زمان دریافت مقاله: ۹۵/۶/۲۱
زمان پذیرش مقاله: ۹۵/۹/۶

جویا: یک سیستم پرسش و پاسخ فارسی

ایمان خانی جزنی
دانشجوی کارشناسی، دانشکدهٔ ریاضی، آمار و علوم کامپیوتر، پردیس علوم، دانشگاه تهران
پست الکترونیکی: imankhanijazani@ut.ac.ir

هدیه ساجدی*
استادیار، دانشکده ریاضی، آمار و علوم کامپیوتر، پردیس علوم، دانشگاه تهران
پست الکترونیکی: hhsajedi@ut.ac.ir

چکیده

سیستم‌های پرسش و پاسخ، زیرشاخه‌ای از پردازش زبان طبیعی و بازیابی اطلاعات محسوب می‌شوند. سیستم‌های پرسش و پاسخ، پرسش را به یک زبان طبیعی (مثلاً فارسی) دریافت کرده و جواب مختصر و دقیق را در اختیار کاربر قرار می‌دهند. بنابراین دیگر لازم نیست کاربر مانند سیستم‌های بازیابی اطلاعات پرسش خود را به کلیدواژه‌ها تبدیل کند و پس از بازیابی، تعداد زیادی سند را مطالعه کند تا به جواب دلخواه خود برسد. پیشرفت‌های قابل توجهی در این زمینه، بخصوص در زبان انگلیسی انجام شده است، اما در زبان فارسی چنین سیستمی طراحی و پیاده سازی نشده است. در صورتی‌که مزایای این‌گونه سیستم‌ها و کاربردهای آن‌ها مانند استفاده به‌عنوان جویشگر با قابلیت ویژه پرسش و پاسخ، دستیارهای همراه و غیره قابل توجه است. در این مقاله، مراحل تولید اولین سیستم پرسش و پاسخ فارسی در دامنه نامحدود و وب‌مبنا به نام «جویا» به همراه مجموعه داده‌های این سیستم، معرفی خواهد شد. جویا به‌صورت کلی شامل زیرمولفه‌های پردازش پرسش، بازیابی اطلاعات و استخراج جواب دقیق می‌باشد. به دلیل نبود مجموعه داده ارزیابی در زبان فارسی، مجموعه داده ارزیابی برای این سیستم تهیه شده است. مجموعه داده ارزیابی شامل ۴۱۲ پرسش متنوع و پاسخ متناظر آن است. سیستم پیشنهادی به ۸۰ درصد صحت دست پیدا کرده است.

واژه‌های کلیدی: سیستم‌های پرسش و پاسخ فارسی، پردازش زبان طبیعی، بازیابی اطلاعات، پردازش پرسش، استخراج جواب

* نویسندهٔ مسئول

۱- مقدمه

سیستم‌های پرسش و پاسخ، پرسش را به یک زبان طبیعی (مثلاً فارسی) دریافت کرده و جواب کوتاه و دقیق را در اختیار کاربر قرار می‌دهند. بنابراین دیگر لازم نیست کاربر مانند سیستم‌های بازیابی اطلاعات پرسش خود را به کلید واژه‌ها تبدیل کند و پس از بازیابی تعدادی زیادی سند را مطالعه کند تا به جواب دلخواه خود برسد. پردازش در سیستم‌های بازیابی اطلاعات غالباً تنها در سطح لغوی انجام می‌شود اما در سیستم‌های پرسش و پاسخ، پرسش کاربر نه تنها در سطح لغوی بلکه در سطح نحوی و معنایی نیز پردازش می‌شود و جواب به زبان طبیعی تولید می‌شود. به همین دلیل سیستم‌های پرسش و پاسخ روش‌های پردازش زبان طبیعی و بازیابی اطلاعات را با هم به‌کار می‌گیرند.

با توجه به تعریفی که از سیستم‌های پرسش و پاسخ ارائه شد، می‌توان این سیستم‌ها را به عنوان یک جویشگر پیشرفته در نظر گرفت که با قابلیت‌های پردازش زبان طبیعی مجهز شده‌اند. این سیستم‌ها کاربردهای بسیاری دارند که می‌توان به استفاده از آن‌ها در خانه‌های هوشمند، دستیارهای صوتی همراه و غیره اشاره کرد. به‌طور کلی می‌توان این‌طور گفت که در سطح کاربردی این سیستم‌ها در مکان‌هایی حضور خواهند داشت که پرسشی موجود باشد، مانند بخش اطلاعات یک سازمان [۱].

سیستم‌های پرسش و پاسخ را می‌توان به شکل‌های مختلفی تقسیم‌بندی کرد. در یک تقسیم‌بندی، سیستم‌های پرسش و پاسخ به دو دسته تقسیم می‌شوند [۲]:

۱- دامنه محدود: به پرسش‌های یک حوزه خاص (مانند اطلاعات در مورد یک کشور خاص، پزشکی، مذهبی و غیره) پاسخ می‌دهد.

۲- دامنه نامحدود: سیستم قابلیت پاسخگویی به هر پرسشی را دارا است. این نوع سیستم‌ها بر هستان‌شناسی‌های عمومی و دانش جهانی تکیه دارند.

ذکر این نکته ضروری است که سیستمی ایده‌آل است که بتواند کمترین وابستگی به محیط پیرامونی را داشته باشد. به این معنا که در تمام زمینه‌ها قابلیت پاسخگویی داشته باشد و در تمام حوزه‌ها مانند فرهنگی، اجتماعی، علمی و غیره قابل استفاده باشد. به همین دلیل نیاز به وجود سیستم‌های دامنه نامحدود مانند جویا در بخش‌های مختلف اجتماع و حتی زندگی روزمره احساس می‌شود.

تقسیم‌بندی دومی که برای سیستم‌های پرسش و پاسخ مطرح می‌شود بر اساس تعداد زبان‌های پشتیبانی شده است. گروهی از این سیستم‌ها تک‌زبانه هستند. به عبارت دیگر، فقط یک زبان خاص را به‌عنوان زبان پرسش در نظر می‌گیرند. گروهی دیگر سیستم‌های چند زبانه هستند که قابلیت درک چندین زبان مختلف را دارا هستند [۳].

می‌توان تقسیم‌بندی دیگری را بر اساس دانش کلی در نظر گرفت. به این صورت که دانش کلی سیستم پرسش و پاسخ می‌تواند بر اساس دانشی محلی باشد. به عبارت دیگر، تمامی اسناد مورد نیاز به صورت محلی ذخیره شده باشند، یا دانش بر اساس محتویات وب باشد. در این‌صورت سیستم پرسش و پاسخ برای دریافت اسناد باید به اینترنت دسترسی داشته باشد که به آن سیستم پرسش و پاسخ وب مبنا می‌گویند.[۴]

در این مقاله اولین سیستم پرسش و پاسخ فارسی با دامنه نامحدود به نام «جویشگر ایرانی» که به اختصار آن را جویا می‌نامیم، معرفی شده است. انگیزه اصلی از طراحی سیستم جویا نیز رسیدن به یک سیستم جامع بر اساس طبقه‌بندی‌های فوق بوده است. جویا از روش‌های مستقل از زبان استفاده می‌کند. این استقلال به معنای عدم به‌کارگیری قواعد زبانی است. جویا یک سیستم پرسش و پاسخ مستقل از زبان است. به‌عبارت دیگر، با صرف هزینه‌ای اندک می‌توان این سیستم را برای زبان‌هایی غیر از فارسی آماده‌سازی کرد. جویا برای کسب دانش و اطلاعات مورد نیاز خود به اینترنت متصل می‌شود. مرجع اصلی دانش جویا، دانش‌نامه آزاد ویکی‌پدیای فارسی است. سیستم پرسش و پاسخ جویا می‌تواند به

پرســش‌هایی از نوع حقیقت[1] پاســخ دهد. جویا دارای سه مولفه پردازش پرسش، بازیابی اطلاعات و استخراج جواب دقیق می‌باشــد. در مولفه پردازش پرســش، هدف تحلیل پرسش است به صورتی که بتوان خواسته پرسش را درک کرد و همچنین پرسش را طوری تغییر داد که مناسب مولفه بازیابی اطلاعات باشد. در مولفه بازیابی اطلاعات اسناد و سپس جملات محتمل بازیابی می‌شود که امید است جواب پرســش کاربر در آن‌ها باشــد. در مولفه استخراج جواب دقیق، سیســتم به دنبال بهترین جواب برای پرسش کاربر می‌گــردد و در انتها بهترین جواب بــه کاربر نمایش داده می‌شود.

برای ارزیابی این سیســتم به مجموعه دادگان پرسش به همراه جواب دقیق آن نیاز است. متاسفانه این مجموعه داده برای زبان فارسی وجود ندارد. به همین دلیل مجموعه داده ای شامل ۴۱۲ پرسش به همراه طبقه پرسش و جواب آن تولید شد. ۴۰ پرســش به‌عنوان مجموعه آزمون برای ارزیابی این سیستم در نظر گرفته شده است. دقت به‌دست آمده برای سیستم پیشنهادی ۸۰ درصد می‌باشد.

نوآوری‌های زیر در این مقاله صورت گرفته است:

۱- پیشنهاد معماری مناســب برای سیستم پرسش و پاسخ وب مبنا و دامنه نامحدود برای زبان فارسی

۲- اســتفاده از الگوریتم‌های یادگیری ماشــین و عدم وابستگی به زبانی خاص

۳- تولید مجموعه داده پرسش- طبقه که شامل پرسش، نوع پرسش و جواب دقیق آن

۴- استفاده از فرهنگ لغت برای اولین بار برای ساخت گراف معنایی

ساختار این مقاله به شرح زیر است: در بخش ۲، کارهای مرتبط با موضوع شــرح داده می‌شود. کارهای مرتبط در دو زیربخش انگلیسی و ســایر زبان‌ها و فارسی بررسی می‌شــوند. در بخش ۳، معماری سیســتم پرسش و پاسخ پیشنهادی جویا و همچنین مجموعه داده‌های مورد استفاده را معرفــی خواهیم کرد. در بخش ۴ مجموعه آزمایش‌ها و تحلیل نتایج حاصل از آن گزارش می‌شــود و در بخش ۵ نتیجه‌گیری و کارهای آتی بیان خواهد شد.

1- factoid

۲- کارهای مرتبط

با گذشت بیش از نیم قرن از ظهور سیستم‌های پرسش و پاسخ تعداد زیادی برنامه کاربردی در این خصوص در ســطح علمی و صنعتی پیاده‌سازی شده‌اند [۵,۶]. کارهای مرتبط در زبان انگلیســی و ســایر زبان‌هــا در زیربخش ۲-۱ و کارهای مرتبط در زبان فارسی در زیر بخش ۲-۲ بررسی خواهند شد.

۲-۱- در زبان انگلیسی و سایر زبان‌ها

با توجه به تحقیقات وسیعی که در زبان انگلیسی انجام شده است، سیســتم‌های پرسش و پاسخ قابل قبولی برای این زبان تولید و توسعه داده شده است و همچنین می‌توان به چندین سیســتم تولید شــده برای دیگر زبان‌ها اشاره کــرد. در ادامــه به معرفی چندین مورد از این سیســتم‌ها می‌پردازیم.

اولین سیســتم پرسش و پاسخ در ســال ۱۹۶۱ به نام BASEBALL نام‌گذاری شد که قادر به جوابگویی پرسش کاربر در حوزه ورزش بیســبال برای اطلاعات یک ســال بــود[۷]. LUNAR رابطی برای داده‌های به‌دســت آمده از تحلیل نمونه ســنگ‌های ماموریت‌های آپولــو به ماه بود که در ســال ۱۹۷۲ ساخته شــد[۸]. سیستم‌های ذکر شده، پرســش‌هایی را بر اساس الگوهای زبان طبیعی که رخداد آن‌هــا در ورودی، مورد انتظار بود، پردازش می‌کردند[۹]. سیســتم‌های BASEBALL و LUNAR از روش‌هایی مانند ELIZA و DOCTOR ، اولین روبات‌های گپ‌زنی، اســتفاده می‌کردند. همچنین می‌توان به سیستم‌های پرسش و پاسخ اولیه دیگری مانند SYNTHEX، LIFER و PLANES اشاره کرد [۱۰]. سیســتم‌های ابتدایی ذکر شده، شامل یک پایگاه داده بودند. به عبارت دیگر بازیابی در اســناد ساخت‌یافته انجام شده است، اما دانش جویا مجموعه اسناد ساخت‌نیافته

می‌باشد.

استارت[2] اولین سیستم پرسش و پاسخ تحت وب که از سال۱۹۹۳ در حال اجرا است، توسط گروه اینفولَب دانشگاه ام‌آی‌تی در سال۲۰۰۴ توسعه داده شده است. استارت در مورد موضوعات مختلفی از قبیل مکان، شخص، فیلم، تعریف فرهنگ لغت و غیره قدرت پاسخگویی به پرسش را دارا است [۱۱].

سیستم پرسش و پاسخ Wolframe-alpha یک موتور محاسباتی دانش یا موتور جواب است که توسط Wolfram Research در سال۲۰۰۹ توسعه داده شده است و می‌تواند به پرسش‌های از نوع حقیقت پاسخگو باشد. یکی از مهم‌ترین و کاربردی‌ترین قسمت‌های این سیستم توانایی آن در حل مسائل ریاضی است که به دلیل وجود Mathematica ، در هسته اصلی این سیستم است.

شرکت آی‌بی‌ام یک کامپیوتر با قابلیت هوش مصنوعی برای پاسخگویی به پرسش‌های کاربر به نام واتسون تولیدکرده است. این سیستم برخلاف بیشتر سیستم‌های پرسش و پاسخ به اینترنت وصل نیست و تعداد زیادی سند در موضوعات مختلف به عنوان دانش جهانی واتسون به صورت محلی ذخیره شده است. این سیستم در سال۲۰۱۱ در مسابقه اطلاعات عمومی معروفی در آمریکا به نام جئوپاردی شرکت کرد و با شکست دادن دو نفر از برندگان قبلی آن مسابقه، جایزه یک میلیون دلاری نفر اول را از آن خود کرد. در سال۲۰۱۳، سیستم پرسش و پاسخ واتسون برای کاربردهای تجاری و صنعتی آماده سازی شد [۵,۶].

ارسطو یک سیستم هوشمند با قابلیت‌های استدلال، یادگیری و خواندن می‌باشد که در سال۲۰۱۳ ارائه شد. هدف از تولید این سیستم، حل پرسش‌های دشوار امتحان علوم در چندین مقطع تحصیلی بود که تمام این امتحانات را گذراند و در نهایت تمرکز این سیستم از حل پرسش‌های امتحان علوم به جوابگویی پرسش‌های کاربر تغییر داده شد. این سیستم نیز در ابتدا محدود به درس علوم بوده و قابلیت استدلال نیز دارد.

درسال ۱۳۸۳ یک سیستم پرسش و پاسخ مبتنی بر هستان‌شناسی برای حوزهٔ مخابرات با قابلیت استخراج و دسته‌بندی خودکار مستندات برای پاسخگویی به پرسش‌های حوزهٔ تخصصی مخابرات فیبر نوری ارائه شده است [۱۲]. این سیستم، پرسش‌های کاربر را به زبان انگلیسی دریافت کرده و به کمک استدلال روی گراف هستان‌شناسی پاسخ دقیق را استخراج کرده و به همراه پاراگراف‌های خلاصه‌سازی شده مرتبط در اختیار کاربر قرار می‌دهد. با تغییر در ساختار هستان‌شناسی این سیستم، می‌توان دامنه پاسخگویی سیستم را تغییر داد. نتایج به‌دست آمده از ۱۰۰ پرسش در دامنه فناوری مخابرات، نشان داده است که این سیستم از دقت و سرعت قابل قبولی برخوردار می‌باشد. این سیستم نیز دامنه محدود است. همچنین این سیستم از هستان‌شناسی استفاده می‌کند که موجب محدود شدن آن می‌شود.

سیستم ارایه شده توسط JIE LIU و همکارانش می‌تواند پرسش‌های حوزه پزشکی به زبان چینی را پاسخگو باشد [۱۳]. این سیستم نیز همانند سیستم پیشین دامنه محدود است.

سیستم پرسش و پاسخ QARAB سیستمی است که دارای سه مؤلفهٔ کلی پردازش پرسش، بازیابی اطلاعات و استخراج جواب است. این سیستم قدرت پاسخگویی به پرسش‌های به زبان عربی، در حوزه محدود به روزنامه A1-RAYA چاپ شده در کشور قطر را دارا است [۱۴]. دانش این سیستم محدود به یک روزنامه خاص است که جامعیت پاسخگویی سیستم را کاهش می‌دهد.

برای اولین بار QA@CLEF-2004 سیستم پرسش و پاسخی را پیشنهاد داده است که زبان‌های متنوع اروپایی مانند هلندی، آلمانی، ایتالیایی، پرتغالی، اسپانیایی، انگلیسی و بلغاری را بتواند هم در مبدأ و هم در مقصد داشته باشد. به‌عبارت دیگر، کاربر بتواند به زبان‌های مختلف اروپایی

2- Start

پرسش بپرسد و به همان زبان مورد نظر، جواب را دریافت کند [۱۵]. این سیستم نیز همانند جویا از روش‌های مستقل از زبان استفاده می‌کند.

جویا همانند سیستم‌های استارت و واتسون دامنه نامحدود است و بر خلاف واتسون اطلاعات خود را از وب اخذ می‌کند. البته در نسخه اول جویا، دانش مبتنی بر ویکی‌پدیای فارسی است. جویا همانند استارت و Wolfram-alpha سوال‌های حقیقت را پاسخگو است و نمی‌تواند پرسش‌های پیچیده مانند واتسون را پاسخگو باشد. جویا دامنه نامحدود است ولی نمی‌تواند پرسش-هایی را که به استدلال منطقی نیاز دارند، پاسخگو باشد. جویا همانند QARAB دارای سه مؤلفهٔ کلی است اما اطلاعات مورد نیاز خود را از انبوه اسناد متنی استخراج می‌کند.

۲-۲- در زبان فارسی

تاکنون در زبان فارسی سیستم پرسش و پاسخ در دامنه نامحدود جهت خدمت رسانی تولید نشده است. با وجود این، برای هر یک از بخش‌های سیستم پرسش و پاسخ، تحقیقاتی انجام شده و نتایج آن‌ها در قالب مقالاتی منتشر شده‌است و همچنین زیرسیستم‌هایی برای زبان فارسی تولید شده است. تعداد محدودی مقاله و پایان نامه در حوزه سیستم‌های پرسش و پاسخ فارسی منتشر شده که به معرفی آن‌ها می‌پردازیم.

در سال ۱۳۸۵ یک سیستم پرسش و پاسخ با دامنه نامحدود برای زبان فارسی ارائه شده است که به نظر می‌رسد با استفاده و بهره گیری از هستان‌شناسی‌های مختلف قادر به پاسخگویی پرسش کاربر است. مقاله یا گزارشی از این سیستم در اختیار عموم قرار ندارد[۱۶]. این سیستم از روش‌های استخراج اطلاعات به منظور ساخت هستان استفاده کرده است. در حقیقت پردازش این سیستم بر روی داده‌های ساخت‌یافته صورت می‌گیرد. این موضوع باعث عدم جامعیت می‌شود، به این دلیل که روابط موجود در هستان محدود است و رابطه‌های خاص و یا رابطه‌های جدید به وجود آمده قابل استخراج نیست و سیستم دوباره باید به روزرسانی گردد.

در سال ۱۳۹۱ سیستم پرسش و پاسخی در دامنه محدود ارائه شده که توانایی پاسخگویی به پرسش‌های در حوزهٔ اطلاعات پرواز را دارد. این سیستم با توجه به معماری ارائه شده، پردازش‌هایی را بر روی پرسش انجام می‌دهد و با ایجاد پرس و جویی به زبان پایگاه داده آن را بر روی پایگاه داده اجرا نموده و جواب ایجاد شده را در یک یا چند سطر به کاربر نشان می‌دهد [۱۷].

در سال ۱۳۹۲ یک سیستم پرسش و پاسخ برای بخش اطلاعات دانشکده مهندسی برق و کامپیوتر ارائه شده است که قادر به پاسخگویی به پرسش‌های آموزشی در دانشکده مهندسی برق و کامپیوتر می‌باشد. در این سیستم پرسش کاربر برای استخراج نوع پرسش، موجودیت‌های با نام، زمان و مفهوم فعل، مورد تحلیل قرار می‌گیرد و سپس پرسش کاربر با استفاده از جایگاه کلمات در پرسش و روابط موجود در هستان‌شناسی به مجموعه ای از سه تایی‌های مرتبط (فاعل، مفعول و فعل) بازنمایی می‌شود. لازم به ذکر است که رویکرد ارائه شده در این سیستم مستقل از دامنه بوده و با تغییر هستان‌شناسی دامنه و داده‌های ورودی آن، برای سایر دامنه‌ها نیز قابل استفاده می‌باشد. این سیستم به دلیل استفاده از ساختارهای هستان‌شناسی فقط دارای دو بخش کلی تحلیل پرسش و استخراج پاسخ است. این سیستم با ۲۴۰ پرسش به عنوان مجموعه پرسش آزمون، صحت ۹۱/۳۴ درصد و بازخوانی ۸۷/۹۲ را کسب کرده است [۱].

دو سیستم پیشنهادی برای اطلاعات پرواز و اطلاعات دانشکده مهندسی برق و کامپیوتر از روش‌های پایگاه داده‌ای و هستان‌شناسی استفاده می‌کنند. بنابراین نمی‌توان از سیستم‌های پیشنهادی در محیط‌های واقعی که پرسش‌ها دامنه مشخصی ندارند استفاده کرد.

در سال ۱۳۹۳ یک پیکره متنی فارسی پرسش و پاسخ که قدمی برای تولید سیستم پرسش و پاسخ در دامنه

محدود به پرسش‌های مذهبی می‌باشد، توسعه داده شده است [۱۸]. داده‌های خام مورد استفاده در این پیکره شامل دو فایل متنی استفتائات آیت الله خامنه‌ای و رساله آیت الله مکارم شیرازی است که توسط مرجع دادگان زبان فارسی فراهم شده است. به منظور توسعه پیکره، برای هر دو فایل متنی ذکر شده، اسنادی به شکل XML ایجاد شده است. فایل XML شامل دسته کلی پرسش، زیرعنوان مربوط به پرسش، متن پرسش و نوع پرسش می‌باشد. این پیکره شامل ۲۱۱۸ پرسش درمورد غیر حقیقت[3]، ۲۰۵۱ پرسش در مورد حقیقت از استفتائات آیت الله خامنه‌ای و ۲۴۵۶ مسئله از کتاب رساله آیت الله مکارم شیرازی می‌باشد. پرسش‌هایی از نوع حقیقت در نظر گرفته می‌شوند که پاسخگویی به آن‌ها نیازمند استدلال نباشد. غالبا پاسخ این نوع پرسش‌ها یک موجویت عددی یا اسمی است. اما پرسش‌های از نوع غیرحقیقت غالبا حاوی پاسخ های طولانی هستند و به دلیل نیاز به استدلال بیشتر، پاسخگویی به آنها دشوارتر است. می‌توان از این پیکره در سیستم‌های پرسش و پاسخ انجمنی[4] در حوزه مذهبی استفاده کرد. اما جویا پرسش‌های موجود در دادگان تولید شده را حفظ نمی‌کند بلکه سعی می‌کند بر اساس الگوریتم‌های پیشنهادی نوع پرسش و طریقه جوابگویی به آن را یاد بگیرد.

جویا کلیه اطلاعات موجود در داده‌های ساخت‌نیافته متنی را مورد بررسی قرار می‌دهد. البته لازم به ذکر است سیستم‌هایی که از هستان‌شناسی استفاده می‌کنند دقت بالاتری دارند اما در شرایط واقعی فراخوانی پایینی دارند. جویا خود را به هستان‌شناسی محدود نکرده است و از روش‌های مبتنی بر بازیابی اطلاعات استفاده می‌کند.

۳- سیستم پیشنهادی پرسش و پاسخ جویا

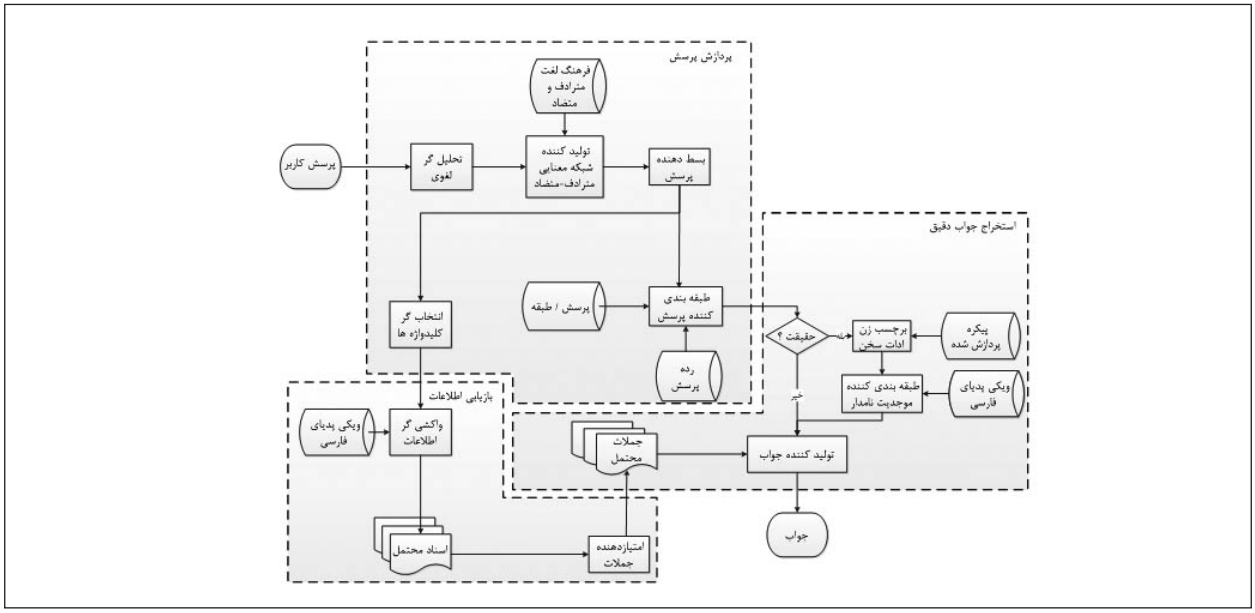
معماری سیستم پیشنهادی پرسش و پاسخ جویا در شکل۱ نمایش داده شده است. جویا به طور کلی دارای سه مولفه پردازش پرسش، بازیابی اطلاعات و استخراج جواب دقیق است.

پردازش پرسش مهمترین مولفه موجود در سیستم‌های پرسش و پاسخ و همچنین سیستم جویا می‌باشد. در سیستم‌های بازیابی اطلاعات، پرسش یا درخواست کاربر در سطح معنایی پردازش نمی‌شود اما در سیستم‌های پرسش و پاسخ، پرسش کاربر نه تنها در سطح لغوی بلکه در سطح معنایی نیز بررسی می‌شود. با رهنمود از تفکر انسانی هنگام جواب‌گویی به پرسش درمی‌یابیم که انسان در ابتدا با بررسی تعداد محدودی از کلمات موجود در پرسش، رده پرسش را تشخیص می‌دهد و بعد از یافتن رده پرسش، رویکردهای متفاوتی برای استخراج جواب در پیش می‌گیرد. در نتیجه یکی از زیرمولفه‌های مهم پردازش پرسش و حساس‌ترین قسمت در سیستم‌های پرسش و پاسخ، رده‌بندی پرسش است. به‌عنوان مثال، پس از پردازش پرسش «نویسنده کتاب شفا کیست؟» مشخص می‌شود که رده پرسش یا نوع جواب مورد نظر، اسم شخص است. پس سیستم پرسش و پاسخ در جملات محتمل باید به دنبال اسم شخص بگردد. در مولفه پردازش پرسش، پرسش کاربر بسط داده شده و پرسش‌هایی هم معنا با پرسش کاربر تولید می‌شود که در سطح لغوی دارای تفاوت‌هایی با پرسش کاربر است. این مولفه در بخش ۳-۱ به طور کامل توضیح داده می‌شود.

بازیابی اطلاعات قسمت میانی سیستم جویا است. پس از مشخص شدن نوع پرسش و بسط پرسش، کلیدواژه‌ها از پرسش استخراج شده و به زیرمولفه واکشی‌گر اطلاعات برای بازیابی اسناد محتمل داده می‌شود. در این مولفه نیز، جملات موجود در اسناد بازیابی شده امتیازدهی می‌شوند و جملات محتمل برای مولفه استخراج جواب دقیق، آماده سازی می‌شوند. این مولفه در بخش ۳-۲ به طور کامل توضیح داده شده است.

مولفه استخراج جواب دقیق با در نظر گرفتن نوع پرسش و جملات محتمل به‌دست آمده، جواب کوتاه و دقیق را استخراج می‌کند. در این مولفه بر اساس رده پرسش

3- Non-factoid
4- Community Question Answering



شکل ۱: معماری سیستم پرسش و پاسخ جویا

به‌دست آمده، رویکردهای متفاوتی برای استخراج جواب دقیق در نظر گرفته می‌شود. این مولفه در بخش ۳-۳ با جزئیات توضیح داده می‌شود.

سیستم پرسش و پاسخ جویا از فرهنگ جامع مترادف و متضاد زبان فارسی [۱۹] برای ساخت شبکه معنایی مترادف و متضاد استفاده می‌کند و همچنین از پیکره برچسب‌گذاری شده بی‌جن‌خان [۲۰] برای برچسب‌گذاری نحوی استفاده می‌کند. مجموعه داده‌هایی نیز برای اولین بار در زبان فارسی جهت هستان‌شناسی پرسش و یادگیری طبقه‌بندی پرسش‌ها نیز، در این سیستم تولید شده است. همچنین اسناد موجود در ویکی‌پدیای فارسی به عنوان دانش جهانی جویا در نظر گرفته شده است. به عبارت دیگر، جویا یک سیستم پرسش و پاسخ وب مبنا است.

۳-۱ معرفی مجموعه داده‌ها

سیستم پرسش و پاسخ جویا از چهار مجموعه داده به نام‌های فرهنگ لغت مترادف و متضاد، رده پرسش، پرسش- طبقه و پیکره پردازش شده استفاده می‌کند که به ترتیب در قسمت های بعد معرفی می‌شوند.

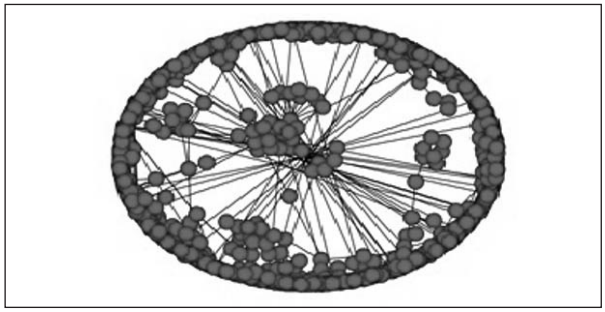
۳-۱-۱ فرهنگ لغت مترادف و متضاد

این فرهنگ لغت متشکل از ۲۰۰۰۰ مدخل، ۲۹۴۰۰ حوزه معنایی و ۱۹۵۰۰۰ واژه تدوین شده است [۱۹]. بر اساس آخرین بررسی‌ها برای اولین‌بار جویا این واژگان را به یک گراف از روابط مترادف و متضاد تبدیل می‌کند که در آن میزان نزدیکی و ارجحیت معنایی واژگان بر اساس وزن یال‌ها مشخص می‌شود. به عبارت دیگر گره‌های این گراف، واژگان این فرهنگ لغت هستند و یال‌ها و وزنشان به ترتیب بیانگر وجود رابطه و میزان نزدیکی از لحاظ معنایی بین دو واژه، می‌باشند. میزان همبستگی ۷۲۴ واژه و ۵۲ واژه از این فرهنگ به ترتیب در شکل۲-الف و ۲-ب نشان داده شده است. برای واژگان مترادف وزن یال‌ها مثبت در نظر گرفته شده و برای واژگان متضاد از وزن منفی، با خط چین مشخص شده، استفاده شده است. شکل ۳ که زیرگراف مربوط به واژه «ماجراجو» است را در نظر بگیرید.

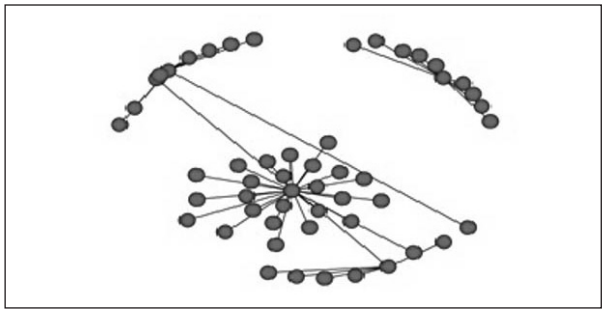
۳-۱-۲ رده بندی[5] پیشنهادی برای پرسش‌ها

رده بندی‌های مختلفی برای نوع پرسش، تولید شده است. از جمله می‌توان از طبقه‌بندی Hermjakob که شامل ۱۸۰ طبقه است و بزرگترین طبقه‌بندی تا کنون می‌باشد،

5- taxonomy



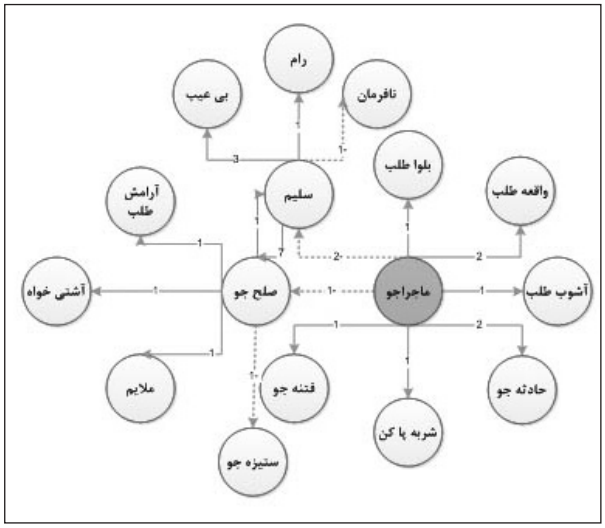
شکل ۲: الف. زیرگراف بدون‌وزن ۷۲۴ واژه



شکل ۲: ب. زیرگراف بدون وزن ۵۲ واژه

نام برد [۲۱]. همچنین می‌توان از طبقه‌بندی Li and Roth نام برد که شامل ۶ درشت‌دانه و ۵۰ ریزدانه است [۲۲]. جدول ۱ این طبقه‌بندی را نشان می‌دهد. Metzler دو درشت‌دانه به نام‌های لیستی و بله خیر را به جدول اضافه کرد [۲۳,۲۴,۲۵].

به دلیل افزایش دقت و پوشش موجودیت‌های مختلف، رده‌های بیشتری در سیستم جویا در نظر گرفته شده است و تغییراتی نسبت به جدول ۱ اعمال شده است. با بررسی پرسش‌های آزمون و طبقه مربوط به آن، ریزدانه‌های جدول ۱ که رخ نداده بودند از ریزدانه‌های سیستم جویا حذف شده‌اند. سیستم جویا شامل ۸ درشت‌دانه و ۵۹ ریزدانه است. به عبارت دیگر سیستم پرسش و پاسخ جویا از درشت‌دانههای طبقه‌بندی Metzler استفاده می‌کند ولی به ریزدانه‌های طبقه‌بندی Li and Roth، ۱۲ ریزدانه اضافه شده است و ۲ ریزدانه برجسته شده در جدول ۱، حذف شده است. کلماتی که در جدول ۲ برجسته شده، ریزدانه‌های اضافه شده و درشت‌دانه‌هایی که به‌صورت کج نشان



شکل ۳: زیرگراف فرهنگ لغت مترادف و متضاد برای واژه «ماجراجو»

جدول ۱: طبقه‌بندی نوع پرسش Li and Roth

درشت‌دانه	ریزدانه
ABBREVIATION	Abbreviation- **expression**
ENTITY	Animal- Body- color- Creative- **Currency**- Disease/ Medicine- language- Event- Food- Instrument- Letter- Other- Plant- Product- Religion- Sport- Substance- Symbol- Technique-Term- Vehicle- Word
Definition– Description– Manner Reason	
City- Country- Mountain- Other- State	
HUMAN	Group- Individual- Title- Description
NUMERIC	Code- Count- Date- Distance- Money- Order- Other- Period- Percent- Speed- Temp– Size– Weight

داده شده از درشت‌دانه‌های اضافه شده در طبقه‌بندی Metzeler گرفته شده است. جدول ۲ طبقه‌بندی نوع پرسش سیستم جویا را نشان می‌دهد.

۳-۱-۳مجموعه داده پرسش-طبقه تولید شده

این مجموعه داده شامل ۴۱۲ پرسش به همراه نوع جواب و همچنین جواب دقیق تحت عنوان پرسش- طبقه

جدول ۲: طبقه‌بندی پیشنهادی نوع پرسش سیستم پرسش و پاسخ جویا

درشت‌دانه	ریزدانه
مخفف	مخفف
موجودیت	حیوان- بدن- رنگ- اثر- بیماری- اتفاق- غذا- وسیله- زبان- حرف- گیاه- محصول- دین- ورزش- ماده- نماد- روش- ماشین- کلمه- بقیه- سوره
توصیف	توصیف- تعریف- حالت- دلیل
مکان	شهر- کشور- کوه- استان -بقیه- رودخانه- شهرستان -قاره- بنای‌تاریخی- سازمان- خیابان- آبشار- اقیانوس
بشری	توصیف- گروه- فرد-شغل- بقیه
عدد	کد- تعداد- تاریخ- فاصله- پول- رتبه- دوره- درصد- سرعت- دما- اندازه- وزن- بقیه
لیست	لیست
بله‌خیر	بله‌خیر

جدول۳: نمونه ای از مجموعه یادگیری پرسش-طبقه تولید شده برای سیستم پرسش و پاسخ جویا

جواب دقیق	پرسش	نوع جواب
۵	زمین از نظر حجم چندمین سیاره در منظومه شمسی است	عدد:رتبه
آلمان	بتهوون اهل کدام کشور است	مکان:کشور
ذیحجه	حضرت علی و فاطمه زهرا چه زمانی ازدواج کردند	عدد:تاریخ
ابن سینا	نویسنده کتاب شفا چه کسی است	بشری:فرد
۱۳۶۱	خرمشهر در چه سالی از وجود نیروهای عراقی آزاد شد	عدد:تاریخ
بری‌بری	بیماری ناشی از کمبود ویتامین B چه نام دارد	موجودیت:بیماری
بله	آیا زین العابدین مراغه ای سیاحت‌نامه ابراهیم بیگ را به تحریر درآورده است	بله‌خیر:بله‌خیر
گلف استریم	به جریان آب گرم اقیانوس اطلس چه میگویند	توصیف:تعریف
غسل ارتماسی	غسلی که در آن یک مرتبه در آب فرو میروند چه نام دارد	موجودیت:دین
۸۰%	حشره ها چه میزان از تعداد جانوران دنیا را تشکیل میدهند	عدد:درصد
نطنز	گلابی سوغات نطنز است یا کرمان	لیست:لیست

گردآوری شده است که درباره موضوعات مختلفی از قبیل سیاسی، فرهنگی، اجتماعی، گردشگری و غیره می‌باشد. قسمتی از این مجموعه داده در جدول ۳ نمایش داده شده است. ۳۷۲ پرسش به همراه نوع جواب برای مرحله یادگیری در نظر گرفته شده و ۴۰ پرسش باقی مانده به همراه نوع جواب و جواب دقیق در مرحله آزمون، برای ارزیابی صحت کار سیستم، مورد استفاده قرار می‌گیرد.

۳-۱-۴پیکره پردازش شده

یکی از اقدامات اساسی و اولیه در حوزه پردازش زبان‌های طبیعی تهیه یک پیکره مناسب می‌باشد. مجموعه پیکره‌های برچسب‌گذاری شده متعددی برای زبان‌های مختلف تاکنون به‌وجود آمده است که از مهمترین آن‌ها می‌توان به پیکره‌ای به نام Penn Treebank اشاره کرد [۲۶].

پیکره پردازش شده [۲۰]، بخشی از پیکره برچسب زده شده بی‌جن‌خان است [۲۶]. این پیکره از برخی اخبار روزنامه‌ها و متون معمولی جمع آوری شده است. یکی از ویژگی‌های بارز این پیکره آن است که هر سند در این مجموعه دارای یک عنوان می‌باشد. به‌عنوان مثال، اسناد تحت عناوین سیاسی، فرهنگی، اقتصادی و غیره دسته‌بندی شده‌اند. قابل ذکر است که در این پیکره ۴۳۰۰ عنوان مختلف وجود دارد. این دسته‌بندی بزرگ نشان دهنده کیفیت بالای این پیکره می‌باشد. در عملیات برچسب‌زنی از عناوین متون صرف نظر شده است زیرا که هدف به‌دست آوردن یک برچسب زننده خودکار است. این پیکره با مجموعه غنی از برچسب‌ها، شامل ۵۵۰ برچسب مختلف، برچسب زنی شده است. این مجموعه برچسب برای برچسب زدن دقیق و جزئی کلمات به کار گرفته می‌شود اما از آنجایی که در برچسب زنی خودکار، هدف مشخص کردن کلمات از نظر نوع کلی آن‌ها می‌باشد و وارد جزئیات نمی‌شود [۴]، معمولا از این تعداد برچسب استفاده نمی‌شود. از طرف دیگر از آنجایی که برچسب زنی خودکار بر اساس یادگیری ماشینی می‌باشد، در نظر گرفتن مجموعه بزرگی از برچسب‌ها روش‌های یادگیری ماشینی را با مشکل مواجه می‌کند و عملاً آن را غیرممکن می‌کند. بنابراین تعداد

برچسب‌ها بر اساس پردازش‌های آماری به ۴۲ نوع کاهش داده شده است.

۳-۲ پردازش پرسش

پردازش پرسش شامل سه زیرمولفه کلی طبقه‌بندی پرسش، بسط پرسش و انتخاب کلیدواژه‌ها می‌باشد[۴,۲۷]. زیرمولفه طبقه‌بندی پرسش نوع پرسش را مشخص می‌کند به‌طوری که رویکرد جستجو برای جواب، در اسناد محتمل تعیین شود. بسط پرسش با استفاده از شبکه معنایی مترادف و متضاد تولید شده، پرسش‌های جدیدی را تولید می‌کند که از نظر معنایی با پرسش کاربر مشابه است اما در سطح لغوی با پرسش کاربر متفاوت است تا بتواند اسناد محتمل بیشتری را بازیابی کند. در زیرمولفه استخراج کلیدواژه‌ها، کلیدواژه‌های بسط داده شده پرسش کاربر برای استفاده در مولفه بازیابی اطلاعات استخراج می‌شود. در زیربخش‌های بعد سه زیرمولفه ذکر شده شرح داده می‌شوند.

۳-۲-۱ طبقه‌بندی پرسش

منظور از طبقه‌بندی پرسش در این سیستم عملیات نگاشت $g: X \rightarrow C$ است. هر نمونه $x \in X$ یک پرسش است که به یکی از n رده نوع پرسش $\{c_i \in C \mid 1 \leq i \leq n\}$ نگاشت می‌شود. هر کدام از این رده‌ها مبتنی بر محدودیت‌های معنایی از یکدیگر تفکیک شده‌اند. طبقه‌بندی در دو سطح بر اساس نوع جواب پرسش در نظر گرفته شده‌است [۲۵].

در بخش ۳-۵ این طبقه‌بندی توصیف شده است. به صورت کلی دو رویکرد قانون‌گرا و یادگیری برای تشخیص طبقه پرسش وجود دارد که می‌توان از روش Hullو Prager نام برد.[۲۸,۲۹] این دو شامل تعداد زیادی قانون برای تشخیص طبقه پرسش هستند اما در روش یادگیری با مجموعه داده‌های عظیم آموزشی روبرو هستیم که با استفاده از الگوریتم های مختلف یادگیری ماشین، رده پرسش به‌دست می‌آید مانند ماشین بردار پشتیبان، الگوریتم ژنتیک و یادگیری بیزی. همچنین می‌توان از روش‌های یادگیری‌قانون‌گرا استفاده کرد. این روش ترکیبی از دو روش فوق است که می‌توان از یکی از کارهای انجام شده با این روش توسط Silva نام برد [۳۰].

روشی که برای سیستم جویا در نظر گرفته شده استفاده از روش یادگیری است که در این روش از یادگیری بیز ساده به همراه ویژگی‌های لغوی استفاده می‌شود.

۳-۲-۲ بسط پرسش

منظور از بسط پرسش در این سیستم تولید پرسش‌های جدید هم‌معنا و غیریکسان در سطح لغوی با پرسش کاربر است. ممکن است کاربر پرسشی را بپرسد ولی توسط قسمت بازیابی اطلاعات، اسنادی بازیابی شوند که جواب در آن‌ها وجود ندارد به دلیل این‌که ممکن است سوال کاربر باعث تولید کلیدواژه‌هایی شود که اسناد مرتبطی را بازیابی نکند و در نتیجه جوابی دقیق را به کاربر گزارش نکند. به عبارت دیگر با بسط پرسش کاربر، کلید واژه‌های جدیدی برای زیرسیستم بازیابی اطلاعات تولید می‌شود تا بتواند تمامی اسناد مرتبط با پرسش کاربر را بازیابی کند. برای پرسش «پایتخت ایران کجاست» تعداد زیادی ترکیب مختلف هم معنا وجود دارد. جویا چهار بهترین پرسش تولید شده را به‌عنوان بسط پرسش در نظر می‌گیرد. آستانه چهار به‌صورت تجربی به‌دست آمده است. به‌عنوان مثال جدول ۴ را در نظر بگیرید.

بسط پرسش توسط شبکه معنایی مترادف و متضاد انجام می‌شود به این صورت که هر یک از کلیدواژه ها به مترادف خود تا حد دو عمق از گراف تبدیل می‌شود. به عبارت دیگر تمامی کلمات با فاصله معنایی کمتر از دو به عنوان کلمات مترادف با کلید واژه‌های پرسش، برای بسط پرسش در نظر گرفته می‌شود. هر چه مجموع فواصل کلمات پرسش بسط داده شده از کلمات پرسش کاربر کمتر باشد، امتیاز بیشتری خواهد گرفت.

۳-۲-۳ انتخاب کلید واژه‌ها

در این زیرمولفه پرسش تبدیل به کلیدواژه‌ها می‌شود تا سیستم بازیابی اطلاعات بتواند برای بازیابی اسناد از آن‌ها استفاده کند. در ابتدا پرسش توسط تجزیه‌کننده

جدول ۴: بسط پرسش «پایتخت ایران کجاست»

پرسش کاربر	بسط پرسش
پایتخت ایران کجاست	پایتخت ایران کجاست دارالسلطنه ایران کجاست پایتخت پارس کجاست دارالسلطنه پارس کجاست

لغوی به کوچکترین واحد با معنای زبان تجزیه می‌شود و در مرحله بعد کلمات پر تکرار و بی‌فایده از پرسش تجزیه شده حذف می‌شود و باقی مانده به‌عنوان کلید واژه‌های محتمل سیستم بازیابی اطلاعات در نظر گرفته می‌شود. چند نمونه از کلمات پرتکرار و بی فایده در جدول ۵ نمایش داده شده است.

تجزیه‌کننده‌ای ساده برای این سیستم در نظر گرفته شده که فقط عمل جداسازی را براساس علائم نگارشی و فضاهای خالی انجام می‌دهد. جویا برای به‌دست آوردن کلمات پرتکرار و ایست واژه‌ها[6] از مجموعه اسناد گردآوری شده دانشنامه آزاد ویکی‌پدیای‌فارسی و مجموعه داده یادگیری معرفی شده در بخش‌های قبل استفاده می‌کند. برای دستیابی به کلمات پرتکرار و ایست واژه‌ها به‌طور دقیق‌تر در پرسش کاربر، از اسناد مذکور و مجموعه داده معرفی شده در بخش‌های قبل با استفاده از تابع امتیازدهی tf-df که تغییر یافته‌ای از روش tf-idf می‌باشد، استفاده شده است[۳۱]. توابع امتیازدهی فوق به ازای یک واژه و یک سند که به ترتیب با t و d نشان داده شده است، در روابط (۱) و (۲) تعریف شده‌اند.

$$tf-idf_{t,d} = \left(1 + log_{10}^{tf_{t,d}}\right) * log_{10}^{N/df_t} \quad (۱)$$

$$tf-df_{t,d} = \left(1 + log_{10}^{tf_{t,d}}\right) * \left(1 + log_{10}^{df_t/N}\right) \quad (۲)$$

در روابط فوق tf و df به ترتیب نشان دهنده فراوانی واژهt در کل مجموعه اسناد Nو فراوانی اسناد شامل این واژه می‌باشد. به این دلیل از روش تغییر یافته tf-df استفاده می‌شود تا کلماتی که هم در تعداد زیادی سند یا پرسش وجود دارند و در هر سند یا پرسش به تعداد زیادی مشاهده شده‌اند، استخراج شوند. همین‌طور که در روابط (۱) و (۲) مشاهده می‌شود این دو معیار تفاوت چندانی با هم ندارند اما با معیار tf-df نتایج بهتری حاصل شده است.

پس از حذف کلمات پرتکرار و ایست واژه‌ها ممکن است کلمات باقی‌مانده، کلیدواژه‌های مناسبی برای سیستم بازیابی اطلاعات نباشند. به همین دلیل جویا از روش n-gram برای دست‌یابی به کلید واژه‌های مناسب استفاده می‌کند. به این صورت که آستانه‌ای به اندازه 3-gram برای جویا در نظر گرفته شده تا بتواند کلمات باقی مانده از پرسش را به صورت تکی، دو تایی و سه تایی به عنوان کلیدواژه در نظر بگیرد. حد آستانه برابر ۳ بر اساس تجربه به دست آمده است. در آزمایش‌ها مشاهده شد اگر این حد آستانه کمتر از ۳ در نظر گرفته شود، اسناد مرتبطی بازیابی نمی‌شود و اگر بیش از ۳ در نظر گرفته شود، زمان زیادی برای پاسخگویی به پرسش صرف می‌گردد. جویا در حقیقت در این زیرمولفه مجموعه کلیدواژه‌های نامزد را تولید می‌کند تا بتواند به بهترین سند موجود دست پیدا کند. سپس سیستم بازیابی اطلاعات اسناد مرتبط با هر کلیدواژه را گزارش می‌کند و هر یک از کلیدواژه‌های به‌دست آمده بر اساس اسناد بازیابی شده با استفاده از روش tf-sf که در رابطه (۳) نشان داده می‌شود امتیازدهی می‌شوند [۳۳]. کلیدواژه‌های با امتیاز بیشتر به‌عنوان محتمل‌ترین کلیدواژه‌ها گزارش می‌شوند. این تابع امتیازدهی در رابطه (۳) تعریف شده است. این تابع امتیازدهی با معیار رابطه (۱) از لحاظ محاسباتی تفاوتی ندارد و این تغییر اسم تنها به دلیل تغییر نحوه استفاده از آن است. در معیار موجود در رابطه (۱) فرکانس در سند اهمیت دارد اما در این معیار فرکانس در جمله حائز اهمیت است. زیرا پس از ارسال کلیدواژه‌ها به سیستم بازیابی اطلاعات مجموعه اسناد برگردانده می‌شود. حال جملات تمامی اسناد به عنوان سند ذکر شده در رابطه (۱) در نظر گرفته می‌شود.

$$tf-sf_{t,d} = \left(1 + log_{10}^{tf_{t,s}}\right) * \left(1 + log_{10}^{sf_s/M}\right) \quad (۳)$$

6- Stop words

جدول ۵. چند نمونه کلمات پرتکرار و ایست واژه‌های زبان فارسی

و	که
در	این
به	را
از	کدام

در رابطه (۳) tf و sf به ترتیب نشان دهنده فراوانی واژه t در کل جملات M در محتمل ترین اسناد بازیابی شده و فراوانی جملات شامل این واژه می‌باشد.

۳-۳ بازیابی اطلاعات

این مولفه را به دو زیرمولفه واکشی اطلاعات و امتیازدهی جملات محتمل تقسیم‌بندی می‌کنیم که هریک به ترتیب در زیربخش‌های بعد شرح داده می‌شوند.

۳-۳-۱ واکشی اطلاعات

دانش و اطلاعات جویا از دانشنامه آزاد ویکی‌پدیای فارسی استخراج می‌شود، به این صورت که کلیدواژه‌های تولید شده در زیربخش قبلی به سیستم بازیابی اطلاعات ویکی‌پدیا داده می‌شود تا محتویات محتمل‌ترین اسناد بازیابی شود. به عبارت دیگر این زیرمولفه پس از تشخیص بهترین کلیدواژه در زیرمولفه پیشین اجرا می‌شود تا بهترین محتویات مرتبط با پرسش کاربر بازیابی شود.

۳-۳-۲ امتیازدهی جملات محتمل

محتمل‌ترین اسناد بازیابی شده در زیربخش قبلی براساس جملات مجزا می‌شوند و به هریک از جملات امتیازی به میزان محتمل بودن حضور جواب در آن‌ها، داده می‌شود. سیستم جویا برای امتیازدهی از ضریب جاکارد [۳۱] استفاده می‌کند. اگر پرسش کاربر را با A و جمله مورد بررسی را B در نظر بگیریم این ضریب با استفاده از رابطه (۴) به‌دست می‌آید:

$$JaccardCoefficient = {|A \cap B|}/{|A \cup B|} \qquad (۴)$$

۳-۳-۳ استخراج جواب دقیق

این مولفه براساس رده پرسش به‌دست آمده از مولفه پردازش پرسش، راهبرد استخراج جواب از مجموعه اسناد محتمل را مشخص می‌کند[۴,۲۵,۲۷]. به‌عبارت دیگر برای هر طبقه پرسش بیان شده در جدول ۴، مشخص می‌کند که چگونه جواب کوتاه و دقیق را به‌دست آورد. اگر درشت‌دانه رده پرسش یکی از درشت‌دانه‌های مخفف، موجودیت، توصیف، مکان، بشری و یا عدد باشد، راهبرد استخراج جواب، به کارگیری برچسب‌گذاری نحوی و طبقه‌بندی موجودیت با نام خواهد بود. اگر رده پرسش از نوع لیست یا بله‌خیر باشد در این نسخه از سیستم، جویا نمی‌تواند جواب دقیق را پاسخگو باشد و تنها مجموعه ای از جملات محتمل گزارش می‌شود. به عبارت دیگر جویا در این نسخه به پرسش‌های در مورد حقیقت پاسخ می‌دهد.

۳-۴ برچسب‌گذاری نحوی

به‌طور کلی برای برچسب‌گذاری سه رویکرد قانون مبنا[7]، آمار مبنا[8] و ترکیبی در نظر گرفته می‌شود که برای هر رویکرد تا کنون روش‌های مختلفی جهت برچسب‌گذاری کلمات در زبان فارسی، انگلیسی و غیره ارائه شده است. یکی از روش‌هایی که در رویکرد آمارمبنا برای برچسب‌گذاری به کار گرفته می‌شود و سیستم پرسش و پاسخ جویا نیز از آن استفاده می‌کند، روش تخمین احتمال بیشینه[9] است.

در مجموعه داده بی‌جن‌خان هر کلمه لزوماً برچسب منحصر به فردی ندارد. به‌عبارت دیگر، یک کلمه می‌تواند نقش‌های متفاوتی در جمله داشته باشد. به همین دلیل در این پژوهش پیش‌پردازش‌هایی بر روی این مجموعه داده صورت گرفت تا کلمات موجود در مجموعه داده دارای برچسب منحصر به فرد باشند به‌طوری‌که این برچسب بیشترین تکرار را برای آن کلمه در کل مجموعه داشته باشد. پس از پیش‌پردازش مجموعه داده آموزشی بی‌جن‌خان این مجموعه داده عظیم به ۷۶۹۷۷ کلمه به همراه برچسب نحوی منحصر به فرد تبدیل شده است. طبق بررسی‌های آماری انجام شده [۲۰] برچسب N_SING که نشان دهنده

7- Rule based
8- Stochastic based
9- Maximum likelihood estimation

اسم مفرد اســت دارای ۸۲۶۵۷۱ فراوانی در مجموعه داده آموزشــی می‌باشد که نسبت به ســایر برچسب‌ها دارای بیشترین فراوانی اســت. با انتساب این برچسب به کلمات ناشناخته، برچســب‌گذاری نحوی به روش تخمین احتمال بیشینه درصد صحت ۹۳/۱۶ را کسب کرده است.

سیستم جویا پس از استخراج جملات محتمل، شروع به برچسب‌گذاری نحوی آن‌ها به روش تخمین احتمال بیشینه می‌کند. تمامی کلماتی که دارای برچسب‌های مرتبط با رده پرسش کاربر می‌باشند به زیرمولفه طبقه‌بندی موجودیت با نام برای استخراج جواب کوتاه و دقیق داده می‌شوند.

۳-۵ طبقه‌بندی موجودیت با نام

روش‌های متفاوتی برای تشــخیص طبقه موجودیت با نــام وجــود دارد که می‌توان از [۳۲] بــه عنوان یک روش فرهنــگ لغت‌مبنــا نام بــرد. روش‌هایی بــرای طبقه‌بندی موجودیت با نام با اســتفاده از ویکی‌پدیا ارائه شده است که می‌توان این دســته را به‌عنــوان روش‌های آمارمبنا در نظر گرفت. همچنین می‌توان به [۳۴] اشــاره کرد که علاوه بر استفاده از نقش نحوی کلمات، از یک ویژگی فهرستی و پیشوند و پسوند کلمات نیز برای طبقه‌بندی استفاده می‌کند. این سیســتم در انتها توســط چندین الگوریتم طبقه‌بندی آموزش داده شده است.

پس از برچســب‌گذاری ادات سخن تمامی اسم‌های یک جمله استخراج می‌شوند و به طبقه‌بندی اسامی برای تعیین نوع اسم داده می‌شود. سیستم جویا برای تعیین نوع اسم از ویکی‌پدیا استفاده می‌کند به این صورت که محتوای هر یک از اســامی را بازیابی می‌کند و رده‌های این اسم را از محتوای به‌دست آمده استخراج می‌کند. تعداد تکرار هریک از کلمات در رده‌ها را به‌دست می‌آورد و پرتکرارترین کلمه را به عنوان یک طبقه اولیه برای اســم در نظر می‌گیرد. اگر تعداد تکرار چندین کلمه نزدیک به هم باشــد تعداد رخداد هر یک از کلمه ها در متن محاسبه می‌شود و پرتکرارترین آن‌ها به عنوان طبقه اولیه انتخاب می‌شــود. با استفاده از شبکه واژگان مترادف و متضاد تولید شده میزان نزدیکی طبقه اولیه با طبقه‌های موجود در جدول ۴ بررسی می‌شود و نزدیک‌ترین طبقه، به‌عنوان طبقه نهایی گزارش می‌شود.

۴- آزمایش‌ها و تحلیل نتایج

برای ارزیابی سیستم پرســش و پاسخ جویا که برای پاســخگویی به پرســش‌ها در تمام مقولات طراحی شده است از ۴۰ پرســش به عنوان آزمون استفاده شده است. نتایج حاصل از ارزیابی مجموعه پرســش‌های جمع آوری شده، صحت پاسخگویی ۸۰٪ را برای سیستم جویا نشان می‌دهد. در این بخش دقت و بازخوانی برای مولفه پردازش سوال، بازیابی اطلاعات و زمان اجرا در زیربخش‌های آتی مورد بررسی قرار می‌گیرد.

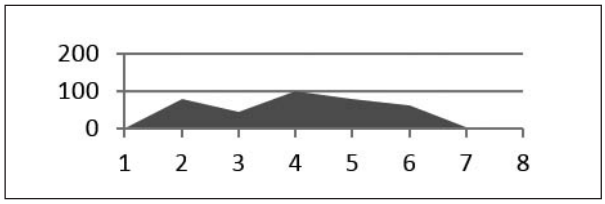
٤-۱- پردازش پرسش

آزمایش مولفه پردازش پرســش بــرای تعیین صحت دســتیابی به طبقه پرسش در جدول ۶ انجام شده است. به دلیل عدم توزیع یکنواخت، برای ارزیابی این سیستم هشت درشــت‌دانه موجود در جدول ۴ بررسی شده‌اند. این عدم توزیع یکنواخت در شکل ۴-الف نمایش داده شده است. در شکل ۴-ب توزیع یکنواختی بر اساس درشت‌دانه‌ها نشان داده شده است. صحت پاسخگویی و بازخوانی را برای هر درشت‌دانه در جدول محاسبه شده است. برای دستیابی به صحت پاسخگویی و بازخوانی این مولفه از میانگین وزنی استفاده شده است. در نتیجه صحت پاسخگویی این مولفه ۵۲٪ می باشد.

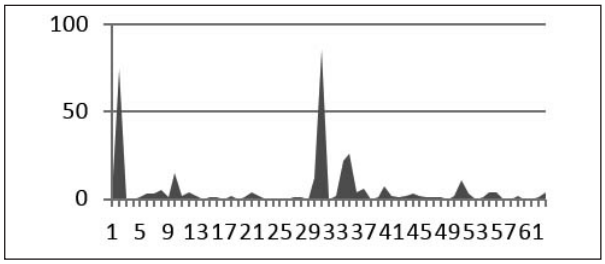
٤-۲- بازیابی اطلاعات

آزمایش مولفه بازیابی اطلاعات برای دستیابی به میزان صحت جملات محتمل بازیابی شده در جدول ۷ انجام شده اســت. چند نمونه از پرسش‌های آزمون سیستم پرسش و پاســخ جویا در جدول ۸ نمایش داده شــده است. صحت پاسخگویی این مولفه ۸۰٪ می باشد.

سیســتم پرســش و پاســخ جویا تحت سیستم عامل ویندوز با ســرعت اینترنــت 256kbps شــرکت مخابرات



شکل ۴: الف. عدم توزیع یکنواخت بر اساس ریزدانه‌ها



شکل ۴: ب. توزیع تقریبا یکنواخت و مناسب برای یادگیری

جدول ۶: آزمایش مولفه پردازش پرسش

تعداد پاسخ درست	تعداد پرسش آزمون	درشت دانه	تعداد پاسخ درست	تعداد پرسش آزمون	درشت دانه
۱	۶	عدد	۸	۱۰	بشری
–	–	لیست	۷	۷	توصیف
۰	۴	مخفف	۰	۶	موجودیت
–	–	بله_خیر	۵	۷	مکان

جدول ۷. آزمایش مولفه بازیابی اطلاعات

تعداد پاسخ درست	تعداد پرسش آزمون	درشت دانه	تعداد پاسخ درست	تعداد پرسش آزمون	درشت دانه
۴	۶	عدد	۱۰	۱۰	بشری
–	–	لیست	۶	۷	توصیف
۲	۴	مخفف	۵	۶	موجودیت
–	–	بله_خیر	۵	۷	مکان

جدول ۸: چند نمونه از پرسش‌های آزمون

بزرگترین حیوان چه نام دارد	دانشگاه تهران در کدام خیابان است
علی اکبر دهخدا بنیان گذار چیست	کهن‌ترین قاره جهان کجاست
رویداد خارجی مشخص و قابل شناسایی چه نام دارد	دیوار چین چند متر طول دارد
ورزشی جسمی و روانی است	روز مهندس در چه تاریخی است

آزمایش شده اســت. متوسط زمان اجرا برای پاسخگویی ۴۰ پرسش آزمایش ۲۳/۶ ثانیه به‌دست آمده است.

۵- نتیجه‌گیری و کارهای آتی

بــا افزایش روزافزون حجم داده‌ها و تنوع ســاختاری آن‌ها وجود سیســتم‌های بازیابی اطلاعات اهمیت ویژه ای پیدا کرده اســت. برای بهبود عملکرد سیستم‌های بازیابی اطلاعات و ارتباط آسان تر کاربران، سیستم‌های پرسش و پاسخ مطرح شدند.

در ایــن مقاله جویا به عنوان یک سیســتم پرســش و پاسخ فارسی که دامنه نامحدود و وب‌مبنا می‌باشد، معرفی گردید. در این سیستم مجموعه داده‌های مفیدی تولید شده و مورد اســتفاده قرار گرفت و همچنین قابل ذکر است که مجموعه داده‌های از پیش ساخته شده‌ای برای اولین بار در سیستم‌های پرسش و پاسخ مورد کاربرد قرار گرفت. این سیستم با هدف اســتقلال از زبان تولید شده است. به این معنا که با صرف هزینه‌ای بســیار کم بتوان از این سیستم برای زبان‌های دیگر اســتفاده کرد. بــه عبارت دیگر کافی

اســـت مجموعه داده‌ها را تغییر داد و از مجموعه داده‌هایی مرتبط با زبان مورد نظر استفاده کرد. در نسخه بعدی این سیستم، در رویکرد پاســـخگویی به پرسش‌ها، از دو رده پرسش لیست و بله خیر در نظر گرفته می‌شود و همچنین پرسش‌هایی که پاسخگویی به آن‌ها نیازمند استدلال است، مورد بررسی قرار خواهد گرفت. تعداد رکوردهای پرسش-طبقه در حال حاضر ۴۱۲ می‌باشد و این پرسش‌ها به طور غیریکنواخت در ریزدانه‌ها توزیع شده‌اند. برای دستیابی به دقت بیشتر در کارهای آینده تعداد نمونه‌های این مجموعه داده افزایش خواهد یافت و به طور یکنواخت در هر ریزدانه توزیع خواهد شـــد. همچنین متوسط زمان اجرا را با بهبود الگوریتم‌ها کاهش خواهیم داد.

قدردانی

از حمایت مالی پارک علم و فناوری دانشـــگاه تهران از این تحقیق در قالب اعتبار شماره ۹۴۰۳۰ قدردانی می‌گردد.

مراجع

[۱]کرمانشـــاهی نیره، عبدالکریمی حسین، شـــمس‌فرد مهرنوش، ساخت یک سیســـتم پرســـش و پاسخ برای کیوســـک اطلاعات دانشـــکده مهندسی برق و کامپیوتر، نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، ۱۳۹۲.

[2]Yao, Xuchen and Benjamin Van Durme, Information extraction over structured data: Question answering with freebase, In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014, pp. 956-966.

[3]Baayen.H, et al. "Advances in Open Domain Question Answering", Published by Springer, P.O. Box 17, 3300 AA Dordrecht, The Netherlands, 2008.

[4]Gupta, P., Gupta, V., A survey of text question answering techniques, International Journal of Computer Applications, 2012, vol. 53, no. 4, pp. 1-8.

[5]Ferrucci, D., et al, Building Watson: an overview of the DeepQA project, AI Magazine, 2010, vol. 31, no. 3, pp. 59–79.

[6]Ferrucci, D., Levas, A., Bagchi, S., Gondek, D. and Mueller, E. T., Watson: Beyond jeopardy!, Artificial Intelligence, 2013,vol. 199, no. 10, pp. 93-105

[7]Green, B.F., Wolf, A.K., Chomsky, C., Laughery, K., BASEBALL: An automatic question answerer, In Proceedings of Western Computing Conference, 1961, vol. 19, pp. 219–224.

[8]Woods, W.A., Kaplan, R.A., Nash-Webber, B., The lunar sciences natural language information system, Technicalreport, Bolt Beranek and Newman Inc.,Cambridge, MA., 1972.

[9]Kolomiyets, O., Moens, M. , A survey on question answering technology from an information retrieval perspective, Information Sciences, 2011, vol. 181, no. 24, pp. 5412-5434.

[10]Paris, C., Towards More Graceful Interaction: A Survey of Question Answering Programs, Columbia University Computer Science Technical Reports, 1985.

[11]Katz, B., Borchardt, G. and Felshin, S. , Natural Language Annotations for Question Answering, In Proceedings of the 19th International FLAIRS Conference , 2006.

[۱۲]حجـــازی محمدرضـــا، میریان حســـین آبادی مریم ســـادات، نشـــاطیان کوروش، افقی بهادررضا، درودی احســـان، سیســـتم پرسش و پاسخ مبتنی بر هستان‌شناســـی برای حوزه مخابرات با قابلیت استخراج و دسته‌بندی خودکار مستندات، نشریه علمی پژوهشی انجمن کامپیوتر ایران، ۱۳۸۳.

[13]Liu, J., Liu, Li-zhen, and Peng, Y., Answer Extraction of Chinese Restricted Domain Question Answering System Based on Ontology, Journal of Computational Information Systems, 2010, vol. 6, no.1, pp. 155-165.

[14]B. Hammou, H. Abu-salem, S. Lytinen, and M. Evens. "QARAB: A question answering system to support the Arabic language". In Proceedings of the workshop on computational approaches to Semitic languages, ACL, 2002.

[15]Perret, Laura. "A question answering system for French", Workshop of the Cross-Language Evaluation Forum for European Languages. Springer Berlin Heidelberg, 2005.

[۱۶]شـــمس فرد مهرنوش، اشراق فائزه و ســـارابی زهرا، ساخت یک سیستم پرســـش و پاسخ به زبان فارسی، دوازدهمین کنفرانس سالانه انجمن کامپیوتر، ۱۳۸۵.

[۱۷]امیر محسن یوسفی واقف. «طراحی و پیاده سازی سیستم پرسش و پاسخ در یک دامنه محدود»، وزارت علوم، تحقیقات و فناوری ، موسسه آموزش عالی غیر انتفاعی و غیر دولتی نبی اکرم، دانشکده فنی، کارشناسی ارشد، ۱۳۹۱.

[۱۸]برشـــبان یاسمن، یوسفی‌نسب حامد، میرروشندل سید ابوالقاسم، رسائل و مسائل: توسعه یک پیکره متنی فارسی پرسش و پاسخ، بیستمین کنفرانس ملی سالانه انجمن کامپیوتر ایران، ۱۳۹۳.

[۱۹]خداپرســـتی فرج‌الله، فرهنگ جامع واژگان مترادف و متضاد زبان فارسی (نسخهٔ رقومی)، دبیرخانه شورای عالی اطلاع‌رسانی، ۱۳۹۱.

[20]Amiri, Hadi, Hojjat, Hosein, Oroumchian, Farhad , Investigation on a Feasible Corpus for Persian POS Tagging, 12th international CSI computer conference, 2007.

[21]Hermjakob, Ulf, Hovy, Eduard and Lin, Chin yew, Automated question answering in webclopedia: a demonstration, In Proceedings of the second international conference on Human Language Technology Research, 2002.

[22]Li, Xin and Roth, Dan, Learning question classifiers, In Proceedings of the 19th international conference on Computational linguistics, 2002.

[23]Loni, Babak, A survey of state-of-the-Art Methods on Question Classification, Journal Article, Literature Survey, Published on TU Delft Repository, 2011.

[24]Metzler, Donald and Croft, W. Bruce, Analysis of statistical question classification for fact-based questions, Information Retrieval, 2005, vol. 8, no. 3, pp. 481-504.

[25]Mollaei, Ali, Rahani, Saeed, Estaji, Azam, Question classification in Persian language based on conditional random fields, Second International eConference Computer and Knowledge Engineering, 2012, pp. 295-300.

[26]Marcus, M., Santorini, B., Ann, M., Building a large an-

notated corpus of English: The Penn Treebank, Computational Linguistics, 1993, vol.19, no. 2, pp. 313-330.
[27]Quarteroni, S., Manandhar, S., Designing an interactive open-domain Question Answering system, Natural Language Engineering, 2009, vol. 15, no. 1, pp. 73-95.
[28]Prager, John, Brown, Eric, Coden, Anni, Radev, Dragomir, Question-answering by predictive annotation, In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000.
[29]Hull, David A., Xerox TREC-8 question answering track report, In Proceedings of the 8th Text Retrieval Conference, 1999.
[30]Silva, Joao, Coheur, Luısa, Mendes, Ana, and Wichert, Andreas, From symbolic to sub-symbolic information in question classification, Artificial Intelligence Review, 2011, vol. 35, no. 2, pp. 137-154.
[31]Manning, Christofer D., Raghavan, Prabhakar, Hinri, An Introduction to Information Retrieval, Cambridge university press, 2009.
[32]Kolali Khormuji, Morteza, Bazrafkan, Mehrnoosh, Persian Named Entity Recognition based with Local Filters, International Journal of Computer Applications, 2014, vol. 100, no. 4.
[۳۳] عربی نرئی، سمیه، وحیدی اصل، مجتبی و مینایی بیدگلی، بهروز. استخراج کلمات کلیدی جهت طبقه‌بندی متون فارسی، اولین کنفرانس داده کاوی ایران، تهران، دانشگاه صنعتی امیرکبیر، موسسه پژوهشی داده پردازان گیت، ۱۳۸۶.
[۳۴]اصفهانی، سید عبدالحمید، راحتی قوچانی، سعید و جهانگیری، نادر، سیستم شناسایی و طبقه‌بندی اسامی در زبان فارسی، پردازش علائم و داده‌ها، ۱۳۸۹.