود و مقدار NAN برای آن لحاظ می‌شود.

در صورتی که کلیدواژه، پس از پیش‌پردازش بیش از یک کلمه داشته باشد (مثلاً عبارتی دو یا سه-کلمه‌ای باشد)، آنگاه برای هر یک از کلمات آن (در صورت موجود بودن) بر اساس مدل کلمه-به-بردار، یک نمایش برداری وجود خواهد داشت و نمایش برداری هر کلیدواژه در نهایت، میانگین بردار کلمات تشکیل‌دهنده آن خواهد بود. بنابراین، بردار نمایش معنایی برای هر کلیدواژه k_l به‌صورت زیر محاسبه می‌شود:

$$V_{sem}(k_l) = \begin{cases} V_{W2V}(k_l), & \text{if } k_l \text{ is a single word and W2V model exists} \\ \frac{1}{I}\sum_{x_i \, in \, k_l} V_{W2V}(x_i), & \text{if } k_l \text{ is a multi word} \\ NAN, & \text{otherwise.} \end{cases} \qquad (۱۱)$$

که I بیانگر تعداد کلمات تشکیل‌دهنده کلیدواژه پس از انجام عملیات پیش‌پردازش است.

اگر بردار کلیدواژه NAN نباشد، آنگاه می‌توان بردار آن را با بردار سند جدید d_{new} مقایسه نمود. هر یک از واژگان سند جدید، پس از پیش‌پردازش، یک نمایش برداری بر اساس مدل آموزش دیده کلمه-به-بردار دارند. واژگانی که در مدل موجود نباشند، در نظر گرفته نمی‌شوند.

کلیدواژه‌هایی که بردار آن‌ها NAN است، مانند داده‌های گم‌شده[1] هستند که به‌طور معمول، برای این داده‌ها از رویکردهایی مانند توزیع آماری داده (García-Laencina, Sancho-Gómez & Figueiras-Vidal 2010) استفاده می‌شود.

اگر واژگان d_{new} را که مدل کلمه-به-بردار هم برای آن‌ها موجود است، داخل مجموعه $\mathcal{R}_{\text{new}}$ قرار گیرد، آنگاه می‌توان شباهت معنایی هر واژه را با کلیدواژه k_l، بر اساس مدل برداری آن‌ها محاسبه کرد و در نهایت، امتیاز هر کلیدواژه بر اساس میانگین این شباهت‌ها حساب شود. بنابراین:

$$\boldsymbol{score_2(k_l, d_{new}) = \frac{1}{|\mathcal{R}_{new}|} \sum_{\mathbf{r}\in\mathcal{R}_{\text{new}}} \cos(\, \boldsymbol{V_{sem}(k_l), V_{W2V}(r)}\,)} \qquad (۱۲)$$

با ضرب دو امتیاز به‌دست‌آمده برای هر کلیدواژه امتیاز نهایی حاصل می‌شود:

$$\boldsymbol{score_{final}(k_l) = score_1(k_l, d_{new}) \times score_2(k_l, d_{new})} \qquad (۱۳)$$

چون هر کلیدواژه کاندید در واقع، خود، کلیدواژه‌ای از اسناد مشابه بازیابی شده است، امتیاز نهایی هر کلیدواژه کاندید بر اساس دو عامل محاسبه می‌شود:

- ◇ میزان شباهت سند حاوی این کلیدواژه به سند مورد نظر ($score_1$)؛
- ◇ میزان شباهت کلیدواژه به محتوای سند مورد نظر ($score_2$).

با رتبه‌بندی کلیهٔ کلیدواژه‌های کاندید بر اساس امتیاز آن‌ها ($score_{final}(k_l)$)، در نهایت، کلیدواژه‌های نهایی از بین کلیدواژه‌های کاندید با توجه به امتیاز نهایی آن‌ها پیشنهاد می‌شوند. یک رویکرد می‌تواند بر اساس تعداد از پیش تعیین شده باشد؛ به این مفهوم که از بین موارد کاندید، تعداد مشخصی از کلیدواژه‌ها که بیشترین امتیاز را دارند، انتخاب شوند. در رویکرد دیگر بر اساس کلیدواژه‌هایی که امتیازشان از یک حد بیشتر است، انتخاب می‌شوند. برای تعیین آستانه امتیاز یا تعداد مشخص کلیدواژه‌ها می‌توان آزمایش‌هایی با مجموعه داده‌های اعتبارسنجی انجام داد.

1. missing data

۵. پیاده‌سازی و ارزیابی

برای پیاده‌سازی روش پیشنهادی از کتابخانه متن‌باز «جنسیم»[1] (Rehurek & Sojka 2010) و «هضم» در «پایتون» استفاده شده است. کتابخانه جنسیم»، یک کتابخانه برای انجام عملیات پردازش متن نظیر نمایه‌سازی اسناد، بازیابی اطلاعات و مدل‌های برداری نمایش سند است.

مجموعه داده این پژوهش متشکل از ۳۱۸۴۰ پایان‌نامه و رساله مستخرج از «پایگاه گنج» است. برای هر سند (یعنی هر پایان‌نامه یا رساله)، اطلاعاتی نظیر عنوان، چکیده، موضوع و کلیدواژه‌های آن در دسترس است. داده‌ها بر اساس موضوعاتشان در سه دستهٔ هنر و ادبیات، فنی و مهندس، و علوم انسانی قرار دارند.

در جدول ۲ تا ۴، نمونه‌هایی از عنوان سند جدید، عناوین اسناد مشابه بازیابی‌شده به همراه کلیدواژه‌های اصلی آن‌ها و تعدادی از کلیدواژه‌های پیشنهادی در سه موضوع مهندسی، هنر و ادبیات، و علوم انسانی گزارش شده است.

جدول ۲. نمونه کلیدواژه‌های پیشنهادی بر اساس رویکرد پیشنهادی در موضوع هنر و ادبیات

عنوان سند جدید	**نگاهی نو به استاپ-موشن در گرافیک دیجیتال**
عنوان اسناد مشابه بازیابی‌شده	مطالعه و تحلیل تیزرها و انیمیشن‌های تبلیغاتی موفق جهان (تکنیک استاپ موشن) قابلیت‌های استاپ موشن در بازنمود فضاهای گروتسک، ذهنی و وهمی تأثیر گرافیک بر سیر تحول تکنیک‌های عکاسی دیجیتال
کلیدواژه‌های پیشنهادی	پویانمایی، تیزر، گرافیک، سوررئالیسم، ارتباطات، بازنمایی، تکامل، تکنولوژی
کلیدواژه‌های اصلی	تکنیک ایست-حرکت، تیزر، تکنیک استاپ-موشن، تیزر تبلیغاتی، پویانمایی، تبلیغات، خلاقیت
عنوان سند جدید	**جستاری در نقاشی معاصر ایران با تأکید بر موانع و چالش‌ها**
عنوان اسناد مشابه بازیابی‌شده	بررسی واقع‌گرایی در نقاشی معاصر ایران (دههٔ ۷۰ و ۸۰ شمسی) نمود زیبایی‌شناسانه هنر اسلامی در نقاشی معاصر ایران نوشتار به مثابه تصویر در نقاشی معاصر ایران
کلیدواژه‌های پیشنهادی	فرانوگرایی، ایران، نقاشی معاصر، واقع‌گرایی، هنر جدید، زیبایی‌شناسی، نقاشی، کلاژ، معماری
کلیدواژه‌های اصلی	هنر معاصر، نقاشی معاصر، نوگرایی، پسانوگرایی، کثرت‌گرایی، ایران، مدرنیت، مدرنیته، مدرنیسم، پست‌مدرنیسم، نقاشی معاصر ایران

1. Gensim

جدول ۳. نمونه کلیدواژه‌های پیشنهادی بر اساس رویکرد پیشنهادی در موضوع علوم انسانی

عنوان سند جدید	رابطه سبک‌های فرزندپروری ادراک‌شده با احساس تنهایی و ترس از ارزیابی منفی در نوجوانان
عنوان اسناد مشابه بازیابی‌شده	بررسی نقش واسطه‌ای ادراک از سبک‌های فرزندپروری والدین در رابطه بین سبک‌های تفکر پدران با سبک‌های تفکر دانش‌آموزان
	بررسی رابطه سبک‌های فرزندپروری ادراک‌شده، طرح‌واره‌های ناسازگار اولیه با احساس تنهایی در نوجوانان
	بررسی رابطه بین سبک‌های فرزندپروری با خلاقیت دانش‌آموزان مقطع متوسطه شهرستان قزوین
کلیدواژه‌های پیشنهادی	والدین، دانش‌آموز، ادراک، تنهایی، نوجوانان، خلاقیت، شیوه فرزندپروری، اقتدارگرایی
کلیدواژه‌های اصلی	شیوه پرورش کودک، احتمال خطر ادراکی، ترس، تنهایی، رابطه والد-کودک، ادراک، نوجوان، سبک‌های فرزندپروری ادراک‌شده
عنوان سند جدید	**سیاست جنایی قضائی ایران در قبال جرائم رایانه‌ای**
عنوان اسناد مشابه بازیابی‌شده	تأثیر پیشگیرانه قانون‌گذاری‌های نظام تقنینی ایران در حوزه جرائم (IT)
	بررسی جرم جعل رایانه‌ای در قلمرو حقوق کیفری ایران
	نقش محیط سایبر در وقوع جرائم
کلیدواژه‌های پیشنهادی	جرم، پیشگیری، جرم‌شناسی، اقتصاد، مدرنیت، اینترنت، امنیت، محرومیت، تمامیت داده، بزه‌کاری، فضای سایبر، فضای مجازی
کلیدواژه‌های اصلی	سیاست کیفری، حقوق جزا، جرم سایبری، رسیدگی قضایی، جرم سایبری یا رایانه‌ای، جرم رایانه‌ای، امنیت داده‌ها، فضای مجازی، رویه قضایی، جرم‌شناسی

جدول ٤. نمونه کلیدواژه‌های پیشنهادی بر اساس رویکرد پیشنهادی در موضوع فنی و مهندسی

عنوان سند جدید	ارائه یک مدل مفهومی دیفرانسیلی به‌منظور شبیه‌سازی عملکرد حرارتی و بهینه‌سازی موتور استرلینگ
عنوان اسناد مشابه بازیابی‌شده	مدل‌سازی دینامیکی و تحلیل عملکرد حالت‌گذرای موتور هیسترزیس دور بالا
	مدل‌سازی ترموهیدرلیکی بازیاب موتور استرلینگ
	طراحی و ساخت موتور استرلینگ نوع آلفا و مقایسه نتایج آزمایشگاهی با پیش‌بینی‌های مدل
کلیدواژه‌های پیشنهادی	بهینه‌سازی، شبیه‌سازی، جریان متلاطم، افت فشار، ترمودینامیک، اصطلاک، بازده، عملکرد، ترموهیدرولیک، موتور استرلینگ، الگوس جریان، سیلندر موتور، نشت، همرفت، جدایش

عنوان سند جدید	ارائه یک مدل مفهومی دیفرانسیلی به‌منظور شبیه‌سازی عملکرد حرارتی و بهینه‌سازی موتور استرلینگ
کلیدواژه‌های اصلی	موتور استرلینگ، تحلیل ترمودینامیکی، تحلیل حرارتی، عملکرد حرارتی، اثر شاتل، تحلیل پلی تروپیک، ترمودینامیک سرعت محدود، ویژگی ترمودینامیکی، شبیه‌سازی، بهینه‌سازی، افت فشار، اصطکاک
عنوان سند جدید	**تأثیر متغیرهای جوشکاری همزن اصطکاکی بر خواص اتصال غیرهمجنس آلیاژهای آلومینیم ۲۰۲۴ به ۷۰۷۵**
عنوان اسناد مشابه بازیابی‌شده	بررسی خواص مکانیکی و ریزساختار جوشکاری آلیاژ آلومینیوم ۲۰۲۴-T۳ به روش جوشکاری اصطکاکی اغتشاشی
	جوشکاری همزن اصطکاکی آلیاژ منیزیم AZ۳۱
	بررسی تأثیر پارامترهای رئوکست روی ریزساختار و خواص مکانیکی آلیاژ Al-۶Si-۲Mg
کلیدواژه‌های پیشنهادی	جوشکاری اصطکاکی اغتشاشی، خواص مکانیکی، استحکام کششی، ریزساختار، آلیاژ آلومینیم، جوشکاری اصطکاکی، جوشکاری همزن اصطکاکی، خمش (شکل‌دهی)، تحلیل عددی، ساختار گلبولی، آلیاژ سیلیسیم
کلیدواژه‌های اصلی	جوشکاری اصطکاکی اغتشاشی، خواص مکانیکی، آلیاژ آلومینیوم، جوشکاری اصطکاکی، سرعت چرخشی، ریزساختار، خمش (شکل‌دهی)، استحکام کششی

۵-۱. تنظیم پارامترها و ارزیابی

روش پیشنهادی از پارامترهای مختلفی تشکیل شده که برخی از آن‌ها مربوط به بخش آموزش و برخی دیگر مربوط به بخش بازیابی و امتیازدهی کلیدواژه‌های کاندید است. پارامترهای بخش آموزش، مربوط به مدل کلمه-به-بردار عبارت‌اند از:

- ◇ اندازه پنجره: به‌طور معمول بین ۵ تا ۸ تنظیم می‌شود، که در اینجا ۸ درنظر گرفته شده است؛
- ◇ ابعاد بردار: بعد بردار برابر عدد ۱۰۰ در نظر گرفته شده است؛
- ◇ نمونه‌برداری: مقداری که به‌صورت معمول در نظر گرفته می‌شود، عدد ۲۵ است؛
- ◇ کمترین فراوانی: به‌صورت معمول بین ۲ تا ۵ تنظیم می‌شود که در اینجا ۵ لحاظ شده است.

در بخش بازیابی اسناد مشابه و امتیازدهی، ۴ پارامتر مختلف وجود دارد:

- ◇ تعداد اسناد مشابه انتخاب‌شده (M)؛
- ◇ وزن روش بازیابی براساس مدل برداری ماتریس سند-کلمه (α)؛
- ◇ تعداد کلیدواژه‌های انتخابی از بین کلیدواژه‌های کاندید (K).

برای تنظیم این سه پارامتر، بر اساس مجموعه داده اعتبارسنجی، آزمایش‌های

متعددی بر اساس معیارهای ارزیابی انجام شده است. برای ارزیابی و تنظیم پارامترها در روش پیشنهادی به‌راحتی نمی‌توان معیار دقت[1] وبازخوانی[2] و معیار F[3] که به‌صورت معمول در این‌گونه از مسائل استفاده می‌شود، به‌کار گرفت. زیرا برخی از کلیدواژه‌های پیشنهادی در فهرست کلیدواژه‌های اصلی نیستند، لیکن با موضوع ارتباط دارند. افزون بر آن، در روش پیشنهادی فهرست حداکثری از کلیدواژه‌ها پیشنهاد می‌شود تا از بین آن‌ها انتخاب شود. این فهرست می‌تواند در بازیابی اطلاعات مورد استفاده قرار گیرد. بنابراین، تعداد کلیدواژه‌های پیشنهادی به‌طور معمول، سه الی چهار برابر بیش از تعداد کلیدواژه‌های اصلی است که بر همین اساس به‌راحتی نمی‌توان دقت را محاسبه نمود. زیرا دقت نسبت تعداد اقلام درست به کل اقلام است که چون تعداد کل اقلام معمولاً سه الی چهار برابر تعداد کل اقلام درست است، این مقدار، معنادار نخواهد بود. در این‌گونه موارد، برای محاسبه دقت، معیارهای دیگری مانند «میانگین حد متوسط دقت»[4] که به‌طور معمول در سیستم‌های پیشنهاددهنده از آن‌ها نیز استفاده می‌شود، به‌کار گرفته می‌شوند. بنابراین، برای ارزیابی از معیار MAP به‌جای معیار دقت استفاده می‌شود.

برای ارزیابی روش پیشنهادی از سه روش زیر استفاده می‌شود:

- ◇ معیار بازخوانی؛
- ◇ معیار MAP: این معیار در سیستم‌های پیشنهاددهنده و بازیابی اطلاعات که تعداد اقلام زیادی به کاربر پیشنهاد می‌شود و ترتیب پیشنهاد نیز اهمیت دارد، استفاده می‌گردد. در این معیار لزوماً با افزایش تعداد کلیدواژه‌های پیشنهادی، مقدار دقت کاهش نمی‌یابد، بلکه آنچه که اهمیت دارد این است که کدام‌یک از موارد پیشنهادشده مرتبط در بالای فهرست پیشنهادی، امتیاز بالاتری دارند (Manning & Raghavan 2009; Shani & Gunawardana 2011).
- ◇ ارزیابی انسانی: افراد متخصص حوزه‌های موضوعی برای تعیین ارتباط یا عدم ارتباط کلیدواژه‌های پیشنهادی به‌کار گرفته شوند و نظرات آن‌ها در نهایت، ملاک ارزیابی قرار گیرد.

برای تنظیم پارامترهای مسئله و بررسی عملکرد روش پیشنهادی با تغییرات پارامترها،

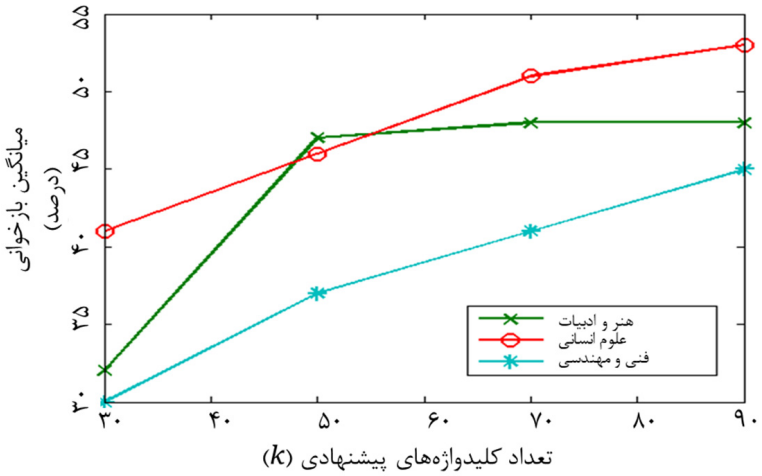
1. precision
2. recall
3. F-measure
4. mean average precision (MAP)

آزمایش‌های مختلفی انجام شده که نتایج آن‌ها در جدول ۵، و شکل ۳، آمده است. با انجام آزمایش‌های مختلف α=0.7 انتخاب شده است.

جدول ۵. بررسی عملکرد روش پیشنهادی با تغییر تعداد کلیدواژه‌های پیشنهادی (α=0.7,M=20)

موضوع	تعداد کلیدواژه‌های پیشنهادی (K)	بازخوانی (درصد)
هنر و ادبیات	۳۰	۳۲
	۵۰	۴۲
	۷۰	۴۸
	۹۰	۴۸
علوم انسانی	۳۰	۴۱
	۵۰	۵۰
	۷۰	۵۱
	۹۰	۵۲
فنی و مهندسی	۳۰	۳۰
	۵۰	۳۷
	۷۰	۴۱
	۹۰	۴۴

۳



شکل ۳. عملکرد روش پیشنهادی با تغییر پارامتر تعداد کلیدواژه‌های پیشنهادی (α=0.7, M=20)

معیار روش MAP براساس پیشنهاد K کلیدواژه، میانگین مقادیر $AP@K$ برای M سند مختلف است. مقدار $AP@K$ (حد متوسط دقت در K) برای یک سند d_i به‌صورت زیر محاسبه می‌شود:

$$AP@N(d_i) = \frac{1}{m}\sum_{k=1}^{K} P(k).rel(k) \qquad (۱۴)$$

که $rel(k)$ در صورتی که کلیدواژه kام مرتبط با d_i باشد، مقدار یک دارد و در غیر این صورت صفر است و m تعداد کل کلیدواژه‌های مرتبط درکل فضای مسئله است. $P(k)$ دقت را برای k کلیدواژه اول نشان می‌دهد که به‌صورت زیر محاسبه می‌شود:

$$P(k) = \frac{\#\{recommended\ keywords\} \cap \{original\ keywords\}}{k} \qquad (۱۵)$$

سرانجام، $MAP@N$ یعنی «میانگین حد متوسط دقت»، به‌صورت زیر محاسبه می‌شود:

$$\boldsymbol{score_{final}(k_l) = score_1(k_l, d_{new}) \times score_2(k_l, d_{new})} \qquad (۱۶)$$

که N در آن تعداد کل اسناد مجموعه آزمایش است. در واقع، $MAP@N$ میانگین مقدار $AP@K$ برای N سند مختلف است. بر همین اساس برای داده‌های آزمایشی، این مقدار محاسبه شده که در جدول ۶، نمایش داده شده است.

جدول ٦. مقدار میانگین حد متوسط دقت (MAP@N) برای K= 50

موضوع	میانگین حد متوسط دقت (MAP@N)
هنر و ادبیات	۲۷/۳
علوم انسانی	۲۷/۵
فنی و مهندسی	۲۲/۲

۵-۲. ارزیابی انسانی

به‌منظور ارزیابی کلیدواژه‌های پیشنهادی بر اساس نظر متخصص، از متخصصان نمایه‌ساز «ایرانداک» استفاده شد. در هر یک از سه حوزه هنر و ادبیات، فنی و مهندسی، و علوم انسانی نظر چهار متخصص دریافت شد. برای دریافت نظرات متخصصان، سامانه ارزیابی مبتنی بر وب طراحی شد تا آن‌ها بتوانند نظراتشان را در آن وارد نمایند. برای هر کلیدواژه پیشنهادی در هر سند، بر اساس ارتباط معنایی بین کلیدواژه و سند، سه گزینه «خیلی کلی است»، «مرتبط» و «غیرمرتبط» طراحی شد. در صورتی که فرد، گزینه «مرتبط» را انتخاب کند، نیاز است که بر اساس میزان ارتباط یکی از اعداد ۱، ۲ یا ۳ (عدد ۱ بیانگر

کمترین ارتباط و عدد ۳ بیانگر بیشترین ارتباط) را انتخاب کند. نتایج به‌دست‌آمده پس از ارزیابی انسانی در جدول ۷، آمده است.

با بررسی امتیاز کلیدواژه‌های مرتبط می‌توان دریافت که تا چه حد این کلیدواژه‌ها مرتبط هستند و از نظر معنایی و مفهومی با سند مورد نظر همخوانی دارند. برای این منظور «مجموع امتیاز کلیدواژه‌های هر سند» نیز محاسبه شده است. سپس، این مقدار با «حداکثر مقدار امتیاز ممکن» مقایسه شده است. به‌عنوان مثال، برای سندی که ۱۰ کلیدواژه مرتبط شناسایی شده، حداکثر امتیاز کلیدواژه‌ها مقدار ۳۰ خواهد بود، زیرا هر کلیدواژه حداکثر ۳ امتیاز از لحاظ میزان ارتباط می‌تواند داشته باشد. برای محاسبه «متوسط میزان ارتباط کلیدواژه‌های پیشنهادی (درصد)» برای این سند، مجموع امتیاز کلیدواژه‌های مرتبط بر مقدار ۳۰ تقسیم، و در ۱۰۰ ضرب شده است. هرچه این مقدار بیشتر باشد، نشان می‌دهد که کلیدواژه‌های مرتبط شناسایی‌شده از نظر معنایی و مفهومی با سند مرتبط‌تر هستند.

جدول ۷. نتایج ارزیابی انسانی

معیار ارزیابی حوزه	متوسط تعداد کلیدواژه‌های خیلی کلی	متوسط تعداد کلیدواژه‌های مرتبط	متوسط مجموع امتیاز کلیدواژه‌ها	متوسط حداکثر امتیاز ممکن برای کلیدواژه‌های مرتبط	متوسط میزان ارتباط کلیدواژه‌های پیشنهادی (درصد)
فنی و مهندسی	۴/۶۹	۱۰/۸۴	۲۲/۸۷	۳۲/۵۲	۷۰/۴۴
علوم انسانی	۱۰/۰۲	۲۱/۰۹	۴۲/۴۶	۶۳/۲۷	۶۷/۰۷
هنر و ادبیات	۷/۶۵	۱۳/۶۶	۲۷/۳۵	۴۰/۹۸	۶۷/۱۷

یافته‌ها

در این پژوهش یک سیستم پیشنهاددهنده برای پیشنهاد کلیدواژه برای مستندات علمی فارسی توسعه یافته است. از آنجا که مسئله استخراج کلیدواژه را نمی‌توان به‌راحتی در قالب قوانین مشخصی مدل‌سازی کرد، سیستم پیشنهاددهنده پیشنهادی بر مبنای استدلال نمونه‌محور عمل می‌کند که در آن با استناد به کلیدواژه‌های اسناد مشابه و میزان تشابه آن‌ها با سند، کلیدواژه‌های مناسب برای سند پیشنهاد می‌شود. در واقع، از دانش ضمنی که در اسنادِ مشابهِ نمایه‌شده وجود دارد، برای یافتن کلیدواژه برای یک سند استفاده می‌شود.

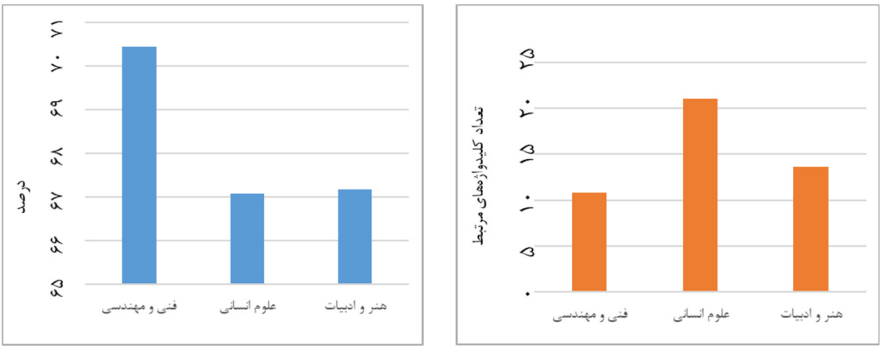
با بررسی نتایج پیاده‌سازی و ارزیابی سیستم پیشنهادی می‌توان گفت که

کلیدواژه‌های اسناد موجود در پیکره‌های تخصصی، همانند مجموعه اسناد «گنج» که در این پژوهش استفاده شده، معمولاً محدود به پنج الی هفت واژه است، در حالی که سیستم پیشنهاددهنده پیشنهادی می‌تواند بیش از این مقدار، واژگان مرتبط ارائه دهد. در حقیقت، بر اساس جدول ۷، می‌توان گفت که به‌طور متوسط حدود ۱۵ کلیدواژه از بین کلیدواژه‌های پیشنهادشده در کل داده‌ها مرتبط هستند. با مقایسه این عدد با متوسط تعداد کلیدواژه‌های اسناد که ۷ است، می‌توان گفت که الگوریتم پیشنهادی در شناسایی کلیدواژه‌های مرتبط از نظر تعداد، توانسته حداقل به اندازه کلیدواژه‌های اصلی سند، کلیدواژه مرتبط بیابد.

کلیدواژه‌های یک سند می‌تواند بیش از تعدادی باشد که در ابتدا به‌صورت دستی توسط نویسنده آن سند یا نمایه‌ساز به آن اختصاص یافته است. در واقع، به‌رغم اینکه در ابتدا تعدادی از واژگان برای یک سند به‌عنوان کلیدواژه شناسایی نشده‌اند، اما سیستم پیشنهاددهنده توانسته تعدادی از آن‌ها را به‌عنوان کلیدواژه‌های مرتبط شناسایی کند تا در نهایت، در فرایند بازیابی اطلاعات نیز از آن‌ها استفاده شود.

نتایج ارزیابی انسانی در جدول ۷، را می‌توان در شکل ۴ (الف) و (ب) برای مقایسه سه حوزه نمایش داد. بر اساس این شکل می‌توان گفت که به‌رغم اینکه در حوزه فنی و مهندسی کلیدواژه‌های کمتری مرتبط شناسایی شده، لیکن میزان ارتباط کلیدواژه‌های مرتبط بیشتر است. به بیان دیگر، در حوزه فنی و مهندسی کلیدواژه‌های تخصصی بیشتری شناسایی شده‌اند. در دو حوزه دیگر، به‌رغم تعداد کلیدواژه‌های مرتبط بیشتر، میزان ارتباط آن‌ها با سند، به‌نسبت حوزه فنی و مهندسی، کمتر است.

از آنجا که سیستم پیشنهادی تنها از کلیدواژه‌های اسناد مشابه برای تخصیص کلیدواژه استفاده می‌کند، در همان حال، ممکن است واژگانی غیرمرتبط نیز در انتها ارائه دهد. این سیستم در کنار سایر روش‌های استخراج کلیدواژه که به‌طور مستقیم از متن سند، کلیدواژه پیشنهاد می‌دهند، می‌تواند اثربخشی بیشتری داشته باشد و تکمیل‌کننده باشد. در این صورت، دانش ضمنی نمایه‌ساز انسانی که از طریق سیستم پیشنهاددهنده استخراج می‌شود، در کنار نتایج روش‌های آماری و هوشمند استخراج کلیدواژه که از متن سند استفاده می‌کنند، می‌تواند چالش قابلیت اطمینان در روش‌های انسانی و نیز اعتبار روش‌های ماشینی را مورد هدف قرار دهد و تا حدی مرتفع سازد.



الف) تعداد کلیدواژه‌های مرتبط **ب) میزان ارتباط کلیدواژه‌های مرتبط**

شکل ٤. تعداد و مقایسه میزان ارتباط کلیدواژه‌های مرتبط در حوزه‌های مختلف

۷. نتیجه‌گیری و پژوهش‌های آتی

در این پژوهش، روشی برای پیشنهاد خودکار کلیدواژه برای مستندات علمی فارسی پیشنهاد شده است که در آن تعداد کلیدواژه‌های بیشتری از تعداد کلیدواژه‌های معمول برای یک سند علمی پیشنهاد می‌شود. در بیشتر پژوهش‌های پیشین برای استخراج خودکار کلیدواژه از مجموعه‌ای از ویژگی‌های آماری و زبانی کلمات در متون برای شناسایی و استخراج عبارت‌های کلیدی استفاده شده است. در روش‌های با ناظر[1] از این ویژگی‌ها برای ساخت مدل و استخراج کلیدواژه استفاده شده و در روش‌های بدون ناظر[2] نیز قوانینی طراحی شده که براساس آنها عبارات کاندید به‌عنوان کلیدواژه، براساس این ویژگی‌ها امتیازدهی می‌شوند و در نهایت، آنهایی که بیشترین امتیاز را دارند به‌عنوان کلیدواژه‌های مناسب انتخاب می‌شوند. براساس مطالعات گذشته، مسئله استخراج و تخصیص خودکار کلیدواژه جزء مسائلی است که نمی‌توان آن را در قالب قوانین مشخص به‌صورت کامل مدل‌سازی کرد. برای این مسائل، روش‌های استدلال نمونه‌محور اثربخش است که در این پژوهش نیز برای طراحی سیستم‌پیشنهاددهنده از آن استفاده شده است. در استدلال نمونه‌محور فرض بر این است که مسائل مشابه می‌توانند راه حل‌های مشابه داشته باشند. در روش پیشنهادی نیز برای پیشنهاد کلیدواژه برای یک سند، از اطلاعات اسناد مشابهی که پیش‌تر نمایه شده‌اند و دارای کلیدواژه هستند، استفاده شده است. مشابه چنین رویکردی در پژوهش‌های پیشین تنها برای نمایه‌سازی بخشی از مقالات پزشکی انگلیسی در پایگاه

1. supervised 2. unsupervised

تخصصی مربوطه استفاده شده و برای مستندات علمی فارسی در قالب استدلال نمونه‌محور به کار گرفته نشده است. در پژوهش‌های پیشین براساس نمایه‌سازی مستندات فارسی از ویژگی‌های آماری و زبانی بهره گرفته شده، لیکن از مشخصات اسناد مشابه و رویکرد استدلال نمونه‌محور استفاده نشده است.

دو بخش اصلی در روش پیشنهادی وجود دارد: بخش آموزش برای ساخت مدل و بخش بازیابی و پیشنهاد کلیدواژه. مجموعه داده استفاده‌شده متشکل از ۳۱۸۴۰ پایان‌نامه و رساله از دانشگاه‌های ایران است که در «پایگاه گنج» متعلق به «ایرانداک» ثبت شده‌اند. برای بهبود عملکرد روش پیشنهادی، مجموعه داده‌ها به سه حوزه موضوعی هنر و ادبیات، علوم انسانی، و فنی-مهندسی تقسیم‌بندی شده و روش پیشنهادی برای هر دسته موضوعی به‌صورت جداگانه پیاده‌سازی شد. برای بررسی اثرگذاری پارامترهای مختلف در روش پیشنهادی، آزمایش‌های مختلفی انجام و عملکرد آن بر اساس شاخص بازخوانی و MAP@K ارزیابی گردید. بر اساس نتایج به‌دست‌آمده می‌توان گفت که آموزش مدل‌های جداگانه برای هر حوزه موضوعی بر روی مجموعه داده‌های موضوعی می‌تواند نتایج بهتری ارائه دهد. از آنجا که نتایج روش پیشنهادی وابسته به تعداد اسناد مشابه بازیابی شده است، الگوریتم بازیابی اسناد مشابه نقش مهمی را در عملکرد روش پیشنهادی دارد.

روش پیشنهادی می‌تواند تعداد کلیدواژه‌های بیشتری را نسبت به سایر روش‌های پیشین ارائه دهد. همچنین، از آنجا که در این روش از کلیدواژه‌های اسناد مشابه برای نمایه‌سازی استفاده می‌شود، چالش‌هایی که به‌طور عام در انتخاب واژگان کنترل‌شده برای نمایه‌سازی وجود دارد، در این روش دیده نمی‌شود؛ زیرا کلیدواژه‌هایی که برای سند در نظر گرفته می‌شوند از بین کلیدواژه‌هایی هستند که پیش‌تر به اسناد مشابه نیز تخصیص داده شده و پیش‌تر از نظر ساختاری و ماهیت کلیدواژه بودن تأیید شده‌اند. لیکن از آنجا که ممکن است این کلیدواژه‌ها نتوانند همه مفاهیم موجود در یک سند را بیان کنند (مثلاً ممکن است مفاهیم جدیدی در سند باشد که در اسناد نمایه‌شده پیشین تحت هیچ عنوان یا مفهوم مشابهی نباشند) می‌توان از سایر منابع نیز برای تکمیل این کلیدواژه‌ها استفاده کرد. بنابراین، کلیدواژه‌های پیشنهادشده در کنار دانش نمایه‌ساز یا سایر روش‌های استخراج کلیدواژه که بر اساس استخراج واژگان مهم از متن سند عمل می‌کنند، می‌توانند به‌کار گرفته شده و به‌عنوان یکی از منابع دانشی در فرایند نمایه‌سازی لحاظ شوند.

برای تحقیقات آتی می‌توان از روش‌های استخراج کلیدواژه از متن که بر مبنای

اطلاعـات آمـاری واژگان کاندیـد درون متـن عمـل می‌کننـد، در کنـار روش پیشـنهادی اسـتفاده کـرد. همچنیـن، از آنجـا کـه عملکـرد روش‌هـای مدل‌سـازی ماننـد کلمه-به-بـردار وابسـته بـه حجـم داده‌هـای آمـوزش اسـت، می‌تـوان از مجموعـه داده بزرگ‌تـر بـا داده‌هـای موضوعـی متنوع‌تـر اسـتفاده نمـود. اسـتفاده از مدل‌هـای مبتنـی بـر یادگیـری عمیـق ماننـد ماننـد «بـرت»[1] به‌جـای مـدل کلمه-به-بـردار می‌توانـد در افزایـش دقـت مؤثـر باشـد.

References

Adomavicius, G., & A. Tuzhilin. 2005. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17 (6): 734–749. https://doi.org/10.1109/TKDE.2005.99

Ahn, J., R. Farzan, & P. Brusilovsky. 2005. Comprehensive personalized information access in an educational digital library. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05), 9–18. https://doi.org/10.1145/1065385.1065388

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

Baltrunas, L. 2008. Exploiting contextual information in recommender systems. *Proceedings of the 2008 ACM Conference on Recommender Systems*, 295–298. Lausanne.

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.

Bohra, A., & N. C. Barwar. 2022. A Deep Learning Approach for Plagiarism Detection System Using BERT. In M. Saraswat, H. Sharma, K. Balachandran, J. H. Kim, & J. C. Bansal (Eds.), *Congress on Intelligent Systems* (pp. 163–174). Springer Nature Singapore.

Bridge, D., M. H. Goker, L. McGinty, & B. Smyth. 2005. Case-based recommender systems. *The Knowledge Engineering Review* 20 (03): 315. https://doi.org/10.1017/S0269888906000567

Burke, R. 2007. Hybrid Web Recommender Systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The Adaptive Web: Methods and Strategies of Web Personalization* (pp. 377–408). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72079-9_12

Firoozeh, N., A. Nazarenko, F. Alizon, & B. Daille. 2020. Keyword extraction: Issues and methods. *Natural Language Engineering* 26 (3): 259–291. https://doi.org/10.1017/S1351324919000457

Frické, M. 2012a. *Logic and the Organization of Information*. Springer. https://doi.org/https://doi.org/10.1007/978-1-4614-3088-9

_____. 2012b. *Logic and the Organization of Information*. Springer. https://doi.org/https://doi.org/10.1007/978-1-4614-3088-9

García-Laencina, P. J., J. L. Sancho-Gómez, & A. R. Figueiras-Vidal. 2010. Pattern classification with missing data: a review. *Neural Computing and Applications*, *19* (2), 263–282. https://doi.org/10.1007/s00521-009-0295-6

Hasan, K. S., & V. Ng. 2014. Automatic Keyphrase Extraction: A Survey of the State of the Art. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1262–1273. https://doi.org/10.3115/v1/P14-1119

1. bidirectional encoder representations from transformers (BERT)

Huang, M., A. Névéol, & Z. Lu. 2011. Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association : JAMIA*, *18* (5), 660–667. https://doi.org/10.1136/amiajnl-2010-000055

Humphreys, P., R. Mcivor, & F. Chan. 2003. Using case-based reasoning to evaluate supplier environmental management performance. *Expert Systems with Applications*, *25*, 141–153. https://doi.org/10.1016/S0957-4174(03)00042-3

Khozani, S. M. H., & H. Bayat. 2011. Specialization of keyword extraction approach to Persian texts. *Proceedings of the 2011 International Conference of Soft Computing and Pattern Recognition, SoCPaR 2011*, 112–116. https://doi.org/10.1109/SoCPaR.2011.6089124

Kian, H., & M. Zahedi. 2011. an Efficient Approach for Keyword Selection; Improving Accessibility of Web Contents By General Search Engines. *International Journal of Web & …*, 2 (4): 81–90.

Klamma, R., P. M. Cuong, & Y. Cao. 2009. You Never Walk Alone: Recommending Academic Events Based on Social Network Analysis. In J. Zhou (Ed.), *Complex Sciences* (pp. 657–670). Berlin Heidelberg: Springer.

Kolodner, J. 1992. An introduction to case-based Reasoning. *Artificial Intelligence Review 6*: 3–34.

Manning, C. D., & P. Raghavan. 2009. An Introduction to Information Retrieval. *Online*, *1*, 1. https://doi.org/10.1109/LPT.2009.2020494

Mao, Y., & Z. Lu. 2017. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. *Journal of Biomedical Semantics*, 8 (1): 15. https://doi.org/10.1186/s13326-017-0123-3

Mehrabi, E., A. Mohebi, & A. Ahmadi. 2021. Improved Keyword Extraction for Persian Academic Texts Using RAKE Algorithm; Case Study: Persian Theses and Dissertations. *Iranian Journal of Information Processing and Management* 37 (1). https://doi.org/10.52547/jipm.37.1.197

Mikolov, T., K. Chen, G. Corrado, & J. Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*. 1–12. https://doi.org/10.1162/153244303322533223

Oh, S., Z. Lei, W. C. Lee, P. Mitra, & J. Yen. 2013. CV-PCR: a context-guided value-driven framework for patent citation recommendation. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management* 2291–2296. New York.

Pay, T., & S. Lucci. 2017. Automatic keyword extraction: An ensemble method. *2017 IEEE International Conference on Big Data (Big Data)*, 4816–4818. https://doi.org/10.1109/BigData.2017.8258552

Recommender Systems Handbook. 2011. In P. B. Ricci, Francesco;Rokach, Lior;Shapira, Bracha; Kantor (Ed.), *Springer*. https://doi.org/10.1007/978-0-387-85820-3

Rehurek, R., & P. Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta.

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

Roul, R. K., & K. Arora. 2019. A nifty review to text summarization-based recommendation system for electronic products. *Soft Computing*, *23*(24), 13183–13204. https://doi.org/10.1007/s00500-019-03861-3

Sayyadiharikandeh, M., M. Ghodsi, & M. Naghibi. 2012. PostRank: A New Algorithm for Incremental Finding of Persian Blog Representative Words. *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics*, 17:1--17:6. https://doi.org/10.1145/2254129.2254152

Shani, G., & A. Gunawardana. 2011. *Evaluating recommendation systems. In Recommender systems handbook* (pp. 257–297). Boston: Springer.

Sharma, C., M. Jain, & A. Aggarwal. 2018. Keyword Extraction Using Graph Centrality and WordNet. In S. Chakraverty, A. Goel, & S. Misra (Eds.), *Towards Extensible and Adaptable Methods in Computing* (pp. 363–372). Springer Singapore. https://doi.org/10.1007/978-981-13-2348-5_27

Smyth, B. 2007. Case-based recommendation. In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds) *The Adaptive Web. Lecture Notes in Computer Science* 4321. Berlin, Heidelberg: 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

Wang, F. A. N., N. Shi, & B. E. N. Chen. 2010. A Comprehensive Survey of the Reviewer Assignment Problem. *International Journal of Information Technology & Decision Making*, 09 (04): 645–668. https://doi.org/10.1142/S0219622010003993

Weber, R. O., K. D. Ashley, & S. Brüninghaus. 2005. Textual case-based reasoning. *The Knowledge Engineering Review* 20 (3): 255–260.

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.

آزاده فخرزاده

دارای مدرک تحصیلی دکتری در رشتهٔ پردازش تصویر از دانشگاه اوپسالای سوئد است. ایشان هم‌اکنون استادیار پژوهشکدهٔ فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

پردازش تصویر، یادگیری ماشین، کلان‌داده‌ها، و یادگیری عمیق از جمله علایق پژوهشی وی است.

مرضیه زرین‌بال

متولد سال ۱۳۶۲، دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی صنایع از دانشگاه صنعتی امیرکبیر است. ایشان هم‌اکنون استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است.

طراحی سیستم‌های اطلاعاتی، بازی‌وارسازی، پردازش تصویر و منطق فازی از جمله علایق پژوهشی وی است.