ارائه روشی جهت بهبود دقت سامانه‌های استخراج آزاد اطلاعات با استفاده از الگوریتم شیرمورچه

سمیه حیدری
گروه کامپیوتر، دانشکده کامپیوتر، موسسه آموزش عالی پویش، قم، ایران

زهره بنائیان
گروه کامپیوتر، دانشکده کامپیوتر، موسسه آموزش عالی پویش، قم، ایران

وحیده رشادت
دانشگاه صنعتی مالک‌اشتر

چکیده
استخراج آزاد اطلاعات روش استخراج مستقل از رابطه است که برای استخراج نمونه‌های رابطه در متون بزرگ مانند وب استفاده می‌شود. در این روش به نوع رابطه خاص اشاره نمی‌شود و برخلاف روش‌های پیشین به مجموعه کوچک از روابط در متن محدود نمی‌شود و همه انواع وابستگی‌های موجود در متن را استخراج می‌کند. ازجمله چالش‌های اصلی سامانه‌های استخراج آزاد اطلاعات این است که این سامانه‌ها، قادر به استخراج تمام روابط نیستند و از طرفی خروجی ناقص و نوفه‌داردارند و نیز ممکن است استخراج اطلاعی را در برنداشته باشند. پیش‌بینی دقیق از خروجی سامانه‌های استخراج اطلاعات الزامی می‌باشد و چالشی اساسی به حساب می‌آید. به دلایلی از جمله تجمیع داده‌ها در پایگاه داده‌ها و بهبود یکپارچگی داده‌ها، استخراج اطلاعات محاوره‌ای، نیاز به تخمین ضریب اطمینان وجود دارد که صحت رابطه استخراج شده بین موجودیت‌ها را نشان می‌دهد.

پژوهش حاضر باهدف افزایش دقت سامانه‌های استخراج آزاد اطلاعات با کمک الگوریتم بهینه‌سازی شیرمورچه انجام شده است. در این مقاله چندین ویژگی پیشنهادی که مبتنی بر جمله و مبتنی بر روابط می‌باشد در جهت بهبود دقت سامانه استخراج آزاد اطلاعات استفاده شده است. از دسته‌بند ماشین بردار پشتیبان برای دسته‌بندی اولیه داده‌ها استفاده شده است. سپس دقت دسته‌بند با الگوریتم شیرمورچه بهینه‌سازی شده است. ارزیابی‌ها نشان می‌دهد که ویژگی‌های استخراج شده موثر بوده است و دقت حاصل از الگوریتم بهینه‌سازی شیر مورچه با بالاترین مقدار دقت برابر با۷۴ می‌باشد و نسبت به روش پایه افزایش قابل توجهی داشته است.

کلمات کلیدی:
پردازش زبان طبیعی، استخراج آزاد اطلاعات، بهبود دقت، الگوریتم شیرمورچه

۱- مقدمه

در دهه‌های اخیر اطلاعات متنی در اینترنت رشد سریعی داشته و بخش قابل توجهی از این اطلاعات (اخبار آنلاین،مقالات علمی وکتب و....) به صورت غیر ساخت‌یافته و ناهمگن می‌باشد واطلاعات غیرساخت‌یافته قابل خواندن، سازماندهی و تحلیل توسط ماشین‌ها نیستند. برای اینکه بتوان از بین این حجم عظیم اطلاعات، انسان را درفهم و یافتن اطلاعات مورد نیاز یاری کرد باید بتوان متن غیرساخت‌یافته را به اطلاعات ساخت‌یافته تبدیل کرد. در نتیجه وجود تکنولوژی استخراج اطلاعات الزامی است. سیستم‌های استخراج اطلاعات با تبدیل اطلاعات به صورت ساخت‌یافته فهم آن را برای ماشین آسان و به انسان در درک بهتر این اطلاعات کمک می‌کنند[۱۱و۱۲].

برای مثال در جمله :

John is a graduate student at the University of Pennsylvania

سامانه استخراج اطلاعات John را بعنوان "person" وthe University of Pennsylvania را بعنوان Universityبرمی‌گرداند، student-of نیز رابطه بین John وUniversity of Pennsylvania می‌باشد. Person وstudent-of برچسب‌هایی است که از قبل توسط طراح سامانه تعیین شده‌اند.

با توجه به مثال بالا متوجه می‌شویم که دو وظیفه اصلی استخراج اطلاعات شامل استخراج موجودیت و استخراج رابطه می‌باشد[۱۱].

با استفاده از سامانه‌های استخراج اطلاعات می‌توان پایگاه دانشی ساخت‌یافته از متون ایجاد کرد. دراین راستا استخراج اطلاعات با روش‌های مبتنی بر یادگیری ماشین (باناظر[1] بدون‌ناظر[2] نیمه‌ناظر[3])تلاش در شناسایی حقایق دارند.

استخراج آزاد اطلاعات[4]نیز روشی است که برای استخراج نمونه‌های رابطه در متون بزرگ مانند وب مورد استفاده می‌باشد و برخلاف روش‌های پیشین استخراج اطلاعات، استخراج همه روابط دلخواه ازجملات موجود در متن را فراهم می‌کند[۱۱].

ازجمله این چالش‌های سامانه‌های استخراج آزاد اطلاعات این است که این سیستم‌ها نیز قادر به استخراج تمام روابط نیستند و از طرفی خروجی ناقص و نوفه‌داردارند و نیز ممکن است استخراج اطلاعی را در برنداشته باشند. از سوی دیگر تخمین ضریب اطمینان می‌تواند کارایی الگوریتم‌های داده‌کاوی را بهبود بخشد که به پایگاه داده‌هایی که توسط سامانه‌های استخراج آزاد اطلاعات ایجاد می‌شوند، بستگی دارد. بنابراین افزایش دقت این سامانه‌ها با روش‌های سریع و کارآمد از اهمیت زیادی برخوردار است و باعث می‌شود که تخمین و استنباط‌های دقیقی از خروجی‌ها بدست آید.

با توجه به اینکه افزایش دقت خروجی و کاهش استخراج‌های نادرست در سامانه‌های استخراج آزاد اطلاعات یک چالش اساسی است. بدین منظور در این پژوهش بهبود دقت خروجی سامانه‌های استخراج آزاد اطلاعات با کمک الگوریتم شیرمورچه مورد ارزیابی قرار می‌گیرد. برای تحقق این هدف و به منظور افزایش نمونه رابطه‌های درست روش پیشنهادی به خروجی سامانه استخراج آزاد اطلاعات اعمال شده و خروجی‌ها بررسی می‌شود. بطوری‌که میزان دقت خروجی بدست‌آمده با روش پیشنهادی بالاتر از دقت روش‌های پیشین خواهد بود.

این مقاله به منظور بهبود دقت خروجی‌های استخراج‌گرهای اطلاعات داده‌های ورودی با کمک چندین ویژگی مختلف دسته‌بندی و برچسب‌گذاری می شوند. در این کار از دسته‌بند ماشین بردار پشتیبان استفاده می‌شود که یک روش یادگیری ماشین و سپس روش الگوریتم بهینه سازی شیرمورچه[5]برای تشخیص دقت خروجی‌ها به کار گرفته می‌شود. در نهایت خروجی‌های حاصل از سامانه‌های استخراج اطلاعات با هدف بهبود دقت به منظور صحیح یا ناصحیح بودن آن‌ها ارزیابی می‌شوند.

---

[1] supervised
[2] Un supervised
[3] Semi supervised
[4] Open information extraction
[5] The Ant Lion Optimizer

این مقاله نوآوری های زیر را دارد:

- چندین ویژگی مهم پیشنهاد شده است که شامل ویژگی‌های مبتنی بر آرگومان‌ها و رابطه و ویژگی‌های مبتنی بر جمله می باشد که می‌توان گفت ویژگی‌های کلیدی برای رسیدن به دقت بالاتر در استخراج‌ها به شمار آیند.
- اینکه چگونه ویژگی‌های پیشنهادی دقت روابط استخراجی را افزایش می‌دهد مورد بررسی قرار گرفته است و برای ارزیابی و بهینه‌سازی نتایج، خروجی سامانه‌های استخراج آزاد اطلاعات با الگوریتم شیرمورچه بهبود داده شده است. الگوریتم شیر مورچه یک الگوریتم بهینه‌سازی فرابتکاری است که با بدست آوردن تابع برازندگی و با الگوگیری از حرکت مورچه‌ها به دنبال بهترین تخمین از تابع هدف می‌باشد و در نتیجه با این روش باعث بهبود قابل توجه دقت خروجی سامانه‌های استخراج آزاد اطلاعات می‌شود.

۲- پیش زمینه و کارهای مرتبط

در این بخش ابتدا توضیح مختصری از استخراج آزاد اطلاعات ارائه خواهد شد و سپس کارهای پیشین و مرتبط بیان خواهد شد.

استخراج اطلاعات یکی از مهم‌ترین وظیفه در متن‌کاوی[1]است و مطالعات گسترده‌ای درشاخه‌های مختلف مرتبط شامل پردازش زبان طبیعی[2]و بازیابی اطلاعات[3]انجام شده‌است. هدف اصلی استخراج اطلاعات یافتن اطلاعات ساخت‌یافته از متن غیرساخت‌یافته یا نیمه‌ساخت‌یافته است. دو وظیفه‌اصلی استخراج اطلاعات تشخیص موجودیت اسمی و استخراج رابطه است. استخراج اطلاعات دارای اهمیت زیادی در بسیاری از نرم‌افزارهای درک متن از جمله هوش وب و موتورهای جستجو است. برای استخراج اطلاعات از اسناد متنی، اکثر سیستم‌های استخراج اطلاعات به مجموعه‌ای از الگوهای استخراج تکیه دارد. الگوهای استخراج براساس محدودیت‌های نحوی و معنایی از موجودیت‌های مورد نظر در جملات زبان طبیعی تعریف می‌شوند[۲].

بیشتر کارهای انجام شده برای تخمین ضریب اطمینان برای استخراج اطلاعات از روش‌های یادگیری استفاده کرده‌اند. آقای شفر و همکارانش[4][۳] با کمک مدل‌های مارکوف پنهان برای تخمین ضریب اطمینان در استخراج اطلاعات استفاده کردند. آنها ضریب اطمینان را برای همه‌ی فیلدها تخمین نزدند بلکه تنها برای هر توکن‌های یگانه این کار انجام شد. آنها اطمینان یک توکن را توسط اختلاف بین احتمال اولین و دومین برچسب‌های محتمل آن تخمین زدند. روش‌های استخراج مبتنی بر قاعده، اطمینان رابراساس پوشش قاعده در دادگان آموزشی تخمین می‌زنند. دیگر زمینه‌هایی که در آنها تخمین اطمینان بکار می‌رود شامل دسته‌بندی اسناد[۴]، که دسته بندها با استفاده از ویژگی‌های اسناد ساخته می‌شوند . تشخیص گفتار[۵] که اطمینانی برای کلمه‌ی تشخیص داده شده توسط لیستی از کلماتی تخمین زده می‌شود که معمولا در تشخیص آنها اشتباه رخ می‌دهد. در ترجمه‌ی ماشینی نیز از روش‌های مختلفی از جمله شبکه‌های عصبی برای یادگیری احتمال ترجمه‌ی صحیح یک کلمه با استفاده از ویژگی‌های متن استفاده می‌شود.

در[۶]یک مدل ترکیبی بنام URNSپیشنهاد شده است که تاثیر اندازه‌ی نمونه، فراوانی و اعتبارسنجی از چندین قاعده‌ی استخراج مجزا، روی احتمال صحیح بودن یک استخراج مورد بررسی قرار گرفته است. این روش از داده‌های برچسب‌زده‌شده‌ی دستی استفاده نمی‌کند. نتایج آزمایش‌ها نشان می‌دهد که این روش نسبت به روش‌های بدون ناظر بهتر عمل می‌کند.

خروجی‌های روش استخراج رابطه‌ی نیمه‌ناظر نوفه‌دار است و نیاز به تخمین کیفیت اطلاعات استخراج شده امری ضروری است. در[۷] روشی برای بهبود پیشنهاد شده است که از الگوریتم بیشینه‌سازی امید ریاضی برای ارزیابی خودکار کیفیت الگوهای استخراجی و سه‌تایی‌های رابطه ی بدست آمده استفاده شده است. موثر بودن این روش روی گستره‌ی وسیعی از روابط بررسی

[1] Text mining
[2] Natural language processing
[3] Information Retrieval
[4] Schaffer & et al

شده است .

جردن اشمیدک و همکاران[1][۹] در مقاله خود مسئله جملات پیچیده‌ای که به عنوان ورودی به سیستم‌های استخراج آزاد اطلاعات تغذیه می‌شوند را مطرح کرده است و روش بازسازی چنین جملاتی را مورد بحث و بررسی قرار داده است. در این کار درباره ساخت مجدد جملات پیچیده برای بهبود استخراج رابطه از متن دلخواه بحث شده است. این مقاله جملات را به صورت قطعه قطعه می‌کند و آن دسته ها را برای تعیین اینکه کدام یک از آنها باید به یک جمله‌ی جزئی تقسیم شود، تحلیل می‌کند. دو استراتژی برای چنین تجزیه و تحلیل ارائه شده است که یکی از تجزیه وابستگی استفاده می‌کند و منجر به افزایش قابل توجه در دقت می‌شود، در حالی که دیگری براساس یک طبقه‌بندی است که از ویژگی‌های تکه‌های جمله استفاده می‌کند. نتایج بررسی‌های جردن نشان می‌دهد که این روش موفق به کاهش زمان پردازش و افزایش دقت روی دو سیستم Reverb و Exemplar می شود. به طوریکه بهبود دقت برای ریورب چشمگیر است وافزایش ۹۵٪ دقت در مجموعه داده NYT-500 دیده می‌شود و۱۱۳٪ برایPENN-100 مشاهده می‌شود. همانطور برای EXEMPLAR، بهبود دقت را مشاهده می‌کنیم، اگر چه تقریبا به اندازه (بین ۵٪ و ۸٪) می‌باشد. همچنین مشاهده می‌شود که روش NB-SR یک ساز و کار بسیار جالب ارائه می‌دهد که باعث افزایش قابل توجه دقت، به ویژه برای ReVerb می‌شود، بدون افزایش هزینه محاسبات نسبت به روش DEP-SR. با توجه به نتایج Exemplar با روشNB-SR ، می‌توان اینگونه نتیجه گرفت که NB-SRیک اثر مثبت کم در افزایش دقت در تمام داده‌ها دارد (هرچند در دقت PENN-100 کاهش قطعی دقت مشاهده شد).

در[۱۰] از مدل میدان‌های تصادفی شرطی استفاده شده است که نوعی مدل گرافیکی است که بطور خودکار فیلدهای رکوردها را برچسب می‌زند. در اینجا رکورد، یک بلاک کامل از اطلاعات شخص و فیلد یک جزئی از آن رکورد است. از چندین روش برای تخمین ضریب اطمینان فیلد و رکورد استفاده شده است. در این روش از یک تخمین زن اطمینان مبتنی بر ریاضی برای سامانه‌های استخراج آزاد اطلاعات با وضعیت متناهی استفاده شده است. سامانه استخراج اطلاعاتی بررسی شده براساس مدل میدان‌های تصادفی شرطی خطی- زنجیری است و نتایج حاصل، بهبود قابل توجهی در دقت تخمین درستی فیلد را نشان می‌دهد.

نیکولاس و همکاران[2][۸] یک رویکرد ساده‌سازی ساده ارائه داده‌اند که به منظور بهبود عملکرد سیستم‌های استخراج آزاد اطلاعات طراحی شده است. جملات پیچیده اغلب چالش‌هایی را برای سیستم‌های استخراج آزاد اطلاعات ایجاد می‌کنند. در این پژوهش یک چارچوب ساده‌سازی ایجاد شده است که مرحله‌ای از پیش‌پردازش را با در نظر گرفتن یک جمله واحد به عنوان ورودی و با استفاده از مجموعه‌ای از قوانین گرامری برای ایجاد یک نسخه ساده‌تر که در سیستم‌های استخراج آزاد اطلاعات استفاده می‌شود، انجام می‌دهد. نتایج ارزیابی‌ها نشان داد که روش‌های استخراج آزاد اطلاعات بالاترین حالت دقت وبا کاهش اطلاعات کمتری در هنگام کار بر روی سیستم‌های پیش پردازش شده توسط چارچوب ساده‌سازی مطرح شده در این مقاله به دست می‌دهد و در واقع چارچوب ساده‌سازی قوانین مبتنی بر قاعده ارائه شده در این پژوهش که ساختار زبان شناختی جملات ورودی را ساده‌تر می‌کند موجب بهبود دقت سیستم‌های استخراج آزاد اطلاعات می‌شود.

۳- روش پیشنهادی

در این بخش روش پیشنهادی برای بهبود دقت خروجی سامانه‌های استخراج آزاد اطلاعات شرح داده خواهد شد. بدین منظور در روش پیشنهادی با کمک الگوریتم شیرمورچه بهبود دقت مورد ارزیابی قرار می‌گیرد. برای تحقق این هدف و به منظور افزایش نمونه رابطه‌های درست روش پیشنهادی به خروجی سامانه استخراج آزاد اطلاعات اعمال شده و خروجی‌ها بررسی می‌شود. بطوری‌که میزان دقت خروجی بدست‌آمده با روش پیشنهادی بالاتر از دقت روش‌های پیشین خواهد بود. در روش پیشنهادی از خروجی سامانه‌های استخراج آزاد اطلاعات استفاده می‌شود. نتایج استخراج‌ها با کمک دسته‌بند ماشین بردار پشتیبان دسته‌بندی می‌شوند و سپس با الگوریتم شیرمورچه دقت دسته‌بندی بدست آمده بهینه‌سازی می‌شود.

[1] Jordan schmidk& etal
[2] Nicolas& etal

الگوریتم شیر مورچه در مرحله اول یک جمعیت اولیه از مورچه‌ها را تولید می‌نماید و این امر بدین معناست که یک مجموعه راه‌حل‌های کاملا تصادفی برای حل مسئله بوجود آمده است. در مرحله دوم مشخص می‌نماید که مقادیری که به موقعیت مورچه داده شده، درست می‌باشد یا خیر. در مرحله سوم که مهم‌ترین بخش یک الگوریتم شیر مورچه است، محاسبه تابع برازندگی مورچه صورت می‌گیرد که تابع شایستگی یک مورچه نشان می‌دهد این راه حل تا چه اندازه بهینه می‌باشد. در روش مطرح شده در این پژوهش از دسته بند ماشین بردار پشتیبان برای تشخیص تابع برازندگی استفاده شده است.

پس از آنکه میزان تابع برازندگی هر مورچه بدست آمد در مرحله چهارم، مورچه‌ها را بر اساس مقدار تابع برازندگی‌شان، از مناسب‌ترین تا نامناسب‌ترین مرتب نموده و در مرحله پنجم بهترین‌ها به عنوان نخبه در نظر گرفته می‌شود. از عملگر انتخاب، برای گزینش بهترین مورچه‌ها استفاده نموده و مورچه‌ای که، بالاترین میزان نخبگی را دارا می‌باشند انتخاب می‌گردند این روال آنقدر ادامه می‌یابد که به نرخ قابلی قبولی از تشخیص وجود هدف دست یافته باشیم که در این صورت مورچه‌ای که بعد از اتمام این فرآیند دارای بالاترین میزان تابع برازندگی باشد به عنوان نخبه‌ترین مورچه انتخاب می‌گردد و راه حلی که این مورچه ارائه می‌دهد به عنوان بهترین جواب مسئله در نظر گرفته می‌شود.

در روش شیرمورچه برای دسته‌بندی اولیه داده‌ها ار دسته‌بند ماشین بردار پشتیبان استفاده شده است و دقت بدست آمده از دسته‌بند ماشین بردار پشتیبان با الگوریتم شیر مورچه بهینه‌سازی می‌شود. روش بردارپشتیبان‌ماشین جزء دسته‌بندهای خطی محسوب می‌شود که می‌تواند با روش‌های بدون‌ناظر داده آموزشی را به بهترین شکل دسته‌بندی کند و هدف اصلی ماشین‌بردارپشتیبان[1] انتخاب جداکننده‌ای که بهترین تفکیک از کلاس‌های مختلف در فضای ویژگی بدست بیاورد می‌باشد. دسته‌بند بردار پشتیبان ماشین روی یک مجموعه داده برچسب زده شده با ویژگی‌های پیشنهادی آموزش داده می‌شود و با بدست آوردن بهترین تفکیک از کلاس‌های مختلف در ویژگی‌های استفاده شده برای رسیدن به ضریب اطمینان مناسب برای استخراج‌ها استفاده می‌شود.

۱-۳- ویژگی‌های استخراج شده از مجموعه داده

چندین ویژگی برای دسته‌بند مورد استفاده در این روش، از دادگان آموزشی استخراج و در نظر گرفته شده است که در ادامه توضیح داده خواهد شد.

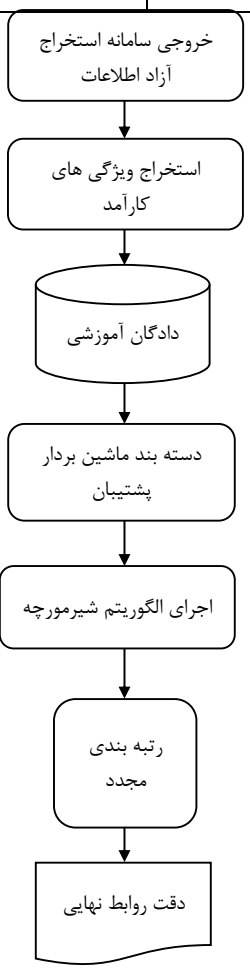
ویژگی‌های مبتنی بر آرگومان‌ها و رابطه: این ویژگی‌ها شامل (موجودیت نامدار، قید ، صفت، ضمیر، فعل، موجودیت عددی، حرف‌اضافه، ضمیرو اسم، اسم و صفت، اسم و عدد، اسم و قید، حرف اضافه در انتهای رابطه)می‌باشد. این ویژگی‌ها بر این‌اساس است که ممکن است هر کدام از این موارد در آرگومان ۱، رابطه استخراج شده و یا آرگومان ۲ موجود باشد.

ویژگی‌های مبتنی بر جمله: این ویژگی‌ها شامل (طول جمله، تعداد کلمات موجود در جمله، تعداد کلمات موجود در بین آرگومان ۱ و ۲، تعداد ضمیرهای موجود در جمله، تعداد آرگومان‌های موجود در یک جمله) می‌باشد. این ویژگی‌ها با توجه به اینکه در اکثر موارد از یک جمله چندین استخراج انجام شده است و فروانی آرگومان وجود دارد در سطح جمله مورد بررسی قرار گرفته است. در مورد طول جمله و تعداد کلمات موجود در جمله هم می‌توان به این مورد اشاره کرد که هر چه این ویژگی مقدار بالاتری داشته باشد احتمال استخراج تعداد روابط نیز بیشتر می‌شود.

[1] Support vector machine

جدول ۱: ویژگی‌های استخراج شده از دادگان آموزشی

نام ویژگی	نوع ویژگی
ویژگی‌های مبتنی بر آرگومان‌ها و رابطه	موجودیت نامدار
	قید
	صفت
	ضمیر
	فعل
	موجودیت عددی
	حرف اضافه
	ضمیرو اسم
	اسم و صفت
	اسم و عدد
	اسم و قید
	حرف اضافه در انتهای رابطه
ویژگی‌های مبتنی بر جمله	طول جمله
	تعداد کلمات موجود در جمله
	تعداد کلمات موجود در بین آرگومان ۱ و ۲
	تعداد ضمیرهای موجود در جمله
	تعداد آرگومان های موجود در یک جمله



شکل۱-ساختار کلی روش پیشنهادی

۴- آزمایش‌ها و ارزیابی

در این بخش تاثیر استفاده از الگوریتم شیرمورچه روی خروجی سامانه‌های استخراج آزاد اطلاعات با کمک ویژگی‌های پیشنهادی ارزیابی و مقایسه شده است.

داده آموزشی استفاده شده در این مقاله از داده‌های موجود و دردسترس می‌باشد[1] این داده آزمایشی از تعداد ۵۰۰جمله نمونه‌برداری شده از وب با استفاده از سرویس لینک تصادفی یاهو ایجاد شده است.این مجموعه داده شامل خروجی استخراجگرهای مختلف از جمله(Reverb و Textrunner) روی ۵۰۰حمله انتخابی است. بعد از اجرای استخراجگر روی جملات، استخراج‌ها بطور مستقل برای تعیین درستی یا نادرستی‌شان بررسی شده‌اند. بررسی‌های انجام شده روی ۸۶ ٪ از استخراج‌ها با معیار توافق k=۰,۶۸ صورت گرفته است. آزمایش‌ها روی زیر مجموعه‌ای از داده‌ها صورت گرفته که دوبرچسب‌زن به توافق رسیده‌اند. قضاوت‌هایی که به عنوان استخراج‌های بی‌معنا برچسب خورده‌اند، نادرست هستند. در این مجموعه، استخراج ها از مجموعه ای از ۱۰۰۰جمله از وب و ویکی پدیا نیز بطور دستی بصورت درست یا نادرست برچسب زده شده‌اند.

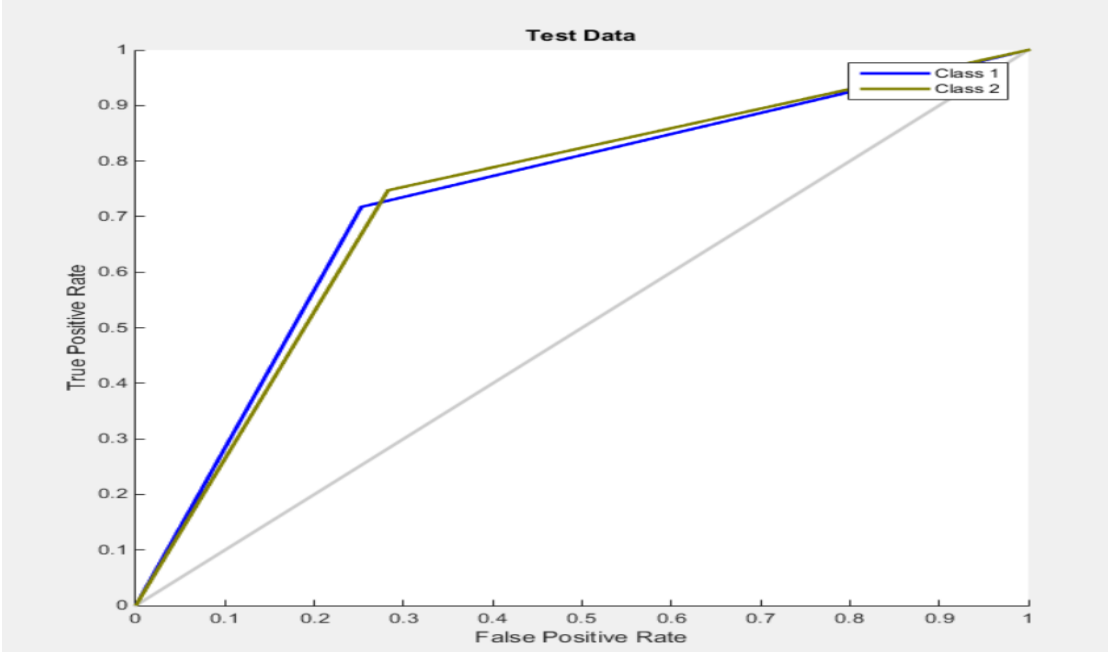
استفاده از دسته‌بند ماشین بردار پشتیبان بعد از خواندن داده‌ها، آنها به نسبت ۷۰به ۳۰به دو دسته آموزش و آزمایش تقسیم می‌شوند و سپس نرخ یادگیری تعیین می‌شود ودسته بند ماشین بردار پشتیبان با داده آموزش، آموزش داده می‌شود تا براساس نرخ خطا، تابع برازندگی تخمین زده شود. در بخش بعدی با داده آزمایش، مورد آزمایش قرار می‌گیرد تا میزان صحت برآورد شده توسط ماشین بردارپشتیبان برآورد گردد و زمانی که تخمین درستی از داده‌ها زده شود مقدار دقت در خروجی نمایش داده می‌شود.

جدول ۲-مقایسه دقت در روش پیشنهادی با الگوریتم شیرمورچه

روش	دقت-Accuracy
روش پایه	۰,۵۰
ماشین بردار پشتیبان	۶۷/۳۰
روش بهینه‌سازی شیرمورچه	۷۴

همانطور که در جدول ۲ مشاهده می‌شود دقت بدست آمده با استفاده از الگوریتم شیر مورچه با مقدار ۷۴ می‌باشد و نسبت به روش پایه افزایش قابل توجه داشته است واین نشان می‌دهد که الگوریتم بهینه‌سازی شیر مورچه بهبود بهتر و قبولی نسبت به روش پایه دارد و خروجی سامانه‌های استخراج آزاد اطلاعات نیز با این روش بهبود داشته است .

[1] . http://reverb.cs.washington.edu

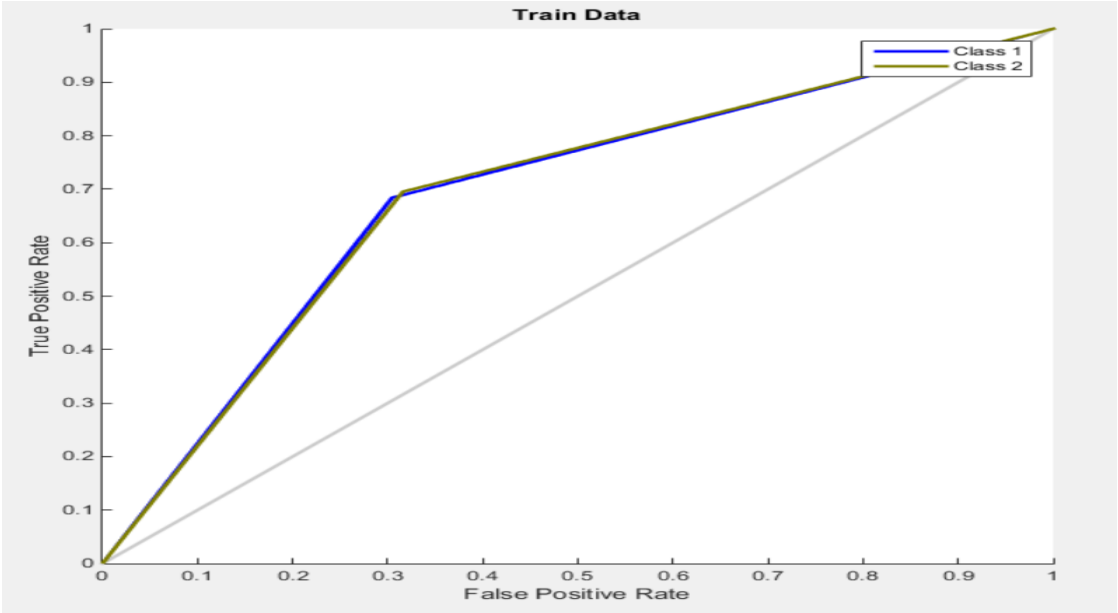


شکل ۲- نمودار ROC الگوریتم شیر مورچه با داده آزمایشی

منحنی roc یا مشخصه عملکرد گیرنده یکی از روش‌های ارزیابی مدل است. در منحنی roc محورهای افقی و عمودی تعریف می‌شوند. محورافقی در این نمودار عبارت است ازمثبت‌هایی که نادرست پیش‌بینی شده است و محور عمودی عبارت است از مثبت‌هایی که به درستی پیش‌بینی شده است. در این نمودار هر چه نقاط به سمت بالا و چپ نزدیک‌تر باشد مناسب‌تر است و مدل پیش‌بینی به حالت ایده‌آل خود نزدیک‌تر است.

همانطور که در شکل ۲ مشاهده می‌شود منحنی مشخصه عملکرد گیرنده دقت مدل پیشنهادی را بررسی کرده است. در این منحنی با رسم مثبت‌های درست پیش‌بینی شده در مقابل مثبت‌هایی که نادرست پیش‌بینی شده است آستانه دقت بدست آمده نشان داده شده است. در شکل ۲ منحنی مشخصه عملکرد گیرنده برای داده آزمایشی نشان داده شده است، همانطور که مشاهده می‌شود بهترین نقطه در این منحنی نقطه۰,۲ =Fp و۰,۷ =Tp می‌باشد که بهترین آستانه دقت برای منحنی را برای داده آزمایشی نشان می‌دهد.

[1] Receiver operating characteristics



شکل ۳- نمودار ROC الگوریتم شیر مورچه با داده آموزشی

برای داده آموزشی نیز بهترین حدآستانه دقت با کمک منحنی مشخه عملکرد گیرنده ارزیابی شده است. همانطور که مشاهده می‌شود خط زیرین منحنی نرخ مثبت‌هایی که نادرست پیش‌بینی شده است را نشان می‌دهد و خط بالای منحنی نرخ مثبت‌هایی که به درستی پیش‌بینی شده‌اند را نشان می‌دهد که بهترین نقطه برای داده آموزشی در این منحنی برابر با۰,۳ =Fp و۰,۷ =Tp می‌باشد. و با مقایسه شکل ۲ و ۳ مشاهده می‌شود که دقت در داده آموزشی افزایش یافته است .

۵- نتیجه‌گیری و کارهای آینده

استخراج اطلاعات فرایند استخراج خودکار داده‌های ساخت‌یافته از متن غیرساخت‌یافته است. یکی از وظایف اصلی در استخراج اطلاعات، استخراج رابطه است که روابط معنایی بین موجودیت‌ها از متون زبان طبیعی را استخراج می‌کند. رویکرد استخراج آزاد اطلاعات نیز این است که روش‌های استخراج اطلاعات را از جهت اندازه و تنوع به مقیاس وب سوق دهد. این روش‌ها اغلب خودناظر هستند و با ایجاد خودکار دادگان آموزشی با استفاده از دسته‌بند و به کمک ویژگی‌های مختلف، روابط را تشخیص می‌دهند. از آنجایی که استخراج آزاد اطلاعات هرگز بطور کامل دقیق نیست، داشتن معیار ضریب اطمینان موثر، برای داشتن خروجی‌های صحیح و دقیق‌تر مفید به نظر می‌رسد تا موجب افزایش یکپارچگی داده‌ها که نیازمند پیش‌بینی دقیق از خروجی‌های صحیح سامانه‌های استخراج اطلاعات می‌باشد. لذا در این پژوهش به ارائه روشی جهت بهبود دقت خروجی سامانه‌های استخراج آزاد اطلاعات با استفاده از روش‌های یادگیری ماشین پرداخته شد. در این کار از دسته‌بند ماشین بردار پشتیبان، برای دسته‌بندی دادگان آموزشی استفاده شده است. مقدار دقت در روش پایه برابر ۰/۵۰ بود که با ارزیابی‌های انجام شده نتایج نشان دادکه دقت بدست آمده از دسته‌بند ماشین‌بردارپشتیبان مقدار ۶۷/۳۰ را نشان می‌دهد و الگوریتم شیرمورچه افزایش قابل قبولی داشت بطوریکه دقت بدست آمده مقدار ۷۴نشان می‌دهد که بهبود حاصل شده است. ونیز دقت در خروجی سامانه استخراج آزاد اطلاعات نسبت به روش پایه بهبود بالایی داشته است.

در آینده با استفاده از ویژگی‌های بیشتر به منظور بهبود کارایی مدل مطرح شده تلاش خواهیم کرد. همانطور که گفته شد در این پژوهش نتایج آزمایش بر روی سامانه ریورب بررسی شد، لذا پیشنهاد می‌شود نتایج بر روی سایر سامانه‌های استخراج آزاد اطلاعات از جمله woe، Ollie، textrunner و .... اعمال شود و خروجی آنها مقایسه شود. علاوه بر این می‌توان از الگوریتم Random Forest برای دسته‌بندی (یکی از روش‌های رده‌بندی است که از ترکیب مجموعه‌ای از درخت‌های تصمیم

تشکیل‌شده است)و مقایسه نتایج مشاهده‌شده توسط آن در جهت افزایش کارایی سامانه‌ها و بهبود نتایج استفاده کرد. همچنین می‌توان روش پیشنهادی را در داده‌کاوی، دسته‌بندی متون و نظیر آنها به کار گرفت. در نهایت برای بهبود نتایج می‌توان آزمایشات را بر روی داده آموزشی بزرگتر انجام داد.

مراجع

[1] J. Piskorski and R. Yangarber, “Information Extraction: Past, Present and Future,” 2013, pp. 23–49.
[2] S. G. Small and L. Medsker, “Review of information extraction technologies and applications,” Neural Comput. Appl., vol. 25, no. 3–4, pp. 533–548, Sep. 2014.
[3]Scheffer T., Decomain C. and Wrobel S., "Active hidden markov models for information extraction", Advances in Intelligent Data Analysis. Springer, pp. 309-318, 2001.
[4]Bennett P. N., Dumais S. T. and Horvitz E., "Probabilistic combination of text classifiers using reliability indicators: Models and results", In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 207-214.
[5]Gunawardana A., Hon H-W. and Jiang L., "Word-based acoustic confidence measures for large-vocabulary speech recognition", In ICSL, 1998
[6]Downey D., Etzioni O. and Soderland S., "Analysis of a probabilistic model of redundancy in unsupervised information extraction", Artificial Intelligence, 174(11): 726-748, 2010.
[7] Agichtein E., "Confidence estimation methods for partially supervised relation extraction", In Proc. of SIAM Intl. Conf. on Data Mining (SDM06), 2006.
[8] C. Niklaus, B. Bermeitinger, and S. Handschuh, “A Sentence Simplification System for Improving Relation Extraction,” pp. 2–5, 2013.
[9] J. Schmidek, “Improving Open Relation Extraction via Sentence Re-Structuring,” pp. 3720–3723.
[10] A. Culotta and A. Mccallum, “Confidence Estimation for Information Extraction,” 2001.
[11] C. C. Aggarwal and C. Zhai, Eds., Mining Text Data. Boston, MA: Springer US, 2012.
[12] M. banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, Open Information Extraction for the Web. University of Washington, 2009.

Somaye Heidary

Iran ,Qom, Pooyesh Higher Education Institute, Faculty Of Computer Engineering

Zoherh Banaiyan

Iran ,Qom, Pooyesh Higher Education Institute, Faculty Of Computer Engineering

Vahideh Reshadat

Malek-Ashtar University of Technology

Abstract

open information extraction is a non-relation-independent extraction method used to extract relation instances in large texts such as the web. In this method, it does not refer to a specific type of relationship, and, unlike previous methods, it does not limit the small set of relation in the text, and extracts all sorts of dependencies in the text. One of the main challenges of open information extraction systems is that these systems are not able to extract all relation and, on the other hand, they are incomplete, and may also not be extracted from the information. careful prediction of the output of information extraction systems is a must-see and a major challenge. For some reason, such as aggregating data in databases and improving data integrity, the extraction of interactive information requires the evaluation of a confidence coefficient that shows the accuracy of the relation extracted between entities.

The purpose of this study was to increase the accuracy of open information extraction systems with the help of Ant Lion optimization algorithm. In this paper, a number of sentence-based and relation-based features are used to improve the accuracy of the information extraction system. A classifier vector machine has been used for initial data categorization. Then, the precision of the group is optimized by the Ant Lion optimization algorithm. The evaluations show that the extracted properties have been effective, and the precision of the outcome optimization algorithm with the highest accuracy value is 74 and has increased significantly compared to the base method.

key words: Natural language processing, open information extraction, Improved accuracy, Ant Lion algorithm