نشریه علمی "پدافند الکترونیکی و سایبری"
سال هشتم، شماره ۳، پاییز ۱۳۹۹، ص ۹۹-۸۳

ارائه یک معماری عامل‌گرا برای کاوش معنایی از داده‌های بزرگ مقیاس در محیط‌های توزیع شده

حسین صابری[1*]، محمدرضا کنگاوری[2]، محمدرضا حسنی آهنگر[3]
۱- مربی دانشگاه جامع امام حسین(ع)، ۲- دانشیار دانشگاه علم و صنعت ایران، ۳- استاد دانشگاه جامع امام حسین(ع)
(دریافت: ۹۷/۰۸/۰۱، پذیرش: ۹۷/۱۲/۱۴)

چکیده

داده‌های بزرگ مقیاس، متشکل از داده‌های حجیم، توزیع شده، پراکنده، ناهمگون و ترکیبی از داده‌های نامتجانس، بی ربط، گمراه کننده، واقعی و غیر واقعی است. بنابراین تجزیه و تحلیل، ایجاد ارزش و بهره‌وری از داده‌ها، همواره چالشی مهم و باز محسوب می شود. بنابراین هدف این پژوهش ارائه یک معماری ائتلافی جدید برای تولید اطلاعات با ارزش برای تصمیم‌گیری از میان انبوه داده‌ها است. معماری پیشنهادی که به اختصار ASMLDE نامیده می‌شود، با هدف توسعه و بهبود داده‌کاوی، کاوش معنایی و تولید قواعد سودمند و با کیفیت از چهار لایه، هفت مؤلفه و شش عامل اصلی تشکیل می‌شود. در معماری پیشنهادی برای جمع‌آوری و استانداردسازی پردازش‌های کیفی و تفسیرهای پیچیده‌تر، از مفهوم‌سازی با فرآیند 4v 's، بینش از حجم و مقیاس داده‌ها در قالب مدل 3V's و درنهایت بینش کیفی مبتنی بر ضخامت داده‌ها استفاده شده است. این معماری با حمایت هستان‌شناسی و عامل‌کاوی، فضاهای بزرگ کاوش را کوچک‌تر و سرعت و کیفیت عملیات داده‌کاوی را به دلیل به‌کارگیری سامانه‌های چند عاملی افزایش می‌دهد. خودکارسازی عملیات کاوش، کاهش پیچیدگی داده‌ها و فرآیندهای کسب‌وکار نیز از مهم‌ترین دستاوردهای معماری پیشنهادی است. به‌منظور ارزیابی معماری پیشنهادی، مجموعه داده‌ای بزرگ مقیاس از دامنه حوادث طبیعی و کلاس هستان‌شناسی زمین لرزه از پایگاه دانش DBpedia مورد استفاده قرار گرفته است. نتایج ارزیابی که حاصل از کاوش قواعد معنایی روی مجموعه داده‌ای ذکر شده است، اثربخشی و قابلیت‌های معماری ASMLDE را در افزایش کیفیت قواعد معنایی کاوش شده متناسب با نیاز کاربر و کوچک‌تر کردن فضای بزرگ داده‌کاوی نسبت به سایر چارچوب‌ها و معماری‌های مشابه نشان می‌دهد.

کلیدواژه‌ها: داده‌های بزرگ مقیاس، کاوش معنایی، هستان‌شناسی، معماری عامل‌گرا

۱- مقدمه[1]

با توجه به نظریه‌های جدید، دانش تنها در فضای ذهنی، فضای فیزیکی و فضای اجتماعی متصور نیست، بلکه ماشین‌ها نیز می‌توانند از طریق شکل‌دهی فضای سایبری و تجزیه و تحلیل داده‌ها، دانش را کشف و بازتاب دهند [۱]. بر اساس گزارش سال ۲۰۱۶ شرکت آی- بی- ام روزانه ۲/۵ اگزا بایت داده (10^{18}) تولید می‌شود که ۹۰ درصد از داده‌های جهان طی سال‌های منتهی به سال ۲۰۱۶ تولید شده است. همچنین شاهد هستید که سازمان‌ها بیشتر به ذخیره‌سازی داده‌ها در مقیاس‌های بزرگ (ترابایت 10^{12}، پتا بایت 10^{15} و سایر) اهتمام کرده و به دلیل عدم سرمایه‌گذاری کافی و اهتمام لازم به ایجاد زیرساخت‌های لازم در تجزیه و تحلیل، کاوش و استخراج دانش مفید از داده‌ها برای تصمیم‌سازی و تصمیم‌گیری با چالش‌های متعددی روبه‌رو هستند. همچنین از ابتدای ظهور داده‌های بزرگ، چالش‌های 3 V's، 5 V's، 10 V's (حجم، سرعت، تنوع، صحت، اعتبار، آسیب پذیری، نوسان، تجسم، ارزش) همواره یک چالش باز در داده‌کاوی محسوب می‌شود[۴-۲]. بنابراین رویارویی با داده‌ها و به تبع آن استخراج اطلاعات و دستیابی به دانش به مهم‌ترین دغدغه متخصصان تبدیل شده است [۳]. به همین سبب، سازمان‌ها به دنبال راهبردهایی برای دستیابی به مزایای رقابتی، تحت آنچه"تجزیه و تحلیل داده‌های بزرگ" نامیده می‌شود هستند [۵].

در نتیجه، عصر داده‌های بزرگ مقیاس، پردازش و تجزیه و تحلیل داده‌های جدید را به ما تحمیل و مدیریت داده‌ها را به چالش کشید. اگر چه تاکنون راه‌حل‌های مختلفی برای تجزیه و تحلیل داده‌های بزرگ مقیاس ارائه شده است، اما حجم و سرعت تولید داده‌ها، به ویژه در سامانه‌های پیچیده و بسیار پویا، با بروز چالش‌های جدیدی همراه شده است که نیاز به مدیریت محتوایی و ساختاری داده‌ها دارد. بنابراین نیاز به یک معماری ائتلافی بر اساس مؤلفه‌های مبتنی بر فناوری معنایی، فناوری عامل‌گرا و

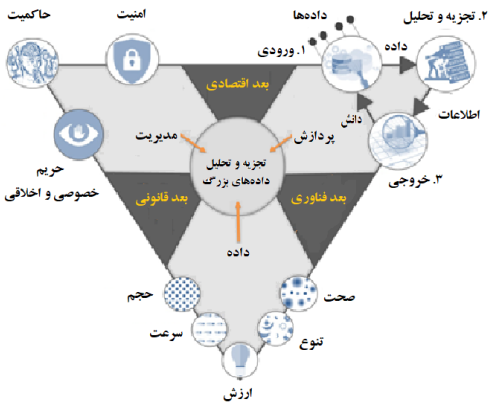
* رایانامه نویسنده مسئول: hsaberi@ihu.ac.ir

الگوریتم‌هایی با قابلیت تجزیه و تحلیل معنایی برای فائق آمدن بر چالش‌های بیان شده ضروری است. در مقاله پیش رو، یک معماری ائتلافی مبتنی بر مؤلفه که به اختصار (اَسم-اِل-دی)[1] نامیده می‌شود، برای مواجه با چالش کاوش از داده‌های بزرگ مقیاس در محیط‌های توزیع شده پیشنهاد شده است. معماری ASMLDE به منظور بالا بردن سطح انتزاع، نظارت و کنترل بر روی فرآیند داده‌کاوی به کاهش فضای کاوش و هرس نمودن فضای داده‌کاوی با استفاده از هستان‌شناسی می‌پردازد. همچنین در طرح این معماری، برای استخراج دانش‌های مفید برای تصمیم‌گیری متناسب با اهداف داده‌کاوی از حمایت فناوری معنایی مبتنی بر هستان‌شناسی و برای افزایش سرعت و کیفیت عملیات کاوش از حمایت فناوری عامل‌گرا مبتنی بر سامانه‌های چند عاملی در محیط‌های محاسباتی توزیع شده استفاده شده است.

سازماندهی این مقاله در ادامه به این شرح است: در بخش (۲)، ادبیات موضوع، تعاریف کلی و مفاهیم پایه‌ای که در این مقاله استفاده می شود معرفی شده و خلاصه‌ای از کارهای مرتبط معرفی می‌گردد. در بخش (۳) کلیات معماری پیشنهادی و در بخش (۴) معماری پیشنهادی شامل مؤلفه‌ها، عامل‌ها و جزئیات معماری پیشنهادی بیان می‌گردد. در بخش (۵)، ارزیابی معماری پیشنهادی مورد بحث قرار می‌گیرد. بخش (۶) نیز به نتیجه‌گیری و همچنین کارهای آتی می‌پردازد.

۲- ادبیات موضوع

اولین قدم در پردازش معنایی داده‌های بزرگ مقیاس ایجاد تفکری جدید نسبت به داده است. در داده‌کاوی معنایی، داده‌ها باید در سطحی از هوشمندی قرار گیرند تا توسط ماشین قابل درک باشند. بنابراین هستان‌شناسی توانایی کمک به فرآیند داده‌کاوی معنایی را از طریق معناشناختی موجود در هستان‌شناسی فراهم می‌کند. درسال‌های اخیر، استفاده از داده-کاوی معنایی[2] با حمایت هستان‌شناسی[3]، و داده‌کاوی توزیع شده با پشتیبانی سامانه‌های چند عاملی[4] به یک حرکت بزرگ در روند داده‌کاوی و مدیریت داده‌های بزرگ تبدیل شده است [۶]. هستان‌شناسی می‌تواند فرآیند کاوش را کنترل و فضای پرس‌وجو را کاهش و یا وادار به کاهش نماید [۷]. سامانه چند عاملی نیز، راه‌حل‌هایی برای حل مسئله تعامل و همکاری در محیط‌های توزیع شده ارائه می‌دهد[۸]. از هستان‌شناسی و عامل‌ها می‌توان در داده‌کاوی برای کاوش قواعد هم‌آیی، طبقه‌بندی، خوشه‌بندی، استخراج اطلاعات و سامانه‌های توصیه‌گر استفاده کرد [۵، ۹ و ۱۰]. اکنون مفاهیم اصلی تجزیه و تحلیل داده‌های بزرگ مقیاس، شدیداً بر روی مفاهیم داده‌کاوی و کاوش‌معنایی متمرکز شده است. در این زمینه مفاهیم مرتبط با موضوع داده‌های بزرگ مقیاس، متشکل از ۱) بعد داده، ۲) بعد زیرساخت فناوری اطلاعات، ۳) بعد روش و ۴) بعد کاربرد است [۱۱]. تجزیه و تحلیل داده‌های بزرگ مقیاس با استفاده از سه مفهوم داده، پردازش و مدیریت در سه بعد اقتصادی، فناوری و قانونی که کلید اصلی تجزیه و تحلیل داده‌های بزرگ را در دست دارند انجام می‌شود [۳]. شکل (۱) مرور کلی تجزیه و تحلیل داده‌های بزرگ را نشان می‌دهد.



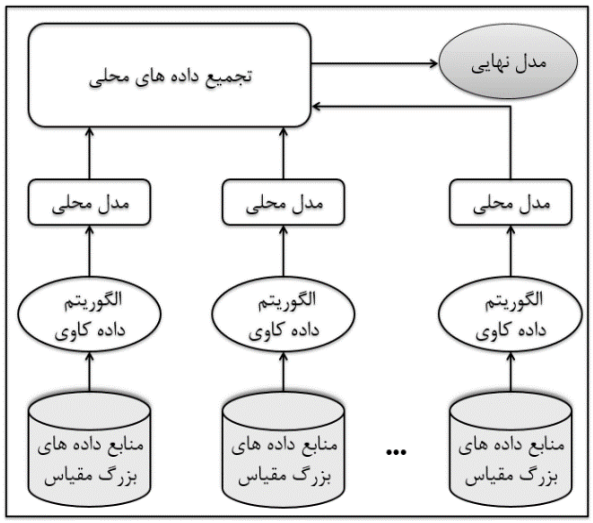
شکل (۱): مرور کلی تجزیه و تحلیل داده‌های بزرگ [۳] .

۲-۱- داده‌کاوی توزیع شده

داده‌کاوی توزیع شده، شاخه‌ای از داده‌کاوی است که چارچوبی برای کاوش داده‌های توزیع شده ارائه می دهد. داده‌کاوی توزیع شده ناشی از نیاز به کاوش روی منابع داده‌ای غیر متمرکز، بدون توجه به مکان فیزیکی آن‌ها است. پردازش در گره‌های توزیع شده متفاوت، چندین مدل محلی ناشی از الگوریتم داده‌کاوی را تولید کرده و سپس به شکل یک مدل سراسری تجمیع شده، دانش سراسری را ارائه می‌کند [۱۲ و ۱۳]. شکل (۲) چارچوب عمومی داده‌کاوی توزیع شده را نشان می‌دهد.

چارچوب داده‌کاوی توزیع شده از داده‌های بزرگ مقیاس، نیازمند یک سامانه تجزیه و تحلیل داده‌ای کامل است که باید مراحل اکتساب، اجتماع، تجزیه و تحلیل و کنش را به‌عنوان فرآیند 4A's برای تجزیه و تحلیل داده‌های بزرگ مقیاس فراهم نماید.

[1] Agent-Based Architecture for Semantic Mining of Large-Scale Data in Distributed Environment (ASMLDE)
[2] Semantic Data Mining
[3] Ontology
[4] Multi Agent Systems



شکل (۲): چارچوب داده‌کاوی توزیع شده [۱۲-۱۴]

۲-۲- هستان‌شناسی

با توجه به تعریف گروبر[1]، هستان‌شناسی مشخصاتی از یک مفهوم و یک دیدگاه منسجم از اطلاعات مدیریت شده در قالب یک مجموعه و به صورت یک لیست صریح و سازمان یافته از تمام اصطلاحات، روابط و اشیائی است که یک دامنه را نشان می‌دهد.

- هستان‌شناسی به‌طور رسمی توسط رابطه $O=(S.L)$ تعریف می‌شود که S ساختار مفهومی مجموعه‌ای از نمونه‌های مفاهیم و $\underline{L}$[2] ساختار واژگانی است. ساختار مفهومی $\underline{S}$ توسط رابطه (۱) تعریف می‌شود [۵ و ۱۵].

$$S = (C.R. \leq . \sigma_R) \quad (۱)$$

  - $C.R$: مجموعه های غیر مجرد، شامل مفاهیم و روابط وابسته هستند. یعنی (C) مجموعه‌ای از مفاهیم و (R) مجموعه‌ای از نام‌های رابطه (روابط کوتاه) نام‌گذاری می‌شود.
  - $\leq_C$: سلسله مراتب مفاهیم را تعریف می‌کند.
  - $\sigma_R: R \rightarrow CxC$: امضای یک رابطه بین مفهوم است.

- ساختار واژگانی $\underline{L}$ شامل تمام برچسب‌هایی است که با مفاهیم و ارتباط مفهوم هستان‌شناسی مرتبط هستند. این تعریف توسط رابطه (۲) بیان می‌شود.

$$L = (L_C, L_R, F_C, F_R) \quad (۲)$$

تابع مفاهیم یک تابع تعریف شده روی مجموعه‌ای از مفاهیم است که توسط رابطه (۳) توصیف می‌شود.

$$\forall l \in L_C . F_C(l) = \{c/c \in C\} \quad (۳)$$

تابع رابطه یک تابع تعریف شده روی مجموعه‌ای از ارتباط‌ها است که توسط رابطه (۴) توصیف می‌شود.

$$\forall l \in L_R . F_R(l) = \left\{\frac{r}{r} \in C\right\} \quad (۴)$$

توابع بیان شده، امکان دسترسی به مفاهیم و روابط تعیین شده توسط یک برچسب را فراهم می کند.

۲-۳- سامانه چند عامله

سامانه چند عامله[3] به‌عنوان یک سامانه کلان متشکل از عوامل مستقل است و اغلب به‌عنوان یک الگو برای طراحی برنامه‌های کاربردی پیچیده به‌کار برده می‌شود. سامانه‌های چندعامله، شکل یافته عامل‌های متعددی است که در تعامل با یکدیگر تشکیل شده است. به‌طور رسمی و دقیق سامانه‌های چندعامله به‌صورت چهارتایی زیر تعریف می‌شود:

$$MAS = <Ag, Env, Org, D> \quad (۵)$$

Ag مجموعه عامل‌های تشکیل دهنده سامانه چندعامله می‌باشد که توسط رابطه (۶) تعریف می‌شود.

$$Ag = \{Ag_1, Ag_2, \ldots, Ag_n\} \quad (۶)$$

Env محیطی است که سامانه چندعامله در آن عمل می‌کند. Org سازمان سامانه چندعامله و D بیان کننده قلمرو سامانه چندعامله است.

۳- کارهای مرتبط

بسیاری از ابزارهای تجزیه و تحلیل داده کنونی که در [۱۶ و ۱۷] بیان شده است، جهت مقابله با چالش‌های در حال اجرا، نتایج رضایتبخش مورد نظر کاربر را با هزینه قابل قبول ارائه می‌دهد. به‌عنوان مثال، ابزار تجزیه و تحلیل کاملاً خودکار بهینه شده مبتنی بر الگوریتم ترکیبی ویژگی عمیق و ابزار تجزیه و تحلیل کاملاً خودکار معرفی شده در مرجع [۱۸] است. اما همچنان جنبه‌های جدیدی از تجزیه و تحلیل داده‌های بزرگ مقیاس مد نظر محققان و پژوهشگران است. در نتیجه، شناخت چالش‌های پیش رو تمدید شده و مرحله نوینی از تجزیه و تحلیل داده‌های بزرگ مقیاس مطرح شده است [۱۹]. در ادامه مهم‌ترین پژوهش‌های انجام شده مرتبط با تحقیق حاضر بیان شده است. در مرجع [۲۰]، چارچوبی برای ارزیابی خودکار قابلیت اطمینان فرآیند کشف دانش ارائه شده است. در این روش، فرآیند کشف دانش بهینه شده و دانش با کیفیت بالاتر تولید می‌شود. در پژوهش [۵]، قاهرا و همکاران استفاده از چارچوب نگاشت-کاهش مبتنی بر هستان‌شناسی را برای کاوش قواعد هم‌آیی[4] از

[1] Grouber
[2] Lexical Structure
[3] Multi Agent System (MAS)
[4] Associaton Rules

داده‌های حجیم ارائه داده‌اند. هستی‌شناسی‌ها اجازه می‌دهد که قواعد هم‌آیی تولید شده را فیلتر کرده و فقط آن‌هایی که مفید هستند را نگهداری کرد. در مقاله [۲۱]، چارچوب ارزیابی برای هستی‌شناسی، توسط سارکی جین و والری مایریک ارائه شده است. روش‌های مورد استفاده در این چارچوب، شامل دو مرحله تأیید و اعتبارسنجی است. تأیید هستان‌شناسی تضمین می کند که هستان‌شناسی به درستی ساخته شده است و اعتبار آن تضمین می کند که هستان‌شناسی برای کاربرد معین شده به صورت صحیح ساخته شده است. در تحقیق [۲۲]، سرجی نادال و همکاران، یکپارچه‌سازی هستان‌شناسی داده‌های بزرگ و یک الگوریتم بازنویسی پرس‌وجو را ارائه داده‌اند که با استفاده از حاشیه نویسی هستان‌شناسی، پرسش‌های روی هستان‌شناسی را به پرس‌وجوی روی منابع تبدیل می‌کند. پژوهش [۱۴]، به رویکردها، روش‌ها و راه حل‌های داده‌کاوی توزیع شده برای کاهش هزینه محاسبات و نیز حفظ حریم خصوصی داده‌ها با توزیع مناسب منابع در سایت‌های توزیع شده می‌پردازد. در مرجع [۲۳] کالیانی و همکاران ، رویکردی برای داده‌کاوی مجموعه اقلام مکرر در محیط هادوپ را از لنز داده‌های بزرگ مقیاس بیان کرده و به سه حوزه چالشی مقیاس پذیری حافظه، تقسیمبندی کارها و متعادل‌سازی بار برای طراحی الگوریتم‌های کاوش مجموعه اقلام مکرر می‌پردازد. در پژوهش [۳]، الهادی بلغاش و همکاران به سامانه‌های چند عاملی برای مدل‌سازی، شبیه‌سازی و حل مشکلات در سامانه‌های پیچیده پرداخته‌اند. هدف این تحقیق بررسی و توصیف چگونگی استفاده از چنین فناوری‌هایی برای داده‌های بزرگ در قالب یک سامانه چند عاملی سازگار[1] است که قابلیت‌های تجزیه و تحلیل پویا را ارائه می‌دهد. این تحقیق در حال حاضر در پروژه نئو کامپوس، دانشگاه تولوز فرانسه استفاده می‌شود. در مطالعه [۲۴]، ونشنگ‌گان و همکاران، مروری بر روش‌های پیشرفته داده‌کاوی توزیع شده، از جمله کاوش مجموعه اقلام مکرر توزیع شده، کاوش توالی مکرر توزیع شده، کاوش مکرر گراف توزیع شده، خوشه بندی توزیع شده، و نگهداری حریم خصوصی داده‌های توزیع شده ارائه شده است. در [۲۵]، مولود باراتیا و همکاران بر این باورند وب معنایی فرصت‌های جدیدی برای تحقیقات داده‌کاوی فراهم می‌کند. داده‌های معنایی وب معمولاً در قالب سه گانه[2] فرمت توصیف منابع[3] شامل موضوع، پیش فرض و شیء نشان داده می‌شوند. پایگاه‌های دانش گسترده به سبک آر - دی- اف شامل صدها میلیون از سه گانه آر - دی - اف است که دانش را در فرمت قابل فهم برای ماشین نشان می‌دهند. این تحقیق رویکردی به نام کاوش قواعد هم‌آیی وب معنایی[4] را معرفی می‌کند که به‌صورت خودکار قواعد هم‌آیی معنایی را از داده‌های آر - دی - اف کاوش می‌کند. در تحقیق [۲۶]، توسط سوربی گوپال آتال و دکتر پی. ان. چاتور پیشنهاد شده است که از ماشین‌های چندگانه برای پردازش داده‌ها و از روش خوشه‌بندی برای جدا کردن با دقت داده‌ها و تولید نتیجه در زمان کمتر استفاده شود. بنابراین داده‌ها را با رویکرد افزایشی و توزیع شده در اکوسامانه هادوپ تقسیم کرده و با کمک تابع نگاشت – کاهش، برنامه‌ها بر روی خوشه‌ها اجرا می‌شوند. در مقاله [۸]، خانم روپالی چیکاله، الگوریتم‌های داده‌کاوی توزیع شده، روش‌ها و روندها را برای کشف دانش از داده‌های توزیع شده در یک روش مؤثر و کارآمد مورد بحث قرار می‌دهد. در مقاله [۲۷]، بهاراتی و همکاران به بحث در مورد توسعه و گسترش روش کاوش قواعد هم‌آیی با هستان‌شناسی به نام آرمو[5] می‌پردازد. آرمو برای یافتن دقیق‌ترین قواعد هم‌آیی مبتنی بر هستان‌شناسی و مجموعه‌های اقلام مکرر مورد استفاده قرار گرفته است. مقاله [۱۲]، به تجزیه و تحلیل تطبیقی چارچوب‌های مبتنی بر عامل برای استخراج قواعد هم‌آیی از منابع داده توزیع شده می‌پردازد. همچنین به معرفی الگوریتم کاوش قواعد هم‌آیی توزیع شده به نام دارم[6] می‌پردازد. هدف این الگوریتم، تولید قواعد از مجموعه داده‌های مختلف در سراسر سایت‌های مختلف جغرافیایی گسترش یافته است. راندمان الگوریتم DARM بسیار وابسته به توزیع داده است. در تحقیق [۷]، پرفسور پریتی-وی- بهاگات و پروفسور پالاس-ام- گورشتتیوار به مرور رویکردهای مبتنی بر هستان‌شناسی برای داده‌کاوی معنایی پرداخته و نقش معانی رسمی در هستان‌شناسی را به روش استخراج اطلاعات پیوند می‌دهد. در پژوهش [۲۸]، الگوریتمی به نام پیما[7] را با استفاده از استراتژی‌های تقسیم‌بندی برای کاوش قواعد هم‌آیی توزیع شده برای مواجه با چالش‌های زمان پاسخ و هزینه ارتباطی بالا ارائه شده است. سیدهنت پاتیل و همکاران در مرجع [۲۹]، به معرفی و مرور سامانه‌های داده‌کاوی عامل‌گرا پرداخته و نقش عامل‌گرایی را در مواجهه و غلبه بر چالش‌های پیش روی داده‌کاوی در یک محیط ناهمگن توزیع شده بیان می‌کنند. مقاله [۳۰]، ضمن بحث در مورد مزایا و اشکالات ارزیابی هستان‌شناسی، به مسئله یافتن یک روش ارزیابی کارآمد هستان‌شناسی پرداخته و روش‌های ارزشیابی را در چهار دسته مبتنی بر استاندارد طلایی، مبتنی بر پیکر، مبتنی بر وظیفه و مبتنی بر معیارها و شاخص‌ها ارائه داده است. مطالعات انجام شده در مرجع [۳۱]، نیز به دو جنبه مهم برای ارزیابی بهتر هستان‌شناسی (صحت و کیفیت) پرداخته و به تمایز بین این دو

[1] Multi-Agent System Adaptive
[2] Triple
[3] Resource Description Format (RDF)
[4] Semantic Web Association Rule Mining
[5] Association Rule Minig with Ontolog
[6] Didtributed Assocation Rule Mining
[7] PEYMA

به‌عنوان راهی برای ارزیابی بهتر هستان‌شناسی پرداخته شده است. همچنین این پژوهش به رویکردها، روش‌ها، معیارها، اندازه‌گیری‌ها و ارزیابی کیفی در هستان‌شناسی می‌پردازد. پژوهش‌های [۳۲ و ۳۳] به استانداردسازی پردازش‌های کیفی خودکار و تفسیرهای پیچیده‌تر و همچنین یکپارچگی مفهوم داده‌های حجیم و داده‌های ضخیم پرداخته شده است. در این دو پژوهش، به چگونگی تجزیه و تحلیل ترکیبی در بهبود تصمیم‌گیری مؤثر پرداخته شده است.

۱-۳- جمع‌بندی کارهای مرتبط

بررسی مطالعات بیان شده نشان می‌دهد، برخی از روش‌ها و پژوهش‌های بیان شده بر مداخله کاربران تأکید بیشتری دارند که به علت حجم زیاد داده‌ها، زمان‌گیر و مستعد خطا هستند. فعالیت‌های پژوهشی انجام شده بر موضوعات ۱) استراتژی‌های تکرار داده برای مدیریت داده‌ها، بهره‌وری و بهبود کارایی دسترسی و زمان پاسخگویی و افزایش قابلیت اطمینان در سامانه‌های داده‌ای توزیع شده ۲) میان‌افزارها و مدیریت منابع، مانند ارتباطات، کشف و زمان‌بندی منابع، امنیت، دسترسی به داده‌ها و تشخیص خطا در محیط محاسباتی توزیع شده ۳) معرفی محیط‌های محاسباتی در پشتیبانی از داده‌کاوی توزیع شده ۴) الگوریتم‌های داده‌کاوی و کشف دانش به روش متمرکز و توزیع شده تمرکز دارد. اما هدف مقاله حاضر، پاسخ به سه سؤال بیان شده زیر در قالب معماری پیشنهادی است.

۱) چگونه می‌توان کاوش از داده‌های بزرگ مقیاس را با استفاده از فناوری معنایی به ویژه هستان‌شناسی متناسب با اهداف کسب‌وکار برای تولید داده‌های ضخیم بهبود بخشید؟

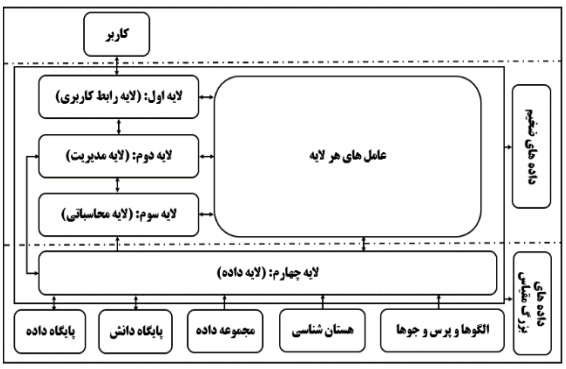
۲) چگونه می‌توان داده‌کاوی و استخراج دانش مفید را از منابع محلی با داده‌های بزرگ مقیاس به سطح سراسری تعمیم داد؟

۳) چگونه می‌توان از عامل‌ها به‌عنوان الگوی محاسباتی جدید در حمایت و بهبود مسائل داده‌کاوی در محیط‌های توزیع شده استفاده کرد؟

بنابراین تحقیق حاضر، راه‌حل خود را برای پاسخ به سؤالات مطرح شده در قالب یک معماری ائتلافی[1] متشکل از ۱) روش‌های مبتنی بر پرس‌وجو (دریافت پرسش از کاربر و پردازش معنایی آن، ۲) روش‌های مبتنی بر پردازش محلی توسط هر گره شبکه، ۳) روش‌های مبتنی بر داده‌کاوی چند عاملی، ۴) روش‌های مبتنی بر هستان‌شناسی ارائه خواهد داد.

۴- کلیات معماری پیشنهادی

به‌منظور تحقق اهداف بیان شده در این پژوهش، از روش معماری عامل‌گرا مبتنی بر سبک معماری لایه‌ای، روش تحلیل و طراحی سامانه‌های چند عاملی و متدلوژی‌های توسعه مهندسی دانش استفاده می‌شود. معماری ASMLDE از نوع ترکیبی غیر متمرکز و مؤلفه‌های آن نیز بر مبنای معماری انتزاعی چهار لایه و عامل‌های تشکیل دهنده آن نیز در هر لایه سازماندهی و تعریف شده است. لایه‌های این معماری در مسیر پاسخ به سؤالات طرح شده برای (۱) چگونگی کاهش فضای کاوش و بالا بردن سطح انتزاع داده جهت شناسایی دانش‌های مهم (۲) تولید دانش مؤثر، معنی‌دار و با کیفیت حاصل از کاوش گره‌های محلی در سطح سراسری (۳) افزایش اعتبار نتایج عمومی داده‌کاوی توزیع شده حاصل از ادغام دانش محلی و استخراج دانش‌های سراسری پیش‌بینی شده است. در معماری ASMLDE برای تجزیه و تحلیل داده‌های بزرگ و دستیابی به داده‌های ضخیم از حمایت داده‌کاوی توزیع شده با تأکید بر فناوری چند عاملی و فناوری معنایی با تأکید بر هستان‌شناسی استفاده می‌شود. معماری ASMLDE از چهار لایه متمایز تشکیل‌شده است. ۱) لایه رابط کاربری، مسئول تأمین تعاملات کاربر است، ۲) لایه مدیریت، تمام پروسه حل مسئله را مدیریت و کنترل می‌کند، ۳) لایه محاسباتی، تأمین کننده عملیات داده‌کاوی و هستان‌شناسی مورد نیاز برای دستیابی به داده‌های ضخیم است، ۴) لایه داده، تأمین کننده انواع داده‌های مورد نیاز لایه محاسباتی (مجموعه داده، پایگاه داده و غیره) از منابع داده‌ای بزرگ مقیاس است. شکل (۳) لایه‌های چهارگانه این معماری را به تفکیک نشان می‌دهد.



شکل (۳): لایه‌های معماری پیشنهادی

۴-۱- لایه رابط کاربری

وظیفه‌ای که توسط کاربر باید انجام شود در یک زبان قابل فهم برای لایه رابط کاربر تعریف می‌شود. این لایه توسط عامل رابط کاربر مدیریت می‌شود. کاربر با در اختیار داشتن مجموعه‌ای از

[1] Coalition Architecture

رابط‌های کاربری (مانند یک پوسته و یا یک رابط کاربری گرافیکی[1]) درخواست‌های خود را توسط یک عامل به لایه محاسباتی ارسال کند. رابط کاربر وظیفه درخواست شده را که در زبان معنایی قابل فهم انسان به هستان‌شناسی است، ترجمه کرده و به لایه محاسباتی انتقال می‌دهد. زبان پرس‌وجوی مبتنی بر هستان‌شناسی می‌تواند در شکل‌گیری و اجرای الزامات کسب‌وکار توسط پرس‌وجوهای ارسال‌شده از سوی کاربر کمک کند. اگر چه می‌توان از هستان‌شناسی برای حمایت و پشتیبانی از تمام مراحل داده‌کاوی و کشف دانش استفاده کرد.

۴-۲- لایه مدیریت

با توجه به توزیع‌شدگی محیط محاسباتی در معماری ASMLDE، مدیریت تبادلات و فعالیت‌های هر لایه به ویژه فرآیند محاسباتی A's 4 توسط عامل‌های هر لایه کنترل، هدایت و مدیریت می‌شود. لایه مدیریت به‌عنوان یک پیوند دهنده بین رابط کاربر و لایه محاسباتی عمل می‌کند. همچنین بهترین روش از روش‌های کاوش موجود را انتخاب، نتایج را جمع‌آوری و اطلاعات آماری مورد نیاز را فراهم می‌کند. این لایه تمام فرآیند حل مسئله را از طریق عامل‌ها کنترل می‌کند.

۴-۳- لایه محاسباتی

لایه محاسباتی متشکل از روش‌های متنوع داده‌کاوی است و توسط عامل محاسباتی خود مدیریت می‌شود. در این لایه فعالیت‌های محاسباتی مربوط به داده‌کاوی و کاوش معنایی مبتنی بر هستان‌شناسی کسب‌وکار انجام می‌شود. لایه محاسباتی از ویژگی دستکاری و بازیابی الگو (ذخیره‌سازی و پرس‌وجوی الگو) پشتیبانی می‌کند. همچنین با دخالت هستان‌شناسی، عملیات کاوش و فیلتر کردن الگوها و خودکارسازی مرحله ارزیابی الگوها متناسب با کسب‌وکار بهبود یافته و از یک زبان (زبان نشانه‌گذاری توسعه‌پذیر[2]) برای ذخیره الگوها و پرس‌وجوها استفاده می‌شود. عملیات داده‌کاوی در این لایه متشکل از روش‌های داده‌کاوی مختلفی است که توسط موتور داده‌کاوی انجام می‌شود. این لایه تأمین کننده داده‌های ضخیم در معماری ASMLED است.

۴-۴- لایه داده

خدمات مدیریت و دسترسی به داده‌های ذخیره شده و همچنین عامل‌های مورد استفاده برای خواندن داده‌ها از منابع متنوع داده‌ای، لایه داده را تشکیل می‌دهد. این لایه شامل پایگاهی از الگوهای متنوع داده‌ای است. اولین پایگاه شامل یک یا چندین پایگاه محلی که داده‌های خام بزرگ مقیاس از جمله مجموعه داده‌های[3] مختلف را در خود ذخیره می‌کند. دومین پایگاه یک یا چندین پایگاه دانش از جمله هستان‌شناسی‌ها را شامل می‌شود که ساختارهای پیچیده و اطلاعات غیر ساخت یافته را در خود نگهداری می‌کند. سومین پایگاه شامل یک پایگاه سراسری است که پرس‌وجوها و الگوهای کاوش شده از داده‌های خام محلی را در خود ذخیره می‌کند.

۵- معماری پیشنهادی

معماری پیشنهاد شده، متشکل از اجزاء و مؤلفه‌هایی است که در محیط تجزیه و تحلیل داده برای حل یک مسئله مشترک با هم تعامل می‌کنند. تعامل سامانه‌های چند عاملی، فناوری معنایی و هستان‌شناسی، نیازهای اصلی معماری پیشنهادی برای ارائه یک معماری عامل‌گرا برای کاوش معنایی از داده‌های بزرگ مقیاس در محیط‌های توزیع شده محسوب می‌شود. با توجه به اهداف طرح شده در این مقاله، تمرکز ما بیشتر بر دستیابی به جنبه‌های منطقی و معنایی مدیریت داده‌ها به‌جای جنبه‌های فیزیکی آن است. اهمیت معماری پیشنهادی این است که نشان می‌دهد، اجرای فرآیندها از نظر منطقی و عملیاتی وجود دارد و می‌توان بدون بروز اختلالی (مثلاً از بین رفتن گره‌های داده‌ای و محاسباتی)، ساختار معنایی را برای داده‌کاوی توزیع شده با پشتیبانی عامل‌ها و هستان‌شناسی‌ها، طراحی، اجرا، پردازش، ذخیره‌سازی و بازیابی کرد. همچنین این فعالیت‌ها می‌تواند توسط ده‌ها عامل محاسباتی دیگر و در سطح رایانه کاربر نیز انجام پذیرد.

در معماری پیشنهادی برای جمع‌آوری و استانداردسازی پردازش‌های کیفی خودکار و تفسیرهای پیچیده‌تر، از ایده و مفهوم‌سازی کارماسفره با فرآیند A's 4، تاگ لانه‌ئی با مدل 3 V's که به بینش از حجم و مقیاس داده‌ها و تریسیا وانگ و لاتزکوتوت که به بینش از ضخامت داده‌ها اشاره دارد، استفاده شده است. داده‌های ضخیم با استفاده از رویکرد کیفی خود از داده‌های بزرگ مقیاس متفاوت است. بر خلاف داده‌های بزرگ مقیاس که از رویکرد کمی استفاده می‌کنند، اطلاعات ضخیم، ابعاد کیفی داده‌ها را ارائه می‌دهد. داده‌های بزرگ مقیاس حقایق را تعریف می‌کند، و داده‌های ضخیم آن‌ها را تشریح و توضیح می‌دهد. همچنین داده‌های ضخیم، معنا و ادراکی را که در فرایند تصمیم‌گیری مورد نیاز است، فراهم می‌کند. داشتن یک تصویر کامل از داده‌های بزرگ مقیاس و داده‌های ضخیم بسیار مهم است، زیرا آن‌ها بینش‌های مختلفی را در مقیاس و عمق داده‌های مختلف تولید می‌کنند.

$$\text{تصویر کامل از داده} = f\left(\text{داده بزرگ مقیاس .داده ضخیم}\right)$$

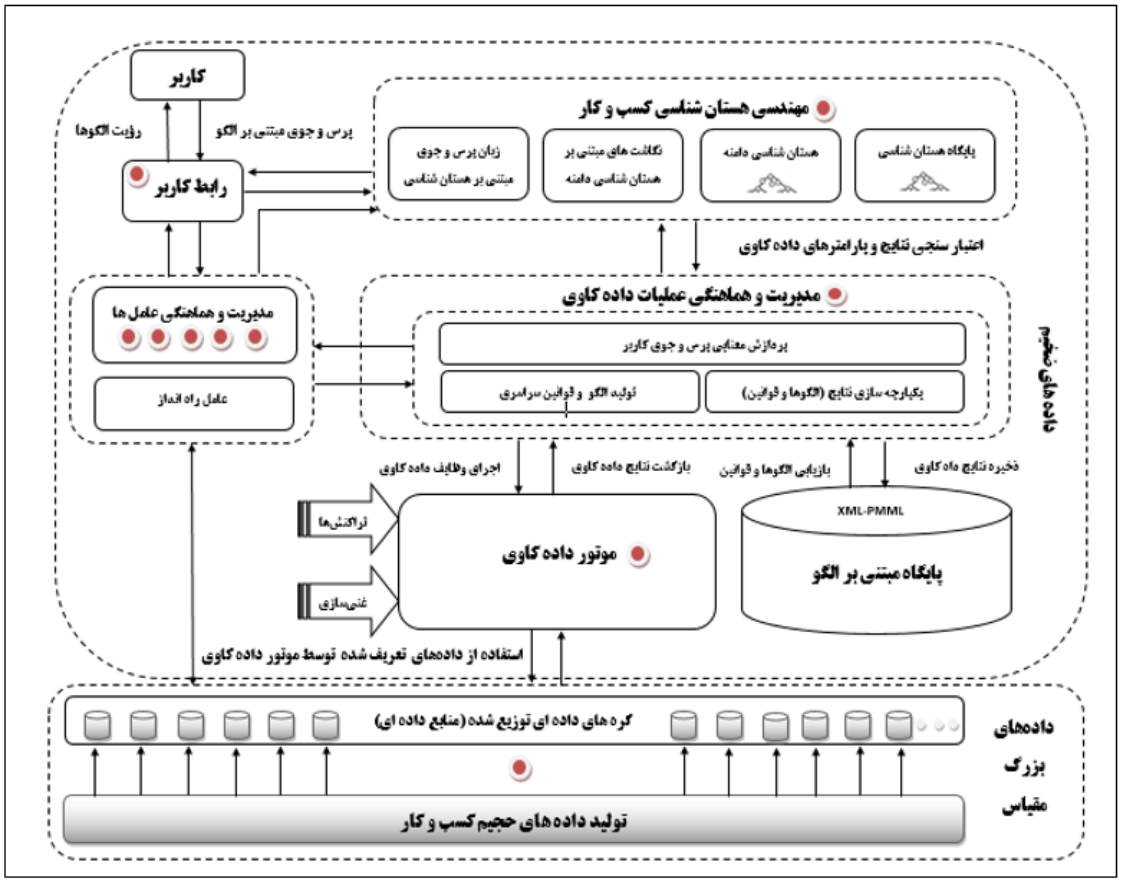
بنابراین در معماری پیشنهادی، داشتن تصویر کاملی از داده‌ها در قالب داده‌های بزرگ و داده‌های ضخیم برای کاوش معنایی نیاز

[1] Graphic User Interface
[2] Extensible Markup Language
[3] Datasets

است. همچنین این معماری با یک راهبرد ائتلافی و ترکیبی ابعاد داده‌های بزرگ مقیاس و داده‌های ضخیم را تحت عنوان (تجزیه و تحلیل مخلوط)[1] باهم ترکیب نموده است. به‌گونه‌ای که برای داده‌کاوی معنایی یکدیگر را تکمیل کنند. شکل (۴) معماری پیشنهادی را به همراه مؤلفه‌های آن نشان می‌دهد.

۵-۱- مؤلفه‌های معماری پیشنهادی

هدف اصلی معماری ASMLED، توسعه یک روش کاربردی جدید بر اساس فناوری‌های معنایی برای کاوش و ادغام داده‌ها از طریق همگرایی هستان‌شناسی، سنجش کیفیت داده‌ها و ادغام با اهداف کسب‌وکار است. همچنین پی‌ریزی یک روش مؤثر و خودکار برای داده‌کاوی از جمله تولید اقلام مکرر و قواعد هم‌آیی با کیفیت و مرتبط از داده‌های بزرگ مقیاس با فیلتر کردن الگوهای اختلالی با پشتیبانی هستان‌شناسی کسب‌وکار و فناوری عامل در یک محیط محاسباتی عامل‌گرای توزیع شده است. معماری پیشنهادی از هفت مؤلفه ۱) مهندسی هستان‌شناسی کسب‌وکار، ۲) رابط کاربر، ۳) مدیریت و هماهنگی عملیات داده‌کاوی، ۴) مدیریت و هماهنگی بین عامل‌ها، ۵) موتور داده‌کاوی، ۶) پایگاه مبتنی بر الگو، ۷) گره‌های داده‌ای توزیع شده (منابع داده‌ای) تشکیل می‌شود. در معماری پیشنهادی دسترسی به داده‌ها بر اساس سه جزء اصلی است. اولین جزء هستان‌شناسی است که رسیدگی به چالش‌های معنایی ارائه شده توسط مجموعه داده‌های بزرگ و یک دیدگاه منسجم از اطلاعات مدیریت شده را فراهم می‌کند. دومین جزء منابع داده‌ای است که مستقل، توزیع شده و ناهمگون هستند. در سامانه‌های دسترسی به داده مبتنی بر هستان‌شناسی، یک هستان‌شناسی برای آشکارسازی اطلاعات به شیوه‌ای مفهومی، با انتزاع از جزئیات سطح داده استفاده می‌شود. به همین سبب در معماری پیشنهادی دسترسی به داده مبتنی بر هستان‌شناسی است. سومین جزء نگاشت(ها) است که داده‌های موجود در منابع داده‌ای را با هستان‌شناسی کسب‌وکار پیوند می‌دهد. ارتباط بین منابع داده‌ای و هستان‌شناسی از طریق نگاشت انجام می‌شود. در ادامه مؤلفه‌های بیان شده در معماری پیشنهادی تشریح خواهد شد.



شکل (۴): معماری پیشنهادی

[1] Mixed Analytics

۵-۱-۱- مهندسی هستان‌شناسی کسب‌وکار

مؤلفه مهندسی هستان‌شناسی کسب‌وکار برای هدایت و ارزش‌گذاری رسیدن به اهداف کسب‌وکار تعریف شده و عملیات داده‌کاوی را حمایت خواهد کرد. مهندسی هستان‌شناسی ادغام داده‌ها از منابع داده‌ای ناهمگن و هستان‌شناسی دامنه، افزایش کارایی، دقت و کیفیت جستجوی معنایی را فراهم می‌کند. وجود این مؤلفه برای درک معانی کسب‌وکار الزامی و برای مسیریابی و جهت دادن به روند داده‌کاوی هدفمندانه پیش‌بینی شده است. توجه به هدف داده‌کاوی و کسب‌وکار، حجم قواعد حاصل از عملیات داده‌کاوی را کاهش داده و در نتیجه، اهداف داده‌کاوی و کسب‌وکار متناسب با نیاز کاربر به شکل مناسب‌تری محقق خواهد شد. هستان‌شناسی را می‌توان به سه روش مختلف در فرآیند عملیات داده‌کاوی دخالت داد. روش اول در مرحله پیش‌پردازش (برای آماده کردن داده برای کاوش)، روش دوم در طول فرایند داده‌کاوی و پرس‌وجوی مبتنی بر الگو و در روش سوم پس از فرآیند داده‌کاوی (برای ارزیابی الگوهای استخراج شده) است. معماری پیشنهادی از هر سه روش برای کاهش فضای کاوش، هرس معنایی و بهبود ارزیابی الگوهای استخراج شده از منابع داده‌ای مختلف استفاده نموده و دانش دامنه توسط هستان‌شناسی پشتیبانی خواهد شد.

۵-۱-۲- رابط کاربر

در این مؤلفه، وظیفه‌ای که باید انجام شود در قالب یک زبان قابل فهم پرس‌وجوی معنایی توسط کاربر تعریف می‌شود. برای این منظور، کاربر درخواست خود را با استفاده از یک زبان مبتنی بر هستان‌شناسی (مانند زبان اسپار-کیو-ال) همراه با انتخاب دامنه هستان‌شناسی مناسب که می‌تواند در شکل‌گیری و اجرای الزامات کسب‌وکار توسط پرس‌وجوها مؤثر واقع شود تعریف می‌کند. این درخواست توسط عامل رابط کاربر و وظایفی را که در زبان قابل فهم انسان به هستان‌شناسی توسط عامل هستان‌شناسی تعریف و حمایت شده است ترجمه و به مدیریت داده‌کاوی ارسال می‌کند. لازم به ذکر است، دسترسی سنتی به داده‌ها اغلب نیاز به پرس‌وجو و دستورالعمل پیچیده‌ای (مانند زبان SQL) دارد که برای متخصصان غیر آشنا به داده، درک آن بسیار سخت است. اما زبان پرس‌وجوی مبتنی بر هستان‌شناسی، همراه با هستان‌شناسی دامنه ارائه شده در معماری می‌تواند در شکل‌گیری و اجرای الزامات کسب‌وکار توسط پرس‌وجوهای ارسال شده از سوی کاربر کمک کند. اگر چه می‌توان از هستان‌شناسی ایجاد شده برای حمایت و پشتیبانی از تمام مراحل داده‌کاوی و کشف دانش استفاده کرد. در معماری پیشنهادی از یک سامانه چند عاملی برای پشتیبانی از عملیات داده‌کاوی (محلی/ توزیع شده) استفاده خواهد شد.

۵-۱-۳- مدیریت و هماهنگی عامل‌ها

این مؤلفه یک فرآیند تعاملی را برای حل مسئله همکاری بین عامل‌ها در محیط‌های توزیع شده با استفاده از زبان‌های ارتباطی عامل (مانندKQML یا FIPA ACL) فراهم می‌کند. این مؤلفه تعامل شش عامل تعریف شده شامل ۱) عامل هستان‌شناسی، ۲) عامل رابط کاربر، ۳) عامل مدیریت و هماهنگی عامل‌ها، ۴) عامل مدیریت و هماهنگی داده‌کاوی، ۵) عامل‌داده‌کاوی و ۶) عامل داده را فراهم می‌کند. وظایف و کاربرد هر یک از عامل‌های شش‌گانه در جدول (۱) بیان شده است. رفتار عامل‌ها به داده‌های مربوط به منابع توزیع شده بستگی دارد. در کاربردهای چند عاملی، عامل‌ها باید فعال و مستقل بوده و محیط خود را درک و به‌طور پویا اقدامات را بر اساس شرایط و ارتباطات مورد نیاز محیط برقرار کنند. همچنین عملیات داده‌کاوی در یک محیط ناهمگن توزیع شده با استفاده از عامل‌های داده‌کاوی انعطاف‌پذیرتر، سازگارتر، قوی‌تر، آسان‌تر و دقیق‌تر می‌شود. بنابراین باید همکاری کاملی بین عامل‌ها و سایر مؤلفه‌ها معماری انجام شود.

جدول (۱): عامل‌های شش‌گانه در معماری ASMLED.

نام عامل	وظایف و کاربرد عامل‌ها
عامل هستان‌شناسی	مسئول تعیین، انتخاب و تحویل هستان‌شناسی متناسب با اهداف کسب‌وکار برای پرس‌وجوی مبتنی بر الگوی ورودی کاربر است.
عامل رابط کاربر	مسئول ارتباط و تعامل با کاربر است.، که شامل پذیرش کار انجام شده به‌عنوان ورودی و ارائه نتایج به‌عنوان خروجی است.
عامل مدیریت و هماهنگی عامل‌ها	مسئول هماهنگی بین سایر عامل‌ها و تکمیل درخواست کاربر است که با اختصاص دادن وظیفه به عامل‌های مختلف، آن‌ها را از طریق عامل راه‌انداز تحریک به اقدام می‌نماید.
عامل مدیریت و هماهنگی داده‌کاوی	مسئول دریافت نتایج کاوش از سایت‌های مختلف، تجمیع الگوهای بازگشت داده شده توسط موتور داده‌کاوی، ارزیابی الگوهای نهایی، ارائه و نمایش الگوهای نهایی استخراج شده را عهده‌دار است.
عامل داده‌کاوی	الگوریتم کاوش را پیاده‌سازی و روش کاوش را بر اساس اطلاعات ارائه شده از قبیل الزامات روش، فرم داده ورودی و غیره، آغاز می‌کند.
عامل داده	مسئول عرضه داده‌ها از منابع مختلف و گره‌های داده‌ای به عامل کاوش است. همچنین عامل داده وظیفه تعیین مسیر و دسترسی به منابع داده‌ای را فراهم می‌کند.

۵-۱-۴- مدیریت و هماهنگ کننده عملیات داده‌کاوی

این مؤلفه برای مدیریت و هماهنگی کاوش، ذخیره‌سازی، پرس‌وجو، نمایه‌سازی و به‌روزرسانی الگوها ارائه شده است. این بخش اگر چه مستقل از مهندسی هستان‌شناسی کسب‌وکار و موتور داده‌کاوی پیش‌بینی شده است، اما در ارتباط و تعامل با این دو است، و قابلیت همکاری، تعامل و تبادل الگو بین منابع داده‌ای مختلف را پشتیبانی و از زبان نشانه‌گذاری توسعه‌پذیر برای ذخیره الگوها در پایگاه الگو استفاده می‌کند. در این بخش، الگوها بر اساس دانش دامنه و هستان‌شناسی کسب‌وکار استخراج می‌شود. بنابراین الگوهای کاوش شده مفید، با اهمیت و مرتبط با کسب‌وکار هستند. کاوش‌های مبتنی بر هستان‌شناسی به استخراج ارزش از داده‌ها و تصمیم‌گیری بهتر در کسب‌وکار کمک می‌کند. این مؤلفه به‌منظور ارائه یک روش مؤثر برای مدیریت حجم بالای الگوها از روش‌های خلاصه‌سازی، یکسان‌سازی و یکپارچه‌سازی (مانند کاهش قواعد هم‌آیی تولید شده غیر مفید) استفاده می‌کند.

۵-۱-۵- موتور داده‌کاوی

در معماری پیشنهادی، موتور داده‌کاوی با تعامل با مدیریت و هماهنگ کننده عملیات داده‌کاوی و بهره‌گیری از روش‌های توصیفی به کاوش الگوها از منابع داده‌ای می‌پردازد. برای این اقدام، الگوریتم داده‌کاوی توزیع شده، به متمرکز نمودن زیرمجموعه‌ای از اقلام داده‌های محلی برای ساختن الگوی سراسری از اقلام و به حداقل انتقال داده برای موفقیت الگوریتم داده‌کاوی توزیع شده نیازمند است.

۵-۱-۶- گره‌های داده‌ای توزیع شده

نقطه شروع همه حرکت‌ها در داده‌کاوی متمرکز یا توزیع شده، منابع داده‌ای متمرکز یا توزیع شده است. بنابراین هماهنگ کننده عملیات داده‌کاوی از منابع داده‌ای به‌صورت محلی و یا سراسری متناسب با روش‌های انتخاب شده عملیات داده‌کاوی و بر اساس نیاز هستان‌شناسی و پرس‌وجوی کاربر استفاده می‌کند.

۵-۲- مراحل عملکرد معماری پیشنهادی

مراحل ده‌گانه معرفی شده در جدول (۲)، تعاملات بین مؤلفه‌ها را در زیرساخت محاسباتی توزیع شده و محیط محاسباتی عامل‌گرا نشان می‌دهد. هدف از بیان این مراحل، کاوش اقلام مکرر و قواعد هم‌آیی معنایی توزیع شده از داده‌های بزرگ مقیاس به صورت خودکار با حمایت هستان‌شناسی دامنه و سامانه‌های چند عاملی در محیط داده‌کاوی توزیع شده است. مراحل و اقدام هر مرحله در ادامه ارائه شده است.

جدول (۲): مراحل کاوش اقلام مکرر و قواعد هم‌آیی معنایی در معماری پیشنهادی (ASMLED).

مراحل	اقدام هر مرحله از عملکرد معماری پیشنهادی
۱	کاربر درخواست خود را از طریق واسط کاربری و در قالب یک مجموعه ویژگی بیان می‌کند.
۲	واسط کاربر درخواست را به عامل هماهنگ کننده در محیط محاسباتی عامل‌گرا ارسال می‌کند.
۳	عامل هماهنگ‌کننده سؤال کاربر را به یک ساختار معنایی در قالب زبان اسپار- کیو- ال (یک زبان پرس‌وجو خاص برای نمایش داده‌های آر - دی - اف است) تبدیل و به کارگزار هستان‌شناسی ارسال می‌کند.
۴	کارگزار هستان‌شناسی، لیست موقتی از تراکنش‌های متناسب با درخواست کاربر را تولید و به‌عنوان نتیجه به عامل هماهنگ کننده در محیط محاسباتی عامل‌گرا بازگشت می‌دهد.
۵	لیست تراکنش‌های تولید شده از اجرای مرحله ۴، فاقد مدل داده‌ای مناسب برای عملیات داده‌کاوی کاوش اقلام مکرر است، بنابراین نیازمند پاکسازی و آماده‌سازی مجدد است و باید از ساختار تراکنشی مناسبی برای کاوش توسط الگوریتم داده‌کاوی برخوردار شود. بنابراین لیست داده‌ای دریافت شده به یک ساختار توصیفی مناسب تبدیل شده و مدل داده‌ای مناسب برای کاوش از روی آن ساخته می‌شود.
۶	در این مرحله درخواست کاربر در قالب یک مدل داده‌ای جدید برای انجام عملیات داده‌کاوی به یک زیر ساخت محاسباتی که متشکل از یک خوشه محاسباتی است ارسال می‌شود.
۷	موتور داده‌کاوی با اجرای الگوریتم داده‌کاوی (مانند آپ- ریو-ری / رشد اقلام تکراری) بر روی مدل داده‌ای توصیفی اخذ شده از مرحله ۵، مجموعه قلام مکرر را تولید می‌کند.
۸	زیر ساخت محاسباتی، مدل داده‌ای تولید شده مبتنی بر ساختار اقلام مکرر مرحله ۶ را دریافت و با اجرای ادامه الگوریتم داده‌کاوی، قوانین هم‌آیی معنایی را تولید می‌کند. نتایج این مرحله به عامل هماهنگ کننده تحویل داده می‌شود.
۹	عامل هماهنگ کننده، قواعد هم‌آیی معنایی تولید شده را جهت تحویل به کاربر به عامل واسط کاربر تحویل می‌دهد.
۱۰	پردازش‌های خفیف‌تر که به‌عنوان پسا پردازش شناخته می‌شود (مانند مرتب سازی و یا ارائه با فرمت خاصی مانند csv.) در این مرحله انجام می‌شود.

۶- ارزیابی معماری پیشنهادی

در این بخش، برای پاسخ به سؤالات تحقیق، معماری پیشنهادی بر اساس روش‌های ارزیابی کیفی و کمی مبتنی بر متدلوژی‌های عامل‌گرا (MASE، PASS، ASPECS) و متدلوژی ترکیبی ARA که برای آن

استفاده از هستان‌شناسی تعریف گردیده به صورت مقایسه‌ای با سایر معماری‌های موجود مورد ارزیابی قرار می‌گیرد. این ارزیابی با تأکید بر جنبه‌های ۱) فرآیند و چرخه حیات، ۲) پوشش مفاهیم، ۳) زبان مدل‌سازی، ۴) عملیاتی بودن ۵) ابزار مدل‌سازی، ۶) ابزار پیاده‌سازی و ۷) جنبه حمایتی متدلوژی‌های عامل‌گرا انجام شده است. در ارزیابی تجربی نیز، بر اساس معماری ASMLED یک زیرساخت محاسباتی توزیع شده و عامل‌گرا پیاده‌سازی و با اعمال هستان‌شناسی بر روی مجموعه داده‌ای زمین لرزه از پایگاه دانش DBpedia با استفاده از معیارهای "حداقل آستانه پشتیبان[1]" و "حداقل آستانه اطمینان[2]"، اقلام مکرر استخراج و کاوش قواعد هم‌آیی معنایی در گام‌های مختلفی تولید و ارزیابی خواهد شد.

۶-۱- ارزیابی کیفی مبتنی بر عامل

این مرحله از ارزیابی به مقایسه معماری پیشنهادی و سایر معماری‌های مشابه مبتنی بر عامل که به‌طور خاص به کاوش اقلام مکرر و کشف قواعد هم‌آیی توزیع شده می‌پردازد اشاره دارد. جدول (۲) مقایسه کیفی مبتنی بر عامل در معماری پیشنهادی ASMLDE با سایر پژوهش‌های انجام‌شده را نشان می‌دهد. در این جدول، مقایسه کاربرد معماری مبتنی بر عامل به همراه نوع و تنوع عامل‌ها بیان شده است. تمام پژوهش‌های مقایسه شده (MADKDS، MADARM، AFARMDD) با معماری ASMLDE از پروژه‌های پژوهشی دانشگاهی محسوب می‌شوند.

۶ -۲- ارزیابی کیفی مبتنی بر ویژگی

در این مرحله از ارزیابی، معیارها و شاخص‌های مختلفی از معماری‌ها و چارچوب‌های مختلف بررسی و ده معیار ارزیابی استخراج گردید. جدول (۳) به معرفی ویژگی‌های ده‌گانه استخراج شده از چارچوب‌های مورد مطالعه و پنج معیار ارزیابی اضافه شده در قالب معماری ASMLED می‌پردازد. ده ویژگی ابتدایی از مجموعه پژوهش‌های انجام شده در معماری‌های مختلف استخراج شده است. پنج ویژگی پایانی بر اساس نیاز مؤلفه‌های معماری پیشنهادی برای ارزیابی در جدول ویژگی‌های مقایسه وارد شده است. در ادامه ارزیابی کیفی مبتنی بر ویژگی به شیوه مقایسه‌ای، معیارهای ارزیابی پانزده گانه تعیین شده برای کاوش قواعد هم‌آیی معنایی در محیط‌های توزیع‌شده برای چارچوب‌های مورد بررسی و مطالعه شده MADKDS، MADARM و FARMDD و همچنین معماری پیشنهادی ASMLED در قالب جدول (۴) توصیف، بررسی و مقایسه شد. ارزیابی مجموعه معیارهای بیان شده، نشان می‌دهد چارچوب‌های عامل‌گرای AFARMDD و MADARM بر اساس برنامه حرکت موازی و سریال چند عاملی استوار است. در حالی که MADKDS از برنامه حرکت موازی و تنها ASMLDE از برنامه حرکت موازی و توزیع شده استفاده می‌کند. همچنین چارچوب MADKDS و معماری ASMLDE دارای یک محیط اجرای عامل است و سایر چارچوب‌ها هیچ محیطی برای آزمایش و اعتبارسنجی ندارند. الگوریتم Apriori در همه چارچوب‌ها برای کاوش مجموع اقلام مکرر استفاده می‌شود. فقط MADARM در مورد مدل هزینه مربوط به کاوش قواعد هم‌آیی بحث می‌کند و سایر چارچوب‌ها فاقد مدل هزینه‌ای هستند. سازوکار حفظ حریم خصوصی برای داده‌های حساس محلی در AFARMDD و ASMLDE وجود دارد در حالی که سایر چارچوب‌ها چنین سازوکاری ندارند. هیچ یک از این چارچوب‌ها به غیر از ASMLED رابط کاربری گرافیکی طراحی شده برای کار با سامانه‌ها نداشته و در کاربردهای واقعی، پروژه‌های توسعه و مطالعات موردی و غیره مورد استفاده قرار نمی‌گیرد. در خصوص سایر معیارها، تنها معماری ASMLDE از هستان‌شناسی برای پرس‌وجو و کاوش قواعد معنایی متناسب با اهداف کاربر برخوردار است. همچنین فقط در این معماری از قابلیت اطمینان در مراحل مختلف داده‌کاوی برای اندازه‌گیری کیفیت مبتنی بر شاخص‌ها و معیارها و اعتبارسنجی تجربی با مجموعه داده‌های آزمایشی یا واقعی و داشتن مطالعه موردی برای اعتبارسنجی تجربی استفاده می‌شود.

۶ -۳- ارزیابی کیفی مبتنی بر هستان‌شناسی

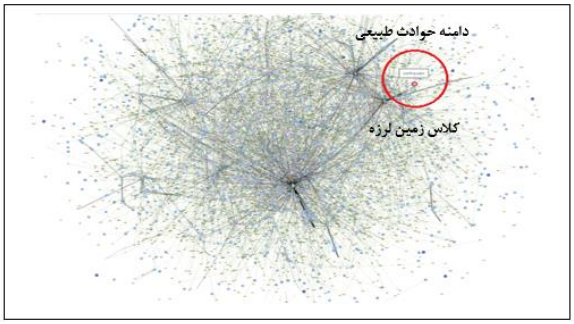
تأکید در ارزیابی کیفی هستان‌شناسی، اطمینان از درستی طراحی و تعریف مفاهیم و ارتباطات معنایی در مؤلفه مهندسی هستان‌شناسی است. برای این ارزیابی، دو معیار اصلی در نظر گرفته شده است. معیار اول سازگاری است. به این مفهوم که عدم وجود تناقض در کلاس‌ها یا عناصر هستان‌شناسی وجود داشته باشد و داده‌های ارائه شده، نتیجه‌های متناقض را ارائه ندهند. معیار دوم کامل بودن است. به‌طوری که هستان‌شناسی هیچ خطایی در حوزه کامل بودن نداشته باشد. جدول (۵) به معیارهای هفت‌گانه برای ارزیابی و اعتبارسنجی کیفی هستان‌شناسی در این معماری اشاره می‌کند. در این تحقیق برای پیاده‌سازی و ارزیابی تجربی از یک هستان‌شناسی با معیارهای بیان شده استفاده شده است. اگر چه برای ارزیابی تجربی، فعالیت‌های مختلفی برای آماده‌سازی و تحقق معیارهای هستان‌شناسی انجام شده است.

۷- ارزیابی تجربی

برای ارزیابی روش پیشنهادی، سناریویی متشکل از یک پیکربندی با ۴ گره محاسباتی (هر یک با ظرفیت پردازشی ۴ هسته ۱۴گیگاهرتز،

[1] Min-Support
[2] Min-Confidence

۱۵۰ گیگابایت ظرفیت ذخیره‌سازی و ۱۰ گیگابایت ظرفیت حافظه اصلی) در نظر گرفته شد. هر گره محاسباتی مبتنی بر چارچوب محاسباتی هادوپ[1] نسخه ۲/۷/۳ و بستر محاسباتی اسپارک[2] نسخه ۲/۴ به‌منظور ایجاد زیرساخت ذخیره‌سازی، محاسباتی و پردازشی و از محیط عامل‌گرای جید[3] به‌منظور پیاده‌سازی عامل‌ها و ساخت محیط چند عاملی راه‌اندازی گردید. هر ۴ گره نقش عامل‌های کارگر و عامل داده را ایفاء می‌کنند. مجموعه داده آزمایشی و هستان‌شناسی معتبر بر اساس معیارهای بیان شده در جدول (۷) انتخاب گردید. در این ارزیابی، اخذ داده‌ها از پایگاه دانش معنایی تا تولید قواعد هم‌آیی معنایی در سه لایه انجام می‌شود. داده‌های معنایی اولیه، رکوردهایی هستند که از داده‌های پیوندی استخراج شده و در قالب سه‌گانه RDF/n3 در پایگاه دانش قرار دارند. به این منظور، مجموعه داده[4] پایگاه دانش معتبر دی بی پدیا[5] نسخه ۲۰۱۷ از انتشار پایدار[6] ایجاد شده در پروژه ویکی‌پدیا و قابل دسترس در وب جهانی به حجم بیش از ۳۰ گیگابایت حاوی میلیون‌ها داده سه‌گانه هستان‌شناسی در قالب RDF/N3 شامل توصیفات متنوع و مختلفی درباره مشاهدات از مفاهیم و پدیده‌های مختلف (از جمله زمین لرزه) به همراه ۵ گیگابایت ایندکس و ۲/۵ مگابایت فایل OWL برای شناخت کلاس‌ها، زیرکلاس‌ها، ارتباطات و جزئیات استفاده از مجموعه داده آزمایشی پایگاه دانش دی بی پدیا پیش پردازش و آماده شد. همچنین فایل هستان‌شناسی دی بی پدیا با استفاده از نرم‌افزار پروتیج[7] قرائت گردید. شناسایی متریک‌ها متشکل از Class، Axiom، Object، Property و روابط موجود در این هستان‌شناسی و متناسب با نیاز معماری آماده‌سازی و پیکربندی گردید. برای این ارزیابی، داده‌های پایگاه دانش دی بی پدیا برای هستان‌شناسی دامنه حوادث طبیعی و کلاس زمین لرزه انتخاب گردید. شکل (۵) کلاس هستان‌شناسی زمین لرزه در پایگاه دانش دی بی پدیا را نشان می‌دهد.

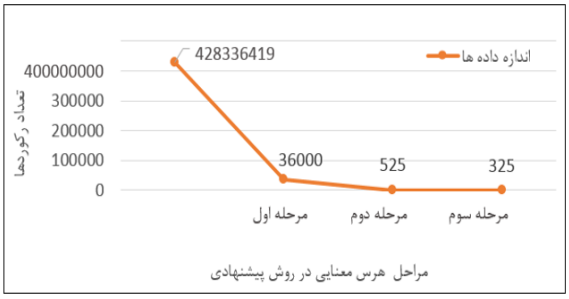


شکل (۵): کلاس هستان‌شناسی زمین لرزه.

[1] Hadoop
[2] Spark
[3] Jade
[4] Dataset
[5] DBpedia
[6] Stable Release
[7] Protégé

اجرای عملیاتی روش پیشنهادی در سناریویی به تعداد ۴۲۸۳۳۶۴۱۹ رکورد از پایگاه دانش دی بی پدیا با صدور پرس‌وجوی معنایی به زبان SPARQL از سوی کاربر آزمون گردید. کوچک سازی فضای کاوش متناسب با هستان‌شناسی دامنه، نیاز کاربر، سطح داده و صفات مرتبط با اهداف کاوش در سه مرحله از هرس معنایی انجام و اقلام مکرر و قواعد هم‌آیی معنایی تولید گردید. در ادامه به این سه مرحله اشاره شده است.

- **هرس اول معنایی:** در هرس اول، پس از اعمال هستان‌شناسی محدود به کلاس زمین لرزه، رکوردهای متناسب با درخواست کاربر فیلتر گردید.
- **هرس دوم معنایی:** در هرس دوم، رکوردهای ناشی از اعمال هستان‌شناسی با مجموعه ویژگی‌های مورد نظر در پرس‌وجوی معنایی کاربر تطبیق داده شد و رکوردهای مبتنی بر هستان‌شناسی زمین لرزه فیلتر گردید.
- **هرس سوم معنایی:** در هرس سوم، رکوردهای استخراج شده در هرس دوم به ساختار تراکنشی قابل محاسبه توسط الگوریتم دادکاوی (FP-Growth) تبدیل و رکوردهای فیلتر شده جهت اجرای الگوریتم آماده گردید. شکل (۶) تعداد رکوردهای ناشی از سه مرحله هرس معنایی و میزان کوچک سازی فضای کاوش را نشان می‌دهد.



شکل (۶): تعداد رکوردها در سه مرحله هرس معنایی.

در ادامه، ارزیابی روش پیشنهادی بر اساس معیارهای مهم ارزیابی، شامل معیار پشتیبان و معیار اطمینان انجام می‌شود.

۱. **معیار پشتیبانی قواعد:** احتمال رخداد آیتم‌های موجود در قواعد در یک موقعیت زمانی است که سودمند بودن[8] قواعد را بیان می‌کند.

$$\text{Support}(A \rightarrow B) = P(A \cap B) \quad (۷)$$

$$\text{Support}(A \rightarrow B) = \text{Support}(B \rightarrow A) \quad (۸)$$

۲. **معیار اطمینان قواعد:** احتمال رخداد آیتم‌های تالی به شرط رخداد آیتم‌های مقدم است که یقین و قطعیت[9] قواعد را بیان می کند.

[8] Usefulness
[9] Certainty

$$\text{Confidence } (A \rightarrow B) = P(B \mid A) \quad (۹)$$

$$\text{Confidence } (A \rightarrow B) = P(B \mid A) = \frac{Support(A \cap B)}{Support\ (A)} \quad (۱۰)$$

موتور داده‌کاوی الگوریتم FP-Growth را در ۹ وضعیت مختلف اجرا کرده و با استخراج اقلام مکرر و قواعد هم‌آیی معنایی در گره‌های محلی، قواعد هم‌آیی معنایی سراسری را تولید می‌کند. جدول (6) وضعیت‌های مختلف در ارزیابی سناریو را نشان می‌دهد.

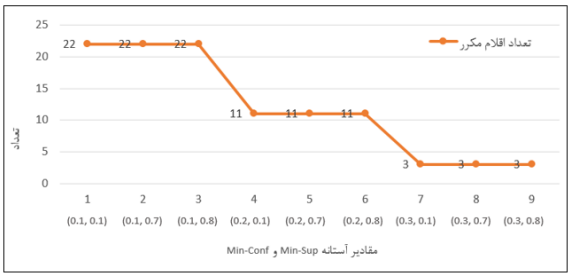
جدول (۲): مقایسه مبتنی بر عامل در معماری ASMLED با سایر پژوهش‌های انجام شده [۱۴-۱۲].

ردیف	نام پژوهش	کاربرد	نوع و تنوع عامل‌ها
۱	**MADKDS** **Mobile Agent-Based Distributed KNowledge Discovery System**	در این چارجوب سامانه کشف دانش توزیع شده مبتنی بر عامل سیار ارائه شده است.	عامل سیار کاوش داده: با استفاده از یک الگوریتم افزایشی برای کاوش مجموعه اقلام‌های مکرر محلی برای تولید پایگاه دانش محلی و بازگرداندن این دانش به مدیر فرآیند کاوش محصور شده است.
			عامل سیار پیش پردازش داده: این عامل برای پیش‌پردازش داده‌های محلی و جمع‌آوری داده‌ها در انبار داده مرکزی است.
			عامل سیار شمارش: برای اسکن پایگاه داده‌های محلی و جمع‌آوری تعداد پشتیبانی برخی از مجموعه اقلام است.
۲	**MADARM** **Multi Agent Distribution Association Rule Miner**	در این چارچوب مدل‌های هزینه نظری را برای کاوش قواعد هم‌آیی مبتنی بر عامل برای داده‌های توزیع شده با استفاده از یک مدل نمونه اولیه پیشنهاد می‌کند.	عامل هماهنگی کاوش قواعد هم‌آیی: برای ایجاد و هماهنگی سایر عامل‌ها در ناحیه مربوط به عامل است.
			کاوشگر قواعد هم‌آیی مبتنی بر عامل سیار: برای انجام وظایف کاوش قواعد هم‌آیی به‌کار می‌رود.
			گزارشگر نتایج مبتنی بر عامل سیار: برای انتقال نتیجه به عامل‌های دیگر به‌کار می‌رود.
			عامل هماهنگی ادغام نتایج: برای ادغام نتایج یا دانش بهینه از روی منابع داده‌ای برای یکپارچه سازی دانش مبتنی بر عامل توزیع شده است.
۳	**AFARMDD** **An Agent-Based Framework for Association Rules Mining of Distributed Data**	در این پژوهش یک چارچوب مبتنی بر عامل برای استخراج قواعد انجمنی‌توزیع شده پیشنهاد شده است. هدف اصلی این مطالعه حفاظت از حریم خصوصی داده‌های محلی از قرار گرفتن در معرض دیگر سایت‌های توزیع شده و کپسوله‌سازی روش‌های موجود ارائه شده است.	عامل پیوند رمز نگاری امن: این عامل برای کپسوله کردن عملیات داده‌کاوی و رمزگذاری عملیات تجمیع امن در هر سایت است.
			عامل پیوند رمز گشایی امن: این عامل برای محاسبه رمزگشایی عملیات تجمیع امن در هر سایت است.
			عامل جمع رمزنگاری: این عامل برای کپسوله کردن رمزگذاری عملیات تجمیع امن در هر سایت است.
			عامل جمع رمزگشایی: این عامل برای محاسبه رمزگشایی عملیات جمع‌آوری امن در هر سایت است.
			عامل انتشار: این عامل برای انتقال مجموعه اقلام مکرر مرتبه K سراسری به هر سایت است.
			عامل مافوق: برای اطلاع دادن به همه سایت‌ها برای خاتمه عملیات کاوش است.
۴	**ASMLDE** **(معماری پیشنهادی)**	در این تحقیق یک معماری عامل‌گرا برای کاوش معنایی از داده‌های بزرگ مقیاس در محیط‌های توزیع شده ارائه شده است.	عامل رابط کاربر: وظیفه گرفتن درخواست‌های ورودی از کاربر و ارائه نتایج نهایی (مانند توصیف اقلام مکرر و یا قواعد انجمنی) به کاربر است.
			عامل مدیریت و هماهنگی بین عامل‌ها: مسئولیت هماهنگی و توزیع وظایف بین عامل‌های مختلف را عهده‌دار است.
			عامل کارگزار هستان‌شناسی: مسئولیت تعیین و ارائه دامنه هستان‌شناسی به سامانه است.
			عامل مدیریت و هماهنگی عملیات داده‌کاوی: تولید اقلام مکرر و قواعد هم‌آیی سراسری از داده‌های محلی و توزیع شده بر روی گره‌های داده‌ای و تعامل با موتور داده‌کاوی و ذخیره‌سازی‌های سامانه را انجام می‌دهد.
			عامل داده‌کاوی: انجام عملیات داده‌کاوی محلی برای استخراج اقلام مکرر و قواعد هم‌آیی و تحویل به عامل هماهنگی و مدیریت داده‌کاوی را انجام می‌دهد.
			عامل داده: تحویل داده‌ها از منابع مختلف به عامل کاوش را فراهم می‌کند.

جدول(۳): معیارهای استخراج شده از پژوهش‌های انجام شده و اضافه کردن معیارهای جدید در معماری ASMLED [۱۲ و ۲۰].

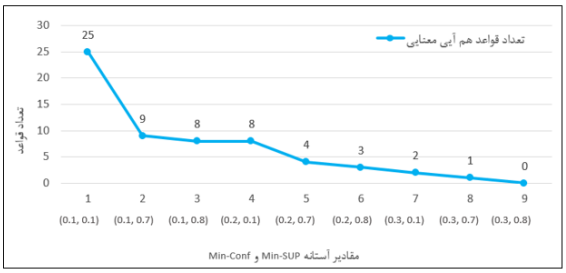
ردیف	معیار ارزیابی	کاربرد معیار
۱	عامل‌ها	زمینه عامل‌ها، اجتماع عامل‌های درگیر در سامانه را نشان می‌دهد.
۲	خط سیر/ برنامه حرکت	موازی یا سریال بودن برنامه حرکت عامل‌ها را به وسیله عامل‌های سیار نشان می‌دهد.
۳	محیط اجرای عامل/ بستر عامل سیار	نشان می‌دهد که کدام محیط اجرای عامل یا بستر عامل سیار برای توسعه سامانه چند عاملی برای کاوش قواعد هم‌آیی توزیع شده استفاده می‌شود.
۴	پیاده‌سازی	نشان می‌دهد که آیا سامانه چند عاملی همراه با زبان مورد استفاده برای اجرا پیاده‌سازی شده است یا فقط یک چارچوب نمونه است.
۵	الگوریتم	نشان‌دهنده الگوریتم مطرح‌شده در مطالعه مانند کاوش قواعد هم‌آیی یا کاوش مجموعه اقلام مکرر است.
۶	مدل هزینه	نشان می‌دهد که آیا مدل هزینه در این مطالعه مورد بحث است.
۷	حفظ حریم خصوصی	نشان می‌دهد که آیا سازوکار حفظ حریم خصوصی برای داده‌های حساس محلی مورد توجه است.
۸	واسط کاربر گرافیکی	برای ویژگی رابط کاربر گرافیکی سامانه چند عاملی است.
۹	مجموعه داده	نشان می‌دهد که آیا هر مجموعه داده (مصنوعی یا واقعی) در اعتبارسنجی آزمایشی استفاده می‌شود.
۱۰	قابل استفاده	نشان می‌دهد از سامانه چند عاملی در برنامه‌های عملی، پروژه‌های توسعه، مطالعات موردی و غیره استفاده می‌شود.
۱۱*	هستان‌شناسی	نشان می‌دهد در معماری ASMLED از هستان‌شناسی برای استخراج اقلام مکرر و کاوش قواعد هم‌آیی معنایی در فرآیند داده‌کاوی استفاده می‌شود.
۱۲*	کاوش معنایی	نشان می‌دهد کاوش معنایی از طریق زبان پرس‌وجوی معنایی متناسب با نیاز و اهداف کاربر در معماری ASMLED انجام می‌شود.
۱۳*	قابلیت اطمینان	قابلیت اطمینان با استفاده از اندازه‌گیری کیفیت مبتنی بر شاخص‌ها و معیارها در سطح داده‌ها در مراحل مختلف فرآیند داده‌کاوی در معماری پیشنهادی انجام می‌شود و قابلیت اطمینان از قابلیت اطمینان اجزای معماری ASMLED محاسبه می‌شود.
۱۴*	اعتبارسنجی تجربی	نشان می‌دهد از یک مجموعه داده‌های آزمایشی یا واقعی با اندازه‌های بزرگ برای اعتبارسنجی معماری ASMLED استفاده می‌شود.
۱۵*	مطالعه موردی	نشان می‌دهد استفاده از مطالعه موردی خاص برای اعتبارسنجی تجربی معماری ASMLED انجام شده است.

شکل (۷)، نتایج حاصل از اجرای سناریوی ارزیابی را مطابق وضعیت‌های تعیین شده برای کاوش اقلام مکرر توسط الگوریتم FP-Growth در مرحله سوم هرس معنایی نشان می‌دهد.



شکل (۷): نتایج حاصل از اجرای مرحله سوم هرس معنایی.

در این مرحله، هدف تولید کمترین قواعد و بهترین توصیف از داده‌ها و عملکرد روش پیشنهادی است. شکل (۸)، قواعد هم‌آیی معنایی تولید شده را در وضعیت‌های مختلف سناریو نشان می‌دهد.



شکل (۸): تعداد قواعد هم‌آیی تولید شده.

جدول (۴): مقایسه کیفی ویژگی‌های معماری ASMLDE و چارچوب‌های مطالعه شده [۱۲-۱۴ و ۲۰].

معیار ارزیابی	چارچوب کاوش قوانین هم‌آیی توزیع شده			
	MADKDS	**AFARMDD**	**MADARM**	**ASMLDE** معماری پیشنهادی
نام و تعداد عامل‌ها در چارچوب‌های مطالعه شده و معماری پیشنهادی	DMMA,PMA,CMA	ESUA,DSUA,ESA, DSA, BA,OA	ARMCA, MAARM, MARR,RICA	UIA, AMCA, OA,DMCMA,DMA,DRA
برنامه حرکت	موازی	سریال و موازی	سریال و موازی	موازی و توزیع شده
محیط اجرای عامل/ عامل سیار	بلی	بلی	خیر	بلی
پیاده‌سازی	بلی	بلی	خیر	بلی
الگوریتم	IAA	Apriori	Apriori -FP-Growth	Apriori - FP-Growth
مدل هزینه	خیر	خیر	بلی	خیر
حفظ حریم خصوصی	خیر	بلی	خیر	بلی
واسط کاربر گرافیکی	خیر	خیر	خیر	بلی
مجموعه داده	خیر	خیر	خیر	بلی
قابل استفاده	خیر	خیر	خیر	بلی
هستان‌شناسی	خیر	خیر	خیر	بلی
کاوش معنایی	خیر	خیر	خیر	بلی
قابلیت اطمینان	خیر	خیر	خیر	بلی
اعتبارسنجی تجربی	خیر	خیر	خیر	بلی
مطالعه موردی	خیر	خیر	خیر	بلی

جدول (۵): معیارهای ارزیابی و اعتبارسنجی هستان‌شناسی در معماری ASMLED [۳۰ و ۳۱].

ردیف	معیار ارزیابی	کاربرد معیار ارزیابی هستان‌شناسی	ردیف	معیار ارزیابی	کاربرد معیار ارزیابی هستان‌شناسی
۱	سازگاری	عدم وجود تضاد بین مفاهیم ارائه‌شده در هستان‌شناسی است.	۵	عمومیت	هستان‌شناسی را می‌توان برای بیش از یک هدف دریک دامنه استفاده کرد
۲	کامل بودن	مفاهیم باید به‌طور کامل تعریف‌شده و هیچ مفهومی از دست نرود.	۶	استحکام	هستان‌شناسی می‌تواند تغییرات آینده را در دامنه پشتیبانی کند.
۳	شفافیت	هیچ اطلاعات غیرضروری نباید در هستان‌شناسی وجود داشته باشد.	۷	غنای‌اطلاعات معنایی	هستان‌شناسی در مفهوم‌سازی غنی و متنوع است.
۴	وضوح	باید شامل مفاهیمی باشد که به‌طور مؤثر دامنه دانش مربوطه را ارائه دهد.			

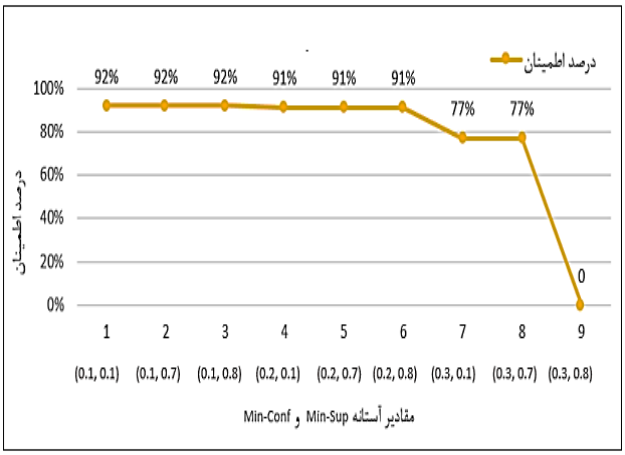
جدول(۶):وضعیت‌های تعریف شده در اجرای سناریو.

درصد اطمینان	تعداد قواعد هم‌آیی معنایی	تعداد اقلام مکرر	مقادیر آستانه		وضعیت
			اطمینان	پشتیبان	
٪۹۲	۲۵	۲۲	۰.۱	۰.۱	۱
٪۹۲	۹	۲۲	۰.۷	۰.۱	۲
٪۹۲	۸	۲۲	۰.۸	۰.۱	۳
٪۹۱	۸	۱۱	۰.۱	۰.۲	۴
٪۹۱	۴	۱۱	۰.۷	۰.۲	۵
٪۹۱	۳	۱۱	۰.۸	۰.۲	۶
٪۷۷	۲	۳	۰.۱	۰.۳	۷
٪۷۷	۱	۳	۰.۷	۰.۳	۸
۰	۰	۳	۰.۸	۰.۳	۹

از آنجایی که معیار پشتیبان، کاربردی بودن قواعد و معیار اطمینان میزان قابلیت اطمینان قواعد را نشان می‌دهد، شکل(۹)، درصد اطمینان قواعد تولید شده را نشان می‌دهد. در ادامه، توصیف قواعد هم‌آیی تولید شده بر اساس توصیف کمیت‌های فرض شده در جدول (۷) انجام شده است. برای بیان دقیق‌تر روش پیشنهادی، نمونه‌ای از قواعد هم‌آیی معنایی ناشی از اجرای سناریوی بیان شده، به همراه درصد اطمینان هر قاعده استخراج شده در جدول (۸) نشان داده شده است. جدول (۹) نیز توصیف نمونه‌ای از قواعد هم‌آیی تولید شده مبتنی بر کلاس هستان‌شناسی زمین‌لرزه را بیان می‌کند.

جدول (۷): توصیف کمیت‌های فرض شده.

ردیف	عنوان توصیف	توصیف کمیت‌های فرض شده
	اندازه زمین‌لرزه	
۱	ضعیف	۱ تا ۲ ریشتر
	متوسط	۳ تا ۵ ریشتر
	قوی	۶ تا ۸ ریشتر
	فاجعه‌آمیز	۹ تا ۱۱ ریشتر
	تلفات زمین‌لرزه	
۲	کم	کمتر از ۱۰۰ نفر
	متوسط	بین ۱۰۰ تا ۱۰۰۰ نفر
	زیاد	بین ۱۰۰۰ تا ۱۰۰۰۰ نفر
	فاجعه‌آمیز	بیشتر از ۱۰۰۰۰ نفر



شکل (۹): اطمینان قواعد هم‌آیی تولید شده.

جدول (۸): نمونه‌ای نمونه‌ای از قواعد هم‌آیی معنایی تولید شده

ردیف	تعداد رکورد انتخاب‌شده مبتنی بر هستان‌شناسی	تعداد تراکنش‌ها پس از پیش‌پردازش	حداقل درجه پشتیبانی	حداقل درجه اطمینان	تعداد قواعد هم‌آیی استخراج‌شده	درصد اطمینان	شماره قواعد هم‌آیی	قوانین هم‌آیی معنایی تولید شده مبتنی بر هستان شناسی دامنه زمین‌لرزه
۱	۵۲۲	۳۲۵	۰,۱	۰,۷	۸	۰,۸۲	۱	[Season->Spring] => [Magnitude->Strong]
						۰,۸۶	۲	[casualties->Moderate] => [Magnitude->Strong]
						۰,۸۲	۳	[Season->Summer, casualties->Few] => [Magnitude->Strong]
						۰,۹۱	۴	[Season->Winter] => [Magnitude->Strong]
						۰,۷۸	۵	[Season->Fall] => [Magnitude->Strong]
						۰,۹۲	۶	[casualties->huge] => [Magnitude->Strong]
						۰,۷۷	۷	[casualties->Few] => [Magnitude->Strong]
						۰,۸۸	۸	[Season->Summer] => [Magnitude->Strong]
۲	۵۲۲	۳۲۵	۰,۲	۰,۷	۴	۰,۸۲	۱	[Season->Spring] => [Magnitude->Strong]
						۰,۹۱	۲	[Season->Winter] => [Magnitude->Strong]
						۰,۷۷	۳	[casualties->Few] => [Magnitude->Strong]
						۰,۸۸	۴	[Season->Summer] => [Magnitude->Strong]

ارزیابی‌های انجام شده، بیان کننده موفقیت روش پیشنهادی در پاسخ به سؤالات تحقیق است. کاهش فضای کاوش، استخراج قواعد هم‌آیی در سطح گره‌های محلی و تولید قواعد هم‌آیی سودمند متناسب با پرس‌وجوی کاربر در سطح سراسری از مهم‌ترین دستاوردهای اجرای عملیاتی روش پیشنهادی است. نتایج به‌دست آمده بیان کننده این واقعیت است، که در سطح مفهومی، هر چه قواعد بر اساس هستان‌شناسی دامنه، کلاس و درخواست کاربر استخراج شود به اهداف کاربر نزدیک‌تر، مرتبط‌تر و کاربردی‌تر است. همچنین قواعد تولید شده در سطح مفهومی بالاتری قرار گرفته و کمک بیشتری به کاربر برای تصمیم‌گیری دقیق‌تر می‌کند. بنابراین با کاوش قواعد هم‌آیی مبتنی بر هستان‌شناسی، سطح مفهومی و معنایی قوانین افزایش و حجم داده‌ها در سطح پردازش و ذخیره سازی به میزان قابل توجهی کاهش می‌یابد.

جدول (۹): توصیف نمونه‌ای از قواعد هم‌آیی معنایی

ردیف	شماره قواعد هم‌آیی	درصد اطمینان	توصیف قوانین هم‌آیی تولید شده مبتنی بر هستان‌شناسی دامنه زمین‌لرزه
			Min-Sup:۰,۱ Min-Conf:۰,۷
۱	۱	۰,۸۲	٪۸۲ درصد زمین‌لرزه‌های با اندازه قوی در فصل بهار رخ داده است.
	۲	۰,۸۶	٪۸۶ درصد زمین‌لرزه‌های با اندازه قوی با تلفات متوسط رخ داده است.
	۳	۰,۸۲	٪۸۲ درصد زمین‌لرزه‌های با اندازه قوی در فصل تابستان و با تلفات کم رخ داده است.
	۴	۰,۹۱	٪۹۱ درصد زمین‌لرزه‌های با اندازه قوی در فصل زمستان رخ داده است.
	۵	۰,۷۸	٪۷۸ درصد زمین‌لرزه‌های با اندازه قوی در فصل پاییز رخ داده است.
	۶	۰,۹۲	٪۹۲ درصد زمین‌لرزه‌های با اندازه قوی با تلفات فاجعه‌آمیز رخ داده است.
	۷	۰,۷۷	٪۷۷ درصد زمین‌لرزه‌های با اندازه قوی با تلفات کم رخ داده است.
	۸	۰,۸۲	٪۸۲ درصد زمین‌لرزه‌های با اندازه قوی در فصل تابستان رخ داده است.
			Min-Sup:۰,۲ Min-Conf:۰,۷
۲	۱	۰,۸۲	٪۸۲ درصد زمین‌لرزه‌های با اندازه قوی در فصل بهار رخ داده است.
	۲	۰,۹۱	٪۹۱ درصد زمین‌لرزه‌های با اندازه قوی در فصل زمستان رخ داده است.
	۳	۰,۷۷	٪۷۷ درصد زمین‌لرزه‌های با اندازه قوی با تلفات کم رخ داده است.
	۴	۰,۸۸	٪۸۸ درصد زمین‌لرزه‌های با اندازه قوی در فصل تابستان رخ داده است.

۸- نتیجه‌گیری

در این مقاله، یک معماری ائتلافی جدید برای کاوش معنایی از داده‌های بزرگ مقیاس در محیط‌های توزیع شده ارائه شد. ائتلاف‌ها و ایده‌های مطرح شده در این معماری، مدیریت داده‌های بزرگ را با استفاده از فناوری‌های معنایی، فناوری‌های محاسباتی، هستان‌شناسی و همچنین سامانه‌های چند عاملی بهبود می‌دهد.

در معماری ASMLDE یک بررسی و تحلیل به‌روز شده از نقش عامل‌ها و هستان‌شناسی در طراحی چارچوب کاوش معنایی از داده‌های بزرگ مقیاس در محیط‌های توزیع شده ارائه گردید. بر اساس این معماری مشخص گردید، مدیریت کارآمد الگوهای استخراج شده با استفاده از روش‌های داده‌کاوی، الگوهای تعریف شده توسط کاربر و هستان‌شناسی دامنه در کاوش و ارزیابی الگوها بسیار تأثیرگذار است. این معماری علاوه بر ارائه یک روش کاربردی، توسعه‌پذیری محیط داده‌کاوی با استفاده از زبان هستان‌شناسی و استانداردهای ارتباط عامل‌ها را نیز فراهم و با افزودن معیارهای جدید به روش‌های موجود، کاوش معنایی، داده‌کاوی توزیع شده و ادغام سامانه‌های کاوش چند عاملی را تؤاماً فراهم نمود.

همچنین این معماری برای غلبه بر چالش‌هایی است که در محیط توزیع شده (مانند پهنای باند محدود، حساسیت به داده‌های محرمانه، محدودیت منابع محاسباتی توزیع شده) بروز می‌کند. از دیگر نتایج این معماری این است که عملیات داده‌کاوی در یک محیط ناهمگن توزیع شده با استفاده از سامانه‌های چند عاملی و هستان‌شناسی انعطاف پذیرتر، سازگارتر، قوی‌تر و آسان‌تر می شود. نتایج این تحقیق در چهار حوزه کلیدی تقسیم می‌شود، ۱) دخالت دادن هستان‌شناسی در فرآیند داده‌کاوی۲) داده‌کاوی معنایی مبتنی بر سامانه‌های چندعاملی، ۳) همکاری عامل‌ها در فرآیند داده‌کاوی و ۴) تجزیه و تحلیل ترکیبی که می‌تواند تصویر کاملی از مقیاس و عمق داده‌ها را ارائه دهد.

۹- کارهای آینده

برای تحقیقات آینده، استفاده از معیارهای کارکردی علاوه بر معیارهای کارآمدی، به‌کارگیری هستان‌شناسی و روش‌های معنایی با پشتیبانی و حمایت الگوریتم‌های یادگیری ماشین به‌منظور ارتقاء اثربخشی و توسعه و تعمیم الگوریتم‌های کاوش و داده‌کاوی معنایی پیشنهاد می‌شود.

۱۰- مراجع

[1] H. Zhuge, "The Complex Link," arXiv preprint arXiv:1805.00434, vol. abs/1805.00434, 2018. [Online]. Available: http://arxiv.org/abs/1805.00434.

[2] A. K. Bhadani and D. Jothimani, "Big Data: Challenges, Opportunities, and Realities," In Effective Big Data Management and Opportunities for Implementation: IGI Global, pp. 1-24, 2016.

[3] E. Belghache, J.-P. Georgé, and M.-P. Gleizes, "Towards an Adaptive Multi-agent System for Dynamic Big Data Analytics," In 2016 Intl. IEEE Conf. on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), IEEE, pp. 753-758, 2016.

[4] H. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. J. C. o. t. A. Shahabi, "Big data and its technical challenges," vol. 57, no. 7, pp. 86-94, 2014.

[5] R. M. Gahar, O .Arfaoui, M. S. Hidri, and N. B. J. P. C. S. Hadj-Alouane, "An Ontology-driven Map Reduce Framework for Association Rules Mining in Massive Data," vol. 126, pp. 224-233, 2018.

[6] B. Eine, M. Jurisch, and W. Quint, "Ontology-based big data management," Systems, vol. 5, no. 3, p. 45, 2017.

[7] P. V. Bhagat and P. M. Gourshettiwar, "A survey paper on ontology-based approaches for semantic data mining," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 3, no. 4, pp. 2137-2141, 2015.

[8] M. R. Chikhale, "Study of Distributed Data Mining Algorithm and Trends," IOSR-JCE, pp. 41-47, 2016. [Online]. Available: www.iosrjournals.org.

[9] D. Dou, H. Wang, and H. Liu, "Semantic Data Mining: A Survey of Ontology-based Approaches," In Proc. of the 2015 IEEE 9th Int. Conf. on Semantic Computing (IEEE ICSC 2015), ۲۰۱۵: IEEE, pp. 244-251, 2015.

[10] V. S. Ms and V. S. Ms and K. Shah, "Performance evaluation of distributed association rule mining algorithms," Procedia Computer Science, vol. 79, pp. 127-134, 2016.

[11] T. Hansmann and P. Niemeyer, "Big Data-characterizing an Emerging Research Field Using Topic Models," In Proc. of the 2014 IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Vol. 01, IEEE Computer Society, pp. 43-53, 2014.

[12] G. S. Bhamra, A. K. Verma, and R. B. Patel, "Agent Based Frameworks for Distributed Association Rule Mining: An Analysis," International Journal in Foundations of Computer Science & Technology (IJFCST), vol. 5, no. 1, pp. 11-22, 2015.

[13] G. S. Bhamra, A. Verma, and R. Patel, "A framework for association rule mining of distributed data," 2015.

[14] S. Urmela and M. Nandhini, "Approaches and Techniques of Distributed Data Mining: A Comprehensive Study," International Journal of Engineering and Technology (IJET), vol. 9, no. 1, p. 69, 2017.

[15] A. G. Touzi, H. B. Massoud, and A. Ayadi, "Automatic ontology generation for data mining using FCA and clustering," arXiv preprint arXiv:1311.1764, 2013.

[16] V. Bhatnagar and S. Srinivasa, Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, December 16-18, 2013,Proceedings. Springer, 2013, vol. 8302.

[17] S. Srinivasa and S. Mehta, Big Data Analytics: Third International Conference, BDA 2014, New Delhi, India, December 20-23, 2014. Proceedings. Springer, 2015, vol. 8883.

[18] J. M. Kanter and K. Veeramachaneni, "Deep feature synthesis: Towards automating data science endeavors," in Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on. IEEE, 2015, pp. 1–10. S. Urmela and M. Nandhini, "Approaches and Techniques of Distributed Data Mining: A Comprehensive Study," International Journal of Engineering and Technology (IJET), vol. 9, no. 1, p. 69, 2017.

[19] W. Fan and A. Bifet, "Mining big data: current status, and forecast to the future," ACM SIGKDD explorations newsletter, vol. 14, no. 2, pp. 1-5, 2013.

[20] H. M. Safhi, B. Frikh, and B. Ouhbi, "Assessing reliability of Big Data Knowledge Discovery process," Procedia computer science, vol. 148, pp. 30-36, 2019.

[21] S. Jain and V. Meyer, "Evaluation and refinement of emergency situation ontology," Int J Inform Educ Technol, vol. 8, no. 10, pp. 713-719, 2018.

[22] S. Nadal, O. Romero, A. Abelló, P. Vassiliadis, and S. S. Nadal, O. Romero, A. Abelló, P. Vassiliadis, and S. Vansummeren, "An integration-oriented ontology to govern evolution in big data ecosystems," Information systems, vol. 79, pp. 3-19, 2019.

[23] B. Jadhav Kalyani, S. Tamhane Manisha, U. Surwase Sonali, and A. P. P. Patil, "A new Approach for Frequent Itemset Data Mining in Hadoop Environment," 2017.

[24] W. Gan, J. C. W. Lin, H. C. Chao, and J. Zhan, "Data mining in distributed environment: a survey," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 6, p. e1216, 2017.

[25] M. Barati, Q. Bai, and Q. Liu, "Mining semantic association rules from RDF data," Knowledge-Based Systems, vol. 133, pp. 183-196, 2017.

[26] S. G. Atal and P. Chatur, "Large scale ontology for semantic web using clustering method over Hadoop," in 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB) , pp. 632-636, IEEE: 2016.

[27] D. A. N. T.Bharathi, "Enhanced Way of Association Rule Mining with Ontology," Int. J. of Engineering and Computer Science, vol. 05, no. 10, pp. 18363-18371, 2016.

[28] A. Ogunde, O. Folorunso, and A. Sodiya, "A partition enhanced mining algorithm for distributed association rule mining systems," Egyptian Informatics Journal, vol. 16, no. 3, pp. 297-307, 2015.

[29] S. Patil, S. Karnik, and V. Sawant, "A Review on Multi-Agent Data Mining Systems," International Journal of Computer Science and Information Technologies, vol. 6, no. 6, pp. 4888-4893, 2015.

[30] J. Raad and C. Cruz, "A survey on ontology evaluation methods," 2015.

[31] H. Hlomani and D. Stacey, "Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey," Semantic Web Journal, vol. 1, no. 5, pp. 1-11, 2014.

[32] C. J. Thompson, "The 'big data'myth and the pitfalls of 'thick data'opportunism: on the need for a different ontology of markets and consumption," Journal of Marketing Management, vol. 35, no. 3-4, pp. 207-230, 2019.

[33] Y. Y. Ang, "Integrating Big Data and Thick Data to Transform Public Services Delivery," 2019.

Journal of Electronical & Cyber Defence
Vol. 8, No. 3, 2020, Serial No. 31

Providing an Agent-Based Architecture for Semantic Mining From Large-Scale Data in Distributed Environments

H. Saberi*, M. R. Kangavari, M. R. Hasani Ahangar

*Imam Hossein Comprehensive University

(Received: 22/09/2019, Accepted: 01/02/2020)

ABSTRACT

Large-scale data may consist of big, distributed, scattered, heterogeneous, irrelevant, misleading, real, and unrealistic data or any combination of them. Therefore, analyzing, creating value and data productivity is always an important and open challenge. Therefore, the purpose of this study is to present a new coalition architecture for generating valuable information for decision making among the masses of data. The proposed architecture, abbreviated ASMLDE, aims to develop and improve data mining and semantic exploration, and to produce useful and high-quality rules consisting of four layers, seven components and six key elements. In the proposed architecture, conceptualization with 4v's process, insight into the volume and scale of data in the form of 3v's model and finally qualitative insight based on data thickness, are used for conceptualization and standardization of qualitative processes and more complex interpretations. This architecture, supported by ontology and agent mining, reduces large search spaces and increases the speed and quality of data mining operations due to the use of multi-agent systems. Automating exploration operations, reducing data complexity and business processes are also important achievements of the proposed architecture. To evaluate the proposed architecture, a large-scale dataset of natural disasters and earthquake ontology classes from the DBpedia knowledge base have been used. The evaluation results obtained by exploring the semantic rules of the mentioned dataset highlight the effectiveness and capabilities of the ASMLDE architecture in enhancing the quality of the semantic rules explored to fit the user need and reducing the large data mining space over other similar frameworks and architectures.

Keywords: Large Scale Data, Semantic Mining, Ontology, Agent-Oriented Architecture

* Corresponding Author Email: hsaberi@ihu.ac.ir