

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



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

کتابخانه مرکزی دانشگاه اصفهان
تیم مرکز

۱۳۸۸ / ۴ / ۶

۱۱۴۹۳۳

کلیه حقوق مادی مرتب بر نتایج مطالعات، ابتکارات
و نوآوری های ناشی از تحقیق موضوع این پایان نامه
متعلق به دانشگاه اصفهان است.

شبهه نگارش پایان نامه
رعایت شده است
تخصصیات تکمیلی دانشگاه اصفهان



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 ۱۳۹۴/۰۲/۰۸ توسط هیئت داوران زیر بررسی و با درجه ... به تصویب نهایی رسید.

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

۲- استاد داور داخل گروه: دکتر احمد برآنی با مرتبه علمی استادیاری امضا.....

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

امضاء مدیر گروه
۱۳۹۴/۰۲/۲۹

بسم الله الرحمن الرحيم

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 یگانه عالم، ای حقیقت مطلق، ای منتهای تحقیق، ای همه هستی در یک کلام، ای خدا

خدایا میخواه که غرور در دل ما ریشه کند

خدایا میخواه 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 تشکر کنم 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 کاستی‌ها 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 هستی‌شناسی مناسب در دامنه مورد بحث و اتصال اطلاعات موجود در صفحات وب به نمونه‌ها و کلاس‌های موجود در هستی‌شناسی می‌باشد. این عمل که حاشیه‌نویسی معنایی اسناد نام دارد، کلیه کلمات، اشیاء و اطلاعات موجود در صفحات وب را به مفاهیم متناظر آنها در هستی‌شناسی مرتبط می‌سازد لذا دامنه معنایی هر کلمه با توجه به کلاس مربوطه مشخص می‌گردد. 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 کامپیوترها تبدیل می‌شوند. در این زمینه خصوصاً با توجه به وجود کلمات هم‌شکل و با معانی متفاوت، با محیطی مبهم روبرو هستیم. لذا معنای کلمات می‌بایست با توجه به زمینه و مفهوم جمله‌ای که در آن بکار رفته است تعیین و رفع ابهام گردد. عمل حاشیه‌نویسی معنایی اسناد در حال حاضر به طور دستی و یا نیمه خودکار توسط عامل انسانی انجام می‌گیرد که عملی وقتگیر، پیچیده، پرهزینه و نادقیق می‌باشد.

یکی از مسائل زیر مجموعه حاشیه‌نویسی خودکار اسناد رفع ابهام معنی کلمات در متن می‌باشد. در این پروژه سعی شده با ارائه راهکاری بر مبنای هوش ماشینی و بکارگیری الگوریتم‌های یادگیری ماشینی با ناظر به خودکار نمودن این عمل وقتگیر توسط کامپیوتر پرداخته شود. سیستم فوق دارای دو فاز آموزش و عملیاتی می‌باشد. در مرحله آموزش سیستم توسط داده‌های آموزشی (صفحات حاشیه‌نویسی شده بصورت دستی) آموزش دیده و سپس در فاز عملیاتی به حاشیه‌نویسی اسناد جدید بصورت خودکار می‌پردازد. در این پایان‌نامه 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 الگوریتم‌ها به ارائه چند مرحله پیش‌پردازش پرداخته شده است. نتایج ارزیابی در این پایان‌نامه نشان خواهد داد که با ادغام روش‌های بر مبنای یادگیری ماشینی و سیستم‌های مبتنی بر دانش می‌توان به بهترین کارایی دست پیدا نمود. همچنین استفاده از روش انتخاب ویژگی‌ها در این مسئله از دقت سیستم می‌کاهد. در صورتی که بوسیله گسترش ویژگی‌های متن، دقت و کارایی سیستم افزایش خواهد یافت.

کلید واژه‌ها:

رفع ابهام معنایی، وب معنایی، حاشیه‌نویسی وب، یادگیری ماشینی، هستی‌شناسی، متن کاوی، گسترش ویژگی‌ها

فهرست مطالب

صفحه	عنوان
۱	فصل ۱: معرفی
۲	۱-۱ مقدمه
۳	۲-۱ تعریف مسئله و اهداف پایان نامه
۴	۱-۲-۱ هدف پایان نامه
۴	۳-۱ روش انجام پروژه
۵	۴-۱ ساختار پایان نامه
۷	فصل ۲: مفاهیم وب معنایی و یادگیری ماشینی
۷	۱-۲ مقدمه
۹	۲-۲ تعریف وب معنایی
۱۳	۳-۲ پیش نیازهای وب معنایی
۱۳	۱-۳-۲ هستی شناسی
۱۵	۱-۳-۲-۱ انواع هستی شناسی
۱۶	۲-۳-۲ کاربردهای هستی شناسی
۱۷	۳-۳-۲ زبانهای هستی شناسی
۱۸	۱-۳-۲-۱ زبان XML
۱۹	۲-۳-۱-۳-۲ RDF: چهارچوب توصیفی منابع
۲۰	۳-۳-۱-۳-۲ OWL: زبان هستی شناسی وب
۲۱	۴-۳-۲ نحوه ایجاد هستی شناسی و مهندسی هستی شناسی
۲۲	۲-۳-۲ فرا داده در وب و حاشیه نویسی وب معنایی
۲۴	۳-۳-۲ جایگاه منطق در وب معنایی
۲۵	۴-۲ عاملها
۲۶	۵-۲ مسائل قابل حل توسط وب معنایی
۲۶	۱-۵-۲ مدیریت دانش
۲۸	۲-۵-۲ تجارت الکترونیک
۲۹	۳-۵-۲ سرویسهای جستجو

عنوان	صفحه
۶-۲ مشکلات مطرح موجود در وب معنایی.....	۳۲
۷-۲ مفاهیم داده کاوی و یادگیری ماشینی.....	۳۳
۱-۷-۲ مقدمه ای بر داده کاوی.....	۳۳
۲-۷-۲ مدل‌های داده کاوی.....	۳۵
۱-۲-۷-۲ رده بندی.....	۳۵
۲-۲-۷-۲ خوشه بندی.....	۳۵
۳-۲-۷-۲ تحلیل پیوند (تحلیل وابستگی داده ها).....	۳۶
۳-۷-۲ تعریف یادگیری ماشینی.....	۳۶
۴-۷-۲ معرفی برخی از الگوریتم‌های یادگیری ماشین و داده کاوی.....	۳۷
۱-۴-۷-۲ شبکه های عصبی.....	۳۷
۲-۴-۷-۲ درخت‌های تصمیم.....	۳۸
۳-۴-۷-۲ استنتاج قوانین.....	۳۸
K۴-۴-۷-۲ نزدیکترین همسایه.....	۳۸
۵-۴-۷-۲ الگوریتم کلاس بند نایویز.....	۳۹
۶-۴-۷-۲ الگوریتم ماشین بردار پشتیبان.....	۴۰
۵-۷-۲ انتخاب ویژگی ها در فرایند یادگیری ماشینی.....	۴۱
۸-۲ جمع بندی.....	۴۲
فصل ۳: کارهای انجام شده پیشین در زمینه حاشیه نویسی معنایی و رفع ابهام معنی کلمات	
۱-۳ مقدمه.....	۴۳
۲-۳ تقسیم بندی کارهای انجام شده در زمینه حاشیه نویسی معنایی وب و یادگیری هستی شناسی...۴۴	۴۴
۳-۳ دسته بندی سکوها و ابزارهای حاشیه نویسی.....	۴۶
۱-۳-۳ ابزارهای حاشیه نویسی دستی.....	۴۶
۲-۳-۳ ابزارهای حاشیه نویسی خودکار و نیمه خودکار.....	۴۸
۱-۲-۳-۳ روش‌های حاشیه نویسی مبتنی بر الگو.....	۴۸
۲-۲-۳-۳ روش‌های حاشیه نویسی خودکار مبتنی بر یادگیری ماشینی.....	۴۹
۴-۳ نگاهی به سکوها و سیستم های حاشیه نویسی معنایی.....	۴۹
۱-۴-۳ سیستم Aero DAML.....	۵۰
۲-۴-۳ سکوی آرمادیلو.....	۵۱

صفحه	عنوان
۵۱	۳-۴-۳ سکوی حاشیه نویسی مدیریت دانش و اطلاعات (KIM).....
۵۲	۴-۴-۳ سیستم ام ان ام.....
۵۳	۵-۴-۳ ابزار موسه.....
۵۴	۶-۴-۳ سیستم آنتومت.....
۵۴	۷-۴-۳ سیستم سمتگ و سکوی سیکر.....
۵۵	۸-۴-۳ سیستم آمیلکیر.....
۵۶	۹-۴-۳ سیستم اسکریم.....
۵۶	۱۰-۴-۳ سیستم مجی پای.....
۵۶	۱۱-۴-۳ سیستم حاشیه نویسی سی پنکو.....
۵۹	۵-۳ دسته بندی کارهای انجام شده در زمینه رفع ابهام معنای کلمات.....
۶۰	۶-۳ جمع بندی.....
۶۱	فصل ۴: طراحی سیستم پیشنهادی جهت رفع ابهام معنای کلمات موجود در متن
۶۱	۴-۱ مقدمه.....
۶۴	۲-۴ تعریف مسئله.....
۶۵	۳-۴ پیش نیاز های ساخت سیستم حاشیه نویسی خودکار وب.....
۶۵	۱-۳-۴ استاندارد های حاشیه نویسی و زبان مناسب بیان فرا داده ها.....
۶۶	۲-۳-۴ هستی شناسی مرتبط با دامنه مورد بحث.....
۶۹	۳-۳-۴ داده های آموزشی و ابزار ذخیره سازی و بازیابی اطلاعات.....
۶۹	۴-۴ معماری سیستم حاشیه نویسی معنایی خودکار سوام.....
۷۰	۱-۴-۴ شرح معماری سیستم سوام.....
۷۳	۲-۴-۴ مراحل پیش پردازش.....
۷۳	۱-۲-۴-۴ جداسازی و شماره گذاری کلمات و جملات موجود در متن.....
۷۳	۲-۲-۴-۴ حذف کلمات اضافه و اعداد از متن.....
۷۴	۳-۲-۴-۴ پیش پردازش کلمات مترادف با کمک پایگاه دانش ووردنت (گسترش ویژگی ها).....
۷۴	۴-۲-۴-۴ ریشه یابی کلمات موجود در متن.....
۷۴	۵-۲-۴-۴ پیش پردازش و ساخت بردار فضای حالت سند.....
۷۵	۶-۲-۴-۴ اعمال صافی عکس بسامد تکرار کلمه در متن ، نرمال سازی و انتخاب ویژگی ها.....
۷۶	۳-۴-۴ مرحله آموزش.....

عنوان	صفحه
مرحله حاشیه نویسی معنایی اسناد.....	۴-۴-۴ ۷۶
جمع بندی فصل.....	۴-۵ ۷۸
فصل ۵: پیاده سازی و ارزیابی نتایج سیستم سوام	
۱-۵ مقدمه.....	۷۹
۲-۵ پیاده سازی سیستم سوام.....	۸۰
۱-۲-۵ مولفه ذخیره و بازیابی اطلاعات و اسناد.....	۸۰
۲-۲-۵ مولفه پیش پردازش.....	۸۱
۳-۲-۵ مولفه یادگیری ماشینی الگوریتم های کلاس بندی متداول.....	۸۵
۴-۲-۵ مولفه الگوریتم کلاس بندی ACC.....	۸۵
۵-۲-۵ مولفه حاشیه نویسی.....	۹۰
۶-۲-۵ مولفه کار با هستی شناسی شبکه واژگان.....	۹۱
۷-۲-۵ مولفه ارزیابی سیستم.....	۹۱
۳-۵ معرفی بستر مورد ارزیابی سیستم سوام.....	۹۲
۴-۵ نتایج حاصل از ارزیابی سیستم.....	۹۶
۵-۵ بررسی نتایج حاصل از آزمایشات ارزیابی سیستم.....	۱۰۳
۶-۵ جمع بندی.....	۱۱۳
فصل ۶: تحلیل نتایج ، نتیجه گیری و کار های آینده	
۱-۶ مقدمه.....	۱۱۴
۲-۶ تحلیل نتایج بدست آمده در پایان نامه.....	۱۱۵
۳-۶ نتیجه گیری.....	۱۱۷
۴-۶ پیشنهادات و کار های آینده.....	۱۱۹
۱: شمای واسط کاربر ارزیابی نتایج سیستم.....	۱۲۱
واژه نامه (انگلیسی به فارسی).....	۱۲۲
منابع و ماخذ.....	۱۲۴

فهرست شکل ها

صفحه	عنوان
۱۰	شکل ۱-۲ سطوح مختلف قابلیت ادراک در داده
۱۲	شکل ۲-۲ طرح لایه ای وب معنایی [۱۳]
۱۸	شکل ۳-۲ نمونه ساختاریک فایل XML
۱۹	شکل ۴-۲ نمونه نحو زبان RDF برای نمایش سه تایی Buddy Belden Owns a Business
۲۰	شکل ۵-۲ نمایش RDF جمله "the people at the meeting were joe, Bob, Susan and Ralph"
۲۰	شکل ۶-۲ نمایش بصری هستی شناسی ذکر شده در شکل ۲-۵
۲۱	شکل ۷-۲ مراحل ایجاد هستی شناسی و مهندسی هستی شناسی [۱۳]
۲۳	شکل ۸-۲ شمای حاشیه نویسی معنایی [۱۸]
۲۳	شکل ۹-۲ نمونه ای از فرا داده های معنایی به زبان XML
۴۱	شکل ۱۰-۲ نحوه عملکرد الگوریتم ماشین بردار پشتیبان
۴۷	شکل ۱-۳ فرایند حاشیه نویسی دستی اسناد وب
۴۸	شکل ۲-۳ فرایند حاشیه نویسی خودکار
۵۷	شکل ۳-۳ فرایند انجام شده در سیستم پنکو [۱۹]
۷۱	شکل ۱-۴ معماری سیستم حاشیه نویسی سوام SWAM
۸۸	شکل ۲-۵ مرحله استخراج مجموعه های کاندید تکرار شونده در ACC [۴۸]، [۳۲]
۸۸	شکل ۳-۵ مرحله استخراج قوانین پیوند در الگوریتم ACC [۴۸]
۸۹	شکل ۴-۵ نمونه ای از قوانین استخراج شده توسط الگوریتم ACC در مرحله آموزش
۹۰	شکل ۵-۵ شمای سیستم حاشیه نویسی سوام به همراه ابزار گیت
۹۲	شکل ۶-۵ فرمت ذخیره یک نمونه از اخبار مجموعه داده رویتر
۹۵	شکل ۷-۵ نمونه مجموعه داده های موجود در کورپوس سنس ایوال
۱۰۴	شکل ۸-۵ تاثیر پارامتر حداقل درجه حمایت بر معیار های فراخوانی و دقت الگوریتم ACC
۱۰۵	شکل ۹-۵ تاثیر پارامتر حداقل درجه اطمینان بر معیار های فراخوانی و دقت الگوریتم ACC
۱۰۷	شکل ۱۰-۵ تاثیر اضافه نمودن واژه های فرا معنی بر دقت سیستم سوام
۱۱۰	شکل ۱۱-۵ توزیع دقت عملکرد سیستم در کلمات مختلف موجود در دیتاست سنس ایوال
۱۱۱	شکل ۱۲-۵ رابطه میان دقت سیستم و تعداد نمونه های موجود در مجموعه داده مورد آزمایش
۱۱۲	شکل ۱۳-۵ روند اثر تعداد معانی معرفی شده در مجموعه داده های سنس ایوال در دقت سیستم

فهرست جدول ها

صفحه	عنوان
۵۸	جدول ۳-۱ جمع بندی خصوصیات ابزارهای حاشیه نویسی [۲۰].
۶۷	جدول ۴-۱ تعداد کلمات موجود در پایگاه دانش هستی شناسی شبکه واژگان در سال ۲۰۰۶ [۴۲].
۶۸	جدول ۴-۲ انواع روابط مربوط به اسم های موجود در سیستم شبکه واژگان (ووردنت) [۹].
۶۹	جدول ۴-۳ انواع روابط مربوط به فعل های موجود در سیستم شبکه واژگان (ووردنت) [۹].
۸۲	جدول ۵-۱ لیست کلمات توقف و کلمات اضافه که از متن حذف می گردند
۹۴	جدول ۵-۲ توزیع کلمات موجود در مجموعه داده سنس ایوال از نظر نقش کلمه [۴۶].
۹۶	جدول ۵-۳ تاثیر پارامتر حداقل درجه حمایت بر پوشش و دقت الگوریتم ACC
۹۷	جدول ۵-۴ تاثیر پارامتر حداقل درجه اطمینان بر پوشش و دقت الگوریتم ACC
۹۸	جدول ۵-۵ نام و گروه طبقه بندی الگوریتم های استفاده شده در سیستم سوام
۹۸	جدول ۵-۶ نتیجه ارزیابی دقت سیستم بدون اعمال مراحل پیش پردازش
۹۹	جدول ۵-۷ تاثیر اضافه نمودن مرحله افزودن واژه های مترادف هایپر نیم بر دقت سیستم
۹۹	جدول ۵-۸ تاثیر اضافه نمودن مرحله حذف واژه های توقف بر کارایی سیستم
۱۰۰	جدول ۵-۹ تاثیر اضافه نمودن مرحله ریشه یابی بر کارایی سیستم
۱۰۰	جدول ۵-۱۰ تاثیر اضافه نمودن صافی TFIDF بر کارایی سیستم
۱۰۰	جدول ۵-۱۱ تاثیر اضافه نمودن انتخاب ویژگی بوسیله صافی بهره اطلاعاتی بر دقت سیستم
۱۰۱	جدول ۵-۱۲ نتایج ارزیابی دقت سیستم توسط الگوریتم ماشین بردار پشتیبان بر داده های سنس ایوال
۱۰۳	جدول ۵-۱۳ اثر اعمال مراحل ترکیبی پیش پردازش بر روی کارایی الگوریتم ماشین بردار پشتیبان در رفع ابهام معنایی کلمات در نقش اسم
۱۰۳	جدول ۵-۱۴ مقایسه بهترین الگوریتم ها پس از اعمال مراحل (اضافه نمودن هایپر نیم + ریشه یابی + حذف واژگان توقف) بر روی کلمات در نقش اسم
۱۰۸	جدول ۵-۱۵ میانگین دقت سیستم توسط الگوریتم ماشین بردار پشتیبان با اعمال مراحل پیش پردازش، بر روی کلمات در نقش های مختلف
۱۰۸	جدول ۵-۱۶ نتایج رفع ابهام تیم های شرکت کننده در مسابقه سنس ایوال ۲ در رفع ابهام کلمات در نقش اسم

فهرست کوتاه نوشت ها

XML	Extensible Markup Language
RDF	Resource Description Framework
OWL	Ontology Web Language
SVM	Support Vector Machine
SWAM	Semantic Web 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
W3C	World Wide Web Consortium
DAML	Darpa Agent Markup Language
B2B	Business to Business
B2C	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 کاربرداز از کاربران دستور گرفته و پس از جستجو در سایت ها اقدام به و رزرو و خرید بلیط هوا پیمای 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.

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. متأسفانه اطلاعات موجود در وب متداول، قابل پردازش و درک برای این نرم افزار های هوشمند نیست. این اطلاعات کاملاً 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 و جوی 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 و یا 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 [۴]، [۲].

با توجه به وجود نا رسایی هایی در وب متداول، پس از ارائه مقاله "تیم برنرلی" در سال ۲۰۰۱ میلادی، وب معنایی پا به عرصه ی دنیای اینترنت گذاشت [۵]. 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 [۶]. با بوجود آمدن زبان XML قادر به نوعی طبقه بندی ساختاری در این اطلاعات شده ولی هنوز این اطلاعات قابل درک و استنتاج توسط کامپیوترها 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^۴ 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 در آن 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 نامرتبط 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.

۲-۱ 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

^۴ 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 نمونه‌ها در هستی‌شناسی نمی‌پردازد، بلکه سیستم فوق جهت تطابق نمونه‌های موجود در هستی‌شناسی شبکه‌واژگان^۹ با کلمات موجود در متن ساخته و ارزیابی شده است، به گونه‌ای که بتواند در معنی کلمات هم شکل رفع ابهام نموده و آنها را حاشیه‌نویسی نماید.

در این سیستم دو دسته الگوریتم جهت یادگیری ماشین بررسی شده است. در روش اول بوسیله روش استخراج قوانین پیوند که یکی از روش‌های داده‌کاوی است الگوریتمی ابتکاری جهت یادگیری سیستم طراحی شده است. این الگوریتم^{۱۰} به گونه‌ای تغییر داده شده تا بتواند توسط داده‌های 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 رده‌بندی^{۱۱} تبدیل نموده و بوسیله ابزار وکا^{۱۲} توسط ۱۰ الگوریتم از روش‌های آموزش با نظارت^{۱۳} یادگیری ماشین تست و ارزیابی شده است. در هر دو روش سیستم بوسیله داده‌های آموزشی حاشیه‌نویسی شده استاندارد مجموعه داده سنس ایوال^{۱۴} مورد تست و ارزیابی قرار گرفته و نتایج الگوریتم‌های مختلف مورد مقایسه قرار گرفته است. سپس راهکارهای پیشنهادی برای بهبود الگوریتم‌های رده‌بندی برای رفع ابهام معنی کلمات بوسیله همین داده‌ها تست و ارزیابی شده و میزان بهبودی الگوریتم‌ها مشخص گردیده است.

۱-۴ ساختار پایان نامه

این پایان‌نامه دارای شش فصل به شرح زیر می‌باشد. پس از فصل اول، در فصل دوم به معرفی وب معنایی و فناوری‌های مربوط به آن مثل هستی‌شناسی و استنتاج از وقایع هستی و جایگاه منطق در وب معنایی پرداخته و سپس مفاهیم کلی آموزش ماشینی و متن‌کاوی را شرح داده ایم. در فصل سوم به کارهای انجام شده قبلی در زمینه حاشیه‌نویسی معنایی، استخراج مفاهیم از متن پرداخته شده است. در فصل چهارم به شرح معماری و طراحی سیستم پیشنهادی پرداخته شده و فصل پنجم 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 راهکارهای آینده جهت بهبود این فناوری پرداخته شده است.