بسم الله الرحمن الرحيم

دانشگاه اصفهان
دانشکده فنی و مهندسی
گروه مهندسی کامپیوتر

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش نرم افزار

رفع ابهام معنای کلمات در حاشیه نویسی خودکار اسناد توسط ادغام روش یادگیری ماشین و روش مبتنی بر دانش

استاد راهنما:

دکتر کامران زمانی فر

پژوهشگر:

بهنام حاجیان

اردیبهشت ماه ۱۳۸۸

دانشگاه اصفهان
دانشکده مهندسی
گروه مهندسی کامپیوتر

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش نرم افزار

آقای بهنام حاجیان تحت عنوان

رفع ابهام معنای کلمات در حاشیه نویسی خودکار اسناد توسط ادغام روش یادگیری ماشین و روش مبتنی بر دانش

در تاریخ ۲۲/۲/۸۸ توسط هیئت داوران زیر بررسی و با درجه عالی به تصویب نهایی رسید.

۱- استاد راهنمای پایان نامه: دکتر کامران زمانی فر با مرتبه علمی استادیاری امضا.....

۲- استاد داور داخل گروه: دکتراحمد براآنی با مرتبه علمی استادیاری امضا.....

۳- استاد داور خارج از گروه: دکتر مازیار پالهنگ با مرتبه علمی استادیاری امضا.....

امضا مدیر گروه

۸۸/۲/۲۹

بسم الله الرحمن الرحیم

سپاسگزاری

سپاس خدای یکتا را که هر چه هست از اوست.

سپاس او را که بی نام و یادش نمی‌توان آغاز کرد و نمی‌توان به پایان برد.

سپاس و حمد بیکران او را که به ما طریق عاشقی آموخت.

سپاس ترا ای یگانهٔ عالم، ای حقیقت مطلق، ای منتهای تحقیق، ای همه هستی در یک کلام، ای خدا

خدایا مخواه که غرور در دل ما ریشه کند

خدایا مخواه که غیر از آنچه تو خواسته ای بخواهیم و غیر از آنچه رضای توست بطلبیم

خدایا ما را آنی به خود وامگذار که به خود نیامده ایم که به خود رویم

خدایا تو میدانی که جز تو از تو نمی‌خواهیم

تنها تو را می‌پرستیم و تنها از تو یاری می‌طلبیم

ما را در این راه تنها مگذار

به مصداق کلام شریف من لم یشکر المخلوق لم یشکر الخالق، بر خود فرض میدانم تا از تلاش تمامی راهنمایان خود در این پروژه بویژه جناب آقای دکتر زمانی فر تشکرکنم و از پروردگار یکتا توفیق روز افزون ایشان را در خدمت به ایران اسلامی مسئلت دارم.

از سروران و عزیزانی که با رهنمودهای خود ما را در یافتن کاستی ها و رفع آنها یاری خواهند داد سپاسگزارم و امیدوارم که شایستگی پذیرش نظرات و پیشنهادات این عزیزان را داشته باشم.

من الله التوفیق

بهنام حاجیان

تقدیم به:

پدر و مادرم که در تمام مراحل زندگی مرا
راهنمایی و پشتیبانی کرده اند .

چکیده

با ظهور وب معنایی به عرصه اینترنت، محیطی قابل درک و استنتاج برای محاسبات کامپیوتری فراهم گردید. در این زمینه اولین گام جهت دستیابی به محیط هوشمند و قابل فهم برای ماشین، ایجاد هستی شناسی مناسب در دامنه مورد بحث و اتصال اطلاعات موجود در صفحات وب به نمونه ها و کلاس های موجود در هستی شناسی می باشد. این عمل که حاشیه نویسی معنایی اسناد نام دارد، کلیه کلمات ، اشیاء و اطلاعات موجود در صفحات وب را به مفاهیم متناظر آنها در هستی شناسی مرتبط می سازد لذا دامنه معنایی هر کلمه با توجه به کلاس مربوطه مشخص می گردد. بدین جهت بسیاری از مشکلات مطرح شده در زمینه بازیابی و جستجوی اطلاعات حل شده و اطلاعات موجود در اسناد به مفاهیمی قابل فهم برای کامپیوتر ها تبدیل می شوند. در این زمینه خصوصا با توجه به وجود کلمات هم شکل و با معانی متفاوت، با محیطی مبهم روبرو هستیم. لذا معنای کلمات می بایست با توجه به زمینه و مفهوم جمله ای که در آن بکار رفته است تعیین و رفع ابهام گردد. عمل حاشیه نویسی معنایی اسناد در حال حاضر به طور دستی ویا نیمه خودکار توسط عامل انسانی انجام می گیرد که عملی وقتگیر، پیچیده ، پرهزینه و نادقیق می باشد.

یکی از مسائل زیر مجموعه حاشیه نویسی خودکار اسناد رفع ابهام معنی کلمات در متن می باشد. در این پروژه سعی شده با ارائه راهکاری بر مبنای هوش ماشینی و بکارگیری الگوریتم های یادگیری ماشینی با ناظر به خودکار نمودن این عمل وقتگیر توسط کامپیوتر پرداخته شود. سیستم فوق دارای دو فاز آموزش و عملیاتی می باشد. در مرحله آموزش سیستم توسط داده های آموزشی (صفحات حاشیه نویسی شده بصورت دستی) آموزش دیده و سپس در فاز عملیاتی به حاشیه نویسی اسناد جدید بصورت خودکار می پردازد. در این پایان نامه از چند الگوریتم مختلف آموزش ماشین استفاده شده است و نتایج عملکرد آنها ارزیابی و با یکدیگر مقایسه شده است. در این ارتباط الگوریتم ماشین بردار پشتیبان بهترین کارایی را از خود نشان داده است. در مرحله بعد جهت افزایش دقت و کارایی این الگوریتم ها به ارائه چند مرحله پیش پردازش پرداخته شده است. نتایج ارزیابی در این پایان نامه نشان خواهد داد که با ادغام روش های بر مبنای یادگیری ماشینی و سیستم های مبتنی بر دانش می توان به بهترین کارایی دست پیدا نمود. همچنین استفاده از روش انتخاب ویژگی ها در این مسئله از دقت سیستم می کاهد. در صورتی که بوسیله گسترش ویژگی های متن، دقت و کارایی سیستم افزایش خواهد یافت.

کلید واژه ها:

رفع ابهام معنایی ، وب معنایی ، حاشیه نویسی وب ، یادگیری ماشینی ، هستی شناسی ، متن کاوی ، گسترش ویژگی ها

فهرست مطالب

عنوان **صفحه**

عنوان **صفحه**

عنوان **.صفحه**

فهرست شکل ها

عنوان **صفحه**

فهرست جدول ها

عنوان **صفحه**

فهرست کوتاه نوشت ها

XML	Extensible Markup Language
RDF	Resource Description Framework
OWL	Ontology Web Langiage
SVM	Support Vector Machine
SWAM	Semantic Wen Annotation by Machine Learning
WSD	Word Sense Disambiguation
ACC	Association Concept Classifier
UML	Unified Markup Language
HTML	Hyper Text Markup Language
FOL	First Order Logic
W۳C	Word Wide Web Consortium
DAML	Darpa Agent Markup Language
B۲B	Business to Business
B۲C	Business to Commerce
API	Application Programming Interface
IDE	Integrated Development Environment
POS	Ports of Speech
VSM	Vector Space Model

فصل ۱

معرفی

امروزه با توجه به گسترش روز افزون هوش مصنوعی در علم کامپیوتر ، بسیاری از عملیات دستی جای خود را به نرم افزار های هوشمند کامپیوتری داده است. بطور کلی ماهیت وجودی هوش به مفهوم جمع آوری اطلاعات، استقراء و تحلیل تجربیات به منظور رسیدن به دانش و یا ارائه تصمیم می باشد. هوش مصنوعی، علم و مهندسی ماشین های هوشمند با بکارگیری از کامپیوتر و الگوگیری از درک هوش انسانی و نهایتا دستیابی به مکانیزم هوشمند می باشد[۱] . در سالهای اخیر این هوشمندی کاملا در عامل های نرم افزاری خود مختار و هوشمند نمایان شده است بطوریکه این عامل ها در بسیاری از موارد جای انسان ها را گرفته و وظایف آنها را انجام می دهند. به عنوان مثال این ابزار می تواند به عنوان یک کارپرداز از کاربران دستور گرفته و پس از جستجو در سایت ها اقدام به و رزرو و خرید بلیط هوا پیما با قیمت مناسب بکند و یا در فعالیت های تجاری مثل یک عامل انسانی اقدام به خرید و فروش و یا معامله نماید. جهت دستیابی به چنین دنیای هوشمندی، نیاز به زیر ساخت هایی جهت تبادل اطلاعات، ابزار های جستجو و دانش زمینه جهت استنتاج می باشد.

۱-۱ مقدمه

امروزه وب به عنوان یکی از مهمترین بسترهای جمع آوری اطلاعات برای عامل های هوشمند معرفی می شود. لذا یکی از دغدغه های دنیای کامپیوتر، ایجاد بستری مناسب در وب جهت ایجاد امکان تعامل بین عامل های هوشمند و دنیای وب می باشد. متاسفانه اطلاعات موجود در وب متداول، قابل پردازش و درک برای این نرم افزار های هوشمند نیست. این اطلاعات کاملا بصورت غیر ساخت یافته در دنیای اینترنت گسترش یافته اند. اغلب این اطلاعات بصورت متنی بوده و به زبان طبیعی، در دنیای وب در دسترس کاربران قرار گرفته اند. ولی زبان های طبیعی غیر قابل فهم و پردازش برای نرم افزار های کامپیوتری می باشند. لذا عامل های هوشمند توانایی تعامل با وب کنونی را نخواهند داشت[۳] ,[۲] . از سوی دیگر ، با توجه به حجم بسیار زیاد اطلاعات موجود در وب مشکلات عمده ای در زمینه بازیابی و جستجوی اطلاعات مورد نظر کاربران پدیدار شده است. در این حیطه موتور های جستجو نیز قادر به ارائه نتایج دقیق و ارزشمندی برای کاربران نمی باشند. موتور های جستجوی متداول تنها قادر به جستجوی اسناد با توجه به وجود یا عدم وجود کلید واژه های موجود در پرس وجوی وارد شده توسط کاربر بوده و در کنار آن تنها می توانند اسناد جدید را که توسط خزشگر[۱] یافت می شوند به پایگاه داده شان افزوده ویا صفحات قبلی را به روز رسانی نمایند. وجود حجم وسیع اطلاعات در وب و ازدیاد صفحات وب[۲] و فقدان عقل سلیم و عدم توانایی موتور های جستجو از تحلیل معنایی صفحات باعث گردیده که نتایج جستجو در این ابزار برای کاربران چندان دقیق و قابل قبول نباشد و امروز شاهد وجود بسیاری از صفحات نا مرتبط با پرس و جوی کاربر در نتایج این ابزار می باشیم[٤] ,[۲] .

با توجه به وجود نا رسایی هایی در وب متداول، پس از ارائه مقاله "تیم برنرزلی" در سال ۲۰۰۱ میلادی، وب معنایی پا به عرصه ی دنیای اینترنت گذاشت [٥]. این محقق و مخترع وب، اهداف و دید دنیای وب را فراتر از صورت متداول آن در دنیای امروز می داند بطوریکه اطلاعات موجود در وب معنایی را قابل پردازش و تحلیل توسط ماشین می خواند[٦] . با بوجود آمدن زبان XML قادر به نوعی طبقه بندی ساختاری در این اطلاعات شده ولی هنوز این اطلاعات قابل درک و استنتاج توسط کامپیوتر ها نبوده و هنوز خلاء معنایی در ساختار اطلاعات موجود در وب احساس می شد. وب معنایی این خلا را توسط حاشیه نویسی اسناد وب و توسعه وب جاری پر

---

[۱] Crawler

[۲] ۳۰ میلیارد صفحه در فوریه سال ۲۰۰۷ طبق آمار نت کرافت

کرده که در آن اطلاعات معنایی بصورت خوش تعریف در قسمتی از پایگاه دانش[۳] کامپیوتر که هستی شناسی[٤] نام دارد تعریف میگردد و قابلیت استنتاج از اطلاعات موجود در وب را به عاملهای هوشمند اهدا نموده است. همراه با این پیشرفت، عامل های هوشمند می توانند با بهره گیری از منطق و استنتاج از قوانین و حقایق تعریف شده در پایگاه دانش و اطلاعات موجود در هستی شناسی و اتصال اطلاعات موجود در وب به مفاهیم تعریف شده درآن اقدام به استنتاج و جوابگویی مناسب به درخواست ها و پرسش های کاربر نمایند[۳] .

همچنین توسط عمل حاشیه نویسی بسیاری از مشکلات موجود در موتور های جستجو با توجه به تعیین و وضوح معنای دقیق برای کلمات حل می گردد. زیرا در وب معنایی، موتور های جستجو می توانند به معنای کلمات توسط فرا داده هایی که معنای کلمات را با توجه به کلاس هستی شناسی دامنه مورد بحث معرفی می کنند پی ببرند و اقدام به حذف صفحات نا مرتبط که شامل کلید واژه های پرس و جو با معنایی غیر از معنای مورد نظر کاربر است نمایند. در نتیجه قادر به نمایش نتایج دقیق تری به کاربران می باشند.

عمل حاشیه نویسی و ساخت هستی شناسی در حال حاضر به صورت دستی و یا ابزار نیمه خودکار انجام می گیرد که عملی وقتگیر و پر هزینه می باشد. این مشکل برای وب سایت های بسیار بزرگ مثل وب سایت های خبری بصورت دستی غیر قابل انجام بوده و نیاز به ابزار هایی جهت خودکار نمودن تولید این فرا داده ها احساس می گردد.

۱-۲ تعریف مسئله و اهداف پایان نامه

با توجه به مشکلات مطرح شده در بالا و پیشنهاد حل آنها توسط وب معنایی، حاشیه نویسی معنایی اسناد وب یکی از مهمترین دغدغه های مطرح شده در ایجاد این ساختار در وب می باشد. حاشیه نویسی معنایی اسناد به معنی اضافه نمودن فرا داده هایی در وب به منظور تعیین معنای اطلاعات موجود در وب و افزودن مفاهیم قابل پردازش و فهم توسط ماشین می باشد [۷]. همانگونه که در بالا ذکر شد این عمل در حال حاضر در محدوده کوچکی از منابع توسط عامل انسانی انجام می گیرد. مسئله تبدیل وب سایتهای جاری به وب سایت های معنایی تنها توسط عمل حاشیه نویسی قابل انجام است که می بایست توسط طراحان و تولید کنندگان وب سایت ها انجام گیرد. مسئله ساخت دستی نمونه ها در هستی شناسی و حاشیه نویسی صفحات نیاز به تخصص در مهندسی هستی

---

[۳] Knowledge Base
[٤] Ontology

شناسی و چگونگی ساخت و نگهداری آن دارد. از طرف دیگر انجام حاشیه نویسی بصورت دستی دقیق نمی باشد و نیاز به اتلاف زمان و هزینه بسیار دارد [۸]. یکی از مسائل مهم در مسئله حاشیه نویسی معنایی اسناد، ارائه روشی با دقت کافی جهت رفع ابهام معنای کلمات هم شکل با معنی متفاوت[۵] است که هدف اصلی این پایان نامه را تشکیل می دهد.

۱-۲-۱ هدف پایان نامه

با توجه به دقت مورد نیاز در رفع ابهام کلمات[۶] موجود در متن، هدف اصلی در این پایان نامه یافتن روشی کارا و با دقت کافی جهت تشخیص رفع ابهام معنای کلمات در جمله ای است که در آن ظاهر شده اند. به عبارت دیگر در این پایان نامه پس از ارائه معماری مناسبی برای حاشیه نویسی اسناد بر مبنای یادگیری ماشین به ارائه راهکار هایی جهت بهبود الگوریتم های رده بندی بوسیله روش های مبتنی بر دانش جهت حل مسئله ی رفع ابهام معنی کلمات پرداخته شده است.

۱-۳ روش انجام پروژه

در این پایان نامه پس از بررسی کار های انجام شده در گذشته بر روی مسئله فوق ، سیستمی ترکیبی بر اساس مدل یادگیری ماشین[۷] ارائه شده تا بتواند اطلاعات متنی موجود در صفحات وب را با توجه به معنای صحیح برآمده از کلمه در هر جمله به مفهوم صحیح و کلاس معنایی مقتضی مربوطه در هستی شناسی اختصاص دهد. به عبارت دیگر سیستم فوق قادر است معنای صحیح کلمات استفاده شده در متن را بوسیله آموزش از صفحاتی که قبلا بصورت دستی حاشیه نویسی شده اند تشخیص داده و به کلاس صحیح موجود در هستی شناسی انتساب داده و حاشیه نویسی کند. سپس با ترکیب روش یادگیری ماشینی و روش مبتنی بر دانش[۸] کارایی و دقت سیستم تا حد قابل قبولی بالا برده شده است.

---

[۵] Polysemy

[۶] Word Sense Disambiguation

[۷] Machine Learning

[۸] Knowledge Based

گفتنی است سیستم فوق به ایجاد نمونه ها در هستی شناسی نمی پردازد، بلکه سیستم فوق جهت تطابق نمونه های موجوددر هستی شناسی شبکه واژگان[۹] با کلمات موجود در متن ساخته و ارزیابی شده است، به گونه ای که بتواند در معنی کلمات هم شکل رفع ابهام نموده و آنها را حاشیه نویسی نماید.

در این سیستم دو دسته الگوریتم جهت یادگیری ماشین بررسی شده است. در روش اول بوسیله روش استخراج قوانین پیوند که یکی از روش های داده کاوی است الگوریتمی ابتکاری جهت یادگیری سیستم طراحی شده است. این الگوریتم[۱۰] به گونه ای تغییر داده شده تا بتواند توسط داده های متنی به استخراج قوانین پیوند میان کلمات جهت آموزش سیستم بپردازد. در روش دوم مسئله انتخاب کلاس مناسب معنایی برای یک کلمه را به مسئله رده بندی[۱۱] تبدیل نموده و بوسیله ابزار وکا[۱۲] توسط ۱۰ الگوریتم از روش های آموزش با نظارت[۱۳] یادگیری ماشین تست و ارزیابی شده است. در هر دو روش سیستم بوسیله داده های آموزشی حاشیه نویسی شده استاندارد مجموعه داده سنس ایوال[۱۴] مورد تست و ارزیابی قرار گرفته و نتایج الگوریتم های مختلف مورد مقایسه قرار گرفته است. سپس راهکار های پیشنهادی برای بهبود الگوریتم های رده بندی برای رفع ابهام معنی کلمات بوسیله همین داده ها تست و ارزیابی شده و میزان بهبودی الگوریتم ها مشخص گردیده است.

۱-٤ ساختار پایان نامه

این پایان نامه دارای شش فصل به شرح زیر می باشد. پس از فصل اول، در فصل دوم به معرفی وب معنایی و فناوری های مربوط به آن مثل هستی شناسی و استنتاج از وقایع هستی و جایگاه منطق در وب معنایی پرداخته و سپس مفاهیم کلی آموزش ماشینی و متن کاوی را شرح داده ایم. در فصل سوم به کار های انجام شده قبلی در زمینه حاشیه نویسی معنایی، استخراج مفاهیم از متن پرداخته شده است. در فصل چهارم به شرح معماری و طراحی سیستم پیشنهادی پرداخته شده و فصل پنج به تست و ارزیابی و نتایج تجربی حاصل از سیستم فوق اختصاص داده

---

[۹] WordNet

[۱۰] Apriori

[۱۱] Classification

[۱۲] WEKA

[۱۳] Supervised Learning

[۱۴] SensEval-۲

شده است. در پایان در فصل ششم به نتیجه گیری و معرفی راهکار های آینده جهت بهبود این فناوری پرداخته شده است.