Automatic Annotation of Images in Persian Scientific Documents Based on Text Analysis Methods

Azadeh Fakhrzadeh*
PhD in Digital Image Processing; Assistant Professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran Email: Fakhrzadeh@irandoc.ac.ir

Mohadeseh Rahnama
MSc in Computer Engineering; Alzahra University; Tehran, Iran; Email: m.rahnama@student.alzahra.ac.ir

Jalal A. Nasiri
PhD in Computer Engineering; Assistant Professor; Faculty of Mathematical Sciences; Ferdowsi University of Mashhad; Mashhad, Iran Email: jnasiri@um.ac.ir

Iranian Journal of Information Processing and Management

Received: 07, Apr. 2021 **Accepted: 16, May 2021**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 37 | No. 3 | pp. 895-918
Spring 2022
https://doi.org/10.35050/JIPM010.2022.839

Abstract: In this paper a new method for annotating images in Persian scientific documents is suggested. Images in scientific documents contain valuable information. In many cases, by analyzing images one can understand the main idea and important results of the document. Due to explosive growth of image data, automatic image annotation has attracted extensive attention and become one of the growing subjects in the literature. Image annotation is the first step in image retrieval methods, in which descriptive tags are assigned to each image.
Here, for image annotation the associated text is used. The caption and the part of the document that includes the reference to the image are considered. None phrases in the associated text are ranked based on five different methods: term frequency, inverse document frequency, term frequency–inverse document frequency, cosine similarity between word embedding of noun phrases in the text and the caption and using both term frequency–inverse document frequency and cosine similarity methods. Image tags in every method are the noun phrases with the highest rank. Suggested methods are evaluated on the test data from Iran scientific information database (Ganj), the main database of Persian scientific documents. Term frequency–inverse document frequency method gives the best results.

Keywords: Image Tagging, Text Analysis, Image Annotation, Image Retrieval, Metadata Extraction, Information Technology

* Corresponding Author

ارائهٔ روشی برای برچسب زدن تصاویر موجود در متون علمی فارسی با استفاده از روش‌های پردازش متن

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

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

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

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۰۱/۱۸ | **پذیرش:** ۱۴۰۰/۰۳/۲۶ | مقاله برای اصلاح به مدت ۷ روز نزد پدیدآوران بوده است.

چکیده: در این مقاله یک روش جدید برای برچسب‌گذاری تصاویر موجود در متون علمی فارسی معرفی می‌شود. در اسناد و مقالات علمی، تصاویر حاوی اطلاعات مهمی هستند و در بسیاری از موارد به‌تنهایی با بررسی آن‌ها می‌توان به ایدهٔ اصلی و یا نتایج مهم مقالهٔ علمی پی برد، بدون اینکه لازم باشد کل مقاله را مطالعه کرد. به‌خاطر رشد روزافزون داده‌های تصویری، بازیابی تصاویر از اسناد علمی توجه زیادی را به خود جلب کرده و به یک موضوع روبه‌رشد در ادبیات تبدیل شده است. اولین قدم در بازیابی تصاویر تخصیص برچسب‌های توصیف‌کننده به هر تصویر است.

در اینجا برای استخراج برچسب تصویر از متن سندی که تصویر به آن تعلق دارد، استفاده شده است. زیرنویس و قسمتی از متن سند که در آن به تصویر مورد نظر اشاره شده است، در نظر گرفته می‌شود. عبارات اسمی در متنِ همراه تصویر با استفاده از پنج روش متفاوتِ فراوانی عبارات در سند، معکوس فراوانی سند، فراوانی کلمه-معکوس فراوانی سند، شباهت کسینوسی عبارات با زیرنویس، و ترکیب روش فراوانی کلمه-معکوس فراوانی سند و شباهت کسینوسی با زیرنویس رتبه‌بندی می‌شوند. در هر

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 ۳ | ص‌ص ۸۹۵-۹۱۸
بهار ۱۴۰۱
https://doi.org/10.35050/JIPM010.2022.839

روش، برچسب‌های انتخابی برای تصویر، عبارات اسمی با رتبهٔ بالاتر در آن روش است. روش‌های معرفی‌شده با استفاده از دادهٔ آزمایشی از پایگاه اطلاعات علمی ایران (گنج) که منبع اصلی اسناد علمی فارسی است، ارزیابی می‌شوند. طبق نتایج به‌دست‌آمده در این تحقیق روش فراوانی کلمه-معکوس فراوانی سند بهترین روش برای برچسب زدن تصاویرموجود در اسناد علمی است.

کلیدواژه‌ها: برچسب زدن تصویر، نشانه‌گذاری تصویر، بازیابی تصویر، پردازش متن، استخراج فراداده، فناوری اطلاعات

۱. مقدمه

محققان در مقالات علمی از نمودارها برای خلاصه کردن نتایج و از فلوچارت‌ها برای بیان گام‌های الگوریتم خود و مقایسهٔ روش‌های مختلف استفاده می‌کنند. به همین دلیل، امروزه، داده‌های تصویری رشد زیادی کرده و به همان نسبت تقاضا برای یک ابزار مؤثر که بتواند این دسته از اطلاعات را بازیابی کند، افزایش پیدا کرده است. دسترسی به این عناصر از یک سند علمی در بسیاری از موارد می‌تواند دورنمایی از ایدهٔ مقاله و نتایج کلی آن را، بدون نیاز به مطالعهٔ متن اصلی مقاله ارائه کند. برای اینکه تصاویر موجود در یک پایگاه داده به‌طور مؤثر بازیابی شوند، نیاز است برای هر تصویر تعدادی برچسب که توصیف‌کنندهٔ آن باشد، در نظر گرفته شود. اکثر روش‌های برچسب زدن تصاویر در ادبیات بر اساس استخراج ویژگی‌های تصویر و استنتاج همبستگی بین این ویژگی‌ها و برچسب‌هایی است که توسط خبرگان انتخاب شده‌اند (Barnard and Forsyth 2001; Jeon, Lavrenko & Manmatha 2003; Wang, Blei & Fei-Fei 2009; Song el al. 2016; Putthividhy, Attias & Nagarajan 2010). تصاویر موجود در اسناد علمی کیفیت پایینی دارند. همچنین، در بسیاری از موارد در اسناد علمی، تصاویر با ویژگی‌های بصری مشابه مطالب متفاوتی را منتقل می‌کنند و باید برچسب‌های متفاوتی برای آن‌ها منظور کرد. چنانچه افزون بر خود تصویر به سندی که تصویر را شامل می‌شود هم دسترسی داشته باشیم، می‌توان از متن سند برای استخراج برچسب بهره برد. به این ترتیب که با استفاده از روش‌های پردازش متن، کلمات کلیدی مرتبط با تصویر را می‌توان از متن همراه تصویر و زیرنویس تصویر استخراج کرد (Leong, Rada & Hassan 2010; Chan, Johar & Hong 2013; Josi, Wartena & Charbonnier 2018). پایگاه اطلاعاتی «گنج» یکی از پایگاه‌های اطلاعاتی اسناد علمی ایران است. این پایگاه دستاوُرد کاربرد فناوری اطلاعات برای مدیریت اطلاعات

علـم و فنـاوری در «پژوهشـگاه علـوم و فنـاوری اطلاعـات ایـران (ایرانـداک)» اسـت. «گنـج» بـا صدهـا هـزار رکـورد، کـه شـامل اطلاعـات پایان‌نامه‌هـا و رسـاله‌ها و پیشـنهادهٔ آن‌هاسـت، بزرگ‌تریـن پایـگاه اطلاعـات علمـی و فنـی کشـور اسـت. ایـن پایـگاه هم‌اکنـون مرجـع بسـیاری از پژوهشـگران ایـران و جهـان اسـت و روزانـه بیـش از ده هـزار کاربـر و ده‌هـا هـزار جسـت‌وجو در آن انجـام می‌گیـرد. در «گنـج» امـکان جسـت‌وجو بـر اسـاس یـک عبـارت متنـی پرس‌وجـو و بازیابـی و نمایـش نتایـج جسـت‌وجو در قالـب فرادادههـای متنـی (عنـوان، چکیـده، کلیـدواژه، پدیـدآور، سـال انتشـار) وجـود دارد. لیکـن، اطلاعـات از تصاویـر موجـود در اسـناد «گنـج» بازیابی نمی‌شـود. فراهم‌کـردن چنیـن امکانـی در «گنـج» به‌عنـوان یـک ارزش افـزوده می‌توانـد آن را از موتورهـای جسـت‌وجوی مشـابه متمایـز سـازد.

اولیـن قـدم در بازیابـی تصاویـر ایجـاد یـک پایـگاه داده از تصاویـر موجـود در «گنـج» اسـت. «فخـرزاده و صدیقـی» روشـی سـاختارمحور کـه مبتنـی بـر چیدمـان و آرایـش فایـل WORD سـند اسـت، بـرای ایجـاد پایـگاه داده از تصاویـر مسـتخرج از اسـناد علمـی ارائـه داده‌انـد. نرم‌افـزار معرفی‌شـده قـادر اسـت تصاویـر و زیرنویـس آن‌هـا و قسـمتی از متـن سـند را کـه شـامل اشـاره بـه تصویـر اسـت، اسـتخراج کنـد (۱۳۹۹). قـدم بعـدی در بازیابـی تصاویـر، اسـتخراج برچسـب‌های توصیف‌کننـده بـرای تصاویـر اسـت. به‌دلیـل پیچیدگی‌هـای موجـود در محتـوای تصاویـر علمـی، در ایـن پژوهـش از قسـمتی از متـن سـند کـه شـامل اشـاره بـه تصویـر اسـت، بـرای برچسـب زدن تصاویـر اسـتفاده می‌شـود. در ایـن تحقیـق، روش‌هـای اسـتخراج خـودکار برچسـب تصاویـر، بـر اسـاس پـردازش زبـان طبیعـی بررسـی و بهتریـن روش معرفـی می‌شـود. در بسـیاری از مـوارد اطلاعـات موجـود در متـن همـراه تصویـر کافـی نیسـت. از طـرف دیگـر، بسـیاری از روش‌هـای پـردازش زبـان طبیعـی در ایـن زمینـه مبتنـی بـر اطلاعـات آمـاری کلمـات اسـت و وقتـی متـن مـورد بررسـی کوتـاه باشـد، بـر روی عملکـرد آن‌هـا اثـر می‌گـذارد. تلاش‌هـای محـدودی بـرای اسـتخراج برچسـب از متـن همـراه تصویـردر مقـالات انگلیسـی انجـام شـده اسـت، امـا از آنجـا کـه زبـان فارسـی ماهیـت و سـاختار متفاوتـی از زبـان انگلیسـی دارد، ایـن روش‌هـا نمی‌تواننـد در زبـان فارسـی بـه‌کار گرفتـه شـوند. تـا آنجـا کـه نویسـندگان ایـن مقالـه اطـلاع دارنـد، پژوهشـی بـرای اسـتخراج برچسـب تصاویـر موجـود در اسـناد علمـی فارسـی در ادبیـات وجـود نـدارد.

ادامـهٔ ایـن پژوهـش بـه ایـن ترتیـب سـازمان یافتـه اسـت کـه ابتـدا، در بخـش دوم، بـا مـروری بـر پیشـینهٔ پژوهـش، مهم‌تریـن مطالعـات صورت‌گرفتـه در حـوزهٔ اسـتخراج برچسـب از تصاویـر بررسـی خواهـد شـد. در بخـش سـوم، ضمـن معرفـی روش پژوهـش، روش‌هـای پیشـنهادی

شرح داده می‌شود. برای بررسی کارایی روش‌های پیشنهادی در بخش چهارم، روش‌های معرفی‌شده با استفاده از یک مطالعهٔ موردی در «پایگاه گنج» ارزیابی و بهترین روش معرفی می‌شود.

۲. پیشینهٔ پژوهش

نویسندگان مقالات علمی از تصاویر برای فراهم‌کردن کمک بصری به‌منظور توضیح بهتر مفاهیم و یا خلاصه‌کردن روش پیشنهادی خود بهره می‌برند. به همین دلیل، موضوع بازیابی اطلاعات از تصاویر به یکی از موضوعات جذاب برای افرادی که در حوزهٔ کتابخانهٔ دیجیتال کار می‌کنند، تبدیل شده و در سال‌های اخیر توجه زیادی را به خود جلب کرده است (Barnard & Forsyth 2001; Jeon, Lavrenko & Manmatha 2003; Wang, Blei & Fei-Fei 2009; Murthy & Manmatha 2015; Johnson, Ballan & Fei-Fei 2015; Yang, Zhang & Xie 2015; Wu et al. 2015; Sohmen et al. 2018; Jobin, Mondal & Jawahar 2019; Morris, Müller-Budack & Ewerth 2020;). اکثر روش‌های موجود در این زمینه به‌دنبال استخراج برچسب‌های مناسب با توجه به محتوای خودِ تصویر هستند (Barnard & Forsyth 2001; Jeon, Lavrenko & Manmatha 2003; Wang, Blei & Fei-Fei 2009). در این تحقیقات سعی بر آن است که با استفاده از ویژگی‌های خودِ تصویر الگویی را پیدا کنند که ویژگی‌های بصری را به برچسب‌های از پیش تعیین‌شده توسط خبرگان مربوط کند. این الگو می‌تواند برای برچسب‌گذاری تصاویر جدید استفاده شود. با به‌وجود آمدن داده‌های آزمایشی عظیم روش‌های مبتنی بر یادگیری عمیق، پیشرفت‌های زیادی در این زمینه مشاهده شده است (Murthy & Manmatha 2015; Mayhew, Chen & Ni 2016; Johnson, Ballan & Fei-Fei 2015; Yang, Zhang & Xie 2015; Wu et. al. 2015;). اکثر این روش‌ها در مورد تصاویری که شامل یک شیء برجسته[1] هستند، عملکرد خوبی دارند، اما در مورد تصاویر موجود در مقالات علمی که در آن تصاویرِ با ساختار مشابه مفاهیم کاملاً متفاوتی را انتقال می‌دهند، عملکرد ضعیفی نشان می‌دهند. نمودارهای خطی و یا نمودارهای پراکندگی در مقالات اگرچه ویژگی‌های بصری یکسان دارند، محتوای علمی متفاوتی را در هر سند ارائه می‌دهند و باید برچسب‌های متفاوتی داشته باشند. با استفاده از اطلاعات موجود در تصاویر می‌توان

1. salient object

گروه اصلی تصاویر (مثل تصاویر طبیعی، نمودار دایره‌ای، نمودار خطی، جدول‌ها، نمودار پراکندگی و غیره) را تعیین کرد. تحقیقاتی انجام شده است که آن‌ها را فقط با تمرکز بر نمودارها و با استفاده از روش‌های پردازش تصویر و یادگیری عمیق گروه‌بندی کرده‌اند (Savva et al. 2011; Tang et al. 2016; Siegel et al. 2016; Zhou and Tan 2000; Prasad et al. 2007). «سیگل» و همکاران، با در نظر گرفتن دو گروه (نمودار یا غیرنمودار) و با استفاده از شبکهٔ عصبی پیچشی[1]، نمودارها را جداسازی کرده‌اند. آن‌ها سپس، با استفاده از روش‌های قانونمند[2] و پردازش متن و تصویر، اطلاعات نمودارها را استخراج کرده و به‌عنوان برچسب در نظر گرفته‌اند (Siegel et al. 2016). «جوبین، موندال و جواهر» برای تصاویر در مقالات علمی ۲۸ گروه معرفی کرده و با استفاده از روش‌های یادگیری عمیق با دقت خوبی توانسته‌اند تصاویر را تقسیم‌بندی کنند (Jobin, Mondal & Jawahar 2019). تا جایی که نویسندگان این تحقیق اطلاع دارند، روش معتبری برای برچسب‌گذاری تصاویر علمی فقط با استفاده از محتوای خودِ تصویر (فراتر از گروه تصویر)، وجود ندارد.

می‌توان از متن سند مربوط به تصویر برای برچسب زدن تصویر بهره برد. «لانگ، رادا و حسن» با استفاده از زیرنویس و توضیح متنی تصاویر و به‌کارگیری تکنیک‌های پردازش زبان طبیعی برچسب تصاویر را استخراج کرده‌اند. آن‌ها از سه نوع مجموعهٔ داده استفاده کرده و سه روش متفاوت برای رتبه‌بندی کلمات در زیرنویس و متن مربوط به تصویر را در نظر گرفته‌اند. در یکی از این روش‌ها از برچسب‌های موجود در انبارهٔ «فلیکر» استفاده کرده‌اند. رتبهٔ هر کلمه در متن مربوط به تصویر با توجه به تکرار آن کلمه در انبارهٔ «فلیکر» تعیین می‌شود. در روش بعدی از داده‌های موجود در «ویکی‌پدیا» و روش گراف استفاده شده است. هر کلمه بر اساس تحلیل معنایی صریح[3] تبدیل به یک بردار می‌شود و به‌عنوان رأس گراف در نظر گرفته می‌شود، سپس، به بقیهٔ کلمات در زیرنویس با یک یال جهت‌دار وصل می‌شود. وزن هر یال با توجه به شباهت استنباطی هر کلمه با کلمهٔ بعدی تعیین می‌شود. با استفاده از وزن‌های محاسبه‌شده و الگوریتم تکرار گراف (Mihalcea and Tarau 2004) رتبهٔ هر کلمه تعیین می‌شود. در روش سوم از روش مدل کردن (McCallum and Li 2006) موضوع متن استفاده کرده‌اند. بدین‌ترتیب، رتبهٔ هر کلمه در زیرنویس با توجه به احتمال تعلق آن‌ها به مدل‌های موضوعیِ پیداشده تعیین

1. convolution nural network
2. rule based
3. explicit semantic analysis

می‌شود. سرانجام، رتبهٔ کلمات با استفاده از این سه روش را به‌عنوان ویژگی کلمات در نظر می‌گیرند و با روش دستگاه بردار پشتیبان، رتبهٔ نهایی هر کلمه تعیین می‌شود (Leong, Rada & Hassan 2010). «ماسون و چارنیاک»، با استفاده از داده‌های Leong, Rada & Hassan (2010) و داده‌های Feng and Lapata (2008) که شامل تصاویر و زیرنویس آن‌هاست، نشان دادند که روش‌های آماری ساده مثل ضریب فراوانی[1] در سند و معکوس فراوانی سند[2]، فراوانی در پیکرهٔ نتایج قابل قبولی در مقایسه با روش‌های پیچده‌تر به‌دست می‌دهند (Mason & Charniak 2012).

تلاش‌های محدودی در استفاده از متن سند در برچسب زدن تصاویر در متون علمی انجام گرفته است. «جوزی، وارتنا و چاربونیر» ابتدا، با استفاده از پردازش‌های ابتدایی زبان‌شناختی عبارات اسمی را از زیرنویس و متنِ اشاره‌شده به تصویر استخراج می‌کنند. سپس، عباراتی را که معکوس فراوانی سند بالاتری دارند، در نظر می‌گیرند (Josi, Wartena & Charbonnier 2018). در مرحلهٔ بعد با استفاده از روش تعبیهٔ کلمه[3]، معادل برداری عبارتِ استخراج‌شده محاسبه می‌شود. عباراتی که معادل برداری آن‌ها شباهت بیشتری به عبارات اسمی زیرنویس داشته باشند، به‌عنوان برچسب انتخاب می‌شوند. انتخاب قسمت مناسبی از متن سند برای برچسب زدن تصویر، نقش مهمی در دقت سیستم دارد. به‌طور معمول، اطلاعات توصیف‌کنندهٔ تصویر، در نزدیکی تصویر قرار دارد (Chan, Johar & Hong 2013).

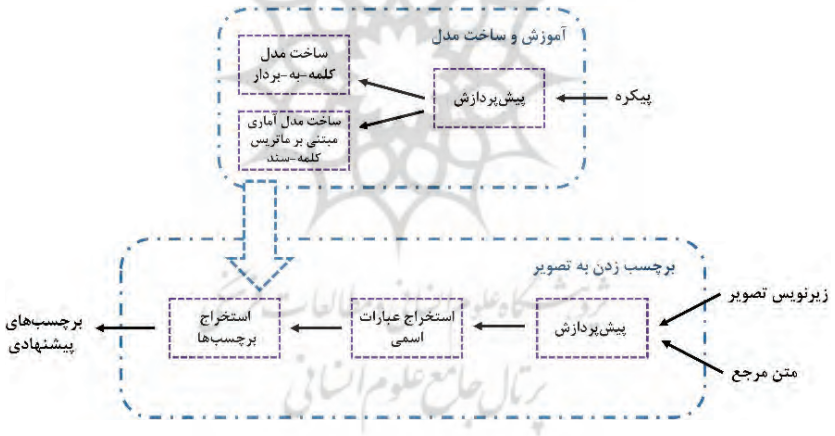
با توجه به پیچیدگی‌های پردازش تصاویر در متون علمی، در این تحقیق از متن سند همراه تصویر برای برچسب زدن تصاویر استفاده می‌شود. روش‌های معرفی‌شده بر اساس پردازش زبان طبیعی را به‌دلیل تفاوت ساختاری زبان فارسی با انگلیسی نمی‌توان در زبان فارسی به‌کار برد. زبان فارسی یکی از زبان‌های هندواروپایی است که توسط بیش از صد میلیون نفر از مردم جهان صحبت شده و به‌عنوان زبان رسمی سه کشور ایران، افغانستان (فارسی دری) و تاجیکستان (فارسی تاجیکی) است. پردازش زبان فارسی به‌دلیل پیچیدگی‌های زبانی، کمبود منابع و مطالعات انجام‌شده در این زبان از دیدگاه محاسباتی کمتر مورد توجه پژوهشگران قرار گرفته است. کاراکترهای یکسان با یونی‌کدهای گوناگون (مانند انواع میم)، وجود کاراکترهای بدون تأثیر در محتوا (مانند کاراکترهای کشسانی)، عدم یکسان‌نویسی فاصله و نیم‌فاصله و علائم نگارشی، کلمات چنددیکته‌ای،

1. term frequency
2. inverse document frequency
3. word embedding

تفاوت ساختار زبان فارسی با لاتین، کلمات جمع مکسر، و ریشه‌یابی پیچیدهٔ کلمات با توجه به ساختار زبان فارسی نمونه‌هایی از چالش‌های پردازش زبان فارسی است که برای متون علمی دوچندان می‌شود. تا آنجا که نویسندگان این مقاله اطلاع دارند، تحقیقی در زمینهٔ برچسب زدن تصاویر موجود در متون علمی فارسی صورت نگرفته است.

۳. روش پژوهش

در این پژوهش با استفاده از روش مطالعات اسنادی، سیستمی برای استخراج برچسب تصاویر موجود در اسناد علمی طراحی شده است. این پژوهش از نظر ماهیت در ردیف پژوهش‌های توسعه‌ای-کاربردی قرار می‌گیرد. شکل ۱، واحدهای مختلف سیستم طراحی‌شده را نشان می‌دهد. این سیستم با تجزیه و تحلیل متنِ همراه تصویر با استفاده از روش‌های پردازش زبان طبیعی، برچسب‌های توصیف‌کنندهٔ تصویر را استخراج می‌کند. عملکرد سیستم با داده‌های آزمایشی بررسی می‌شود.



شکل ۱. روش پیشنهادی برای استخراج برچسب تصویر

۳-۱. پیش‌پردازش

در مرحلهٔ اول، نرمال‌سازی یا یکسان‌سازی انجام می‌شود. هدف از نرمال‌سازی یکسان کردن کلمات موجود در متون است. از آنجا که اسناد مختلف از کدگذاری‌های مختلف استفاده می‌کنند، کلماتی که از نظر نوشتاری تفاوت دارند اما در واقع یک کلمه هستند، با نرمال‌سازی یکسان می‌شوند. در مرحلهٔ بعد، کلمات ریشه‌یابی می‌شوند. در

اینجـا از یـک ریشـه‌یابی سـاده اسـتفاده شـده اسـت:

- ◇ حذف شناسه‌های فعل فارسی با در نظر گرفتن زمان فعل
- ◇ حذف پیشوندهای فعل فارسی؛ مانند: «می»، «ب» و ...
- ◇ حـذف پسـوندهای مالکیـت؛ماننـد: «م»، «ت»، «ش»، «یـم»، «یـت»، «یـش»، «یتـان»، «یشـان»، «یمـان»، «مـان»، «تـان»، «شـان» و «ان»
- ◇ حذف نشانه‌های جمع فارسی و عربی؛ مانند: «ها»، «ات»، «های»، «ان»، «هایی» و «ین».
- ◇ حذف «ی» میانجی

پـس از آن بـا اسـتفاده از واحدسـاز، متـن بـه واحدهـای تشـکیل‌دهندهٔ آن یعنـی کلمـات تبدیـل می‌شـود. مراحـل پیش‌پـردازش روی تمـام متـون موجـود در پیکـره اعمـال می‌شـود. در جـدول ۱، نمونـه‌ای از متـن همـراه یـک تصویـر و پردازش‌شـدهٔ آن مشـاهده می‌شـود.

۳-۲. استخراج عبارات اسمی

پـس از پیرایـش اولیـه، تمـام عبـارات اسـمی یک کلمـه‌ای[1]، دوکلمـه‌ای[2] و سـه کلمه‌ای[3] متـن تمـام اسـناد موجـود در پیکـره اسـتخراج می‌شـود. در متـن اسـناد، کلمـات رایجـی وجـود دارنـد کـه بـار معنایـی کمـی دارنـد. ایـن کلمـات ایسـت‌واژه نامیـده می‌شـوند. ایسـت‌واژه‌های فارسـی فهرسـتی از افعـال فارسـی، مصـدر، حـروف ربـط، حـروف اضافـه و انـواع قیدهـای زبـان فارسـی هسـتند. افـزون بـر آن، کلمـات رایجـی کـه ممکـن اسـت در توضیحـات تصاویـر ذکـر شـوند، بـه ایـن فهرسـت اضافـه شـده‌اند. واژه‌هایـی چـون «شـکل»، «نقشـه»، «نمـودار»، «تصویـر» و یـا لغاتـی کـه در حـوزهٔ مـورد بررسـی رایـج تلقـی می‌شـوند، همچـون کلمـات «مگاپاسـکال» یـا «نرم‌افـزار متلـب» بـه ایـن فهرسـت افـزوده شـده‌اند. عباراتـی کـه شـامل ایسـت‌واژه‌ها هسـتند، حـذف می‌شـوند.

1. unigram 2. bigram 3. trigram

جدول ۱. نمونه‌ای از یک متن ورودی و متن پیش‌پردازش‌شده

متن ورودی	پیش‌پردازش
شکل 14- نرخ خطای بیت برای رشته با طول 10 شکل 12- سیگنال اصلی در رشته شبه‌تصادفی حال با استفاده از مدولاسیون PSK باینری سیگنال را مدوله کرده وبه یک کانال با نویز سفید گوسی جمع شونده ارسال می کنیم سیگنال به نویز کانال را dB5 در نظر می گیریم. همان‌طور که در شکل 13 مشاهده می کنید، نمودار قرمز سیگنال رسیده به گیرنده می‌باشد که با نویز بسیار زیادی جمع شده است و سیگنال آبی سیگنال قبل از ارسال به کانال است. سپس با استفاده از دمدولاتور PSK باینری سیگنال ضرب شده در رشته شبه تصادفی را آشکار می کنیم. اگر طول سیگنال اصلی را T (در اینجا 10) در نظر بگیریم، برای آشکارسازی بعد از ضرب دوباره سیگنال شبه‌تصادفی در سیگنال دمدوله شده انتگرالی روی بازه T میگیریم.	شکل ۱۴ نرخ خطا بیت برای رشته با طول ۱۰ شکل ۱۲ سیگنال اصلی در رشته شبه‌تصادفی حال با استفاده از مدولاسیون PSK باینری سیگنال را مدوله کرده و به یک کانال با نویز سفید گوسی جمع‌شو ارسال کرد&کن سیگنال به نویز کانال را 5dB در نظر گرفت&گیر همان‌طور که در شکل ۱۳ مشاهده کرد &کن نمودار قرمز سیگنال رسیده به گیرنده بود&باش که با نویز بسیار زیادی جمع شد&شو و سیگنال آبی سیگنال قبل از ارسال به کانال است. سپس با استفاده از دمدولاتور PSK باینری سیگنال ضرب شده در رشته شبه‌تصادفی را آشکار کرد&کن اگر طول سیگنال اصلی را T دراینجا ۱۰ در نظر گرفت&گیر برای آشکارسازی بعد از ضرب دوباره سیگنال شبه‌تصادفی در سیگنال دمدوله شده انتگرالی روی بازه T گرفت&گیر

۳-۳. استخراج مدل‌های کلمات از پیکره

بعد از استخراج عبارات اسمی از متن مرجع، این عبارات رتبه‌بندی می‌شوند. عبارات با رتبهٔ بالاتر به‌عنوان برچسب تصویر در نظر گرفته می‌شود. همان‌طور که در شکل ۱، نشان داده شده، برای این منظور، ابتدا باید دو مدل فراوانی کلمه-معکوس فراوانی سند[1] و کلمه-به-بردار[2] از پیکره استخراج شود. با استفاده از مدل tf_idf عباراتی کلیدی از متن همراه تصویر استخراج می‌شود و با استفاده از مدل کلمه-به-بردار عبارات کلیدی که شباهت بیشتری به زیرنویس دارند، در نظر گرفته می‌شود. مدل tf_idf به‌صورت زیر محاسبه می‌شود:

فرمول ۱
$$tf\_idf(t,d,D) = tf(t,d).idf(t,D)$$

فرمول ۲
$$tf(t,d) = f_{t,d}$$

فرمول ۳
$$idf(t,D) = log\frac{N}{|d \in D: t \in d|}$$

t، واژهٔ مورد نظر، d، سند مربوط، $f_{t,d}$، فراوانی واژهٔ t در سند d است. N، تعداد کل اسناد

1. term frequency-inverse document frequency (tf_idf)
2. word2vec

در پیکرهٔ D است. معکوس فراوانی سند idf برای یک واژه، لگاریتم معکوس کسر اسنادی است که حاوی آن واژه است و نشان‌دهندهٔ اطلاعاتی است که آن واژه شامل می‌شود (Spärck 1972). خروجی این روش برای یک پیکره، یک ماتریس عددی است که در آن هر سند به‌صورت بُرداری از وزن‌های tf_idf متناظر با واژه‌هایش مشخص می‌شود. کلمه یا عبارتی که وزن متناظر با آن در ماتریس tf_idf بیشتر است، عبارت بااهمیت‌تری در سند مربوط محسوب می‌شود.

عبارات کلیدی می‌توانند بر اساس شباهت آن‌ها به زیرنویس هم رتبه‌بندی شوند. برای محاسبهٔ شباهت بین کلمات نیاز است که مدل عددی کلمات را با استفاده از روش‌های تعبیهٔ کلمات[1] استخراج کرد. مدل کلمه-به-بردار از روش‌های کارآمد تعبیهٔ کلمات است که با استفاده از روش یادگیری عمیق و بدون نظارت محاسبه می‌شود (Mikolov et al. 2013). در این روش کلمات پیکره به‌عنوان ورودی دریافت می‌شود و ماتریسی تولید می‌شود که تعداد سطرهای آن برابر تعداد کل کلمات پیکره است و هر سطر، بردار معادل هر کلمه را شامل می‌شود. بردار متناظر کلمات با در نظر گرفتن شباهت معنایی و مفهومی آن‌ها محاسبه می‌شود. روش کلمه-به-بردار از دو مدل معماری شبکه استفاده می‌کند. در مدل CBOW[2] با دانستن مجموعهٔ محدودی از کلمات اطراف y، کلمهٔ y پیش‌بینی می‌شود. به‌عنوان مثال، در عبارت «گربه بالای درخت»، کلمهٔ «بالای» کلمهٔ مرکزی و «گربه» و «درخت» کلمات محتوا هستند. در مدل skip-gram با دانستن کلمهٔ y، کلمات اطراف آن پیش‌بینی می‌شود. در اینجا از مدل CBOW استفاده می‌شود.

در روش CBOW ورودی الگوریتم، کلمات محتواست و خروجی y همان کلمهٔ مرکزی است که می‌خواهیم پیش‌بینی کنیم. ورودی و خروجی به‌صورت بردار one-hot هستند. دو ماتریس $\boldsymbol{v} \in \boldsymbol{R}^{\{|V|\times n\}}$ و $\boldsymbol{R}^{\{n\times|V|\}}$ را ایجاد می‌کنیم. n یک عدد دلخواه است که اندازهٔ فضای تعبیه را مشخص می‌کند و $|V|$ تعداد کلمات موجود در پیکره است. v_i ستون i از ماتریس v، بردار تعبیهٔ n بُعدی متناظر با کلمهٔ ورودی w_i است. u_j ردیف j از ماتریس خروجی U، بردار تعبیهٔ n بُعدی متناظر با کلمهٔ مرکزی (کلمهٔ هدف) w_j است. خلاصه‌ای از الگوریتم در زیر آمده است.

۱. در مجموعهٔ کلمات محتوا با طول C، برای هر کلمهٔ محتوا یک بردار one-hot تولید

1. word embeding
2. ontinues bag of words

می‌کنیم:

$$(x_{1k}, x_{2k}, x_{3k}, \dots, x_{Ck} \in R^{|v|}) \qquad \textbf{فرمول ٤}$$

۲. بردارهای تعبیه را برای کلمات محتوا به‌دست می‌آوریم:

$$(v_{1k} = vx_{1k}, v_{2k} = vx_{2k}, \dots, v_{Ck} = v\, x_{Ck} \in R) \qquad \textbf{فرمول ٥}$$

۳. میانگین بردارهای تعبیه را حساب می‌کنیم:

$$\hat{v} = \frac{v_{1k} + v_{2k} + \cdots + v_{Ck}}{C} \qquad \textbf{فرمول ٦}$$

۴. بردار امتیازها را به‌دست می‌آوریم، $z = U\hat{v} \in R^{|v|}$ از آنجا که ضرب داخلی بردارهای مشابه بالاتر است، کلمات مشابه نزدیک هم قرار می‌گیرند تا امتیاز بالاتری به‌دست بیاید.

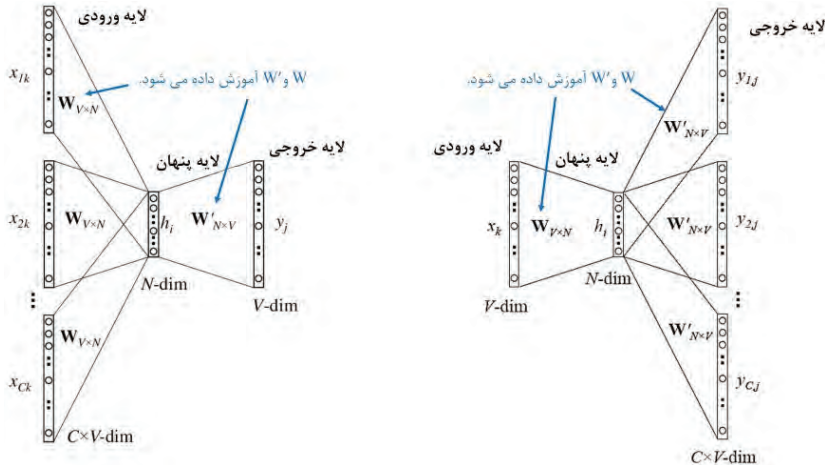
۵. با تابع softmax بردار امتیازها را به احتمال تبدیل می‌کنیم:

$$\hat{y} = softmax(z) \in R^{|v|} \qquad \textbf{فرمول ۷}$$

۶. مطلوب آن است که بردار احتمالاتی که توسط شبکهٔ عصبی تولید می‌شود $\hat{y} \in R^{|V|}$ با برادر صحیح احتمال $y \in R^{|V|}$ تطابق داشته باشد. بردار صحیح احتمال همان بردار one-hot، معادل کلمهٔ مرکزی است. تابع هدف با استفاده از آنتروپی متقاطع بین y و $\hat{y}$ به‌صورت زیر محاسبه می‌شود:

$$minimize\ J = -u_c^T \hat{v} + log \sum_{j=1}^{|V|} exp(u_j^T \hat{v}) \qquad \textbf{فرمول ۸}$$

با بهینه‌سازی تابع هدف با استفاده از دادهٔ آزمایشی، بردارهای تعبیهٔ کلمهٔ مرکزی و کلمات محتوا به‌دست می‌آید.



شکل ۲. شبکهٔ عصبی CBOW در چپ و شبکهٔ عصبی Skip-gram در راست نشان داده شده است (Rong 2014)

۳-٤. استخراج برچسب‌ها

پس از محاسبهٔ مدل‌های کلمات، با استفاده از نرم‌افزار استخراج تصاویر (فخرزاده و صدیقی ۱۳۹۹) فایل XML وُرد تجزیه شده و تصاویر، زیرنویس و قسمتی از متن سند که در آن به تصویر مورد نظر اشاره شده (متن مرجع)، استخراج می‌شود. پس از اعمال پیش‌پردازش یک‌کلمه‌ای و ترکیب‌های دوکلمه‌ای و سه‌کلمه‌ای کلمات زیرنویس و متن مرجع را تشکیل می‌دهیم. با استفاده از پیکره، مدل‌های tf، idf، tf-idf و کلمه-به-بردار را محاسبه می‌کنیم. در اینجا پنج روش مختلف برای استخراج برچسب‌ها در نظر گرفته می‌شود.

- در روش tf، مجموع یک‌کلمه‌ای‌ها، دوکلمه‌ای‌ها و سه‌کلمه‌ای‌های متن مرجع را بر اساس فراوانی آن‌ها در سند مربوط رتبه‌بندی می‌کنیم.
- در روش idf، مجموع یک‌کلمه‌ای‌ها، دوکلمه‌ای‌ها و سه‌کلمه‌ای‌های متن مرجع را بر اساس معکوس فراوانی آن‌ها در پیکره رتبه‌بندی می‌کنیم.
- در روش tf-idf مجموع یک‌کلمه‌ای‌ها، دوکلمه‌ای‌ها و سه‌کلمه‌ای‌های متن مرجع را بر اساس نسبت فراوانی آن‌ها در سند مربوط به معکوس فراوانی آن‌ها در پیکره رتبه‌بندی می‌کنیم.
- در روش شباهت کسینوسی، مجموع یک‌کلمه‌ای‌ها، دوکلمه‌ای‌ها و سه‌کلمه‌ای‌های

متن مرجع را بر اساس شباهت کسینوسی با زیرنویس رتبه‌بندی می‌کنیم. با استفاده از مدل کلمه-به-بردار که از پیکره ساخته شده، معادل برداری زیرنویس و یک‌کلمه‌ای و دوکلمه‌ای موجود در متن مرجع را به‌دست می‌آوریم. معادل برداری عبارات چندکلمه‌ای، میانگین بردارهای متناظر با کلمات آن‌هاست. شباهت کسینوسی بین بردار معادل زیرنویس V_c با بردار معادل عبارت استخراج‌شده از متن مرجع V_R، یک معیار برای سنجش شباهت با زیرنویس محسوب می‌شود. شباهت کسینوسی در واقع، نسبت ضرب داخلی دو بردار به حاصل‌ضرب اندازهٔ دو بردار است و به‌صورت زیر حساب می‌شود:

فرمول ۹

$$cosine\_similarity(V_C, V_R) = \frac{V_c . V_R}{|V_c||V_R|}$$

- در روش شباهت کسینوسی و tf-idf، مجموع یک‌کلمه‌ای‌ها، دوکلمه‌ای‌ها و سه‌کلمه‌ای‌ها را ابتدا بر اساس tf-idf رتبه‌بندی می‌کنیم. سپس، مجموعه‌ای از برچسب‌های برتر را انتخاب می‌کنیم و دوباره بر اساس شباهت کسینوسی به زیرنویس رتبه‌بندی می‌کنیم.

تعداد برچسب‌های خودکار پذیرفته‌شده در هر روش، با بهینه‌سازی میانگین معیار F_1، با استفاده از داده‌های آزمایشی تعیین می‌شود. به این ترتیب که در هر روش تعدادی از برچسب‌های خودکار که به ازای آن‌ها میانگین معیار F_1 بیشینه است، در نظر گرفته می‌شود.

٤. ارزیابی روش‌های پیشنهادی

برای بررسی کارایی روش‌های پیشنهادی در برچسب زدن تصاویر، از یک مطالعهٔ موردی در «پایگاه اطلاعات علمی ایران (گنج)» کمک گرفته شده است. این پایگاه مرجع اصلی دسترسی به تمام‌متن پایان‌نامه‌ها و رساله‌های تحصیلات تکمیلی در ایران است. سی‌صد سند فنی-مهندسی در فرمت وُرد را به‌صورت تصادفی از پایگاه دادهٔ «گنج» انتخاب کردیم. سپس، با استفاده از کد استخراج تصویر از فایل ورد، تصاویر این فایل‌ها را به همراه زیرنویس و متن مرجع استخراج کردیم. این کد قادر به استخراج تصاویر از فایل‌هایی است که فرمت استاندارد وُرد را دارند. تعدادی از فایل‌ها خالی از تصویر بودند و تعدادی از آن‌ها فرمت استاندارد نداشتند. برای استخراج متن مرجع هر تصویر، پاراگرافی که در آن به تصویر مورد نظر اشاره شده، به همراه دو پاراگراف قبل از آن و دو پاراگراف

بعد از آن در نظر گرفته شده است. در شکل ۳، نمونه‌ای از یک تصویر به همراه متن مرجع آن دیده می‌شود. از مجموعهٔ تصاویر استخراج‌شده ۱۰۰ تصویر به‌صورت تصادفی انتخاب شده است و بر اساس محتوای زیرنویس و متن مرجع برچسب موضوعی خورده‌اند. این برچسب‌ها توسط سه خبرهٔ مستقل در حوزهٔ مهندسی و فنی انتخاب شده است. این ۱۰۰ تصویر برای محاسبهٔ خطای سیستم استفاده می‌شود. انتخاب برچسب مناسب فقط با در نظر گرفتن یک برش کوتاه از متن سند در بعضی موارد بسیار چالش‌برانگیز بوده است. در شکل ۳، پاراگراف‌های قبل و بعد از متن مرجع، زیرنویس‌های جدول‌ها و تصاویر اطراف آن است و به همین دلیل، قابلیت توصیفی متن مرجع کم است.

شکل ۳. نمونه‌ای از یک تصویر همراه با زیرنویس و متن مرجع. برچسب‌هایی که توسط روش tf-idf استخراج شده نیز نمایش داده شده است (از پایان‌نامهٔ کارشناسی ارشد اسلامی ۱۳۹۱)

٤-١. معیارهای ارزیابی

برای ارزیابی بازخوانی از معیار McCarthy and Navigli (2007) استفاده کردیم. برای محاسبهٔ این معیار تصاویر را تعدادی خبرهٔ مستقل برچسب می‌زنند و هر برچسب خودکاری که استخراج می‌شود بر اساس فراوانی آن در لیست برچسب خبرگان وزن‌دهی می‌شود. اگر فرض کنیم $\boldsymbol{H}=\{\boldsymbol{h_1},\boldsymbol{h_2},\boldsymbol{h_3}\}$ مجموعهٔ خبرگان باشد و $\boldsymbol{I}=\{\boldsymbol{i_1},\boldsymbol{i_2},\boldsymbol{i_3},..\}$ مجموعهٔ تصاویری باشد که توسط خبرگان برچسب خورده‌اند، $\boldsymbol{T_j}$ مجموعهٔ تمام برچسب‌هایی است که توسط خبرگان، برای تصویر $\boldsymbol{i_j}$ در نظر گرفته شده و $\boldsymbol{a_j}$ مجموعهٔ برچسب‌هایی

است که توسط سیستم برای تصویر i_j انتخاب شده است. بازخوانی به‌صورت زیر تعریف می‌شود:

فرمول ۱۰

$$R = \frac{\sum_{a_j:i_j \in I} \frac{\sum_{word \in a_j} freq_{word}}{|T_j|}}{|I|}$$

در معادلهٔ بالا $freq_{word}$ فراوانی کلمه $word$ از مجموع برچسب‌های خودکار a_j در مجموع برچسب‌های دستی است. دقت سیستم به‌صورت زیر محاسبه می‌شود:

فرمول ۱۱

$$P = \frac{\sum_{a_j:i_j \in I} \frac{\sum_{word \in a_j} \{1 \; if \; word \in T_j\}}{|a_j|}}{|I|}$$

با استفاده از معیار بازخوانی و دقت معیار F_1 به‌صورت زیر محاسبه می‌شود:

فرمول ۱۲

$$P = \frac{\sum_{a_j:i_j \in I} \frac{\sum_{word \in a_j} \{1 \; if \; word \in T_j\}}{|a_j|}}{|I|}$$

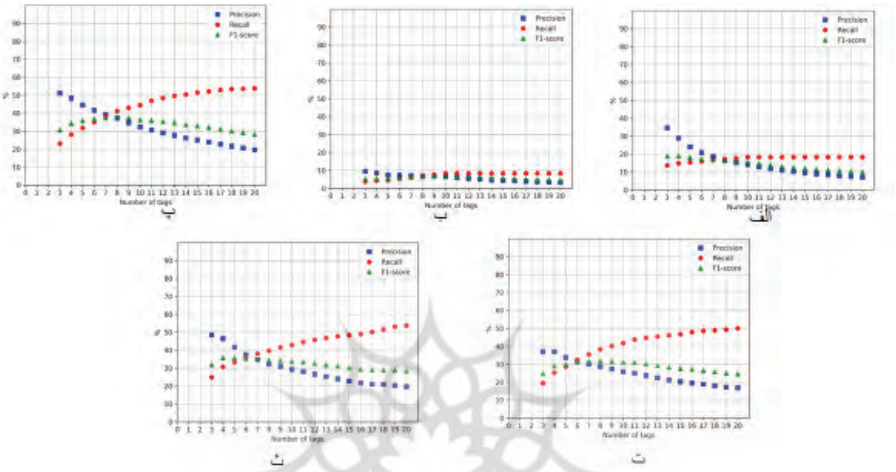
٤-۲. پیاده‌سازی و بررسی نتایج

به‌منظور پیاده‌سازی از زبان برنامه‌نویسی «پایتون» نسخهٔ ۳/۶ و کتابخانه‌های «هضم»[1]، NLTK، «جنسیم»[2] (Rehurek & Sojka 2010) و «پارسیوار»[3] (Salar et al. 2018) استفاده شده است. با استفاده از مدل کلمه-به-بردار CBOW معادل برداری کلمات پیکره استخراج شد. طول بردار برابر با ۳۰۰ و پارامترهای اندازهٔ پنجره و کمینهٔ رخداد کلمات در سند هر دو برابر با ۵ در نظر گرفته شد. یک‌کلمه‌ای ودوکلمه‌ای و سه‌کلمه‌ای‌های موجود درپیکره استخراج شده و مدل tf و tf_idf محاسبه شد. بر اساس پنج روش پیشنهادشده، یک‌کلمه‌ای ودوکلمه‌ای و سه‌کلمه‌ای‌های موجود در متن مرجع رتبه‌بندی شدند. در مشاهدات اولیه با در نظر گرفتن یک‌کلمه‌ای‌ها، عبارات نامرتبط زیادی در خروجی مشاهده شد و تأثیر منفی در کارکرد الگوریتم داشت. همچنین، در مجموع برچسب‌هایی که توسط خبرگان انتخاب شده بود، تعداد خیلی محدودی برچسب یک‌کلمه‌ای وجود داشت. به همین دلیل، برای این مسئله تنها دوکلمه‌ای‌ها و سه‌کلمه‌ای‌ها در نظر گرفته شد.

در هر روش ۲۰ برچسب برتر که رتبهٔ بالاتری دارند، انتخاب شدند. بر اساس مشاهدات انجام‌شده بعد از ۲۰ برچسب، موارد مرتبط به‌ندرت دیده شده است. از میان این بیست

1. Hazm 2. Gensim 3. Parsivar

برچسب، تعدادی از برچسب‌های برتر که میانگین معیار F_1 به ازای آن‌ها بیشینه شود، انتخاب شدند. نمودار میانگین دقت و بازخوانی و معیار F_1 برای پنج روش معرفی‌شده، و برای بیست برچسب اول در شکل ۴، نمایش داده شده است. بهترین دقت و بازخوانی و معیار F_1، مربوط به پنج روش در جدول ۲، نشان داده شده است.



شکل ٤. مقادیر دقت (آبی)، بازخوانی (قرمز) و معیار F_1 برای پنج روش معرفی‌شده. الف) tf، ب) idf، پ) tf-idf، ت) شباهت کسینوسی، و ث) tf-idf و شباهت کسینوسی

همان‌طور که در جدول ۲، می‌بینیم روش tf-idf بیشترین معیار F_1 را دارد. هرچه تعداد برچسب‌های انتخابی توسط سیستم را افزایش دهیم، بازخوانی سیستم بهتر می‌شود، اما به‌دلیل ظاهر شدن مثبت‌های اشتباه، دقت سیستم و به همان نسبت معیار F_1 کاهش پیدا می‌کند. انتظار اولیهٔ ما این بود که کلمات موجود در زیرنویس بهترین توصیف‌کننده‌های تصویر باشند. عباراتی که tf-idf بالاتری دارند، در واقع کلمات کلیدی متن هستند. بر این اساس، همراستا با روش Josi, Wartena & Charbonnier (2018) می‌توان ابتدا تعداد زیادی عبارات را که tf-idf بالاتری دارند، انتخاب کرد و در نهایت، تعداد محدودی از آن‌ها را که شباهت کسینوسی بیشتری به زیرنویس دارند، به‌عنوان برچسب تصویر در نظر گرفت. اما نتایج آزمایشات نشان داد که در نظر گرفتن شباهت کسینوسی با زیرنویس، اگرچه نتایج قابل مقایسه‌ای با tf-idf دارد، باعث بهبود خروجی آن نشده است. دلیل اصلی این امر آن است که در خیلی از موارد زیرنویس‌های انتخاب‌شده توسط نویسندهٔ سند دقیق نیستند.

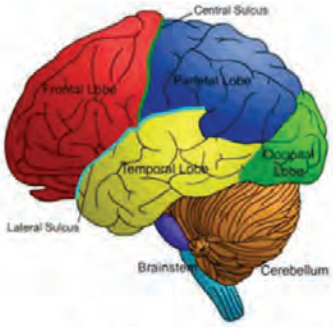
جدول ۲. مقایسهٔ پنج روش پیشنهادی

روش	صحت درصد	بازخوانی درصد	F_1 درصد
tf	۱۵	۱۷	۱۵
idf	۶	۸	۷
tf-idf	۳۷	۴۱	۳۸
Cosine similarity	۲۹	۳۸	۳۱
tf-idf +Cosine similarity	۴۲	۳۳	۳۶

شکل ۵، یک نمونه از تصاویر با زیرنویس همراه با بیست برچسب انتخابی توسط سیستم با دو روش مختلف را نشان می‌دهد. اولین پاراگراف عباراتی را که tf-idf بالاتری دارد، نشان می‌دهد. این عبارات بار دیگر بر اساس شباهت کسینوسی رتبه‌بندی شده و در پاراگراف دوم نمایش داده شده است. کلماتی که درشت‌نمایی شده‌اند، عباراتی هستند که توسط خبرگان نیز به‌عنوان برچسب در نظر گرفته شده‌اند. اعداد داخل پرانتز نشان‌دهندهٔ تعداد خبرگانی هستند که برچسب مورد نظر را انتخاب کرده‌اند. کلمهٔ «نمای جانبی» به اشتباه در زیرنویس دوبار تکرار شده است و این باعث شده است که در شباهت کسینوسی با زیرنویس، عباراتی که شامل {نما و جانبی} هستند، رتبهٔ بالاتری از عباراتی را به‌دست آورند که شامل کلمهٔ «مغز» هستند؛ در حالی که در این تصویر، از نظر سه خبره، «قشر مغز ولوب» عبارات توصیف‌کنندهٔ تصویر است و «نما» و «جانبی» کلمات کم‌اهمیتی بوده و توصیف‌کنندهٔ تصویر نیستند. در شکل ۶، خروجی سیستم با سه روش tf-idf، شباهت کسینوسی و ترکیب tf-idf و شباهت کسینوسی نمایش داده شده است. در روش شباهت کسینوسی، عباراتی که شامل «جانب» است، رتبه بالایی به‌دست آورده‌اند، در حالی که «جانب» نسبت به کلمات دیگر در زیرنویس ارزش اطلاعاتی کمتری دارد. با ترکیب شباهت کسینوسی با tf-idf نیز بهبودی در خروجی tf-idf ایجاد نمی‌شود.

گاهی، زیرنویس یک توضیح کلی از تصویر را ارائه می‌دهد و توضیح دقیق‌تر تصویر در متن مرجع می‌آید. برای مثال، در شکل ۳، زیرنویس «نتایج حاصل از ردیابی مسیر تحت نویز سفید» است و در توضیحات تکمیلی در متن مرجع همراه تصویر، چنین به نظر می‌آید که در این تصویر نتایج کنترلر برای بازوی مکانیکی ماهر در مسیر تحت

نویز سفید بررسی شده است. بنابراین، هر سه خبره بازوی مکانیکی ماهر و کنترلر[1] PID را به‌عنوان برچسب در نظر گرفته‌اند که با خروجی سیستم بر اساس روش tf-idf تطابق دارد.



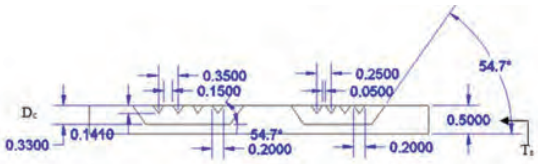
شکل ۴-۱ نمای جانبی از نمای جانبی لوبهای قشر مغز

قشر مغز(۳) ،فعالیت مغزی، **لوب پیشانی(۲)** ،**لوب آهیانه‌ای(۲)**، **لوب گیجگاهی(۱)** ،نما جانبی، **نما جانبی لوبها(۲)** ،جانبی لوبهای قشر، **لوبهای قشر مغز(۳)**، جانبی لوبها، لوبهای قشر، نواحی قشرمخ، تحقیقات BCI، عادی بدن ایجاد، لوب پیشانی لوب، قسمت خاص نسبت،ناحیه اصلی شامل،حالت عادی بدن، شامل لوب پیشانی،ا صلی شامل لوب

جانبی لوبها، **قشر مغز(۳)**، نما جانبی،**نما جانبی لوبها (۲)**،جانبی لوبهای قشر، لوبهای قشر، **لوبهای قشر مغز(۳)** ،**لوب پیشانی(۲)**، **لوب گیجگاهی(۱)**،فعالیت مغزی، **لوب آهیانه‌ای(۱)**، عادی بدن ایجاد،لوب پیشانی لوب،نواحی قشرمخ ،قسمت خاص نسبت،ناحیه اصلی شامل،حالت عادی بدن،شامل لوب پیشانی،اصلی شامل لوب،تحقیقات BCI

شکل ۵. برچسب‌های استخراج‌شده توسط سیستم. پاراگراف اول: روش tf-idf، پاراگراف دوم: روش شباهت کسینوسی (از پایان‌نامهٔ کارشناسی ارشد حاتمی ۱۳۹۲)

1. proportionaliIntegral derivative (PID)



(ب)

شکل 5-3 ابعاد میز نوری سیلیکونی الف) دید از بالای میز نوری ب) دید از جانب میز نوری (ابعاد در مقیاس میلی‌متر می‌باشند)

شکل 5-2 : شماتیک میز نوری سیلیکونی به همراه المان‌های نصب شده بر روی آن

ابعاد VCSEL و دیود نوری معرفی شده در فصل دوم به طور خلاصه در جدول 5-1 مشخص شده‌اند. ابعاد حفره‌ها و شیارهای V شکل متناسب با ابعاد VCSEL، فوتودایود و فیبر نوری مشخص شده‌اند. به عنوان مثال طول حفره‌ی مربوط به آرایه‌ی VCSEL باید از 965 μm (طول VCSEL) و طول حفره‌ی گیرنده باید از 1350 μm (طول چهار دیود نوری با فاصله‌ی 50 μm از یکدیگر) بزرگتر باشد. برای راحت حرکت دادن و چسباندن المان‌ها درون حفره‌ها، طول و عرض حفره‌ها حدود 100 μm از هر طرف بزرگتر در نظر گرفته شد.

شکل 5-3 ابعاد میز نوری سیلیکونی طراحی شده بر اساس ابعاد المان‌های مشخص شده در جدول 5-1 را نشان می‌دهد (تمام ابعاد برحسب میلی‌متر مشخص شده‌اند). شکل 5-3 دید از بالای میز نوری را نشان می‌دهد که در این شکل ابعاد حفره‌ها و شیارها، مشخص شده است. شکل 5-3 دید از کنار میز نوری را نشان می‌دهد که در این شکل زاویه‌ی دیواره‌ای حفره، عرض شیارها و عمق حفره‌ها و شیارها مشخص شده است. براساس این شکل فاصله‌ی بین شیارهای V‌شکل در طرف فرستنده برابر با 250 μm (فاصله‌ی بین دو VCSEL در آرایه‌ی VCSEL) و در طرف گیرنده برابر با 350 μm می‌باشد. ابعاد میز نوری طراحی شده در جدول 5-2 مشخص شده است. به دلیل زاویه‌ی واگرایی کم VCSEL (حدود 14درجه) و همچنین به دلیل اینکه فضای انتشار بین VCSEL و فیبر هرا می‌باشد همانند آنچه در شکل 4-5 نشان داده شده است، سطح زیرین فیبر مانند یک لنز عمل کرده که باعث تمرکز پرتوها بالاتر از محور نوری فیبر می‌شود. این تاثیر هنگام استفاده از پوشش رزین بین VCSEL و فیبر نوری با ضریب شکست بیشتر از هرا کاهش می‌یابد. شکل 5-4 مشخصات فیبر و شیار v شکل را مشخص می‌کند. براساس این شکل مقدار بالاآزدگی فیبر از شیار (q) برابر با 29.5μm می‌باشد و فاصله‌ی بین لیزر و سطح بالایی فیبر برابر با 34.5 μm می‌باشد. پارامترهای شکل در جدول 5-2 مشخص شده‌اند.

برچسبها:
روش tf-idf: میز نوری(3)،نوری سیلیکونی(1)،میز نوری سیلیکونی(3)،ابعاد میز نوری(3)،میز نوری ابعاد،فیبر نوری،دیود نوری(1)،آرایه VCSEL(1)
روش شباهت کسینوسی: میز نوری(3)،نوری ابعاد،)،نوری سیلیکونی(1)،جانب میز نوری، میز نوری سیلیکونی(3)، ابعاد دیود نوری،ابعاد المان‌های
ترکیب tf-idf وشباهت کسینوسی: میز نوری(3)،نوری سیلیکونی(1)،میز نوری سیلیکونی(3)،ابعاد میز نوری(3)،میز نوری ابعاد،فیبر نوری،دیود
نوری(1)،شماتیک میز نوری

شکل ٦. مقایسهٔ برچسب‌های استخراجی بر اساس روش‌های tf-idf، شباهت کسینوسی، و روش ترکیب tf-idf و شباهت کسینوسی (از پایان‌نامهٔ افیونی اکبری ۱۳۹۱)

۵. نتیجه‌گیری

تصاویر، حاوی اطلاعات مهمی از اسناد علمی هستند. به همین دلیل، امروزه تحقیقات در زمینهٔ بازیابی تصاویر از اسناد علمی از تحقیقاتِ روبه‌رشد است. برای بازیابی تصاویر ابتدا باید یک پایگاه داده از تصاویر و برچسب‌های توصیف‌کنندهٔ آن‌ها ایجاد شود. به‌دلیل پیچیدگی‌های محتوایی تصاویر علمی، در این تحقیق از زیرنویس و قسمتی از متن سند که شامل اشاره به تصویر است، برای برچسب زدن تصاویر استفاده شد. ابتدا، پیش پردازش‌های رایج متن (نرمال‌سازی، ریشه‌یابی) بر روی متنِ همراهِ تصویر اعمال شد. سپس، عبارات اسمی که شامل همهٔ ترکیبات یک‌کلمه‌ای، دوکلمه‌ای، و سه‌کلمه‌ای متن مرجع است، استخراج شد. در نهایت، عبارات اسمی استخراج‌شده با استفاده از پنج روشِ فراوانی کلمه، معکوس فراوانی سند، فراوانی کلمه-معکوس فراوانی سند، شباهت کسینوسی عبارات با زیرنویس و ترکیب روش فراوانی کلمه-معکوس فراوانی سند و شباهت کسینوسی رتبه‌بندی شدند. نتایج دادهٔ آزمایشی نشان داد که روش فراوانی کلمه-معکوس

فراوانی سند عملکرد بهتری نسبت به سایر روش‌ها دارد.

یکی از مشکلات روش پیشنهادی و البته، تمام روش‌های مشابه که بر اساس مدل‌سازی آماری طراحی شده، این است که در خروجی نهایی، عبارات مثبت اشتباهی مشاهده می‌شود که در خیلی موارد بار معنایی کمی دارند. در صورت دسترسی به پیکرهٔ عظیم داده در حوزهٔ تخصصی مورد نظر، می‌توانیم مدل‌های کلماتی قابل اعتماد بهتری بسازیم. مدل کلماتیِ بهتر خطای سیستم را کاهش می‌دهد. همچنین، دسترسی به یک بانک دادهٔ عظیم از واژه‌های کلیدی موجود در حوزهٔ مربوط می‌تواند در فیلتر کردن عبارات بی‌معنا کمک‌کننده باشد.

در روش ارائه‌شده، با بررسی متنی که به تصویر اشاره می‌کند، برچسب‌های تصویر استخراج شده و محتوای خودِ تصویر در نظر گرفته نشده است. از جملهٔ پیشنهادها برای تحقیقات آتی این است که محتوای خودِ تصویر هم بررسی شود. از آنجا که تنوع تصاویر در اسناد علمی زیاد است و ویژگی‌های یک گروه با گروه دیگر از تصاویر خیلی متفاوت است. بنابراین، در قدم اول باید تصاویر را گروه‌بندی کرد و سپس، با استفاده از روش‌های پردازش تصویر اطلاعات مناسب در خودِ تصاویر را برای برچسب زدن در نظر گرفت.

۶. قدردانی

این مقاله مستخرج از طرح پژوهشی است که با حمایت‌های مادی و معنوی «پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)» به انجام رسیده است. همچنین، نویسندگان از «مرکز فناوری اطلاعات ایرانداک» برای فراهم‌کردن دادهٔ آزمایشی برای این پژوهش و از آزمایشگاه متن‌کاوی و یادگیری ماشین «ایراندک» قدردانی می‌کنند.

فهرست منابع

اسلامی، علی. ۱۳۹۱. طراحی کنترلر شبکه ویولتی برای بازوی مکانیکی متحرک. پایان‌نامهٔ کارشناسی ارشد. دانشگاه تبریز.

افیونی اکبری، شیرین. ۱۳۹۱. طراحی و پیاده‌سازی میز نوری سیلیکونی جهت کوپلینگ VCSEL و دیود نوری PIN به فیبر نوری چند مود. پایان‌نامه کارشناسی ارشد. دانشگاه اصفهان.

حاتمی، رویا. ۱۳۹۲. طبقه‌بندی سیگنال‌های EEG ناشی از تصور حرکتی به کمک تلفیق فیلترهای فضایی فرکانسی. پایان‌نامه کارشناسی ارشد. دانشگاه صنعتی شاهرود.

فخـرزاده، آزاده، و امیرحسـین صدیقـی. ۱۳۹۹. *ارائـه روشـی سـاختارمحور بـرای ایجـاد پایـگاه داده از تصاویـر مستخرج از اسناد علمـی؛ مـورد مطالعـه: پایـگاه اطلاعـات علمـی ایـران (گنـج)*. پژوهشـنامهٔ پـردازش و مدیریـت اطلاعـات ۳۵ (۳): ۷۲۹-۷۵۴.

References

Barnard, K & D. Forsyth. 2001. Learning the semantics of words and pictures. *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*. (pp. 408-415). *IEEE*. doi: 10.1109/ICCV.2001.937654. Vancouver, BC, Canada.

Bratasanu, D., I. Nedelcu, M. Datcu.. 2011. The semantic gap for satellite image annotation and automatic mapping applications. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 4 (1)204–193 :.

Chan, Ch. Sh., A. Johar, & Jer Lang Hong. 2013. Contextual information for image retrieval systems. 2013. 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Shenyang, China. pp. 863-867 2013, doi: 10.1109/FSKD.2013.6816315.

Feng, Y. and M. Lapata. 2008. Automatic image annotation using auxiliary text information. Proceedings of ACL-08: HLT, pages 272–280.

Jeon J., V. Lavrenko, and R. Manmatha. 2003. Automatic image annotation and retrieval using cross-media levance models. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada.

Jobin, K. V., A. Mondal, and C. V. Jawahar. 2019. Docfigure: A dataset for scientific document figure classification. 13th IAPR International Workshop on Graphics Recognition. Sydney, NSW, Australia.

Johnson, L. Ballan, & L. Fei-Fei. 2015. Love thy neighbors: Image annotation by exploiting image metadata.2015. IEEE International Conference on Computer Vision. Santiago, Chile (pp. 4624–4632). IEEE.

Josi F., C. Wartena, and J. Charbonnier. 2018. Text-Based Annotation of Scientific Images Using Wikimedia Categories. DEXA 2018. *Communications in Computer and Information Science* 903: 243-253.

Leong, C. W, M. Rada, and S. Hassan. 2010. Text Mining for Automatic Image Tagging. Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China.

Li.Wei, and A. McCallum. 2006. Pachinko allocation: DAG-structured mixture models of topic correlations, Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA.

Mason, R., and E. Charniak. 2012. Apples to Oranges: Evaluating Image Annotations from Natural Language Processing Systems. Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies, june, 2012. Canada (pp172-181). Association for Computational Linguistics.

Mayhew, M. B., B. Chen, & K. S. Ni. 2016. Assessing semantic information in convolutional neural network representations of images via image annotation. IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA (pp. 2266–2270). IEEE.

McCarthy, D., and R. Navigli. 2007. Evaluations (SemEval-2007 Task 10: English Lexical Substitution Task). Proceedings of the Fourth International Workshop on Semantic. Prague,Czech Republic; USA (pp.48-53) Association for Computational Linguistics.

Mihalcea R., and P. Tarau. 2004. Textrank: Bringing order into texts. Proceedings of Empirical Methods in Natural Language Processing. Barcelona, Spain.

Mikolov, T., K. Chen, G. Corrado, & J. Dean, 2013. Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

Morris. D, E. Müller-Budack, & R. Ewerth. 2020. SlideImages: A Dataset for Educational Image Classification. Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. *Lecture Notes in Computer Science*, vol12036. Cham. Portugal: Springer. https://doi.org/10.1007/978-3-030-45442-5_36

Murthy, V. N., S. Maji, & R. Manmatha. 2015. Automatic image annotation using deep learning representations. ACM on International Conference on Multimedia Retrieval. 2015, New York, NY, USA (pp. 603 – 606).

Prasad, V. S. N., B. Siddiquie, J. Golbeck, and L. S. Davis. 2007. Classifying computer generated charts. 2007 International Workshop on Content-Based Multimedia Indexing (CBMI). Talence, France.

Proceedings of the Eleventh International Conference on Language Resources and Evaluation ({LREC} Miyazaki, Japan.

Putthividhy D., H. T. Attias, and S. S. Nagarajan. 2010. Topic regression multi-modal latent dirichlet allocation for image annotation. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA.

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

Rong, X. 2014. word2vec Parameter Learning Explained (cite arxiv:1411.2738)

Salar, B., A. Roshanfekr, U. Zafarian, & H. Asghari. 2018. A Language Processing Toolkit for Persian.

Proceedings of the Eleventh International Conference on Language Resources and Evaluation ({LREC} Miyazaki, Japan.

Savva, M., N. Kong, A. Chhajta, L. Fei-Fei, M. Agrawala, and J. Heer. 2011. Revision: Automated classification, analysis and redesign of chart images. In Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST '11). New York, NY, USA (pp: 393–402). Association for Computing Machinery. DOI: https://doi.org/10.1145/2047196.2047247

Siegel, N., Z. Horvitz, R. Levin, S. Divvala, and A. Farhadi. 2016. Figureseer: parsing result-figures in research papers. *ECCV, 2016*. Lecture Notes in Computer Science, vol 9911. Cham. Amsterdam, Netherlands: Springer. https://doi.org/10.1007/978-3-319-46478-7_41

Sohmen, L., J. Charbonnier, I. Blümel, C. Wartena, and L. Heller. 2018. Figures in scientific open access publications. Digital Libraries for Open Knowledge, 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings. (E. Méndez, F. Crestani, C. Ribeiro, G. David, and J. C. Lopes, eds.), vol. 11057 of Lecture Notes in Computer Science, pp. 220–226, Springer.

Song, L., M. Luo, J. Liu, L. Zhang, B. Qian, M. H. Li, & Q. Zheng. 2016. Sparsemulti-modaltopical coding for image annotation, *Neurocomputing* 214 (C) 162–174. https://doi. org/10.1016/j. neucom.2016.06.005.

SpärckJones, K. 1972. A statistical interpretation of term specificity and its Application in retrieval. *Journal of Documentation*. Vol. 28 (1).11-21 :

Tang, B., X. Liu, J. L., M. Song, D. Tao, S. Sun, and F. Dong. 2016. Deepchart: Combining deep convolutional networks and deep belief networks in chart classification. *Signal Processing* 124: 156-161.

Wang, C., David Blei, and Li Fei-Fei. 2009. Simultaneous image classification and annotation. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA.

Wu, J., Y. Yu, C. Huang, & K. Yu. 2015. Deep multiple instance learning for image classification and auto-annotation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA (pp. 3460–3469) IEEE.

Yang, Y., W. Zhang, & Y. Xie. 2015. Image automatic annotation via multi-view deep representation, *Journal of Visual Communication Image Representation* 33 (8): 368–377.

Zhou, Y. P., and C. L. Tan. 2000. Hough technique for bar charts detection and recognition in document images. Proceedings 2000 International Conference on Image Processing (Cat. No.00CH 37101) (ICIP), Vancouver, BC, Canada.

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.

محدثه رهنما

متولد ۱۳۷۴ و دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی در دانشگاه الزهراست.

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

جلال‌الدین نصیری

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

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