رویکردی با ناظر در استخراج واژگان کلیدی اسناد فارسی با استفاده از زنجیره‌های لغوی

عطیه شریفی* و محمّدامین مهدوی

گروه مهندسی کامپیوتر، دانشگاه بین المللی امام خمینی، قزوین، ایران

چکیده

واژگان کلیدی، واژگان اصلی و کانونی یک متن و مضمون اصلی مطلب هستند. تهیهٔ این واژگان به روش سنّتی نیازمند صرف زمان و هم‌چنین دانش تخصّصی راجع به موضوع متن است. از آن‌جا که واژگان کلیدی کاربردهای فراوانی در به‌کارگیری مستندات الکترونیکی دارند، شناسایی روش‌های خودکار و بهبودیافته برای استخراج این دسته از واژگان همیشه مورد توجه بوده است. رویکرد پژوهش حاضر یک روش باناظر برای استخراج واژگان کلیدی است که در آن با استفاده از زنجیره‌های لغوی واژگان متن، ویژگی‌های جدیدی برای هر واژه استخراج شده است. در ایجاد زنجیره‌های لغوی سعی بر شکل‌گیری روابط بین معنای واژگان بوده‌ایم، ازاین‌رو در مدل ارائه‌شده «فارس‌نت» نقش مهمی در ایجاد آنها ایفا می‌کند. داده‌های مورد ارزیابی در این پژوهش مقالات علمی پژوهشی نشریات فارسی هستند. نتایج به‌دست‌آمده نشان می‌دهد که استفاده از روابط معنایی بین واژگان در کنار ویژگی‌های آماری، عملکرد مناسبی را در استخراج واژگان کلیدی از مقالات نتیجه می‌دهد.

واژگان کلیدی: استخراج واژگان کلیدی، اسناد فارسی، یادگیری باناظر، زنجیرهٔ لغوی، فارس‌نت

Supervised approach for keyword extraction from Persian documents using lexical chains

Atieh Sharifi* & M.Amin Mahdavi

Department of Computer Engineering, Imam Khomeini International University, Qazvin, Iran

Abstract

Keywords are the main focal points of interest within a text, which intends to represent the principal concepts outlined in the document. Determining the keywords using traditional methods is a time consuming process and requires specialized knowledge of the subject. For the purposes of indexing the vast expanse of electronic documents, it is important to automate the keyword extraction task. Since keywords structure is coherent, we focus on the relation between words. Most of previous methods in Persian are based on statistical relation between words and didn't consider the sense relations. However, by existing ambiguity in the meaning, using these statistic methods couldn't help in determining relations between words. Our method for extracting keywords is a supervised method which by using lexical chain of words, new features are extracted for each word. Using these features beside of statistic features could be more effective in a supervised system. We have tried to map the relations amongst word senses by using lexical chains. Therefore, in the proposed model, "FarsNet" plays a key role in constructing the lexical chains. Lexical chain is created by using Galley and McKeown's algorithm that of course, some changes have been made to the algorithm. We used java version of hazm library to determine candidate words in the text. These words were identified by using POS tagging and Noun phrase chunking. Ten features are considered for each candidate word. Four features related to frequency and position of word in the text and the rest related to lexical chain of the word. After extracting the keywords by the classifier, post-processing performs for determining Two-word key phrases that were not obtained in the previous step. The dataset used in this research was chosen from among Persian scientific

* Corresponding author

* نویسندهٔ عهده‌دار مکاتبات • تاریخ ارسال مقاله: ۱۳۹۶/۹/۱۲ • تاریخ آخرین بازنگری: ۱۳۹۷/۱/۱۰ • تاریخ پذیرش: ۱۳۹۷/۲/۲۶

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]

[DOI: 10.29252/jsdp.15.4.95]

papers. We only used the title and abstract of these papers. The results depicted that using semantic relations, besides statistical features, would improve the overall performance of keyword extraction for papers. Also, the Naive Bayes classifier gives the best result among the investigated classifiers, of course, eliminating some of the features of the lexical chain improved its performance.

Keywords: Keyword Extraction, Persian Document, Supervised Learning, Lexical Chain, FarsNet

۱- مقدّمه

در بسیاری از حوزه‌های پردازش زبان طبیعی از واژگان کلیدی برای نمایش اسناد استفاده می‌شود. به مجموعه‌ای از واژگان یا عبارات کوتاه که بیان‌کنندهٔ منظور اصلی یک متن باشند، واژگان کلیدی متن گفته می‌شود [1].

یک عبارت کلیدی از چند کلمه تشکیل شده است که این واژگان در یک ساختار گرامری کنار یکدیگر قرار گرفته‌اند و معنای مشخّصی دارند. عبارات کلیدی در مقایسه با واژگان کلیدی، بیشتر بازگوکنندهٔ محتوای متن هستند. برای مثال، عبارت کلیدی «شبکهٔ عصبی» نمی‌تواند به‌صورت جداگانه معنای مورد نظر را برساند. واژگان و عبارات کلیدی علاوه‌بر آن که باعث کشف مفهوم و ایدهٔ اصلی یک متن می‌شوند، در حوزه‌های مختلف دیگری مانند بازیابی اطلاعات، طبقه‌بندی متون، خلاصه‌سازی، ترجمهٔ متن و سامانه‌های هوشمند پاسخ به پرسش کاربرد دارند [2].

زنجیره‌های لغوی وابستگی معنایی بین واژگان متن را نشان می‌دهند. به بیان ساده، زنجیره‌های لغوی فهرستی از واژگان متن هستند که با یکدیگر ارتباط معنایی دارند [3]. ارتباط بین واژگان می‌تواند از طریق شبکه‌های معنایی مانند وردنت[1] و روابط تعریف‌شده در آن به‌دست آید [4]. به‌دلیل این‌که در شبکه‌های معنایی روابط بین معانی شکل می‌گیرد نه خود واژگان، رفع ابهام معنایی واژگان یکی از نیازمندی‌های ایجاد زنجیره‌های لغوی با استفاده از این شبکه‌ها است. الگوریتم‌های مختلفی برای رفع ابهام معنایی واژگان با استفاده از وردنت وجود دارد [4, 5].

در این پژوهش، از نسخه دوم فارس‌نت[2] [6] به‌عنوان هستان‌شناسی[3] لغوی برای ایجاد زنجیرهٔ لغوی استفاده شده است. فقط آن دسته از عبارات نامزدی که در فارس‌نت وجود دارند، برای ایجاد زنجیرهٔ لغوی استفاده و بقیه حذف می‌شوند. زنجیرهٔ لغوی ایجادشده یک پیاده‌سازی از روش *گالی* و *مک‌کیوان* است [7] که البته تغییراتی در آن داده شده است. این تغییرات مربوط به نوع و فاصلهٔ روابط بین معانی و شیوهٔ تعیین رتبه برای هر معنی است که به آنها اشاره خواهیم کرد.

رویکرد پژوهش حاضر استفاده از روشی باناظر برای استخراج واژگان کلیدی است. ایدهٔ اصلی در این روش‌ها دسته‌بندی دودویی واژگان متن به دو دستهٔ واژگان کلیدی و غیر کلیدی است [8, 9]. برای این منظور، ویژگی‌هایی برای هر کلمه در نظر می‌گیریم که بین واژگان کلیدی و غیر کلیدی متمایزکننده باشند. برخی از این ویژگی‌ها آماری و تعدادی از آنها با استفاده از زنجیرهٔ لغوی ساخته‌شده به‌دست می‌آیند.

محدودهٔ بررسی این مسئله بر روی مقالات علمی‌ـ پژوهشی نشریات فارسی است. از آن‌جایی‌که پردازش تمام متن مقاله کاری زمان‌بر است و مهم‌تر این‌که با پردازش کل متن، احتمال تعیین واژگان نامرتبط زیادی به‌عنوان واژهٔ کلیدی وجود دارد، در این پژوهش فقط از عنوان و چکیده مقالات استفاده کردیم. البته، گفتنی است که در چکیدهٔ مقالات، بسیاری از واژگان کلیدی مقاله وجود دارد. در بخش دوم این نوشتار، پیشینهٔ پژوهش بیان و در بخش سوم به معرّفی روش پیشنهادی پرداخته می‌شود. بخش چهارم به نتایج پژوهش اختصاص دارد. درنهایت، بخش پنجم دربرگیرندهٔ جمع‌بندی و پیشنهادهاست.

۲- پیشینهٔ پژوهش

روش‌های استخراج واژگان کلیدی به چهار دستهٔ اصلیِ آماری، زبان‌شناسی، یادگیری ماشین، و روش‌های ترکیبی تقسیم می‌شوند [10].

روش‌های آماری برای به‌کارگیری، بسیار ساده هستند. این روش‌ها نیازمندی‌های اندکی دارند. از همه مهم‌تر، نیازی به داده‌های آموزشی ندارند [10]. روش‌های آماری بر ویژگی‌های غیرزبانی متن؛ مانند فراوانی واژگان و مکان قرارگیری کلمه در سند تمرکز می‌کنند. روش‌های «بسامد کلمه-معکوس بسامد سند[4]» و «ماتریس هم‌رخدادی» جزء روش‌های آماری هستند [3]. با استفاده از معیار بسامد واژه-معکوس بسامد سند، واژگانی از یک سند که بسامد بالایی در متن دارند و هم‌چنین در اسناد دیگر (پیکرهٔ[5] متن) کمتر

[1] WordNet
[2] http://dadegan.ir/catalog/farsnet
[3] Ontology

[4] Term frequency-inverse document frequency (TF-IDF)
[5] corpus

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]

استفاده شده‌اند، امتیاز بالاتری می‌گیرند؛ درنتیجه، این معیار میزان خاص‌بودن کلمه در پیکرهٔ متن و کلیدی‌بودن آن را در متن نشان می‌دهد. در این روش، به روابط بین واژگان توجهی نمی‌شود. جدا از بااهمیت‌بودن هر یک از واژگان کلیدی یک متن، باید بین مجموعهٔ واژگان کلیدی پیوستگی هم وجود داشته باشد. در روش ماتریس هم‌رخدادی، واژگانی که نرخ هم‌رخدادی آنها با سایر واژگان متن بیشتر باشد، میزان وابستگی آنها با واژگان متن بیشتر است و درنتیجه واژهٔ کلیدی هستند. دو واژهٔ هم‌رخداد هستند، اگر هر دو در یک محدوده[1] باشند. در این روش نیازی به پیکرهٔ متن نداریم. برای محاسبهٔ نرخ هم‌رخدادی هر واژه از آزمون مربع کای پیرسون[2] استفاده می‌شود [3].

روش‌های زبان‌شناسی از ویژگی‌های زبانی واژگان، جمله‌ها و اسناد مانند ساختار لغوی، نحوی و معنایی استفاده می‌کنند. این ویژگی‌ها برای افزایش کیفیت و گاهی برای حذف واژگان نامناسب استفاده می‌شوند. استفاده از اصطلاح‌نامه‌ها به‌منظور تحلیل معنایی واژگان، برچسب‌گذاری نحوی[3] و ریخت‌شناسی[4] جزء رویکردهای زبان‌شناسی محسوب می‌شود [2]. *وانگ* و همکارانش از وردنت برای استخراج واژگان کلیدی استفاده و با استفاده از برچسب‌گذاری نحوی، اسامی متن را استخراج کردند و سپس معانی موجود برای این اسامی در وردنت ، به‌عنوان رئوس گراف در نظر گرفته شده است. یال‌های گراف با استفاده از روابط بین معانی در وردنت ایجاد شد و با توجه به اهمیت ارتباط برای هر یال وزنی تعیین کردند. با استفاده از معیار رتبهٔ صفحه[5] برای هر رأس وزنی تعیین شد و سپس رفع ابهام معنایی برای هر اسم صورت گرفت. بدین صورت که معنای با وزن بیشتر به‌عنوان معنای درست انتخاب شد و باقی معانی حذف و با تعیین مجدد وزن هر رأس با استفاده از معیار رتبهٔ صفحه، رئوسِ با وزن بیشتر به‌عنوان واژگان کلیدی تعیین شدند [1].

روش‌های یادگیری ماشین به دو دستهٔ روش‌های باناظر و بی‌ناظر تقسیم می‌شوند. سامانه‌های بی‌ناظر اغلب با استفاده ازخوشه‌بندی یا تبدیل متن به یک گراف، کار تحلیل بین واژگان را انجام می‌دهند. در روش‌های مبتنی بر گراف، متن به‌صورت یک گراف در نظر گرفته می‌شود که واژگان متن رئوس آن و ارتباط بین واژگان، یال‌های گراف را تشکیل می‌دهد. [11]. در مقاله [12] از یک روش گروهی برای استخراج واژگان کلیدی متن استفاده شده است. سه روش بی‌ناظر در استخراج واژگان کلیدی در نظر گرفته شده است که هر سه روش مبتنی بر گراف هستند و از هم‌رخدادی برای تعیین روابط بین واژگان استفاده می‌کنند. واژگان نامزد متن از تجمیع واژگان نامزد استخراج‌شده توسط هر یک از این سه روش به‌دست می‌آیند؛ سپس، برخی از این واژگان بر اساس یک الگو حذف و رتبهٔ واژگان باقیمانده بر اساس معیاری مشخص دوباره حساب می‌شود. درنهایت، واژگانی که رتبهٔ آنها از میانگین رتبهٔ کل واژگان بیشتر باشد به‌عنوان واژگان کلیدی انتخاب می‌شوند. نتایج نشان می‌دهد تجمیع این سه روش عملکرد بهتری نسبت به هر کدام از روش‌ها دارد.

در روش‌های باناظر بر خلاف روش‌های بی ناظر نیاز به یک مجموعه‌دادهٔ برچسب‌گذاری‌شده داریم. که در اینجا برچسب داده‌ها کلیدی یا غیرکلیدی‌بودن واژگان است. یا به‌عبارتی دیگر، نیاز به مجموعه‌ای از متون با واژگان کلیدی مشخص داریم. سامانه با تحلیل این مجموعه از داده‌ها یاد می‌گیرد که چگونه واژگان کلیدی را از غیرکلیدی تشخیص دهد [13]. با توجه به این‌که رویکرد پژوهش حاضر نیز روشی باناظر است، در ادامه به برخی از کارهای انجام‌شده با استفاده از این رویکرد می‌پردازیم.

GenEx و KEA دو سامانهٔ باناظر هستند که در سال ۱۹۹۹ توسط *ترنی* و *ویتن* مطرح شدند. ویژگی‌های مهم واژگان در این دو روش برای دسته‌بندی به دو گروه کلیدی و غیرکلیدی، بسامد و مکان قرار گیری آنها در سند است. ترنی از الگوریتم ژنتیک و ویتن از الگوریتم بیز برای یادگیری استفاده کرده است. این دو روش مبنای مهمی برای توسعه سایر روش‌ها هستند [10]. در سال ۲۰۰۳، *هولث* با استفاده از دانش زبان‌شناسی، علاوه بر خصوصیات آماری، خصوصیات نحوی را نیز ارائه داده است. نتایج هولث نشان می‌دهد که قطعه‌بندی عبارات اسمی[6] دقت بالاتری نسبت به چند گرام‌ها[7] دارد و اضافه‌کردن برچسب نحوی به خصوصیات یک عبارت نتیجه‌ی بهتری را حاصل می‌کند [14]. *هاکوهن* و همکارانش در سال ۲۰۰۵ از شیوهٔ یادگیری دیگری برای استخراج واژگان کلیدی مقالات علمی انگلیسی استفاده کردند. آن‌ها نشان دادند که J45 (نوع توسعه‌یافتهٔ C4.5) بهترین نتیجه را به‌دست می‌آورد. ویژگی‌های استخراج‌شده از واژگان، مربوط به بسامد و بیشتر مکان قرارگیری آنها در متن می‌شد؛ به‌خصوص واژگانی که در بخش‌های چکیده، مقدمه و نتیجه‌گیری قرار داشتند، امتیاز بالاتری می‌گرفتند [15]. در

[1] window
[2] Pearson's chi-squared test
[3] Part Of Speech Tagging (POS)
[4] Morphology
[5] Page Rank
[6] Noun phrase chunk (NP chunk)
[7] N-grams

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]

سال ۲۰۰۶ یک روش جدید به نام KEA++ ارائه شد. در این روش اطلاعات معنایی با استفاده از اصطلاح‌نامه مربوط به آن موضوع، به هر یک از عبارت‌ها اضافه می‌شود [16]. *ارکان* [5] روشی با استفاده از زنجیره‌های لغوی ارائه داده که در آن از ویژگی‌های استخراج‌شده از زنجیره‌های لغوی به‌منظور تعیین واژگان کلیدی استفاده شده است. برای مثال، تعداد اعضای زنجیره‌ای که کلمه در آن قرار دارد به‌عنوان یک ویژگی برای کلمه در نظر گرفته شده است [5]. در مقاله [17] استخراج واژگان کلیدی با استفاده از میدان تصادفی شرطی[1] یا به‌اختصار CRF مطرح شده است. CRF یک روش برچسب‌گذاری جملات است که می‌تواند از خصوصیات سند به‌طور مؤثرتر استفاده کند و استخراج واژگان کلیدی را به‌عنوان عمل برچسب‌گذاری رشته‌ای در نظر می‌گیرد. نتایج نشان می‌دهد این مدل عملکرد بهتری نسبت به ماشین‌های یادگیری از قبیل ماشین بردار پشتیبان و رگرسیون خطی چندگانه بر روی متون چینی دارد. هم‌چنین پژوهش‌هایی برای اضافه‌کردن ویژگی‌هایی جدید با استفاده از وابستگی‌های نحوی بین واژگان [18] و شبکه‌های استنادی [19] نیز صورت گرفته است. در مقالهٔ [20] از گراف حاصل از هم‌رخدادی بین واژگان متن ویژگی‌هایی مانند درجهٔ رأس و مجموعهٔ درجات رئوس مجاور برای هر کلمه استخراج شده و با استفاده از دسته‌بند بیز ساده، نتیجهٔ مطلوبی به‌دست آمده است؛ در این مقاله، از چکیدهٔ مقالات علمی برای ارزیابی الگوریتم استفاده شده است.

روش‌های دیگر استخراج واژگان کلیدی، ترکیبی از شیوه‌های یادشده در بالا و یا استفاده از برخی اطلاعات اکتشافی مانند مکان کلمه، طول، چیدمان، برچسب‌های Html اطراف کلمه و یا ساختار سند Html است [10]. در[21] از ویکیپدیا برای استخراج واژگان کلیدی متن استفاده شده است. به این صورت که معیار «بسامد کلمه- معکوس بسامد سند» برای هر واژهٔ متن در هر صفحه‌ای از ویکیپدیا که واژه در آن وجود دارد، محاسبه می‌شود. این مجموعه اعداد به‌عنوان بردارِ متنِ[2] هر واژه در نظر گرفته شده است؛ سپس، با استفاده از شباهت کسینوسی بین این بردارهای متن، واژگان خوشه‌بندی و واژگان نمونهٔ هر خوشه به‌عنوان واژهٔ کلیدی در نظر گرفته می‌شود.

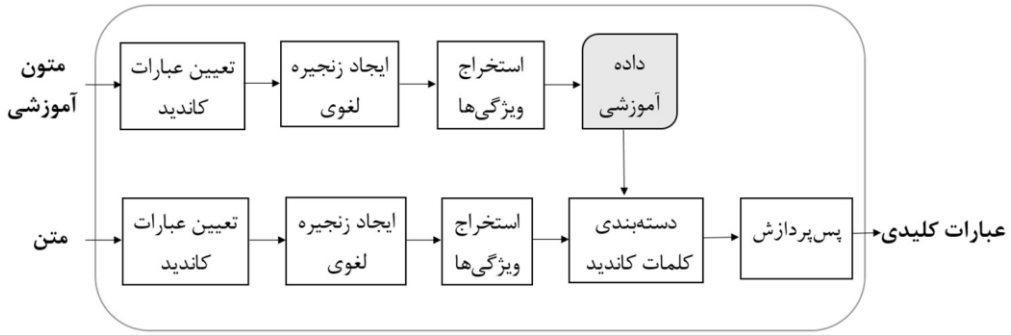
کارهای انجام‌شده در زمینهٔ استخراج واژگان کلیدی اسناد فارسی بیش‌تر مبتنی بر روش‌های آماری بوده و بیش‌تر مقالات راجع به خلاصه‌سازی متون فارسی هستند که البته پژوهش‌های انجام‌شده در آنها می‌تواند در استخراج واژگان کلیدی مفید باشند. در زمینهٔ استخراج واژگان کلیدی فارسی، *سمیهٔ عربی* و همکاران از ترکیبی از روش‌های الهام‌گرفته از وردنت و الگوریتم پورتر[3] تطبیق‌یافته با زبان فارسی و روش لان[4] برای استخراج واژگان کلیدی متن استفاده کرده‌اند [22]. سامانهٔ ارائه‌شده توسط *محمدی* و *آنالویی* با ترکیب روش فازی و رخداد هم‌زمان، واژگان بامعناتری را پیشنهاد می‌دهد. هم‌چنین در این شیوه واژگان کلیدی دوواژه‌ای نیز استخراج می‌شوند [23]. *احمدی* و *حسینی‌خواه* از شبکهٔ رقابتی LVQ[5] و شبکهٔ عصبی MLP[6] برای استخراج واژگان کلیدی استفاده کرده‌اند. نتایج، حاکی از آن است که شبکهٔ عصبی MLP نتیجه بهتری نسبت به LVQ دارد [24]. *راد* و همکارانش با استفاده از یک اصلاح‌نامه، برای هر واژه روابط هم‌ارزی، همبسته و سلسله‌مراتبی را مشخص کرده‌اند؛ سپس با استفاده از یک روش وزن‌دهی برای هر واژهٔ وزنی تعیین شده است و واژگان با وزن بالاتر به‌عنوان واژگان کلیدی در نظر گرفته شده‌اند [25].

درحالی‌که پژوهش‌های مشابه به‌منظور استخراج واژگان کلیدی مانند استفاده از شبکه‌های معنایی و روش‌های یادگیری ماشین بر روی زبان‌های اروپایی صورت پذیرفته است، نیاز است در زبان فارسی نیز این کار صورت پذیرد. ازآنجایی‌که واژگان کلیدی یک متن ساختاری منسجم دارند در رویکردهای استخراج واژگان کلیدی به روابط بین واژگان نیز توجه می‌شود. بیش‌تر روش‌های انجام‌شده در فارسی مبتنی بر روابط آماری بین واژگان بوده‌اند و به معنای واژگان توجه‌ای نداشته‌اند. با وجود ابهامات معنایی واژگان، رویکردهای آماری نمی‌توانند خیلی مشخص‌کنندهٔ روابط بین آنها باشند. در این پژوهش، بر آن شدیم که روشی بر اساس تحلیل معنایی واژگان ارائه دهیم تا بتوانیم واژگان را بر اساس معنایشان در متن و نه فقط شکل ظاهری‌شان، بررسی کنیم. این رویکرد با استفاده از پایگاه دادهٔ لغوی «فارس‌نت» انجام گرفته است. استفاده از فارس‌نت هیچ محدودیتی برای سامانه ایجاد نمی‌کند و تنها زمان پردازش را نسبت به زمانی‌که از روش‌های آماری استفاده می‌کنیم، بالاتر می‌برد؛ اما درمقابل، کارایی سامانه را بهبود می‌بخشد.

[1] CRF; Conditional Random Fields
[2] Context Vector

[3] Porter
[4] Luhn
[5] LVQ ; learning Vector Quantization
[6] MLP; Multi-Layer Perceptron

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]



(شکل-۱): ساختار روش پیشنهادی
(Figure-1): Structure of proposed method

۳- ساختار روش پیشنهادی

همان‌طورکه پیش‌تر بیان کردیم، ساختار کلی روش پیشنهادی رویکردی باناظر است که برخی ویژگی‌های واژگان در آن با استفاده از زنجیرهٔ لغوی ساخته‌شده از متن استخراج می‌شوند. ساختار کلی روش پیشنهادی در شکل (۱) نشان داده شده است.

متون آموزشی موجود در شکل، مجموعه‌ای از مقالات علمی پژوهشی فارسی هستند. همان‌طورکه در مقدمه بیان کردیم، فقط از عنوان، چکیده و واژگان کلیدی هر مقاله استفاده شده است.

متن ورودی در شکل، مقاله‌ای با واژگان کلیدی نامشخص است که برای استخراج واژگان کلیدی باید عنوان و چکیدهٔ این مقاله به سامانه داده شود. به‌دلیل این‌که متون آموزشی ما مجموعه‌ای از مقالات علوم رایانه است، برای عملکرد بهتر سامانه، از مقالات همین حوزه به‌عنوان ورودی استفاده می‌کنیم.

با توجه به این که روش پیشنهادی یک روش باناظر است، مراحل تعیین واژگان نامزد، ایجاد زنجیرهٔ لغوی و استخراج ویژگی‌ها هم برای متن ورودی و هم برای متون آموزشی انجام می‌پذیرد. با پردازش متون آموزشی، دادهٔ آموزشی ایجاد می‌شود. پس از آن دسته‌بند با استفاده از ویژگی‌های استخراج‌شده از واژگان متن و دادهٔ آموزشی ایجادشده، واژگان کلیدی متن را استخراج می‌کند. با انجام یک پس‌پردازش بر روی واژگان استخراج‌شده، عبارات کلیدی متن به‌دست می‌آیند. عبارات کلیدی استخراج‌شده در این سامانه می‌توانند شامل یک و یا دو کلمه باشند. این سامانه به زبان جاوا پیاده‌سازی شده است. در بخش‌های بعد به شرح هر یک از مراحل روش پیشنهادی می‌پردازیم.

۳-۱- تعیین واژگان نامزد

قبل از انجام پردازش‌های لازم جهت استخراج واژگان کلیدی، باید یک پیش‌پردازش بر روی متن ورودی انجام شود. در این مرحله، متن ورودی تبدیل به مجموعه‌ای از واژگان می‌شود. از طرفی، در عنوان و چکیدهٔ مقالات، واژگان زیادی وجود دارد که پردازش همه آنها دقّت سامانه را به‌شدّت کاهش می‌دهد؛ درنتیجه، زیرمجموعه‌ای از مجموعه واژگان متن که مشخصات خاصی دارند، جهت پردازش تعیین می‌شوند که به آنها واژگان نامزد می‌گوییم. تعیین واژگان نامزد در پنج مرحله انجام می‌شود که شامل: هنجارسازی متن، تقطیع واژگان، برچسب‌گذاری نحوی، تعیین عبارات اسمی و ریشه‌یابی هستند. به‌جز بخش تعیین عبارات اسمی در باقی موارد از نسخهٔ جاوای کتابخانهٔ هضم[1] برای پیاده‌سازی استفاده شده است. در ادامه، به شرح هر یک از این مراحل می‌پردازیم.

۱) **هنجارسازی متن:** گاهی نویسه‌های به‌کاررفته در دو واژهٔ یکسان، با هم متفاوت هستند؛ این باعث می‌شود که هنگام شمردن واژه‌ها، دو واژه یکسان با املاهای متفاوت، به‌عنوان دو واژهٔ مختلف در نظر گرفته شوند. برای مثال «زیست شناسی» و «زیست‌شناسی» باید به یک شکل تبدیل شوند

۲) **تقطیع واژگان:** متن ورودی به مجموعه‌ای از واژگان تبدیل می‌شود.

۳) **برچسب گذاری نحوی:** برچسب نحوی هر یک از واژگان متن تعیین می‌شود.

۴) **تشخیص گروه‌های اسمی:** با برچسب‌گذاری نحوی، الگوهای خاصی از عبارات متن استخراج می‌شود. از آن‌جایی که واژگان کلیدی بیشتر شامل اسامی و صفت‌ها هستند، الگوهای استخراج‌شده شامل عبارات تک‌واژه‌ای از نوع اسم یا صفت و هم‌چنین عبارات دوواژه‌ای هستند که در عبارات

[1] https://github.com/sobhe/hazm

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]

دوواژه‌ای، واژه نخست از نوع اسم و واژه دوم می‌تواند اسم و یا صفت باشد. البته عبارات تک‌واژه‌ای از نوع صفت در صورتی در نظر گرفته می‌شوند که قبل از آنها واژه‌ای از نوع اسم قرار گرفته باشد. به‌عنوان مثال اگر قبل از یک صفت واژه‌ای از نوع صفت قرار گرفته باشد، آن را در نظر نمی‌گیریم. این الگوها به‌صورت زیر هستند:

- اسم (مفرد/جمع) اسم (مفرد/جمع)
- اسم (مفرد/جمع) صفت
- اسم (مفرد/جمع)
- صفت -> درصورتی‌که قبل از آن اسم باشد.

۵) **ریشه‌یابی واژگان:** عباراتی که در مرحلهٔ قبل، از الگوهای نحوی تعیین‌شده، به‌دست آمدند، ریشه‌یابی می‌شوند. برای مثال "زنجیره" و "زنجیره‌ها" دو واژه با حالت‌های مختلف نوشتاری هستند؛ درصورتی‌که ریشه‌یابی انجام نگیرد این دو واژه به‌عنوان دو واژهٔ جدا مورد پردازش قرار می‌گیرند. درنهایت این عبارات ریشه‌یابی‌شده به‌عنوان عبارات نامزد متن در نظر گرفته می‌شوند.

۲-۳- ایجاد زنجیرهٔ لغوی

هستان‌شناسی لغوی استفاده‌شده برای ایجاد زنجیرهٔ لغوی، نسخهٔ دوم فارس‌نت است. فقط آن دسته از عبارات نامزدی که در فارس‌نت وجود دارند، برای ایجاد زنجیرهٔ لغوی استفاده و بقیه حذف می‌شوند. به همین دلیل، در مرحلهٔ قبل تنها عبارات یک یا دوواژه‌ای را به‌عنوان عبارات نامزد در نظر گرفتیم؛ چون که عباراتی با طول سه واژه یا بیش‌تر، اغلب در فارس‌نت وجود ندارند. زنجیرهٔ لغوی ایجادشده یک پیاده‌سازی از روش گالی و مک‌کیوان است [7] که البته تغییراتی در آن داده شده است. این تغییرات مربوط به نوع و فاصله روابط بین معانی و شیوهٔ تعیین رتبه برای هر معنا است که به آنها اشاره خواهیم کرد.

در این الگوریتم مشابه روش گالی و مک‌کیوان عملیات رفع ابهام قبل از ایجاد زنجیره‌های لغوی و به‌صورت مجزا انجام می‌گیرد. این الگوریتم در سه مرحله انجام می‌شود. در مرحلهٔ نخست، تمام تفسیرهای ممکن از واژگان نامزد، مشابه روش *سیلبر* و *مک‌کوی* [26] با استفاده از فرازنجیره‌ها[1] ایجاد می‌شوند. در مرحلهٔ دوم، رفع ابهام معنایی واژگان نامزد صورت می‌گیرد و درست‌ترین معنی برای هر کلمه نامزد مشخص می‌شود. درنهایت، در مرحلهٔ سوم، زنجیره‌های لغوی ایجاد می‌شوند. مرتبه زمانی این الگوریتم خطی است و برحسب تعداد واژگان نامزد یکتا است.

[1] metachain

۱-۲-۳- مرحلهٔ نخست

مشابه روش سیلبر و مک‌کوی هر فرازنجیرهٔ ایجادشده یک معنای برتر و یا عنوان دارد. برترین معنای فرازنجیره، یک معنا در فارس‌نت است که با تمام واژگان موجود در فرازنجیره، ارتباط معنایی داشته باشد [26]. دو معنا با یکدیگر مرتبط‌اند، اگر یکسان باشند یا بین آنها روابط ابرمفهوم[2]/ زیرمفهوم[3]، هم‌پایه[4] و مرتبط-است-با[5] برقرار باشد. اضافه‌شدن ارتباط معنایی مرتبط-است-با یکی از تفاوت‌های الگوریتم حاضر با الگوریتم گالی و مک‌کیوان است. برای ذخیرهٔ فرازنجیره‌ها از یک جدول درهم‌سازی استفاده شده است که در آن کلیدها شناسهٔ معانی در فارس‌نت هستند و مقدار هر کلید، فهرستی از واژگان آن فرازنجیره است. در نسخهٔ دوم فارس‌نت ۲۰۴۳۲ مفهوم معنایی وجود دارد که این عدد می‌تواند یک کران بالا برای هر متن باشد.

در مرحلهٔ نخست، هر واژهٔ یکتا (واژگان تکراری بررسی نمی‌شوند) در هر فرازنجیره‌ای که با آن مرتبط است، قرار می‌گیرد. که این امر با جستجوی هر معنا ممکن s برای واژه و اضافه‌کردن واژه به فرازنجیره‌هایی با شناسه s، ابرمفهوم یا زیرمفهوم s ، هم‌پایه s و مرتبط(مرتبط-است-با) s صورت می‌گیرد. این الگوریتم در شکل (۲) نشان داده شده است.

```
for all candidate words w in the document do
  for all senses s of w do
    Insert w into metachain hashtable at index s
    for all hyponyms c of s do
      Insert w into metachain hashtable at index c
    end for
    for all hypernyms p of s do
      Insert w into metachain hashtable at index p
      for all hyponyms c of p do
        Insert w into metachain hashtable at index c
      end for
    end for
    for all related-to c of s do
      Insert w into metachain hashtable at index c
    end for
  end for
end for
```

(شکل-۲): شبه‌کد ایجاد فرازنجیره

(Figure-2): pseudocode for metachain creation

[2] hypernym
[3] hyponym
[4] coordinate
[5] Related-to

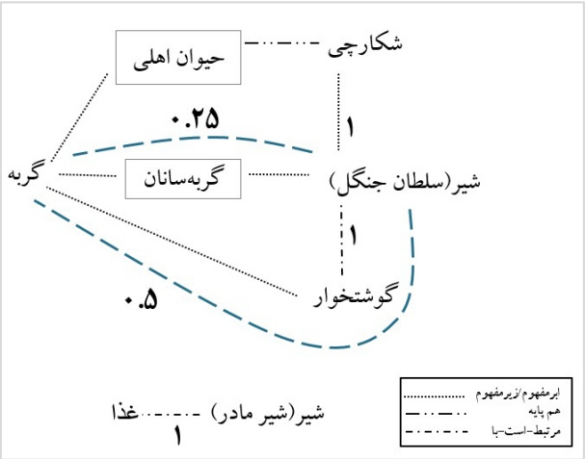
[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]

[DOI: 10.29252/jsdp.15.4.95]

سطرهای جدول (۱) از جدول فرازنجیره‌های ایجادشده برای متنی شامل واژگان نامزد «شیر»، «شکارچی»، «گربه»، «گوشتخوار» و «غذا» برگزیده شده‌اند. هر یک از این واژگان به‌عنوان ورودی به الگوریتم شکل (۲) داده و جدول فرازنجیره‌ها به‌مرور تکمیل می‌شود. ستون نخست شناسه فرازنجیره‌ها را مشخص می‌کند که درواقع همان کلیدها در جدول درهم‌سازی هستند. این شناسه‌ها یک معنا در فارس‌نت هستند که تعریف معانی آنها در ستون دوم آورده شده است. ستون سوم مقادیر کلیدها در جدول درهم‌سازی هستند. درواقع ستون سوم فهرستی از واژگان متن را نشان می‌دهد که معنای آنها با عنوان فرازنجیره مرتبط هستند. این واژگان با استفاده از روابط مورد نیازی که در فارس‌نت تعریف شده‌اند، به فهرست عبارات هر فرازنجیره اضافه می‌شوند. این روابط در شکل (۳) نشان داده شده است.

(جدول-۱): نمونه‌ای از جدول فرازنجیره
(Table-1): Sample metachain table

شناسه فرازنجیره	معنی(عنوان فرازنجیره)	عبارات فرازنجیره		
۱۴۰۰۹	شیر : "سلطان جنگل"	شیر	شکارچی	گوشتخوار
۱۴۰۰۵	شیر: " شیر مادر"	شیر	غذا	
۱۰۷۹۴	شکارچی: "حیوان شکارگر"	شیر	شکارچی	
۱۴۱۱۲	گربه‌سانان	شیر	گربه	
۱۲۷۲۰	گوشتخوار: "گروهی از پستانداران"	شیر	گربه	گوشتخوار
۱۴۱۰۶	گربه	گربه	گوشتخوار	
۱۰۷۷۱	حیوان اهلی	شکارچی	گربه	
۱۲۷۱۴	غذا: "ماده غذایی"	شیر	غذا	



(شکل-۳): روابط معانی در فارس‌نت برای واژگان جدول (۱)
(Figure-3): Senses relation in FarsNet for words in table (1)

۳-۲-۲- مرحلهٔ دوم

بعد از ایجاد فرازنجیره‌ها، باید رفع ابهام معنایی صورت گیرد تا برای هر واژه فقط یک معنا، آن هم معنایی که در متن مورد نظر است، به‌دست آید. البته ممکن است که یک واژه در بخش‌های مختلف متن معانی مختلفی داشته باشد که ما در اینجا از آن صرف نظر می‌کنیم و برای رخدادهای مختلف یک واژه در متن، فقط یک معنا در نظر می‌گیریم. بعد از تعیین معانی درست تمام واژگان، معانی دیگر واژگان از فهرست عبارات جدول فرازنجیره‌ها (مقادیر کلیدها) حذف می‌شوند. توجه به این نکته ضروری است که به‌روزرسانی جدول فرازنجیره‌ها بعد از تعیین معانی درست تمام واژگان و نه هر کلمه است. چون، اگر بعد از تعیین معنای هر واژه جدول به‌روزرسانی شود، این تغییر در فرازنجیره‌ها منتشر شده و به‌نوعی معانی به واژگان دیگر تحمیل می‌شوند [27]. تعیین معنای درست برای هر واژه، با تخصیص رتبه به هر یک از معانی آن و انتخاب معنای با بالاترین رتبه صورت می‌پذیرد.

یکی دیگر از تفاوت‌های الگوریتم حاضر با روش گالی و مک‌کیوان این است که در روش گالی و مک‌کیوان معانی با یکدیگر فقط ارتباط مستقیم دارند [7]؛ ولی در روش حاضر معانی می‌توانند به‌وسیلهٔ کلمه‌ای خارج از واژگان متن، با طول رابطهٔ دو، با یکدیگر ارتباط داشته باشند. همان‌طورکه در شکل (۳) مشخص است، معانی «شکارچی» و «گربه» با استفاده از یک معنای میانی که جزء واژگان متن نیست، به یکدیگر متصل شده‌اند.

برای تعیین رتبهٔ یک معنا، از تعداد روابط به‌طول یک و دوی آن با سایر معانی، استفاده می‌شود. شیوهٔ تعیین رتبه نیز یکی دیگر از تفاوت‌های الگوریتم حاضر با روش گالی و مک‌کیوان است. در اینجا این فرضیه را مطرح کردیم که معنای درست یک واژه در متن آن معنایی است که گستردگی روابط بیش‌تری با سایر معانی دارد. رتبهٔ هر معنا، به‌ازای هر معنا که به‌طور مستقیم با آن در ارتباط باشد باضافهٔ یک می‌شود و به‌ازای هر معنا که با فاصلهٔ دو با آن در ارتباط باشد، درصورتی‌که معنای میانی جزء معانی واژگان متن باشد، به‌اضافهٔ ۰/۵ و در غیر این صورت به‌اضافهٔ ۰/۲۵ می‌شود. این شیوهٔ امتیازدهی در شکل (۳) برای دو معنای مختلف واژهٔ شیر با عنوان «سلطان جنگل» و «شیر مادر» مشخص شده است که درکل با مجموع امتیازها، رتبهٔ ۲/۷۵ برای معنای نخست و رتبه یک برای معنای دوم به‌دست می‌آید. درنتیجه «سلطان جنگل» به‌عنوان معنای درست برای واژهٔ شیر در نظر گرفته

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]

می‌شود. پس از تعیین رتبهٔ تمام معانی، معانی نادرست از فهرست عبارات جدول فرازنجیره‌ها حذف می‌شوند. در جدول (۱) فقط برای واژهٔ شیر دو معنا آورده شده و باقی واژگان تنها یک معنا دارند. درنتیجه، تنها معنای «شیر مادر» جزء معانی حذفی است و از فهرست عبارات فرازنجیره‌هایی با شناسه ۱۴۰۰۵ و ۱۲۷۱۴ حذف می‌شود.

۳-۲-۳- مرحلهٔ سوم

در این مرحله زنجیره‌های لغوی ایجاد و برای این منظور فرازنجیره‌هایی که لغات مشترک دارند، با یکدیگر ترکیب می‌شوند که البته واژگان تکراری در نظر گرفته نمی‌شوند. این فرآیند بر روی زنجیره‌های ادغام‌شده ادامه می‌یابد تا زمانی که هیچ دو زنجیره‌ای عضو مشترک نداشته باشند. جدول (۱) پس از رفع ابهام در مرحلهٔ دوم، برای تعیین زنجیره‌های لغوی در مرحلهٔ سوم مورد پردازش قرار می‌گیرد که درنهایت واژگان متن در دو زنجیره قرار می‌گیرند که در جدول (۲) نشان داده شده است.

(جدول-۲): فهرست زنجیره‌های لغوی برای واژگان جدول (۱)
(Table-2): List of lexical chains for words in table (1)

شماره زنجیره	عبارات زنجیره
۱	شکارچی, شیر, گربه, گوشتخوار
۲	غذا

۳-۳- تعیین ویژگی‌های واژگان نامزد

برای هر واژهٔ نامزد ده ویژگی در نظر گرفته می‌شود. به غیر از ویژگی‌های یک، دو، سه و نه بقیه از زنجیرهٔ لغوی گرفته شده‌اند. به این ترتیب، درصورت وجود نقص‌هایی در فارس‌نت و یا الگوریتم ایجاد زنجیره، چهار ویژگی غیروابسته به زنجیره می‌توانند برای عبارات با اهمیت تأثیرگذار باشند. این ده ویژگی عبارت اند از:

۱. **بسامد واژه:** تعداد دفعاتی که واژه در متن تکرار شده است.

۲. **نخستین رخداد واژه:** از تقسیم تعداد واژگانی که قبل از نخستین رخداد واژه در متن آمده‌اند بر تعداد کل واژگان نامزد به‌دست می‌آیند. این ویژگی به این خاطر است که واژگان کلیدی اغلب در ابتدای متن ظاهر می‌شوند.

۳. **آخرین رخداد واژه:** از تقسیم تعداد واژگانی که قبل از آخرین رخداد واژه در متن آمده‌اند، بر تعداد کل واژگان نامزد به‌دست می‌آید. این ویژگی به این خاطر است که واژگان کلیدی علاوه‌بر ابتدای متن در انتهای متن نیز زیاد به کار برده می‌شوند.

۴. **رتبهٔ واژه در زنجیره:** رتبه‌ای که در مرحلهٔ دوم برای معنای درست کلمه به‌دست آمد، به‌عنوان رتبهٔ واژه در نظر گرفته می‌شود. این رتبه میزان ارتباط واژه با سایر واژگان زنجیره را نشان می‌دهد.

۵. **رتبهٔ زنجیرهٔ واژه:** مجموع رتبه‌های واژگان زنجیره‌ای که واژهٔ مورد نظر در آن قرار دارد.

۶. **محدودهٔ زنجیرهٔ لغوی:** میزانی از متن را که زنجیره پوشش می‌دهد، نشان می‌دهد. زنجیره‌ای با اهمیت است که واژگان آن، محدودهٔ بیش‌تری از متن را پوشش دهند. برای محاسبه آن از رابطه زیر استفاده می‌شود [5].

$$\text{محدوده زنجیره لغوی} = (\text{آخرین رخداد واژگانی که عضو زنجیره هستند})\text{بیشینه} - (\text{نخستین رخداد واژگانی که عضو زنجیره هستند})\text{کمینه} \quad (۱)$$

۷. **درصد پوشش زنجیره:** درصد واژگان نامزدی را که در زنجیره وجود دارد، نشان می‌دهد. فرض بر این است زنجیره‌هایی که واژگان بیش‌تری را از متن پوشش می‌دهند، ارتباط نزدیک‌تری با موضوع متن دارند. برای محاسبهٔ درصد پوشش زنجیره از رابطهٔ زیر استفاده می‌شود:

$$\text{درصد پوشش زنجیره} = \frac{\text{تعداد واژگان نامزد درون زنجیره}}{\text{تعداد کل واژگان نامزد}} \quad (۲)$$

۸. **درصد پوشش واژه:** این ویژگی میزان اهمیت واژه نامزد نسبت به کل واژگان نامزد را نشان می‌دهد که به‌صورت زیر محاسبه می‌شود.

$$\text{درصد پوشش واژه} = \frac{\text{رتبهٔ واژه در زنجیره}}{\text{رتبه زنجیرهٔ واژه}} \times \text{درص پوشش زنجیره} \quad (۳)$$

۹. **وجود واژه در عنوان متن:** این ویژگی به این خاطر است که بسیاری از واژگان کلیدی در عنوان متن وجود دارند. در صورت وجود واژهٔ نامزد در عنوان متن، این ویژگی مقدار یک و در غیر این صورت مقدار صفر را می‌گیرد.

۱۰. **وابستگی زنجیره با عنوان متن:** زنجیره‌ای که تعداد واژگان بیشتری را از عنوان پوشش دهد، اهمیت بیشتری دارد. برای محاسبهٔ وابستگی زنجیره با عنوان متن از رابطهٔ زیر استفاده می‌شود:

$$\text{وابستگی زنجیره با عنوان متن} = \frac{\text{تعداد واژگان عنوان درون زنجیره}}{\text{تعداد کل واژگان عنوان}} \quad (۴)$$

پس از تعیین این ده ویژگی، بردار ویژگی واژه نامزد ایجاد می‌شود. برای متون آموزشی مقدار این بردار با بررسی کلیدی‌بودن یا کلیدی‌نبودن واژه در متن به‌دست می‌آید.

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]

درصورتی‌که واژه در متن کلیدی باشد، مقدار یک و در غیر این صورت مقدار ۱- را می‌گیرد. مجموعه زوج های (x,v)، که در آن x بردار ویژگی و v مقدار آن است، برای واژگان متون آموزشی به‌دست می‌آیند و در فایلی به نام دادهٔ آموزشی ذخیره می‌شوند. برای متنی با واژگان کلیدی نامشخص نیز این بردارهای ویژگی محاسبه می‌شوند؛ ولی مقدار این بردارها نامشخص است.

۴-۳- دسته‌بندی واژگان نامزد

در این مرحله دسته‌بند با استفاده از دادهٔ آموزشی ایجادشده، مقدار بردارهای ویژگی متن ورودی را تعیین می‌کند که این مقدار کلیدی یا کلیدنبودن واژه نامزد را نشان می‌دهد. واژگانی که دسته‌بند مقدار بردار ویژگی آنها را یک تعیین کند به‌عنوان واژگان کلیدی در نظر گرفته می‌شوند. از مجموعه دسته‌بندهای موجود در کتابخانه وکا[1] برای دسته‌بندی واژگان نامزد استفاده شده است.

۵-۳- پس‌پردازش

تعداد زیادی از واژگان نامزد تعیین‌شده، شامل عبارات دوواژه‌ای "اسم،اسم" و یا "اسم،صفت" هستند؛ اما بیش‌تر این عبارات در فارس‌نت وجود ندارد و نادیده گرفته می‌شوند. درنتیجه بسیاری از واژگان کلیدی استخراج‌شده از مرحلهٔ قبل را عبارات تک‌واژه‌ای تشکیل می‌دهد. جستجو برای عبارات تک‌واژه‌ای در فارس‌نت، نتیجه بهتری نسبت به عبارات دوواژه‌ای دارد. از آن‌جایی که در مرحلهٔ تعیین واژگان نامزد، هر جزء از عبارات دوواژه‌ای را هم به‌عنوان کلمهٔ نامزد در نظر گرفتیم، پس از تعیین واژگان کلیدی، آن دسته از عبارات دوواژه‌ای نامزد که هرجزء آن به‌عنوان واژهٔ کلیدی تعیین شده‌اند، جایگزین اجزای خود می‌شوند. درنتیجه، بسیاری از عبارات دوواژه‌ای کلیدی که در مرحله قبل میسر نشدند، در این مرحله به‌دست می‌آیند.

۴- نتایج پژوهش

داده‌های این پژوهش از مجلات علمی پژوهشی *پردازش علائم و داده‌ها* و *ماشین بینایی و پردازش تصویر* جمع‌آوری شده‌اند. این مجموعه شامل ۱۰۲ مقاله با موضوعات پردازش متن،گفتار و تصویر است که عنوان، چکیده و واژگان کلیدی هر مقاله جمع‌آوری شده است.

[1] Weka

به‌منظور ارزیابی سامانه طراحی‌شده، از نرم‌افزار وکا استفاده کرده‌ایم. دادهٔ آموزشی به‌دست آمده، شامل ۵۸۴۲ نمونه آموزشی است که ۶۱۶ عدد از آنها واژگان کلیدی هستند و ۵۲۲۶ از آنها غیر کلیدی هستند. به‌منظور ایجاد نتیجهٔ بهتر، داده‌ها پالایش شده و داده‌های پیوسته تبدیل به داده‌های گسسته شده‌اند. این داده‌ها با استفاده از دسته‌بندهای مختلفی که در جدول (۳) آورده شده است، مورد ارزیابی قرار گرفتند. به این صورت که ۶۶٪ از داده‌ها به‌عنوان داده آموزش و ۳۴٪ آنها را به‌عنوان داده‌های آزمون در نظر گرفتیم. معیار بازخوانی[2] ، معیار دقت[3] و معیار F[4] (میانگین هم‌ساز) برای هر یک از این دسته‌بندها در جدول (۳) نشان داده شده است.

از کتاب‌خانه libsvm برای ماشین بردار پشتیبان استفاده شده است، که بهترین نتیجهٔ آن با استفاده از هستهٔ توابع پایهٔ شعاعی[5] به‌دست آمده است. ماشین بردار پشتیبان بالاترین دقّت را در بین دسته‌بندهای مورد ارزیابی دارد، درحالی‌که کم‌ترین بازخوانی نیز مربوط به همین دسته‌بند می‌شود؛ از این رو معیار F بالایی ندارد.

(جدول-۳): ارزیابی عملکرد روش پیشنهادی
(Table-3): Performance evaluation of proposed method

نوع دسته‌بند	دقت	بازخوانی	معیار F
بیز ساده	46.8	57.9	51.8
بگینگ - بیز ساده	47.2	57.9	52
آدابوست - بیز ساده	59.7	35.1	44.2
درخت تصمیم (j48)	0.65	37.6	47.6
ماشین بردار پشتیبان(libsvm)	**67.7**	32.2	43.6
بیز ساده (بدون ویژگی‌های رتبه زنجیره کلمه، درصد پوشش زنجیره، درصد پوشش کلمه و وابستگی زنجیره با عنوان متن)	48	**58.9**	**52.9**

با حذف ویژگی‌های رتبهٔ زنجیرهٔ واژه، درصد پوشش زنجیره، درصد پوشش واژه و وابستگی زنجیره با عنوان متن، به بالاترین معیار F و بازخوانی با استفاده از دسته‌بند بیز ساده رسیدیم، که نشان می‌دهد این چهار ویژگی تأثیرگذاری زیادی در نتایج به‌دست‌آمده ندارند.

جدول (۴) نتایج به‌دست‌آمده را از روش‌های مختلف باناظر در استخراج واژگان کلیدی نشان می‌دهد. در بخش ۲

[2] Recall
[3] Precision
[4] F-Measure
[5] Radial basis function (RBF)

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]

مختصری راجع به هر یک از این روش‌ها توضیح داده شد. فقط روش‌هایی که معیارهای ارزیابی آنها با پژوهش انجام شده یکسان بود، در جدول گذاشته شده‌اند.

(جدول-۴): مقایسه روش پیشنهادی با سایر روش‌های با ناظر

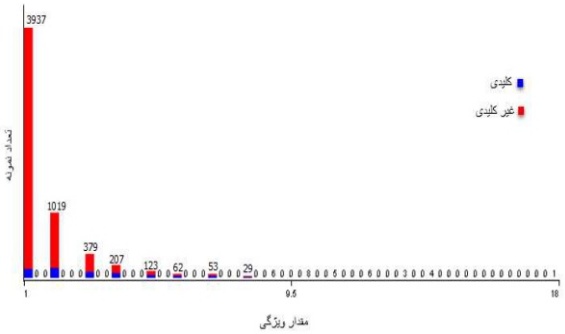
(Table-4): Comparison of the proposed method with other supervised methods

روش	دقت	بازخوانی	معیار F
[14]	38.9	54.8	45.5
[15]	**84.1**	59.46	**69.67**
[16]	28.3	26.1	25.2
[17]	66.37	41.96	51.25
[18]	26.40	34.15	29.78
[19]	21.3	41.3	28
[20]	76.32	**62.88**	68.95
روش پیشنهادی	48	58.9	52.9

۴-۱- ارزیابی ویژگی‌های واژگان

در این بخش به بررسی هر یک از ده ویژگی در تعیین واژگان کلیدی می‌پردازیم. محور افقی در همه نمودارهای این بخش، مقادیر ویژگی مورد نظر و محوری عمودی تعداد واژگان کلیدی و غیرکلیدی در داده آموزشی است.

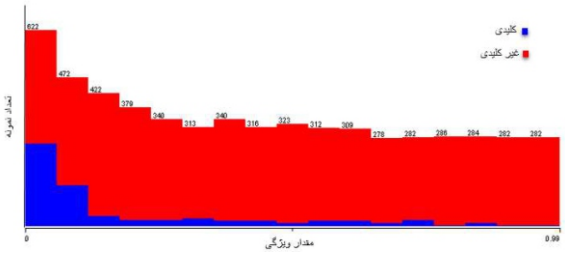
با توجه به نمودار شکل (۴)، احتمال کلیدی‌بودن واژگان در بسامدهای بالا بیش‌تر از غیرکلیدی‌بودن آن است. اما همهٔ واژگان با بسامد پایین غیرکلیدی نیستند. بیشتر واژگان نامزد دارای بسامد نزدیک به یک بوده‌اند که بخش زیادی از آن را واژگان غیر کلیدی تشکیل می‌دهند؛ اما تعداد قابل ملاحظه‌ای از واژگان کلیدی نیز در این بسامد قرار دارند. این پدیده بیان‌گر این است که واژه‌ای ممکن است، یک‌بار در متن آورده شده باشد، ولی کلیدی باشد.



(شکل-۴): تعداد واژگان کلیدی و غیرکلیدی در دادهٔ آموزشی برحسب مقدار «بسامد واژه»

(Figure-4): Number of keywords and non-keywords in training data according to value of «word frequency»

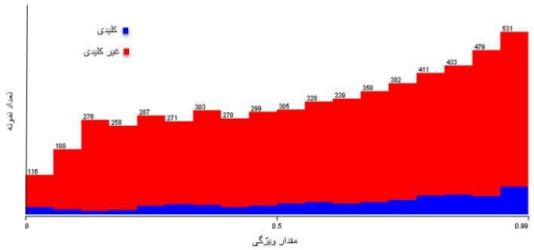
با توجه به نمودار شکل (۵)، حجم زیادی از واژگان کلیدی را واژگانی تشکیل می‌دهند که نخستین رخداد آنها در ابتدای متن باشد. از طرفی، تعداد زیادی از واژگان نامزدی که در ابتدای متن می‌آیند، غیرکلیدی هستند. به همین دلیل، نمی‌توان احتمال بالایی برای کلیدی‌بودن واژگان ابتدایی متن در نظر گرفت.



(شکل-۵): تعداد واژگان کلیدی و غیرکلیدی در دادهٔ آموزشی برحسب مقدار«نخستین رخداد کلمه»

(Figure-5): Number of keywords and non-keywords in training data according to value of « first occurrence of word »

با توجه به نمودار شکل (۶)، حجم زیادی از واژگان کلیدی را واژگانی تشکیل می‌دهند که آخرین رخداد آنها در انتهای متن باشد. از طرفی، تعداد زیادی از واژگان نامزدی که در انتهای متن می‌آیند، غیرکلیدی هستند. به همین دلیل، نمی‌توان احتمال بالایی برای کلیدی‌بودن واژگان انتهایی متن در نظر گرفت.



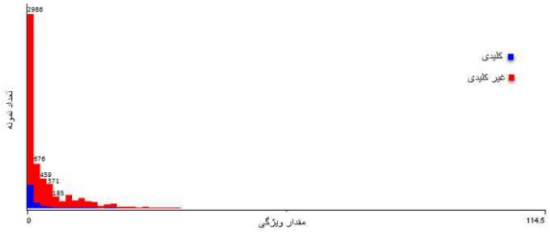
(شکل-۶): تعداد واژگان کلیدی و غیرکلیدی در داده آموزشی برحسب مقدار «آخرین رخداد کلمه»

(Figure-6): Number of keywords and non-keywords in training data according to value of « last occurrence of word »

با توجه به نمودار شکل(۷)، حجم زیادی از واژگان کلیدی، دارای رتبهٔ بین صفر تا یک هستند و رتبه‌های خیلی بالا مربوط به واژگان غیرکلیدی می‌شوند. واژگانی مانند «روش، حل، کار» در بسیاری از مقالات ارزیابی‌شده قراردارند. این واژگان، اغلب، با یکدیگر ارتباط زیادی دارند و حجم زیادی از واژگان با رتبه‌های زیاد را این نوع از واژگان تشکیل می‌دهند. به همین دلیل، رتبه‌های زیاد به واژگان غیرکلیدی اختصاص می‌یابد؛ اما برای رتبهٔ پایین واژگان کلیدی دلایل مختلفی وجود دارد که در ادامه به آنها اشاره می‌کنیم:

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]

[DOI: 10.29252/jsdp.15.4.95]

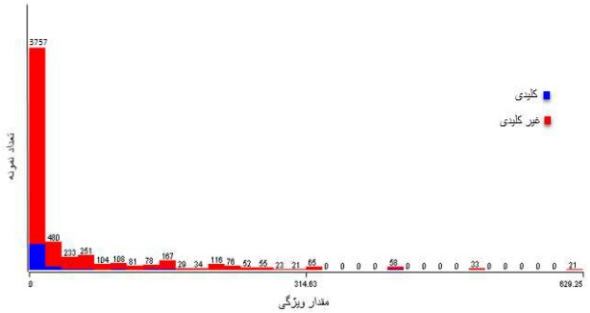


(شکل-۷): تعداد واژگان کلیدی و غیرکلیدی در دادهٔ آموزشی برحسب مقدار «رتبه واژه در زنجیره»

(Figure-7): Number of keywords and non-keywords in training data according to value of «word score in chain »

- هنجارسازی، ریشه‌یابی، تقطیع واژگان و برچسب‌گذاری نادرست باعث ایجاد واژگان نادرست و درنتیجه روابط نادرست میان آنها شود. برای مثال، اگر واژه «پیاده سازی» به‌صورت «پیاده‌سازی» نرمال نشود، این کلمه به‌عنوان دو واژه «پیاده» و «سازی» در نظر گرفته می‌شود. در تقطیع واژگان، اگر واژهٔ مرکب «زد و خورد» تشخیص داده نشود، این واژه به واژگان مجزا تبدیل می‌شود. ابزاری که در این پروژه به‌منظور تقطیع واژگان استفاده شده است، قادر به تشخیص اسامی مرکب نیست؛ یا در برچسب‌گذاری نحوی اگر فعل مرکب «قرار گرفته است» به‌درستی تشخیص داده نشود، واژهٔ «قرار» به‌عنوان اسم و «گرفته است» فعل در نظر گرفته می‌شود. این باعث ورود بسیاری از واژگان نادرست به مجموعهٔ واژگان نامزد می‌شود. نقص در ریشه‌یابی، نیز، باعث ایجاد خطاهایی می‌شود. برای مثال ریشه‌یاب استفاده‌شده در این پژوهش قادر به ریشه‌یابی واژگان با نشانه‌های جمع عربی نیست. به‌عنوان مثال، واژهٔ «کلمات» بعد از ریشه‌یابی تغییری نمی‌کند و این واژه با نشانهٔ جمع «ات» در فارس‌نت موجود نیست. درنتیجه، به‌هیچ‌وجه مورد پردازش قرار نمی‌گیرد.
- نقص فارس‌نت یکی دیگر از دلایل رتبهٔ پایین برای برخی واژگان کلیدی است. ارتباط برخی از واژگان با استفاده از روابط تعریف‌شده در الگوریتم پیشنهادی میسر نیست. برای مثال، بین واژهٔ «زبان» و «فارسی» هیچ ارتباطی در فارس‌نت وجود ندارد. هم‌چنین، در برخی مواقع رابطه ابرمفهوم/زیرمفهوم به‌صورت یک‌طرفه بیان شده است. درصورتی‌که، این رابطه دو طرفه است. برای مثال «رنگ» ابرمفهوم «آبی» است؛ ولی «آبی» زیرمفهوم رنگ نیست. معانی ناقص برخی واژگان یکی دیگر از اشکالات فارس‌نت است. برای مثال، واژهٔ «درخت» فقط یک معنا در فارس‌نت دارد؛ درصورتی‌که این واژه می‌تواند به‌عنوان معنای «نمودار» هم استفاده شود.
- رفع ابهام معنایی نادرست موجب ایجاد روابط غلط میان واژگان می‌شود. درنتیجه، بالابردن دقت الگوریتم رفع ابهام موجب بهبود نتایج می‌شود. برای مثال اگر در عبارت «آشکارسازی گل در ویدیو مسابقات فوتبال» واژهٔ «گل» به‌اشتباه رفع ابهام شود، آن‌گاه بین این واژه و واژهٔ «فوتبال» ارتباطی برقرار نمی‌شود.

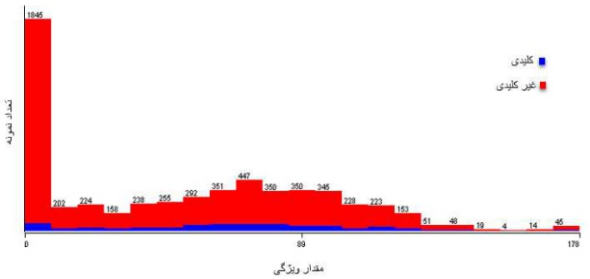
با توجه به نمودار شکل (۸)، زنجیره‌هایی با رتبه‌های خیلی بالا بیش‌تر شامل واژگان غیر کلیدی هستند، که با توجه به نتایج به‌دست‌آمده از ویژگی قبلی، این نتیجه دور از انتظار نیست.



(شکل-۸): تعداد واژگان کلیدی و غیرکلیدی در دادهٔ آموزشی برحسب مقدار «رتبه زنجیرهٔ واژه»

(Figure-8): Number of keywords and non-keywords in training data according to value of «word chain score »

با توجه به نمودار شکل (۹)، احتمال کلیدی‌بودن یک عبارت، با زیادشدن محدودهٔ زنجیرهٔ لغوی بالا می‌رود؛ چون، بیش‌تر واژگان غیرکلیدی محدودهٔ زنجیرهٔ لغوی پایین دارند.

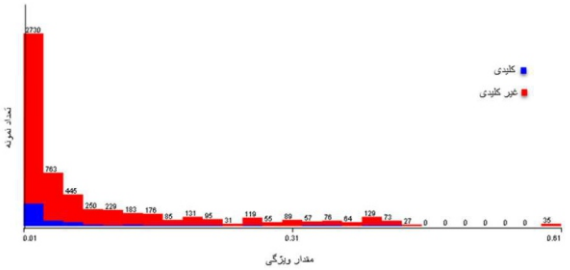


(شکل-۹): تعداد واژگان کلیدی و غیرکلیدی در داده آموزشی برحسب مقدار «محدوده زنجیرهٔ لغوی»

(Figure-9): Number of keywords and non-keywords in training data according to value of «lexical chain span »

با توجه به نمودار شکل (۱۰)، بیشتر واژگان کلیدی در زنجیره‌های کوتاه قرار گرفته‌اند. البته، حجم زیادی از واژگان غیرکلیدی هم در زنجیره‌های کوتاه هستند و این تمایز بین واژگان کلیدی و غیرکلیدی را دشوار می‌کند.

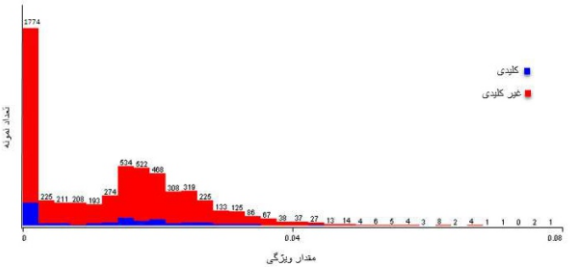
[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]



(شکل-۱۰): تعداد واژگان کلیدی و غیرکلیدی در دادهٔ آموزشی برحسب مقدار «درصد پوشش زنجیره»

(Figure-10): Number of keywords and non-keywords in training data according to value of «coverage percentage of chain »

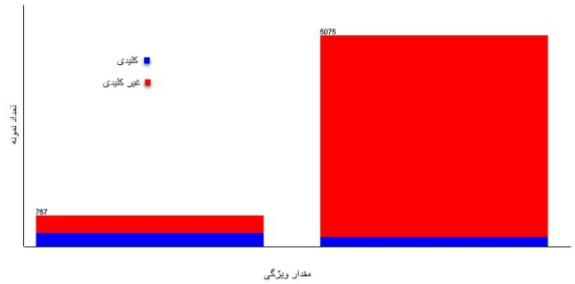
با توجه به نمودار شکل (۱۱)، با افزایش مقدار درصد پوشش واژه، نسبت تغییرات تعداد واژگان کلیدی و غیرکلیدی به‌طورتقریبی یکسان بوده است. به همین دلیل، این معیار بین واژگان کلیدی و غیرکلیدی خیلی متمایزکننده نیست.



(شکل-۱۱): تعداد واژگان کلیدی و غیرکلیدی در داده آموزشی برحسب مقدار «درصد پوشش واژه»

(Figure-11): Number of keywords and non-keywords in training data according to value of «coverage percentage of word »

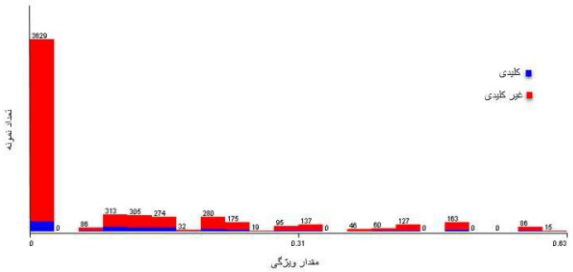
با توجه به نمودار شکل (۱۲)، درصد زیادی از واژگان عنوان متن جزء واژگان کلیدی هستند. این معیار نتیجه مطلوبی را در تمایز بین واژگان کلیدی و غیرکلیدی ایفا می‌کند.



(شکل-۱۲): تعداد واژگان کلیدی و غیرکلیدی در دادهٔ آموزشی برحسب مقدار «وجود واژه در عنوان»

(Figure-12): Number of keywords and non-keywords in training data according to value of «being the word in title »

با توجه به نمودار شکل (۱۳)، تعداد زیادی از واژگان غیرکلیدی در زنجیره‌هایی قرار گرفتند که واژگان عنوان کمتر در آنها وجود دارد؛ اما با افزایش این مقدار، نسبت تغییرات تعداد واژگان کلیدی و غیرکلیدی به‌طورتقریبی یکسان است. به همین دلیل این معیار خیلی نمی‌تواند در تعیین واژگان کلیدی تأثیرگذار باشد.



(شکل-۱۳): تعداد واژگان کلیدی و غیرکلیدی در دادهٔ آموزشی برحسب مقدار «وابستگی زنجیره با عنوان متن»

(Figure-13): Number of keywords and non-keywords in training data according to value of «chain dependency with text title»

در جدول (۵)، تاثیر هر یک از ویژگی‌های «رتبهٔ زنجیرهٔ واژه»، «درصد پوشش زنجیره»، «درصد پوشش واژه» و «وابستگی زنجیره با عنوان» بر نتایج حاصل از دسته‌بند بیز ساده بررسی شده است. حذف سه ویژگی نخست تغییری در نتیجه ایجاد نمی‌کند و حذف ویژگی «وابستگی زنجیرهٔ باعنوان» نتایج را بهبود می‌بخشد.

(جدول-۵): تأثیر ویژگی‌های «رتبهٔ زنجیرهٔ واژه»، «درصد پوشش زنجیره»، «درصد پوشش واژه» و «وابستگی زنجیرهٔ باعنوان» بر نتایج حاصل از دسته‌بند بیز ساده

(Table 5): impact of «word chain score», «coverage percentage of chain» , « coverage percentage of word» and «chain dependency with text title » features on result of Naive Bayes classifier

دسته‌بند	دقت	بازخوانی	معیار F
بیز ساده (تمام ویژگی‌ها)	46.8	57.9	51.8
بیز ساده (بدون ویژگی رتبه زنجیره کلمه)	46.8	57.9	51.8
بیز ساده (بدون ویژگی درصد پوشش زنجیره)	46.8	57.9	51.8
بیز ساده (بدون ویژگی درصد پوشش کلمه)	46.8	57.9	51.8
بیز ساده (بدون ویژگی وابستگی زنجیره با عنوان متن)	**48**	**58.9**	**52.9**

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]
[DOI: 10.29252/jsdp.15.4.95]

۵- جمع‌بندی و پیشنهادها

در این مقاله، استخراج واژگان کلیدی متون فارسی با استفاده از رویکردی باناظر مورد بررسی قرار گرفت. واژگان نامزد متن با استفاده از اطلاعات نحوی و کتاب‌خانه هضم مشخص شدند. برای استخراج ویژگی‌ها از زنجیرهٔ لغوی استفاده شد که این زنجیره با استفاده از هستان‌شناسی فارس‌نت و الگوریتم گالی و مک‌کیوان ایجاد و هم‌چنین، برای دریافت نتایج بهتر تغییراتی در این الگوریتم داده شد. برخی ویژگی‌های استخراج‌شده از واژگان، مربوط به زنجیرهٔ لغوی و برخی از ویژگی‌ها نیز مربوط به اطلاعات آماری واژه بودند. با توجه به نتایج به‌دست‌آمده ویژگی‌های معنایی در کنار ویژگی‌های آماری باعث بهبود نتایج می‌شوند. هم‌چنین دسته‌بند بیز ساده بهترین نتیجه را در بین دسته‌بندهای مورد ارزیابی حاصل کرد. که البته حذف برخی ویژگی‌های حاصل از زنجیره‌ی لغوی باعث بهبود عملکرد آن نیز شد.

دقّت ابزارهای هنجارسازی، تقطیع واژگان، برچسب‌گذاری نحوی و ریشه‌یابی واژگان در دقت سامانهٔ پیشنهادی تأثیرگذار هستند. ازاین‌رو، استفاده از ابزارهای کارآمدتر پردازش زبان فارسی می‌تواند دقت الگوریتم پیشنهادی را بهبود بخشد.

حجم بالای واژگان غیرکلیدی در مقابل واژگان کلیدی، کار تمایز بین آنها را دشوار می‌کند. به‌نظر می‌آید با کم‌کردن تعداد واژگان نامزد بتوان به‌شکلی مشکل نامتوازن‌بودن داده‌ها را رفع کرد. درواقع باید میزان واژگان غیرکلیدی مورد پردازش را کاهش دهیم. این امر می‌تواند با درنظرگرفتن واژگان نامزدی با بسامد و یا ویژگی مشخص دیگری صورت گیرد. با توجه به خطی‌بودن الگوریتم زنجیرهٔ لغوی، با کاهش تعداد واژگان نامزد، می‌تواند باعث بهبود اجرای برنامه نیز شد.

هم‌چنین، می‌توانیم از ترکیب چندین دسته‌بند با مجموعه‌ویژگی‌های مختلف استفاده کنیم و واژگان کلیدی نهایی را با استفاده از رأی اکثریت به‌دست آوریم. در این صورت ممکن است، کاستی‌های برخی ویژگی‌ها و یا دسته‌بندها جبران شود.

از آنجا که داده‌های مورد پردازش در این پژوهش مقالات علمی بودند، تعدادی از واژگان تخصصی مقالات در فارس‌نت وجود ندارند و مورد پردازش قرار نگرفتند. می‌توان با استفاده از پیکره‌ها و یا اصطلاح‌نامه‌های تخصصی، ارتباط بین این گونه واژگان را نیز به‌دست آورد.

۶- مراجع 6- References

[1]J. Wang, J. Liu and C. Wang, "Keyword Extraction Based on PageRank," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer Berlin Heidelberg*, 2007.

[2]X. Li and F. Song, "Keyphrase Extraction and Grouping Based on Association Rules," in *FLAIRS Conference*, Hollywood, Florida, 2015.

[3] B. Lott, "Survey of keyword extraction techniques," UNM Education, 2012.

[4] R. Nelken and S. M. Shieber, "Lexical chaining and word-sense-disambiguation," *School of Engineering and Applied Sciences*, Harvard University, Cambridge ,Technical Report TR-06-07, MA, 2007.

[5] G. Ercan, "Automated text summarization and keyphrase extraction," M.S. thesis, bilkent university, Ankara, Turkey, 2006.

[6] M. Shamsfard, "Towards Semi Automatic Construction of a Lexical Ontology for Persian," in *sixth International Conference on Language Resources and Evaluation*, Morocco, 2008.

[7] M. Galley and K. McKeown, "Improving word sense disambiguation in lexical chaining," *IJCAI,* vol. 3, pp. 1486-1488, 2003.

[8] k. Hasan and v. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," in *ACL*, 2014.

[9] C. Wu, M. Marchese and J. Jiang, "Machine Learning-Based Keywords Extraction for Scientific Literature," *Journal of Universal Computer Science,* vol. 13, no. 10, pp. 1471-1483, 2007.

[10] S. Beliga, "Keyword extraction: a review of methods and approaches," University of Rijeka, Department of Informatics, Rijeka, 2014.

[11] S. beliga, A. Mestrovic and S. Martincic, "An overview of graph-based keyword extraction methods and approaches," *Journal of information and organizational sciences,* vol. 39, no. 1, pp. 1-20, 2015.

[12] T. Pay and S. Lucci, "Automatic Keyword Extraction: An Ensemble Method," in *2017 IEEE International Conference on Big Data*, Boston, 2017.

[13] M. Johansson and P. Lindstrom, "Keyword Extraction using Machine Learning," M.S. thesis, Gothenburg University, Gothenburg, Sweden, 2010.

[14] A. Hulth, "Combining machine learning and natural language processing for automatic keyword extraction," Ph.D. dissertation, Stockholm University, Stockholms, Sweden, 2004.

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]

[DOI: 10.29252/jsdp.15.4.95]

[15] Y. HaCohen-kerner, Z. Gross and A. Masa, "Automatic extraction and learning of keyphrases from scientific articles," in *International Conference on Intelligent Text Processing and Computational Linguistics. Springer Berlin Heidelberg*, 2005.

[16] O. Medelyan and I. H. Witten, "Thesaurus based automatic keyphrase indexing," in *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries. ACM*, 2006.

[17] C. Zhang, H. WANG, Y. LIU, D. WU, Y. LIAO and B. WANG, "Automatic Keyword Extraction from Documents Using Conditional Random Fields," *Computational Information Systems,* vol. 4, no. 3, pp. 1169-1180, 2008.

[18] M. Krapivin, A. Autayeu, M. Ma, E. Blanzieri and N. Segata, "Keyphrases extraction from scientific documents: improving machine learning approaches with natural language processing," in *International Conference on Asian Digital Libraries. Springer Berlin Heidelberg*, 2010.

[19] C. Caragea and F. Bulgarov, "Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach," in *Empirical Methods in Natural Language Processing (EMNLP)*, Doha, 2014.

[20] O. Alqaryouti, T. A. Farouk, A. R. Nabhan and K. Shaalan, "Graph-Based Keyword Extraction," in *Intelligent Natural Language Processing: Trends and Applications*, Springer, Cham, 2018, pp. 159-172.

[21] Z. Liu and P. Liu, "Clustering to Find Exemplar Terms for Keyphrase Extraction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009.

[۲۲]س. عربی نرئی, م. وحیدی اصل و ب. مینایی بیدگلی, "استخراج واژگان کلیدی جهت طبقه‌بندی متون فارسی," در *اولین کنفرانس داده‌کاوی ایران*, دانشگاه صنعتی امیرکبیر, ۱۳۸۶.

[22] S. Arabi Narei, M.Vahidi Asl and B.Minaei Bidgoli, "Keyword extraction for persian text classification,"in *First Iran Data Mining Conference*, Amir kabir university ,2007.

[۲۳] م. محمدی جنقرا و م. آنالویی, "استخراج واژگان کلیدی اسناد فارسی," در *سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران*, جزیره کیش - انجمن کامپیوتر, دانشگاه صنعتی شریف, ۱۳۸۶.

[23]M. Mohammadi Janghara and M.Analouei , " keyword extraction from persian documents", in *13th Annual Conference of Computer Society of Iran*, kish island- computer society, Sharif University of Technology, 2008.

[۲۴]ع. احمدی و ط. حسینی خواه, "استخراج واژگان کلیدی یک متن با استفاده از شبکه‌های عصبی," در *دهمین کنفرانس بین المللی مهندسی صنایع*, دانشگاه امیرکبیر, ۱۳۹۲.

[24] A. Ahmadi and T. Hoseinikhah, "Keyword Extraction from a text using Neural Network," in *Tenth international industrial engineering conference*, Amirkabir University, 2014.

[۲۵] ف. راد, ح. پروین, آ. دهباشی و ب. مینایی, "ارائه روشی جدید برای شاخص‌گذاری خودکار و استخراج واژگان کلیدی برای بازیابی اطلاعات و خوشه‌بندی متون," *نشریه پردازش علائم و داده‌ها*, جلد ۱۳, شماره ۱, صفحه ۸۷-۱۰۰, ۱۳۹۵.

[25]F. Rad, H. Parvin, A. Dehbashi, B. Minaei, "A New Method for Automatic Indexing and Extracting Keywords for Information Retrieval and Clustering of Texts", *Journal of Signal Processing and Data*, Volume 13, No. 1, page 87-100, 2017.

[26]H. G. Silber and K. F. McCoy, "Efficiently computed lexical chains as an intermediate representation for automatic text summarization," *Association for Computational Linguistics,* vol. 28, no. 4, pp. 487-496, 2002.

[27]M. Enss, "An investigation of word sense disambiguation for improving lexical chaining," M.S. thesis, Waterloo University, Waterloo, Canada, 2006.

[28]X. Li, "Keyphrase Extraction and Grouping Based on Association Rules," M.S. thesis, Guelph University, Guelph, Canada, 2014.

[29]B. Lott, "Survey of keyword extraction techniques," December, 2012.

[30]S. Beliga, "Keyword extraction: a review of methods and approaches," unpublished, 2014.

عطیه شریفی مدرک کارشناسی خود را در رشتهٔ مهندسی رایانه گرایش نرم‌افزار در سال ۱۳۹۱ از دانشگاه سمنان و کارشناسی ارشد خود را در همین رشته در سال ۱۳۹۵ از دانشگاه بین‌المللی امام خمینی (ره) دریافت کرد. زمینه‌های مورد علاقهٔ ایشان پردازش زبان طبیعی، داده‌کاوی و هستان‌شناسی است.

نشانی رایانامهٔ ایشان عبارت است از:

a.sharifi@edu.ikiu.ac.ir
sharifiatieh@gmail.com

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]

[DOI: 10.29252/jsdp.15.4.95]

محمد امین مهدوی استادیار گروه مهندسی رایانه دانشگاه بین‌المللی امام خمینی (ره) قزوین است. زمینه‌های پژوهشی مورد علاقهٔ ایشان متمرکز بر پردازش زبان فارسی، مدیریت دانش، بیوانفورماتیک و تحلیل گفتمان است. زمینهٔ اصلی پژوهش علمی ایشان روش‌های نمادین در تحلیل‌های مورفولوژیکی و نحوی فارسی و هم‌چنین تحلیل گفتمان در فارسی است.

نشانی رایانامهٔ ایشان عبارت است از:

mahdavi@eng.ikiu.ac.ir

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]

[DOI: 10.29252/jsdp.15.4.95]

[Downloaded from jsdp.rcisp.ac.ir on 2025-12-19]

[DOI: 10.29252/jsdp.15.4.95]