Ontology of Terms Present in the Titles of Scientific Articles with the Usage Relationship between Them

Soroush Mobasheri[1]

Abstract

Purpose: The question of what methods have been used by researchers to solve a given scientific problem is an important question for a researcher who intends to address that scientific problem. Important concepts and methods in any field of science are mentioned in the technical terms of that field. Therefore, the "usage" relationship between technical terms is of considerable importance. The purpose of this research is to establish a method that recognizes the usage relationship between terms present in the titles of scientific articles and to include those terms and the usage relationship between them in an ontology. Identifying and recording the usage relationship between terms, when implemented on a large scale, reveals what approaches are examined by researchers to address a given task.

Method: This research is applied in purpose and qualitative in nature. The primary data are the titles of articles in a selected scientific journal, a specified time period. We propose a method which detects technical terms present in the title of scientific articles, finds two relations between those terms, namely the hyponym-hypernym and usage relations, and inserts the result into an ontology. First, the noun phrases present in the title of the article are detected. These noun phrases are the technical terms, with some reservations. Our method focuses on titles in which two technical terms are related through a connector, which semantically indicates usage. Such cases are inserted into the ontology after detection. In addition to the usage relation, a hyponymy relation is also proposed based solely on the syntactic structure of each found noun phrase. Appropriate inference rules are designed in the ontology, based on which the usage relation between terms is inherited from the hyponym to the hypernym. Although the usage relationship between the article title components can also be extracted using large language models, performing this for a large number of article titles is costly. In addition, it will be necessary to analyze the response of the language models. Also, the need for inclusion in the ontology will remain.

Findings: To evaluate the method, the titles of one year of articles from a scientific journal were used. After implementing the method, the ontology was examined. In 69% of the titles which contained the pattern of this study, both noun phrase, i.e. the used term and the using term, were completely extracted. In 17% of the titles with the pattern, one or both noun phrase, i.e. at least one of the two terms, were incompletely extracted, so that the extracted part is semantically correct but does not contain the entire meaning of the title. In 14% of the titles with the pattern, at least one of the two noun phrase, i.e., the technical used term or the using term, was extracted incorrectly.

Conclusion: Since the proposed method is automatic, it can be applied to a large number of scientific articles. To increase accuracy, it is suggested that other connectives be considered in the analysis of the article title, in addition to the connectives that reflect the usage relationship.

Keywords

Term, Usage Relationship, Ontology, Information Extraction, Article Title, Hyponymy

Citation: Mobasheri, S. (2025). Ontology of Terms Present in the Titles of Scientific Articles with the Usage Relationship between Them. *Librarianship and Information Organization Studies*, 36(4), 219-250.

Doi: 10.30484/nastinfo.2025.3876.2351

Article Type: Research Article
Article history:
Received: 3 Aug. 2025
Revised: 1 Oct. 2025
Accepted: 17 Oct. 2025
Available online: 22 Dec. 2025

1. Ph.D., Department of Computer Engineering, NT.C., Islamic Azad University, Tehran, Iran;
soroush.mobasheri@iau.ac.ir

Publisher: National Library and Archives of I.R. of Iran
© The Author.

Introduction

In various scientific fields, there are problems and requirements that researchers try to solve or improve existing answers to. When a researcher wishes to become familiar with the research literature on a certain problem, one of their first questions would be what solutions researchers have tried to address that problem. It is clear that knowing the answer to this question will be helpful in the early steps to study the research literature on the subject.

This study is a step towards addressing this requirement. Specifically, we investigate the hypothesis that syntactic analysis of the titles of scientific articles can lead to the extraction of usage relationships between the scientific terms present in them.

Purpose

The purpose of this research is to establish a method that recognizes the usage relationship between terms present in the titles of scientific articles and to include those terms and the usage relationship between them in an ontology. Identifying and recording the usage relationship between terms, when implemented on a large scale, reveals which approaches researchers have examined to address a given task.

Method

This research is applied in purpose and qualitative in nature. The primary data are the titles of articles in a selected scientific journal, a specified time period. We propose a method which detects technical terms present in the title of scientific articles, finds two relations between those terms, namely the hyponym-hypernym and usage relations, and inserts the result into an ontology.

For any given title such as *x*, the noun phrases contained in x are found through syntactic analysis. Each of these noun phrases is a candidate for a scientific term with some caveats. Then, the conjunction between two consecutive noun phrases is considered. If the conjunction indicates a usage relationship between the two noun phrases before and after it, this relationship is extracted and noted along with those two noun phrases. For example, if in examining the title of the article *x*, eee ciiii aaiinn "*y* for *z*" ss see,, nncccc h *y* and *z* are two noun phrases, it is found that the term *y* is used in the term *z*.

In this case, this relationship is selected for inclusion in the ontology add eee eeeeeeeee eIn *x, y* is used for *z*" ss eeeeeeee nn eee llllll y. eee ceeeee cssss siiii eered nn ... s ddddy are "rrr", "ggggggg, "ii a" add "ggggggg".

Also, for each noun phrase consisting of several words, a hypernymy relationship is established and included in the ontology. For example, the noun phrase "fast Fourier transform" generalizes to "Fourier transform", and the latter generalizes to "transform". As arrrrrr tttt acce, eee wwo tttt tttt ttt ss "mnage eegaaaaa aå"" add "..... .. . aaaaa aå"" aee hynnny"" "" " "egaaaaa aå"".

In the ontology, there are inference rules that provide for the transmission of the usage through the hypernymy relation, from the specific term to the general term. Strictly, if *y* is used in *z*, *y* is generalized to y_1 and *z* to z_1, the inference rules provide that y_1 is used in *z*, *y* is used in z_1, and y_1 is used in z_1.

Findings

To evaluate the method, the titles of the articles in the year 2024 of the *International Journal of Machine Vision* – one of the leading journals in the field of Computer Vision – were used. The proposed algorithm was implemented in Python 3.13. An ontology was constructed in OWL 2 and viewed in Protégé 5.5.0. The HermiT 1.4.3.456 reasoning engine was used for reasoning.

After implementing the method, the constructed ontology was examined. In this evaluation, in 69% of the titles that contain the pattern of interest in this study, both noun phrases, i.e. the used term and the using term, were properly extracted. In 17% of the titles containing the pattern, one or both of the noun phrases, i.e. at least one of the two terms, were incompletely extracted, so that the extracted part is semantically correct but does not contain the entire meaning of the title. In 14% of the titles, at least one of the two noun phrases was incorrectly extracted. The error in extracting the usage relationship is mainly due to the presence of more than two noun phrases that are related by more than one connector. In other words, in relatively long titles with more than one connector, incomplete or unrelated noun aaaaee aa y be exaaac.... eee ooo ceeeee ciiees "a"""" add """""""""" """ee the highest frequency.

The ontology formed in this experiment has 504 terms, of which 108 terms are without generalization, meaning that they are not in a hyponymy relationship with any more general term. Recall that the ontology was formed based on the titles of the sample set under test. It is expected that if the method is applied to a larger set, a larger ontology will be obtained.

As an example, it was investigated which terms are employed by eee rrrm “eecogttt ”””””” ”ccrr ggggooeee rrr med llllll ll y? eee eesult is nnnnn nn Figrr e .. It can ee eeen tttt t recogttt ”””””” ”” eeeeeeee rrmm eee sycccciic geneaaiizaiinn ff 6 rrrrr rrr,,, hhhh hs “ooooo oocogttt ..... . ach rrrm nn eee gggt eeee ee hhe iigrr e ss dddd nn a hynnnym eemm ff “recggttt ,,,,, eecce eeey aee all used in “eecogttt bbbb bb rratttt iii ty ee eeee oyuuuu uuuuuh geneaaiizaii...

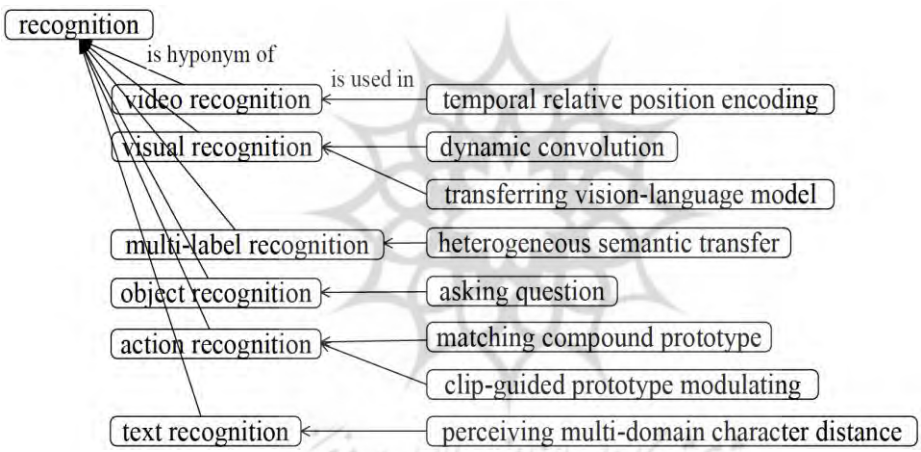


Figure 1. Terms used by hyponyms of “recognition”

Conclusion

The method proposed in this study can be extended to a large set of titles. The method has positive features for scalability. First, since the titles of research articles are freely available, collecting the input collection does not require any expense. Second, leading publishers have provided software interfaces for collecting meta-content. Therefore, collecting a large collection of article titles can be done mechanically.

To increase accuracy, it is suggested that connectives, other than those used in the present work, be considered in the analysis of article titles. Another suggestion would be to form a large corpus of scientific

article titles in the desired fields, which could greatly help in building ontologies on a real scale.

Acknowledgements

The author would like to thank anonymous referees for their constructive comments.

Ethical Considerations

The author avoided data fabrication, falsification, and plagiarism, and any form of misconduct.

Conflict of Interest

The author declares that there is no conflict of interest.

Declaration of AI Use

In the preparation of this article, the GPT-5 system (OpenAI, 2025) was used solely for experimental performance comparison with the proposed method. This system played no role in the writing of the article or in the generation of the research results. Full responsibility for the final content of the article rests with the author.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی

هستی‌نگاری اصطلاحات حاضر در عنوان مقاله‌های علمی با رابطهٔ کاربرد بین آن‌ها

سروش مبشری[1]

۱. دکتری، گروه مهندسی کامپیوتر، واحد تهران‌شمال، دانشگاه آزاد اسلامی، تهران، ایران؛
soroush.mobasheri@iau.ac.ir

چکیده

هدف: این پرسش که برای حل یک مسالهٔ معین علمی چه روش‌هایی توسط پژوهش‌گران به‌کار گرفته شده است، برای پژوهش‌گری که قصد دارد به آن مسالهٔ علمی بپردازد پرسشی مهم است. مفاهیم و روش‌های مهم در هر حوزه‌ای از علم، در اصطلاحات تخصصی آن حوزه ذکر می‌شوند. پس رابطهٔ «کاربرد» بین اصطلاحات تخصصی دارای اهمیتی شایان است. هدف این پژوهش تنظیم روشی است که رابطهٔ کاربرد را بین اصطلاحات حاضر در عنوان مقاله‌های علمی تشخیص دهد و آن اصطلاحات و رابطهٔ کاربرد بین آن‌ها را در یک هستی‌نگاری درج کند. تشخیص و ثبت رابطهٔ کاربرد بین اصطلاحات، هرگاه در مقیاسی کلان پیاده‌سازی شود، آشکار می‌سازد که پژوهش‌گران برای یک هدف مفروض، چه رویافت‌هایی را آزموده‌اند.

روش: این پژوهش به لحاظ هدف، کاربردی و از نظر ماهیت داده‌ها، کیفی است. داده‌های اولیه عبارتند از عنوان مقاله‌های یک نشریهٔ علمی خاص دلخواه، طی یک دورهٔ زمانی مشخص دلخواه. در این پژوهش روشی پیشنهاد می‌کنیم که اصطلاحات تخصصی حاضر در عنوان مقاله‌های علمی را تشخیص می‌دهد، دو رابطهٔ خاص-عام و کاربرد را بین آن اصطلاحات می‌یابد و نتیجه را در یک هستی‌نگاری درج می‌کند. نخست عبارت‌های اسمی حاضر در عنوان مقاله مشخص می‌شوند. این عبارت‌های اسمی با ملاحظاتی، اصطلاحات تخصصی هستند. این روش بر عنوان‌هایی متمرکز است که در آن‌ها دو اصطلاح تخصصی توسط یک رابط به هم مربوط شده‌اند و آن رابط، به لحاظ معنایی نشان‌دهندهٔ کاربرد است، به این معنی که یکی از آن اصطلاحات، در دیگری به‌کار رفته است. چنین مواردی پس از تشخیص، در هستی‌نگاری درج می‌شوند. افزون بر رابطهٔ کاربرد، رابطهٔ خاص-عام نیز تنها بر پایهٔ ساخت نحوی هر گروه اسمی یافته شده پیشنهاد می‌شود. افزون بر آن، قاعده‌های استنتاج مناسب در هستی‌نگاری طراحی شده است که بر پایهٔ آن‌ها رابطهٔ کاربرد بین اصطلاحات، از خاص به عام به ارث برده می‌شود. اگرچه رابطهٔ کاربرد بین اجزای عنوان مقاله را می‌توان از مدل‌های کلان زبانی نیز استخراج کرد، انجام این کار برای شماری کلان از عنوان‌های مقالات، مستلزم هزینه است. گذشته از این، لازم خواهد بود که پاسخ مدل‌های کلان زبانی نیز تحلیل شود و افزون بر آن، نیاز به درج در هستی‌نگاری به قوت خود باقی خواهد ماند.

یافته‌ها: برای ارزیابی روش، عنوان مقاله‌های یک سال از یک نشریهٔ علمی به‌کار رفته است. پس از اجرای روش، هستی‌نگاری تشکیل شده معاینه شده است. در این ارزیابی در ۶۹٪ از عنوان‌هایی که حاوی الگوی مورد نظر این پژوهش هستند، هر دو گروه اسمی یعنی اصطلاح به‌کاررفته و اصطلاح به‌کاربرنده به طور کامل استخراج شده‌اند. در ۱۷٪ عنوان‌های دارای الگو، یک یا هر دو گروه اسمی یعنی دست‌کم یکی از دو اصطلاح به‌کاررفته و به‌کاربرنده، ناقص استخراج شده‌اند به طوری که بخش استخراج شده از نظر معنایی صحیح است اما در بردارندهٔ همهٔ معنی عنوان نیست. در ۱۴٪ عنوان‌های دارای الگو، دست‌کم یکی از دو گروه اسمی یعنی اصطلاح تخصصی به‌کاررفته یا به‌کاربرنده نادرست استخراج شده است.

نتیجه‌گیری: از آن‌جا که روش پیشنهاد شده مکانیزه است، می‌توان آن را برای شماری کلان از مقاله‌های علمی به‌کار برد. برای افزایش دقت، پیشنهاد می‌شود حروف رابط دیگر افزون بر حروف رابطی که نسبت کاربرد را منعکس می‌سازند، در تحلیل عنوان مقاله در نظر گرفته شوند.

کلیدواژه‌ها

اصطلاح، رابطهٔ کاربرد، هستی‌نگاری، استخراج اطلاعات، عنوان مقاله، خاص-عام

استناد: مبشری، سروش (۱۴۰۴). هستی‌نگاری اصطلاحات حاضر در عنوان مقاله‌های علمی با رابطهٔ کاربرد بین آن‌ها. *مطالعات کتابداری و سازماندهی اطلاعات*، ۳۶(۴)، ۲۱۹-۲۵۰.

Doi: 10.30484/nastinfo.2025.3876.2351

فصلنامهٔ مطالعات کتابداری و سازماندهی اطلاعات، ۳۶ (٤)، زمستان ۱٤۰٤

نوع مقاله: پژوهشی

تاریخ دریافت: ۱٤۰٤/۰۵/۱۲

دریافت بازنگری: ۱٤۰٤/۰۷/۰۹

تاریخ پذیرش: ۱٤۰٤/۰۷/۲۵

تاریخ انتشار: ۱٤۰٤/۱۰/۰۱

ناشر: سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران

© نویسنده

مقدمه

در حوزه‌های گوناگون علم، مساله‌ها یا نیازمندی‌هایی مطرح است که پژوهش‌گران کوشش می‌کنند آن مساله‌ها را حل کنند یا

پاسخ‌های موجود را بهبود بخشند. تلاش برای حل هر مساله یا پاسخ به هر پرسش، امری ادامه‌دار است. چه‌بسا از پس راه‌حل‌هایی که تاکنون برای مساله‌ای معین شناخته شده است، راه‌حلی تازه ابداع شود یا راه‌حل‌های موجود بهبود یابند تا پاسخ مطلوب‌تری به‌دست آید.

هنگامی که شخص پژوهش‌گر قصد آشنایی با ادبیات پژوهشی مساله‌ای معین را داشته باشد یکی از نخستین پرسش‌های او این است که پژوهش‌گران برای حل آن مساله چه راه‌حل‌هایی را آزموده‌اند؟ روشن است که دانستن پاسخ این پرسش برای آغاز مطالعهٔ ادبیات پژوهشی موضوع، کمک‌کننده خواهد بود. پژوهش حاضر گامی است برای پاسخ به این پرسش.

معمولا یک مسالهٔ مورد نظر در یک دامنهٔ علمی، با اصطلاحی شناخته‌شده خوانده می‌شود. برای نمونه چند مساله از مسائل حوزهٔ «بینایی ماشین» عبارتند از «رده‌بندی تصویر»، «قطعه‌بندی تصویر» و «تشخیص صورت». روش‌ها و راه‌حل‌ها معمولا اسم مشخص ندارند مگر این که برجسته و ممتاز باشند. روش‌های دیگر عملا با کمک اجزای ممتاز خود نامیده می‌شوند. مثلا ممکن است در یک الگوریتم، از «تبدیل سریع فوریه» برای حل مسالهٔ «قطعه‌بندی تصویر» استفاده شده باشد. چنین پژوهشی را می‌توان در مقاله‌ای با عنوان «قطعه‌بندی تصویر به کمک تبدیل سریع فوریه» گزارش کرد. البته چنین عنوانی بازتاب دهندهٔ

تنها یک جنبه از الگوریتم پیشنهاد شده خواهد بود. اما این جنبهٔ اعلام شده، از دید پژوهش‌گر، شایستگی آن را داشته است که نشان اصلی الگوریتم ساختهٔ او باشد.

بررسی عنوان مقاله‌های علمی، دست‌کم در برخی از حوزه‌ها نشان می‌دهد که عنوان‌هایی از آن دست که در بالا بدان اشاره شد، از الگویی نسبتا رایج پیروی می‌کنند. لذا با بررسی خودکار چنین عنوان‌هایی می‌توان رابطهٔ کاربری یک روش یا حوزه را در حل یک مساله استخراج کرد.

روشن است که شرح کامل هر راه‌حلی را باید در بدنهٔ گزارش علمی جست‌وجو کرد که البته خودکارسازی آن کاری بسیار دشوار است. اما جان‌مایهٔ روش را معمولا در عنوان مقاله می‌توان یافت. عنوان مقاله‌های علمی، گنجینهٔ اصطلاحات تخصصی است. عنوان، تقریبا همیشه با دقت زیاد تنظیم می‌شود و در ساختمان آن عبارت‌هایی دقیق به‌کار می‌رود. تقریبا عنوان همهٔ مقاله‌های علمی حاوی عبارت‌هایی است که عمیقا و در حد امکان به‌روشنی موضوع مقاله را اعلام می‌کنند. از این رو توجه به عنوان مقاله‌ها برای استخراج اصطلاحات تخصصی، کاری امیدبخش است.

حجم بسیار زیاد اطلاعات علمی، ایجاد هر گونه ساختار در درون این پیکرهٔ بزرگ را با اهمیت می‌سازد. هدف‌ها و روش‌ها، دو گروه از نهادهای هر حوزه‌ای از علم است (مبشری، ۱۴۰۳). لذا دانستن این که برای نزدیک شدن به هر یک از هدف‌های یک حوزه از علم، چه روش‌هایی توسط محققان به‌کار گرفته شده است سودمند خواهد بود. پژوهش حاضر کوششی است برای آماده ساختن این امکان. هدف این است که بتوان به پرسش‌هایی از این دست که «برای انجام یک عمل معین، چه روش‌هایی توسط محققان آزموده شده است؟» پاسخ داد. روشن است که پاسخ کامل به چنین پرسشی مستلزم تحلیل پیکرهٔ عظیم علم مکتوب است. با این همه در پژوهش حاضر به این موضوع پرداخته شده است که با مراجعه به تنها عنوان مقاله‌های علمی، چه پاسخی می‌توان به پرسش بالا داد؟

اگرچه پرسش بالا را می‌توان از سیستم‌های هوشمند مجهز به مدل‌های کلان زبانی نیز پرسید. با این همه پاسخی که از چنین سیستم هوشمند دریافت می‌شود معمولا مبتنی بر دلایلی است که آن دلایل در دسترس کاربر قرار ندارد. پس ضمن مفید بودن، برای کاربر مشخص نیست که پاسخ دریافت‌شده با تکیه بر چه پیکره‌ای از فرهنگ مکتوب علمی به دست آمده

است؟ یک مبنای مشخص استنتاجی مانند آن‌چه که از روش پیشنهادی ما به دست می‌آید، می‌تواند به‌کاربران برای رسیدن به نتیجهٔ مطلوب کمک کند.

پیشینهٔ پژوهش

استخراج اطلاعات[1] از متن، حوزه‌ای از پردازش زبان طبیعی است برای یافتن خودکار ساختارهای موجود در متن مکتوب. در این حوزه که دست‌کم به دههٔ ۱۹۷۰ میلادی بازمی‌گردد (Cowie & Lehnert, 1996) از جمله کوشش می‌شود که نهادهای حاضر در متن و رابطه‌های ذکر شده بین آن‌ها تشخیص داده شود. نهادها اساسا گروه‌های اسمی هستند. رابطهٔ بین نهادها توسط فعل جمله یا به اصطلاح منطق، با محمول جمله اعلام می‌شود.

در استخراج اطلاعات باز، همان هدف دنبال می‌شود و افزون بر آن، هیچ پیش‌فرضی روی مجموعهٔ واژگان یا مجموعهٔ رابطه‌های موجود در نظر گرفته نمی‌شود. این آزادی در طرح مساله، قابلیت توسعهٔ نامحدود روش‌ها را به ارمغان می‌آورد. اما پیچیدگی‌های زبان طبیعی، کار استخراج اطلاعات باز را بغرنج می‌سازد. فیدر[2] و همکاران (۲۰۱۱) به دشواری ناشی از رابطه‌هایی پرداخته‌اند که در زبان انگلیسی متناظر با یک فعل تک‌واژه‌ای نیستند و در نتیجه، حضور آن رابطه‌ها را نمی‌توان از یک واژهٔ تنها تشخیص داد.

شریف (۱۳۸۸) به استخراج روابط معنایی از متن فارسی و میزان پیدایی آن‌ها پرداخته است. در آن پژوهش مشخص شد که نیمی از روابط به صورت کاملا تلویحی برقرار هستند و هیچ نشانهٔ صریحی در متن ندارند. در نتیجه تشخیص خودکار روابط معنایی موجود در متن، تنها با تکیه بر نشانهٔ صریح، رضایت‌بخش نیست.

شمس‌فرد و جعفری‌نژاد (۱۳۹۱) با سه روش، (۱) مبتنی بر ریخت‌شناسی و تحلیل لغوی، (۲) مبتنی بر تعمیم با استفاده از پایگاه‌دادهٔ ساختار وابستگی افعال و (۳) برچسب‌زنی نقش‌های معنایی، به یافتن روابط معنایی میان فعل و اجزای دیگر جمله در متون فارسی پرداخته‌اند.

عنوان مقاله‌های علمی عمدتا جملهٔ کامل نبوده، حاوی هیچ فعلی نیستند. در عوض،

1 Information Extraction (IE)
2 Fader

عنوان‌ها معمولا عبارتند از چند گروه اسمی که با رابط‌هایی به هم مربوط شده‌اند. هر یک از این رابط‌ها به شکل یک *حرف اضافه* یا *حرف ربط*، رابطهٔ بین دو گروه اسمی پیش و پس از خود را بیان می‌کند. لذا ساختمان نحوی عنوان‌ها به طور کلی از ساختمان نحوی جمله‌های زبان ساده‌تر است.

در خصوص عبارت‌های تخصصی استخراج شده از عنوان مقاله‌ها، می‌توان پرسید که آیا این اصطلاحات، لزوما کلیدواژه هستند و اگر بله، آیا از مجموعه‌ای بسته و کنترل‌شده از کلیدواژه‌ها می‌آیند؟

این که آیا امکان‌پذیر است که مجموعهٔ کلیدواژه‌ها در هر حوزه‌ای از علم، مجموعه‌ای کنترل شده و بسته باشد مورد تردید است. در درزی خلردی و رضوی (۱۳۹۷) نشان داده شده است که در مجموعهٔ مورد مطالعهٔ آن پژوهش، تنها کم‌تر از نیمی از کلیدواژه‌ها با مجموعهٔ کنترل شده مطابقت داشته‌اند. با این همه روشن است که تدوین، آگاهی‌رسانی و رعایت چنین مجموعه‌های کنترل‌شده‌ای از کلیدواژه‌ها می‌تواند گامی موثر در راستای ساختارمند کردن علم مکتوب باشد (غنی پور تفرشی و حاجی زین العابدینی، ۱۳۹۷). استخراج، تنظیم و ساختار دادن به مجموعهٔ کلیدواژه‌ها در حوزه‌های گوناگون علمی می‌تواند در میان‌مدت به تشکیل چنین مجموعه‌های کنترل شده و رعایت بیش‌تر آن‌ها در بین پژوهش‌گران کمک کند.

بهاری ورزنه (۱۴۰۱) نشان داده است که برخلاف ادعای موتورهای جست‌وجوی معنایی، جست‌وجو بر پایهٔ کلیدواژه‌ها دست‌کم به اندازهٔ جست‌وجو بر اساس عبارت معنایی موفق است. این بررسی مجددا اهمیت کلیدواژه‌ها را یادآوری می‌کند. هدایتی و حسنی آهنگر (۱۴۰۰) روش‌های مهم استخراج کلیدواژه‌ها را از متن تشریح کرده‌اند. آن‌ها به عبارت‌های اسمی برای تشخیص کلیدواژه‌ها توجه نموده‌اند.

در متن مقاله‌ها، مهم‌ترین ویژگی قابل استفاده در تشخیص کلیدواژه‌ها بسامد زیاد آن‌ها است (Nomoto, 2022)، (Xu & Zhang, 2021)، (Ding & Luo, 2021). علاوه بر متن، توجه به قسمت‌هایی ویژه از اثر مکتوب می‌تواند یافتن کلیدواژه‌ها را تسهیل کند. از جمله در ریاحی‌نیا و همکاران (۲۰۲۲) تشخیص کلیدواژه بر پایهٔ فهرست مطالب کتاب‌های درسی است. در اسماعیلی آدرگانی و فراهی (۱۳۹۲) از بخش‌های مهم مقاله‌های علمی مانند چکیده و بخش نتیجه‌گیری

برای استخراج کلیدواژه‌ها استفاده شده است.

پس تشخیص کلیدواژه‌ها از متن عمدتا مربوط به بسامد زیاد، حضور در بخش‌های مهم متن مانند چکیده یا بخش نتیجه‌گیری، ماهیت نحوی پاره‌های سخن مانند گروه اسمی، یا شباهت با عبارت‌های عضو مجموعه‌های کنترل شده است.

روش پژوهش

در میان اجزای مختلف مقاله‌های علمی، عنوان مقاله ویژگی‌هایی دارد که تحلیل زبانی آن را با اهمیت می‌سازد.

عنوان مقاله رایگان در دسترس است. افزون بر این، برخی از ناشران عمده واسطهای نرم‌افزاری در اختیار می‌گذارند که به کمک آن‌ها می‌توان فهرست عنوان‌ها را گرد آورد. در نتیجه امکان انجام عملیات روی پیکرهٔ بزرگ عنوان‌ها عملی است.

≠ عنوان بی‌گمان دربردارندهٔ اصطلاحات تخصصی است.

≠ عنوان به لحاظ نحوی ساختاری ساده‌تر از جمله دارد زیرا بدون فعل است. درواقع عنوان ریسه‌ای از گروه‌های اسمی است که با رابطهایی به هم مربوط شده‌اند. اگرچه همین رابطها نیز ممکن است ساختارهای پیچیده‌ای را بسازند، نبود فعل، از پیچیدگی مضاعف ساختار عنوان جلوگیری می‌کند.

بنابراین، در پژوهش حاضر عنوان مقاله‌های علمی مورد بررسی قرار گرفت. راهبرد پژوهش کنونی این است که با استفاده از تحلیل نحوی عنوان مقاله‌های علمی، رابطهٔ کاربرد بین دو اصطلاح تخصصی حاضر در عنوان استخراج شود. با این کار، قصد تشخیص کلیدواژه‌ها را نداریم. با این همه روشن است که عبارت اسمی حاضر در عنوان مقالهٔ علمی در بازتاب کردن موضوع مقاله حایز اهمیت است. بنابراین ضمن این که این پژوهش برای استخراج کلیدواژه‌ها هدف‌گذاری نشده است، احتمال این که اصطلاحات به‌دست آمده عملا کلیدواژه باشند زیاد است. به‌هرروی، مقایسه‌ای بین این اصطلاحات با مجموعه‌ای کنترل شده از کلیدواژه‌ها صورت نمی‌گیرد.

به کمک روش پیشنهادی که روشی کیفی است، می‌توان از درون عنوان مقاله‌های علمی، اصطلاحات تخصصی و رابطه‌های خاص-عام و کاربرد را بین آن‌ها تشخیص داد و به طور

خودکار در هستی‌نگاری درج کرد. اطلاعات به دست آمده از نوع «برای انجام عمل x در پژوهش y از روش z استفاده می‌شود» می‌باشد که در آن x و z دو اصطلاح تخصصی هستند و y عنوان یک مقالهٔ علمی است. روش پیشنهاد شده را می‌توان برای پیکره‌ای عظیم، متشکل از عنوان مقاله‌های علمی به‌کار برد و گزاره‌هایی پرشمار از این دست را نتیجه گرفت.

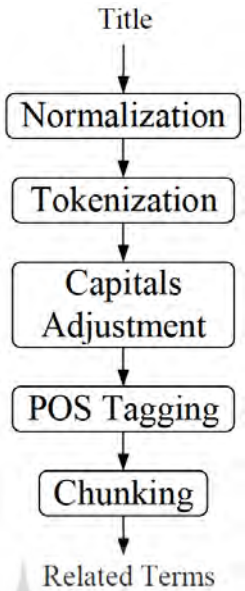
منظور از «برای انجام عمل x از روش z استفاده می‌شود» این است که عمل x یکی از هدف‌های شناخته شده در یک دامنهٔ علمی است و پژوهش‌گران برای نیل به این هدف یا به‌بود روش‌های نیل به آن، از روش z استفاده کرده‌اند. در نتیجه اصطلاح تخصصی x یک هدف و اصطلاح تخصصی z یک روش یا ابزار است. دیگر روابط مانند تاثیر یا تخصیص مورد نظر نیست. برای نمونه در عنوان Fast Fourier Transform for Image Segmentation بین دو گروه اسمی (مشخص شده با زیرخط رابطهٔ کاربرد برقرار است. اما در عنوان A New Model for Medical Imagary بین دو گروه اسمی رابطهٔ کاربرد برقرار نبوده، در عوض رابطهٔ تخصیص وجود دارد که مورد نظر ما نیست.

به طور خلاصه روش پیشنهادی عبارت است از تفکیک عنوان مقاله به گروه‌های اسمی و توجه به برخی از رابط‌های بین آن‌ها. کم‌وبیش هر گروه اسمی به عنوان یک اصطلاح تخصصی در نظر گرفته می‌شود.

گروه‌های اسمی توسط رابط‌هایی به هم مربوط می‌شوند. برخی از این رابط‌ها معنایی مشخص دارند. برای نمونه رابط for نشان می‌دهد که عبارت اسمی پیش از آن، به منظور کمک به عبارت اسمی پس از آن به‌کار رفته است. در روش پیشنهادی ما این رابطه بین عبارت‌های تخصصی آشکار و در هستی‌نگاری درج می‌شود. گام‌های اصلی این روش به طور خلاصه چنین هستند (شکل ۱):

- ≠ پیش‌پردازش، توکن‌سازی و تنظیم حروف بزرگ،
- ≠ برچسب‌زدن،
- ≠ قطعه‌بندی و استخراج اصطلاحات تخصصی و رابطهٔ کاربرد بین آن‌ها،
- ≠ درج اصطلاحات تخصصی و رابطهٔ بین آن‌ها در هستی‌نگاری.

این گام‌ها به‌ترتیب، در بخش‌های پیش رو تشریح می‌شود.



شکل ۱. گام‌های کلان روش پیشنهادی

پیش‌پردازش، توکن‌سازی و تنظیم حروف بزرگ

ورودی روش، مجموعه‌ای از عنوان مقاله‌های علمی است. پس از پیش‌پردازش و توکن‌سازی، لازم است بزرگ یا کوچک بودن حرف اول توکن‌ها تنظیم شود. در بسیاری از نشریه‌های علمی، واژه‌های عنوان با حرف بزرگ آغاز می‌شود. هم‌چنین در تمامی نشریه‌ها نخستین واژهٔ عنوان با حرف بزرگ آغاز می‌شود. این آرایش ممکن است موجب شود که الگوریتمی که کار برچسب زدن به پاره‌های سخن را انجام می‌دهد[1]، به‌خطا، آن واژه‌ها را اسم خاص در نظر بگیرد. پس لازم است که حرف اول واژه‌های حاضر در عنوان، به حرف کوچک تبدیل شوند.

با این همه در دو دسته از توکن‌ها، بهتر است از تبدیل حرف بزرگ به کوچک پرهیز شود. یک دسته واژگانی هستند که درواقع اسم خاص می‌باشند. مانند نام دانشمندان. برای نمونه Fourier در اصطلاح Fourier transform اسم خاص است و اگر با حرف کوچک آغاز شود، ممکن است موتور برچسب‌زن، نوع آن را به‌درستی نشناسد. دستهٔ دوم از توکن‌هایی که بهتر است حروف آن‌ها بدون تغییر باقی بمانند، علامت‌های اختصاری هستند که معمولا با

1 Part of speech (PoS) tagger

حروف تمام‌بزرگ تنظیم می‌شوند.

برچسب‌زدن

پس از تنظیم حروف بزرگ، فهرست توکن‌ها که حاصل کار توکنایزر روی عنوان است، به برچسب‌زن سپرده می‌شود. نتیجهٔ آن فهرست دوگانه‌های توکن-برچسب است.

نمونهٔ ۱. عنوان زیر را در نظر می‌گیریم:

AO Qjj cct-vvvll Imgge Rrrr ssttt tt iff f rr High-vvvll Viaaal Rccggii tinn

این عنوان پس از پیش‌پردازش، توکنیاز شدن و برچسب‌گذاری به فهرست دوگانه‌های توکن-برچسب زیر می‌رسد:

[(an, DT), (object-level, JJ), (image, NN), (representation, NN), (for, IN), (high-level, JJ), (visual, JJ), (recognition, NN)]

برخی از حروف اضافه از قبیل for دلیلی هستند بر حضور یک رابطهٔ معین بین دو اصطلاح تخصصی. از این رو لازم است برچسب این حروف اضافه، متمایز از حروف اضافهٔ دیگر باشد. پس فهرست توکن-برچسب به دست آمده مرور می‌شود و برچسب حروف اضافهٔ مورد نظر، با برچسب خاص جایگزین می‌گردد. در پژوهش حاضر، به حروف اضافهٔ for، using، via و through برچسب‌های خاص داده شده است.

نمونهٔ ۲. در فهرست توکن-برچسب‌های نمونهٔ ۱، برچسب حرف اضافهٔ for از IN به FOR تبدیل می‌شود.

نتیجه در ذیل نشان داده شده است. در فهرست جدید، برچسب توکن پنجم تبدیل شده است:

[(an, DT), (object-level, JJ), (image, NN), (representation, NN), (for, FOR), (high-level, JJ), (visual, JJ), (recognition, NN)]

قطعه‌بندی و استخراج اصطلاحات تخصصی و رابطهٔ کاربرد بین آن‌ها

در گام بعد، فهرست توکن‌های برچسب زده، قطعه‌بندی F۱[1] می‌شود. خردترین قطعه‌ای که برای منظور ما اهمیت دارد، گروه اسمی F۲[2] است. گروه اسمی با عبارت منتظم

1 Chunk
2 Nominal

<JJ|RB>*<NN|NNP>+ توصیف شده است و متشکل از تعداد دل‌خواه قید یا صفت و به‌دنبال آن دست‌کم یک نام عام یا نام خاص است. یادآوری می‌کنیم که گروه اسمی برخلاف عبارت اسمی F۳[1] با ضمیر یا حرف تعریف آغاز نمی‌شود.

باری. از یک گروه اسمی متشکل از n توکن، n اصطلاح تخصصی استخراج می‌شود. نخستین اصطلاح، از توکن پایانی گروه اسمی تشکیل می‌شود. دومین اصطلاح، از ۲ توکن پایانی گروه اسمی تشکیل می‌شود. به همین ترتیب تا n-امین اصطلاح که از همهٔ n توکن گروه اسمی تشکیل می‌شود. در چنین رشته‌ای از اصطلاحات تخصصی، هر اصطلاح، خاص‌تر از اصطلاح پیش از خود است.

به بیان صوری، $[c_1,,, \ c_n]$ یک گروه اسمی باشد. از این گروه اسمی، اصطلاحات ذیل استخراج می‌شود:

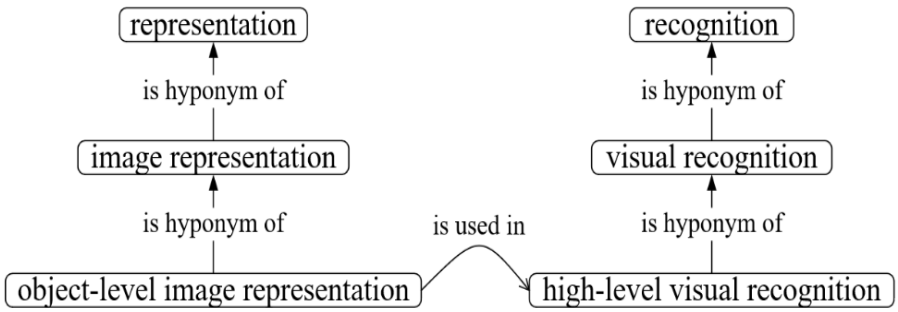
$$t_i = [c_i, \ldots, c_n], \quad 1 \le i \le n$$

هم‌چنین رابطهٔ خاص–عام، بین این اصطلاحات به صورت ذیل اعلام می‌شود:

$$\text{hyponymy} = \{(t_i, t_{i-1}) : 2 \le i \le n\}$$

نمونهٔ ۳. عنوان یادشده در نمونه‌های ۱ و ۲ حاوی دو گروه اسمی است که عبارتند از object-level image representation و high-level visual recognition. از گروه اسمی نخست که شامل ۳ توکن است ۳ اصطلاح representation، image representation و object-level image representation استخراج می‌شود. از گروه اسمی دوم نیز که شامل ۳ توکن است ۳ اصطلاح recognition، visual recognition و high-level visual recognition استخراج می‌شود. هم‌چنین چنان‌که گفته شد رابطهٔ خاص–عام بین اصطلاحات متوالی هر رشته برقرار می‌شود (شکل ۲).

1 Noun phrase



شکل ۲. اصطلاحات نمونهٔ ۳

افزون بر رابطهٔ خاص–عام، از برخی از حروف ربط بین گروه‌های اسمی می‌توان رابطهٔ کاربرد را نتیجه گرفت. در پژوهش حاضر رابطهٔ کاربرد بین گروه‌های اسمی که با یکی از حروف اضافهٔ for، using، via و through به یکدیگر مربوط شده‌اند استخراج شده است. معنی هر یک از این الگوهای نحوی در جدول ۱ آمده است.

جدول ۱. الگوهای نحوی به‌کار رفته در پژوهش حاضر و معنی آن‌ها

الگوی نحوی	معنی
x for *y*	*x* در *y* به‌کار رفته است.
x using *y*	*y* در *x* به‌کار رفته است.
x via *y*	*y* در *x* به‌کار رفته است.
x through *y*	*y* در *x* به‌کار رفته است.

برای یافتن الگوهای نحوی جدول ۱ لازم است عبارت‌های منتظم آن الگوها برای الگوریتم قطعه‌بندی به‌کار رود. برای نمونه الگوی نحوی *x* for *y* با عبارت منتظم

JJ J|RBNNNNNNNNOOOOOOOOJJ J|RBNNNNNNNN+++

مطابقت داده می‌شود. این عبارت منتظم شامل دو نسخه از عبارت منتظم گروه اسمی می‌باشد همراه با حرف ربط <FOR> بین آن‌ها.

نمونهٔ ٤. در عنوان نمونهٔ ۳، الگوی *x* for *y* حاضر است. حضور این الگو به صورت یک رابطهٔ کاربرد بین دو اصطلاح تخصصی تفسیر می‌شود (رابطهٔ is used in در شکل ۱).

رابطهٔ کاربرد بین اصطلاحات استخراج شده از دو گروه اسمی، بین خاص‌ترین اصطلاحات برقرار می‌شود. ماهیت این رابطه اقتضا می‌کند که از خاص به عام، به ارث برسد.

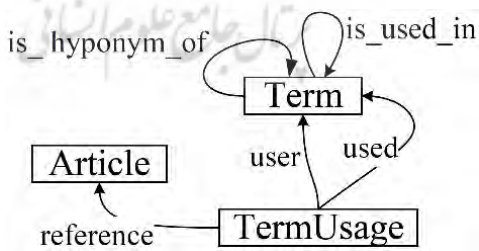
یعنی اگر x در y به‌کار رود و x_1 و y_1 به‌ترتیب تعمیم x و y باشند، در این صورت هم‌چنین x در y_1، x_1 در y و x_1 در y_1 نیز به‌کار می‌رود. برای تامین این ارث‌بری قاعده‌های استنتاج در هستی‌نگاری طراحی شده‌اند که در بخش آینده تشریح خواهد شد.

اگر یک حرف ربط polysemous باشد، یعنی در کاربردهای گوناگون به معنی‌های گوناگون به کار رود، این مساله می‌تواند روی روش پیشنهادی ما تاثیر منفی بگذارد. برای نمونه حرف ربط for ممکن است به معنی کاربرد باشد مانند Fourier transform for image segmentation و نیز به معنی تخصیص به کار رود مانند A quantitative model for vehicle traffic.

درج در هستی‌نگاری

در این بخش هستی‌نگاری میزبان اصطلاحات و رابطه‌های استخراج شده را تشریح می‌کنیم. شکل ۳ کلاس‌های هستی‌نگاری را نشان می‌دهد. هر اصطلاح استخراج شده به عنوان یک نمونه از کلاس Term درج می‌شود. رابطهٔ is hyponym of بین دو اصطلاح، رابطهٔ خاص-عام است که پیش از این به آن پرداختیم.

برای رابطهٔ کاربرد، افزون بر اصطلاح «به‌کار رفته» و اصطلاح «به‌کار برنده» نهاد دیگری به عنوان مرجع در نظر گرفته شده است که در آن مقاله‌ای که این کاربرد را منعکس کرده است درج شود. به این ترتیب، کاربرد یک اصطلاح در یک اصطلاح دیگر، یک رابطه با ۳ نهاد است: اصطلاح به‌کاررفته، اصطلاح به‌کاربرنده و مقالهٔ مرجع. بدین سبب برای بازنمایی آن، به کلاس مستقلی نیاز است (کلاس TermUsage در شکل ۳).



شکل ۳. کلاس‌ها و رابطه‌های هستی‌نگاری برای درج اصطلاحات تخصصی و رابطهٔ کاربرد بین آن‌ها

افزون بر کلاس‌ها، برای پاسخ دادن به این پرسش که کدام روش‌ها برای حل یک مسالهٔ معین

آزموده شده است به قاعده‌های استنتاج نیاز است. این قاعده‌های استنتاج در شکل ۴ آمده است. قاعدهٔ استنتاج نخست از دو رابطهٔ used و user در یک فرد از کلاس TermUsage نتیجه می‌گیرد که یک اصطلاح در یک اصطلاح دیگر به‌کار رفته است. به عبارت دیگر از TermUsageی مانند x، اصطلاح y را به‌عنوان «به‌کاررفته» و اصلاح z را به‌عنوان «به‌کاربرنده» اعلام کند، نتیجه گرفته می‌شود که y در z به کار رفته است. قاعده‌های استنتاج دوم و سوم سرایت کاربرد را از راه رابطهٔ خاص–عام، از اصطلاح خاص به اصطلاح عام تامین می‌کنند. هم‌چنین رابطهٔ uses به‌عنوان وارون رابطهٔ is_used_in بین اصطلاحات تعریف شده است که برای اختصار، در شکل ۳ نیامده است. کلاس‌های نشان داده شده در شکل ۳ تنها کلاس‌های هستی‌نگاری اصطلاحات هستند. هر اصطلاحی که از عنوان مقالهٔ پژوهشی استخراج می‌شود به عنوان یک نمونه یا فرد از کلاس Term درج می‌شود. هم‌چنین مشخصات مقاله نیز به عنوان یک نمونه یا فرد از کلاس Article درج می‌شود.

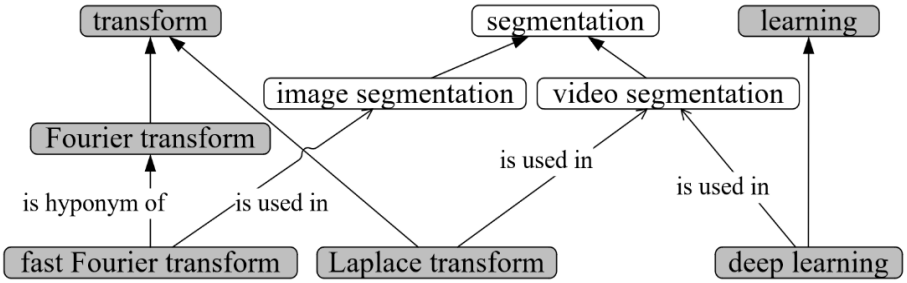
```
S1: TermUsage(?x) ^
    termUsed(?x, ?y) ^
    termUser(?x, ?z) -> isUsedIn(?y, ?z)

S2: isUsedIn(?x, ?y) ^
    isHyponymOf(?y, ?z) -> isUsedIn(?x, ?z)

S3: isUsedIn(?x, ?y) ^
    isHyponymOf(?x, ?z) -> isUsedIn(?z, ?y)
```

شکل ٤ . قاعده‌های استنتاج دربارهٔ کاربرد یک اصطلاح تخصصی در یک اصطلاح تخصصی دیگر

نمونهٔ ۵. مجموعهٔ متشکل از ۳ عنوان فرضی زیر را در نظر می‌گیریم:

1. Fast Fourier Transform for Image Segmentation
2. Laplace Transform for Video Segmentation
3. Deep Learning for Video Segmentation

با اجرای الگوریتم پیشنهادی روی این مجموعه، اصطلاحات و رابطه‌های نشان داده‌شده در شکل ۵ استخراج و در هستی‌نگاری درج می‌شوند.



شکل ۵ . اصطلاحات و رابطه‌های استخراج شده از نمونهٔ ۵

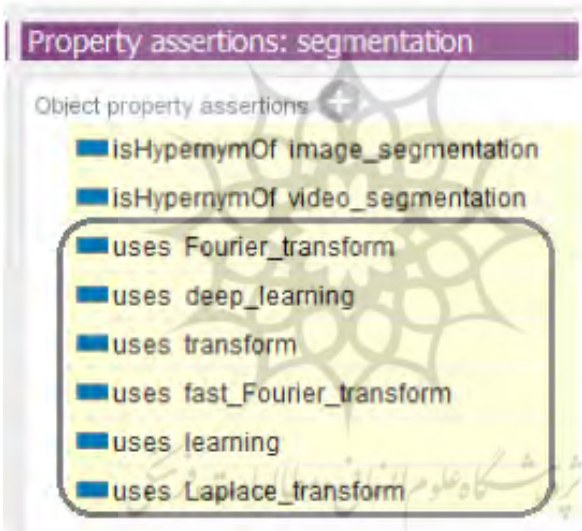
همان‌طور که ملاحظه می‌شود، fast Fourier transform رابطهٔ خاص-عام با اصطلاح کلی‌تر Fourier transform و این اصطلاح نیز رابطهٔ خاص-عام با اصطلاح کلی‌تر transform دارد. به همین ترتیب اصطلاحات خاص دیگر مانند Laplace transform و deep learning. هم‌چنین ملاحظه می‌شود که دو اصطلاح Fourier transform و Laplace transform در رابطهٔ خاص-عام با یک اصطلاح کلی‌تر مشترک یعنی transform هستند. از این نمونه روشن می‌شود که همان‌طور که مطلوب است، مفاهیم و اصطلاحات، با روش پیشنهادی ما درخت‌واره تشکیل می‌دهند.

از هر یک از ۳ عنوان مورد بررسی یک رابطهٔ کاربرد استخراج می‌شود. از عنوان نخست، رابطهٔ کاربرد بین fast Fourier transform و image segmentation درج می‌شود. از عنوان دوم رابطهٔ کاربرد بین Laplace transform و video segmentation درج می‌شود. از عنوان سوم رابطهٔ کاربرد بین deep learning و video segmentation درج می‌شود. این ۳ رابطه در شکل ۵ صریحا با لبه نشان داده شده است.

اکنون با راه‌اندازی موتور استنتاج روی هستی‌نگاری تشکیل شده، قاعده‌های استنتاجی که در بالا ذکر شدند نتیجه می‌دهند که رابطهٔ is used in هم از گره به‌کار رفته و هم از گره به‌کار برنده، روی درخت‌واره به سوی ریشه (به سوی بالا) سرایت می‌کند. در نتیجه نه تنها مثلا fast Fourier transform در image segmentation کاربرد دارد، هم‌چنین Fourier transform و transform نیز در image segmentation کاربرد دارند و افزون بر این‌ها این هر ۳ اصطلاح، هم‌چنین در segmentation نیز کاربرد دارند. به همین ترتیب، نه تنها Laplace transform در video segmentation کاربرد دارد، هم‌چنین transform نیز در

video segmentation کاربرد دارد و افزون بر این‌ها ای هر ۲ اصطلاح، در segmentation نیز کاربرد دارند. باز به همین ترتیب، هم deep learning و هم learning در video segmentation و هم‌چنین در segmentation کاربرد دارند.

با توجه به این توضیحات، پاسخ این هستی‌نگاری به این پرسش که «چه مفاهیم و روش‌هایی در segmentation کاربرد دارند؟» عبارت خواهد بود از مجموعهٔ اصطلاحاتی که در شکل ۵ با رنگ خاکستری نشان داده شده‌اند. همان‌گونه که در شکل ۶ مشاهده می‌شود، هستی‌نگاری پس از به‌کار افتادن موتور استدلال، به درستی همهٔ این ۶ اصطلاح را در پاسخ به پرسش مذکور نام می‌برد. یادآوری می‌شود که رابطهٔ uses به عنوان وارون رابطهٔ is used in تعریف شده است.



شکل ٦ . رابطه‌های کاربرد، استنتاج شده برای اصطلاح segmentation در نمونهٔ ۵

یافته‌ها

در این بخش تجربهٔ انجام شده معرفی می‌شود. نخست، نتیجهٔ آزمودن روش پیشنهادی بر مجموعه‌ای از عنوان‌های مقاله‌های پژوهشی عرضه می‌گردد. در ادامه، روش پیشنهادی برای استخراج اطلاعات با روش مبتنی بر مدل‌های بزرگ زبانی جایگزین و یافته‌ها مقایسه می‌شود.

تجربه با روش پیشنهادی

الگوریتم پیشنهادی در زبان Python 3.13 پیاده‌سازی شده است. هستی‌نگاری به زبان OWL 2 تشکیل و در محیط Protégé 5.5.0 مشاهده شده است. موتور استدلال HermiT 1.4.3.456 برای انجام استدلال به‌کار گرفته شده است.

روش پیشنهادی را می‌توان به‌طور بالقوه روی پیکره‌ای بزرگ از عنوان مقاله‌های پژوهشی پیاده کرد. با این همه در پژوهش حاضر به عنوان نمونه، برای آزمایش و ارزیابی روش، نشریهٔ علمی International Journal of Computer Vision[1] که یکی از نشریه‌های شاخص در حوزهٔ بینایی ماشین است در نظر گرفته شده است. دلیل انتخاب حوزهٔ بینایی ماشین این است که نگارنده با این دامنه آشنا است و ارزیابی عبارت‌های استخراج شده به عنوان اصطلاحات تخصصی برای وی امکان‌پذیر بوده است. عنوان مقاله‌های سال ۲۰۲۴ این نشریه همراه با مشخصات اصلی ماخذ مانند نام و شمارهٔ نشریه و سال انتشار، در یک جدول گردآوری شدند[2]. این جدول، پیکرهٔ حاوی اطلاعات خام به‌کار رفته در روش پیشنهادی است و قابل گسترش است. به این معنی که تنها با افزودن عنوان مقاله‌های دیگر به این جدول، می‌توان پیکرهٔ به‌کار رفته در روش پیشنهادی را گسترش داد.

مجموعهٔ یاد شده پس از حذف یادداشت‌های اصلاح و یادداشت‌های سردبیر، شامل ۲۸۸ عنوان مقاله است. از این شمار، ۱۱۸ عنوان شامل الگوهای یاد شده در پژوهش حاضر هستند و توسط روش پیشنهادی ما استخراج و در هستان‌شناسی درج شده‌اند.

بررسی نشان داد که در ۸۱ عنوان یعنی ۶۹٪ عنوان‌هایی که الگو در آن‌ها یافته شده است، هر دو گروه اسمی یعنی اصطلاح به‌کاررفته و اصطلاح به‌کاربرنده به طور درست و کامل استخراج شده‌اند.

در ۲۰ عنوان یعنی ۱۷٪ عنوان‌های دارای الگو، یک یا هر دو گروه اسمی یعنی دست‌کم یکی از دو اصطلاح به‌کاررفته و به‌کاربرنده، ناقص استخراج شده‌اند به طوری که بخش

---

1 IJCV

۲ فهرست عنوان‌ها، هم‌چنین هستی‌نگاری تشکیل شده طی این تجربه در نشانی https://github.com/smobasheri/TitleTerms در دسترس است.

استخراج شده از نظر معنایی صحیح است اما در بردارندهٔ همهٔ معنی عنوان نیست.

نمونه‌ای از این‌ها عنوان ذیل است:

A Nonlinear, Regularized, and Data-independent Modulation for Continuously Interactive Image Processing Network

گروه‌های اسمی استخراج شده از این عنوان عبارتند از گروه اسمی data-independent modulation که درست ولی ناکامل است و گروه اسمی continuously interactive image processing network که درست و کامل است. علت ناکامل بودن گروه اسمی نخست، پیچیدگی ناشی از عطف چند صفت یعنی nonlinear، regularized و data-independent می‌باشد.

در ۱۷ عنوان یعنی ۱۴٪ عنوان‌های دارای الگو، دست‌کم یکی از دو گروه اسمی یعنی اصطلاح تخصصی به‌کار رفته یا به‌کار برنده نادرست است. نمونه‌ای از این عنوان‌ها در ذیل می‌آید.

Towards Unified Defense for Face Forgery and Spoofing Attacks via Dual Space Reconstruction Learning

گروه‌های اسمی استخراج شده از این عنوان عبارتند از گروه اسمی dual space reconstruction learning که درست و کامل است و گروه اسمی spoofing attack که نادرست است زیرا با غایب بودن عبارت unified defenece معنی عنوان به درستی بازتاب نیافته است. در این نمونه حضور بیش از یک رابط (درواقع ۳ رابط حاضر هستند) مانع از موفقیت روش شده است.

نتیجهٔ بررسی در جدول ۲ آمده است.

جدول ۲. آمار درستی اصطلاحات استخراج شده از عنوان‌های حاوی الگوهای مورد نظر

درستی گروه‌های اسمی استخراج شده	فراوانی نسبی
هر دو گروه اسمی درست و کامل است	۶۹٪
دست‌کم یکی از دو گروه اسمی درست ولی ناقص است	۱۷٪
دست‌کم یکی از دو گروه اسمی نادرست است	۱۴٪

در بین ۴۰ گروه اسمی استخراج شده معنی ناقص یا نادرست داشته‌اند، علت بررسی شده و نتیجه در جدول ۳ آمده است. ملاحظه می‌شود که حدود ۹۰٪ از خطاها به علت حضور حرف

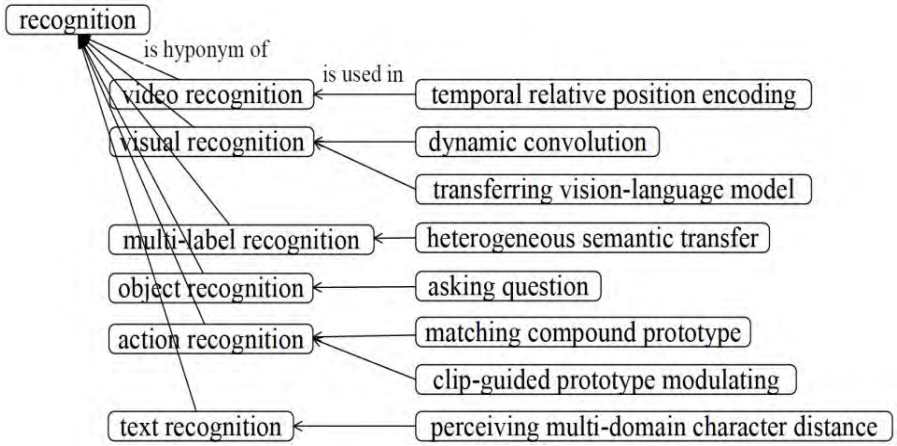
ربط دوم است، یعنی حرف ربطی افزون بر آن که در روش پیشنهادی به کار گرفته می‌شود.

جدول ۳. آمار علت استخراج گروه اسمی با معنی ناقص یا نادرست

علت استخراج گروه اسمی با معنی ناقص یا نادرست	فراوانی نسبی
حرف ربط دوم and	۳۵٪
حرف ربط دوم with	۱۸٪
حرف ربط دوم in	۸٪
حرف ربط دوم of	۸٪
حضور هم‌زمان for و via	۸٪
حروف ربط on، under، to و against	۱۳٪
علت‌های دیگر: for به معنی تخصیص، عنوان به صورت پرسش، ضعف در دیکته	۱۰٪

هستی‌نگاری تشکیل شده در این تجربه دارای ۵۰۴ اصطلاح است که ۱۰۸ اصطلاح، بدون تعمیم هستند به این معنی که در رابطهٔ خاص–عام با هیچ اصطلاح عام‌تر از خود قرار ندارند. یادآوری می‌شود که این هستی‌نگاری تنها بر پایهٔ عنوان‌های مجموعه‌ی نمونهٔ مورد آزمایش تشکیل شده است. انتظار می‌رود که در صورت اجرای روش روی یک مجموعهٔ کلان، هستی‌نگاری کلان‌تری حاصل شود.

به عنوان نمونه، بررسی شد که بنا بر هستی‌نگاری تشکیل شده، اصطلاح recognition از چه اصطلاح‌هایی بهره می‌گیرد؟ نتیجه در شکل ۷ نشان داده شده است. ملاحظه می‌شود که این اصطلاح، از تعمیم نحوی ۶ اصطلاح دیگر مانند video recognition به دست آمده است. اصطلاحات واقع در سمت راست شکل ۷ در اصطلاحات خاص‌تر از recognition کاربرد دارند.



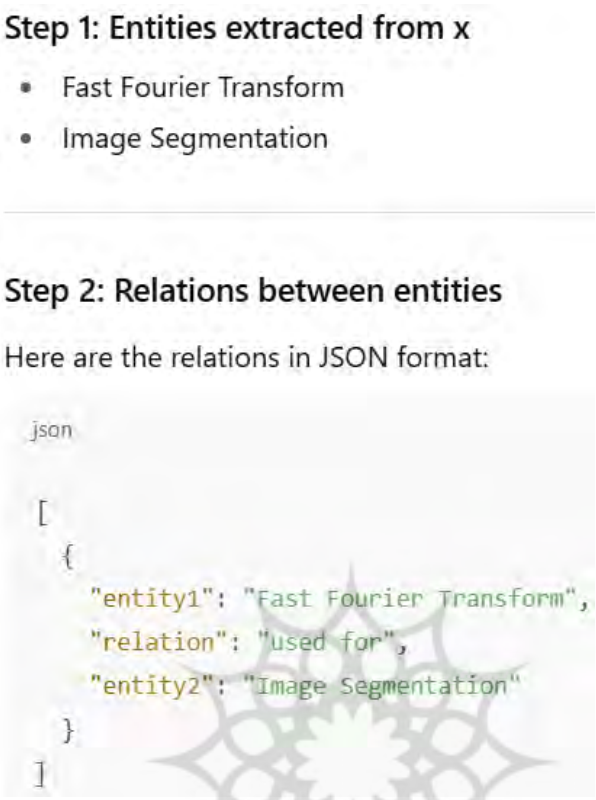
شکل ۷. اصطلاحات دارای کاربرد در recognition و اصطلاحات خاص‌تر از آن

تجربه با سامانهٔ هوشمند مبتنی بر مدل بزرگ زبانی

همان مجموعه از عنوان‌ها که برای آزمودن روش پیشنهادی به کار رفته است، با استفاده از سامانهٔ GPT-5 متعلق به OpenAI بررسی شد (OpenAI, 2025). شکل ۸ نمونه‌ای از پرسش مطرح شده و پاسخ سامانه را نشان می‌دهد. مشاهده می‌شود که سامانه نهادهای حاضر در عبارت را به‌درستی تشخیص داده و رابطهٔ بین آن دو را نیز به‌درستی استخراج نموده است. در این تجربه، GPT-5 نهادها و رابطه‌های حاضر در تقریبا همهٔ عنوان‌ها را به‌درستی تشخیص داد، مگر فقط در یک عنوان سنگین و پیچیده. لذا خطای این سامانه در این تجربه، عملا کم‌تر از ۱٪ بوده است.

هم‌چنین از GPT-5 خواسته شد که پیچیدگی محاسباتی عمل استخراج نهادها و رابطه‌ها را برای روش مبتنی بر LLM و روش کلاسیک با هم مقایسه کند. بنا بر برآورد این سامانه تعداد عملیات توسط سامانه ۴۲۰۰۰ برابر تعداد عملیات در روش کلاسیک NLP است. دقت و حجم عملیات دو روش، در جدول ۴ مقایسه شده است.

For the text between triple curly braces, named as x, do the following:
1. Extract all the entities from x, and list them. 2. Extract all the relations within the above-mentioned entties. Report the found relations in a json format with the following fields: "entity1", "relation", "entity2"
{{{Fast Fourier Transform for Image Segmentation}}}



شکل ۸ . یک نمونه از پرامپت تحویل شده به GPT-5 و پاسخ آن. نهادها و رابطهٔ میان آن‌ها به‌درستی استخراج شده است.

جدول ٤ . دقت و پیچیدگی محاسبات در روش پیشنهادی و روش مبتنی بر GPT-5

	روش پیشنهادی	روش مبتنی بر GPT-5
دقت	۶۹٪	۹۹٪
پیچیدگی محاسبات	۰,۰۰۱ TFLOPs[1]	۴,۲ TFLOPs

نتیجه‌گیری

در پژوهش حاضر روشی نحوی برای تشکیل خودکار هستی‌نگاری از اصطلاحات تخصصی

1 Tera (1012) Floating Point OPerations

حاضر در عنوان مقاله‌های علمی پیشنهاد شده است. پرسش رقابتی این هستی‌نگاری این است: «برای انجام عمل x از چه روش‌هایی استفاده شده است؟»

در روش پیشنهادی رابطهٔ خاص-عام بین اصطلاحات از یک نسبت سادهٔ صوری بین عبارت‌های آن اصطلاحات به دست می‌آید. هم‌چنین رابطهٔ کاربرد بین اصطلاحات تخصصی از الگوهای نحوی مشخصی در عنوان مقاله‌ها استخراج می‌شود. گذشته از این، رابطهٔ کاربرد از نقطه‌ای که برپا شده است، در درخت‌وارهٔ اصطلاحات به سوی اصطلاحات عام‌تر به ارث می‌رسد. اگر روش پیشنهادی ما برای بخشی قابل توجه از مقاله‌های علمی در یک دامنهٔ علمی مورد نظر اجرا شود، می‌تواند یک هستی‌نگاری غنی از اصطلاحات آن دامنه با ربط کاربرد بین آن‌ها ایجاد کند.

نمونهٔ ۵ و توضیح شکل ۷ نشان می‌دهد که قاعده‌های استنتاج تنظیم شده در هستی‌نگاری، به‌درستی مشخص می‌کنند که روش‌های به‌کار رفته برای انجام یک هدف مورد نظر کدام‌ها هستند؟ در این نمونه هم‌چنین سودمندی برپا کردن رابطهٔ خاص-عام با روش پیشنهادی آشکار می‌شود. فرض کنیم افزون بر اصطلاح visual recognition، اصطلاح voice recognition نیز که از یک عنوان فرضی استخراج شده است در هستی‌نگاری اصطلاحات درج شود. روشن است که این دو اصطلاح، دو گونه از recognition به معنی *بازشناسی* هستند. اگر همان‌گونه که در نوشتهٔ حاضر پیشنهاد شده است هر دو اصطلاح به recognition تعمیم یافته باشند، در این صورت در هستی‌نگاری، اصطلاح recognition والد هر دو اصطلاح خواهد بود. اما اگر این تعمیم مغفول مانده باشد، در آن صورت این دو اصطلاح، والد مشترکی در هستی‌نگاری نخواهند داشت که به‌روشنی نادرست، یا دست‌کم ناقص خواهد بود.

از آن‌جا که روش پیشنهادی ما کاملا مبتنی بر نحو است، قادر به تشخیص رابطه‌های معنایی نیست. با این همه چون اصطلاحات را تا سطوح انتزاعی تعمیم می‌دهد (شکل ۵ و شکل ۷)، از این طریق امکان ملحق شدن به روش‌های مبتنی بر هستی‌نگاری را دارد.

روش پیشنهاد شده در این پژوهش قابل گسترش به مجموعه‌ای کلان از عنوان‌ها است. این روش برای مقیاس‌پذیری ویژگی‌هایی مثبت دارد. نخست، از آن‌جا که عنوان مقاله‌های پژوهشی به‌رایگان در اختیار هستند، گردآوری مجموعهٔ ورودی مستلزم صرف هزینه نیست. دوم، امروزه ناشران مطرح جهان، برای گردآوری فرامحتوا، واسط نرم‌افزاری ارائه کرده‌اند. لذا

گردآوردن مجموعه‌ای کلان از عنوان مقاله‌ها به‌طور مکانیزه قابل انجام است.

رابطه‌های خاص–عام و کاربرد را می‌توان به کمک سیستم‌های هوشمند مبتنی بر مدل‌های کلان زبانی مانند GPT-5 نیز استخراج کرد. تجربهٔ پژوهش حاضر نشان می‌دهد که دقت در روش مبتنی بر LLM بسیار بیش‌تر از روش پیشنهادی ما و عملا نزدیک صددرصد است. با این حال، حجم عملیات در روش مبتنی بر LLM به مقیاس ۱۰هزار (درواقع ۴۲هزار) بار سنگین‌تر از روش پیشنهادی ما است.

در تجربهٔ ما نسخهٔ رایگان GPT-5 به کار رفته است. به علت محدودیت‌های اعمال شده توسط این نسخه از سامانه، در هر نشست، بیش از حدود ۲۰ دستور پذیرفته نمی‌شود و برای دستورهای بعدی لازم است مدت معینی صبر کرد. این تجربه نشان می‌دهد که استخراج اطلاعات بر پایهٔ LLM عملا در مقیاس وسیع امکان‌پذیر نیست.

گذشته از این، عمل درج در هستی‌نگاری ضروری خواهد بود زیرا پرسش‌هایی مانند این که «برای هدف x چه روش‌هایی به‌کار رفته است؟» در صورتی پاسخ مناسب می‌یابند که شماری عظیم از عنوان‌ها مورد بررسی قرار گرفته و افزون بر آن، رابطه‌های استخراج شدهٔ کاربرد، در هستی‌نگاری درج شده باشد. افزون بر درج در هستی‌نگاری، استدلال مبتنی بر قاعده نیز که در روش پیشنهادی ما گنجانده شده است، به یافتن رابطه‌های کاربرد بین اصطلاحات کمک می‌کند.

جدول ۳ نشان می‌دهد که ۹۰٪ از خطاهای روش پیشنهادی در مواردی رخ می‌دهند که عنوان مورد بررسی حاوی بیش از یک حرف ربط باشد. پس یکی از پیشنهادها برای کارهای آینده، طراحی الگوهای پیچیده‌تر برای تحلیل عنوان مقاله‌ها است، به‌طوری که عنوان‌های حاوی بیش از یک رابط، به‌ویژه عنوان‌های حاوی رابط‌های and و with را شامل شوند.

پیشنهاد دیگر تشکیل پیکره‌ای کلان از عنوان مقاله‌های علمی در حوزه‌های مورد نظر است که می‌تواند به ساخت هستی‌نگاری‌هایی در مقیاس واقعی بسیار کمک کند.

پیوست: تجربه با مقاله‌های علمی فارسی

روش پیشنهاد شده در پژوهش حاضر در واقع با الگوبرداری از تجربه‌های پیشین که روی مقاله‌های علمی انگلیسی به‌دست‌آمده بود تنظیم شد. با این همه این روش روی شماری از

عنوان‌های مقاله‌های علمی فارسی نیز تجربه شد. در این پیوست، حاصل این تجربه گزارش می‌شود.

«نشریهٔ مهندسی برق و مهندسی کامپیوتر ایران»[1] متعلق به جهاد دانشگاهی به عنوان نمونه به کار رفته است. عنوان مقاله‌های سال ۱۴۰۳ این نشریه شامل ۵۵ عنوان گردآوری و با روش پیشنهادی بررسی شد. تعداد ۳۹ عنوان، حاوی ساخت نحوی مورد استفاده در روش ما هستند.

از این شمار، ۳ عنوان در واقع حاوی رابطهٔ کاربرد نیستند اما با روش پیشنهادی، به‌خطا رابطهٔ کاربرد در آن‌ها اعلام می‌شود. برای نمونه در عنوان «طراحی بهینه و تحلیل محدودکننده جریان خطای مبتنی بر راکتور سری متغیر با هسته هوایی» حضور حرف «با» موجب خطای روش پیشنهادی می‌شود زیرا در این‌جا «با» نشان‌دهندهٔ رابطهٔ احتوا است نه کاربرد. حروف ربط «با» و «برای» می‌توانند موجب خطای روش مورد نظر شوند.

از ۳۶ عنوان که درواقع حاوی رابطهٔ کاربرد هستند، در ۱۵ عنوان هر دو اصطلاح، یعنی اصطلاح به‌کاررفته و اصطلاح به‌کاربرنده به‌درستی استخراج شدند. در ۱۱ عنوان، یک یا هر دو اصطلاح به طور ناقص استخراج شدند و در ۱۰ عنوان، یک یا هر دو اصطلاح به طور نادرست استخراج شدند. این آمار در جدول ۵ آمده است. مشاهده می‌شود که دقت این روش برای عنوان‌های فارسی کم‌تر از عنوان‌های انگلیسی است.

جدول ۵. آمار درستی اصطلاحات استخراج شده از عنوان‌های فارسی حاوی الگوهای مورد نظر

درستی گروه‌های اسمی استخراج شده	فراوانی نسبی
هر دو گروه اسمی درست و کامل است	۳۸٪
دست‌کم یکی از دو گروه اسمی درست ولی ناقص است	۲۸٪
دست‌کم یکی از دو گروه اسمی نادرست است	۲۶٪
رابطهٔ کابرد به خطا اعلام شده است	۸٪

عمدهٔ عوامل بازدارندهٔ روش پیشنهادی در عنوان‌های فارسی عبارتند از:

۱. حضور رابط‌های دیگر علاوه بر رابط مورد استفادهٔ روش پیشنهادی.

1 http://ijece.org

۲. برخی از رابط‌ها چندمعنایی هستند. در تجربهٔ حاضر، «با» و «برای» چنین هستند.

۳. حضور واژه‌هایی که حاوی اطلاعاتی درباره‌ی ابزار یا هدف پژوهش نیستند. برای نمونه در عنوان فرضی «ارائه مدلی برای x به منظور y»، بخش نخست یعنی «ارائه مدلی» عملا حاوی اطلاعاتی از ابزار نیست. یعنی بیان نمی‌کند که چه ابزار ریاضی یا غیرریاضی به y کمک کرده است؟

۴. رابط‌هایی که از چند واژه تشکیل می‌شوند مانند «با استفاده از» معمولا در توکن‌سازی، در چند توکن منعکس می‌شوند. لازم است که پس از برچسب‌زنی، این رابط‌ها یافته شده و تنها یک برچسب (نه چند برچسب) دریافت کنند.

۵. پیشنهاد برپاکردن رابطهٔ خاص–عام در گروه‌های اسمی فارسی نیاز به اصلاح دارد. برای نمونه اگرچه گروه اسمی fast Fourier transform مطابق شکل ۵ به Fourier transform و سپس به transform تعمیم می‌یابد، در فارسی گروه اسمی «تبدیل سریع فوریه» ترتیبی متفاوت از انگلیسی دارد و نمی‌توان آن را به‌ترتیب به «تبدیل سریع» و «تبدیل» تعمیم داد. لذا به علت حضور ساختار توام اضافهٔ مرکب و صفت، برای برقراری صحیح رابطهٔ تعمیم، ضروری است که به زیردرخت وابستگی گروه اسمی توجه کرد.

اعلام استفاده از هوش مصنوعی

در تهیه این مقاله، از سامانهٔ GPT-5 (OpenAI, 2025) ، فقط برای مقایسهٔ تجربی عملکرد با روش پیشنهادی استفاده شده است. این سامانه نقشی در نگارش مقاله یا تولید نتایج پژوهش نداشته است. مسئولیت کامل محتوای نهایی مقاله بر عهدهٔ نویسنده است.

منابع

اسماعیلی آدرگانی، زهرا و فراهی، احمد (۱۳۹۲). استخراج کلیدواژه‌های مقالات فارسی بر اساس تحلیل معنایی. *هفتمین کنفرانس ملی مهندسی برق با محوریت انرژی های نو*. https://civilica.com/doc/244182.

بهاری ورزنه، حسین (۱۴۰۱). مقایسه عملکرد بازیابی اطلاعات موتورهای جستجوی معنایی و کلیدواژه‌ای بر اساس جستجوی عبارتی. *فصلنامه مطالعات دانش پژوهی*.

https://civilica.com/doc/1568953.

درزی خلردی، صغرا و رضوی، علی‌اصغر (۱۳۹۷). همخوانی کلیدواژه‌های مقاله‌های مجلات دانشگاه علوم کشاورزی و منابع طبیعی ساری با اصطلاحنامه کب. *مجله دانش‌شناسی*. https://civilica.com/doc/1603395.

شریف، عاطفه (۱۳۸۸). مهندسی خودکار هستی‌شناسی: امکان‌سنجی استخراج روابط معنایی از متون فارسی و تعیین میزان پیدایی آن‌ها. *کتابداری و اطلاع‌رسانی* ۱۲(۲)، ۲۶۳-۲۴۳.

شمس‌فرد، مهرنوش و جعفری‌نژاد، فاطمه (۱۳۹۱). استخراج روابط معنایی میان فعل و وابسته‌های آن از متون زبان فارسی. *فصلنامه پازند* ۸(۳۰)، ۷۱-۵۳.

غنی پور تفرشی، مریم و حاجی زین العابدینی، محسن (۱۳۹۷). تاثیر مستندسازی کلیدواژه‌ها با استفاده از اصطلاحنامهٔ پزشکی فارسی بر بازیابی مقالات از پایگاه‌های ایران مدکس و مدلیب. *پژوهش های کتابخانه های دیجیتالی و هوشمند*. https://civilica.com/doc/1015990.

مبشری، سروش (۱۴۰۳). *مدلی برای بازنمایی فراعلم جهت کار با هستان‌شناسی‌های علمی*. پایان‌نامه دکتری. گروه هوش مصنوعی، رباتیک و رایانش شناختی. دانشکده مهندسی و علوم کامپیوتر. دانشگاه شهید بهشتی.

هدایتی، سعید و حسنی آهنگر، محمدرضا (۱۴۰۰). استخراج تعابیر روشی جهت استخراج کلیدواژه ترکیبی در اسناد متنی با استفاده از شبکه عصبی کانولوشن. *هشتمین کنفرانس ملی پژوهش های کاربردی در علوم برق، کامپیوتر و مهندسی پزشکی*. https://civilica.com/doc/1234162.

References

Bahari-Varzaneh, H. (2022). Comparing the Information Retrieval Performance of Semantic and Keyword Search Engines based on Phrase Search, *Quarterly Journal of Scholarly Studies*. https://civilica.com/doc/1568953. [In Persian]

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM*, *39*(1), 80-91.

Darzi-Khalardi, S. & Razavi, A. (2019). Consistency of keywords in journal articles of Sari University of Agricultural Sciences and Natural

Resources with the Kab Thesaurus, *Journal of Knowledge Studies*. https://civilica.com/doc/1603395. [In Persian]

Ding, H., & Luo, X. (2021). Attentionrank: Unsupervised keyphrase extraction using self and cross attentions. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1919–1928.

Esmaeli-Adargai, Z. & Farahi, A. (2013). Extracting Keywords from Persian Articles based on Semantic Analysis, *The 7th National Conference on Electrical Engineering with a Focus on New Energies*. https://civilica.com/doc/244182. [In Persian]

Fader, A., Soderland, S., & Etzioni, O. (2011). Identifying relations for open information extraction. In *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1535-1545).

Ghanipour-Tafreshi, M. and Haji-Zeinolabedini, M. (2019). The effect of documenting keywords using Persian medical thesaurus on retrieving articles from Iran MedEx and MedLib databases, *Digital and Smart Libraries Researches*. https://civilica.com/doc/1015990. [In Persian]

Hedayati, S. and Hasani-Ahangar, M. (2021). Extracting Interpretations: A Method for Extracting Compound Keywords in Text Documents Using Convolutional Neural Networks, *8th National Conference on Applied Research in Electrical, Computer and Medical Engineering*. https://civilica.com/doc/1234162. [In Persian]

Mobasheri, S. (2025). *A Meta-science Representation Model for Handling Scientific Ontologies*. PhD thesis. Department of Artificial Intelligence, Robotics and Cognitive Computing. Faculty of Computer Science and Engineering. Shahid Beheshti University. [In Persian]

Nomoto, T. (2022). Keyword extraction: a modern perspective. *SN Computer Science*, *4*(1), 92.

OpenAI. (2025). *ChatGPT (GPT-5)* [Large language model]. OpenAI. https://chat.openai.com/

RiahiNia, N., Shadanpour, F., Borna, K., & Montazer, G. A. (2022). Automatic keyword extraction using Latent Dirichlet Allocation topic mllll igg: mmihrrity with gll nnn ttaaaard ddd ssrrs' vvll ttt i... *Human Information Interaction*, *9*(3). http://hii.khu.ac.ir/article-1-3069-fa.html [In Persian]

Shamsfard, M. and Jafarinejad, F. (2012). Extracting Semantic Relations between Verbs and their Arguments from Persian Texts. *Pazand Quarterly*, *8*(30), 53-71. [In Persian]

Sharif, A. (1388). Automated Ontology Engineering: Feasibility study of extracting semantic relationships from Persian texts and determining their occurrence rate. *Library and Information Sciences*. [In Persian]

Xu, Z., & Zhang, J. (2021). Extracting keywords from texts based on word frequency and association features. *Procedia Computer Science*, *187*, 77–82.

پژوهشگاه علوم انسانی و مطالعات فرهنگی
پرتال جامع علوم انسانی