Optimizing Semantic Information Retrieval by Labeling and Ontology

H. Jafari Pavarsi[1] | N. Hariri[2] | M. Alipour-Hafezi[3]
F. Babalhavaeji[4] | M. Khademi[5]

Purpose: To optimize the semantic information retrieval by labeling and ontology methods.

Methodology: This applied research has been done with the approach of content analysis. 313 Persian articles on the subject of information retrieval were collected in a database with subject search capabilities for both pre-test and post-test groups. After labeling 5700 words with the help of Ferdowsi University of Mashhad's software for natural language processing software, the ontology of concepts and their semantic relations were designed and implemented in protégé software. The accuracy of the retrieved results was measured in two stages before and after the test.

Findings: The significance level of Z test, in terms of statistical and reliability of 0.99, showed a significant difference between the accuracy of the retrieved related results in the two groups of pre-test and post-test. Therefore, these tools are acceptable.

Conclusion: Tow methods of natural language processing and ontology optimize semantic information retrieval.

Received: 28, May 2019
Accepted: 23, Oct. 2019

DOI: 10.30484/NASTINFO.2019.2247.1866

1. PhD Candidate, Knowledge and Information Science; Science and Research Branch; Islamic Azad University; Tehran, Iran, hmdh.jfr@gmail.com
2. PhD in Knowledge and Information Science; Professor; Department of Science and Research Branch; Islamic Azad University; Tehran, Iran (Corresponding author), nadjlahariri@gmail.com
3. PhD in Knowledge and Information Science; Assistant Professor; Department of Knowledge and Information Science; Allameh Tabataba'i University; Tehran, Iran, meh.hafezi@gmail.com
4. PhD in Knowledge and Information Science; Associate Professor; Science and Research Branch; Islamic Azad University; Tehran, Iran, f.babalhavaeji@gmail.com
5. PhD in Applied Mathematics; Associate Professor, Department of Tehran-South Branch; Islamic Azad University; Tehran, Iran, dr.maryam.khademi@gmail.com

Keywords:

Semantic information retrieval, Labeling, Natural language processing, Protégé, Ontology

NASTINFO

ارتقای بازیابی معنایی اطلاعات با استفاده از برچسب‌گذاری و هستان‌شناسی

حمیده جعفری‌پاورسی[۱] | نجلا حریری[۲] | مهدی علیپورحافظی[۳]
فهیمه باب‌الحوائجی[۴] | مریم خادمی[۵]

هدف: بهینه‌سازی بازیابی معنایی اطلاعات با استفاده از روش‌های برچسب‌گذاری و هستان‌شناسی.

روش‌شناسی: این پژوهش کاربردی با رویکرد تحلیل محتوا انجام شده است. ۳۱۳ مقاله فارسی در موضوع بازیابی اطلاعات در یک پایگاه اطلاعاتی با قابلیت‌های جستجوی موضوعی برای دو گروه پیش‌آزمون و پس‌آزمون گردآوری شد. پس از برچسب‌گذاری ۵۷۰۰ واژه به‌کمک نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد، هستان‌شناسی مفاهیم و روابط معنایی آنها در محیط پروتژ طراحی و پیاده‌سازی شد. دقت نتایج بازیابی‌شده در دو مرحله پیش و پس‌آزمون سنجیده شد.

یافته‌ها: سطح معناداری آزمون Z، به‌لحاظ آماری و اطمینان ۰/۹۹، تفاوت معناداری را میان میزان دقت نتایج مرتبط بازیابی‌شده در دو گروه پیش‌آزمون و پس‌آزمون نشان داد. بنابراین، این ابزارها کارایی پذیرفتنی دارند.

نتیجه‌گیری: دو روش پردازش زبان طبیعی و هستان‌شناسی به ارتقای بازیابی معنایی اطلاعات منجر می‌شود.

کلیدواژه‌ها

بازیابی معنایی اطلاعات، برچسب‌گذاری، هستان‌شناسی، پروتژ، پردازش زبان طبیعی

دریافت: ۹۸/۰۵/۱۲ پذیرش: ۹۸/۰۹/۱۰

۱. دانشجوی دکترای علم اطلاعات و دانش‌شناسی؛ دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران
hmdh.jfr@gmail.com
۲. دکترای علم اطلاعات و دانش‌شناسی؛ استاد؛ دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران (نویسنده مسئول)
nadjlahariri@gmail.com
۳. دکترای علم اطلاعات و دانش‌شناسی؛ استادیار؛ دانشگاه علامه طباطبائی؛ تهران، ایران
meh.hafezi@gmail.com
۴. دکترای علم اطلاعات و دانش‌شناسی؛ دانشیار؛ دانشگاه آزاد اسلامی؛ واحد علوم و تحقیقات؛ تهران، ایران f.babalhavaeji@gmail.com
۵. دکترای ریاضی کاربردی؛ دانشیار؛ دانشگاه آزاد اسلامی؛ واحد تهران جنوب؛ تهران، ایران
dr.maryam.khademi@gmail.com

مطالعات ملی کتابداری و سازماندهی اطلاعات، دوره سی‌ویکم، شماره اول، بهار ۱۳۹۹، ص ۱۸-۳۸.

DOI: 10.30484/NASTINFO.2019.2247.1866

مقدمه

متخصصان علم اطلاعات، در فهرست‌نویسی و نمایه‌سازی، کلیدواژه‌ها را از فهرست مندرجات و چکیده‌ها برمی‌گیرند و سپس با مراجعه به اصطلاح‌نامه‌ها و سرعنوان‌های موضوعی برای اثر موضوع تعیین می‌کنند. نرم‌افزارهای کتابخانه‌ای نیز قادر به تحلیل تمام‌متن نیستند. سامانه‌های بازیابی اطلاعات نیز روابط میان مدارک و حتی روابط میان مدارک و نویسندگان و معانی واژگان را درک نمی‌کنند. بنابراین، ضروری است از ابزارها و روش‌های دیگر برای ارتقای کارکرد این نظام‌ها برای افزودن بر دقت و جامعیت برونداد استفاده کرد. شیوه‌های گوناگون در نگارش واژه در زبان فارسی، اختلاف میان زبان کنترل‌شده و فولکسونومی، هم‌آوایی، شمول معنایی، ابهام، و تشابه واژگانی برخی از مشکلات در بازیابی مدارک فارسی محسوب می‌شوند (Macgregor & McCulloch, 2006).

هوش مصنوعی و فناورانه و تأثیرپذیری علم اطلاعات از آنها برای یکپارچه‌سازی، میان‌کنش‌پذیری، انتقال و اشتراک اطلاعات با کشورهای دیگر، قرابت حوزه‌های علمی و ظهور علوم میان‌رشته‌ای، ناکافی‌بودن ابزارهای سنتی نظیر اصطلاح‌نامه و رده‌بندی برای سازماندهی دانش، رشد روزافزون اطلاعات، تغییر ماهیت نیاز اطلاعاتی کاربران به کشف دانش به‌جای اطلاعات و گسترش علائق آنان به اشتراک دانش با زبان مشترک، و نیز کمبود وقت آنان، نیاز به نظام‌های بازیابی محتوایی و معنایی به‌جای نظام‌های واژه‌محور را تشدید کرده است. در جستجوی اطلاعات از نظام‌های بازیابی، مشکلاتی ازقبیل انتخاب کلیدواژه‌های مناسب برای جستجو، چگونگی پرکردن اطلاعات فرم جستجو، انتخاب فیلترهای جستجوی مناسب، ناآشنایی با عملگرهای جستجو و کاستی‌های رفتار اطلاع‌یابی (اخوتی، رحیمی، و ذوالعلی، ۱۳۹۳؛ فرج‌پهلو و شهبازی، ۱۳۸۱)، ناآگاهی از رده‌بندی، نظام‌ها و واژگان نمایه‌سازی، و قطعیت‌نداشتن و ارتباط و انسجام نتایج بازیابی اطلاعات وجود دارد.

هستان‌شناسی‌ها با تعریف ساختار مفهومی برای داده‌های ساختارنیافته (متون) و ابزارهای هوش مصنوعی به یادگیری معنا به‌وسیله کلاس‌ها، مفاهیم، و روابط معنایی کمک می‌کنند. رویکرد نظام‌های بازیابی اطلاعات امروزی از تطبیق صرف کلیدواژه‌ها و توصیفگرمحوری به مفهوم محتوا و داده‌محوری تغییر یافته است. کاربران به‌دنبال اطلاعات دقیق‌تر هستند که در کمترین زمان در اختیارشان گذاشته شود. این جز به‌کمک ابزارها و زبان‌های وب معنایی میسر نیست. روابط معنایی در هستان‌شناسی‌ها

بسیار غنی‌تر و متغیرتر است. با پیشرفت ابزارهای وب معنایی، دست‌اندرکاران با واردکردن خصیصه‌های معنایی به قطعیت ربط بیشتر نتایج بازیابی‌شده از پایگاه‌های اطلاعاتی کمک کرده‌اند و با امکان‌پذیری تعریف درخواست جستجوی کاربر با استفاده از این خصیصه‌ها، سطح اثربخشی نظام‌های بازیابی اطلاعات را از نگاه کاربر افزایش داده‌اند. ما این کار را ارتقای بازیابی معنایی می‌نامیم.

در این پژوهش، با استفاده از ابزارهای موجود روابط معنایی میان مفاهیم و مشابهت‌سازی پایگاه اطلاعاتی نمونه پژوهش با نظام‌های اطلاعاتی معنایی کوشیده‌ایم بازیابی اطلاعات را ارتقا دهیم. در اینجا از ترکیبی از روش‌های پردازش زبان طبیعی و هوش مصنوعی در پایگاه‌های اطلاعاتی که مدعی اجرای جستجوی کاربر به‌صورت معنایی و محتوایی هستند استفاده می‌شود تا پرس‌وجوی کاربر در بین مفاهیم و روابط ارائه‌شده در نظام به‌شکل نحوی و معنایی تحلیل و نتایج تحویل شود.

هدف این پژوهش، ارتقای بازیابی نظام‌های اطلاعاتی کتابخانه‌های دیجیتالی با استفاده از روش‌های ترکیبی پیشنهادی برچسب‌زنی و هستان‌شناسی است. چالش‌های موجود در پردازش زبان طبیعی اغلب شامل شناسایی گفتار، درک، و تولید زبان طبیعی است.

«پردازش زبان طبیعی»[1] در برنامه‌های نرم‌افزاری امروزی بخش‌بندی جمله، برچسب‌زنی نقش دستوری، و تحلیل و استخراج موجودیت‌ها یک راه‌حل رایج است. جستجوی متن با اصطلاحات کنترل‌نشده نتایج ناهمگون به‌دست می‌دهد؛ زیرا هرکس برای مفهوم مشابه واژه مختلف به‌کار می‌برد. ویژگی‌های زبان طبیعی مشکلات عمده‌ای را بر نظام‌های سنتی سازماندهی تحمیل می‌کند (کفاشان و فتاحی، ۱۳۹۰). زبان‌های طبیعی بسیار متنوع و پیچیده هستند. موانع اصلی برای درک زبان طبیعی به‌وسیله ماشین و چالش‌های پیاده‌سازی که وجود دارد عبارت‌اند از: «برچسب‌گذاری ادات سخن»[2]؛ «بخش‌بندی متن‌ها»[3]؛ «ابهام‌زدایی معنایی کلمه»[4]؛ «ابهام نحوی»[5]؛ «ورودی‌های ناقص یا غیرمعمول»[6]؛ و «کنش‌های کلامی»[7] (نعمتی‌شمس‌آباد، ۱۳۹۰).

از نخستین تلاش‌ها در زبان فارسی، تولید برچسب‌گذار برای یک پیکره فارسی توسط عاصی و حاجی عبدالحسینی است. اما بخش خودکار نرم‌افزار آن نمی‌تواند بر واژه‌های کم‌تکرار برچسب بگذارد و دقت آن برای صفت‌ها و قیدها کم است (فیضی‌درخشی، فیروزی، و رحیمی، ۱۳۹۳). سامانه پردازش متن و مترجم چندزبانه آزمایشگاه هوش مصنوعی و یادگیری ماشین دانشکده مهندسی برق و کامپیوتر در آزمایشگاه سپهر دانشگاه تبریز با نسخه ۳/۱ (آزمایشگاه سیستم‌های پردازش هوشمند

1. Natural Language Processing (NLP)
2. Part- of speech tagging
3. Text segmentation
4. Uation sense W
5. Syntactic ambiguity
6. Imperfect or irregular input
7. Speech acts

رایانه‌ای (سپهر) دانشگاه تبریز، ۱۳۹۷) تلاش دیگری است. مناسب‌تر از این دو، نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد[1] نسخه ۱/۶/۱ تنها نرم‌افزار رایگان دسترس‌پذیر و استفاده در زمان انجام این پژوهش بود که عملیات برچسب‌زنی، توکن‌کردن، جداسازی، و ساختار درختی آن را در سطح مقبولی انجام می‌دهد.

با مشخص‌شدن نقش هر واژه در متن، پژوهشگر نسبت به استخراج مفاهیم، مستندسازی آنها، غنی‌سازی، و ترسیم روابط معنایی آنها با سایر مفاهیم موجود در قالب هستان‌شناسی و در محیط پروتژ[2] که از ابزارهای منبع‌باز دسترس‌پذیر است اقدام کرد.

هدف از کنترل واژگانی و پیوند داده‌ها در نظام‌های نوین سازماندهی دانش و مبتنی بر هستان‌شناسی، کاهش ابهام از زبان طبیعی در زمان توصیف و بازیابی مفاهیم است.

بخشی از برچسب‌گذاری گفتار، «برچسب‌گذاری ادات سخن»[3] است که فرایند تخصیص قسمت‌هایی از برچسب گفتار به هر علامت، مانند اسم، فعل، و صفت است. برچسب‌گذاری اجزای واژگانی فرایند انتساب اجزای کلام یا واحد زبانی مناسب (فعل، اسم و...) به هر واژه در یک جمله با زبان طبیعی است (Zhang, Fleyeh, Wang, & Lu, 2019). برچسب‌گذاری بخشی مهم در پردازش زبان طبیعی و برای بسیاری از کاربردهای پردازش زبان سودمند است. این فرایند، اغلب اولین مرحله در پردازش زبان است و پس از آن پردازش‌های دیگر ازجمله بررسی واژه‌ها انجام می‌شود.

روش‌های به‌کارگرفته‌شده به‌وسیله الگوریتم‌های برچسب‌گذار از دو روش برای تگ‌گذاری استفاده می‌کنند: به خود واژه رجوع و با بررسی خصوصیات آن تگ مناسب را پیش‌بینی می‌کنند؛ یا واژه‌های موجود در همسایگی واژهٔ مدنظر نیز بررسی و احتمال رخداد یک تگ متناسب با تگ‌های موجود در همسایگی آن را محاسبه می‌کنند و تگ با بیشترین احتمال انتخاب می‌شود. چالش در برچسب‌گذاری، پیداکردن نقش واژه‌های ناشناخته است. هنگامی‌که یک جمله به‌عنوان ورودی برای یافتن نقش واژه‌های آن به یک الگوریتم داده می‌شود، درصورتی‌که واژه‌ها در پیکره باشد، الگوریتم‌ها با دقت نقش آن را به‌کمک واژه نقش‌گذاری‌شده در پیکره پیدا و بر آن برچسب می‌گذارد. به اینها، «واژه‌های شناخته‌شده»[4] می‌گویند. واژه‌های ناشناخته در پیکره وجود ندارد و الگوریتم می‌باید به روشی به‌کمک پیکره موجود تگ مناسب را پیش‌بینی کند.

1. NLP tools
2. Protégé
3. Part of speech tagging: (POS tagging)
4. Known words

یافتن نقش واژه‌های ناشناخته چالش اصلی الگوریتم‌هاست. نتایج تگ‌گذاری واژه‌ها نشان از کم‌دقتی تگ‌گذاری برای «واژه‌های ناشناخته»[1] است (خوشحال، ۱۳۹۳). مدل‌ها، الگوریتم‌ها، و روش‌های برچسب‌گذاری در زبان‌ها دو دسته‌اند: ۱) رهیافت‌های آماری که از پیکره‌های برچسب‌گذاری‌شده[2] بهره می‌جویند و ۲) رهیافت‌های غیرآماری و مبتنی بر قانون که بر یادگیری ماشینی و دانش بشری استوارند. ازجمله آنها «مدل مخفی مارکوف»[3] است که احتمال رخداد ترتیب برچسب‌ها و احتمال‌های مربوط به رخداد کلمه‌ها نشان‌دهنده پارامترهاست (الهی‌منش و مینایی، ۱۳۹۰)، «نظام‌های ماکزیمم آنتروپی»[4]، «برچسب‌گذاری مبتنی بر تبدیل»[5]، و «نظام‌های مبتنی بر حافظه»[6] (محسنی و مینایی‌بیدگلی، ۱۳۸۶؛ ۱۳۸۸) هستند.

هستان‌شناسی‌ها از سه منبع دانش یعنی ابراصطلاح‌نامه، شبکه معنایی (گونه‌ها و روابط معنایی[7])، و واژه‌نامه‌های تخصصی و ابزارهای واژگانی استفاده می‌کنند. بدین ترتیب، تمام مفاهیم درون اصطلاح‌نامه مطابق با سطح کلی خویشاوندی یا نسبی خود در شبکه معنایی مقوله‌بندی می‌شود (زاهدی، امین، کریمی، و علی‌بیک، ۱۳۹۲).

با تعریفی که فرهنگ‌نامه وبستر از پیشوند «متا» به‌معنی «بسیار جامع و فراگیر» ارائه کرده است می‌توان ابراصطلاح‌نامه را جامع و فراگیرنده اصطلاح‌نامه‌ها، منابع واژگانی، و رده‌بندی‌ها و به‌عبارت گویاتر آن را «ابراصطلاح‌نامه» نامید. درواقع، ابراصطلاح‌نامه از معانی یا مفاهیم تشکیل شده و در اصل هدف آن برقراری پیوند میان اسامی و رویکردهای متفاوت مفاهیم یکسان و شناسایی روابط مفید میان مفاهیم غیریکسان است. تمامی مفاهیم درون ابراصطلاح‌نامه دست‌کم به یک گونه معنایی درون شبکه معنایی در این نظام منتسب شده‌اند (ولی‌نژادی، آزاده، حری، شمس‌اردکانی، و امیرحسینی، ۱۳۸۷).

گروه پردازش زبان طبیعی در آزمایشگاه فناوری وب طی سال‌های ۱۳۹۱ تا ۱۳۹۳، نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد را تهیه کرده است. این نرم‌افزار در چهار بخش «اجرای ریشه‌یابی کلمات»، «نرمال‌سازی متن و تشخیص واژه‌های عامیانه»، «برچسب‌زنی نقش کلمات»، و «پارسر فارسی» دیده شده است. در بخش ریشه‌یابی واژه‌ها می‌توان متن مدنظر خود را درج یا فایل مدنظر را فراخوانی کرد. یکی دیگر از خواص این بخش، شمارش واژه‌های متن مدنظر است. در قسمت نرمال‌سازی به تشخیص و اصلاح واژه‌های عامیانه می‌پردازد. برچسب‌زنی واژه‌ها با نقش‌های ذیل همراه است:

1. Unknown words
2. Annotated corpora
3. Hidden Markov model
4. Maximum entropy systems
5. Transformation-based tagger
6. Memory-based systems
۷. گونه‌ای معنایی، مفاهیم درون یک ابراصطلاحنامه را به یکدیگر مرتبط و روابط معنایی، میان گونه‌های معنایی رابطه برقرار می‌کند.

جدول ۱. فهرست نقش برچسب کلمات

برچسب	نقش	برچسب	نقش
NP	گروه اسمی	QW	کلمات پرسشی
VP	گروه فعلی	Khas	اسم خاص
VP2	جزء اول فعل دوبخشی	Zamir	ضمیر
ADVP	گروه قیدی	Amr	حرف امر
ADJP	گروه صفتی یا مسندی	MuchMany	حروف کمیتی
HP	حرف یا اصطلاحی عام	Delimeter	علائم جداکننده لغات
AP	حرف ربط	SentenceSpliter	علائم جداکننده جملات
PP	حرف اضافه	Pss	حروف ربط مجاز بین دو جمله غیرپایه و پیرو

در بخش پارسر امکان درج ورودی و فراخوانی از فایل الحاقی فراهم آمده است که با فشار دکمه آغاز عملیات تگ‌های جزئی، کلی، ساختار جدولی، و درختی نیز مشاهده‌پذیر است.

این پژوهش درصدد پاسخ‌گویی به پرسش‌های زیر است:

- از چه ابزارها و روش‌هایی می‌توان برای پردازش زبان طبیعی مدارک استفاده کرد؟
- چه روابط معنایی میان مفاهیم وجود دارد؟
- از چه روش‌هایی برای ارزیابی نظام‌های بازیابی معنایی اطلاعات استفاده می‌شود؟
- آیا روش‌های پردازش زبان طبیعی و هستان‌شناسی به ارتقای بازیابی معنایی اطلاعات می‌انجامد؟

برای طراحی سامانه‌های پردازش زبان طبیعی متون زبان فارسی و نحوه برچسب‌گذاری اجزای آنها و ترسیم نقشه‌های مفهومی این تلاش‌ها انجام شده است:

در سامانه دنا با به‌کارگیری نظریه وابستگی مفهومی شنگ[1] جمله‌های فارسی به شبکه‌ای از مفهوم‌ها و روابط میان آنها تبدیل می‌شوند. گام‌های پردازش عبارت‌اند از: واکافت واژه‌ای[2]، واکافت ساخت‌واژی، واکافت نحوی، واکافت معنایی، و استنتاج (شهابی و صراف‌زاده، ۱۳۸۰).

خون‌سیاوش (۱۳۸۹) با هدف شناسایی مفاهیم مستتر در دامنه معنایی و متون و اسناد، برای استفاده در نمایه‌سازی و بهبود عملکرد نظام‌های بازیابی اطلاعات، دامنه معنایی متن را با استفاده از دامنه معنایی مفاهیم تعریف‌شده در پایگاه دانش نظام شناسایی کرد. در طراحی او سپس مفاهیم مستتر در دامنه معنایی متن استخراج و براساس ارتباط معنایی که با متن (مفاهیم موجود) دارد رده‌بندی می‌شود. مفاهیم موجود

1. Tak Zhang (Conceptual dependency)(CD)
2. Lexer Analysis

در صدر رده‌بندیِ ذکرشده مهم‌ترین مفاهیم مستتر در دامنه معنایی متن تلقی و به نمایه متن افزوده می‌شود تا در زمان پرس‌وجوها با نمایه مدنظر قرار بگیرند. پیاده‌سازی ایده او به ابداع دو روش اکتشافی درباره مهندسی دانش و هستی‌شناسی و دیگری پردازش زبان طبیعی انجامید. در زمینه هستی‌شناسی، روش جدیدی برای نمایش مفاهیم به‌وسیله بُردار معنایی در فضای n بُعدی دامنه و برای نگاشت متن به پایگاه دانش نظام، مفهوم هسته‌های معنایی متن براساس زنجیره‌های معنایی ارائه و از آن استفاده شد.

نیشابوری (۱۳۸۹) با هدف وارسی دستوری و نگارشی در زبان فارسی ابتدا روش نشانه‌گذاری هیبریدی با کارایی مناسب برای زبان فارسی ساخت و سپس به بررسی و پیاده‌سازی وارسی‌کننده‌ها به روش‌های آماری و قانون‌محور پرداخت. او برای بهبود کارایی آن از مزایای جداکننده عبارت‌ها در زبان فارسی استفاده کرده است.

صنعت‌جو و فتحیان (۱۳۹۰) با هدف گسترش ابزار معنایی در بازنمون دانش به‌روش تحلیل حوزه، نمونه اولیه‌ای از هستان‌شناسی در قلمرو «نمایه‌سازی» با عنوان ASFAOnt در نسخه ۴.۴.۳ نرم‌افزار پروتژ ساختند و به‌کمک پرسشنامه ساختاریافته و روش انتومتریک، روش بلندفکرکردن در زمان جستجو و مصاحبه پس از انجام جستجو، کاربردپذیری هستان‌شناسی را در مقایسه با اصطلاح‌نامه سنجیدند و دریافتند کارآمدی هستی‌شناسی در بازنمون دانش از *اصطلاح‌نامه اصفا* بیشتر است.

خوشحال (۱۳۹۳) با استفاده از روش داده‌کاوی کوشیده است با بهبود دقت در برچسب‌گذاریِ واژه‌های ناشناخته، دقت نظام‌های خودکار برچسب‌گذار را بهبود بخشد. وی روشِ خود را بر دو پیکره UPC و دادگان ارزیابی کرد و نتایج خوبی برای واژه‌های ناشناخته به‌دست آورد. او دقت برچسب‌گذاری واژه‌های ناشناخته را در پیکره UPC برابر با ۸۵/۲۲ درصد و برای پیکره دادگان برابر با ۸۲/۸۶ درصد یافت.

زرداری (۱۳۹۵) ساختار مفهومی علم اطلاعات و دانش‌شناسی را با رویکرد هستی‌نگاشتی به‌روش مث‌آنتالوجی عرضه کرده است. ساختار تاکسونومی آن از رده‌بندی دهدهی دیوئی اقتباس و برای استخراج فرهنگ لغت از متون و مدخل‌های *دائره‌المعارف کتابداری و اطلاع‌رسانی* گرفته شده و از نسخه ۵ بتای نرم‌افزار پروتِژ برای رسمی‌سازی، استخراج ساختار مفهومی، و رمزگذاری هستی‌نگاری استفاده کرده است. وی همه عناصر هستی‌نگاری اعم از عناصر لغوی، مفاهیم، و روابط بین آنها را در نرم‌افزار تعریف کرد. یافته اصلی پژوهش او ۳۱۴ کلاس با ۲۲۴ نوع رابطه، ۵۶۳۳ نمونه مستقر در کلاس‌ها، و ۲۶۴۱۳ اصل موضوعی برگرفته از *دائره‌المعارف کتابداری و اطلاع‌رسانی* در قالب چند گراف مصوّرسازی شده است. وی مدعی

است هستی‌نگاری‌اش در سطح معناداری ۰/۰۰۰۵ از صحت کلی و همچنین صحت اجزای مختلف برخوردار است.

یادگاری (۱۳۹۶) با این استدلال که بیشتر کارهای انجام‌شده در زبان فارسی و دیگر زبان‌ها از نوع خلاصه‌سازی استخراجی است و به خلاصه‌سازی چکیده چندان نپرداخته‌اند مدعی است روشی برای تولید خلاصه‌سازی چکیده و بررسی زبان‌شناختی زبان فارسی به‌شکل پیوسته و منسجم عرضه کرده است.

دو پژوهشگر ویژگی‌های شناسایی اسناد عربی را با استفاده از شبکه عصبی برگشتی بررسی کردند. شناسایی زبان اسناد عربی برمبنای فرکانس نامه با استفاده از شبکه عصبی برگشتی و استفاده از داده‌های مجموعه‌ای از اسناد زبان عربی، فارسی، اردو، و پشتو که با الفبای عربی نوشته می‌شوند نشان داد میانگین خطای متوسط مربع خطای زبان شناسایی سند عربی براساس الگوریتم انتخاب فرکانس نامه کمتر از الگوریتم پنجره است. همچنین استفاده از شبکه‌های عصبی با روش‌های مناسب انتخاب ویژگی، عملکرد شناسایی زبان را افزایش می‌دهد (Selamat & Ng, 2008).

پژوهشی دیگر با هدف تدوین روشی جدید برای بازیابی روابط بالقوه انجام شد که به ساخت هستان‌شناسی در داده‌های رشته ریاضیات در پایگاه کتابخانه دانشگاهی ووهان[1] انجامید (Lou & Qiu, 2014).

پژوهشگران دیگری نیز برای بهینه‌سازی پردازش زبان عربی با ویژگی‌های خاص آن برای تولید موتور جستجوی معنایی، قرآن را انتخاب کردند. آنها برای تعیین مفهوم معنایی واژه‌های قرآن از هوش مصنوعی برای ایجاد هستی‌شناسی قرآنی استفاده کردند تا معنای واژه‌ها و روابط آنها را نشان دهد. از این روش برای هر مفهوم به‌منظور غنی‌سازی پرس‌وجو استفاده می‌شود (Beirade, Azzoune, & Eddine Zegour, 2019).

اما، درباره بهبود نتایج بازیابی اطلاعات با استفاده از ابزارهای معنایی (نظیر شبکه‌های معنایی و هستان‌شناسی‌ها) و توجه به کارکرد آن در نظام‌های معنایی پایگاه اطلاعاتی مدارک فارسی و پیاده‌سازی هستان‌شناسی و سنجش میزان دقت و بهبود نظام در رشته علم اطلاعات و دانش‌شناسی کاری انجام نشده است. در پردازش زبان طبیعی زبان فارسی نیز تنها گام‌های نخست و بعضاً نظری انجام شده است. در هستان‌شناسی زبان فارسی نیز مطالعات روبه افزایش اما مشابه است. در پیشینه‌های خارجی این فعالیت‌ها از چندی پیش آغاز شده و روبه پیشرفت است اما از دیدگاه متخصصان کامپیوتر و زبان‌شناس.

1. Wuhan University Library

پژوهش حاضر تلاش کرده است تا ضمن استفاده از نرم‌افزار پردازش زبان، جداسازی عبارت‌های اسمی موجود در مدارک حوزه بازیابی اطلاعات نمونه مطالعه‌شده این پژوهش و آماده‌سازی آنها برای ورود به مرحله مستندسازی واژگان با ابزارهای اصطلاح‌نامه، سرعنوان‌های موضوعی فارسی، و فارس‌نت را انجام دهد. همچنین مفاهیم و روابط میان آنها را استخراج و هستان‌شناسی را با استفاده از نسخه بتای پروتژ طراحی کند. تا اینکه تمامی این روش‌ها و فرایندها به بهبود و ارتقای بازیابی معنایی اطلاعات منتج شوند.

روش‌شناسی

جامعه پژوهش در اینجا ۳۱۳ مقاله حوزه بازیابی اطلاعات برگرفته از پایگاه اطلاعاتی نورمگز است. لزوم محیطی ثابت (بدون حضور عوامل مداخله‌گر) به طراحی پایگاه اطلاعاتی با امکانات جستجو منجر شد. این پایگاه در قالب مرحله پیش‌آزمون در اختیار ۳۰ نفر از دانشجویان دکترای رشته علم اطلاعات و دانش‌شناسی قرار گرفت تا علاوه بر جستجوی نیاز اطلاعاتی خویش از میان ۱۰ رکورد نخست بازیابی‌شده، قضاوت خود را درباره ربط رکوردهای بازیابی‌شده به‌صورت کاملاً مرتبط، نسبتاً مرتبط، و غیرمرتبط اعلام کنند تا دقت نتایج بازیابی پیش‌آزمون محاسبه شود.

در مرحله بعد با توجه به شکاف نیاز اطلاعاتی کاربر و نظام‌های اطلاعاتی تلاش شد با به‌کارگیری دو روش پردازش زبان طبیعی و تدوین هستان‌شناسی مدارک، پایگاه طراحی‌شده به‌شیوه معنایی نمایه‌سازی و مصوّر شود تا رضایت کاربران با ارائه نتایج مرتبط‌تر تأمین شود. برای تشخیص و جداسازی نقش واژه‌های موجود در متن هریک از مقالات، تحلیل متن مقالات از طریق پارسر فارسی در نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد انجام شد. سپس شبه‌اصطلاح‌نامه‌ای متشکل از ۵۷۰۰ عبارت اسمی (با روابط اعم، اخص، مرتبط، و به‌جای) مستخرج از پارسر فارسی نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد از طریق فارس‌نت[1] و اصطلاح‌نامه‌های مربوط ازجمله *اصطلاح‌نامه فرهنگی فارسی (اصفا)*؛ *اصطلاح‌نامه فنی و مهندسی*؛ *اصطلاح‌نامه فرهنگ، ارتباطات، اطلاعات*؛ *اصطلاح‌نامه نما (نظام مبادله اطلاعات علمی-فنی)*؛ و همچنین *سرعنوان‌های موضوعی فارسی* ساخته شد. در گام بعدی، روابط معنایی میان مفاهیم ترسیم و در قالب هستان‌شناسیِ مفاهیم مستخرج مستند متشکل از ۱۳۳ کلاس، مجموع ۱۸ نوع رابطه، و ۸۰ نمونه مستقر در کلاس‌ها در محیط پروتژ طراحی و

۱. یکی از ابزارهای روزآمد با بیش از ۱۰۰۰۰ مجموعه هم‌معنا شامل شبکه اسامی، صفات، و افعال است (باقربیگی و شمس‌فرد، ۱۳۹۰).

مصوّر شد. در مرحله بعد، هستان‌شناسی طراحی‌شده با کمک زبان سی‌شارپ کدنویسی شد تا تمامی مفاهیم و روابط آنها در پایگاه اطلاعاتی طراحی‌شده پیاده‌سازی شود.

پس از پیاده‌سازی هستان‌شناسی در پایگاه اطلاعاتی طراحی‌شده، از همان دانشجویان پیش‌آزمون درخواست شد با همان کلیدواژه‌ها، نیاز اطلاعاتی خویش را جستجو و از میان ۱۰ رکورد نخست بازیابی‌شده، میزان ربط آنها را با نیاز اطلاعاتی خویش برای محاسبه و مقایسه اعلام کنند.

نمونه پایگاه اطلاعاتی طراحی‌شده پس از پیاده‌سازی هستان‌شناسی در شکل ۱ نشان داده شده است.

شکل ۱. نمونه پایگاه اطلاعاتی طراحی‌شده پس از پیاده‌سازی هستان‌شناسی

پایگاه اطلاعاتی طراحی‌شده دو بخش دارد: ۱) نتایج بازیابی اطلاعات در مرحله پیش‌آزمون (شکل ۱) حاوی عنوان مقالات بازیابی‌شده و ۲) نتایج بازیابی اطلاعات مرحله پس‌آزمون.

یافته‌ها

- **ابزارها و روش‌های استفاده‌شده برای پردازش زبان طبیعی مدارک**

مدل‌های مخفی مارکوف از پرکاربردترین روش‌های استفاده‌شده برای برچسب‌گذاری اجزای واژگانی کلام است. اساس مدل مخفی مارکوف تابعی احتمالی از زنجیر مارکوف است. زنجیره/ فرایند/ مدل مارکوف به‌نام خالق آن آندرئی مارکوف[1] دو خاصیت افق محدود و مستقل از زمان دارد؛ بدین صورت که برچسبِ

1. Andrei A. Markov

یک واژه تنها وابسته به برچسب واژه قبلی است (افق محدود). این وابستگی طول زمان تغییر نمی‌کند (= مستقل از زمان‌بودن). در دو مدل رایج برچسب‌گذاری‌های مبتنی بر مدل مخفی مارکوف (bigram و trigram) تعداد پارامترهای مدل متناسب با تعداد واژه‌ها در واژگان و تعداد برچسب‌های موجود در مجموعه برچسب است (دانشگاه علم و صنعت ایران، ۱۳۸۸).

روش دیگر، برچسب‌گذاری مبتنی بر حافظه است که از روش‌های یادگیری ماشینی است و با ناظر عمل می‌کند. این روش یادگیری سعی می‌کند با یادگیری اطلاعات از روی نمونه‌های قبلی، راجع به نمونه‌های جدید تصمیم‌گیری کند. این روش برای یافتن کلاسِ نمونه جدید از روش نزدیک‌ترین همسایه (K) استفاده می‌کند که روشی معروف در شناسایی آماری الگوست (دانشگاه علم و صنعت ایران، ۱۳۸۸).

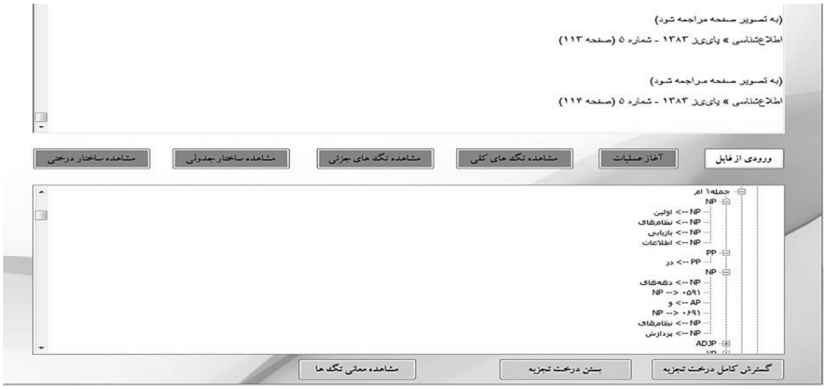
ازآنجاکه نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد در عملیات پارسر فارسی، برچسب واژگان را با توجه به واژه‌های مجاور و نمونه‌های قبلی تشخیص و انجام می‌دهد و از روش‌های پردازش زبان طبیعی چون قطعه‌بندی و نرمال‌سازی متن، تحلیل ساخت‌واژی، برچسب‌زنی جزء کلام، پارس[1] فارسی (تجزیه)، «تحلیل معنایی»[2]، و مدل‌سازی زبانی (یادگیری ماشین) استفاده می‌کند، می‌توان گفت از روش تلفیقی مارکوف و برچسب‌گذاری مبتنی بر حافظه استفاده می‌کند. جدول ۲ حاوی فهرست برچسب‌های متداول پیکره متون است.

جدول ۲. فهرست برچسب‌های اجزای مدارک (دانشگاه علم و صنعت ایران، ۱۳۸۸)

برچسب	توصیف برچسب	برچسب	توصیف برچسب
N	اسم	Alpha-per	حرف الفبای فارسی
NP	گروه اسمی	Alpha-eng	حرف الفبای انگلیسی
OH	حرف ندا	ADJ	صفت
OHH	منادی	ADV	قید
P	حرف اضافه	AR	کلمات عربی
PP	گروه حرف اضافه‌ای	CON	حرف ربط
PRO	ضمیر	DELM	جداکننده
PS	جمله‌واره	DET	حرف تعریف
QUA	سور	IF	ادات شرط
RA	حرف اضافه معرفه‌ای (را)	INT	حرف صوت
SPEC	کیفیت‌نما	MORP	تکواژ
SUBJ	موضوع متن	MS	علامت ریاضی
V	فعل		

نمونه پارسر و ساختار درختی در نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد در شکل ۲ نشان داده شده است.

1. Parsing
2. Syntax analysis



شکل ۲. نمونه پارسر و ساختار درختی در نرم‌افزار پردازش زبان طبیعی دانشگاه فردوسی مشهد

طی مرحله برچسب‌گذاری، اجزای واژگانی هریک از مدارک به‌تفکیک (اسم، فعل، صفت، ضمیر و...) تحلیل و فهرست‌برداری شد. سپس جایگاه هریک از واژه‌ها از اصطلاحات، روابط اعم، اخص، مرتبط، و جایگزین آنها نیز استخراج شد.

- **روابط معنایی میان مفاهیم**

روابط معنایی اجزای واژگانی مستخرج از برچسب‌گذاری هریک از مدارک به‌تفکیک تحلیل و در قالب روابط دودویی فهرست‌برداری شد. به‌دلیل طولانی‌بودن این فهرست به‌تفکیک هر مدرک به نمونه‌ای از آن بسنده می‌کنیم.

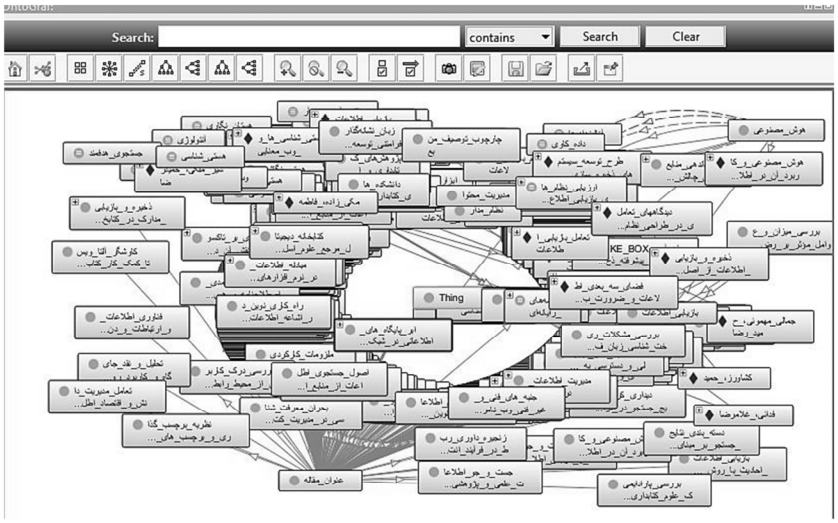
جدول ۳. گزیده‌ای از فهرست روابط معنایی میان واژگان و مفاهیم مدارک

مفهوم اول	مفهوم دوم	نوع رابطه معنایی
حریری، نجلا	ربط در مدل سنتی و مدل‌های تعاملی بازیابی اطلاعات	نویسنده
ربط در مدل سنتی و مدل‌های تعاملی بازیابی اطلاعات	حریری، نجلا	عنوان مقاله
حریری، نجلا	نشریه اطلاع‌شناسی	چاپ مقاله
ربط در مدل سنتی و مدل‌های تعاملی بازیابی اطلاعات	نشریه اطلاع‌شناسی	عنوان مقاله در مجله
بازیابی اطلاعات	اوسی‌الاسی (نظام بازیابی اطلاعات)	بخشی از ... است
ربط	معنی (روان‌شناسی)	بخشی از ... است
ربط	معنی (فلسفه)	بخشی از ... است
طراحی نظام کاربرمدار	طراحی نظام	مرتبط است
طراحی نظام کاربرمدار	طراحی نظام	بخشی از ... است
فناوری	فناوران	مرتبط است
فناوری	فناوری و علوم تجربی	بخشی از ... است

مفهوم اول	مفهوم دوم	نوع رابطه معنایی
فناوری هسته‌ای	فناوری	نوعی از ... است
علم رایانه	فناوری	نوعی از ... است
فناوری کنترل	فناوری	نوعی از ... است
کشاورزی	فناوری	نوعی از ... است
مهندسی	فناوری	نوعی از ... است
فناوری فضایی	فناوری	نوعی از ... است
فناوری اطلاعات و ارتباطات	فناوری	نوعی از ... است
ریزفناوری	فناوری	نوعی از ... است
فناوران	فناوری	مرتبط است
نیازهای اطلاعاتی	اطلاعات	مرتبط است
نیازهای اطلاعاتی	اطلاعات	بخشی از ... است
منابع اطلاعاتی	وسایل ارتباطی	بخشی از ... است
انگیزش	روان‌شناسی عاطفی	مرتبط است
انگیزش	روان‌شناسی عاطفی	بخشی از ... است
انگیزش آگاه	انگیزش	نوعی از ... است
استنباط	اصول فقه	مرتبط است
استنباط	اصول فقه	بخشی از ... است
محتوا	پیام	بخشی از ... است
برنامه‌نویسان کامپیوتر	داده‌پردازی	بخشی از ... است
پیش‌بینی	علوم تجربی	مرتبط است
پیش‌بینی	علوم تجربی	بخشی از ... است
دانش	فلسفه	بخشی از ... است
قضاوت	احکام شرعی	بخشی از ... است
راهبردها	مدیریت راهبردی	بخشی از ... است
رابطه مترادف	رابطه هم‌ارز	بخشی از ... است
نیم‌عمر منابع	منابع اطلاعاتی	مرتبط است
نیم‌عمر منابع	منابع اطلاعاتی	بخشی از ... است
فورد	نظریه‌پرداز	است
رابینز	نظریه‌پرداز	است
ساراسویک	نظریه‌پرداز	است
بیتز	نظریه‌پرداز	است
بلکین	نظریه‌پرداز	است
بروکس	نظریه‌پرداز	است
مدل توت چینی بیتز	مدل بازیابی اطلاعات	نوعی از ... است
مدل شناختی انگورسن	مدل بازیابی اطلاعات	نوعی از ... است
مدل اپیزودی بلکین	مدل بازیابی اطلاعات	نوعی از ... است
مدل طبقه‌ای ساراسویک	مدل بازیابی اطلاعات	نوعی از ... است
مدل تعاملی اسپینک	مدل بازیابی اطلاعات	نوعی از ... است

از جدول ۳ می‌توان دریافت بیشتر روابط به‌کاررفته «بخشی از»، «مرتبط است»، «عنوان مقاله»، «عنوان مقاله در مجله»، «چاپ مقاله»، و «نویسنده» است.

پس از پیاده‌سازی مفاهیم و روابط معنایی مستخرج از مراحل قبلی، با استفاده از نسخه ۵/۰ نرم‌افزار پروتژ، هستان‌شناسی داده‌های جمع‌آوری‌شده متشکل از ۱۳۳ کلاس، مجموع ۱۸ نوع رابطه، و ۸۰ نمونه مستقر در کلاس‌ها ترسیم شد. گراف حاصل از آن در شکل ۳ آمده است.



شکل ۳. گراف هستان‌شناسی طراحی‌شده مدارک حوزه بازیابی اطلاعات با نرم‌افزار پروتژ

- **روش‌های استفاده‌شده برای ارزیابی نظام‌های بازیابی معنایی اطلاعات**

پژوهش حاضر مقیاس ضریب دقت را به‌عنوان عنصر و مؤلفه ارزیابی نظام بازیابی اطلاعات، سنجیده است. برای تجزیه و تحلیل داده‌ها و محاسبه میزان ربط نتایج بازیابی از فرمول دقت استفاده کرده‌ایم:

$$\text{ضریب دقت} = \frac{\text{مدارک مرتبط بازیابی شده}}{\text{کل مدارک بازیابی شده}}$$

ربط به‌عنوان معیار ارزیابی نظام‌های بازیابی اطلاعات و کارایی کاوش با ملاک‌های جامعیت و مانعیت سنجیده می‌شود. ما ربط را به‌شکل عینی و ذهنی سنجیده‌ایم. تأکید ما در شکل عینی بر نظام‌های بازیابی اطلاعات و در شکل ذهنی بر داوری کاربران بوده است. دیدگاه نخست، محتوای مدرک را مدنظر قرار می‌دهد و ربط براساس هم‌خوانی موضوع سند با موضوع درخواست متقاضی تعیین می‌شود. در دیدگاه دوم متغیرهای دیگری مانند روزآمدی، دسترس‌پذیری، و کیفیت منبع اهمیت

دارد. این متغیرها بر برداشت استفاده‌کننده از ارتباط مدارک بازیابی‌شده تأثیر می‌گذارند (حریری، ۱۳۷۷).

بنابراین، در این پژوهش از دیدگاه کاربرمدار و ربط عینی استفاده‌کننده نظام بازیابی اطلاعات استفاده شده است؛ زیرا در مرحله نخست، کاربر دقت نتایج بازیابی‌شده را سنجید و در پایان با مؤلفه‌های بازیابی اطلاعات محاسبه و پس از ارتقای نظام بازیابی اطلاعات مقایسه شد تا میزان ارتقا مشخص شود. ازآنجاکه برای قضاوت درباره ربط از سوی جستجوگران، تنها به ۱۰ نتیجه نخست بازیابی بسنده شد، مقیاس مناسبی برای سنجش میزان جامعیت (بازیافت) در دست نبود و اصرار بر آن مقیاسی غیرواقعی و نامنطبق با نمونه آماری به‌دست می‌داد. از این‌رو، به سنجش میزان دقت (مانعیت) به‌عنوان ملاک ارتقا و بهبود نظام اکتفا کردیم.

- **استفاده از روش‌های پردازش زبان طبیعی و هستان‌شناسی در ارتقای بازیابی معنایی اطلاعات**

از جدول ۴ برمی‌آید که به‌کارگیری ابزارهای پردازش زبان طبیعی و هستان‌شناسی در نظام‌های بازیابی اطلاعات به بهبود نتایج بازیابی اطلاعات و افزایش میزان دقت (مانعیت) کمک شایان می‌کند.

جدول ۴. مقایسه فراوانی و میزان دقت پیش‌آزمون و پس‌آزمون نتایج مرتبط نمونه آماری

شماره دانشجو	موضوع	دقت پیش‌آزمون نتایج مرتبط	میزان دقت پیش‌آزمون نتایج مرتبط (به درصد)	دقت پس‌آزمون نتایج مرتبط	میزان دقت پس‌آزمون نتایج مرتبط (به درصد)
۱	مدیریت دانش	۱	۱۰	۲	۲۰
۲	هستی‌شناسی	۲	۲۰	۵	۵۰
۳	مصوّرسازی اطلاعات	۳	۳۰	۳	۳۰
۴	رفتار اطلاع‌یابی	۱	۱۰	۴	۴۰
۵	نمایه‌سازی معنایی	۲	۲۰	۴	۴۰
۶	اقتصاد اطلاعات	—	—	۳	۳۰
۷	ربط	۳	۳۰	۱۰	۱۰۰
۸	هرمنوتیک	۲۰	۲۰	۲۰	۲۰
۹	متن‌کاوی	۲۰	۲۰	۲۰	۲۰
۱۰	کتابخانه دیجیتالی	—	—	۹	۹۰
۱۱	معماری اطلاعات	۱	۱۰	۲	۲۰
۱۲	موتورهای کاوش	۱	۱۰	۱	۱۰۰

شماره دانشجو	موضوع	دقت پیش‌آزمون نتایج مرتبط	میزان دقت پیش‌آزمون نتایج مرتبط (به درصد)	دقت پس‌آزمون نتایج مرتبط	میزان دقت پس‌آزمون نتایج مرتبط (به درصد)
۱۳	فراداده	۲	۲۰	۲	۲۰
۱۴	هوش مصنوعی	۲	۲۰	۳	۳۰
۱۵	علم‌سنجی	—	—	۱	۱۰
۱۶	روان‌شناسی	—	—	۰	—
۱۷	رابط کاربر	—	—	۵	۵۰
۱۸	مجموعه‌سازی	—	—	۰	—
۱۹	نرم‌افزار کتابخانه‌ای	—	—	۷	۷۰
۲۰	تحلیل لاگ	۱	۱۰	۱	۱۰
۲۱	بازاریابی اطلاعات	۳	۳۰	۰	—
۲۲	برچسب‌زنی	۱	۱۰	۱	۱۰
۲۳	سواد اطلاعاتی	—	—	۶	۶۰
۲۴	تحلیل متن	—	—	۰	—
۲۵	آرشیو	—	—	۲	۲۰
۲۶	راهبردهای جستجو	۲	۲۰	۱۰	۱۰۰
۲۷	جامعیت و مانعیت	۲	۲۰	۶	۶۰
۲۸	اف‌آربی‌آر	—	—	۲	۲۰
۲۹	سیرنتیک	۲	۲۰	۱	۱۰
۳۰	ارزیابی نظام‌های بازیابی اطلاعات	—	—	۲	۲۰

پیش از هر آزمون، از آزمون شاپیرو ویلک[1] برای بررسی طبیعی‌بودن یا نبودن توزیع داده‌ها استفاده شد. این آزمون نشان داد به‌سبب طبیعی‌نبودن توزیع داده‌ها باید از آزمون‌های ناپارامتریک استفاده کرد.

جدول ۵. آزمون Z برای بررسی دقت حاصل از نتایج مرتبط بازیابی‌شده

نتایج مرتبط بازیابی‌شده پیش‌آزمون و پس‌آزمون	آزمون Z	سطح معناداری
	-۴/۸۰۶	۰/۰۰۰

آزمون آماری ویل کاکسون[2] (آزمونی ناپارامتریک برای مقایسه دو گروه وابسته) با استفاده از نرم‌افزار SPSS 22 انجام شد. مطابق جدول ۵، نتایج این آزمون نشان می‌دهد با توجه به سطح معناداری آزمون Z به‌لحاظ آماری با اطمینان ۰/۹۹ تفاوت معناداری میان دقت حاصل از نتایج مرتبط بازیابی‌شده پیش‌آزمون و پس‌آزمون وجود دارد.

1. Shapiro Wilk
2. Wilcoxon

جدول ۶. میانگین دقت حاصل از نتایج مرتبط بازیابی‌شده

دقت نتایج مرتبط بازیابی‌شده	میانگین
پیش‌آزمون	۱
پس‌آزمون	۱۶

برای بررسی اینکه دقت کدام‌یک از نتایج مرتبط بازیابی‌شده پیش‌آزمون و پس‌آزمون بالاتر است، یافته‌های جدول ۶ نشان می‌دهد میانگین دقت نتایج بازیابی‌شده مرتبط پس‌آزمون (۱۶) از میانگین دقت نتایج بازیابی‌شده مرتبط پیش‌آزمون (۱) بالاتر است.

نتیجه‌گیری

برای ساختارمندکردن اطلاعات، بهبود جستجوها، و نمایش معانی و محتوای اطلاعات نیاز به فناوری جدیدی است تا بتوان بین اطلاعات موجود و سایر اطلاعات اتصال برقرار کرد و معنای صریح در حیطه ارائه اطلاعات حاصل کرد. زمان آن فرا رسیده است تا موتورهای جستجو به‌جای جستجوی کلیدواژه‌ها، مضامین و محتوا را از منابع اطلاعاتی استخراج کنند. ابزارهای معنایی می‌توانند بازیابی اطلاعات را متحول کنند.

پس از مقایسه و ارزیابی نتایج پژوهش دریافتیم بین دقت نتایج مرتبط بازیابی‌شده در مرحله پیش‌آزمون و پس‌آزمون تفاوت معناداری وجود دارد. بنابراین، دو روش به‌کارگرفته‌شده (پردازش زبان طبیعی و هستان‌شناسی) به ارتقای بازیابی معنایی اطلاعات منجر می‌شود. ما تلاش کردیم با استفاده از ابزارهای داخلی پردازش زبان فارسی و تولید هستان‌شناسی و ترسیم روابط غنی‌شده و به‌دست‌دادن الگو، زمینه را برای پیاده‌سازی آن در سامانه‌های تخصصی جستجوی اطلاعات با کمک ابزارهای هوش مصنوعی و خودکارسازی این روش فراهم کنیم. بدین ترتیب، جستجوگر و طراح پایگاه اطلاعاتی از بازبینی متن کامل الحاقی منبع مدنظر به فراداده‌ها و جستجوی تمام‌متن آن بی‌نیاز و به افزایش سطح معنایی نتایج بازیابی اطلاعات و دقت آن منجر می‌شود.

این پژوهش با محدودیت‌هایی مواجه بود: به‌روزنبودن اصطلاح‌نامه‌ها و ابزارهای کمکی دیگر در مستندسازی واژگان و ترسیم روابط میان آنها و دیگری بومی‌سازی‌نشدن نرم‌افزار طراحی هستان‌شناسی برای زبان فارسی. پشتیبانی وبی، استنتاج‌ها و همچنین نقصان‌های نگارشی رسم‌الخط فارسی و نشانه‌گذاری‌ها نیز ما را با مشکل روبه‌رو کرد.

مآخذ

آزمایشگاه سیستم‌های پردازش هوشمند رایانه‌ای (سپهر) دانشگاه تبریز (۱۳۹۷). *راهنمای نصب، راه‌اندازی و استفاده برنامه نرم‌افزار جامع پردازش متن دانشگاه تبریز*. ویرایش ۳. بازیابی ۷ آذر ۱۳۹۸، از http://ece.tabrizu.ac.ir/

اخوتی، مریم؛ رحیمی، مژگان؛ و ذوالعلی، فرزانه (۱۳۹۳). بررسی تأثیر عوامل زمینه‌ای بر فرایند رفتار اطلاع‌جویی دانشجویان کارشناسی ارشد دانشگاه علوم پزشکی کرمان در وب. *پردازش و مدیریت اطلاعات*، ۳۰ (۲)، ۴۱۹-۴۴۵.

الهی‌منش، محمدحسین؛ مینایی، بهروز (۱۳۹۰). *برچسب‌گذاری ادات سخن متون فارسی به‌کمک مدل مخفی مارکوف*. *ره‌آوردنور*، ۱۰ (۳۴). ۱۰۲-۱۰۶.

باقربیگی، سمیه؛ شمس‌فرد، مهرنوش (۱۳۹۰). *روشی نوین در ساخت نیمه‌خودکار شبکه‌ی واژگانی افعال فارسی*. *نامه فرهنگستان*، ۱۲ (۱)، ۱۰۸-۱۶۱.

حریری، نجلا (۱۳۷۷). مفهوم «ربط» در بازیابی از نظام‌های اطلاعاتی. *فصلنامه کتاب*، ۹ (۲)، ۷-۱۷.

خوشحال، مصطفی (۱۳۹۳). *ارائه یک سیستم برچسب‌گذاری خودکار اجزای واژگانی کلام برای متون فارسی*. پایان‌نامه کارشناسی ارشد، دانشگاه شیراز، شیراز.

خون‌سیاوش، احسان (۱۳۸۹). *ارائه یک روش نمایه‌سازی معنایی برپایه هستی‌شناسی برای نمایه‌سازی متون و اسناد علمی*. پایان‌نامه کارشناسی ارشد، دانشگاه اصفهان، اصفهان.

دانشگاه علم و صنعت ایران. (۱۳۸۸). *مطالعه و بررسی ابزارهای برچسب‌دهی خودکار به‌منظور به‌کارگیری در پیکره متنی زبان فارسی*. دبیرخانه شورای عالی اطلاع‌رسانی. بازیابی ۲۱ اردیبهشت ۱۳۹۸، از https://www.prosody.ir/%D9%85%D9%82%D8%A7%D9%84%D9%87-%D9%87%D8%A7

زاهدی، راضیه؛ امین، غلامرضا؛ کریمی، مهرداد؛ و علی‌بیک، محمدرضا (۱۳۹۲). روش‌شناسی ایجاد هستی‌شناسی مبتنی بر نظام زبان واحد پزشکی؛ مطالعه موردی: هستی‌شناسی گیاهان دارویی ایران. *کتابداری و اطلاع‌رسانی*، ۱۶ (۳)، ۸۱-۱۰۰.

زرداری، سولماز (۱۳۹۵). مهندسی هستی‌نگاری علم اطلاعات و دانش‌شناسی براساس «دائره‌المعارف کتابداری و اطلاع‌رسانی». پایان‌نامه کارشناسی ارشد، دانشگاه شهید چمران، اهواز.

شهابی، امیرشهاب؛ صراف‌زاده، امیرحسین (۱۳۸۰). *ترجمه ماشینی زبان فارسی: راهکارها و موانع*. *تازه‌های علوم شناختی*، ۳ (۱ و ۲)، ۹-۱۴.

صنعت‌جو، اعظم؛ فتحیان، اکرم (۱۳۹۰). مقایسه کارآمدی اصطلاحنامه و هستی‌شناسی در بازنمون دانش (طراحی و ساخت نمونه هستی‌شناسی اصفا). *پژوهشنامه کتابداری و اطلاع‌رسانی، ۱* (۱)، ۲۱۹-۲۴۰.

فرج‌پهلو، عبدالحسین؛ شهبازی، مهری (۱۳۸۱). عوامل مؤثر در استفاده از پایگاه‌های اطلاعاتی: بررسی نگرش‌ها و عملکرد دانشجویان دوره تحصیلات تکمیلی. *فصلنامه کتاب، ۱۳* (۴). ۵۴-۷۱.

فیضی‌درخشی، محمدرضا؛ فیروزی، فرهنگ؛ و رحیمی، مهدی (۱۳۹۳، ۲۸ و ۲۹ آبان). *مقایسه کارهای انجام‌شده برای برچسب‌گذاری ادات سخن زبان فارسی*. مقاله ارائه‌شده در سومین همایش ملی زبان‌شناسی رایانشی، تهران. بازیابی ۷ آذر ۱۳۹۸، از https://www.researchgate.net/publication/269107861_mqaysh_karhay_anjam_shdh_bray_brchsb_gdhary_adat_skhn_zban_farsy

کفاشان، مجتبی؛ فتاحی، رحمت‌الله (۱۳۹۰). نظام‌های نوین سازماندهی دانش: وب معنایی، هستی‌شناسی و ابزارهای سازماندهی دانش عینی. *کتابداری و اطلاع‌رسانی*، ۱۴ (۲)، ۴۵-۷۰.

محسنی، مهدی؛ مینایی‌بیدگلی، بهروز (۱۳۸۶). مدل مارکوف مرتبه دو برای برچسب‌گذاری پیکره زبان فارسی. *مجموعه مقالات دانشگاه علامه طباطبایی*، ۲۲۰، ۵۹۱-۶۰۳.

محسنی، مهدی؛ مینایی‌بیدگلی، بهروز (۱۳۸۸). سیستم برچسب‌گذاری اجزای واژگانی کلام در زبان فارسی. *پردازش علائم و داده‌ها، ۶* (۲)، بازیابی ۱۶ فروردین ۱۳۹۸، از https://www.magiran.com/paper/878748

نعمتی‌شمس‌آباد، حسنعلی (۱۳۹۰). *متن‌کاوی و وب‌کاوی*. بازیابی ۷ آذر ۱۳۹۸، از http://www.farabar.net/wp-content/uploads/2016/12/@Farabar_BI-Text-Web-Mining.pdf

نیشابوری، مرتضی (۱۳۸۹). *ارائه یک روش کارا در پردازش زبان‌های طبیعی برای ساخت وارسی کننده‌های دستوری و نگارشی در زبان فارسی*. پایان‌نامه کارشناسی ارشد، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات، تهران.

ولی‌نژادی، علی؛ آزاده، فریدون؛ حری، عباس؛ شمس‌اردکانی، محمدرضا؛ و امیرحسینی، مازیار (۱۳۸۷). *طرح ادغام سرشاخه خوشه طب سنتی ایران در ساختار ابراصطلاحنامه «نظام زبان واحد پزشکی (UMLS)»*. *پیاورد سلامت*، ۲ (۳)، ۶۷-۷۴.

یادگاری، الهام (۱۳۹۶). *روشی جدید برای خلاصه‌سازی تک سند فارسی با استفاده از پردازش زبان طبیعی*. پایان‌نامه کارشناسی ارشد، مؤسسه آموزش عالی صفاهان، اصفهان.

Beirade, F., Azzoune, H., & Eddine Zegour, D. (2019). Semantic query for Quranic ontology. *Journal of King Saud University - Computer and Information Sciences*. Retrieved May 28, 2019, from https://doi.org/10.1016/j.jksuci.2019.04.005

Lou, W., & Qiu, J. (2014). Semantic information retrieval research based on co-occurrence analysis. *Online Information Review*, *38* (1), 4-23.

Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organization and resource discovery tool. *Library Review*, *55* (5), 291-300.

Selamat, A., & Ng, C.-C. (2008). Arabic script language identification using letter frequency neural networks. *International Journal of Web Information Systems, 4* (4), 484 -500.

Zhang, F., Fleyeh, Hasan,, Wang, Xinru, Lu, Minghui (2019). Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99, 238–248. Retrieved March 24, 2019, from https://doi.org/10.1016/j.autcon.2018.12.016

استناد به این مقاله:

جعفری‌پاورسی، حمیده؛ حریری، نجلا؛ علیپورحافظی، مهدی؛ باب‌الحوائجی، فهیمه؛ و خادمی، مریم (۱۳۹۹). ارتقای بازیابی معنایی اطلاعات با استفاده از برچسب‌گذاری و هستان‌شناسی. *مطالعات ملی کتابداری و سازماندهی اطلاعات*، ۳۱ (۱)، ۱۸–۳۸.