روش نوین انطباق هستی‌شناسی با استفاده از پیکره‌های متنی

بعثت کسایی[1] | کارشناس ارشد، مهندسی نرم‌افزار

مسعود رهگذر[2] | دکتری، علوم کامپیوتر

علیرضا وظیفه‌دوست*| دانشجوی دکتری، مهندسی نرم‌افزار

دریافت: ۱۳۹۰/۱۰/۱۲ | پذیرش: ۱۳۹۱/۰۳/۰۱

چکیده: امروزه، استفاده از هستی‌شناسی برای مقاصد گوناگونی در حال گسترش است. اما، در مواقعی به خاطر تفاوت‌هایی که در ساخت هستی‌شناسی در مراکز مختلف وجود دارد، امکان تبادل دانش بین دو هستی‌شناسی میسر نیست. برای حل این مشکل، روش‌های مختلفی برای انطباق هستی‌شناسی ارائه شده است که برخی از آنها مبتنی بر فنون یادگیری ماشین است. انطباق هستی‌شناسی یا به عبارت دیگر تشخیص مفاهیم متناظر در هستی‌شناسی‌های مختلف دارای کاربردهای متنوعی است. در این مقاله، یک روش جدید ارائه شده است که با بهره‌گیری از یادگیری ماشین و نیز پیکره‌های متنی به عنوان منبع دانش از شباهت‌های معنایی بین هستی‌شناسی‌ها جهت انطباق استفاده می‌کند.

کلیدواژه‌ها: انطباق هستی‌شناسی، یادگیری ماشین، پیکره متنی، روش نایوبیز، شباهت معنایی

فصلنامهٔ علمی پژوهشی
پژوهشگاه علوم و فنّاوری اطلاعات ایران
شاپا(چاپی) ۸۲۲۳-۲۲۵۱
شاپا(الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISA، SCOPUS و ISC
http://jipm.irandoc.ac.ir
دوره ۲۸ | شمارهٔ ۳ | ص ص ۸۰۷-۸۲۷
بهار ۱۳۹۲

نوع مقاله: پژوهشی

1. besat_k@yahoo.com
2. rahgozar@ut.ac.ir
*vazifehdst@ut.ac.ir

۱. مقدمه

هستی‌شناسی[۱] در عرصه‌های مختلفی از وب معنایی و محیط‌های چندعامله[۲] گرفته تا بازیابی اطلاعات و مدیریت دانش استفاده می‌شود. به طور مختصر، به مفاهیم سازمان‌یافته در یک قالب مشخص هستی‌شناسی اطلاق می‌شود. در واقع وقتی روابط بین مفاهیم شناسایی و از این مفاهیم و روابط استخراج می‌شود یک ساختار گراف‌گونه تشکیل گردد. این گراف نشانگر یک هستی‌شناسی است. روابط بین مفاهیم یک هستی‌شناسی می‌توانند در دو دسته کلی قرار گیرند: سلسله مراتبی[۳] و غیرسلسله مراتبی[۴].

انطباق هستی‌شناسی‌ها[۵] بر یافتن شباهت‌ها و تناظرات بین مفاهیم موجود در هستی‌شناسی‌های مختلف متمرکز است(Lin 2007) . عمل انطباق در نهایت یا به یافتن شباهت‌های ظاهری مانند شباهت لغوی بین مفاهیم ختم می‌شود و یا به یافتن شباهت‌های معنایی که در هر دو صورت به منظور برقراری ارتباط بین دو هستی‌شناسی مجزا مفید است. با یافتن روابط بین هستی‌شناسی‌ها، ضروری نیست برای جدید هستی‌شناسی جدیدی ایجاد کنیم و از هستی‌شناسی‌های موجود استفاده می‌کنیم. از جمله حوزه‌هایی که انطباق هستی‌شناسی دارای کاربرد ملموس است می‌توان به نظام‌های چندعامله (Ghamrawy 2009) و وب معنایی[۶] (Doan et al.2003) اشاره نمود.

یکی از کاربردهای مهم هستی‌شناسی برطرف نمودن ناهمگونی‌های موجود در داده‌هاست، اما از طرفی خود هستی‌شناسی‌ها نیز دارای ناهمگونی‌هایی هستند (Ehrig, Euzenat, and Castro 2004). برای مثال، هستی‌شناسی‌ها می‌توانند با زبان‌های متفاوتی بیان شوند، مانند "او دبلیو ال"[۷] یا "آر دی اف"[۸] و هر زبان، نحو[۹]، نمایش منطقی[۱۰] و دیگر ویژگی‌های خاص خود را دارد و این عامل باعث بروز ناهمگونی در هستی‌شناسی‌ها می‌گردد. البته داشتن زبان مشترک نیز تضمینی برای همگون بودن هستی‌شناسی‌ها نیست و ممکن است برای یک مفهوم خاص از واژه‌های مختلفی در هستی‌شناسی‌ها استفاده شود. این ناهمگونی‌ها می‌توانند در سطح زبان و یا در سطح هستی‌شناسی بروز نماید. ناهمگونی سطح زبان به‌دلیل نحوهای متفاوت زبان‌ها ایجاد می‌گردد. از طرفی، ممکن است از واژه‌های یکسان برای بیان مفاهیم متفاوت استفاده شود. به عنوان مثال، کلمه "کنداکتور"[۱۱] ممکن است هم در هستی‌شناسی مربوط به موزیک و هم در هستی‌شناسی مربوط به مهندسی برق بیان شود. با اینکه واژه یکسان است، اما دارای مفاهیم متفاوتی است به این پدیده ناهمگونی در سطح هستی‌شناسی اطلاق می‌گردد.

---

1. Ontology
2. multi-agent systems
3. taxonomic
4. non taxonomic
5. ontology Matching
6. semantic Web
7. OWL
8. RDF
9. syntax
10. logical representation
11. conductor

در ادامه مقاله، به مروری کلی بر کارهای مرتبط انجام‌شده در این حوزه تمرکز شده است و سپس، روش پیشنهادشده با جزئیات بیان می‌گردد. در نهایت، نحوه پیاده‌سازی و ارزیابی از نتایج به‌دست‌آمده بیان می‌گردد.

۲. پیشینه پژوهش

تاکنون روش‌های متعددی به منظور انجام انطباق هستی‌شناسی ارائه شده است. از بین روش‌های ارائه‌شده، روش‌های مبتنی بر یافتن شباهت‌های لغوی و ساختاری Dieng and Hug (Cohen, Ravikumar, and Fienberg2003; Bach, Kuntz, and Gandon 2004*)*; 1998; Maedche and Staab 2002 شناخته‌شده‌تر هستند. این مقاله، بیشتر به روش‌هایی می‌پردازد که شباهت‌ها را بر اساس نمونه‌ها[1] تشخیص می‌دهند. در روش‌های مبتنی بر شباهت نمونه‌ای در حقیقت، شباهت دو مفهوم بر اساس شباهت بین نمونه‌های دو مفهوم سنجیده می‌شود.

لازم به اشاره است که استفاده از پیکره‌های متنی[2] و روش‌های یادگیری ماشین در انطباق هستی‌شناسی رویکرد جدیدی نیست. پیش از این، کارهایی توسط ناتاراجان (Natarajan 2005) و دون و دیگران (Doan et al. 2003) برای یافتن انطباق بین هستی‌شناسی‌ها بر اساس شباهت نمونه‌ای و روش‌های یادگیری ماشین انجام گرفته است. این کارها به ترتیب از روش یادگیری نایوبیز[3] و فرمول شباهت جاکارد[4] برای محاسبه میزان شباهت بین دو مفهوم استفاده کرده‌اند. در واقع، این روش‌ها میزان شباهت بین دو مفهوم را بر اساس میزان اشتراک بین نمونه‌های دو مفهوم و با استفاده از فرمول جاکارد محاسبه می‌کنند.

روش استخراج ویژگی‌های مفاهیم برای استفاده در فرآیند یادگیری ماشین می‌تواند آثار مثبت و یا منفی بر کیفیت یادگیری داشته باشد. در روش‌های پیشین، ویژگی‌های مورد نیاز از یک منبع محدود اطلاعاتی که محتویات خود هستی‌شناسی در اختیار می‌گذارد، تأمین می‌شود. واضح است که دانش محدود ما در مورد یک نمونه (مفهوم) می‌تواند اثر سوء بر کیفیت یادگیری داشته باشد. سو و گولا ویژگی‌های با ارزش یک مفهوم را از برخی مستندات مرتبط به آن استخراج کرده‌اند که این مسأله باعث افزایش شناخت ما از مفاهیم و در نتیجه، بهبود کیفیت قضاوت ما در مورد میزان شباهت میان دو مفهوم می‌شود (Su and Gulla. 2006). به‌طور مشابه، در روش پیشنهادی مقاله حاضر نیز ویژگی‌های مورد نیاز برای انجام یادگیری از یک پیکره متنی در دامنه اطلاعاتی یکسان با هستی‌شناسی استخراج می‌گردد. این روش باعث

1. instance 2. textual corpus 3. naive bayes 4. jaccard similarity formula

می‌شود کیفیت ویژگی‌های نسبت داده شده به یک مفهوم بهتر و در نتیجه یادگیری بهتر انجام شود و نتایج انطباق قابل قبول‌تری نیز حاصل گردد.

به دلیل برخی شباهت‌های روش پیشنهادی در مقاله حاضر با روش ارائه‌شده توسط سو و گولا، در این قسمت به بیان تفاوت‌های این دو روش می‌پردازیم.

به‌طور کلی، نحوه استخراج ویژگی‌ها در روش پیشنهادی این مقاله چند تفاوت اساسی با روش مطرح‌شده توسط سو و گولا دارد. اولین تفاوت از شیوه ساخت پیکره‌های متنی ناشی می‌شود. در روش پیشنهادی سو و گولا، پیکره متنی به صورت دستی ساخته شده است، در حالی که در مقاله حاضر این پیکره با استفاده از موتور جستجوی برخط[1] و با کمترین دخالت انسانی ساخته شده است. دومین تفاوت در نحوه محاسبه مقدار یک ویژگی است. سو و گولا به سادگی و فقط با استفاده از پارامتر tf.idf مقادیر ویژگی‌ها را محاسبه می‌کنند در حالی که در روش پیشنهادی مقاله حاضر از مقدار هم‌تکراری هر ویژگی با مفهوم مرتبط با آن، برای محاسبه مقدار یک ویژگی استفاده شده است. پارامتر tf.idf فقط میزان اهمیت یک کلمه را در مشخص نمودن یک مستند نشان می‌دهد و در حالی که هم‌تکراری، میزان اهمیت یک ویژگی در شناخت یک مفهوم را مشخص می‌نماید. به همین دلیل می‌توان ثابت نمود که روش پیشنهادی مقاله حاضر برای محاسبه مقدار ویژگی‌ها دارای ارزش اطلاعات بیشتری است.

آخرین تفاوت روش ارائه‌شده در مقاله حاضر با روش ارائه‌شده توسط سو و گولا در نحوه محاسبه شباهت است. در هر دو روش از فرمول جاکارد برای محاسبه شباهت استفاده شده است، اما در مقاله حاضر از یادگیری ماشین، به منظور سنجش دقیق‌تر شباهت دو مفهوم استفاده شده است که این روش به ما کمک می‌کند با در نظر گرفتن شباهت مفهومی موجود بین دو مفهوم فقط بر اساس شباهت ظاهری تصمیم نگیریم. در حقیقت، سو و گولا از فرمول جاکارد برای محاسبه شباهت بر اساس ظاهر استفاده نموده‌اند که در موارد متعددی ممکن است به نتایج و تصمیمات اشتباه منجر شود. برای مثال، بر اساس شباهت رشته‌ای ممکن است دو واژه word و world بسیار شبیه به حساب آیند، در حالی که هیچ ارتباط معنایی با هم ندارند.

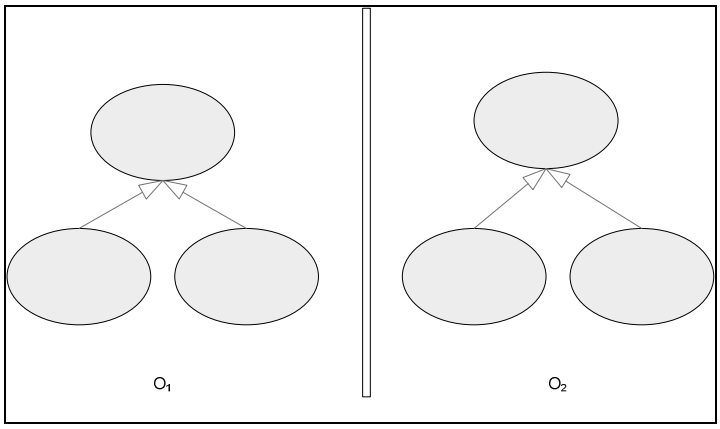
در ادامه، به شرح روش پیشنهادی مقاله حاضر در خصوص استفاده از پیکره متنی در انطباق هستی‌شناسی پرداخته می‌شود و توضیحاتی در خصوص نحوه پیاده‌سازی ارائه می‌گردد. در نهایت، به ارائه نتایج حاصل از پیاده‌سازی این روش و نیز ارزیابی و بررسی نتایج به‌دست‌آمده پرداخته می‌شود.

---

1. online

۳. روش پژوهش

هر مفهوم موجود در هستی‌شناسی حامل میزانی از دانش حوزه آن هستی‌شناسی است. مفاهیم موجود در هستی‌شناسی‌ها از طریق برخی ارتباطات استاندارد و تعریف‌شده به هم مرتبط شده‌اند. دو مورد از مهم‌ترین ارتباطات موجود در هستی‌شناسی روابط "IS-A" و "Part-of" است. در ادامه مقاله، به مفاهیمی که با رابطه "IS-A" به مفهوم A مرتبط شده‌اند، نمونه‌های مفهوم A اطلاق می‌گردد. همچنین، فرض بر این است که مقدار شباهت بین دو مفهوم A و B از دو هستی‌شناسی مختلف به میزان اشتراک نمونه‌های دو مفهوم مرتبط است. البته لازم به اشاره است که اشتراک نمونه‌ها بر پایه معنی استخراج می‌گردد و نه ظاهر آنها. به عنوان مثال، در دو هستی‌شناسی شکل ۱، آلفا-قنطورس می‌تواند نمونه‌ای از ستاره باشد به خاطر اینکه خصوصیاتی مشترک با خورشید که نمونه دیگری از ستاره است، دارد.



شکل ۱. آلفا-قنطورس به دلیل خصوصیات مشترک با خورشید می‌تواند نمونه ای از ستاره باشد.

برای محاسبه میزان شباهت از فرمول جاکارد به شرح زیر استفاده شده است:

$$Jaccard - Sim(A,B) = \frac{P(A \cap B)}{P(A \cup B)} = \frac{P(A,B)}{P(A,B)+P(A,\bar{B})+P(\bar{A},B)}$$ فرمول ۱

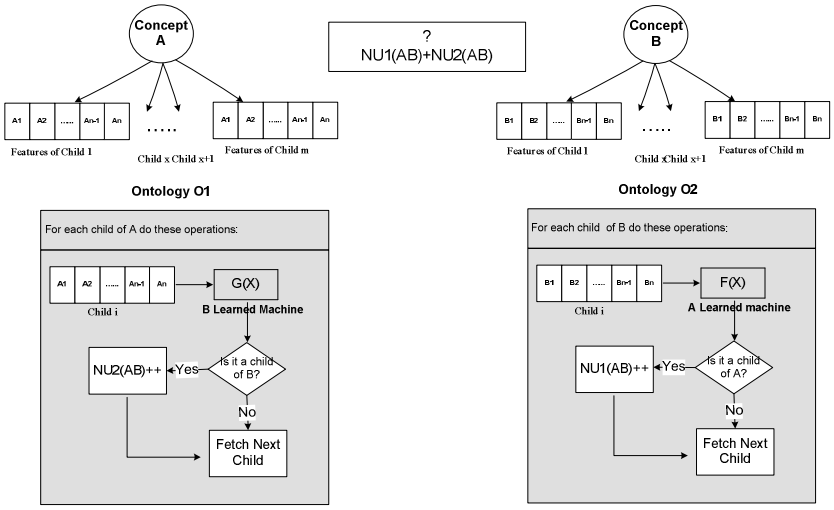
این فرمول از چندین مؤلفه مختلف تشکیل شده است، اما تمامی مؤلفه‌ها به روش یکسانی محاسبه می‌شوند. بدین دلیل، برای نمونه فقط نحوه محاسبه یکی از اجزا را در این قسمت توضیح می‌دهیم. نحوه محاسبه P(A,B) در فرمول ۲ آمده است.

$$P(A,B) = \frac{N(U_1^{A,B}) + N(U_2^{A,B})}{N(U_1) + N(U_2)}$$ فرمول ۲

در محاسبه و تخمین مؤلفه‌های مختلف این معادله از یادگیری ماشین استفاده می‌کنیم. فرض کنیم $N(U_1^{A,B})$ نشان‌دهنده تعداد نمونه‌هایی از هستی‌شناسی O_1 باشد که می‌تواند به هر دو مفهوم A (از هستی‌شناسی O_1) و B (ازهستی‌شناسی O_2) متعلق باشد.

به همین ترتیب $N(U_2^{A,B})$ نیز با جایگذاری O_2 به جای O_1 معنی مشابهی دارد. برای محاسبه $N(U_1^{A,B})$ نیازمند دانستن تعداد نمونه‌های مشترک بین دو مفهوم A و B هستیم. فرض کنید نمونه s را از هستی‌شناسی O_1 انتخاب کرده‌ایم. فهمیدن اینکه آیا s متعلق به نمونه‌های A هست یا خیر راحت است، زیرا این مفاهیم هر دو در یک هستی‌شناسی هستند. مشکل زمانی پیش می‌آید که می‌خواهیم در مورد امکان عضویت s به مجموعه نمونه‌های مفهوم B از هستی‌شناسی O_2 اظهار نظر کنیم. در واقع، s شاید به ظاهر و در نگاه اول جزء نمونه‌های مفهوم B نباشد، اما گاهی می‌توان به دلیل شباهت معنایی بالایی که این نمونه با نمونه‌های مفهوم B دارد، s را جزء نمونه‌های B محسوب کرد. البته این روش ممکن است نتایجی معکوس در مورد شباهت‌های ظاهری (مانند شباهت رشته‌ای)، بین نمونه‌های B و s داشته باشد، بدین معنی که شباهت‌های ظاهری را نادیده بگیرد. بعد از تعیین عضویت یک نمونه به مفهوم A در هستی‌شناسی اول باید در خصوص عضویت آن به مفهوم B در هستی‌شناسی دوم تصمیم‌گیری کنیم و در این رابطه نیازمند استفاده از فنون یادگیری ماشین هستیم. فرض کنیم در مورد نمونه‌های جاری مفهوم B و نیز ویژگی‌های آن اطلاعات کافی در دست داریم. پس می‌توانیم از این اطلاعات به عنوان مجموعه آموزشی[1] ماشینی که دارای قابلیت تشخیص و تمایز بین دو دسته Z_1 و Z_2 است، استفاده نماییم. Z_1 نشان‌دهنده آن دسته از مفاهیمی است که می‌توانند به عنوان نمونه‌های مفهوم B در نظر گرفته شوند و Z_2 نشانگر آن دسته از مفاهیمی است که در زمره نمونه‌های B قرار نمی‌گیرند. فرآیند توضیح داده‌شده، در شکل ۲ به صورت تصویری نمایش داده شده است. حال پس از انجام فرآیند یادگیری ماشین ما می‌توانیم هر نمونه دلخواهی مانند s از هستی‌شناسی O_1 را به ماشین مرتبط با مفهوم B داده و ماشین در مورد عضویت آن به نمونه‌های B تصمیم بگیرد. در صورتی که جواب مثبت بود می‌توانیم به تعداد نمونه‌های مشترک بین A و B اضافه کنیم (شکل ۲).

1. training set

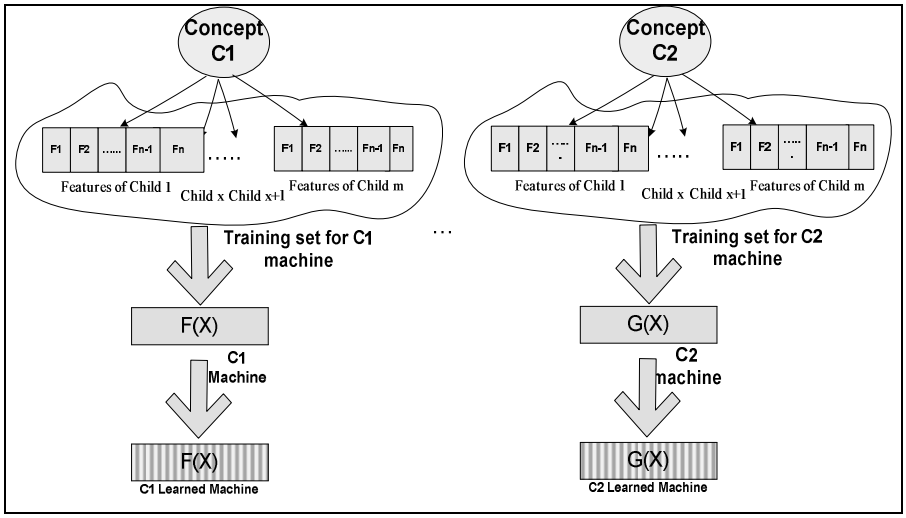


شکل۲. فرآیند یادگیری برای مفهوم A از هستی‌شناسی اول و مفهوم B از هستی‌شناسی دوم

در کارهای انجام‌شده قبلی ویژگی‌های مختلفی از مفاهیم در فرآیند یادگیری استفاده می‌شود. برای مثال، دون و دیگران از دو نوع یادگیرنده استفاده کردند، که هر کدام از ویژگی‌های مختلفی از مفاهیم را به منظور یادگیری استفاده می‌نماید و یادگیرنده سوم نتایج این دو را با هم ترکیب می‌کند. یادگیرنده اول از تعداد واژه‌های موجود در محتوای متنی[1] یک مفهوم استفاده می‌نماید و دومین یادگیرنده از نام کامل یک مفهوم که از کنار هم قرار دادن متوالی نام مفاهیم از ریشه تا مفهوم مورد بحث درست می‌شود به عنوان ویژگی در فرآیند یادگیری استفاده می‌نماید (Doan et al. 2003).

پژوهش ماخول و دیگران نیز ادامه‌ای است بر کار انجام‌شده توسط دون و دیگران که از برخی ویژگی‌ها مانند نام کامل استفاده نموده است، اما علاوه بر آن، یادگیری بر پایه نزدیک‌ترین همسایه نیز مورد کاربرد است. در این مقاله، از یک پیکره متنی در فرآیند انطباق هستی‌شناسی‌ها استفاده شده است. بدین معنی که شباهت مستنداتی که مفاهیم مختلف در آنها دیده می‌شود نشانه‌ای برای شباهت خود مفاهیم است. ایده جدید ارائه‌شده در این مقاله استخراج ویژگی‌های یک مفهوم از یک پیکره متنی است. مفهوم ویژگی در این مقاله به واژه‌های درون یک پیکره متنی اشاره دارد. مقدار یک ویژگی (واژه)، همان میزان هم‌تکراری آن واژه با مفهومی از هستی‌شناسی را که مورد بحث است نشان می‌دهد (Makhoul et al. 1999).

1. textual content



شکل ۳. چگونگی محاسبهء مولفه‌های فرمول جاکارد با استفاده از یادگیری ماشین

مقدار هم‌تکراری می‌تواند با استفاده از فرمول ۳ محاسبه شود.

$$Co - occurrence\ \ Value(A, B) = \frac{N_d(A,B)}{N_d(A)}$$ فرمول ۳

در این معادله، N_d (A, B) تعداد مستنداتی از پیکره متنی را که مفهوم A از هستی‌شناسی مورد نظر و واژه B از پیکره متنی مرتبط با هم آمده‌اند، نشان می‌دهد. N_d (A) نیز اشاره دارد به تعداد کل مستندات موجود در این پیکره متنی که فقط حاوی مفهوم A است. ایده مطرح‌شده این است که هم‌تکراری بیشتر نشان‌دهنده ارتباط قوی‌تری بین یک مفهوم و ویژگی مربوط است.

جدول ۱. نمایش ویژگی‌ها و مقادیر آنها برای سه مفهوم فرضی

Peace	Expression	…..	Date	Name	Title	Address	ویژگی / مفهوم
0.01	0	…..	0.21	0.65	0.20	0.75	Author
0	0	…..	0.11	0.60	0.12	0.81	Publisher
0	0.1	…..	0.66	0.70	0.80	0.40	Journal

حال با داشتن جدولی از مفاهیم مانند جدول ۱ و نیز ویژگی‌های آنها می‌توانیم انواع روش‌های یادگیری موجود را به‌کار ببریم. در این مقاله، از روش نایوبیز که روشی ساده و مؤثر است، بهره جسته‌ایم. طبقه‌بند[1] بیز بر اساس نظریه بیز، احتمال قرارگیری یک مفهوم در یک دسته را محاسبه می‌نماید. فرض کنیم که X نشان‌دهنده یک بردار از ویژگی‌های یک مفهوم است مانند: $X(x_1,x_2,x_3,.....x_k)$. حال می‌خواهیم احتمال شرطی P(H|X) را محاسبه نماییم. H نشاندهنده یک متغیر تصادفی است که احتمال تعلق یک مفهوم به دسته c را نشان می‌دهد. این احتمال شرطی می‌تواند به صورت فرمول ۴ نوشته شود.

$$P(H = c|X) = \frac{P(X|H=c)P(H=c)}{P(X)}$$ فرمول ۴

در این فرمول، P(H) احتمال وقوع دسته c است. می‌توان ثابت کرد که به فرض گسسته بودن مقادیر ویژگی‌ها و نیز مستقل بودن متغییرهای تصادفی و اینکه فقط دو دسته Z_1 و Z_2 داشته باشیم، فرمول ۴ را می‌توان به صورت زیر نوشت (Nilsson1990):

$$g(X)=\sum_{i=1}^{d} x_i \log\left[\frac{p_i(1-q_i)}{q_i(1-p_i)}\right] + \sum_{i=1}^{d} \log\left(\frac{1-p_i}{1-q_i}\right) + \log\left[\frac{p(1)}{1-p(1)}\right]$$ فرمول ۵

اگر $g(x) > 0$ آنگاه X متعلق به دسته ۱ و اگر $g(x) < 0$ آنگاه X متعلق به دسته ۲ خواهد بود. d نشان‌دهنده تعداد ویژگی‌هاست و p_i و q_i نیز به صورت فرمول ۶ محاسبه می‌شوند $(i = 1, \ldots, d)$:

$$p(x_i = 1|1) \triangleq p_i$$
$$p(x_i = 0|1) \triangleq 1 - p_i$$
$$p(x_i = 1|2) \triangleq q_i$$
$$p(x_i = 0|2) \triangleq 1 - q_i$$

فرمول ۶

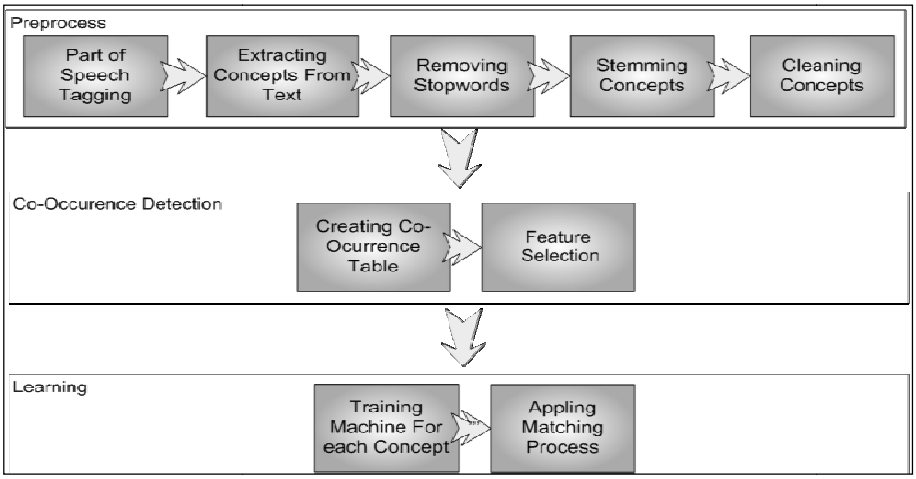
بنابراین، کل فرآیند یادگیری ماشین به یافتن مؤلفه‌های این ضرایب از مجموعه آموزشی کاهش می‌یابد که به آسانی قابل پیاده‌سازی است.

1. classifier

٤. پیاده‌سازی و نتایج

در این مقاله، آزمایش‌ها بر روی دو هستی‌شناسی (هستی‌شناسی‌های شماره ۱۰۱و۳۰۴) از منابع ارائه‌شده در کنفرانس سالانه OAIE[1] (Ontology Alignment Initiative Evaluation) انجام شده است. هر دو هستی‌شناسی در حوزه کتاب‌شناختی است و در کل، ۷۵ مفهوم در دو هستی‌شناسی موجود است. برای ایجاد پیکره متنی از موتور جستجوی Google استفاده شده است. برای این منظور، عبارت زیر به ازای هر مفهوم موجود در هستی‌شناسی‌ها جستجو شده است.

ConceptName + Bibliographic Information + filetype:doc



شکل ٤. فرآیند پیاده‌سازی روش پیشنهادی

در نهایت، ۵۰ مستندی که در بالاترین سطح رتبه‌بندی برگردانده شده بودند انتخاب و در داخل پیکره متنی گنجانده شده است. با این کار سعی کرده‌ایم مستنداتی که به حوزه کتاب‌شناختی مربوط می‌شود و توسط گوگل در رتبه بالاتری رده‌بندی شده‌اند، به عنوان مستندات مرتبط جمع‌آوری کنیم و در پیکره متنی استفاده نماییم. در نهایت، یک پیکره متنی شامل ۵۰۵ مستند با حجم ۱۶ مگابایت حاصل شد.

برای انجام انطباق، ۳ مرحله کلی طراحی شده است (شکل ۴). مرحله اول پیش‌پردازش اطلاعات است. در این مرحله، مستندات موجود در پیکره متنی برای استفاده آماده‌سازی

۱. کنفرانس OAIE سالانه برگزار میشود و به بررسی و ارزیابی الگوریتم‌های انطباق ارائه‌شده در کنفرانس می‌پردازد.

می‌شوند. در مرحله پیش‌پردازش از ابزار GATE (Cunningham and Wilks 1996) استفاده کرده‌ایم. مراحل پیش‌پردازش اطلاعات به طور کلی شامل زدن برچسب‌های نحوی[1]، استخراج مفاهیم از متن و انتقال آنها به پایگاه داده، حذف Stopwords، ریشه‌یابی مفاهیم، پاک‌سازی مفاهیم است.

در گام بعد، میزان هم‌تکراری واژه‌های موجود در پیکره متنی با مفاهیم موجود در هستی‌شناسی‌ها محاسبه می‌شود. به منظور کاهش پیچیدگی برنامه و کم اثر نمودن ویژگی‌های بی‌اهمیت، یک زیرمجموعه از بین تمامی ویژگی‌های استخراج‌شده از پیکره متنی که میزان هم‌تکراری آنها با مفاهیم هستی‌شناسی‌ها محاسبه شده است، انتخاب می‌کنیم. برای انتخاب ویژگی‌های مفید دو پارامتر در نظر گرفته شده است. اولین پارامتر ZeroParam و دومین پارامتر Entropy نامگذاری شده است که در ادامه به شرح آنها می‌پردازیم.

٤-١. نحوه محاسبه ZeroParam

ویژگی‌های انتخاب‌شده در مورد همه مفاهیم و نه فقط دسته کوچکی از آنها باید مفید باشد. برخی از مفاهیم فقط با تعداد محدودی از ویژگی‌ها هم‌تکراری دارند و اگر تعداد زیادی از این ویژگی‌ها در مجموعه انتخابی موجود نباشند، با مفهومی مواجه خواهیم شد که برای بیشتر ویژگی‌هایش مقدار صفر دارد. این مسأله به فرآیند یادگیری غیردقیقی منجر خواهد شد. پس بهتر است که از شرایطی که در آن تعداد زیادی از ویژگی‌ها برای تعداد زیادی از مفاهیم مقدار صفر دارد، پرهیز کنیم. این مسأله ما را به تعریف یک عامل یا شاخص که میزان اهمیت یک ویژگی به ازای هر مفهوم را مشخص می‌کند سوق می‌دهد.

این پارامتر میزان اهمیت هر ویژگی غیرصفر را با توجه به مفهوم مرتبط نشان می‌دهد. بنابراین، فرمول ۷ را به ازای هر مفهوم ارائه کرده‌ایم که میزان اهمیت را محاسبه می‌نماید.

$$ConceptZeroParam(C) = \frac{Count\ of\ all\ features}{Count\ of\ NonZero\ Features}$$ فرمول ۷

صورت کسر نشانگر تعداد کل ویژگی‌هاست و مخرج کسر نشانگر تعداد ویژگی‌های غیرصفر مفهوم c است. این پارامتر میزان اهمیت یک ویژگی نوعی با مقدار غیرصفر را در ازای یک مفهوم اندازه‌گیری می‌کند. حال مقدار ZeroParam برای یک ویژگی به صورت فرمول ۸ محاسبه می‌گردد.

1. part of speech tagging

فرمول ۸ $FeatureZeroParam(f) = \sum_{c \epsilon Concepts} ConceptZeroParam\ (c)$

Where valueof (f) is not zero

فرمول ۸ نشان‌دهنده یک جمع ساده بر روی کل ConceptZeroParams موجود است، در جاهایی که ویژگی f مقدار غیرصفر دارد. با این پارامتر قادر هستیم ویژگی‌هایی را که برای بیشتر مفاهیم مقدار غیرصفر دارد، انتخاب نماییم.

Entropy .۲-٤

به نظر می‌رسد در فرآیند یادگیری ماشین ویژگی‌هایی که مقادیر متنوع‌تری به ازای هر مفهوم دارند، ویژگی بهتری برای یادگیری و انجام عمل طبقه‌بندی[1] هستند. بنابراین، ما به دنبال انتخاب ویژگی‌هایی هستیم که بیشترین میزان بی‌نظمی و یا تنوع را در مقادیرشان دارند. برای اندازه‌گیری این تنوع، ما از فرمول استاندارد سنجش آنتروپی (Shannon 1948) به صورت فرمول ۹ استفاده می‌کنیم.

فرمول ۹ $Entropy\ (f) = \sum_{d \in Distinct\ Values\ of\ f} P(d)\ log_2 \frac{1}{P(d)}$

در این معادله P(d) احتمال وقوع یک مقدار معین d بین تمامی مقادیر قابل قبول برای ویژگی f است. بر اساس این معادله، هر چه میزان تنوع مقادیر یک ویژگی بیشتر باشد، آنتروپی آن بالاتر خواهد بود. حال با استفاده از این پارامترها یک پارامتر ترکیبی برای هر ویژگی تعریف می‌گردد (فرمول ۱۰).

فرمول ۱۰ $FeatureQuality\ (f) = Entropy(f) * FeatureZeroParam(f)$

در نهایت، ویژگی‌هایی با بیشترین مقدار FeatureQuality برای استفاده در فرآیند یادگیری انتخاب می‌شوند. با به‌کارگیری فرمول ۱۰ بر داده‌های موجود، ۵۰۰ ویژگی مناسب برای استفاده در این پژوهش انتخاب گردید که نمونه‌ای از این ویژگی‌ها در جدول ۲ نشان داده شده است.

1. classification

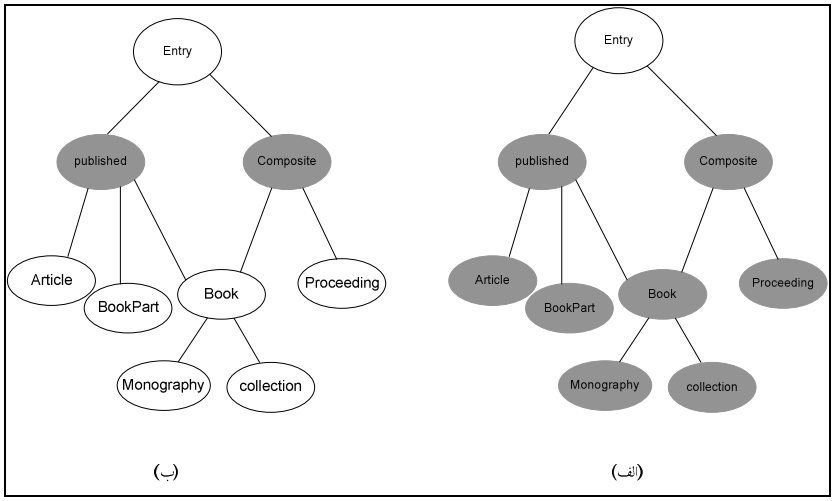
٤-٣. انتخاب فرزندان یک مفهوم

موضوع دیگری که باید در زمان پیاده‌سازی مورد توجه قرار گیرد، روش تعریف مجموعه فرزندان یک مفهوم است. در این خصوص از دو روش می‌توان استفاده کرد(شکل ۵). در روش اول، تمامی مفاهیم زیر درخت منشعب‌شده از یک مفهوم به عنوان فرزندان آن مفهوم در نظر گرفته می‌شود. در روش دوم، فقط مفاهیمی که با رابطه IS-A و به صورت مستقیم به یک مفهوم مرتبط هستند، به عنوان فرزندان آن مفهوم در نظر گرفته می‌شوند. واضح است که نتایج حاصل از این دو روش متفاوت خواهند بود. در این مقاله، روش دوم پیاده‌سازی شده است. در کارهای آتی روش اول نیز پیاده‌سازی شده است و نتایج مقایسه خواهند شد.

جدول ۲. تعدادی از ویژگی‌های انتخاب شده

Term Semmed	ZeroParam	Entropy	Feature Quality	TermSemmed	ZeroParam	Entropy	Feature Quality
bibliographi	1135.4804	3.5554	4037.0311	rang	893.6693	3.4965	3124.7243
item	1135.4804	3.5554	4037.0311	present	883.4833	3.5264	3115.4831
ht ml	1135.4804	3.5554	4037.0311	nation	883.4833	3.5264	3115.4831
citat	1135.4804	3.5264	4004.1165	manag	883.4833	3.5264	3115.4831
date	1135.4804	3.5264	4004.1165	brown	876.1202	3.5264	3089.5183
proceed	1135.4804	3.5264	4004.1165	new	876.1202	3.4965	3063.3636
author	1135.4804	3.5264	4004.1165	master	876.1202	3.4965	3063.3636
year	1135.4804	3.5264	4004.1165	volum	860.5474	3.5554	3059.5480
juli	1135.4804	3.5264	4004.1165	demonstr	867.2892	3.5264	3058.3768
issu	1135.4804	3.5264	4004.1165	nd	867.2892	3.5264	3058.3768
collabor	1135.4804	3.5264	4004.1165	august	855.0406	3.5554	3039.9692
inform	1135.4804	3.5264	4004.1165	martin	855.0406	3.5554	3039.9692

در نهایت، شبیه‌ترین مفاهیم از دو هستی‌شناسی برای انطباق انتخاب می‌گردند. برای هر کدام از انطباق‌های شناسایی‌شده، عددی در نظر گرفته می‌شود که میزان دقت انطباق را نشان می‌دهد.



شکل ۵. الف) انتخاب کل زیردرخت به عنوان فرزندان یک مفهوم، ب) انتخاب فرزندان مستقیم

در این مقاله، میزان شباهت حاصل از فرمول جاکارد به‌عنوان میزان دقت رابطه استفاده شده است. یک نمونه از انطباق‌های یافت‌شده در جدول ۳ نشان داده شده است. با بررسی داده‌های موجود در جدول ۳ تفاوت روش پیشنهادی با روش‌های مبتنی بر شباهت ظاهری مانند شباهت لغوی به طور کامل، مشهود است. برای مثال، میزان شباهت محاسبه شده بین دو مفهوم (composite و book) از میزان شباهت محاسبه‌شده بین دو مفهوم (part و part) بیشتر است. زیرا در حوزه دو هستی‌شناسی مورد انطباق در این پژوهش دو مفهوم (composite و book) از نظر معنایی بیشتر به هم شبیه هستند تا دو مفهوم part که در دو هستی‌شناسی جداگانه و با دو بار معنایی متفاوت مورد استفاده قرار گرفته‌اند.

۵. تجزیه و تحلیل یافته‌ها

روش‌های مختلفی برای ارزیابی کیفی نتایج آزمایش‌ها موجود است. یک روش ارائه یک انطباق استاندارد است (R) که نتایج حاصل از الگوریتم پیشنهادی (A) باید با نتایج ارائه‌شده در این مرجع سنجیده شود.

به منظور ارزیابی نتایج به‌دست‌آمده، سه معیار شناخته‌شده recall، precision و f-measure محاسبه شده است (Evzenate, Marc ,and Castro Raul 2005). در این قسمت، به توصیف این معیارها از دید انطباق هستی‌شناسی‌ها می‌پردازیم. فرض کنید A یک انطباق داده شده و R یک انطباق مرجع شکل ۶ است معیار precision برای A به شکل فرمول ۱۱ محاسبه می‌گردد.

فرمول ۱۱ $$P(A,R)=\frac{|A\cap R|}{|A|}$$

```
<?xml version='1.0' encoding='utf-8' standalone='no'?>
<!DOCTYPE rdf:RDF SYSTEM "align.dtd">
<rdf:RDF xmlns='http://knowledgeweb.semanticweb.org/heterogeneity/alignment'
xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
xmlns:xsd='http://www.w3.org/2001/XMLSchema#'>
<Alignment>
<xml>yes</xml>
<level>0</level>
<type>**</type>
<onto1>http://www.example.org/ontology1</onto1>
<onto2>http://www.example.org/ontology2</onto2>
<map>
<Cell>
<entity1 rdf:resource='http://www.example.org/ontology1#reviewedarticle'/>
<entity2 rdf:resource='http://www.example.org/ontology2#article'/>
<measure rdf:datatype='&xsd;float'>0.6363636363636364</measure>
<relation>=</relation>
</Cell>
</map>
<map>
<Cell>
<entity1 rdf:resource='http://www.example.org/ontology1#journalarticle'/>
<entity2 rdf:resource='http://www.example.org/ontology2#journalarticle'/>
<measure rdf:datatype='&xsd;float'>1.0</measure>
<relation>=</relation>
</Cell>
</map>
</Alignment>
</rdf:RDF>
```

شکل ٦. نمونه‌ای از انطباق مرجع R

|A ∩ R| نشان‌دهنده تعداد انطباق‌های صحیح تشخیص داده‌شده و |A| نشان‌دهنده تعداد کل انطباق‌های یافت شده است. لازم به اشاره است که یک انطباق مرجع مانند R دربرگیرنده انطباق‌های صحیح و پذیرفته شده است. در این مقاله، انطباق مرجع ارائه‌شده توسط (OAIE 2010) مورد استفاده قرار گرفته است (Benchmark test Library of OAIE 2010). همچنین، از ابزارهای استاندارد این کنفرانس برای محاسبه recall، precision و f-measure بهره گرفته‌ایم.

جدول ۳. نمونه‌ای از نتایج به‌دست‌آمده از اجرای روش پیشنهادی

Concept1	Concept2	Similarity
reference	Entry	0.33015873
book	Composite	0.55
informal	Informal	0.537378115
part	Part	0.457746479
academic	mastersthesis	0.493055556
academic	Phdthesis	0.493055556
misc	Misc	1
report	deliverable	0.493055556
report	Techreport	0.493055556
motionpicture	motionpicture	1
journal	Journal	0.986486486
conference	Journal	0.486666667
address	Entry	0.025345622
institution	Institution	0.473333333
institution	School	0.473333333
institution	Publisher	0.473333333

Recall از معادله فرمول ۱۲ محاسبه می گردد.

فرمول ۱۲ $$R(A,R) = \frac{|A \cap R|}{|R|}$$

در اینجا $|R|$ تعداد کل انطباق‌های قابل شناسایی را نشان می‌دهد. معیار f-measure از ترکیب دو معیار recall و precision به شکل فرمول ۱۳ حاصل می گردد.

فرمول ۱۳ $$M_\alpha(A,R) = \frac{P(A,R).R(A,R)}{(1-\alpha)\times P(A,R)+\alpha \times R(A,R)}$$

دراین مقاله، برای α مقدار ۰.۵ در نظر گرفته شده است. جدول ۴ نشان‌دهنده میزان بهبود یافتگی یا تخریب نتایج نظام پیشنهادی در مقایسه با نظام‌های دیگر است. میزان تخریب یا بهبود با DoI[1] به صورت فرمول ۱۴ محاسبه شده است.

1. Destruction or Improvement

$$DoI = \frac{the\ result\ of\ Proposed\ System - The\ result\ of\ SystemX}{the\ result\ of\ System\ X} \times 100$$ فرمول ۱۴

جدول ۴ بر اساس رابطه فرمول ۱۴ محاسبه و تنظیم شده است. این جدول نشان می‌دهد برای نمونه نظام پیشنهادی این پژوهش در مقایسه با نظام Edna در معیار precision نزدیک به ۳۱ درصد بهبود ایجاد کرده است و یا نظام پیشنهادی در مقایسه با نظام MapPSO در معیار Recall نزدیک به ۱۷ درصد تخریب داشته است. متوسط میزان (بهبود یا تخریب) DoI نظام پیشنهادی برای معیار precision، ۱۶/۵۹- درصد، برای معیار recall ۵۰/۵۶+ درصد و برای معیار fmeasure، ۱۳/۴۱+ درصد نسبت به نظام‌های دیگر است.

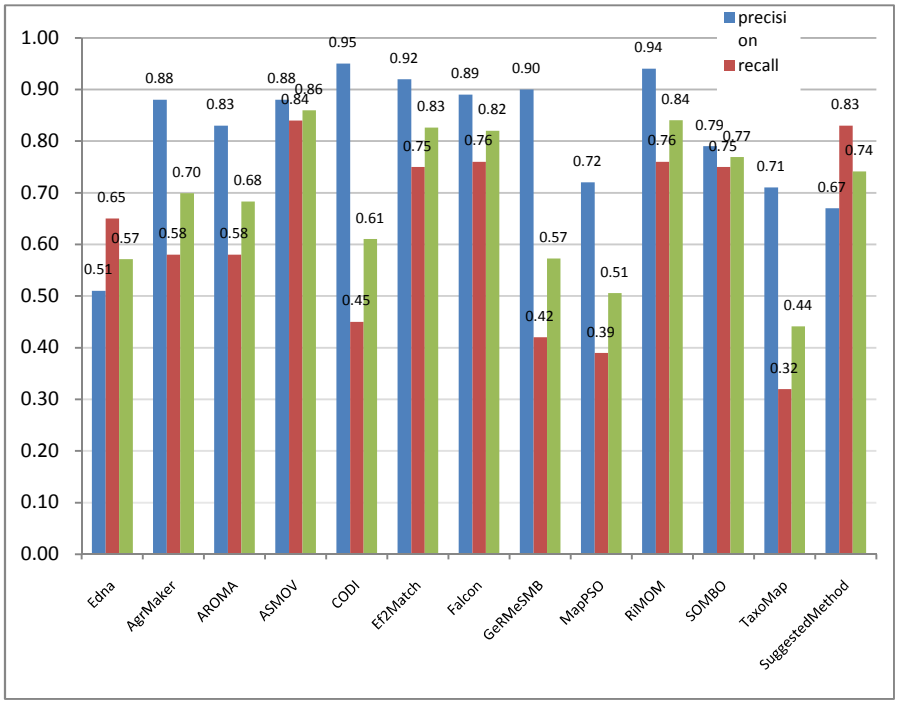
جدول ٤. نتایج ارزیابی

Systems	Measures	DoI	Systems	Measures	DoI
Edna	precision	+31.37	GeRMeSMB	precision	-25.56
	recall	+27.69		recall	+97.62
	fmeasure	+29.73		fmeasure	+29.46
AgrMaker	precision	-23.86	MapPSO	precision	-6.94
	recall	+43.10		recall	+112.82
	fmeasure	+6.05		fmeasure	-46.55
AROMA	precision	-19.28	RiMOM	precision	-28.72
	recall	+43.10		recall	+9.21
	fmeasure	+8.59		fmeasure	-11.78
ASMOV	precision	-23.86	SOMBO	precision	-15.19
	recall	-1.19		recall	+10.67
	fmeasure	-13.74		fmeasure	-3.67
CODI	precision	-29.47	TaxoMap	precision	-5.63
	recall	+84.44		recall	+159.38
	fmeasure	+21.41		fmeasure	+68.07
Ef2Match	precision	-27.17	Falcon	Precision	-24.79
	recall	+10.67		Recall	+9.21
	fmeasure	-10.27		Fmeasure	-9.56

٦. نتیجه‌گیری

با توجه به افزایش روزافزون تولید هستی‌شناسی‌ها در حوزه‌های مختلف و کاربرد مهم آنها در ذخیره‌سازی دانش بشری، بهبود و توسعه روش‌های بهره‌برداری از این ابزار بسیار مورد بحث و بررسی است. یکی از مباحث مهم تفاوت بین هستی‌شناسی‌هاست که وجود این تفاوت‌ها باعث بروز مشکلاتی در برقراری ارتباط بین هستی‌شناسی‌ها و در نتیجه، استفاده از آنها می‌شود. همین مسأله باعث ایجاد یک شاخه جدید در مباحث مربوط به هستی‌شناسی به نام

انطباق هستی‌شناسی‌ها شده است. در واقع، هدف از انطباق هستی‌شناسی، کشف تشابهات بین هستی‌شناسی‌هاست و نتیجه این فرآیند تعیین مشخصات شباهت‌های موجود بین هستی‌شناسی‌هاست. این مشخصات تولیدشده می‌تواند به عنوان ورودی برای فرآیندهای نگاشت یا ترکیب به‌کار رود. در این پژوهش، بررسی‌های کاملی بر روی کارهایی که تاکنون در این زمینه انجام گرفته است، انجام شد و در نهایت، تمرکز اصلی این پژوهش بر روش‌هایی که از یادگیری ماشین بهره جسته‌اند، قرار گرفت. با بررسی دقیق این روش‌ها به این نتیجه رسیدیم که تمام روش‌ها در نوع و تعداد ویژگی‌های مورد استفاده برای آموزش ماشین دارای نقصان هستند. در واقع، این روش‌ها ویژگی‌های مورد استفاده برای آموزش ماشین را از داخل خود هستی‌شناسی استخراج می‌نمایند که این منبع بسیار محدود است. با توجه به وجود منابع عظیم اطلاعات دیجیتالی در وب، در این پژوهش از این منابع استفاده گردید و ویژگی‌های مورد نظر با رعایت برخی معیارها از پیکره‌های متنی استخراج‌شده از وب، ایجاد گردید.



شکل ۷. مقایسه نتایج نظام پیشنهادی (SuggestedMethod) با نظام‌های موجود (OAIE 2010)

روش پیشنهادی بر استفاده از یادگیری ماشین با روش نایوبیز متمرکز است. در این روش، هر مفهوم به عنوان یک ماشین در نظر گرفته می‌شود و با استفاده از یک مجموعه آموزشی[1] (مفاهیمی که فرزند این مفهوم هستند) آموزش داده می‌شود و سپس، با استفاده از ماشین آموزش داده‌شده و فرمول شباهت جاکارد میزان شباهت دو مفهوم و در نتیجه انطباق یا عدم انطباق آنها مشخص می‌شود. با توجه به اینکه این پژوهش به دنبال یافتن انطباق بر اساس روابط معنایی بین مفاهیم موجود در هستی‌شناسی‌های مختلف است، به نظر می‌رسد نتایج به‌دست‌آمده امیدوارکننده است. به عنوان مثال، انطباق صحیح بین دو مفهوم (Reference و Entry) یا (Institution و School) که هیچ شباهت لغوی و ظاهری با هم ندارند، ولی در حوزه هستی‌شناسی‌های مورد بحث هم معنی هستند، یافت شده است.

دو موضوع بر نتایج آزمایش‌های انجام‌شده در این پژوهش اثر منفی داشته است. نخست، آزمایش‌های این پژوهش بر روی هستی‌شناسی‌هایی با اندازه کوچک (تعداد مفاهیم محدود) انجام‌شده است و این بدان معنی است که تعداد نمونه‌های استفاده‌شده در فرآیند یادگیری نیز محدود بوده و این امر بر میزان دقت نتایج تأثیر منفی داشته است. دوم، محک‌های[2] ارائه‌شده در OAIE توجه زیادی به روابط مفهومی موجود بین هستی‌شناسی‌ها نداشته و این دلیلی است که راه‌حل پیشنهادی این مقاله بهترین جایگاه را در بین نظام‌های ارائه‌شده در کنفرانس سالانه OAIE نداشته است (شکل ۷). این نظام‌ها در مقالات (Zhang et al. 2010)، (Hu et al. 2010)، (Noessner and Niepert 2010)، (Cruz, Stroe ,and Caci 2010)،(Xu, Wang, and Cheng 2010) (Shvaiko, Euzenat and Giunchiglia)، (Jean-Mary, Shironoshita, and Kabuka 2010) 2010، (Hamdi, Safar ,and Reynaud 2010)، ارائه شده است.

از نتایج به‌دست‌آمده می‌توان استنباط کرد که در صورت رفع این دو موضوع روش پیشنهادی دارای کارایی بالایی است. روش پیشنهادی این مقاله در حوزه‌های مختلفی قابل توسعه است که در ادامه به برخی از این حوزه‌ها اشاره می‌گردد. در این مقاله، روش انتخاب فرزندان به صورت مستقیم پیاده‌سازی شده است. به این معنی که برای آموزش ماشین (به ازای هر مفهوم) طبق شکل ۵ (ب) فقط فرزندان مستقیم انتخاب شده‌اند. پیشنهاد می‌گردد به منظور توسعه نظام در آینده طبق شکل۵ (الف) کل زیردرخت مربوط به یک مفهوم به عنوان فرزندان در نظر گرفته شود و برای آموزش ماشین استفاده گردد. پیشنهاد دیگر در مورد روش استفاده شده برای یادگیری ماشین است. در پیاده‌سازی انجام‌شده برای روش نایوبیز مقدار ویژگی‌ها با در نظر گرفتن یک مقدار آستانه به مقادیر گسسته ۰ یا ۱ تبدیل شده است و از روش نایوبیز

1. training set 2. benchmarks

گسسته استفاده‌شده است که به منظور توسعه می‌توان مقادیر را دست نخورده نگه داشت و از روش نایوبیز پیوسته به منظور آموزش ماشین استفاده نمود. در این مقاله، فقط از یک روش یادگیری (نایوبیز) استفاده‌شده است که می‌توان در کارهای آتی از روش‌های دیگر یادگیری ماشین استفاده کرد و با طراحی یک یادگیر چندگانه[1] مناسب، نتایج یادگیرها را ترکیب نمود.

در روش پیشنهادی فقط از یک روش مبتنی بر شباهت‌های مفهومی به منظور انطباق استفاده شده است که این مورد نیز قابل توسعه است. بدین شکل که می‌توان روش‌های مبتنی بر ساختار و شباهت‌های لغوی و روش‌های معنایی دیگر مانند استفاده از یک منبع اطلاعاتی استاندارد (مانند ورد نت[2]) را با اعمال وزن‌های مناسب به هر روش، با روش پیشنهادی تلفیق نمود و دقت نتایج به دست آمده را بهبود داد.

۷. منابع

Bach, Thanh-Le, Rose Dieng-Kuntz, and Fabien Gandon. 2004. On ontology matching problems (for building a corporate semantic web in a multi-communicates organization). In *Proceedings of the 6th International Conference on Enterprise Information Systems(ICEIS). Porto, Portugal* 236-243.

Benchmark test library of OAIE2010. http://oaei.ontologymatching.org/2010/benchmarks (accessed 25 Feb. 2010).

Bock, Jürgen, Peng Liu and Jan Hettenhausen. 2010. MapPSO results for OAEI 2010. In *Proceedings of the 9th International Semantic Web Conference(ISWC). Shanghai, China* 180-186.

Cohen, William W., Pradeep D. Ravikumar, and Stephen E. Fienberg. 2003. A comparison of String distance metrics for name-matching tasks. In *Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03). Acapulco, Mexico* 73-78.

Cruz, Isabel , Cosmin Stroe, Michele Caci.2010. Using AgreementMaker to Align Ontologies for OAEI-2010. In *Proceedings of the 9th International Semantic Web Conference (ISWC). Shanghai, China* 118-125.

Cunningham, Hamish and Yorick Wilks.1996. GATE-a General Architecture for Text Engineering. In *Proceedings of the 16th Conference on Computational Linguistics(COLING96). Copenhagen, Denmark* 1057-1060.

Dieng, Rose and Stefan Hug. 1998. Comparison of personal ontologies represented through conceptual graphs. In *Proceedings of the 13th ECAI. Brighton, UK* 341-345.

Doan, AnHai, Jayant Madhavan, Robin Dhamankar, Pedro Domingos, and Alon Halevy. 2003. Learning to Match Ontologies on the Semantic Web. *The VLDB Journal* 12 (4): 303-319.

Ehrig, Marc , Jerome Euzenat, Raúl García Castro. 2004. Specification of a benchmarking methodology for alignment techniques. Technical report.

Euzenat, Jérôme, Marc Ehrig, and Raúl García Castro. 2005. Towards a methodology for evaluating alignment and matching algorithms.France: http://oaei.inrialpes.fr (accessed 20 Feb. 2010).

Ghamrawy, Sally M.2009. Dynamic ontology mapping for communication in distributed multi-agent intelligent system. In *Proceedings of the International Conference on Networking and Media Convergence(ICNM 2009), Cairo,Egypt* 103-108.

Hamdi, Fayçal , Brigitte Safar and Chantal Reynaud. 2010. TaxoMap alignment and refinment modules: Result for OAEI 2010. In *Proceedings of the 9th International Semantic Web Conference (ISWC). Shanghai, China* 212-219.

1. Metalearner 2. WordNet

Hu, Wei , Jianfeng Chen, Gong Cheng, and Yuzhong Qu. 2010. Object Coref & falcon-AO: Results for OAEI 2010. In *Proceedings of the 9th International Semantic Web Conference (ISWC). Shanghai, China* 158-165.

Jean-Mary, Yves R., E. Patrick Shironoshita, Mansur R. Kabuka. 2010. ASMOV: results for OAEI 2010. In *Proceedings of the 9th International Semantic Web Conference (ISWC). Shanghai, China* 126-133.

Lin, Feiyu. 2007. State of the Art Automatic Ontology Matching. Jönköping University. School of Engineering. Research Report. Jönköping, Sweden.

Maedche, Alexander, Steffen Staab. 2002. Measuring Similarity between Ontologies.

Makhoul, John, Francis Kubala, Richard Schwartz, and Ralph Weischedel.1999. Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop. Herndon.*

Natarajan, Pradeep. 2005. A machine learning approach to ontology matching.Technicalreport :http://www-cf.usc.edu/~pnataraj/CS548.rtf (accessed 18 Feb. 2010).

Nilsson, Nils J. 1990. *The Mathematical Foundations of Learning Machines*. Morgan Kaufmann Publishers Inc.USA.

Noessner, Jan and Mathias Niepert . 2010. CODI: combinatorial optimization for data Integration –result for the OAEI 2010. *The 9th International semantic web conference*.

Shannon, Claude E.1948. A mathematical theory of communication. *bell system technical journal* (27): 379–423, 623-656.

Shvaiko, Pavel , Jérôme Euzenat and Fausto Giunchiglia. 2010. Ontology Matching (OM 2010). *The 9th International Semantic Web Conference*.

Su, Xiaomeng, and Jon Atle Gulla. 2006. An information retrieval approach to ontology mapping, Data & Knowledge Engineering. Application of natural language to information systems (NLDB04): 47-69.

Xu, Peigang , Yadong Wang, and Liang Cheng. 2010. Alignment results of SOBOM for OAIE 2010. *The 9th International Semantic Web Conference*.

Zhang, Xiao , Qian Zhong, Juanzi Li, and Jie Tang. 2010. RiMOM results for OAEI 2010. *The 9th International Semantic Web Conference*.

A New Method for Ontology Matching by Using Textual Corpus

Besat Kassaie[1]
Ms of Software Engineer

Maseud Rahgozar[2]
Ph.D. of Computer Sciences

Alireza Vazifedoost*
Ph.D. Student in Software Engineering

Iranian Journal of Information Processing & Management

Iranian Research Institute Iranian
For Science and Technology
ISSN 2251-8223
eISSN 2251-8231
Indexed in LISA, SCOPUS & ISC
Vol.28 | No.3 | pp: 807-827
Spring 2013

Abstract: The aim of ontology matching is to find similarities or matches between concepts of different ontologies. There are many new applications which need a sort of ontology matching. Some examples comprises of semantic web applications, multi agent systems, applications mash up and so on. One may be interested in either finding lexical similarity or semantic similarity, but in the both cases, the result of such a matching process can be useful for relating distinct ontologies. Leveraging ontology matching system enables us to reuse existing ontologies in new applications and save costs by eliminating the need for developing new ontologies. Among current algorithms proposed for matching onologies applying machine learning techniques is a promising one. However, there are some problems regarding the results of these methods which are mainly due to poor features used in learning process.

In this paper we propose a new method in which a text corpus is used as the source of knowledge in conjunction with a machine learning method to find matching between two ontologies .The main objective in this new method is to find similarity of two concepts based on similarity of their instances. We show how contextual knowledge hidden in domain specific documents can help us to boost the machine learning methods by providing enough features. Also we show how taking benefit from this knowledge transcends the current approaches merely detect lexical similarity by either recognizing semantic similarity of concepts.

Keywords: ontology matching, machine learning, text corpus, naïve bayes method, semantic similarity

1. besat_k@yahoo.com 2. rahgozar@ut.ac.ir
*Corresponding author: vazifehdst@ut.ac.ir