مجله علمی «نوآوری‌های فناوری اطلاعات و ارتباطات کاربردی»
سال اول، شماره ۴، فصل زمستان ۱۴۰۰؛ ص۳۳تا۴۳

بهره‌برداری از بازخورد ارتباط در جست‌وجوی نمودار دانش

زهرا مسعودی نژاد[1*] — دانشجوی دکتری
حسن نادری[2] — دانشیار دانشگاه علم و صنعت
عبدالباقی قادرزاده[3] — استادیار دانشکده مهندسی کامپیوتر،دانشگاه ازاد اسلامی، سنندج، ایران

(دریافت: ۱۴۰۰/۰۸/۰۱، پذیرش:۱۴۰۱/۰۸/۰۹)

چکیده

نیاز به نرم‌افزار و یا به‌صورت کلی سیستمی که بتواند به‌صورت روزانه اطلاعات مهمی (که فرد احساس می‌کند ممکن است در آینده به آنها نیاز پیدا کند) را در خود ذخیره کند و در زمان لازم این اطلاعات را در اختیار فرد قرار دهد، لازم و ضروری به نظر می‌رسد. سعی بر این است تا در این مقاله با ارائه سیستمی به انسان کمک شود تا با ذخیره کامل اطلاعات در گذشت زمان بتواند در هر زمانی که نیاز به اطلاعاتی داشت (که به دلیل مشغله‌های ذهنی قادر به یادآوری آنها نیست) بتواند با یادآوری بخشی از اطلاعات موردنیاز خود به‌تمامی اطلاعات خود به‌صورت کامل دست یابد. در سیستم پیشنهادی اطلاعات در ابتدا در گراف دانشی ذخیره می‌شود، این گراف دانش دارای ارتباطاتی بین نودها است که به دلیل وجود همین ارتباط‌ها و با استفاده از روش‌های جست‌وجوی کلیدی بهترین و کامل‌ترین پاسخ به کاربر برگردانده می‌شود. کاربر به‌صورت روزمره اطلاعات مهم خود را که می‌شنود و یا می‌خواند و گمان می‌کند ممکن است در آینده به آنها نیاز پیدا کند را به‌صورت کامل در این سیستم ذخیره می‌کند و در زمان لازم فقط با یادآوری بخشی از اطلاعات به کل آن اطلاعات دسترسی پیدا می‌کند. پس از ذخیره داده‌ها در گراف دانش (که دارای ارتباط‌های خاصی بین نودها است) ذخیره می‌شود و با روش‌های جست‌وجوی کلیدی اطلاعات را بازیابی می‌کند. روش ارائه شده کمک بهینه‌ای به بازیابی اطلاعات ازدست‌رفته انسان خواهد کرد.

کلیدواژه‌ها: نمودار دانش، بازخورد ارتباط، حافظه انسان

Utilize communication feedback in search of knowledge charts

Z.Masoudinejhad[1]*, H.Naderi[2] , A.B. Ghaderzadeh[3]
department of computer engineering, Islamic azad university of sanandaj, sanandaj, iran, Masoudinezhad.z@iaud.ac.ir
department of computer- university of science & Technology,Iran,Tehran, naderi@iust.ac.ir
department of computer, Islamic azad university of sanandaj, Sanandaj, Iran
(Received: 2021/October/23; Accepted: 2022/July/31)

Abstract

The need for a software or a system in general that can store important information (which a person feels may be needed in the future) on a daily basis and provide this information to the person at the necessary time, It seems necessary and essential. In this article, an attempt is made to help humans by providing a system so that by fully storing information over time, whenever they need information (which they are unable to recall due to mental preoccupations), they can recall a part of it,Get all the information you need in full. In the proposed system, information is initially stored in a knowledge graph, this knowledge graph has connections between nodes, which return the best and most complete answer to the user due to the existence of these connections and using key search methods to be. On a daily basis, the user stores his important information, which he hears or reads and thinks he may need in the future, in this system, and when necessary, just by reminding him. A part of the information gets access to the whole information. On a daily basis, the user stores his important information, which he hears or reads and thinks he may need in the future, in this system, and when necessary, just by reminding him. A part of the information gets access to the whole information.

Keywords : knowledge graph, feedback communication, human memory

*Corresponding Author E-mail: Masoudinezhad.z@iaud.ac.ir

۱. مقدمه

با توجه به اینکه افراد در گذر زمان اطلاعاتی را کـه برایشـان اهمیت دارد و ممکن است در آینـده بـه آنهـا نیـاز پیـدا کننـد (درگذشته به ذهن سپرده‌اند) ممکن است بـه فراموشـی بسـپارند، نیاز به سیستمی که بتواند در حل این مشکل به آنها کمـک کنـد لازم و ضروری است.

سیستم پیشنهادی به این صورت اسـت کـه اطلاعـات کـاربر در گراف دانشی (که ساختار آن به‌گونه‌ای است که اطلاعات کاربران در آن ذخیره شده و با استفاده از ارتباطاتی که در این گراف بـین اطلاعات ذخیره شده وجود دارد) تعریف می‌شـود و همچنـین بـا استفاده از روش‌های جستجو بهترین و کامل‌ترین پاسخ به کـاربر برگردانده می‌شود. سیستم پیشنهادی ما در این تحقیق با عنـوان knowledge GCF (Knowledge graph with effective communication feedback) معرفی می‌شود و از اینجا بـا نـام knowledge GCF صدا زده می‌شود. گراف‌هـای دانـش در وب معنایی به‌صورت نـوعی، از داده‌هـای لینـک شـده اسـتفاده مـی-کنند[۱].

داده‌های گراف جستجو از قبیل اطلاعات، اجتماع و گراف‌های دانش همیشه یک مورد چالش‌برانگیز بوده است. از یک طرف، بـا توجه به طرح‌های پیچیده و توصیف‌های مختلف اطلاعـات، بـرای کاربران ساخت ساختارهای پرس‌وجویی مانند SPARQL (بدون صرف ساعت‌ها جهت درک طرح) بسیار مشکل است[۲]. از طرف دیگـر، تکنیـک‌هـای جسـت‌وجـوی غیرسـاخت‌یافتـه ماننـد جست‌وجوی کلیدی به‌اندازه کافی بیانگر کشـف سـاختار داده‌هـا تاحدامکان واضح و رسا نیستند. گراف‌های دانش شامل منبع غنی از اطلاعات که شامل گره‌ها به‌عنوان موجودیت‌ها و لبه‌ها به‌عنوان روابط بین موجودیـت‌هـا، مـی‌باشـد. در سـال‌هـای اخیـر شـاهد شـکوفایی گـراف‌هـای دانـش در مقیـاس بـزرگ از قبیـل DBpedia [۳]، Freebase [۴]، Google's Knowledge [۵] Graph و YAGO[۶] هستیم. فرم‌هـای پـرس‌وجـوی دیگـری برای گراف‌های دانش وجود دارد که می‌توانند آنهـا را بـه پـرس-وجوی گراف تبدیل کننـد، از قبیـل logic query [۷] (پـرس-وجوی منطقـی)، natural language Query[۸] (پـرس‌وجـوی زبـان طبیعـی) و exemplary query[۹ و ۱۰] (پـرس‌وجـوی نمونه‌ای). پرس‌وجوی گراف می‌تواند به‌عنـوان پـرس‌وجـوی حـد واسط برای رسیدن به پاسخ کاربر ایفای نقش کند.

با توجه به اینکه اطلاعاتی که روزمره با آنها سروکار داریم (از قبیل نام افراد، محل زندگی یک دوسـت، نویسـنده یـک کتـاب و مانند آن) و معمولاً این اطلاعات را به ذهن می‌سپاریم، باگذشـت زمان و هنگامی که به این اطلاعات نیاز داریم، بـه دلیـل مشـغله-های فراوان ذهنی، ممکن است که اطلاعات را به‌صـورت کامـل و با تمام جزئیات به‌خاطر نیاوریم. در این مقاله سعی بر ایـن اسـت که با استفاده از گراف دانش و ذخیره اطلاعات در این گراف و بـا استفاده از روش‌های جست‌وجوی کلیدی، فقط با واردکردن چنـد کلمه از اطلاعاتی که ذخیره شده به‌تمامی اطلاعات ذخیـره شـده دست پیدا کنیم. اطلاعات در ابتدا در گراف دانشـی ذخیـره مـی-شود، این گراف دانش دارای ارتباطاتی بـین نودهـا اسـت کـه بـه دلیل وجود همین ارتباط‌ها و با استفاده از روش‌های جست‌وجوی کلیدی بهترین و کامل‌ترین پاسخ به کاربر برگردانده می‌شـود. در واقع استفاده از این سیستم به این صورت است که کاربر به‌صورت روزمره اطلاعات مهم خود را که می‌شنود و یا می‌خوانـد، و گمـان می‌کند ممکن است در آینده به آنها نیاز پیدا کنـد را بـه‌صـورت کامل در این سیستم ذخیـره مـی‌کنـد و در زمـان لازم فقـط بـا یادآوری بخشی از اطلاعات به کل آن اطلاعات دسترسی پیدا می کند. از طرف دیگر پس از ذخیره داده‌ها توسط کاربر اطلاعـات در سیستم در گراف دانش که ساختار آن به‌گونه‌ای اسـت کـه دارای ارتباط‌های خاصی بین نودها است، ذخیره می‌شود که با اسـتفاده از روش‌های جست‌وجوی کلیدی اطلاعات را بـه‌صـورت کامـل در اختیار کاربر قرار می‌دهد. به‌صورت مختصر مخزنی از اطلاعات که بـه آن لقـب گـراف دانـش داده مـی‌شـود، بـه‌تـدریج در حـال اضافه‌کردن مفاهیم و جزئیات سودمند به جدولی از لینک‌هاسـت. از این فناوری می‌توان برای جستجو درباره افراد خاص و یـا همـه آن چیزهایی که شما به دنبال آن هسـتید، اسـتفاده کـرد. گـراف دانش به‌عنوان یک پایگـاه‌داده بـزرگ مـی‌توانـد بـه کـاربر اجـازه مـرتبط کـردن مطالـب و مفـاهیم مربـوط بـه افـراد، آدرس‌هـا و موضوعات ذخیره شده دیگـر را بـا هـم بدهـد. در واقـع کـاربر بـا واردکردن اطلاعات خود به‌تدریج و ایجاد گراف دانش و همچنـین ایجاد جدولی از لینک‌ها بین مطالب ذخیـره شـده ایـن امکـان را فراهم می‌کند تا در آینده با استفاده از روش‌هـای مختلـف و کـارا که جهت جست‌وجوی اطلاعـات وجـود دارد، اطلاعـات موردنیـاز خود را که امکان فراموشی در حافظه بلندمدت داشـت را بتوانـد بازیابی کند. هر چه روش ذخیره اطلاعات و ایجاد گراف دانـش و لینک بین داده‌ها از کارایی بالاتری برخوردار باشـد، قطعـاً گـراف دانش پربارتر و منسجم‌تری خواهیم داشت، همچنین طراحی یک روش جستجو با کارایی بالا نیز می‌تواند در بازیابی اطلاعات بهتـر کمک قابل توجهی به کاربر کند. در حال حاضر در نـرم‌افزارهـای جستجو، جستجو بیشتر بر اساس کلمات و عبارات است و نه معنا و ارتباط بین واژه‌ها. با توجه به اینکه انسان نیاز دارد در هر زمـان اطلاعات مهم خود را در جایی ذخیره کند و سپس در زمـانی کـه به آنها نیاز دارد و به فراموشی سپرده شده است، فقط با یـادآوری کلمات کلیدی آن اطلاعـات بـه‌تمـامی اطلاعـات موردنیـاز خـود به‌صورت کامل دست یابد.

*Corresponding Author E-mail: Masoudinezhad.z@iaud.ac.ir

در این مقاله اطلاعات در گراف دانش ذخیره شده و سپس با استفاده از روش‌های جست‌وجوی کلیدی و همچنین ارتباط بین آنها در گراف دانش جست‌وجوی کامل‌تر و بهتری را بر روی دانش نهفته در حافظه کاربر به وی خواهیم داد.

ساختار مقاله به این صورت است که در ابتدا پیشینه تحقیق مورد بررسی قرار می‌گیرد، سپس در مرحله بعد روش پیشنهادی ارائه می‌گردد و در انتها ارزیابی روش پیشنهادی را خواهیم داشت.

۲. پیشینه تحقیق

در ادامه در این بخش به پیشینهٔ[s] این تحقیق می‌پردازیم و چند نمونه گراف دانش را بررسی می‌کنیم.

۲-۱ Google's Knowledge Graph (گراف دانش گوگل)

گراف دانش گوگل به‌صورت عمومی در سال ۲۰۱۲ معرفی شد و این در حالی بود که واژه گراف دانش در حال پدید آمدن بود. اطلاعات گوگل در مورد ساخت گراف دانش به‌صورت نسبتاً محرمانه است، تنها منابع خارجی اندکی وجود دارد که بعضی از مکانیزم‌های جریان اطلاعات درون گراف دانش بر پایه اطلاعات را مورد تجزیه و تحلیل قرار می‌دهد. با توجه به این مطالب، می‌توان فرض کرد که منابع اینترنتی نیمه ساختیافته اصلی، از قبیل ویکی پدیا، و همچنین ساختارهای نشانه‌گذاری شده (مانند schema.org Microdata[۱۱]) بر روی صفحات وب و محتوای شبکه‌های اجتماعی آنلاین گوگل پلاس، در ساختار گراف دانش مشارکت دارند. بر این اساس[۳]، گراف دانش گوگل حاوی ۱۸ میلیون وضعیت در مورد ۵۷۰ میلیون موجودیت و همچنین طرح آن شامل ۱۵۰۰ نوع موجودیت و ۳۵۰۰۰ نوع رابطه است.

۲-۲ Google's Knowledge Vault (گراف دانش وات)

جهش دانش پروژه دیگری است که توسط گوگل انجام شد. در این پروژه دانش از منابع متفاوتی از قبیل سندهای متنی، جداول HTML و یادداشت‌های ساخت‌یافته در وب با Microdata و MicroFormats، استخراج می‌شود. حقایق استخراج شده با استفاده از دو مقدار مطمئن استخراج‌کننده و همچنین احتمال-های قبلی برای وضعیت‌ها که با استفاده از گراف دانش Freebase محاسبه می‌شوند، ترکیب می‌شوند. با توجه به این اجزا، یک مقدار مطمئن برای هر حقیقت محاسبه می‌شود و فقط حقایق مطمئن به Knowledge Vault وارد می‌شوند. بر همین اساس[۱۲]، Knowledge Vault حاوی حدوداً ۴۵ میلیون موجودیت و ۲۷۱ میلیون وضعیت می‌باشد و در حال استفاده از ۱۱۱۰ نوع موجودیت و ۴۵۰۰ نوع رابطه می‌باشد.

۲-۳ Yahoo! 's Knowledge Graph (گراف دانش یاهو)

همانند گوگل، یاهو نیز دارای گراف دانش مربوط به خود است که برای جست‌وجوی اطلاعات از آن استفاده می‌شود. گراف دانش بر روی دو منبع داده عمومی (به‌عنوان‌مثال ویکی‌پدیا و Freebase)، و همچنین منابع تجاری در حوزه‌های مختلف، ساخته می‌شود. این گراف به‌عنوان پایه و اساس برای منابع مختلف استفاده می‌شود و نمایش‌دهنده منابع تکاملی از قبیل ویکی‌پدیا، برای بروز رسانی‌های ثابت است. گراف دانش یاهو شامل حدوداً ۳,۵ میلیون موجودیت و ۱,۴ بیلیون رابطه است. طرح آن که با schema.org هماهنگ است، حاوی ۲۵۰ نوع موجودیت و ۸۰۰ نوع رابطه است[۱۳].

۲-۴ Microsoft's Satori

Satori، معادل مایکروسافت به گراف دانش گوگل است. اگرچه تقریباً هیچ اطلاعات عمومی بر روی ساختار آن وجود ندارد، ولی طرح آن و یا نسخه داده‌ای Satori موجود است. این گراف شامل ۳۰۰ میلیون موجودیت و ۸۰۰ میلیون رابطه در سال ۲۰۱۲ می‌باشد و فرمت نمایش داده‌های آن به‌صورت RDF است .

۲-۵ Facebook's Entities Graph (گراف موجودیت فیس‌بوک)

اگرچه اکثریت داده‌ها در شبکه اجتماعی آنلاین فیس‌بوک، به‌صورت ارتباط بین افراد مشاهده می‌شود، فیس‌بوک بر روی استخراج گراف دانش که شامل انواع مختلفی از اشخاص است، نیز کار می‌کند. اطلاعات افراد به‌صورت اطلاعات شخصی تهیه می‌شود (به‌عنوان‌مثال شهر محل زندگی آنها، مدرسه‌ای که در آن درس‌خوانده)، و همچنین علاقه‌مندی‌های آنها (فیلم‌ها، کتاب‌ها، سرگرمی‌ها و...) تهیه می‌شود. غالباً موجودیت‌هایی نمایش داده می‌شوند که می‌توانند ارتباط افراد با دیگر افراد را نشان دهند. اگرچه ارقام عمومی زیادی درباره موجودیت‌های گراف فیس‌بوک وجود ندارد، ولی گفته شده است که این گراف حاوی بیش از ۱۰۰ میلیارد ارتباط بین اشخاص است.

جدول (۱) به‌صورت خلاصه، خصوصیات گراف‌های دانش را نشان می‌دهد[۱۴]. می‌توان مشاهده کرد که گراف‌ها در اندازه‌گیری-های پایه از قبیل تعداد موجودیت‌ها و روابط و همچنین اندازه طرحی که انها استفاده کرده اند (به عنوان مثال کلاسها و رابطه-ها)، متفاوت هستند. علاوه بر این تفاوت‌ها، می‌توان به این نتیجه

رسید که گراف‌های دانش تفاوت زیادی در خصوصیات دیگر از جمله میانگین درجه گره‌ها، چگالی و یا اتصالات دارند. در این جدول، نمونه ها بیانگر تعداد نمونه ها و یا مفاهیم A-box تعریف شده در گراف است. Facts ، بیانگر تعداد وضعیت‌ها در مورد این نمونه‌ها و یا کلاسهای تعریف شده در طرح می‌باشد. types، نشان دهنده تعداد نوع‌های متفاوت کلاسهای تعریف شده در طرح و در نهایت روابط بیانگر تعداد روابط مختلف تعریف شده در طرح است.

جست‌وجوی کلیدواژه یک رابط ساده اما کاربرپسند برای بازیابی اطلاعات از ساختارهای داده پیچیده فراهم می‌کند. ازآنجایی‌که بسیاری از مجموعه داده‌های زندگی واقعی توسط اشکال و نمودارها نشان داده می‌شوند، جست‌وجوی کلیدواژه به یک مکانیسم جذاب برای داده‌های انواع مختلف تبدیل شده است. این جستجو، مکانیسم بازیابی اطلاعات برای داده‌های موجود در وب است. همچنین ثابت می‌کند که یک مکانیسم مؤثر برای یک داده نیمه‌ساختاریافته و ساختاریافته، به‌خاطر رابط کاربرپسند آن است [۱۵]. اخیراً، پردازش و پرس‌وجو بر روی داده‌های ساختاریافته گراف، توجه فزاینده‌ای را به خود جلب کرده است، چرا که بسیاری از کاربردها توسط گراف ساختاریافته و تولید اطلاعات آن هدایت می‌شوند. جستجو در داده‌های XML مسئله‌ای ساده‌تر از گراف‌های بدون طرح‌واره آزاد است. آن‌ها اساساً محدود به ساختارهای گرافی هستند که در آن هر نود تنها یک مسیر ورودی منفرد دارد. این ویژگی فرصت‌های بهینه‌سازی زیادی را فراهم می‌کند. اطلاعات اتصال نیز می‌توانند به طور مؤثری کدگذاری شده و نمایه‌گذاری شده باشند. به‌عنوان‌مثال، در XRank، فهرست معکوس "دیویی" به‌عنوان مسیرهای شاخص استفاده می‌شود به‌طوری‌که یک موج رادیویی می‌تواند بدون پیمایش نمودار ارزیابی شود. جستجو در پایگاه‌های اطلاعاتی رابطه‌ای توجه زیادی را به خود جلب کرده است. از نظر مفهومی، یک پایگاه‌داده به‌عنوان یک گراف برچسب‌دار در نظر گرفته می‌شود که در آن مجموعه داده‌ها در جدول‌های مختلف به‌عنوان گره‌های متصل شده از طریق روابط کلیدی خارجی در نظر گرفته می‌شوند. توجه داشته باشید که نمودار ساخت این روش معمولاً ساختار منظمی دارد؛ چون الگو ارتباطات گره را محدود می‌کند. جدا از رویکرد جست‌وجوی گراف در مخزن‌های اطلاعات، ساختارهای DBXplor و DISCOVER به حالات معین ملحق شده و آن‌ها را ارزیابی می‌کنند و به‌شدت بر الگوی پایگاه‌داده و تکنیک‌های پردازش جستجو در RDBMS تکیه می‌کنند. تعداد زیادی تحقیق بر روی جست‌وجوی کلیدواژه‌های نیمه‌ساختاریافته و ساختاریافته در مورد جست‌وجوی گراف در سال‌های اخیر مطرح شده است. الگوریتم‌های شناخته‌شده شامل جست‌وجوی توسعه وارونه، جست‌وجوی دوسویه، تکنیک‌های برنامه‌نویسی پویا DPBF و BLINKS هستند. اخیراً، تحقیقی که جست‌وجوی کلیدواژه‌ای را برای گراف‌ها بر روی حافظه خارجی گسترش می‌دهد، پیشنهاد شده است. به دلیل ساختار اساسی گراف، جست‌وجوی کلیدواژه بر روی داده‌های گراف بسیار پیچیده‌تر از جست‌وجوی کلیدواژه بر روی اسناد است. چالش‌ها سه جنبه دارند، به این ترتیب که چگونه می‌توان معنای جست‌وجوی شهودی را برای جست‌وجوی کلمات کلیدی در مورد گراف‌ها، چگونگی طراحی استراتژی‌های رتبه‌بندی معنی‌دار برای پاسخ‌ها، و چگونگی طراحی الگوریتم‌های کارآمد که معناشناسی و استراتژی‌های رتبه‌بندی را اجرا می‌کنند را طراحی کرد. چالش‌های باقی‌مانده در حوزه جست‌وجوی کلیدواژه بر روی گراف وجود دارد. یک حوزه که از اهمیت ویژه‌ای برخوردار است، چگونگی ارائه یک موتور جست‌وجوی معنایی برای داده‌های گراف است. این گراف بهترین نمایشی است که ما برای اطلاعات پیچیده؛ مانند دانش انسانی، پویایی فرهنگی و فرهنگی و غیره داریم. برای مثال، برخی از تحقیقات اخیر، ناگا و همکارانش به امکان ایجاد یک موتور جست‌وجوی معنایی پرداخته‌اند. بااین‌حال، تحقیق ناگا مبتنی بر کلمات کلیدی نیست که پیچیدگی را برای شکل‌دهی به یک موج ارسالی مطرح می‌کند. چالش مهم دیگر این است که اندازه گراف به طور قابل‌توجهی از حافظه بزرگ‌تر است. بسیاری از الگوریتم‌های جست‌وجوی کلمات کلیدی گراف مبتنی بر حافظه هستند که به این معنی است که آن‌ها نمی‌توانند گراف‌ها را مثل ویکی‌پدیا انگلیسی که بیش از ۳۰ میلیون رابط دارد را کنترل کنند. استخراج اطلاعات مبتنی بر عبارت ClausIE36F[۱۶] ، از دانش زبان‌شناسی درباره گرامر زبان انگلیسی برای تشخیص عبارات در جمله ورودی بهره می‌برد و سپس نوع هر عبارت را بر طبق تابع گرامری مشخص می‌کند. براین‌اساس،ClausIE می‌تواند استخراج بادقت بالا را انجام دهد. سیستم ClausIE مبتنی بر تجزیه وابستگی و یک مجموعه کوچک لغوی مستقل از دامنه مانند افعال copular، جمله به جمله بدون هیچ پس‌پردازشی (برای حذف استخراج‌ها بادقت پایین) عمل می‌کند و به هیچ داده آموزشی نیاز ندارد و دارای دقت و فراخوانی بالایی نسبت به سیستم‌های دیگر می‌باشد.

بعد از تشخیص عبارات، یک یا چند گزاره از هر عبارت با توجه به نوع آن، استخراج می‌شود. خطاهای استخراج در ClausIE به علت درخت‌های تجزیه نادرست است. به طور چشمگیری آهسته‌تر از همه تکنیک‌های بالا می‌باشد؛ اما استخراج‌هایی باکیفیت بالا تولید می‌کند که حتی می‌تواند به‌عنوان داده آموزشی برای سیستم‌هایی مانند R2A2 مورد استفاده قرار گیرد.

در ClausIE نوع عبارات و افعال از نظر لازم و متعدی و یا افعال ترکیبی و کمکی و... بررسی می‌شود. اطلاعات در مورد عبارت از طریق تجزیه وابستگی، و اطلاعات دو مورد انواع فعل از یک مجموعه کوچک لغوی مستقل از دامنه فراهم می‌شود. بدین‌وسیله با اطلاعات به دست آمده، می‌توان نوع عبارت را مشخص کرد. چالش‌های این سیستم وابستگی به گرامر و الگوهای مشخص برای عبارات زبان انگلیسی، استخراج موارد غیرحقیقی (واقعیت‌های اشتباه)، هزینه زمانی بالای آن می‌باشد.

سیستم ReVerb از محدودیت‌های نحوی و محدودیت‌های لغوی استفاده کرده است. بر اساس محدودیت‌های نحوی، عبارات رابطه‌ای باید با الگوی زیر (الگوی برچسب POS) مطابق باشد. این الگو، عبارات رابطه‌ای را محدود به عبارت فعلی ساده یا عبارت فعلی به همراه حرف‌اضافه یا فعل به همراه یک یا چند اسم، صفت و مختوم به حرف‌اضافه می‌کند. اگر چندین تطبیق در جمله امکان داشته باشد، بلندترین تطابق انتخاب می‌شود [22].

V | VP | VWP
V= verb particle? Adv?
W= (noun | adj | adv | pron | det)
P = (prep | particle | inf.marker)

در جدول (۱) انواع گراف‌های دانش ذکر شده در بالا و چند گراف دانش دیگر و خصوصیات آنها به‌صورت خلاصه بیان شده است.

جدول (۱): خصوصیات گراف های دانش [۱۷]

نام گراف	نمونه‌ها	تعداد fact	تعداد type	رابطه
DBpedia(English)	۴۸۰۶۱۵۰	۱۷۶۰۴۳۱۲۹	۷۳۵	۲۸۱۳
YAGO	۴۵۹۵۹۰۶	۲۵۹۴۶۸۷۰	۴۸۸۴۶۹	۷۷
Freebase	۴۹۹۴۷۸۴۵	۳۰۴۱۷۲۲۶۳۵	۲۶۵۰۷	۳۷۷۸۱
Wikidata	۱۵۶۰۲۰۶۰	۶۵۹۹۳۷۹۷	۲۳۱۵۷	۱۶۷۳
NELL	۲۰۰۶۸۹۶	۴۳۲۸۴۵	۲۸۵	۴۲۵
OpenCyc	۱۱۸۴۹۹	۲۴۱۳۸۹۴	۴۵۱۵۳	۱۸۵۲۶
Google's Knowledge Graph	۵۷۰۰۰۰۰۰۰	۱۸۰۰۰۰۰۰۰۰	۱۵۰۰	۳۵۰۰۰
Google's Knowledge Vault	۴۵۰۰۰۰۰۰	۲۷۱۰۰۰۰۰۰	۱۱۰۰	۴۴۶۹
Yahoo! Knowledge Graph	۳۴۴۳۷۴۳	۱۳۹۱۰۵۴۹۹۰	۲۵۰	۸۰۰

برای کامل‌کردن دانش در گراف دانش دو روش اصلی وجود دارد: یکی این است که اطلاعات دانش به‌صورت دستی ایجاد شود و راه دیگر، تکمیل‌کردن به‌وسیله یادگیری نمایش دانش[۱] است. کامل‌کردن گراف دانش به‌صورت ساختگی (دستی) بسیار کند و زمان بر است و همچنین هنگام رویارویی با داده‌های مقیاس بزرگ و گسترش داده‌ها بسیار دشوار می‌شود. بنابراین، برای کامل‌کردن دانش در مقیاس بزرگ، افراد تمایل دارند تا از یادگیری نمایش دانش برای تکمیل‌کردن دانش و ایجاد گراف دانش استفاده کنند. مدل‌های نمایش داده سنتی، از قبیل TransE و RESCAL و SME و LFM وجود دارند که پایه و اساس آنها وجود دانش در گراف دانش است.

با توجه با اینکه دانش به‌صورت گسترده و روزبه‌روز در زمینه‌های حرفه‌ای و تخصصی با سرعت در حال ظهور است و افراد می‌توانند با سرعت دانش مورد نیاز خود را بر اساس نیازشان شناسایی کنند؛ ولی ماشین‌آلات نمی‌توانند. مشکلات بسیاری در سازماندهی و استفاده از گراف‌های سنتی وجود دارد از قبیل نادقیق بودن نمایش گراف‌ها که دستیابی به آن را دشوار می‌کند. ساختار آشفته و به‌هم‌ریخته دانش‌ها در گراف دانش باعث می‌شود که پرس‌وجوها بسیار کند و زمان‌بر باشند. بنابراین باید به دنبال راهی بود تا بتواند دانش‌ها را به‌صورت منظم و ساخت‌یافته در گراف دانش جای دهد تا بتوانیم در زمان کمتر به پاسخ پرس‌وجوهای موردنظر دست پیدا کنیم.

۳. روش پیشنهادی

روش پیشنهادی از چارچوبی شبیه به چارچوب RDF استفاده می‌کند که اطلاعات را در گراف دانش ذخیره کرده و سپس با ذخیره‌سازی اطلاعات نودها در جداول بانک اطلاعات، امکان بازیابی و جستجو را فراهم می‌کند. سیستم پیشنهادی دارای سه بخش استخراج ساختار، ذخیره‌سازی داده‌ها و بازیابی داده‌ها است. این بخش‌ها دارای وظایف معینی هستند که در ادامه هر یک به تفکیک شرح داده می‌شود.

بخش استخراج ساختار: در این بخش مجموعه‌دادهٔ مورد نظر مورد پردازش قرار گرفته و ساختار موجودیت‌ها از آن استخراج می‌شوند. منظور از ساختار استخراجی از موجودیت‌ها، الگوهایی از

[۱] Knowledge representation learning

صفت‌هایی است که موجودیت‌ها توسط آنها توصیف شده‌اند.
بخش ذخیره‌سازی: در این بخش داده‌ها مورد پردازش قرار می‌گیرد. از هر سند در هر لحظه یک موجودیت و الگوی صفاتش استخراج می‌شود. سپس با توجه به الگوی مورد نظر، موجودیت استخراجی در جدول مناسب ذخیره می‌شود.
بخش بازیابی: در این بخش ابتدا از پرس‌وجوی مورد نظر، الگوی صفت مورد پرس‌وجو استخراج و سپس با توجه به الگوی مورد نظر جداولی که آن الگو را شامل می‌شوند، اطلاعات خود را در نتایج پرس‌وجو نمایش می‌دهند.
روش پیشنهادی با استفاده از viewها توانسته است اجرای پرس‌وجوها را به طور میانگین سریع‌تر نماید. چرا که برای جست‌وجوی هر الگوی مورد پرس‌وجو به جدولی متناسب با آن رجوع می‌کند.
بنابراین سیستم می‌تواند به کاربر در اجرای پرس‌وجوها و یا تعریف موجودیت‌های جدید روی مجموعه داده مورد نظر کمک کند، و بدین‌وسیله سبب بهبود عملکرد انسان شود.

۱-۳) ایجاد knowledge GCF

در این مقاله سعی شده است که از چارچوبی شبیه به چارچوب RDF استفاده شده است .RDF چارچوبی است که امکان قابل‌استفاده بودن، قابل کشف بودن و دسترسی آسان را به داده‌ها می‌دهد و از همه مهم‌تر قابلیت تجمیع از چندین منابع ناهمگون را داراست. RDF امکان ایجاد پیوندها و روابط معنادار را بین مفاهیم ایجاد می‌کند. حافظه انسان روزانه حجم بسیار زیادی از اطلاعات را در خود جای می‌دهد بنابراین این حجم بسیار زیاد از داده‌ها نیاز به یک روش مقیاس‌پذیر، جهت شاخص‌گذاری RDF دارد. یک سیستم شاخص‌گذاری RDF از دو بخش ذخیره و بازیابی داده تشکیل می‌شود. با توجه به حجم روزافزون داده‌های RDF، مهم‌ترین انتظاری که از یک سیستم شاخص‌گذاری می‌رود مقیاس‌پذیری آن، هم در ذخیره و هم در بازیابی داده‌ها است. این سیستم‌ها باید در عین ذخیرهٔ داده‌های حجیم RDF، با ارائه یک واسط کاربری مناسب برای پرس‌وجوهای پیشرفته، امکان دسترسی به چندین منبع داده را به کاربران در کم‌ترین زمان ممکن بدهند. بنابراین دو چالش اساسی در پیاده سازی یک سیستم مقیاس‌پذیرِ شاخص‌گذاری RDF مطرح می‌شود. چالش اول، انتخاب نوع سیستمی است که برای ذخیره و مدیریت شاخص استفاده می‌شود که در این مقاله از گراف دانش استفاده شده است و چالش دوم، انتخاب نوع شما (مدل داده) که برای ذخیره و بازیابی داده‌های RDF به کار گرفته می‌شود که برای روبرویی با این مسئله نیز از شمای مبتنی برگراف استفاده شد که در ادامه با جزئیات بیشتری شرح داده می‌شود.

یک سیستم شاخص‌گذاری از دو بخش ذخیره و بازیابی داده تشکیل می‌شود. با توجه به این مسئله سیستم پیشنهادی در بخش ذخیرهٔ داده‌ها در دو مرحله کار خود را انجام می‌دهد. بنابراین معماری پیشنهادی از سه بخش اصلی استخراج ساختار داده، ذخیره‌سازی و بازیابی داده‌ها تشکیل می‌شود. در ادامه این سه بخش و اجزای درون آن شرح داده می‌شود.

بخش ذخیرهٔ داده، در پی یافتن ساختار، در بین داده‌های ورودی به حافظه است. ساختار استخراجی در نهایت، شمای ذخیره‌سازی داده‌ها را تشکیل می‌دهد.

قبل از شروع ساختن نمودارهای دانش، مهم است که بدانیم چگونه اطلاعات یادانش در این نمودارها جاسازی شده است.

اولین قدم برای ایجاد نمودار دانش تقسیم سند متنی به جملات است. سپس، فقط آن جملاتی را که در آنها دقیقاً ۱ موضوع و ۱ شی وجود دارد ، کوتاه خواهیم کرد. به عنوان مثال، متن نمونه زیر :

" علی احمدی، بازیکن ایران، به جایگاه برتری رسید و در آخرین رده‌بندی انفرادی مردان، بالاتر از ۱۲۹ تیم برتر حرفه ای قرار گرفت. علی مسابقات مالزی جونیور را برد. او اولین بازی خود را در برابر جوزف از انگلیس انجام داد. ایران ست اول را برد. "

پاراگراف فوق را به جملات تقسیم می‌کنیم:

- علی احمدی، بازیکن ایران، در آخرین رده‌بندی انفرادی مردان، بالاتر از ۱۲۹ نفر حرفه ای قرار گرفت.
- علی مسابقات مالزی جونیور را برد.
- او اولین بازی خود را در برابر جوزف از انگلیس انجام داد.
- ایران ست اول را برد.

از میان این چهار جمله، جمله‌های دوم و چهارم را کوتاه می‌کنیم؛ زیرا هر یک از آنها دارای ۱ فاعل و ۱ مفعول است. در جمله دوم، "علی" فاعل و مفعول "مسابقات " است. در جمله چهارم ، فاعل "ایران" است و "ست اول" مفعول است: .ماشین باید فاعل و مفعول را در جملات چند کلمه ای بشناسد
یک «سیستم مبتنی بر پایگاه دانش» از دو زیرسیستم تشکیل شده است: ۱- پایگاه دانش که نشان‌دهندهٔ واقعیت‌هایی در مورد جهان می‌باشند، و ۲- یک ماشین استنتاج که در مورد این واقعیت‌ها استنتاج می‌کنند و از قواعد برای استنتاج واقعیت‌های جدید استفاده می‌کند.
هدف از پایگاه‌داده، ذخیرهٔ داده‌های بزرگ به‌صورت جداول داده‌ای می‌باشد. ۲ نیازمندی اصلی برای ساخت یک پایگاه‌داده، ۱-

پشتیبانی از چندین کاربر توزیع شده برای دسترسی به یک داده (multiple, distributed user)، و پشتیبانی از انجام تراکنش‌ها (Transaction) دو حالت انجام قطعی یا بازگشت به حالت قبل می‌باشد.

یک پایگاه دانش، این نیازمندی‌های طراحی را ندارد، در عوض به داده‌های ساختاریافته نیاز دارد که ممکن است نشانگرهایی (pointer) شیئ‌های دیگر داشته باشند. قابل ذکر است که نمایش ایده‌آل برای یک پایگاه دانش مدل شیئی (Object Model) می‌باشد که در این مدل، کلاس، زیر کلاس و نمونه وجود دارد.

داده‌های پایگاه دانش برای سیستم‌های خبره، برای رسیدن به جواب خاصی استفاده می‌شدند: رسیدن به تشخیص پزشکی، طراحی یک مولکول، رسیدن به پاسخ در موارد اضطراری و یا هدف این پژوهش بازیابی اطلاعات شبیه حافظه انسان.

با استفاده از یک پایگاه‌داده، یک نمودار دانش را ایجاد خواهیم کرد. پایگاه‌داده مجموعه‌ای از اطلاعات سامان‌یافته است که بر اساس ترتیب و قواعدی مشخص در کنار یکدیگر نگهداری می‌شوند. مدیریت اطلاعات ذخیره شده در پایگاه‌داده توسط کاربران معمولاً از طریق سیستم مدیریت پایگاه داده یا Database Management System صورت می‌گیرد DBMSها ابزارها و مکانیزم‌های مختلفی را برای ایجاد و مدیریت دیتابیس‌ها در اختیار ما قرار می‌دهند.

در این پژوهش از پایگاه‌داده رابطه‌ای sql server برای ذخیره-سازی اطلاعات کاربر استفاده شد.

برای طراحی پایگاه‌داده‌ها در سطح انتزاعی پایین‌تر از سطح مدل‌سازی، به یک ساختار داده‌ای از یک مدل داده‌ای نیاز است و اساساً همین مدل داده‌ای تأمین‌کننده محیط انتزاعی است. در پایگاه‌داده رابطه‌ای بالاخص در محیط انتزاعی مورد استفاده کاربر، رابطه نمایشی جدولی دارد و اساساً پایگاه‌داده رابطه‌ای مجموعه‌ای است از تعدادی نوع جدول. مفاهیم ساختار جدولی عبارت‌اند از: سطر، جدول و ستون.

هر جدول از نظر محتوای داده‌ای مجموعه‌ای است از نمونه‌های متمایز از انواع سطرها و هر سطر نیز مجموعه‌ای از مقادیر است که هر کدام از یک مجموعه برگرفته شده‌اند. به هر یک از عناصر سطر یک ستون گویند. لازم است ذکر شود که در ساختار جدولی، تنها عنصر ساختاری اساسی همین مفهوم نوع جدول است .

معمولاً هر جدول (یا رابطه)، مربوط به یک نوع موجودیت (نظیر محصول، کارمند، دانشجو و ...) می‌باشد و هر ردیف از آن نمایانگر نمونه‌ای از این نوع موجودیت (نظیر محصولی با نام و مدل مشخص) و ستون‌ها هم مقادیر ویژگی‌ها (نظیر قیمت) را برای نمونه‌های مختلف موجودیت موردنظر نگه‌داری می‌کنند. مقادیر مجاز برای هر ویژگی می‌تواند توسط یک دامنه محدود شود. برای مثال ویژگی جنسیت می‌تواند به مجموعه دامنه {"مرد" و "زن"} محدود شود.

برای تعیین برچسب وابستگی برای اطلاعات ثبت شده توسط کاربر از کتابخانه عمومی spacy برای این کار استفاده خواهیم کرد.

```
1  import re
2  import pandas as pd
3  import bs4
4  import requests
5  import spacy
6  from spacy import displacy
7  nlp = spacy.load ('en_core_web_sm')
8  from spacy.matcher import Matcher
9  from spacy.tokens import Span
10 import networkx as nx
11 import matplotlib.pyplot as plt
12 from tqdm import tqdm
13 pd.set_option ('display.max_colwidth' , 200)
14 %matplotlib inline
```

spaCy یک ابزار یا کتابخانه رایگان برای پردازش زبان طبیعی بوده که با استفاده از زبان برنامه‌نویسی Python و Cython توسعه داده شده است. ابزار spaCy می‌تواند از اسنادی که با زبان‌های English, German, Spanish, Portuguese, French, Italian, Dutch and multi-languageنوشته شده، استفاده گردد و انواع عملیات متن‌کاوی را بر روی این اسناد انجام دهد.

مهم‌ترین ویژگی‌های ابزارspaCy :

- توکن‌سازی بدون تخریب
- شناسایی موجودیت نام‌گذاری شده
- پشتیبانی از "Alpha tokenization" در بیش از ۲۵ زبان
- مدل‌های آماری در بیش از ۸ زبان
- بردارهای کلمهٔ از پیش آزمایش شده
- برچسب‌گذاری بخشی از گفتار
- تجزیه وابستگی با برچسب
- تقسیم‌بندی جمله مبتنی بر نحو
- طبقه‌بندی متن
- بصری‌سازی داخلی برای نحو و موجودیت‌های نام‌گذاری شده
- ادغام یادگیری عمیق

یکی دیگر از مزایای ابزار spaCy استفاده از افزونه‌هایی مانند Thinc ، sense2vec و displaCy بوده که امکان بصری‌سازی برخی از عملیات متن‌کاوی را برای کاربران

فراهم می‌آورد. به‌عنوان‌مثال، کد زیر به طـور بصـری، رویـهٔ Dependency را بین واژه‌های یک سند نشان می‌دهد.

به طور مثال می‌توان قطعه کـد بـالا روی اطلاعـات فرضـی پیاده‌سازی کرد.

```
impot  spacy
nlp – spacy.load(‘en_core_web_sm’)
doc = nlp ("علی احمدی مسابقات مالزی جونیور را برد .")
for tok in doc:
   print(tok.text, " ..." ,  tok.dep_)
```

خروجی کد بالا به‌صورت زیر است.

```
علی  ... nsubj
احمدی  ... amod
مسابقات  ... compound
مالزی  ... compound
جونیور  ... compound
را .... aux
برد ... Root
.  ...  punct
```

و یا اطلاعات دیگری همچون

```
impot  spacy
nlp – spacy.load(‘en_core_web_sm’)
doc = nlp ("ایران  ست اول را برد")
for tok in doc:
   print(tok.text, " ..." ,  tok.dep_)
```

خروجی کد بالا بدین صورت است:

```
ایران  ... amod
ست  ... compound
اول  ... npadvmod
را .... aux
برد ... Root
```

۲-۳) طراحی گراف دانش

برای ساختن نمودار دانش، مهم‌ترین مؤلفه‌ها، گره‌ها و لبه-های بین آنها است. این گره‌ها، موجودیت‌هایی هستند که در جملات پایگاه داد حضور دارند. لبه‌ها، روابطی هستند که این موجودات را به یکدیگر متصل می‌کنند. ما این عناصر را به شکلی بدون نظارت استخراج می‌کنیم، یعنی از دستور زبان جملات استفاده خواهیم کرد. ایده اصلی این است که یک جمله را پشت سر بگذارید و فاعل و مفعول را به‌محض مواجهه با آنها استخراج کنید . سپس، تابعی را برای استخراج فاعل و مفعول (اشخاص) از یک جمله ایجاد کرده‌ایم. جهت استخراج روابط بین موجودیت‌ها از کد زیر استفاده می‌کنیم.

```
1   def get_relation(sent):
2      doc =  nlp(sent)
3       #Matcher  class  object
4       matcher  = Matcher(nlp.vocab)
5       #define  the  pattern
6       Pattern = [{‘DEP’ : ‘ROOT’} ,
7                  {‘DEP’ : ‘prep’ ,’OP’ :”?”}
8                  {‘DEP’ : ‘agent’ ,’OP’ :”?”}
9                  {‘DEP’ : ‘ADJ’ ,’OP’ :”?”}]
10   Matcher.add(“ matching _1” , None ,
     pattern)
11   Matches = matcher(doc)
12   K= len(matches) - 1
13    span = doc[matches[k][1] : matches[k][2]]
14    return(span.text)
```

الگوی تعریف شده در تابع سعی می‌کند کلمه ROOT یا فعل اصلی موجود در جمله را پیدا کند. پس از شناسایی ROOT ، این الگو بررسی می‌کند که آیا آن را با یک جمله (prep)یا کلمه عامل دنبال می‌کنید. یا نه. اگر بله پس آن به کلمه ROOT اضافه می‌شود. به همین ترتیب، روابط را از همه اطلاعات به دست می‌آوریم سرانجام یک نمودار دانش را از موجودیت‌های استخراج شده (جفت فاعل - مفعول) و گزاره‌ها (رابطه بین موجودات) ایجاد خواهیم کرد.

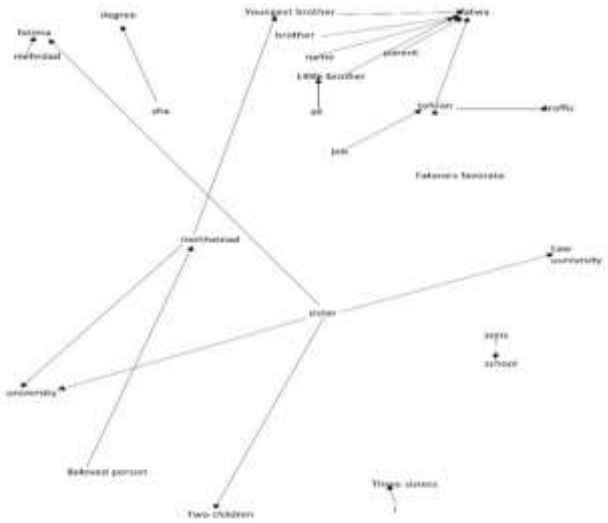
```
1   # extract subject
2   Source = [i[0] for i in entity_pairs]
3   # extract object
4    Target = [i[1] for i in entity_pairs]
5   kg_df = pd.DataFrome({‘source’:source,
    ‘target’:target, ‘edge’ : relation})
```

در مرحله بعد، ما از کتابخانه networkx برای ایجاد شبکه‌ای از این dataframe استفاده خواهیم کرد. گره‌ها نشان‌دهندهٔ موجودات هستند و لبه‌ها یا اتصالات بین گره‌ها روابط بین گره‌ها را نمایان می‌کنند.

به‌عبارت‌دیگر، رابطه بین هر جفت گره متصل دوطرفه نیست، بلکه فقط از یک گره به دیگری است. برای استخراج هر کلمه‌ای که می‌خواهیم کافی است لبه مربوطه را مطابق

آن در نظر بگیریم. درنهایت گراف دانش متن زیر به‌صورت شکل (۱) طراحی شد .

شکل (۱): گراف دانش



۴. نتایج و ارزیابی

ارزیابی یک مدل دسته‌بندی می‌تواند بر اساس نمونه‌های آموزشی و آزمایشی صورت گیرد. برای ارزیابی باید برچسبی که مدل دسته‌بندی به آن دسته حمله نسبت‌داده‌شده، مقایسه شود. وقوع حالات مختلف برای دسته‌ها با توجه به مجموعه داده‌های ورودی برای دسته‌بندی با مهم‌ترین معیار برای تعیین کارایی یک الگوریتم دسته‌بندی معیار Accuracy می‌باشد. این معیار دقت کل یک دسته‌بندی را محاسبه می‌کند. این معیار نشان‌دهنده این موضوع است که چند درصد از کل مجموعه داده‌ها به‌درستی دسته‌بندی شده است. رابطه زیر نحوه محاسبه معیار درستی را نشان می‌دهد.

فرمول (۱)

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N}$$

زمانی که ارزش False Negetive بالا باشد، معیار Recall معیار مناسبی خواهد بود. برای یک دسته، که از میان تمامی دسته‌های حمله‌های متعلق به آن دسته، به‌درستی دسته‌بندی‌شده است. نحوه محاسبه این معیار در رابطه زیر نشان‌داده‌شده است.

فرمول (۲)

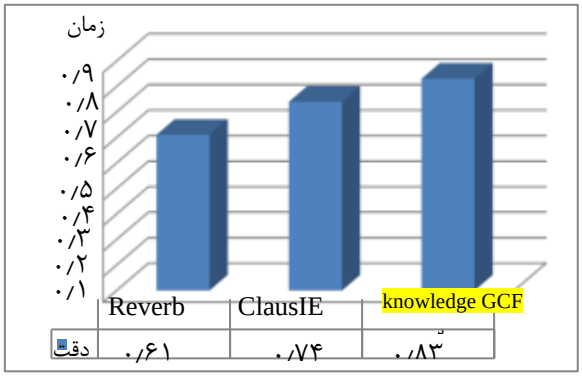
$$recall = \frac{T_P}{T_P + T_N}$$

معیار F-measure در مواردی استفاده می‌شود که نتوان اهمیت ویژه‌ای را برای هر یک از دو معیار Recall و Precision نسبت به یکدیگر قائل شد. رابطه زیر نحوه محاسبه این معیار را نشان می‌دهد.

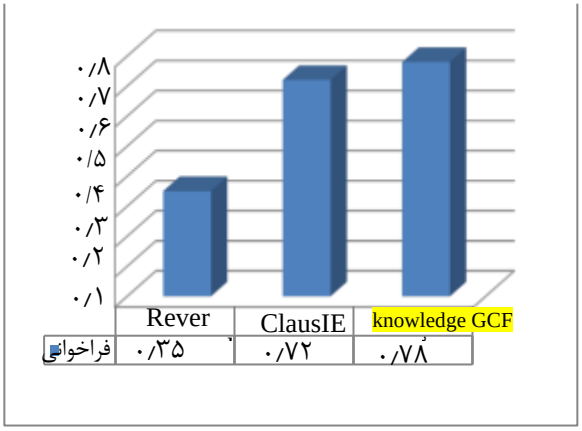
فرمول (۳)

$$F - MEASURE = \frac{2T_P}{2T_P + F_P + F_N}$$

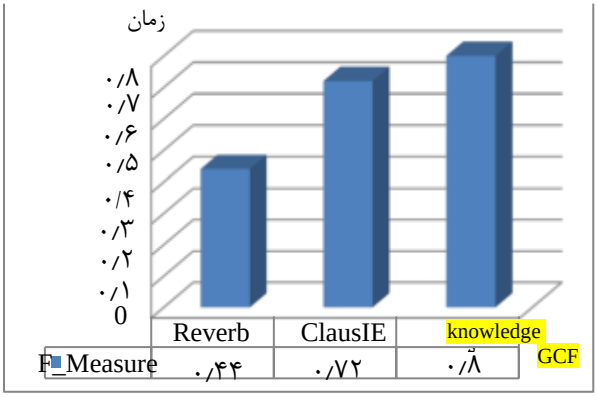
در بین سیستم‌های ذکر شده در بخش پیشینه، سیستم پیشنهادی در راستای سیستم‌های ClausIE [۱۲] و Reverb [۱۱] است. در نتیجه این سیستم‌ها جهت مقایسه و ارزیابی انتخاب شده‌اند.



شکل (سیستم ۲): ارزیابی و مقایسه دقت knowledge GCF



شکل(۳): ارزیابی و مقایسه فراخوانی knowledge GCF



شکل (۴): ارزیابی و مقایسه knowledge GCF f_measure

همان‌طور که در بخش کارهای پیشین بیان شد، سیستم ClausIE دقت و فراخوانی بالاتری نسبت به سایر سیستم‌های استخراج آزاد دیگر دارد. ولی یکی از چالش‌های این سیستم، کند بودن آن است. بعضی از حقایقی که در این سیستم استخراج می‌شوند، با یکدیگر هم‌پوشانی دارند. از بعضی جملات، حقایق اضافی و اشتباه به دست می‌آید، مثلاً صفت ملکی به‌تنهایی در قسمت نهاد قرار می‌گیرد . ولی در knowledge GCF ما با کدهایی که جهت استخراج حقایق بیان شد، فاعل، مفعول، نهاد، فعل و مانند آن به‌درستی مجزا شده و همین امر سرعت بازیابی اطلاعات را با توجه به اطلاعات تفکیک شده بالا می‌برد.

۵. نتیجه گیری

در ابتدا نحوه پیاده‌سازی knowledge GCF شرح داده شده است. پیاده‌سازی knowledge GCF در دو بخش ثبت اطلاعات در پایگاه‌داده توسط کاربر و استخراج حقایق در قالب RDF بیان شده است. سپس در بخش ارزیابی بعد از بیان معیارهای ارزیابی knowledge GCF در سه بخش تولید حقایق، ایجاد ارتباطات و ذخیره شمای گراف در قالب RDF و تولید خروجی گراف دانش با سیستم‌های دیگر مقایسه و ارزیابی شده است. نتایج حاصل از ارزیابی نشان می‌دهد که knowledge GCF در استخراج حقایق و ایجاد گراف دانش موفق بوده و نتایج مطلوبی را به دست آورده است. در راستای بهبود نتایج این تحقیق می‌توانیم موارد ذیل را پیشنهاد دهیم.

- استفاده از متغیرهای بیشتر به‌منظور مدل‌سازی بر روی مجموعه‌داده‌ها.
- پیاده‌سازی مدل پیشنهادی در این تحقیق بر روی مجموعه داده‌های دیگر و گزارش نتایج به دست آمده در تحقیقات آتی.
- می‌توان با استفاده از پایگاه‌داده‌های بزرگ‌تر، تعداد رکوردها را افزایش و دقت الگوریتم‌ها را نیز افزایش داد.

۶. مراجع

[1] B. R. Herganna, *an introduction to learning theories*, translated by Ali Akbar Seif, Ch IV, Tehran, Doran, p.423 ،و1397(in persian).

[2] B. R. Herganeh, previous, p. 433,1397(in persian).

[3] Rita. L. Atkinson, *The field of psychology*, translated by Mehdi Ganji, Ch. 10, p. 421,2013(in persian).

[4] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – The Story So Far. International journal on semantic web and information systems, 5(3):1–22, 2009.

[5] Douglas B Lenat. CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11):33– 38, 1995.

[6] Samuel Sarjant, Catherine Legg, Michael Robinson, and Olena Medelyan. “All you can eat” ontology-building:Feeding Wikipedia to Cyc. In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, pages 341–348, Piscataway, NJ, 2009. IEEE Computer Society.

[7] Denny Vrandeˇci´c and Markus Krötzsch. Wikidata: a Free Collaborative Knowledge Base. Communications of the ACM, 57(10):78–85, 2014.

[8] Jens Lehmann, Robert Isele,Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 6(2), 2013.

[9] George A Miller. WordNet: a lexical database for English.Communications of the ACM, 38(11):39–41, 1995. http: //dx.doi.org/10.1145/219717.219748.

[10] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M.Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. In Conference on Innovative Data Systems Research,2015.

http://www.cidrdb.org/cidr2015/
Papers/CIDR15_Paper1.pdf.

[11] Andrew Carlson, Justin Betteridge, Richard C Wang, Estevam R Hruschka Jr, and Tom M Mitchell. Coupled semisupervised learning for information extraction. In Proceedings of the third ACM international conference on Web search and data mining, pages 101–110, New York, 2010.

[12] L.Terveen and D.W.McDonald.Social matching: A framework and research agenda. ACM Trans. Comput-Hum. Interact., 12(3),2005.

[13] W.Eberle and L. Holder. Applying graph-based approaches to insider threat detection. In Proc.of the 5th Annual Workshop on Cyber Security and Information Intelligence Research , pages 206-208,2009.

[14] B. Zong , R. Raghhavendra, M. Srivatsa, X. Yan, A. K. Singh, and KW Lee. Cloud service placement via subgraph matching . In ICDE,2014.

[15] L. Del Corro, R. Gemulla, “ClausIE: clause-based open information extraction”, In Proceedings of the 22nd international conference on World Wide Web, pp. 355-366, International World Wide Web Conferences Steering Committee, 2013.

[16] O. Etzioni, A. Fader, J. Christensen, S. Soderland, M. Mausam, “Open information extraction: The second generation”, In Proceedings of the TwentySecond international joint conference on Artificial Intelligence-Volume Volume One, pp. 3-10, AAAI Press, 2011.

[17] Heiko Paulheim: Knowledge Graph Refinement:A Survey of Approaches and Evaluation Methods, Published 2017 in Semantic Web DOI:10.3233/SW-160218.