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

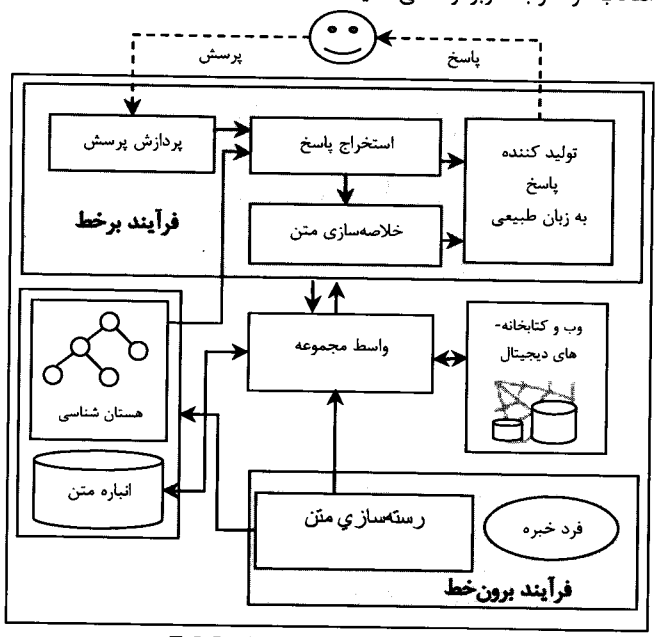
با توجه به طرحی که در نظر داریم، می‌توان بخشهایی چون هستان شناسی، مولفه بازیابی اطلاعات و مولفه‌های اتصال به اینترنت را از اجزاء حتمی معماری TeLQAS دانست. هرچند تکمیل معماری کنونی TeLQAS یک روند تکاملی داشته و آنچه در اینجا ارائه می‌شود حاصل چندین بار نمونه‌سازی، آزمایش و پالایش طرح بوده است، اما آنچه 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.

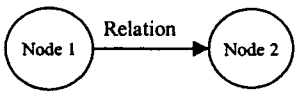
۲) طراحی و ایجاد محیطی جهت بیان و ارتباط با هستان شناسی به نحوی که سایر زیرسیستم‌ها بتوانند بدون در نظر داشتن جزئیات ذخیره سازی، در بالاترین سطح انتزاع به عناصر هستان شناسی دسترسی داشته باشند. جهت تحقق بخشیدن به این نیاز، یک معماری سه لایه با ویژگیهایی منحصر بفرد تولید شده که در ادامه به شرح آن می‌پردازیم.

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:

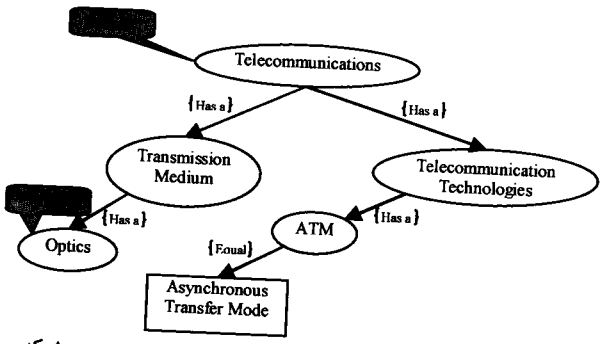
- ❖ *حوزه*[۲۳]: این گره نمایانگر یک حوزه می‌باشد. به عنوان مثال مقوله مخابرات[۲۴] یک حوزه است. یک حوزه مفهومی بسیار کلی و عام است که در بالاترین نقطه یک هستان شناسی قرار می‌گیرد. یک حوزه معمولا خود دارای چند زیر-هستان شناسی است. به همین ترتیب "هنر" نیز یک حوزه محسوب می‌شود.
- ❖ *زیرحوزه*[۲۵]: این گره 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.



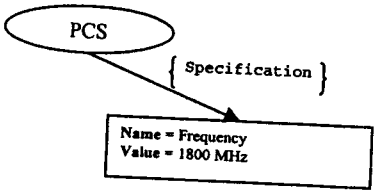
شکل۳- قسمتی از گراف هستان شناسی شامل حوزه مخابرات و مدیریت شبکه

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 با 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 می‌شود 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 مفهوم یک نام وجود داشته باشد. از آنجا که معمولا حوزه‌ها تخصصی هستند (مانند مخابرات)، همواره تعریف مشخص و متمایزی برای هر مفهوم حوزه وجود دارد این ترتیب ابهامی در این خصوص که دو مفهوم با یک نام ولی با معانی مختلف 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 هستند و باید در هم ادغام شوند. به این ترتیب اتصالات بین زیرحوزه‌ای پدید می‌آید 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 مخابرات

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

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:

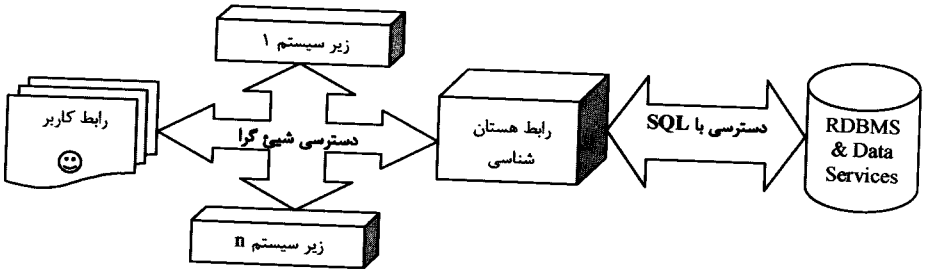
۱) لازم است بتوان در مدل ارائه شده کل دانش حوزه را بر اساس مجموعه‌ای از گره‌ها و روابط بیان کرد بگونه‌ای که تمامی عملیات متداول در پروسه‌های بازیابی اطلاعات بر روی مدل جدید نیز قابل اجرا باشد.
۲) تمام موجودیتهایی که این رابط ارائه می‌دهد باید بصورت شئ‌گرا باشد.
۳) خدمات رابط باید مستقل از روش ذخیره‌سازی 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.

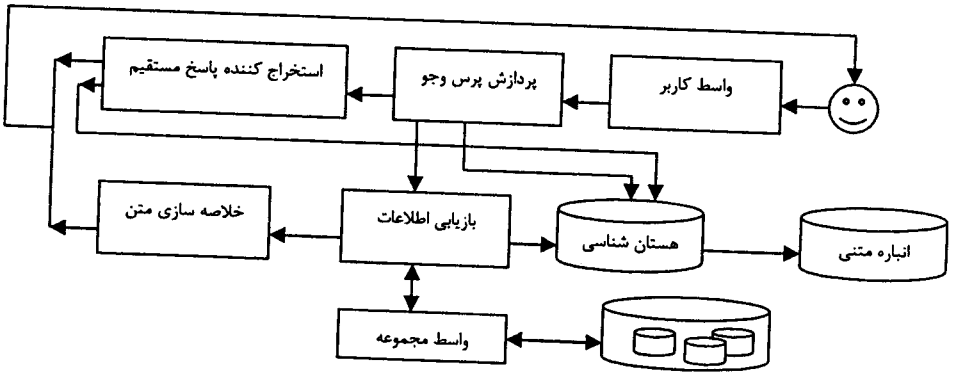
از آنجاکه ساختار هستان شناسی‌ها عموما سلسله مراتبی می‌باشند، بسیاری از استانداردها و نرم‌افزارهایی که در این رابطه تهیه شده‌اند، از یک فرم سلسله‌مراتبی برای ذخیره‌سازی داده استفاده می‌کنند که در این میان می‌توان از فرمتهای RDF و XML نام برد[۱۱]. هرچند این نوع ساختمان داده بسیاری از مسائل طراحی و پیاده‌سازی 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 استخراج نموده به کاربر نمایش بدهیم.

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

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

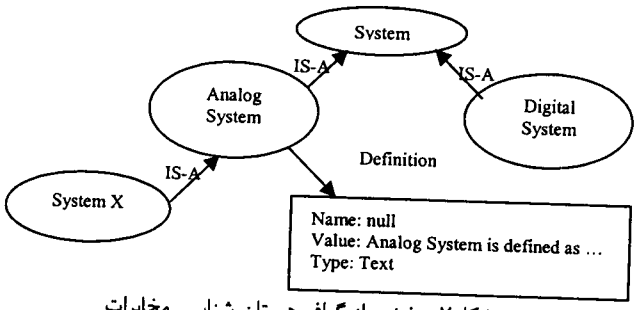
باز هم گره مورد تأکید به سادگی و مانند قبل تعیین می‌گردد؛ Cable X و Cable Y. اما از آنجا که هستان شناسی دارای نوع ارتباطی Similarity و Difference نیست، باید به نحوی از مرتبط ترین نوع ارتباطی موجود استفاده شود. بنابراین نیاز به پردازشی بیشتر برای بدست آوردن شباهت و تفاوت این دو گره وجود دارد. با مراجعه 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 مفهـوم فرزند تلقی گـردد، 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.

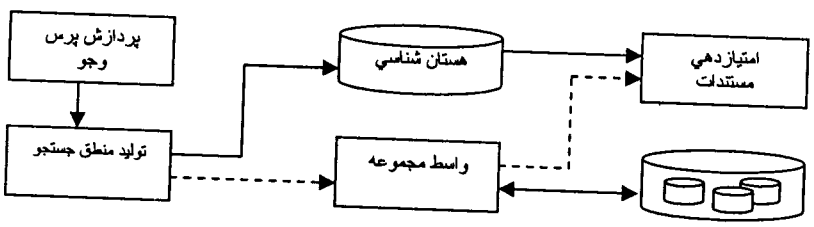


شکل۷- بخشی از گراف هستان شناسی مخابرات

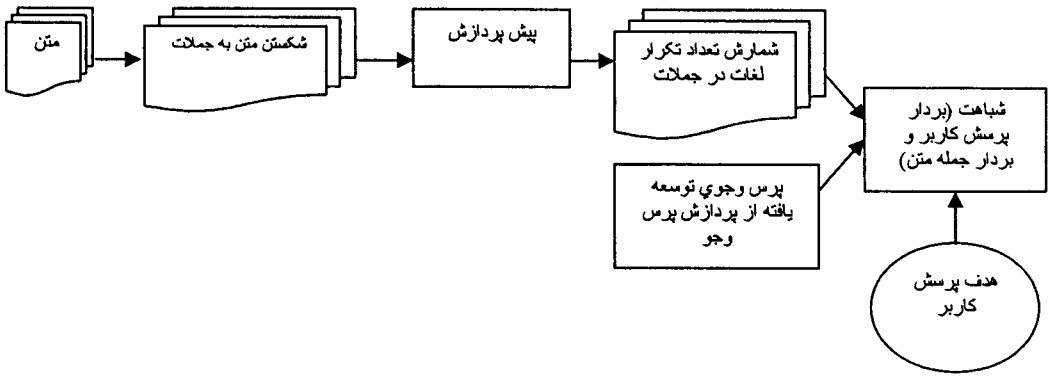
۴-۳ مؤلفه بازیابی و امتیازدهی به مستندات

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