مروری بر سیستم‌های پرسش‌وپاسخ به زبان طبیعی روی داده‌های پیوندی

مهدی بخشی[۱]، محمدعلی نعمت‌بخش[۲]، مهران محسن‌زاده[۳]

[۱] گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد شهربابک، شهربابک
mb@shahrbabakiau.ac.ir

[۲] گروه مهندسی نرم افزار، دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان
nematbakhsh@eng.ui.ac.ir

[۳] گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد علوم و تحقیقات، تهران
mohsenzadeh@srbiau.ac.ir

چکیده

با توجه به سرعت افزایش دانش ساخت‌یافته معنایی در وب داده، نیاز به فراهم کردن روش‌هایی برای جستجوی این حجم زیاد اطلاعات ساخت‌یافته و ناهمگن، بیش از پیش حس می‌شود. در سالیان اخیر تلاش زیادی برای ارائه واسطهایی مبتنی بر زبان طبیعی برای جستجوی داده‌های معنایی انجام شده است. توسعه این واسطها دارای این مزیت است که آنها می‌توانند از قدرت توصیف‌گری مدل داده و زبان پرس‌وجوی وب معنایی، همزمان با مخفی‌سازی پیچیدگی‌شان از دید کاربر، استفاده کنند. در این مقاله پس از ذکر چالش‌های موجود، به مرور دسته‌های مختلف سیستم‌های پرسش‌وپاسخ روی داده‌های پیوندی می پردازیم و نحوه عملکرد هر سیستم در مراحل مختلف را بیان کرده و همچنین به ذکر مزایا و معایب آنها می‌پردازیم.

کلمات کلیدی

سیستم پرسش‌وپاسخ، داده پیوندی، RDF، زبان پرس‌وجوی SPARQL

۱- مقدمه

وب معنایی توسط تیم برنرزلی در سال ۲۰۰۱ ارائه شد [1]. هدف اصلی وب معنایی که وب داده نیز گفته می‌شود، ارائه یک چارچوب مشترک است که در آن داده بتواند به اشتراک گذاشته شود و توسط ماشین قابل تفسیر باشد. یکی از روش‌های تحقق این هدف، داده پیوندی است که در آن، برخلاف وب سنتی که اسناد به یکدیگر متصل می‌شدند، داده‌ها بر اساس چند قانون ساده به‌صورت معنایی به یکدیگر ارتباط داده می‌شوند [2]. داده پیوندی از چارچوب توصیف منابع RDF که یک مدل داده بر اساس سه‌تایی‌ها است به‌منظور انتشار اطلاعات پیرامون موجودیت‌ها و روابط بین آنها استفاده می‌کند.

حجم دانش ساخت‌یافته و ناهمگن روی وب به شدت در حال افزایش است. وجود این حجم زیاد از دانش، فرایند جستجو و پرس‌وجوی آن را چالش برانگیز ساخته است [3,4]. برای دسترسی به منابع اطلاعاتی در وب، نیاز اطلاعاتی کاربر بر اساس روش‌های مختلفی می‌تواند بیان شود. این روش‌ها می‌تواند شامل استفاده از کلیدواژه‌ها، جملات زبان طبیعی یا پرس‌وجوی ساخت‌یافته باشد. یکی از معایب استفاده از دستورات زبان‌های پرس‌وجوی ساخت‌یافته مثل SPARQL، عدم آشنایی کاربران عادی با نحو این دستورات و واژگان خاص پایگاه داده‌ها است [5]. به همین دلیل، استفاده از زبان‌های طبیعی برای بیان نیاز کاربر به یکی از روش‌های متداول تبدیل شده است. در چنین مواردی کاربر می‌تواند نیاز خود را به صورت سنتی از طریق

کلیدواژه‌ها بیان کند یا از جملات کامل استفاده کند. در مواردی که کاربر به جای جملات کامل از کلیدواژه‌ها استفاده می‌نماید، وظیفه سیستم بازیابی در تعیین نیت کاربر سخت‌تر خواهد بود [6].

نوعی از سیستم‌های بازیابی اطلاعات که اجازه می‌دهند کاربر نیاز خود را به صورت عبارات یا سوالات کامل زبان طبیعی مطرح کند، سیستم پرسش-وپاسخ نامیده می‌شوند [7]. همانطور که در شکل (۱) مشخص است عملکرد یک سیستم پرس‌وجو روی داده‌های RDF شامل مراحل آنالیز سوال، نگاشت عبارات سوال به مفاهیم پایگاه شامل موجودیت‌ها و کلاس‌ها و روابط، رفع ابهام از نگاشت انجام شده و ساخت پرس‌وجوی ساخت‌یافته نهایی است [8].

در این مقاله در بخش مقدمه ضرورت ایجاد واسطهای زبان‌طبیعی روی داده‌های پیوندی برای پرس‌وجو بیان شده است. در بخش دوم به بیان چالش‌های اساسی در مورد این سیستم‌ها پرداخته می‌شود. در بخش سوم مروری بر روش‌های مورد استفاده در مراحل مختلف سیستم‌های پرسش-وپاسخ روی داده‌های پیوندی با توجه به دسته‌بندی کلی روش‌ها بیان می‌شود و کارهای مهم‌تر در سال‌های اخیر در این دسته‌بندی قرار می‌گیرند. بخش چهارم به صورت خلاصه به مقایسه این روش‌ها و ذکر مزایا و معایب هر دسته می‌پردازد و در آخر در فصل پنجم نتیجه‌گیری بیان می‌شود.

۲- چالش‌ها

در این قسمت با توجه به مراجع [3,9] به چالش‌های موجود در مورد سیستم-های پرسش‌وپاسخ روی داده‌های پیوندی پرداخته شده و هر کدام از این چالش‌ها به اختصار توضیح داده شده است.

یکی از این چالش‌ها اختلاف لغوی است. کلماتی که در پرس‌وجوی کاربر بیان می‌شوند، ممکن است از نظر ظاهری با برچسب‌شان در وب داده متفاوت باشند، ولی معنی یکسانی داشته باشند. این مشکل به دلیل وجود کلمات تعدد معانی ایجاد می‌شود و روی فراخوان سیستم تاثیر منفی دارد.

اختلاف ساختاری در سوال زبان طبیعی و منبع وب داده از آنجا به‌وجود می‌آید که در سطح دادنه‌بندی مفهومی بین زبان و طرح واژگانی در پایگاه داده تفاوت وجود دارد.

یکی دیگر از این چالش‌ها وجود ابهام در اجزا سوال است. از مهمترین دلایل ابهام وجود کلمات هم آوا است که دو یا چند کلمه فرم ظاهری یکسان دارند ولی دارای معانی متعدد هستند. این مشکل روی دقت سیستم تاثیر منفی می‌گذارد.

چندزبانی بودن وب داده از آنجا به وجود می‌آید که منابع RDF را می-توان همزمان به چندین زبان طبیعی با استفاده از برچسب‌های زبان توصیف کرد. بنابراین از بیش از یک زبان می‌توان برای بازیابی آنها استفاده کرد.

توزیع شدگی چالش مهم دیگر است. اطلاعات درخواستی یک پرس‌وجو ممکن است در چند پایگاه مختلف در داده‌های پیوندی توزیع شده باشند.

اکثر سیستم‌های موجود برای پاسخ دادن به سوالات پیچیده دچار مشکل هستند. در این سوالات پاسخ سوال بایستی از بعضی محدودیت‌ها و قواعد پیروی کند. این محدودیت‌ها باعث ایجاد سوالات تجمعی و مقایسه‌ای و برترین و غیره می‌شود.

۳- دسته بندی سیستم‌های پرسش‌وپاسخ روی داده‌های پیوندی

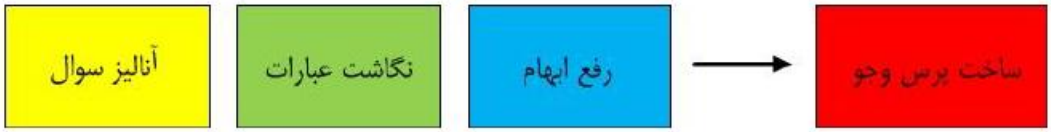
برای غلبه بر چالش‌هایی که در مورد سیستم‌های پرس‌وجو روی داده‌های پیوندی وجود دارد، گستره‌ای از روش‌ها بیان شده است. این روش‌ها از بعضی دیدگاه‌ها متفاوت هستند. از جمله طراحی شدن برای یک دامنه خاص، صرف نظر کردن از طرح واژگانی در آنتولوژی، آنالیز سوال بر اساس روش‌های زبان-شناسی عمیق، استفاده از روش‌های آماری یا اکتشاف گراف و استفاده از منابع و ابزارهای متفاوت. در این بخش، مروری بر این سیستم‌ها با استفاده از دسته‌بندی کلی مرجع [3] ارائه خواهد شد و مزایا و معایب آن‌ها بیان خواهد شد.

۳-۱- زبان‌های طبیعی محدود شده

در سیستم Ginseng [10]، ورودی کاربر زبان طبیعی کنترل شده است. واژه‌ها و ساختارهای موجود در آنتولوژی برای ساخت مجموعه لغات و گرامر استفاده می‌شود و ورودی کاربر به همین مجموعه لغات و گرامر محدود می‌شود. آنتولوژی‌ها در این سیستم می‌تواند به‌صورت دستی با استفاده از کلمات هم‌معنی، غنی شود. این نوع سیستم به کاربر اجازه بیان همه سوالی نمی‌دهد و کاربر در این زمینه با محدودیت جدی روبرو است.

۳-۲- گرامرهای رسمی

در این روشها، با استفاده از گرامرهای زبانی برای اجزای لغوی زبان، نحو و معنا در نظر گرفته می‌شود و معنای پرسش کاربر با استفاده از معناشناسی ترکیبی از ترکیب معنای اجزای آن بدست می‌آید. سیستم ORAKEL [11]، سوالات ورودی را به منطق مرتبه اول ترجمه می‌کند. این سیستم از یک مجموعه لغات برای نگاشت دقیق اجزای زبان طبیعی به موجودیت‌های آنتولوژی استفاده می‌کند. برای ساخت مجموعه لغات مستقل از دامنه، از افراد خبره باید استفاده شود. تفاوت سیستم Pythia [12] با سیستم قبل در این است که این سیستم مبتنی بر آنتولوژی است و سوال را بر اساس یک واژه‌نامه با دامنه خاص که به صورت اتوماتیک ایجاد می‌شود تجزیه می‌کند و براساس معناشناسی ترکیبی به یک پرسش ساخت‌یافته تبدیل می‌کند. مزیت این روش‌ها پرداختن به سوالات با هر سطح از پیچیدگی است و عیب آن عدم توانایی در پردازش پرسش ورودی در صورتی‌که نتواند توسط گرامر تولید شود است.



شکل (۱): اجزا و مراحل سیستم‌های پرسش‌وپاسخ [8]

۳-۳- نگاشت ساختارهای زبان‌شناختی به ساختارهای معنایی

این روش‌ها سعی می‌کنند بر خلاف روش‌های گرامری از راهکارهای سطحی‌تری استفاده کنند و ساختارهای زبانی را به سه‌تایی‌های معنایی نگاشت کنند. آنها بیشتر بر محاسبه شباهت بین عناصر سوال و سه‌تایی‌ها در پایگاه دانش تکیه دارند. در کل این روشها نسبت به روش‌های دیگر استوارترند و در حالت‌های مختلف بیشتری قادر به پاسخ‌گویی هستند، اما نیازمند یک تطابق یک‌به‌یک بین ساختار نحوی سوال و نمایش معنایی پایگاه دانش هستند [3].

سیستم CASIA [13] در سه مرحله کار ایجاد دستورات ساخت‌یافته برای پاسخ دادن به سوال را انجام می‌دهد. در مرحله تحلیل سوال با استفاده از ابزارهای زبان شناختی مثل پارسر وابستگی، سوال زبان طبیعی را به سه‌تایی‌های سوال مطابق با ساختار سه‌تایی‌های منبع در می‌آورد و سپس در مرحله نگاشت به منابع با توجه به ساختار سه‌تایی‌های سوال سعی می‌کند موضوع و هدف را به کلاس و موجودیت و روابط را به خاصیت‌ها نگاشت کند. برای نگاشت از منابع متفاوتی استفاده می‌کند و در مرحله آخر با توجه به سه‌تایی‌های حاصل و نوع سوال که می‌تواند بله یا خیر و شمارشی و معمولی و غیره باشد و با استفاده از قالب‌های از قبل طراحی شده برای هر نوع سوال، پرسش ساخت‌یافته حاصل می‌شود.

سیستم RTV [14] مدل‌سازی معنایی لغوی را با استنتاج آماری در یک معماری پیچیده ترکیب می‌کند تا تفسیر زبان طبیعی را به سه مرحله متفاوت تقسیم نماید: (۱) انتخاب عبارات در سوال مثل گزاره، آرگومان و خاصیت (۲) پیدا کردن منابع معادل در آنتولوژی از طریق رفع ابهام مشترک از تمام کاندیداها و (۳) گردآوری پرس‌وجوی نهایی روی سه‌تایی‌های RDF. این معماری از مدل مخفی مارکوف برای انتخاب سه‌تایی‌های مناسب آنتولوژی بر اساس ماهیت گراف RDF استفاده می‌کند و برای هر پرس‌وجو یک مدل مخفی مارکوف تولید می‌کند که جواب آن، رفع ابهام مشترک بین تمام عبارات جمله است. در مرحله آخر باتوجه به منابع سه‌تایی حاصل از مرحله قبل و در نظر گرفتن دامنه و محدوده در خاصیت موجود بهترین ترتیب سه‌تایی برای ایجاد SPARQL نهایی انتخاب می‌شود.

سیستم SINA [15] در اصل برای ایجاد یک سوال SPARQL فدرال از یک سوال کوتاه یا کامل بر روی مجموعه‌ای از پایگاه‌های مجزای داده‌پیوندی طراحی شده است. در این سیستم سعی شده که (۱) با استفاده از مدل مخفی مارکوف بهترین منابع از پایگاه‌های مختلف برای عبارات موجود در سوال طی مراحل قطعه‌بندی و رفع‌ابهام حاصل شود و (۲) یک روش برای ایجاد سوال ساخت‌یافته با استفاده از منابع نگاشت شده به عبارات حاصل از مرحله قبل و استفاده از دانش پیش‌زمینه از داده‌های سه‌تایی و استنتاج دامنه و برد و ساختار لینک‌دهی در پایگاه‌ها ارائه شود. برای تعیین بهترین مسیر در بین حالات مخفی برای مشاهده حالات قابل رویت از الگوریتم ویتربی استفاده می‌شود و باتوجه به منابع حاصل شده از مراحل قبل و طی دو مرحله گراف سوال حاصل می شود.

در سیستم Intui2 [16]، برای تفسیر جمله پرسشی زبان طبیعی از درخت تجزیه نحوی ایجاد شده از آن استفاده می‌شود. فرض بر این است که هر سوال از مجموعه‌ای از زیرقسمت‌های نحوی ایجاد شده است. هر زیر قسمت نحوی از نظر نحوی به یک زیردرخت تجزیه نحوی و از نظر معنایی به یک حداقل قسمت از متن که می‌تواند به‌عنوان یک URI که شناسه منبع در وب داده است، یک سه‌تایی یا یک پرس‌وجوی پیچیده تفسیر شود، مطابقت می‌کند. برای تفسیر زیرقسمت نحوی والد می‌توان زیرقسمت‌های فرزند را ترکیب کرد. برای ساخت پرسش ساخت‌یافته تفاسیر بازگشتی از زیرقسمتهای نحوی استخراج می‌شوند و به این ترتیب الگوهای نحوی ایجاد می‌شوند و سپس این الگوهای نحوی به سه‌تایی‌های منبع نگاشت می‌شوند.

سیستم DEANNA [17] شامل مراحل (۱) تشخیص عبارات و نوع آنها و (۲) نگاشت عبارات به منابع در پایگاه و تشکیل سه تایی‌های کاندید و (۳) رفع ابهام از نحوه نگاشت عبارات به منابع و (۴) گروه بندی منابع برای تشکیل سه‌تایی‌ها و ساخت پرسش ساخت‌یافته نهایی، است. هسته تشکیل دهنده این سیستم روشی برای رفع‌ابهام از نگاشت عبارات به منابع و پرکردن سه‌تایی‌های کاندید است. برای این منظور از روشی مبتنی بر رفع ابهام اشتراکی که رفع‌ابهام از یک عبارت روی نگاشت بقیه عبارات تاثیر دارد، استفاده می‌کند. ابتدا گراف رفع ابهام که شامل حالات مختلف نگاشت به همراه نودها و یال‌های وصل‌کننده آنها است، تشکیل می‌شود و سپس با استفاده از روش بهینه‌سازی برنامه‌ریزی صحیح خطی و مشخص کردن محدودیت‌های مختلف و حل کردن این مسئله بهینه‌سازی، گراف رفع ابهام خلوت شده که نگاشت نهایی را نشان می‌دهد، بدست می‌آید.

سیستم Xser [18]، کار ساخت پرسش ساخت‌یافته از سوال زبان طبیعی را در یک پارادایم متوالی در دو مرحله شناخت ساختار قصد سوال که مستقل از پایگاه دانش است و مقداردهی این ساختار به مفاهیم پایگاه دانش که وابسته به پایگاه دانش است انجام می‌دهد. در مرحله اول ابتدا کار ساخت یک تشخیص دهنده نوع عبارت، انجام می‌شود. سپس یک پارسر معنایی جهت شناخت وابستگی‌های گزاره- آرگومان بین عبارات سوال ایجاد می‌شود. در مرحله دوم که وابسته به یک پایگاه دانش است ابتدا بر اساس نوع عبارت بعضی کاندیدها از منابع پایگاه به عبارت استخراج می‌شوند. سپس کار رفع ابهام از نگاشت‌های موجود شامل نودهای عبارت برچسب و روابط وابستگی با آیتم‌های مطابق‌شان در پایگاه دانش انجام می‌شود.

بیشتر سیستم‌های پرسش‌وپاسخ قبل از CASIA V2 [19] سعی می‌کنند که بر اساس یک چهارچوب متوالی شامل شناخت عبارات سوال زبان طبیعی، نگاشت آنها به منابع داده‌پیوندی، گروه‌بندی منابع به صورت سه‌تایی و تشکیل پرسش SPARQL، عمل کنند که این باعث می‌شود هر مرحله عمل رفع‌ابهام را فقط در آن مرحله انجام دهد و نتایج آن به مرحله بعد انتقال پیدا کنند. در عوض این روش کل مراحل را با هم در نظر می‌گیرد و کار تشخیص عبارات و نگاشت به منابع پایگاه و گروه‌بندی آنها را در یک مرحله مشترک رفع‌ابهام می‌کند. این سیستم این کار را با مدل کردن مسئله به وسیله روش شبکه منطقی مارکوف انجام می‌دهد.

در سیستم Intui3 [20] ابتدا هر سوال زبان طبیعی ورودی توسط ابزارهای زبان‌شناختی پیش‌پردازش می‌شود و سیستم بر اساس اطلاعات برچسب‌ها بعضی قطعه‌ها را تجزیه می‌کند و بعضی دیگر را ترکیب می‌کند و قطعات جدیدی ایجاد می‌کند. سپس بر اساس هر قطعه‌ای یک یا چند تفسیر را برای آن قطعه در نظر می‌گیرد و قطعات را از چپ به راست در نظر می‌گیرد و تفاسیر آنها را ترکیب می‌کند و حالات مختلف آنها را در نظر می‌گیرد. حال با دو روش امتیازدهی شباهت رشته‌ای و مبتنی بر WordNet [21] به آنها

امتیاز می‌دهد و بهترین را انتخاب می‌کند و کار ترکیب تفاسیر را تا انتهای قطعات انجام می‌دهد.

در سیستم gAnswer [22] دو قسمت برخط و برون‌خط ارائه شده است. در قسمت برون‌خط یک دیکشنری عبارات که روابط در جمله و معادل خاصیت آنها در پایگاه را مشخص می‌کند، ایجاد می‌شود. حال در قسمت برخط در مرحله فهمیدن سوال یک گراف پرسش معنایی حاصل می‌شود و پس از انتصاب کاندیدهای مختلف به نودها و یال‌های گراف، در مرحله ارزیابی سوال برای رفع ابهام از کاندیدها و به دست آوردن گراف سوال SPARQL مسئله را به پیدا کردن زیر گرافی یکریخت از گراف RDF مدل می‌کند و چون این یک مسئله رام نشدنی است از قواعدی برای کوچک کردن فضای جستجو استفاده می کند که در کل باعث اثربخشی و بهروری مطلوب سیستم می‌شود.

سیستم SemGraphQA [23] بر پایه تبدیل گراف طراحی شده است. سیستم ابتدا در مرحله نگاشت معنایی، تمام عبارات که قادر به نگاشت به منابع پایگاه هستند را پیدا کرده و سپس به نگاشت عبارات به موجودیت‌های پایگاه می‌پردازد. در مرحله بیان معنایی سوال یک گراف ساختاری با در نظر گرفتن نحوه اتصال وابستگی بین اجزا جمله ایجاد می‌کند. سیستم بعد از اعمال قواعدی، گراف‌های ساختاری-معنایی که هر کدام بیانگر تفسیر متفاوت سوال است را بدست می‌آورد. سپس با توجه به اعمال چند قاعده چند گراف معنایی حاصل می‌شود. سپس آنها را به ترتیب نمره مرتب می‌کند و سوال SPARQL مربوط به هر کدام را روی پایگاه دانش اجرا می‌کند تا به جواب برسد.

سیستم QAnswer [24] از ویکی‌پدیا برای نگاشت بهتر موجودیت‌ها، کلاس‌ها و روابط به منابع پایگاه استفاده می‌کند. این سیستم ابتدا یک گراف وابستگی از سوال زبان طبیعی می‌سازد و بعد از آن سعی می‌کند هر عبارت در آن را به منابع پایگاه نگاشت کند و به این صورت گراف‌های متعددی ایجاد می‌کند. بعد از انتصاب منابع کاندید به هر عبارت در گراف وابستگی، به گراف‌ها بر اساس میزان تطابق هر عبارت با منبع پایگاه و همچنین وجود سه-تایی‌های گراف در پایگاه، امتیاز می‌دهد و گراف با بالاترین امتیاز به آخرین مرحله می‌رود. در مرحله آخر برای ساخت سوال SPARQL بر اساس یک الگوریتم بازگشتی بر اساس پیمایش گراف، سه‌تایی‌ها ساخته می‌شوند.

۳-٤- مبتنی بر الگو

با وجود آنکه سیستم‌های مبتنی بر الگو یکی از سریع‌ترین روش‌ها برای توسعه سیستم‌های پرسش‌وپاسخ هستند، اما اقبال به سوی این سیستم‌ها در چند سال اخیر کمتر بوده است. دلیل آن احتمالا به خاطر انعطاف‌پذیری کم در مقابل سیستم‌های با دامنه باز است [9]. مزیت این دسته توانایی پاسخ دادن به بعضی سوالات تجمعی که سیستم‌های دیگر معمولا در این مورد با مشکل مواجه می‌شوند است. همچنین دارای این عیب هستند که برای پاسخ دادن به سوالاتی که بین سوال و داده‌ها اختلاف ساختاری وجود دارد با مشکل مواجه می‌شوند [3]. در این قسمت مروری بر این روش‌ها ارائه خواهد شد.

سیستم TBSL [25] یک شرح وفادارانه از ساختار معنایی سوال ارائه می‌کند. برای این منظور سیستم ابتدا یک الگوی SPARQL راکه به طور کامل ساختار داخلی سوال کاربر را منعکس می‌کند، تولید می‌کند و سپس این الگو را با URI مربوط به منابع که از روش‌های آماری برای تشخیص موجودیت‌ها و روابط بدست آمده، پر می‌کند. سیستم BELA [26] بر اساس سیستم TBSL و برای نگاشت بهتر عبارات سوال و منابع پایگاه با ایجاد یک پارادایم خطاولهای توسعه‌یافته است. ویگو ژنگ و همکارانش در [27] روشی برای تولید اتوماتیک الگوها، با مقایسه گراف‌های سوال و گراف‌های پرسش-های SPARQL بیان کرده‌اند.

۳-٥- اکتشاف گراف

این سیستم‌ها معمولا به جای تولید گراف سوال ساخت‌یافته و ایجاد پرسش SPARQL، در پی مسیرهایی در گراف وب داده هستند که بیشترین شباهت را با سوال زبان طبیعی دارند و به پاسخ منتهی می‌شوند. مشکل اصلی این روش‌ها استخراج گراف بزرگ داده‌ها است [3]. در این قسمت مروری کلی بر این سیستم‌ها ارائه می‌شود.

سیستم Treo [28] کار پردازش سوال را با استخراج موجودیت‌های محوری در سوال شروع می‌کند و آنها را به عنوان نقطه شروع برای الگوریتم جستجوی انتشاری در نظر می‌گیرد. با شروع از این موجودیت‌ها، الگوریتم کار پیمایش نودهای همسایه که بیشترین شباهت را با عناصر سوال داشته باشند، انجام می‌دهد. این روش مجموعه‌ای از مسیرهای سه‌تایی مرتب شده که از موجودیت محوری شروع می‌شوند و به منبع نهایی که جواب مورد نظر است ختم می‌شوند، به عنوان خروجی تولید می‌کند. سیستم SemSek [29] در مقایسه با سیستم قبل سعی می‌کند روش بهتری برای نگاشت عناصر سوال به مفاهیم پایگاه ارائه کند.

سیستم Graph Traversal [30] از بین مراحل استخراج ساختارهای روابط معنایی در سوال زبان طبیعی و رفع ابهام از آنها با مقداردهی آنها با توجه به داده‌های پایگاه بیشتر به دومی پرداخته است. این سیستم سعی می‌کند به جای تولید الگوهای پرسش SPARQL، مسیرهایی با توجه به ساختار توپولوژیکی سوال زبان طبیعی در گراف داده ایجاد کند و بنا به بعضی قواعد به آنها امتیاز دهد و موجودیت‌های منتهی به مسیرهای با امتیاز بالاتر را به عنوان جواب انتخاب کند.

۳-٦- یادگیری ماشین

تعدادی از تحقیقات اخیر موضوع پرسش‌وپاسخ روی داده‌های پیوندی را به صورت یک مسئله یادگیری ماشین مورد توجه قرار می‌دهند و سعی می‌کنند با فهمیدن معنای دقیق یک سوال زبان طبیعی، آن را به یک فرم منطقی جهت تبدیل به یک پرسش ساخت‌یافته، درآورند [31]. به عنوان مثال می‌توان به [32] و [33] اشاره کرد که با داشتن پایگاه و جفت‌های سوال‌وجواب سعی در آموزش یک تجزیه‌کننده معنایی برای ماشین دارند. سیستم [34] نیز یک الگوریتم تحت‌نظارت استاندارد برای آموزش یک تجزیه‌کننده معنایی را با یک الگوریتم برای نگاشت عبارات سوال زبان طبیعی و مفاهیم آنتولوژی، ترکیب می‌کند. معمولا فرایند آموزش سیستم با برچسب زدن به داده‌ها فرایندی زمانبر است ولی دارای این مزیت است که معمولا به دقت بالایی منجر می-شود.

٤- مقایسه روش‌ها

همانطور که در مقاله در مورد سیستم‌های پرسش‌وپاسخ مختلف عنوان شد، هر دسته از روش‌ها از رویکرد متفاوتی برای رسیدن به پاسخ استفاده می‌کند. در جدول (۱) به طور خلاصه رویکرد به کار رفته در دسته‌های مختلف سیستم‌های پرسش‌وپاسخ روی داده‌های پیوندی و مزیت و عیب هر دسته ذکر شده است.

٥- نتیجه گیری

با توجه به سرعت افزایش دانش ساخت‌یافته معنایی در وب داده، نیاز به فراهم کردن روش‌هایی برای جستجوی این حجم زیاد اطلاعات ساخت‌یافته، بیش از پیش حس می‌شود. در سالیان اخیر تلاش زیادی برای ارائه واسط‌هایی مبتنی بر زبان طبیعی برای جستجوی داده‌های معنایی تحت مدل داده RDF انجام شده است. در این مقاله پس از بیان چالش‌ها در این حوزه، مروری بر سیستم‌های پرسش‌وپاسخ روی داده‌های پیوندی و همچنین نحوه عملکردشان و ابزارهای مورد استفاده در مراحل مختلف با توجه به دسته‌بندی این سیستم-ها ارائه شده است و مزایا و معایب آنها ذکر شده است.

جهت گیری کارهای آینده در این حوزه بیشتر به سمت حل چالش‌های اختلاف لغوی و ساختاری، ابهام و توزیع شدگی و چندزبانی و بیان سوالات پیچیده خواهد بود.

جدول (۱): مقایسه انواع سیستم‌ها در دسته‌های مختلف

دسته	رویکرد	مزیت	عیب
زبان‌های طبیعی محدود شده	استفاده از ساختارها و لغات آنتولوژی برای پرس‌وجو	×	عدم پاسخگویی به سوالاتی که لغاتشان در آنتولوژی موجود نیست
گرامرهای رسمی	تعریف نحو و معنا برای اجزای لغوی زبان و استفاده از معناشناسی ترکیبی	پرداختن به سوال با هر سطح از پیچیدگی	عدم پاسخگویی به سوالاتی که توسط گرامر قابل تولید نیستند
نگاشت ساختارهای زبان‌شناختی به ساختارهای معنایی	نگاشت ساختارهای سوال زبان طبیعی به سه‌تایی‌های معنایی	استواری در حالات مختلف	نیاز به تطابق یک‌به‌یک بین ساختار سوال و داده سه‌تایی
مبتنی بر الگو	استخراج ساختار پرس‌وجوی ساخت‌یافته از ساختار سوال و پرکردن جاهای خالی با URI مناسب	توانایی پاسخ به سوالات تجمعی	عدم پاسخگویی به سوالاتی که بین سوال و داده‌ها اختلاف ساختاری وجود دارد
اکتشاف گراف	پیدا کردن مسیرهایی در گراف داده که به پاسخ سوال منتهی می‌شوند	اطلاع از داده‌ها	هزینه بازیابی گراف بزرگ داده
یادگیری ماشین	استفاده از یادگیری ماشین جهت تبدیل سوال به فرم منطقی قابل تبدیل به پرس‌وجوی ساخت‌یافته	دقت بالا به قیمت آموزش سیستم	زمانبر بودن فرایند آموزش سیستم یا برچسب زدن به داده‌ها

مراجع

[1] T. Berners-Lee, J. Hendler, and O. Lassila, *"The Semantic Web,"* Sci. Am., vol. 284, no. 5, pp. 34–43, 2001.

[2] C. Bizer, T. Heath, and T. Berners-Lee, *"Linked Data - The Story So Far"* Int. J. Semant. Web Inf. Syst., vol. 5, pp. 1–22, 2009.

[3] C. Unger, A. Freitas, and P. Cimiano, *"An introduction to Question Answering over Linked Data,"* Reasoning on the Web in the Big Data Era, LNCS 8714, Springer International Publishing, pp. 100-140, 2014.

[4] P. Cimiano and U. Bielefeld, *"Is Question Answering fit for the Semantic Web ?: a Survey,"* Semantic web, vol. 2, no. 2, pp. 125–155, 2011.

[5] A. Freitas, E. Curry, J.G Oliveira, and S. O Riain,*"Querying Heterogeneous Datasets on the Linked Data Web: Challenges, Approaches, and Trends,"* IEEE Internet Computing, pp. 24–33, 2012.

[6] E. Kaufmann and A. Bernstein, *"Evaluating the Usability of Natural Language Query Languages and Interfaces to Semantic Web Knowledge Bases,"* J. Web Semantics: Science, Services, and Agents on the World Wide Web, vol. 8, 2010, pp. 377–393.

[7] O. Kolomiyets, and M.F Moens, *"A Survey on Question Answering Technology from an Information Retrieval Perspective,"* Information Sciences, 2011.

[8] D. Diefenbach, K. Singh, and P. Maret, *"Core Techniques of Ontology-Based Question Answering Systems: a Survey,"* Semantic Web Journal, pp. 1-20, 2015.

[9] K. Höffner, S. Walter, E. Marx, J. Lehmann, A. Ngonga, and R. Usbeck, *"Overcoming challenges of semantic question answering in the semantic web,"* Semantic Web Journal, pp. 1-21, 2015.

[10] A. Bernstein, E. Kaufmann, and C. Kaiser, *"Querying the semantic web with ginseng: A guided input natural language search engine,"* in 15th Workshop on Information Technologies and Systems, Las Vegas, pp. 112–126, 2005.

[11] P. Cimiano, P. Haase, J. Heizmann, M. Mantel, and R. Studer, *"Towards portable natural language interfaces to knowledge bases-The case of the ORAKEL system,"* Data Knowl. Eng., vol. 65, no. 2, pp. 325–354, 2008.

[12] C. Unger and P.Cimiano. *"Pythia: Compositional meaning construction for ontology-based question answering on the semantic web,"* In Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems, Springer, NLDB 2011, pp. 153-160, 2011.

[28] A. Freitas, J.G. Oliveira, S. O'Riain, E. Curry, and J.C. Pereira da Silva, *"Querying linked data using semantic relatedness: a vocabulary independent approach,"* In Data & Knowledge Engineering, pp. 126-141, 2013.

[29] N. Aggarwal, P. Buitelaar, *"A System Description of Natural Language Query over DBpedia,"* in: Proc. of Interacting with Linked Data (ILD 2012), pp. 96-99, (2012).

[30] C. Zhu, K. Ren, X. Liu, H. Wang, Y. Tian, and Y. Yu, *"A Graph Traversal Based Approach to Answer Non-Aggregation Questions Over DBpedia."* CoRR abs/1510.04780, (2015).

[31] K. Liu, J. Zhao, S. He, and Y. Zhang, *"Question Answering over knowledge Bases,"* In Natural Language Processing, pp. 26-35, 2015.

[32] J. Berant, A. Chou, R. Frostig and P. Liang. "Semantic parsing on Freebase from question-answer pairs," In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013.

[33] T. Kwiatkowski, E. Choi, Y. Artzi, and L. Zettlemoyer. *"Scaling semantic parsers with on-the-fly ontology matching."* In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1545-1556, 2013.

[34] Q. Cai, and A. Yates, *"Large-scale semantic parsing via schema matching and lexicon extension,"* In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2013.

[13] S. He, S. Liu, Y. Chen, G. Zhou, K. Liu and J. Zhao, *"Casia@qald-3: A question answering system over linked data,"* In Work. Multilingual Question Answering over Linked Data (QALD-3), pp. 2-9, 2013.

[14] G. Cristina, B. Valentina and B. Roberto, *"A hmm-based approach to question answering against linked data,"* In Work. Multilingual Question Answering over Linked Data (QALD- 3), pp. 1-13, 2013.

[15] S. Shekarpour, E. Marx, A-C.NgongaNgomo and S. Auer, *"Sina: Semantic interpretation of userqueries for question answering on interlinked data,"* Web Semantics: Science, Services and Agents on the World Wide Web, pp. 39–51, 2015.

[16] C. Dima, *"Intui2: A prototype system for question answering over linked data,"* Proceedings of the Question Answering over Linked Data lab (QALD-3) at CLEF, pp. 1-12, 2013.

[17] M. Yahya, K. Berberich, S. Elbassuoni, M.Ramanath, V. Tresp, and G.Weikum. *"Natural language questions for the web of data,"* In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 379–390, 2012.

[18] K.Xu, Y. Feng, and D. Zhao. *"Xser@ qald- 4: Answering natural language questions via phrasal semantic parsing."* In Conference and Labs of the Evaluation Forum (CLEF), 2014.

[19] S. He, Y. Zhang, K. Liu and J. Zhao, *"Casia@ v2: Amln-based question answering system over linked data,"* In Conference and Labs of the Evaluation Forum (CLEF), 2014.

[20] C. Dima, *"Answering natural language questions with intui3,"* In Conference and Labs of the Evaluation Forum (CLEF), 2014.

[21] G. A. Miller. *"WordNet: A lexical database for English,"* Communications of the ACM, 38(11):39–41, 1995.

[22] Lei Zou, Ruizhe Huang, Haixun Wang, JefferXu Yu, Wenqiang He, and Dongyan Zhao. *"Natural language question answering over rdf: a graph data driven approach,"* In Proceedings of the 2014 ACM SIGMOD international conference on Management of data, pages 313–324, 2014.

[23] R. Beaumont, B.Grau, and A.Ligozat, *"Semgraphqa@ qald-5: Limsi participation at qald-5@ clef. CLEF,"* In Conference and Labs of the Evaluation Forum (CLEF), 2015.

[24] S.Ruseti, A.Mirea, T.Rebedea and S.Trausan-Matu, *"Qanswer-enhanced entity matching for question answering over linked data,"* In Conference and Labs of the Evaluation Forum (CLEF), 2015.

[25] C. Unger, L. Bühmann, J. Lehmann, A.C. NgongaNgomo, D. Gerber and P. Cimiano, *"Template-based Question Answering over RDF data,"* In Proceedings of the 21st international conference on World Wide Web, pp. 639–648, 2012.

[26] S. Walter, C.Unger, P. Cimiano, and D. Bar, *"Evaluation of a layered approach to question answering over linkeddata,"* in Proc. of the 11th International Semantic WebConference (ISWC 2012), 2012.

[27] W. Zheng, L. Zou, X. Lian, J. Xu Yu, S. Song, and D. Zhao, *"How to Build Templates for RDF Question/Answering: An Uncertain Graph Similarity Join Approach,"* ACM SIGMOD International Conference on Management of Data, SIGMOD, pp. 1809-1824, 2015.