Journal of Electronical & Cyber Defence
Vol. 9, No. 1, 2021, Serial No. 33

Providing an Ontology-Based Method for Exploring the Association Rules in Multi-Agent Distributed Environments

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

*Imam Hossein Comprehensive University

(Received: 17/02/2020, Accepted: 05/08/2020)

ABSTRACT

Distributed association rules mining is one of the most important data mining methods that extracts the inter dependence of data items from decentralized data sources, regardless of their physical location and is based on the process of extracting repeated items. When exploration algorithms are implemented on large-scale data, a large number of recurring items are produced, many of which are irrelevant, ambiguous, and unusable for the business, thus causing a challenge called "combination explosion ."In this paper, a new coalition method based on distributed data mining and domain archeology, abbreviated to DARMASO, is proposed to address this challenge. This method uses three algorithms: the DARMASOMAIN algorithm to guide and control the process of exploration and aggregation of universal rules, the DARMASOPRU algorithm to reduce and prune the data and the DARMASOINT algorithm to explore and aggregate the rules of all the generated data sources. DARMASO uses a map-reduce-based distributed computational model in a multi-agent distributed environment. It also provides a practical way for semantic mining of large-scale data sets. This method filters out the association rules of generality based on the purposes of data mining as well as the needs of the user and only produces and maintains useful rules. Reducing the scope of exploration and filtration of rules is achieved through the process of semantic pruning in the form of removing inappropriate candidates from the set of frequent items and producing association rules of utility. The implementation is performed using a data set from the scope of natural disasters and the earthquake class. It also improves the speed and quality of rule extraction and generates practical, reliable, logical, quality and valuable rules to support decision-making amid the masses of data.

Keywords: Association Rules, Ontology, Multi Agent Systems, Mapping-Reduction

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

نشریه علمی "پدافند الکترونیکی و سایبری"
سال نهم، شماره ۱، بهار ۱۴۰۰، ص ۱۷-۱

علمی – پژوهشی

ارائه یک روش مبتنی بر هستان‌شناسی برای کاوش قواعد هم‌آیی در محیط‌های توزیع‌شده چندعاملی

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: ۱۳۹۸/۱۱/۲۸، پذیرش: ۱۳۹۹/۰۵/۱۵)

چکیده

کاوش قواعد هم‌آیی توزیع‌شده یکی از روش‌های مهم داده‌کاوی است که وابستگی بین اقلام داده‌ای را از منابع داده‌ای غیرمتمرکز، بدون توجه به مکان فیزیکی آن‌ها و بر مبنای فرآیند استخراج اقلام مکرر استخراج می‌کند. هنگامی‌که الگوریتم‌های کاوش روی داده‌های بزرگ‌مقیاس اجرا می‌شوند، مقدار زیادی اقلام مکرر تولید می‌گردد که بسیاری از آن‌ها غیرمرتبط، مبهم و غیرقابل استفاده برای کسب و کار است و سبب بروز چالشی به نام " انفجار ترکیبی" خواهد شد. در این مقاله یک روش ائتلافی جدید مبتنی بر داده‌کاوی توزیع شده و هستان‌شناسی دامنه که به‌اختصار DARMASO نامیده می‌شود برای برخورد با این چالش پیشنهاد شده است. این روش از سه الگوریتم به نام ARMASOMAIN جهت هدایت و کنترل فرآیند کاوش و تجمیع قواعد هم‌آیی، DARMASOPRU برای کاهش و هرس داده‌ها و الگوریتم DARMASOINT برای کاوش و تجمیع قواعد هم‌آیی تولیدشده از منابع داده‌ای توزیع‌شده استفاده می‌کند. DARMASO از یک الگوی محاسباتی توزیع‌شده مبتنی بر چارچوب نگاشت-کاهش در محیط توزیع‌شده چندعاملی استفاده می‌کند. همچنین یک روش عملی را برای کاوش معنایی از مجموعه داده‌های بزرگ‌مقیاس فراهم می‌کند. این روش، قواعد هم‌آیی را مبتنی بر اهداف داده‌کاوی و نیاز کاربر فیلتر کرده و فقط قواعد مفید را تولید و نگهداری می‌کند.کاهش فضای کاوش و فیلترسازی قواعد، با فرآیند هرس معنایی در قالب حذف نامزدهای نامناسب از مجموعه اقلام مکرر و تولید قواعد هم‌آیی سودمند حاصل می‌شود. پیاده‌سازی با استفاده از یک مجموعه داده‌ای از دامنه حوادث طبیعی و کلاس زمین‌لرزه انجام شده است. نتایج ارزیابی با استفاده از معیارهای کمی و کیفی نشان می‌دهد، الگوریتم‌های ارائه‌شده در DARMASO ، فضای کاوش را به میزان قابل توجهی کاهش می‌دهد. همچنین سرعت و کیفیتِ استخراج قواعد را بهبود بخشیده و قواعدِ کاربردی، مطمئن، منطقی، با کیفیت و ارزشمندی را برای پشتیبانی از تصمیم‌گیری از میان انبوه داده‌ها تولید می‌کند.

کلیدواژه‌ها: قواعد هم‌آیی، هستان‌شناسی، سامانه‌های چندعاملی، نگاشت-کاهش

۱- مقدمه

رشد چشمگیر داده‌ها از طریق جهش در حجم و دامنه آن موجب ایجاد فرصت‌های جدید شده است [۱]. همچنین داده‌ها به مواد اولیه ارزشمند برای تولید در سازمان‌ها تبدیل شده است [۲]. لذا تمرکز مدیران و راهبران فناوری اطلاعات بر تجزیه و تحلیل داده‌ها ضروری است. تجزیه و تحلیل حجم و دامنه داده‌ها یک وظیفه حیاتی در مدیریت و مهندسی داده‌ها است. لذا، سامانه‌های تجزیه و تحلیل داده باید با طراحی مناسب، قادر به حل مسائل 10 Vs (حجم، سرعت، تنوع، صحت، اعتبار، آسیب‌پذیری، نوسان، تجسم، ارزش، تغییرپذیری) بوده و به‌طور مؤثر به ایجاد تعادل بین اهداف و هزینه پردازش بپردازند [۳].

استخراج ارزش از داده‌ها، به‌عنوان یک مرحله مهم در تجزیه و تحلیل داده‌های بزرگ‌مقیاس محسوب می‌شود. در این زمینه پژوهشگران معتقدند که رویکردهای معنایی مبتنی‌بر هستان‌شناسی[1] یک چارچوب عملی برای رسیدگی به چالش‌های ارائه‌شده در مجموعه داده‌های بزرگ‌مقیاس فراهم می‌کند [۴]. در این میان، تولید قواعد هم‌آیی[2] یک فرآیند تکراری و تعاملی است که شامل چندین مرحله از انتخاب و آماده‌سازی داده‌ها برای تفسیرِ نتایج و استخراجِ دانش از فرآیند داده‌کاوی است.

روش‌های پیشنهادی برای استخراج قواعد هم‌آیی بر اساس چهار مرحله ۱) آماده‌سازی داده‌ها، ۲) استخراج مجموعه اقلام مکرر از ویژگی‌ها، ۳) تولید قواعد هم‌آیی و ۴) تفسیر نتایج است.

[1] Ontology
[2] Association Rules

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

ناشر: دانشگاه جامع امام حسین (ع)

در مواجهه با چهار مرحله قبلی مشکلات مختلفی بروز می‌کند. اول، تولید زمان پاسخ قواعد هم‌آیی[1] به زمان استخراج مجموعه اقلام مکرر[2] بستگی دارد. دوم، جریان داده‌های بزرگ مقیاس، موجب تولید تعداد قابل توجهی از اقلام مکرر[3] می‌شود که عمدتاً اضافه و بی اهمیت هستند که حذف آن‌ها می‌تواند به‌طور قابل توجهی تعداد قواعد پردازش را کاهش دهد[۳]. به این ترتیب، از کاوش مجموعه اقلام مکرر برای تولید قواعد هم‌آیی، به‌عنوان یک وظیفه مهم برای استخراج اطلاعات و داده‌کاوی استفاده می‌شود. برای پاسخ دادن به مشکلات بیان شده، یک روش جدید با نام DARMASO[4] برای تولید قواعد هم‌آیی مفید بر اساس مجموعه اقلام مکرر توزیع‌شده در چارچوب نگاشت-کاهش[5] و محیط توزیع‌شده چندعاملی[6] پیشنهاد و با استفاده از هستان‌شناسی (شامل هرس معنایی[7] مبتنی بر هستان‌شناسی) بهبود می‌یابد. ایده اصلی DARMASO این است که استفاده از یک دامنه هستان‌شناسی به عنوان یک حامی هرس معنایی برای حذف برخی از الگوهای اقلام مکرر، به‌منظور کاهش تعداد مجموعه اقلام مکرر و تولید قواعد هم‌آیی مناسب اطمینان حاصل شود. این اقدام به کاهش کمی و افزایش کیفی مجموعه اقلام مکرر و به تبع آن قواعد هم‌آیی می‌انجامد. هرس معنایی از هستان‌شناسی، گام مهمی است که در هر دو وظیفه نگاشت و کاهش معرفی می‌شود و به‌عنوان پیش‌پردازش و مراحل بعد از پردازش به‌منظور تولید و نگهداری قواعد هم‌آیی مفید و مهم به کار می‌رود. DARMASO کاربرد مفاهیم جدید مبتنی بر هستان‌شناسی و عامل‌گرایی را در حوزه مدیریت داده‌های بزرگ‌مقیاس فراهم می‌کند.

هدف اصلی راه‌حل‌های مبتنی بر هستان‌شناسی، حمایت از فرآیند برخورد با داده‌های ناهمگون و دسترسی به داده‌های مرتبط با توجه به افزایش اندازه آن است [۴]. همچنین رویکردهای مبتنی بر هستان‌شناسی و سامانه‌های چندعاملی نقش اساسی در فرایند داده‌کاوی و کاوش معنایی ایفاء می‌کند. عامل‌ها نیز فرآیندهای مؤثر کشف دانش، شامل انتخاب داده[8]، استخراج داده[9]، پیش‌پردازش داده[10] و ادغام داده[11] را پشتیبانی و بر توسعه و تغییر داده‌ها در منابع داده‌ای و سایت‌های توزیع‌شده و کاوش الگوها و ادغام آن‌ها نظارت خواهد کرد[۱]. سازماندهی این مقاله در ادامه به شرح زیر است: در بخش۲، تعاریف کلی و مفاهیم پایه‌ای معرفی می‌گردد. در بخش۳، مروری بر کارهای مرتبط در حوزه کاوش قواعد هم‌آیی انجام شده و نقاط ضعف و قوت آن‌ها بیان می‌شود. در بخش۴، روش پیشنهادی بیان شده و جزئیات چارچوب نگاشت-کاهش و محیط توزیع‌شده چندعاملی مبتنی بر هستان‌شناسی برای کاوش قواعد هم‌آیی تشریح می‌گردد. بخش ۵، نتایج محاسباتی DARMASO را مورد بحث قرار می‌دهد. و در بخش آخر، نتیجه‌گیری از بحث انجام شده و پیشنهادهایی در قالب کارهای آتی ارائه می‌شود.

۲- ادبیات تحقیق

اولین قدم در پردازش معنایی داده‌های بزرگ‌مقیاس ایجاد تفکری جدید نسبت به داده است. در داده‌کاوی معنایی، داده‌ها باید در سطحی از هوشمندی قرار گیرند تا توسط ماشین قابل درک باشند. لذا هستان‌شناسی توانایی کمک به فرآیند داده‌کاوی معنایی[12] را از طریق معناشناختی موجود در هستان‌شناسی فراهم می‌کند. در سال‌های اخیر، استفاده از داده‌کاوی معنایی با حمایت هستان‌شناسی و داده‌کاوی توزیع‌شده با پشتیبانی سامانه‌های چندعاملی به یک حرکت بزرگ در داده‌کاوی و مدیریت داده‌های بزرگ‌مقیاس تبدیل شده است [۴]. هستان‌شناسی می‌تواند فرآیند کاوش را کنترل و فضای پرس و جو را کاهش دهد [۵]. سامانه چندعاملی نیز، راه‌حل‌هایی برای حل مسئله تعامل و همکاری در محیط‌های توزیع‌شده ارائه می‌دهد[۶]. از هستان‌شناسی و عامل‌ها می‌توان در داده‌کاوی برای کاوش قواعد هم‌آیی، طبقه‌بندی، خوشه‌بندی، استخراج اطلاعات و سامانه‌های توصیه‌گر[13] استفاده کرد [۳، ۷ و ۸].

۲-۱- مجموعه‌های اقلام مکرر

بر اساس تعریف آگراوال[14]،[۹] مجموعه اقلام مکرر[15]، شکلی از الگوی مکرر است که اقلام آن به‌طور مکرر در یک مجموعه داده ظاهر می‌شوند. کشف مجموعه اقلام مکرر منجر به کشف وابستگی و روابط میان اقلام در مجموعه داده‌های رابطه‌ای و تراکنشی می‌شود. فرض کنید I مجموعه‌ای از اقلام به‌صورت:

$$I = (i_1; i_2; \dots; i_n) \quad (۱)$$

[1] Response Times of the ARs Generation
[2] Times of the Frequent Itemsets
[3] Frequent Pattern
[4] Distributed Association Rule Mining in Multi Agent Environment with Semantic Ontologies (DARMASO)
[5] MapReduce Framework
[6] Multi Agent Distributed Environment
[7] Semantic Pruning
[8] Data Selection
[9] Data extraction
[10] Data Preprocessing
[11] Data Integration
[12] Semantic Data Mining `
[13] Recommender Systems
[14] Agrawal
[15] Frequent Itemsets

و D مجموعه‌ای از تراکنش‌ها است. به‌طوری که هر تراکنش T مجموعه‌ای از اقلام به‌صورت $T \subseteq I$ است. هر تراکنش دارای یک شناسه به نام شناسه تراکنش[1] است . فرض کنید A یک مجموعه از اقلام باشد، یک تراکنش T حاوی A است اگر و فقط اگر $A \subseteq T$ باشد. یک قاعده وابستگی، مفهومی است به فرم $A \Rightarrow B$، جایی که $A \subset I$، $B \subset I$ و $A \cap B = \emptyset$ است. قاعده $A \Rightarrow B$ در مجموعه داده D با درجه پشتیبانی S قرار دارد، جایی که S[2] درصد تراکنش‌هایی در D است که شامل $A \cup B$ می‌باشد و با احتمال P(A∪B) نشان داده‌شده در (۲) به‌دست می‌آید. قاعده فرم $A \Rightarrow B$ در مجموعه داده D دارای درجه اطمینان C هست، جایی که C[3] درصد تراکنش‌های شامل A در D می‌باشد که شامل B نیز هست و با احتمال شرطی P(B|A) نشان داده‌شده در (۳) به‌دست می‌آید.

$$S = (A \Rightarrow B) = P(A \cup B) \qquad (۲)$$

$$C = (A \Rightarrow B) P(A|B) \qquad (۳)$$

قواعدی که هر دو آستانه پشتیبانی حداقل[4] و آستانه اطمینان حداقل[5] را دارا باشند قواعد قوی[6] نامیده می‌شود [۱۰]. چالش اصلی در کاوش مجموعه اقلام مکرر، تولید تعداد زیادی از مجموعه اقلام با آستانه حداقل پشتیبانی است. اگر یک مجموعه اقلام مکرر باشد، هر یک از زیر مجموعه‌های آن نیز مکرر هستند. به عبارت دیگر، یک مجموعه اقلام مکرر بزرگ، شامل تعداد زیادی مجموعه اقلام مکرر کوچک‌تر است. الگوریتم‌هایی که همه مجموعه اقلام مکرر را کشف می‌کنند، متحمل مشکل انفجار ترکیبی[7] می‌شوند. برای حل این مشکل مفاهیم مجموعه اقلام مکرر بسته[8] و مجموعه اقلام مکرر بیشینه[9] معرفی شده‌اند. از آنجایی که کاوش مجموعه اقلام مکرر از نظر محاسباتی گران است، راهبرد های توزیع و موازی‌سازی آن بر استفاده حافظه، متعادل‌سازی بار و هزینه های ارتباطی تأثیرگذار است [۳، ۱۱ و ۱۲].

۲-۲- قواعد هم‌آیی

یک قاعده هم‌آیی[10]، بیان یک شکل X⇒Y است. جایی که X و Y مجموعه‌های اقلام و $X \cap Y = [1]$ است. چنین قاعده‌ای بیان می‌کند که اگر تراکنش شامل تمام اقلام در X باشد، این تراکنش شامل تمام اقلام در Y نیز می‌باشد. X بدنه یا مقدم و Y به‌عنوان رأس یا تالی نامیده می‌شود. پشتیبانی از یک قاعده هم‌آیی $X \Rightarrow Y$ در D، حمایت از $X \cup Y$ در D است و به همین ترتیب، تکرار قاعده، توالی یا تکرار $X \cup Y$ است. اطمینان یا دقت قاعده هم‌آیی X⇒Y در D، احتمال شرطی داشتن Y در یک تراکنش با توجه به این است که X در آن تراکنش وجود دارد. یعنی:

$$\text{Confedence}\left(X \Rightarrow Y, D\right) = P(Y|X) = \frac{\text{Support}\left(X \cup Y, D\right)}{\text{Support}\left(X, D\right)} \qquad (۴)$$

اگر $P(Y \mid X)$ از یک آستانه اطمینان حداقل γ، حاوی $0 \leq \gamma \leq 1$ فراتر رود، قاعده مطمئن[11] نامیده می‌شود. کاوش قواعد هم‌آیی توزیع‌شده نیز، استخراج دانش و قواعد از منابع داده‌ای توزیع شده است و به‌طور خاص بر معماری توزیع‌شده متمرکز و نیازمند ادغام دانش تولیدشده از منابع مختلف داده‌ای است [۱، ۳ و ۱۳].

۲-۳- هستان‌شناسی

هستان‌شناسی یکی از پیشرفت‌های مؤثر در مهندسی دانش و تعریف صریح و بدون ابهام یک مفهوم است [۵]. با توجه به تعریف گروبر[12] [۳]، هستان‌شناسی مشخصاتی از یک مفهوم و یک دیدگاه منسجم از اطلاعات مدیریت‌شده در قالب یک مجموعه و به‌صورت یک لیست صریح و سازمان‌یافته از تمام اصطلاحات، روابط و اشیائی است که نمایش یک دامنه را نشان می‌دهد. هستان‌شناسی توسط رابطه $O = (S.L)$ تعریف می‌شود. ساختار مفهومی توسط رابطه (۵) تعریف می‌شود [۳، ۱۴].

$$S = (C.R. \leq . \sigma_R) \qquad (۵)$$

واژگان شامل برچسب‌هایی است که با مفاهیم و ارتباط مفهوم هستان‌شناسی مرتبط هستند. این تعریف توسط رابطه (۶) تعریف می‌شود.

$$L = (L_C . L_R F_C . F_R) \qquad (۶)$$

تابع مفاهیم یک تابع تعریف‌شده روی مجموعه‌ای از مفاهیم است که توسط رابطه (۷) توصیف می‌شود.

$$\forall l \in L_C . F_C(l) = \left\{\frac{c}{c} \in C\right\} \qquad (۷)$$

و تابع رابطه یک تابع تعریف‌شده روی مجموعه‌ای از ارتباط‌ها است که توسط رابطه (۸) توصیف می‌شود.

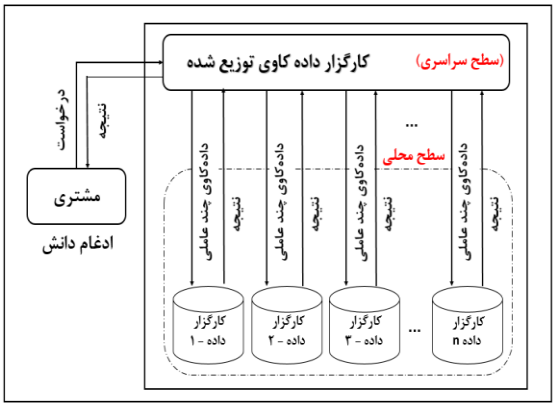
[1] Transaction Identification (TID)
[2] Support
[3] Confedence
[4] Min_Support
[5] Min_Confedence
[6] Strong Rule
[7] Combinatorial Explosion
[8] Closed Frequent Itemset (CFI)
[9] Maximal Frequent Itemset (MFI)
[10] Association Rule
[11] Confident Rule
[12] Gruber

$$\forall l \in L_R . F_R(l) = \left\{\frac{r}{r} \in C\right\} \quad (۸)$$

توابع بیان‌شده، امکان دسترسی به مفاهیم و روابط تعیین‌شده توسط یک برچسب را فراهم می‌کند [۱۵ و ۱۶].

۲-۴- داده‌کاوی توزیع‌شده عامل‌گرا[1]

آینده داده‌کاوی، مبتنی بر محاسبات داده‌ای در مکان‌های مختلف جغرافیایی است که به نام داده‌کاوی توزیع‌شده یا داده‌کاوی جمعی[2] نامیده می‌شود [۱۶]. داده‌کاوی توزیع‌شده بر کاوش از منابع داده‌ای توزیع‌شده تأکید دارد و عمدتاً شامل دو روش اطلاعات توزیع شده و محاسبات توزیع شده است. چارچوب عامل‌گرایی نیز به‌عنوان حوزه‌ای مؤثر در محیط‌های توزیع‌شده شناخته شده و نقش کلیدی برای توسعه عامل‌گرایی دارد. معماری‌های مبتنی بر عامل، سازوکاری کلیدی است که مستقل از اجزاء، کارایی کارآمد را در محیط‌های پویا، باز و واقعی پشتیبانی می‌کند. لذا داده‌کاوی توزیع‌شده عامه‌گرا، بر اساس جمع‌آوری، پردازش داده‌ها و اجرای کد منبع در سطح محلی یا سراسری انجام می‌شود. الگوریتم‌های داده‌کاوی توزیع‌شده بر داده‌های موجود در سایت‌های ناهمگون به‌عنوان مدل محلی اعمال خواهد شد و در نهایت، نتیجه محاسبات انجام‌شده داده‌کاوی برای مدل‌سازی سراسری تجمیع می‌شود. داده‌کاوی توزیع‌شده عامل‌گرا، انتقال و جابجایی کد به جای داده‌ها و ارائه روشی مقیاس‌پذیر و کارآمد است. محدودیت‌های آن نیز شامل استفاده ناکارآمد از منابع محاسباتی و دخالت جزئی کارگزار داده‌کاوی توزیع‌شده در مراحل داده‌کاوی است [۱۷]. شکل (۱) معماری داده‌کاوی توزیع‌شده عامل‌گرا را نشان می‌دهد.



شکل (۱): معماری داده‌کاوی توزیع شده عامل‌گرا [۱۷و ۱۸]

[1] Agent -based Distributed Data Mining
[2] Collective Data Mining

سامانه‌های چندعاملی[3]، شکل یافته عامل‌های مختلفی است که در تعامل با یکدیگر تشکیل و فرآیند کاوش را انجام می‌دهند. استفاده از سامانه‌های چندعاملی در داده‌کاوی توزیع شده، اجازه انتخاب پویا از منابع، جمع‌آوری داده‌ها و همچنین مقیاس‌پذیری، امنیت، قابلیت اطمینان، تعامل‌پذیری و خودمختاری فرآیند استخراج داده‌ها را فراهم می‌کند[۱۸]. MAS به‌صورت چهارتایی در رابطه (۹) تعریف می‌شود.

$$MAS = <Ag, Env, Org, D> \quad (۹)$$

$Ag = \{Ag_1, Ag_2, \ldots, Ag_n\}$ مجموعه عامل‌های تشکیل‌دهنده سامانه چندعاملی هست. Env محیطی است که سامانه چندعاملی در آن عمل می‌کند. Org سازمان سامانه چندعاملی و D قلمرو سامانه چندعاملی است.

۲-۵- نگاشت- کاهش[4]

برای کشف دانش از داده‌های بزرگ‌مقیاس و افزایش مقیاس پذیری، موازی‌سازی[5]، یک راه‌حل برای مقابله با داده‌های بزرگ‌مقیاس است، به همین دلیل بسیاری از الگوریتم ها با استفاده از چارچوب محاسباتی نگاشت- کاهشبا هم موازی شده‌اند. مدل پردازشی برنامه نگاشت- کاهش با استفاده از دو تابع Map() و Reduce() نوشته شده و داده‌ها به‌صورت جفت مقدار (کلید و ارزش)[6] نمایش داده می‌شوند. این دو تابع شامل:

- تابع Map (): جفت مقدار ورودی را می‌گیرد و یک جفت کلید / ارزش میانجی ارائه می‌دهد. سپس تمام جفت‌های میانجی به ترتیب با کلید میانجی معمولی I گروه‌بندی می‌شوند و به تابع کاهش می‌رسند.

$$Map(k_1.v_1) \rightarrow list(k_2.v_2)) \quad (۱۰)$$

- تابع Reduce (): یک کلید میانجی I و مجموعه‌ای از مقادیر برای آن کلید را می‌پذیرد.

$$\left(Reduce(k_2.list(v_2)\right) \rightarrow list(v_3) \quad (۱۱)$$

داده‌های ورودی در مجموعه‌ای از پارتیشن‌ها و در یک چارچوب پردازش توزیع‌شده به نام فایل‌سامانه توزیع‌شده[7] ذخیره می‌شوند [۳]، [۱۲] و [۱۳].

[3] Multi Agent System (MAS)
[4] Map-Reduce
[5] Parallelization
[6] Key-Value
[7] Hadoop Distributed File System (HDFS)

۲-۶- پژوهش‌های مرتبط

در سال‌های اخیر، راه‌حل‌های نوینی برای کاوش و تجزیه و تحلیل داده‌های بزرگ‌مقیاس ارائه شده است. با این حال، تلاش‌های محدودی برای دستیابی به نتایج عملی از همگرایی شیوه‌ها و فن‌ها انجام‌شده است. رویکردهای پیشنهادی این مقاله که در ادامه به آن پرداخته می‌شود، ارائه یک روش ائتلافی متشکل از داده‌کاوی توزیع‌شده، عامل‌گرایی و هستان‌شناسی (شامل زبان و سازوکارهای استنتاج) است. این رویکرد در ساختار سامانه‌های اطلاعاتی و داده‌های بزرگ‌مقیاس و پیچیده منجر به درجه بالایی از ادغام، مدیریت و اتوماسیون داده‌ها و فرآیندها بر اساس هستان‌شناسی و در نهایت داده‌کاوی هدفمند خواهد شد. پژوهش‌های مرتبط با این مقاله در چهار بخش بیان می‌شود. دسته اول، به معماری‌ها و الگوریتم‌های داده‌کاوی توزیع‌شده و دسته دوم به استفاده از هستان‌شناسی در داده‌کاوی می‌پردازد. دسته سوم به انواع معماری عامل‌ها در داده‌کاوی و دسته چهارم به نحوه مدیریت و کاوش از داده‌های بزرگ‌مقیاس می‌پردازد.

مقاله [۲] به ارائه چارچوبی می‌پردازد که به‌طور خودکار قابلیت اطمینان فرآیند داده‌کاوی را ارزیابی می‌کند. قابلیت اطمینان یک نگرانی اساسی در تجزیه و تحلیل داده‌های بزرگ‌مقیاس است که در بسیاری از پروژه‌ها نادیده گرفته می‌شود. این چارچوب به اطمینان از اعتبار هر مرحله فرآیند داده‌کاوی و همچنین به شناسایی و تجدید نظر در مراحل غیرقابل اطمینان کمک می‌کند. با اجرای این روش، فرآیند داده‌کاوی بهینه‌شده و قواعد با کیفیت بالاتری تولید می‌شود. مقاله [۲۱] به رویکردها و روش‌های داده‌کاوی متمرکز و توزیع شده پرداخته است. این مقاله به معرفی چارچوب‌ها و سامانه‌های داده‌کاوی توزیع‌شده با استفاده از معماری عامل‌گرا برای کاوش قواعد هم‌آیی توزیع‌شده می‌پردازد. در این مقاله سامانه‌های چندعاملی پادما[1]، جَم[2]، بودهی[3] و پاپیروس[4] معرفی شده است. همچنین مقایسه کیفی چارچوب‌های کاوش قواعد هم‌آیی عامل‌گرا شامل MADKDS، AFARMDD و MADARM بیان شده است. نویسنده معتقد است، رویکردهای متمرکز به کاهش پیچیدگی محاسباتی، صحت و بهبود کارایی و رویکردهای توزیع شده به کاهش پیچیدگی و زمان محاسبات می‌انجامد.

در مقاله [۱۸] نویسنده به نقش عامل‌کاوی برای غلبه بر چالش‌های داده‌کاوی در یک محیط ناهمگون توزیع شده پرداخته است. نتیجه ارائه شده این است که داده‌کاوی در یک محیط ناهمگون و توزیع شده با استفاده از عامل‌های داده‌کاوی، انعطاف‌پذیر، سازگار، قوی و آسان می‌شود و عامل کاوی و سامانه‌های چندعاملی به یک حرکت بزرگ در داده‌کاوی تبدیل شده است. مقاله [۲۰] یک روش مقیاس‌پذیر برای حل مسئله کاوش اقلام مکرر در حافظه‌های غیرفرار[5] به نام PevFP-tree ارائه می‌کند. در این روش، کاوش اقلام مکرر موجود مانند FP-tree عمومی با استفاده از مزایای عملکرد متقارن چند پردازنده‌ای در معماری محاسبات موازی ادغام و تقویت شده است. در [۱۹] به معرفی یک الگوریتم به نام B-Min و یک ساختار داده‌ای جدید به نام B-Table با هدف بهبود الگوریتم‌های کاوش اقلام مکرر کلاسیک Apriori، FP-Growth و H-Mine می‌پردازد. الگوریتم‌های پیشنهاد شده در دو جهت تحقیقاتی شامل ۱) روش‌های موازی و نگاشت-کاهش در قالب الگوریتم PB-mine و ۲) فشرده‌سازی داده‌ها برای کاوش اقلام مکرر از داده‌های بزرگ در قالب یک نسخه فشرده ازB-Mine به نام EB-mine معرفی شده است. در [۴] به کاربرد مفاهیم مبتنی‌بر هستان‌شناسی در مدیریت داده‌های بزرگ پرداخته و یک روش جدید بر اساس فناوری‌های معنایی و هستان‌شناسی ارائه می‌دهد. این روش به چگونگی ادغام داده‌ها از طریق همگرایی هستان‌شناسی، سنجش کیفیت داده‌ها و ادغام فرآیند کسب و کار می‌پردازد. مقاله [۱۷] به هفت رویکرد، روش و چارچوب عامل‌گرا برای داده‌کاوی توزیع‌شده پرداخته و روش‌های عامل‌کاوی را برای توسعه داده‌کاوی و افزایش دقت در داده‌های کاوش شده بیان می‌کند. در این مقاله به معرفی چارچوبی به نام CoLe2 مبتنی بر عامل‌های همراه در رویکردهای داده‌کاوی توزیع‌شده پرداخته شده است. مقاله [۱۳] به معرفی دو چالش کار با داده‌های بزرگ‌مقیاس پرداخته است. چالش اول، اندازه مجموعه داده[6] است که بسیار سریع‌تر از حافظه در دسترس یک ایستگاه کاری افزایش می‌یابد. چالش دوم، زمان محاسباتی مورد نیاز برای یافتن یک راه‌حل است. در این مقاله الگوی محاسباتی داده‌های بزرگ‌مقیاس، شامل الگوریتم‌های موازی، حافظه مشترک، حافظه توزیع (عبور پیام و نگاشت–کاهش) ارائه شده است. جدول (۱) به مقایسه پژوهش‌های مرتبط با این مقاله پرداخته، مزایا و معایب آن‌ها را به‌اختصار بیان می‌کند.

[1] PADMA
[2] JAM
[3] BODHI
[4] Papyrus

[5] Non-volatile Memories
[6] Dataset

جدول (۱): بررسی و مقایسه پژوهش‌های مرتبط

پژوهش	سال	مزایا	معایب
[۱۷] S. Urmela and M. Nandhini	۲۰۱۹	▪ ارائه یک چارچوب خودکار برای ارزیابی قابلیت اطمینان در فرآیند کشف دانش ▪ سازگاری با داده‌های بدون ساختار	▪ عدم اجرای یک مورد مطالعاتی
[۱] D. Patel and J. Shah	۲۰۱۷	▪ سازگاری با داده‌های بدون ساختار ▪ داشتن یک مدل فرمال برای چارچوب ارائه‌شده ▪ اطمینان از قواعد تولیدشده	• تعیین روش‌های مناسب برای استفاده از داده‌ها • دقت قابل قبول در یک زمان معقول • عدم اجرای یک مورد مطالعاتی
[۴] B. Eine, M. Jurisch, and W. Quint	۲۰۱۷	▪ پشتیبانی از ارزیابی و کیفیت بهتر نتایج حاصل از ادغام و یکپارچه‌سازی داده‌ها ▪ تسهیل و افزایش سریع‌تر ادغام منابع جدید داده‌ای و ارزیابی نتایج حاصل از یکپارچه‌سازی داده‌ها ▪ مدیریت اطلاعات پیچیده و روابط بین آن‌ها	• عدم اجرای نمونه اولیه و مطالعه کاربردی برای ارزیابی اثربخشی و قابلیت استفاده مجدد • عدم ارزیابی خودکار هستان‌شناسی با مدیریت داده‌های بزرگ‌مقیاس • بروز مشکلات کارآیی و پیچیدگی فرآیند استدلال مبتنی بر هستان‌شناسی
[۶] M. R. Chikhale	۲۰۱۷	▪ ارائه یک طبقه‌بندی سه سطحی مناسب از الگوریتم‌های داده‌کاوی توزیع‌شده ▪ بیان معماری ساده چارچوب داده‌کاوی برای هر طبقه ▪ بهبود سرعت اجرا در مقایسه با الگوریتم داده‌کاوی دیگر	• پردازش محدود عامل‌ها • محدودیت دسترسی به داده‌های محلی (حریم خصوصی) • دشواری ادغام و آماده‌سازی داده‌ها • عامل با قابلیت یادگیری باید از الگوریتم یادگیری و استدلال تغذیه شود.
[۱۵] D. A. Koutsomitropoulos and A. K. Kalou	۲۰۱۷	▪ پرداختن به جزئیات داده‌کاوی توزیع شده ▪ پرداختن به نحوه پردازش داده‌ها در سطح محلی و سراسری	• کاهش پیچیدگی محاسباتی ، بهبود دقت و کارایی در پردازش سراسری • کاهش پیچیدگی فضا و زمان محاسبه در پردازش محلی
[۲۰] Y. Lin, P.-C. Huang, D. Liu, and L. Liang	۲۰۱۷	▪ کاهش هزینه سربار برای ادغام چندین درخت اقلام مکرر محلی برای ساخت درخت اقلام مکرر سراسری به‌منظور افزایش کارآیی و بهره‌وری در کاوش اقلام مکرر	• محدودیت مقیاس‌پذیری روش‌های کاوش از مجموعه داده بسیار بزرگ و تعداد اقلام تولید شده از مجموعه کلان‌داده‌ها
[۷] D. Dou, H. Wang, and H. Liu	۲۰۱۷	▪ طراحی، پیاده‌سازی و اجرای عملیاتی الگوریتم فشرده‌سازی داده‌ها برای الگوریتم پیشنهادشده ▪ رفع نگرانی سطح کاربر در مورد جزئیات پیاده‌سازی با بهره‌گیری از مدل پردازشی نگاشت- کاهش	• عدم انتخاب و استفاده از مجموعه داده‌ای بزرگ در ارزیابی تجربی الگوریتم‌های پیشنهادشده
[۱۶] A. Soylu et al.	۲۰۱۵	▪ پشتیبانی از معماری مبتنی بر عامل برای داده‌کاوی ▪ کاهش زمان محاسبات مبتنی بر معماری عامل‌گرا ▪ افزایش سرعت محاسبات مبتنی بر معماری عامل‌گرا ▪ الگوهای بهینه‌سازی سامانه‌های چندعاملی	• کارآیی پایین • دقت داده‌های استخراج‌شده
[۲۱] Bhamra, Gurpreet S Verma, Anil K Patel, Ram B	۲۰۱۵	▪ ارائه چارچوب‌ها و معماری‌های مبتنی بر عامل با پشتیبانی از استقلال پلتفرم در کاوش قواعد هم‌آیی ▪ کاهش پیچیدگی محاسباتی در چارچوب‌های مبتنی بر عامل ▪ صحت و کارایی بهتر مبتنی بر عامل در کاوش قواعد هم‌آیی	• بیشتر معماری‌ها و چارچوب‌های معرفی‌شده بر پردازش متمرکز استوار است. • تنها مدل نمونه اولیه ارائه‌شده و فاقد اجرا با استفاده از مجموعه داده‌های واقعی است. • ارزیابی و نتایج این مقاله فاقد اعتبار سنجی تجربی است.
[۱۱] D. Apiletti et al.	۲۰۱۴	▪ حل مسئله کاوش اقلام مکرر با استفاده از روش‌های موازی ▪ شناسایی سه چالش مقیاس‌پذیری حافظه، تقسیم کار و توازن بار در طراحی الگوریتم برای استخراج الگوی مکرر	• مقیاس‌پذیری در سطح داده‌های بزرگ‌مقیاس • اعتبارسنجی در مقیاس داده‌های بزرگ‌مقیاس
[۲۲] صابری. حسین و همکاران	۱۳۹۸	▪ ارائه یک معماری ائتلافی مبتنی بر هستان‌شناسی به نام ASMLED برای بهبود عملیات داده‌کاوی ▪ ارائه روشی برای خودکارسازی عملیات کاوش، کاهش پیچیدگی داده‌ها و فرآیندهای کسب و کار	• عدم ارزیابی برخی از مؤلفه‌های معماری ارائه‌شده
[۲۳] مهدوی بصیر. حبیب الله و همکاران	۱۳۹۸	▪ ارائه یک الگوی داده‌کاوی مبتنی بر هستان‌شناسی اهداف برای بهبود معیارهای زمانی و حافظه‌ای در داده‌کاوی	▪ توجه صرف به کاهش داده‌ها در سطح تراکنش‌ها
[۲۴] حیدری یزدی. اشرف السادات	۱۳۹۲	▪ کاوش بر روی داده‌های معنایی ▪ کاوش قوانین انجمنی از روی نمونه داده‌های معنایی	▪ عدم استفاده از مفاهیم موجود در هستان‌شناسی و وجود نداشتن انسجام معنایی در تمام مراحل
[۱۰] فرزان یار. زهرا	۱۳۹۱	▪ توسعه الگوریتم‌های کشف مجموعه اقلام مکرر به محیط‌های همتا به همتای بزرگ‌مقیاس	-

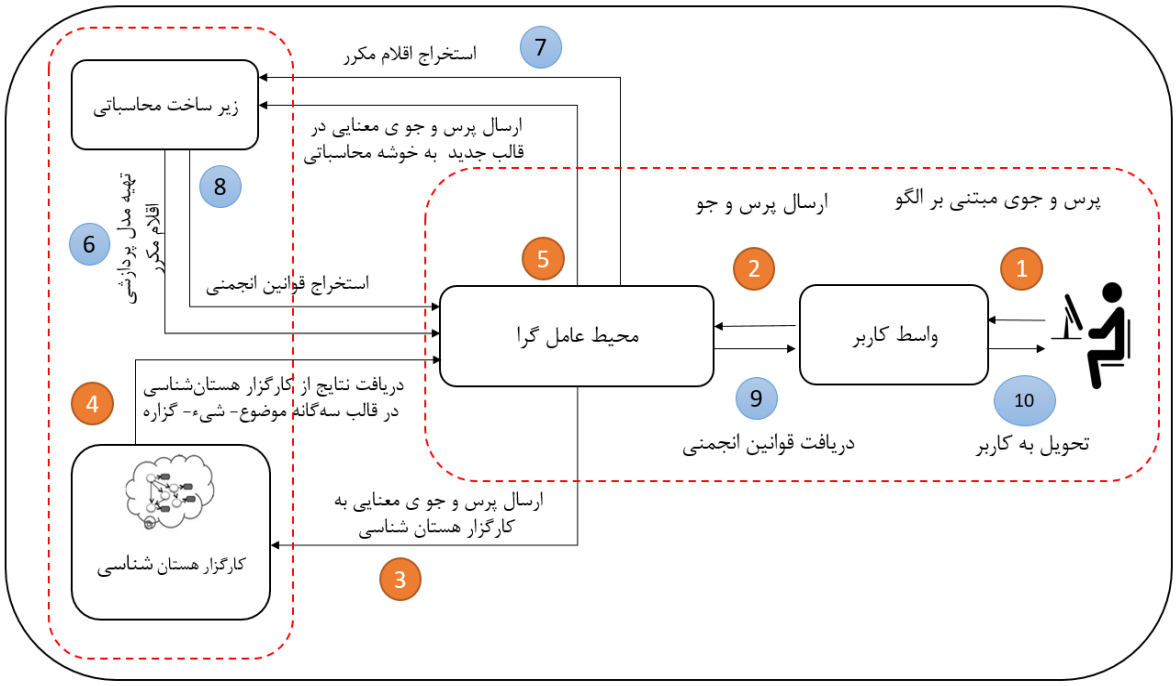
بررسی پژوهش‌های بیان‌شده نشان می‌دهد، برخی از پژوهش‌های بیان‌شده بر مداخله کاربران تأکید بیشتری دارند که به علت حجم زیاد داده‌ها، زمان‌گیر و مستعد خطا هستند. پژوهش‌های انجام‌شده بر موضوعات؛ ۱) راهبردهای تکرار داده برای مدیریت داده‌ها، بهره‌وری و بهبود کارایی دسترسی و زمان پاسخگویی و افزایش قابلیت اطمینان در سامانه‌های داده‌ای توزیع‌شده، ۲) میان‌افزارها و مدیریت منابع، مانند ارتباطات، کشف و زمان‌بندی منابع، امنیت، دسترسی به داده‌ها و تشخیص خطا در محیط محاسباتی توزیع‌شده، ۳) معرفی محیط‌های محاسباتی در پشتیبانی از داده‌کاوی توزیع‌شده و ۴) الگوریتم‌های داده‌کاوی و کشف دانش به روش متمرکز و توزیع‌شده تمرکز دارد.

هدف اساسی مقاله حاضر، ارائه یک روش ائتلافی متشکل از ۱) روش مبتنی بر پرس‌وجو (دریافت پرسش از کاربر و پردازش معنایی آن، ۲) روش مبتنی بر پردازش محلی توسط هر گره شبکه، ۳) روش مبتنی بر هستان‌شناسی کسب و کار و ۴) روش مبتنی بر داده‌کاوی چندعاملی در قالب روش DARMASO است.

۳- روش پیشنهادی

روش پیشنهادی که DARMASO نامیده می‌شود، بر مبنای تولید اقلام مکرر برای یافتن قواعد هم‌آیی معنایی مؤثر از داده‌های بزرگ‌مقیاس از سه لایه مختلف (کاربر- محاسبات- داده) تشکیل شده است. اخذ داده‌ها از پایگاه دانش معنایی تا مرحله تولید قواعد هم‌آیی معنایی، در این سه لایه انجام می‌شود. این رویکرد به‌طور کامل مطابق با موازی‌سازی مبتنی‌بر چارچوب محاسباتی نگاشت-کاهش و سیستم چندعاملی مبتنی چارچوب عامل‌گرایی است. نتیجه این رویکرد، تولید قواعد هم‌آیی معنایی در سطح سراسری از مجموعه اقلام مکرر محلی در یک محیط توزیع‌شده عامل‌گرا است. در این روش، داده‌ها در مقیاس بزرگ و در بین پارتیشن‌های موجود در منابع داده‌ای توزیع‌شده قرار دارند و از قواعد تولیدشده سراسری برای تصمیم‌های سراسری و از قواعد تولیدشده محلی برای تصمیم‌های محلی استفاده می‌شود.

ایده اصلی در روش DARMASO این است که به‌منظور کاهش فضای کاوش و همچنین تعداد مجموعه اقلام مکرر تولید شده، هرس الگوهای غیر مرتبط از محاسبه مجموعه‌های اقلام مکرر محلی توسط هستان‌شناسی انجام شود. در این رویکرد، ابتدا هر کاندیدایی که عناصر آن به لحاظ معنایی نزدیک به یک مفهوم هستان‌شناسی و در سلسله مراتب مفاهیم آن است باقی‌مانده و سایر آن‌ها از محاسبه مجموعه اقلام مکرر حذف می‌شود. در انتها نیز فقط قواعد هم‌آیی مفید نگه‌داری می‌شود. اهمیت روش پیشنهادی این است که اجرای فرآیندها از نظر منطقی و عملیاتی با کمترین اختلال (مثلاً از بین رفتن گره‌های داده‌ای و محاسباتی) وجود دارد و می‌توان ساختار معنایی را برای داده‌کاوی توزیع‌شده با پشتیبانی عامل‌ها و هستان‌شناسی، طراحی، اجرا، پردازش، ذخیره‌سازی و بازیابی کرد. شکل (۲) فرآیند کاوش قواعد هم‌آیی معنایی در DARMASO را نشان می‌دهد.



شکل (۲): فرآیند DARMASO

۳-۱- مراحل DARMASO

اجرای گام به گام روش DARMASO به ۱۰ مرحله تقسیم می‌شود. همه مراحل و تعاملات لازم بین مؤلفه‌های این رویکرد مبتنی بر هستان‌شناسی کسب‌وکار و در زیرساخت محاسباتی توزیع‌شده و محیط محاسباتی عامل‌گرا انجام می‌شود. فعالیت هر مرحله به تفکیک در جدول (۲) بیان شده است.

۳-۲- تعریف‌ها و مفاهیم پایه در DARMASO

کاوش اقلام مکرر توزیع‌شده به تولید قواعد هم‌آیی از مجموعه اقلام مکرر محلی در یک محیط توزیع‌شده می‌پردازد. داده‌ها در پارتیشن‌های مختلف در قالب منابع داده توزیع‌شده قرار دارند و قواعد سراسری برای تصمیم‌های سراسری و قواعد محلی برای تصمیم‌های محلی استفاده می‌شود. در ادامه تعاریف موردنیاز برای بیان الگوریتمی و درک بهتر ارائه خواهد شد. تعاریف و مفاهیم پایه برای توصیف روش DARMASO در جدول (۳) بیان‌شده است.

۳-۳- الگوریتم‌های DARMASO

الگوریتم‌های DARAMASO شامل ۱) الگوریتم فرآیند کاوش و تجمیع قواعد هم‌آیی به نام DARMASOMAIN، ۲) الگوریتم کاهش و هرس معنایی داده‌ها به نام DARMASOPRU، ۳) الگوریتم استخراج و تجمیع قواعد هم‌آیی به نام DARMASOINT است که بر اساس روندنمای شکل (۳) در شکل‌های (۴-۶) بیان شده است.

جدول (۲): مراحل DARMASO

مراحل	اقدام هر مرحله از عملکرد معماری پیشنهادی
۱	کاربر درخواست خود را از طریق واسط کاربری و در قالب یک مجموعه ویژگی[1] بیان می‌کند.
۲	واسط کاربر درخواست را به عامل هماهنگ‌کننده در محیط محاسباتی عامل‌گرا ارسال می‌کند.
۳	عامل هماهنگ‌کننده سؤال کاربر را به یک ساختار معنایی در قالب زبان اسپار- کیو- ال[2](یک زبان پرس و جو 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 هماهنگ‌کننده در محیط محاسباتی عامل‌گرا بازگشت می‌دهد.
۵	لیست داده‌ای دریافت شده از مرحله ۴ به یک ساختار توصیفی مناسب تبدیل و مدل داده‌ای مناسبی 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[3]، رشد اقلام تکراری[4]، ریلیم[5]، دی‌فین[6]غیره) برروی مدل داده‌ای توصیفی اخذشده از مرحله ۵، مجموعه اقلام مکرر را تولید می‌کند.
۸	زیر ساخت محاسباتی، مدل داده‌ای تولیدشده 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 هماهنگ‌کننده تحویل داده می‌شود.
۹	عامل هماهنگ‌کننده، قواعد هم‌آیی معنایی تولیدشده را جهت تحویل به کاربر به عامل واسط کاربر تحویل می‌دهد.
۱۰	پردازش‌های خفیف‌تر که به‌عنوان پسا پردازش شناخته می‌شود (مانند مرتب‌سازی و یا ارائه با فرمت خاصی مانند csv. و غیره) در این مرحله انجام می‌شود.

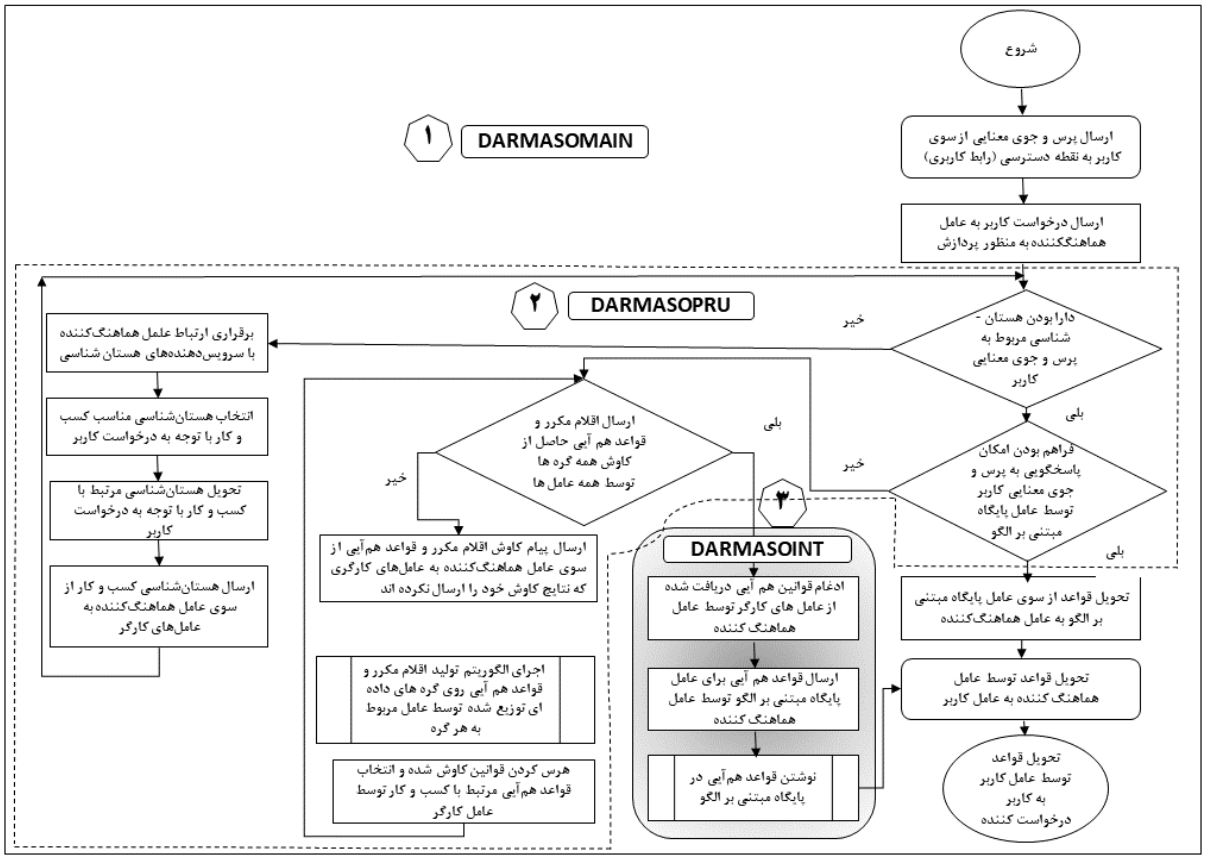
[1] Feature Set
[2] SPARQL
[3] Apriori
[4] FP-growth
[5] Recursive Elimination (RElim)
[6] DFIN

جدول (۳): تعریف‌ها و مفاهیم پایه در روش DARMASO

ردیف	نشانه	تعریف
۱	$Ag_{coordinator}$	عامل هماهنگ کننده
۲	$Ag_{repo-manager}$	عامل مدیر مخزن اقلام مکرر
۳	$Ag_{Ont-Server}$	عامل کارگزار هستان‌شناسی
۴	$Ag^{\square}_{worker}\ or\ Ag^{S}_{w} = \{Ag^{S}_{w1}, Ag^{S}_{w2}, Ag^{S}_{wn} \mid S = \{Active\ or\ A, Disable\ or\ D\}\}$	عامل‌های داده‌کاوی/کارگر
۵	$LSD = \left\{\left\{\sum_{i=1}^{m} D_i \mid \forall\ D_i, 1 < i < m, D_i\right\}\right\}$	مجموعه داده بزرگ‌مقیاس که از تجمیع تعداد مشخصی گره (m گره) از گره های داده‌ای توزیع شده ساخته می‌شود.
۶	$D_i = \left\{\left\{\sum_{j=1}^{k} En_j \mid \forall\ En_j, 1 < j < k, En_j\right\}\right\}$	هر گره داده (به غیر از مجموعه داده بزرگ مقیاس) متشکل از داده‌های موجودیت (k موجودیت) است.
۷	$Op^{Ag^{i}_{w}}_{Type} = \left\{Type \in \{M, Ont, OntS\} \middle\vert \begin{matrix} M = \{Mining\}, OntS = \{Ontology\ Selection\}, \\ Recieve\ Ont = \{Recieve\ Ontology\}, WK = \{Write\ Knowledge\ on\ Repository\ \}, \\ RK = \{Read\ Knowledge\ from\ Repository\}, OBDR = \{Ontology - Based\ Data\ Reduction\} \end{matrix}\right\}$	هریک از عامل‌ها، عملیاتی را انجام می‌دهند. بسته به نوع عملیات (s) و این که کدام عامل(Ag^{i}_{w}) این عملیات را انجام می‌دهد به وسیله نشانه $Op^{Ag^{i}_{w}}_{s}$ متمایز و تعریف می‌شود.
۸	$Ont^{BT}_{L}, BT = \{Earthquake, Car, \ldots\}\ and\ L = \{Global, Local\}$	اعمال هستان شناسی روی قواعد هم‌آیی استخراج شده از اقلام مکرر و یا داده‌ها می بایست متناسب با هستان‌شناسی کسب و کار(BT) و با مد نظر قرار دادن سطح هستان شناسی(L) انجام شود که با نشانه Ont^{BT}_{L} تعریف می‌شود.
۹	$FP_{D_i}\ or\ FP(D_i)$	مجموع اقلام مکرر تولید شده حاصل از کاوش گره داده D_i
۱۰	$AR_{D_i}\ or\ AR(D_i)$	لیستی از قواعد هم‌آیی استخراج شده و هرس شده از اقلام مکرر تولید شده در گره داده D_i را نشان می‌دهد.
۱۱	$AR_{Total}\ or\ AR(Total)$	لیستی از قواعد هم‌آیی هرس شده بر اساس هستان شناسی کسب و کار را نشان می‌دهد.
۱۲	$SAR_{Total}\ or\ SAR(Total)$	وضعیت لیستی از قواعد هم‌آیی هرس شده بر اساس هستان شناسی کسب و کار (خالی، ناقص،کامل) را نشان می‌دهد.
۱۳	$SM_{Ag}(O)$	ارسال یک خروجی (O)، از طریق پیام توسط یک عامل (Ag) را نشان می‌دهد.

۳-۴- روندنمای DARMASO

در شکل (۳) روندنمای DARMASO در سه بخش و به‌صورت گام به گام بیان شده است.



شکل (۳): روندنمای DARMASO

۳-۴-۱- الگوریتم DARMASOMAIN

Send User's Query to User Interface
Assign User's Query to $Ag_{coordinator}$
If (Query_ Result==true) then
Run($Op_{RK}^{Ag_{repo-manager}}$ **)**
$Ag_{coordinator}$ = $SM_{Ag_{repo-manager}}$ (AR_{Total})
User= $SM_{Ag_{coordinator}}$ (AR_{Total})
End if
Else While (Query_ Result ==False) do
While (Ontology_ Selection==False) do
$Ag_{Ont-Server}$ = $SM_{Ag_{coordinator}}$ ($Op_{OntS}^{Ag_{coordinator}}$)
Run ($Op_{OntS}^{Ag_{coordinator}}$)
For i=1, Ag_{wi}^{A}**.counts() do**
Ag_{wi}^{A}= $SM_{Ag_{coordinator}}$ ($Op_{Recieve\_Ont}^{Ag_{wi}^{A}}$)
Ag_{wi}^{A} = Run ($Op_{OBDR}^{Ag_{wi}^{A}}$)
End For
End While
Case SAR_{Total} **do**
Case Empty
For i=1, Ag_{wi}^{A}**.counts() do**
Ag_{wi}^{A}=$SM_{Ag_{coordinator}}$ ($Op_{M}^{Ag_{w}^{i}}$)
Run($Op_{M}^{Ag_{w}^{i}}$**)**
End For
Case Incomplete
i= Ag_{w}^{D}**.counts**
For i=1, Ag_{wi}^{D}**.counts() do**
Ag_{wi}^{D}=$SM_{Ag_{coordinator}}$ ($Op_{M}^{Ag_{wi}^{D}}$)
Run($Op_{M}^{Ag_{wi}^{D}}$**)**
End For
Case Complete
For i=1, Ag_{w}^{A}**.counts() do**
$AR_{Total} = AR_{Total} + AR_{D_i}$
End For
End Case
$Ag_{repo-manager}$= $SM_{Ag_{coordinator}}$ (AR_{Total})

For i=1, Ag_{w}^{A}**.counts() do**
$Ag_{coordinator}$ = SM_{Ag} (AR_{D_i})
Run($Op_{WK}^{Ag_{repo-manager}}$**)**
Query_ Result=true
End While
User= $SM_{Ag_{coordinator}}$ (AR_{Total})
End Else
Finish

شکل (۴): الگوریتم DARMASOMAIN

۳-۴-۲- الگوریتم کاهش و هرسمعنایی DARMASOPRU

Start
Input: Failure to Select a Business Ontology Appropriate to the User's Question
Output: Select Business Ontology, Send it to User Nodes, and Reduce Data Accordingly
While (Ontology_Selection==False) do
$Ag_{Ont-Server}$ = $SM_{Ag_{coordinator}}$ ($Op_{OntS}^{Ag_{coordinator}}$)
Run ($Op_{OntS}^{Ag_{coordinator}}$)
For i=1, Ag_{wi}^{A}**.counts() do**
Ag_{wi}^{A} = $SM_{Ag_{coordinator}}$ ($Op_{Recieve\_Ont}^{Ag_{wi}^{A}}$)
Ag_{wi}^{A} = Run ($Op_{OBDR}^{Ag_{wi}^{A}}$)
End For
End While
End

شکل (۵): الگوریتم DARMASOPRO

۳-۴-۳. الگوریتم ادغام و تجمیع قواعد DARMASOINT

Start
Input: Decreased Ontology-based Data in Distributed Data Node
Output: Extraction and Aggregation of Rules from Reduced Ontology Based Data on Distributed Data Nodes
Case SAR_{Total} **do**
Case Empty
For i=1, Ag_{wi}^{A}**.counts() do**
Ag_{wi}^{A} = $SM_{Ag_{coordinator}}$ ($Op_{M}^{Ag_{w}^{i}}$)
Run($Op_{M}^{Ag_{w}^{i}}$**)**
End For
Case Incomplete
i= Ag_{w}^{D}**.counts**
For i=1, Ag_{wi}^{D}**.counts() do**
Ag_{wi}^{D} = $SM_{Ag_{coordinator}}$ ($Op_{M}^{Ag_{wi}^{D}}$)
Run($Op_{M}^{Ag_{wi}^{D}}$**)**
End For
Case Complete
For i=1, Ag_{w}^{A}**.counts() do**
$AR_{Total} = AR_{Total} + AR_{D_i}$
End For
End Case
End

شکل (۶): الگوریتم DARMASOINT

۴- پیاده‌سازی DARMASO

پیاده‌سازی DARMASO و فرآیند عملیاتی آن نیازمند زیرساخت‌های محاسباتی و ابزارهایی است که هرکدام دربردارنده بخشی از مشخصه‌های مهم این مقاله است. از آنجایی که کاوش مجموعه اقلام مکرر از نظر محاسباتی پرهزینه است، راهبردهای توزیع‌شده عامل‌گرا و موازی‌سازی بر استفاده بهینه حافظه، متعادل‌سازی بار پردازشی و هزینه‌های ارتباطی تأثیر می‌گذارد. به همین منظور در پیاده‌سازی DARMASO از روش محاسباتی نگاشت-کاهش و محیط محاسباتی عامل‌گرا استفاده شده است.

برای انجام و پشتیبانی عملیات داده‌کاوی از داده‌های بزرگ‌مقیاس، باید از روش‌ها و الگوریتم‌های کاوش توزیع‌شده استفاده نمود. یکی از چالش‌های کاوش از داده‌های توزیع‌شده، محاسبات حجیم و به دنبال آن نیاز به زیرساخت‌های محاسباتی متناسب است. اگر چه فراهم‌سازی این زیرساخت نیازمند هزینه فراوان است، اما در اختیار داشتن آن موجب افزایش کیفیت سرویس (عملکرد) سامانه‌های کاوش می‌شود.

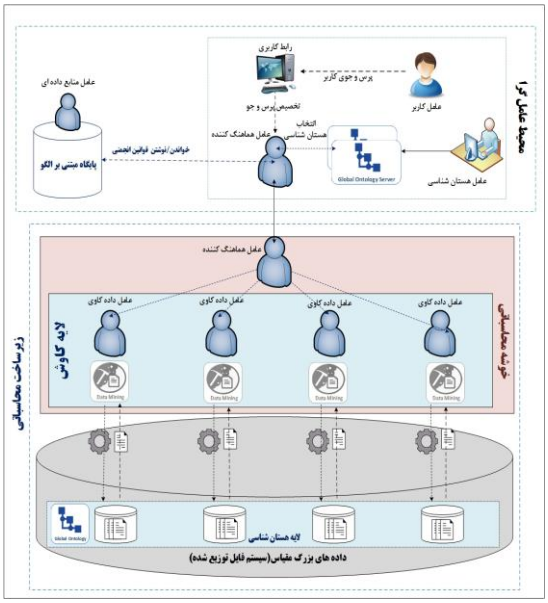
اولین ویژگی روش پیشنهادیDARMASO استفاده از هستان‌شناسی کسب‌وکار و حذف داده‌های غیر مرتبط و یا کمتر مرتبط با اهداف کسب‌وکار از گره‌های توزیع‌شده‌ای است که داده‌ها در آن ذخیره‌شده است. این امر موجب کاهش میزان داده ارسال‌شده به موتور کاوش می‌گردد و با کاهش حجم محاسبات، نیاز کمتر به منابع محاسباتی و همچنین افزایش عملکرد موتور کاوش را امکان‌پذیر می‌کند.

دومین ویژگی روش DARMASO حمایت صحیح از اهداف کسب‌وکار است. الگوریتم‌های کاوش مبتنی بر قواعد هم‌آیی معمولاً اقلام مکرر بسیاری تولید می‌نمایند. بدیهی است درصورتی‌که همه داده‌های مورد کاوش با اهداف کسب‌وکار مرتبط نباشد، کسب‌وکار به‌درستی مورد حمایت قرار نمی‌گیرد. به همین دلیل است که در این مقاله برای رفع این چالش، الگوریتم‌های کاوش اقلام مکرر، قواعد هم‌آیی و هستان‌شناسی توأماً به خدمت گرفته‌شده است.

با توجه به اینکه هستان‌شناسی بر اساس یک الگوی مشخص ارتباطات موجود میان داده‌ها را توصیف می‌نماید، در این مقاله، الگوریتم‌های کاوش اقلام مکرر توزیع‌شده توسط هستان‌شناسی حمایت می‌شود. در همین راستا دو الگوریتم DARMASOPRO و DARMASOINT نسبت به هرس داده‌های غیر مرتبط با اهداف کسب‌وکار و تجمیع قواعد هم‌آیی معنایی کاوش شده اقدام می‌کند. از آنجایی‌که DARMASO داده‌های بزرگ‌مقیاس را مورد کاوش قرار می‌دهد و داده‌ها در گره‌های مختلف توزیع‌شده است، بنابراین هستان‌شناسی توسط عامل‌های مرتبط روی همه گره‌ها به‌صورت جداگانه و هم‌زمان اعمال می‌گردد. شایان‌ذکر است عامل‌ها در این مقاله از پیچیدگی‌های توزیع‌شدگی کاسته و موجب افزایش سرعت در کاوش می‌گردد.

۴-۱- معماری عملیاتی DARMASO

در DARMASO ، هادوپ-اسپارک[1] یک زیرساخت نرم‌افزاری است که اجازه می‌دهد، روش نگاشت- کاهش را به شیوه‌ای قابل اجراء در خوشه‌ای از ماشین‌ها اجرا کرد. چارچوب نرم‌افزاری عامل‌گرای جید[2] نیز محیط چندعاملی را برای تعریف و تعامل عامل‌ها فراهم می‌نماید. شکل (۷) معماری عملیاتی DARMASO را نشان می‌دهد. فرآیند کاوش و تحویل قواعد کاوش شده متناسب با پرس‌وجوی کاربر توسط مجموعه‌ای از عامل‌ها در یک محیط محاسباتی عامل گرا پیاده‌سازی شده است. در این معماری هر عامل برحسب نوع وظیفه‌ای که بر عهده دارد (مانند کاهش داده‌ها متناسب با هستان‌شناسی کسب‌وکار و کاوش قواعد هم‌آیی) به ایفای نقش می‌پردازد. داده‌ها به‌صورت پراکنده در محیط محاسباتی توزیع‌شده ذخیره‌شده و هر گره داده‌ای دارای یک عامل کارگر است. در محیط محاسباتی عامل‌گرا، عامل هماهنگ‌کننده‌ای وجود دارد که وظیفه هماهنگی کلیه عامل‌ها با یکدیگر را برای به سرانجام رساندن فرآیند کاوش و تحویل قواعد متناسب با پرس‌وجوی کاربر به عهده دارد.



شکل (۷): معماری عملیاتی DARMASO

[1] Hadoop-Spark
[2] Jade

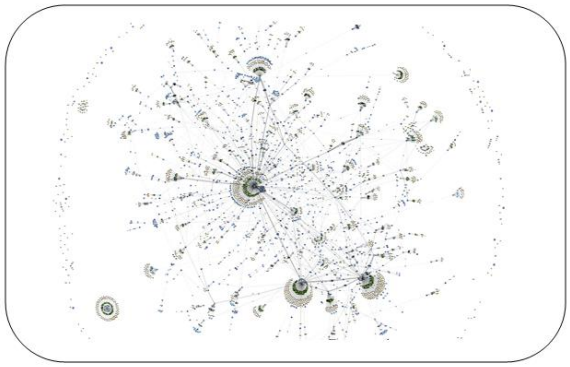
ازجمله کارهای مهم عامل هماهنگ‌کننده، دریافت پرس‌وجوی معنایی از کاربر، بررسی وجود و یا عدم وجود هستان‌شناسی متناسب با آن، دریافت هستان‌شناسی کسب‌وکار متناسب با پرس‌وجوی کاربر، مدیریت فرآیند کاوش قواعد هم‌آیی از داده‌ها، ادغام قواعد هم‌آیی کاوش شده از داده‌ها و تحویل قواعد متناسب با پرس‌وجو به کاربر است. عامل‌های دیگری همچون عامل مهندسی هستان‌شناسی و عامل پایگاه مبتنی بر الگو (پایگاه دانش) نیز در روش پیشنهادی وجود دارند که خدمات دیگری را انجام می‌دهند. عامل مهندسی هستان‌شناسی، عامل هماهنگ‌کننده را در انتخاب و دریافت هستان‌شناسی کسب‌وکار متناسب با پرس‌وجوی کاربر یاری می‌کند. عامل پایگاه مبتنی بر الگو نیز با خواندن قواعد موجود در پایگاه و همچنین نوشتن قواعد جدید در پایگاه دانش، عامل هماهنگ‌کننده را در تحویل سریع‌تر الگوها و قواعد درست به کاربر یاری می‌دهد. اجرای معماری عملیاتی روش پیشنهادی شامل ۱) آماده‌سازی مجموعه داده‌ها مبتنی بر هستان‌شناسی، ۲) آماده‌سازی محیط محاسباتی توزیع‌شده، ۳) آماده‌سازی محیط عامل‌گرای جید و ۴) ورودی‌ها و خروجی‌های معماری عملیاتی DARMASO انجام پذیرفت.

۴-۱-۱- آماده‌سازی مجموعه داده‌ها مبتنی بر هستان‌شناسی

در این مرحله، مجموعه داده[1] پایگاه دانش معتبر دی بی پدیا[2] نسخه ۲۰۱۷ از انتشار پایدار[3] ایجادشده به حجم ۲۰ گیگابایت حاوی میلیون‌ها داده سه‌گانه هستان‌شناسی در قالب RDF/N3 به همراه ۵ گیگابایت ایندکس و ۲٫۵ مگابایت فایل OWL برای شناخت کلاس‌ها، زیر کلاس‌ها، ارتباطات و جزئیات استفاده از مجموعه داده آزمایشی پایگاه دانش دی بی پدیا جهت پیاده‌سازی پیش‌پردازش و آماده گردید. برای آماده‌سازی، فایل هستان‌شناسی دی بی پدیا با استفاده از نرم‌افزار پروتیج[4] قرائت گردید. شناسایی متریک‌ها متشکل از اصول[5]، کلاس‌ها، اشیاء[6]، ویژگی‌ها[7] و روابط موجود در این هستان‌شناسی مورد بازبینی و مطالعه قرار گرفت. مجموعه داده‌ای آماده‌شده، اجازه استخراج ساختاری از محتوای اطلاعات ایجادشده را فراهم کرده و اجازه می‌دهد به‌طور معنایی روابط و خواص منابع مجموعه داده‌ها ازجمله پیوندها به سایر مجموعه‌های مرتبط کاوش شود.

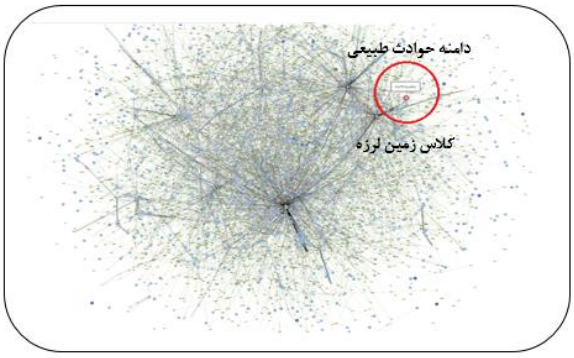
[1] Data Set
[2] DBpedia
[3] Stable Release
[4] Protégé
[5] Axioms
[6] Objects
[7] Property

شکل (۸) تصویر کاملی از گستردگی هستان‌شناسی دی بی پدیا نشان می‌دهد.



شکل (۸): هستان‌شناسی دی بی پدیا

در ادامه، شکل (۹) نقطه کلاس هستان‌شناسی زمین‌لرزه را در پایگاه دانش دی بی پدیا نشان می‌دهد.



شکل (۹): کلاس هستان‌شناسی زمین‌لرزه

کلاس هستان‌شناسی زمین‌لرزه پایگاه دانش DBpedia متشکل از هشت نگاشت ویژگی[8]، شامل ویژگی الگو[9] و ویژگی هستان‌شناسی[10] است. جزئیات کلاس و مشخصه‌های انتخاب‌شده مورد مطالعه هستان‌شناسی زمین‌لرزه در جدول (۶) نشان داده شده است. استخراج داده‌های مورد نظر با برنامه‌نویسی و به کمک کتابخانه jena (یک کتابخانه جاوای متن‌باز است که برای پردازش داده‌های RDF در وب معنایی و داده‌های پیوندی به‌کار می‌رود. این کتابخانه علاوه بر امکان ایجاد مدل‌های مختلف هستان‌شناسی شامل RDFS و OWL، پردازش پرس‌وجوهای زبان SPARQL و استنتاج مبتنی بر قواعد را نیز حمایت می کند) صورت گرفته است. سپس با هرس مشخصه‌های مورد نظر، رویدادهای زمین‌لرزه جداسازی شده است. مشخصه‌های انتخاب‌شده در جدول (۴) نشان داده شده است.

[8] Property Mapping
[9] Template Property
[10] Ontology Property

جدول (۴): الگوی انتخاب‌شده از پایگاه دانش دی- بی- پدیا

ردیف	ویژگی هستان‌شناسی	ویژگی الگوی تعریف‌شده
۱	foaf:name	نام کشور وقوع زمین‌لرزه
۲	foaf:depiction	توصیف زمین‌لرزه وقوع یافته
۳	date	تاریخ وقوع زمین‌لرزه
۴	time	زمان وقوع زمین‌لرزه
۵	duration	مدت زمان زلزله
۶	location	محل جغرافیایی وقوع زمین‌لرزه
۷	damage	میزان خسارت زلزله
۸	casualties	میزان تلفات زلزله

به‌منظور کشف و نمایش دقیق ارتباط بین موجودیت‌های مختلف در هستان‌شناسی زمین‌لرزه و همچنین ارتباط با سایر هستان‌شناسی‌ها، این هستان‌شناسی در ویرایشگر پروتیج قرائت و اجزاء آن به همراه ارتباطات کامل آن استخراج گردید.

۴-۱-۲- آماده‌سازی محیط محاسباتی توزیع‌شده

از آنجایی که کاوش مجموعه اقلام مکرر از نظر محاسباتی پرهزینه است، راهبردهای توزیع‌شدگی، عامل‌گرایی و موازی‌سازی بر استفاده بهینه حافظه، متعادل‌سازی بار پردازشی و هزینه‌های ارتباطی تأثیر می‌گذارد. به همین منظور در پیاده‌سازی DARMASO از روش پردازشی نگاشت-کاهش مبتنی بر چارچوب نرم‌افزاری موازی و توزیع‌شده هادوپ[1] و محیط محاسباتی عامل‌گرای مبتنی بر چارچوب نرم‌افزاری جید استفاده شده است. هادوپ متشکل از دو بخش اصلی است که وظیفه پردازش و ذخیره‌سازی داده‌های بزرگ‌مقیاس را بر عهده دارد. بخش اول که سامانه فایل توزیعی هادوپ[2] نام دارد که وظیفه ذخیره‌سازی داده‌ها (پایگاه هستان‌شناسی، پایگاه مبتنی بر الگو، داده‌های بزرگ‌مقیاس) را بر عهده دارد. بخش دوم خوشه محاسباتی که شامل عامل‌های کارگر (به ازای هر گره داده یک عامل) است و این بخش وظیفه انجام پردازش یا محاسبات گوناگون روی داده‌های ذخیره‌شده در سامانه فایل توزیع‌شده هادوپ را بر عهده دارد. اعمالی همچون خواندن، کاوش و نوشتن داده‌ها توسط عامل‌های کارگر در این بخش انجام می‌شود. زیرساخت محاسباتی محیط عامل‌گرا نیز به‌صورت عملیاتی بر روی چهار گره ماشین مجازی که هر گره مجازی در یک ماشین فیزیکی متفاوت قرار دارد پیاده‌سازی شده است. این چهار گره یک خوشه را تشکیل می‌دهند که در این خوشه یک گره اصلی[3] (مدیر خوشه) و سایر خوشه‌ها نقش گره‌های تابع را دارا هستند. در هر گره بستر محاسباتی توزیع‌شده هادوپ نسخه ۲,۷,۳ به‌منظور پشتیبانی از بخش اول زیرساخت محاسباتی و بستر محاسباتی اسپارک[4] نسخه ۲,۴ به‌منظور پشتیبانی از بخش دوم زیرساخت محاسباتی راه‌اندازی شده است. هر چهار گره، نقش عامل کارگر و گره داده‌ای را ایفاء می‌کنند. همچنین گره اصلی، نقش عامل هماهنگ‌کننده را نیز ایفاء می‌نماید. پیکره‌بندی ماشین‌های مجازی برای گره‌های زیرساخت محاسباتی متشکل از چهار گره (هر گره دارای ۱۰گیگابایت ظرفیت حافظه اصلی، ۱۵۰گیگابایت ظرفیت ذخیره‌سازی و چهار هسته ۱۴ گیگاهرتز ظرفیت پردازشی) است.

۴-۱-۳- آماده‌سازی محیط عامل گرای جید

در پیاده‌سازی DARMASO از محیط اجرایی اینتلی‌جی آیدیا[5] برای پیاده‌سازی محیط چندعاملی و ارتباط با محیط محاسباتی هادوپ اسپارک استفاده شده است. عامل‌های طراحی‌شده (عامل هماهنگ‌کننده، عامل پایگاه مبتنی‌بر الگو، عامل مهندسی هستان‌شناسی و عامل کاربر) در بستر محیط عامل‌گرای جید پیاده‌سازی شده است و این عامل‌ها بر اساس وظایفی که بر عهده دارند، با محیط محاسباتی هادوپ تعامل می‌کنند. ارتباط بین محیط جید و هادوپ که از نوآوری‌های پیاده‌سازی DARMASO نیز محسوب می‌شود، توسط عامل تعریف‌شده، از طریق تبادل و ارسال پیام بین دو محیط جید و اکوسیستم هادوپ فراهم می‌شود. درخواست‌های سطح کاربر توسط عامل تعریف‌شده در محیط جید برای اجرا به عامل اسپارک در اکوسیستم محاسباتی هادوپ تحویل داده می‌شود.

۵- ارزیابی و نتایج محاسباتی

در این بخش بر اساس روش‌های ارزیابی کمی و کیفی، DARMASO ارزیابی و اعتبارسنجی شده است. این ارزیابی و اعتبارسنجی برای مؤلفه هستان‌شناسی به‌صورت کیفی و برای سایر مؤلفه‌های روش پیشنهادی به‌صورت کمی و تجربی می‌باشد.

۵-۱- ارزیابی هستان‌شناسی

ارزیابی هستان‌شناسی به روش پارامتری، اطمینان از این است که مفاهیم و ارتباطات معنایی به‌درستی طراحی و تعریف شده است. در این مقاله برای ارزیابی هستان‌شناسی، دو معیار اصلی و پنج معیار پشتیبان در نظر گرفته شده است.

- معیار اول سازگاری[6] است. به این مفهوم که تناقض در کلاس‌ها یا عناصر هستان‌شناسی وجود نداشته و داده‌های ارائه‌شده نتیجه‌های متناقض را ارائه ندهند.

[1] Hadoop Infrastructure
[2] Hadoop Distributed File System
[3] Master Node
[4] Spark
[5] Intellij Idea
[6] Consistency

- معیار دوم کامل بودن[1] هستان‌شناسی است، به‌طوری‌که هستان‌شناسی هیچ خطایی در حوزه کامل بودن نداشته باشد. معیارهای شفافیت، وضوح، عمومیت، استحکام و غنای اطلاعات معنایی از دیگر معیارهای پشتیبان در این ارزیابی است.

در نتیجه با در نظر گرفتن معیارهای معرفی‌شده و تطبیق آن با هستان‌شناسی موجود مانند، YAGO، WikiData، WordNet، پایگاه دانش DBpedia به دلیل برخورداری از معیارهای هفت‌گانه بیان‌شده انتخاب و با تطبیق دادن با سناریوی مورد نظر در مراحل پیاده‌سازی مورد استفاده قرار گرفت.

۵-۲- ارزیابی تجربی

در این بخش به ارزیابی تجربی DARMASO پرداخته می‌شود. ارزیابی بر اساس سه معیار پشتیبان، معیار اطمینان و معیار بالابری[2] قواعد انجام می‌شود.

- **معیار پشتیبان قواعد**: احتمال رخداد آیتم‌های موجود در قواعد در یک موقعیت زمانی است. این معیار میزان کاربردی بودن و سودمند بودن[3] قواعد را بیان می‌کند.

$$\text{Support}\ (A \rightarrow B) = P(A \cap B) \qquad (۱۲)$$

$$\text{Support}(A \rightarrow B) = \text{Support}(B \rightarrow A) \qquad (۱۳)$$

- **معیار اطمینان قواعد:** احتمال رخداد آیتم‌های تالی به‌شرط رخداد آیتم‌های مقدم است که قابلیت اطمینان، یقین و قطعیت[4] قواعد را بیان می‌کند.

$$\text{Confidence}\ (A \rightarrow B) = P(B \mid A \qquad (۱۴)$$

$$\text{Confidence}\ (A \rightarrow B) = P(B \mid A) = \frac{Support(A \cap B)}{Support\ (A)} \qquad (۱۵)$$

- **معیار بالابری قواعد:** میزان هم اتفاقی بین ویژگی‌ها را در نظر می‌گیرد و میزان رخداد تکی بخش تالی قواعد را در محاسبات دخالت می‌دهد. این معیار میزان منطقی بودن و کارآیی قواعد هم‌آیی را بیان می‌کند.

$$\text{Lift}\ (A \rightarrow B) = \frac{Confedence(A \rightarrow B)}{Support\ (B)} \qquad (۱۶)$$

- **سطح مفهومی قواعد:** هر چه قوانین به اهداف نزدیک‌تر باشد در سطح مفهومی بالاتری قرار گرفته و کمک بیشتری به تصمیم‌گیری می‌کند. لذا با اجرای DARAMSO، سطح مفهومی و معنایی قواعد افزایش می‌یابد.

در ارزیابی تجربی، داده‌های پایگاه دانش دی بی پدیا برای دامنه حوادث طبیعی و کلاس زمین‌لرزه انتخاب گردید. اجرای عملیاتی DARMASO در سناریویی به تعداد ۴۳۸,۳۳۶,۴۱۹ رکورد اولیه در سه مرحله برای کاهش فضای کاوش و استخراج اقلام مکرر و تولید قواعد هم‌آیی معنایی مفید انجام شده است.

- در مرحله اول کاهش، پس از اعمال هستان‌شناسی، ۳۶۰۰۰ رکورد معنایی از دامنه حوادث طبیعی فیلتر شد.
- در مرحله دوم کاهش، ابتدا رکوردهای استخراج‌شده مرحله قبل به یک ساختار داده‌ای قابل‌پردازش توسط الگوریتم کاوش اقلام مکرر تبدیل گردید. در ادامه این رکوردها با در نظر گرفتن حداقل آستانه پشتیبان و پرس‌وجوی معنایی کاربر و صفات درخواست شده تطبیق داده شد. مجموعاً ۱۱ الگو با مجموع ۱۸۴۹ تکرار، متناسب با پرس‌وجوی معنایی کاربر و مبتنی بر کلاس هستان‌شناسی زمین‌لرزه فیلتر گردید. جدول (۵) نمونه‌ای از مجموعه اقلام مکرر سراسری کاوش شده در این مرحله را نشان می‌دهد.
- در مرحله سوم کاهش، با در نظر گرفتن حداقل آستانه اطمینان، الگوریتم کاوش اجرا و قواعد هم‌آیی معنایی سراسری استخراج گردید. نمونه این قواعد که بر اساس درخواست کاربر، کمیت‌های توصیف شده در جدول (۶) و معیارهای سه‌گانه ارزیابی کاوش شده، در جدول (۷) نشان داده شده است.

جدول (۵): نمونه‌ای از اقلام مکرر کاوش شده در کاهش دوم

الگوهای مکرر کاوش شده در مرحله دوم کاهش	تعداد تکرار هر الگوی مکرر	مرحله دوم کاهش (مجموع تکرار الگوی کاوش)	مرحله دوم کاهش (تعداد الگوهای کاوش شده)	مرحله اول کاهش (تعداد رکوردهای حوادث)	رکوردهای معنایی اولیه (تعداد کل رکوردها)
[Min-Support:.۰۲]					
[Magnitude->Strong]	۴۴۵	۱۸۴۹	۱۱	۳۶۰۰۰	۴۳۸,۳۳۶,۴۱۹
[casualties->Few]	۲۳۵				
[casualties->Few, Magnitude->Strong]	۱۸۳				
[Season->Spring]	۱۴۴				
[Season->Spring, Magnitude->Strong]	۱۱۹				
[Season->Summer]	۱۳۳				
[Season->Summer, Magnitude->Strong]	۱۱۸				
[Season->Winter]	۱۲۳				
[Season->Winter, Magnitude->Strong]	۱۱۳				
[Season->Fall]	۱۲۱				
[casualties->Moderate]	۱۱۵				

[1] Completeness
[2] Lift
[3] Usefulness
[4] Certainty

جدول (۶): توصیف کمیت‌ها

ردیف	عنوان کمیت	توصیف کمیت
۱	**اندازه زمین‌لرزه**	
	ضعیف	۱ تا ۲ ریشتر
	متوسط	۳ تا ۵ ریشتر
	قوی	۶ تا ۸ ریشتر
	فاجعه‌آمیز	۹ تا ۱۱ ریشتر
۲	**تلفات زمین‌لرزه**	
	کم	کمتر از ۱۰۰ نفر
	متوسط	بین ۱۰۰ تا ۱۰۰۰ نفر
	زیاد	بین ۱۰۰۰ تا ۱۰۰۰۰ نفر
	فاجعه‌آمیز	بیشتر از ۱۰۰۰۰ نفر
۳	**سودمندی قواعد**	
	ارتباط منفی	کمتر از ۱ یا ۱>
	کاملاً بی ربط	مساوی یک یا ۱-
	ارتباط مثبت	بزرگتر از یک یا ۱<

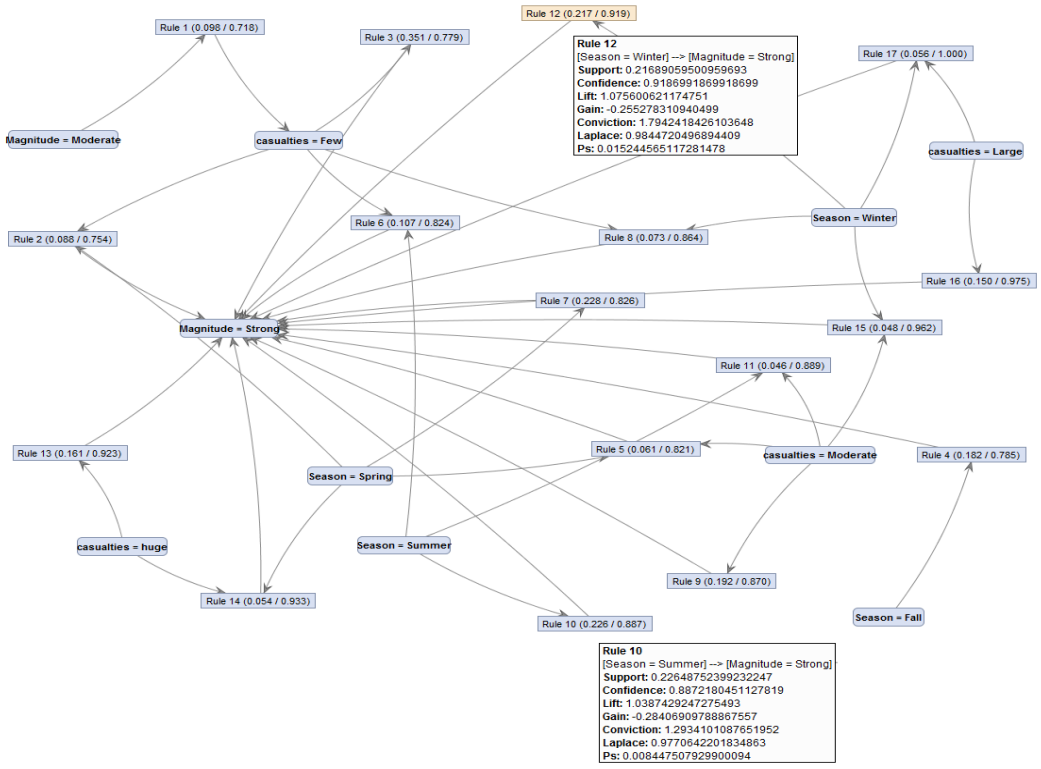
بر اساس ارزیابی انجام‌شده در گراف شکل (۱۰)، قاعده دوم با داشتن ارتباط مثبت، درصد پشتیبانی ۰,۲۲، درصد اطمینان ۰,۹۲، و درصد بالابری ۱,۰۸و قاعده چهارم با داشتن ارتباط مثبت، درصد پشتیبانی ۰,۲۳، درصد اطمینان ۰,۸۹ و درصد بالابری ۱,۰۴ به‌عنوان دو قاعده ارزشمند استخراج گردید. در این ارزیابی، انتخاب قاعده هم‌آیی نهایی که از اطمینان و صحت بالاتری برخوردار باشد، بر اساس معیار بالابری و معیار پشتیبانی هر قاعده انجام می‌پذیرد. لذا قاعده‌ای که بالابری کمتری دارد، چون پشتیبانی بهتری دارد، از صحت بالاتری نیز برخوردار است. بنابراین، از قابلیت اطمینان بالاتری نیز برخوردار است. بر این اساس قاعده دوم به‌عنوان قاعده ارزشمند نهایی انتخاب گردید. یعنی:

$$[\textbf{Season} \rightarrow \boldsymbol{Winter}.\textbf{Magnitude} \rightarrow \boldsymbol{Strong}] \qquad (۱۷)$$

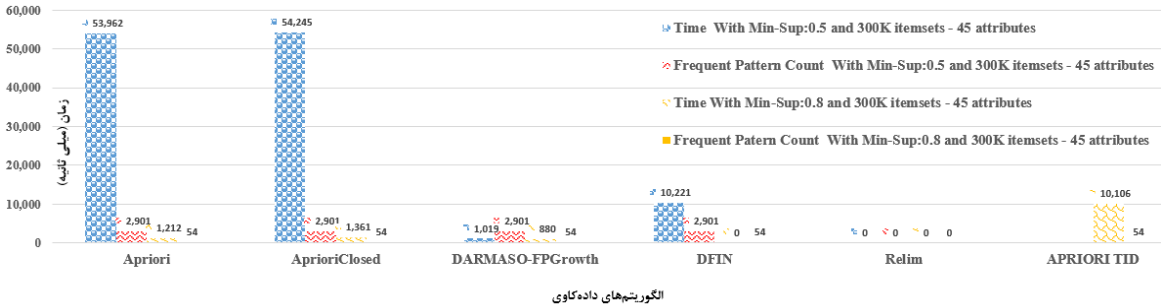
برای ارزیابی روش پیشنهادی، سناریوهای مختلفی توسط الگوریتم‌های FP-Growth، Apriori، AprioriClosed، DFIN، Relim و AprioriTID مورد آزمون قرار گرفت. بر اساس این ارزیابی، DARMASO با اجرای الگوریتم رشد اقلام مکرر بهترین عملکرد از نظر مقیاس‌پذیری، زمان اجرا، کاوش اقلام مکرر و تولید قواعد هم‌آیی معنایی را نسبت به سایر الگوریتم‌های آزمون شده نشان می‌دهد. نتایج حاصل از این آزمون در شکل (۱۱) نمایش داده شده است. کاهش فضای کاوش، استخراج الگوهای مکرر در سطح گره‌های محلی و تولید قواعد هم‌آیی سودمند متناسب با پرس و جوی کاربر در سطح سراسری از مهم‌ترین دستاوردهای اجرای روش ائتلافی DARMASO است. بر اساس نتایج تجربی، استفاده از هستان‌شناسی در داده‌کاوی، فرآیند کاوش را هدایت و کنترل نموده و فضای پرس و جو را کاهش و یا وادار به کاهش می‌کند. همچنین فناوری سامانه‌های چندعاملی با به اشتراک‌گذاری تنها کد و نه داده‌ها در میان سایت‌های توزیع‌شده، سبب کاهش ترافیک شبکه، زمان محاسبات، هزینه حافظه، پیچیدگی داده‌ها و افزایش کارایی و مقیاس‌پذیری خواهد شد. همچنین هزینه استفاده از منابع داده-ای، تابعی خطی از تعداد منابع داده‌ها با شیب بسیار کم است. در مقابل روش سنتی با منابع داده‌ای کمتر، از زمان پاسخگویی مناسب‌تری برخوردار است. اما با افزایش منابع داده‌ها، هزینه استفاده از منابع داده‌ای در روش ائتلافی نسبت به روش سنتی کمتر و مقرون به‌صرفه‌تر و هزینه روش سنتی به‌صورت نمایی افزایش خواهد یافت. شکل (۱۲) نمودار مقایسه فرآیندهای مبتنی بر اجرای روشDARMASO و روش سنتی مبتنی بر تجمیع داده‌ها را نشان می‌دهد.

جدول (۷): توصیف قواعد هم‌آیی معنایی کاوش شده

سودمندی قواعد	درصد منطقی بودن قواعد	درصد اطمینان قواعد	نمونه الگوهای کاوش شده درکاهش سوم	کاهش سوم (تعداد قواعد هم‌آیی)	کاهش دوم (تعداد الگوهای کاوش شده)
			[Min-Support:0.2]-[Min- confidence:0.7]		
ارتباط منفی	۰,۹۷	۰,۸۳	**[Season->Spring] => [Magnitude->Strong]** ٪۸۳ درصد زمین‌لرزه‌های با اندازه قوی در فصل بهار رخ داده است.	۴	۱۱
ارتباط مثبت	۱,۰۸	۰,۹۲	**[Season->Winter] => [Magnitude->Strong]** ٪۹۲ درصد زمین‌لرزه‌های با اندازه قوی در فصل زمستان رخ داده است.		
ارتباط منفی	۰,۹۱	۰,۷۸	**[casualties->Few] => [Magnitude->Strong]** ٪۷۸ درصد زمین‌لرزه‌های با اندازه قوی با تلفات کم رخ داده است.		
ارتباط مثبت	۱,۰۴	۰,۸۹	**[Season->Summer] => [Magnitude->Strong]** ٪۸۹ درصد زمین‌لرزه‌های با اندازه قوی در فصل تابستان رخ داده است.		



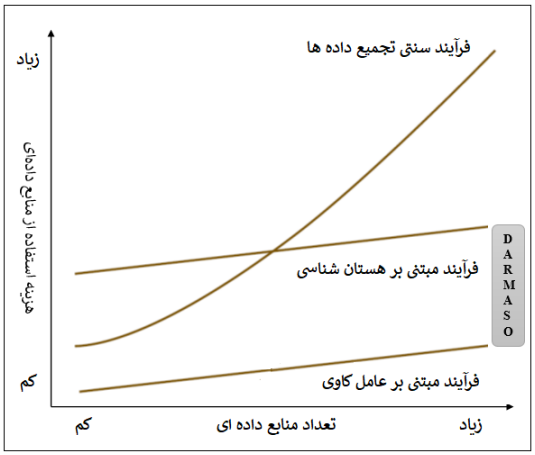
شکل (۱۰): گراف قواعد هم‌آیی کاوش شده



شکل (۱۱): اجرای الگوریتم‌های مختلف و مقایسه با روش پیشنهادی DARMASO

۶- نتایج و بحث

در این مقاله یک روش ائتلافی به نام DARMASO پیشنهاد گردید. در این مقاله نشان داده شد که استفاده از هستان‌شناسی و عامل‌گرایی ائتلاف ارزشمندی را برای پشتیبانی از کاهش فضای کاوش، استخراج اقلام مکرر و تولید قواعد هم‌آیی ارزشمند و مفید فراهم می‌کند. نتایج به‌دست‌آمده اثبات‌کننده این واقعیت است، که اولاً در سطح مفهومی، هر چه قواعد بر اساس هستان‌شناسی دامنه، کلاس و درخواست کاربر استخراج شود، به اهداف مورد نظر کاربر نزدیک و مرتبط است. ثانیاً با حمایت هستان‌شناسی قواعد تولیدشده در سطح مفهومی و وضوح معنایی بالاتری قرار گرفته و کمک بیشتری به تصمیم سازی و تصمیم‌گیری دقیق‌تر در کسب و کار می‌کند. ثالثاً سطح مفهومی



شکل ۱۲. نمودار مقایسه DARMASO و روش سنتی

و معنایی قوانین کاوش شده افزایش یافته و با پشتیبانی عامل‌ها حجم داده‌ها در سطح پردازش و ذخیره‌سازی به میزان قابل توجهی کاهش می‌یابد. همچنین رویکردهای مبتنی بر هستان‌شناسی در مواجه با منابع داده‌ای زیاد، هزینه کمتری داشته و قابلیت همکاری برای محیط‌های باز و بزرگ را تضمین می‌کند. همچنین برای بررسی کارایی و عملکرد روش ائتلافی DARMASO با شش الگوریتم مختلف برای ارزیابی زمان اجرا و مقیاس‌پذیری مقایسه گردید. آزمایش‌های گسترده روی چندین مجموعه داده، کارآمدی DARMASO را نسبت به روش سنتی تأیید می‌کند. همچنین روش پیشنهادی، توانایی اداره معنایی داده‌های بزرگ را برای کاوش اقلام مکرر و تولید قواعد هم‌آیی مفید و ارزشمند، با استفاده از الگوریتم‌های مختلف دارد.

۷- پژوهش‌های آینده

در پژوهش‌های آینده، تکمیل روش ائتلافی DARMASO با استفاده از روش‌های هوشمند یادگیری ماشین مبتنی بر قواعد پیشنهاد می‌شود. روش یادگیری مبتنی بر قواعد، کشف ارتباط‌ها و قواعد هم‌آیی سودمند و ارزشمند را از میان انبوه داده‌ها توسعه داده و سبب اثربخشی، تعمیم الگوریتم‌های کاوش و داده‌کاوی معنایی خواهد شد.

۸- مراجع

[1] D. Patel and J. Shah, "Jade Agent Framework for Distributed Data Mining and Pattern Analysis," International Journal of Computer Applications, vol. 975, p. 8887, 2017.

[2] 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.

[3] R. M. Gahar, O. Arfaoui, M. S. Hidri, and N. B. Hadj-Alouane, "An Ontology-driven MapReduce Framework for Association Rules Mining in Massive Data," Procedia Computer Science, vol. 126, pp. 224-233, 2018.

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

[5] P. V. Bhagat, P. M. J. I. J. O. R. Gourshettiwar, I. T. i. Computing, and Communication, "A survey paper on ontology-based approaches for semantic data mining," vol. 3, no. 4, pp. 2137-2141, 2015.

[6] M. R. Chikhale, "Study of Distributed Data Mining Algorithm and Trends," IOSR Journal of Computer Engineering (IOSR-JCE), pp. 41-47, 2016.

[7] D. Dou, H. Wang, and H. Liu, "Semantic data mining: A survey of ontology-based approaches," In Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015), IEEE, pp. 244-251, 2015.

[8] V. S. Ms and K. J. P. C. S. Shah, "Performance evaluation of distributed association rule mining algorithms," vol. 79, pp. 127-134, 2016.

[9] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, 1993, pp. 207-216.

[10] Z. Farzanyar, "Development of Algorithms for Detecting Frequent Items Set to Large-Scale Peer-to-Peer Environments with Flow Data Attitudes," Iran University of Science and Technology, 1391. (In Persian)

[11] D. Apiletti, E. Baralis, T. Cerquitelli, P. Garza, F. Pulvirenti, and L. Venturini, "Frequent itemsets mining for big data: a comparative analysis," Big data research, vol. 9, pp. 67-83, 2017.

[12] D. Xia, X. Lu, H. Li, W. Wang, Y. Li, and Z. Zhang, "A MapReduce-based parallel frequent pattern growth algorithm for spatiotemporal association analysis of mobile trajectory big data," Complexity, 2018.

[13] D. C. Anastasiu, J. Iverson, S. Smith, and G. Karypis, "Big data frequent pattern mining," In Frequent pattern mining: Springer, pp. 225-259, 2014.

[14] 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.

[15] D. A. Koutsomitropoulos and A. K. Kalou, "A standards-based ontology and support for Big Data Analytics in the insurance industry," ICT Express, vol. 3, no. 2, pp. 57-61, 2017.

[16] A. Soylu et al., "Ontology-based Visual Querying with OptiqueVQS: Statoil and Siemens Cases," 2016.

[17] S. Urmela and M. Nandhini, "Approaches and Techniques of Distributed Data Mining: A Comprehensive Study," International Journal of Engineering and Technology (IJET), 2017.

[18] 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.

[19] F. Jiang, "Efficient frequent pattern mining from big data and its applications," Ph.D. Thesis, Department of Computer ScienceThe University of Manitoba Winnipeg, Manitoba, Canada, 2014.

[20] Y. Lin, P.-C. Huang, D. Liu, and L. Liang, "Scalable frequent-pattern mining on nonvolatile memories," In 2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE, pp. 578-583, 2017.

[21] Bhamra, Gurpreet S Verma, Anil K Patel, Ram B, "Agent Based Frameworks for Distributed Association Rule Mining: An Analysis," International Journal in Foundations of Computer Science & Technology (IJFCST), 2015.

[22] H. Saberi, M. R. Kangavari, and M. R. H. Ahangar, "Provide an agent-oriented architecture for semantic exploration of large-scale data in distributed environments," Scientific-Research Journal of Electronic and Cyber Defense, 1398. [Online]. Available: https://ecdj.ihu.ac.ir, In Persian.

[23] H. M. Basir, H. Saberi, and M. A. Javadzadeh, "Provide a method for selecting data based on the ontology of the organization's goals," presented at the Fifth National Conference on Defense Science and Engineering, Tehran, 1398. (In Persian)

[24] A. H. Yazdi, "Extract association rules from semantic data streams," Computer Engineering, Ferdowsi University of Mashhad, 1392. [Online]. Available: http://www.um.ac.ir, In Persian

.

Journal of Electronical & Cyber Defence
Vol. 9, No. 1, 2021, Serial No. 33

Providing an Ontology-Based Method for Exploring the Association Rules in Multi-Agent Distributed Environments

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

*Imam Hossein Comprehensive University

(Received: 17/02/2020, Accepted: 05/08/2020)

ABSTRACT

Distributed association rules mining is one of the most important data mining methods that extracts the inter dependence of data items from decentralized data sources, regardless of their physical location and is based on the process of extracting repeated items. When exploration algorithms are implemented on large-scale data, a large number of recurring items are produced, many of which are irrelevant, ambiguous, and unusable for the business, thus causing a challenge called "combination explosion ."In this paper, a new coalition method based on distributed data mining and domain archeology, abbreviated to DARMASO, is proposed to address this challenge. This method uses three algorithms: the DARMASOMAIN algorithm to guide and control the process of exploration and aggregation of universal rules, the DARMASOPRU algorithm to reduce and prune the data and the DARMASOINT algorithm to explore and aggregate the rules of all the generated data sources. DARMASO uses a map-reduce-based distributed computational model in a multi-agent distributed environment. It also provides a practical way for semantic mining of large-scale data sets. This method filters out the association rules of generality based on the purposes of data mining as well as the needs of the user and only produces and maintains useful rules. Reducing the scope of exploration and filtration of rules is achieved through the process of semantic pruning in the form of removing inappropriate candidates from the set of frequent items and producing association rules of utility. The implementation is performed using a data set from the scope of natural disasters and the earthquake class. It also improves the speed and quality of rule extraction and generates practical, reliable, logical, quality and valuable rules to support decision-making amid the masses of data.

Keywords: Association Rules, Ontology, Multi Agent Systems, Mapping-Reduction

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