Feasibility Study of Ontological Development Using Semi-Automatic Method based on Lexical Frequency Analysis: A Case Study of "Glaucoma"

Somayeh Tamjid
PhD Candidate in Knowledge and Information Science; Department of Communication and Knowledge Sciences; Islamic Azad University; Science and Research Branch; Tehran, Iran Email: s.tamjid@srbiau.ac.ir; tamjid.s@iums.ac.ir

Fatemeh Nooshinfard*
PhD in Knowledge and Information Science; Associate Professor; Department of Communication and Knowledge Sciences; Islamic Azad University; Science and Research Branch; Tehran, Iran Email: nooshinfard@srbiau.ac.ir; f.nooshinfard@gmail.com

Molouk Sadat Hosseini Beheshti
PhD in Linguistics; Associate Professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran; Email: Beheshti@irandoc.ac.ir

Nadjla Hariri
PhD in Knowledge and Information Science; Professor; Department of Communication and Information Science; Islamic Azad University; Science and Research Branch; Tehran, Iran Email: N_hariri@srbiau.ac.ir

Fahimeh Babolhavaeji
PhD in Knowledge and Information Science; Associate Professor; Department of Communication and Knowledge Sciences Islamic Azad University; Science and Research Branch; Tehran, Iran; Email: f.babalhavaeji@gmail.com

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 39 | No. 1 | pp. 131-156
Autumn 2023
https://doi.org/jipm.39.1

Received: 13, Mar. 2022 **Accepted: 21, Aug. 2022**

Abstract: Following recent trends in information management systems, conventional word-based information retrieval methods are changing to concept-based approaches by means of the broad application of ontologies. More specifically, the use of ontologies for knowledge management is significant in the medical sciences and human disease domains due to the diversity and necessity of information sharing between numerous data repositories such as medical records, health record

* Corresponding Author

systems, and so on. Furthermore, ontologies make natural language processing approaches more feasible by reducing semantic ambiguity and making concepts comprehensible to computer-based deductions. In this research, a semi-automated approach for ontology development is proposed, which assists in identifying structural components of an ontology and determining possible relations between them based on scientific text records. The proposed approach, in a general view, includes the gathering of a large volume of technical data in text format, processing, and extraction of results with a minimal contribution of human-based supervision. The processing stage is coded in Matlab code named TmbOnt_Alfa and applies two main techniques including word frequency and Lexico-Synactic patterns analysis, to identify concepts and relations, respectively. The role of the human supervisor is narrowed to entering target terms, eliminating unnecessary outputs, and finalizing the ontology structure. In order to evaluate the efficiency of the proposed method, a case study for ontological development in the field of glaucoma has been conducted, and results are compared with medical subject headings of MESH descriptors, the Persian medical thesaurus, ontology of diseases, and Bioassay ontology (BAO).

According to results, the developed ontology, when compared by Glaucoma entry, covered 80% of the medical titles in Mesh, 100% of the medical terms developed in the Persian Medical Thesaurus, and 100% of the Persian medical descriptors. Moreover, the resultant ontology structure is compatible with more than 90% of the same ontology represented in Bioassay and 57% of the ontology of diseases (DO). It also proposed an average of 30% more terms for existing ontological structures.

Keywords: Ontology, Text Mining, Information Representation, Glaucoma, Eye Disease, Medical Thesaurus, Protégé

امکان‌سنجی توسعه هستی‌شناسی به روش نیمه‌خودکار مبتنی بر تحلیل بسامد واژگان: مطالعه موردی بیماری «گلوکوم»

سمیه تمجید
دانشجوی دکتری علم اطلاعات و دانش‌شناسی؛
دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران؛
s.tamjid@srbiau.ac.ir

فاطمه نوشین‌فرد
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛
دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران؛
پدیدآور رابط nooshinfard@sriau.ac.ir

ملوک‌السادات حسینی بهشتی
دکتری زبان‌شناسی؛ دانشیار؛
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)؛
تهران، ایران beheshti@irandoc.ac.ir

نجلا حریری
دکتری علم اطلاعات و دانش‌شناسی؛ استاد؛
دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران؛
N_hariri@srbiau.ac.ir

فهیمه باب‌الحوائجی
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛
دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران؛
f.babalhavaeji@gmail.com

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۱۲/۲۲ | پذیرش: ۱۴۰۱/۰۵/۳۰ | مقاله برای اصلاح به مدت ۵۳ روز نزد پدیدآوران بوده است.

چکیده: تغییر رویکرد نظام‌های اطلاعاتی از پردازش واژه به پردازش مفهوم، موجب توجه به هستی‌شناسی‌ها شده است. در علوم پزشکی و بیماری‌های انسان به لحاظ وجود تنوع در اصطلاحات و لزوم اشتراک اطلاعات از طریق نرم‌افزارهای مختلف مانند پرونده‌های پزشکی، سامانه‌های ثبت سوابق بهداشتی و ... به‌کارگیری هستی‌شناسی‌ها ضروری به نظر می‌رسد. در پژوهش حاضر، رویکردی نیمه‌خودکار برای توسعه هستی‌شناسی پیشنهاد شده است که می‌تواند با استفاده از ابزارهای متن‌کاوی، شناسایی مؤلفه‌های ساختاری هستی‌شناسی و تعیین نسبی روابط را از متون علمی تسهیل کند. مدل پیشنهادی در قالب کد نرم‌افزاری با نام اختصاری TmbOnt_Alfa ارائه شده است. این کد با استفاده از رابط کاربر،

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۹ | شمارهٔ ۱ | ص‌ص ۱۳۱-۱۵۶
پاییز ۱۴۰۲
https://doi.org/jipm.39.1

فایل متنی ورودی را فراخوانی کرده و پس از پردازش بر اساس تنظیمات، اصطلاحات کلیدی برای توسعهٔ هستی‌شناسی را استخراج می‌کند. به‌منظور ارزیابی کارایی روش پیشنهادی، مطالعه موردی در حوزهٔ بیماری «گلوکوم» با داده‌های متنی مشتمل بر ۱۰،۰۰۰ چکیده مقاله از «پابمد» بر مبنای جست‌وجوی واژگانی تهیه گردید. پس از مراحل پردازش، مفاهیم و ساختار سلسله‌مراتبی هستی‌شناسی حاصل در نرم‌افزار «پروتژ» وارد شد. سرانجام، سنجش قیاسی هستی‌شناسی توسعه‌یافته با سرعنوان‌های موضوعی پزشکی «مش»، «اصطلاحنامه وتوصیفگرهای پزشکی فارسی» و «هستی‌شناسی بیماری‌ها» و «هستی‌شناسی زیست‌پزشکی» نشان داد که میانگین دقت مفاهیم و میانگین دقت مکانی مفاهیم بیش از ۷۰ درصد با هستی‌شناسی‌های بازنمایی‌شده در پایگاه‌های معتبر هستی‌شناسی بیماری‌های انسانی انطباق داشته و به‌طور میانگین بیش از ۳۰ درصد واژگان جدید برای افزودن به دامنه را فراهم کرده است.

کلیدواژه‌ها: هستی‌شناسی، متن‌کاوی، بازنمایی دانش، «گلوکوم»، بیماری‌های انسان، پروتژ

۱. مقدمه

امروزه استفاده از هستی‌شناسی جهت بازنمایی و بازیابی اطلاعات و دانش در حوزه‌های مختلف به صورتی روزافزون مورد توجه قرار گرفته است. گسترش مرزهای دانش و رشد شتابان علوم جدید، پرداختن به ایده‌های ایجاد و یا به‌روزرسانی هستی‌شناسی‌ها را ضروری کرده است. در حوزه‌های پیشگام و به‌ویژه در حوزهٔ پزشکی، توسعه هستی‌شناسی‌های جدید و بهبود مستمر هستی‌شناسی‌های موجود، با تلاشی مضاعف توسط جوامع تحقیقاتی دنبال می‌شود. طبق مطالعات، اصطلاحنامه‌ها در دو بعد مفاهیم و روابط بین اصطلاحات و انتخاب یک اصطلاح به‌عنوان مرجح و برقراری ارتباط بین آن و سایر اصطلاحات در قالب اعم/ اخص/ مرتبط، دارای محدودیت‌هایی هستند (صنعت‌جو، دیانی و فتحیان دستگردی ۱۳۸۹). به نظر «سورگل» و همکاران فقدان انتزاع مفهومی، پوشش معنایی محدود و پردازش خودکار محدود جزو محدودیت‌های اصطلاحنامه‌هاست (Soergel et al. 2004). همین عوامل تجدید نظر در ایجاد و استفاده از ابزارهای معناشناسی و بازنمون دانش با استفاده از فناوری‌های نوین را ضروری ساخته است.

هستی‌شناسی یکی از ابزارهای معناشناسی برای غلبه بر مشکلات تعامل با سیستم‌هاست. هستی‌شناسی یک توصیف رسمی، صریح و روشن از مفاهیم در حوزه‌ای خاص است. امروزه، محققان زیادی تأثیر هستی‌شناسی را برای ساخت و تعامل بهتر در سیستم‌های اطلاعات تائید و تصدیق می‌کنند. در اصطلاح امروزی، هستی‌شناسی به‌عنوان ابزاری برای توصیف «مفاهیم و روابط بین آن‌ها برای محققانی که نیاز به اشتراک

اطلاعـات در یـک دامنـه دارنـد» تعریـف شـده اسـت (Noy and McGuinness 2001). به‌طـور خلاصـه، بـا حرکـت از تاکسـونومی‌ها و اصطلاحنامه‌هـا بـه سـوی الگـوی مفهومـی و نظریـه منطقـی در برقـراری ارتبـاط بیـن مفاهیـم، بـه هستی‌شناسـی‌ها می‌رسـیم (حسـینی بهشـتی ۱۳۹۲). ایـن تأثیـر در حـوزه سـلامت، بـا توجـه بـه تعـدد و پیچیدگـی اصطلاحـات تخصصـی و لـزوم دسـتیابی بـه تعامـلات بیـن سیسـتمی در بازیابـی دانـش می‌توانـد بسـیار کارآمـد باشـد. بـه همیـن دلیـل، شـاهد رشـد روزافـزون هستی‌شناسـی‌ها در حـوزه زیست‌پزشـکی و سـلامت بـه نسـبت سـایر حوزه‌هـا هسـتیم (Noy et al. 2009). توسـعهٔ هستی‌شناسـی در ایـن حـوزه کمـک می‌کنـد کـه اطلاعـات گسسـتهٔ به‌دسـت‌آمده از بطـن فعالیت‌هـای تجربـی و نظـری قابلیـت بازیابـی بهتـری یافتـه و بـه دانـش قابـل ادراک توسـط سـامانه‌های نرم‌افـزاری تبدیـل شـود. در مـدت کوتـاه مواجهـه بـا بیمـاری «کوویـد ۱۹»، محققـان بـا سـاخت چندیـن هستی‌شناسـی در سراسـر دنیـا تـلاش کردنـد تـا بـا ایجـاد مفاهیـم مشـترک بتواننـد داده‌هـای بیمـاری را بـه اشـتراک بگذارنـد. هستی‌شناسـی‌های "VIDO"، "CIDO"، "IDO-COVID-19"، "CODO" همگـی طـی سـال‌های ۲۰۲۰ و ۲۰۲۱ ایجـاد شـده‌اند (Babcock et al. 2021 و Dutta and DeBellis 2020). از منظـر اقتصـادی، اهمیـت هستی‌شناسـی از میـزان بودجـه تخصیـص یافتـه بـه آن در جوامعـی کـه هستی‌شناسـی‌ها را توسـعه می‌دهنـد، مشـهود اسـت. اختصـاص بودجـه ۴/۴ میلیـون دلاری در سـال ۲۰۰۹ بـرای «هستی‌شناسـی ژن»، تخصیـص بودجـه ۱۸/۸ میلیـون دلاری بـرای پنـج سـال کار در حـوزهٔ «هستی‌شناسـی زیست‌پزشـکی» و نیـز تخصیـص ۳۲/۴ میلیـون دلار در سـال ۲۰۰۳ و پـس از آن ۶ میلیـون دلار بودجـهٔ سـالیانه توسـط «کتابخانـه ملـی پزشـکی» بـرای پـروژهٔ "SNOMED-CT" می‌توانـد منعکس‌کننـدهٔ بخشـی از حجـم سـرمایه‌گذاری در ایـن حـوزه باشـد (Liu, Hogan and Crowley 2011). به‌طـور معمـول، توسـعهٔ مـداوم و نگه‌داشـت مؤثـر هستی‌شناسـی‌ها توسـط عامـل انسـانی انجـام می‌شـود کـه زمان‌بـر بـوده و مسـتعد اشـتباهات متعـددی اسـت. امـروزه، تکنیک‌هـای نیمه‌خـودکار بازیابـی دانـش در حـال تبدیـل شـدن بـه ابـزاری بـرای کمـک بـه عوامـل انسـانی اسـت. اتخـاذ چنیـن رویکـردی می‌توانـد فعالیت‌هـای تکـراری و خسـته‌کننده را بـه اقدامـات نظارتـی تغییـر داده و منجـر بـه کاهـش زمـان و بهبـود کیفیـت شـود.

مزیت‌هـای حاصـل از به‌کارگیـری هستی‌شناسـی‌ها در نرم‌افزارهـای جسـت‌وجو و بازیابـی اطلاعـات، توسـعه‌دهندگان هستی‌شناسـی در حـوزهٔ پزشـکی را بـر آن داشـت تـا از ایـن مفهـوم بـرای تسـهیل ادغـام داده‌هـای زیست‌پزشـکی مرتبـط بـا بیماری‌هـای انسـان بهـره

ببرنـد. بـرای مثـال، از سـال ۲۰۱۲، پایـگاه داده پـروژهٔ «هستی‌شناسـی بیماری‌هـا»[1]، ۱۹۲ بـار به‌روزرسـانی شـده و طـی ایـن فراینـد، حـدود ۷۶۰ اصطـلاح جدیـد، کـه برخـی از آن‌هـا در تعاریـف بـه‌کار رفته‌انـد، بـدان اضافـه شـده اسـت (Kibbe et al. 2015). یکـی از مشـکلات توسـعه هستی‌شناسـی از ایـن واقعیـت ناشـی می‌شـود کـه ایجـاد هستی‌شناسـی بـر پایـه اصطلاحنامـه، هرچنـد در شناسـایی اصطلاحـات مرجـع به‌عنـوان ریشـه مفیـد اسـت، امـا بسـیار هزینه‌بـر بـوده و به‌طـور کامـل بـه عامـل انسـانی متخصـص وابسـته اسـت و امـکان اشـتباه در مراحـل مختلـف دور از انتظـار نیسـت (Kless et al. 2016). بـرای رفـع ایـن مشـکل در ایـن پژوهـش از طراحـی هستی‌شناسـی به‌صـورت نیمه‌خـودکار و بـا روش متـن‌کاوی و پـردازش زبـان طبیعـی اسـتفاده شـده و مفاهیـم و اصطلاحـات حاصـل بـا هستی‌شناسـی‌ها و اصطلاحنامه‌هـای پزشـکی معتبـر مقایسـه شـده اسـت. مطالعـه مـوردی بـرای امکان‌سـنجی توسـعهٔ هستی‌شناسـی روی موضـوع گلوکـوم[2] انجـام شـده اسـت. گلوکـوم بـر اسـاس مطالعـات، دومیـن عامـل نابینایـی افـراد بـالای ۵۰ سـال اعـلام شـده و مهم‌تریـن دلیـل نابینایـی قابـل کنتـرل در افـراد بالـغ اسـت (معصومـی و همـکاران ۱۳۹۱). جسـت‌وجو در پایگاه‌هـای اطلاعاتـی نشـان‌دهندهٔ رونـد صعـودی تولیـدات علمـی در مـورد ایـن بیمـاری اسـت (شـکل ۲). ایـن در حالـی اسـت کـه سـرعنوان‌های پزشـکی صرفـاً پنـج زیرشـاخه بـرای آن در نظـر گرفتـه، ولـی ایـن پنـج مـورد به‌طـور کامـل نمی‌توانـد سـاختار دانـش ایـن حـوزه را دربرگیـرد (شـکل ۵). ایـن پژوهـش در صـدد توسـعهٔ هستی‌شناسـی جهـت بازنمـون دامنـه دانـش در حیطـهٔ گلوکـوم، از طریـق متـن‌کاوی جدیدتریـن مقـالات منتشرشـده در پایگاه‌هـای اطلاعاتـی اسـت. رجـوع بـه متـن منـدرج در مقـالات تـازه انتشـاریافته می‌توانـد مفاهیـم جدیـد مـورد اسـتفاده در حـوزهٔ دانـش بـر اسـاس آخریـن مطالعـات در حـال انجـام در مـرز دانـش ایـن حـوزه را شناسـایی کنـد کـه در توسـعه و به‌روزرسـانی هستی‌شناسـی بسـیار مؤثـر خواهـد بـود.

مقالـه حاضـر در شـش بخـش اصلـی سـازماندهی شـده اسـت. در بخـش دوم، پیشـینهٔ پژوهـش بررسـی شـده اسـت. بخـش سـوم، بـه روش و رویکردهـای خـودکار موجـود بـرای توسـعهٔ هستی‌شناسـی پرداختـه اسـت. در بخـش چهـارم، جزئیـات مطالعـهٔ مـوردیِ انجام‌شـده در حـوزهٔ گلوکـوم ارائه‌شـده و در بخـش پنجـم، نتایـج به‌دسـت‌آمده بحـث شـده اسـت. سـرانجام، در بخـش ششـم، نتیجه‌گیـری و پیشـنهادات بـرای پژوهش‌هـای آتـی ارائـه شـده اسـت.

1. disease ontology (DO)
2. glaucoma

۲. پیشینه پژوهش

توسعه و غنی‌سازی هستی‌شناسی از بطن متون انبوه، رویکردی منطقی است که طی سال‌های اخیر توسط بسیاری از محققان مورد اقبال قرارگرفته است. «میسیکوف، ولاردی و فابریانی» با استفاده از تکنیک متن‌کاوی و تحلیل تشابه متن، دستیاری برای توسعه هستی‌شناسی را در جهت کمک به متخصصان هستی‌شناسی ارائه کردند (Missikoff, Velardi and Fabriani 2003). انگیزه اصلی این تحقیق، کاهش زمان، کاهش هزینه و همسان‌سازی تفاسیر مبتنی بر رویکردهای انسانی بود. «جیانگ و تان» روشی را برای استخراج مفاهیم و روابط معنایی از مجموعه متون دامنه ارائه دادند. آن‌ها یک سیستم جدید به نام CRCTOL برای استخراج دانش معنایی اسناد متنی در دامنه موضوعی خاص طراحی کردند که به جای استخراج مفهوم و روابط به‌صورت سطحی، دانش معنایی غنی در قالب هستی‌شناسی را استخراج می‌کرد (Jiang and Tan 2005). «شارلت، بچیمونت و ژالنت» با استفاده از ابزار پردازش زبان طبیعی به نام «SYNTEX»، روشی را توسعه داده‌اند که هستی‌شناسی پزشکی را از درون گزارش‌های متنی استخراج می‌کند (Charlet, Bachimont and Jaulent 2006). «کیان، لان و لیجان» با استفاده از پیکره متنی و روش‌های پردازش زبان طبیعی خودکار، روشی برای ساخت هستی‌شناسی کشاورزی پویا بر اساس متن‌کاوی متون نمایه‌شده با اصطلاحنامه AGROVOC پیشنهاد کردند (Qian, Lan and Lijun 2007). «لیو، هوگان و کرولی» روش‌های پردازش زبان طبیعی[1] را مورد بررسی قرار داده و بر ضرورت این روش جهت پردازش حوزه‌های گستردهٔ اطلاعات امروزی تأکید کرده‌اند. نتیجه بررسی آن‌ها برتری رویکردهای نیمه‌خودکار موجود نسبت به رویکردهای کاملاً خودکار را نشان داد (Liu, Hogan and Crowley 2011). «واچر و شرودر» یک مولد نیمه‌خودکار هستی‌شناسی را با هدف استخراج اصطلاحات، تعاریف و روابط والد–فرزند از وب، «پابمد»[2] و یا فایل PDF ایجاد کردند. این طرح در قالب افزونه‌ای برای «پروتژ»[3] نهایی شد. عملکرد آن به این صورت است که عبارات اسمی را با توجه به طول و دفعات تکرار می‌یابد و شواهدی در خصوص شناسایی آن به‌عنوان یک اصطلاح، تعریف و یا رابطه معتبر ارائه می‌کند (Wächter & Schroeder 2010). «سیه» و همکاران طرح توسعه هستی‌شناسی نیمه‌خودکار از دانشنامه‌های یک حوزه موضوعی معین ارائه

1. natural language processing (NLP) 2. Pubmed 3. Protégé

کردند. رویکرد آن‌ها محدود به فضای دانشنامه چاپی بود و با شروع از واژه‌نامه و سپس افزودن زیرکلاس‌ها و روابط، تلاش کردند تا مفاهیم و روابط بین آن‌ها را استخراج کنند (Hsieh et al. 2011). «فابیان، واچر و شرودر» روشی را برای توسعهٔ هستی‌شناسی موجود، از طریق جست‌وجوی اینترنتی واژگان موجود هستی‌شناسی و متن‌کاوی متون بدون ساختار ارائه دادند (Fabian, Wächter and Schroeder 2012). «ساتسارونیس» و همکاران روشی نیمه‌خودکار برای استخراج تعاریف متداول از مفاهیم زیست‌پزشکی به‌منظور دستیابی به روابط بین مفاهیم ارائه دادند (Tsatsaronis et al. 2013).

«شیاکس» و همکاران روش جمع‌آوری سریع جزئیات تاکسونومی را با روش متن‌کاوی اجرا کردند. آن‌ها زیرساخت متن‌کاوی «آلویس»[1] را روی مجموعهٔ متون به‌دست‌آمده از جست‌وجو در «پابمد» و حوزه میکروبیولوژی مواد غذایی به‌کار گرفتند (Chaix et al. 2019). «شی، جین و سیه» روشی نیمه‌خودکار برای توسعه هستی‌شناسی پایه با استفاده از یک استراتژی ترکیبی شامل روش بالا به پایین (غنی‌سازی هستی‌شناسی موجود) و روش پایین به بالا (کشف مفاهیم و روابط از مجموعه اسناد خاص دامنه) پیشنهاد دادند (Chi, Jin and Hsieh 2019). به‌دلیل پیچیدگی بسیار زیاد در اصطلاحات و روابط تاکسونومی یا سلسله‌مراتبی در حوزه‌های فنی، ایدهٔ توسعهٔ کاملاً خودکار هستی‌شناسی هنوز عملیاتی به نظر نمی‌رسد. این است که در مقاله حاضر رویکرد نیمه‌خودکار اتخاذ شده و مشابه این انتخاب در مقالات (Hearst 1992، Missikoff, Velardi, and Fabriani 2003، Jiang and Tan 2005، Zhang et al. 2007، Hsieh et al. 2011، Chi, Jin and Hsieh 2019) مورد تأکید قرار گرفته است. بر اساس مقالات بررسی‌شده، اتخاذ رویکرد نیمه‌خودکار در راستای توسعه و بهبود مستمر هستی‌شناسی توسط بسیاری از محققان مورد توجه قرار گرفته است. چنین به نظر می‌رسد که این رویکرد همچنان برای به‌روزرسانی هستی‌شناسی بر مبنای اسناد و انتشاراتی که حجم انبوهی از متن را دربردارند، ترجیح بیشتری داشته باشد. از سوی دیگر، هر حوزه موضوعی با وجود اشتراک با دیگر حوزه‌های موضوعی، دارای مفاهیم، اصول، الگوها و نظریه‌های یکتاست. این مسئله به‌ویژه در حوزه‌های تخصصی پزشکی ملموس‌تر است. بنابراین، لازم است فعالیت‌هایی به‌منظور تحلیل حوزه صورت گیرد و ضمن شناسایی مفاهیم و روابط آن‌ها، چارچوبی برای بیان مفاهیم حوزه ارائه شود. بر این اساس، در مقالهٔ

1. Alvis

حاضر، برای انتخاب مفاهیم هستی‌شناسی از متن‌کاوی انتشارات علمی معتبر با رویکرد نیمه‌خودکار استفاده شده است.

گفته می‌شود که ایجاد هستی‌شناسی در هر حوزه نیازمند احاطه کامل موضوعی است؛ اما اغلب متخصصان علم اطلاعات و دانش‌شناسی، بر اساس چشم‌انداز و مأموریت شغلی، در مراکز اطلاعاتی تخصصی امروزی مانند پزشکی، صرفاً با احاطه نسبی به موضوع با مبحث هستی‌شناسی‌ها مواجهه دارند. بنابراین، طرح حاضر با هدف ارائه راه‌حلی سریع و کاربردی برای توسعه هستی‌شناسی بر مبنای اطلاعات متنی انبوه حاصل از جست‌وجوی هدفمند بسامد واژگان در پایگاه‌های دانش ارائه شده است تا با تسلط نسبی عامل انسانی، بیشترین پوشش موضوعی به دامنهٔ تحت بررسی را فراهم سازد. بخش بعد به مرور رویکردهای خودکار به‌کاررفته برای توسعهٔ هستی‌شناسی پرداخته و سپس، روش مقاله فعلی در طرح توسعه هستی‌شناسی نیمه‌خودکار ارائه شده است. در ادامه، به تشریح جزئیات مطالعه موردی که برای توسعه هستی‌شناسی «گلوکوم» انجام شده، پرداخته‌ایم.

۳. روش

این پژوهش از نوع بنیادی-کاربردی است که با هدف طراحی هستی‌شناسی به‌عنوان ابزار بازنمون دانش در یکی از حوزه‌های تخصصی پزشکی و بیماری‌های چشم (گلوکوم) یا آب سیاه انجام می‌گردد. بخش بنیادی آن نرم‌افزار کدنویسی‌شده‌ای است که با نام "TmbOnt_Alfa" معرفی شده است و می‌توان از آن برای پژوهش‌های دیگر نیز استفاده کرد. بخش کاربردی آن بالفعل بوده و می‌توان آن را برای سازماندهی، آموزش و ... به‌کار بست. بنابراین، هستی‌شناسی به‌دست‌آمده نه‌تنها قابلیت کاربردی دارد، بلکه از ابزار متن‌کاوی طراحی‌شده نیز می‌توان در گسترش هستی‌شناسی‌های سایر حوزه‌های پزشکی و غیرپزشکی هم استفاده کرد.

طبق نتایج حاصل از پیشینه‌ها، در پژوهش حاضر رویکرد نیمه‌خودکار در استخراج مفاهیم و روابط هستی‌شناسی اتخاذ شده است. بدین منظور، یک رابط کاربری نرم‌افزاری با نام "TmbOnt_Alfa" در بستر نرم‌افزار «متلب»[1] ایجاد شده که قادر است پردازش‌های لازم به روی متن انبوه را، بدون محدودیت در حجم و فرمت متن، به‌صورت متمرکز انجام دهد.

1. Matlab

فراهم‌سازی متن انبوه می‌تواند با استفاده از موتورهای جست‌وجوی متداول مانند «پابمد» انجام شود. نرم‌افزار "TmbOnt_Alfa"، بلوک فایل متنی را به‌عنوان ورودی دریافت کرده، محتوای آن را مطابق تنظیمات تعریف‌شده توسط کاربر پردازش نموده و واژگان مؤثر را بر اساس تعدد تکرار و میزان مرتبط بودن آن‌ها با کلمه سرشاخهٔ هستی‌شناسی، رتبه‌بندی می‌کند. بر اساس نتایج ارائه‌شده توسط این نرم‌افزار، کاربر خبره می‌تواند کلمات پیشنهادشده را بررسی نموده و در خصوص اینکه آیا این عبارت از نظر متنی مربوط به کلمه اصلی است یا خیر، تصمیم‌گیری کند. این رویکرد، افزون بر امکان ایجاد هستی‌شناسی جدید و مبنا می‌تواند به‌منظور ارتقای هستی‌شناسی‌های موجود و جهت به‌روزرسانی آن، مطابق با انتشارات علمی حوزهٔ دانش مربوط مورد استفاده قرار گیرد. در این پژوهش، طرح پیشنهادی در قالب مطالعهٔ موردی برای توسعه هستی‌شناسی «گلوکوم» به‌کار گرفته شده است. بلوک فایل متنی در این مطالعه، صرفاً از طریق جست‌وجوی هدفمند واژهٔ "Glaucoma" در «پابمد»، فراهم شده است. در شکل ۱، طرح کلی رویکرد توسعهٔ نیمه‌خودکار هستی‌شناسی نشان داده شده است. این طرح مشتمل بر سه ماژول اصلی شامل ماژول‌های پیش‌پردازش، ماژول پردازش اصلی و ماژول رابط کاربری است. ماژول پیش‌پردازش شامل زیرماژول توکن‌ساز (واحدسازی)[1]، برچسب‌گذار، نرمال‌ساز کلمات، حذف‌کنندهٔ علائم نگارشی، حروف اضافه و محدودیت طول یا تعداد کاراکتر کلمات است. فایل دادهٔ متنی، به‌صورت بلوک داده و از طریق رابط کاربری وارد فرایند پردازش می‌شود. زیرماژول توکن‌ساز، آرایهٔ متنی[2] را تبدیل به مجموعه کلمات متسلسل می‌کند و این امکان فراهم می‌آید که با استفاده از برنامه کامپیوتری، بروی تک‌تک کلمات کار پردازشی صورت پذیرد. توسط زیرماژول برچسب‌گذار، برای هر کدام از کلمات توکن‌سازی‌شده، بر اساس گرامر و موقعیت کلمه در جمله و ساختار آن، برچسب گفتاری معینی (فعل، فاعل، صفت و ...) تخصیص داده می‌شود. زیرماژول نرمال‌ساز، کلمات برچسب‌گذاری‌شده را به‌صورت ریشه‌ای آن بازمی‌گرداند تا تغییرات ناشی از اعمال زمان در افعال و صورت جمع/ منفرد کلمات را حذف و همسان نماید. زیرماژول حذف‌کننده علائم نگارشی، مواردی از قبیل نقطه، ویرگول، ممیز و ... را از ساختار جمله حذف می‌کند. در ادامه، زیرماژول حذف‌کننده حروف اضافه، اقدام به حذف حروف اضافه از قبیل "the, at, of, ..." می‌نماید. سرانجام، این امکان برای کاربر فراهم شده است که در صورت نیاز، بتواند تعداد

1. tokenized
2. text string

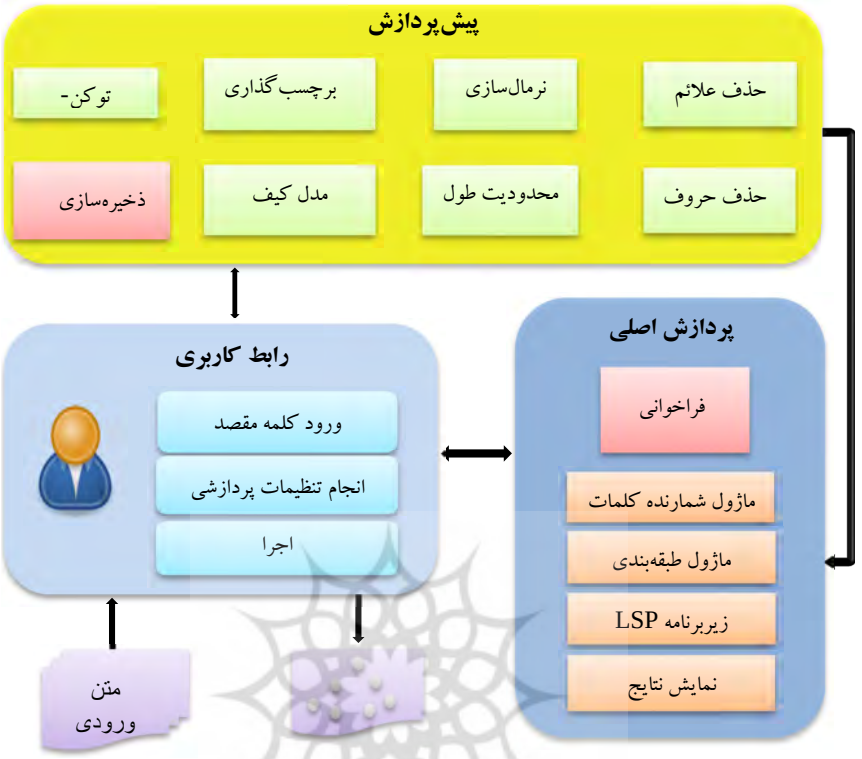
حداقلی کاراکترهای کلمات را معین کند تا کلماتی که طول کوتاهی دارند، از فرایند پردازش اصلی حذف گردند. کلیه این زیرماژول‌ها به شکل انتخابی برای کاربر تدارک دیده شده تا در صورت صلاحدید بتواند در پردازش‌های خاص، برخی از زیرماژول‌ها را تغییر دهد یا عملیاتی نکند. هرچند که حذف هر یک از مراحل پیش‌پردازش منجر به افزایش حجم داده و به‌دنبال آن، منجر به افزایش زمان پردازش خواهد شد. پس از اتمام عملیات پیش‌پردازش و در فرایند پردازش تحت نظارت کاربر، ساختار سلسله‌مراتبی مفاهیم و برخی روابط هستی‌شناسی معین می‌شود. طرح پیشنهادی در عین سادگی، دسترس‌پذیری، و مقرون‌به‌صرفه بودن، در ارزیابی با هستی‌شناسی‌ها و اصطلاحنامه‌های موجود پزشکی، ساختاری قابل قبول ارائه می‌کند و امکان به‌روزرسانی و توسعهٔ ساختار هستی‌شناسی‌های موجود را بر مبنای آخرین گزارشات تخصصی فراهم می‌آورد. این طرح می‌تواند توسط کاربر غیرمتخصص با دانش دامنهٔ مورد بررسی که اطلاعاتی کلی در خصوص هستی‌شناسی دارد نیز به‌کار گرفته شود.

۳-۱. گردآوری داده از دامنه

بر اساس موضوع هستی‌شناسی، دانش ارزشمند باید از منابع معتبر از قبیل دایره‌المعارف‌ها، کتب مرجع، گزارش‌های علمی و یا مقالات علمی معتبر گردآوری شود. انتخاب منبع جامع برای ورودی فرایند پردازش می‌تواند ساختار هستی‌شناسی کامل و جامع را به‌دست دهد. امروزه، به لطف مجلات تحت وب و پایگاه‌های اطلاعاتی علمی، امکان استفادهٔ گسترده از مقالات علمی برای اعتباربخشی به دادهٔ متنی امکان‌پذیر است.

در پژوهش حاضر، شناسایی دامنه صرفاً با به‌کارگیری کلمهٔ سرشاخهٔ هستی‌شناسی[1] و تجمیع فایل مقالات علمی حول این کلیدواژه صورت می‌پذیرد. بدین نحو که فایل متنی انبوه مورد نیاز برای ورودی فرایند متن‌کاوی، با انتخاب کلیدواژهٔ اصلی دامنه هدف در «پابمد» و ذخیره‌سازی مرتبط‌ترین نتایج در قالب تک‌فایل متنی تهیه و استخراج می‌گردد. تأمین متن ساده و خالص (عاری از آدرس‌های اینترنتی، جداول، تصاویر و ...) منجر به کاهش زمان پردازش می‌شود. از این رو، برای سهولت پردازش، لازم است فایل متنی خام در ساده‌ترین قالب و به فرمت «اسکی»[2] تهیه شود.

1. Target OR Seed
2. ASCII
2.



شکل ۱. طرح کلی توسعهٔ هستی‌شناسی نیمه‌خودکار

۳-۲. پردازش داده

رویکردهای توسعهٔ هستی‌شناسی خودکار بایستی بتوانند راهکار مناسبی برای دو چالش اصلی شناسایی مفاهیم و شناسایی روابط ارائه دهند (Liu, Hogan and Crowley 2011). این راهکارها را می‌توان به دو دستهٔ کلی شامل روش‌های نمادین[1] و یا روش‌های آماری تفکیک کرد. فارغ از نوع راهکار، به‌طور کلی استخراج مفاهیم و مترادف‌ها نسبت به استخراج روابط با چالش کمتری مواجه است (Liu, Hogan and Crowley 2011; Maedche, Pekar & Staab 2003).

روش نمادین اغلب مبتنی بر رویکرد تطابق الگوی واژگانی است که توسط Hearst (1992) ارائه شده و سعی دارد رابطه اصطلاحات را از طریق سبک‌های واژگانی که از قبل

1. symbolic

تبیین شده‌اند، شناسایی کند. شکل کلی و بنیادین الگوهای ترکیبی[1] به‌صورت قوانین «اگر-آنگاه» در جدول ۱، نشان داده شده است.

از سوی دیگر، در روش‌های آماری از خوشه‌بندی و یا یادگیری ماشین برای شناسایی مفاهیم جدید و یافتن روابط آن‌ها استفاده می‌شود (Alfonseca & Manandhar 2002). روش‌های نمادین اگرچه موفقیت بیشتری در شناسایی رابطه بین مفاهیم در قیاس با روش‌های آماری نشان می‌دهند، اما روش‌های آماری به‌دلیل توسعه منظم تکنیک‌های یادگیری ماشین می‌توانند در راستای توسعهٔ سیستم‌های خودکار کارایی بهتری داشته باشند (Liu, Hogan and Crowley 2011).

در تحقیق حاضر، ابزاری جهت پردازش متن با شیوه ترکیبی آماری-نمادین ارائه شده است. ماژول پردازش اصلی، داده‌هایی را که پردازش اولیه شده‌اند، به‌صورت مستمر فراخوانی کرده و مطابق تنظیمات تعریف‌شده توسط کاربر، به‌صورت نهایی پردازش می‌کند. ماژول اصلی مشتمل بر زیربرنامه‌های پردازش اصلی است که عملکرد جست‌وجو، خوشه‌بندی و LSP را اجرا می‌کند. این ماژول، در انتهای فرایند پردازش، قابلیت ارائه گزارش‌ها و تولید نمودار را نیز دارد.

جدول ۱. برخی از الگوهای واژگان در قالب قوانین منطقی اگر-آنگاه (Hearst 1992)

ردیف	نمونه الگوهای واژگانی
۱	**IF** ***NP such as {NP_1 (and|or) NP_2, ..., NP_i, ..., NP_n}*** **Then** ***NP_i ISA NP*** **مثال:** **… eye disease such as cataracts, glaucoma, macular degeneration, retinal detachment, and vision loss due to diabetes and high blood pressure is that …**[1] ◇ **Glaucoma ISA eye disease** ◇ **Macular degeneration ISA eye esaesid**

1. lexica-syntactic patterns (LSP)

1. https://williamsburgeye.com/patient-education/eye-disease/

ردیف	نمونه الگوهای واژگانی
۲	**IF** ***such NP as {NP_1 (and|or) NP_2, ..., NP_i, ..., NP_n}*** **Then** ***NP_i ISA NP*** **مثال:** **… works by such authors as Herrick, Goldsmith, and Shakespeare… [17]** **◇ Herrick ISA author** **◇ Goldsmith ISA author**
۳	**IF** ***{NP_1 (and|or) NP_2, ..., NP_i, ..., NP_n} or/and other NP*** **Then** ***NP_i ISA NP*** **مثال:** **This medication is used alone or with other medications to treat high pressure inside the eye due to glaucoma or other eye diseases (e.g., ocular hypertension)[1].** **◇ Glaucoma ISA eye disease**
۴	**IF** ***NP including {NP_1 (and|or) NP_2, ..., NP_i, ..., NP_n}*** **Then** ***NP_i ISA NP*** **مثال:** **◇ This suggests that a diabetic retinopathy screening program needs to detect and report other eye disease, including glaucoma and macular disease.[2]** **◇ Glaucoma ISA eye disease** **◇ Macular ISA eye disease**
۵	**IF** ***NP specially {NP_1 (and|or) NP_2, ..., NP_i, ..., NP_n}*** **Then** ***NP_i ISA NP*** **مثال:** **… most European countries, specially France, England, and Spain. [17]** **◇ France ISA European country** **◇ England ISA European country**

۳-۲-۱. شناسایی اصطلاحات

در متن‌های حجیم، شناسایی اصطلاحات و مفاهیم به‌طور معمول توسط روش‌های مبتنی بر متن‌کاوی یا روش‌های پردازش زبان طبیعی انجام می‌شود (Alfonseca &

1. https://www.rxlist.com/fdb/drugs/60722/pilosol-ophthalmic-eye-drug.htm
2. https://www.ncbi.nlm.nih.gov/pubmed/27328169

Manandhar 2002 و Chi, Jin, and Hsieh 2019). فرایند شناسایی اصطلاح در رویکرد حاضر، توسط یک زیربرنامه هدفمند محاسبه بسامد واژگانی انجام می‌شود. این زیربرنامه برای یافتن پیش‌واژه‌های اصطلاح مورد جست‌وجو توسعه داده شده است. زیربرنامه مذکور، به‌عنوان بخشی از برنامه اصلی، پیش‌واژه‌ها را بر اساس میزان تکرار و نزدیک بودن به واژهٔ اصلی، اولویت‌بندی می‌کند. در واقع، بسامد واژگانی که به‌عنوان پیش‌واژه کلمه هدف قرار گرفته‌اند، محاسبه می‌شود. این است که در این پژوهش، افزون بر توجه به بسامد واژگان، کیفیت واژگان و ارتباط آن‌ها که سرانجام، چیدمان ساختار هستی‌شناسی را شکل می‌دهد نیز یکی از مهم‌ترین مراحل پردازش بوده است. طبق پژوهش «احمدی»، محل استخراج واژه در مدرک، توجه به مسائل زبانی و ترکیب واژگان، و ارتباط معنایی از جمله موارد مطرح در انتخاب واژه با کیفیت است (۱۳۹۴). مراحل مذکور به نحوی برنامه‌ریزی شده است که واژه‌ها از چکیده، که عصاره مطالب مدرک را دربردارد، استخراج شده و پس از برچسب‌گذاری[1] بر اساس همبستگی با واژهٔ هدف و با در نظر گرفتن پیش‌واژه یا پس‌واژه ارزش‌گذاری گردند.

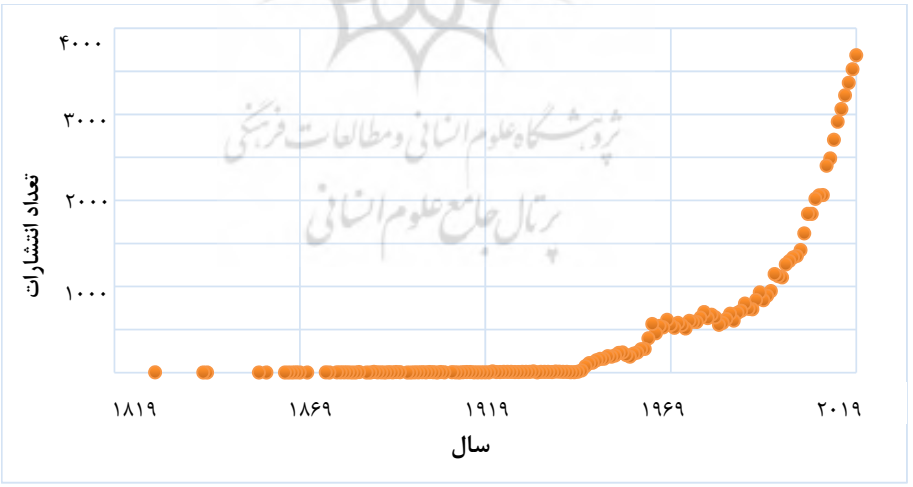
۳-۲-۲. شناسایی روابط

فرایند شناسایی ارتباط در این پژوهش بر مبنای دو استراتژی مجزا شامل شناسایی واژه پسین با استفاده از منطق جست‌وجوی پس‌رونده[2] و شناسایی واژه پیشین با استفاده از منطق جست‌وجوی پیش‌رونده نسبت به واژهٔ هدف استوار است. رویکرد شناسایی واژه پسین مبتنی بر پردازش اسم مرکب است که پیش‌تر توسط Morin (2004) مورد بحث قرار گرفته است. فرض اصلی این رویکرد بر این است که اصطلاحات حاصل از ترکیب چند واژه با یک واژهٔ اصلی، اشاره به واژه اخص‌تری نسبت به شکل منفرد واژهٔ اصلی دارند. برای مثال اصطلاح "Open-angle Glaucoma" اشاره به اصطلاح اخص‌تر در قیاس با اصطلاح "Glaucoma" دارد و با استفاده از جست‌وجوی پس‌رونده نسبت به واژهٔ هدف "Glaucoma" در بین عباراتی که در ترکیب با این واژه در متن‌ها وجود دارد، شناسایی می‌شود. در همین حال، رویکرد پیش رو، متکی به روش استخراج رابطه LSP از Hearst (1992) است که در جدول ۱، ارائه شده است. این دو رویکرد، به‌صورت دو زیربرنامه مجزا برای شناسایی روابط، کدنویسی شده‌اند.

1. tagging
2. backward chaining

٤. مطالعهٔ موردی: «گلوکوم»

طرح پیشنهادشده برای توسعهٔ هستی‌شناسی، از جهت ارزیابی قابلیت‌های عملکردی آن در شناسایی اصطلاحات و روابط بین آن‌ها، در دامنهٔ اطلاعاتی «گلوکوم» مورد بررسی قرار گرفت. «مرکز کنترل و پیشگیری از بیماری‌ها»[1] در صفحه بیماری‌های چشم بیان کرده است که بیش از ۴/۲ میلیون آمریکایی با سن ۴۰ سال و بیشتر رسماً نابینا هستند. «گلوکوم» دومین عامل نابینایی بعد از آب مروارید در دنیاست. بیماری «گلوکوم» ناشی از آسیب ساختاری به عصب‌های بینایی است و دارای الگوی معینی در آسیب رأس عصب‌های بینایی است (Foster et al. 2002). بهبود هستی‌شناسی‌های موجود در این زمینه، به جامع‌ترشدن هستی‌شناسی و فراگیرتر شدن آن بر اساس آخرین یافته‌های پژوهشی کمک شایانی خواهد کرد. جست‌وجوی اولیه در «پابمد» حاکی از حجم وسیع پژوهش‌های انجام‌شده طی سال‌های اخیر در این حوزه است. در نمودار ارائه‌شده در شکل ۲، حجم انتشارات این حوزه بر اساس سال نمایش داده شده که نشانگر رشد شتابان این دسته از انتشارات طی سال‌های اخیر است. حجم انتشارات و میانگین نرخ رشد آن به‌وضوح رشد شتابان مرزهای علمی حوزهٔ تخصصی «گلوکوم» را نمایش می‌دهند و از این منظر می‌تواند گزینه مناسبی برای بررسی امکان توسعهٔ هستی‌شناسی آن بر مبنای پژوهش‌های علمی باشد.



شکل ۲. توزیع فراوانی انتشارات با محوریت «گلوکوم» بر اساس سال

1. Centers for Disease Control and Prevention (CDC)

٤-١. ایجاد فایل متن انبوه[1]

در طرح پیشنهادی، تهیه فایل متن انبوه از بین پژوهش‌های علمی معتبر و جدید صورت گرفته است که با به‌کارگیری کلیدواژه اصلی در جست‌وجوی هدفمند این دسته از انتشارات به‌دست می‌آید. در مطالعهٔ موردی حاضر، جست‌وجوی عبارت "Glaucoma" در «پابمد»، بیش از ۷۰،۰۰۰ نتیجه را گزارش داد که بیش از ۱۴،۰۰۰ مورد دارای خلاصه مقاله و یا متن کامل در این پایگاه اطلاعاتی بودند. از این میان، ۱۰،۰۰۰ رکورد مرتبط‌ترِ نخست که بالغ بر ۳،۰۰۰،۰۰۰ عبارت و اصطلاح است، برای تأمین فایل متن انبوه تجمیع گردید. این فایل با فرمت txt ذخیره شد تا در پردازش توسط نرم‌افزار "TmbOnt_Alfa" مورد استفاده قرار گیرد.

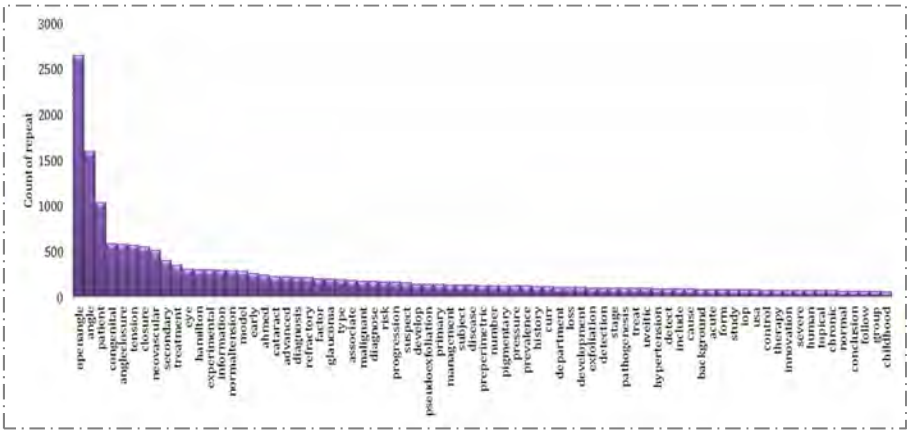
٤-٢. پردازش مقدماتی

فایل متن انبوه در رابط کاربر "TmbOnt_Alfa"، وارد فرایند پردازش گردید. ماژول پردازش مقدماتی با استفاده از تنظیمات پیش‌فرض اقدام به اعمال زیربرنامه‌های نرمال‌ساز، حذف‌کننده علائم نگارشی و حذف‌کننده حروف اضافه نمود. افزون بر این، در این مرحله، طول اصطلاحات نیز بر مبنای حداقل طول دو کاراکتر تعیین شد و برچسب‌های گفتاری برای استفاده‌های آتی اضافه گردید. خروجی مرحله پردازش مقدماتی به‌منظور استفاده مکرر در فازهای بعدی، به‌صورت فایل متغیر در فضای «متلب» ذخیره شد.

٤-٣. پردازش اصلی

سطح اول ساختار هستی‌شناسی با درج موجودیت هدف، یعنی واژهٔ «گلوکوم» شکل گرفت. این ماژول مبتنی بر این فرض است که بیشترین شانس عضویت در لایهٔ اول هستی‌شناسی به پیش‌واژه‌های سطح اول موجودیت هدف اختصاص دارد. بدین ترتیب، تمام پیش‌واژه‌های منسوب به واژهٔ هدف، تجمیع شده و برای ارزیابی توسط متخصص بازنمایی شدند. در خصوص واژهٔ «گلوکوم»، خروجی حاصل از این ماژول در شکل نشان داده شده است.

1. corpus file



شکل ۳. نتیجه جست‌وجوی پیش‌واژه لایه اول برای موجودیت «گلوکوم»

با توجه به توزیع فروانی پیش‌واژه‌ها، می‌توان با تقریب مناسبی اظهار کرد که بعد از ۱۰۰ رکورد اول، میزان بروز اصطلاحات پیش‌واژه در سطح فراوانی بسیار کمتری نسبت به رکوردهای اولیه هستند. با این حال، به جهت حصول اطمینان حداکثری برای از دست ندادن هیچ موجودیت احتمالی، در اینجا ۲۵۰ کلمه برتر برای تولید سطح اول هستی‌شناسی انتخاب گردید. در این مرحله عامل انسانی می‌تواند واژگان با بسامد بالا را به‌عنوان گزینه‌های بعدی سرشاخه انتخاب کند. این رویه تا زمانی که به محدودیت تعیین‌شده توسط کاربر (۲۵۰ مورد در این پژوهش) و شانس ناچیز برای انتخاب شدن (بر مبنای قضاوت کاربر) برسد، ادامه می‌یابد.

٤-۳-۱. شناسایی روابط با استفاده از LSP

افزون بر رویکرد تحلیل بسامد واژگان، که در خصوص واژهٔ «گلوکوم» شاکلهٔ کلی ساختار هستی‌شناسی را برپا کرد، کارایی شیوهٔ LSP نیز بررسی گردید تا امکان شناسایی روابط بیشتر بین واژگان ارزیابی شود. در به‌کارگیری این رویکرد، افزون بر الگوهای LSP نمایه‌شده در جدول، الگوهای مضاعفی نیز متناسب با واژهٔ گلوکوم که در اثنای جست‌وجو در متون تخصصی غالب به نظر می‌رسیدند، انتخاب و مطابق با جدول به فرایند جست‌وجو افزوده شدند.

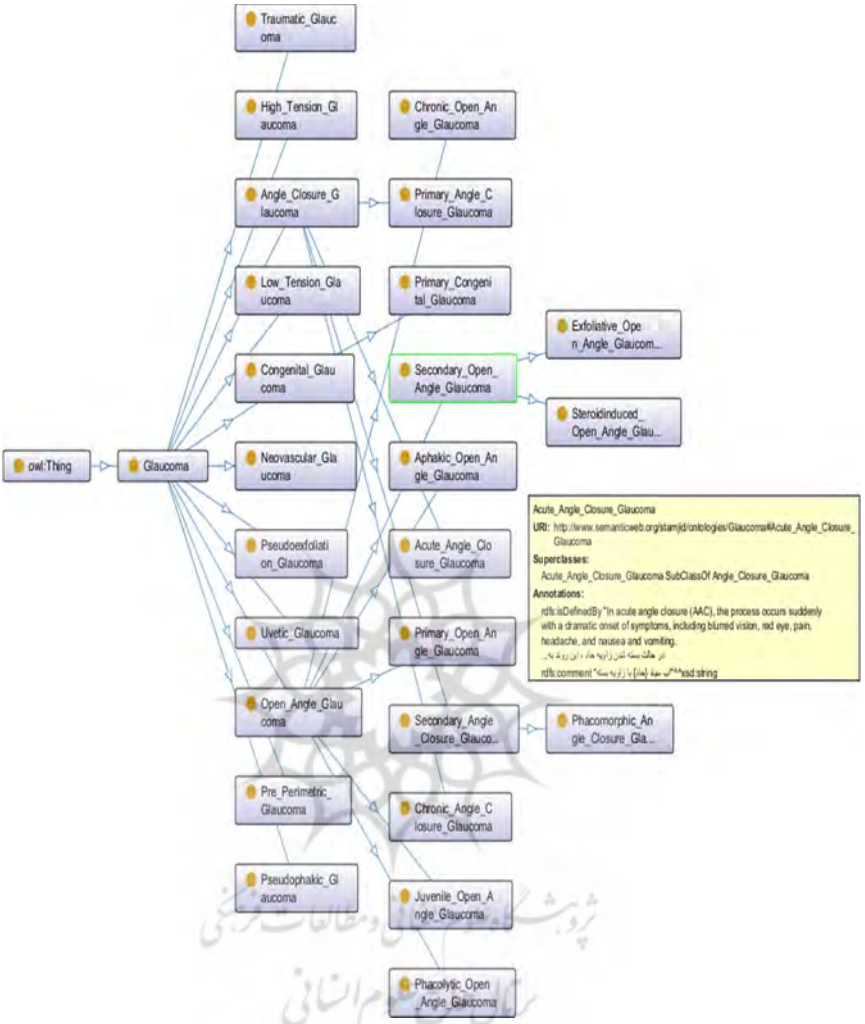
جدول ۲. الگوهای مضاعف LSP به‌کاررفته برای فاز مطالعهٔ موردی

الگو	مثال
NP1 is a type of NP	Pigmentary glaucoma is a type of secondary open-angle glaucoma…
NP1 also known as NP	Normal-tension glaucoma (NTG), also known as low tension or normal pressure glaucoma.
NP1 is a form of NP	Steroid-induced glaucoma is a form of secondary open angle glaucoma that results from the use of steroids.
NP1 is the most common type of NP	Exfoliative glaucoma is the most common type of secondary open-angle glaucoma worldwide.
NP also referred to as NP1, NP2 or NP3…	Childhood glaucoma, also referred to as congenital glaucoma, pediatric, or infantile glaucoma, occurs in babies and young children.

جست‌وجوی الگوهای LSP از طریق زیربرنامه افزوده‌شده با همین نام به برنامه اصلی صورت می‌گیرد که این زیربرنامه کل ساختار فایل توکن‌شده را برای یافتن الگوهای مشابه با الگوی معرفی‌شده پالایش می‌کند.

٤-۳-۲. نهایی‌سازی ساختار واژگان

ساختار اولیه ایجادشده نیازمند انجام بررسی توسط متخصص حوزه دانش است تا بتواند پس از پالایش و اعمال تغییرات لازم، به‌صورت نهایی تکمیل گردد. مفاهیم حاصل بر اساس روابط شناسایی‌شده شامل Is a, also known, Is a form of, Is the most common type, also referred to,... در نرم‌افزار «پروتژ ۵» وارد گردید. به‌منظور قابلیت استفاده هستی‌شناسی در زبان فارسی، معادل فارسی مفاهیم در قسمت "comment" نرم‌افزار «پروتژ» وارد شد و همین مفاهیم برای قیاس با مراجع فارسی مورد استفاده قرار گرفت. تعاریف انگلیسی و فارسی هر مفهوم در قسمت "definition" وارد شد.



شکل ٤. ساختار نهایی‌شده هستی‌شناسی در «پروتژ»

۵. تحلیل نتایج

رویکرد پیشنهادشده در مقاله حاضر توانسته است موضوع تنوع و پیچیدگی بسیار بالای واژگان و مفاهیم در پالایش حجم انبوه داده‌های متنی را رفع نماید. زمانی که منبع داده‌ها بانک مقالات علمی معتبر باشد، دشواری کار متن‌کاوی با توجه به ثقیل‌تر بودن ادبیات به‌کاررفته در چنین متن‌هایی به مراتب بیشتر می‌شود. با وجود این، در صورت غلبه بر این چالش می‌توان انتظار داشت که ساختار حاصل بتواند دامنه بیشتری

از واژگان تخصصی و علمی را پوشش داده و به کامل‌تر شدن هستی‌شناسی‌های موجود کمک مؤثری نماید. بر اساس نتایج حاصل از مرحله ارزیابی علمی، استفاده از ماژول پردازش اولیه و به‌کارگیری فیلترهای مناسب در آن، کارایی بسیار مناسبی در جهت حذف محاسبات غیرضروری نشان داد که منجر به کاهش بسیار محسوس حجم و زمان محاسبات گردید. در خصوص مطالعهٔ موردی صورت‌گرفته برای استخراج مفاهیم بیماری «گلوکوم»، حدود ۱۰،۰۰۰ خلاصه مقاله از بانک اطلاعاتی «پابمد» وارد تحلیل شد که مشتمل بر حدود ۳/۸ میلیون آیتم توکن‌شده بود. ماژول پردازش اولیه این حجم از داده را به حدود ۱/۸۷ میلیون کاهش داد که به مفهوم کاهش ۵۰ درصدی حجم محاسبات غیرضروری است. نکتهٔ مهم این است که استفاده از چنین فیلتری می‌تواند به حذف عناصری منجر شود که در پردازش LSP کارایی دارند. بنابراین، در صورت به‌کارگیری الگوریتم LSP، با علم به افزایش زمان پردازش، لازم است از به‌کارگیری فیلترهایی که حروف اضافه را حذف می‌کنند، اجتناب شود. در شناسایی واژگان و روابط بین آن‌ها استفاده از رویکرد بسامد و رتبه‌بندی واژگان، در قیاس با روش LSP توانست کارایی بهتری داشته باشد و نیز توانست با شناسایی تعداد قابل قبولی از مفاهیم و روابط ساختار کلی هستی‌شناسی را توسعه دهد. این در حالی است که روش LSP صرفاً توانست لایه‌های اولیه و بالادستی هستی‌شناسی را پوشش دهد. اتکا به نیروی انسانی در مرحلهٔ نهایی‌سازی ساختار و بررسی شاخه‌های مترادف ضروری است و اگر بتوان این موضوع را با برنامه‌نویسی جامع‌تر (حداقل برای آشکارسازی و تلفیق شاخه‌های مترادف) مرتفع نمود، گام بلندی در جهت توسعهٔ تمام‌خودکار هستی‌شناسی خواهد بود.

به‌منظور ارزیابی میزان انطباق نتایج به‌دست‌آمده با ساختارهای معتبر موجود، معیاری کمّی برای قیاس با هستی‌شناسی‌های بازنمایی‌شده برای بیماری گلوکوم در بانک‌های هستی‌شناسی بیماری‌های انسانی توسعه داده شد. مقایسه با ساختار «سرعنوان‌های پزشکی مش»[1] و «اصطلاحنامه و توصیفگرهای پزشکی فارسی» نیز انجام گردید. این قیاس بین مفاهیم، اصطلاحات و جایگاه سلسله‌مراتبی آن‌ها حاصل از برنامه کدشده، با مفاهیم، اصطلاحات و جایگاه هر اصطلاح در ساختار سلسله‌مراتبی هستی‌شناسی‌های معتبر موجود و اصطلاحنامه پزشکی انجام شده است. مطابق رابطهٔ (۱)، درجهٔ انطباق با به‌کارگیری

1. MESH Subject Headings

فرمولی ساده که بر اساس میزان انطباق و موقعیت انطباق، امتیازدهی می‌کند، به مقیاس کمّی تبدیل شد. این قیاس هم تشابه در واژگان و هم تشابه در موقعیت واژه در ساختار هستی‌شناسی را مدنظر قرار داده است. اینکه دو واژه در دو ساختار درختی تحت قیاس، در موقعیت‌های همسان قرار گیرند، مؤید آن است که روابط نیز به‌درستی شناسایی شده‌اند. با ذکر این نکته که در خود ساختارهای ارائه‌شده در پایگاه‌های معتبر نیز اختلافاتی وجود دارد، قیاس انجام‌شده توانسته است کلیات ساختار را صحه‌گذاری کرده و تائید نماید که رویکرد جدید توانسته گامی در مسیر سهولت ایجاد و توسعه نیمه‌خودکار هستی‌شناسی بردارد. با استفاده از این رابطه، میزان انطباق با «مش» ۸۰ درصد، انطباق و پوشش با «اصطلاحنامه پزشکی فارسی» ۱۰۰ درصد و در قیاس با ساختارو واژگان «توصیفگرهای پزشکی فارسی» ۱۰۰ درصد محاسبه گردید. همچنین، در مقایسه با هستی‌شناسی بازنمایی‌شده در «هستی‌شناسی بیماری‌های بیماری‌های انسانی»[1] میزان انطباق ۵۷ درصد و در قیاس با هستی‌شناسی نمایه‌شده در «هستی‌شناسی BioAssay»، ۹۱ درصد به‌دست آمد. افزون بر آن، رویکرد پیشنهادی توانست به‌طور میانگین حدود ۳۰ درصد به دامنهٔ واژگان ساختار هستی‌شناسی موجود بیفزاید. این توسعهٔ قابل توجه به‌دلیل توانمندی این شیوه برای پردازش حجم انبوه از متن‌های تخصصی است و با توجه به اینکه این متون در فضای ادبیاتی دانشگاهی و بر اساس به‌روزترین نتایج محققان تحریر می‌شود، می‌تواند نویدبخش قابلیت نسبی روش پیشنهادی برای بهبود هستی‌شناسی‌های موجود باشد. همچنین بر اساس نتایج به‌دست‌آمده، استفاده از این رویکرد با درجهٔ مناسبی از اطمینان می‌تواند برای تهیه فرم نخست ساختار هستی‌شناسی در دامنه‌های متنوع و خاص مورد استفاده قرار گیرد. بر اساس نتایج حاصل از فاز تجربی، تهیه ساختار هستی‌شناسی گلوکوم با میزان انطباق ذکرشده صرفاً با صرف زمانی حدود پنج ساعت و توسط نیروی انسانی با سطح دانش و مهارت متوسط صورت پذیرفت که با توجه به حجم دادهٔ پردازش‌شده و کیفیت نتایج، می‌تواند راهبردی مناسب برای توسعه اولیهٔ ساختار هستی‌شناسی و نیز نگه‌داشت ساختارهای موجود با صرف حداقلی منابع باشد. این دستاورد در قیاس با کار مشابه (Chi, Jin and Hsieh 2019) که به‌کارگیری ۱۶ نفر طی دو ساعت را گزارش داده‌اند، وابستگی به نیروی انسانی را تا شش برابر کاهش داده است. علت اصلی این بهبود، امکانات سطح

1. Human Disease Ontology (DOID)

بالای ماژول‌های پردازش متن به‌کار گرفته‌شده در نرم‌افزار TmbOnt_Alfa و عدم نیاز به بررسی متون ورودی به تحلیل توسط نیروی انسانی خبره است. سرانجام، با توجه به اینکه کار حاضر در بستر کدنویسی «متلب» توسعه داده شده که مجهز به ماژول‌های پردازشی بسیار گسترده در حوزهٔ هوش مصنوعی و یادگیری ماشین است، می‌توانند بر خلاف کارهای مشابه که از بسترهای متن‌کاوی صرف استفاده نموده‌اند (مانند کار انجام‌شده در پژوهش Chaix et al. (2019) که از بستر مستقل متن‌کاوی Alvis استفاده شده)، انعطاف بسیار بالایی در توسعه‌های آتی داشته باشد.

$$\boldsymbol{Compatibility\ Degree} = \frac{\boldsymbol{a} * \mathbf{3} + \boldsymbol{b} * \mathbf{2}}{\boldsymbol{all} * \mathbf{3}} * \mathbf{100} \quad (۱)$$

a: واژه‌های مشابه در موقعیت‌های همسان

b: واژه‌های مشابه در موقعیت‌های غیرهمسان

شکل ۵. ساختار درختی واژه «گلوکوم» در «مش»

٦. نتیجه‌گیری

در این مقاله، راهکاری جهت توسعهٔ هستی‌شناسی به روش نیمه‌خودکار ارائه گردید که بستری مناسب برای ایجاد یک هستی‌شناسی جدید و یا توسعهٔ هستی‌شناسی موجود

بـر مبنـای انتشـارات معتبـر علمـی نمایه‌شـده در پایگاه‌هـای اطلاعـات علمـی فراهـم می‌کنـد. در ایـن روش، متـن انبـوه ورودی در قالـب فایـل متنـی فراخوانـی شـده و بـر اسـاس تنظیمـات کاربـر پـردازش می‌شـود. حاصـل پـردازش اولیـه، به‌صـورت اطلاعـات میانـی ذخیـره می‌گـردد و کاربـر خبـره می‌توانـد بـدون نیـاز بـه صـرف زمـان بـرای پـردازش مجـدد، و صرفـاً بـا وارد نمـودن کلمـهٔ مـورد نظـر، نسـبت بـه فراخوانـی محتـوای پیش‌پردازش‌شـده و اسـتخراج کلمـات مرتبـط بـر اسـاس اولویـت اقـدام نمایـد. فرمـت فایـل ورودی بـرای متـن می‌توانـد به‌صـورت txt، pdf و یـا سـایر فرمت‌هـای متنـی بـوده و کاربـر انسـانی صرفـاً بـرای تصمیم‌گیـری و مراحـل نهایی‌سـازی درگیـر کار می‌شـود. برتـری رویکـرد نیمه‌خـودکار در توسـعه و بهبـود هستی‌شناسـی‌ها نسـبت بـه روش‌هـای دسـتی کـه صرفـاً مبتنـی بـر عامـل انسـانی هسـتند، از نظـر صرفه‌جویـی در هزینـه و زمـان قابـل توجـه اسـت. بـر مبنـای ایـن رویکـرد و در توسـعه‌های آتـی، ایـن امـکان وجـود خواهـد داشـت کـه بـا کدنویسـی فعالیت‌هـای کاربـر انسـانی، امـکان ایجـاد و بررسـی رویکردهـای کامـلاً خـودکار فراهـم گـردد.

بـر اسـاس یافته‌هـای حاصـل از مطالعـهٔ مـوردی توسـعهٔ هستی‌شناسـی «گلوکـوم»، هستی‌شناسـی توسـعه داده‌شـده توسـط رویکـرد پیشـنهادی به‌طـور میانگیـن بیـش از ۷۰ درصـد بـا هستی‌شناسـی‌های بازنمایی‌شـده در پایگاه‌هـای معتبـر هستی‌شناسـی بیماری‌هـای انسـانی[1] انطبـاق داشـت و به‌طـور میانگیـن بیـش از ۳۰ درصـد واژگان جدیـد بـرای افـزودن بـه دامنـهٔ واژگان آن‌هـا را فراهـم کـرده اسـت. بسـتر فراهم‌شـده در پژوهـش حاضـر، صرفـاً محـدود بـه دامنـهٔ اطلاعاتـی خاصـی نبـوده و بـرای طراحـی و توسـعه هستی‌شناسـی در هـر حـوزه‌ای قابـل اسـتفاده اسـت. همچنیـن، هستی‌شناسـی به‌دسـت‌آمده بـر خـلاف هستی‌شناسـی‌هایی کـه صرفـاً بـرای یـک کارکـرد طراحـی می‌شـوند، قابلیـت اسـتفاده در کاربردهـای مختلـف ماننـد فهرست‌نویسـی، نمایه‌سـازی، آمـوزش و ... را دارد.

بـرای پژوهش‌هـای آتـی، پیشـنهاد می‌شـود کـه ایجـاد هستی‌شناسـی در سـایر حوزه‌هـای موضوعـی بـا اسـتفاده از رویکـرد متـن‌کاوی، همچنیـن قیـاس هستی‌شناسـی‌های به‌دسـت‌آمده از روش متـن‌کاوی بـا هستی‌شناسـی‌هایی کـه مفاهیـم آن‌هـا صرفـاً از اصطلاحنامه‌هـا حاصـل شـده‌اند، در دسـتور کار قـرار گیـرد. افـزون بـر آن، بـه نظـر می‌رسـد کـه بـر مبنـای رویکـرد پیشـنهادی در پژوهـش حاضـر، امـکان حـذف نقـش ناظـر انسـانی (یـا رسـاندن آن بـه حداقـل

1. bio-ontologies

ممکن) می‌تواند بیشتر قابل اجرا باشد. این ایده می‌تواند گام مؤثری در راهبرد به‌سوی تولید کاملاً خودکار ساختارهای هستی‌شناسی در پژوهش‌های آینده باشد.

فهرست منابع

احمدی، حمید. ترسیم و تحلیل شبکه مفهومی و هستی‌شناسی ساختار دانش حوزه علم سنجی ایران بر اساس رویکرد تحلیل حوزه. رساله جهت دریافت درجه دکتری. دانشگاه چمران اهواز. ۱۳۹۴.

حسینی بهشتی، ملوک‌السادات. ۱۳۹۲. *ساخت‌واژه: اصطلاح‌شناسی و مهندسی دانش*. تهران: پژوهشگاه علوم و فناوری اطلاعات ایران؛ چاپار.

فتحیان دستگردی، اکرم. ۱۳۸۹. مقایسه کارآمدی اصطلاحنامه و هستی‌شناسی در بازنمون دانش و بازیابی مفاهیم. پایان‌نامه جهت دریافت درجه کارشناسی ارشد، دانشگاه فردوسی مشهد، دانشکده علوم تربیتی و روان‌شناسی.

معصومی، رحیم، امین معصومی گنجگاه، حبیب اوجاقی، عیسی بنازاده. ۱۳۹۱. توزیع فراوانی علل اختلالات بینائی در افراد بالای ۴۰ سال مراجعه‌کننده به درمانگاه چشم بیمارستان علوی طی سال‌های ۸۵-۱۳۸۴. *مجله دانشگاه علوم پزشکی و خدمات بهداشتی درمانی اردبیل* ۱۲ (۲): ۱۶۶-۱۷۲.

References

Alfonseca, E, & S. Manandhar. 2002. An Unsupervised Method for General Named Entity Recognition and Automated Concept Discovery. *Proc 1st Int Conf Gen WordNet Mysore India* 69: 1–9. *https://doi.org/2682189*

Babcock, S, J. Beverley, L. G. Cowell, & B. Smith. 2021. The Infectious Disease Ontology in the age of COVID-19. *J Biomed Semantics*. 2021 Jul 18; 12 (1):13. doi: 10.1186/s13326-021-00245-1. PMID: 34275487; PMCID: PMC8286442. *https://pubmed.ncbi.nlm.nih.gov/34275487/*

Chaix, E, L. Deléger, R. Bossy, & C. Nédellec. 2019. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiol* 81: 63–75. *https://doi.org/10.1016/j.fm.2018.04.011*

Charlet, J., B. Bachimont, & M. C. Jaulent. 2006. Building medical ontologies by terminology extraction from texts: An experiment for the intensive care units. *Comput Biol Med* 36: 857–870. *https://doi.org/10.1016/j.compbiomed.2005.04.012*

Chi, N. W., Y. H. Jin, & S. H. Hsieh. 2019. Developing base domain ontology from a reference collection to aidinformationretrieval.*AutomConstr*100:180–189.*https://doi.org/10.1016/j.autcon.2019.01.001*

Dutta, B., & M. DeBellis. 2020. CODO: an ontology for collection and analysis of COVID-19 data. *arXiv preprint arXiv:2009.01210*. *https://doi.org/10.48550/arXiv.2009.01210*

Fabian, G., T. Wächter, & M. Schroeder. 2012. Extending ontologies by finding siblings using set expansion techniques. 28: 292–300. *https://doi.org/10.1093/bioinformatics/bts215*

Foster, P. J., R. Buhrmann, H. A. Quigley, & GJ Johnson. 2002 The definition and classification of glaucoma in prevalence surveys. *Br J Ophthalmol* 86: 238–242. *https://doi.org/10.1136/bjo.86.2.238*

Hearst, MA. 1992. Automatic acquisition of hyponyms from large text corpora. *Proc 14th Conf Comput Linguist* 23–28. *https://doi.org/https://doi.org/10.3115/992133.992154*

Hsieh, SHang-hsieh, Lin Hsien-Tang, Chi NaiWen, kuang wu Chou, and Ken Yu Lin. 2011. Enabling the development of base domain ontology through extraction of knowledge from engineering domain handbooks. *Adv Eng Informatics* 25: 288–296. *https://doi.org/10.1016/j.aei.2010.08.004*

Jiang, X, & AH Tan. 2005. Mining ontological knowledge from domain-specific text documents. *Proc - IEEE Int Conf Data Mining,* ICDM 665–668. *https://doi.org/10.1109/ICDM.2005.97*

Kibbe, A. Warren, Cesar Arze, Victor Felix, Elvira Mitraka, Evan Bolton, Gang Fu, J. Christopher, Mungall Janos, X Binder, James Malone, Drashtti Vasant, Helen Parkinson, and Lynn M. Schriml. (2015) Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 43: D1071–D1078. *https://doi.org/10.1093/nar/gku1011*

Kless, Daniel, Ludger Jansen, & S. Milton. 2016. A content-focused method for re-engineering thesauri into semantically adequate ontologies using OWL. *Semantic Web,* vol. 7, no. 5, pp. 543-576. *https://content.iospress.com/articles/semantic-web/sw194*

Liu, K, WR Hogan, & RS Crowley. 2011. Natural Language Processing methods and systems for biomedical ontology learning. *J Biomed Inform* 44: 163–79. *https://doi.org/10.1016/j.jbi.2010.07.006*

Maedche, A., V. Pekar, & S. Staab. 2003. Ontology Learning Part One — on Discovering Taxonomic Relations from the Web. *Web Intell* 301–319. *https://doi.org/10.1007/978-3-662-05320-1_14*

Missikoff, M., P. Velardi, & P. Fabriani. 2003. Text mining techniques to automatically enrich a domain ontology. *Appl Intell* 18: 323–340. *https://doi.org/10.1023/A:1023254205945*

Morin, E., & C. Jacquemin. 2004 Automatic acquisition and expansion of hypernym links. *Comput Hum* 38: 363–396. *https://doi.org/10.1007/s10579-004-1926-2*

Qian, Wang, Tao Lan, Zhu Lijun. 2007. Approach to ontology construction based on text mining, *New Zealand Journal of Agricultural Research* 50(5): 1383-1391. *https://doi.org/10.1080/00288230709510426*

Soergel, D., Boris Lauser, Anita Liang, & Frehiwot Fisseha. 2004. Reengineering thesauri for New Application: the AGROVOC Example. *Journal of Digital Information*, vol 4: 4. Article No. 257. *https://www.fao.org/3/af234e/af234e.pdf* (accessed Nov. 19, 2021)

Tsatsaronis, George, Petrova Alina, Kissa Maria, Yue Ma, Felix Distel, FranzBaader, and Michael Schroeder. 2013. Learning Formal Definitions for Biomedical Concepts. OWLED. *https://www.researchgate.net/publication/244484656_Learning_Formal_Definitions_for_Biomedical_Concepts* (accessed Aug. 11, 2019)

Wächter T, & M. Schroeder. 2010. Semi-automated ontology generation within OBO-Edit. Bioinformatics 26: 88–96. *https://doi.org/10.1093/bioinformatics/btq188*

Zhang Xiaodan, Liping Jing, Xiaohua Hu, and Xiaoua Zhou. 2007. A Comparative Study of Ontology Based Term Similarity Measures on PubMed Document Clustering. *Adv Databases Concepts, Syst Appl* 115–126. *https://doi.org/10.1007/978-3-540-71703-4_12*

سمیه تمجید

متولد ۱۳۶۱ دارای مدرک دکتری در رشته علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون در کتابخانه مرکزی دانشگاه علوم پزشکی ایران مشغول فعالیت است. کتاب «راهنمای کاربردی جست‌وجو در پایگاه‌های اطلاعاتی» از جمله آثار مشترک ایشان است.

جست‌وجو و بازیابی اطلاعات، هستی‌شناسی، مدیریت فناوری اطلاعات از علایق پژوهشی ایشان است.

فاطمه نوشین‌فرد

دارای مدرک تحصیلی دکتری در رشته علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون دانشیار گروه علم ارتباطات و دانش‌شناسی واحد علوم و تحقیقات تهران دانشگاه آزاد اسلامی است.

سیبرنتیک، مدیریت دانش، رفتار اطلاع‌یابی، بازیابی اطلاعات، علم‌سنجی، کتابخانه‌های دیجیتال، و بازاریابی از جمله علایق پژوهشی ایشان است.

ملوک‌السادات حسینی بهشتی

دارای مدرک تحصیلی دکتری در رشته زبان‌شناسی همگانی از دانشگاه تهران است. ایشان هم‌اکنون دانشیار پژوهشکده علوم اطلاعات پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) است. کتاب «ساختواژه، اصطلاح‌شناسی و مهندسی دانش» از جمله آثار ایشان است.

زبان‌شناسی، هستی‌شناسی، اصطلاح‌شناسی و مهندسی دانش از جمله علایق پژوهشی ایشان است.

نجلا حریری

دارای مدرک تحصیلی دکتری در رشته علم اطلاعات و دانش‌شناسی است. وی هم‌اکنون استاد گروه علم ارتباطات و دانش‌شناسی واحد علوم و تحقیقات تهران دانشگاه آزاد اسلامی است. کتاب «اصول و روش‌های پژوهش کیفی» از آثار ایشان است.

فهیمه باب‌الحوائجی

متولد سال ۱۳۵۵، دارای مدرک تحصیلی دکتری در رشته علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون دانشیار گروه علم ارتباطات و دانش‌شناسی واحد علوم و تحقیقات تهران دانشگاه آزاد اسلامی است.

اقتصاد اطلاعات، معماری اطلاعات و ذخیره و بازیابی اطلاعات از جمله علایق پژوهشی ایشان است.

پژوهش‌نامه
پردازش و مدیریت اطلاعات