استخراج، ادغام و تقویت موجودیت برای حفظ و نگهداری محتوای وب اجتماعی

صورت‌جلسه دومین کارگاه بین‌المللی آرشیوهای رقمی معنایی (SAD ۲۰۱۲)

نویسندگان: استفان دیتز، دیانا می‌نارد، الن دمی‌دوا، توماس ریس، ویم پیترز، کاترین دوکا، یانیس استاوراکاس | مترجم: اعظم‌السادات حسینی

■ چکیده

با افزایش سرعت تحولات محتوای وب، به‌ویژه رسانه اجتماعی، حفظ و نگهداری وب و تحولات آن در طول زمان به چالش بسیار مهمی تبدیل شده است. تحلیل معنایی محتوای وب، مناسب دیدگاه موجودیت است و برای تنظیم و سازمان‌دهی منابع وب بر اساس اشیاء اطلاعاتی مرتبط با آن‌ها به‌کار می‌آید. بنابراین چالش اساسی استخراج، شناسایی و برقراری ارتباط میان تعداد کثیری از منابع گوناگون و نامتجانس وب است که در آن‌ها ماهیت و کیفیت محتوا بسیار متنوع است. بااینکه در این فرایند از منبع غنی ابزارهای استخراج اطلاعات استفاده‌شده است، یکپارچه‌سازی، ادغام و تقویت داده‌هایی که به‌طور خودکار استخراج‌شده نیز برای تضمین کیفیت و عدم ابهام داده‌های تولیدی، لازم است. در این مقاله رویکردی بر اساس چرخه تعاملی بهره‌برداری از داده‌های وب در موارد زیر ارائه می‌شود:

۱- آرشیو هدفمند اشیاء وب؛ ۲- استخراج و شناسایی موجودیت؛ ۳- برقراری ارتباط میان موجودیت‌ها.

هدف طولانی‌مدت این کار، حفظ و نگهداری محتوای وب در طول زمان و امکان رهیابی و تحلیل آن بر اساس داده‌های دارای ساختار مناسب آر. دی. اف. درباره موجودیت‌هاست.

کلیدواژه‌ها

استخراج دانش، داده مرتبط/ پیوندی، ادغام و تقویت داده‌ها، غنی‌سازی داده‌ها، آرشیو کردن وب، شناسایی موجودیت.

استخراج، ادغام و تقویت موجودیت برای حفظ و نگهداری محتوای وب اجتماعی

صورت‌جلسه دومین کارگاه بین‌المللی آرشیوهای رقمی معنایی (SAD ۲۰۱۲)[۱]

استفان دیتز، دیانا می‌نارد، الن دمی‌دوا، توماس ریس، ویم پیترز، کاترین دوکا، یانیس استاوراکاس

مترجم: اعظم‌السادات حسینی[۲]

۱- مقدمه

با توجه به‌سرعت فزاینده تحول محتوای وب، آرشیو مناسب و حفظ و نگهداری وب یکی از ضرورت‌های فرهنگی به‌شمار می‌آید. در کنار چالش‌های معمولی و رایج در حفظ و نگهداری رقمی مثل زوال حامل یا واسط، کهنگی و از رواج افتادن فنّاوری، اعتبارسنجی و مسئله تمامیت و یکپارچگی (داده)، برای حفظ و نگهداری وب باید مسائلی مثل اندازه واقعی و رشد فزاینده محتوای وب را نیز مدنظر قرار داد. این مورد به‌ویژه در خصوص محتوای تولیدشده توسط کاربر و رسانه اجتماعی که شاخصه آن تنوع فراوان، کیفیت متغیر و عدم تجانس و گوناگونی است، بیشتر به‌کار می‌آید. سازمان‌های آرشیوی به‌جای پیروی از راهبرد جمع‌آوری همه موارد، می‌کوشند تا آرشیوهای متمرکزی را با یک موضوع خاص به‌وجود آورند که تنوع و گوناگونی اطلاعات موردعلاقه عمومی را نشان دهد. بنابراین آرشیوهای متمرکز بیشتر درباره موجودیت‌های معرف یک موضوع یا علاقه عمومی مثل اشخاص، سازمان‌ها و مکان‌ها هستند. بنابراین استخراج موجودیت‌ها[۳] از میان محتوای آرشیوشده وب در رسانه‌های خاص اجتماعی، ازجمله چالش‌های اساسی است که برای انجام جستجو و رهیابی معنایی در آرشیوهای وب و ارزیابی ارتباط دسته اشیاء وب برای یک خزش[۴] متمرکز مطرح است.

بااینکه برای استخراج اطلاعات از متون رسمی ابزارهای زیادی وجود دارد، رسانه اجتماعی چالش‌های خاصی را برای کسب دانش ایجاد می‌کند. این چالش‌ها در بخش سوم

۱.این مقاله ترجمه‌ای است از:
Entity extraction and consolidation for social web content preservation. (Proceedings of the 2nd International Workshop on Semantic Digital Archives- SDA (2010

۲.مترجم اداره‌کل اطلاع‌رسانی و ارتباطات سازمان اسناد و کتابخانه ملی ایران

3. Entities

4. Crawl

به‌طور ویژه و با جزئیات مطرح‌شده است. این امر نیازمند مجموعه‌ای از راهبردها و فنون خاص برای تقویت، ادغام، غنی‌سازی، رفع ابهام و پیونددادن داده‌های استخراجی است. با بهره‌گیری از دانش موجود مثل داده باز/ آزاد مرتبط[1] [۱] می‌توان اطلاعات منسوخ و تجزیه‌شده[2] را اصلاح، جبران و رفع ابهام کرد. بااینکه روش‌های ادغام و تقویت‌داده[3] به‌طور سنتی از شناسایی موجودیت‌های نامدار[4] مستقل است، تمامیت و یکپارچگی هماهنگ آن‌ها با جریان‌های کاری هم پیوند[5] بسیار اهمیت دارد. به‌این‌ترتیب، منبع داده‌ای که به‌طور خودکار استخراج‌شده غنی‌تر خواهد شد. هم‌چنین ضرورت افزایش تنوع استخراج دانش در دسترس عموم و مناسب کاربر نهایی و ابزارهای شناسایی موجودیت‌های نامدار مثل: DBpedia Spotlight[6]، GATE[7]، Open Calais[8]، Zemanta[9] بر اهمیت این امر می‌افزاید.

در این مقاله، رویکرد یکپارچه استخراج و ادغام دانش ساختاری درباره موجودیت‌ها برای محتوای آرشیوشده وب را معرفی می‌کنیم. این دانش در آینده برای تسهیل جستجوی معنایی آرشیوهای وب و راهنمایی بیشتر خزش به‌کار می‌رود. این کار در پروژه یکپارچه‌سازی آرکومم[10] که تحت حمایت مالی جامعه اروپاست، توسعه یافت. توجه داشته باشید، بااینکه جنبه‌های موقت مربوط به تحول دانش و اصطلاحات برای حفظ و نگهداری وب ضروری است، اما همچنان در دست بررسی است [۲۴] و در این مقاله ذکری از این کار به میان نیامده است.

۲- کارهای مرتبط

شناسایی موجودیت یکی از کارهای اصلی در استخراج اطلاعات است و شناسایی موجودیت‌های نامدار و استخراج اصطلاحات را نیز در برمی‌گیرد. شناسایی موجودیت ممکن است شامل سامانه‌های قانون‌محور [۱۳] یا روش‌های فراگیری ماشینی [۱۴] باشد. استخراج اصطلاح مستلزم شناسایی و پالایش (فیلتر) اصطلاحات منتخب به‌منظور شناسایی اصطلاحات یا موجودیت‌های مربوط به این حوزه است.

هدف اصلی در شناسایی اصطلاحات خودکار، تعیین اصطلاح‌بودن یک واژه یا چند واژه است که حوزه و هدف نهایی را مشخص می‌کند. بسیاری از روش‌های استخراج اصطلاح، ترکیبی از پالایش زبان‌شناسی (مثل توالی احتمالی برچسب‌های مقولات دستوری) و اقدامات آماری (مثل:tf.idf) را به‌کار می‌گیرند [۱۵] [۱۶] تا اهمیت هر اصطلاح منتخب برای هر سند در مجموعه مشخص شود [۲۳].

تقویت و ادغام داده باید محدوده‌های متفاوتی را از غنی‌سازی[11] گرفته تا شفاف‌سازی موجودیت/ هویت برای رفع ابهام تا خوشه‌بندی[12] و ایجاد ارتباط به‌منظور ادغام و تقویت داده‌های متفاوت و ناهم‌خوان، پوشش دهد. به‌علاوه، پیش‌بینی و کشف پیوند برای انجام

1. Linked Open Data
2. Degraded
3.Consolidation
4. Named Entity Recognition (NER)
5. Unified Workflows
6. http://spotlight.dbpedia.org
7. http://gate.ac.uk/
8. http://www.openclais.com/
9. http://www.zemanta.com/
10. Arcomem; http://www.arcomem.eu
11. Enrichment
12. Clustering

خوشه‌بندی و برقراری ارتباط منابع غنی‌شده داده، لازم است. روش‌های مختلفی برای شفاف‌سازی موجودیت پیشنهاد شده است که در آن‌ها از روابط میان موجودیت‌ها [۷]، نگارش زنجیره شباهت‌ها[1] [۶] و دگرگونی‌ها و تغییرات [۹] استفاده شده است. بازنگری بیشتر کارهای مهم در این زمینه در [۸] یافت می‌شود. برخلاف روش‌های برقراری ارتباط میان موجودیت‌ها در این مقاله، خوشه‌بندی متون مستندات از بُردارهای شاخص برای نمایش و ارائه مستندات بر اساس اصطلاحات موجود استفاده کرده است [۱۰] [۱۱] [۱۲].

الگوریتم‌های خوشه‌بندی، شباهت میان مستندات را سنجیده و هر مستند را بر اساس این شباهت به خوشه مناسب تخصیص می‌دهد. هم‌چنین رویکردهای بُردارمحور برای جانمایی هستی‌شناسی (آنتولوژی)[2] و دسته‌های داده مورداِستفاده قرار می‌گیرد [۲] [۳]. خوشه‌بندی و برقراری ارتباط میان موجودیت‌ها برخلاف خوشه‌بندی متن از دانش زمینه‌ای و پیشینه دسته‌های داده مرتبط برای برقراری ارتباط میان موجودیت‌هایی که قبلاً استخراج‌شده‌اند، بهره می‌گیرد. بنابراین، کشف پیوند، موضوع اساسی دیگری است که باید موردِتوجه قرار گیرد. خلاصه‌سازی ترسیم‌ها، پیوندهایی را در ترسیم‌های حاشیه‌نویسی آر. دی. اف.[5] پیش‌بینی می‌کند. تحقیق دقیق درباره روش‌های پیش‌بینی پیوند در شبکه‌های پیچیده (ترکیبی) و اجتماعی در [۴] و [۵] ارائه شده است.

۳- چالش‌ها و رویکرد کلی

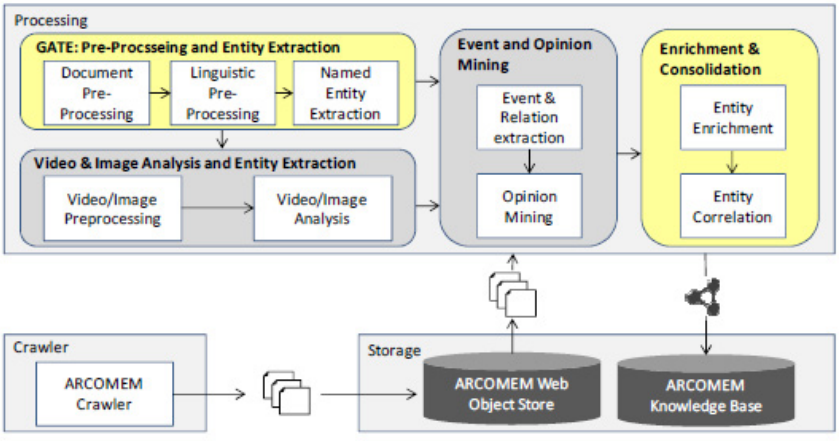
ARCOMEM از رویکرد موردی پیروی می‌کند که اساس آن بر ایجاد آرشیوهای متمرکز وب است. ما از مخزن مستند محتوای وب خزش‌شده و پایگاه دانش ساختاری آر. دی. اف. را جایگذاری کرده‌ایم که حاوی فراداده موجودیت‌های کشف و شناسایی‌شده در محتوای آرشیوی است. آرشیوداران می‌توانند مشخصات خزش (اساساً شامل دسته‌های منتخب از موضوعات و موجودیت‌های مربوط) را تعیین یا اصلاح کنند. خزشگر[3] هوشمند قادر است درباره اهداف خزش بیاموزد و راهبرد خزش را در شرایط شتاب بهسازی کند. این امر به‌ویژه برای خزش‌های طولانی‌مدت با موضوعات وسیع‌تر مثل انتخابات یا بحران مالی، مهم است که در آن‌ها موجودیت‌ها به‌دفعات تغییر می‌کنند و درنتیجه هماهنگ‌سازی منظم ویژگی‌های خزش لازم است. برنامه‌های کاربردی کاربر نهایی با بهره‌گیری از فراداده‌ای که دربارهٔ موجودیت‌ها و موضوعات به‌صورت خودکار استخراج‌شده، امکان جستجو و مرور آرشیوها را برای کاربران فراهم می‌کند.

استخراج صحیح و کارآمد موجودیت‌ها از محتوای آرشیوی وب برای پالایش راهبرد خزش و رهیابی[4] و هدایت آرشیو وب بسیار اهمیت دارد. رسانه‌های اجتماعی به‌دلیل ماهیت

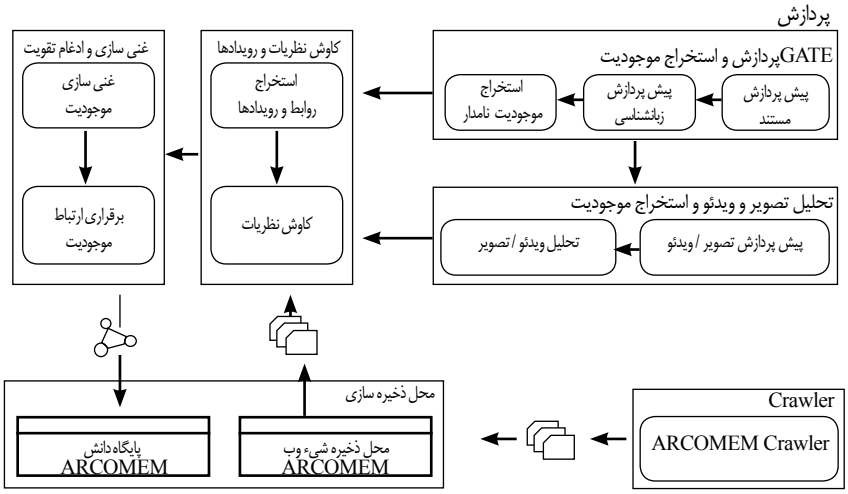
1. String similarity metrics
2. Ontology
3. Crawler
4. Navigation
5. RDF

زوال‌پذیر متن، به‌ویژه در خصوص پربسامد (توئیت)، با چالش‌های متعددی برای ابزارهای تحلیل زبانی مواجه‌اند. در یک تحقیق، Stanford NER tagger (برچسب‌زن نام موجودیت‌های استن‌فورد) هنگام کاربرد برای مجموعه توئیت‌ها[1] از F1 %۹۰/۸ به %۴۵/۸۸ کاهش یافت [۱۷].

در این تحقیق [۱۹] هم‌چنین برخی از مشکلات کاربرد برچسب‌زنی سنتی POS، chunking و روش‌های شناسایی موجودیت‌های نامدار در خصوص توئیت‌ها ارائه‌شده است، درحالی‌که ابزارهای شناسایی زبان نوعاً برای جملات کوتاه کارآمد نیستند. مشکلات در اثر تلفظ، املاء و دستور زبان نادرست، واژه‌های ساختگی، برچسب‌های درهم‌وبرهم و تکراری (hashtag)، @ و emoticonها، بزرگ‌سازی حروف و املای واژه‌ها به شیوهٔ غیرمعمول (تکرار حروف در یک واژه برای تأکید یا گفتار متن). ازآنجایی‌که نمونه نمایی[2]، برچسب‌زنی POS و مقابله با فهرست نام‌های جغرافیایی[3] کلیدهای شناسایی موجودیت‌های نامدار هستند، حل مشکلات زیر از اهمیت زیادی برخوردار است: گزینش روش‌هایی مثل انطباق نمونه نماها[4]، کاربرد فنون عادی‌سازی SMS، بازآموزی شناسایی‌کنندگان زبان[5]، کاربرد تقابل غیر موردی[6] برای موارد خاص، به‌کارگیری فنون سطحی به‌جای تقطیع و تجزیه کامل[7] و کاربرد اشکال انعطاف‌پذیرتر مقابله[8]. استخراج و غنی‌سازی موجودیت تحت پوشش دسته‌ای از اجزاء اختصاصی است که ضمیمه زنجیره پردازش اختصاصی می‌باشد (شکل۱) که شناسایی موجودیت‌های نامدار و «ادغام تقویت» را (غنی‌سازی، خوشه‌بندی، رفع ابهام) به‌عنوان بخشی از جریان کاری هماهنگ، مدیریت و جابه‌جا می‌کند.



1. Tweet
2. Tokenisation
3. Gazetter
4. Tokeniser
5. Language identifiers
6. Case – in sensitive matching
7. Full parsing
8. Matching



شکل ۱- استخراج موجودیت و زنجیره پردازش ادغام و تقویت

محل ذخیره‌سازی ACROMEM که شامل محل ذخیره شیء و پایگاه دانش است دو نوع داده را جابه‌جا و مدیریت می‌کند: الف) داده میانی، به‌شکل اشیاء وب که نشان‌دهندهٔ محتوای اصلی جمع‌آوری‌شده با خزشگر است؛ ب) داده نیمه‌ساختاری به‌شکل سه‌گانه‌های آر. دی. اف.[1] (زیرنویس‌های اشیاء وب). محل ذخیره بر اساس راه‌حل توزیعی شکل‌گرفته و الگوی[2] MapReduce [۹] را با پایگاه‌داده (دادگان) No SQL ترکیب می‌کند و بر اساس H Base[3] قابل‌درک و تشخیص است (دراین‌باره به منبع [۲۵] نیز مراجعه کنید). الگوی داده ARCOMEM[4] طرح RDF را برای نشان‌دادن نیازهای اطلاعاتی برای کسب، خزش و نگهداری دانش ارائه می‌کند، (برای جزئیات بیشتر به منبع [۲۰] مراجعه کنید). در الگوی ACROMEM، «موجودیت» هم موجودیت‌های نامدار سنتی و هم اصطلاحات چند واژه‌ای و منفرد را در برمی‌گیرد: تشخیص هر دو با استفاده از ابزارهای GATE صورت می‌گیرد. GATE از میان سایر ابزارهای NLP انتخاب‌شده است اولاً به‌دلیل پوشش و شمول آن، گسترش‌پذیری و انعطاف‌پذیری آن و ثانیاً به‌دلیل دارا بودن دامنه وسیعی از اجزاء NPL که در خصوص درخواست‌های پروژه به‌آسانی قابل‌اصطلاح است برخلاف ابزارهایی مثل Open Calais و DBPedia Spotlight که حوزه و گستره محدودتری دارند. بااینکه داده استخراجی در فرآیند استخراج رده‌بندی و برچسب‌زنی می‌شود، اما ۱) نامتجانس و ناهم‌خوان یعنی دارای پیوندهای نامناسب، ۲) مبهم و ۳) ارائه‌کننده اطلاعات محدود است. علت این امر، استخراج داده از طریق اجزاء مختلف و در چرخه‌های پردازش مستقل است زیرا این ابزارها امکان

1. http://www.w3.org/RDF/
2. Paradigm
3. Apache Foundation, The Apache H Base Project: http://hbase.apache.org/
4. http://www.gate.ac.uk./ns/ontologies/acromem-datamodel.rdf

اجرای مرجع مشترک برای موجودیت‌های غیر هم‌زمان در مستندات چندگانه را ندارند. مثلاً در یک چرخه خاص، اجزاء تحلیل متن ممکن است موجودیتی را از اصطلاح «ایرلند» کشف کند، درحالی‌که در چرخه‌های بعدی موجودیت‌های برگرفته از اصطلاح «جمهوری ایرلند» یا اصطلاح آلمانی «ایرلند» ممکن است همراه با موجودیت «دوبلین» استخراج شود. همه این‌ها به‌عنوان موجودیت‌های «جانمایی» رده‌بندی و در محل ذخیره‌سازی داده به‌عنوان موجودیت‌های جداگانه بر اساس الگوی داده ذخیره می‌شوند؛ بنابراین غنی‌سازی و ادغام و تقویت (شکل ۱) سه هدف را دنبال می‌کند: الف) غنی‌سازی موجودیت‌ها با دانش عمومی موجود و مربوط؛ ب) رفع ابهام؛ پ) شناسایی هم‌بستگی داده مثل آنچه در بالا نشان داده شد. دستیابی به این اهداف در سایه بازنمایی/ نگاشت موجودیت‌های منفرد در مفاهیم داخل دادگان‌های مرجع (غنی‌سازی) و بهره‌برداری از ترسیم‌های مرتبط برای کشف همبستگی‌ها انجام‌پذیر است؛ بنابراین، دادهٔ در دسترس عموم را از ابرداده باز/ آزاد مرتبط[1] برداشت می‌کنیم. این ابر مقدار زیادی داده دارای دامنه خاص و داده مستقل از دامنه‌ها و حوزه‌ها را ارائه می‌کند (انتشار ۳۱ بیلیون تریپل[2] یعنی بیانیه‌های آر. دی. اف.[3]).

۴- اجرا و به‌کارگیری

برای شناسایی موجودیت از نسخه اصلاح‌شده ANNIE [۱۸] استفاده می‌کنیم تا اشاره‌های مربوط به شخص، مکان، سازمان، تاریخ، زمان، پول و درصد را پیدا کنیم. در این راه از زیرگروه‌های سازمان مثل گروه و حزب سیاسی نیز استفاده کردیم و اصلاحات متعددی را برای برطرف‌کردن مشکلات خاص رسانه اجتماعی مثل انگلیسی نادرست انجام دادیم (برای جزئیات بیشتر به منبع [۲۱] مراجعه کنید). چارچوب استخراج موجودیت به اجزاء زیر تقسیم می‌شود (اجزاء GATE در شکل ۱ ارائه شده است) که به‌صورت متوالی در مجموعه مستندات اجرا می‌شود:

پیش‌پردازش مستند (تحلیل قطع و قالب مستند، شناسایی و کشف محتوا)

پیش‌پردازش زبان‌شناسی (شناسایی زبان، نمونه‌نمایی، برچسب‌زنی POS و...)

استخراج موجودیت نامدار: استخراج اصطلاح (ایجاد فهرست رتبه‌بندی اصطلاحات و آستانه‌ها) و شناسایی موجودیت‌های نامدار (فهرست نام‌های جغرافیایی، دستور زبان اصلی و هم‌مرجع‌ها/ مرجع مشترک)

ما برای استخراج اصطلاح، از نسخه سازگارشده Term Raider[4] استفاده کردیم. این نسخه عبارات اسمی[5] را به‌عنوان اصطلاحات منتخب در نظر گرفته است (همان‌طورکه در پیش‌پردازش زبان‌شناسی مشخص شد) و آن‌ها را با نظمی بر اساس سه کارکرد مختلف

1. Linked Open Data Cloud
2. Triples
3. http://lod-cloud.net/state
4. http://gate.ac.uk/projects/arcomem/TermRaider.htmll
5. NPS

امتیازی رتبه‌بندی کرده است: ۱- Tf.idf اصلی، ۲- Tf.idf افزایشی که امتیاز tf.idf همه زیرنام‌های[1] اصطلاح منتخب را مدنظر قرار دهد، ۳- امتیاز Kyoto بر اساس [۲۲] که تعداد زیرنام‌های اصطلاح منتخب موجود در متن را موردِتوجه قرار می‌دهد. این کارکردها برای عرضه و نمایش مقادیر میان ۰ تا ۱۰۰ قاعده‌مند شده‌اند. به‌منظور جلوگیری از تکثیر و تکرار، اصطلاح مناسب و منتخب اگر با موجودیت‌های نامدار فعلی هماهنگ باشد یا ازجمله آن‌ها باشد موجودیت در نظر گرفته نمی‌شود. هم‌چنین، ما امتیاز آستانه را مقرر می‌کنیم و اصطلاح مناسب بالاتر از آن را ارزشمند می‌دانیم. این آستانه، مؤلفه‌ای است که به‌صورت دستی در هر زمانی قابلِ‌تغییر است— درحال‌حاضر ارزش افزایشی معادل ۴۵ دارد یعنی فقط اصطلاحاتی با امتیاز ۴۵ یا بیشتر در فرآیندهای بعدی موردِاستفاده قرار می‌گیرند.

استخراج موجودیت، داده آر. دی. اف. را به‌وجود می‌آورد که NES و اصطلاحات مطابق با الگوی داده ACROMEM را توصیف می‌کند. این اطلاعات به پایگاه دانش منتقل‌شده و مستقیماً توسط اجزاء و عوامل غنی‌سازی و ادغام و تقویت خلاصه و جذب می‌شود (شکل۱). آخرین مرحله الف) برچسب موجودیت و ب) نوع موجودیت را موردِاستفاده قرار می‌دهد تا داده استخراجی را گسترش دهد، رفع ابهام کند و هم‌بسته سازد. توجه داشته باشید که برچسب رویداد/ موجودیت ممکن است مستقیماً به برچسب یک شاخه[2] منحصربه‌فرد در دسته داده ساختاری مربوط باشد (چنانکه درباره موجودیت نوع شخص برچسب زده شد «Angela Markel») و درعین‌حال ممکن است به بیش از یک شاخه/ مفهوم مرتبط باشد چنانکه بیشتر رویدادهای دسته داده ما این‌گونه است. مثلاً رویدادی با برچسب «جین کلاودتریچت در نشست ECB نطق اصلی را ایراد می‌کند»[3] احتمالاً با پیوندهایی به مفاهیم نشان‌دهنده ECB و جین کلاوترچت غنی‌سازی می‌شود. مراحل زیر اساس رویکرد ما را تشکیل می‌دهد (در شکل ۱ نشان داده شده است):

مرحله ۱- غنی‌سازی موجودیت

مرحله اول- الف) ترجمه: ما زبان برچسب موجودیت را تعیین می‌کنیم و در صورت لزوم، آن را با استفاده از خدمات ترجمه برخط به انگلیسی برمی‌گردانیم.

مرحله اول- ب) غنی‌سازی: هم‌مرجع‌سازی (ایجاد مرجع مشترک) با موجودیت‌های درون دسته‌های داده مرجع.

مرحله ۲- هم‌بستگی (برقراری ارتباط) با موجودیت و خوشه‌بندی.

به‌منظور غنی‌سازی این موجودیت‌ها، جستارهایی بر روی پایگاه‌های دانش خارجی (بیرونی) انجام دادیم. رویکرد غنی‌سازی فعلی ما از DBPedia[4] و Freebase[5] به‌عنوان دسته‌های داده مرجع استفاده می‌کند، اگرچه پیش‌بینی می‌شود این رویکرد با دسته‌های داده بیشتر و

1. Hyponyms
2. Node
3. Jean Claude Trichet gives Keynote at ECB summit.
4. http://dbpedia.org/
5. http://www.freebase.com/

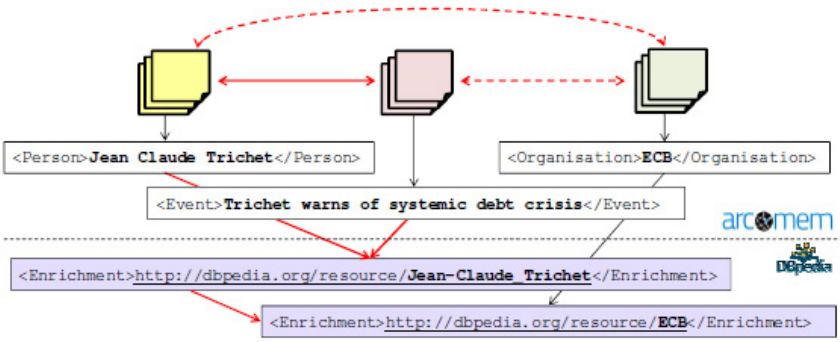
دارای حوزه خاص گسترش باید مثلاً انواع مربوط به رویداد خاص. DBpedia و Freebase به‌دلیل اندازه بزرگ، دستیابی به فنون رفع ابهام (که برای برچسب‌های متعدد چندزبانه در هر دو دسته داده برای داده‌های منفرد سودمند هستند) و سطح اتصال و ارتباط هر دو دسته داده (که بازیابی گنجینه اطلاعات مرتبط با اقلام خاص را ممکن می‌سازد) بسیار مناسب هستند. در مورد DBpedia ما از خدمات DBpedia Spotlight استفاده می‌کنیم که هماهنگی تقریبی با سطح اطمینان مناسب را در فاصله ۰ و ۱ میسر می‌سازد. در بخش ۶ ما به‌عنوان بخشی از ارزیابی‌مان، به‌طور تجربی سطح اطمینان ۰/۶ را انتخاب کردیم که بهترین سطح دقت و تعادل و فراخوانی را ارائه می‌کند. توجه داشته باشید که Spotlight، قابلیت‌های شناسایی موجودیت‌های نامدار را در تکمیل GATE عرضه می‌کند. بااین‌حال، همه این موارد فقط در مواردی کارآمد و مفید هستند که موجودیت‌ها/ رویدادها در شکل اتمی نباشند، چنانکه در خصوص رویدادهای متشکل از توصیفات متنی آزاد مطرح است.

Freebase حاوی حدود ۲۲ میلیون موجودیت و بیش از ۳۵۰ میلیون حقیقت (نکته/ داده) در ۱۰۰ حوزه و دامنه است. جستارهای کلیدواژه‌ای در Freebase به‌دلیل اندازه و ساختار دسته داده مبهم هستند. ما برای کاهش ابهام جستارها، از Freebase API و انواع محدودشده موجودیت‌هایی استفاده کردیم که با استفاده از نگاشت‌ها و بازآفرینی‌های تعریف‌شده ACROMEM با انواع[1] موجودیت‌های Freebase هماهنگ شود. مثلاً ما به‌جای نوع تعریف‌شده «people/ person» در Freebase از نوع تعریف‌شده «person» در ACROMEM استفاده کردیم و به‌جای انواع «location/ continent»، «location/ location» و «location/ country» در Freebase، نوع «location» در ACROMEM را به‌کار بردیم. مثلاً موجودیت ACROMEM برای نوع «person» با برچسب «Angela Markel» در جستار Freebase MQL طوری نگاشته می‌شود که یک موجودیت Freebase منحصربه‌فرد را با متوسط «/m/0jl0g» بازیابی کند. در خصوص هم‌بستگی داده، با دو نوع ارتباط مستقیم و غیرمستقیم مواجه هستیم. لطفاً توجه داشته باشید که «هم‌بستگی و ارتباط» به‌معنی تساوی یا معادل‌بودن نیست (برای مثال: شبیه جعد: مانند) بلکه صرفاً به‌معنای داشتن سطحی از ارتباط است.

شکل ۲ هر دو مورد ارتباط مستقیم و غیرمستقیم را نشان می‌دهد. ارتباط و هم‌بستگی مستقیم به‌وسیله غنی‌سازی مشترک و مساوی مشخص می‌شود یعنی همه موجودیت‌ها یا رویدادهایی که غنی‌سازی مشابه و مشترک داشته باشند، هم‌بسته و مرتبط هستند و بنابراین خوشه‌بندی می‌شوند. ارتباط مستقیم در میان موجودیت نوع «person» با برچسب «JeanClaude Trichet» و رویداد «تریچت درباره بحران سیستمی/ سازگانی وام/ بدهی

1. Type

هشدار می‌دهد» کاملاً واضح و قابل‌مشاهده است. به‌علاوه، غنی‌سازی‌های بازیابی‌شده، موجودیت‌های ARCOMEM و اشیاء هم‌بسته وب را با دانش هم‌بسته می‌سازد یعنی ترسیم داده، در دسته داده مرجع هم‌بسته موجود است.



شکل ۲- غنی‌سازی و مثال ارتباط هم‌بستگی: اشیاء وب ACROMEM، موجودیت‌ها/رویدادها، غنی‌سازی‌های DBpedia هم‌بسته و ارتباطات شناسایی‌شده

برای مثال، منبع DBpedia بانک مرکزی اروپا[1] حقایق بیشتری را (مثلاً، رده‌بندی به‌عنوان سازمان، اعضای آن یا رؤسای سابق) در شکل ساختاری و درنتیجه قابل‌پردازش با ماشین ارائه می‌کند. بهره‌برداری از ترسیمات دسته‌های داده مرجع، امکان شناسایی ارتباط و هم‌بستگی مستقیم بیشتری را فراهم می‌کند. بااینکه رویکردهای زبان‌شناسی یا نحوی در کشف و شناسایی رابطه میان دو غنی‌سازی بالا (ECB و Trichet) و درنتیجه موجودیت‌ها و اشیاء وب مرتبط موفق نیست، اما تحلیل ترسیم DBpedia به ما کمک می‌کند تا رابطه نزدیک میان این دو را آشکار کنیم (تریچت رئیس سابق ECB). بنابراین با بررسی رایانه‌ای ارتباط غنی‌سازی‌ها ما می‌توانیم ارتباطات غیرمستقیم را برای ایجاد رابطه (خط dash) میان رویدادها/ موجودیت‌های خیلی مرتبط استفاده کنیم، (به غیر از تساوی و معادل‌بودن).

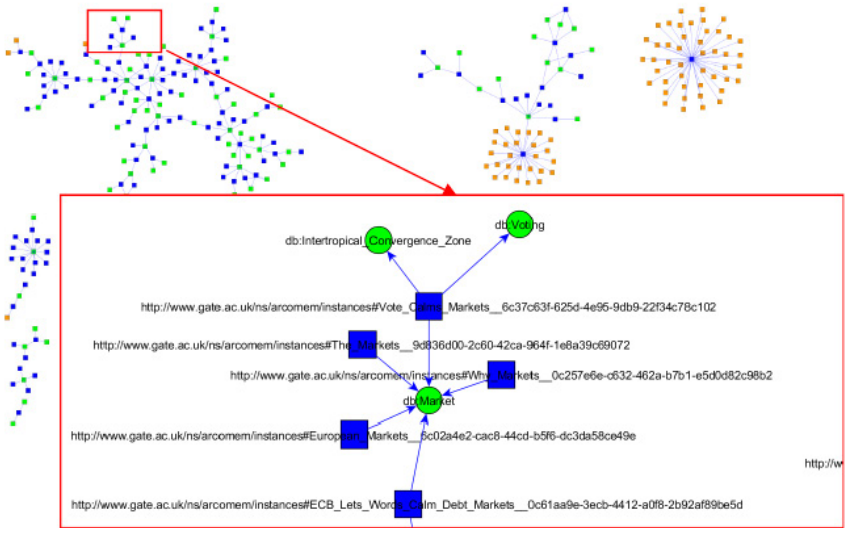
اقدام فعلی ما به شناسایی ارتباطات مستقیم محدود می‌شود، درحالی‌که تجربه بعدی ما بر سازوکارهای تحلیل ترسیم استوار است و هدف از آن محاسبه خودکار روابط معنایی موجودیت‌ها در دسته‌های داده مرجع به‌منظور کشف روابط غیرمستقیم است. بااینکه در یک تریمرگ همه مفاهیم به‌نوعی باهم مرتبط هستند، بررسی رهیابی و هدایت ترسیم صحیح و روش‌های تحلیل برای کشف روابط غیرمستقیم اما معنی‌دار در میان منابع موجود در دسته‌های داده مرجع و به‌تبع آن موجودیت‌های ACROMEM و اشیاء وب، چالش اساسی در پژوهش است.

1. European Central Bank: http://DBpedia.org/resource/ECB

۵- نتایج و ارزیابی

ما در تجربه خود دسته داده متشکل از اشیاء آرشیوی آلمانی و انگلیسی را به‌کار گرفتیم و نمونه‌ای از خزش‌های مربوط به بحران مالی[۱] را ایجاد کردیم. محتوای انگلیسی، ۳۲ پست فیس‌بوک، ۴۱۰۰۰ توئیت و ۸۰۰ نظریه کاربر در شبکه بحران یونان (greekcrisis.net) را شامل می‌شد. محتوای آلمانی شامل داده آرشیوی مجلس اطریش[۲] حاوی ۳۲۶ مستند (بیشتر به‌صورت پی.دی.اف. و برخی به‌صورت اچ. تی. ام. ال.) بود.

حاصل تجربه استخراج و غنی‌سازی ما، ارزیابی دسته داده[۳] متشکل از ۹۹۵۶۹ موجودیت منفرد شامل انواع «رویداد»، «مکان»، «پول»، «سازمان»، «شخص» و «زمان» بود. با به‌کارگیری رویه‌های فوق‌الذکر، به غنی‌سازی ۱۳۵۸ موجودیت در دسته داده خود با استفاده از DBpedia (۴۸۴ موجودیت) و Freebase (۹۷۵ موجودیت) دست‌یافتیم: در مجموع، ۵۲۹۱ غنی‌سازی در Freebase و ۴۹۱ غنی‌سازی در DBpedia. این غنی‌سازی‌ها ۵۸۰۱ جفت غنی‌سازی موجودیت: ۵۰۳۹ با Freebase و ۴۹۲ با DBpedia را به‌وجود می‌آورد.



شکل ۳- خوشه‌ها و ترسیمات ایجادشده ARCOMEM

۵-۱ ارزیابی استخراج موجودیت

ما ارزیابی‌های مقدماتی را روی اجزاء مختلف و متعدد تحلیل متن انجام داده‌ایم. مجموعه کوچکی از ۲۰ Freebase را (به زبان انگلیسی) از دسته داده توصیف‌شده در بالا با موجودیت‌های نامدار به‌صورت دستی حاشیه‌نویسی کرده‌ایم تا مجموعه‌ای با استاندارد طلایی به‌وجود آوریم. این مجموعه

۱. بخشی از خزش‌های آرشیوی در این نشانی موجود است: http://collections.europarchive.org/arcomem

2. http://www.parliament.gv.at/

۳. پایانه SPARQL دسته داده ما (موجودیت‌های استخراج‌شده و غنی‌سازی‌ها) در نشانی زیر موجود است:
http://arcomem.13s.uni-hannover.de:9988/openrdf-sesame/repositories/arcomem-rdf?query.

شامل ۹۳ نمونه از موجودیت‌های نامدار است. ما برای ارزیابی Term Raider از دسته بزرگ‌تری با ۸۰ مستند از دسته داده بحران مالی استفاده کردیم که Term Raider از آن ۱۰۰۳ اصطلاح مناسب و منتخب را ایجاد کرد (این اصطلاحات از ادغام نتایج سه سامانه امتیازدهی مختلف به‌دست آمد) سه شخص حاشیه‌نویس اصطلاحات ارزشمند را از این فهرست انتخاب کردند و ما توانستیم استاندارد طلایی ۳۱۵ را تولید کنیم. هر اصطلاح مناسب و منتخب در این استاندارد حداقل توسط دو حاشیه‌نویس انتخاب‌شده است (۲۲۱ اصطلاح توسط دو نفر و ۹۴ اصطلاح توسط هر سه نفر انتخاب شده است). بااینکه توافق درباره حاشیه‌نویسی‌ها، سطح پایین است اما برای استخراج اصطلاح که کاری موضوعی است، طبیعی می‌باشد. بااین‌وجود، در آینده رهنمودها و دستورالعمل حاشیه‌نویسی را سخت‌تر می‌کنیم و برای دستیابی به توافق بیشتر حاشیه‌نویس‌ها را تحت آموزش قرار خواهیم داد.

ما برای ارزیابی تشخیص موجودیت‌های نامدار، حاشیه‌نویسی‌های سامانه را با استاندارد طلایی مقایسه کردیم. این سامانه به‌دقت ۸۰% و فراخوانی ۶۸% در کار کشف و شناسایی موجودیت‌های نامدار دست یافت (یعنی کشف وجود یا عدم وجود یک موجودیت صرف‌نظر از نوع آن). در امر تشخیص نوع (تشخیص «نوع» صحیح موجودیت مثل شخص، سازمان، مکان و...)، این سامانه با دقت ۹۸/۸% و فراخوانی ۹۸/۵% عمل می‌کند. درمجموع (برای ترکیب این دو کار) امتیاز تشخیص موجودیت نامدار به ۷۹% دقت و ۶۷% فراخوانی می‌رسد. بااین‌وجود نتایج کمی پایین است زیرا این کار در واقع شامل کشف و شناسایی جمله نیز می‌شود. طبیعتاً کشف جمله ازنظر درستی و دقت ۱۰۰% است (یا کافی است) اما در این مورد تحت تأثیر مسئله شناسایی زبان قرار می‌گیرد زیرا در این کار ما فقط کشف و شناسایی موجودیت را برای جملاتی انجام داده‌ایم که آن‌ها را مرتبط دانسته‌ایم (به زبان کار و به بخشی از مستند این کار و متن اصلی نامه‌های کاربران). ۲۶ حاشیه‌نویسی سامانه‌های مفقود، در این مستند خارج از جملات حاشیه‌نویسی‌شده قرارگرفته است. با خارج‌کردن این‌ها درصد فراخوانی از ۶۸% به ۸۳/۹% برای کشف موجودیت نامدار (در جدول به‌صورت کشف موجودیت نامدار تنظیم‌شده آمده است) و برای کل کار از ۶۷% به ۷۳/۵% افزایش‌یافته است (در جدول به‌صورت شناسایی کامل موجودیت نامدار (تنظیم‌شده) نشان داده شده است).

کار	دقت	فراخوانی	F۱
کشف و شناسایی موجودیت نامدار	%۸۰	%۶۸	%۷۴
کشف و شناسایی موجودیت نامدار (تنظیم‌شده)	%۸۰	%۸۳/۹	%۸۱/۹
تعیین نوع	%۹۸/۸	%۹۸/۵	%۹۸/۶
شناسایی کامل موجودیت نامدار	%۷۹	%۶۷	%۷۲/۵
شناسایی کامل موجودیت نامدار (تنظیم‌شده)	%۷۹	%۸۲/۱	%۸۰/۵

جدول۱- نتایج ارزیابی شناسایی موجودیت‌های نامدار

ما برای شناسایی بیشتر، خروجی Term Raider برای هر سامانه امتیازدهی را در سطوح مختلف فهرست رتبه‌بندی با دسته اصطلاحات استاندارد طلایی مقایسه کردیم (شکل۴). برای اصطلاحات بالاتر از آستانه به‌دقت ۳۱% و فراخوانی ۹۰% برای tf.idf، ۷۳% دقت و ۵۰% فراخوانی برای tf.idf افزایشی و ۶۳% دقت و ۱۷% فراخوانی برای امتیاز Kyoto دست‌یافتیم. ما برای سایر پردازش‌ها فقط از اصطلاحاتی استفاده می‌کنیم که از طریق tf.idf افزایشی امتیازی بالاتر از آستانه داشته باشند.

۲-۵ ارزیابی غنی‌سازی و هم‌بستگی ارتباط (ارتباط)

ما برای این کار دسته‌ای از موجودیت‌های غنی‌شده دوتایی را به‌صورت تصادفی انتخاب کردیم. ارزیابی به‌صورت دستی و توسط ۶ داور شامل دانشجویان و پژوهشگران فارغ‌التحصیل علوم رایانه‌ای انجام شد. داوران به هر جفت از موجودیت‌های غنی‌شده امتیازی دادند: صفر برای «نادرست» و ۱ برای درست. آن غنی‌سازی درست تلقی می‌شد که تا حدودی ابعاد خاص یک موجودیت/ رویداد را معرفی کند؛ یعنی غنی‌سازی لازم نیست کاملاً با موجودیت منطبق باشد. مثلاً غنی‌سازی‌های مربوط به (http://dbpedia.org/resource/Doctor_title) و http://dbpedia.org/page/Angela_Markel و غنی‌سازی موجودیت نوع «شخص» با برچسب «Dr. Angela Markel» هر دو به‌عنوان «درست» رتبه‌بندی می‌شوند. این به‌خاطر موجودیت‌ها و رویدادهایی است که به غنی‌سازی چندگانه مربوط هستند و هر کدام وجه خاصی از رویداد/موجودیت منبع را غنی‌سازی می‌کنند. هر جفت موجودیت غنی‌سازی‌شده، حداقل به ۳ داور نشان داده شد و برای کاهش خطا و انحراف متوسط امتیازات آن‌ها در نظر گرفته شد. اگر برچسب یک موجودیت برای یکی از داوران مفهوم نداشت، فرض را براین گذاشتیم که در مرحله استخراج اشتباهی رخ داده است. دراین‌خصوص ما از داوران خواستیم موجودیت مذکور را فاقد ارزش اعلام کنند و آن را از برنامه ارزیابی خارج کنند.

ما متوسط امتیازات مربوط به جفت‌های غنی‌سازی موجودیت را محاسبه کردیم و متوسط امتیاز هر «نوع» موجودیت را نیز به‌دست آوردیم. جدول ۴ امتیاز متوسط جفت‌های غنی‌سازی شده موجودیت‌ها را با استفاده از DBpedia و Freebase برای انواع مختلف موجودیت ACROMEM ارائه می‌کند.

نوع موجودیت	امتیاز متوسط DBpedia	امتیاز متوسط Freebase	مجموع امتیاز متوسط
مکان	۰/۹۴	۰/۹۴	۰/۹۴
پول	۰/۶۳	-	۰/۶۳
سازمان	۰/۹۳	۱	۰/۹۷
شخص	۰/۷۲	۰/۸۹	۰/۸
زمان	۱	-	۱
جمع کل	۰/۸۴	۰/۹۴	۰/۸۹

جدول۲- نتایج ارزیابی غنی‌سازی

رویکرد مقدماتی خوشه‌بندی به‌سادگی با رویدادها/ موجودیت‌های دارای غنی‌سازی مشترک مرتبط شد. در مجموع ما ۱۰۱۳ خوشه با متوسط ۲/۸۵ موجودیت یعنی حداقل ۲ و حداکثر ۱۱۲ موجودیت، ایجاد کردیم. غنی‌سازی‌های مبهم خوشه‌های زائد ایجاد می‌کردند و به رفع ابهام مجدد نیاز داشتند. مثلاً موجودیت مکان با برچسب «Berlin» (به‌درستی) با دو نشانی مربوط به مکان‌های مختلف برلین غنی‌سازی می‌شد اما برای رفع ابهام باید خوشه‌های زائد برطرف می‌شد. در پایان ما از روش‌های تحلیل ترسیم برای کشف و شناسایی تقریب غنی‌سازی‌های مربوط به یک شیء استفاده کردیم. مثلاً، سنجش ارتباط دو موجودیت مکانی «برلین» و «Angela Markel» که برای حاشیه‌نویسی و توضیح شیء مشابه در وب به‌کار می‌رود. ما را برای رفع ابهام از غنی‌سازی‌ها کمک می‌کند.

۶ مباحثه و کارهای آتی

ما در این مقاله، راهبرد فعلی استخراج و غنی‌سازی موجودیت طبق پروژه ARCOMEM ارائه کرده‌ایم با این هدف که پایگاه دانش بزرگی را از دانش ساختاری مربوط به محتوای آرشیوی نامتجانس وب ایجاد کنیم. ما بر اساس زنجیره پردازش یکپارچه، کار غنی‌سازی و ادغام و تقویت موجودیت را به‌عنوان فعالیتی ضمنی در جریان کار استخراج اطلاعات انجام دادیم.

نتایج استخراج موجودیت، امتیازات قابلِ‌توجهی را برای این نوع داده رسانه اجتماعی نشان داد که روش‌ها و فنون NLP برپایه آن عمل می‌کنند. بااین‌حال، کار فعلی بر جابه‌جایی و مدیریت بهتر انگلیسی دون‌پایه و زوال‌یافته متمرکز است (نمونه‌نمایی، شناسایی زبان و ...) به‌ویژه توئیت‌ها که استخراج موجودیت را بهبود بخشیده‌اند. نتایج غنی‌سازی، غنی‌سازی‌هایی با کیفیت بالا را بیان می‌کند. نتایجی که از DBpedia Spotlight به‌دست آمد، فراخوانی پایین‌تری را ارائه کرد اما به‌دلیل شاخص رفع ابهام Spotlight، غنی‌سازی‌های واضح‌تر و بهتری را معرفی کرد. به‌عبارت‌دیگر، کلیدواژه‌هایی با سازگاری تقریبی دقت نتایج را کاهش می‌دهد. کار آتی ما، پیش‌بینی جهات مختلف بهبود کیفیت نتایج غنی‌سازی است. برای مثال، استفاده از جستارهای ساختاری DBpedia برای محدودکردن انواع موجودیت مشابه رویکرد موردِاستفاده برای Freebase. ما هم‌چنین زیرنمونه‌های (زیرنوع‌های)[1] موجودیت را برای افزایش سازگاری «نوع‌ها» معرفی کردیم. به‌علاوه، بااینکه حفظ و نگهداری محتوای وب در طول زمان جنبه‌ای موقت و گذرا دارد، تحول موجودیت‌ها و اصطلاحات و رفع ابهام (وابسته‌به‌زمان) زمینه‌های پژوهشی مهمی هستند که درحال‌حاضر در دست بررسی می‌باشند [۲۴]. با وجودی‌که رویکرد فعلی ادغام و تقویت داده فقط روابط مستقیم میان موجودیت‌های دارای یک غنی‌سازی را کشف و شناسایی می‌کند، تلاش اصلی ما به بررسی سازوکارهای تحلیل ترسیم اختصاص یافته

1. Subtype

است. بنابراین هدف ما بهره‌برداری از دانش رمزی‌شده در ترسیمات مرجع به‌منظور شناسایی خودکار روابط معنایی میان موجودیت‌های متفاوت استخراج‌شده از چرخه‌های پردازش مختلف است. با توجه به افزایش کاربرد ابزارهای شناسایی موجودیت‌های نامدار خودکار و دسته‌های داده مرجع مثل DBpedia، WordNet یا Freebase، نیاز به ادغام و تقویت اطلاعات وب که به‌طور خودکار استخراج می‌شود نیز بیشتر می‌شود و ما در کار خود تلاش می‌کنیم این امر را میسر سازیم.

منابع و مراجع:

Bizer, C., Heath, Berners-Lee, T. (2009) Linked data – The Story So Far. Special Issue on Linked data, International Journal on Semantic Web and Information Systems.

Dietze, S., and Domingue, J. (2008) Exploiting Conceptual Spaces for Ontology Integration, Workshop: Data Integration through Semantic Technology (DIST2008) Workshop at 3rd Asian Semantic Web Conference (ASWC) 08, Bangkok Thailand.

Dietze, S., Gugliotta, A., and Domingue, J. (2009) Exploiting Metrics for Similarity – based Semantic Web Service Discovery, IEEE 7th International Conference n Web Services (ICWS2009), Los Angles, CA, USA.

Lu, L., Zhou,: Link prediction in complex networks: a survey, Physica A 390 (2011), 1150-1170.

Hasan, M. A., Zaki, M. J.: A survey of link prediction in social network. In C. Aggrwal, editor, Social Network Data Analytics, pages 243-276. Springer.

Cohen, W. W., Ravikumar, P. D., Fienberg< S.E. A comparison of string distance metrics for name-matching tasks. In IIWeb, 2003.

Dong, X., Halevy, A., Madhavan, J., Reference reconciliation in complex information spaces. In SIGMOD, 2005.

Elmagarmid, A. K., Ipeirotis, P. G., Verykios, V.S., Duplicate record detection: A survey. TKDE, 19(1), 2007.

Tejada, S., Knoblock, C. A., Minton, S., Learning domain-independent string transformation weights for high accuracy object identification. In KDD, 2002.

Boley, D., Principal Direction Divisive Partitioning. Data Mining and Knowledge Discovery, 2(4), 1998.

Broder, A., Glassman, S., Manasse, M., Zweig, G., Syntactic Clustering of the Web. In Proceedings of the 6th International World Wide Web Conference, pages 1997.

Hotho, A., Maedche, A., Staab S., Ontology-based Text Clustering. In Proceeding of the IJCAI Workshop on \ Text Learning: Beyond Supervision", 2001.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y., Named Entity Recognition from Diverse Text Types. Recent Advances in Natural Language Processing 2001 Conference, Tzigov Chark, Bulgaria, 2001.

Li, Y., Bontcheva, K., Cunningham, H., Adapting SVM for Data Sparseness and Imbalance: A Case Study on Information Extraction. Natural Language Engineering, 15 (02), 241-271,2009.

Buckley, C., G. Salton, G., Term-Weighting Approaches in automatic text retrieval. Information Processing and Management, 24(5):513-523, 1988.

Maynard, D., Li, Y., Peters, W., NLP techniques for term extraction and ontology population. In: Buitelaar, P. and Cimiano, P. (eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp. 171-199. IOS Press, Amsterdam (2008).

Lui, M., Baldwin, T., 2011. Cross-domain feature selection for language for language identification. In Proceeding of 5th International Joint Conferences on Natural Language Processing, pages 553-561, November.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Application. Proceedings of the 40th Anniversary Meeting of the Association for Comptutational Linguistics (ACL02).

Ritter, A., Clark, S., Mausam, Etzioni, O., 2011. Named entity recognition in Twits: An experimental study. In proc. Of Empirical Methods for Natural Language Processing (EMNLP), Edinburgh, UK

Risse, T., Dietze, S., Peters, W., Doka, K., Stavrakas, Y., Senellart, P., Exploiting the Social media. In Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012, May 2012, Istanbul, Turkey.

Maynard, D., Bontcheva, K., Rout, D., Challenges in developing opinion mining tools for social media. In Proceedings of@NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012, May 2012, Istanbul, Turkey.

Bosma, W., Vossen, P., 2010 Bootstrapping languageneutral term extraction. In 7th Language Resources and Evaluation Conferences (LREC), Valleta, Malta.

Deane, P. A nonparametric method for extraction of candidate phrasal terms, In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005.

Tahmasebi, N., Risse, T., Dietze, S. (2011) Towards Automatic Language Evolution Tracking: A Study on Word Sense Tracking, Joint Workshop on Knowledge Evolution and Ontology Dynamics 2011 (EvoDin2011), at the 10th International Semantic Web Conferences (ISWC2011), Bonn, Germany.

Weiss, C., Karras, P., Bernstein, A., Hexastore: sextuple indexing for semantic web data management. Proceeding of the VLDB Endowment, 1(1): 1008-1019, 2008.