علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۲، شماره ۱و۳ (الف)، بهار و پائیز ۱۳۸۳
صفحات ۱-۱۷
مقاله عادی

سیستم پرسش و پاسخ مبتنی بر هستان شناسی برای حوزه‌ی مخابرات با قابلیت استخراج و دسته‌بندی خودکار مستندات

محمودرضا حجازی مریم سادات میریان حسین‌آبادی کوروش نشاطیان بهادررضا افقی
احسان درودی

گروه کاربردهای فناوری اطلاعات، پژوهشکده فناوری اطلاعات، مرکز تحقیقات مخابرات ایران، تهران، ایران

چکیده

در این مقاله، یک سیستم پرسش و پاسخ[1] مبتنی بر هستان شناسی[2] که برای حوزه مخابرات طراحی و نمونه‌سازی شده‌است، مورد بررسی قرار می‌گیرد. این سیستم نمونه TeLQAS[3] نام داشته و از دو فرآیند نسبتا مستقل برخط و برون‌خط تشکیل شده است. در بخش برخط، سیستم پرسشهای کاربران را به زبان انگلیسی دریافت کرده و به کمک استدلال روی گراف هستان شناسی پاسخ دقیق استخراج و به همراه پاراگرافهای خلاصه‌سازی شده مرتبط در اختیار کاربران قرار می‌دهد. در بخش برون خط، سیستم با استفاده از یک مکانیسم رسته‌سازی متن، مستندات مرتبط به مفاهیم حوزه را از مجموعه‌های موجود، نظیر وب و انباره‌ی متن داخلی، بصورت اتوماتیک استخراج کرده و طبقه‌بندی می‌کند. نتایج بدست آمده از به‌کارگیری این سیستم برای پاسخ به سوالات آزمایشی در حوزه تخصصی مخابرات فیبر نوری گواه عملکرد چشمگیر آن است، ضمن اینکه دقت سیستم با طرح پرسشهای بیشتر افزایش می‌یابد. گسترش این سیستم به سایر حوزه‌های تخصصی با ایجاد هستان شناسی مربوطه به سادگی امکان‌پذیر است.

کلمات کلیدی: سیستم پرسش و پاسخ، بازیابی اطلاعات، هستان شناسی، خلاصه‌سازی متن، کلاسه‌بندی

۱- مقدمه

گرچه سیر پیشرفت فن‌آوریهای بازیابی اطلاعات از رشد نسبتا خوبی در حوزه علوم کامپیوتر برخوردار بوده‌است ولی هنوز فاصله زیادی با توقعات کاربران اطلاعات دارد. هم‌اکنون بیشتر سیستم‌های بازیابی اطلاعات که قابلیت استفاده عملی در مقیاس کلان را دارند، به صورت موتورهای جستجو و در قالب ترکیبی از عامل‌های نمایه‌سازی[4] هستند. این دسته از سیستم‌ها (به عنوان مثال Google) عمل استخراج مستندات را بر اساس کلمات کلیدی مورد نظر کاربر انجام داده و مجموعه بزرگی از مستندات را که از لحاظ کلیدواژه‌ای شانس بیشتری را برای مرتبط بودن با نیاز کاربر دارند، به او ارائه می‌کنند. در نهایت این کاربر است که باید با مرور این مجموعه مستندات، جواب اصلی خود را (در صورت وجود) استخراج نماید. خیلی از اوقات مستندات بازیابی شده تفاوت اساسی با منظور اصلی کاربر دارند و با توجه به این موضوع که در اصل، نیاز اطلاعاتی کاربران بصورت یک پرسش است و نه کلمات کلیدی، لازم است تا کاربران تجربه و مهارت کافی در تبدیل یک سؤال به چند کلمه کلیدی را داشته باشند. در مقابل این فن‌آوری، یک سیستم پرسش و پاسخ باید قادر باشد تا سوالات کاربران را بصورت یک پرسش در زبان طبیعی دریافت کرده و با حداقل افزونگی و حداکثر دقت، پاسخ را تولید نماید.

تا بحال تلاش زیادی در جهت ساخت سیستم‌های پرسش و پاسخ به عمل آمده و نمونه‌های زیادی نیز تولید شده است. هرچند تابحال اکثر آنها در یک مقیاس عمده بکار نرفته‌اند ولی به پیشرفتهای خوبی در این زمینه نائل آمده‌اند. از یک دیدگاه می‌توان سیستم‌های پرسش و پاسخ را از نظر حوزه فعالیتشان به دو دسته عمومی و تخصصی تقسیم کرد. سیستم‌های حوزه عمومی[5] که هر ساله تحولات آنها در همایش TREC[6] منعکس می‌شود، طبق تعریف باید قادر باشند تا با رجوع به یک مجموعه متنی بزرگ از پیش تعیین شده، سوالات عمومی مربوط به آن را پاسخ گویند. در مقابل، سیستم‌های پرسش و پاسخ حوزه‌های خاص[7]، همانطور که

منابع خاص یک حوزه تخصصی مانند اطلاعات موجود در یک کتابخانه دیجیتال محلی، کاربرگهای فنی[۸] و امثال آن نیز استفاده کنند.

کارگزار MELISA[۹] یک نمونه خوب برای سیستم‌هایی است که در حوزه تخصصی کار می کنند[۱]. این سیستم برای حوزه پزشکی پیشنهاد شده است. MELISA مبتنی بر هستان شناسی است و به گونه ای طراحی شده که قابلیت تطبیق با منابع پزشکی را داشته باشد. مهمترین ویژگی‌های طراحی این سیستم، استفاده از یک معماری سه لایه بصورت انتزاعی[۱۰]، بکارگیری هستان شناسی، تعریف چند مدل پرس و جوی جداگانه و نیز تعریف چند اپراتور تجمع[۱۱] است. این روش، مبتنی بر توسعه یک سیستم با قابلیت آموزش برای استخراج اطلاعات است. سیستم ارائه شده، برای انجام عملیات نیاز به دو ورودی دارد: اولین ورودی، یک هستان شناسی از مفاهیم و ارتباطات بین آنهاست و دومی، مجموعه ای است از داده های آموزشی شامل ناحیه های مشخص شده توسط متون Hyper Text که نمونه هایی[۱۲] از مفاهیم هستان شناسی را مشخص می‌کنند.

در این مقاله، سیستم پرسش و پاسخ پیشنهادی TeLQAS مورد بررسی قرار می گیرد. این سیستم می تواند به پرسشهای مطرح شده در حوزه تخصصی مخابرات پاسخ دهد. پایگاه اطلاعات سیستم[۱۳]، مفاهیم و واقعیات اطلاعات مربوط به حوزه تخصصی مورد نظر است که در این سیستم، در قالب گراف هستان شناسی ذخیره شده و شامل مفاهیم، ارتباطات بین آنها و همچنین مستندات مرتبط است. مزیت برجسته سیستم پیشنهادی ما در مقایسه با سیستم های پرسش و پاسخی که در حوزه های عمومی فعالیت می‌کنند (سیستم هایی که فاقد یک پایگاه دانش تخصصی در یک حوزه خاص هستند[۲,۳,۴])، استفاده از یک هستان شناسی بعنوان محور اصلی کلیه فعالیت های آن است. به این ترتیب سیستم می‌تواند بصورت معنایی[۱۴] و با دقتی بالاتر به سوالات کاربران در حوزه ای که هستان شناسی برای آن طراحی شده است، پاسخ دهد. علاوه بر آن، در صورت موجود نبودن مستندات مناسب (یا کافی) در مورد سؤال کاربر در پایگاه داده، سیستم می‌تواند مستنداتی را با استفاده از تعدادی جویشگر استخراج کرده و پس از رسته-سازی، در صورتی که این مستندات مناسب باشند، آنها را برای استفاده های بعدی، به پایگاه اطلاعات اضافه کند. شایان ذکر است در صورت موجود بودن هستان شناسی های مناسب، می‌توان آنها را جایگزین هستان شناسی فعلی کرده، زمینه تخصصی سیستم را عوض کرد یا بسادگی با اضافه کردن آن به هستان شناسی فعلی قدرت و زمینه کاربردی آن را افزایش داد.

۲- معماری سیستم TeLQAS

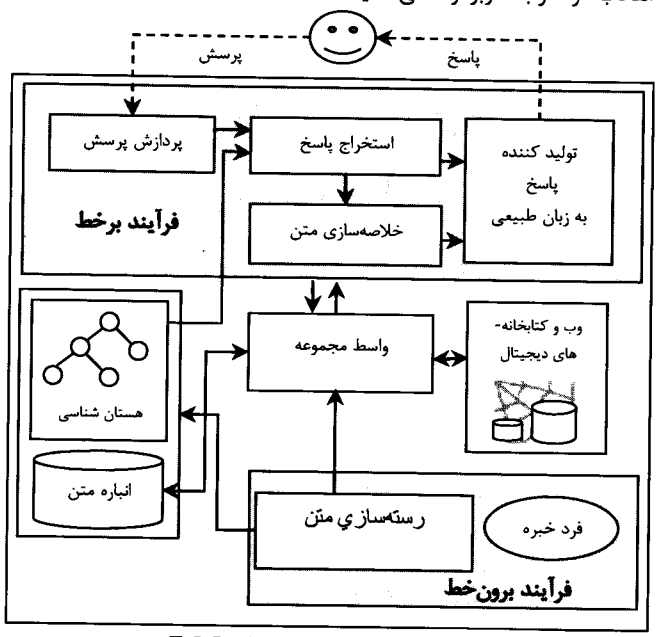
با توجه به طرحی که در نظر داریم، می‌توان بخشهایی چون هستان شناسی، مولفه بازیابی اطلاعات و مولفه‌های اتصال به اینترنت را از اجزاء حتمی معماری TeLQAS دانست. هرچند تکمیل معماری کنونی TeLQAS یک روند تکاملی داشته و آنچه در اینجا ارائه می‌شود حاصل چندین بار نمونه‌سازی، آزمایش و پالایش طرح بوده است، اما آنچه چهارچوب و اسکلت اصلی آنرا تعیین می‌کند، تعریف وظایف این سیستم است. به عنوان یک سیستم پرسش و پاسخ، انتظار داریم این سیستم خدمات زیر را ارئه دهد:

۱) پرسش کاربران را به زبان طبیعی دریافت کرده و پس از پردازش و رجوع به پایگاه دانش سیستم بتواند پاسخ متناظر با آن را تولید کند.

۲) در صورت فراهم نبودن اطلاعات لازم در هستان شناسی بتواند با رجوع به اینترنت، مجموعه مستندات محلی و سایر منابع، مناسبترین بند حاوی پاسخ را در اختیار کاربر قرار دهد.

۳) قابلیت رسته‌بندی خودکار مستندات را داشته باشد یعنی بتواند با گذشت زمان و جمع آوری اطلاعات بیشتر عملکرد خود را بهبود بخشد.

در شکل ۱، نمای کلی TeLQAS آورده شده است. یکی از نوآوریهای مهم در طراحی TeLQAS استفاده توامان از بسیاری از روش‌ها و تکنیکهای بازیابی اطلاعات است که تابحال عمدتا با آنها بصورت مستقل برخورد شده است. از نظر اجرای فرایندها می‌توان TeLQAS را به دو بخش برخط[۱۵] و برون‌خط[۱۶] تقسیم کرد. در شکل ۱، این دو بخش مشخص شده‌اند. فرایند بخش برخط با فراخوانی کاربر (یعنی ارائه پرسش) به اجرا در می‌آید. همانطور که جلوتر به تفصیل خواهد آمد، پرسش کاربر پس از پردازش و استخراج نوع آن، به بخش استخراج پاسخ فرستاده خواهد شد. اگر پاسخ به صورت استدلالی و از پایگاه اطلاعات (به صورت مستقیم) استخراج شود، پاسخ نهایی را مولفه تولید کننده زبان طبیعی سنتز خواهد کرد و چنانچه با استفاده از منابع متنی اینترنتی یا محلی در صدد پاسخگویی به پرسش کاربر باشیم، واحد خلاصه سازی متن، مناسبترین جملات را انتخاب کرده و به کاربر ارائه می نماید.



شکل۱- نمای کلی سیستم TeLQAS

خدمات برخی از مولفه‌های TeLQAS مانند واسط مجموعه، هستان شناسی و انباره متون در هر دو بخش برخط و برون خط استفاده می‌شود. به همین دلیل در شکل ۱ آنها را بصورت مستقل از فرآیندها آورده‌ایم. وظیفه مولفه «واسط مجموعه» اتصال به اینترنت و سایر منابع محلی جهت استخراج مستندات مربوط به حوزه تخصصی است. این واکشی مستندات ممکن است بنابر در خواست فرآیند برخط، برون خط و یا جزئی از روال عادی کار این مولفه باشد. مستندات بدست آمده توسط این مولفه، در انباره مستندات[۱۷] سیستم ذخیره می‌شود. مولفه هستان شناسی زیرساخت لازم برای ذخیره‌سازی و نمایش مفاهیم یک حوزه تخصصی و روابط بین آنها را ارائه می‌کند. دسترسی سایر مولفه‌های سیستم به هستان شناسی و انباره مستندات از طریق این مولفه صورت می‌گیرد.

مهمترین مولفه زیرسیستم برون خط، مولفه رسته‌سازی خودکار متون است، این زیرسیستم دارای قابلیت انتخاب خودکار ویژگی و نیز قابلیت یادگیری با استفاده از مجموعه‌ای از مستندات آموزشی دارد. وظیفه اصلی این زیرسیستم طبقه‌بندی خودکار متون استخراج شده توسط واسط مجموعه است. در ادامه به شرح این دو زیر سیستم و مولفه‌های مربوط به آنها پرداخته، وظایف و نحوه عملکرد هر یک از آنها را توضیح می‌دهیم. ولی در ابتدا مولفه‌ی هستان شناسی را که نقش محوری بین این دو زیر سیستم دارد بررسی میکنیم.

۳- بخش هستان شناسی در TeLQAS

بخش زیادی از فعالیتهای کنونی در حوزه بازیابی اطلاعات در خصوص استفاده از ساختارهای مفهومی در سیستم‌های پرسش/پاسخ است[۵]. از جمله

شناخته‌شده‌ترین ساختارهای مفهومی می‌توان از هستان شناسی و گراف مفاهیم یادکرد. اینگونه ساختارها می‌توانند به عنوان یک پایگاه دانش درونی برای یک سیستم پرسش/پاسخ، نگاشتی از واقعیات دنیای بیرون به اطلاعات بازیابی شده باشند.

در علوم کامپیوتر، تعاریف مختلفی برای هستان شناسی ارائه شده است. در حوزه هوش مصنوعی، هستان شناسی مجموعه‌ای از تعاریف رسمی برای مفاهیم یک حوزه مورد نظر و روابط بین آنها می‌باشد[۷]. استفاده از هستان شناسی و شبکه‌های معنایی یکی از مهمترین راه‌های غنی‌سازی سیستم‌های بازیابی اطلاعات به ویژه سیستم‌های پرسش/پاسخ است. در سالهای اخیر، گراف هستان شناسی بعنوان یکی از راهکارهای مناسب در بازنمایی حوزه های کاربری مورد استفاده قرار گرفته است. یک هستان شناسی به همراه نمونه‌هایی که برای کلاسهایش تعریف شده، تشکیل یک پایگاه دانش را برای حوزه مربوطه می‌دهد. در بعضی از دیدگاه-ها، جایی که هستان شناسی خاتمه می‌یابد، پایگاه دانش شروع می‌شود. علت این امر محدود کردن هستان شناسی به تعریف کلاسها و خصوصیات آنهاست. در بسیاری از مدل‌های امروزی، محدودیتی برای هستان شناسی متصور نیستیم و همواره سعی بر آن است که هستان شناسی در یک روند تکاملی و به مرور زمان به سوی یک پایگاه دانش سوق پیدا کند. برای استفاده از هستان شناسی درسیستمهای بازیابی اطلاعات لازم است تا سه نیاز اصلی در این خصوص برآورده شود:

۱) ایجاد هستان شناسی از طریق استخراج مفاهیم حوزه مورد نظر و تشخیص روابط حاکم بر آنها در قالب یک گراف. برای انجام این کار، نیاز به کارشناسانی است که علاوه بر آشنایی با حوزه مورد نظر، با اصول ایجاد هستان شناسی نیز آشنا باشند. در TeLQAS این کار توسط متخصصین حوزه‌های مخابرات سیار و مخابرات نوری جهت ایجاد دو هستان شناسی در زمینه‌های مذکور انجام شده است.

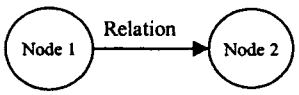
۲) طراحی و ایجاد محیطی جهت بیان و ارتباط با هستان شناسی به نحوی که سایر زیرسیستم‌ها بتوانند بدون در نظر داشتن جزئیات ذخیره سازی، در بالاترین سطح انتزاع به عناصر هستان شناسی دسترسی داشته باشند. جهت تحقق بخشیدن به این نیاز، یک معماری سه لایه با ویژگیهایی منحصر بفرد تولید شده که در ادامه به شرح آن می‌پردازیم.

۳) ایجاد واسط (واسطهای) مناسب جهت ورود و نمایش هستان شناسی. در TeLQAS علاوه بر ایجاد واسطهای گرافیکی کاربر، با طراحی مبدل‌های لازم این امکان فراهم شده است که بتوان از ابزارهای استانداردی که در این زمینه وجود دارد نیز استفاده کرد. در حال حاضر کلیه هستان شناسی‌های مورد نیاز TeLQAS می‌توانند در محیط Protégé [۸] ایجاد و ویرایش شوند.

۳-۱ زیرساخت هستان شناسی

آنچه باعث شده تا TeLQAS دارای زیرساختی منحصر بفرد برای ایجاد هستان شناسی باشد این است که هستان شناسی TeLQAS باید همزمان چندین منظور مختلف را برآورده سازد. اول اینکه باید بتواند هر نوع اطلاعاتی را که یک هستان شناسی در حالت عادی ممکن است دارا باشد شامل شود و نقش یک پایگاه دانش را در سیستم اجرا کند. دوم آنکه اطلاعات این هستان شناسی باید به گونه‌ای عرضه شود که شکلی بهینه برای پردازش یک سیستم پرسش/پاسخ مانند TeLQAS داشته باشد و بالاخره آنکه در سیستم‌های پرسش و پاسخ و بازیابی اطلاعات[۱۸] دسترسی به مستندات بر اساس ارتباطشان با مفاهیم هستان شناسی از اهمیت به سزایی برخوردار است[۹]. هستان شناسی TeLQAS این خدمت را به صورتی کاملا یکپارچه[۱۹] با هستان شناسی ارائه می‌کند به صورتی که مرز بین هستان شناسی و انباره مستندات[۲۰] نامحسوس است.

زیرساخت هستان شناسی در TeLQAS بیشتر به یک وب معانی[۲۱] نزدیک است و همین امر باعث سادگی و تفسیرپذیری کم هزینه آن می‌شود. هستان شناسی TeLQAS یک گراف است که در آن مفاهیم نقش رئوس[۲۲] را بازی می‌کنند و روابط بین مفاهیم بصورت لبه‌ها ظاهر می‌شوند. همانطور که نشان خواهیم داد، این مدل می‌تواند هر گونه اطلاعاتی را که ممکن است یک هستان شناسی یا پایگاه دانش در بر داشته باشد، ذخیره کند. در شکل۲ دو گره همراه با رابطه بین آنها دیده می‌شود.



شکل ۲- نمونه‌ای از ارتباط بین دو گره؛ ساختار فوق اساس هر نوع اطلاعاتی است که در هستان شناسی TeLQAS ذخیره می‌شود

گره -یکی از دو عنصر کلیدی است که در هستان شناسی TeLQAS- خود دارای چهار نوع مختلف می‌باشد:

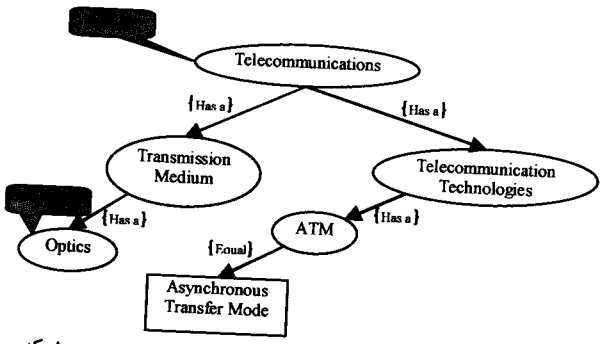
- ❖ *حوزه*[۲۳]: این گره نمایانگر یک حوزه می‌باشد. به عنوان مثال مقوله مخابرات[۲۴] یک حوزه است. یک حوزه مفهومی بسیار کلی و عام است که در بالاترین نقطه یک هستان شناسی قرار می‌گیرد. یک حوزه معمولا خود دارای چند زیر-هستان شناسی است. به همین ترتیب "هنر" نیز یک حوزه محسوب می‌شود.
- ❖ *زیرحوزه*[۲۵]: این گره مبین یک زیرهستان شناسی[۲۶] (یا دقیقتر یک زیر-حوزه) است. این گره، نقطه شروع یک زیرحوزه را در حوزه مورد نظر تعیین می‌کند. زیرحوزه‌های یک حوزه می‌توانند بطور موازی و همزمان توسط گروه‌های تخصصی مربوطه تولید شوند.
- ❖ *مفهوم*[۲۷]: نمایانگر یک مفهوم در حوزه مربوطه است. مفاهیم در مدل‌های دیگر هستان شناسی به صورت کلاس یا نمونه[۲۸] نمایش داده می‌شوند. نام مفاهیم در یک حوزه یکتا هستند. بنابراین ممکن است که یک مفهوم در چندین زیرحوزه ظاهر شود.
- ❖ *صفت خاصه*[۲۹]: این گره به صورت یک زوج 'name=value'می‌باشد. از این نوع گره برای تعیین مشخصات یک مفهوم استفاده می‌شود. به عنوان مثال اگر لازم است تا برای یک مفهوم که یک وسیله ارتباطی است میزان پهنای باند را تعیین کنیم، از گره نوع صفت خاصه استفاده می‌کنیم و آنرا به این شکل نمایش می‌دهیم: 'Bandwidth=8MHz'.. به این ترتیب این نوع گره، نقش اسلات[۳۰] و مقدار منتسب به آن را برای نمونه‌ها خواهد داشت. نکته دیگری که شایان ذکر است این است که صفتهای خاصه مانند برگهای گراف هستان شناسی هستند و فاقد هرگونه انشعاب خروجی می‌باشند و تنها یک رابطه ورودی از طرف یک مفهوم دیگر دارد. مقادیر صفات خاصه می‌توانند بصورت درونی[۳۱] (یک مقداری واقعی) و یا بصورت خارجی[۳۲] یعنی اشاره‌کننده به یک مستند خارجی (مثلا از نوع hyper text) باشند.

عنصر کلیدی بعدی رابطه[۳۳] است که ارتباط بین دو گره از هستان شناسی را مشخص می‌کند. این رابطه مبین ارتباط معنایی بین این دو گره است. در هستان شناسی TeLQAS، روابط دارای انواع مختلفی هستند. بعضی از این انواع در سایر هستان شناسی‌ها نیز وجود دارد مانند رابطه نوع (is-a) و رابطه شمولیت (has-a). بعضی دیگر از روابطی که در هستان شناسی TeLQAS وجود دارند، به جهت سهولت در روند استنتاج و آماده‌سازی پاسخ کاربران پدید آمده‌اند. از جمله روابط uses، affects، causes که به عنوان مثال برای بیان روابطی چون استفاده کردن یک مفهوم (شئ) از مفهومی دیگر، علت پدیدار شدن رخدادی یا تاثیرگذاری بر آن می‌باشند. ضمنا یک نوع رابطه تحت عنوان specification پیش‌بینی شده است

که به کمک آن بتوان خصوصیات و مشخصات مفاهیم را تعیین نمود. این نوع رابطه اخیر بیشتر با گره نوع صفت خاصه در ارتباط است.

۳-۱-۱ گراف هستان شناسی

گراف هستان شناسی در TeLQAS مجموعه‌ای از مفاهیم و روابط بین آنها می‌باشد که دانش یک حوزه را بازنمایی می‌کند. همانطور که اشاره شد، به ازای هر حوزه یک گره از نوع حوزه در هستان شناسی وجود دارد. این گره فاقد رابطه ورودی است. معمولا از این گره چند انشعاب خروجی به سایر مفاهیم و هستان شناسی‌ها وجود دارد. در شکل ۳، قسمتی از دیاگرام هستان شناسی حوزه مخابرات نمایش داده شده است. در ریشه‌ی این گراف یک گره به نام 'Telecommunications' وجود دارد که از نوع حوزه بوده و مبین حوزه مخابرات است. در این شکل، گره 'Optics' یک گره از نوع هستان شناسی است که هستان شناسی فیبر نوری را دربر می‌گیرد. سایر گره‌هایی که در داخل دایره (با بیضی) نشان داده شده‌اند، مفهوم هستند. مفهوم ATM یک معادل دارد (که در واقع شکل غیر مخفف آن است). این معادل با رابطه Equal به یک گره صفت خاصه منتهی می‌شود. در شکل مذکور صفات خاصه با مستطیل نشان داده شده‌اند.



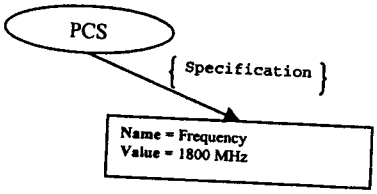
شکل۳- قسمتی از گراف هستان شناسی شامل حوزه مخابرات و مدیریت شبکه

همانطور که قبلا نیز اشاره شد، هستان شناسی TeLQAS می‌تواند علاوه بر آنچه سایر هستان شناسی‌ها ذخیره می‌کنند، هرگونه دانشی را که یک حوزه ممکن است دربر گیرد، ذخیره کند. برای تکمیل اثبات این امر، در ادامه حالاتی از ارتباطات گرافی را که نیاز به توجه خاص دارند و نیز قواعدی کلی که باید در گراف رعایت گردد را، مورد بررسی قرار می‌دهیم:

❖ **ایجاد مفاهیم معادل[۳۴] با مفهوم اصلی:** هر مفهوم هستان شناسی ممکن است با چند مفهوم دیگر هم ارز باشد. در طراحی هستان شناسی‌های غیر از TeLQAS توصیه می‌شود که به ازای مفاهیم هم‌ارز، کلاس‌ها یا نمونه‌های جداگانه ایجاد نکنیم. در اینجا نیز این موضوع معتبر است. ولی از آنجاکه در TeLQAS ما با مفاهیم سروکار داریم و نه کلاسها یا نمونه‌ها، مفاهیم معادل نیز به صورت یک مفهوم در هستان شناسی ظاهر می‌شوند اما تنها یک رابطه ورودی از نوع 'Equal' دارند. از این روابط نمی‌توان روابط یا گره‌هایی دیگر منشعب کرد. وجود این دسته از مفاهیم صرفا به دلیل آگاهی سیستم از نامهای مختلف یک مفهوم است و گرنه برای ثبت هر نوع اطلاعات دیگر باید از مفهوم اصلی که مقصد این نوع رابطه است، استفاده شود. بدیهی است اگر یک مفوم دارای چندین مفهوم معادل باشد، برای هریک از این مفاهیم معادل یک انشعاب از مفهوم اصلی با نوع رابطه 'Equal' ایجاد می‌گردد. اینکه از چندین نام برای یک مفهوم، کدامیک را به عنوان مفهوم اصلی انتخاب کنیم، اهمیت چندانی ندارد و می‌تواند به صورت قراردادی توسط خود طراح تعیین شود.

❖ **ایجاد تعاریف برای مفاهیم:** به عنوان یک پایگاه دانش که قرار است در یک سیستم پرسش و پاسخ بکار رود، لازم است تا هستان شناسی تعریفی نیز از مفاهیمی که در بر می‌گیرد، داشته باشد. رابطه Definition به همین منظور تعبیه شده است. هر چند که گره مقصد این رابطه می‌تواند یک مفهوم نیز باشد، اما معمولا مقصد این رابطه یک گره از نوع صفت خاصه است که نام آن تهی و مقدار آن رشته‌ای (یا آدرس مستندی) است که مفهوم مذکور را تشریح می‌کند.

❖ **ایجاد مشخصه برای مفاهیم هستان شناسی:** خیلی از اوقات می‌خواهیم مشخصه‌های یک مفهوم را در هستان شناسی ثبت کنیم. همانطورکه اشاره شد برای ثبت مشخصه‌های یک مفهوم لازم است تا به ازای هر مشخصه یک گره از نوع صفت مشخصه ایجاد کنیم و سپس این گره‌ها را با رابطه 'specification' به مفهوم مورد نظر متصل سازیم (یعنی مفهوم مذکور باید در مبدا رابطه قرار گیرد و گره‌های از نوع صفت مشخصه در مقصد رابطه قرار گیرند). هنگام ایجاد گره صفت مشخصه نام مشخصه را در فیلد 'name' و ارزش آن را در فیلد 'value' ذخیره می‌کنیم. به عنوان مثال اگر مفهومی با نام PCS داریم که بسامد کاری آن ۱۸۰۰ مگاهرتز است، این اطلاعات را مانند شکل۴ ثبت می‌کنیم.



شکل ۴- نحوه نمایش و ثبت خصوصیات یک مفهوم فرضی به نام PCS

۳-۱-۲ قواعد حفظ سازگاری در گراف هستان شناسی

برای آنکه ساختار مفهومی گراف هستان شناسی همواره در وضعیتی قابل قبول و منطقی باشد، قواعدی وضع شده که به اهم آنها در این قسمت اشاره خواهیم کرد:

❖ **پیوستگی و اتصال گراف هستان شناسی:** تمامی قسمتهای یک هستان شناسی باید با یکدیگر در ارتباط باشند. این امر نه تنها از نظر مفهومی و منطقی لازم نظر می‌رسد، بلکه بعضی از محاسبات و تفسیرهایی که بر روی گراف هستان شناسی صورت می‌گیرد نیز بر اساس این پیش فرض طراحی شده‌اند. بنابراین حذف یک رابطه یا افزودن بعضی از گره‌ها و روابط باعث یک عدم پیوستگی گراف شوند، باید از ادامه عملیات جلوگیری به عمل آید (این کنترل بگونه‌ای پایگاه داده هستان شناسی تعبیه شده است که بصورت خودکار انجام می‌ش البته یک استثنا در این مورد وجود دارد و آن مربوط است به گره‌های ذاتا آزاد هستان شناسی که در ادامه آنرا شرح می‌دهیم.

❖ **یکتایی نام مفاهیم در یک حوزه:** هیچگاه در یک حوزه نمی‌تواند دو مفهو یک نام وجود داشته باشد. از آنجا که معمولا حوزه‌ها تخصصی هستند (مانند - مخابرات)، همواره تعریف مشخص و متمایزی برای هر مفهوم حوزه وجود دار این ترتیب ابهامی در این خصوص که دو مفهوم با یک نام ولی با معانی مخ ممکن است وجود داشته باشد، پدید نمی‌آید. اگر در دو زیرحوزه مختلف از حوزه، به دو مفهوم با نامهای یکسان اشاره شد، این دو مفهوم در واقع یکی ه و باید در هم ادغام شوند. به این ترتیب اتصالات بین زیرحوزه‌ای پدید می‌آ کاملا معتبر و مفید هستند. بدیهی است که در حوزه‌های مختلف ممکن واژه‌هایی وجود داشته باشند که به مفاهیمی مختلف در هر حوزه اشاره باشند. مانند واژه مدولاسیون در حوزه موسیقی و همین واژه در حوزه مخابرا

❖ **صفات خاصه به عنوان گره‌های انتهایی گراف:** همانطور که قبلاً نیز به شد، صفات خاصه به عنوان گره‌های انتهایی گراف هستان شناسی ه

البته این امر به این معنی نیست که تنها صفات خاصه می‌توانند به عنوان گره‌های انتهایی گراف واقع شوند، در واقع انتهایی بودن صفات خاصه شرطی لازم است. بنابراین صفات خاصه نمی‌توانند به عنوان مبداء هیچ رابطه‌ای منظور شوند. در عین حال تنها یک رابطه ورودی به یک صفت خاصه وجود دارد. به عنوان مثال اگر دو مفهوم در هستان شناسی دارای یک مشخصه مشترک باشند نمی‌توان از هر دو مفهوم مذکور، یک رابطه specification به یک صفت خاصه وصل کرد. بلکه در این حالت باید دو گره صفت خاصه یکسان ایجاد شود و سپس هر مفهوم جداگانه به یکی از این صفات خاصه متصل گردد. هرچند این عمل ممکن در نگاه اول تنها باعث اتلاف فضا به نظر آید، اما سهولت بسیار زیادی در اجرای عملیاتی چون دوباره تعریفی[۳۵] و اصلاح مقادیر فعلی اسلات بوجود می‌آورد.

❖ **مفاهیم آزاد در هستان شناسی:** گفتیم که گراف یک هستان شناسی باید کاملا پیوسته و متصل باشد اما گاهی بعضی از مفاهیم که معمولا خیلی متداول نیز هستند، مستقیما مربوط به هیچ یک از هستان شناسی‌های حوزه نمی‌شوند (در عین حال در کل حوزه بسیار رایج هستند). شاید بتوان گفت که مفهوم 'signal' در حوزه مخابرات از این دسته مفاهیم است. برای آنکه چنین مفاهیمی را در حوزه پوشش دهیم، می‌توانیم آنها را بصورت آزاد (یعنی بدون هیچ رابطه ورودی و هیچ اتصالی با گراف هستان شناسی) ایجاد کنیم. اما مفاهیم آزاد دارای محدودیت‌هایی هستند. از جمله اینکه تنها روابطی را می‌توان از آنها منشعب کرد که به یک گره صفت خاصه منتهی شوند. واضح است اگر لازم باشد مفاهیم دیگری را در ادامه این مفهوم اضافه کنیم، باید ابتدا مفهوم اصلی را به طریقی به گراف هستان شناسی متصل سازیم.

۳-۲ طراحی و پیاده‌سازی یک معماری چند لایه جهت بهره‌گیری از هستان شناسی و انباره مستندات

دسترسی به اطلاعات و دانش ذخیره شده در یک هستان شناسی، از طریق یک رابط برنامه‌نویسی میانی به نام رابط هستان شناسی[۳۶] انجام می‌شود. به علت ساختار وب‌گونه هستان شناسی، اجرای عملیاتی نظیر جستجو، پیمایش، بروز رسانی و انجام محاسبات بسیار وقتگیر و پرهزینه است. در اینجا نشان خواهیم داد که چگونه با یک ساختار چند لایه پیشنهادی علاوه بر فائق آمدن بر مشکلات ناشی از پرهزینه بودن عملیات بر روی هستان شناسی، می‌توان ساختار و عناصر هستان شناسی را همانطور که هستند (بصورت انتزاعی و وب‌گونه) و بدور از هرگونه جزئیات فنی و برنامه‌نویسی در اختیار سایر زیرسیستم‌ها قرار داد. هدف دیگری که در این معماری چند لایه دنبال می‌کنیم، پنهان سازی جزئیات مربوط به ذخیره سازی از خدمات رابط هستان شناسی است. هرچند رابطی که در اینجا توضیح داده می‌شود، ابتدا به منظور استفاده در TeLQAS توسعه داده شده است، اما ساختار عام و جامع آن اجازه می‌دهد که آنرا در سیستم دیگری که محتاج به استفاده از هستان شناسی است بکار گرفت. تاکنون بیشتر مطالعاتی که در این زمینه انجام شده معطوف به زیرساختها و معماری ذخیره سازی بوده است و کمتر موضوع یک رابط با اهداف مذکور بررسی شده است. در واقع اغلب تلاشهایی که در بهبود یکی از نماهای منطقی و یا فیزیکی[۳۷] هستان شناسی به عمل آمده است، باعث دشوارتر شدن شرایط در نمای دیگر شده است[۱۰, ۶] از طرفی بدون استفاده از یک رابط مناسب جهت دسترسی به هستان شناسی، تمام رویه‌های[۳۸] مورد نیاز برای این امر بصورت کدنویسی در زیرسیستم‌های استفاده کننده توسعه داده خواهد شد و بنابراین نگهداری و بروزرسانی این رویه بسیار دشوار خواهد بود. اهم اهدافی که در طراحی این رابط مد نظر بوده، بصورت زیر است:

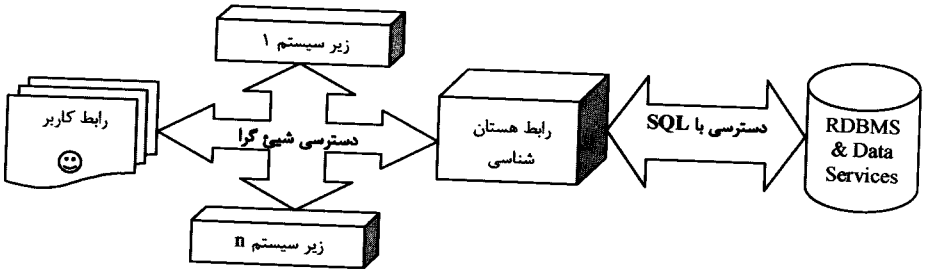
۱) لازم است بتوان در مدل ارائه شده کل دانش حوزه را بر اساس مجموعه‌ای از گره‌ها و روابط بیان کرد بگونه‌ای که تمامی عملیات متداول در پروسه‌های بازیابی اطلاعات بر روی مدل جدید نیز قابل اجرا باشد.
۲) تمام موجودیتهایی که این رابط ارائه می‌دهد باید بصورت شئ‌گرا باشد.
۳) خدمات رابط باید مستقل از روش ذخیره‌سازی و پردازش داخلی ارائه شوند.
۴) سرعت در اجرای عملیات، حفاظت، چند کاربری و پشتیبانی از تراکنش داده‌ها از ویژگیهای مورد انتظار در این رابط هستند.

برای برآورده سازی این انتظارات، یک معماری چند لایه طراحی و توسعه داده شده است که شمای کلی آن در شکل ۵ نشان داده شده است. در شکل مذکور هستان شناسی و رابط آن در لایه خدمات داده‌ای[۳۹] و لایه میانی نشان داده شده‌اند. جلوتر توضیح خواهیم داد که چگونه این رابط با استفاده از یک نگاشت مناسب از لایه خدمات داده‌ای برای ذخیره‌سازی و استخراج اطلاعات استفاده می‌کند.

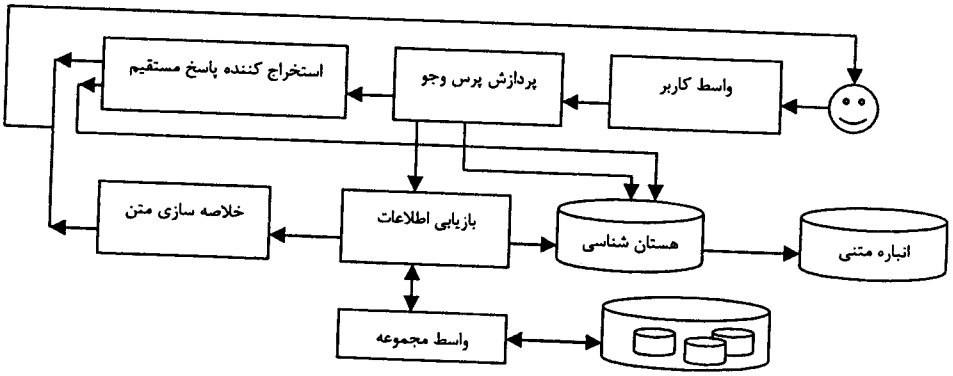
از آنجاکه ساختار هستان شناسی‌ها عموما سلسله مراتبی می‌باشند، بسیاری از استانداردها و نرم‌افزارهایی که در این رابطه تهیه شده‌اند، از یک فرم سلسله‌مراتبی برای ذخیره‌سازی داده استفاده می‌کنند که در این میان می‌توان از فرمتهای RDF و XML نام برد[۱۱]. هرچند این نوع ساختمان داده بسیاری از مسائل طراحی و پیاده‌سازی هستان شناسی را سهولت می‌بخشد، اما هیچیک از اهداف مورد نظر ما (با آنچه در بندهای ۳ و ۴ در بالا اشاره شد) را تحقق نمی‌بخشند. بنابراین ما در لایه خدمات داده‌ای از مدل رابطه‌ای و یک RDBMS استفاده کرده‌ایم. تمام اطلاعات هستان شناسی در یک پایگاه داده بصورت رابطه‌ای ذخیره می‌شوند. به این ترتیب برای عملیاتی نظیر جستجو و کار با داده‌ها از امکانات این مدل نظیر شاخص‌بندی، مدیریت تراکنش‌ها، حفاظت و امثال آن استفاده می‌شود. به این ترتیب بستر مناسبی نیز جهت استفاده همزمان چند زیرسیستم از هستان شناسی پدید می‌آید.

در طراحی لایه میانی سعی بر آن بوده است که تمامی عناصر هستان شناسی و خدمات مرتبط با آنها در قالب اشیاء متناظر با آنها ارائه گردد. این لایه ماهیت رابطه‌ای اطلاعات ذخیره شده در بخش خدمات داده‌ای هستان شناسی را پنهان می‌کند. کلیه خصوصیات[۴۰] و رویه‌های مرتبط با موجودیت‌های هستان شناسی به عنوان اعضای اشیاء مربوطه آمده‌اند. این اعضاء با اعمال یک نگاشت مناسب وظیفه خود را به فرم رابطه‌ای انجام داده و نتایج را در اختیار زیر سیستم استفاده کننده قرار می‌دهند. همچنین در این لایه کلاسها و خصوصیاتی جهت عملیات بازیابی اطلاعات تعبیه شده است. کلاسهایی نظیر کلاس مستند[۴۱] و خصوصیاتی نظیر نمونه‌های مرتبط[۴۲] (با یک مفهوم) از این نوع هستند. علاوه بر این توضیحات، ویژگیهای دیگری نیز در طراحی کلاسها در نظر گرفته شده که مهمترین آنها به شرح زیر است:

۱) طراحی کلاسهای رابط هستان شناسی با استفاده از تکنیکهای چندریختی[۴۳] به نحوی انجام شده است که می‌توان تمام عناصر هستان شناسی را بصورت انتزاعی، اشیائی از نوع گره[۴۴] و رابطه دید. که البته در عین حال می‌توان در صورت لزوم به صورت کاملا اختصاصی با این اشیاء برخورد کرد (مانند اشیاء از نوع صفت خاصه و زیردامنه که مشتقاتی از کلاس گره هستند).
۲) استفاده از کلاسهای مجموعه‌ای[۴۵] که تمام عملیات مربوط به مجموعه‌ها نظیر درج، حذف، مرتب‌سازی و از این قبیل را پشتیبانی می‌کنند.
۳) مصرف بهینه حافظه با استفاده از مکانیسم دو مرحله‌ای اسکان دادن اطلاعات[۴۶] به این معنی که در مرحله اول ایجاد، اشیاء تنها یک مشخصه که منحصرا معرف آنها در پایگاه داده هستان شناسی می‌باشد، با خود حمل می‌کنند؛ سپس هرگاه به یکی از خصوصیات شئ مورد نظر دسترسی پیدا شد، اطلاعات شی مزبور از پایگاه داده بازیابی می‌شوند.



شکل ۵- معماری سیستم TeLQAS از نظر دسترسی به هستان شناسی



شکل۶- نحوه تعامل زیرسیستمهای موجود در بخش برخط سیستم

۴) کاهش ترافیک شبکه با استفاده از ارتباطات ماندگار و ذخیره سازی اطلاعات اشیاء در حافظه (مرحله دوم فرایند بالا) جهت عدم رجوع دوباره به پایگاه هستان شناسی.

۵) مدیریت هوشمند خطاهای زمان اجرا به نحوی که خطا در ارتباط، نوع درخواست، شرایط کنونی هستان شناسی و امثال آن باعث شکسته شدن روند سیستم نشده و منابع تحت اختیار رابط هستان شناسی مانند شکل عادی به سیستم باز خواهند گشت.

۴- فرآیندهای بر خط: پردازش پرس و جو، خلاصه‌سازی متن و بازیابی پاسخ

نمای کلی از بخش بر خط TeLQAS در شکل ۶ آمده است. کار سیستم در این بخش، با ارسال پرسش کاربر به سیستم آغاز می گردد. پس از بررسی صحت املایی کلمات پرسش کاربر، عملیات پردازش پرس و جو آغاز می شود تا از پرسش کاربر، اطلاعات مورد نیاز برای جستجو استخراج شده و بخشهایی که استفاده خاصی در جستجو ندارند، از مجموعه اطلاعات مفید حذف شوند. که این عملیات همگی در زیرسیستم پردازش پرس و جو صورت می گیرد. از عمده ترین وظایف بخش پردازش پرس و جو، استخراج کلمات و عبارات کلیدی پرسش کاربر، تعیین هدف سوال وی و در نتیجه تعیین نوع ارتباطی مورد نظر و همچنین یافتن مفاهیم مرتبط به سوال وی از لحاظ همسایگی‌های موجود در هستان شناسی است تا با اضافه نمودن این کلمات به مجموعه مفاهیمی که برای بازیابی استفاده میشود، شانس پیدا کردن مستندات مرتبط با سوال کاربر افزایش یابد.

همانطور که اشاره شد، زیرسیستم پردازش پرس و جو، موظف است که هدف سوال کاربر را تعیین نماید که علت آن در مرحله اول، نگاشت سوال کاربر به شاخه (هایی) از هستان شناسی است که به هـدف پرسش وی ارتبـاط بیشتـری دارند و

درگام بعدی، کمک به زیرسیستم خلاصه ساز متن جهت استخراج خلاصه های مناسبتر است.

در صورتی که، حالت اول انجام پذیرد (یعنی سوال کاربر را بتوان به شاخه (های) خاصی از هستان شناسی مرتبط کرد)، می توان پاسخ سوال را بطور دقیق یا با تقریب خوبی بصورت مستقیم از هستان شناسی استخراج نمود. به بیان دیگر، دانستن منظور کاربر از پرسش مطرح شده به دقیقتر نمودن محدوده جستجوکمک کرده منجر به یافتن جوابی می گردد که مد نظر اوست. اما در صورتی که استخراج جوابهای مناسب و کافی مستقیما و به تنهایی از هستان شناسی ممکن نباشد، پس از استخراج متنهای مرتبط در زیرسیستم بازیابی اطلاعات، این مستندات در زیرسیستم خلاصه ساز متن مورد کاوش قرار می گیرد که در این حالت نیز، دانستن هدف پرسش کاربرمی تواند سیستم را در یافتن جوابهای بهتر یاری نماید. برای مثال فرض کنید که کاربر سوال " What is Transmission Loss?" را مطرح نموده است. با توجه به ساختار زبان طبیعی[۴۷] که این سوال بر اساس آن مطرح شده است، کلاسه کننده موجود در بخش پردازش پرس و جو می تواند تعیین نماید که هدف این سوال یافتن تعریفی برای مفهوم Transmission Loss است. چنانچه هستان شناسی ما دقیقا شامل مفهوم Transmission Loss باشد و آن گره هم خود دارای نوع ارتباط Definition باشد، مرتبطترین پاسخ، اصل یافتن گره‌ای است که با نوع ارتباط Definition از گره Transmission Loss منشعب می‌گردد. البته در صورتی که برای این مفهوم Transmission (Loss هیچ ارتباطی با نوع Definition نداشته باشیم، طبیعی است که جستجو در داخل متون مرتبط و با توجه به نوع سوال که Definition تشخیص داده شده است، سعی می شود بهترین جملاتی از متن که شرایط زبانی بیان تعریف را دارند و با مفهوم Transmission Loss هم مرتبط هستند را استخ نموده به کاربر نمایش بدهیم.

بر اساس آنچه که در بالا ذکر شد، مولفه بازیابی اطلاعات، موظف است که در ابـ

مستندات مرتبط با مفاهیمی راکه بخش پردازش پرس و جو آنها را تعیین نموده است، از انبار متون سیستم استخراج نموده و آنها را بر اساس عوامل مختلفی مانند میزان ارتباط به مفهوم (که یا توسط کارشناس حوزه هستان شناسی و یا زیر سیستم رسته ساز متن تعیین می‌شود)، میزان دوری یا نزدیکی به مفهومی (مفاهیمی) که کاربر در ارتباط با آن (آنها) سوال کرده است و همچنین نوع ارتباط مفهوم را با مفاهیم مدنظر کاربر، امتیازدهی نماید و سپس مستنداتی را که امتیازشان ازآستانه امتیاز تعیین شده در این مرحله بیشتر باشد، برای ادامه عملیات مشخص سازد.

در صورتی‌که، در این مرحله، مستندی با امتیاز کافی در انبار متون سیستم وجود نداشته باشد ویا تعداد مستندات کافی نباشد، نیاز به استخراج مستندات متنی از وب و از طریق واسط مجموعه می‌باشد. در این حالت یک پرس و جوی توسعه یافته که تلفیق مناسبی از مفاهیم کلیدی و واژه های معادل آنها است، به واسط مجموعه ارسال می گردد تا بازیابی از مجموعه ها (وب، کتابخانه‌های دیجیتال و ...) صورت پذیرد. این مستندات بازیابی شده، پس از خلاصه سازی در صورتی که پاسخهای مناسبی را شامل شوند، به مجموعه مستندات سیستم اضافه می‌شوند تا در بازیابی‌های بعدی مورد استفاده قرار گیرد.

پس از اینکه عمل بازیابی مستندات متنی از انباره مستندات سیستم یا از طریق واسط مجموعه انجام شد، نوبت به مولفه خلاصه ساز متن می‌رسد که با توجه به نیاز اطلاعاتی کاربر، خلاصه مطلوبی برای او تهیه کند و به بیان دیگر، پاسخ دقیقی را به سؤال وی (به شکل چند جمله مرتبط) از متون بازیابی شده استخراج نموده، به وی ارائه دهد.

برای روشنتر شدن مطلب، همان پرسش نمونه قبلی را در نظر بگیرید. سیستم در مؤلفه پردازش پرس و جو و توسعه آن سعی می‌کند که تعدادی از مفاهیم مرتبط (از لحاظ مترادفها و همسایه های نزدیک در هستان شناسی) را بیابد. مثلا در این ارتباط، مفاهیم Mechanism of Loss، Optical Loss، و Bending Loss به همراه اولویت نسبی که در حوزه مخابرات با کلمات موجود در سوال کاربر دارند، پیدا می‌شوند. اکنون مولفه بازیابی با در اختیار داشتن مفاهیم بالا، سعی در یافتن مستندات مرتبط می‌کند و تعدادی از مستنداتی را که با مفاهیم مرتبط با Transmission Loss ارتباط بیشتری دارند. این ارتباط در قالب عدد Relevance که توسط استخراج شده توسط زیرسیستم رسته‌ساز متن) تعیین شده است، استخراج نموده و پس از انجام امتیازدهی برای خلاصه سازی در اختیار خلاصه ساز متون می‌گذارد تا در نهایت با ارائه بهترین پاسخهای متنی به کاربر، عملیات جستجو خاتمه یابد.

۴-۱ زیرسیستم پردازش پرس وجو

زیرسیستم پردازش پرس و جو، همانطور که در شرح کلی بخش برخط ارائه شد، وظیفه استخراج اطلاعات مفید در رابطه با پرسش کاربر را به عهده دارد تا با یافتن مفاهیم مرتبط به آنها از داخل هستان شناسی، یک پرس و جوی توسعه یافته حاصل شود. بطور کلی اجزاء اصلی این سیستم شامل موارد زیر می‌باشد؛

- **مؤلفه بررسی صحت املائی کلمات:** به منظور حصول اطمینان از صحت املایی کلمات پرسش کاربر، پیش از اینکه عملیات جستجو با پرسش او آغاز گردد، از این مولفه استفاده می‌شود.
- **مؤلفه تعیین نقش کلمات**[۴۸]: این مولفه که یکی از رایجترین عناصرکاربردهای حوزه پردازش زبان طبیعی است، برای تعیین نقش کلمات پرسش کاربر و به منظور جداسازی اجزاء مهم و کلیدی از اجزائی که بار اطلاعاتی چندانی برای دخالت در فرآیند بازیابی ندارند، در سیستم قرار گرفته است. به کمک خروجی این مؤلفه، می توان اسامی، افعال اصلی و صفات را از افعال کمکی، حروف اضافه و ربط و دیگر اجزاء ساختاری جمله تفکیک نمود.
- **مؤلفه تعیین کلمات کلیدی:** با استفاده از خروجی مولفه تعیین نقش کلمات، اسمها و افعال اصلی که کلمات کلیدی پرسش کاربر به حساب می آیند، قابل تعیین هستند.
- **مؤلفه تعیین عبارات کلیدی:** باز هم به کمک مولفه تعیین نقش کلمات، می‌توان عباراتی مانند صفت و موصوف و نیز مضاف و مضاف الیه را تشخیص داد تا به کمک عبارات (نه فقط تک واژه ها) بر دقت جستجو افزود.
- **مؤلفه کلاسه کننده برای تعیین هدف سوال:** از آنجا که که هیچ یک از مولفه های دیگر موجود در سیستم قادر نخواهند بود که از پرس وجوی کاربر، اضافه بر استخراج لغات کلیدی، اطلاعات دیگری از قبیل نوع و هدف را استخراج نمایند، و با توجه به اینکه ساختار پرس وجوی کاربر می تواند حاوی اطلاعاتی باشد که نادیده گرفتن آنها از دقت پاسخ سیستم بکاهد، این مؤلفه در سیستم تعبیه شده است. هدف دیگر از وجود این مولفه را می توان در بهبود پرسش هایی دانست که از ساختار جمله مناسبی برخوردار نیستند و در اصطلاح جملات نویزی نامیده می شوند. نکته مهمی که در نحوه طراحی این کلاسه کننده مدنظر بوده این است که صرفا ساختار زبانشناسانه جملات انگلیسی مدنظر قرار گرفته و از استفاده از لغات خاص حوزه اجتناب شده است. همچنین در طراحی این کلاسه کننده قانون-پایه[۴۹] به گونه ای عمل شده است که حتی الامکان و در درجه اول، انواع پرسشهایی که در تناظر مستقیم با انواع ارتباطات مطرح در هستان شناسی هستند شناسایی شوند (در[۱۲] یک سیستم پرسش و پاسخ مشابه قانون پایه معرفی شده است که برای درک مطالب در آزمونهای زبان انگلیسی طراحی شده‌است). سپس انواعی از سوالات رایج این حوزه که عیناً متناظر با نوع ارتباط خاصی نیستند مورد توجه قرارمی‌گیرند. انواع ارتباطات مطرح در هستان شناسی به همراه نوع پرسش مرتبط با آنها در جدول ۱ آورده شده است.

جدول ۱- چند نمونه از سؤالاتی که نوع ارتباطی متناظر با آنها عیناً در هستان شناسی وجود دارد

Is-A	What kind / type
Has-A	Which Part
Measured-with	Measurement Unit
Made of / Is a Part of	Which components
Causes / Caused by Affects / Affected By	Cause/Effect
Definition	Definition
Used in / Used by	Usage
Related to	-
Synonym	Synonym, Abbreviation, Acronym
Specified-by	Specification / Characterization

به جز موارد بالا، کلاسه کننده انواع دیگری از سوالات را که دقیقاً مرتبط با نوع خاصی از ارتباط در هستان شناسی نیستند نیز مشخص می کند که از آن جمله می توان به Similar / Different و Advantage / Disadvantage اشاره کرد.

۴-۲ زیرسیستم استخراج مستقیم پاسخ به کمک هستان شناسی

همانطور که پیشتر هم ذکر شد، اولین مرحله در پاسخگویی به سؤال کاربر،

استفاده مستقیم از هستان شناسی برای یافتن پاسخ مستقیم به سؤال کاربر است، بدون اینکه به متن مراجعه ای انجام گیرد. در این مؤلفه به کمک نتایج حاصل از پردازش پرس و جو، ابتدا یک گره (یا در برخی حالات خاص با توجه به نوع سوال کاربر، دو یا چند گره) به عنوان گره مورد تأکید[۵۰] کاربر برای جستجو تعیین می‌گردد. سپس بر اساس نوع سؤال کاربر یک نوع ارتباط مطلوب[۵۱] که از گره موردنظر خارج می‌گردد را مدنظر قرار می‌دهیم. بهترین حالتی که می توان متصور شد آن است که این گره مورد تأکید عیناً دارای آن نوع ارتباط مطلوب باشد که در این صورت پاسخ همان گره ای است که در سوی دیگر این ارتباط واقع است. این مسأله با ارائه چند نمونه پرسش روشن می‌گردد. در نخستین قدم پرسش زیر در نظر گرفته می‌شود؛

Question 1: What is the definition of absorption loss?
Focused Concept: Absorption loss
Type of Question: Definition

در این حالت، مؤلفه جستجوی مستقیم پاسخ به کمک هستان شناسی، محتوای گره ای که با نوع ارتباطی Definition از Absorption Loss منشعب می شود، را به عنوان پاسخ بر می‌گرداند. در درصورتی که پرسشی به شکل زیر مطرح گردد،

Question 2: What is the Effect of absorption loss?
Focused Concept: Absorption loss
Type of Question: Cause

سیستم با توجه به وجود گره Absorption loss و نیز نوع ارتباطی Causes، Caused-by درگراف هستان شناسی، پاسخ را به صورت مستقیم می یابد. با پیچیده تر شدن پرسش به فرم زیر،

Question 3: What are the similarities of Cable X and Cable Y?
Focused Concepts: Cable X and Cable Y
Type of Question: Similarity and Differences X & Y

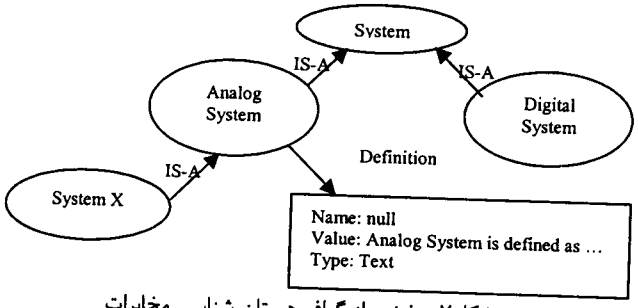
باز هم گره مورد تأکید به سادگی و مانند قبل تعیین می‌گردد؛ Cable X و Cable Y. اما از آنجا که هستان شناسی دارای نوع ارتباطی Similarity و Difference نیست، باید به نحوی از مرتبط ترین نوع ارتباطی موجود استفاده شود. بنابراین نیاز به پردازشی بیشتر برای بدست آوردن شباهت و تفاوت این دو گره وجود دارد. با مراجعه به Specification این دو مفهوم در گراف هستان شناسی، مشخص میگردد که شباهت این دو گره در میزان SP1 بوده و تفاوتشان در هر یک از دو مورد SP2 وSP3 می‌باشد. بنابراین، با ارائه Specification این دو گره به کاربر نیاز اطلاعاتی وی برآورده خواهد شد. از سوی دیگر، با مطرح شدن پرسشی بصورت زیر،

Question 4: What is System X?
Focused Concept: System X
Type of Question: Definition

مشاهده می‌شود که تعریفی عیناً برای System X در گراف هستان شناسی شکل ۷ وجود ندارد. اما در صورتی که استفاده از نوع ارتباطی IS-A مجاز باشد و تعریفی که برای مفهوم پدر (در اینجا Analog System) وجود دارد با کمی کاهش امتیاز، قابل در نظر گرفتن برای مفهـــوم فرزند تلقی گـــردد، می توان پاسخی مستقیم به کمک هستان شناسی به کاربر ارائه نمود.

در ادامه مثالهایی که ذکر شد، لازم به توضیح است که این زیر سیستم، پاسخهای استخراج شده را امتیازدهی نیز می‌کند. این امتیازدهی که در[۱۳] به طور کامل تشریح شده است، به طور خلاصه به این ترتیب است که میزان شباهت میان نام گره مورد تاکید با کلمات کلیدی پرسیده شده توسط کاربر، به عنوان پارامتر اول تصمیم گیری و نیز میزان شباهت میان نوع پرسش تعیین شده برای سوال کاربر و نوع ارتباطی متناظری که برای گره مورد تاکید وجود دارد، به عنوان پارامتر دوم در نظر گرفته می شود.

سپس حاصلضرب این دو پارامتر به عنوان امتیاز نهایی برای پاسخ استخراج شده لحاظ می شود. چنانچه این امتیاز از آستانه امتیاز پاسخهای مستقیم (که به طور تطبیقی و بر اساس میانگینی از پارامترهایی که ذکر شد، محاسبه می گردد) بالاتر باشد، یک پاسخ مستقیم را پاسخ قابل قبول تلقی خواهیم نمود و در غیر اینصورت آن را از مجموعه پاسخهایی که به کاربر نمایش داده خواهند شد، حذف می‌کنیم.

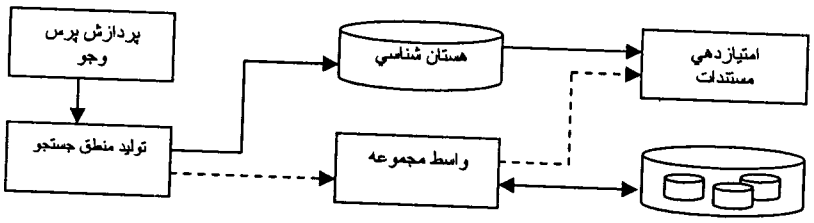


شکل۷- بخشی از گراف هستان شناسی مخابرات

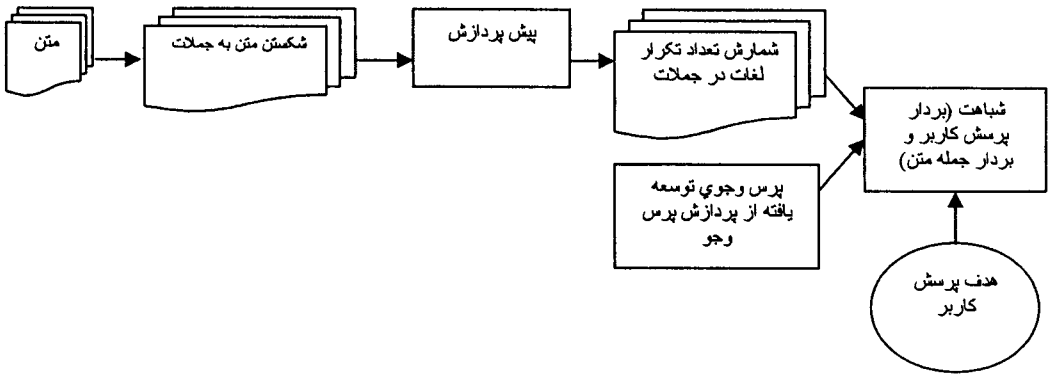
۴-۳ مؤلفه بازیابی و امتیازدهی به مستندات

پس از اینکه کار جستجوی مستقیم در هستان شناسی توسط مولفه پاسخ دهی مستقیم تکمیل شد و امتیاز پاسخهای به دست آمده به آستانه قابل قبولی که در بالا تشریح شد، نرسید یا به هر دلیلی کاربر درخواست نمود که پاسخهای بیشتری دریافت نماید، مولفه بازیابی و امتیازدهی مستندات وظیفه دارد که تعدادی از مرتبط ترین مستندات به مفاهیم موجود در پرسش کاربر را بازیابی نموده و پس از یک امتیازدهی اولیه برای تولید خلاصه به بخش خلاصه سازی ارجاع دهد تا بهترین جملات از داخل متن استخراج گردد.

در زیرسیستم بازیابی اطلاعات که اجزاء آن درشکل ۸ نشان داده شده است مستندات مرتبط به هر یک از مفاهیمی که به عنوان مفاهیم مرتبط از ماحصل بخش پردازش پرس و جو به دست آمده اند، بازیابی می‌شوند و پس از امتیازدهی به زیرسیستم خلاصه ساز متن ارائه می‌شوند تا خلاصه مرتبطی از آنها تهیه گردد در صورتی که هستان شناسی شامل متون مرتبط و مناسب نباشد (از لحاظ امتیا یا تعداد)، استخراج متون ازطریق واسط مجموعه صورت می‌گیرد که به عنوا زیرسیستم دیگری در بخش بعدی شرح داده شده است. درهرصورت، خروجی ای زیرسیستم تعدادی مستند امتیازدهی شده است که برای خلاصه سازی در اختی زیرسیستم خلاصه ساز متن قرار می‌گیرد.



شکل۸- اجزاء زیرسیستم بازیابی اطلاعات



شکل۹- بلوک دیاگرام زیرسیستم خلاصه ساز متن

۴-۴ زیرسیستم خلاصه ساز متن

پس از دریافت متون بازیابی شده و پرس و جوی پردازش شده به ترتیب از زیر سیستمهای بازیابی اطلاعات و پردازش پرس و جو، عمل خلاصه سازی متون آغاز می شود. برای خلاصه سازی متون از روش امتیازدهی به جملات استفاده شده است که در این روش با استفاده از لغات پرس و جو که شامل نوع و هدف پرس و جوی کاربر نیز می شود، ویژگیهای آماری برای هر جمله مشخص می شود. بدیهی است با استفاده از نوع پرسش که در زیرسیستم پردازش پرس و جو تعیین گردید، چنانچه پرس و جوی کاربر با توجه به ساختار زبانی جملات، به رسته پرسشی در مرحله بعد، با دریافت پرس و جوی توسعه یافته از مولد پرس و جو، فرکانس هر لغت از پرس و جوی توسعه داده شده، در هر جمله متن و همچنین در پرس وجوی کاربر محاسبه می‌شود. سپس، بر اساس فرمول امتیازدهی (۱) به جملات، شباهت بین پرس و جو و جملات محاسبه می‌شود. در ضمن، اگر پرس و جوی کاربر رسته پرسشی خاصی را مشخص کند (دارای نوع سوال معینی باشد)، این ویژگیهای زبانی نیز وزندهی می‌شوند. این زیرسیستم دارای ماجولهای استخراج ویژگی و استخراج نتایج است که در ماجول اول، پارامتر TF[۵۲] برای تمام کلمات موجود محاسبه می‌گردد. این پارامترکه تعداد تکرار کلمه را در جمله یا پرس و جو مشخص می‌کند، در بخش محاسبه امتیاز از اهمیت بالایی برخوردار است. در ماجول دوم، با توجه به رابطه زیر، امتیاز نهایی برای جملات استخراج شده، تعیین می‌شود و در نهایت این نتایج به ترتیب امتیازات، مرتب شده و جهت نمایش به کاربر در اختیار واسط کاربر قرار می‌گیرد؛

$$Score(s_i) = \lambda \sum_{s \in S} w_s * .Sim(q, S_i) + (1-\lambda) * \sum_{l \in L} w_l * (L_l . S_i) \quad (۱)$$

در فرمول (۱)، S مجموعه ویژگیهای آماری، L مجموعه ویژگیهای زبانی، q پرس وجوی کاربر و w وزن ویژگیها را مشخص می‌کند و *Sim* بصورت زیر محاسبه می‌شود:

$$Sim(q, S_K) = \sum_{w \in S_k, q} tf(wi, q).tf(w_i, s_k)\left(1 - \frac{\log(df(w_i)+1)}{\log(n+1)}\right)^2 \quad (۲)$$

۵- فرآیند برون خط: غنی سازی هستان شناسی

استخراج پاسخهای مناسب در واکنش به پرسشهای زبان طبیعی از میان توده وسیعی از مستنداتی که بصورت برخط قابل دسترسی هستند، از طریق شناسایی جوابهای مورد انتظار در درون این متنها امکان پذیر می‌گردد. این موضوع در بسیاری از سیستمهای پرسش و پاسخ با بهره گیری از امکانات خلاصه سازی متن به اجرا در می‌آید، که در بخشهای قبلی به آن اشاره شده است. به این ترتیب مشخصی متعلق باشد، در نظر گرفتن آن، نقش مهمی در تهیه خلاصه مفیدتر و مرتبط تر، ایفا می‌کند. در واقع، با معلوم بودن رسته پرسش و سمانتیک جوابی که متناظر با این رسته است، این زیر سیستم قادر خواهد بود که امتیازدهی به جملات مستندات بازیابی شده را به نحو کاملتری انجام دهد. درشکل۹، بلوک دیاگرام کلی این زیرسیستم مشخص شده است. همانطوری که در شکل نیز دیده می‌شود، ابتدا یک سری پیش پردازش برروی تک تک جملات متون بازیابی شده انجام می‌گیرد که شامل تعیین جملات متن، حذف حروف اضافه و افعال کمکی و ضمائر و تشخیص ریشه هر کلمه است.

مسئله دقت پاسخها در کنار زمان سپری شده برای حصول این نتایج، از اهمیت خاصی برخوردار می‌باشد، که بطور مستقیم کارایی کلی سیستم پاسخگویی را تحت تأثیر قرار می‌دهد. در سیستم TeLQAS یک بخش برون خط طراحی و پیاده سازی شده است، که انگیزه و هدف اصلی آن به مسئله مذکور در ارتباط با کارایی این سیستم بازمی‌گردد. این بخش با استفاده از مکانیسمی که در زیرسیستم رسته ساز متن[۵۳] اجرا می‌گردد، اقدام به رسته بندی مستندات متنی بازیابی شده توسط واسط مجموعه از منابع اطلاعاتی متفاوت موجود در وب می‌نماید. این رسته بندی براساس مفاهیم موجود در هستان شناسی سیستم صورت می‌گیرد، که مرتبط با حوزه مفاهیم شناسایی شده در دامنه فناوری مخابرات ساخته شده است[۱].

۵-۱ جایگاه و وظایف زیر سیستم رسته‌ساز متن

زیرسیستم رسته‌ساز متن، زیرسیستم‌اصلی بخش برون‌خط سیستم TeLQAS است[۱۴]. مهمترین وظیفه این زیرسیستم، طبقه‌بندی مستنداتی است که از طریق واسط مجموعه[۵۴] در اختیار آن قرار می‌گیرد[۱۵]. این مستندات که ممکن است از منابع مختلفی گردآوری/تشکیل شده باشند، با هدف ارتباط‌شان با حوزه مخابرات و هستان شناسی‌های تعریف شده در آن استخراج می‌شوند. زیرسیستم رسته‌ساز متن پس از دریافت این مستندات در خصوص ارتباطشان با حوزه مخابرات (هستان شناسی سیستم) تصمیم‌گیری می‌کند و بررسی می‌کند که مستند با کدامیک از مفاهیم هستان شناسی سیستم مرتبط دارد[۱۶].

برای روشن‌تر شدن جایگاه زیرسیستم رسته‌ساز و وظایفی که به عهده دارد، می‌توان به شکل۱۰ توجه کرد. در یک پروسه برون‌خط، زیرسیستم رسته‌ساز متن، مجموعه‌ای از مستندات را از بخش واسط مجموعه دریافت می‌کند. این مجموعه مستندات در یک صف پردازشی قرار می‌گیرند. پس از رسیدگی به هر درخواست، اعم از اینکه مستند واکشی شده برای حوزه و هستان شناسی در دست بررسی، مناسب بوده است یا خیر، نتیجه کار به واسط مجموعه اطلاع داده می‌شود. بر اساس همین اطلاعات بازخوردی[۵۵]، واسط مجموعه می‌تواند در خصوص میزان ارتباط منابع دراختیار (مانند سایت‌های وب) با مفاهیم هستان شناسی قضاوت

کند. اگر زیرسیستم رسته‌ساز تشخیص دهد که مستندی با یک یا چند مفهوم از هستان شناسی ارتباط دارد، این ارتباط و درجه آن را (که عددی بین صفر تا یک است) در هستان شناسی ذخیره می‌کند. علاوه بر خدمات اخیر، همانطور که در شکل ۱۰ نیز نشان داده شده است، این زیرسیستم می‌تواند پاسخگوی درخواستهای فوری و بلادرنگ بخش برخط سیستم TeLQAS نیز باشد. از دیگر تعاملات زیرسیستم رسته‌ساز متن با سایر اجزای سیستم که در شکل ۱۰ نشان داده شده است، می‌توان از ارتباطات این زیر سیستم با انباره مستندات، جهت دریافت مستندات آموزشی و استفاده از بعضی خدمات تکمیلی هستان شناسی، جهت ثبت ساختار بردار ویژگی و مراکز ثقل مفاهیم، نام برد.

۵-۲ معماری زیرسیستم رسته‌ساز متن

بنابر آنچه در خصوص وظایف زیرسیستم رسته‌ساز متن گفته شد، می‌توان عنوان کرد که عمده فعالیت این زیرسیستم حول محور کلاسه‌بندی مستندات[۵۶] بر اساس کلاسهایی از پیش تعیین شده که همان مفاهیم هستان شناسی هستند، می‌باشد. بنابراین یکی از اجزای اصلی این زیرسیستم یک کلاسه کننده است. معمولا مبحث کلاسه‌بندی، شرح موضوعات دیگری را نیز می‌طلبد، که از جمله می‌توان به مساله تشکیل بردار ویژگی[۵۷]، انتخاب ویژگی[۵۸] و همچنین، آموزش کلاسه کننده اشاره نمود.

از آنجا که در بیشتر روش‌های کلاسه‌بندی موجود (که البته همه‌منظوره هستند)، ورودی سیستم یک نقطه در فرافضایی n بعدی[۵۹] است، لازم است تا هر موجودیتی که قرار است کلاسه‌بندی شود، بصورت یک n تایی مرتب یا به تعبیری، یک بردار n عنصری بیان شود. در بیشتر متدولوژی‌هایی که به منظور تشکیل بردار متناظر با یک مستند بکار می‌روند، پیش‌فرضی برای قالب مستند یا اطلاعاتی تکمیلی همراه مستند قائل نیستند و در نتیجه مستند را به صورت یک فایل متنی ساده می‌بینند. در این روشها، ملاک تشکیل بردار ویژگی، بسامد واژه‌ها یا عبارات داخل مستند است[۱۷]. اما، محاسبه بسامد هر واژه یا عبارتی در متن، باعث ایجاد فضاهایی با بیشتر از صد هزار بعد می‌شود که انجام امور کلاسه‌بندی را از نظر زمان پردازش و حافظه مورد نیاز، بسیار پرهزینه و گاهی غیر ممکن می‌سازد. این امر، ما را ملزم می‌سازد تا از وجود یک مولفه انتخاب ویژگی، جهت کاهش ابعاد فضای ورودی مساله، بهره گیریم. لازم به تذکر است که فرایند انتخاب ویژگی قبل از شروع کار کلاسه‌کننده انجام می‌شود و نتیجه این فرایندکه لیستی از واژه‌ها و عبارات است، در تشکیل بردار ویژگی متناظر با یک مستند بکار می‌رود.

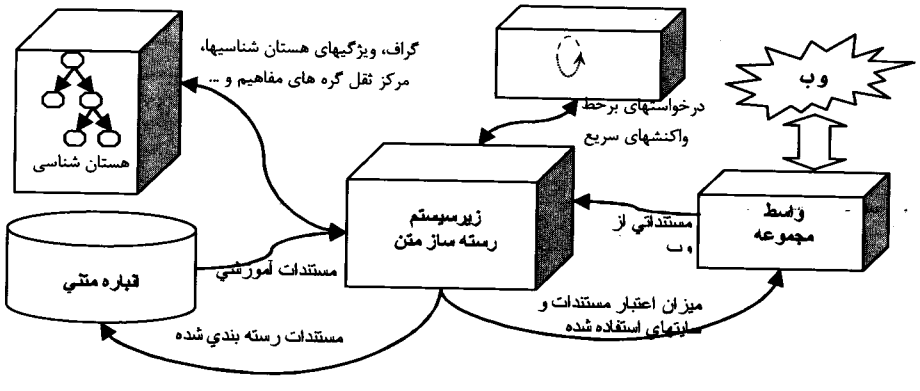
با توجـــه به آزمایشـاتی که یانگ[۶۰] و پدرسـون[۶۱] در سال ۱۹۹۷ جهت بـــررسی و مقایسه انواع روشهای انتخاب ویژگی انجام دادند، روش بهره اطلاعاتی[۶۲]، بالاتر دقت و کارایی را در میان سایر روشهای انتخاب ویژگی داراست[۱۸].

در روش بهره اطلاعاتی، امتیاز یک واژه یا عبارت در قبال رسته‌هایی که در سیـ تعریف شده است و مستندات آموزشی مربوط به هریک از آنها، با فرمول محاسبه می‌شود[۱۹] (پایین صفحه) که در این رابطه c نشان دهنده طبقه کلاس مستند است. m تعداد کل کلاس های تعریف شده برای سیستم را نـ می دهد و t نمایانگر یک کلمه یا عبارت است که معرف ویژگی می باشد. با تو به این معیار، برای هر ویژگی (واژه یا عبارت) یک امتیاز بدست می آید. در نها N ویژگی‌ای که بالاترین امتیاز را دارند، انتخاب می کنیم.

هر چند که فرایند انتخاب ویژگی به منظور کاهش ابعاد مساله و در نتیجه سـ کردن کار کلاسه کننده بکار می‌رود، اما، با وجود مجموعه‌ای بالغ بر صدها واژه در هر زبان، اجرای الگوریتم و محاسبات مربوط به انتخاب ویژگی پردازشی زمان‌بر و پر هزینه (از نظر نیاز به منابع سیستم) خواهد بود. در هرز بسیاری از واژه ها بسیار عمومی هستند و می توانند در هر متنی بکار روند. بد است که چنین واژه‌هایی هیچگاه توسط الگوریتم های انتخاب ویژگی، انتـ نخواهند شد. بنابراین، با وارد کردن چنین اطلاعاتی در پردازش انتخاب ویژ تنها زمان پردازش افزایش خواهد یافت. پس، مناسبتر است که قبل از انتـ ویژگی، با یک پیش پردازش چنین عبارات یا واژه هایی از متن حذف شوند. چ پیش پردازشی، کلمات توقف[۶۳] مانند افعال کمکی، ضمایر و کلماتی که در حوزه خیلی عمومی هستند (مانند کلمه "سیستم" در حوزه های مهندسی، عا ...) را از مجموعه کلمات حذف می کند. ضمنا، در صورت نیاز می توان قبل انتخاب ویژگی عمل ریشه یابی[۶۴] را انجام داد تا تمام مشتقات واژه ها در قالب ورودی به مولفه انتخاب ویژگی وارد شوند.

مولفه دوم زیرسیستم رسته‌ساز متن که عمده وظایف زیرسیستم را می‌دهد، یک کلاسه‌کننده است. برای آنکه کلاسه‌کننده بتواند مطابق با انتظاری از آن داریم ایفای نقش کند، لازم است تا با مجموعه‌ای از مستنداتی که توسط خبرگان حوزه تخصصی مربوطه، رسته‌بندی شده‌اند (به یک یا چند ما از هستان شناسی منتسب شده‌اند)، آموزش ببیند پس از آموزش، کلاسه می‌تواند مستندات دریافتی را در حوزه مورد نظر، کلاسه‌بندی کند. شایان است؛ مولفه کلاسه کننده برای دسترسی به مستندات آموزشی (در حین آمو و برای انتساب و واکشی سایر مستندات، از خدمات ویژه‌ای که بدین منظو هستان شناسی تعبیه شده است[۲۰]، استفاده می‌کند.

$$\Pr(t)\sum_{i=1}^{m}\Pr(c_i \mid t)\log\Pr(c_i \mid t)+\Pr(\bar{t})\sum_{i=1}^{m}\Pr(c_i \mid \bar{t})\log\Pr(c_i \mid \bar{t}) \qquad (۳)$$



شکل ۱۰- ارتباط زیرسیستم رسته‌ساز متن با سایر اجزای سیستم TeLQAS و خدماتی که به آنها ارائه می‌کند

با توجه به آزمایشاتی که در سال ۲۰۰۰ توسط هان [۶۵] و کاریپیس[۶۶] جهت مقایسه انواع کلاسه کننده‌ها انجام شد، کلاسه‌کننده مبتنی بر مرکز ثقل[۶۷] در مقایسه با سایر کلاسه‌کننده‌های رایج در امور کلاسه‌بندی متون نظیر بیز[۶۸]، C4.5، نزدیکترین همسایه‌ها[۶۹]، از کارایی و دقت عمل بیشتری برخوردار است [۲۱] به همین دلیل نیز، در سیستم TeLQAS ، کلاسه‌کننده مبتنی بر مرکز ثقل، در زیرسیستم رسته‌ساز متن انتخاب شده است. همانطور که از نام این کلاسه‌کننده پیداست، در این روش برای هر کلاس (یا مفهوم هستان شناسی)، یک مرکز ثقل در نظر می‌گیریم که گرانیگاه آن مفهوم، در فضای ویژگی مستندات است. برای مثال، اگر مستند d بر اساس بسامد بعضی از واژه‌ها، به بردار ویژگی معادلش تبدیل شود، یعنی بصورت فرمول (۴):

$$d_{tf} = (tf_1, tf_2, ..., tf_n) \qquad (۴)$$

آنگاه، مرکز ثقل کلاس C با فرمول (۵) قابل محاسبه است.

$$C = \frac{1}{|S|}\sum_{d \in S} d \qquad (۵)$$

که در این رابطه، S تعداد مستندات آموزشی‌ای است که به کلاس (مفهوم) C اختصاص داده شده است. در واقع، محاسبه رابطه فوق، حکم آموزش کلاسه‌کننده را دارد که از نظر پردازشی، بسیار کم هزینه‌تر از روشهای آموزشی مبتنی بر بهینه‌سازی (مانند روش‌های آموزش شبکه‌های عصبی) می‌باشد. پس از مرحله آموزش، برای آنکه کلاسه‌کننده تعیین کند که یک مستند متعلق به چه کلاسی است، کسینوس زاویه بین بردار ویژگی مستند با بردار مرکز ثقل تک تک مفاهیم را محاسبه می‌کند و در نهایت مستند به کلاسی تعلق دارد که کسینوس زاویه بین بردار ویژگی مستند با کلاس مربوطه، بزرگترین مقدار را داشته باشد. کسینوس زاویه مذکور با فرمول (۶) محاسبه می‌شود:

$$\cos(d, C) = \frac{d \cdot C}{\| d \|_2 * \| C \|_2} = \frac{d \cdot C}{\| C \|_2} \qquad (۶)$$

علاوه بر سهولت و کم هزینه بودن آموزش، کلاسه‌کننده مبتنی بر مرکز ثقل، دارای مزیت مهم دیگری نیز هست که آن عدم حساسیت به اندازه مستندات است. در این کلاسه‌کننده ملاک شباهت بردارها، کوچکی زاویه بین آنهاست و نه فاصله بین نوک پیکان آنها (برای نمونه، فاصله اقلیدسی). بنابراین، اندازه یا بزرگی یک بردار تاثیری در اندازه زاویه ندارد.

۶ - واسط مجموعه در TeLQAS

واسط مجموعه، به عنوان گذرگاه ارتباطی TeLQAS با مجموعه های اطلاعاتی مختلف در هر دو وضعیت برون خط و بر خط (زمانی که سیستم نتواند پاسخهای مناسبی را از پایگاه دانش خود استخراج کند)، وظیفه استخراج مستندات متنی را به‌عهده دارد. این مستندات جهت پردازشهای بعدی در بخشهای برخط و برون خط مورد استفاده قرار می‌گیرند (بخشهای ۴ و ۵ مقاله). بطور کلی، وظایف واسط مجموعه در دو وضعیت یاد شده به صورت زیرخلاصه می‌گردد؛

- در حالت بر خط:
  ۱) پیدا کردن مستندات مناسب برای پرس و جوهای توسعه یافته[۷۰] از پرسش کاربر،
  ۲) استخراج متن مستند متناظر با یک URL مشخص شده.
- در حالت برون خط:
  ۱) جمع آوری مستندات متنی متنوع و مرتبط ازمجموعه های اطلاعاتی مختلف،
  ۲) ارائه این مستندات به زیرسیستم رسته ساز متن جهت دسته بندی آنها بر حسب مفاهیم موجود در هستان شناسی.

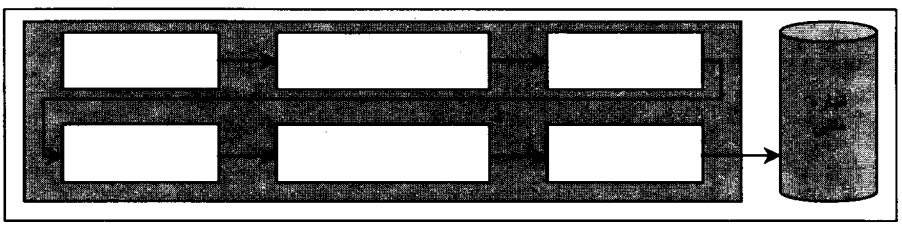
واسط مجموعه از یک سو با هردو بخش برخط و برون خط سیستم در حال تعامل بوده و از سوی دیگر به انواع مجموعه های منابع اطلاعاتی دسترسی دارد. تنوع این منابع، استراتژیهای متفاوتی را برای بهره گیری از اطلاعات موجود در هریک سبب شده است. واسط مجموعه بطور سلسله مراتبی با مجموعه‌های اطلاعاتی تخصصی و موتورهای جستجوی عمومی در ارتباط بوده و به طرق مختلف از امکانات آنها استفاده می‌نماید[۱۵]. در حال حاضر تنها مجموعه اطلاعاتی تخصصی که واسط مجموعه به آن دسترسی دارد، یک مجموعه تخصصی از مستندات مربوط به دامنه فناوری مخابرات در قالب XML می‌باشد، که در TeLQAS تدوین شده است.

۶-۱ معماری واسط مجموعه

از آنجا که در یک پروسه کامل از فعالیتهای واسط مجموعه، چندین زیرپروسه اجرا می‌گردند، واسط مجموعه به شکل چند ماجول جداگانه در نظر گرفته شده است. به این ترتیب، ماجولهای انتخاب مجموعه[۷۱]، مدیریتور[۷۲]، ادغام نتایج[۷۳]، پالایش نتایج[۷۴]، واکشی کننده و مبدل مستند[۷۵]، و ذخیره سازی، اجزاء اصلی بدنه واسط مجموعه را تشکیل می‌دهند. شکل ۱۱ نحوه ارتباط این ماجولها و سایر اجزاء مرتبط را به تصویرکشیده است. هریک از این ماجولها خود حاوی یک یا چند تابع عملیاتی درونی بوده، قادر است بصورت مستقل به اجرای نقش مورد انتظار بپردازد. وظیفه هر کدام از این ماجولها در قسمت بعدی تشریح می‌گردد.

۶-۲ سناریوی عملیاتی واسط مجموعه

فرآیند واسط مجموعه بدین صورت آغاز می‌شود که در ابتدا، پرس وجوهای مطرح شده توسط متخصصان توسعه دهنده‌ی هستان شناسی یا بخش برخط از راه یک مؤلفه پردازنده پرس و جو مورد برخی پردازشها قرار گرفته و به قالب مورد نظر سیستم تبدیل می‌گردد. پرس وجوهای حاصل از این مرحله را پرس وجوهای توسعه یافته می‌نامیم. در درصورتی‌که پرس وجوی مطرح شده اولیه به



شکل ۱۱- دیاگرام عملیاتی واسط مجموعه

فرم $UQK_1 + UQK_2 + ... + UQK_m$ باشد، پرس وجوی توسعه یافته به قالب زیر حاصل خواهد شد؛

$$Extended\ \ Query = [(UQK_1\ \ or\ \ Syn_1^1\ \ or\ \ ...\ \ or\ \ Syn_n^1)\ \ and...and\ \ (UQK_m\ \ or\ \ Syn_1^m\ \ or\ \ ...\ \ or\ \ Syn_n^m)]$$

که در آن UQK_m بیانگر کلمه کلیدی شماره m از پرس وجوی کاربر بوده و Syn_j^i نشانگر کلمه مترادف شماره j برای کلمه شماره i از پرس وجوی کاربر می‌باشد. فعالیت دیگری که درست پس از این مرحله به اجرا می‌رسد، یافتن پرس وجوهای مشابه قبلی ارسال شده به واسط مجموعه است. این امر با دو هدف زیر انجام می‌گیرد:

۱) تعیین ملاک مقایسه مجموعه های متفاوت در ماجول انتخاب مجموعه،

۲) بدست آوردن معیاری برای پالایش نتایج نهایی حاصل شده.

توابع گوناگونی برای شباهت یابی بین پرس وجوها معرفی شده اند که از میان تمام آنها، رابطه مورد استفاده، یک تابع وابسته به کلمات کلیدی موجود در پرس وجوها بفرم فرمول (۷) می‌باشد[۲۲] ؛

$$Sim_{keywords}(p,q) = \frac{KN(p,q)}{Max[kn(p), kn(q)]} \qquad (۷)$$

که $KN(p,q)$ تعداد کلمات کلیدی مشترک میان دو پرس وجوی p و q بوده و $kn(p)$ نیز تعداد کلمات کلیدی پرس وجوی p را نشان می‌دهد. به این ترتیب پرس وجوهای مشابه با پرس وجوی مطرح شده فعلی (در صورت وجود) مشخص شده و در مراحل بعدی مورد استفاده قرار می‌گیرند. پرس وجوهای توسعه یافته نهایی سپس در اختیار ماجول انتخاب مجموعه قرار می‌گیرند. این ماجول نخستین ماجول درونی واسط مجموعه می‌باشد که درخواستهای رسیده را بررسی می‌نماید. این ماجول تنها زمانی فعال می‌گردد که دسترسی به سایر مجموعه های اطلاعاتی (نظیر وب) مورد نیاز باشد. برای این منظور، با محاسبه وضعیت عمومی مجموعه های مختلف در ارتباط با پرس وجوهای مشابه قبلی، از راه ملاحظه بازخورد میزان اعتبار مستندات بازیابی شده برای این پرس وجوها، ماجول انتخاب مجموعه می‌تواند تصمیم گیری مؤثری در گزینش مجموعه های مختلف، چه در مرحله‌ی انتخاب نوع مجموعه و چه در سطح برگزینی یک مجموعه، از خود نشان دهد. راهکارهای فراوانی برای موضوع انتخاب مجموعه ها در سیستمهای بازیابی اطلاعات مطرح می‌باشند[۲۳]، در حالی‌که در این پروژه با افزایش تعداد پرس وجوهای مطرح شده در بخش برون خط، بطور تدریجی یک پایگاه قواعد[۷۶] برای انجام وظیفه اصلی این ماجول ساخته می‌شود.

پس از تعیین مجموعه های مناسب، مدیریتور به فراخوانی این مجموعه ها و دریافت مستندات مرتبط با پرس وجوی توسعه یافته مشغول می‌گردد. عملیات این ماجول در صورتی‌که دسترسی به مجموعه هایی مانند وب لازم باشد، پس از واکشی هایی که ماجول مبدل انجام می‌دهد، یکی از وقت‌گیرترین عملیاتی است که در طول یک چرخه کامل اجرای واسط مجموعه صورت می‌پذیرد. دلیل اصلی این امر محدودیتهای بهره گیری از اتصالات اینترنت برای دستیابی به چنین منابعی می‌باشد. در نهایت، محصول کاری این ماجول، تعدادی مستند متنی حاصل از مجموعه‌های داخلی (مستندات تخصصی گردآوری شده در درون سیستم) یا تعدادی از لیستهای حاوی مستندات متنی بازیابی شده از مجموعه های مختلف خارجی (مجموعه هایی نظیر وب که در خارج از سیستم قرار گرفته اند) خواهد بود که به دلیل تعدد این لیستها نیاز به ماجول دیگری برای ادغام آنها احساس می‌گردد.

به این ترتیب، وظیفه ماجول ادغام نتایج، تهیه لیستی نهایی از مستندات بازیابی شده خواهد بود. در این رهگذر نیز متدهای متفاوتی می‌توانند مورد استفاده باشند[۲۴]، که از میان آنها روش Round Robin برگزیده شده است. ملزومات اصلی یک متد ادغام نتایج را ارضاء می‌نماید که عبارتند از:

- سازگاری بین لیستی[۷۷]: به این مفهوم که اگر رتبه مستند x ازرتبه در یک لیست بازیابی شده بیشتر باشد و این دو مستند درلیست بگیرند، باید همچنان رتبه مستند x از رتبه مستند y بیشتر باشد.
- نتیجه قطعی[۷۸]: این الزام متضمن این امر خواهد بود که برای پرس یکسان و بدون تغییر مجموعه ها، نتایج و ترتیب آنها ثابت باشند.

روش انتخاب شده برای ادغام، بصورت چرخشی تمام مستندات لیستهای واکشی کرده و در لیست ادغام شده جای می‌دهد[۲۴]. این عملیات تا تمام لیستها خالی شوند ادامه خواهد یافت. پس از تعیین لیست نهایی در درصورتی که پرس وجوی مشابهی در سیستم یافته شده باشد، لازم نتایج حاصل برای پرس وجوی فعلی در برابر نتایج حاصل برای پرس مشابه پالایش گردد. این موضوع عدم بازیابی، ذخیره و پردازش مستندات یا غیر مفید را تضمین خواهد نمود. ماجول پالایش نتایج با همین انگیزه مجموعه پیش بینی و طراحی شده است.

لازم به ذکر است که تا بدین مرحله هرجا که به مستندات بازیابی شده است، در حقیقت آدرسURL نماینده آن در واسط مجموعه مورد نظر است. به تعویق انداختن واکشی محتویات مستندات تا این لحظه به سرعت عملیاتی که تا این مرحله انجام گرفته کمک چشم‌گیری داشته همین دلیل در زمان تبدیل مستند بدست آمده از مجموعه های خارجی متنی مورد نیاز، ابتدا باید محتویات مستند از وب درخواست و واکشی رویدادها در ماجول واکشی و تبدیل مستند انجام می‌گیرند. متأسفانه این از سرعت بسیار پایینی در مقایسه با سایر وظایفی که به عهده واسط گذارده شده برخوردار است. البته باید درنظر داشت که یکی از انگیزه گیری از زیرسیستم برون خط به همین موضوع نیز بازمی‌گردد، چراکه برون خط بسیاری از نیازهای کاربران توسط مدیر پیش بینی و پردازش به همین علت، این تأخیر در واکشی محتویات مستندات وب در حد معقو نگران کننده بنظر نمی‌رسد. پس از واکشی محتویات یک مستند در قالب که با استفاده از تکنیک Google Cache صورت می‌پذیرد، لازم اس محتویات در قالب شبیه به XML تدوین شده در سیستم قرار گیرند. متن واکشی شده یک مستند در قالب مذکور در زیر قابل مشاهده است:

```
ument>
    <section id="1">
            In the standard reinforcement-
            learning model، an agent is
            connected to its environment via
            perception and action، as depicted in
            Figure.
    </section>
    ...
    <section id="10">
            Some aspects of reinforcement
            learning are closely related to search
            and planning issues in artificial
            intelligence.
    </section>
ument>
```

بطور کلی چهار قالب متنی در واسط مجموعه پشتیبانی می‌شوند:

۱) قالب متنی شبیه به XML کامل

۲) قالب متنی شبیه به XML ساده

۳) قالب متنی ساده
۴) قالب SGML [۲۵]

به این ترتیب مستندات متنی مرتبط با پرس وجوی مطرح شده در وضعیت برون خط تهیه شده و در پایگاه محلی سیستم ذخیره می‌گردند. این مستندات در آینده توسط زیرسیستم رسته ساز متن کلاس بندی شده و به گره های مرتبط از هستان شناسی سیستم تخصیص می‌یابند.

۳-۶ سلسله مراتب دسترسی به مجموعه‌ها در واسط مجموعه

از آنجا که انواع متفاوتی از منابع اطلاعاتی قابل دسترسی در مقابل واسط مجموعه قرار دارد، ضروری است که این واسط بتواند بگونه ای مؤثر و کارآمد یک مکانیسم مدیریت دسترسی به این مجموعه ها را پیاده سازی نماید. به این علت که در سیستم پرسش و پاسخ TeLQAS نوعی روش سلسله مراتبی در فراخوانی واسط مجموعه درپیش گرفته می‌شود، بهترین راهکار مدیریت دسترسی منابع مختلف توسط واسط مجموعه پیگیری همین مکانیسم در برخورد با منابع موجود و قابل دسترسی می‌باشد. از این‌رو، واسط مجموعه در یک چنین راهکاری، برای یافتن مستندات متنی مورد نیاز در واکنش به تحریک بخشهای برخط و برون‌خط، ابتدا به مجموعه مستندات تخصصی گردآوری شده در درون سیستم مراجعه می‌نماید. در درصورتی‌که نتایج مورد نظر از این منبع تأمین گردند، برخی از فعالیتها در سایر ماجولهای واسط مجموعه یا انجام نشده و یا نقش آنها بسیار کمرنگ خواهد شد. فعالیتهایی نظیر انتخاب مجموعه و واکشی و تبدیل مستند از این نوع می‌باشند.

در شرایطی که نتایج مورد نیاز از مجموعه تخصصی محلی بدست نیایند، دسترسی به مجموعه‌های خارجی صورت خواهد گرفت. این دسترسی از دیدگاه واسط مجموعه و کل سیستم چندان دلخواه نیست، چراکه منجر به فعال شدن ماجولهای یاد شده و ایجاد تأخیر بسیار بیشتر در پاسخگویی به پرس وجوی مطرح شده می‌گردد. البته بخش بزرگی از این تأخیر در کنترل واسط مجموعه نبوده و به محدودیتهای شبکه و سخت افزار باز می‌گردد.

۷- سنجش عملکرد سیستم

اصولا تعیین یک معیار مناسب جهت سنجش نحوه عملکرد یک سیستم بازیابی اطلاعات به ویژه از نوع پرسش و پاسخ، امری دشوار است. مهمترین معضل بر سر تعیین این معیار، مساله غیرقطعی بودن نظر کاربران در خصوص کیفیت یک پاسخ است. این امر برای سیستم TeLQAS دشواری بیشتری به همراه دارد. چراکه همانگونه که قبلا نیز بدان اشاره کردیم این سیستم از نظر میزان استفاده از یک هستان شناسی و پایگاه دانش حوزه تخصصی، ماهیتی منحصربفرد دارد و بنابراین مقایسه آن در قالب معیارهای از پیش تعیین شده برای سایر سیستم بازیابی اطلاعات مقدور نمی‌باشد. همانطور که پیشتر هم اشاره شد، در سالهای اخیر، TREC [۲۶] بعنوان معتبر ترین همایش برای ارزیابی سیستم های پرسش و پاسخ معرفی شده است. در این همایش هر ساله معیاری برای سنجش مزیت سیستم های پرسش و پاسخ تعیین می‌شود و بر اساس آن سیستم های شرکت کننده ارزیابی می‌شوند. اگرچه این معیارها برای ارزیابی سیستم هایی پیشنهاد شده اند که در حوزه های عمومی فعالیت می‌کنند، ولی می‌توان با اعمال پاره‌ای از تغییرات، آنها را برای ارزیابی سایر سیستم های پرسش و پاسخ نیز بکار برد.

برای ارزیابی عملکرد TeLQAS که برای یک حوزه خاص یعنی فناوریهای مخابرات طراحی شده نیز از یکی از معیارهای TREC که در سال ۲۰۰۲ معرفی گردید، استفاده شده است. در این روش، قبلا لیستی از سؤال هایی که ممکن است کاربران یک سیستم پرسش و پاسخ مطرح کنند آماده می‌شود. پس از آن این لیست به سیستم های شرکت کننده داده می‌شود (۵۰۰ سؤال برای TREC 2002). سپس جوابها باید با روش های تماما خودکار از مجموعه متونی که در اختیار سیستم است استخراج شده و جهت ارزیابی در اختیار داوران قرار گیرند. در نهایت امتیاز سیستم براساس معیار وزن اطمینان[۷۹] به صورت فرمول (۸) تعریف می‌شود:

$$ConfidenceWeight = \frac{1}{Q}\sum_{i=1}^{Q}\frac{Number\ of\ correct\ answers\ for\ the\ i\ first\ questions}{i} \quad (۸)$$

که Q نمایانگر تعداد سوالات می‌باشد. ارزیابی پاسخ ها در TREC بشکل دسته ای[۸۰] انجام میشود و فرمول بالا بگونه ای طراحی شده است که نسبت به جای پاسخ های صحیح برگردانده شده در لیست پاسخ ها حساس است. بدین شکل که هرچه پاسخ های صحیح مربوط به ۵۰۰ سوال مطرح شده در رده های بالاتر لیست برگردادنده شده قرار داشته باشند تاثیر بیشتری در افزایش وزن اطمینان آن سیستم دارند. این نحوه ارزیابی برای سامانه های واقعی مناسب نیست چراکه پرسش های در این سیستم ها بصورت برخط و نه دسته ای وارد سیستم میشوند. بنابراین با توجه به ماهیت کاربری TeLQAS و اینکه این سیستم جهت استفاده برخط طراحی شده است، برای محاسبه‌ی وزن اطمینان TeLQAS، معیار فوق بدون توجه به ترتیب سوالات و به‌صورت میانگینی از حالات تصادفی قرار گرفتن سوالات با یک توزیع یکنواخت[۸۱] در لیست پاسخ های برگردادنده شده محاسبه شد. به منظور ارزیابی TeLQAS از ۱۰۰ پرسش به زبان طبیعی که مرتبط به حوزه کاربری سیستم بود استفاده شد. در جمع آوری این پرسشها سعی برآن بوده است که رسته های مختلف توضیح داده شده در بخش ۴-۱ در محتویات این پرسشها قرارگیرند. تعدادی از این پرسشها در زیر آمده اند:

- What is fiber optic?
- What causes photon loss?
- What are the different types of multiple access techniques?
- Is Reflection the same as Optical Return Loss?
- What types of glass exist?
- How do you measure Photon Loss?
- What ATM stands for?
- What are the specifications of Submarine Links?

پرسشهای آزمایشی سیستم همگی از حوزه تخصصی سیستم انتخاب شده اند اما برخی از آنها (حدود ۲۰٪) در گراف آنتولوژی وجود نداشتند و برای پاسخگویی به آنها سیستم نیازمند مراجعه به متون و سپس خلاصه سازی آنها بود ولی حدود ۸۰٪ سوالات به نوعی پرسشی را در مورد مفاهیم موجود در آنتولوژی مطرح می‌کردند.

در همایش TREC معمولا سیستم ها از نظر زمان جوابگویی ارزیابی نمی‌شوند، ولی از آنجایی که TeLQAS به منظور استفاده عملی برخط طراحی شده است مؤلفه زمان نیز نقش مهمی در میزان رضایت کاربران خواهد داشت. به همین دلیل سیستم از نظر زمانی هم مورد ارزیابی قرار گرفته است. برای این منظور زمان turn-around به‌صورت فرمول (۹) محاسبه می‌گردد.

$$Turn\text{-}around\ Time = AnswerCompletionTime - QuestionSubmitionTime \quad (۹)$$

به منظور ارزیابی سیستم، کاربران (داوران) جوابهای سیستم را در چهار کلاس زیر دسته‌بندی می‌کنند:

۱) جوابهای درست: جوابهای دقیق[۸۲] و جوابهای قابل قبول[۸۳]
۲) پاسخهای نادقیق

۳) جوابهای پشتیبانی نشده (جواب درست است ولی در مجموعه متون چنین جوابی ذکر نشده است.)

۴) پاسخهای اشتباه

البته برای استفاده از معیار وزن اطمینان که در بالا معرفی شد، نیاز به یک طبقه‌بندی مطلق (یعنی درست یا نادرست) داریم. به همین منظور ما کلاسهای دوم، سوم و چهارم را به عنوان نادرست و کلاس اول را به عنوان درست در نظر می‌گیریم.

در جدول ۳ کارایی سیستم در سه حالت عملیاتی مختلف نشان داده شده است. سطر اول بیانگر نتیجه ارزیابی سیستم در حالتی است که پاسخها با استنباط مستقیم از روی گراف آنتولوژی به دست می‌آیند. سطر دوم و سوم به ترتیب به استفاده از مستندات رسته بندی شده محلی (که توسط مفاهیم هستان شناسی به آنها اشاره شده است) و دسترسی به وب به عنوان منبع اطلاعاتی دیگر مربوط می‌باشند. البته در ۰ نتایج عملکرد سیستم در حالت مبتنی بر خلاصه سازی (سطر دوم) نشان داده شده است ولی استنتاج مستقیم پاسخ و نیز استفاده از وب، دو مد عملیاتی جدیدتری هستند که در ۰ به آنها اشاره‌ای نشده است.

جدول۳- نتایج ارزیابی سیستم

Turn-around Time	وزن اطمینان	مد عملیاتی سیستم
حدود ۱ ثانیه	بالای ۹۰٪	استنباط مستقیم از هستان شناسی
بین ۱ و ۱۰ ثانیه	بین ۸۰٪ و ۹۰٪	خلاصه سازی مستندات محلی رسته بندی شده / مجموعه تخصصی مسندات
بالای ۱۰ ثانیه	زیر ۵۰٪	خلاصه سازی مجموعه وب

این نتایج نشانگر عملکرد قابل توجه وکارایی چشمگیر سیستم TeLQAS به عنوان یک سیستم پرسش و پاسخ است. همچنین مقایسه این نتایج با آنچه در ۰ منتشر شده نشان می‌دهد که با اضافه شدن مولفه‌ی استنباط معرفی شده، کارایی سیستم به میزان قابل توجهی بهبود پیدا کرده و در کنار آن، زمان turn-around بدست آمده از استدلال مستقیم از هستان شناسی نیز از کاهش یافته است.

۸- نتیجه گیری و کارهای آینده

در این مقاله سیستم پرسش و پاسخ TeLQAS که در حوزه فناوری مخابرات طراحی و پیاده سازی شده مورد بررسی قرار گرفت. این سیستم، مبتنی بر یک گراف هستان شناسی حاوی مفاهیم مرتبط با دامنه فناوری مخابرات بوده و می‌تواند پرسشهای زبان طبیعی کاربر را به زبان انگلیسی پردازش کرده و پاسخهای مناسب را از طریق یک واسط در اختیار وی قرار دهد. در کنار این بخش برخط، که مستقیما با کاربر در تعامل است، جهت بهبود کارایی سیستم از دو منظر سرعت و دقت پاسخگویی، پردازشهایی بصورت برون خط صورت می‌گیرد که طی آن، مستندات و متون بازیابی شده توسط بخش واسط مجموعه سیستم، رسته بندی شده و به مفاهیم مرتبط در گراف هستان شناسی منتسب می‌گردند. این سیستم می‌تواند کار پاسخگویی به پرسشهای کاربر را بصورت مستقیم از هستان شناسی، یا به شکل غیر مستقیم از بخشهایی از مستندات منتسب به مفاهیم هستان شناسی، که توسط یک زیرسیستم خلاصه سازی متن شناسایی و استخراج می‌گردند، به انجام برساند. همچنین این سیستم دارای یک بخش واسط مجموعه، برای برقراری ارتباط و دسترسی به اطلاعات متنی موجود در مجموعه های اطلاعاتی موجود، نظیر وب، می‌باشد که در هردو وضعیت برخط و برون خط فعال است. نتایج بدست آمده از آزمایش سیستم با ۱۰۰ پرسش در دامنه فناوری مخابرات، نشان داده است که این سیستم در مقایسه با سیستمهای مشابه موجود از دقت و سرعت قبولی برخوردار می‌باشد. این آزمایشها با استفاده از مجموعه مستندات تخ جمع آوری شده در پروژه صورت گرفته و به همین دلیل، در نخستین توسعه و بهبود این سیستم، استفاده از داده های استاندارد موجود در م هایی همانند TREC مورد توجه خواهد بود، تا بدین طریق امکان محا مقایسه دقیقتر و استانداردتر کارایی سیستم با سایر سیستمهای مشابه ایجاد برای بهبود و تکمیل عملکرد TeLQAS در بخش برخط در گام بعدی جای منطق استدلالی موجود برای استخراج پاسخ از هستان شناسی با منطق اس مقبول[۲۸] است. منطق استدلالی مقبول سعی در تشخیص الگوهایی د انسان‌ها هنگام نتیجه‌گیری بر اساس شواهد ناقص بدست آمده از محیط خود، از آنها استفاده می‌کنند. در این منطق از الگوهای استدلالی استفاده م که در دیگر منطق های کلاسیک وجود ندارند مثل، شباهت و عدم شباهت مفاهیم و تعمیم و تخصیص حقایق.

مرحله بعدی جایگزینی روش خلاصه سازی فعلی با یک روش بهتر است ترکیب خلاصه ها کیفیت خلاصه نهایی بهبود یابد. همچنین با افزودن Query Navigation به بخش برخط، کاربر در ارائه پرسشهای بهتر و دریافت پاسخهای مرتبطتر راهبری می شود و بالاخره با اضافه کردن مدلسازی کاربر، TeLQAS قادر می گردد که با شناخت بهتر کاربر وی دریافت پاسخ بهتر یاری دهد.

از سوی دیگر، بمنظور غنی‌تر نمودن منابع قابل دسترسی توسط واسط م طراحی و پیاده‌سازی یک کراولر تخصصی برای فناوری مخابرات مورد نظ گرفته است. این کراولر قادر خواهد بود که یک مجموعه تخصصی از مستند مرتبط با دامنه کاری سیستم را تهیه کرده و به این ترتیب منبع اطلاعاتی مناسبی را در اختیار واسط مجموعه بگذارد. این موضوع می‌تواند وابستگی و برخط سیستم را به مجموعه‌های اطلاعاتی خارجی، نظیر وب، تا حد بسیا کاهش داده و کارایی کلی سیستم را از دیدگاه دقت و سرعت پاسخگوی بخشد.

قدردانی و تشکر

این پژوهش به عنوان یک پروژه تحقیقاتی در مرکز تحقیقات مخابرات ایر شده است. نویسندگان این مقاله مراتب قدردانی و سپاس خویش را از آق کامبیز بدیع، مدیر پژوهشکده فناوری اطلاعات و دکتر نصراله مقدم تحقیقات مخابرات ایران، بخاطر ارائه ایده‌ها و راهنمایی‌های بی‌دریغ می‌دارند.

مراجع

M. Abasolo and M. Gomez, "MELISA: An Ontology-sed Agent for Information Retrieval in Medicine," *:DL 2000 Workshop on the Semantic Web Lisbon*, rtugal, 2000.

Moldovan, S.Harabagiu, R. Gîrju, P. Morărescu, F. cătuşu, A.Novischi, A. Bădulescu, and O. Bolohan, c Tools for Question Answering," *Proceedings of the h Text REtrieval Conference (TREC-2002)*, ithersburg, MD, pp. 144-155,2002.

Yang, T. S. Chua and S. Wang, "Modeling Web owledge for Answering Event-based Questions," *12th l World Wide Web conference (WWW '03)*, Hungary 13.

Conf. on Information and Knowedg Engineering'03, vol. 2, pp. 500-504, 2003.

[15] M.R. Hejazi, K. Neshatian, and B.R. Ofoghi, "A System Developed for Automatic Extraction and Categorization of Telecommunication Literatures Used in TeLQAS," *Proceedings of International Symposium on Telecommunications*, pp. 571-575, Isfahan, Iran, 2003.

[16] K. Neshatian, and M. R. Hejazi, "Text Categorization and Classification in Terms of Multi-Attribute Concepts for Enriching Existing Ontologies," *Proceedings of 2nd WITID Conference on Information Technology and its Disciplines*, pp. 43-48, Kish Island, Iran, 2004.

[17] Y. Yang, and X. Liu, "A Re-Examination of Text Categorization Methods," *Proceedings of 22nd annual international ACM SIGIR Conference on Research and Development in information retrieval*, pp. 42-49, Berkeley, California, 1999.

[18] Y. Yang, and J.O. Pederson, "A Comparative Study on Feature Selection in Text Categorization," *Proceedings of the International Conference on Machine Learning*, 1997.

[19] T. Mitchell, *Machine Learning*, McGRAW- Hill, 1997.

[20] K. Neshatian, and M.R. Hejazi, "An Object Oriented Ontology Interface for Information Retrieval Purposes in the Domain of Telecommunications," *Proceedings of International Symposium on Telecommunications*, pp. 677-681, Isfahan, Iran, 2003.

[21] E. H.Han, and G. Karypis, "Centroid-Based Document Classification: Analysis & Experimental Results," *The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases*, Lyon, France, 2000.

[22] J. I. R Wen, J. Y. Nie, and H. J. Zhang, "Query Clustering Using User Logs," *ACM Transactions on Information Systems*, vol. 20, no. 1, 2002.

[23] D. Hawking, and D. Paul Thistlewaite, "Methods for Information Server Selection," *ACM Transactions on Information Systems*, vol. 17, no. 1, 1999.

[24] R. L. Yager, and A. Rybalov, "On the Fusion of Documents from Multiple Collection Information Retrieval Systems," *Journal of the American Society for Information ScienceV.* 49, Issue 13, pp. 1177 - 1184, 1998.

[25] R. Baeza-Yates, and B. Ribeiro-Neto, *Modern Information Retrieval*, ACM Press Inc., New York, 1999.

[26] E. M. Voorhees, "Overview of the TREC 2002 Question Answering Track," *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*, 2002.

[۲۷] م.ر. حجازی، م. میریان، ا. جلالی، ب. عبدالهی و س. بابازاده، "TeLQAS : سیستم پرسش و پاسخ مبتنی بر هستان شناسی برای فناوریهای مخابراتی" هشتمین کنفرانس انجمن کامپیوتر ایران، مشهد، ۱۳۸۱.

[4] B. Magniti, M. Negri, R. Prevete, and H. Tanev, "Mining Knowledge from Repeated Co-occurrences: DIOGENE," *Proceedings of the 11th Text Retrieval Conference (TREC-2002)*, 2002.

[5] M. Montes-y-Gómez, A. López-López and A. Gelbukh, "Information Retrieval with Conceptual Graph Matching," *Proceedings of DEXA-2000, 11th International Conference and Workshop on Database and Expert Systems Applications*, Greenwich, England, Lecture Notes in Computer Science, Springer, 2000.

[6] S. Aitken, and S. Reid, "Evaluation of an Ontology-Based Information Retrieval Tool," *Workshop on the Applications of Ontologies and Problem-Solving Methods*, (eds) A. Gómez-Pérez, V.R. Benjamins, N Guarino, and M.Uschold, European Conference on Artificial Intelligence, Berlin, 2000.

[7] T.R. Gruber and G.R. Olsen, "An Ontology for Engineering Mathematics," Technical Report KSL-94-18, Knowledge Systems Laboratory, Stanford University, Palo Alto, CA. 1994.

[8] N. F. Noy, R. W. Fergerson, and M. A. Musen, "The Knowledge Model of Protege-2000: Combining interoperability and flexibility," *2th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000)*, Juan-les-Pins, France, 2000.

[9] G. S. Mann, "Fine-Grained Proper Noun Ontologies for Question Answering," *SemaNet'02: Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

[10] P. Mitra, M. Kersten, and G. Wiederhold, "A Graph-Oriented Model for Articulation of Ontology Interdependencies," *Proceedings of 7th Conference on Exlending Database Technologies*, Konstanz, Germany, 2000.

[11] W. N. Borst, *Construction of Engineering Ontologies*, PhD thesis, University of Twente, Enschede, 1997.

[12] E. Riloff, and M. Thelen. "A Rule-based Question Answering System for Reading Comprehension Tests," *ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Seattle, WA, 2000.

[13] M. S. Mirian, M. R. Hejazi, and E. Darrudi, "Finding Answers through a Heuristic Reasoning Mechanism For an Ontology-based Question Answering System," *Proceedings of 2nd WITID Conference on Information Technology and its Disciplines*, pp. 30-35, Kish Island, Iran.

[14] M. R. Hejazi, M. S. Mirian, K. Neshatian, A. Jalali, and B. R. Ofoghi, "TeLQAS: A Telecommunication Literature Question/Answering System Benefits from a Text categorization Mechanism," *Proceedings of Int.*

[28] A. Collins, and R. S. Michalski, "The Logic of Plausible Reasoning: A Core Theory," *Cognitive Science*, vol. 13, pp. 1-49, 1989.

[1] Question Answering
[2] Ontology
[3] Telecommunication Literature Question Answering System
[4] Indexing Agents
[5] Open-Domain
[6] Text Retrieval Conference: http://trec.nist.gov
[7] Specific-Domain
[8] Priori Knowledge
[9] Technical Data Sheets
[10] MEdical Literature Search Agent
[11] Abstract
[12] Aggregation
[13] Instances
[14] Information Base
[15] Semantic
[16] Online
[17] Offline
[18] Document Warehouse
[19] Information Retrieval
[20] Integrated
[21] Document Warehouse
[22] Semantic Web
[23] Vertexes
[24] Domain
[25] Telecommunications
[26] Sub-Domain
[27] Sub-Ontology
[28] Concept
[29] Instance
[30] Attribute
[31] Slot
[32] Internal
[33] External
[34] Relation
[35] Equal Concepts
[36] Overriding
[37] Ontology Programming Interface
[38] Logical or Physical View
[39] Procedure
[40] Data Services
[41] Properties
[42] Document Class
[43] Related Instances
[44] Polymorphism
[45] Node Object
[46] Collection Classes
[47] Two Phase Population
[48] Natural Language
[49] Part of Speech Tagger (POST)
[50] Rule-based Classifier
[51] Focused Node (Focused Concept)
[52] Desired Relation
[53] Term Frequency
[54] Text Categorization Subsystem
[55] Collection Interface
[56] Feedback Information
[57] Document Classification
[58] Feature Vector Construction
[59] Feature Selection
[60] Hyper-Space
[61] Yiming Yang
[62] Jan O. Pederson
[63] Information Gain
[64] Stop Words
[65] Stemming
[66] Eui-Hong(Sam) Han
[67] George Karypis
[68] Centroid-based Classifier
[69] Bayesian
[70] K Nearest Neighbor
[71] Extended Queries
[72] Collection Selection

iation
lts Fusion
lts Refinement
ment Fetch and Conversion
Base
List Consistency
rministic Results
idence Weight
orm
Answers
ible Answers

محمودرضا حجازی مدرک کارشناسی و کارشناسی ارشد خود را به ترتیب در سالهای ۱۳۷۱ و ۱۳۷۴ در رشته مهندسی برق و مخابرات از دانشگاههای صنعتی شریف و تهران دریافت کرد. سپس، طی سالهای ۱۳۷۷ تا ۱۳۸۲ در پژوهشگاه فناوری اطلاعات مرکز تحقیقات مخابرات ایران، ابتدا بعنوان محقق و سپس مدیر پروژه به فعالیتهای تحقیقاتی در حوزه کاربردهای پیشرفته اطلاعات ادامه داد. در حال حاضر، وی در حال گذراندن دوره دکتری در پردازش تصویر در دانشکده مخابرات و فناوری اطلاعات دانشگاه علوم و گوانگجو (GIST) کره جنوبی است و حوزه تخصصی تحقیقات وی در ار روشها واستانداداهای آنالیز و فشرده سازی تصویر و ویدئو و نیز بازیابی مح محتوای تصاویر می باشد.

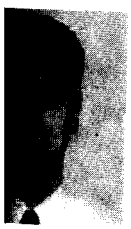
آدرس پست الکترونیکی نامبرده عبارتست از :

jazi@gist.ac.kr

مریم سادات میریان حسین آبادی مدرک کارشناسی و کارشناسی ارشد خود را در رشته مهندسی کامپیوتر در دانشکده فنی، دانشگاه تهران اخذ نمود و در حال حاضر دانشجوی دکترای هوش مصنوعی و رباتیک در دانشگاه تهران است. وی به تحقیقات در زمینه پردازش زبان طبیعی، بازیابی اطلاعات و یادگیری چندعامله علاقه مند است و در حال حاضر در پژوهشکده کاربردهای اطلاعات مرکز تحقیقات مخابرات ایران به عنوان عضو تیم مدیریت طرح سازمان الکترونیکی مشغول به تحقیق است.

آدرس پست الکترونیکی نامبرده عبارتست از :

n@ut.ac.ir



کوروش نشاطیان تحصیلات کارشناسی ارشد خود را در رشته هوش مصنوعی و رباتیک به پایان رسانده است و هم اکنون دانشجوی دکتری نرم افزار در دانشگاه آزاد اسلامی واحد علوم و تحقیقات می‌باشد. زمینه تحقیقاتی وی داده‌کاوی و محاسبات نرم می‌باشد. او در حال حاضر سرپرستی یک تیم پژوهشی را در پژوهشگاه ارتباطات و فناوری اطلاعات مرکز مخابرات ایران به عهده دارد.

آدرس پست الکترونیکی نامبرده عبارتست از :

itrc.ac.ir

بهادررضا افقی مقاطع کارشناسی و کارشناسی ارشد را در دانشگاه آزاد اسلامی بترتیب در رشته های مهندسی نرم افزار و هوش مصنوعی به پایان رسانده است و هم اکنون دانشجوی دکترای دانشگاه Ballarat استرالیا در رشته‌ی فناوری اطلاعات است. زمینه های تحقیقاتی مورد علاقه‌ی ایشان بازیابی اطلاعات (بطور خاص، پرسش و پاسخ)،Semantic Structures/Analysis و Web User Modelling است.

آدرس پست الکترونیکی نامبرده عبارتست از :

br_ofoghi@itrc.ac.ir

احسان درودی دانشجوی کارشناسی ارشد دانشگاه تهران در رشته مهندسی نرم افزار است. زمینه ای تحقیقاتی مورد علاقه وی بطور عام شامل بررسی چگونگی فرایند تفکر و استدلال در انسان و ماشین، و بطور خاص تئوری استدلال مقبول انسانی است. وی هم اکنون در صندوق حمایت از پژوهشگران مشغول بکار است و تجربه حضوری دو ساله را نیز در مرکز تحقیقات مخابرات ایران دارد.

آدرس پست الکترونیکی نامبرده عبارتست از :

e.darody@ece.ut.ac.ir