Designing Semiautomatic System in Ontology Structure by Co-occurrence Word Analysis and C-value Method (Case Study: The Field of Scientometrics of Iran)

Hamid Ahmadi
PhD Condidate in Knowledge and Information Science;
Chamran University of Ahwaz;
Corresponding Auther hamid_ahmadi@razi.ac.ir

Farideh Osareh
PhD in Knowledge and Information Science;
Professor; Shahid Chamran University;
Library and Information Science Department osareh.f@gmail.com

Molouk Sadat Hosseini Beheshti
PhD in Linguistic; Assistant Professor; Iranian Research Institute for Information Science and Technology (IranDoc); Tehran, Iran;
beheshti@irandoc.ac.ir

Gholamreza Heidari
PhD in Knowledge and Information Science; Assistant Professor;
Chamran University; Library and Information Science Department;
ghrhaidari@gmail.com

Iranian Journal of Information Processing and Management

Iranian Research Institute for Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 33 | No. 1 | pp. 1-30
Autumn 2017

Received: 10, Jan. 2016 **Accepted: 03, Aug. 2016**

Abstract: Ontologies are the means of expression of formal concepts and relations in the specific regions. It have recently tried to design the learning methods and automation process of constructing of Ontology. Whereas Ontology containing concepts and the relations, extracting of concepts, the semantic relations among concept are important.

Constructing of various Ontology for various regions and different applications are expensive and time-consuming processe. Automation of this prose is important step. The lack of knowledge such as treasures or database of knowledge domains, will make it difficult to gain knowledge of ontology in different domains.

in present study a semi-automatic method was suggested in order to gain knowledge in the Iran scinetometrics domains. this method can extract information of this domain and processing exiting knowledge for constricting Ontology in a proses., therefore, at first, the documents of Domain were collected and then automated indexing by text mining

methods. text maing then, it was in the next step by using C-value method, main concepts were extract., than by using k-means clustering, relater documents were clustered, and based on TF-IDF method, main concepts were estrect for each cluster. Finally by using co-word analyses the hierarchy of concepts were extract and related ontology were construct. The results showed that this method in compare with other has had a lot accuracy in ontology building learning.

Keywords: Ontology, Scientometrics of Iran, Co-occurrence Word Analysis, C-value Method, Clustering Documents

طراحی سامانهٔ نیمه‌خودکار ساخت هستی‌شناسی به‌کمک تحلیل هم‌رخدادی واژگان و روش C-value

(مطالعهٔ موردی: حوزهٔ علم‌سنجی ایران)[1]

حمید احمدی
دانشجوی دکتری علم اطلاعات و دانش‌شناسی؛ دانشگاه شهید چمران اهواز؛
پدیدآور رابط hamid_ahmadi@razi.ac.ir

فریده عصاره
دکتری علم اطلاعات و دانش‌شناسی؛ استاد؛ علم اطلاعات و دانش‌شناسی؛ دانشگاه شهید چمران اهواز osareh.f@gmail.com

ملوک‌السادات حسینی بهشتی
دکتری زبان‌شناسی همگانی؛ استادیار؛ پژوهشکدهٔ مدیریت دانش؛ پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک) beheshti@irandoc.ac.ir

غلامرضا حیدری
دکتری علم اطلاعات و دانش‌شناسی؛ استادیار؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه شهید چمران اهواز ghrhaidari@gmail.com

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹٤/۱۰/۲۰ پذیرش: ۱۳۹۵/۰۵/۱۲ مقاله برای اصلاح به مدت ۱۴ روز نزد پدیدآوران بوده است.

چکیده: هستی‌شناسی‌ها ابزار بیان رسمی مفاهیم و روابط موجود در قلمرویی خاص هستند. در سال‌های اخیر تلاش‌های زیادی برای طراحی روش‌های یادگیری و خودکارسازی فرایند ساخت هستی‌شناسی انجام گرفته است. از آنجا که هستی‌شناسی را مجموعهٔ مفاهیم و روابط آن می‌دانیم، استخراج مفاهیم و روابط معنایی میان این مفاهیم از اهمیت بسیاری برخوردار است. ساخت انواع هستی‌شناسی برای انواع قلمروها و کاربردهای گوناگون، فرایندی پرهزینه و زمان‌بَر بوده و خودکارسازی این فرایند گام مهمی در رفع آن است. فقدان دانش پایه مانند اصطلاح‌نامه‌ها یا پایگاه‌های دانش حوزه‌ها، اکتساب دانش برای ساخت هستی‌شناسی آن حوزه‌ها را مشکل خواهد کرد. در این پژوهش روشی نیمه‌خودکار برای اکتساب دانش در حوزهٔ علم‌سنجی ایران ارائه شده که قادر است اطلاعات

فصلنامه | علمی پژوهشی
پژوهشگاه علوم و فناوری اطلاعات ایران
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۳ | شمارهٔ ۱ | ص‌ص ۱۸۷-۲۱۸
پاییز ۱۳۹۶

۱. برگرفته از پایان‌نامهٔ دکتری؛ دانشگاه شهید چمران اهواز

این حوزه را استخراج کرده و در فرایندی، دانش موجود را برای ساخت هستی‌شناسیِ آن پردازش کند. بدین‌منظور، ابتدا اسناد مرتبط با حوزهٔ مورد نظر گردآوری شده و به روشِ متن‌کاوی، نمایه‌سازی خودکار گردید. سپس، در مرحلهٔ بعدی با استفاده از روش C-value مفاهیم اصلی استخراج شد. آن‌گاه اسناد مربوطه با استفاده از روش خوشه‌بندی k-means، خوشه‌بندی شدند و برای هر خوشه با محاسبهٔ وزن مفاهیم، بر اساس روش TF-IDF مفاهیم کلیدی مناسب استخراج گردید. در پایان، با استفاده از روش تحلیل هم‌رخدادی واژگان، سلسله‌مراتب مفاهیم حوزه استخراج شده و هستی‌شناسی مربوطه ساخته شد. نتایج به‌دست آمده نشان می‌دهد که این روش در مقایسه با روش‌های مشابه دقت بسیاری در یادگیری ساخت هستی‌شناسی داشته است.

کلیدواژه‌ها: هستی‌شناسی، حوزهٔ علم‌سنجی ایران، هم‌رخدادی واژگان، روش C-value، خوشه‌بندی اسناد دلفی

۱. مقدمه و بیان مسئله

در سال‌های اخیر، گسترش اطلاعات مسائلی را در ارتباط با سازگاری نظام‌های سنتی در مدیریت علم و دانش به‌وجود آورده، و لزوم توجه به محیط‌های نوینِ پردازش اطلاعات در بازنمون و مدیریت هوشمند، آن را اجتناب‌ناپذیر کرده است. در این راستا، روش‌های مهندسی دانش، نظیر هستی‌شناسی‌ها[1] و شبکه‌های مفهومی[2] روزبه‌روز در حال گسترش بوده و توانایی منحصربه‌فردی در استخراج، تحلیل و مدل‌سازی به‌عنوان پایگاه‌های بزرگ دانش مفهومی و ابزارهای اصلی حفظ و تبادل دانش میان سیستم‌های مختلف داشته و در سیستم‌های هوشمند و مبتنی بر دانش نقشی اساسی ایفا می‌کنند.

در شرایط کنونی، خودکارسازی پردازش اطلاعات به‌واسطهٔ سامانه‌های اطلاعاتی در بازیابی و طراحی ساخت پایگاه‌های دانش و به‌طور کلی، در تحلیل حوزه‌های علمی از اهمیت بالایی برخوردارند. در این میان ابزارهای مهندسی دانش با ایجاد شبکه‌ای از مفاهیم و روابط میان آن‌ها قادرند دانش حوزه‌های علمی را شناسایی و توصیف کرده و به‌عنوان ابزاری معناشناختی برای ایجاد پایگاه دانش حوزه‌ها مؤثر و مفید باشند و در نهایت، می‌توان آن‌ها را در مدیریت علم و دانش حوزه‌ها به‌کار برد.

یکی از ابزارهای معناشناختی و روابط معنایی که قادر است مفاهیم و روابط میان آن‌ها را به‌صورت دقیق‌تر نمایش دهد، هستی‌شناسی‌ها هستند. هستی‌شناسی به‌عنوان

1. ontology
2. conceptual network

یک ابزار در مهندسی دانش مطرح است و برای بیان مفاهیم و روابط معنایی در ایجاد رابطه‌ها، نمونه‌ها، نمایش تصویری مفاهیم و ساختار مفهومی کارکرد اساسی دارد.

هستی‌شناسی، یک علم در فلسفه است که به مطالعهٔ آنچه موجود است و آنچه باید فرض شود که وجود دارد، به‌منظور دستیابی به یک توصیف متقاعدکننده از واقعیت می‌پردازد. فلسفه، اولین حوزهٔ دانش است که در آن، مفهوم هستی‌شناسی به‌کار گرفته شده است. کاربرد این مفهوم، ریشه در نظرات و آراء ارسطو دارد که اکنون به‌عنوان حوزهٔ متافیزیک شناخته شده و در ابتدا به مطالعهٔ ماهیت وجود و در مرحلهٔ بعدی به مختصات حقیقت وجود اشیاء و پدیده‌ها می‌پردازد. در حال حاضر، هستی‌شناسی گرایشی از فلسفه است که هدفش تعریف وجود در نظام هستی است که با انواع و ساختارهای اشیاء، پدیده‌ها، مختصات رخدادها، فرایندها و روابط میان اجزا و رده‌بندی جزء و کل سروکار دارد. از دههٔ ۹۰ میلادی، هستی‌شناسی‌ها در حوزهٔ دانش هوش مصنوعی اهمیت بیشتری یافتند، زیرا رویکرد خاصی به مهندسی دانش، پردازش زبان و بازنمایی دانش داشت (حسینی بهشتی ۱۳۹۲). به‌دلیل نیاز به ابزاری برای طبقه‌بندی مفاهیم، متخصصان هوش مصنوعی در دههٔ ۹۰ میلادی با وام‌گیری مفهوم هستی‌شناسی از رشتهٔ فلسفه به توصیف مفاهیم پرداختند. از این طریق، مفهوم هستی‌شناسی به حوزه‌های دیگر مانند علم اطلاعات و دانش‌شناسی، وب معنایی، سیبرنتیک، زبان‌شناسی، پردازش زبان طبیعی و ... نیز راه یافت.

«گروبر» هستی‌شناسی را مفهومی می‌داند که به ایجاد فهم مشترک از حوزه‌ها اشاره دارد و شامل مجموعه‌ای از مفاهیم، روابط، کارکردها، اصول بدیهی و نمونه‌هاست (Gruber 1993). امروزه، هستی‌شناسی‌ها به‌عنوان ابزارهای معنایی در حوزه‌های مختلفی همچون سامانه‌های بازیابی اطلاعات معنایی، سامانه‌های مدیریت دانش معنایی، فهرست‌های معنایی و غیره کاربرد دارند. ویژگی‌های بارز هستی‌شناسی‌ها همچون قابلیت استنتاج، ایجاد ارتباط و میانکنش‌پذیری بین سامانه‌های اطلاعاتی، حمایت از پردازش زبان طبیعی، فهم پرسش جست‌وجو و غیره توجه پژوهشگران را به استفاده از این ابزار در ایجاد کتابخانه‌های دیجیتال معنایی جلب نموده است.

هستی‌شناسی‌ها ابزار بیان رسمی مفاهیم و روابط موجود در یک قلمرو خاص هستند. در حوزهٔ علم اطلاعات و دانش‌شناسی، هستی‌شناسی به‌عنوان ابزار معناشناسی به‌کارگرفته شده است که قادر است مفاهیم و روابط میان آن‌ها را به‌صورت دقیق‌تر نمایش دهد.

«نشاطی» (۱۳۸۶) دلایل ساخت هستی‌شناسی را به‌طور خلاصه در موارد زیر می‌داند:

الف- اشتراک اطلاعات میان انسان و کارگزارهای خودکار: هدف اصلی از ساخت هستی‌شناسی، مشترک‌سازی اطلاعات میان انسان و کارگزارهای خودکار است؛

ب- استفادهٔ مجدد از دانشِ حوزه‌های مختلف: استفادهٔ مجدد از دانشِ حوزه‌های مختلف، از جنبه‌های گوناگون قابل بررسی است و؛

ج- بیان صریح فرضیات حوزه‌ها: هنگامی که دانش حوزه‌های مختلف به‌صورت صریح بیان شده باشد، تغییرات آن ساده‌تر است و مستلزم تغییر در کد برنامهٔ کاربردی نیست.

علاوه بر این، جداسازی دانش دامنه از دانش عملیاتی و وسیله‌ای برای آنالیز دانش دامنه، از دلایل توسعهٔ هستی‌شناسی‌ها متصور است (حسینی بهشتی ۱۳۹۲).

در سال‌های اخیر، توسعهٔ هستی‌شناسی‌ها از یک کار آزمایشگاهی در آزمایشگاه هوش مصنوعی به یک کار با کاربردهای واقعی تبدیل شده است (Kruz 2007). پژوهش در زمینهٔ هستی‌شناسی، رشد بسیاری در حوزهٔ رایانه داشته است که در زمینه‌هایی نظیر وب معنایی[1]، موتورهای جست‌وجو[2]، تجارت الکترونیکی، پردازش زبان طبیعی، مهندسی دانش، استخراج و بازیابی اطلاعات، طراحی پایگاه‌های اطلاعاتی[3]، سامانه‌های چندکارگزاره، و کتابخانه های دیجیتال کاربرد دارد.

با عنایت به این‌که لازمهٔ مدیریت هدفمند علم و دانش، حصول شناخت و ارزیابی مستدلی از وضعیت کنونی یک حوزهٔ علمی است، در پژوهش حاضر، مدلی مفهومی تحت عنوان «سامانهٔ نیمه‌خودکار ساخت هستی‌شناسی» ارائه گردید تا برون‌دادهای علمی این حوزه به‌صورت نیمه‌خودکار استخراج و مورد پردازش قرار گیرد و از این طریق به درک صریحی از جهان آن با توصیف و تشریح مفاهیم و روابط این حوزه در ایران پرداخته و شناخت بیشتری حاصل شود. در مدل مفهومی از مدل خاصی پیروی نمی‌شود، اما از چارچوب مطالعهٔ پژوهش‌های مرتبط با یادگیری هستی‌شناسی استفاده به عمل می‌آید. بنابراین، مسئلهٔ اصلی این پژوهش، ارائهٔ مدلی مفهومی برای ساخت هستی‌شناسی به‌صورت نیمه‌خودکار است که حاصل آن مهندسی دانش حوزهٔ علم‌سنجی

1. semantic web
2. search engine
3. database design

ایران خواهد بود. در مقالهٔ حاضر، روش یا مدل خاصی برای استخراج و مدل‌سازی مفاهیم حوزهٔ علم‌سنجی ایران تحت عنوان هستی‌شناسی حوزه به‌عنوان مطالعهٔ موردی ارائه شد و در پایانِ این مدل مورد ارزیابی نیز قرار گرفت. به سخن دیگر، تعدادی اسناد در حوزهٔ علم‌سنجی ایران وجود دارد که این پژوهش قصد دارد، مفاهیم اصلیِ بیانگر این حوزه و ارتباط معنایی میان آن‌ها را از طریق ساخت یک هستی‌شناسی مشخص نماید. بدین‌منظور لازم است هستی‌شناسی آن در این حوزه طراحی و ترسیم شود که دربرگیرندهٔ مفاهیم و ارتباط معنایی میان آن‌ها باشد و در نهایت، به‌واسطهٔ آن، درکی از شناخت حوزهٔ مورد پژوهش حاصل شود و به‌عنوان یک سند راهبردی در برنامه‌ریزی، سیاست‌گذاری، آینده‌نگری و آینده‌پژوهی این حوزه از آن استفاده شده و نیز به‌عنوان ابزاری برای تحلیل و ترسیم ساختار دانش سایر حوزه‌ها استفاده شود. بنابراین، هدف پژوهش حاضر، مدل‌سازی مفهومی دانش حوزهٔ علم‌سنجی در ایران است که به‌واسطهٔ آن، درکی از وجود مفاهیم حوزه و روابط سلسله‌مراتبی آن همراه با خصوصیات و ویژگی‌های معنایی آن فراهم می‌گردد.

۲. پیشینهٔ پژوهش

در سال‌های اخیر پژوهش‌های بسیاری در خارج از کشور در زمینهٔ یادگیری و کاربردهای هستی‌شناسی در بهبود کارآمدی، بازیابی اطلاعات و شخصی‌سازی فرایند جست‌وجو انجام شده است. برخی پژوهش‌ها نیز به کاربرد طرح‌های تبدیل اصطلاح‌نامه به هستی‌شناسی پرداخته‌اند. در این میان می‌توان به پژوهش‌های Kawtrakul (2005), Soergel (2005) Kless (2012), Huang (2007), و در ایران به پژوهش‌های «خسروی و وظیفه‌دوست» (۲۰۰۷) و «حسینی بهشتی» (۱۳۹۳) به‌عنوان نمونه‌ای از پژوهش‌های تبدیل اصطلاح‌نامه به هستی‌شناسی اشاره کرد. «حسینی بهشتی و اژه‌ای» با هدف شکل‌گیری هستی‌شناسی در حوزهٔ علوم پایه، اصطلاح‌نامه‌هایی که پیش از این در پژوهشگاه علوم و فناوری اطلاعات ایران در حوزه‌های مختلف شیمی، فیزیک، زیست‌شناسی، زمین‌شناسی و ریاضی تدوین شده بود، مبنای ساخت هستی‌شناسی قرار داد. «صنعت‌جو و فتحیان» نیز در پژوهشی در زمینهٔ هستی‌شناسی، به مقایسهٔ کارآمدی اصطلاح‌نامه و هستی‌شناسی در بازیابی مفاهیم موضوعی پرداختند. هستی‌شناسی پژوهشگران مذکور از نوع هستی‌شناسی دامنه است و شامل واژگان مربوط به مفاهیم یک حوزه و روابط آن است و هدف از آن،

مدل‌کردن یک حوزهٔ خاص است (۱۳۹۰). «شریف» در دو پژوهش بازنمایی دانش سازمان از طریق فناوری هستی‌شناسی و امکان‌سنجی مهندسی خودکار ساخت هستی‌شناسی را مورد مطالعه قرار داد (۱۳۸۸ الف و ب).

برخی دیگر از پژوهش‌ها برای خوشه‌بندی و ساخت مفاهیم در پایگاه اطلاعاتی مورد مطالعه قرار گرفته‌اند. در این زمینه می‌توان به پژوهش‌های Chen (2008), Chuang (2008) و «حورعلی» (۱۳۹۰) اشاره کرد. «حورعلی» در پژوهشی روشی خودکار برای یادگیری هوشمند هستی‌شناسی در تمام سطوح یادگیری ارائه کرد که می‌توان از آن در کاربردها و حوزه‌های مختلف استفاده نمود. در این روش، نیازی به وجود هستی‌شناسی‌های عمومی یا تخصصی اولیه و واژگان معنایی از پیش تعریف شده نبوده و پایگاه دانش اولیهٔ آن تنها شامل مجموعه‌ای از متون ورودی است (۱۳۹۰). پژوهش حاضر در این دسته قرار می‌گیرد و چون از روشی نیمه‌خودکار بهره می‌برد، تا حدودی دقت بیشتری نسبت به پژوهش‌های نامبرده در ساخت مفاهیم دارد. علاوه بر آن، در این پژوهش از روش تحلیل هم‌واژگانی، که فنی در تحلیل محتوا و یکی از روش‌های علم‌سنجی است، استفاده شده است.

برخی فعالیت‌های انجام‌شده در ایران عموماً متمرکز بر «وردنت»[1] یا شبکهٔ واژگانی بوده است؛ مانند پژوهش «شمس‌فرد» (۱۳۸۱، ۱۳۹۱) و «فدایی، قادر، و فیلی» (۱۳۹۱). «وردنت» یا شبکهٔ واژگانی شامل مجموعه‌هایی از واژه‌های هم‌معناست که با روابطی مشخص مانند شمول، جزءواژگی[2]، اشتقاق و غیره به هم مرتبط شده‌اند و از آن در بررسی مفهومی واژگان توسط رایانه در شاخه‌های مختلف زبان طبیعی استفاده می‌شود.

بنابراین، در سال‌های اخیر، تلاش‌های بسیاری برای طراحی و خودکارسازی مدارک علمی در ترسیم و طراحی آن در جهان انجام گرفته است. نتایج پژوهش‌های مورد بررسی حاکی از آن است که از هستی‌شناسی جهت شناسایی مفاهیم و روابط بین این مفاهیم و خوشه‌بندی اطلاعات حوزه‌های علمی به‌صورت خودکار استفاده شده و همچنین، بیشتر پژوهش‌ها در جهت یادگیری روش ترسیم هستی شناسی صورت گرفته است.

مدل ساخت هستی‌شناسی پژوهش حاضر از نوع هستی‌شناسی دامنه است با این تفاوت که از دانش پایه و موجود استفاده نمی‌کند. به‌عبارتی، داده‌های مورد استفاده در

1. WordNet
2. meronomy

این هستی‌شناسی از نوع داده‌های ساختارنیافته است که از متون مدارک حوزهٔ خاصی مانند علم‌سنجی به روش‌های خودکار استخراج شده و پایهٔ مفهومی هستی‌شناسی مورد پژوهش بر مبنای آن شکل گرفته است. مزیت عمدهٔ روش هستی‌شناسی در این پژوهش، ارائهٔ مدلی مفهومی است که به نوعی، هم یادگیری هستی‌شناسی را آموزش می‌دهد و هم، حوزهٔ مورد مطالعه را مدل‌سازی مفهومی می‌کند.

معمولاً برای به‌دست‌آوردن دانش پایه و تعیین روابط سلسله‌مراتبی مفاهیم در ساخت هستی‌شناسی حوزه‌های علمی از سه روشِ (۱) هستی‌شناسی‌های قبلی، ۲) اصطلاح‌نامه‌های موجود و ۳) داده‌های غیرساختاریافته استفاده می‌شود. اصطلاح‌نامه‌ها با توجه به این‌که یک منبع دانش سازماندهی‌شدهٔ متخصصان است و روابط معنایی اعم و اخص را به ما می‌دهد، منبع مناسبی در ساخت هستی‌شناسی حوزه‌های علمی هستند. هستی‌شناسی‌های قبلی نیز در حوزه‌های مربوطه پایهٔ مناسبی برای گسترش آن هستی‌شناسی هستند. مشکل عمده، نبودِ این دو دانش پایه در بعضی حوزه‌هاست. به‌کارگیری روش‌های نوین برای استخراج و سپس تعیین روابط سلسله‌مراتبی مفاهیم آن دسته از حوزه‌هایی که فاقد دانش پایه هستند، از اهمیت زیادی برخوردار است. مزیت روش پژوهش حاضر در این است که مدلی ارائه می‌کند که هم، منجر به استخراج مفاهیم زیادی از مدارک غیرساختاریافته می‌گردد و هم، مانع دخالت انسان در انتخاب و مفهوم‌سازی دانش پایه نمی‌شود. در نتیجه، از این لحاظ که به‌صورت نیمه‌خودکار عمل می‌کند، نسبت به روش‌هایی که فقط از روش خودکار استفاده می‌کنند، دانش معنایی با جامعیت و مانعیت بهتری برای ساخت هستی‌شناسی استخراج می‌نماید.

۳. روش پژوهش

این پژوهش در صدد است مدلی ارائه دهد که بتوان شبکهٔ اطلاعات مبتنی بر هستی‌شناسی حوزهٔ علم‌سنجی ایران را ارزیابی کرد. به‌عبارتی، مهندسی دانش حوزه علم سنجی است. برای این کار، پژوهش در سه مرحله صورت گرفت. در مرحلهٔ اول، از طریق یادگیری مدل، مفاهیم حوزهٔ علم‌سنجی ایران شناسایی شد و در مرحلهٔ دوم بر اساس آن، هستی‌شناسی حوزه ساخته شد و در مرحلهٔ سوم، به‌منظور غنی‌سازی هستی‌شناسی، سامانهٔ مدل پیشنهادی پژوهش مورد ارزیابی و سنجش قرار گرفت. در مرحلهٔ اول، از روش C-value برای یادگیری مفاهیم استفاده شد. در مرحلهٔ دوم، از تحلیل هم‌رخدادی واژگان و در مرحلهٔ

سوم، از روش‌های آماری استفاده شد.

تحلیل هم واژگانی که بر اساس هم‌رخدادی واژگان در اسناد مدارک عمل می‌کند، فنی در تحلیل محتواست که قادر است روابط پنهان ایده‌ها و الگوهای هر حوزهٔ علمی را روشن کند (He 1999). «کینگ» تحلیل هم‌واژگانی را به‌عنوان جانشینی برای تحلیل هم‌استنادی در نظر می‌گیرد (King 1987). «لیدسدورف» (۱۹۹۱) این فن را برای حوزهٔ علم‌سنجی به‌کار برده است (Leydesdorff 1991). «وایتاکر» تحلیل هم واژگانی روشی کارآمد برای تحلیل محتوا می‌داند که در میزان ارتباط بین واژگان کلیدی در داده‌های متنی تأثیرگذار است و فضای واژگان کلیدی را به مجموعه‌ای از گراف‌های شبکه‌ای محدود می‌کند که به‌طوری تأثیرگذار به تشریح قوی‌ترین ارتباط موجود بین توصیفگرها می‌پردازد (Whittaker 1989). تحلیل هم‌واژگانی نمونه‌ای از روش مدل‌سازی گرافیکی است که در آن از ایده‌های مربوط به تحلیل رابطه استفاده می‌شود (Kaufman & Rousseeuw 1990).

در این پژوهش از روش تحلیل هم‌رخدادی واژگان به‌منظور استخراج سلسله‌مراتب مفاهیم و ترسیم هستی‌شناسی استفاده شده است. سلسله‌مراتب مفاهیم، اطلاعات را به رده‌هایی ساختاردهی می‌کند تا امکان جست‌وجو، استفادهٔ مجدد، و درک آن‌ها آسان شود. سامانه‌های دانش‌بنیاد[1] با مشکل اکتساب دانش و به‌ویژه مدل‌سازی دانش حوزه مواجه هستند که در این موارد استخراج سلسله‌مراتب مفاهیم می‌تواند راهگشا باشد. نتایج برخی پژوهش‌ها نشان می‌دهد که رخداد برخی واژه‌ها به معنای رخداد دیگر واژه‌ها در جملات، پاراگراف یا اسناد مشابه است و رابطهٔ مستقیمی بین آن دو کلمه وجود دارد. این نظریه با نظریهٔ هم‌مکانی مرتبط است و «تحلیل هم‌رخدادی واژگان»[2] نامیده می‌شود. این تحلیل یکی از روش‌های استخراج سلسله‌مراتب مفاهیم است. دو کلمه را در صورتی «هم‌مکان» می‌گویند که در یک پاراگراف، جمله یا سند با یکدیگر رخ دهند یا نزدیک به هم بیشتر از حد تصادف ظاهر شوند.

جامعیت و مانعیت اصطلاحات، تأثیر به‌سزایی در کیفیت هستی‌شناسی نهایی حوزه‌ها خواهد داشت. یک اصطلاح مناسب باید دو ویژگی «توصیف» و «تمایز» را دارا

1. knowledge-based system
2. terms co-occurrence analysis

باشد. توصیف، به این معناست که اصطلاح مناسب باید محتوای اطلاعاتی یک سند را به‌درستی بیان کند و تمایز، یعنی اصطلاحی مناسب است که یک سند را از سندهای دیگر متمایز سازد (Syafullah 2010&Salim). روش‌های بسیاری جهت تبدیل اصطلاحات به مفاهیم (مفهوم‌سازی) و به‌طور کلی، یادگیری واژگان وجود دارد. بسیاری از روش‌ها بر مبنای روش‌های زبانی و تعدادی دیگر بر مبنای روش‌های آماری هستند. روش C-value از جمله روش‌هایی است که ترکیبی از روش‌های زبانی و آماری را برای استخراج واژه‌های ترکیبی به‌کار می‌گیرد (Frantzi, Ananiadou & Mima 2000). از جمله مزیت‌های این روش آن است که نسبت به روش‌هایی که تنها از مقدار فراوانی به‌منظور استخراج واژه‌ها استفاده می‌کنند، دقت بیشتری دارد و مزیت عمدهٔ دیگر آن، توانایی استخراج واژه‌های تودرتو و ترکیبی است. این روش واژه‌های چندکلمه‌ای اسناد را با استفاده از روش‌های زبانی و آماری استخراج می‌کند.

جامعهٔ پژوهش حاضر شامل چکیدهٔ ۲۵۸۵ مدرک تولیدشدهٔ پژوهشگران ایرانی است که در قالب کتاب‌ها، مقاله‌ها، پایان‌نامه‌های تحصیلی و طرح‌های پژوهشی حوزهٔ علم‌سنجی در داخل و خارج از ایران اعم از چاپی و الکترونیکی به زبان فارسی و سایر زبان‌ها از بدو پیدایشِ این حوزه در ایران (۱۳۶۱) تولید شده است. برای گردآوری داده‌ها از پایگاه‌های داخلی و خارجی و سایر پایگاه‌ها استفاده گردید و همچنین، علاوه بر جست‌وجوهای تخصصی، مجله‌های علم اطلاعات و دانش‌شناسی و کتابشناسی علم‌سنجی ایران به‌طور کامل مرور شد. پژوهش در هفت مرحله انجام گرفت:

در مرحلهٔ اول، مفاهیم و واژگان اولیه از مدارک استخراج شدند. در این مرحله، چکیده و عنوان کلیهٔ مدارک، تهیه شد و سپس، به‌صورت یک فایل txt درآمد و به روش متن‌کاوی، نمایه‌سازی خودکار شد. برای انجام این کار از نرم‌افزار داده‌کاوی «دانش‌نگار سند» استفاده شد. اساس کار این نرم‌افزار بدین‌صورت است که به‌طور خودکار، بر اساس برخی واژه‌ها و نشانه‌ها و با زبان «عبارات الگودار[1]»، شکاف‌هایی در متن ایجاد می‌کند. احتمال دارد واژه‌ها و عباراتی که در بین این شکاف‌ها قرار می‌گیرند، اصطلاحاتی باشند که دربردارندهٔ مفهوم باشند. این ابزار عباراتی را که حاوی مفهوم نیستند خارج کرده و سپس، واژه‌ها و عبارات مفهومی را تعیین می‌کند (توکلی‌زاده راوری ۱۳۹۴).

۱. در زبان‌های برنامه‌نویسی است که اکثر زبان‌های سطح بالا آن را پشتیبانی می‌کنند. این زبان بر الگو استوار است (Regular expression)

با به‌کارگیری این مدل نمایه‌سازی بر روی اسناد، در مرحلهٔ اولیه بیش از ۳۰۰۰ واژه-مفهوم استخراج شد.

در مرحلهٔ دوم، پیش‌پردازش زبانی مفاهیم صورت گرفت. هدف از این مرحله، تبدیل واژگان به مفاهیم و به‌طور کلی مفهوم‌سازی است و از طریق روش زبانی C-value صورت گرفت. در روش‌های زبانی عمدتاً از روش‌های پردازش متن نظیرِ علامت‌گذاری[1]، توکن‌ساز[2]، برچسب‌گذاری بخشی از کلام[3]، پالایهٔ زبانی[4]، تحلیل‌گر نحوی[5]، ریشه‌یابی زبانی[6] و شناسایی موجودیت‌های اسمی[7] (نام اشخاص، مکان‌ها، سازمان‌ها و غیره) استفاده می‌شود (Ruiz-Martinez & Valencia 2011). روش زبانی C-value شامل برچسب‌گذاری بخشی از کلام، پالایهٔ زبانی و سیاههٔ کلمات بازدارنده است. برچسب‌گذاری بخشی از کلام، اختصاص‌دادن برچسب‌های گرامری (نظیر اسم، صفت، فعل، حرف اضافه، ضمیر و غیره) به هر واژه در متن است (حسینی بهشتی ۱۳۹۲). در روش زبانی C-value سه پالایهٔ زبانی زیر در نظر گرفته شده است:

1. Noun+Noun
2. (Adj|Noun)+Noun
3. ((Adj| Noun) +((Adj|Noun) *(Noun preposition)?) (Adj|Noun) *)Noun

با اِعمال این پالایه‌ها و درنظرگرفتن لیست کلمات توقف، در نهایت، حدود ۱۸۰۳ مفهوم در این مرحله به‌دست آمد.

در مرحلهٔ سوم، مفاهیم بااهمیت انتخاب شد. در این مرحله از روش آماری C-value برای انتخاب مهم‌ترین مفاهیم در شبکهٔ مفهومی و هستی‌شناسی حوزه استفاده شد. روش آماری C-value، روشی برای استخراج واژه‌های چندکلمه‌ای است که هدف آن بهبود استخراج واژه‌های تودرتو است. واژه‌های تودرتو آن‌هایی هستند که با واژه‌های طولانی‌تر ظاهر می‌شوند و ممکن است به‌تنهایی در متن رخ ندهند (Frantzi, Ananiadou & Mima 2000).

1. tokenizing
2. tokenizer
3. part of speech (POS) tagging
4. linguistic filter
5. syntactic analyzer (parser)
6. stemmer
7. name entity recognition (NER)

به‌عنوان مثال اصطلاح «ضریب همکاری» تودرتو است که با مفاهیمی نظیر «ضریب همکاری علمی» و «ضریب همکاری گروهی» در متن رخ می‌دهند. واژه‌های چندکلمه‌ای دارای معنای متمایزتر و مشخص‌تری نسبت به تک‌واژه‌ها در ساخت هستی‌شناسی است و به‌طور کلی، برای مدل‌سازی دانش یک حوزهٔ علمی مناسب‌ترند. مقدار آماری C-value به رشته‌های مفهومی، مقادیری را نسبت می‌دهد و بر این اساس، آن‌ها را در سیاههٔ خروجی رتبه‌بندی می‌کند. به‌منظور محاسبهٔ مقدار C-value رشتهٔ a، دو حالت زیر در نظر گرفته می‌شود:

الف- اگر a رشته‌ای با بیشترین طول باشد، یا تودرتو نباشد، مقدار C-value با استفاده از فراوانی کلی آن در متن و طول آن بر اساس رابطهٔ زیر به‌دست می‌آید که در آن |a| طول رشتهٔ a و f(a) فراوانی رخداد آن در متن است:

$$\boldsymbol{C-value(a) = log_2|a|.(f(a))}$$

ب- اگر a، رشته‌ای تودرتو باشد، باید بررسی شود که آیا بخشی از واژه‌ها با طول بلندتر است یا نه. اگر چنین باشد، برای محاسبهٔ مقدار C-value باید فراوانی آن به‌عنوان یک رشتهٔ تودرتو و تعداد واژه‌های طولانی‌تر محاسبه شود. در این حالت مقدار C-value بر اساس رابطهٔ زیر محاسبه می‌شود:

$$C-value(a) = log_2|a|.(f(a) - \frac{1}{P(T_a)}\sum_{b\in T_a} f(b))$$

که در آن |a| طول رشتهٔ a، f(a) فراوانی رخداد رشتهٔ a در متن، T_a مجموعهٔ رشته‌های نامزد استخراج‌شده شامل a، $P(T_a)$ تعداد عناصر T_a و $\sum_{b\in T_a} f(b)$ مجموع فراوانی‌هایی است که a در رشته‌های طولانی‌تر رخ می‌دهد (همان). به‌منظور درک بهتر نحوهٔ محاسبه، به نمونهٔ زیر اشاره می‌شود:

چنانچه دو مفهوم مانند «ضریب تأثیر» با تعداد رخداد ۲۰ و «ضریب تأثیر مجلات» با رخداد ۸ داشته باشیم. ارزش c-value آن‌ها بر اساس محاسبهٔ فرمول فوق به‌ترتیب، ۱۶/۵ و ۱۱/۰۶ است:

$Log_2|3|.8=11.06$ و $log_2|2|.(20 - \frac{۱}{۲}(8))=16/5$

بدین‌ترتیب، با انجام محاسبات فوق، در این مرحله در حد آستانهٔ مقدار عددی (۲)، ۶۵۳ مفهوم برای ادامهٔ کار در ساخت هستی‌شناسی انتخاب گردید. آستانهٔ مورد نظر

به‌صورت تجربی در نظر گرفته شده که باعث انتخاب مفاهیم بااهمیت‌تری می‌شود.

در مرحلهٔ چهارم، ماتریس با هدف خوشه‌بندی اسناد و مدارک حوزه ساخته شد. این ماتریس نوعی ماتریس نامتقارن است که در آن سطرها دربرگیرندهٔ اسناد، و ستون‌ها دربرگیرندهٔ مفاهیم حوزه است. این ماتریس، یک ماتریس ۶۵۳×۲۵۸۵ است. عناصر این ماتریس ۰ یا ۱ هستند و بیانگر این است که واژهٔ مورد نظر در آن سند بوده است یا خیر. ماتریس‌های مورد نظر از طریق نرم‌افزار Ravar_Matrix.exe ساخته شد.

در مرحلهٔ پنجم، خوشه‌بندی و وزن‌دهی به مفاهیم هر خوشه انجام گرفت. در این پژوهش از طریق روش meansk- با استفاده از نرم‌افزار شبکهٔ عصبی MATLAB خوشه‌بندی‌ها صورت گرفت. عمل وزن‌دهی به‌منظور رتبه‌بندی، تعیین مفاهیم متناظر و همچنین، تعیین روابط سلسله‌مراتبی در هستی‌شناسی حوزه صورت گرفت. برای این منظور، از روش فراوانی «واژه- معکوس فراوانی سند، TF-IDF» (Leydesdorff and Welbers 2011) استفاده شد. TF بیانگر نسبت مدارک دربرگیرندهٔ آن مفهوم در بین تمامی مدارک است. برای محاسبهٔ آن‌ها ابتدا فراوانی واژه ***i*** در مدرک j (F_{ij}) محاسبه می‌شود و با هنجارکردن آن در تمامی مجموعه، مقدار TF به‌دست می‌آید؛ یعنی رابطه‌های زیر:

$$IDF_i = log\ (N/n_i) \qquad TF_{ij} = F_{ij} / max\ (F_{ij})$$

$TF\text{-}IDF = TF_{ij} * IDF_i$ اساس IDF بر این است که واژه‌هایی که در مدارکِ بسیاری ظاهر می‌شوند، بیشتر بیانگر موضوع خاص هستند. به همین دلیل، برای محاسبهٔ آن ابتدا تعداد اسنادی که دربرگیرندهٔ واژه ***i*** هستند (n_i) و تعداد کل اسناد در مجموعهٔ (N) مشخص شده، سپس IDF به‌صورت زیر محاسبه می‌شود:

این رابطه نشان‌دهندهٔ حاصل‌ضرب TF در IDF و بیانگر اهمیت یک واژه یا مفهوم در مدرک بوده و می‌توان بر اساس آن، مفاهیم موجود در مدارک را بر حسب میزان اهمیت آن‌ها رتبه‌بندی کرد.

مرحلهٔ ششم، شامل ساخت سلسله‌مراتب مفهومی هستی‌شناسی (رده‌بندی) است. روابط رده‌بندی به‌طوری گسترده برای سازماندهی دانش هستی‌شناسی‌ها با استفاده از روابط خاص/ عام به‌کار می‌روند و اغلب رابطهٔ «هست» را شامل می‌شوند. به‌منظور یادگیری روابط رده‌بندی از روش‌های مختلفی نظیر روش‌های مبتنی بر الگو، روش‌های آماری و روش‌های یادگیری ماشین استفاده شده است (Liu, Hogan & Crowley 2011). در این پژوهش برای ساخت روابط سلسله‌مراتبیِ مفاهیم حوزه از روش آماری هم‌رخدادی

واژگان استفاده شده است. «سندرسون و کرفت» در سال ۱۹۹۲ تعریف رده‌بندی را بدین‌صورت ارائه کردند: واژهٔ T_1 خاص‌تر از واژه T_2 است. اگر T_2 در همهٔ اسنادی که T_1 رخ دهد، ظاهر شود، این رویکرد را می‌توان به شکل زیر $P(x|y) > P(y|x)$ تعمیم داد: واژهٔ x شامل واژهٔ y است اگر داشته باشیم:

که $P(x|y)$ به‌شکل زیر تعریف می‌شود: $P(x|y) = \frac{n(x , y)}{n(y)}$

n(x,y) تعداد اسنادی است که x و y با هم رخ می‌دهند و n(y) تعداد اسنادی است که شامل y هستند (Sanderson and Croft 1992 نقل در Cimiano 2006).

بدین‌ترتیب، مطابق با خوشه‌بندی مرحلهٔ قبل، مفاهیم متناظر هر خوشه که دارای وزن بالاتری است به‌عنوان سطح یک هستی‌شناسی علم‌سنجی منظور گردید و سایر مفاهیم به شرط آن محاسبه شده‌اند. میزان آستانهٔ درنظر گرفته‌شده ۰/۰۵ است. بنابراین، اگر احتمال شرطی محاسبه‌شده از میزان آستانه بیشتر باشد، آن مفاهیم در زیرمفهوم سطح یک افزوده می‌شود؛ در غیر این‌صورت دومین مفهوم که دارای هم‌رخدای بیشتری بوده در نظر گرفته شده است و این رویّه تا زمانی ادامه می‌یابد که همهٔ مفاهیم در سلسله‌مراتب هستی‌شناسی جای گیرند. آستانهٔ مورد نظر برای رده‌بندی به‌صورت تجربی انتخاب شده و در این حد، روابط سلسله‌مراتبی قابل قبولی استخراج می‌کند. در این رده‌بندی، مفهوم «علم‌سنجی» که قصد رده‌بندی آن را داریم در رأس قرار داده شد و سایر مفاهیم در زیر این مفهوم جای گرفته‌اند. به‌عنوان نمونه، طبق رابطهٔ ۷، احتمال شرطی مفهوم «خودپیوندی وب» به شرط مفهوم «ضریب تأثیر وب» به‌صورت زیر محاسبه می‌شود:

$$P(\text{خودپیوندی وب} \mid \text{ضریب تأثیر وب}) =$$

$$\frac{\text{(تعداد اسناد شامل مفاهیم خودپیوندی وب و ضریب تأثیر وب)}}{\text{(تعداد اسناد شامل مفهوم ضریب تأثیر وب)}} = \frac{5}{54} = 0.09$$

به‌دلیل این‌که میزان این احتمال شرطی بیشتر از حد آستانهٔ ۰/۰۵ است، مفهوم *«خودپیوندی وب به‌عنوان فرزند مفهوم ضریب تأثیر وب»* در سلسله‌مراتب هستی‌شناسی قرار می‌گیرد.

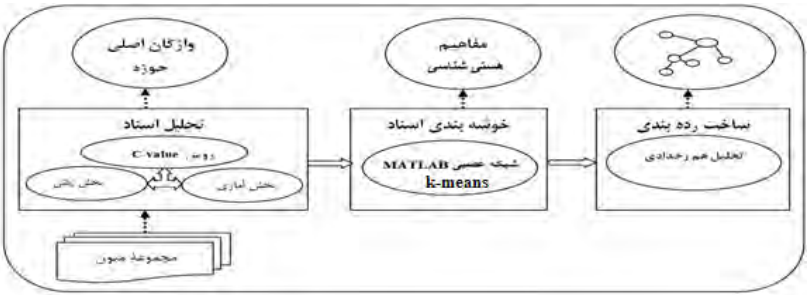
مرحلهٔ هفتم، شامل ترسیم هستی‌شناسی (آنتوگراف)[1] است. در این مرحله مطابق

1. Anthropology

آنچه که به‌دست آمده، استخراج و ترسیم روابط سلسله‌مراتبی هستی‌شناسی حوزه از طریق نرم‌افزار protégé انجام شد. این نرم‌افزار یکی از قدرتمندترین ابزارهای ایجاد و مدیریت هستی‌شناسی به‌شمار می‌آید. نرم‌افزار فوق، توسط «دانشگاه استنفورد» ساخته شده و مجموعه‌ای غنی از ساختارهای مدل‌سازی دانش ارائه می‌کند و از زبان‌های OWL و RDF پشتیبانی می‌نماید. با استفاده از این زبان‌ها، کاربران قادرند به ساخت و مدیریت هستی‌شناسی مبادرت ورزند. کاربران همچنین می‌توانند هستی‌شناسی مبتنی بر این دو زبان را بارگذاری، ذخیره، کلاس‌بندی، تعیین خصیصه‌ها، نمونه‌ها و قواعد سبک هستی‌شناسی را ویرایش کنند. ایجاد هستی‌شناسی و ترسیم آنتوگرافی حوزهٔ علم‌سنجی ایران با این نرم‌افزار انجام گرفت. نمونه‌ای از رده‌بندی حوزه با این نرم‌افزار در شکل ۲ مشاهده می‌شود.

مرحلهٔ هشتم و پایانی، شامل ارزیابی سامانهٔ مدل پیشنهادی ساخت هستی‌شناسی است. محتوای هستی‌شناسی‌ها نیز باید قبل از به‌کارگیری و مهندسی مجدد مورد ارزیابی قرار گیرند. در واقع، پس از ساخت هستی‌شناسی، ارزیابی هستی‌شناسی به‌منظور تضمین اطمینان از برآوردن هدف مد نظر ضروری است. به‌منظور ارزیابی نتایج روش ساخت هستی‌شناسی این پژوهش از دو رویکرد استفاده شد. ابتدا، از خبرگان دانش حوزهٔ مورد نظر به‌منظور تعیین دقت نتایج حاصل استفاده شد. سپس، روش به‌کار برده‌شده در این پژوهش با روش‌های مشابه در سایر پژوهش‌ها مقایسه شد. معیار ارزیابی توسط خبرگان، میانگین نظرات آنان می‌باشد.

به‌طور کلی، فرایند ساخت هستی‌شناسی در این پژوهش شامل مراحل: (۱) تحلیل اسناد، (۲) خوشه‌بندی اسناد، (۳) استخراج سلسله‌مراتب مفاهیم و ساخت هستی‌شناسی و (۴) ارزیابی هستی‌شناسی است. معماری سامانهٔ پیشنهادی در شکل ۱، مشاهده می‌شود.



شکل ۱. چرخهٔ یادگیری معماری سامانهٔ ساخت هستی‌شناسی

٤. تجزیه و تحلیل یافته‌ها

با توجه به مراحل اجرای این پژوهش که پیش‌تر در روش‌شناسی پژوهش بیان گردید، در این قسمت به ذکر نمونه‌هایی از یافته‌های به‌دست‌آمده با روش‌های مذکور پرداخته می‌شود. یافته‌های این پژوهش طیف وسیعی از داده‌ها را شامل می‌شود که به‌دلیل محدودیت فضای مقاله امکان نمایش تمام آن‌ها نیست و تنها چند نمونه از مفاهیم استخراج‌شده، خوشه‌های به‌دست‌آمده، و تصاویری از ترسیم هستی‌شناسی ارائه می‌گردد. ابتدا، نمونه‌ای از مفاهیم استخراج‌شده بر اساس پالایه‌های زبانی نمایش داده می‌شود.

الف. نمونه‌ای از مفاهیم استخراج‌شده با اعمال پالایه‌های زبانی روش C-value

بر اساس این پالایه‌ها، با برچسب‌گذاری اصطلاحاتی که در مرحلهٔ نمایه‌سازی به‌دست آمده بودند، واژه‌ها به مفاهیم یا اصطلاحاتی ترکیبی تبدیل شدند. با اِعمال این پالایه‌ها و درنظرگرفتن لیست کلمات توقف، در نهایت، حدود ۱۸۰۳ مفهوم به‌دست آمد. نمونه‌ای از مفاهیم پالایه‌شده در شکل زیر مشاهده می‌شود.

^تحلیل شبکه اجتماعی;شاخص مرکزیت بینابینی شبکه;شاخص مرکزیت شبکه;شاخص مرکزیت نزدیکی شبکه;شاخص های تحلیل شبکه اجتماعی;شاخص های شباهت مدرک;شاخص های علم سنجی;شاخص های هم نویسندگی;
^روابط علمی;شاخص های همکاری علمی;;همکاری علمی منطقه ای;
^ارزیابی علم;
^تحلیل استنادی;نشانه گذاری آثار علمی;همبستگی استنادی;
^ارزیابی وب سایت;رتبه بندی وب سایت;شبکه های کتاب محور;
^آموزش روش های علم سنجی; روش های علم سنجی; مطالعه نظری علم سنجی;نقشه جامع علمی;نیاز سنجی اطلاعات;
^آینده پژوهی;;
^همکاری علمی بین المللی;
^تحلیل روند پژوهش;تحلیل شبکه اجتماعی;;داده کاوی اطلاعات;کشف دانش;متن کاوی اطلاعات;
^ارزیابی تطبیقی فعالیت های پژوهشی;;رتبه بندی مؤسسه های پژوهشی;سنجش توانمندی علم;قطب علمی; جبهه پژوهش;
^سنجش تولیدات علمی;ارزیابی کیفی تولیدات علمی;پژوهش کیفی;;
^پژوهش کیفی;;شاخص های ارزیابی کیفی پژوهش;نفوذ مدارک;جبهه پژوهش
^ارزیابی وب سایت;پیوند وب;تراکم واژگان; هم پیوندی وب;
^استناد درون متنی;تحلیل استنادی; تحلیل محتوای کتاب;
^آسیب شناسی پژوهش;ارزیابی علم;تأثیر خود استنادی;خود استنادی;رتبه بندی نویسندگان;شاخص g;شاخص های علم سنجی;شاخص هرش;
^ارزیابی فعالیت های پژوهشی;اولویت های پژوهش;

شکل ۲. نمونه‌ای از مفاهیم استخراج‌شده پس از اعمال پالایه‌های زبانی

ب. نمونه‌ای از مفاهیم استخراج‌شده با اعمال روش آماری C-value

با اعمال محاسبات آماری بر روی تعداد ۱۸۰۳ مفهوم که در مرحلهٔ زبانی روش مذکور حاصل شده بود و درنظرگرفتن آستانهٔ عدد ۲، در این مرحله تعداد ۶۵۳ مفهوم بااهمیت به‌دست آمد. همان‌گونه که پیش‌تر توضیح داده شد، مبنای محاسبهٔ این روش

آماری، طول کلمات و فراوانی رخداد آن در اسناد است. بدین‌ترتیب، ۶۵۳ مفهوم واژه در ساخت هستی‌شناسی مدّ نظر قرار گرفت. تعدادی از مفاهیم استخراج‌شده در این مرحله در جدول ۱ آورده شده است.

جدول ۱. نمونه‌ای از مفاهیم استخراج‌شده بر اساس روش آماری C-value

رتبه	مفاهیم	مقدار c-value	رتبه	مفاهیم	مقدار c-value
۱	ارزیابی تولیدات علمی	۴۶۱/۳۶	۱۱	همکاری علمی	۹۵/۵
۲	تولید علم	۱۷۵/۵	۱۲	ترسیم ساختار علم	۹۴/۸
۳	ارزیابی وب‌سایت	۱۶۴/۳۲	۱۳	روابط هم‌نویسندگی	۹۳/۲۲
۴	تحلیل استنادی	۱۵۱	۱۴	مطالعهٔ نظری علم‌سنجی	۹۱/۶۴
۵	ارزیابی فعالیت‌های پژوهشی	۱۴۶/۹۴	۱۵	شاخص‌های علم‌سنجی	۹۰/۰۶
۶	تحلیل استنادی پایان‌نامه‌ها	۱۳۴/۷۲	۱۶	ضریب تأثیر	۹۰
۷	ارزیابی تطبیقی تولیدات علمی	۱۳۰	۱۷	ارزیابی مجلات	۸۸
۸	پایگاه wos	۱۲۲	۱۸	ضریب تأثیر مجلات	۸۲/۶۲
۹	پایگاه isi	۱۰۶	۳۰۹	چالش استناد وبی	۴/۷۴
۱۰	تحلیل استنادی مجلات	۹۸/۲۷	۶۵۳	هنجارهای علمی	۲

ج. خوشه‌بندی اسناد

خوشه‌بندی مفاهیم حوزه در پژوهش حاضر، با هدف یافتن گروه‌هایی از اسناد و مفاهیمی که با یکدیگر دارای مشابهت هستند و نیز ایجاد روابط سلسله‌مراتبی مفاهیم در ساخت هستی‌شناسی انجام گرفت. نتایج به‌دست‌آمده حاکی از تشکیل ۱۸ خوشهٔ تقریباً منسجم در حوزهٔ علم‌سنجی است. در جدول شماره ۲، خوشه‌های شکل‌گرفته به همراه تعداد اسناد هر خوشه ارائه شده است.

جدول ۲. نتایج خوشه‌بندی شبکهٔ عصبی MATLAB

خوشه	تعداد اسناد	خوشه	تعداد اسناد	خوشه	تعداد اسناد	خوشه	تعداد اسناد
۱	۳۶	۶	۴۳	۱۱	۲۰	۱۶	۶
۲	۳۵	۷	۱۰۳	۱۲	۲۲	۱۷	۴
۳	۴۷	۸	۲۳	۱۳	۲۰	۱۸	۲
۴	۷۵	۹	۴۲	۱۴	۳۷	---	----
۵	۳۱	۱۰	۳۳	۱۵	۱۹	----	-----

د. وزن‌دهی به مفاهیم هر خوشه

برای هر یک از مفاهیم موجود در اسناد خوشه‌ها، وزن TF-IDF مطابق آنچه که در روش پژوهش شرح داده شد، محاسبه گردید و مفاهیمی که بیشترین مقدار را داشتند، به‌عنوان نام آن خوشه (مفهوم هستی‌شناسی) انتخاب شد. در صورتی که مفهومی از لحاظ معنایی نمایندهٔ کل مفاهیم خوشه نبود، از مفهوم کلی‌تری استفاده شد. در جدول شماره ۳، نحوهٔ محاسبهٔ TF-IDF برخی از واژه‌های خوشهٔ اول آورده شده است.

جدول ۳. مقدار TF-IDF برای برخی از واژه‌های خوشهٔ اول

مفاهیم	فراوانی واژه در سند (F_{ij})	فراوانی واژه (TF)	تعداد اسناد دربرگیرندهٔ واژه (ni)	معکوس فراوانی سند (IDF)	TF-IDF
اخلاق علمی	۴۲	$F_{ij}/max(F_{ij})=1$	۴۲	log (۲۵۸۵/۴۲) = ۱/۷۹	۱/۷۹
آسیب‌شناسی پژوهش	۳۹	۰/۹۳	۳۹	۱/۸۲	۱/۶۹
عامل تولید علم	۳۲	۰/۷۶	۳۲	۱/۹۱	۱/۴۵
موانع تولید علم	۲۹	۰/۶۹	۲۹	۱/۹۵	۱/۳۶
آسیب‌شناسی علم	۱۸	۰/۴۳	۱۸	۲/۱۶	۰/۹۲
داوری علمی	۱۶	۰/۳۸	۱۶	۲/۲۱	۰/۸۴
چالش تولید علم	۱۵	۰/۳۶	۱۵	۲/۲۴	۰/۸۰
سرقت علمی	۱۵	۰/۳۶	۱۵	۲/۲۴	۰/۸۰

ه. ساخت هستی‌شناسی (رده‌بندی مفاهیم)

برای استخراج یا یادگیری روابط رده‌بندی در هستی‌شناسی‌ها از روش‌های مختلفی

نظیر روش‌های مبتنی بر الگو، روش‌های آماری، و روش یادگیری ماشین استفاده شده است (Liu 2011). همان‌گونه که در روش‌شناسی پژوهش تشریح گردید، در این پژوهش برای تعیین روابط رده‌بندی بین مفاهیم حوزهٔ علم‌سنجی، از روش آماری هم‌رخدادی واژگان استفاده شد و برای روابط غیررده‌بندی «بخشی از» و «نمونه‌ای از» از نظرات خبرگان علم‌سنجی بهره برده شد.

هستی‌شناسی ساخته‌شده، علاوه بر روابط رده‌ای و کلاس‌بندی‌ها، شامل ویژگی‌ها، خصوصیات و نوع مقادیر هر یک از رده‌ها نیز می‌شود که از طریق نرم‌افزار Protégé ایجاد و مدیریت شده است. در شکل شماره ۳، نمایی از مدیریت هستی‌شناسی حوزهٔ مورد مطالعه در این نرم‌افزار نمایش داده شده است. اما، همان‌گونه که بیان شد، مهم‌ترین و اساسی‌ترین کار یک هستی‌شناسی دامنه یا حوزه، تعیین روابط نظام سلسله‌مراتب مفهومی است. روابط رده‌بندی در هستی‌شناسی معمولاً بر اساس رابطهٔ «هست» یا is a تعیین می‌شود و برای روابط غیررده‌بندی روش‌های خاصی به‌کار برده می‌شود. در پژوهش حاضر، علاوه بر روابط is a، از دو نوع روابط غیررده‌بندی دیگر مانند روابط غیررده‌بندی "Part of" به معنای بخشی و "Instance of"، در هستی‌شناسی نهایی بر اساس نظرات خبرگان حوزه استفاده شد.

به‌منظور یادگیری رابطهٔ is a، از شبکهٔ هم‌رخدادی واژگان استفاده شد. برای محاسبهٔ احتمالات شرطی (همان‌طور که در روش‌شناسی پژوهش تشریح شد) داشتن احتمالات پیشین ضروری است که آستان آن برای تمامی محاسبات به‌صورت تجربی، عدد ۰/۰۵ در نظر گرفته شده است. برای انتخاب این عدد، آزمایش‌های متعددی صورت گرفت، به‌طوری که در صورت انتخاب عددی بیش از این، مفاهیم ارزشمندی حذف می‌شد و انتخاب عددی کمتر از این نیز باعث انتخاب مفاهیم غیرمرتبطی در رده‌بندی این حوزه می‌گشت. بنابراین، این عدد برای این حوزه، مناسب تشخیص داده شد. برای یادگیری روابط is a رویّهٔ زیر در نظر گرفته شده است:

مفهوم علم‌سنجی که قصد ساخت هستی‌شناسی آن را داریم، در رأس قرار گرفته و سایر مفاهیم در زیر این مفهوم جای گرفته‌اند. ترتیب استخراج رده‌بندی، بر حسب وزن مفاهیم اصلی (هستی‌شناسی سطح ۱) است. در این مرحله، ۱۸ مفهوم به‌عنوان مفاهیم اصلی در این روابط جای گرفته‌اند که در شکل ۴، نشان داده شده است. بدین‌منظور، ابتدا آن دسته از مفاهیمی که در هر خوشه دارای بیشترین وزن بوده به‌عنوان مفهوم سطح

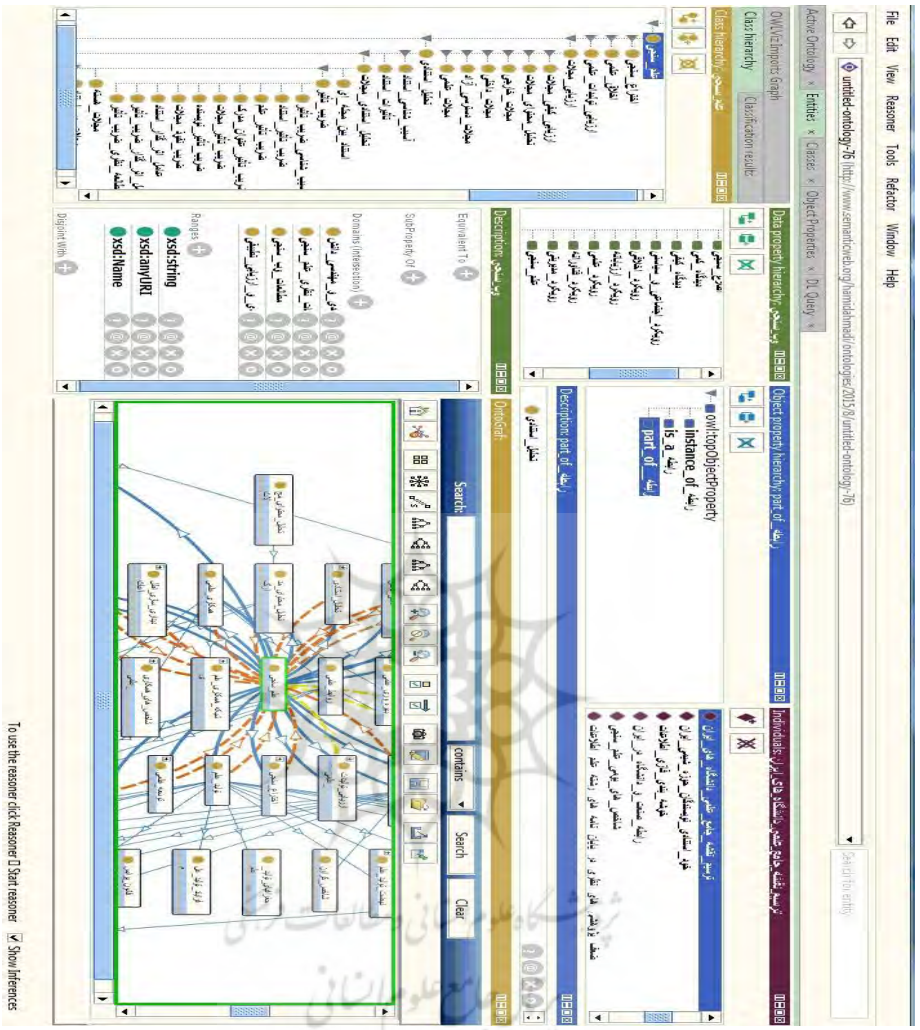
۱ هستی‌شناسی قرار گرفته و سایر مفاهیم که در خوشه‌های خود یا سایر خوشه‌ها جای گرفته‌اند، در صورتی که با آن مفهوم هم‌رخدادی داشتند، به شرط آن محاسبه شد و هستی‌شناسی سطح ۲ را تشکیل داده‌اند. در این مرحله تعدادی مفهوم، زیر مفاهیم اصلی قرار گرفتند. در بین مفاهیم باقیمانده، مفهوم یا مفاهیمی که بیشترین وزن را دارند، یافت شده و احتمال شرطی مفاهیم دیگری به شرط آن محاسبه شده است. بدین‌ترتیب، سایر سطوح هستی‌شناسی تکمیل گردید. این رویّه تا هنگامی که تمام مفاهیم در سلسله‌مراتب مفهومی جای بگیرند، ادامه پیدا کرده است.

در شکل شماره ۴، ردهٔ اصلی هستی‌شناسی (آنتوگراف یا هستان‌نگار) حوزهٔ علم‌سنجی ایران نمایش داده شده است. هستی‌شناسی سطح ۱، زیرردەهای اصلی علم‌سنجی ایران را به نمایش گذاشته است. در این کلاس‌ها ردهٔ تحلیل استنادی، تولید علم، و مطالعات نظری، به‌ترتیب، دارای بیشترین رده و زیررده هستند و رده‌های اختراع‌سنجی، پژوهش زنان و تحلیل محتوای مدارک، کمترین رده و زیررده را در این حوزه دارند.

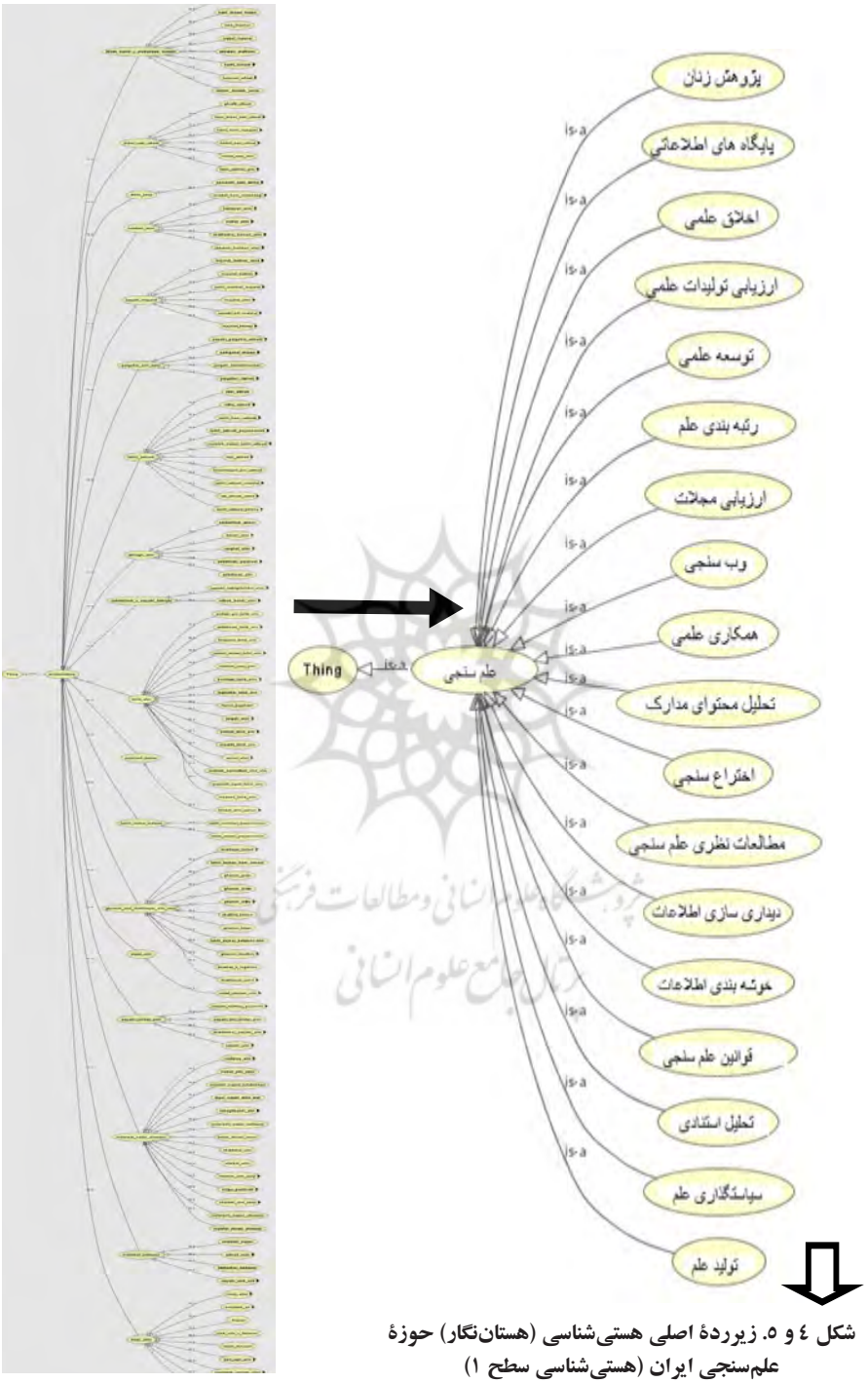
در شکل شماره ۳، نمایی کلی از تولید و مدیریت هستی‌شناسی حوزه با نرم‌افزار Protégé مشاهده می‌شود. به‌دلیل محدودیت فضای مقاله، امکان نمایش تمام رده‌های حوزه وجود ندارد و تنها ردهٔ کلی و ردهٔ تولید علم به‌عنوان نمونه نمایش داده شده است. هستی‌شناسی ساخته‌شده از طریق نرم‌افزار Protégé با آدرس[1] در آینده قابل رؤیت خواهد بود.

در شکل ۳، استخراج و ترسیم روابط سلسله‌مراتبی هستی‌شناسی حوزه از طریق نرم‌افزار protégé دیده می‌شود. شکل ۴ و ۵، رده‌های سطح ۱ هستی‌شناسی حوزه را نمایش می‌دهند و در شکل شماره ۶، ردهٔ «تولید علم» که یکی از نمونه‌های رده‌بندی حوزهٔ علم‌سنجی است، با روابط is a ترسیم شده است.

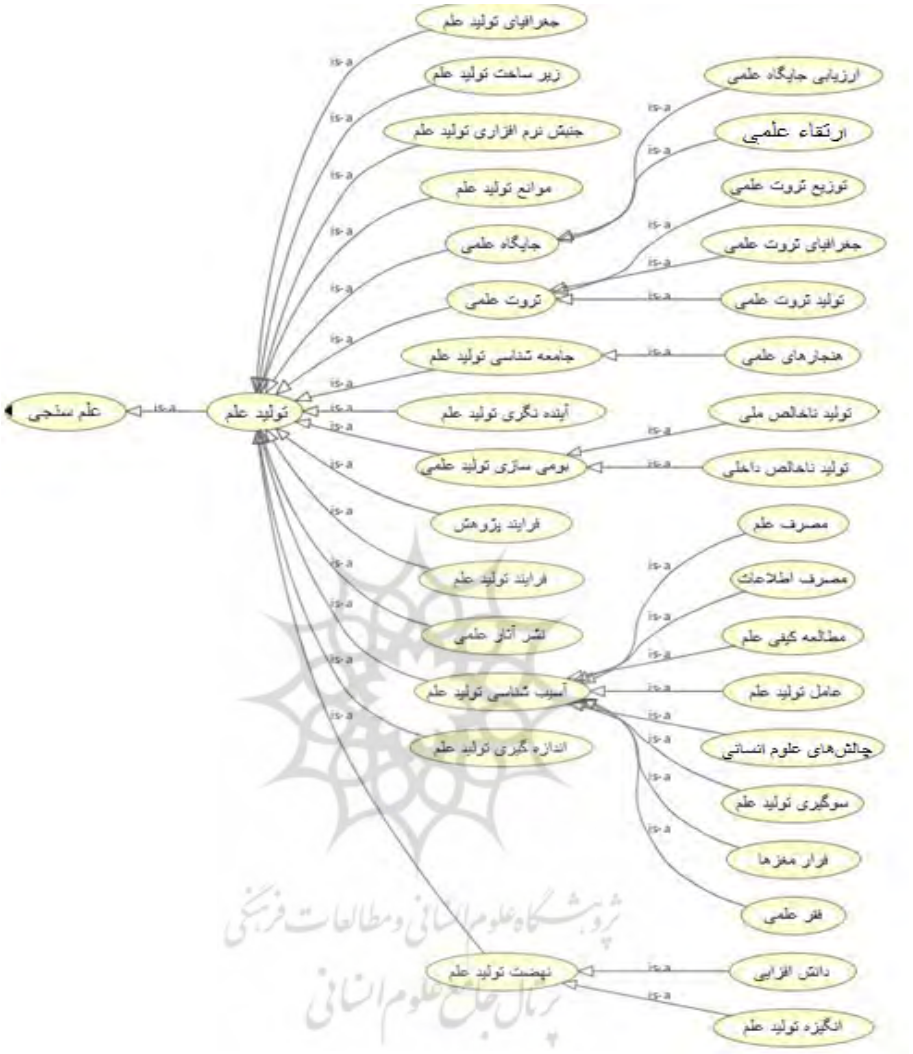
1. http://www.sementicweb.org/scientometrics/ontologies/2015/hamid-ahmadi-iran



شکل ۳. بخشی از کلاس‌بندی مفاهیم هستی‌شناسی حوزهٔ علم‌سنجی ایران با نرم‌افزار Protégé



شکل ٤ و ٥. زیردهٔ اصلی هستی‌شناسی (هستان‌نگار) حوزهٔ علم‌سنجی ایران (هستی‌شناسی سطح ۱)



شکل ٦. زیردهٔ تولید علم هستی‌شناسی (هستان‌نگار) حوزهٔ علم‌سنجی ایران

و. ارزیابی سامانه هستی‌شناسی

به‌منظور ارزیابی نتایجِ روش ساخت هستی‌شناسی از دو رویکرد بهره گرفته شد. ابتدا، از خبرگان دانش حوزهٔ مورد نظر به‌منظور تعیین دقت نتایج به‌دست‌آمده استفاده شد. سپس، روش به‌کار برده‌شده در این پژوهش با روش‌های مشابه در پژوهش‌های دیگر مقایسه شد. نتایج دو روش مورد استفاده در ادامه تشریح می‌شود.

۱) نتایج ارزیابی خبرگان

Precision (C_P) =A/A+B

Precision (C_L_P) =C/C+D

برای ارزیابی نتایج این پژوهش و مقایسهٔ هستی‌شناسی به‌دست‌آمده با نتایج واقعی، از نظرات ده نفر از خبرگان این حوزه برای ارزیابی دقت هستی‌شناسی ساخته‌شده استفاده شد. بدین‌منظور، دو نوع روش ارزیابی مد نظر قرار گرفت. دقت مفاهیم[1] که بیانگر دقت کلمات کلیدی است، از طریق روش پیشنهادی پژوهش به‌دست آمد و دقت مکانی مفاهیم[2] که هم، بیانگر دقت کلمات کلیدی (مفاهیم) انتخابی و هم، نشان‌دهندهٔ دقت مکان کلمات کلیدی در سلسله‌مراتب روابط هستی‌شناسی است. روابط مورد استفاده برای این دو شاخص به شرح زیر است که متغیرهای A، B، C و D به‌ترتیب زیر تعریف می‌شوند (Chen, Liang, Pan 2008).

A: واژه‌ها (مفاهیم) که سامانه (مدل مفهومی روش پژوهش) آن‌ها را تولید و خبره تأیید می‌کند.

B: واژه‌ها (مفاهیم) که سامانه آن‌ها را تولید کرده، اما خبره تأیید نمی‌کند.

C: واژه‌ها (مفاهیم) که سامانه آن‌ها را تولید و خبره مکان آن‌ها را در سلسله‌مراتب هستی‌شناسی تأیید می‌کند.

D: واژه‌ها (مفاهیم) که سامانه آن‌ها را تولید و خبره مکان آن‌ها را در سلسله‌مراتب هستی‌شناسی تأیید نمی‌کند.

همچنین، به‌منظور بهره‌گیری از نظرات خبرگان در تعیین نوع ارتباط مفاهیم به یکدیگر (غنی‌سازی هستی‌شناسی ساخته‌شده)، از خبرگان خواسته شد نوع ارتباط مفاهیم به یکدیگر را با استفاده از یکی از روابط Is-a ، Pat-of و Instance of تعیین کنند. بدین‌منظور، از نظرات خبرگان در تعیین روابط غیرردهای هستی‌شناسی حوزهٔ علم‌سنجی ایران در این پژوهش بهره جسته و پس از استخراج هستی‌شناسی اولیه، هستی‌شناسی اصلی بر اساس این سه رابطه مجدداً ویرایش گردید.

1. concept precision (C_P)
2. concept location precision (C_L_P)

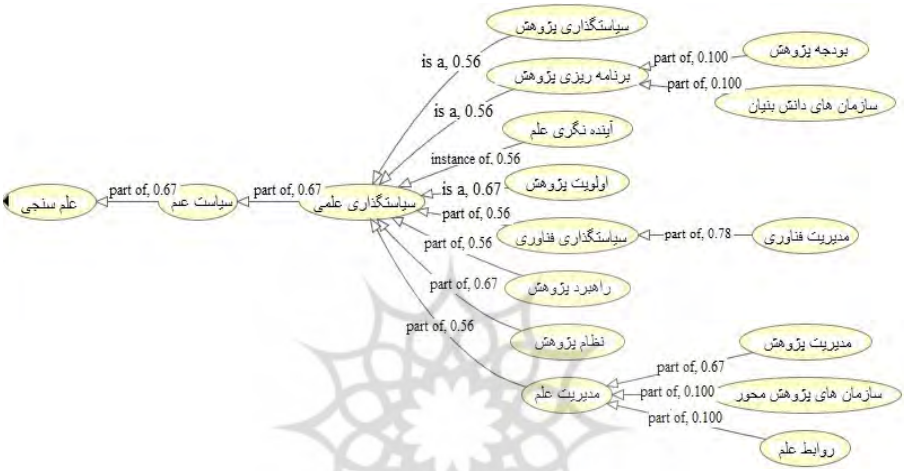
جدول ٤. نتایج ارزیابی هستی‌شناسی توسط خبرگان

متغیرهای ارزیابی / نظر خبرگان	A	B	C	D
خبرهٔ ۱	۶۲۷	۲۲	۴۱۶	۱۰
خبرهٔ ۲	۶۲۵	۲۴	۴۱۰	۱۶
خبرهٔ ۳	۵۹۴	۵۷	۳۹۰	۳۶
خبرهٔ ۴	۶۱۷	۳۱	۴۱۱	۱۵
خبرهٔ ۵	۵۷۹	۷۰	۴۰۸	۱۸
خبرهٔ ۶	۶۰۷	۴۲	۴۰۱	۲۵
خبرهٔ ۷	۵۸۹	۶۰	۳۹۰	۳۶
خبره ۸	۶۲۰	۲۶	۳۳۰	۹۶
خبره ۱۰	۶۲۳	۲۶	۳۸۰	۴۶
خبره ۱۱	۵۹۸	۵۱	۴۰۰	۲۶
میانگین نظرات	۶۱۰/۸	۳۸	۳۹۵/۵	۳۰/۵
		Average C_P=۰/۹۳۵	Average C_L_P=۰/۹۲۸	

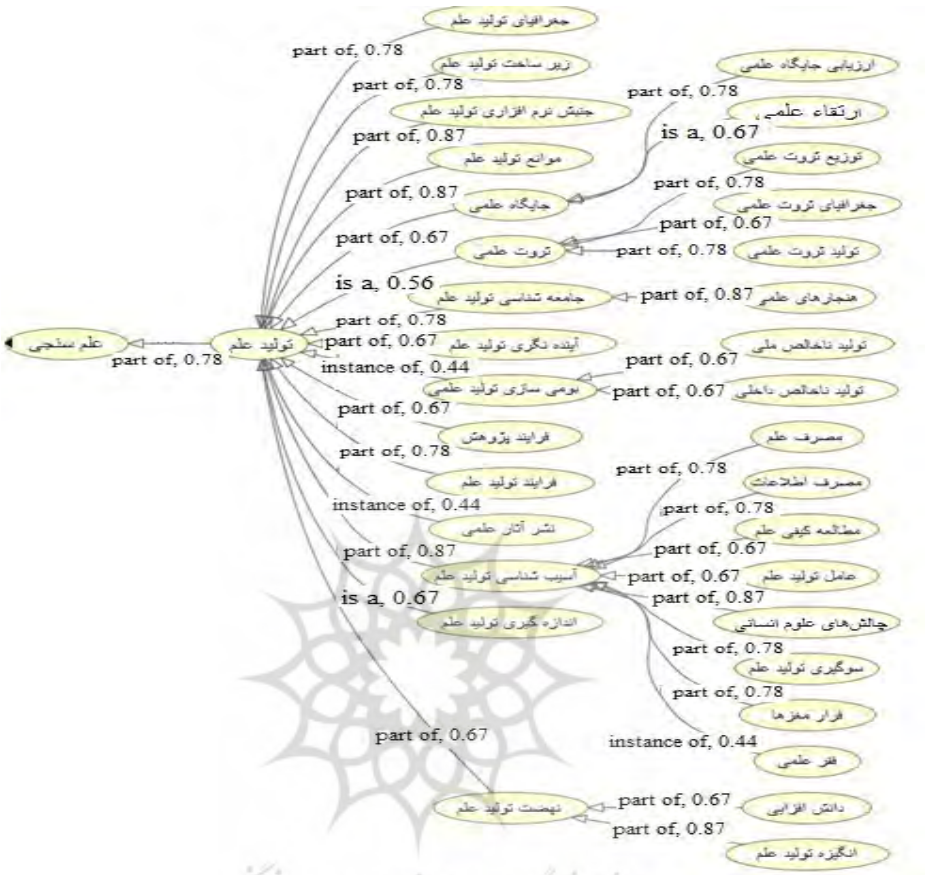
در جدول ۴، نتایج ارزیابی خبرگان آورده شده است. به این ترتیب، میانگین دقت مفاهیم (Average C_P) ده خبره برابر با ۰/۹۳۵ و میانگین دقت مکانی مفاهیم (Average C_L_P) ده خبره برابر با ۰/۹۲۸ به‌دست آمد.

در این پژوهش علاوه بر ارزیابی روابط رده‌ای میان مفاهیم، از دو نوع رابطهٔ غیررده‌ای استفاده گردید. بدین‌ترتیب، در ارزیابی سومی از متخصصان حوزه خواسته شد، روابط رده‌بندی را تأیید یا رد کنند و چنانچه نوع رابطهٔ به‌دست‌آمده (is a) را تأیید نمی‌کنند، از دو نوع رابطهٔ غیررده‌بندی مانند (Part of و Instance Of) در تعیین نوع ارتباط میان مفاهیم استفاده کنند. نتایج نظرات به‌دست آمده مجدداً در سامانهٔ نرم‌افزاری protge اعمال گردید. اعمال نظرات متخصصان حوزه شامل نوع ارتباط مفاهیم و درصد این رابطه است. چنانچه رابطه بین دو مفهوم بر اساس میانگین نظرات خبرگان، نوع رابطه "Part of" تشخیص داده شد، میزان این رابطه نیز با درصد مشخص گردیده است. مثلاً مفهوم «فرار مغزها» تا ۷۸ درصد بخشی از مفهوم «آسیب‌شناسی تولید علم» محسوب می‌شود. بدین‌ترتیب،

تمامی مفاهیمی که در هستی‌شناسی اولیه بر اساس روابط رده‌ای استخراج شدند، در مرحلهٔ غنی‌سازی بر اساس نظرات خبرگان حوزه بازنگری شده و علاوه بر روابط رده‌بندی، سایر روابط غیررده‌بندی میان مفاهیم اعمال گردید. در شکل ۷ و ۸، ردهٔ «سیاست‌گذاری علم» و ردهٔ «تولید علم» به‌عنوان نمونه، پس از ارزیابی خبرگان به شکل‌های زیر غنی‌سازی شدند.



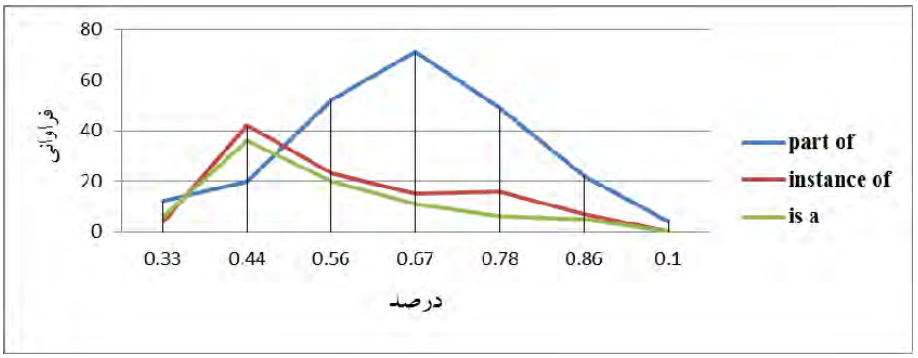
شکل ۷. زیردهٔ سیاست‌گذاری علم هستی‌شناسی (هستان‌نگار) حوزهٔ علم‌سنجی ایران بر اساس نظرات خبرگان



شکل ۸. زیردهٔ تولید علم هستی‌شناسی (هستان‌نگار) حوزهٔ علم‌سنجی ایران بر اساس نظرات خبرگان

نتایج ارزیابی خبرگان علم‌سنجی در مورد تأیید و تعیین روابط میان مفاهیم حوزه نشان می‌دهد که رابطهٔ غیرده‌بندی "Part of" بیشتر از دو رابطهٔ دیگر در نظام سلسله‌مراتب مفهومی هستی‌شناسی حوزه حاکم است. از ۴۰۸ رابطه، ۲۳۰ مفهوم، رابطهٔ "Part of" با هم دارند. در این میان میزان این رابطه‌ها متغیرند، به‌طوری که ۱۴۲ رابطه بیش از ۷۰ درصد و ۸۴ رابطه از ۳۳ درصد تا ۵۶ درصد متغیر هستند. از میان رابطه‌های غیرده‌بندی رابطهٔ "Instance of" به معنای «نمونه‌ای از» در رتبهٔ دوم قرار دارد و به‌طور کلی، از ۴۰۸ رابطه، ۱۰۷ رابطه از این نوع هستند. اما میزان آن متغیر است و این میزان در حد متوسط می‌باشد. رابطهٔ «is a» که از قبل، توسط مدل مفهومی این پژوهش تعیین شده بود و از نوع رده‌بندی است، بر اساس نظر خبرگان به میزان کمتری تأیید گردید و در مجموع، فقط تعداد ۸۴

مفهوم دارای رابطهٔ "is a" هستند و میزان آن هم به‌طور میانگین حدود ۵۰ درصد است. برای عینیت بیشتر در نمودار شماره ۱، سه رابطهٔ فوق نمایش داده شده است.



نمودار ۱. فراوانی و درصد فراوانی نوع ارتباط مفاهیم هستی‌شناسی حوزهٔ علم‌سنجی ایران

۲) نتایج مقایسهٔ روش پیشنهادی با روش‌های دیگر

در کار مشابهی که توسط Pan & Chen, Liang (2008)، Chuang and Chen (2008) و «حورعلی» (۱۳۹۰) در حوزهٔ نوشیدنی‌ها و فناوری اطلاعات انجام شده، میانگین ارزیابی دقت مفاهیم و دقت مکانی مفاهیم این پژوهش در مقایسه با آن‌ها مطابق جدول زیر مشاهده می‌شود.

جدول ۵. مقایسهٔ دقت روش پیشنهادی با روش‌های مشابه

دقت / روش	چن	چانگ	حورعلی	پیشنهادی
میانگین دقت مفاهیم	۰/۷۹	۰/۸۵	۰/۹۱	۰/۹۳
میانگین دقت مکانی مفاهیم	۰/۷۴	۰/۷۵	۰/۸۶	۰/۹۲

همان‌گونه که آمارهای این جدول نشان می‌دهد، میانگین دقت مفاهیم و میانگین دقت مکانی مفاهیمِ روش پیشنهادی نسبت به روش‌های مشابه بیشتر است.

۵. نتیجه‌گیری

امروزه با آمدن فناوری‌های اطلاعاتی و بهره‌گیری از فنون و الگوریتم‌های آن، روابط معنایی میان مفاهیم را می‌توان با ایجاد هستی‌شناسی در سطح وسیع‌تر و با داده‌های

بزرگ‌تری تحلیل و بررسی کرد تا حوزه‌های علمی از طریق آن مدل‌سازی مفهومی شوند. هستی‌شناسی‌های خاص یک حوزه، ابزارهای کارآمدی در کمک به کاربران برای جست‌وجو و بازیابی اطلاعات و مدل‌کردن دانش حوزه و همچنین، ابزارهایی مفید در حوزهٔ وب معنایی هستند. بنابراین، ساخت سریع و کارآمد هستی‌شناسی‌ها گام مهمی در جست‌وجوی محتوای حوزه در پایگاه‌های دانش و ابزاری برای تحلیل حوزه هستند. در این مقاله فرایند جدیدی برای ساخت هستی‌شناسی با استفاده از روش C-value، تحلیل هم‌رخدادی واژگان، روش وزن‌دهی TF-IDF و خوشه‌بندی اطلاعات ارائه شده که قادر است مفاهیم مرتبط در حوزهٔ دانش مورد نظر را به کاربران ارائه دهد. از روش C-value به‌منظور استخراج واژه‌های اصلی، از شبکهٔ عصبی MATLAB به روش k-means به‌منظور خوشه‌بندی اسناد، از روش TF-IDF برای انتخاب مفهوم متناظر با خوشه‌های استخراج‌شده و از تحلیل هم‌رخدادی واژگان در استخراج سلسله‌مراتب و ساخت هستی‌شناسی استفاده گردید. به نظر می‌رسد که از ترکیب روش‌های مختلفی که ذکر گردید، در نهایت، می‌توان در ساخت هستی‌شناسی حوزه‌ها بهره برد. در ارزیابی هستی‌شناسی ساخته‌شده دو رویکرد، بهره‌گیری از خبرگان حوزه و مقایسه با روش‌های مشابه در نظر گرفته شد و میانگین نظرات خبرگان، ملاک ارزیابی دقت هستی‌شناسی قرار گرفت. به این ترتیب، میانگین دقت مفاهیم برابر با ۰/۹۳ و میانگین دقت مکانی مفاهیم برابر با ۰/۹۲ به‌دست آمد. همچنین، دقت روش پیشنهادی برای ساخت هستی‌شناسی، نسبت به روش‌های مشابه مقایسه شد و مشاهده گردید که میانگین دقت مفاهیم و میانگین دقت مکانی مفاهیمِ روش پیشنهادی، در هر دو حالت، از روش‌های مشابه بیشتر است. به نظر می‌رسد که این دقت، حاصل روش نیمه‌خودکار این پژوهش نسبت به روش خودکار آن در پژوهش‌های مشابه است. در ساخت هستی‌شناسی‌ها روش ساخت هستی‌شناسیِ استخراج‌شده دارای ضعف‌هایی از این قبیل است که تنها روابط is a را تولید می‌نماید و سایر روابط رده‌بندی و غیررده‌بندی را استخراج نمی‌کند. در پژوهش حاضر، غیر از روابط "is a" از روابطی مانند "Part of" به معنای «بخشی از» و "Instance of" به معنای «نمونه‌ای از» استفاده شد. این روابط، بر اساس نظرات خبرگان علم‌سنجی ایران اعمال گردید. نتایج این ارزیابی حاکی از آن است که بیشترین روابط میان مفاهیم حوزهٔ علم‌سنجی ایران از نوع رابطهٔ غیررده‌بندی "Part of" و "Instance of" است. به‌عبارتی، رابطهٔ معنایی مفاهیم در هستی‌شناسی علم‌سنجی ایران در حدود ۸۰ درصد از نوع جزءواژگی است. در نهایت، باید گفت که روش

نیمه‌خودکاری که در این پژوهش به‌کار گرفته شد، روشی مناسب برای تعیین روابط سلسله‌مراتبی مفاهیم در ساخت هستی‌شناسی پایه است و تا حدود زیادی در حوزه‌هایی که فاقد اصطلاح‌نامه است، روشی کارآمد برای تعیین روابط رده‌بندی و غیررده‌بندی است و از طریق آن می‌توان روابط معنایی میان مفاهیم را تا حدود زیادی شناخت. مشکل اصلی در ایجاد هستی‌شناسی‌ها به‌دست‌آورن داده‌های باکیفیت است. تاکنون روش‌های گوناگونی برای استخراج داده‌ها اجرا شده است. در این پژوهش، روش جدیدی ارائه گردید که باعث شد داده‌ها به‌صورت خودکار استخراج شده، سپس، در یک فرایند زبانی و آماری، داده‌های باکیفیت به‌دست آید.

فهرست منابع

توکلی‌زاده راوری، محمد. ۱۳۹۴. مدل دومرحله‌ای شکاف- گلچین برای نمایه‌سازی خودکار متون فارسی. *تحقیقات اطلاع‌رسانی و کتابخانه‌های عمومی* ۲۱ (۱): ۱۳-۴۰.

حسینی بهشتی، ملوک‌السادات. ۱۳۹۲. *ساخت‌واژه، اصطلاح‌شناسی و مهندسی دانش*. تهران، نشر چاپار.

_____، و فاطمه اژه‌ای. ۱۳۹۳. طراحی و پیاده‌سازی هستی‌شناسی علوم پایه بر اساس مفاهیم و روابط موجود در اصطلاح‌نامه‌های مرتبط. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۰ (۳): ۶۷۷-۶۹۶.

حورعلی، مریم. ۱۳۹۰. یادگیری هوشمند هستان‌نگار برای بسط پرسمان در جست‌وجوی معنایی (مورد پژوهی: کتابخانه دیجیتالی). پایان‌نامه منتشرشده دکتری. دانشگاه تربیت مدرس.

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

_____. ۱۳۸۸ ب. *بازنمایی دانش سازمانی پژوهشگاه اطلاعات و مدارک علمی ایران با استفاده از فناوری هستی‌شناسی*. همایش ملی مدیریت دانش و علوم اطلاعات، تهران.

شمس‌فرد، مهرنوش. ۱۳۹۱. *ساخت نیمه‌خودکار وردنت فارسی: از واژه تا واژهستان‌شناسی*. مجموعه‌مقالات دومین همایش زبان‌شناسی رایانشی. تهران. انتشارات اهورا. انجمن زبان‌شناسی ایران، ۲۲۲-۲۰۲.

_____، و احمد عبدالله‌زاده. ۱۳۸۱. ساخت هستان‌شناسی از روی متون زبان طبیعی. تهران: مقاله منتشرشده در *هشتمین کنفرانس سالانه انجمن کامپیوتر ایران*. بازیابی شده از: http://www.civilica.com. (دسترسی در ۱۳۹۳/۳/۲۰)

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

فدایی، مرضیه، حمیدرضا قادر، هشام فیلی. ۱۳۹۱. *تولید خودکار وردنت فارسی با استفاده از روش‌شناختی و کاربرد آن در بازیابی اطلاعات*. مجموعه‌مقالات دومی هم‌اندیشی زبان‌شناسی رایانشی. تهران: انتشارات

اهـورا، انجمـن زبان‌شناسـی ایـران.

نشـاطی، محمـود. ۱۳۸۶. یادگیـری هستی‌شناسـی از انباردهـی متنـی. پایان‌نامـهٔ کارشناسـی ارشـد، دانشـگاه صنعتـی شـریف.

Chen, R. C., J. Y. Liang, and R. H. Pan. 2008. Using recursive ART network to construction domain ontology based on term frequency and inverse document frequency. *Expert Systems with Applications* 34 (1): 488–501.

Chuang, C. H., and R. C. Chen. 2008. Automating construction of a domain ontology using a projective adaptive resonance theory neural network and Bayesian network. *Expert Systems* 25 (4): 414-430.

Cimiano, P. 2006. Ontology Learning and Population from Text Algorithms. Evaluation and Applications. Computational Linguistics, The Association for Computer Linguistics, 45, 888–895.

Cruz, L. F. 2007. A visual tool for ontology alignment to enable geospatial interoperability. Journal of Visual Languages and Computing, 18, 230–254. from: doi:10.1016/j.jvlc.2007.02.005. (accessed March 8, 2015).

Frantzi, K., S. Ananiadou, and H. Mima. 2000. Automatic recognition of multi- word terms. *International Journal of Digital Libraries* 3 (2): 115-130.

Gruber, T. R. 1993. A translation approach to portable ontologies. Appeared in Knowledge Acquisition, 5 (2), 199–220. from: http://lisas.de/~david/brunel/references/Gruber_ontolingua-kaj-1993.pdf. (accessed April 6, 2014).

He, Q. 1999. Knowledge discovery throght coword analysis. *Library Trends* 48 (1): 133-159.

Huang, Jin-Xia; Shin, Ji-Ae; & Choi, Key-Sun. 2007. Enriching Core Ontology with Domain Thesaurus through Concept and Relation Classification. Proc. OntoLex, ISWC.

Kaufman L., & Rousseeuw P. J. (1990). Finding Groups in Data: AnIntroduction to Cluster Analysis. John Wiley & Sons, Retrieved 2014, May 12, from: http://onlinelibrary.wiley.com/book/10.1002/9780470316801.

Kawtrakul, A. et al. (2005) Automatic term relationship cleaning and refinement for AGROVOC. In Workshop on The Sixth Agricultural Ontology Service, Workshop "Ontologies: the more practical issues and experiences", July 25-28. Vila Real, Portugal. Retrieved from ftp://ftp.fao.org/docrep/fao/008/af240e/af240e00.pdf (accessed Nov. 2, 2015).

Khosravi, F.; Vazifedoost, A. 2007. *Creating a Persian Ontology through Thesaurus: Reengineering for Organizing the Digital Library of the National Library of Iran*. In Building An Information Society For All: Proceedings of the International Conference on Libraries, Information and Society, ICOLIS 2007, June 26-27. Petaling Jaya, Malaysia: 41-53. Retrieved from: http://dspace.fsktm.um.edu.my/xmlui/handle/1812/285 (accessed June 2015)

King, J. 1987. A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science* 13 (5): 261–276.

Kless, D., L. Jansen, J. Lindenthal, & J. Wiebensohn. 2012. in Information Systems. Proceedings of the Seventh International Conference (FOIS 2012) , IOS Press, 2012, 133-146.

Leydesdorff, L. 1991. In Search of Epistemic Networks. *Social Studies of Science* 21110-75 :(1) .

_____, & K. Welbers. 2011. The semantic mapping of words and co-words in contexts,. *Journal of Informetrics* 5 (3): 417-461.

Liu, K., W. R. Hogan, and C. S. Crowley. 2011. Natural Language Processing Methods And Systems For Biomedical Ontology Learning. *Journal Of Biomedical Informatics* 44:163–179.

Soergel, D. et.al. (2004). Reengineering Thesauri for New Applications: the AGROVOC Example. Journal of Digital Information, 4 (4). Retrieved from http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel (accessed Sept. 17, 2015)

Syafullah, M., and N. Salim. 2010. Improving Term Extraction Using Particle Swarm Optimization techniques. *Journal of computing*, Retrieved September 17, 2015, from http://arxiv.org/ftp/arxiv/papers/1002/1002.4041.pdf

Whittaker, J. 1989. Creativity and conformity in science: Titles, keywords and co-word analysis. Social Studies of Science 19 (3): 473-496.

حمید احمدی

متولد سال ۱۳۵۰، دارای مدرک تحصیلی دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی از دانشگاه شهید چمران اهواز است. ایشان هم‌اکنون در دانشگاه رازی کرمانشاه با گروه علم اطلاعات همکاری دارد.

علم‌سنجی، مهندسی دانش، فناوری اطلاعات و برنامه‌ریزی راهبردی از جمله علایق پژوهشی وی است.

فریده عصاره

دارای مدرک تحصیلی دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی از دانشگاه نیوسات ولز استرالیاست. ایشان هم‌اکنون استاد گروه علم اطلاعات و دانش‌شناسی و مدیر قطب مدیریت دانش دانشگاه شهید چمران اهواز است.

علم‌سنجی، اطلاع‌سنجی، کتاب‌سنجی، وب‌سنجی، ذخیره و بازیابی اطلاعات و نظام‌های اطلاعاتی، همکاری‌های علمی ملی و بین‌المللی، کتابخانه‌های دیجیتالی، فهرست‌نویسی و طبقه‌بندی از جمله علایق پژوهشی وی است.

ملوک‌السادات حسینی بهشتی

دارای مدرک تحصیلی دکتری در رشتهٔ زبان‌شناسی همگانی از دانشگاه تهران است. ایشان هم‌اکنون استادیار پژوهشکدهٔ مدیریت دانش پژوهشگاه علوم و فناوری اطلاعات ایران است.

زبان‌شناسی، اصطلاح‌شناسی، مهندسی دانش، از جمله علایق پژوهشی وی است.

غلامرضا حیدری

دارای مدرک تحصیلی دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی از دانشگاه شهید چمران اهواز است. ایشان هم‌اکنون استادیار گروه علم اطلاعات و دانش‌شناسی دانشگاه شهید چمران اهواز است.

مبانی و نظریه‌های اطلاعات و دانش- مدیریت دانش- علم‌سنجی و مطالعات علم- روش‌شناسی پژوهش، آموزش، پژوهش و کارآفرینی در اطلاعات و دانش از جمله علایق پژوهشی وی است.