

خوشه بندی مبتنی بر چند عامل: به سوی داده کاوی چند-عامله‌ی عمومی

چکیده

یک چارچوب برای داده کاوی مبتنی بر چند عامل در این نوشته توضیح داده شده است. این قالب یا چارچوب شامل مجموعه عواملی است که برای انجام یک وظیفه داده کاوی با هم همکاری می کنند. مفهوم بنیادی ای که اساس این چهارچوب را تشکیل می دهد این است که باید از داده کاوی عمومی پشتیبانی کند. چشم انداز و هدف مورد نظر در واقع سیستمی است که به روش اُرگانیک (سازمانی یا ساختمانی) رشد کند. موضوع اساسی برای تسهیل این رشد، محیط ارتباط لازم برای پشتیبانی از تعامل عامل ها است. این موضوع تا حدودی توسط ماهیت طراحی پیشنهادی و تا حدودی هم توسط آنتولوژی قابل گسترش فراهم شده است. انگیزه بعدی این است که هنوز بهترین الگوریتم خوشه بندی شناسایی نشده است و در نتیجه یک روش مبتنی بر عامل می تواند برای شناسایی آن مورد استفاده قرار گیرد. این کاربرد سعی در نشان دادن پتانسیل داده کاوی چند عامله دارد.

کلمات کلیدی: داده کاوی چند عامله , خوشه بندی مبتنی بر عامل

مقدمه

فواید سیستم های چند عامله با توجه به رایانش (محاسبات) مشارکتی توزیع شده به خوبی درک شده اند. به طور کلی این مقاله به دنبال این است تا نشان دهد چگونه مزایای عمومی سیستم های چند عامله در داده کاوی , مثل داده کاوی چند عامله , می تواند مورد استفاده قرار گیرد.

داده کاوی چند عامله می تواند برای رسیدگی و حل برخی از مشکلات و مسائل داده کاوی عمومی پشتیبانی ای ارائه دهد, مسائلی مانند:

1. در نظر گرفتن اندازه مجموعه داده ها: استخراج کنندگان داده تمایل دارند تا همه چیز را استخراج کنند; آن ها می خواهند متن, عکس ها, فیلم و داده های چند-رسانه ای را به خوبی داده های جدولی استخراج کنند. روش های داده کاوی برای استخراج مجموعه داده های جدولی به خوبی مشخص شده اند, اما مجموعه داده های بزرگتر, مجموعه های پیچیده تر (عکس و فیلم), داده های امروزی تر (مانند گراف ها, شبکه ها, درخت ها و غیره) نیاز به استخراج شدن دارند. منابع پردازش این داده ها بسیار مهم هستند; روش داده کاوی چند عامله شاید برای این موضوع راه حل ارائه دهد.

2. امنیت و حفاظت داده: مسائل قانونی و اقتصادی مرتبط با امنیت داده و محافظت از آن اهمیت روزافزونی پیدا کرده اند. ایده به اشتراک گذاشتن داده ها در یک منبع داده واحد اغلب اوقات کاربردی نیست یا تنها زمانی کاربردی است که ابتدا پیش پردازش مناسب و ناشناس ماندن (anonimizayion = گمنام ماندن افرادی که داده ها در مورد آن ها توضیحاتی ارائه می دهد) انجام شود. داده کاوی چند عامله مکانیسمی برای پشتیبانی امنیت داده ارائه می دهد.

3. مناسب بودن الگوریتم های داده کاوی: یک موضوع جالب در رابطه با تحقیقات داده کاوی انجام شده این است که برای بسیاری از وظایف داده کاوی (مانند خوشه بندی) شواهد اندکی برای بهترین الگوریتم استخراج مناسب انواع داده ها وجود دارد. حتی در مورد داده های جدولی نسبتاً ساده, در چهارچوب خوشه بندی, نیز الگوریتم واحدی که در همه ی موارد بهترین خوشه ها را ارائه دهد, وجود ندارد.

چشم انداز نشان داده شده در این مقاله درباره ی داده کاوی چند عامله, یک زیرساختی را فراهم می کند که به آژانس های داده کاوی این اجازه را می دهد تا در کنار هم به انجام امور اختصاصی استخراج داده بپردازند. اگرچه چشم انداز ارائه شده راه حل های بالقوه ای را برای مسائل بالا ارائه داد, ولی مسائلی هستند که با دستیابی به این چشم انداز به راحتی برطرف می شوند:

1. طبیعت نامساوی و غیرمتجانس داده کاوی: ماهیت بالقوه امور داده کاوی ای که در مورد آن ها خیال پردازی می کنیم و آرزوی آن را داریم این است که داده کاوی چند عامله بتواند به طور گسترده ای عمل کند.

2. رشد ارگانیک: برای اینکه این روش داده کاوی به معنای واقعی کلمه کاربردی و مفید باشد، باید بتواند به شکل ارگانیک رشد کند.

بنابراین این روش داده کاوی باید از امکاناتی پشتیبانی کند که اجازه گنجایش عوامل اضافی در چهارچوب مد نظر را به شیوه ad hoc (موردی) بدهد. بنابراین مکانیسم ارتباطی پشتیبانی کننده از داده کاوی چند عامله یک موضوع کلیدی است. این مکانیسم باید از تعامل مناسب عوامل پشتیبانی کند؛ بنابراین عوامل موجود ممکن است وظایف مختلف داده کاوی را بر عهده بگیرند و با توجه به این موضوع، عوامل بیشتری می توانند توسط کاربر نهایی وارد سیستم بشوند.

این مقاله یک چهارچوب قابل استفاده عمومی را برای داده کاوی چند عامله توضیح می دهد که از طریق شناسایی ماهیت داده کاوی قابل گسترش از ارتباط مذکور در پاراگراف قبل پشتیبانی می کند. برای ایجاد توجه و تمرکز برای کار توضیح داده شده، یک روند خوشه بندی نشان داده شده است. انگیزه و هدف این روند خوشه بندی این است که تعدادی از الگوریتم های خوشه بندی را به کار بگیرد و نتیجه ای را که بهترین مجموعه خوشه ها را ایجاد کرده است را انتخاب کند. اگر چه این روند مشکلات و مسائل عمومی تر داده کاوی چند عامله را که در بالا توضیح داده شد را به تصویر می کشد.

2. کارهای قبلی (سابقه کارهای انجام شده)

این بخش اطلاعات زمینه ای را در مورد موضوع توضیح داده شده را برای شما بیان می کند. این بخش خود به دو بخش کوچکتر تقسیم می شود. قسمت اول جدیدترین و به روز ترین مطالعه و بازیابی انجام شده در مورد داده کاوی چند عامله را بیان می کند. قسمت دوم نیز اطلاعات پایه ای مربوط به کاربرد خوشه بندی داده را برای ترسیم کار توضیح داده شده در این مقاله ارائه می دهد.

2.1 داده کاوی چند عامله

در نوشته ها به تعدادی از روش های مبتنی بر چند عامل شامل مواردی در مورد خوشه بندی اشاره شده است. این موارد شامل PADMA, PAPYRUS, JABAT هستند. PADMA برای ایجاد خوشه های بر اساس سلسله مراتب در زمینه طبقه اسنادی مورد استفاده قرار می گرفت. خوشه های محلی در سایت های محلی ایجاد شده و بعد در سایت مرکزی برای تولید خوشه های جهانی مورد استفاده قرار می گیرند. PAPYRUS یک سیستم چند عامله خوشه بندی است که در آن هم داده ها و هم نتایج بین عوامل مرتبط با استراتژی

سیستم چند عامله منتقل می شوند. JABAT یک سیستم چند عامله برای خوشه بندی توزیع شده و هم چنین خوشه بندی توزیع نشده است (موجود در الگوریتم K-means). JABAT در این مقاله مورد توجه است زیرا این سیستم از علم آنتولوژی (علمی که به بررسی معنا و ارتباط کلمات در حوزه مورد استفاده می پردازد) برای تعریف کردن واژگان و علم معانی، برای تبادل پیام بین عوامل استفاده می کند. هیچ کدام از این سیستم های داده کاوی استفاده از داده کاوی چند عامله برای تعیین بهترین خوشه ها را پشتیبانی نمی کنند.

مثال دیگری برای سیستم داده کاوی چند عامله، سیستم پیشنهادی توسط Baazaoui Zghal است. سیستم پیشنهادی توسط او، متمرکز بر داده های پراکنده جغرافیایی بود که از تکنیک های استخراج الگو برای ایجاد پایگاه اطلاعات استفاده می کرد. این پایگاه سپس برای پشتیبانی از تصمیم گیری توسط کاربری مورد استفاده قرار می گیرد.

کارهای انجام شده بر روی داده کاوی چند عامله بسیار کم هستند. یک مثال EMADS (چهارچوب یا قالب داده کاوی چند عامله گسترش پذیر) است. EMADS از دو روش استخراج داده استفاده می کند: استخراج قاعده ارتباط (association rule = جملات اگر-سپس که بیان کننده احتمال نتیجه کاری در صورت انجام آن هستند) و دسته بندی. این سیستم می تواند بهترین دسته بندی کننده با بالاترین دقت را با توجه به مجموعه داده مورد نظر، فراهم کند. یکی از معایب این سیستم استفاده از پروتکل های ثابت است در حالیکه داده کاوی چند عامله از یک روش مبتنی بر آنتولوژی قابل دسترس تر استفاده می کند. اگرچه کار توضیح داده شده در این مقاله از برخی از ایده های EMADS نیز استفاده می کند.

2.2 خوشه بندی داده

یک سناریوی خوشه بندی برای به تصویر کشیدن اجرای داده کاوی چند عامله عمومی پیشنهاد شده، در این مقاله استفاده شده است. هدف، شناسایی بهترین مجموعه خوشه های در یک مجموعه داده ی مد نظر است. این تظاهرات شامل سه عامل خوشه بندی هستند که هر یک دارای الگوریتم خوشه بندی متفاوتی است:

K-means

DBSCAN

K nearest neighbor (نزدیکترین همسایه k)

الگوریتم k-means یک تکنیک خوشه بندی پارتیشنال است که پارامتری تحت عنوان K (تعداد خوشه های مد نظر) را در نظر می گیرد و سپس مجموعه داده هایی را که به آن می دهند را در خوشه های K تقسیم بندی می کند. شباهت خوشه ها با در نظر گرفتن ارزش میانگین اعضای هر خوشه محاسبه می شود. این الگوریتم به شکل زیر عمل می کند:

- (1) امتیازهای تصادفی K را به عنوان مراکز انتخاب می کند که به آن ها سنتروئید می گویند،
 - (2) هر جزء را به نزدیکترین سنتروئید مرتبط می کند، بر اساس فاصله بین اجزاء و سنتروئید ها،
 - (3) زمانی که همه اجزاء به یک خوشه مرتبط شدند سنتروئید جدیدی را برای هر خوشه محاسبه می کند و دوباره گام دوم به بعد را برای آن طی می کند تا خوشه ها به همگرایی برسند.
- یکی از معایب این الگوریتم این است که باید تعداد خوشه ها را از قبل تعیین کرد.

الگوریتم نزدیکترین همسایه K از یک حد آستانه (ترشولد) استفاده می کند تا نزدیکترین همسایه را تعیین کند. اگر فاصله بین یک آیتم و نزدیکترین جزء کمتر از آستانه باشد، این آیتم باید در همان خوشه ای قرار بگیرد که نزدیکترین جزء مذکور در آن قرار دارد. زمانی که فاصله بیشتر از آستانه باشد یک خوشه جدید ایجاد می شود. آن گاه ارزش t به طور چشمگیری بر تعداد خوشه ها اثر می گذارد.

DBSCAN یک الگوریتم خوشه بندی مبتنی بر دانسیته (چگالی) است که خوشه هایی را با حداقل سائز و آستانه دانسیته داده شده ایجاد می کند. این ویژگی به الگوریتم اجازه می دهد تا مشکل داده های خرج از محدوده را مدیریت کند و آن هم به وسیله اطمینان از عدم حضور هیچ گونه داده خارج از مرکزی در یک خوشه. تعداد نهایی خوشه ها نیز توسط الگوریتم مشخص می شود.

برای ارزیابی اهداف از مقیاس F در این مقاله استفاده شده است تا نتایج خوشه بندی را نسبت به هم مقایسه کنیم. این مقیاس به طور گسترده ای در زمینه بازیابی اطلاعات استفاده می شود. این روش چگونگی مطابقت طبقه بندی خوشه ها را با طبقه بندی کلاس عرضه شده خارجی را اندازه گیری می کند. مقیاس F احتمال اینکه عضوی از یک خوشه متعلق به پارتیشن خاصی باشد (precision) را با تا چه حد یک خوشه شامل تمام اشیاء یک پارتیشن خاص است (recall)، ترکیب می کند. اگر $C = \{C_1, C_2, \dots, C_k\}$ نتیجه یک خوشه بندی باشد آن گاه $P = \{P_1, P_2, \dots, P_k\}$ پارتیشن ground-truth خواهد بود. دقت یک خوشه با توجه به دقت پارتیشن j به صورت مقابل است:

$$precision(i, j) = |C_i \cap P_j| / |C_i|$$

معیار ریکال برای یک خوشه i نیز با توجه به پارتیشن Z به صورت زیر محاسبه می شود:

$$recall(i, j) = |Ci \cap Pj| / |Pj|$$

با توجه به موارد بالا مقیاس F یک خوشه به صورت زیر محاسبه می شود:

$$F_{ij} = \frac{2 \times precision(i, j) \times recall(i, j)}{precision(i, j) + recall(i, j)}$$

و مقیاس F کلی برای یک خوشه نیز به شکل زیر محاسبه می شود:

$$F = \sum_{i=1}^m \frac{|P_i|}{|P|} \times \max(F_{ij})$$

با توجه به سناریوی خوشه بندی استفاده شده در این مقاله، بهترین الگوریتم خوشه بندی الگوریتمی است که مقیاس F کلی بالاتری نسبت به مجموعه داده های مورد نظر، داشته باشد.

3. قالب داده کاوی چند عامله

قالب داده کاوی چند عامله پیشنهادی دارای 5 نوع عامل است که هر کدام ممکن است چند زیر واحد مرتبط داشته باشند. این عوامل عبارتند از:

1. عامل کاربر: یک رابط کاربری گرافیکی بین کاربر و عوامل کار فراهم می کند.
2. عوامل وظایف داده کاوی: عملکرد داده کاوی را تسهیل می کند. سه زیر واحد متناظر با سه عملکرد داده کاوی سطح بالا شناخته شده است: استخراج قاعده ارتباط (association rule mining)، دسته بندی و خوشه بندی. این عوامل مسئول مدیریت و برنامه ریزی درخواست داده کاوی هستند. این عوامل، عامل های مناسب دیگر لازم برای کامل کردن داده کاوی را با رجوع به سرویس "صفحه زرد" شناسایی می کنند.
3. عامل داده: عواملی هستند که هر یک دارای متادیتا (فراداده) راجع به یک مجموعه داده به خصوص هستند که اجازه دسترسی به داده را فراهم می کند. میان این عوامل و مجموعه داده ها رابط یک به یک برقرار است.

4. عوامل استخراج: مسئولیت انجام استخراج و تولید نتایج را برعهده دارند. نتیجه در نهایت به عوامل وظایف داده کاوی فرستاده می شوند.

5. عوامل بررسی کننده صحت: عواملی هستند که صحت نتایج داده کاوی را ارزیابی و بررسی می کنند. زیر واحد های زیادی برای این عوامل وجود دارد که مشابه زیر واحد های مورد دو است: استخراج قاعده ارتباط (association rule mining), دسته بندی و خوشه بندی.

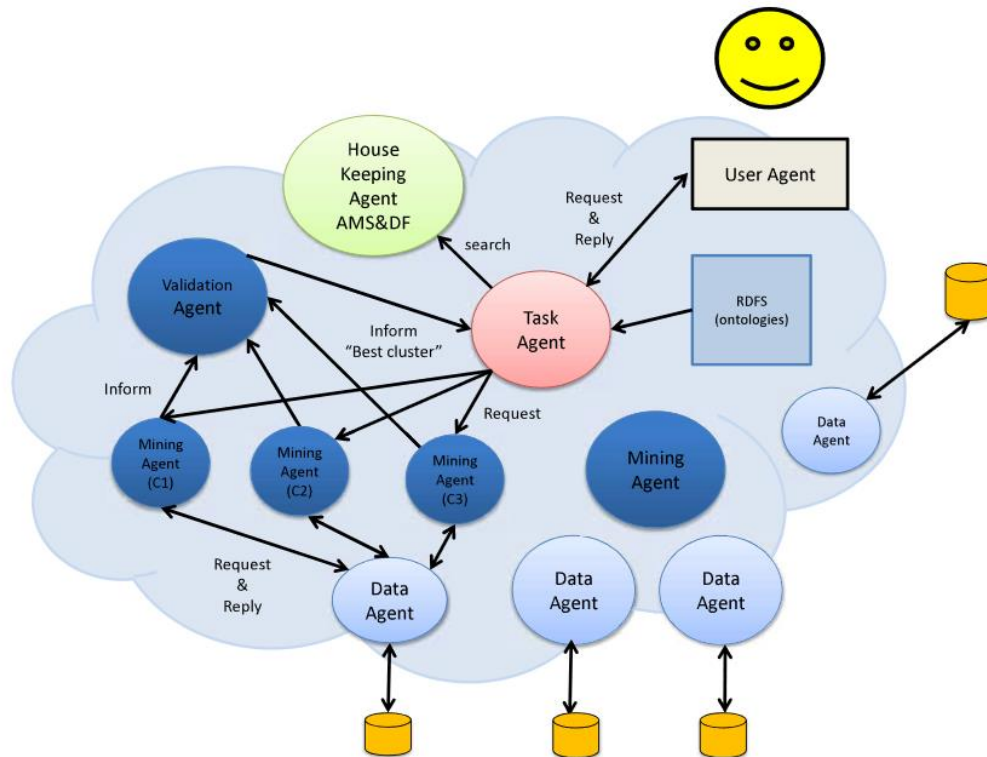
در برخی موارد ممکن است یک عامل خاص زیر واحد های ثانویه مرتبط با خود به منظور اصلاح اعمال مشخصی از داده کاوی چند عامله, داشته باشد.

قالب داده کاوی چند عامله با استفاده از JADE (محیط توسعه عامل جاوا), یک ابزار توسعه عامل تثبیت شده, مورد تایید FIPA, اجرا شد. این ابزار تعدادی عامل داخلی اضافی برای قالب داده کاوی چند عامله فراهم می کند. این عامل های داخلی اضافه عبارتند از: AMS (سیستم مدیریت عامل) که مسئول کنترل چرخه سایر عوامل پلتفرم هستند, DF (تسهیل کننده دایرکتوری) که سرویس "صفحه زرد" را فراهم می کند تا هر عامل بتواند ظرفیت خود را ثبت کند و به عوامل وظایف داده کاوی اجازه دهد تا عوامل مناسب را در زمان برنامه ریزی داده کاوی شناسایی کنند.

شکل یک مثالی از پیکربندی عوامل را برای داده کاوی چند عامله نشان می دهد. این پیکربندی شامل مثال های متفاوتی از عوامل شناخته شده مذکور هستند. این شکل همچنین پیکربندی یک عامل را به منظور دست یابی به خوشه بندی داده ها به تصویر می کشد. جریان کنترل توسط یک عامل کاربر آغاز می شود که یک عامل وظیفه خاص را به وجود می آورد و در این مورد این عامل وظیفه خاص, یک عامل خوشه بندی است. عامل وظیفه ساخته شده با عامل های داخلی (housekeeping) تعامل داشته و با همکاری هم تعیین می کنند که کدام عوامل در دسترس هستند تا وظیفه خواسته شده را انجام دهند و سپس از میان آن ها عوامل مورد نظر را انتخاب می کنند. در مثال شکل 1, 5 عامل انتخاب شده اند: سه عامل

استخراج (C1, C2, C3), یک عامل بررسی صحت و یک عامل داده. در این مثال عامل وظیفه با سه عامل استخراج ارتباط برقرار کرده که با عامل داده تعامل و همکاری دارند (ولی می توانند به طور مساوی با چند عامل داده تعامل داشته باشند). عوامل استخراج نتایج را به عامل بررسی صحت می دهند که آن هم (با کمک عامل ثانویه خود) نتایج را پردازش می کند و نتایج نهایی را از طریق عامل وظیفه به عامل کاربر می

دهد. توجه داشته باشید که عامل وظیفه برخلاف سایر عوامل پس از انجام کار متوقف می شود.



شکل 1. شماتیک سیستم داده کاوی چند عامله برای خوشه بندی

4. ارتباطات درون عامل

یکی از جنبه های مهم در هر سیستم چند عاملی این است که عوامل آن بتوانند به شیوه ای موثر ارتباط برقرار کنند. گسترش زبان های ارتباط عوامل تحت تاثیر فلسفه ی نظریه اعمال گفتاری قرار گرفته است ولی همچنان یک زبان ارتباطی عوامل که به صورت قطعی در مورد همه ی کاربرد ها مناسب باشد وجود ندارد. اگرچه چند زبان ارتباطی محبوب برای استفاده در سیستم های چند عاملی وجود دارد، که دو مورد برجسته آن ها عبارتند از: KQML (زبان دستکاری و پرس و جوی اطلاعات) و قالب تبادل اطلاعات مرتبط با آن (KIF) و FIPA ACL (زبان ارتباط عامل فیپا).

KQML به عنوان یک زبان ارتباطی بیرونی مبتنی بر پیام است که پیام ها را در گروه های خاصی طبقه بندی می کند، که به آنها Performative می گویند، تا یک فرمت مشترک برای تبادلات فراهم کند.

بر عکس KQML, KIF تنها در پی ایجاد بستری برای ارتباط درونی است مانند اطلاعات قابل اجرا در یک زمینه خاص. اگرچه KQML در بین توسعه دهنده های عوامل محبوب و دارای نفوذ است و پایه تعدادی از کارهای انجام شده در زمینه ارتباطات را تشکیل داده است، انتقاد های بسیاری در زمینه های مختلف در مورد آن وجود دارد؛ مانند قابلیت همکاری، حذف برخی انواع پیام ها برای مدیریت برخی عبارات و کمبود معنانشناسی (semantics).

انتقاد ها بر این زبان موجب ایجاد و گسترش یک زبان جداگانه ولی شبیه آن شد که تحت عنوان زبان ارتباط عامل فیفا شناخته می شود که هدف آن ایجاد و تثبیت یک زبان ارتباط استاندارد قابل استفاده برای عوامل مستقل است. این زبان از لحاظ ترکیب و هماهنگی بخش های مختلف مشابه KQML است و همچنین مانند آن از زبان بیرونی استفاده می کند تا انتقال پیام را از محتوای داخلی مستقل امکان پذیر سازد که در هر زبان منطقی مناسب قابل بیان است. برای زبان بیرونی، این زبان جدید 22 گروه از پیام ها (performatives) را ارائه می دهد تا بتواند انواع مختلف پیام هایی که می توانند میان عوامل تبادل شوند را، از هم تمیز دهد. نمونه هایی از این گروه از پیام ها (performatives) inform ها هستند که پیام ها را از یک عامل به عامل دیگر می فرستند و نمونه دیگری نیز request ها هستند که از عوامل می خوتهند تا کار مخصوصی را انجام دهند. گروه های اختصاصی این زبان را می توانید در منبع شماره {15} مشاهده کنید. برای قرار نگرفتن در معرض برخی انتقاد ها راجع به KQML، توسعه دهندگان زبان ارتباط عامل فیفا یک دایره معانی رسمی جامع را برای این زبان ارائه کرده اند. این دایره معانی (semantics) از طریق تعریف یک زبان رسمی به نام زبان معنانشناسی، از نظریه اعمال (کنش های) گفتاری استفاده کرده است. زبان معنانشناسی ارائه خواسته ها، اختیارات و اعمال ممکن برای عملکرد را فراهم می کند. برای اطمینان از مطابقت عواملی که از این زبان استفاده می کنند، زبان معنانشناسی محدودیت ها و پیش فرض هایی در قالب فرمول برای هر پیامی در زبان ارتباط عامل فیفا دارد.

با وجود مزایایی که این زبان جدید نسبت به KQML فراهم کرده اما خود از انتقادها در امان نمانده است و ما این موضوع را در بخش 6 مورد بررسی قرار می دهیم. حال مکانیسم ارتباطی استفاده شده در قالب داده کاوی چند عامله خود را بررسی می کنیم. همانطور که اشاره شد، ما کار خود را به وسیله ابزار JADE انجام دادیم که استفاده از زبان ارتباط عامل فیفا را ارتقا بخشید تا تعامل میان عوامل را تسهیل کند. عوامل از یک مکانیسم انتقال پیام غیر همزمان استفاده می کنند (پیام زبان ارتباط عامل = ACLMessages) تا پیام ها را در

زیرساخت رایانه تبادل کنند. از طریق زبان محتوا یک عبارت از فرستنده به گیرنده می رسد. بیان محتوی ممکن است به روش های مختلف رمز گذاری شده باشد. JADE سه نوع رمزگذاری زبان محتوا را ارائه می دهد:

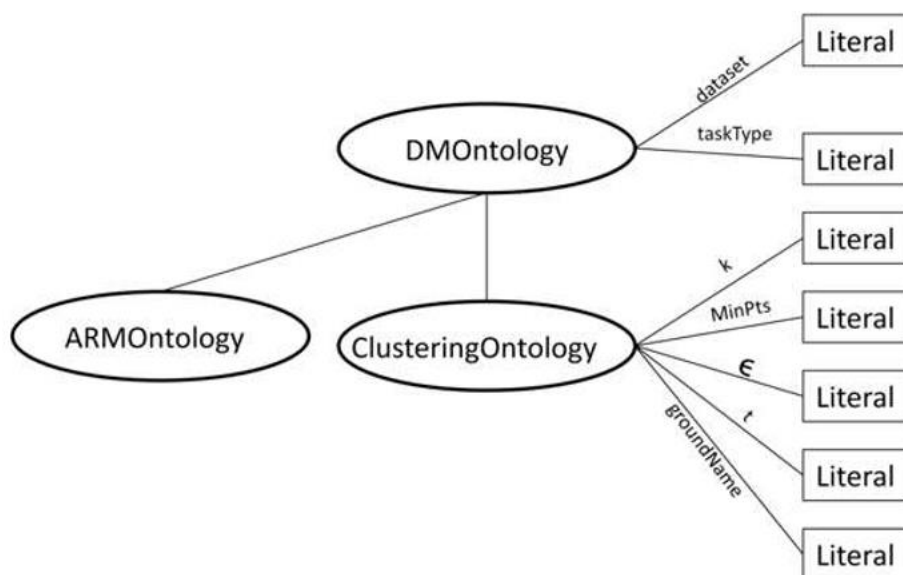
1. SL (زبان معنانشناسی): این زبان محتوا یک زبان محتوای رمزگذاری شده قابل خواندن برای انسان است و مناسب برای زمانی است که عوامل از توسعه دهنده های متفاوتی می آیند و روی پلتفرم های متفاوتی اجرا می شوند و باید با هم در ارتباط باشند.
 2. LEAP (پلتفرم عامل گسترش پذیر سبک): این زبان محتوا یک زبان محتوای بایت-کدک غیر قابل خواندن برای انسان است. بنابراین نسبت به مورد قبل سبک تر است و برای کاربردهایی که محدودیت حافظه دارند مناسب است.
 3. تعریف شده توسط کاربر: این زبان های محتوا با زبان های مدیریت شونده توسط منابع سازگاری دارند.
- فیفا یک زبان محتوای خاص را معرفی نکرده است اما توصیه می کند زمان برقراری ارتباط با عوامل AMS و DF از زبان SL استفاده شود. در قالب داده کاوی چند عامله پیشنهاد شده توسط ما , از زبان SL استفاده شده است زیرا قالب داده کاوی چند عامله ما عوامل مختلفی را در بر میگیرد که ممکن است از توسعه دهنده های مختلفی آمده باشند و در پلتفرم های متفاوتی اجرا شوند.
- در سیستم داده کاوی چند عامله ما, عوامل از طریق ایجاد یک درخواست و پاسخ به درخواست ها با هم ارتباط برقرار می کنند. همانطور که اشاره شد, تبدلات ارتباطاتی مطابق با زبان ارتباط عامل فیفا انجام می شوند. برای مثال جدول 1 را در نظر بگیرید که قطعه ای از یک ارتباط را با توجه به یک درخواست استخراج خاص فرستاده شده از عامل وظیفه به عامل داده کاوی, نشان می دهد. در این مثال Performative (گروه پیام ها) استفاده شده, request است و جزئیاتی نظیر کدام عامل فرستنده است, کدام دریافت کننده است, زبان مورد استفاده و گفتگوی انجام شده میان عوامل نیز آورده شده است. با توجه به محتوای واقعی پیام عبارتی تحت عنوان طرح چارچوب توضیح منابع (RDFS) به میان می آید. مطابق آنچه که آنتولوژی موارد را توضیح می دهد و معنا می کند, ویژگی ها و روابط مربوط به کار مورد نظر در دسترس است و بنابراین نیاز به ارجاع به جایی دارد که ارتباط در حال اجرا است. بنابراین آنتولوژی

RDFS این امکان را فراهم می کند تا واژگان معنا یافته و محتوای پیام RDF که می تواند توسط این آنتولوژی ایجاد شود , یک علم معناسازی را برای پیام های رد و بدل شده میان عوامل را فراهم کند. همانطور که در مثال جدول یک دیده می شود, محتوای پیام در RDF همانطور که از آنتولوژی گرفته شده است, بیان شده است و نوع کار نیز بیان شده است(خوشه بندی),مجموعه داده مورد استفاده نیز بیان شده است,ارزش ها و ریزنمرات و پارامتر ها نیز موجود است. عامل استخراج که این پیام را دریافت کرده است, می تواند متعاقبا پاسخ مناسب را برای عامل وظیفه ای که این پیام را فرستاده است, ارسال کند.

شکل دو جزئیاتی از آنتولوژی داده کاوی چند عامله عمومی را در موضوع کاربرد خوشه بندی داده ارائه می دهد. در این قالب مثال آنتولوژی به فرمت RDF آمده است.

5.نمایش خوشه بندی داده

برای به تصویر کشیدن عملکرد قالب داده کاوی چند عامله پیشنهاد شده یک کاربرد خوشه بندی در این بخش ارزیابی شده و توضیح داده شده است. سناریو در واقع این بود که کاربر نهایی می خواست که بهترین مجموعه خوشه ها را از میان داده های مورد نظر شناسایی کند. سه عمال استخراج با سه الگوریتم متفاوت فراهم شدند: K-means,KNN,DBSCAN



شکل 2. آنتولوژی داده کاوی چند عامله برای کار خوشه بندی

مقیاس F که در بخش دو توضیح داده شد نیز در این مورد برای مقایسه و شناسایی بهترین خوشه بندی مورد استفاده قرار گرفته است) هرچند هر سیستم ارزیابی کیفیت خوشه بندی دیگری نیز می تواند مورد استفاده قرار گیرد). یک عامل بررسی صحت نیز فراهم شد تا مقیاس FI را در مورد خوشه های خاص داده شده محاسبه کند. برای این منظور، یک عامل بررسی صحت ثانویه نیز که به پارتیشن ground-truth) مجموعه ای از مدارک شناسایی شده و پیشین از خوشه های از پیش تعریف شده) دسترسی داشتن استفاده شد تا مقیاس F را اندازه گیری کند. به یاد بیاورید که نتیجه ای که مقیاس F بزرگتری داشته باشد بهتر است. مطابق محتوای چارچوب داده کاوی چند عامله پیشنهاد شده، فرآیند خوشه بندی به شکلی که در ادامه می آید انجام گرفت. فرآیند با درخواست یک کاربر نهایی آغاز شد. کاربر جزئیات مجموعه داده را در اختیار گذاشت: یک پارتیشن ground-truth و پارامترهای لازم برای سه الگوریتم خوشه بندی استفاده شده. پارامترها به شکل زیر بودند:

تعداد خوشه خواسته شده (K)، حد آستانه KNN (t)، حداقل اندازه DBSCAN (minPts) و آستانه دانسیته. عامل کاربر یک عامل وظیفه مناسب ایجاد می کند. زمانی که عامل وظیفه ایجاد شد به تعامل با عامل DF (عامل دایرکتوری) می پردازد تا بتواند عوامل استخراج و داده و صحت مناسبی را ایجاد کند. درخواست استخراج اصلی برای عوامل استخراج شناسایی شده فرستاده می شود. نتایج خوشه بندی از عوامل استخراج برای عامل بررسی صحت فرستاده می شود؛ جایی که بهترین خوشه ها شناسایی می شوند. این نتیجه نهایی از طریق عوامل وظیفه و کاربر، به کاربر می رسد. در ادامه عامل وظیفه نابود می شود.

```
(request
:sender usert1
:receiver DBSCANAgent1
:content
<rdf:RDF
  xmlns:j.0="http://protege.stanford.edu/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://somewhere/ClusteringOntology">
    <j.0:Eps>5.0</j.0:Eps>
    <j.0:MinPts>3</j.0:MinPts>
    <j.0:dataset>usert1</j.0:dataset>
    <j.0:taskType>clustering</j.0:taskType>
  </rdf:Description>
</rdf:RDF>
:language SL
:conversation-id data mining
)
```

جدول 1. مثالی از پیام درخواست

برای ارزیابی اهداف از 10 مجموعه داده از مخزن ماشین یادگیری داده UCI استفاده شد. جدول دو نتایج را برای این ده مجموعه داده که روش داده کاوی چند عامله در مورد آن ها استفاده شد، نشان می دهد. جدول اندازه مجموعه داده (به صورت ویژگی ها و مدارک داده) تعداد خوشه ایجاد شده و مقیاس F اندازه گیری شده برای هر سه الگوریتم استفاده شده و ارزش های پارامترهای مورد نیاز را در اختیار ما می گذارد.

توجه داشته باشید که در الگوریتم K-means، برای تصدیق تعداد خوشه ها، ما از اعداد داده شد در مخزن UCI استفاده کردیم. الگوریتم KNN از آستانه (t) استفاده می کند تا مشخص شود که یک آیتم در یک خوشه خاص قرار می گیرد یا خیر. این آستانه که نزدیکترین اعداد را به اعداد مخزن داده UCI ایجاد کرد، در این مطالعه انتخاب شد و الگوریتم KNN بهترین فرصت را برای خوشه بندی داده های مورد نظر داشت. پارامترهای آستانه دانسیته و حداقل اندازه که در DBSCAN استفاده شدند، به روش مشابه پارامتر آستانه KNN ارزیابی شدند.

جدول سه بهترین نتایج رجوع داده شده به کاربر را نشان می دهد. نتایج توسط عامل صحت ایجاد شده بودند. جدول سه از این دیدگاه پشتیبانی می کند که بهترین الگوریتم خوشه بندی مستقل و واحد وجود ندارد و از انگیزه این مطالعه نیز پشتیبانی می کند.

موارد مذکور نشان دادند که اولاً داده کاوی چند عامله دارای تطبیق پذیری است. عامل های جدید می توانند به قالب اضافه شده و مطابق آن چه که در آنتولوژی تعریف و معنا می شوند، عمل کنند. یک تکنیک استخراج جدید، تکنیک ارزیابی DM، یا یک مجموعه داده می تواند با استفاده از قالب ها و الگو های موجود برای عوامل، با کاربران دیگر به اشتراک گذاشته شود. هدف این است که با به کارگیری آنتولوژی سیستم بتواند به شکل ارگانیک رشد کند. همچنین سناریوی خوشه بندی نشان می دهد که داده کاوی چند عامله چگونه می تواند بهترین الگوریتم را برای یک استفاده خاص (در این مورد خوشه بندی) انتخاب کند.

شماره	داده	تعداد ویژگی	تعداد ریکال	K-means		KNN			DBSCAN		
				تعداد کلاس	مقیاس F	آستانه	تعداد کلاس	مقیاس F	حداقل اندازه	دانشیه	تعداد کلاس
1	عدسی (لنز)	4	24	3	4.69	1	1	14.4	1	1	4
2	گیاه زنبق و سوسن	4	150	3	44.26	1	4	29.41	1	5	4
3	باغ وحش	18	101	7	10.72	2	9	8.1	2	4	7
4	شراب	13	178	3	40.53	135	4	30.05	10	2700	7
5	قلب	13	270	2	78.97	94	3	59.94	3	55	3
6	اشریشیا کلای	7	336	8	38.39	0.45	6	35.23	1	0.4	7
7	انتقال خون	4	748	2	265.36	1100	6	86.17	15	120	8
8	دیابت سرخپوست های هندی	8	768	2	246.84	135	4	128.41	10	300	5
9	مخمر	8	1484	10	58.84	0.35	9	66.95	2	0.5	9
10	اتومبیل	6	1782	4	176.70	1.45	5	195.95	2	35	4

جدول 2. نتایج خوشه بندی انجام شده در قالب داده کاوی چند عامله

شماره	مجموعه داده	مقیاس F کلی	بهترین الگوریتم
1	عدسی (لنز)	14.4	KNN
2	گیاه زنبق و سوسن	44.26	K-means
3	باغ وحش	11.06	DBSCAN
4	شراب	40.53	K-means
5	قلب	78.97	K-means

DBSCAN	44.89	اشریشیا کلای	6
K-means	265.36	انتقال خون	7
K-means	246.84	دیابت سرخپوست های هندی	8
DBSCAN	89.4	مخمر	9
DBSCAN	226.93	اتومبیل	10

حدول 3. بهترین نتایج خوشه بندی در قالب داده کاوی چند عامله

6. نتیجه گیری

در این مقاله یک قالب داده کاوی چند عامله عمومی توضیح داده شد. هدف از این قالب این است که از داده کاوی چند عامله عمومی، از طریق ایجاد مکانیسمی برای رشد ارگانیک، پشتیبانی کند. این موضوع تا حدودی توسط طراحی قالب داده کاوی چند عامله تسهیل می شود و تا حدودی هم با به کارگیری آنتولوژی. استفاده از این قالب با استفاده از سناریو خوشه بندی داده ترسیم شد. این سناریو کارآیی و تاثیر این قالب را نشان داد.

آنتولوژی استخراج داده، که در بخشی از این مقاله ارائه شد، هنوز کامل نیست. در واقع می توان گفت که هرگز هم کامل نخواهد شد. اگرچه آنتولوژی امروزی قابل گسترش است و محققان در حال افزایش وسعت آن هستند. همچنین جنبه هایی از مکانیسم های ارتباطی وجود دارند که قصد داریم در مطالعات بعدی به آن ها بپردازیم. همانطور که در بخش 4 اشاره شد، با وجود بهبودی که زبان ارتباط عامل فیفا در نسخه قبلی پدید آورد، با انتقاد های فراوانی رو به رو شد. نقطه نظر برخی از این انتقاد ها این است تحقیق و بررسی صحت معنانشناسی آن دشوار است، زبان و قواعد ساختار های کوچک و محدودی را به وجود می آورند؛ مثلاً قواعدی برای جلوگیری از رفتار های در هم گسیخته وجود ندارد، و مورد دیگر اینکه زبان ارتباط داده برای به دست آوردن و خریدن گفتگو ها طراحی شد و برای کاربرد های دیگر مانند داده کاوی مناسب نیست. اگرچه یک انتقاد قابل توجه دیگر از دید ما در مورد این زبان فیفا وجود دارد و آن هم ظرفیت های مورد تردید بسیار محدود است. در آینده امیدواریم که عوامل ما بتوانند در گفتگوهای (دیالوگ های) پرمعنا تری در مورد اینکه بهترین خوشه بندی کدام است و چرا، شرکت کنند (برای مثال بتوانند در مورد اینکه کدام الگوریتم خوشه بندی در یک مورد خاص بهتر است و دلیل آن، بحث کنند). ما قصد داریم در آینده این موضوع را به روش مباحثه مورد بررسی قرار بدهیم که در نهایت به روش ارتباطی پر معنا تری میان عوامل مستقل برسیم.

