Librarianship and Information Organization Studies
eISSN: 2783-4646

Effectiveness of Semantic Tagging in Sense Disambiguation of Specialized Homographs from the Perspective of False Drop in Retrieving Scientific Texts

Mina Rezaei Dinani[1], Masoumeh Karbala Aghaei Kamran[2], VahidReza Mirzaeian[3]

[1]. Ph.D. Candidate in Knowledge and Information Science, Alzahra University, Tehran, Iran; mina.rezaei.d@gmail.com

[2]. Associate Professor, Department of Knowledge and Information Science (Corresponding Author), Alzahra University, Tehran, Iran (Corresponding author); mkamran@alzahra.ac.ir

[3]. Assistant Professor, Department of English, Faculty of Literature, Alzahra University, Tehran, Iran; mirzaeian@alzahra.ac.ir

Abstract

Purpose: The key problem in achieving efficient and user friendly retrieval when specialized homographs are searched is the development of a search mechanism to guarantee delivery of minimal irrelevant information (false drop=0). This paper has solved the problem through the implementation of a corpus-based approach using semantic tagging. The aim has been to optimize information retrieval system's performance using semantic tagging of specialized homographs to decrease false drop.

Method: This research was conducted experimentally and employed one of the three methods of word sense disambiguation. The research sample consisted of 442 scientific articles of two groups ie, experimental group and the control group. The control group had 221 full-text articles without tags and the experimental group included the same number articles, but manually tagged and placed in the proposed retrieval system to measure the effectiveness of tags in disambiguating specialized homographs and decreasing false drop.

Findings: While retrieval in the control group was with false drops due to the semantic ambiguity of specialized homographs, tagging specialized homographs in the full text of articles in the experimental group had a direct effect on decreasing false drop. The level of significance of the Wilcoxon signed-rank test (P = 0.0001, Z = -5.909) showed that the rate of false drop of retrieval results after using the tagged specialized corpus in the information retrieval system was significantly different. Assessment of negative and positive rankings showed that the rate of false drop of the results after using the tagged specialized text corpus decreased significantly and reached its minimum level of 0.

Conclusion: The rate of false drop in the research findings is an evidence of acceptable tagging effectiveness in Sense Disambiguation of specialized homographs and its effective role in optimizing the information retrieval system to minimize false drop of the results. Accordingly, the corpus-based approach of the information retrieval system, while providing an opportunity for full-text retrieval could prevent false drop and save the user time and energy. Semantic tags are valuable for disambiguation of specialized homographs, but require high quality training data. Overall, the results show that well-structured training data can play a very important role to improve disambiguation. This research experimentally and analytically reveals that this approach, compared to keyword search, achieves a significantly better degree of false drop. The technique employed can be applied to the problem of information retrieval in all languages.

Keywords: Specialized Homographs, Sense Tagging, False Drop, Text Corpus, Sense Disambiguation

Article Type: Research Article

Article history: Received: 19 Jul. 2021; Received in revised form: 10 Aug. 2021; Accepted:19

Citation:

Rezaei Dinani, M., Karbala Aghaei Kamran, M., & Mirzaeian, V. (2022). Effectiveness of semantic tagging in sense disambiguation of specialized homographs from the perspective of false drop in retrieving scientific texts. *Librarianship and Information Organization Studies*, *33*(1), 107-124. Doi: 10.30484/NASTINFO.2021.2914.2063

Publisher: *National Library and Archives of I.R. of Iran*

Doi: 10.30484/NASTINFO.2021.2914.2063

مطالعات کتابداری و سازماندهی اطلاعات

تأثیر برچسب‌گذاری معنایی در رفع ابهام هم‌نویسه‌های تخصصی از نظر ریزش کاذب در بازیابی متون علمی

مینا رضایی دینانی[۱]، معصومه کربلاآقایی کامران[۲]، وحیدرضا میرزاییان[۳]

[۱] دانشجوی دکتری علم اطلاعات و دانش‌شناسی (گرایش بازیابی اطلاعات)، دانشگاه الزهرا، تهران، ایران؛ mina.rezaei.d@gmail.com

[۲] دانشیار گروه علم اطلاعات و دانش‌شناسی، دانشگاه الزهرا، تهران، ایران (نویسندهٔ مسئول)؛ mkamran@alzahra.ac.ir

[۳] استادیار گروه زبان انگلیسی، دانشکده ادبیات، دانشگاه الزهرا، تهران، ایران؛ mirzaeian@alzahra.ac.ir

چکیده

هدف: مسئلهٔ اصلی در بازیابی مؤثر و کاربرمدار هم‌نویسه‌های تخصصی، توسعه فرایندی است که بازیابی اطلاعات نامرتبط را به حداقل برساند (ریزش کاذب= ۰). در این مقاله سعی شده با به‌کارگیری رویکرد پیکره‌مدار با استفاده از برچسب‌گذاری معنایی، بازیابی مدارک حاوی هم‌نویسه‌های تخصصی ارزیابی و با راهبردهای موجود (بدون برچسب‌گذاری) مقایسه و یافته‌ها آشکار شود. هدف بهینه‌سازی کارایی نظام بازیابی اطلاعات در کاهش ریزش کاذب بازیابی متون علمی با استفاده از روش رفع ابهام معنایی از هم‌نویسه‌های تخصصی به کمک برچسب‌گذاری معنایی بوده است.

روش: پژوهش به‌دلیل ماهیتش به روش تجربی انجام شد. روش تجربی از روش‌های سه‌گانهٔ رفع ابهام معنایی (بانظارت، نیمه‌نظارتی و بدون نظارت) بوده و روشی بانظارت به‌شمار می‌رود. جامعهٔ پژوهش را ٤٤٢ مقالهٔ علمی در دو گروه گواه و آزمون تشکیل دادند. گروه گواه (پایه) شامل ۲۲۱ متن کامل مقالهٔ بدون برچسب و گروه تجربی (آزمون) شامل همان ۲۲۱ مقاله، اما دارای برچسب بود که ٤٦ هم‌نویسهٔ تخصصی آن‌ها به روش دستی برچسب‌گذاری شد و در نظام بازیابی پیشنهادی قرارگرفتند و برای بررسی کارآیی برچسب‌ها در رفع ابهام معنایی، از هم‌نویسه‌های تخصصی و کاهش ریزش کاذب آزموده شدند.

یافته‌ها: بازیابی در مقاله‌های گروه گواه به‌دلیل ابهام معنایی هم‌نویسه‌های تخصصی، با ریزش کاذب همراه بود؛ درحالی‌که برچسب‌گذاری هم‌نویسه‌های تخصصی در متن کامل مقاله‌های گروه تجربی، تأثیر مستقیمی در کاهش ریزش کاذب داشت. سطح معنی‌داری آزمون رتبه‌های علامت‌دار ویلکاکسون (P = ۰/۰۰۰۱، Z = −۵/۹۰۹) نشان داد که میزان ریزش کاذب نتایج بازیابی بعد از به‌کارگیری پیکرهٔ تخصصی برچسب‌گذاری‌شده در نظام بازیابی اطلاعات به‌نسبت قبل، تفاوت معناداری داشت. بررسی رتبه‌های منفی و مثبت نشان داد که میزان ریزش کاذب نتایج بازیابی بعد از به‌کارگیری پیکرهٔ تخصصی برچسب‌گذاری‌شده به میزان معناداری کاهش یافته است.

نتیجه‌گیری: حد ریزش کاذب در یافته‌های پژوهش، گواه عملکرد قابل قبول برچسب‌گذاری در رفع ابهام معنایی هم‌نویسه‌های تخصصی است. همچنین بیانگر نقش مؤثر آن در بهینه‌سازی نظام بازیابی اطلاعات برای به‌حداقل‌رساندن ریزش کاذب نتایج است. بنابراین، رویکرد پیکره‌مدار نظام بازیابی اطلاعات، ضمن فراهم‌آوردن بستر بازیابی تمام‌متن، زمینهٔ جلوگیری از ریزش کاذب و صرفه‌جویی در وقت و انرژی کاربران را فراهم خواهد کرد. گفتنی است برای رفع ابهام معنایی هم‌نویسه‌های تخصصی، برچسب‌ها منابع ارزشمندی‌اند، اما این مستلزم بهره‌مندی از مجموعهٔ آموزش باکیفیت است. نتایج پژوهش نشان می‌دهد که داده‌های آموزشی، که به‌خوبی ساختاربندی شده باشند، نقش بسیار مهمی در بهبود رفع ابهام معنایی هم‌نویسه‌های تخصصی ایفا می‌کنند. این پژوهش به‌صورت تجربی و تحلیلی نشان داد که رویکرد پیکره‌مدار در مقایسه با جست‌وجوی مبتنی بر کلیدواژه، به‌طور معناداری سطح ایده‌آلی از ریزش کاذب را به‌دست می‌دهد. روش به‌کاررفته برای رفع ابهام معنایی هم‌نویسه‌های تخصصی در همهٔ زبان‌ها کاربرد دارد.

کلیدواژه‌ها: هم‌نویسه‌های تخصصی، برچسب‌گذاری معنایی، ریزش کاذب، پیکرهٔ متنی، رفع ابهام معنایی

نوع مقاله: پژوهشی

تاریخ دریافت: ۱٤۰۰/۰٤/۲۸؛ دریافت آخرین اصلاحات: ۱٤۰۰/۰۵/۱۹؛ پذیرش: ۱٤۰۰/۰۵/۲۸

استناد:

رضایی دینانی، مینا، کربلاآقایی کامران، معصومه و میرزائیان، وحیدرضا (۱٤۰۰). تأثیر برچسب‌گذاری معنایی در رفع ابهام هم‌نویسه‌های تخصصی از نظر ریزش کاذب در بازیابی متون علمی. *مطالعات کتابداری و سازماندهی اطلاعات*، ۳۳(٤)، ۱۲٤-۱۰۷.

Doi: 10.30484/NASTINFO.2021.2914.2063

ناشر: *سازمان اسناد و کتابخانه ملی جمهوری اسلامی ایران*

Doi: 10.30484/NASTINFO.2021.2914.2063

مقدمه

زبان طبیعی در مقایسه با زبان برنامه‌نویسی یا زبان صوری پدیده‌ای چندوجهی است که در کانالی ارتباطی برای انتقال مفهوم میان افراد جامعه به‌کار می‌رود (قیومی، ۱۳۹۸). در نظام‌های بازیابی خودکار، حل چالش‌های پیش‌روی پژوهشگران پردازش زبان طبیعی تا حد زیادی مستلزم سطح بالایی از درک محیط پیرامون و حالات انسان برای رایانه است. چالش عمده در این زمینه، ماشینی‌کردن فرایند درک و برداشت مفاهیم از متون مربوط به زبان طبیعی انسان است. جست‌وجوی اطلاعات از نظام‌های بازیابی مشکلاتی را در پی دارد؛ از قبیل انتخاب کلیدواژه‌های مناسب برای جست‌وجو، چگونگی پرکردن فرم اطلاعات جست‌وجو، انتخاب فیلترهای جست‌وجوی مناسب، ناآشنایی کاربران با عملگرهای جست‌وجو، کاستی‌های رفتار اطلاع‌یابی، ناآگاهی از رده‌بندی، نظام‌ها و واژگان نمایه‌سازی و قطعیت‌نداشتن، ارتباط و انسجام نتایج بازیابی (جعفری پاورسی و همکاران، ۱۳۹۹).

در نظام‌های بازیابی اطلاعات، روند سنتی و متداول بازیابی مبتنی‌بر جست‌وجوی کلیدواژه است (Khan et al., 2004). تعدادی از پژوهشگران در پژوهش خود تأثیر کلیدواژه را در بازیابی اطلاعات بسیار مهم ارزیابی کرده‌اند (دلیخون، ۱۳۹۵؛ شهبازی و شاهینی، ۱۳۹٤؛ نوروزی و هماوندی، ۱۳۹٤؛ گل‌تاجی و بذرگر، ۱۳۸۹؛ عبدالهی و جوکار، ۱۳۸۸؛ عبدالهی‌نورعلی، ۱۳۸٦؛ Lewandowski, 2008 ; Lazarinis, 2007 ; Zhang & Lin, 2007). علاوه بر پژوهش‌هایی که بر اهمیت کلیدواژه در جست‌وجوهای عمومی تأکید می‌ورزند، برخی پژوهش‌ها به جست‌وجوهای تخصصی معطوف و در آن‌ها تأکید شده که پژوهشگران نیز به‌طور معمول، بازیابی مدارک و اسناد علمی مانند مقاله‌های مجلات، همایش‌ها و پایان‌نامه‌ها را در نظام‌های بازیابی خودکار اطلاعات، با نوشتن کلیدواژه انجام می‌دهند و جست‌وجوی کلیدواژه‌ای را به نوع ترکیبی آن و جست‌وجوی ساده را به استفاده از راهبردهای جست‌وجو ترجیح می‌دهند (طباطبایی جعفری، ۱۳۹۰؛ عبدالهی و جوکار، ۱۳۸۸؛ شاپوری، ۱۳۷۹؛ Spink et al., 2001 Schutze, 2014; زرداری ۱۳۹۵؛ یوسفی‌راد، ۱۳۸۸) در تداوم پژوهش‌های مربوط به این مبحث بیان می‌دارند که مهم‌ترین دلیل نارسایی نظام‌های بازیابی اطلاعات، کلیدواژه‌مداربودن فرایند بازیابی اطلاعات است. ازاین‌رو، مهم‌ترین مبحث در این زمینه، تحلیل و درک رایانه از واژگان ورودی است.

این محدودیت‌ها و مقتضیات، طراحان نظام‌های بازیابی اطلاعات را ناگزیر به جایگزینی بازیابی واژه‌محور با بازیابی محتوایی می‌سازد. یکی از ملزومات گذر از بازیابی واژه‌محور به بازیابی محتوایی، رفع ابهام معنایی از کلیدواژه‌های ورودی است. رفع ابهام معنایی از کلمه، قابلیت تعیین معنای واژگان متن است که در بازیابی خودکار اطلاعات با استفاده از روش محاسباتی و از طریق ماشین انجام می‌شود. ازجملهٔ این واژگان، هم‌نویسه‌ها هستند که باعث ایجاد ابهام معنایی در بازیابی می‌شوند و درصورت رفع‌نکردن ابهام معنای آن‌ها، ریزش کاذب و بازیابی نتایج غیرمرتبط را به همراه خواهند داشت. در تعریف این قسم از کلمه‌ها گفتنی است هم‌نویسه‌ها[1] آن دسته از کلمات چندمعنا هستند که به‌صورت یکسان نوشته می‌شوند؛ یعنی واژگانی که شکل نوشتاری یکسان، اما معنای متفاوتی از یکدیگر دارند (Hearst, 1991). همانند متون عمومی، در متون علمی نیز هم‌نویسه‌های تخصصی وجود دارد. هم‌نویسه‌های تخصصی در رشته‌های علمی مختلف، شکل نوشتاری یکسانی دارند، اما از مفهوم و تعریف منحصر به آن رشته برخوردارند. درحالی‌که متخصص ممکن است بتواند به لطف سوابق حرفه‌ای خود معنای صحیح هم‌نویسه را در مقالهٔ علمی تعیین کند، برای ماشین، شناسایی معنای هم‌نویسهٔ تخصصی، به‌سادگی انسان نیست و روش‌های خودکار بازیابی اغلب بدون داشتن چنین دانشی نمی‌توانند اصطلاحات را به‌درستی از هم تفکیک کنند (Prokofyev et al., 2013). در نتیجه این مسئله می‌تواند موجب ابهام زیادی در درک متن و ریزش کاذب شدید به‌ویژه در جست‌وجوهای تخصصی شود (مینایی بیدگلی و همکاران، ۱۳۸٦، نقل در ستوده و هوشیار، ۱۳۹۷). اقلام بازیابی‌شدهٔ نامرتبط

---

[1] . در این پژوهش، از میان سه واژهٔ مترادف هم‌نویسه، هم‌نگاره و جناس تام، واژهٔ هم‌نویسه برای پرداختن به موضوع استفاده می‌شود.

و ناخواسته در جست‌وجو را ریزش کاذب یا پارازیت می‌گویند. نتایج بازیابی هم‌نویسه‌ها به‌طور معمول می‌تواند ضمن پراکندگی موضوع‌های همانند و اجتماع موضوع‌های بی‌ربط (ریزش کاذب)، کارایی نظام را تا حد درخور توجهی کاهش دهد. یکی از راه‌های جلوگیری از ریزش کاذب ناشی از وجود هم‌نویسه‌ها، استفاده از توضیحگر است. زمانی که دو یا چند مقوله با کلمه‌ای ثابت بیان شوند، می‌توان از توضیحگر در داخل پرانتز استفاده کرد. مثال: حرکت (بازیگری)، حرکت (فلسفه)، حرکت (فیزیک). توضیحگر درون پرانتز در مقابل موضوع یا توصیفگر برای تعیین حدود معنی، برطرف‌کردن شبهه و اخص‌کردن موضوع می‌آید. توضیحگر در واقع به معنی «از نظر» است (سلطانی و فانی، ۱۳۷۳، نقل در یوسفی، ۱۳۷٦). این نوع رفع ابهام معنایی از هم‌نویسه‌ها، نوعی روش نظارتی[1] است.

در این روش، که شامل مراحل تحت نظارت آموزش و آزمون است، ابتدا ویژگی‌های متن که می‌تواند نظام بازیابی اطلاعات را در استخراج معنی صحیح یاری رساند، شناسایی می‌شود. سپس نوبت به آموزش می‌رسد که نظام بازیابی اطلاعات را قادر می‌سازد این ویژگی‌ها را هنگام بازیابی شناسایی و استخراج کند. مجموعه آموزش یک گروه لغت ابهام‌زدایی شده است که به‌صورت حاشیه‌نویسی معنایی[2] برای آموزش در دسترس است و دارای ویژگی‌های نحوی و لغوی است. در این روش، کلمات جدید براساس متن‌هایی که قبلاً تعریف شده‌اند، در طبقه مناسب خود قرار می‌گیرند. در این پژوهش مجموعه‌ای از هم‌نویسه‌های تخصصی به همراه برچسب‌هایشان، برای آموزش در اختیار نظام بازیابی قرار می‌گیرد.

از دیگر تسهیلاتی که برای رفع ابهام معنایی از هم‌نویسه‌های تخصصی با روش نظارتی ضروری به نظر می‌رسد، مجموعه‌ای از آزمایش یا آزمون است که فرایند بازیابی متون مرتبط را در جست‌وجوی هم‌نویسه‌ها تسهیل می‌کند.

استفاده از مجموعه آزمون، یک استاندارد واقعی ارزیابی است و با وجود قدمت زیاد، هنوز ابزاری بسیار ارزشمند برای تحقیقات بازیابی است. در بیشتر تحقیقات منتشرشده، اثربخشی بازیابی اطلاعات با استفاده از مجموعه آزمون و معیارهای ارزیابی مرتبط ارزیابی شده است (Sanderson, 2010). بنابراین، استفاده از مجموعه‌های آزمون، به استاندارد عملی ارزیابی بدل شده است (حریری و همکاران، ۱۳۹۳). این مجموعه‌ها، شتاب‌دهندهٔ پژوهش‌های بازیابی اطلاعات و چکیده‌ای از محیط عملیاتی بازیابی‌اند که امکان اعتبارسنجی و مقایسهٔ اثربخشی رویکردهای موجود بازیابی اطلاعات را با رویکردهای نوین بازیابی فراهم می‌آورند. به این ترتیب، ابزاری برای پژوهشگر فراهم می‌شود که فواید راهبردهای مختلف بازیابی را در یک مجموعهٔ آزمایشگاهی شناسایی و کشف کند. این مجموعه، جست‌وجوی کاربر را در محیطی عملیاتی شبیه‌سازی می‌کند و طیف گسترده‌ای از انواع اسناد از قبیل متن، موسیقی، گفتار، تصاویر، فیلم، ساختارهای شیمیایی و غیره را پوشش می‌دهد (Voorhees, 1998; Sanderson, 2010). داده‌های آموزش، نمایانگر داده‌های آزمون‌اند و داده‌های آزمون، نمایانگر مشکل واقعی‌اند (Kessler, 2012). در این پژوهش مجموعه آزمون شامل پیکره‌ای از متون علمی است.

در تعریفی ساده از پیکره می‌توان گفت مجموعهٔ بزرگی از متون معتبر نوشتاری و یا گفتاری آوانویسی شده است که طبق معیارهای خاصی در قالب الکترونیکی برای هدف مشخصی، جمع‌آوری و ذخیره شدند (Bowker, 2018). برچسب‌زنی موضوعی متون در پیکره (مجموعهٔ آزمون) و توضیحگرها (مجموعهٔ آموزش) امری مهم در بازیابی اطلاعات و نوعی دسته‌بندی یا طبقه‌بندی در زبان طبیعی است. از طریق این نوع برچسب‌گذاری، طبقه‌ها یا ویژگی هم‌نویسه‌ها مشخص شده و از یک‌دیگر متمایز می‌شوند. در این پژوهش، تعیین موضوع هم‌نویسه‌های تخصصی با استفاده از برچسب‌گذاری انجام شد تا نظام بازیابی اطلاعات را هنگام پردازش اطلاعات در تمایز میان هم‌نویسه‌های تخصصی توانمند ساخته و سطح اثربخشی عملکرد نظام بازیابی ارتقا یابد. پرسشی که در اینجا مطرح می‌شود این است که استفاده از مجموعهٔ آموزش و پیکرهٔ

1. Supervised
2. Sense-Annotated

برچسب‌گذاری‌شده در بازیابی اطلاعات به چه میزان در کاهش ریزش کاذب نتایج بازیابی هم‌نویسه‌های تخصصی تأثیرگذار است؟ و به عبارت دیگر، آیا میان میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های تخصصی قبل و بعد از برچسب‌گذاری، تفاوت معناداری وجود دارد یا خیر؟ پژوهش حاضر برای پاسخ‌گویی به این سؤال طرح‌ریزی شده است.

مطالعات در موضوع تأثیر پردازش زبان طبیعی در بازیابی اطلاعات بر سه محور کلی متمرکز است. ۱) بررسی دشواری‌های نگارشی؛ ۲) آزمایش تأثیر روش‌ها، تکنیک‌ها و یا ابزارهای خاص؛ ۳) طراحی و آزمایش روش‌ها، تکنیک‌ها، الگوریتم‌ها و یا ابزارهای خاص (ستوده و هنرجویان، ۱۳۹۱).

پیشینه‌های مربوط به آزمایش تأثیر روش‌ها، تکنیک‌ها و یا ابزارهای خاص در اثربخشی بازیابی

در این محور (که پژوهش حاضر نیز در این دسته است)، پژوهش‌هایی قرار می‌گیرند که تأثیر تکنیک‌ها، ابزار و روش‌های خاص در اثربخشی بازیابی اطلاعات را آزمایش، بررسی و تحلیل کرده‌اند. در ادامه به پیشینه‌های خارج و داخل کشور پرداخته و یافته‌های آن‌ها تشریح می‌شود.

برخی پژوهشگران در پژوهش خود تأثیر برچسب‌های لغوی، املایی و دستوری و بافتار (واژگان و برچسب‌ها) اطراف هم‌نویسه را در رفع ابهام معنایی هم‌نویسه‌ها به اثبات رساندند (Hearst, 1991; Tesprasit et al., 2003).

کریم‌پور و همکاران نشان دادند که برچسب‌زنی اجزای کلام همراه با ریشه‌یابی، به افزایش دقت نتایج بازیابی منتهی می‌شود (Karimpour et al., 2005). الکسوپولو دریافت هستی‌شناسی ساختاریافته تأثیر بسزایی در بهبود تفکیک مفاهیم دارد (Alexopoulou 2009). هسی و شیائوشیائو ابهام‌زدایی از هم‌نویسه‌ها را از طریق شبکهٔ معنایی نام‌ها و در بستر وردنت[1] مغولستان انجام دادند (Hasi & Xiaoxiao, 2003). گورمن، مازوفسکی و نیکلاو نشان دادند نظام‌های ترکیبی رفع ابهام معنایی هم‌نویسه‌ها (مبتنی بر قانون و یادگیری ماشینی) دقت بیشتری دارند و با استفاده از آن‌ها به‌طرز چشمگیری خطاهای حاصل از ابهام معنایی هم‌نویسه‌ها کاهش می‌یابد (Gorman et al., 2018).

جلالی (۱۳۸۷) در فرایند بازیابی ۲۴۰۰۰ مفهوم در مستندات پزشکی روش‌های بازیابی دقیق در هستان‌شناسی قلمرو پزشکی را تعیین کرد. سلطانی و فیلی (۱۳۸۷) برای رفع ابهام از کلمات چندمعنا بیان داشتند که استفاده از پیکرهٔ متنی بدون برچسب و گراف وابستگی معنایی راهکار مناسبی برای افزایش کارایی و دقت بازیابی است. مدرس خیابانی (۱۳۹۱) در پژوهش خود بر اهمیت پیکره‌های زبانی در مطالعات زبان‌شناختی، تأکید کرد. صدقی (۱۳۹۲) با استفاده از پیکرهٔ کوچک برای رفع ابهام هم‌نویسه‌ها دریافت استفاده از الگوریتم لیست تصمیم، دقت بازیابی را دو تا سه درصد افزایش می‌دهد. عرب (۱۳۹۴) با استفاده از مدل فضای برداری، تأثیر مثبت غنی‌سازی موضوع را در تعیین معنی نهایی واژه به اثبات رساند. ستوده و هوشیار (۱۳۹۷) در پژوهش خود برای ابهام‌زدایی معنایی هم‌نویسه‌ها با مقایسهٔ پنج بافتار متنی شامل استناد متنی، ارجاع، عنوان ارجاع، عنوان و متن مقاله دریافتند بافتار متن و عنوان ارجاع، نقش معناداری در بهبود نتایج بازیابی دارند. خیرمند پاریزی و نورمندی‌پور (۱۳۹۵) طی پژوهش خود صحت روش یادگیری نظارتی مبتنی بر نمونه را تا ۸۸/۳۱ درصد نشان دادند. انبایی‌فریمانی و همکاران (۱۳۹۸) با بررسی تأثیر استفاده از پیکره در شناسایی مفهوم پنهان نشان دادند استفاده از الگوریتم نزدیکترین K همسایه و معیار شباهت واگرای «کولبک لیبلر»[2] صحت سازمان‌دهی را به میزان ۸۲/۵ درصد می‌رساند. این روش پیچیدگی فرایند سازمان‌دهی و بازیابی متون مطالعه‌شدهٔ پژوهش را کاهش داده است. علی‌پوری حافظی و همکاران (۱۳۹۸) با استفاده از رویکرد پیکره‌بنیاد در رفع ابهام معنایی از واژگان هم‌آوا- هم‌نویسهٔ زبان فارسی نشان دادند امکان رفع ابهام معنایی از واژگان هم‌نویسه با استفاده از همایندهای آن‌ها وجود دارد.

---

1. WordNet
2. Kullback–Leibler

ازجمله پیشینه‌های فارسی واژه‌پژوهی متون دانشگاهی با رویکرد پیکره‌بنیاد، پژوهش ذوالفقارکندری و همکاران (۱۳۹۹) است که در آن با استفاده از روش بسامدشماری در پیکرهٔ متون پزشکی به استخراج خودکار واژگان پایه علوم پزشکی اقدام کردند. نتایج این پژوهش نشان داد استفاده از روش‌های آماری در استخراج واژگان پایهٔ علوم پزشکی به‌طور خودکار از توانایی بالایی برخوردار است. با استفاده از این روش می‌توان واژگان تخصصی را برای اهداف پژوهشی و تهیهٔ دانشنامه‌های تخصصی به‌کار برد.

مطالعه پژوهش‌های مرتبط نشان می‌دهد با توجه به اینکه کاربران اصلی نظام‌های بازیابی مدارک و مقالات علمی، دانشجویان، اساتید و پژوهشگران هستند و در عصر اطلاعات، اطلاعات علمی ـ فنی، زیربنای هرگونه تصمیمات جدی اجرایی و تحقیقاتی به‌شمار می‌روند (حسینی‌بهشتی، ۱۳۸۲)، سهولت و دقت دسترسی این گروه از کاربران به مقالات علمی در فرایند مسیریابی دقیق و درست تولید دانش و توسعهٔ ناشی از آن، اهمیت زیادی دارد. ازاین‌رو درصورتی که ابهام معنایی اصطلاح‌هایی که هم‌نویسه‌های تخصصی هستند رفع نشود، نمایانی مدارک مرتبط کاهش و نمایانی مدارک غیرمرتبط افزایش می‌یابد و این مسئله می‌تواند در کیفیت عملکرد نظام بازیابی اطلاعات تأثیر سوء داشته باشد. مطالعهٔ پیشینه‌های داخل و خارج کشور نشان می‌دهد در زمینهٔ پاسخ به این مسئله که برچسب‌گذاری معنایی به چه میزان می‌تواند با رفع ابهام معنایی از هم‌نویسه‌های تخصصی به کاهش ریزش کاذب منجر شود، اطلاعات کافی وجود ندارد. همچنین تاکنون پژوهشی که به موضوع واژگان تخصصی متون علمی و به‌طور ویژه ابهام معنایی هم‌نویسه‌های تخصصی و دانشگاهی بپردازد و رویکرد پیکره‌مدار به این موضوع داشته باشد مشاهده نشده است؛ بنابراین این پژوهش در زمرهٔ معدود مطالعات واژه‌پژوهی دانشگاهی به‌شمار می‌رود که به‌دلیل نقش مهم و تعیین‌کنندهٔ هم‌نویسه‌های تخصصی در مسیریابی دقیق و کامل پژوهش‌های علمی، با هدف برون‌رفت از چالش ابهام معنایی هم‌نویسه‌های تخصصی، به عرضهٔ راهکار و آزمودن تأثیر آن در کاهش ریزش کاذب در بازیابی اطلاعات می‌پردازد.

روش پژوهش

این پژوهش به موضوع رفع ابهام معنایی هم‌نویسه‌های تخصصی، رویکردی پیکره‌مدار دارد و بنابراین در زمرهٔ پژوهش‌های کاربردشناسی پیکره‌ای یا کاربردشناسی تجربی قرار می‌گیرد.

ازآنجاکه در این پژوهش، تحلیل‌های واژگانی مربوط به پردازش زبان طبیعی مطرح می‌شود، در مراحل نخست تحقیق برای شناسایی هم‌نویسه‌ها در متون از روش مشاهدهٔ مستقیم و تحلیل واژه استفاده شد. نوع تحلیل در این مرحله ریخت‌شناسی هم‌نویسهٔ تخصصی بود و در مراحل مختلف شناسایی و تجمیع، ذخیره، پردازش، بازیابی و مقایسهٔ نتایج در نمونه‌مقاله‌های شش رشتهٔ علمی ریاضیات، علوم زمین، علوم زیستی، شیمی، فیزیک و جامعه‌شناسی استفاده شد.

در الگوریتم پیشنهادی پژوهش طبق استاندارد Semeval 2020 [1]، سه مرحلهٔ مجزا برای رفع ابهام معنایی هم‌نویسه‌های تخصصی فارسی پیش‌بینی شده است:

۱. مرحلهٔ آموزش؛ ۲. مرحلهٔ آزمون؛ ۳. مرحلهٔ ارزیابی.

در پژوهش حاضر اقدامات انجام‌شده در هر مرحله به شرح زیر است:

۱) مرحلهٔ آموزش: برای تهیهٔ مجموعهٔ آموزش و برای استنتاج معانی از هم‌نویسه‌های تخصصی، از اصطلاح‌نامه‌های (أخذشده از نظام اصطلاح‌نامه‌های علمی و فنی پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)) استفاده شد. این اصطلاح‌نامه‌ها عبارت است از:

[1]. Semantic Evaluation 2020

- اصطلاح‌نامهٔ شیمی/ تألیف تقی رجبی و دیگران]
- اصطلاح‌نامهٔ علوم‌زیستی/ تألیف اسماعیل اکبری، ملوک‌السادات حسینی‌بهشتی و مهرداد نوروزی اقبالی؛
- اصطلاح‌نامهٔ جامعه‌شناسی/ تألیف باربارا بوث و میشل بلر؛
- اصطلاح‌نامهٔ ریاضیات/ تألیف ملوک‌السادات حسینی‌بهشتی، وفایی، سعیده و مهرداد نوروزی‌اقبالی؛
- اصطلاح‌نامهٔ علوم زمین/ تألیف مهری صدیقی، ملوک‌السادات حسینی‌بهشتی و مهرداد نوروزی‌اقبالی؛
- اصطلاح‌نامهٔ فیزیک/ تألیف مریم نوروزی‌اقبالی، ملوک‌السادات حسینی‌بهشتی و مهرداد نوروزی‌اقبالی.

برای شناسایی و پالایش هم‌نویسه‌های تخصصی از این اصطلاح‌نامه‌ها، از نرم‌افزار اکسل استفاده شد تا امکان تطبیق اصطلاح‌های تخصصی فراهم شود. سپس هم‌نویسه‌های شناسایی‌شده، به همراه برچسب موضوعی تعیین‌شده، به‌منزلهٔ مجموعهٔ آموزش در مسیر[1] مدنظر، در نظام بازیابی اطلاعات ذخیره شد.

۲) مرحلهٔ آزمون؛ این مجموعهٔ آزمون فراهم و برچسب‌گذاری شد تا از معنای واژهٔ مدنظر در متون علمی رفع ابهام کند. پیکرهٔ متنی که برای انجام پژوهش ساخته شد، مجموعهٔ آزمون را تشکیل می‌دهد.

جونز و والر در زمینهٔ اندازهٔ مناسب پیکره (Jones & Waller, 2015) معتقدند پیکرهٔ کوچک هم می‌تواند به اندازهٔ پیکره بزرگ مؤثر باشد و این به هدف و اصولی که پشتوانهٔ ساخت آن پیکره است بستگی دارد. پیکره‌هایی که به‌صورت دستی برچسب‌گذاری می‌شوند می‌توانند منبع ارزشمندی برای پژوهش‌های پیکره‌مدار باشند (Stefanowitsch, 2006). در این پژوهش، به‌دلیل نوع روش جست‌وجو و برچسب‌گذاری دستی و با هدف اطمینان از دقت فرایندها، مجموعه‌ای محدود بررسی شد. این تصمیم همچنین باعث شد حصول نتیجه، مستلزم صرف وقت و انرژی افزون بر توان پژوهشگر نباشد. به‌دلیل اینکه پاسخ به پرسش پژوهش با استفاده از هیچ‌یک از پیکره‌های شناخته‌شدهٔ موجود فارسی امکان‌پذیر[2] نیست، پژوهشگر با کمک متخصص نرم‌افزار، ناگزیر به ایجاد پیکرهٔ تخصصی بود. همان‌طور که از تقسیم‌بندی انواع پیکره به‌دست اتکینز، کلییر و استلر در سال ۱۹۹۲ برمی‌آید، پیکرهٔ متنی این پژوهش از نوع نمونه‌ای است و از تمام متون علمی دانشگاهی و علمی ـ پژوهشی، تنها نمونه‌ای از مقالات علمی شش رشتهٔ تخصصی ریاضیات، شیمی، فیزیک، علوم زیستی، جامعه‌شناسی و علوم زمین انتخاب شد. این پیکره، پیکره‌ای باز است که داده‌های آن بعد از جمع‌آوری قابلیت افزایش دارد. همچنین از نوع هم‌زمانی است و در مدت معینی جمع‌آوری می‌شود. پیکره‌ای منفرد و تک‌زبانه است که به زبان فارسی تعلق دارد. سایر تقسیم‌بندی‌های اتکینز، کلییر و استلر در این پژوهش مصداق نمی‌یابد. برای پاسخ به پرسش پژوهش، این پژوهش در زمرهٔ پژوهش‌های تجربی قرار می‌گیرد. طرح پژوهش تجربی شامل دو گروه آزمودنی گروه تجربی[3] و گروه کنترل[4] است.

۱) گروه گواه، شاهد و یا کنترل، شامل ۲۲۱ مقاله‌ای است که از شش رشتهٔ علمی ریاضیات، علوم زمین، علوم زیستی، شیمی، فیزیک و جامعه‌شناسی جمع‌آوری شده و به‌صورت خام در پیکره و تحت آزمون اولیه برای ارزیابی میزان ریزش کاذب نتایج بازیابی قرار گرفته‌اند؛

۲) گروه تجربی، همان مقاله‌ها اما این‌بار با برچسب‌های موضوعی است و برای ارزیابی تأثیر برچسب‌گذاری پیکرهٔ متنی به‌منزلهٔ متغیر مستقل در میزان ریزش کاذب مقالات علمی به‌منزلهٔ متغیر وابسته، تحت آزمون ثانویه برای ارزیابی نتایج مرحلهٔ دوم بازیابی قرار گرفتند.

داده‌های زبانی نمونه پژوهش از وب‌گاه‌های مجله‌های علمی کشور جمع‌آوری و در پیکره گنجانده شد.

---

1. Directory
2. Atkins, Clear & Ostler
3. Experimental Group
4. Control Group

برای استفاده از پیکره در این پژوهش کاربردشناسی، اطلاعات کاربردشناختی لازم به آن‌ها افزوده شد و بر روی این پیکره‌ها، حاشیه‌نویسی کاربردشناختی[1] انجام شد. برچسب موضوعی که در برچسب‌گذاری پیکرهٔ این پژوهش استفاده شد، شامل عناوین شش رشتهٔ علمی بود که به هم‌نویسه‌های تخصصی نمونهٔ پژوهش در مجموعهٔ آموزش و آزمون تعلق گرفت و وابستگی علمی آن هم‌نویسه را دربارهٔ رشته‌ای خاص نشان می‌داد.

برای تهیهٔ پیکره، تمام اسناد نوشتاری مقاله‌ها در قالب مایکروسافت ورد[2]، پی دی اف[3] و... به قالب تکست[4] تبدیل شدند؛ هم‌نویسه‌های آن‌ها شناسایی و برچسب‌گذاری شد و فایل نهایی با کدگذاری یونیکد[5] (سازگار با زبان فارسی) به‌منزلهٔ پیکره (مجموعهٔ آزمون) در مسیر داده‌های نظام بازیابی اطلاعات استفاده شد.

۳) مرحلهٔ ارزیابی: در سومین مرحله از رفع ابهام معنایی واژه‌ها یعنی در مرحلهٔ ارزیابی، ریزش کاذب نتایج حاصل از بازیابی در دو گروه گواه و تجربی محاسبه شد. پس از مقایسه، آزمون، تحلیل و تفسیر داده‌های هر دو گروه، یافته‌های پژوهش آشکار شد.

برای پاسخ به پرسش پژوهش، قضاوت ربط نتایج بازیابی‌شده به‌صورت دودویی و توسط شش متخصص موضوعی از شش رشتهٔ علمی مدنظر انجام شد. سپس برای کسب اطمینان، توسط پژوهشگر صحت‌سنجی شد. برای هر رشتهٔ علمی حداقل ۵ هم‌نویسهٔ تخصصی و در مجموع ٤٦ واژهٔ تخصصی در قالب ١٦ هم‌نویسهٔ تخصصی به‌منزلهٔ کلیدواژه‌های جست‌وجوی متون علمی در شش رشتهٔ مدنظر تعیین شد. این تعداد هم‌نویسه از میان ۸۹۹ هم‌نویسهٔ تک‌واژه‌ای گزینش شدند که از مقایسه و تطبیق اصطلاح‌نامه‌های شش رشتهٔ مدنظر به‌دست آمد. جدول ۱ هم‌نویسه‌های مدنظر پژوهش را به همراه رشته‌های دربردارندهٔ آن نشان می‌دهد.

1. Pragmatic Nnnotation
2. Microsoft Word
3. PDF
4. Txt
5 .Unicode

جدول ۱- هم‌نویسه‌های انتخابی پژوهش

جرم	صفحه	پیچش	پروانه	آرایه	چشمه	برگ	حلال
فیزیک شیمی علوم زمین جامعه‌شناسی	ریاضی شیمی علوم زمین جامعه‌شناسی	ریاضی فیزیک علوم زمین علوم‌زیستی	ریاضی فیزیک علوم زیستی	ریاضی فیزیک علوم زمین	فیزیک شیمی علوم زمین	ریاضی علوم زمین علوم زیستی	ریاضی شیمی علوم زمین
درخت	**قطر**	**بازیابی**	**تورم**	**تقلب**	**دوران**	**یال**	**پلاسما**
ریاضی علوم زمین علوم‌زیستی	ریاضی علوم زمین جامعه‌شناسی	فیزیک شیمی علوم زمین	فیزیک جامعه‌شناسی	فیزیک جامعه‌شناسی	ریاضی علوم زمین	ریاضی علوم زمین	فیزیک علوم‌زیستی

جدول۲ تعداد هم‌نویسه‌های تخصصی منتخب پژوهش را به تفکیک رشته‌های علمی نشان می‌دهد.

جدول ۲- تعداد هم‌نویسه‌های تخصصی منتخب پژوهش در شش رشتهٔ علمی

رشته‌ها	علوم زمین		ریاضیات		فیزیک		شیمی		علوم زیستی		جامعه‌شناسی		جمع	
مقادیر	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد
	۱۲	۲۶/۱	۱۰	۲۱/۷	۹	۱۹/۵	۵	۱۰/۹	۵	۱۰/۹	۵	۱۰/۹	٤٦	۱۰۰

همان‌طور که جدول ۲ نشان می‌دهد، مجموع واژگان تخصصی انتخاب‌شده در شش رشتهٔ علمی مدنظر، ٤٦ واژه است اما به‌دلیل همپوشانی و به عبارت بهتر، هم‌نویسگی آن‌ها در دو رشته، سه رشته یا چهار رشته، همچنان که جدول ۱ نشان می‌دهد، ۱٦ واژهٔ منحصربه‌فرد به‌شمار می‌روند.

یافته‌های پژوهش

پاسخ به پرسش پژوهش شامل یک مرحلهٔ ارزیابی سه بخشی است که پژوهشگر انجام داد؛

در بخش اول و دوم ارزیابی، میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های تخصصی رشته‌های علمی مدنظر در دو گروه گواه و تجربی محاسبه شد.

برای محاسبه و مقایسهٔ ریزش کاذب نتایج در دو نظام اولیه و پیشنهادی از فرمول به‌کار رفته در پژوهش میرجود و دیگران (۱۳۹٤) استفاده شد.

رابطهٔ (۱)

$$\text{ریزش کاذب} = \frac{\text{تعداد مدارک بازیابی شدهٔ نامرتبط}}{\text{تعداد مدارک بازیابی‌شده}}$$

در شرایط ایده‌آل، مقدار ریزش کاذب نتایج برابر با ۰ و در بدترین حالت ممکن برابر ۱ خواهد بود.

جدول‌های ۳ تا ۸ میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های مدنظر پژوهش را به تفکیک شش رشتهٔ علمی در نظام بازیابی اولیه و نظام بازیابی پیشنهادی مقایسه می‌کند.

جدول ۳ میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های مدنظر رشتهٔ زیست‌شناسی را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ۳- میزان ریزش کاذب نتایج بازیابی‌شده در نمونهٔ پژوهش رشتهٔ زیست‌شناسی در نظام بازیابی اولیه و پیشنهادی

هم‌نویسه	نظام بازیابی اولیه			نظام بازیابی پیشنهادی		
	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شده نامرتبط	میزان ریزش کاذب	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب
پیچش	۲۱	۱٦	۰/۷٦	۵	۰	۰
پروانه	۱۵	۱۰	۰/٦۷	۵	۰	۰
برگ	٤٤	۲۸	۰/٦٤	۱٦	۰	۰
درخت	۲۰	۱۲	۰/٦	۸	۰	۰
پلاسما	۱۲	۷	۰/۵۸	۵	۰	۰
مجموع	**۱۱۲**	**۷۳**	**۰/٦۵**	**۳۹**	۰	۰

همچنان که جدول ۳ نشان می‌دهد، میانگین ریزش کاذب بازیابی پنج هم‌نویسهٔ مدنظر در متون زیست‌شناسی از ٦۵ درصد در نظام بازیابی اولیه به کمترین حد آن، یعنی ۰ در نظام پیشنهادی رسیده است.

جدول ٤ میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های مدنظر رشتهٔ ریاضیات را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ٤- میزان ریزش کاذب نتایج بازیابی‌شده در نمونهٔ پژوهش رشتهٔ ریاضیات در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	نظام بازیابی اولیه			نظام بازیابی پیشنهادی		
	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب	کل مدارک بازیابی شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب
صفحه	۲٦	۱۹	۰/۷۳	۷	۰	۰
پیچش	۲۱	۱٦	۰/۷٦	۵	۰	۰
پروانه	۱۵	۱۰	۰/٦۷	۵	۰	۰
آرایه	۱۵	۱۰	۰/٦۷	۵	۰	۰
برگ	٤٤	۳۹	۰/۸۹	۵	۰	۰
حلال	۲۹	۲٤	۰/۸۳	۵	۰	۰
درخت	۲۰	۱٤	۰/۷	٦	۰	۰
قطر	۳۵	۲۹	۰/۸۳	٦	۰	۰
دوران	۱۵	۹	۰/٦	٦	۰	۰
یال	۳٤	۲۸	۰/۸۲	٦	۰	۰
مجموع	**۲۵٤**	**۱۹۸**	**۰/۷۸**	**۵٦**	۰	۰

براساس جدول ٤، میانگین ریزش کاذب بازیابی ۱۰ هم‌نویسه مدنظر در متون ریاضیات از ۷۸ درصد در نظام بازیابی اولیه به کمترین حد آن یعنی ۰ در نظام پیشنهادی رسیده است.

در جدول ۵ میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های مدنظر رشتهٔ علوم زمین در نظام بازیابی اولیه و نظام بازیابی پیشنهادی آمده است.

جدول ۵- میزان ریزش کاذب نتایج بازیابی‌شده در نمونهٔ پژوهش رشتهٔ علوم زمین در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	نظام بازیابی اولیه			نظام بازیابی پیشنهادی		
	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب
جرم	۳۳	۲۶	۰/۷۹	۷	۰	۰
صفحه	۲۶	۲۰	۰/۷۷	۶	۰	۰
پیچش	۲۱	۱۶	۰/۷۶	۵	۰	۰
آرایه	۱۵	۱۰	۰/۶۷	۵	۰	۰
چشمه	۲۳	۱۴	۰/۶۱	۹	۰	۰
برگ	۴۴	۳۹	۰/۸۹	۵	۰	۰
حلال	۲۹	۲۴	۰/۸۳	۵	۰	۰
درخت	۲۰	۱۵	۰/۷۵	۵	۰	۰
قطر	۳۵	۲۹	۰/۸۳	۶	۰	۰
بازیابی	۲۰	۱۴	۰/۷	۶	۰	۰
دوران	۱۵	۱۰	۰/۶۷	۵	۰	۰
یال	۳۴	۱۹	۰/۵۶	۱۵	۰	۰
مجموع	**۳۱۵**	**۲۳۶**	**۰/۷۵**	**۷۹**	**۰**	**۰**

جدول ۵ نشان می‌دهد میانگین ریزش کاذب بازیابی ۱۲ هم‌نویسهٔ مدنظر در متون علوم زمین از ۷۵ درصد در نظام بازیابی اولیه به کمترین حد آن یعنی ۰ در نظام پیشنهادی رسیده است.

میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های مدنظر رشته جامعه‌شناسی را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی در جدول ۶ آمده است.

جدول ۶- میزان ریزش کاذب نتایج بازیابی‌شده در نمونهٔ پژوهش رشتهٔ جامعه‌شناسی در نظام بازیابی اولیه و پیشنهادی

هم‌نویسه	نظام بازیابی اولیه			نظام بازیابی پیشنهادی		
	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب
جرم	۳۳	۲۳	۰/۷	۱۰	۰	۰
صفحه	۲۶	۲۱	۰/۸۱	۵	۰	۰
قطر	۳۵	۳۰	۰/۸۶	۵	۰	۰
تورم	۱۳	۷	۰/۵۴	۶	۰	۰
تقلب	۱۰	۵	۰/۵	۵	۰	۰
مجموع	**۱۱۷**	**۸۶**	**۰/۷۴**	**۳۱**	**۰**	**۰**

همان‌طور که جدول ۶ نشان می‌دهد، میانگین ریزش کاذب بازیابی پنج هم‌نویسهٔ مدنظر در متون جامعه‌شناسی از ۷۴ درصد در نظام بازیابی اولیه به کمترین حد آن یعنی ۰ در نظام پیشنهادی رسیده است.

میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های مدنظر رشتهٔ شیمی در نظام بازیابی اولیه و نظام بازیابی پیشنهادی در جدول ۷ مشاهده می‌شود.

جدول ۷- میزان ریزش کاذب نتایج بازیابی‌شده در نمونهٔ پژوهش رشتهٔ شیمی در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	نظام بازیابی اولیه			نظام بازیابی پیشنهادی		
	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شده نامرتبط	میزان ریزش کاذب	کل مدارک بازیابی شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب
جرم	۳۳	۲۸	۰/۸۵	۵	۰	۰
صفحه	۲۶	۲۱	۰/۸۱	۵	۰	۰
چشمه	۲۳	۱۸	۰/۷۸	۵	۰	۰
حلال	۲۹	۱۸	۰/۶۲	۱۱	۰	۰
بازیابی	۲۰	۱۲	۰/۶	۸	۰	۰
مجموع	**۱۳۱**	**۹۷**	**۰/۷۴**	**۳۴**	۰	۰

بر اساس داده‌های جدول ۷، میانگین ریزش کاذب بازیابی پنج هم‌نویسهٔ مدنظر در متون شیمی از ۷۴ درصد در نظام بازیابی اولیه به کمترین حد آن یعنی ۰ در نظام پیشنهادی رسیده است.

جدول ۸ میزان ریزش کاذب نتایج بازیابی هم‌نویسه‌های مدنظر رشتهٔ فیزیک را در نظام بازیابی اولیه و نظام بازیابی پیشنهادی نشان می‌دهد.

جدول ۸- میزان ریزش کاذب نتایج بازیابی‌شده در نمونه پژوهش رشتهٔ فیزیک در نظام بازیابی اولیه و نظام بازیابی پیشنهادی

هم‌نویسه	نظام بازیابی اولیه			نظام بازیابی پیشنهادی		
	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب	کل مدارک بازیابی‌شده	تعداد مدارک بازیابی‌شدهٔ نامرتبط	میزان ریزش کاذب
جرم	۳۳	۲۶	۰/۷۹	۷	۰	۰
پیچش	۲۱	۱۵	۰/۷۱	۶	۰	۰
پروانه	۱۵	۱۰	۰/۶۷	۵	۰	۰
آرایه	۱۵	۱۰	۰/۶۷	۵	۰	۰
چشمه	۲۳	۱۷	۰/۷۴	۶	۰	۰
بازیابی	۲۰	۱۵	۰/۷۵	۵	۰	۰
تورم	۱۳	۸	۰/۶۲	۵	۰	۰
تقلب	۱۰	۵	۰/۵	۵	۰	۰
پلاسما	۱۲	۶	۰/۵	۶	۰	۰
مجموع	**۱۶۲**	**۱۱۲**	**۰/۶۹**	**۵۰**	۰	۰

نتایج جدول ۸ نشان می‌دهد که میانگین ریزش کاذب بازیابی ۹ هم‌نویسهٔ مدنظر در متون فیزیک از ۶۹ درصد در نظام بازیابی اولیه به کمترین حد آن یعنی ۰ در نظام پیشنهادی رسیده است.

در بخش سوم ارزیابی، نتیجهٔ حاصل از دو بخش قبلی، با استفاده از آزمون ناپارامتریک رتبه‌های علامت‌دار ویلکاکسون، مقایسه و به پرسش پژوهش پاسخ داده شد.

نتایج آزمون ویلکاکسون برای مقایسهٔ میزان ریزش کاذب بازیابی هم‌نویسه‌های تخصصی با به‌کارگیری دو نظام بازیابی اولیه و پیشنهادی در جدول ۹ آمده است.

جدول ۹- رتبه‌ها: مقایسهٔ میزان ریزش کاذب نتایج بازیابی دو نظام بازیابی

		تعداد	میانگین رتبه‌ها	جمع رتبه‌ها
نظام بازیابی پیشنهادی ـ نظام بازیابی اولیه	رتبه‌های منفی*	٤٦	۲۳/۵۰	۱۰۸۱/۰۰
	رتبه‌های مثبت**	۰	۰/۰۰	۰/۰۰
	گره‌ها***	۰		
	جمع	٤٦		
Z = -۵/۹۰۹، سطح معنی‌داری = ۰/۰۰۰۱				
* نظام بازیابی پیشنهادی > نظام بازیابی اولیه				
** نظام بازیابی پیشنهادی < نظام بازیابی اولیه				
*** نظام بازیابی پیشنهادی = نظام بازیابی اولیه				

سطح معنی‌داری آزمون رتبه‌های علامت‌دار ویلکاکسون (Z = -۵/۹۰۹ ،P = ۰/۰۰۰۱) نشان می‌دهد میزان ریزش کاذب نتایج بازیابی بعد از به‌کارگیری پیکرهٔ تخصصی برچسب‌گذاری‌شده در نظام بازیابی اطلاعات نسبت به قبل از آن تفاوت معناداری دارد. بررسی رتبه‌های منفی و مثبت نشان می‌دهد میزان ریزش کاذب نتایج بازیابی بعد از به‌کارگیری پیکرهٔ تخصصی برچسب‌گذاری‌شده به میزان معنی‌داری کاهش یافته است.

نتیجه‌گیری

مطالعهٔ پژوهش‌ها نشان می‌دهد عملکرد نظام‌های بازیابی به‌طور معمول مبتنی بر شکل واژهٔ ورودی است و تأکید بر ویژگی ظاهری کلیدواژه صرف‌نظر از نقش معنایی آن در متن، به ریزش کاذب در بازیابی نتایج منجر می‌شود. غلبه بر چالش‌های پردازش زبان طبیعی هنگام بازیابی اطلاعات مستلزم بررسی علمی پدیده‌های زبانی و چاره‌جویی فناورانه است. این پژوهش در راستای پژوهش مدرس خیابانی (۱۳۹۱) از رویکرد پیکره‌محور برای برون‌رفت از چالش هم‌نویسگی اصطلاح‌های تخصصی بهره گرفت. نتایج پژوهش همسو با یافته‌های برخی پژوهشگران (عرب، ۱۳۹٤؛ ستوده و هوشیار، ۱۳۹۷؛ Hearst, 1991; Tesprasit et al., 2003; Karimpour et al., 2005) و نشان داد که غنی‌سازی موضوع از طریق برچسب‌گذاری در تعیین معنی نهایی واژه، تأثیر مثبت دارد و یادگیری ماشینی و استفاده از روش رفع ابهام معنایی نظارتی می‌تواند عملکرد نظام بازیابی را در کاهش ریزش کاذب بهبود بخشد. رویکرد پیکره‌مدار و برچسب‌گذاری معنایی هم‌نویسه‌های تخصصی در این پژوهش، در هیچ‌یک از پیشینه‌های اشاره‌شده برای رفع ابهام معنایی هم‌نویسه‌های تخصصی استفاده و آزمون نشده بود.

آزمون داده‌ها در پاسخ به سؤال پژوهش نشان می‌دهد استفاده از رویکرد پیکره‌مدار در نظام بازیابی اطلاعات و برچسب‌گذاری معنایی هم‌نویسه‌های تخصصی، ریزش کاذب را در بازیابی متون علمی حاوی این هم‌نویسه‌ها به‌طور معناداری کاهش می‌دهد. به نظر می‌رسد کارایی بهتر نظام بازیابی با استفاده از این رویکرد، به‌دلیل توانمندسازی نظام بازیابی در تمایز بین هم‌نویسه‌های تخصصی و شناسایی نقش‌های معنایی آن‌ها با استفاده از برچسب‌های معنایی است که در مجموعهٔ آزمون و آموزش برای آن‌ها در نظر گرفته شده است. این رویکردی است که می‌تواند نظام بازیابی اطلاعات را هنگام بازیابی متون حاوی هم‌نویسه‌های تخصصی از بازیابی واژه‌محور به بازیابی محتوامحور سوق دهد. از مزایای استفاده از این رویکرد علاوه بر فراهم کردن امکان بازیابی تمام‌متن، غلبه بر موانعی مانند ناآگاهی کاربران از هم‌نویسگی اصطلاح تخصصی جست‌وجوشده، تمایل نداشتن کاربران به استفاده از راهبردهای جست‌وجو و پایبندی آن‌ها به اصل کمترین کوشش و توجیه‌پذیری هزینه ـ سودمندی بازیابی را می‌توان ذکر کرد.

در این پژوهش با هماهنگ‌سازی فرایند ذخیره و بازیابی متون علمی حاوی هم‌نویسه‌های تخصصی با نتایج مدنظر کاربران در بازیابی اطلاعات، ریزش کاذب به حد بهینهٔ آن یعنی صفر رسید. رویکرد پیکره‌مدار این پژوهش و استخراج زمینهٔ تخصصی متون علمی و اختصاص آن به هم‌نویسه‌های موجود در متون می‌تواند در آینده زمینه‌ساز عرضهٔ راهکارهای جدید برای چالش‌های پردازش زبان فارسی و طراحی و پیاده‌سازی نظام‌های بازیابی اطلاعات محتوامحور به‌صورت کاربردی شود.

پیشنهاد می‌شود از روش‌های نیمه‌نظارتی و بدون نظارت برای مقایسهٔ نتایج بازیابی با روش نظارتی به‌کار رفته در این پژوهش استفاده شود تا روش کارا و مؤثر شناسایی و معرفی شود. با توجه به اینکه روش استفاده‌شده محدود به زبان فارسی نیست و به سایر زبان‌ها قابلیت تعمیم دارد، آزمون آن برای بررسی ریزش کاذب در متون علمی سایر زبان‌ها نیز شایان توجه است. پژوهش‌های بیشتر پیرامون رضایت‌مندی کاربران از ربط نتایج بازیابی‌شده و زمان صرف‌شده برای جست‌وجو با استفاده از دو رویکرد موجود و پیشنهادی پژوهش و مقایسهٔ آن‌ها، ابعاد بیشتری از موضوع را تبیین می‌کند.

منابع

اکبری، اسماعیل، حسینی بهشتی، ملوک‌السادات و نوروزی‌اقبالی، مهرداد (۱۳۸٤). *اصطلاح‌نامه علوم‌زیستی*. تهران: مرکز اطلاعات و مدارک علمی ایران.

انبایی‌فریمانی، سعیده، طباطبایی، حمید و کفاشان‌کاخکی، مجتبی (۱۳۹۸). جستاری بر فرایند سازمان‌دهی و بازیابی متون وبی مبتنی بر تجمیع مفاهیم معنایی در راستای سازمان‌دهی دانش. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳٤(٤)، ۱۸۷۹–۱۹۰٤.

بوث، باربارا و بلر، میشل (۱۳۸۲). *اصطلاح‌نامه جامعه‌شناسی*. ترجمه مهوش معترف. تهران: مرکز اطلاعات و مدارک علمی ایران.

جعفری پاورسی، حمیده، حریری، نجلا، علیپورحافظی، مهدی، باب الحوائجی، فهیمه و خادمی، مریم (۱۳۹۹). ارتقای بازیابی معنایی اطلاعات با استفاده از برچسب‌گذاری و هستان‌شناسی. فصلنامه *مطالعات کتابداری و سازماندهی اطلاعات*، ۳۱(۱)، ۱۸–۳۸.

جلالی، وحید (۱۳۸۷). *بازیابی معنایی اطلاعات با استفاده از بسط مفاهیم حاصل از جست‌وجوی مبتنی بر کلیدواژه*. پایان‌نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران.

حریری، نجلا، باب‌الحوائجی، فهیمه، فرزندی‌پور، مهرداد و نادی راوندی، سمیه (۱۳۹۳). معیارهای ارزیابی ربط در نظام‌های بازیابی اطلاعات: دانسته‌ها و ندانسته‌ها. *پردازش و مدیریت اطلاعات*، ۳۰(۱)، ۱۹۹–۲۲۱.

حسینی‌بهشتی، ملوک‌السادات (۱۳۸۲). کاربرد اصطلاح‌شناسی و واژه‌گزینی در نمایه‌سازی ماشینی و بازیابی اطلاعات. *کتابداری و اطلاع‌رسانی*، ۱۸ (۳)، ۳۱–٤٤.

حسینی بهشتی، ملوک‌السادات؛ وفایی، سعیده و نوروزی اقبالی، مهرداد (۱۳۹۳). *اصطلاح‌نامه ریاضیات*. تهران: مرکز اطلاعات و مدارک علمی ایران.

خیرمند پاریزی، منیر و نورمندی‌پور، رضا (۱۳۹۵). رفع ابهام معنایی کلمات فارسی با استفاده از رویکرد نظارت‌شده الگوریتم‌های IBL. *علوم رایانشی*، ۱(۲)، ٦۳–۷۳.

دلیخون، لیلا (۱۳۹۵). *بررسی راه‌های گسترش پرسش کاربران در موتورهای جست‌وجو و پایگاه داده‌های تخصصی: مطالعه موردی دانشجویان کارشناسی ارشد فنی و مهندسی دانشگاه الزهرا (س)*. پایان‌نامه کارشناسی ارشد. دانشکده علوم تربیتی و روانشناسی. دانشگاه الزهرا.

ذوالفقار کندری، زهره، میانگاه، طیبه، روشن، بلقیس و وکیلی‌فرد، امیررضا (۱۳۹۹). بررسی تکنیک‌های بهبود عملکرد روش‌های بسامدشماری پیکره‌بنیاد در استخراج خودکار واژگان (مورد مطالعه: واژگان پایه علوم پزشکی). *پژوهشنامه پردازش و مدیریت اطلاعات*. ۳۵(٤)، ۱۰۳۹–۱۰٦٤.

رجبی، تقی، غریبی، حسین، حسینی بهشتی، ملوک‌السادات و نوروزی اقبالی، مهرداد (۱۳۸۳). *اصطلاح‌نامه شیمی*. تهران: مرکز اطلاعات و مدارک علمی ایران.

زرداری، سولماز (۱۳۹۵). *مهندسی هستی‌نگاری علم اطلاعات و دانش‌شناسی بر اساس دائرةالمعارف کتابداری و اطلاع‌رسانی*. رساله دکتری، دانشکده علوم تربیتی و روانشناسی، دانشگاه شهید چمران اهواز.

ستوده، هاجر و هنرجویان، زهره (۱۳۹۱). مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آنها بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات. کتابداری و اطلاع‌رسانی، *۱۵*(٤)، ۵۹-۹۲.

ستوده، هاجر و هوشیار، مژگان (۱۳۹۷). بررسی نقش انواع بافتار هم نویسه‌ها در تعیین شباهت بین مدارک. *پژوهشنامه پردازش و مدیریت اطلاعات*، *۳۳*(۳)، ۱۱۸۳-۱۲۰۶.

سلطانی، محمود و فیلی، هشام (۱۳۸۷). *استفاده از تکنیک ابهام‌زدایی معنایی واژگان در بازیابی بین زبانی اطلاعات*. چهاردهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، انجمن کامپیوتر ایران. تهران، دانشگاه صنعتی امیرکبیر.

شاپوری، سودابه (۱۳۷۹). مشکلات جست‌وجوی موضوعی استفاده‌کنندگان از فهرست رایانه‌ای کتابخانه مرکزی دانشگاه فردوسی مشهد. *کتابداری و اطلاع‌رسانی*، ۳(۲)، ٤۹-٦۸.

شهبازی، مهری و شاهینی، شبنم (۱۳۹٤). بررسی میزان کارآیی پایگاه‌های اطلاعاتی مگ ایران، نورمگز و اس.آی.دی. در بازیابی و ربط مباحث علم اطلاعات و دانش‌شناسی با استفاده از کلیدواژه‌های آزاد و مقایسه آنها از نظر میزان استفاده از کلیدواژه‌های مهارشده. پژوهشنامه *پردازش و مدیریت اطلاعات*. ۳۱(۲)، ٤۳۱-٤۵٤.

صدقی، فاطمه (۱۳۹۲). *رفع ابهام از هم‌نویسه‌ها در متون فارسی با روش‌های نیمه نظارتی*. پایان‌نامه کارشناسی ارشد. گروه مهندسی کامپیوتر ـ هوش مصنوعی. دانشکده فنی ـ مهندسی. دانشگاه الزهرا.

صدیقی، مهری، حسینی‌بهشتی، ملوک‌السادات و نوروزی‌اقبالی، مهرداد (۱۳۸٤). *اصطلاح‌نامه علوم زمین*. تهران: مرکز اطلاعات و مدارک علمی ایران.

طباطبایی جعفری، زهرا (۱۳۹۰). *بررسی شیوه‌های بسط پرسش در رفتار جست‌وجوی اطلاعاتی کاربران در موتورهای جست‌وجو: مطالعه در میان دانشجویان تحصیلات تکمیلی علوم کتابداری و اطلاع‌رسانی دانشگاه‌های سراسری شهرتهران*. پایان‌نامه کارشناسی ارشد. دانشکده ادبیات و علوم انسانی. دانشگاه قم.

عبدالهی نورعلی، محمدصادق (۱۳۸٦). *کندوکاو مسائل ریخت‌شناسی زبان فارسی در بازیابی اطلاعات از جست‌وجوگرهای وب*. پایان‌نامهٔ کارشناسی ارشد، گروه علم اطلاعات و دانش‌شناسی. دانشکدهٔ علوم تربیتی و روان‌شناسی. دانشگاه شیراز

عبدالهی‌نورعلی، محمدصادق و جوکار، عبدالرسول (۱۳۸۸). چالش‌های شیوهٔ نگارش زبان فارسی در بازیابی اطلاعات از موتورهای کاوش وب. *مطالعات تربیتی و روانشناسی دانشگاه فردوسی مشهد*. *۱۰*(۲)، ٦۷-۹۰.

عرب، میثم (۱۳۹٤). *استفاده از روابط پنهان بین کلمات در رفع ابهام معنایی کلمات*. پایان‌نامه کارشناسی ارشد. گروه مهندسی کامپیوتر. واحد بین‌الملل دانشکده فنی ـ مهندسی. دانشگاه شیراز.

علی‌پوری حافظی، حامد، مولودی، امیرسعید و بیات، محمدکریم (۱۳۹۸). *رفع ابهام معنایی از واژگان هم‌آوا – هم‌نویسه فارسی: رویکرد پیکره‌بنیاد*. دومین کنفرانس بازیابی تعاملی اطلاعات، تهران.

قیومی، مسعود (۱۳۹۸). تعیین خودکار معنای واژه‌های فارسی با استفاده از تعبیه معنایی واژه. *پژوهشنامه پردازش و مدیریت اطلاعات*، *۳۵*(۱)، ۵۰-۲۵.

گل‌تاجی، مرضیه و بذرگر، سعیده (۱۳۸۹). بررسی مشکلات ریخت‌شناسی زبان فارسی در سه پایگاه اطلاعاتی مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری، پژوهشگاه اطلاعات و مدارک علمی ایران و جهاد دانشگاهی. *کتابداری و اطلاع‌رسانی*، *۱۳*(۲)، ۱۹۱-۲۱٤.

مدرس خیابانی، شهرام (۱۳۹۱). بررسی پیکره‌بنیاد واژه‌های هم‌معنی. *پاژند*، ۳۰(۸)، ۸۵-۱۰۵.

میرجود، سیدحسین، قیاسی، میترا، دلیری، سعی، کوچکی‌نژاد، لیلا و عباسیان جوشقانی، آمنه (۱۳۹٤). مقایسه دقت موتورهای جست‌وجوی عمومی و تخصصی در بازیابی تصاویر پزشکی. *توسعهٔ آموزش جندی شاپور*، ٦(۲)، ۱۳۱-۱۳۸.

نوروزی، یعقوب و هماوندی، هدی (۱۳۹٤). بررسی مشکلات جست‌وجو و بازیابی تصاویر در موتورهای کاوش برگزیده مبتنی بر ویژگی‌های نگارشی زبان فارسی. *کتابداری و اطلاع‌رسانی*، *۵*(۲)، ۲۲۲-۲۰٦.

نوروزی اقبالی، مریم، حسینی بهشتی، ملوک‌السادات و نوروزی اقبالی، مهرداد (۱۳۸۵). *اصطلاح‌نامه فیزیک*. تهران: مرکز اطلاعات و مدارک علمی ایران.

یوسفی، احمد (۱۳۷٦). ریزش کاذب در ذخیره و بازیابی اطلاعات. *پژوهشنامه پردازش و مدیریت اطلاعات*، ۱۳(۱)، ۱۰-۱۹.

یوسفی‌راد، ابراهیم (۱۳۸۸). آر. دی. اف: الگویی برای توصیف منابع در وب معنایی. فصلنامه *کتاب*، ۲۰(۳)، ۹-۲۲.

References

Abdollahi NorAli, M. S. (2007). *Survey on morphological difficulties of Persian language in information retrieval from web search tools*. M.S thesis, Library and Information science. Faculty of Education and Psychology, Shiraz University. Doi: 10.22067/FE.V10I2.2112. [In Persian].

Abdollahi NorAli, M. S. & Jokar, A. (2009). Survey on morphological difficulties of Persian Language in Information retrieval from Web Search Engines. *Educational and Psychological Studies*. *10*(2), 67-90 [In Persian].

Akbari, E., Hosseini Beheshti, Moluksadat & Noroozi Eghbali, Mehrdad (2005). *Thesaurus of Biological Science*. Tehran, Iranian Research Institute for Information Science and Technology [In Persian].

Alexopoulou, D. (2009). Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. Automation meets accuracy. *BMC Bioinformatics*, *10*(1), 28. https://doi.org/10.1186/1471-2105-10-28

Alipouri Hafezi, H., Moloudi, A. & Bayat, M.K. (2019). *Sense disambiguation of the Persian homographs: The corpus-based Approach.* Interactive Information Retrieval Conference, Tehran. [In Persian]

Anbaee Farimani, S., Tabatabaee, H. & Kaffashan Kakhki, M. (2019). An Investigation into the Process of Organizing and Retrieving Web Texts based on the Integration of Semantic Concept in order to Organize Knowledge. *Iranian Journal of Information Processing & Management*; 34 (4), 1879-1904. [In Persian].

Arab, M. (2016). *Using hidden relations between words in word sense disambiguation*. M.S thesis, Computer engineering, Artificial intelligence. Shiraz University. [In Persian]

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, *7*(1), 1-16. Doi: 10.1093/llc/7.1.1

Booth, B. & Blair, M. (1992). *Thesaurus of sociological indexing terms*. Tehran, Iranian Research Institute for Information Science and Technology. [In Persian]

Bowker, L (2018). Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research. *Library Hi Tech*, *36*(2), 358-371. https://doi.org/10.1108/LHT-12-2017-0271

Delikhoun, L. (2016). *A Survey of Query Expansion (QE) of Users in Search Engines and Specialized Databases: A Case Study of Engineering Graduate Student at Alzahra University*. Thesis for Master, Knowledge and Information Science. Faculty of Education and Psychology, Alzahra University. [In Persian]

Ghayoomi, M. (2019). Identifying Persian Words' Senses Automatically by Utilizing the Word Embedding Method. *Iranian Journal of Information Processing & Management*; *35*(1), 25-50. [In Persian]

Goltaji, M. & Bazregar, S. (2010). Investigating the morphological problems of Persian language in three databases of ISC, Irandoc & Jahad institute. *Library and Information Sciences*, *13*(2), 191-214. [In Persian]

Gorman, K., Mazovetskiy, G. & Nikolaev, V. (2018). Improving homograph disambiguation with supervised machine learning. *LREC* (11), 1349-1352.

Hariri, N., Babalhavaeji, F., Farzandipour, M. & Nadi Ravandi, S. (2014). Evaluation Criteria of Information Retrieval Systems: What We Know and What We Do Not Know. *Iranian Journal of Information Processing & Management*; *30*(1)، 199-221. [In Persian]

Hasi, Y. Z., & Xiaoxiao, X. (2013). *Research on the homonyms disambiguation based on Mongolian nouns semantic network*. 6th International conference on intelligent networks and intelligent systems (ICINIS). Doi**:** 10.1109/ICINIS.2013.69

Hearst, M. A (1991). *Noun homograph disambiguation using local context in large text corpora*. Proceedings of the 7st Annual conference of the University of Waterloo Centre for the new OED and text research, Berkeley, 185-188.

Hoseini Beheshti, M. S. (2003). Application of terminology and word selection in machine indexing and information retrieval. *Library and Information science*, 18 (3), 31-44. [In Persian]

Hoseini Beheshti, M. S., Vafaei, S. & Noroozi Eghbali, M. (2015). *Thesaurus of mathematic*. Tehran, Iranian Research Institute for Information Science and Technology. [In Persian]

Jafari Pavarsi, H., Hariri, N., Alipour Hafezi, M., Babalhavaeji, F., & Khademi, M. (2020). Optimizing Semantic Information Retrieval by Labeling and Ontology. *Journal of National Studies on Librarianship and Information Organization*, *31*(1), 18-38. Doi: 10.30484/NASTINFO.2019.2247.186 .[In Persian]

Jalali, V. (2009). *Semantic Information retrieval using result concepts of a keyword based query*. Msc Thesis. Department of Computer Engineering and Information Technology, AmirKabir University of Technology. [In Persian]

Jones, C. & Waller, D. (2015). *Corpus linguistics for grammar. A guide for research*. London, Routledge.

Karimpour, R., Ghorbani, A., Pishdad, A., Mohtarami, M., AleAhmad, A., Amiri, H., & Oroumchian, F. (2008). *Using Part of Speech Tagging in Persian Information Retrieval*. http://ceur-ws.org/Vol-1174/CLEF2008wn-adhoc-KarimpourEt2008.pdf

Kessler, W. (2012). *Evaluation of Text Classification*. Retrieved 20 March, 2020 from: http://www.ims.unistuttgart.de/institut/mitarbeiter/kesslewd/lehre/sentimentanalysis12s/ml_evaluation.pdf.

Khan, L., McLeod, D., & Hovy, E. (2004). Retrieval effectiveness of an ontology-based model for information selection. The VLDB Journal— *The International Journal on Very Large Data Bases*, 13(1), 71-85. Doi: 10.1007/s00778-003-0105-1

Kheirmand Parizi, M. & Nourmandipoor, R. (2016). Persian word sense disambiguation with using of supervised method (IBL algorithms). *Computing Science Journal (CSJ)*, *1*(2), 63-73. [In Persian]

Lazarinis, F. (2007). Evaluating the searching capabilities of e-commerce web sites in a non-English language: A Greek case study, *Online Information Review*, *31*(6), 881-891. https://doi.org/10.1108/14684520710841829

Lewandowski, D. (2008). Problems with the use of Web search engines to find results in foreign languages. *Online Information Review*, *32*(4), 668-672. Doi: 10.1108/14684520810914034

Modarres Khiabani, S. (2010). Corpus-based approach on synonyms. *Pazhand*, *30*(8), 85-105. [In Persian]

Mirjoud, S. Ho., Ghiasi, M., Daliri, S., Kouchakinezhad, L., & Abasian Joshaghani, A. (2015). Comparison of the accuracy of general search engines and specialized search engines in retrieve medical images. *Educational Development of Jundishapur*, *6*(2), 131-138. [In Persian]

Norouzi, Y. & Homavandi, H. (2015). Survey of Image Search and Retrieval Problems in *5*(2), 206-222. [In Persian]

Noroozi Eghbali, M., Hoseini Beheshti, M. S & Noroozi Eghbali, M. (2007). *Thesaurus of physics*. Tehran, Iranian Research Institute for Information Science and Technology. [In Persian]

Prokofyev, R., Demartini, G., Boyarsky, A., Ruchayskiy, O., & Cudré-Mauroux, P. (2013*). Ontology-based word sense disambiguation for scientific literature. Advances in information retrieval*. 35th European conference on IR research, ECIR 2013. Berlin, Germany: Springer.594-605. Doi: 10.1007/978-3-642-36973-5_50

Rajabi, T., Gharibi, H., Hosseini Beheshti, M. & Noroozi Eghbali, M. (2004). *Thesaurus of Chemistry*. Tehran, Iranian Research Institute for Information Science and Technology. [In Persian]

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval*. *4*(4), 247-375. http://www.nowpublishers.com/articles/foundations-and-trends-in-informationretrieval/ INR-009 (accessed May 22, 2020). Doi: 10.1561/1500000009

Schutze, H. (2014). *Introduction to Inforation Retrieval: Relevance Feedback and Quary Extention*. http://www.cis.uni-muenchen.de/~hs/teach/13s/ir/pdf/09expand.pdf

Sedghi, F. (2013). *Homograph Disambiguation in Persian Context with Semi-Supervised Methods*. Msc Thesis. Faculty of computer engineering Alzahra University. [In Persian]

Sedighi, M., Hoseini Beheshti, M. S., & Noroozi Eghbali, M. (2004). *Thesaurus of geosciences*. Tehran, Iranian Research Institute for Information Science and Technology. [In Persian].

Shahbazi, M. & Shahini, S. (2016). Study of the the efficacy Magiran, Noormags and SID database in retrieval and relevance of Information Science and Knowledge subject by free keywords and Compare them in terms of the use of controlled keywords. *Iranian Journal of Information Processing & Management*, 31(2), 431-454. [In Persian]

Shapoori, S. (2000). Problems of subject search for users of the computer catalog of the Central Library of Ferdowsi University of Mashhad. *Library and Information Science, 3*(2), 49-68. [In Persian]

Soltani, M., & Faili. H. (2009). *Using of word sense disambiguation technique in cross language Information retrieval*. 14 th annual International CSI Computer Conference. Computer society of Iran, Tehran. AmirKabir University of Technology. [In Persian]

Sotoudeh, H. & Honarjooyan, Z. (2012). A review of the difficulties of the Persian language in the digital environment and their effects on the effectiveness of automatic text processing and information retrieval. *Library and Information Science, 15*(4), 59-92. [In Persian]

Sotoudeh, H. & Houshyar, M. (2018). The Role of Different Types of Homograph Contexts in Measuring Documents Similarities. *Iranian Journal of Information Processing & Management, 33*(3), 1195-1220. [In Persian]

Spink, A., Wolfram, D., Jansen, M. B., & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology*. 52(3), 226-234. Doi:10.1002/1097-4571(2000)9999:9999<::AID-ASI1591>3.3.CO;2-I

Stefanowitsch, A. (2006). *Corpus-based approaches to metaphor and metonymy. In Corpus-Based Approaches to Metaphor and Metonymy*. Ed. By Anatol Stefanowitsch and Stefan Th. https://doi.org/10.1515/9783110199895

Tabatabaie Jafari, Z. (2011). *A Survey to Query Expansion (QE) in Information Searching Behavior in Search Engines: A study of LIS graduate student Tehran states university*. Thesis for Degree of Master of ART (MA), Library & Information science. Faculty of Humanities, The University of Qom. [In Persian]

Tesprasit, V., Charoenpornsawat, P., & Sornlertlamvanich, V (2003). *A contextsensitive homograph disambiguation in Thai text-to-speech synthesis*. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Thailand. 2,103-105. Doi: 10.3115/1073483.1073518

Voorhees, E. M. (1998). *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*. In C. J.Van rijs Bergen W. Bruce Croft,Alistair Moffat, Proceedings of the 21st Annual International ACM Sigir Conference on Research and Development in Information Retrieval, Pages315-323. Australia, University of Melbourne https://doi.org/10.1145/290941.291017

Yousefi, A. (1997). False drop in Information storage and retrieval. *Iranian Research Institute for Information Science and Technology (IranDoc), 13*(1), 10-19. [In Persian]

Yousefi Rad, E. (2009). R.D.F: A model for resource description in semantic web. *National Studies on Librarianship and Information Organization, 20*(3), 9-22. [In Persian]

Zardary, S. (2016). *Ontology engineering of knowledge and information science based on Encyclopedia of Library and Information Science*. Ph.D. degree, Knowledge and Information Science Department. Faculty of Education and Psychology, Shahid Chamran University of Ahwaz. [In Persian]

Zhang, J. & Lin, S. (2007). Multiple language supports in search engines, *Online Information Review, 31*(4), 516-532. Doi:10.1108/14684520710780458

Zolfaghar, Z., Mosavi Miangah, T., Rovshan, B. & Vakilifard, A. R. (2020). A Study on the Improved Techniques of Corpus-based Frequency Approaches in Automatic Term Extraction (ATE) (The Case Study: Basic Medicine Vocabulary). *Iranian Research Institute for Information Science and Technology (IranDoc), 35*(4), 1039-1064. [In Persian]