علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۴ شماره ۱ (الف)، ۱۳۸۵
صفحات ۱۱-۱۸
مقاله کوتاه

نشریه علمی پژوهشی انجمن کامپیوتر ایران

یک سیستم پردازش درخواست مبتنی بر واژگان شناختی برای پاسخ دهی به پرسشها

احمد عبدالله زاده قربان خردمندیان

آزمایشگاه سیستمهای هوشمند، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیر کبیر، تهران ،ایران

چکیده

در این مقاله یک سیستم پاسخ دهی به پرسش مبتنی بر واژگان شناختی معرفی شده است. در واقع با توجه به اهمیت روزافزون بازیابی اطلاعات مورد نیاز از منابع اطلاعاتی عظیم و با توجه به قابلیتهای تکنیکهای پردازش زبان طبیعی، این سیستم پیشنهاد می شود. همچنین به دلیل عدم پردازش مفهوم توسط اکثر سیستمهای موجود، در این سیستم واژگان شناختی به عنوان یک راهکار برای شناسایی مفاهیم و روابط صریح موجود بین مفاهیم برای یک قلمرو مشخص در راستای پردازش مفهوم معرفی شده است. کارایی سیستم پیشنهادی در پاسخ دهی به یک مجموعه مدارک مرتبط با خوشه بندی مورد بررسی قرار گرفت و مشخص شد که استفاده از واژگان شناختی در افزایش دقت پاسخگویی نقش بسزایی دارد و همچنین آزماش با واژگان با ابعاد مختلف نشان داد که هر چه واژگان گسترده تر باشد به همان میزان دقت پاسخ گویی نیز افزایش خواهد یافت.

واژه‌های کلیدی: پردازش زبان، پاسخ دهی به پرسشها، واژگان شناختی، شناخت مفهومی، دسته بندی پرسشها

۱- مقدمه

از آنجاییکه اطلاعات برخط[1] با حجم بسیاروسیعی هم اکنون موجود است، نیاز به سیستمهای اتوماتیک پاسخ دهی به پرسشها بسیار محسوس میگردد. در واقع ما نیازمند سیستمهایی هستیم که به کاربران این امکان را بدهند که با هر زبانی از آنها پرسش شود و سیستم نیز بتواند با همان زبان پاسخ گو باشد. واسط کاربر های فعلی، با استفاده از زبانهای غیر رسمی و غیر طبیعی، فقط لیستی از اسناد بازیابی شده را به عنوان پاسخ به کاربر ارائه می کنند.

برای پاسخگویی به پرسشها، یک سیستم باید پرسش را آنالیز کرده که احتمالاً این آنالیز در زمینه مورد محاوره صورت می گیرد. سیستم باید یک و یا بیشتر پاسخ، با توجه به مشورت با منابع برخط موجود، فراهم کند و همچنین در یک غالب مناسب آنرا برای کاربر ارائه کند. کنفرانسها و کارگاههای آموزشی زیادی هم اکنون بر روی جنبه های مختلف زمینه تحقیقاتی پاسخ دهی به پرسشها متمرکز شده اند. با شروع از سال۱۹۹۹، کنفرانس بازیابی متن موسوم به TREC (Text Retrieval Conference) یک شاخه ای نیز برای ارزیابی سیستمهای پاسخ دهی به پرسشها ایجاد کرده است و برای ارزیابی سیستمها نیز از اسناد موجود در متن های موضوعی مفصل خود TREC استفاده می کند [1].

سیستمهایی که در TREC شرکت می کنند اغلب بعد از دریافت پرسش مراحل مختلفی را برای ارائه پاسخ مناسب انجام می دهند [1,2,5,9,11]. در مرجع [1] به این مراحل، به شرح زیر عنوان شده است:

۱- کلمات کلیدی پرسش ورودی را از نظر املائی بررسی می کنند.

۲- اجزا و ترکیبات جمله مشخص و همچنین مفاهیم موجود شناسایی و وابستگی بین مفاهیم استخراج می شود. که در حقیقت، وابستگی مفهومی اجزاء جمله را بازنمایی می گردد. مثلاً برای پرسش How much could you rent a Volkswagen bug in 1966? که آنرا Q1 می نامیم، یک وابستگی دودویی بین مفاهیم rent و 1966 در نظر گرفته می شود.

۳- با استفاده از روابط و وابستگی مفهومی استخراج شده نوع پاسخ مشخص می شود. برای مثال برای پرسش Q1 با توجه به وابستگی بین How much و rent پاسخ باید از نوع Money باشد.

۴- با انتخاب کلمات کلیدی موجود در پرسش ورودی، امکان جستجو در مجموعه مدارک موجود و بازیابی مدارک مناسب فراهم می گردد. برای Q1 می توانیم

کلمات Volkswagen و bug را به عنوان کلمات کلیدی برای جستجو در مجموعه مدارک انتخاب کنیم.

۵- در این قسمت مجموعه مدارک موجود برای یافتن کلمات گام چهارم مورد جستجو قرار می گیرند و مدارک و پاراگرافهایی که این کلمات کلیدی در آنها تکرار شده باشند استخراج می شوند.

۶- همه پاراگرافها ممکن است که مناسب نباشند در این گام پاراگرافهای نا مرتبط فیلتر می شوند. مثلاً آنهایی که 1966 را در خود ندارند باید حذف شوند.

۷- از بین پاراگرافها و مدارک باقی مانده بایستی پاسخهای کاندید مناسب انتخاب شوند. برای Q1 باید پاسخی تولید شود که از نوع Money باشد. برای مثال اگر در پاراگرافهای استخراج شده کلماتی نظیر $1 و یا USD 520 موجود باشند مناسب خواهند بود.

۸- بعد از انتخاب پاسخهای مناسب در این فاز پاسخها بر اساس معیار های تعریف شده، میزان مرتبط بودنشان مشخص شده و رتبه بندی می شوند.

۹- پاسخ تولید خواهد شد. پاسخی مانند: rent a Volkswagen bug for $1 a day برای پرسش Q1 تولید خواهد شد.

با بررسی این سیستمها متوجه می شویم که آنها فقط متکی بر کلمات موجود در پرسش ارائه شده هستند و ممکن است که در بعضی مواقع پاسخ مناسبی یافت شود. اما در حالت کلی در جاهایی که کلمات کلیدی موجود در پرسش در مجموعه مدارک موجود یافت نشوند، نمی توان پاسخ صحیحی استخراج کرد. مهمترین مشکل این سیستمها عدم توانایی در پردازش مفهوم است[1]. این سیستمها پس از دریافت پرسش، با استفاده از تکنیکهای تطبیق پذیری کلمات مدارک با کلمات پرسش به بازیابی مدارک می پردازند. به عبارات دیگر سیستم هیچ درکی از معنا و مفهوم پرسشهای ارائه شده در قالب ورودی سیستم نداشته و در عملیات پردازش بیشتر متکی بر تکنیکهای آماری هستند. عملکرد این سیستمها جهت سازماندهی و بازیابی بهینه مدارک و اطلاعات با فرآیند ساخت نمایه ها۲ آغاز میشود. اعمال منطق محاسباتی به مدارک موجود، دیدگاههای منطقی خاصی بوجود میاورند. آنچه که واضح است غیر قطعی بودن و عدم تطابق این دیدگاهها با محتوای واقعی و معنایی آنهاست. طبیعتاً این فاصله در دقت و سازماندهی و بازیابی اطلاعات و در نهایت در پاسخ گویی به پرسشهای کاربران تأثیر منفی فراوان دارد. با این دیدگاه بازنمایی انجام شده توسط سیستم باید حداکثر نزدیکی را به محتوا داشته باشد اما روشهای آماری و احتمالاتی صرف ، قادر به انجام این مهم نیستند.

با رشد بی سابقه سیستمهای اطلاعاتی برخط، مثل اینترنت که مدارک بزبان طبیعی و تمام متن را سازماندهی میکند ضرورت و اهمیت وجود یک منبع زبان مثل سیستمهای طبقه بندی، سرعنوان موضوعی، اصطلاحنامه[3] و یا هستان شناسی[4] (که در این مقاله واژگان شناختی نامیده شده است) مشخص و واضح شده است. در کارهای انجام شده در [12] و [13] سعی شده است که از هستان شناسی برای پردازش مفهوم استفاده شود. در این دو کار آنچه که به عنوان هستان شناسی بکار فته است بیشتر شبیه تزاروس یا اصطلاحنامه است و ساختاری مانند WordNet که در [12] در بخش توسعه پرسش ارائه شده است و استفاده از واژه های مترادف با کلمات کلیدی پرسش را نیز برای جستجو در پایگاه مدارک پیشنهاد می کند، با آنچه که ازآن به عنوان هستان شناسی (یا واژگان شناختی که در این مقاله مورد استفاده قرار گرفته است) یاد میشود، تفاوت دارد.

در این مقاله یک سیستم پاسخ دهی به پرسشها معرفی می شود که برای غلبه بر مشکل یاد شده در این سیستم از واژگان شناختی استفاده شده است.

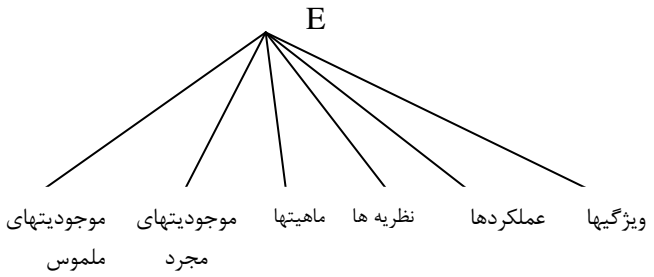
۲- واژگان شناختی

واژگان شناختی در زمینه های متعددی در علوم کامپیوتر بکار می رود. یک تعریف جامع و رایج برای تعریف آن عبارت است از:

یک واژگان شناختی یک توصیف واضح فرمال مفاهیم در ارتباط با هم در یک محدوده مشخص می‌باشد [4]. منظور از ارتباط مفاهیم، یک دید انتزاعی از یک قلمرو می باشد که مفاهیم مناسب برای نمایش آن قلمرو را شناسایی می کند. این دید باید دانشی را منعکس کند که بصورت اشتراکی مورد استفاده قرار گیرد. یک واژگان شناختی مفهوم سازی را با تشخیص مفاهیم و روابط واضح بین انها توسط زبانهای فرمال ارائه می کند[3].

۲-۱ طرح و مدل مفهومی موجودیتهای هستی در واژگان مفهومی

چارچوب شناختی واژگان دارای یک نقطه آغازین است که با E نمایش داده میشود. بقیه موجودیتهای هستی زیر نقطه آغاز و متصل به آن قرار داده می‌شوند [۱۴]:



شکل۱. چارچوب هستی شناختی موجودیتها و مفاهیم قابل سازماندهی در واژگان شناختی[۱۴]

نکته قابل توجه آن است که ماهیتها و نظریه ها خود زیر مجموعه موجودیتهای مجرد یا انتزاعی هستند، اما اتخاذ این دیدگاه سبب می شود تا در صورت ازدیاد موجودیتها و وجود روابط گسترده میان آنها، یک شبکه معنایی پیچیده بوجود آید که کار سازماندهی و پیاده سازی را دچار پیچیدگی بیشتری می کند. برای رفع ابهام ، ارائه وانتقال خردمندانه مفهوم کامل یک واژه به پرسشگر لازم است که موارد زیر در ساختار واژگان لحاظ شود:

۱- کلیه نظریه های موجود در مورد مفهوم مورد نظر

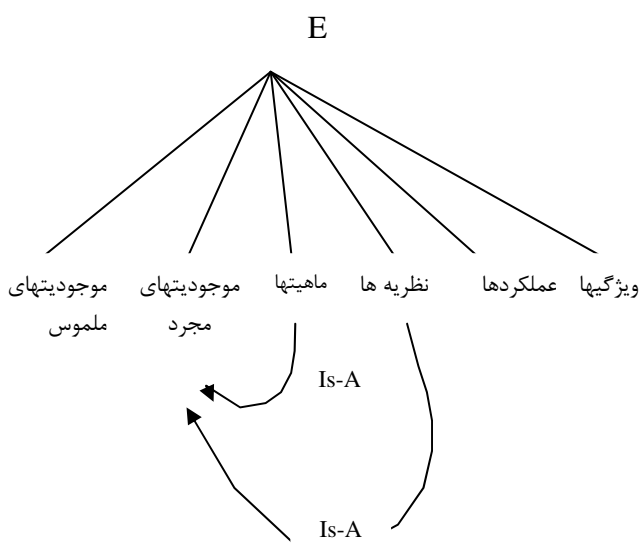
۲- برداشتهای رایج در مورد مفهوم مورد نظر

۳- ماهیت مفهوم مورد نظر

۴- عملکرد و ویژگی مفهوم موردنظر

تنها در این شرایط است که جمیع معانی و روابط موجود میان مفهوم مورد نظر پرسشگر و سایر مفاهیم راهبردی مرتبط در قسمتهای دیگر شبکه های معنایی در قالب ارائه و بازنمایی خرد به پرسشگر منتقل می شود.

برای جلوگیری از گستردگی بیش از حد با استفاده از انعطاف پذیر بودن شبکه های معنایی و قدرت آنها در بازنمایی روابط بین مفاهیم، زیاد بودن لایه ها را میتوان با دو یا چند ارتباط مشخص کرد و از مقدار لایه ها کاست:



شکل ۲. انعطاف پذیری در ساختار واژگان شناختی[۱۴]

معماری فوق در پیاده سازی ماشینی و کد کردن روابط بین مفاهیم، مخصوصاً به هنگام ازدیاد تعداد گره ها و لایه ها بسیار مؤثر بوده و از پیچیدگی و حجم کار می کاهد. در مورد عملکردها و ویژگی ها نیز همین موضوع صادق است.

۲-۲ برقرار کردن ارتباطات مفهومی بین گره ها

پس از ساخت پیکره اصلی، روابط بین مفهومی را می توان در سه سطح مشخص کرد:

۱- سطح سلسله مراتبی(طبقه ای) متناوب

۲- سطح سلسله مراتبی غیر متناوب

۳- سطح غیر سلسله مراتبی(غیر طبقه ای)

منظور از سطح طبقه ای متناوب سطحی است که در آن گره ها دقیقاً پشت سر هم قرار گرفته و یک سلسله مراتب شمول را تشکیل داده اند. سطح طبقه ای غیر متناوب شامل ارتباط دو یا چند گره است که در یک درخت مفهومی در شاخه های غیر متناوب، (شاخه هایی که در سطوح مختلف درخت قرار دارند و ارتباط آنها با کمان مشخص شده است) قرار گرفته و با یکدیگر ارتباط معنایی دارند. سطح غیر سلسله مراتبی که ساختار اصلی واژگان شناختی بر مبنای سازماندهی این نوع رابطه استوار شده است و بدنه اصلی شبکه معنایی را تشکیل می دهد عبارتست از روابط خاص و موجود میان دو حوزه کاملاً متفاوت که در شکل گیری ابعاد مختلف مفاهیم چند بعدی شرکت دارند، بگونه ای که نادیده گرفتن این ابعاد شناخته شده به یک جنبه نگری و استخراج دانش ناقص و دانش بد منتهی میشود. بر عکس کشف و استخراج این روابط موجب تفسیرها و تبیین های بخردانه درباره ابعاد مفاهیم مذکور بوده یا افق های جدیدی در این حوزه های مشترک خلق می کند و راه حلهای مناسبی را برای برخی مشکلات خاص ارائه میکنند. به این گونه روابط که از نظر استراتژیک اهمیت خاصی داشته و معمولاً از دید عموم پنهان و معمولاً به دانش ذهنی گروه خاصی انحصار یافته اند در واژگان شناختی "ارتباط خرد" گفته می شود. لازم به توضیح است در بیشتر موارد ارتباطات و مفهوم های خردگرا بسیار بیشتر از آنکه تصور می شود ساده و غیر پیچیده هستند اما خصوصیت بارز آنها دسترس ناپذیری ذهنی و همان سادگی و چند بعدی بودن است که کشف آنها ذهن را از درک واقعیت تکه ای به واقعیت کل آن پدیده نزدیکتر میکند.

۳- معماری سیستم پیشنهادی

در شکل (۳) معماری سیستم پیشنهادی نشان داده شده است. همانطور که در شکل پیداست سیستم از اجزای مختلف تشکیل شده است که در ادامه هر کدام بصورت جداگانه تشریح می شود.

۳-۱ مؤلفه پیش پردازش کلمات

در این بخش پرسش ورودی از نظر املائی مورد بررسی قرار می گیرد و در صورت لزوم اصلاحات مورد نیاز انجام می شود. برای این کار لازم است که ازپایگاه داده واژگان زبان انگلیسی استفاده کرد. بعد از انجام تغییرات احتمالی پرسش ورودی برای قسمت بعد فرستاده می شود.

۳-۲ مؤلفه آنالیز نحوی

آنالیز نحوی بر روی پرسش ورودی انجام می گیرد و درستی و یا نادرستی آن مشخص می شود. در اینجا از دانش گرامرهای محاسباتی استفاده می شود. برای پردازش نحوی پرسشها، از گرامرهای محاسباتی Context Free به همراه ویژگیها (features) استفاده شده است. برای مثال قسمتی از این نوع گرامر استفاده شده در سیستم پیشنهادی به زبان پرولوگ در زیر آمده است.

```
s(X):- append(A,B,X),np(A),vp(B),num_np(A,N),num_vp(B,N).
num_np(X,N):- append(A,B,X),num_pref(A,N),num_np1(B,N).
num_np(X,N):- num_noun(X,N).
num_np(X,N):- num_pron(X,N).
num_np(X,N):- append(A,B,X),num_np4(A,N),num_pps(B,N).
np(X):- append(A,B,X),pref(A),np1(B).
np(X):- noun(X).
np(X):- pron(X).
np(X):- append(A,B,X),np4(A),pps(B).
num_pps(X,N):- append(A,B,X),num_pp(A,N),num_pps1(B,N).
pps(X):- append(A,B,X),pp(A),pps1(B).
```

اما برای آنالیز نحوی درست و تطابق بین اجزای جمله، برای قواعد گرامرها ویژگیهایی نیز تعریف شده است که این ویژگیها شامل: مفرد و جمع بودن، اول شخص، دوم شخص و سوم شخص بودن، فعل متعدی و یا لازم بودن، مؤنث و مذکر بودن می باشد.

۳-۳ مؤلفه ساخت بازنمایی پرسش

بعد از اینکه در گام قبل اجزای پرسش ورودی از نظر نحوی شناسایی شدند، در این فاز مفاهیم موجود در پرسش ورودی و همچنین روابط و وابستگی مفهومی موجود بین این مفاهیم استخراج می شود. همچنین کلمات قابل حذف[۵] ، حروف اضافه و حروف تعریف بدلیل بی نقش بودنشان حذف می شوند . منظو ر از وابستگی دودویی نیز چیزی شبیه به گراف مفهومی می باشد. به عبارت دیگر گرافی ترسیم می شود که فعل جمله پرسشی، نقش محوری ایجاد می کند. برای مثال برای پرسش Q1 که در قسمتهای قبل مورد بررسی قرار گرفت می توان وابستگی های دودویی آنرا بصورت شکل ۴ نشان داد.

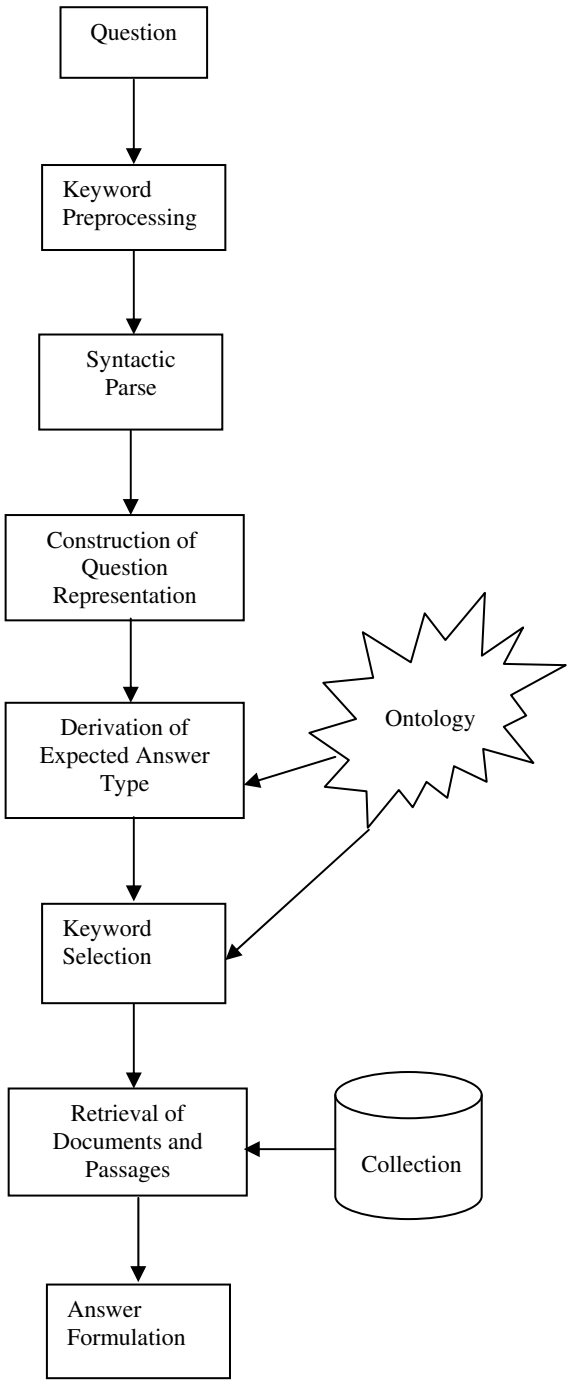
برای بدست آوردن گراف مفهومی، بعد از انجام آنالیز نحوی و استخراج درخت پارس، برگهایی از درخت پارس که اسم، فعلهای غیر کمکی، صفت و قید باشند انتخاب می شوند و برچسب use میگیرند. با پیمایش پایین به بالای درخت پارس، برچسب برگها به نودهای پدر منتشر می شود. بر حسب نقش نحوی نود پدر، برچسب یکی از نودهای فرزند را انتخاب می کند و منتشر می کند و نیز در گراف مفهومی مورد نظر نود فرزند برنده به نود فرزند دیگر متصل می شود. پیمایش پایین به بالا تا آنجا انجام می شود که نود ریشه نیز دارای برچسب شود[11] .

۳-۴ مؤلفه استخراج نوع پاسخ مورد انتظار

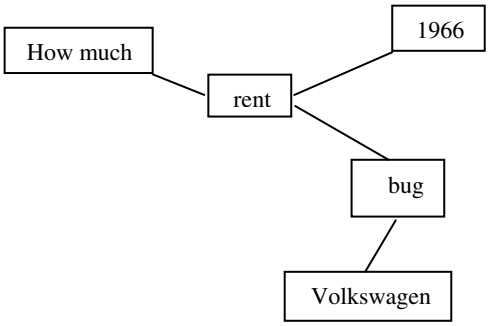
در این قسمت باید نوع پاسخی که برای پرسش داده شده مناسب است مشخص شود. برای اینکار ما از یک واژگان شناختی برای طبقه بندی کردن انواع پرسشها استفاده کرده ایم. کلمه پرسشی[۶] هر پرسش (مانند:...who, what, where) و همچنین اولین عبارات اسمی [۷] موجود در پرسش داده شده میتوانند کلاس پرسشها را مشخص کنند.

طبقه بندی سلسله مراتبی که انواع پرسشها را مشخص می کند به صورت یک گراف است که در این گراف هر نود متناظر با یک نوع پرسش و همچنین پاسخ مناسب با آن در نظر گرفته شده است. هر نود دارای پنج نوع دانش می باشد که عبارتند از: نمایش معنایی پرسش، نوع پرسش، نوع پاسخ، مرکز توجه پرسش[۸] و کلمات کلیدی پرسش ارائه شده. برای نمایش هر دانش یک از attribute که نوع دانش و یک value که مقادیر ممکن آن دانش را نشان می دهد، استفاده میشود. شکل۵ بعضی از نودهای بالای این سلسله مراتب را نشان می دهد. برای روشن شدن پنج دانشی که ذکر شد، برای نمونه نوع Currency که برای دلالت کردن بر پول و موجودی بکار می رود را در نظر بگیرید، یک بازنمایی معنایی که برای پرسش این نوع پاسخ مطرح می شود می تواند به شکل۶ باشد. همانطور که در این بازنمایی مشخص است، نوع پرسش و نوع پاسخ Currency می باشد. مرکز توجه این پرسش نیز از نوع پول و مقدار می باشد. برای مثال پرسشی مانند: How much did Manchester united spend on players in 1993? قابل نگاشت شدن بر روی بازنمایی شکل ۶ است. بدین ترتیب که Manchester United با Author of action، spend با value word، players با object of action و 1993 با timestamp دااری ارتباط مفهومی هستند[11].

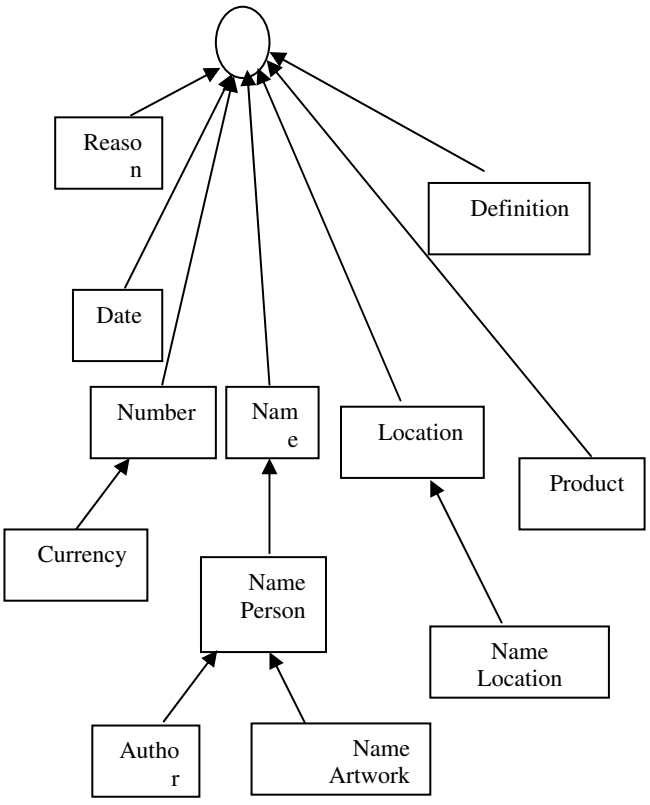
پس با نگاشت پرسش طرح شده بر روی بازنمایی نحوی شکل ۶، روشن است که پاسخ این پرسش باید نشان دهنده مقدار و یا پول باشد.



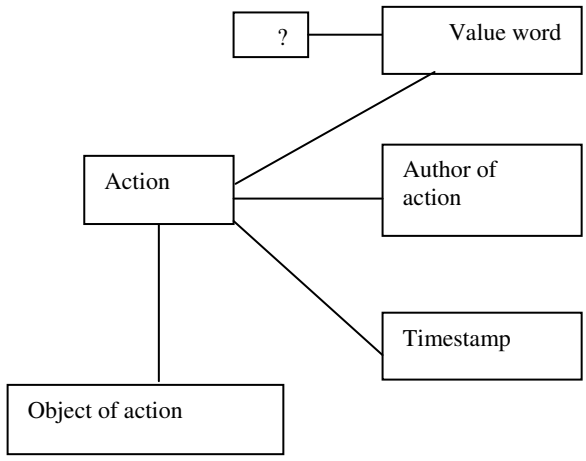
شکل ۳. معماری سیستم پیشنهادی



شکل ۴. نمایش وابستگی های دودویی پرسش Q1



شکل ۵. نمایش سطوح بالای کلاس بندی سلسله مراتبی



شکل ۶. بازنمایی نحوی پرسش ممکن درباره موجودی و پول

در اینجاست که شدیداً لزوم یک واژگان شناختی احساس می شود، به عبارت دیگر فهم این نکته که spend و value word دارای ارتباط مفهومی نزدیکی هستند باید از قبل مشخص شده باشد و آنچه که در این رابطه به ما کمک فراوانی خواهد کرد و راهگشاست طراحی یک واژگان شناختی مناسب می باشد.

۵-۳ مؤلفه انتخاب کلمات کلیدی

بعد از اینکه در مرحله قبل نوع پاسخ و مرکز توجه پرسش مورد نظر مشخص شد، در این مرحله برای جستجوی پاسخهای مناسب، لازم است که کلمات کلیدی انتخاب شوند که از آنها برای جستجو در مجموعه مدارک موجود استفاده شود. اولین کلمات کلیدی که مناسب به نظر می رسند، کلماتی هستند که در بازنمایی معنایی پرسش مطرح شده قرار دارند. برای نمونه برای پرسش داده شده در قسمت قبل، کلمات: Manchester United، spend، players،1993 در دید نخست انتخاب می شوند. اما ممکن است که پاسخی که ما به دنبال آن هستیم با این کلمات موجود نباشد و با کلمات دیگری درمجموعه مدارک موجود توصیف شده باشد، دوباره نیاز به یک واژگان شناختی احساس می شود. به عبارت دیگر برای فهم این موضوع که کلمات:spend، buy، rent و invest دارای مفاهیم نزدیک هستند و می توان از آنها نیز به عنوان کلمات کلیدی برای یافتن پاسخهای مناسب استفاده کرد، نیاز به داشتن واژگان شناختی مناسب را آشکار می سازد. پس علاوه بر کلمات کلیدی موجود در بازنمایی نحوی پرسش ورودی، کلماتی با مفاهیم نزدیک به آنها را که واژگان شناختی معرفی می کند نیز می توان به عنـوان کلمات کلیدی مرتبط برای استخراج پاسخهای مورد نظر بکار برد[6].

۶-۳ مؤلفه بازیابی مدارک و مجموعه جملات

حال بعد از انتخاب کلمات کلیدی مناسب، می توان برای یافتن مدارک و پاراگرافهایی که شامل این کلمات کلیدی باشند اقدام کرد. این فرآیند همان رویه ای است که در موتورهای جستجو بکار می رود. یعنی مدارکی که این کلمات در آنها موجودباشند استخراج شده و جملات و یا پاراگرافهایی که این کلمات با هم در آنجا ذکر شده باشند به عنوان پاسخ در نظر گرفته شده و به خروجی منتقل می-شود [8].

۴- بازنمایی دانش

لازم است که درباره نحوه بازنمایی دانش به جهت اهمیت آن در سیستم پیشنهادی توضیحاتی ارائه شود. همانطور که مطلع هستید روشهای بازنمایی متعددی چون: منطق (Logic)، فریم(Frame)، شبکه معنایی (Semantic Network)، گراف مفهومی(Conceptual Graph) ، وابستگی معنایی (Conceptual Dependency) در ارتباط با بازنمایی دانش وجود دارند. در سیستم پیشنهادی ما از منطق درجه اول برای بازنمایی دانش استفاده کرده ایم. فرض کنید که یک پرسش مانند: "the red ball is on the table" به سیستم ارائه شده است، برای تبدیل آن به فرم منطقی پس از آنالیز نحوی، می توان آنرا اینگونه نمایش داد: On(red(ball),table)، که()On و ()red به عنوان predicateهای دو و یک آرگومانی می باشند و اشیاء ball، table نیزبه عنوان آرگومانهای آنها محسوب می شوند. دلیل استفاده ما از منطق نیز برای بازنمایی نیز الف)- سادگی منطق، ب)- نزدیکی آن به زبانهای برنامه نویسی نظیر پرولوگ و ج)- زبانهای فرمال برای برقرای ارتباط بین سیستمهای هوشمند مختلف مانند KIF(Knowledge Interchange Format) نیز با منطق بازنمایی شده اند و بنابراین استفاده از آن دانشهای عمومی نیز در سیستم پیشنهادی ساده خواهد بود.

۵- نتایج آزمایشات انجام شده

برای ارزیابی نقش واژگان شناختی در سیستمهای پاسخگویی، دو حالت ارزیابی دستی و اتوماتیکی مبتنی بر سیستمهای کامپیوتری در نظر گرفته شده است.

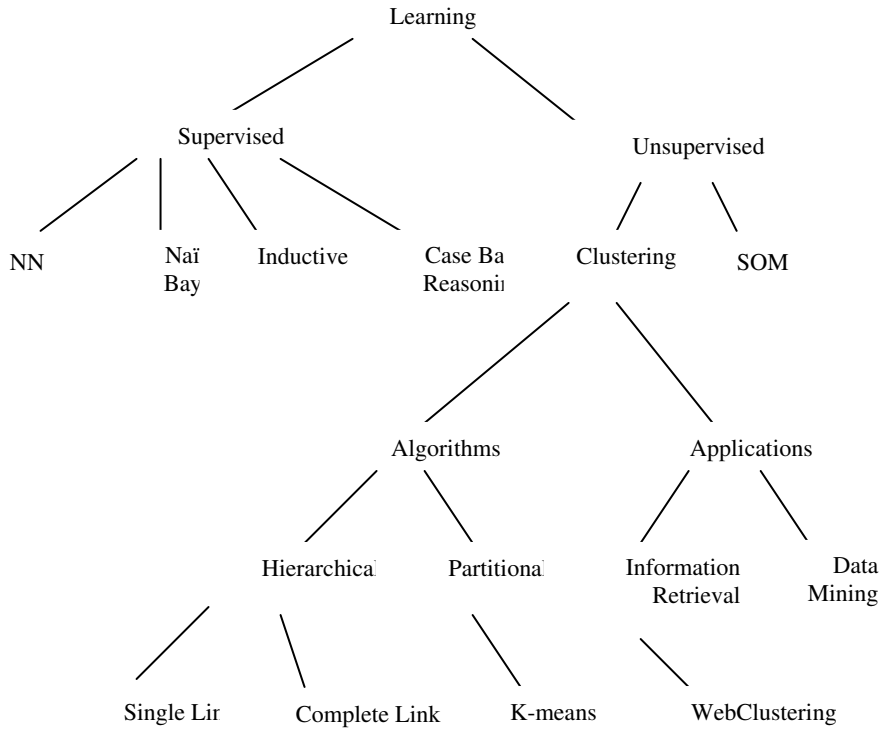
۵-۱ ارزیابی دستی

در این حالت، مدارکی با موضوعات: فلسفه و کتابداری، معرفت شناسی، کتابداری (اطلاعات، ذخیره و بازیابی، سازماندهی، سواد اطلاعاتی،...)، فلسفه و کامپیوتر (فلسفه ذهن، هوش مصنوعی، سیستمهای خبره)، دانش و شیوه های سازماندهی آن، روشهای بازنمایی دانش، هستی شناسی در فلسفه و ... انتخاب شد. بعد از نمایه سازی این مدارک، از بیست شرکت کننده (دانشجویان رشته کتابداری و کامپیوتر) خواسته شد که پرسشهای خود را در قالب کلید واژه یا پرسش به زبان طبیعی برای دستیابی به مدارک مورد نظرشان ارائه کنند. برای ارزیابی کارآیی واژگان شناختی، عملیات بازیابی مدارک دو بار انجام شد، یکبار از اصطلاحنامه و بار دیگر از واژگان شناختی استفاده شد که نتایج این آزمایش در جدول ۱ نشان داده شده است[۱۴]. در جدول ۱ از پارامترهای ارزیابی مختلفی برای استفاده از واژگان شناختی در پاسخگویی به منظور بازیآبی مدارک در یک سیستم کتابخانه بهره برده شده است. پارامتر انتقال راهبردهای استراتژیک، ازمفاهیم موجود در حوزه کتابداری است و قدرت یک منبع را در انتقال استراتژیک نشان می دهد.

همانطور که در جدول نشان داده شده، تفاوت واژگان و تزاروس در رفع ابهام قابل ملاحظه است، در ایجاد ابهام تفاوت فاحشی ندارند، در بازیابی مدارک مناسب نیز اختلاف چشمگیری ملاحظه می گردد، در ارتقاء دانش کاربر و همچنین اصلاح پرسش نیز، واژگان شناختی قابلیت بالایی ارائه می کند. پارامتر انتقال راهبردهای استراتژیک نیز بیانگر این است که استفاده از تزاروس مناسب نیست.

۵-۲ ارزیابی اتوماتیکی

در این حالت، سیستم کامپیوتری تشریح شده در بخشهای قبل پیاده سازی و برای آزمایش سیستم پیشنهادی، مجموعه مدارکی درباره خوشه بندی و کاربردهای آن در نظر گرفته و پرسشهایی از این مجموعه مورد نظر به عمل آمد، برای پاسخ دادن به پرسشها دو حالت مورد بررسی قرار گرفت، یکبار بدون استفاده از واژگان شناختی بار دیگر پاسخها با بکار بردن یک واژگان شناختی که برای همین پرسشها طراحی شده است استخراج شدند.



شکل ۷. نمایش سطوح بالای واژگان شناختی طراحی شده برای قلمرو یادگیری و خوشه بندی

جدول۱. کارآیی واژگان شناختی در بازیابی اطلاعات و رفع ابهام [۱۴]

	رفع ابهام	ایجاد ابهام	بازیابی مدارک مرتبط	بازیابی مدارک مناسب	ارتقاء دانش کاربر	ایجاد قابلیت اصلاح پرسش	انتقال راهبردهای استراتژیک
تزاروس	٪۱۰	٪۲۰/۵	٪۴۶/۵	٪۲۲	٪۳۷	٪۳۵	منفی
واژگان شناختی	٪۴۷/۲	٪۵	٪۵۷	٪۳۴/۵	٪۵۷	٪۵۷	مثبت

در شکل ۷ سطـوح بالای واژگان شنـاختی در نظـر گرفته شده بـــرای قلمـرو یادگیری و خوشه بندی نشان داده شده است. مجموعه مدارک استفاده شده برای ارزیابی سیستم پیشنهادی، متشکل از پنجاه وب سایت مرتبط با موضوع خوشه بندی می باشد.

برای ارزیابی سیستم پیشنهادی پرسشهای متفاوتی مطرح و به سیستم ارائه شد و در دو حالت بدون استفاده از واژگان شناختی و استفاده از واژگان شناختی، پاسخها توسط سیستم تولید شد که چند نمونه از آنها ذکر می شود.

پرسش ۱:

Question1:What is the application of clustering algorithms?

الف)- پاسخ بدون استفاده از واژگان شناختی:

Answer: Clustering algorithms can be applied to group like objects.

ب)- پاسخ با استفاده از واژگان شناختی:

Answer: 1)- The unsupervised adaptive clustering algorithm is used for data mining…

2)- Motivated by the benefits in organizing the documents in Web search engines, we consider the problem of automatic Web page classification. We employ the clustering techniques to make the classification process automatic.

پرسش ۲ :

Question1:How is the automatic grouping of similar objects performed in data mining?

الف)- پاسخ بدون استفاده از واژگان شناختی:

Answer: It is now clear that automated tools must be developed to group similar objects for application of data mining.

ب)- پاسخ با استفاده از واژگان شناختی:

Answer: Cluster analysis, or automatic classification, is a multivariate statistical technique that seeks to identify groups, or clusters, of similar objects in a multi-dimensional space.

برای اینکه سیستم پیشنهادی در پاسخ دهی به پرسشهای متعدد دارای عملکرد خوبی باشد، لازم است که واژگان شناختی در نظر گرفته شده مفاهیم حوزه های مختلف را پوشش دهد، به عبارت دیگر نودهای گراف واژگان شناختی می بایست افزایش یابد. در واقع ما در آزمایش سیستم پیشنهادی، شاهد این واقعیت بودیم که هر چه تعداد نودهای واژگان شناختی افزایش می یابد به همان میزان قدرت پاسخگویی سیستم نیز افزایش می یابد. برای نشان دادن این موضوع نموداری تهیه شده است که در آن میزان دقت پاسخگویی در مقابل افزایش تعداد نودهای واژگان شناختی ترسیم شده است. منظور از دقت پاسخگویی تعداد پاسخهای صحیحی است که انتظار داریم سیستم آن پاسخها را تولید کند. به عبارت دیگر ما با دانستن پاسخ صحیح هر پرسش ارائه شده از میان مجموعه مدارک موجود، کارایی سیستم را در تعداد پاسخهایی که مشابه با پاسخهای ما برای پرسشهای طرح شده ارائه میکند ارزیابی می کنیم.

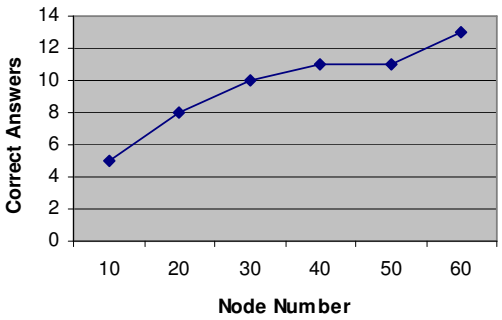
group similar objects for application of data mining.

ب)- پاسخ با استفاده از واژگان شناختی:

Answer: Cluster analysis, or automatic classification, is a multivariate statistical technique that seeks to identify groups, or clusters, of similar objects in a multi-dimensional space.

برای اینکه سیستم پیشنهادی در پاسخ دهی به پرسشهای متعدد دارای عملکرد خوبی باشد، لازم است که واژگان شناختی در نظر گرفته شده مفاهیم حوزه های مختلف را پوشش دهد، به عبارت دیگر نودهای گراف واژگان شناختی می بایست افزایش یابد. در واقع ما در آزمایش سیستم پیشنهادی، شاهد این واقعیت بودیم که هر چه تعداد نودهای واژگان شناختی افزایش می یابد به همان میزان قدرت پاسخگویی سیستم نیز افزایش می یابد. برای نشان دادن این موضوع نموداری تهیه شده است که در آن میزان دقت پاسخگویی در مقابل افزایش تعداد نودهای واژگان شناختی ترسیم شده است. منظور از دقت پاسخگویی تعداد پاسخهای صحیحی است که انتظار داریم سیستم آن پاسخها را تولید کند. به عبارت دیگر ما با دانستن پاسخ صحیح هر پرسش ارائه شده از میان مجموعه مدارک موجود، کارایی سیستم را در تعداد پاسخهایی که مشابه با پاسخهای ما برای پرسشهای طرح شده ارائه میکند ارزیابی می کنیم.

برای ارزیابی دقت سیستم پیشنهادی در مقابل افزایش نودهای واژگان شناختی، ۲۰ پرسش به سیستم پیشنهادی داده شده و برای این پرسشها واژگان شناختی با تعداد نودهای مختلف در نظر گرفته شده است و در هر حالت میزان پاسخهای صحیح سیستم پیشنهادی محاسبه شده است و در نهایت نمودار تعداد پاسخهای صحیح در مقابل افزایش تعداد نودها مانند شکل ۸ ترسیم شده است.



شکل ۸. نمودار تعداد پاسخهای مورد انتظاردر مقابل افزایش تعداد نودهای گراف واژگان شناختی

همانطور که در شکل پیداست با افزایش تعداد نودهای گراف واژگان شناختی، میزان دقت سیستم پیشنهادی در ارائه پاسخهای مورد انتظار افزایش خواهد یافت.

۶- نتیجه گیری

در این مقاله یک سیستم پاسخ دهی به پرسش زبان انگلیسی پیشنهاد شده است. با توجه به مشکلات ذکر شده سیستمهای دیگر که فقط مبتنی بر کلمات کلیدی هستند و اسنادی را که حاوی این کلمات هستند بر اساس روشهای آماری استخراج می کنند، استفاده از واژگان شناختی در سیستم پیشنهادی معرفی شده است.

در واقع مهمترین مشکل سیستمهای موجود در پردازش مفاهیم می باشد. واژگان شناختی که در رابطه با یک قلمروخاص، همه مفاهیم مناسب برای نمایش آن قلمرورا شناسایی و روابط واضح بین انها را نیز پوشش می دهد می تواند بر این مشکل غلبه کند. برای ارزیابی سیستم پاسخ دهی به پرسش پیشنهادی، مجموعه مدارک متشکل از ۵۰ وب سایت در مورد خوشه بندی مورد استفاده قرار گرفت. بیست پرسش از این سیستم به عمل آمد و پاسخهای آنها مورد بررسی قرار گرفت که حاکی از افزایش دقت بواسطه بکار بردن واژگان شناختی می باشد. همچنین از روی نمودار شکل ۸ مشخص شد که هر چه واژگان شناختی گسترده تر و به عبارت دیگر دارای نودهای زیادی باشد به همان میزان دقت سیستم پیشنهادی در ارائه پاسخهای مورد انتظار نیز افزایش خواهد یافت.

مراجع

[1] D. Moldovan, M. Pasca, S. Harabagiu, and M. Surdeanu, "Performance Issues and Error Analysis in an Open-Domain Question Answering system," *ACM Transactions on Information Systems,* Vol. 21, No. 2, Pages 133–154, 2003.

[2] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton, "Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering," *SIGIR'03*, 2003.

[3] N. F. Noy, "Semantic Integration: A Survey Of Ontology-Based Approaches", *SIGMOD Record*, Vol. 33, No. 4, 2004.

[4] Annika Flycht-Eriksson, "Design of Ontologies for Dialogue Interaction and Information Extraction," *In Proceedings of the 39th Annual Meeting of the Association*

for Computational Linguistics, *Workshop on Open-Domain Question Answering*, 2001.

[5] Jamie Callan, "Open Domain Question Answering," http:// www.cs.cmu.edu, 2004.

[6] Alessandro Moschitti, "Answer filtering via Text Categorization in Question Answering Systems," *In Proceedings ACM SIGIR 2001*, pp. 366–374, ACM Press, 2001.

[7] Matthew W. Bilotti, Boris Katz, and Jimmy Lin, "What Works Better for Question Answering: Stemming or Morphological Query Expansion?," *In Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*, 2004

[8] Hang Cui, Min-Yen Kan, Tat-Seng Chua, Jing Xiao, "A Comparative Study on Sentence Retrieval for Definitional Question Answering," *The Twelfth Text REtrieval Conference (TREC 2003) Notebook*, pp. 54-63, 2003.

[9] Ellen M. Voorhees, "Overview of the TREC 2003 Question Answering Track," *In TREC 2003,* 2003

[10] S. Narayanan and S. Harabagiu, "Question Answering based on semantic structures," *SIGMOD Record*, Vol. 33, No. 4, December 2004

[11] S. M. Harabagiu and M. A. Parca and S. J. Maiorano, "Experiments with Open-Domain Textual Question Answering," *In Proceedings of the 18th International Conference on Computational Linguistics,* pp. 292–298, , 2000.

[12] M. Vargas-Vera, E. Motta and J. Domingue, "AQUA: An Ontology-Driven Question Answering System " *AAAI Symposium on New Directions of Question Answering*, Stanford University, 2003.

[13] Remi Zajac, " Towards Ontological Question Answering," http://www.citeseer.ifi.unizh.ch/558911.html.

[۱۴] ف. دادرس، *طراحی و سنجش فرا واژگان شناختی، رویکردی جدید در بازنمایی دانش ومعرفت*، پایان نامه کارشناسی ارشد، دانشگاه آزاد اسلامی، دانشکده علوم انسانی، گروه کتابداری و اطلاع رسانی، پاییز ۱۳۸۲.

قربان خردمندیان در سال ۱۳۷۹ از دانشگاه اصفهان در رشته مهندسی کامپیوتر گرایش نرم افزار فارغ التحصیل شد و در سال ۱۳۸۱ نیز کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی را در دانشگاه صنعتی امیرکبیر به پایان رساند و در بهمن سال ۱۳۸۱ در آزمون دکترای مهندسی کامپیوتر دانشگاه صنعتی امیرکبیرپذیرفته شد و هم اکنون نیز مشغول به انجام پایان نامه می باشد.موضوع پایان نامه کارشناسی ارشد درباره استفاده از تکنیکهای بینایی کامپیوتربرای ردیابی خودروها به منظور تشخیص تخلفات رانندگی بوده است که به سرپرستی دکتر محمد رحمتی انجام شده است. پایان نامه وی در مقطع دکترا، روی موضوع: یادگیری تقویتی سلسله مراتبی با کشف اتوماتیک ساختار سلسله مراتبی مسئله، متمرکز شده است که به سرپرستی دکتر محمد رحمتی می باشد.

زمینه های تحقیقاتی مورد علاقه وی شامل: یادگیری تقویتی، داده کاوی، بازیابی اطلاعات، پردازش زبان طبیعی و عاملهای هوشمند می باشد. پانزده مقاله از وی در کنفرانسهای بین المللی به چاپ رسیده است و چندین مقاله نیز به مجلات معتبر خارجی ارسال شده است.

آدرس پست الکترونیک: kheradmandian@ce.aut.ac.ir

احمد عبدالله زاده مدرک کارشناسی خود را در سال ۱۳۵۴ در رشته حسابداری از دانشگاه تهران و مدرک کارشناسی ارشد را در سال ۱۳۵۹ در علوم کامپیوتر از دانشگاه West Coast University, Los Angeles کشور آمریکا دریافت نمود. ایشان تحصیلات دوره دکترا خود را با تحقیقات انجام شده در موضوع مرتبط با هوش مصنوعی و پایگاه داده هوشمند در سال ۱۹۹۰در دانشگاه بریستول کشور انگلستان به پایان رسانده است. دکتر عبدالله زاده فرصت مطالعاتی در سالهای ۲۰۰۲-۲۰۰۰ در دانشگاه Orsayپاریس و Maryland University at college parkبعنوان استاد مدعو گذرانده است. ایشان در حال حاضر بعنوان دانشیار دانشکده مهندسی کامپیوتر و فناوری اطلاعات مشغول امور آموزشی و پژوهشی میباشد و موضوعات مورد علاقه ایشان عبارت است از: بازیابی اطلاعات، تکنیکهای هوش مصنوعی، سیستم خبره، پردازش زبان طبیعی، سیستمهای تصمیم یار، بازنمایی دانش و مهندسی نرم افزار.

آدرس پست الکترونیک: ahmad@ce.aut.ac.ir