یک روش کاربردی برای جمع‌آوری توده متون جامع مورد نیاز برای تولید هستان‌شناسی حوزه فازی

احسان شریفی[1]، محمود دی‌پیر[2]

۱مربی دانشکده مهندسی کامپیوتر، دانشگاه هوایی شهید ستاری، sharifi@ssau.ac.ir

۲ استادیار دانشکده مهندسی کامپیوتر، دانشگاه هوایی شهید ستاری

تاریخ دریافت: ۹۲/۱۰/۲ تاریخ پذیرش: ۹۴/۱۰/۲

چکیده

هستان‌شناسی، توصیفی از یک حوزه در قالب یک ساختار قابل فهم توسط انسان و قابل خواندن توسط ماشین می‌باشد که از مفاهیم، صفات، روابط و قواعد تشکیل شده است. هستان‌شناسی حوزه، نوع خاصی از هستان‌شناسی است که به منظور بازنمایی دانش مرتبط با یک حوزه کاربردی خاص مورد استفاده قرار می‌گیرد. اما بدلیل طبیعت غیر دقیق دانش موجود در جهان واقع، هستان‌شناسی‌های فازی گزینه مناسب‌تری نسبت به هستان‌شناسی‌های محض برای مدلسازی معنایی جهان واقع می‌باشند. هستان‌شناسی‌های فازی از منطق نرم بجای منطق سخت برای بازنمایی دانش بهره می‌برند. کیفیت هستان‌شناسی فازی، رابطه مستقیم با جامعیت توده متونی دارد که برای ساخت هستان‌شناسی از آن استفاده می‌کنیم. ما در این مقاله یک روش کاربردی برای جمع‌آوری توده متون جامع مورد نیاز برای تولید هستان‌شناسی حوزه فازی پیشنهاد می‌کنیم. این روش شامل مراحل ایجاد مدل حوزه و جمع‌آوری توده متون با استفاده از خزشگر وب تاکیدی می‌باشد. درانتها نیز به منظور ارزیابی کارایی هستان‌شناسی به بررسی کاربرد هستان‌شناسی حوزه فازی در فرآیند توصیف و همتایابی وب‌سرویس‌های معنایی می‌پردازیم.

کلیدواژه

هستان‌شناسی حوزه فازی، مدل حوزه، خزشگر وب تاکیدی، همتایابی وب‌سرویس معنایی

مقدمه

واژه هستان‌شناسی[1] دارای تعاریف متفاوتی در دو حوزه فلسفه و علوم کامپیوتر می‌باشد. هستان‌شناسی از نظر فلسفی ریشه در متافیزیک دارد. متافیزیک را ریشه علم فلسفه می‌دانند و آنرا از الهیات جدا می‌کنند. ارسطو متافیزیک را فلسفه اول نامید و هستان‌شناسی را به عنوان یکی از اجزای متافیزیک به صورت مطالعه وجود و هستی تعریف نمود[2]. در حوزه علوم کامپیوتر پر ارجاع‌ترین تعریف از هستان‌شناسی توسط گروبر[3] ارائه شد. بر اساس این تعریف، هستان‌شناسی توصیفی[4] صریح[5] از یک ادراک[6] است. ادراک یک نمای ساده شده و انتزاعی[7] از حوزه‌ای است که قصد داریم آنرا برای یک هدف مشخص بازنمایی[8] کنیم. این تعریف توسط استادر[9] بدین صورت تکمیل گردید: هستان‌شناسی یک توصیف صریح و صوری[10] از یک ادراک اشتراکی[11] می‌باشد. در واقع هستان‌شناسی درک مشترکی از مفاهیم موجود در یک حوزه[12] به همراه روابط بین آنها ارائه می‌نماید که برای حل مشکل چندمعنایی موجود در اکثر کاربردها مفید می‌باشد. اما هستان‌شناسی‌های کلاسیک دارای یک ایراد ذاتی می‌باشند. آنها برای نمایش مفاهیم غیر دقیق موجود در اکثر حوزه‌ها مناسب نیستند. مفاهیم غیر دقیقی نظیر نزدیک و بلند یا مفهومی نظیر هاورکرافت[13] که ترکیبی از هواپیما، قایق و اتوموبیل است توسط هستان‌شناسی کلاسیک قابل توصیف نیست. بدین منظور، هستان‌شناسی‌های فازی پا به عرصه ظهور نهادند. هستان‌شناسی‌های فازی از منطق فازی برای بازنمایی دانش مبهم و همچنین تسهیل استدلال بر روی آن بهره می‌برند. طی سالیان اخیر هستان‌شناسی‌های فازی در زمینه‌های مختلفی نظیر بازیابی

1 Ontology
2 The study of being and existence
3 Gruber
4 Specification
5 Explicit
6 Conceptualization
7 Abstract
8 Represent

9 Studer
10 Formal
11 Shared
12 Domain
13 Hovercraft

اطلاعات، موتورهای جستجو و مباحث مرتبط با وب معنایی مورد استفاده قرار گرفته‌اند. ایجاد غیرخودکار هستان‌شناسی فازی از یک ساختار سلسله مراتبی از مفاهیم و یا هستان‌شناسی‌های غیرفازی موجود فرآیندی خسته‌کننده و مشکل بوده و اغلب نیازمند حضور افراد خبره در این زمینه می‌باشد. لذا معرفی یک روش کارآمد به منظور بازیابی خودکار هستان‌شناسی حوزه فازی امری مطلوب می‌باشد. دیدگاه‌های متنوعی برای تولید خودکار هستان شناسی فازی در طی سالیان اخیر معرفی شده است. ایراد اصلی اکثر دیدگاه‌های موجود، عدم وجود راهبردی مشخص برای جمع‌آوری توده متون[۱۴] جامع مرتبط با حوزه مورد نظر می‌باشند. توده متون شامل کلیه منابعی است که برای تولید هستان‌شناسی حوزه فازی از آنها بهره می‌بریم. کیفیت و جامعیت یک هستان‌شناسی، رابطه مستقیم با جامعیت توده متونی دارد که برای ساخت هستان‌شناسی از آن استفاده می‌کنیم. لذا ارائه یک روش کاربردی برای تولید توده متون جامع می‌تواند باعث ارتقای کیفیت هستان‌شناسی فازی گردد.

در این مقاله در ابتدا شاخص‌ترین فعالیت‌های انجام شده در خصوص تولید هستان‌شناسی حوزه فازی را با تمرکز بر روی نحوه جمع‌آوری توده متون مورد بررسی قرار می‌دهیم. سپس به معرفی روش پیشنهادی برای جمع‌آوری توده متون جامع پرداخته و در ادامه به ارزیابی این روش می‌پردازیم.

کارهای انجام شده

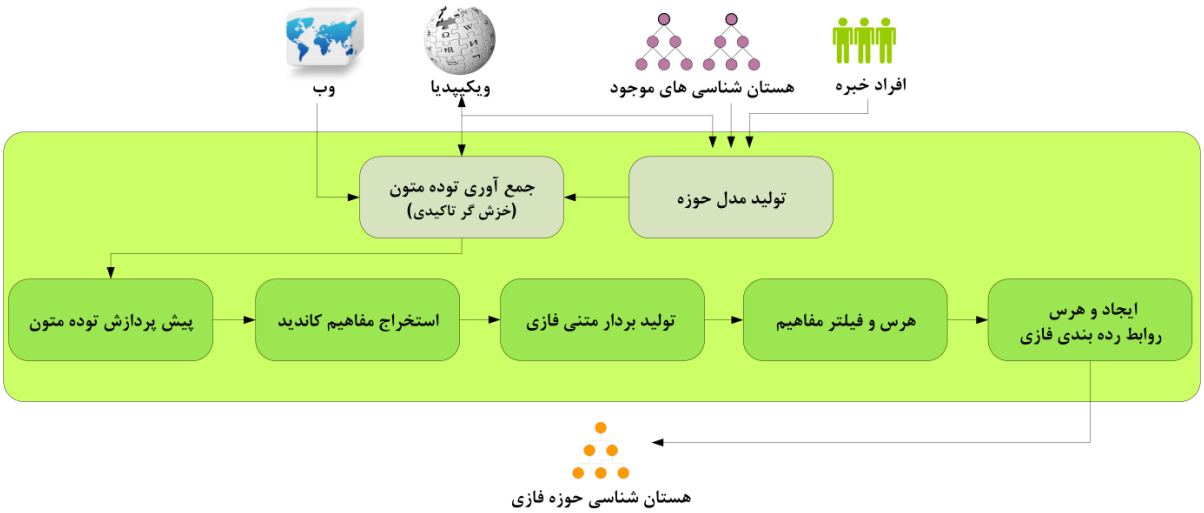
در زمینه تولید هستان‌شناسی فازی تلاش‌های متعددی طی سالیان اخیر صورت گرفته است. پس از بررسی مقالات و پژوهش‌های صورت گرفته در این زمینه، شش دیدگاه شاخص در این زمینه انتخاب گردید که در ادامه به اجمال مورد بررسی قرار می‌گیرند. معیار انتخاب بر اساس میزان ارجاع به مقالات مذکور و کاربردی بودن دیدگاه مربوطه می‌باشد.

یکی از اولین دیدگاه‌های مطرح در زمینه ایجاد هستان شناسی فازی توسط سیمیانو و همکاران[۱۵] در سال ۲۰۰۵ معرفی شد. آنها یک الگوریتم برای آموزش خودکار رده‌بندی[۱۶] معرفی نمودند که توسط آن امکان تشخیص سلسله‌مراتب مفاهیم از توده متون فراهم گردید. این الگوریتم بر اساس تحلیل مفهوم صوری[۱۷] بنا شده بود. FCA یک متد اصولی برای بازیابی روابط تلویحی بین اشیائی است که توسط مجموعه‌ای از ویژگی‌ها توصیف گردیده‌اند. ایده محوری در FCA زمینه صوری[۱۸] است که بیانگر ویژگی‌ها و خصوصیات مهم و مشترک بین مجموعه اشیاء یک کلاس می‌باشد.

این دیدگاه فرض می‌کند که توده متون موجود بوده و لذا هیچ روشی برای جمع‌آوری آن ارائه نمی کند[۱]. دیدگاه بعدی، ارائه یک هستان‌شناسی فازی توسط گسترش یک هستان‌شناسی حوزه محض بود که به منظور تلخیص اخبار مورد استفاده قرار گرفت و توسط لی و همکاران[۱۹]، در سال ۲۰۰۵ معرفی گردید. در این دیدگاه به عنوان یک روش نیمه خودکار، ابتدا یک هستان‌شناسی حوزه در مورد رویدادهای خبری مختلف توسط فرد خبره تعریف می‌گردد. سپس با استفاده از یک مکانیزم پیش‌پردازش سند، واژگان معنادار موجود در توده متون اخبار بر اساس فرهنگ لغات چینی که توسط فرد خبره پیش‌تعریف گردیده است استخراج می‌گردد. واژه‌های معنادار بر اساس رویدادهای خبری و توسط دسته‌بند واژه دسته‌بندی می‌گردند. توده متون مورد نیاز برای تولید این هستان‌شناسی فازی، اخبار منتشر شده در سطح وب می‌باشد که به طور متناوب توسط یک عامل بازیاب[۲۰] از وب‌سایت‌های خبری کشور تایوان جمع‌آوری می‌گردد[۲]. یکی دیگر از مهم‌ترین چارچوب‌ها توسط سو و همکاران[۲۱] در سال ۲۰۰۶ معرفی گردید. در این چارچوب که FOGA نامیده می‌شود هستان‌شناسی فازی به صورت یک چهارتایی به شکل $F_O = (C, A^C, R, X)$ تعریف می‌گردد که C مجموعه مفاهیم، A^C مجموعه صفات و R بیانگر مجموعه روابط رده‌بندی و غیررده‌بندی بوده و X بیانگر مجموعه قواعد هستان‌شناسی می‌باشد. تحلیل مفهوم صوری فازی، خوشه‌بندی مفهومی فازی[۲۲] و تولید روابط رده‌بندی[۲۳] بخش‌های اصلی این چارچوب را تشکیل می‌دهند. در این دیدگاه فرض شده است که یک پایگاه داده از اطلاعات غیر دقیق به عنوان توده متون از ابتدا موجود می‌باشد[۳]. دیدگاه بعدی توسط لائو و همکاران[۲۴] در سال ۲۰۰۹ معرفی گردید. هستان‌شناسی فازی در این دیدگاه یک شش‌تایی به شکل $Ont = <X, A, C, R_{XC}, R_{AC}, R_{CC}>$ می‌باشد که در آن X مجموعه‌ای از اشیاء، A مجموعه‌ای از صفاتی که اشیاء را توصیف می‌کنند و C مجموعه‌ای از مفاهیم می‌باشند. سه نوع رابطه رده‌بندی فازی R_{XC}، R_{AC} و R_{CC} موجود بین این سه مجموعه نیز هستان‌شناسی را تکمیل می‌نماید. در این دیدگاه در ابتدا عملیات پیش‌پردازش متن بر روی توده متون صورت می‌پذیرد.

14 Corpus
15 Cimiano et al
16 Taxonomy
17 Formal Concept Analysis (FCA)
18 Formal Context

19 Lee et al.
20 Retrieval agent
21 Tho et al
22 Fuzzy Conceptual Clustering
23 Hierarchical Relation Generation
24 Lau et al.



شکل ۱. روش پیشنهادی برای تولید هستان‌شناسی حوزه فازی

سپس با استفاده از تکنیک پنجره‌سازی توده متون، کل متن از ابتدا تا انتها توسط پنجره‌هایی با طول مشخص مورد بررسی قرار می‌گیرد. در حین این فرآیند اطلاعات آماری مرتبط با رخدادهای توام بین الگوهای نحوی خاص موجود در پنجره محاسبه می‌گردد. این الگوها به طور بالقوه نماینده مفاهیم موجود در توده متن می‌باشند. سپس به منظور حذف مفاهیم غیر مرتبط به حوزه مذکور به بررسی نحوه رخداد مفاهیم در سایر حوزه‌ها پرداخته و مفاهیمی که با فرکانس تکرار بالا در سایر حوزه‌ها تکرار گردیده‌اند از لیست مفاهیم کاندید این حوزه حذف می‌گردند. این روش به دلیل خاص منظوره بودن، توده متون مورد نیاز خود را از میان پیام‌های آنلاین، پست‌های وبلاگ، ایمیل‌ها و اتاق‌های گفتگو بدون هیچ‌گونه راهبرد خاصی جمع‌آوری می‌نماید[۴].

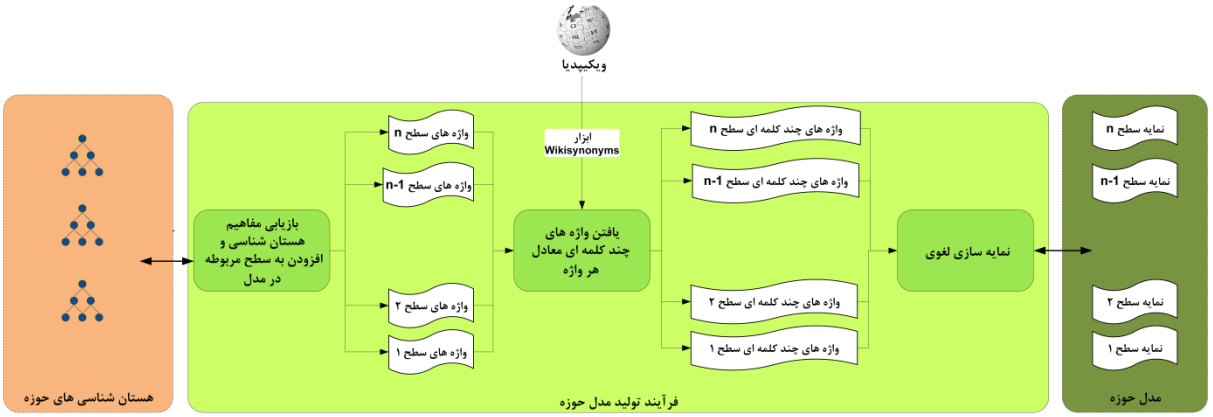
یاگوئینوما و همکاران[۲۵] در سال ۲۰۱۱ تلاش نمودند به منظور مجتمع‌سازی داده‌ها از هستان‌شناسی‌های فازی استفاده نمایند. آنها سیستم DISFOQuE را برای مجتمع‌سازی داده بر پایه منطق فازی ارائه نمودند. این دیدگاه نیز روش مشخصی برای تولید توده متون جامع پیشنهاد ننموده است[۵]. دیدگاه مطرح بعدی توسط ژو و همکاران[۲۶] در سال ۲۰۱۳ ارائه گردید. این دیدگاه برای تولید هستان‌شناسی فازی در حوزه تحقیقات علمی از دو هستان‌شناسی فازی به نامهای T2FO[۲۷] و IVFO[۲۸] در قالب دو شش‌تایی به فرم $O_{2F} = \{C,A^C,R,A^R,H,X\}$ بهره می‌برد که در آن C مجموعه‌ای از مفاهیم، A^C مجموعه‌ای از صفات مرتبط با هر مفهوم، A^R بیانگر مجموعه صفات رابطه R، H بیانگر رده‌بندی مفاهیم و X مجموعه قواعد می‌باشد. رابطه R در T2FO بیانگر روابط فازی نوع دو[۲۹] بوده که شامل روابط رده‌بندی و غیررده‌بندی می‌باشد. اما رابطه R در IVFO بیانگر روابط فازی فاصله‌ای[۳۰] می‌باشند. این دیدگاه نیز برای تولید و جمع‌آوری توده متون دارای راهبرد مشخصی نمی باشد[۶].

معرفی روش پیشنهادی

در این مقاله، چارچوب ارائه شده توسط لائو و همکاران[۴] با معرفی روشی برای تولید توده متون جامع، توسعه می‌یابد. در این روش، با ایجاد یک مدل حوزه[۳۱] و استفاده از یک خزشگر تاکیدی[۳۲] ، امکان جمع‌آوری توده متون جامع مرتبط با حوزه مورد نظر فراهم می‌گردد. خزشگر تاکیدی با استفاده از یک لیست از واژگان و یا هستان‌شناسی‌های مرتبط با حوزه به جمع‌آوری صفحات مرتبط به آن حوزه خاص می‌پردازد. رفتار این نوع خزشگر در حالت کلی مشابه خزشگرهای معمولی است با این تفاوت که پس از بررسی صفحات، فقط در صورت مطلوب بودن اقدام به ذخیره‌سازی و شاخص‌گذاری آنها می‌نماید و در غیر اینصورت از صفحات مذکور و پیوندهای آنها صرف نظر می‌نماید. لذا کیفیت توده متون جمع‌آوری شده توسط خزشگر رابطه مستقیم با مدل مرجعی دارد که میزان مطلوبیت توده متون با توجه به آن سنجیده می‌شود. در تولید مدل حوزه، از سه منبع هستان‌شناسی‌های موجود، دایره‌المعارف ویکیپدیا و دانش افراد خبره بهره می‌بریم. با استفاده از دانش افراد خبره و همچنین هستان‌شناسی‌های حوزه موجود، مفاهیم موجود در حوزه مورد نظر تعیین و سطوح مختلف مدل ایجاد می‌شود.

25 C. A. Yaguinuma et al
26 Xue et al
27 Type-2 Fuzzy Ontology
28 Interval Valued Fuzzy Ontology
29 Type-2 Fuzzy Relation

30 Interval Fuzzy Relations
31 Domain model
32 Focused Crawler



شکل ۲. تولید مدل حوزه

سپس با استفاده از دایره‌المعارف ویکیپدیا مقالات مرتبط با مفاهیم مذکور بازیابی‌شده و پس از انجام فرآیند نمایه‌سازی لغوی[۳۳] مدل تکمیل می‌گردد. در ادامه فرآیند تولید مدل حوزه به تفصیل بیان می‌شود. شکل۱ نمای کلی روش پیشنهادی را نمایش می‌دهد. این روش طی هفت مرحله، فرآیند تولید هستان‌شناسی فازی حوزه را بیان می‌نماید. در ادامه به بررسی تفصیلی دو مرحله اول که بیانگر نوآوری این مقاله می‌باشند می‌پردازیم. سپس سایر مراحل به اختصار مورد بررسی قرار می‌گیرد. برای آشنایی بیشتر با دیدگاه لائو و همکاران مراجعه به مقالات مربوط به آن پیشنهاد می‌شود.

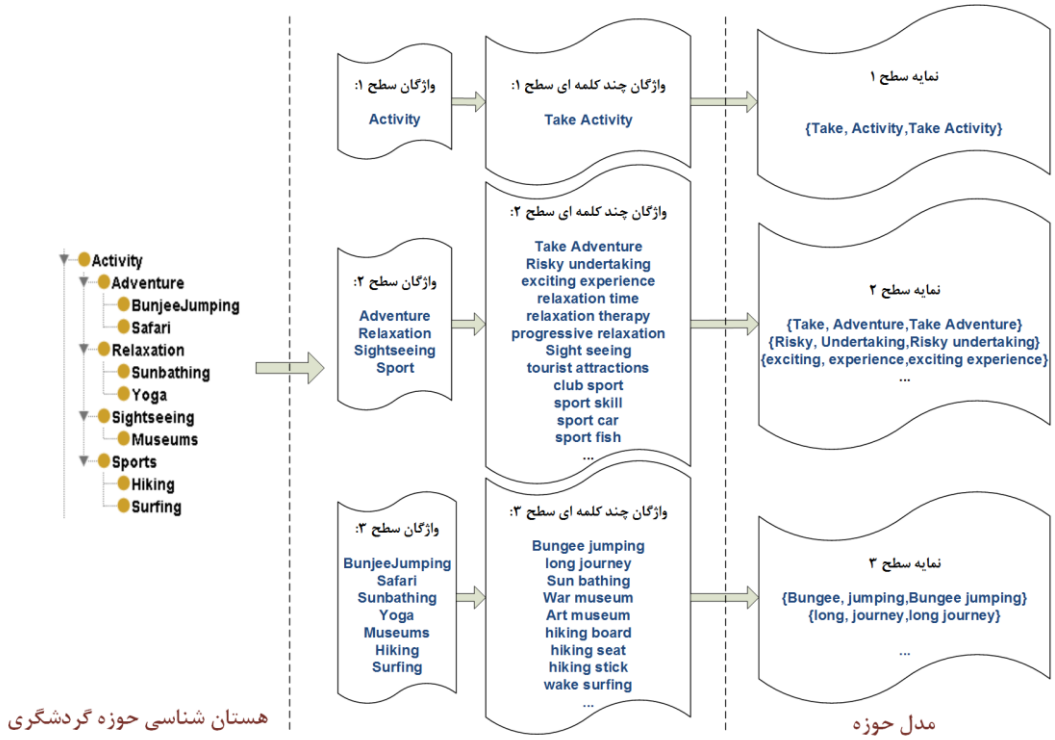
تولید مدل حوزه

به منظور تولید مدل حوزه، در ابتدا یک لیست اولیه از واژگان مرتبط به حوزه مورد نظر با استفاده از هستان‌شناسی‌های حوزه موجود تولید می‌گردد. به منظور بازیابی هستان‌شناسی‌های حوزه، از موتور جستجوی معنایی[۳۴] Swoogle استفاده می‌شود. در ادامه مفاهیم موجود در هستان‌شناسی‌های مذکور بازیابی می‌شوند. در این مرحله می‌توانیم به منظور افزایش دقت، مفاهیم غیرمرتبط به حوزه مورد نظر را با استفاده از دانش افراد خبره فیلتر نماییم. هر مفهوم پس از ریشه‌یابی[۳۵] با توجه به شماره رده‌ای که در هستان‌شناسی مبداء دارد به سطح متناظر خود در مدل افزوده می‌شود. شماره رده مفاهیم موجود در ریشه ۱ و فرزندان آن ۲ و به همین ترتیب سایر مفاهیم موجود در هر رده شماره‌گذاری می‌گردند. در صورتی‌که یک واژه[۳۶] تکراری بخواهد به سطحی از مدل اضافه شود، با توجه به شماره رده واژه مذکور، حالت‌های مختلفی رخ می‌دهد. اگر واژه تکراری با شماره رده n بخواهد به سطحی از مدل با شماره سطح n اضافه شود آن واژه حذف می‌گردد. اگر واژه جدید با شماره رده فرضی n، قبلا در یکی از سطوح مدل با شماره سطح m وارد شده باشد در صورتی‌که n>m باشد واژه قبلی از سطح مربوطه در مدل حذف گردیده و به سطح n منتقل می‌شود. اما در حالتی که n<=m باشد واژه جدید به مدل اضافه نشده و از آن صرف نظر می‌شود.

در مرحله بعد برای هر یک از سطوح مدل، فرآیند نمایه‌سازی لغوی اجرا می‌گردد. نمایه[۴۰] مرتبط با یک واژه به صورت تمامی ترکیبات خطی[۴۱] ممکن از زیر رشته‌های موجود در آن تعریف می‌گردد. اما از آنجائی‌که فرآیند نمایه‌سازی لغوی بر روی واژگان چندکلمه‌ای عینیت می‌یابد می‌بایست ابتدا واژگان چندکلمه‌ای[۴۲] مرتبط با هریک از واژه‌های تک‌کلمه‌ای مدل بازیابی شده و سپس فرآیند نمایه‌سازی لغوی اجرا گردد. بدین منظور از ابزار WikiSynonyms[۴۳] بهره می‌بریم. این ابزار برای یافتن واژه‌های مترادف یک کلمه از دایره‌المعارف ویکیپدیا بهره می‌برد. این ابزار به صورت یک Api در اختیار برنامه‌نویسان می‌باشد. در این مرحله مترادف‌های چندکلمه‌ای تمامی واژه‌های سطوح مختلف مدل با استفاده از این ابزار بازیابی می‌گردد. به منظور بالا بردن کیفیت مدل حوزه می‌توانیم از دانش افراد خبره به منظور حذف واژه‌های غیرمرتبط بهره ببریم. شکل۲ فرآیند تولید مدل حوزه را نمایش می‌دهد. جدول۱ به عنوان مثال بخشی از واژه‌های چندکلمه‌ای مترادف واژه Computer به همراه نمایه‌های لغوی آن‌را نمایش می‌دهد. شکل ۳ نیز بخشی از فرآیند تولید یک مدل حوزه را بر اساس هستان‌شناسی حوزه گردشگری[۴۴] نمایش می‌دهد.

33 Lexical Profiling
34 http://swoogle.umbc.edu
35 Stemming
36 در ادامه مفهوم ریشه‌یابی شده را واژه term اطلاق می کنیم

40 Profile
41 Linear Combinations
42 Multiword Terms
43 http://wikisynonyms.ipeirotis.com/page/api
44 Travel



شکل ۳. تولید مدل حوزه گردشگری

جدول ۱ . نمایه‌های لغوی مربوط به واژه computer

واژه‌های چندکلمه‌ای	نمایه لغوی
Computer system Digital computer Computing device Electronic computer Hardware system	Computer, System, Computer system Digital, computer, Digital computer Computing, device, Computing device Electronic, computer, Electronic computer Hardware, system, Hardware system

به عنوان نمونه در شکل۳، واژه Activity در هستان‌شناسی متناظرش در رده ۱ قرار دارد و لذا وقتی به مدل افزوده می‌شود در سطح ۱ مدل قرار می‌گیرد. سپس لیست واژگان چندکلمه‌ای مترادف آن از دایره‌المعارف ویکیپدیا بازیابی و فرآیند نمایه‌سازی لغوی بر روی آن اجرا می‌گردد.

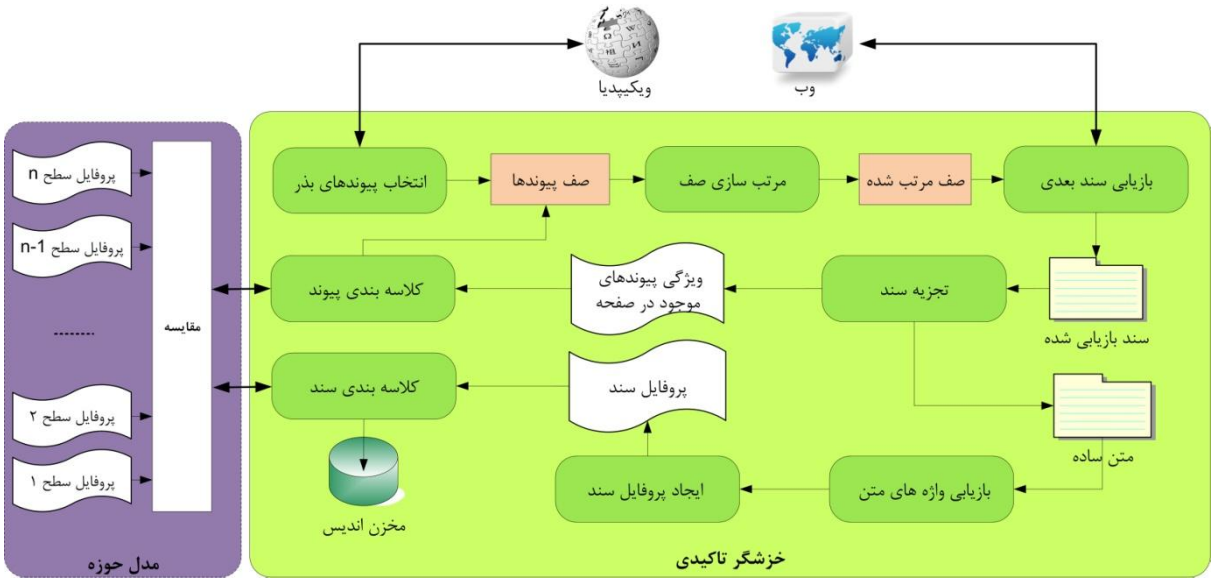
جمع‌آوری توده متون توسط خزشگر تاکیدی

ما به منظور جمع‌آوری توده متون مرتبط به حوزه مورد نظر نیازمند استفاده از یک خزشگر تاکیدی بودیم. لذا پس از بررسی پژوهش‌های انجام شده در این زمینه و ابزارهای موجود، یک نمونه خزشگر تاکیدی انتخاب گردید. این خزشگر توسط گرینوود[۴۵] و در دانشگاه منچستر طراحی و پیاده‌سازی گردیده است[۷].

در این خزشگر، صفحات وب موجود در رده‌بندی‌های مختلف فهرست‌های عمومی وب نظیر Yahoo Directory[۴۶], Open Directory Project[۴۷] یا Wikipedia Category به عنوان توده متون آموزشی به خزشگر ارائه گردیده و به سطوح سلسله مراتبی متفاوتی تقسیم شده و سپس به شکل لغوی نمایه می‌گردند. ما به منظور بکارگیری خزشگر گرینوود در روش پیشنهادی این مقاله تغییراتی در آن اعمال نمودیم. بدین منظور بجای اینکه صفحات وب موجود در رده‌های مختلف فهرست‌های عمومی مورد پردازش قرار گرفته و واژه‌های کلیدی موجود در آنها بازیابی گردند از مدل حوزه ایجاد شده در مرحله قبل برای شروع فرآیند خزش بهره می‌بریم. به منظور شروع فرآیند خزش از پیوندهای بذر[۴۸] استفاده می‌گردد. لذا در ابتدا به جمع‌آوری پیوندهای بذر مورد نیاز با استفاده از دایره‌المعارف ویکیپدیا می‌پردازیم. بدین منظور از ابزار JWPL برای ارتباط با ویکیپدیا استفاده می‌نماییم. به منظور کنترل فرآیند خزش تعداد کل صفحاتی که ملاقات می‌شوند محدود به یک حد مشخص می‌باشد که این حد آستانه توسط فرد خبره تعیین می‌گردد. فرآیند انجام خزش بدین شکل است که در ابتدا، پیوندهای بذر بازیابی شده در مرحله قبل درون یک صف قرار می‌گیرند.

45 Greenwood

46 http://dir.yahoo.com
47 http://www.dmoz.org
48 Seed



شکل ۴. مدل فرآیند خزشگر تاکیدی

سپس در هر تکرار از فرآیند خزش، صفحه وب مرتبط با پیوند موجود در ابتدای صف بازیابی و محتویات آن مورد ارزیابی قرار گرفته و عملیات دسته‌بندی[۴۹] صفحه صورت می‌پذیرد. ایده اصلی فرآیند دسته‌بندی صفحات، اندازه‌گیری میزان تشابه لغوی بین سطوح مدل و صفحاتی است که در فرآیند خزش بازدید می‌شوند. همچنین تمامی پیوندهای موجود در این صفحه بر اساس خصوصیات واژگانی خود امتیازدهی گردیده و به انتهای صف افزوده می‌شوند. هنگام دسته‌بندی پیوندهای موجود در هر صفحه، مشخصاتی نظیر متن پیوند مدنظر می‌باشد. معمولا این متن، توصیفات کوتاهی در مورد صفحه مورد نظر ارائه می‌دهد.

این فرآیند تا زمانی که صف خالی گردد و یا حد خزش فرا رسد ادامه می‌یابد. همچنین متنی که در حوالی پیوند مزبور واقع شده است شامل اطلاعات دقیقی در مورد صفحه‌ای است که توسط این پیوند قابل دسترسی می‌باشد. هنگامی‌که یک پیوند در یک صفحه یافت می‌شود، کلمات اطراف این پیوند تا یک طول مشخص استخراج و نمایه می‌گردند. سپس میزان تشابه لغوی این کلمات با سطوح مختلف رده‌بندی مورد ارزیابی واقع می‌گردد. شکل ۴ مدل فرآیند این خزشگر را نمایش می‌دهد.

پیش پردازش توده متون

پس از جمع‌آوری توده متون مرتبط با حوزه مورد نظر توسط خزشگر تاکیدی وارد مرحله پیش‌پردازش می‌شویم. این مرحله شامل حذف کلمات ایست[۵۰]، برچسب‌گذاری ادات سخن[۵۱] و ریشه‌یابی کلمات[۵۲] می‌باشد. در صورتی‌که توده متن شامل مستندات با قالب HTML باشد قبل از مراحل فوق نیازمند حذف تگ‌های HTML نیز می‌باشیم.

استخراج مفاهیم کاندید

در زبان‌شناسی محاسباتی یک واژه از یک یا چند کلمه تشکیل شده است و اگر این واژه حامل معنی قابل تشخیصی با توجه به متن و حوزه‌ای که در آن ظاهر شده است باشد به عنوان یک مفهوم مطرح می‌باشد. لذا به منظور کاهش تولید مفاهیم نویزی صرفاً به بررسی الگوهای زبانی خاص در یک حیطه مشخص می‌پردازیم. مراحل اصلی عملیات استخراج مفاهیم شامل پنجره‌سازی متن[۵۳]، استخراج الگوهای نحوی مورد نظر و محاسبه اطلاعات آماری[۵۴] بین واژه‌های متن می‌باشد. فرآیند کار در پنجره‌سازی بدین شکل است که در این بخش هر یک از سندهای موجود در توده متون را توسط پنجره‌هایی مجازی با اندازه‌های δ واژه از ابتدای هر سند مورد پردازش قرار می‌دهیم. بر اساس آزمون‌های صورت گرفته توسط لائو، اندازه موثر این پنجره‌ها ۵ تا ۱۰ واژه می‌باشد. پس از اتمام پردازش در هر یک از این پنجره‌های مجازی، این پنجره را به اندازه یک کلمه به سمت راست شیفت می‌دهیم تا اینکه کل سند مورد پردازش قرار گیرد. این فرآیند برای تمام اسناد موجود در توده متن تکرار می‌گردد. به منظور استخراج الگوهای نحوی مورد نظر، با استفاده از مفهوم عبارات با قاعده[۵۵] صرفا به استخراج الگوهای نحوی خاصی که بیانگر مفاهیم کاندید موجود در هر پنجره می‌باشند می‌پردازیم. پردازش الگوهای نحوی خاص باعث عدم تولید مفاهیم نویزی می‌شود. نوع الگوی

49 Classify
50 Stop Words Removal
51 POS Tagging
52 Word Stemming

53 Text Windowing
54 Statistical data
55 Regular Expersion

انتخابی به تعداد واژه‌های تشکیل‌دهنده مفاهیم کاندید وابسته می‌باشد. تعداد واژه‌ها از یک واژه تا هر تعداد واژه مد نظر کاربر قابل تغییر می‌باشند. به منظور محاسبه اطلاعات آماری بین واژه‌های متن بر روی اطلاعات وابستگی بین واژه‌ها و جمع‌آوری اطلاعات آماری مرتبط با الگوهای نحوی مرحله قبل متمرکز می‌شویم. در واقع در این مرحله اطلاعات آماری مرتبط با مفاهیم کاندید نظیر فرکانس تکرار آنها در هر پنجره و شماره پنجره‌هایی که در آنها ظاهر می‌گردند محاسبه و ذخیره می‌گردد.

تولید بردار متنی فازی

پس از استخراج مفاهیم کاندید و واژه‌های دارای فرکانس تکرار بالا، بردار متنی برای هر مفهوم کاندید محاسبه می‌شود. معیارهای مختلفی نظیر NGD[56]، JA[57]، CP[58]، KL[59]، ECH[60] برای محاسبه درجه عضویت یک واژه به مفهوم کاندید معرفی شده است. معیار مورد استفاده در دیدگاه لائو[61] BMI می‌باشد[۸]. $\mu_{ci}(t_j)$ بیانگر میزان وابستگی واژه t_j به مفهوم کاندید c_i می‌باشد که بر اساس روش BMI محاسبه می‌گردد. رابطه ۱ نحوه محاسبه را نمایش می‌دهد. در این فرمول $Pr(t_i, t_j)$ بیانگر احتمال توام ظاهر شدن هر دو واژه t_i و t_j در یک پنجره می‌باشد. $Pr(t_i)$ نیز بیانگر احتمال ظاهر شدن واژه t_i در پنجره متنی می‌باشد. این احتمال بر اساس رابطه $\frac{|w_t|}{|w|}$ محاسبه می‌گردد که در آن $|w_t|$ بیانگر تعداد پنجره‌هایی است که شامل واژه t بوده و $|w|$ بیانگر کل تعداد پنجره‌های ساخته شده از کل توده متن می‌باشد.

$$\mu_{C_i}(t_j) \approx BMI(t_i, t_j) = \beta \times \left[\begin{array}{l} P_r(t_i,t_j)\log_2\left(\frac{P_r(t_i,t_j)+1}{P_r(t_i)P_r(t_j)}\right) \\ +\, P_r(\neg t_i,\neg t_j)\log_2\left(\frac{P_r(\neg t_i,\neg t_j)+1}{P_r(\neg t_i)P_r(\neg t_j)}\right) \end{array} \right] - (1-\beta)\times \left[\begin{array}{l} P_r(t_i,\neg t_j)\log_2\left(\frac{P_r(t_i,\neg t_j)+1}{P_r(t_i)P_r(\neg t_j)}\right) \\ +\, P_r(\neg t_i,t_j)\log_2\left(\frac{P_r(\neg t_i,t_j)+1}{P_r(\neg t_i)P_r(t_j)}\right) \end{array} \right] \quad (۱)$$

در این متد برای محاسبه میزان وابستگی بین واژه‌ها، احتمالات مرتبط با وجود و عدم وجود واژه‌ها در هر پنجره مورد بررسی قرار می‌گیرد. عامل وزنی β نیز برای کنترل میزان اهمیت دو نوع ملاک موجود در این رابطه استفاده می‌گردد (وجود یا عدم وجود). $\mu_{ci}(t_j)$ در واقع یک ساز و کار محاسباتی برای تولید رابطه R_{AC} در هستان‌شناسی فازی می‌باشد.

هرس و فیلتر مفاهیم

اگر درجه عضویت یک واژه به مفهوم از یک حد آستانه مشخص ζ کمتر باشد آنرا از بردار متنی وابسته به آن مفهوم حذف می‌نماییم. این عمل معادل برش α می‌باشد. از طرف دیگر اگر تمامی واژه‌های متعلق به بردار متنی یک مفهوم از حد آستانه کمتر باشند مفهوم کاندید مورد نظر از لیست مفاهیم نهایی حذف می‌گردد. این حدود آستانه بر اساس حدود آستانه‌ای انتخاب شده در دیدگاه لائو تعیین شده‌اند. به منظور حذف مفاهیم کاندید نامرتبط به حوزه مورد نظر، به بررسی و مقایسه رخداد مفهوم مورد نظر در سایر حوزه‌ها و حوزه مورد نظر خود می‌پردازیم. بر اساس این دیدگاه مفاهیمی که با بسامد تکرار بالایی در یک حوزه خاص تکرار شوند عموما مفاهیم خاص آن حوزه می‌باشند. مفاهیم کاندیدی که میزان ارتباط آنها به دامنه مورد نظر از یک حد آستانه مشخص کمتر بود از لیست نهایی حذف می‌گردند.

ایجاد و هرس روابط رده‌بندی فازی

در این مرحله رابطه رده‌بندی بین مفاهیم محاسبه می‌گردد. برای محاسبه درجه رده‌بندی بین دو مفهوم از یک روش که به طور موفق در فرآیند آنالیز تصاویر مورد استفاده قرار گرفته است بهره گرفته می‌شود. در این دیدگاه دو مفهوم مشابه‌اند هرگاه تشابه ساختاری آنها زیاد بوده و مقدار این تشابه نزدیک به یک باشد. حال با استفاده از میزان تشابه ساختاری مطرح شده در مرحله قبل درجه رده‌بندی بین مفاهیم محاسبه می‌گردد. فرض می‌کنیم $Spec(c_x,c_y)$ نمایانگر میزان خاص بودن مفهوم c_x نسبت به مفهوم c_y باشد. به منظور محاسبه $Spec(c_x,c_y)$ بر اساس تشابه ساختاری، ابتدا یک مفهوم مشترک c_g را که از اشتراک دو مفهوم مورد نظر حاصل می‌گردد محاسبه می‌کنیم. این اشتراک بر اساس میزان تشابه واژه‌های موجود در بردارهای متنی دو مفهوم محاسبه می‌گردد. در نهایت، میزان رابطه رده‌بندی فازی بین این دو مفهوم به شکل زیر محاسبه می‌گردد.

$$c_g = c_x \cap c_y$$
$$\mu_{R_{CC}}(c_x,c_y) \approx Spec(c_x,c_y) = \begin{cases} 0 & if\ SSIM(c_y,c_g) > SSIM(c_x,c_g) \\ \dfrac{SSIM(c_x,c_g) - SSIM(c_y,c_g)}{SSIM(c_x,c_g)} & otherwise \end{cases} \quad (۲)$$

56 Normalized Google Distance
57 Jaccard
58 Cconditional probability
59 Kullback-Leibler divergence
60 Expected Cross Entropy
61 Balanced Mutual Information

در رابطه ۲، $SSIM\ (c_x,c_y)$ بیانگر تشابه ساختاری دو مفهوم c_x و c_y بوده که مقدار این تشابه از حاصلضرب سه پارامتر تشابه بر اساس وابستگی معنایی[۶۲] ، تشابه بر اساس واریانس معنایی[۶۳] و تشابه بر اساس اجزاء ساختاری[۶۴] محاسبه می‌گردد[۹].

پس از محاسبه روابط رده‌بندی فازی بین کلیه مفاهیم می‌بایست صرفا روابط رده‌بندی که مقدار آنها از یک حد آستانه λ بیشتر هستند به مجموعه روابط افزوده و سایر روابط را هرس نمود. همچنین می‌بایست بین دو مفهوم روابط زیر به طور همزمان برقرار باشد تا آن‌را به عنوان یک رابطه به تاکسونومی بیفزاییم:

$$Spec(c_x,c_y) > Spec(c_y, c_x)\ ,\ Spec\ (c_x,c_y) > \lambda \qquad (۳)$$

بهترین مقدار برای λ بر اساس آزمون‌های تجربی مختلف تعیین می‌گردد. اگر دو رابطه زیر بین هر دو مفهومی برقرار باشد یک رابطه تساوی بین دو مفهوم استخراج می‌گردد:

$$Spec(c_x,c_y) = Spec(c_y, c_x)\ ,\ Spec\ (c_x,c_y) > \lambda \qquad (۴)$$

همچنین در طی یک مرحله دیگر، روابط تاکسونومی تکراری حذف می‌گردند.

$$\mu_{c\times c}\,(c_1,c_2) \le \min\,(\{\mu_{c\times c}\,(c_1,c_i)\ ,\ldots,\ \mu_{c\times c}\,(c_i,c_2)\ \}) \qquad (۵)$$

اگر رابطه بالا برقرار باشد و همچنین c_1، c_i، ...، c_2 یک مسیر از c_1 تا c_2 را تشکیل دهند آنگاه رابطه $R(c_1,c_2)$ حذف می‌گردد، زیرا از سایر روابط قابل استنباط می‌باشد.

ارزیابی

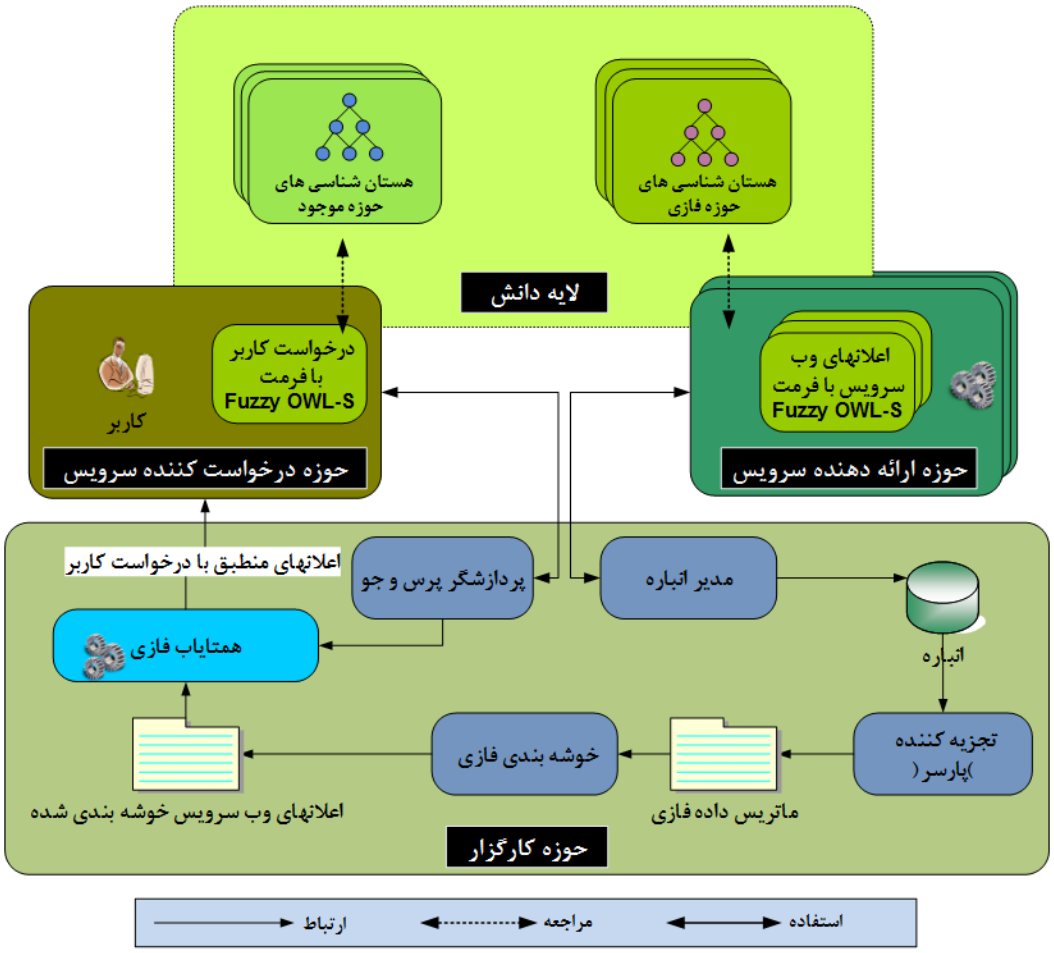
برای ارزیابی کارایی هستان‌شناسی‌ها روش‌های متفاوتی ارائه شده است[۱۰]. یکی از کاربردی‌ترین روش‌ها، بکارگیری هستان‌شناسی مربوطه در یک شاخه خاص نظیر پردازش زبان طبیعی و یا جمع‌آوری مستندات می‌باشد. ما در این مقاله از معیار کاربرد هستان‌شناسی حوزه فازی در فرآیند توصیف و همتایابی وب‌سرویس‌های معنایی بهره برده‌ایم. بدین منظور، در ابتدا با انتخاب یک همتایاب مناسب، نتایج حاصل از همتایابی با استفاده از هستان‌شناسی فازی و غیرفازی را مقایسه می‌نماییم. هدف ما در این بخش ارزیابی کارایی فرآیند همتایابی وب‌سرویس‌های معنایی با توجه به توصیف آنها توسط هستان‌شناسی‌های حوزه فازی می‌باشد. سپس در ادامه به مقایسه نتایج حاصل از همتایابی با استفاده از هستان‌شناسی‌های فازی تولید شده توسط دیدگاه لائو و دیدگاه پیشنهادی می‌پردازیم. نتایج این ارزیابی می‌تواند بیانگر میزان موفقیت روش پیشنهادی این مقاله در ارتقاء عملکرد دیدگاه لائو باشد. به منظور بررسی و ارزیابی فرآیند توصیف و همتایابی فازی وب‌سرویس‌های معنایی نیازمند استفاده از یک معماری مناسب بودیم تا بتوانیم با استفاده از آن قابلیت‌های این نحوه توصیف را مورد ارزیابی قرار دهیم. پس از بررسی نقاط قوت و ضعف الگوریتم‌های همتایابی موجود، دیدگاه ارائه شده توسط فنزا و همکاران برای استفاده در این مقاله انتخاب گردید[۱۱].

ایده اصلی این دیدگاه براساس خوشه‌بند[۶۵] فازی وب سرویس معنایی است که در آن از مولتی‌ست فازی[۶۶] برای بازنمایی اعلان‌ها[۶۷] و درخواست‌های[۶۸] وب‌سرویس و برای خوشه‌بندی از الگوریتم C-means فازی بهره برده است. ورودی الگوریتم خوشه‌بندی فازی ماتریس داده فازی می‌باشد. در الگوریتم خوشه‌بندی معمولی هر داده به یک و فقط یک خوشه اختصاص داده می‌شود. اما در خوشه‌بندی فازی هر داده با درجه‌های مختلفی از تعلق به خوشه‌های مختلف وابسته می‌باشد. در این دیدگاه، ابتدا اعلان‌های وب‌سرویس ارائه شده توسط ارائه‌دهندگان[۶۹] مختلف توسط مولتی‌ست فازی بازنمایی گرده و سپس با استفاده از الگوریتم C-means فازی خوشه‌بندی می‌گردند. ما در ابتدا هستان‌شناسی حوزه فازی تولید شده توسط روش پیشنهادی را به لایه دانش معماری فنزا اضافه نمودیم. دیدگاه‌های متنوعی برای بازنمایی هستان‌شناسی فازی توسط زبان OWL مطرح شده است. در این پژوهش از زبان FuzzyOWL2 که توسط بابیلو و استراسیا در سال ۲۰۱۱ مطرح گردیده است بهره برده‌ایم[۱۲]. از مزایای اصلی این زبان ارائه یک پلاگین برای ایجاد و مدیریت آن در محیط نرم‌افزار *protégé* می‌باشد. شکل ۵ اجزای مختلف این معماری را نمایش می‌دهد. برای آشنایی بیشتر با دیدگاه فنزا و مباحث مقدماتی این تحقیق به منابع [۱۳]و [۱۴]می‌توان مراجعه نمود.

در این مقاله ارزیابی نتایج حاصل از همتایابی از جنبه کارایی[۷۰] مورد بررسی قرار می‌گیرد. لذا در ابتدا مجموعه داده‌های مورد استفاده و ویژگی‌های آن مورد بررسی قرار می‌گیرد و سپس معیارها و پارامترهای ارزیابی معرفی می‌شوند. در ادامه، آزمون‌های انجام شده و نتایج آن‌ها گزارش می‌شود. در پایان نیز نتایج آزمون‌ها مورد بررسی و تحلیل قرار می‌گیرند.

61 Semantic Coherence
62 Semantic Variance
64 Component Structure

65 Clustering
66 Fuzzy Multiset
67 Advertisements
68 Requests
69 Providers
70 Effectiveness



شکل ۵. چارچوب فنزا برای توصیف فازی وب‌سرویس‌های معنایی

مجموعه داده‌ها

تاکنون مجموعه داده‌های آزمایشی استانداردی برای ارزیابی همتایاب‌های معنایی معرفی نشده است. اما اکثر دیدگاه‌ها از مجموعه آزمایشی OWLS-TC [۷۱] برای ارزیابی کارایی الگوریتم‌های همتایابی بر اساس OWL-S بهره می‌برند.

همچنین مسابقه جهانی انتخاب سرویس معنایی [۷۲] نیز به منظور ارزیابی همتایاب‌های شرکت‌کننده در این مسابقه از این مجموعه آزمایشی بهره می‌برد. اما از آنجائی‌که در مورد کاربرد ما، اعلان‌های وب‌سرویس‌های معنایی [۷۳] و درخواست‌های واصله از سوی کاربران می‌توانند توسط هستان‌شناسی‌های فازی توصیف گردند می‌بایست مجموعه آموزشی OWLS-TC نیز با استفاده از یک یا چند هستان‌شناسی حوزه فازی بازنویسی گردد. با بررسی صورت گرفته، حوزه گردشگری [۷۴] به عنوان حوزه مورد نظر برای گسترش یک هستان‌شناسی فازی انتخاب گردید. حدود هشتاد درصد از مفاهیم موجود در اعلان‌های وب‌سرویس‌معنایی حوزه گردشگری توسط هستان‌شناسی Travel.owl توصیف می‌گردند. لذا با تولید یک هستان‌شناسی فازی برای این حوزه، امکان توصیف فازی درصد بالایی از اعلان‌های وب‌سرویس موجود در این حوزه فراهم می‌گردد. اعلان‌های موجود در سایر حوزه‌ها بدون تغییر و کماکان براساس هستان‌شناسی‌های محض می‌باشند. همچنین مجموعه اعلان‌های مرتبط مربوط به نمونه درخواست‌های حوزه گردشگری نیز با توجه به هستان‌شناسی فازی حوزه بازنویسی می‌گردد.

تولید هستان‌شناسی فازی حوزه گردشگری

ما در ابتدا و به منظور تولید هستان‌شناسی حوزه فازی با استفاده از روش پیشنهادی نیازمند جمع‌آوری توده متون مرتبط با حوزه گردشگری بودیم. بدین منظور، در ابتدا با مراجعه به هستان‌شناسی Travel.owl مفاهیم موجود در آن بازیابی و پس از یافتن مفاهیم چند کلمه‌ای معادل آن مدل حوزه ایجاد گردید. سپس با شروع از آدرس‌های معرفی شده به عنوان بذر و با استفاده از خزشگر تاکیدی در مدت حدود ۴ ساعت موفق به گردآوری حدود ۴۰۰۰ صفحه مرتبط به حوزه مذکور شدیم. در نهایت

71 OWL-S test collection
72 International Semantic Service Selection Contest (S3 Contest)
73 Semantic Web Services Advertisement
74 Travel

براساس میزان امتیاز صفحات بازیابی شده توسط خزشگر تاکیدی، تعداد ۵۰۰ صفحه که دارای بالاترین امتیاز بودند به منظور افزودن به توده متون نهایی انتخاب گردیدند. به منظور حذف مفاهیم غیر مرتبط با حوزه گردشگری، عملیات خزش به منظور جمع‌آوری صفحات مرتبط با حوزه‌های غذا[۷۵] ، آموزش[۷۶] و پزشکی[۷۷] صورت پذیرفت. بدین منظور نیز در هر حوزه ۱۰۰ صفحه که دارای بالاترین امتیاز بودند انتخاب شد. پس از جمع‌آوری توده متون مرتبط با حوزه گردشگری، فرآیند تولید هستان‌شناسی حوزه فازی اجرا گردید. با استفاده از توده متون بازیابی شده یک مجموعه از مفاهیم به همراه بردار متنی آنها به همراه روابط رده‌بندی فازی بین آنها بازیابی گردید.

در ادامه، هستان‌شناسی فازی مرتبط با حوزه گردشگری با استفاده از دیدگاه لائو و همکاران نیز تولید گردید. همانگونه که قبلا ذکر شد دیدگاه لائو شامل دو مرحله ابتدایی مربوط به دیدگاه پیشنهادی این مقاله نبوده و به منظور ساخت هستان‌شناسی فازی راهبرد مشخصی ندارد. لذا به منظور توده متون مورد نیاز، از یک خزشگر معمولی استفاده نمودیم.

بدین منظوربا استفاده از پیوندهای بذر بازیابی شده در مرحله قبل فرآیند خزش را آغاز نموده و تعداد ۱۰۰۰ صفحه بازیابی گردید. این صفحات را به عنوان توده متون مورد نیاز برای تولید هستان‌شناسی فازی در دیدگاه لائو مورد استفاده قرار دادیم. در نهایت هستان‌شناسی‌های فازی تولید شده توسط روش پیشنهادی و لائو توسط زبان FuzzyOWL2 و نرم‌افزار *protégé* بازنمایی گردید.

معیارهای ارزیابی کارایی

کارایی همتایاب‌های معنایی بر اساس قدرت آنها در بازیابی نتایج مرتبط با درخواست مطرح‌شده سنجیده می‌شود. ستسوس و همکاران[۷۸] یک دیدگاه کلی برای ارزیابی کارایی همتایاب‌های معنایی معرفی نموده‌اند[۱۵].

بر اساس این دیدگاه، موتور همتایاب درجه تشابه را برای هر اعلان A_i در برابر درخواست R به صورت $e(R, A_i)$ محاسبه می‌کند. حال به منظور ارزیابی کارایی همتایاب، درجه تشابهی به صورت $r(R, A_i)$ که توسط فرد خبره تعیین می‌گردد نیز می‌بایست وجود داشته باشد. بردارهای e و r بدین شکل تعریف می‌گردند:

$$r: Q \times A \rightarrow W \qquad (۱)$$

$$e: Q \times A \rightarrow W$$

که در آن Q مجموعه‌ای از درخواست‌های موجود، A مجموعه‌ای از اعلان‌های موجود و W مجموعه‌ای از مقادیری است که بیانگر درجه تشابه (r) یا درجه تطابق (e) بین یک درخواست از Q و یک اعلان از A می‌باشد. r و e می‌توانند انواع مختلفی از مقادیر نظیر دودویی ($\{0,1\}$)، اعداد صحیح ($[0,1]$) یا طبقه‌بندی ($\{exact, plugin, subsume, \ldots\}$) را کسب نمایند. بر این اساس، ارزیابی همتایاب با دقتی که بردار e توسط بردار r تخمین زده شود تعیین می‌گردد. حال اگر مجموعهٔ درجه تشابه‌های تعیین شده توسط افراد خبره به صورت دودویی موجود باشد می‌توان از معیارهای مورد استفاده در حوزه بازیابی اطلاعات به منظور ارزیابی نتایج حاصل از همتایابی استفاده نمود. معیارهای دقت[۷۹] و یادآوری[۸۰] از معیارهای استاندارد در حوزه بازیابی اطلاعات می‌باشند و زمانی‌که مجموعه‌های مرتبط به صورت دودویی بیان شده باشند قابل استفاده می‌باشند.

معیار دقت نشان‌دهنده نسبت تعداد سرویس‌های مناسبی ارائه شده توسط همتایاب به کل سرویس‌هایی است که توسط همتایاب ارائه شده‌اند. در مقابل معیار یادآوری نشان‌دهنده تعداد سرویس‌های مناسب ارائه شده توسط همتایاب به کل سرویس‌های مناسب موجود در مخزن می‌باشد. لذا می‌توان میزان دقت این همتایاب را بر اساس سطوح مختلف یادآوری یکبار با در نظر گرفتن روابط رده‌بندی فازی و یکبار بدون در نظر گرفتن درجات فازی بین مفاهیم مورد مقایسه قرار داد.

همچنین به منظور ارزیابی دیدگاه پیشنهادی به مقایسه تعداد اشتباه‌های مثبت[۸۱] و اشتباه‌های منفی[۸۲] بر اساس درجات بازیابی مختلف و تعداد خوشه مشخص در دو همتایاب می‌پردازیم. اشتباه‌های مثبت زمانی رخ می‌دهد که همتایاب معنایی تعدادی از اعلان‌های نامرتبط با نمونه درخواست را به عنوان نتیجه همتایابی ارائه نماید. در مقابل، اشتباه‌های منفی زمانی رخ می‌دهند که همتایاب معنایی نتواند تعدادی از اعلان‌های مرتبط با نمونه درخواست را به عنوان نتیجه همتایابی به کاربر ارائه نماید. هر چه الگوریتم انعطاف‌پذیرتر باشد تعداد اشتباه‌های مثبت افزایش یافته و تعداد اشتباهای منفی کاهش می‌یابد. بدیهی است که نتیجه مطلوب، نتیجه‌ای است که در آن تعداد اشتباه‌های مثبت و منفی به حداقل رسیده و نمودارهای مرتبط با آنها به یکدیگر نزدیک‌تر باشند.

پارامترهای ارزیابی

اولین پارامتری که حائز اهمیت است تعداد خوشه‌ها در فرآیند خوشه‌بندی فازی می‌باشد. انتخاب تعداد خوشه‌های مناسب تاثیر

75 Food
76 Education
77 Medical
78 Tsetsos et al.

79 Precision
80 Recall
81 False positives
82 False negatives

مستقیم بر فرآیند همتایابی دارد. پارامتر مهم دیگر، میزان درجه صحت فرآیند بازیابی سرویس[۸۳] می‌باشد و توسط دو پارامتر α و h تعیین می‌گردد. اعلان‌های وب‌سرویسی که در یک همسایگی از پیش تعریف‌شده با مولتی‌ست متناظر با درخواست کاربر می‌باشند به عنوان اعلان‌های وب‌سرویس همتای آن انتخاب می‌گردند. این همسایگی با عنوان منطقه تاثیر[۸۴] درخواست شناخت شده و توسط دو پارامتر α و h تعیین می‌گردد.

آزمون‌های انجام شده

در اولین آزمون به مقایسه و ارزیابی نتایج حاصل از همتایاب فنزا و همتایاب پیشنهادی در این مقاله می‌پردازیم. توجه به این نکته ضروری است که در ادامه منظور از همتایاب پیشنهادی همان همتایاب فنزا است که از افزودن هستان‌شناسی‌های فازی تولید شده توسط روش پیشنهادی به لایه دانش همتایاب فنزا ایجاد شده است.

در دومین آزمون به مقایسه و ارزیابی نتایج حاصل از همتایاب پیشنهادی و همتایاب لائو می‌پردازیم. همتایاب لائو نیز به طور مشابه از افزودن هستان‌شناسی‌های فازی تولید شده توسط دیدگاه لائو به لایه دانش همتایاب فنزا تولید می‌گردد. در ابتدا قصد داریم که پاسخ سوال زیر را بیابیم:

- آیا توصیف فازی ویژگی‌های کارکردی وب‌سرویس‌های معنایی، فرآیند همتایابی را ارتقاء می‌بخشد؟

برای پاسخ به این سوال نشان خواهیم داد که فرآیند همتایابی توسط همتایاب پیشنهادی با استفاده از هستان‌شناسی حوزه فازی باعث ارتقای نتایج خواهد شد.

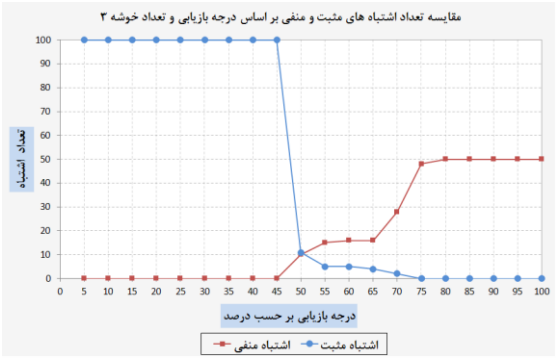
حال بر اساس هدف ذکر شده بالا فرضیه زیر را مطرح می‌نماییم:

- توصیف معنایی ویژگی‌های کارکردی وب‌سرویس‌ها توسط هستان‌شناسی حوزه فازی می‌تواند باعث بهبود فرآیند همتایابی وب‌سرویس‌های معنایی گردد.

ما به منظور آزمون این فرضیه، نتایج کسب شده از همتایاب فنزا و همتایاب پیشنهادی را در یک سناریوی مشابه درخواست وب‌سرویس مورد مقایسه قرار می‌دهیم. سپس این نتایج با مجموعه‌های مرتبط تعریف شده در OWLS-TC مقایسه می‌گردد تا مشخص گردد که نتایج کدام همتایاب با نتایج موجود در این مجموعه‌ها تشابه بیشتری دارد. در ابتدا یک نمونه درخواست را به همتایاب فنزا و نمونه درخواست معادل آن را که توسط هستان شناسی فازی بازنویسی شده را به همتایاب پیشنهادی عرضه می‌نماییم. توجه به این نکته حائز اهمیت است که به منظور تعیین مجموعه وب‌سرویس‌های مرتبط با نمونه درخواست از مجموعه مرتبط معرفی شده در OWLS-TC بهره برده‌ایم. در این مرحله تمامی اعلان‌های وب‌سرویسی که به شکل دودویی مرتبط با نمونه درخواست ارائه شده شناخته شده‌اند به مجموعه سرویس‌های مرتبط افزوده می‌گردند. بدین منظور برای تمام نمونه درخواست‌ها، فرآیند همتایابی را توسط همتایاب پیشنهادی و فنزا و بر اساس تعداد خوشه‌های مختلف تکرار می‌نماییم.

شکل‌های ۶ و۷ میزان اشتباه‌های مثبت و منفی را بر اساس یک نمونه درخواست از حوزه گردشگری[۸۵] و با بکارگیری همتایاب‌های پیشنهادی و فنزا نمایش می‌دهند. در این آزمون تعداد خوشه‌ها برابر ۳ می‌باشد. جدول‌های ۲ تا ۴ مشخصات نمونه درخواست مورد نظر را به همراه بازنمایی فازی آن، یکبار بدون در نظر گرفتن روابط رده‌بندی فازی و بار دیگر با در نظر گرفتن این روابط توسط هستان‌شناسی‌های حوزه فازی تولید شده توسط روش پیشنهادی و روش لائو نمایش می‌دهند. با بررسی آزمون‌های فوق برای سایر نمونه درخواست‌ها به این نتیجه می‌رسیم که میانگین میزان اشتباه‌های منفی در همتایاب پیشنهادی در مقایسه با همتایاب فنزا کمتر می‌باشد.

سپس نمودارهای دقت و یادآوری بر اساس درجه‌های بازیابی مختلف مورد بررسی قرار می‌گیرد. شکل‌های ۸ و۹ نمودارهای دقت و یادآوری همتایاب‌های پیشنهادی و فنزا را برای درجه‌های بازیابی ۵۰ و ۵۵ نمایش می‌دهند.



شکل ۶. بررسی تعداد اشتباه‌های مثبت و منفی حاصل از همتایابی نمونه درخواست توسط همتایاب پیشنهادی (تعداد ۳ خوشه)

83 Precision degree of service retrieval
84 Action Scope

85 FCityCountry_Hotel_Service.owls

جدول ۲. بازنمایی نمونه درخواست توسط مولتی‌ست فازی بدون در نظر گرفتن روابط رده‌بندی فازی

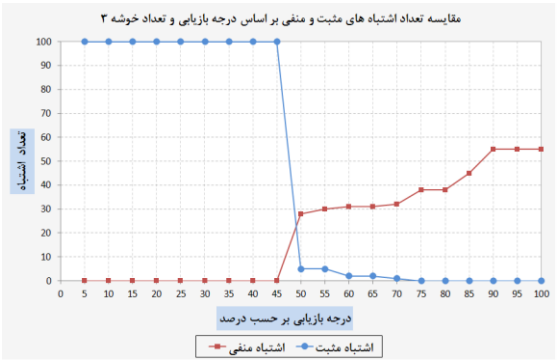
time-measure		duration		urbanarea		quality		accommodation		activity		Title		preparedfood	
0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
retailstore		financing		government		geopolitical-entity		destination		quantity		food		generic-agent	
0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
missile		giving		comedyfilm		ruralarea		film		price		legal-agent		videomedia	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

جدول ۳. بازنمایی نمونه درخواست توسط مولتی‌ست فازی با در نظر گرفتن هستان‌شناسی حوزه فازی پیشنهادی

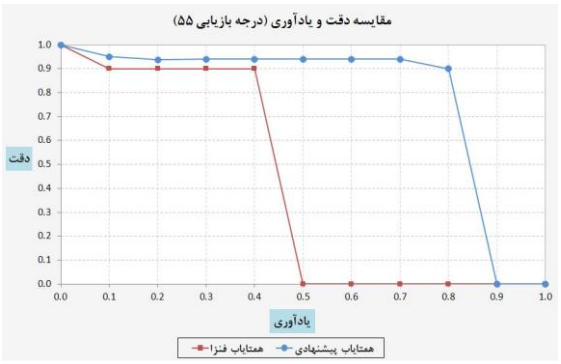
time-measure		duration		urbanarea		quality		accommodation		activity		Title		preparedfood	
0.0	0.0	0.0	0.0	0.58	0.0	0.0	0.0	0.95	0.0	0.0	0.0	0.0	0.0	0.0	0.0
retailstore		financing		government		geopolitical-entity		destination		quantity		food		generic-agent	
0.0	0.0	0.0	0.0	0.0	0.0	0.57	0.0	0.55	0.0	0.0	0.0	0.0	0.0	0.0	0.0
missile		giving		comedyfilm		ruralarea		film		price		legal-agent		videomedia	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

جدول ۴. بازنمایی نمونه درخواست توسط مولتی‌ست فازی با در نظر گرفتن هستان‌شناسی حوزه فازی لائو

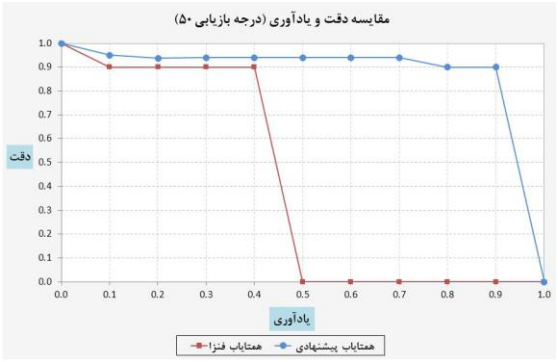
time-measure		duration		urbanarea		quality		accommodation		activity		Title		preparedfood	
0.0	0.0	0.0	0.0	0.48	0.0	0.0	0.0	0.73	0.0	0.0	0.0	0.0	0.0	0.0	0.0
retailstore		financing		government		geopolitical-entity		destination		quantity		food		generic-agent	
0.0	0.0	0.0	0.0	0.0	0.0	0.43	0.0	0.42	0.0	0.0	0.0	0.0	0.0	0.0	0.0
missile		giving		comedyfilm		ruralarea		film		price		legal-agent		videomedia	
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0



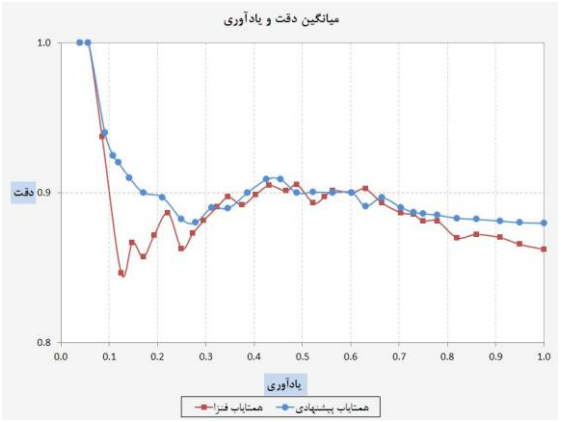
شکل ۷. بررسی تعداد اشتباه‌های مثبت و منفی حاصل از همتایابی نمونه درخواست توسط همتایاب فنزا (تعداد ۳ خوشه)



شکل۹. مقایسه دقت و یادآوری همتایاب پیشنهادی و فنزا (درجه بازیابی ۵۵)



شکل۸. مقایسه دقت و یادآوری همتایاب پیشنهادی و فنزا (درجه بازیابی ۵۰)



شکل ۱۰. نمودار میانگین- میکرو

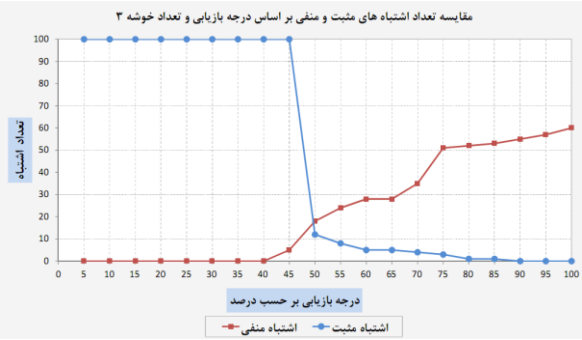
به منظور ارزیابی نهایی، به بررسی معیار میانگین- میکرو برای برای تمام نمونه درخواست‌های حوزه گردشگری با در نظر گرفتن دو پارامتر تعداد خوشه و درجه بازیابی می‌پردازیم. شکل ۱۰ یک نمونه از منحنی مذکور را برای درجه بازیابی ۵۵ و تعداد خوشه ۳ نمایش می‌دهد. همانگونه که مشاهده می‌گردد منحنی مربوط به همتایاب پیشنهادی نتیجه بهتری را نسبت به همتایاب فنزا ارائه می‌نماید. نتایج مشابهی برای سایر مقادیر مرتبط با درجه بازیابی و تعداد خوشه مشاهده می‌شود.

در ادامه قصد داریم که پاسخ سوال زیر را بیابیم:

- آیا همتایاب پیشنهادی نسبت به همتایاب لائو، کیفیت فرآیند همتایابی را ارتقاء می‌بخشد؟ برای پاسخ به این سوال نشان خواهیم داد که فرآیند همتایابی توسط همتایاب پیشنهادی در مقایسه با همتایاب لائو باعث ارتقای نتایج خواهد شد. حال بر اساس هدف ذکر شده بالا فرضیه زیر را مطرح می‌نماییم:

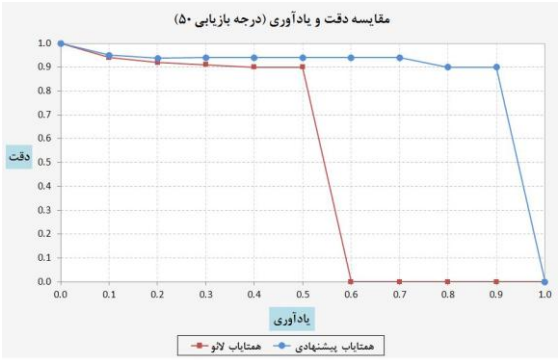
- همتایاب پیشنهادی در مقایسه با همتایاب لائو فرآیند همتایابی را ارتقاء می‌بخشد.

مشابه بخش اول و به منظور آزمون این فرضیه، نتایج کسب شده از همتایاب لائو و همتایاب پیشنهادی را در یک سناریوی درخواست وب‌سرویس مورد مقایسه قرار می‌دهیم. شکل ۱۱ میزان اشتباه‌های مثبت و منفی را بر اساس نمونه درخواست ذکر شده و با بکارگیری همتایاب لائو نمایش می‌دهد.

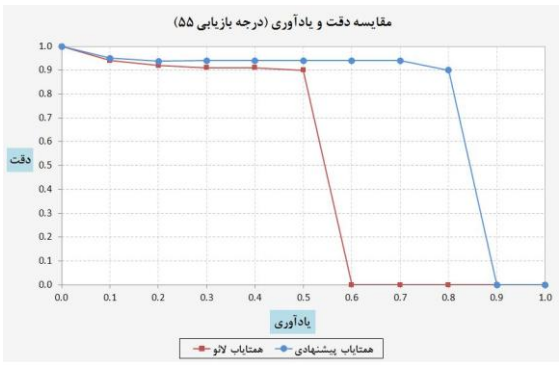


شکل ۱۱. بررسی تعداد اشتباه‌های مثبت و منفی حاصل از همتایابی نمونه درخواست توسط همتایاب لائو (تعداد ۳ خوشه)

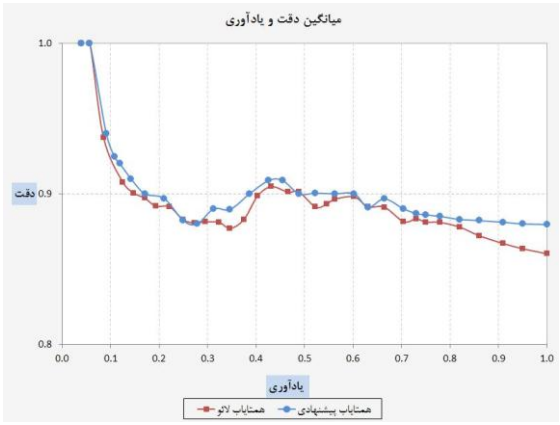
همانگونه که مشاهده می‌شود میانگین میزان اشتباه‌های منفی در همتایاب پیشنهادی در مقایسه با همتایاب لائو کمتر می‌باشد. شکل‌های ۱۲ و ۱۳ نیز به طور مشابه نمودارهای دقت و یادآوری دو همتایاب را برای دو درجه بازیابی ۵۰ و ۵۵ نمایش می‌دهند.



شکل۱۲. مقایسه دقت و یادآوری همتایاب پیشنهادی و لائو (درجه بازیابی ۵۰)



شکل۱۳. مقایسه دقت و یادآوری همتایاب پیشنهادی و لائو (درجه بازیابی ۵۵)



شکل ۱۴. نمودار میانگین- میکرو

همانگونه که مشاهده می‌شود در این بخش نیز نتایج حاصل از همتایاب پیشنهادی نسبت به همتایاب لائو بهبود یافته است. شکل ۱۴ نیز نمودار میانگین میکرو مربوط به این دو همتایاب را نمایش می‌دهد. این نمودار مشابه آزمون قبل با مقدار ۵۵ برای درجه بازیابی و تعداد خوشه ۳ صورت گرفته است. در این نمودار نیز برتری نتایج همتایاب پیشنهادی محسوس می‌باشد.

نتیجه‌گیری

در این پژوهش در ابتدا یک مدل حوزه بر اساس هستان‌شناسی‌های حوزه موجود ایجاد گردید. سپس با استفاده از یک خزشگر تاکیدی که از این مدل حوزه بهره می‌برد به جمع‌آوری توده متون مرتبط با حوزه پرداختیم. سپس با استفاده از

چارچوب پیشنهادی لائو و همکاران و استفاده از توده متون جمع‌آوری شده مرتبط با حوزه، فرآیند تولید هستان‌شناسی حوزه فازی صورت پذیرفت. به منظور ارزیابی کیفیت هستان‌شناسی حوزه تولید شده توسط این روش، به بررسی نحوه تاثیر آن در فرآیند توصیف ویژگی‌های کارکردی وب‌سرویس‌های معنایی پرداختیم. نتایج حاصل از ارزیابی بیانگر این واقعیت بود که فرآیند همتایابی وب‌سرویس‌های معنایی در حالتی که اعلان‌های وب‌سرویس توسط هستان‌شناسی حوزه فازی توصیف گردیده‌اند نتایج بهتری ارائه می‌نماید. این نتایج با هستان‌شناسی فازی حوزه تولید شده توسط روش لائو و همکاران مقایسه گردید و بهبود محسوسی مشاهده شد. در حقیقت ارتقای کیفیت تولید هستان‌شناسی حوزه فازی توسط روش پیشنهادی تاثیر مستقیم و مثبتی بر نتایج حاصل از همتایابی وب‌سرویس‌های معنایی گذاشته است. مهمترین جنبه تاثیرگذاری را می‌توان از این منظر مد نظر قرار داد که با آشکار شدن روابط رده‌بندی فازی جدید بین مفاهیم موجود در هستان‌شناسی حوزه فازی امکان توصیف اعلان‌های وب‌سرویس جدید فراهم می‌گردد.

در واقع، بکارگیری هستان‌شناسی‌های حوزه فازی با کیفیت این امکان را فراهم می‌سازد تا ارائه‌دهندگان وب‌سرویس بتوانند ویژگی‌های کارکردی وب‌سرویس‌های مدنظرشان را با توجه به روابط رده‌بندی فازی بین مفاهیم با قابلیت انعطاف بیشتری اعلان نمایند. از طرف دیگر این امکان برای درخواست‌کنندگان وب‌سرویس نیز فراهم می‌گردد که به شکلی منعطف بتوانند درخواست وب‌سرویس خود را توصیف نمایند. نتیجه منطقی این دو پیشرفت، ارتقای فرآیند همتایابی وب‌سرویس‌های معنایی می‌باشد.

مرجع‌ها

[1]Cimiano P, Hotho A, Staab S. “Learning concept hierarchies from text corpora using formal concept analysis”. Journal of Artificial Intelligence Research, 2005, Volume 24, pp. 305–339.

[2]Lee Ch, Jian Z, and Huang L. “A Fuzzy Ontology and Its Application to News Summarization”. In: IEEE Transactions on Systems, Man and Cybernetics, 2005, Part B 35.5. pp. 859–880.

[3]Tho Q, Hui S, Fong A, and Hoang Cao T. “Automatic fuzzy ontology generation for semantic web”. In: IEEE Transactions on Knowledge and Data Engineering, 2006, volume 18(6), pp. 842–856.

[4]Lau R.Y.K., Song D, Li Y, Cheung C.H., Hao J.X. “Towards A Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning”. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(6), pp. 800-813.

[5]Yaguinuma, C.A., et al., “A Fuzzy Ontology-Based Semantic Data Integration System,” Journal of Information & Knowledge Management. 2011, 10(3): 285-299.

[6]Na Xue, Suling Jia, Jinxing Hao and Qiang Wang, “Scientific Ontology Construction Based on Interval Valued Fuzzy Theory under Web 2.0”. Journal of Software, AUGUST 2013, VOL. 8, NO. 8.

[7]Greenwood M, Nenadic G, “Lexical profiling of Existing Web Directories to Support Fine-Grained Topic-Focused Web Crawling”. In: Proc. of the Corpus Profiling for IR and NLP Workshop, 2008, London.

[8]Koller D and Sahami M. “Hierarchically classifying documents using very few words”. In: Douglas H. Fisher, editor, Proceedings of ICML-97, 14th International Conference on Machine Learning, Nashville, Tennessee. Morgan Kaufmann Publishers, SanFrancisco, California, 1997, pp. 170–178.

[9]Wang X, Vitvar T, Kerrigan M, and Toma I. “A qos-aware selection model for semantic web services”. In: Asit Dan and Winfried Lamersdorf, editors, ICSOC, 2006, volume 4294 of Lecture Notes in Computer Science, Springer, pp. 390-401.

[10]Hartmann J, Spyns P, Giboin A. “Methods for ontology evaluation”, Research Project#: IST-2004-507482IST, Commission of the European Communities, 2005.

[11]Fenza G, Loia V, Senatore S. “A hybrid approach to semantic web services matchmaking”. International Journal of Approximate Reasoning, 2008, Volume 48, pp. 808–828.

[12]Bobillo F, Straccia U. “Fuzzy Ontology Representation using OWL2”. International Journal of Approximate Reasoning, 2011, 52(7), pp.1073-1094.

[13]Bobillo F, Sharifi E, Ebadzadeh M.M, ASkari. Moghadam R. “On Fuzzy Semantic Web Services”. In: Proceedings of the 19th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2010), Barcelona (Spain), 2010, pp. 2418-2423.

[14]Sharifi E, ASakri. Moghadam R, Bobillo F, Ebadzadeh M.M. “A Fuzzy Framework for Semantic Web Service Description, Matchmaking, Ranking and Selection”. In: Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2011), China, 2011, vol1,pp.621-625.

[15]Tsetsos V, Anagnostopoulos Ch, and Hadjiefthymiades S. “On the evaluation of semantic web service matchmaking systems”. In: Proceedings of the European Conference on Web Services, Washington, DC, USA. IEEE Computer Society,2006, pp. 255–264.