دانشگاه علامه طباطبائی
دانشکده آمار، ریاضی و رایانه

پایان‌نامه کارشناسی ارشد

رشته علم داده‌ها

عنوان

تبدیل داده‌های بدون ساختار اینترنت اشیاء به داده‌های ساختاریافته با استفاده از الگوریتم‌های یادگیری ماشینی

استاد راهنما

دکتر فرزاد اسکندری

استاد مشاور

دکتر محمد بامنی‌مقدم

استاد داور

دکتر رضا پورطاهری

پژوهش‌گر

سجاد حاجی‌زاده

پاییز ۱۴۰۲


بسم الله الرحمن الرحيم

منشور اخلاق پژوهش

به یاری از خداوند سبحان و اعتقاد به این که عالم محضر خداوند است و همواره ناظر به اعمال انسان و به منظور پاس‌داشت مقام بلند دانش و پژوهش و نظر به اهمیت جایگاه دانشگاه در اعتلای فرهنگ و تمدن بشری ما دانشجویان دانشکده‌های دانشگاه علامه طباطبائی متعهد می‌گردیم اصول زیر را در انجام فعالیت‌های پژوهشی مد نظر قرارداده و از آن تخطی نکنیم:

۱. اصل حقیقت‌جویی: تلاش در راستای پی‌جویی حقیقت و وفاداری به آن و دوری از هرگونه پنهان‌سازی حقیقت.

۲. اصل رعایت حقوق: التزام به رعایت کامل حقوق پژوهشگران و پژوهیدگان (انسان، حیوان و نبات) و سایر صاحبان حق.

۳. اصل مالکیت مادی و معنوی: تعهد به رعایت کامل حقوق مادی و معنوی دانشگاه و کلیه همکاران پژوهشی.

۴. اصل منافع ملی: تعهد به رعایت مصالح ملی و در نظر داشتن پیشبرد و توسعه کشور در کلیه مراحل پژوهش.

۵. اصل رعایت انصاف و امانت: تعهد به اجتناب از هرگونه جانب‌داری غیر علمی و حفاظت از اموال، تجهیزات و منابع در اختیار.

۶. اصل رازداری: تعهد به صیانت از اسرار و اطلاعات محرمانه افراد، سازمان‌ها و کشور و کلیه افراد و نهادهای مرتبط با تحقیق.

۷. اصل احترام: تعهد به رعایت حریم‌ها و حرمت‌ها در انجام تحقیقات و رعایت جانب نقد و خودداری از هرگونه حرمت شکنی.

۸. اصل ترویج: تعهد به رواج دانش و اشاعه نتایج تحقیقات و انتقال آن به همکاران علمی و دانشجویان به غیر از مواردی که منع قانونی دارد.

۹. اصل برائت: التزام به برائت‌جویی از هرگونه رفتار غیر حرفه‌ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را به شائبه‌های غیر علمی می‌آلایند.

نام و نام خانوادگی: سجاد حاجی‌زاده

امضا:

پاییز ۱۴۰۲

تعهدنامه‌ی اصالت پایان‌نامه

عنوان پایان‌نامه: تبدیل داده‌های بدون ساختار اینترنت اشیاء

به داده‌های ساختاریافته با استفاده از الگوریتم‌های یادگیری ماشینی

پژوهش‌گر: سجاد حاجی‌زاده
شماره‌ی دانشجویی: ۴۰۰۱۳۱۴۱۰۱۱
استاد راهنما: دکتر فرزاد اسکندری

این‌جانب **سجاد حاجی‌زاده** دانش آموخته مقطع تحصیلی کارشناسی ارشد رشته‌ی **رشته علم داده‌ها** از دانشکده‌ی علوم ریاضی و رایانه دانشگاه علامه طباطبائی هستم و از پایان‌نامه خود در پاییز ۱۴۰۲دفاع نموده‌ام، متعهد می‌شوم:

۱. این پایان‌نامه حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و در مواردی که از دستاوردهای علمی و پژوهشی دیگران (اعم از مقاله، کتاب، پایان‌نامه و غیره) استفاده نموده‌ام، مطابق ضوابط و رویه موجود، نام منبع مورد استفاده و سایر مشخصات آن را در فهرست مربوط ذکر و درج کرده‌ام.

۲. این پایان‌نامه قبلا برای دریافت هیچ مدرک تحصیلی (هم سطح، پایین‌تر یا بالاتر) در سایر دانشگاه‌ها و موسسات آموزش عالی ارائه نشده است.

۳. چنانچه بعد از فراغت از تحصیل، قصد استفاده از هرگونه بهره‌برداری اعم از چاپ کتاب، ثبت اختراع و ازین دست موارد از این پایان‌نامه را داشته باشم، از حوزه معاونت پژوهشی دانشگاه علامه طباطبائی مجوزهای مربوطه را اخذ نمایم.

۴. چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می‌پذیرم و دانشگاه مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و در صورت ابطال مدرک تحصیلی‌ام هیچ‌گونه ادعایی نخواهم داشت.

نام و نام خانوادگی: سجاد حاجی‌زاده
امضا:
پاییز ۱۴۰۲

دانشگاه علامه طباطبائی
دانشکده آمار، ریاضی و رایانه

پایان‌نامه کارشناسی ارشد

تبدیل داده‌های بدون ساختار اینترنت اشیاء به داده‌های ساختاریافته با استفاده از الگوریتم‌های یادگیری ماشینی

پژوهش‌گر: سجاد حاجی‌زاده

استاد راهنما: دکتر فرزاد اسکندری امضاء:

استاد مشاور: دکتر محمد بامنی‌مقدم امضاء:

استاد داور: دکتر رضا پورطاهری امضاء:

تقدیم به پدر مادر و مهربانم که در سختی‌ها و دشواری‌های زندگی، همواره یاوری دلسوز و فداکار و پشتیبانی محکم و مطمئن برایم بوده‌اند.

سپاس‌گزاری

سپاس خدای را که هر توفیقی در گرو عنایت اوست. اکنون که با یاری او توانسته‌ام تلاشی هر چند ناچیز را در راه کسب دانش به انجام رسانم، بر خود لازم می‌دانم از استاد راهنمای بزرگوارم، جناب آقای دکتر فرزاد اسکندری که به پایان رساندن این تحقیق جز با راهنمایی‌های پدرانه و هدایت‌های بی‌دریغ ایشان میسر نبود، قدردانی نمایم.

امیدوارم بتوانم از عهده ادای حق این عزیزان برآیم.

پاییز ۱۴۰۲

چکیده

درسال‌های اخیر اینترنت اشیاء (IoT) بسیار مورد توجه پژوهش‌گران قرارگرفته است. این فناوری به کمک حسگرهایی، اطلاعات را از محیط اطراف خود دریافت می‌کند و با شبکه‌ها و دستگاه‌های دیگر ارتباط برقرار می‌کند و داده‌های گردآوری شده را به سیستم‌های مرکزی ارسال می‌کند. این فناوری یکی از ابزارهای مهم و قدرتمند درجهت ارتقای امنیت، کیفیت زندگی و بهره‌وری به‌کارگرفته می‌شود و در صنایع مختلفی چون سلامت و پزشکی، کشاورزی، خودروسازی، آب و هوا و غیره به‌عنوان ابزاری برای بهبود کارایی و کاهش هزینه‌ها استفاده می‌شود. با استفاده از اینترنت اشیاء می‌توان از این اطلاعات برای تحلیل و پیش‌بینی رفتارهایی که در آینده ممکن است رخ دهد، استفاده‌کرد و در نتیجه امکان پیش‌بینی مشکلات و راه‌حل‌هایی برای آن وجود دارد.

یکی از چالش‌های موجود در زمینه اینترنت اشیاء پردازش داده‌ها و کاربردهای آن است. این چالش به دلیل وجود داده‌های بدون ساختار ایجادشده از طریق حسگرها، داده‌های شبکه‌های اجتماعی و تارنما (وب) که توسط شرکت‌ها و سازمان‌ها گردآوری شده‌اند، به وجودآمده است. با توجه به تعداد بالای داده‌های بدون ساختار موجود در حوزه‌های مختلف درصورتی که داده‌های بدون ساختار به داده‌های ساختاریافته تبدیل شوند این امکان را به ما می‌دهد که بتوانیم از آن‌ها به نحو بهتری استفاده کنیم، الگوهایی را که درآن‌ها وجود دارد شناسایی کنیم و برای پیش‌بینی رفتار آینده، تصمیم‌گیری‌های بهتری بگیریم.

در این پایان‌نامه، ما به بررسی فرایندهای مختلف تبدیل داده‌های بدون ساختار به داده‌ساختار یافته و استفاده از روش‌های یادگیری ماشین به منظور کشف الگوها از داده بدون ساختار اینترنت اشیاء می‌پردازیم.

واژگان کلیدی: *الگوریتم‌های یادگیری ماشین، داده‌های بدون ساختار، داده‌های ساختاریافته، رده‌بندی، حسگرهای اینترنت اشیاء.*

فهرست مطالب

فهرست تصاویر

نمادها و علامت‌های اختصاری

اینترنت اشیاء	IoT
کشف دانش در پایگاه داده‌ها	KDD
کا نزدیکترین همسایه	KNN
ماشسن بردار پشتیبان	Svm

فصل ۱

کلیات پژوهش

۱.۱ مقدمه

در دهه‌های اخیر، با پیشرفت فراوان فناوری اطلاعات و ارتباطات، نحوه‌ی تولید، جمع‌آوری، و مدیریت داده‌ها به یکی از چالش‌های اساسی در حوزه تکنولوژی تبدیل شده است. امروزه داده‌ها به عنوان یک دارایی ارزشمند شناخته می‌شوند و نقش بسیار مهمی در تصمیم‌گیری‌ها و توسعه‌ی فناوری ایفا می‌کنند. با این حال، این حجم عظیم از داده‌ها همراه با ویژگی‌های مختلف و بدون ساختار،چالش‌هایی در زمینه مدیریت، تحلیل، و بهره‌برداری از آنها ایجاد کرده است. یکی از مسائل اساسی موجود در حوزه داده، وجود داده‌های بدون ساختار است. این داده‌ها ممکن است از منابع مختلفی چون رسانه‌های اجتماعی، وب، سنسورها و دستگاه‌های مختلف جمع‌آوری شوند. عدم ساختار مشخص و یکنواخت در این داده‌ها باعث ایجاد چالش‌هایی در تحلیل و استفاده از آنها می‌شود. از این رو، توسعه روش‌ها و الگوریتم‌های موثر برای استخراج اطلاعات مفید از داده‌های بدون ساختار از اهمیت بسیاری برخوردار است. همچنین، با گسترش اینترنت اشیا، داده‌ها از منابع متنوعی چون حسگرها، دستگاه‌های هوشمند، و سیستم‌های مختلف جمع‌آوری می‌شوند. این حجم عظیم از داده‌ها، افزایش پیچیدگی تحلیل و مدیریت داده را به همراه داشته است. در این زمینه، امکان استفاده بهینه از داده‌های اینترنت اشیاء و تحلیل صحیح آنها، نقش مهمی در پیشرفت فناوری اطلاعات و ارتباطات ایفا می‌کند. در این پژوهش، ما به بررسی فرایندهای مختلف تبدیل داده‌های بدون ساختار به داده‌ساختار یافته و استفاده از روش‌های یادگیری ماشین به منظور کشف الگوها از داده بدون ساختار اینترنت اشیاء می‌پردازیم.

۲.۱ بیان مسئله

درسال‌های اخیر اینترنت اشیاء (IoT) بسیار مورد توجه پژوهش‌گران قرارگرفته است. این فناوری به کمک حسگرهایی، اطلاعات را از محیط اطراف خود دریافت می‌کند و با شبکه‌ها و دستگاه‌های دیگر ارتباط برقرار می‌کند و داده‌های گردآوری شده را به سیستم‌های مرکزی ارسال می‌کند. این فناوری یکی از ابزارهای مهم و قدرتمند درجهت ارتقای امنیت، کیفیت زندگی و بهره‌وری به‌کارگرفته می‌شود و در صنایع مختلفی چون سلامت و پزشکی، کشاورزی، خودروسازی، آب و هوا و غیره به‌عنوان ابزاری برای بهبود کارایی و کاهش هزینه‌ها استفاده می‌شود. با استفاده از اینترنت اشیاء می‌توان از این اطلاعات برای تحلیل و پیش‌بینی رفتارهایی که در آینده ممکن است رخ دهد، استفاده‌کرد و در نتیجه امکان پیش‌بینی مشکلات و راه‌حل‌هایی برای آن وجود دارد.

یکی از چالش‌های موجود در زمینه اینترنت اشیاء پردازش داده‌ها و کاربردهای آن است. این چالش به دلیل وجود داده‌های بدون ساختار ایجادشده از طریق حسگرها، داده‌های شبکه‌های اجتماعی و تارنما (وب) که توسط شرکت‌ها و سازمان‌ها گردآوری شده‌اند، به وجودآمده است. با توجه به تعداد بالای داده‌های بدون ساختار موجود در حوزه‌های مختلف درصورتی که داده‌های بدون ساختار به داده‌های ساختاریافته تبدیل شوند این امکان را به ما می‌دهد که بتوانیم از آن‌ها به نحو بهتری استفاده کنیم، الگوهایی را که درآن‌ها وجود دارد شناسایی کنیم و برای پیش‌بینی رفتار آینده، تصمیم‌گیری‌های بهتری بگیریم.

در این پژوهش، ما به بررسی فرایندهای مختلف تبدیل داده‌های بدون ساختار به داده‌ساختار یافته و استفاده از روش‌های یادگیری ماشین به منظور کشف الگوها از داده بدون ساختار اینترنت اشیاء می‌پردازیم.

۳.۱ پیشینه پژوهش

با افزایش حجم داده‌ها و تنوع منابع تولید داده،چالش‌های زیادی در زمینه مدیریت و تجزیه و تحلیل داده‌های بزرگ بدون ساختار به‌وجود آمده‌است که پژوهشگران راهکارهایی برای برطرف کردن این چالش‌ها ارائه کرده‌اند.برخی از پژوهش‌ها در این زمینه، به بررسی روش‌های موجود برای پردازش و مدیریت داده‌های بدون ساختار، تبدیل داده‌های بدون ساختار به داده‌های ساختاری و به‌دست آوردن اطلاعات معنی‌دار از داده‌ها می‌پردازند.

عابدین و همکاران (۲۰۱۰) به ارائه تحقیقات در مورد شناسایی داده‌های بدون ساختار ، استخراج و طبقه‌بندی صفحات وب که سپس به سند ساختاریافته به زبان نشانه‌گذاری (XML) تبدیل می‌شود پرداخته که بعداً در یک پایگاه داده چندرسانه‌ای ذخیره می شود.

(گاندی و مدیا، ۲۰۱۶). پژوهشی با موضوع استخراج اطلاعات از منابع ناساختار و نیمه‌ساختار در زمینه فناوری اطلاعات با استفاده ازsemantic Web، انجام داده‌است. در این پژوهش، نویسنده با ارائه یک سیستم پیشنهادی، اطلاعات مفید و ساختاری

موردنیاز از منابع مختلف مثل مجلات علمی معتبر در زمینه فناوری اطلاعات را استخراج کرده است. با استفاده از semantic Web، داده‌ها به صورت ساختاری و با معنا استخراج شده و معنای داده‌ها در طول استخراج حفظ شده‌است. هدف این پژوهش، کاهش زمان و هزینه‌های مرتبط با خواندن مجلات به‌طور کامل بوده‌است که در نتیجه می‌تواند به عنوان یک راه‌حل کارآمد در استخراج اطلاعات از منابع ناساختار و نیمه‌ساختار، به ویژه در زمینه فناوری اطلاعات، مورد استفاده قرارگیرد.

میشرا و میسره (۲۰۱۷) به بررسی تکنیک‌هایی برای تجزیه وتحلیل داده‌های بدون ساختار برای استخراج اطلاعات معنی‌دار پنهان در داده‌های بزرگ پرداخته‌است.

سامبرکار و همکاران (۲۰۱۸) به ارائه یک روش پیشنهادی برای تبدیل داده‌های بدون ساختار کشاورزی به داده های نیمه‌ساختاریافته یا ساختاریافته با استفاده از Couchbase ابزار NOSQL ارائه شده است. Couchbase پایگاه داده سند با معماری توزیع شده‌است که مقیاس پذیری ، عملکرد و در دسترس بودن را فراهم می‌کند.

تکلی (۲۰۱۶) در این مقاله، نویسنده بررسی کرده‌است که چگونه می توان از تحلیل و تمایز داده‌های نیمه ساختار یافته در قالب XML برای برنامه های هوشمند استفاده کرد. او به بررسی روش‌های مختلفی برای تحلیل و تمایز داده‌های XML و نیمه ساختاریافته پرداخته‌است و همچنین به بررسی کاربردهای مختلفی که از تحلیل داده‌های XML می‌توان داشت، از جمله خوشه‌بندی داده‌ها، یادگیری انطباقی و ... همچنین، نویسنده به بررسی چالش‌های مرتبط با تحلیل داده‌های XML پرداخته‌است.

آرتایلز و همکاران (۲۰۰۴) یک روش برای تشخیص معنای کلمات بر اساس شباهت میان کلمات در فضای متنی ارائه کرده‌است.

عبدالله و کمسوریه (۲۰۱۳) پژوهشی در مورد فرآیند نقشه برداری داده‌های بدون‌ساختار به داده‌های ساختاری انجام داده‌است در این مقاله چهار (۴) فرآیند اصلی که شامل استخراج، طبقه بندی، توسعه مخازن و نقشه‌برداری داده‌ها با قصد کمک به تولید داده‌ها و اطلاعات جدید است که ساختاریافته‌تر، جامع‌تر، جمعی و متنوع‌تر در مباحث هستند تا نیازهای سازمانی را ارائه دهند. خروجی از این مطالعه اطمینان از فرآیند نقشه‌برداری داده‌های بدون‌ساختار به داده‌های ساخت‌یافته، داده‌های بدون ساختار را به عنوان دارایی‌های تجدید پذیر مفید، آگاه و معنادار برای خدمت به عملکردهای سازمانی نشان می‌دهد.

روسو و همکاران (۲۰۱۳) به تبدیل داده‌های بدون‌ساختار و نیمه‌ساختاریافته به دانش پرداخته‌است. استخراج دانش، فرایند ایجاد دانش از داده‌های ساختاری، بدون‌ساختار و نیمه‌ساختاریافته است. این مقاله به بررسی امکانات استخراج دانش از داده‌های بدون‌ساختار و نیمه‌ساختار یافته می‌پردازد. نظریه‌ها و ابزارهای استخراج دانش در زمینه نوظهور، کشف دانش در پایگاه داده‌ها (KDD) مورد بررسی قرارگرفته و روند کلی KDD چند مرحله‌ای بیان شده است. درادامه، برنامه‌های اخیر KDD در دنیای واقعی به صورت خلاصه بیان‌شده و چالش‌هایی که برای تحقیقات و توسعه آینده در سیستم‌های KDD وجود دارد، بررسی شده است.

۴.۱ هدف پژوهش

هدف از انجام این پژوهش آشنایی با روش‌های مختلف تبدیل داده‌های بدون ساختار به داده ساختار یافته و استفاده از الگوریتم‌های یادگیری ماشینی برای کشف الگو از داده‌های ناساختار حسگرهای اینترنت اشیاء می‌باشد.

۵.۱ چشم‌انداز

فصل‌بندی مطالب این پایان‌نامه به صورت زیر طراحی و نگارش شده است.

فصل ۱ از کلیات این پژوهش و چیستی انجام آن برخوردار است؛ بدین گونه که به بیان مسئله و مرور ادبیات مربوط به آن و هدف از این پژوهش پرداخته شد.

فصل۲ با هدف ارائه چارچوب و مبانی نظری این پژوهش جمع‌آوری خواهد شد تا زمینه برای بررسی موضوع مورد مطالعه این پایان‌نامه فراهم شود. بدین‌ترتیب که در ابتدا با انواع دسته‌بندی داده‌ها آشنا خواهیم شد و سپس به بررسی الگوریتم‌های مختلف یادگیری ماشین می‌پردازیم.

فصل۳ در این فصل به بررسی مبانی نظری روش‌های مختلف تبدیل داده‌های بدون ساختار به داده‌های ساختاریافته خواهیم‌پرداخت.

فصل۴ که آخرین فصل پایان‌نامه می‌باشد، به پیاده‌سازی یک روش مطرح‌شده در فصل ؟؟، بر روی داده بدون‌ساختار اینترنت اشیاء خواهدشد.بدین گونه که از طریق مجموعه داده‌های استانداردی که از پایگاه‌های داده‌ی معتبر استخراج شده‌است،این الگوریتم‌ها را پیاده‌سازی خواهیم‌کرد و نتایج خروجی از این پیاده‌سازی‌ها مورد تجزیه و تحلیل قرار خواهد گرفت و در نهایت میزان دقت آن‌ها مورد بحث قرار می‌گیرد.

فصل ۲

مبانی نظری پژوهش

۱.۲ مقدمه

اینترنت اشیاء به عنوان یک پدیده فناورانه پیشرو، با جمع آوری تخصص‌های گوناگون از حوزه‌های متنوع نظیر الکترونیک، کامپیوتر، و آمار، به دنبال فراهم آوردن راهکارهای جامع برای تأمین نیازهای پیچیده امروز جامعه می‌باشد. اینترنت اشیاء را با تقسیم به دو حوزه ی سخت افزار و نرم افزار میتوان مورد مطالعه قرار داد که در بخش نرم افزار اینترنت اشیاء به منظور تحلیل داده های مربوط به آن تنها دو رویکرد سطح بالاتر باید مورد استفاده قرار گیرند در دل رویکردهای تحلیلی دادههای اینترنت اشیاء، پیش پردازش داده ها و روشهای یادگیری ماشین و یادگیری عمیق از الزامات تحلیل .هستند.در این فصل ضمن آشنایی با فناوری اینترنت اشیاء و کاربردهای آن به بررسی روش های مختلف تحلیل دادههای این فناوری می‌پردازیم.

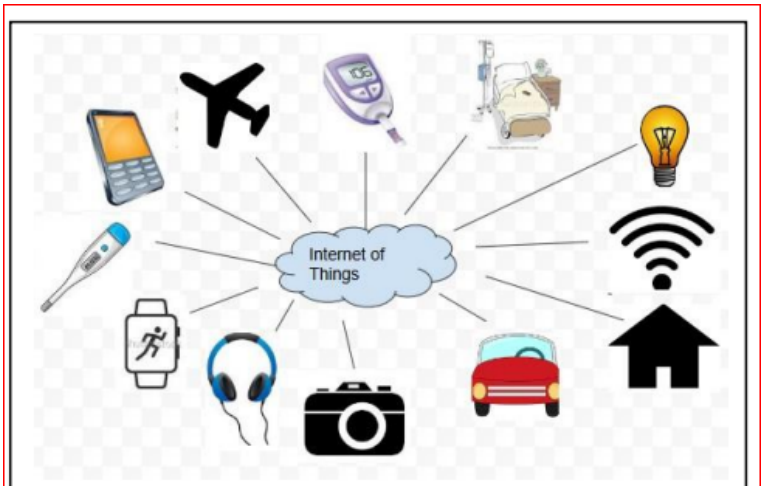
۲.۲ اینترنت اشیاء

۱.۲.۲ تعریف

مفهوم اینترنت اشیاء یا اینترنت چیزها برای اولین بار توسط کوین اشتون در سال ۱۹۹۹ معرفی شد. او جهانی را توصیف کرد که هر چیزی از جمله اشیاء و چیزها،هویت دیجیتالی پیدا کنند و از طریق کامپیوتر و موبایل هوشمند کنترل و مدیریت شوند. این مفهوم برای اولین بار در سال ۲۰۰۵ به طور رسمی توسط اتحادیه بین المللی مخابرات تعریف شده است. اینترنت مفهومی بود که تعامل بین انسان‌ها را مستقل از موقعیت جغرافیایی ممکن ساخت و اینترنت اشیاء این کار را برای اشیاء با هم و انسان و اشیاء انجام می‌دهد. تعریف دیگری از اینترنت اشیا (IoT) به شبکهای از دستگاههای الکترونیکی متصل به هم و به اینترنت گفته می شود که به منظور

جمع آوری، ارسال و پردازش داده های مرتبط با این دستگاه‌ها ایجاد شده است. این دستگاه‌ها می توانند در موارد مختلفی مانند خانه هوشمند، شهر هوشمند، صنعت هوشمند و سلامت هوشمند استفاده شوند .

پنج بخش اصلی در اکوسیستم اینترنت اشیاء شامل اشیاء(سنسورها)، ارتباطات، پلتفرم(جمع آوری داده)، درگاه‌های کاربردی و تحلیل و پردازش داده‌ها می‌باشد که در این پایان‌نامه هدف ما بخش تحلیل و پردازش داده‌ها می‌باشد. عزیزی و امینی (۲۰۲۱)



شکل ۱.۲: سنسورهای اینترنت اشیاء

۲.۲.۲ اهمیت و کاربرد

اینترنت اشیاء با دارابودن قابلیت بسیار بالا برای بهره‌ور نمودن کسب و کارها در حوزه‌های مختلف از جمله صنایع به عنوان انقلاب آتی در فناوری اطلاعات و ارتباطات معرفی شده‌است. این بهره‌وری در زمنیه بروز نوآوری و ارائه قابلیت‌های نو برای کسب‌و کارها است. صنایع مختلف در خصوص اینترنت اشیاء واکنش‌های مختلفی را نشان داده‌اند اما آنچه واضح است این است که اینترنت اشیاء در تمامی کسب‌و کارها و صنایع دارای کاربرد است. این کاربردها در برخی صنایع مانند بهداشت و حوزه سلامت و یا حمل و نقل پیشرفت چشمگیری داشته اما در صنایع دیگر همچون کشاورزی و دامداری در حال توسعه است در واقع تولید داده‌ها بر مبنای اینترنت اشیاء از ارکان اصلی در حوزه مه‌داده‌ها و علم داده‌ها خواهد بود. لذا استفاده از مفاهیم و مدل‌های آماری که در علم داده‌ها مورد استفاده قرار می‌گیرند به خوبی می‌توانند در این‌گونه داده‌ها مورد استفاده قرار گیرند.

۳.۲.۲ چالش ها

چالش های مختلفی در زمینه اکوسیستم اینترنت اشیاء وجود دارد اما به صورت عمده می‌توان این موارد را به ۲ لایه فیزیکی و نرم افزاری تقسیم‌بندی کرد. به دلیل وجود داشتن انواع متنوع و مختلف از سنسورها در این زمینه باعث تولید بی وقفه داده و ایجاد

داده‌های مختلف با ساختارهای متفاوت می‌شود که تمرکز این پایان‌نامه به بررسی روش‌های مختلف تحلیل این داده‌ها و انجام تحلیل بر روی یک نوع داده اینترنت اشیاء می‌باشد.

۳.۲ ساختار داده‌ها

کلان داده‌ها می‌توانند به اشکال مختلف، از جمله داده‌های ساختاریافته و غیرساخت‌یافته مانند داده‌های مالی، فایل‌های متنی، فایل‌های چندرسانه‌ای باشند. برخلاف بسیاری از تحلیل‌های سنتی داده‌ها که توسط سازمان‌ها انجام می‌شود، بیشتر داده‌های بزرگ ماهیتی بدون ساختار یا نیمه ساختار یافته دارند که برای پردازش و تجزیه و تحلیل به تکنیک‌ها و ابزارهای مختلفی نیاز دارد. محیط‌های محاسباتی توزیع‌شده و معماری‌های پردازش انبوه موازی (MPP) که دریافت و تجزیه و تحلیل داده‌های موازی را امکان‌پذیر می‌کنند، رویکرد ترجیحی برای پردازش چنین داده‌های پیچیده هستند.

شکل ۲.۲ چهار نوع ساختار داده را نشان می‌دهد که ۸۰ تا ۹۰ درصد رشد داده‌های آینده از انواع داده‌های غیر ساختاریافته است. اگرچه این چهار نوع ساختار با هم متفاوت هستند اما معمولا با یکدیگر مخلوط می‌شوند.به عنوان مثال یک سیستم مدیریت پایگاه داده رابطه‌ای کلاسیک(RDBMS) ممکن است گزارش‌های تماس را برای یک مرکز تماس پشتیبانی نرم‌افزار ذخیره کند.RDBMS ممکن است ویژگی های تماس‌های پشتیبانی را به عنوان داده‌های ساختار یافته ذخیره کند اما ممکن است علاوه‌بر این موضوع سیستم دارای داده‌های بدون ساختار،شبه یا نیمه ساختاریافته باشد مانند تاریخچه چت مشتری، فایل صوتی مکالمات تلفنی، تصویر رسید مشتری و غیره باشد.

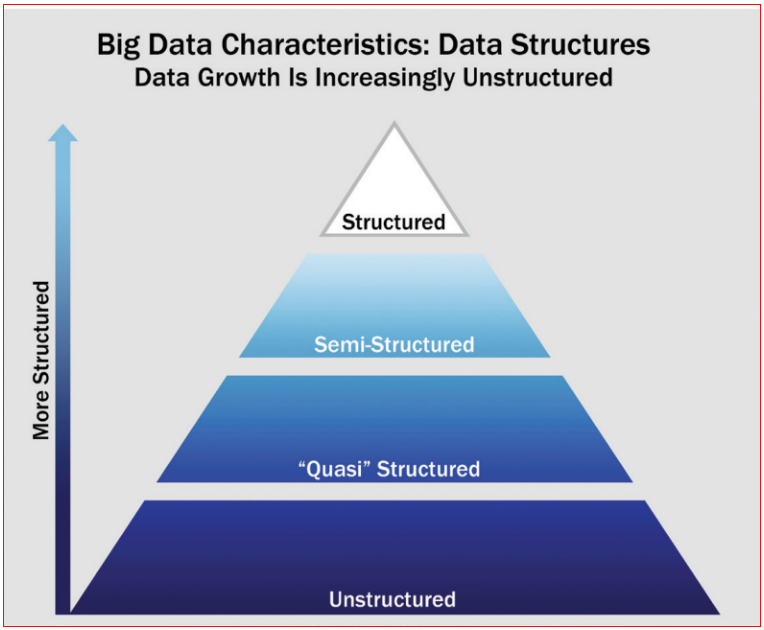
روش‌های شناخته‌شده و مرسوم و متداولی برای تجزیه و تحلیل داده‌های ساختار یافته وجود دارد اما تکنیک‌های متفاوتی برای مقابله چالش‌های تجزیه و تحلیل داده‌های نیمه ساختاریافته، شبه ساختاریافته، و داده‌های بدون ساختار مورد نیاز است. در ادامه به معرفی و بررسی هرکدام از این نوع داده‌ها می‌پردازیم. وایلی (۲۰۱۵)

۱.۳.۲ داده ساختاریافته

داده‌های ساختاریافته [1] داده‌هایی هستند که فرمت مناسبی دارند و به راحتی می‌توان اطلاعات از آن‌ها استخراج کرد. داده‌های ساختاریافته را می‌توان در ستون‌ها و ردیف‌ها ذخیره کرد. این داده‌ها به راحتی توسط ابزارهای داده‌کاوی مورد استفاده قرار می‌گیرد و می‌تواند در سیستم مدیریت پایگاه داده رابطه‌ای (RDBMS) ذخیره شود. عمدتا پایگاه داده رابطه‌ای سنتی برای ذخیره داده‌های ساختاریافته استفاده می‌شود. داده‌های تراکنش، فایل‌های CSV، فایل‌های اکسل از انواع داده‌های ساختار یافته می‌باشند.

ورما و همکاران (۲۰۲۰)

[1] Structured data



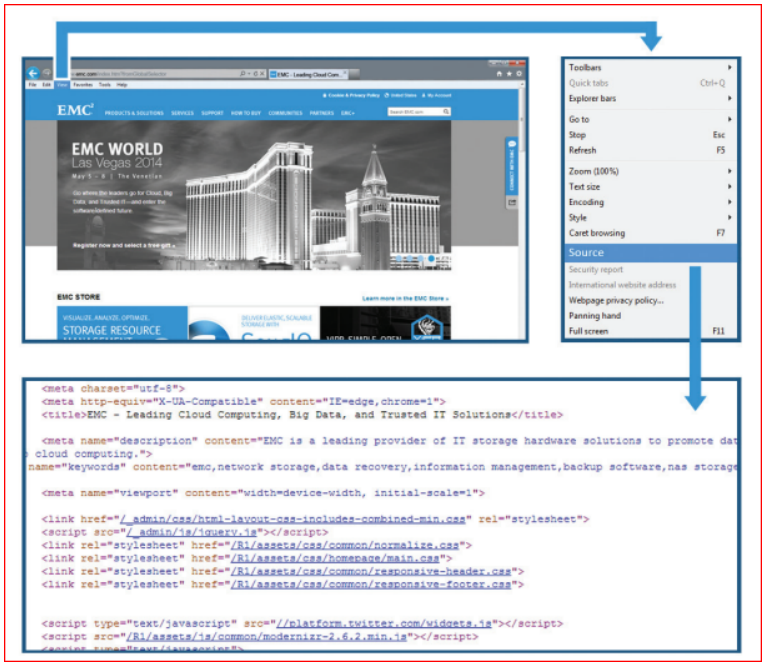
شکل ۲.۲: رشد داده‌ها

SUMMER FOOD SERVICE PROGRAM 1]				
(Data as of August 01, 2011)				
Fiscal Year	Number of Sites	Peak (July) Participation	Meals Served	Total Federal Expenditures 2]
	----------Thousands----------		--Mil.--	---Million $---
1969	1.2	99	2.2	0.3
1970	1.9	227	8.2	1.8
1971	3.2	569	29.0	8.2
1972	6.5	1,080	73.5	21.9
1973	11.2	1,437	65.4	26.6
1974	10.6	1,403	63.6	33.6
1975	12.0	1,785	84.3	50.3
1976	16.0	2,453	104.8	73.4
TQ 3]	22.4	3,455	198.0	88.9
1977	23.7	2,791	170.4	114.4
1978	22.4	2,333	120.3	100.3
1979	23.0	2,126	121.8	108.6
1980	21.6	1,922	108.2	110.1
1981	20.6	1,726	90.3	105.9
1982	14.4	1,397	68.2	87.1
1983	14.9	1,401	71.3	93.4
1984	15.1	1,422	73.8	96.2
1985	16.0	1,462	77.2	111.5
1986	16.1	1,509	77.1	114.7
1987	16.9	1,560	79.9	129.3
1988	17.2	1,577	80.3	133.3
1989	18.5	1,652	86.0	143.8
1990	19.2	1,692	91.2	163.3

شکل ۳.۲: مثال داده ساختاریافته

۲.۳.۲ داده نیمه‌ساختار یافته

داده نیمه ساختار یافته [۲] از ساختار رسمی ندمل داده‌های مربوط به پایگاه داده رابطه‌ای که شامل سطر و ستون هستند پیروی نمی‌کنند. پیشرفت در استفاده از اینترنت حضور داده‌های نیمه‌ساختاری راا افزایش می‌دهد.فایل‌های داده متنی با الگوی قابل تشخیص که تجزیه را امکان‌پذیر میکند مانند فایل‌های داده زبان نشانه‌گذاری توسعه‌یافته [۳] یا فایل‌های نمادگذاری اشیاء در جاوااسکریپت [۴] از این نوع داده‌ها می‌باشند.



شکل ۴.۲: مثال داده نیمه ساختاریافته

۳.۳.۲ داده شبه ساختاریافته

داده‌های شبه ساختاریافته [۵] داده‌های متنی با قالب‌های داده نامنظم که می‌توانند با تلاش، ابزار و زمان قالب‌بندی شوند به عنوان مثال، داده‌های جریان کلیک وب که ممکن است حاوی ناسازگاری در مقادیر و قالب‌های داده باشد از این نوع داده‌ها می‌باشد. شکل ۵.۲ را ببینید.

[2] Semi-structured data
[3] XML
[4] JSON
[5] Quasi-structured data

شکل ۵.۲: مثال داده شبه ساختاریافته

۴.۳.۲ داده بدون ساختار

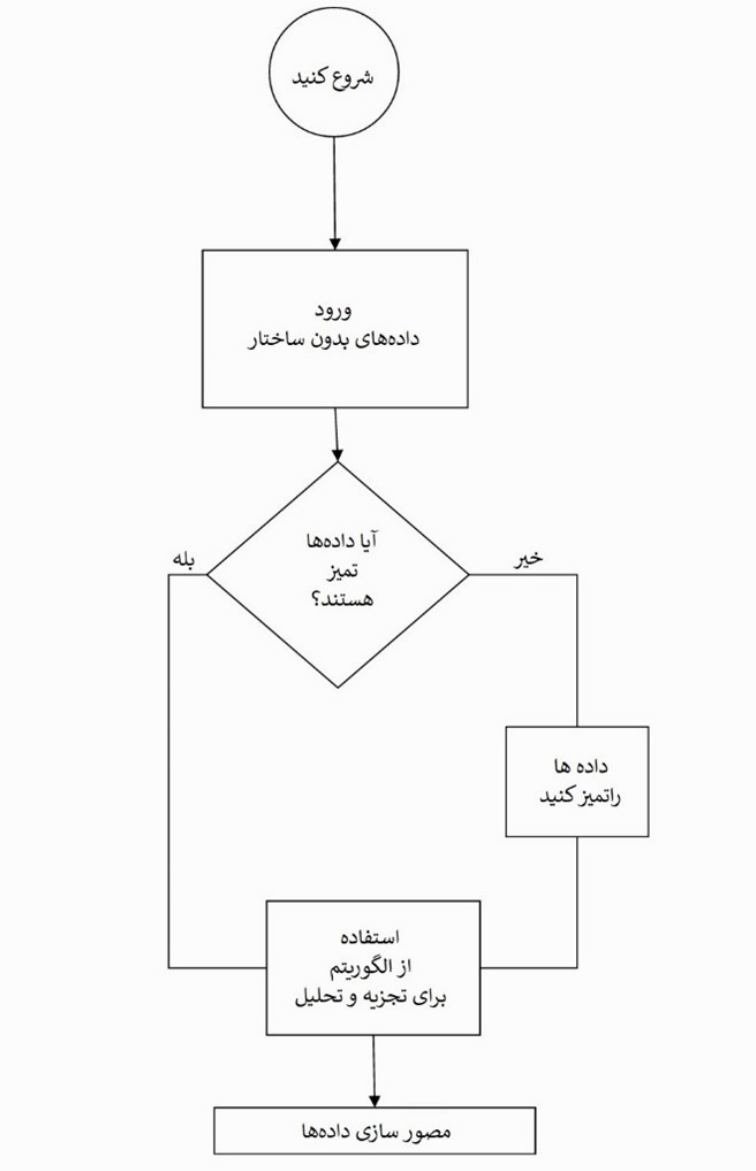
داده بدون ساختار [۶] داده‌هایی هستند که مدل مناسبی ندارند و یا به روش مناسب سازماندهی نشده‌اند.درک و تجزیه و تحلیل این داده‌ها به دلیل اینکه در قالب جدولی نیستند سخت است.انواع این داده‌ها عبارتند از: اسناد متنی، اسناد قابل حمل [۷] تصاویر، صوت و فایل‌های ویدئو باشند.

شکل ۶.۲: ویدئویی در مورد سفر قطب جنوب نمونه‌ای از داده بدون ساختار

[6] Unstructured data
[7] PDF

۴.۲ فرآیند تحلیل داده اینترنت اشیاء

مقدار داده تولید شده توسط سنسورهای اینترنت اشیاء بسیار زیاد است و به طور مداوم تولید می‌شود. داده‌های تولید شده به صورت بدون ساختار هستند و این داده‌ها باید به شکل ساختاری برای استخراج دانش از آن تبدیل شوند.سیستم پیشنهادی با استفاده از الگوریتم های یادگیری ماشین و یادگیری عمیق، داده‌های بدون ساختار را به داده‌های ساختاریافته تبدیل می‌کند. شکل ۷.۲ نمودار فرآیند سیستم پیشنهادی جهت تحلیل داده بدون ساختار است. ابتدا داده‌ها را از منابع مد نظر وارد می‌کنیم، سپس باید پاکسازی داده‌ها برای هر مجموعه داده انجام شود. پاکسازی داده‌ها شامل استاندارد سازی داده‌ها، بررسی موارد گمشده و داده‌های دورافتاده و مرتب کردن داده‌ها می‌باشد. سپس داده‌ها را به داده‌های آموزشی و آزمایشی تقسیم می‌کنیم. داده‌های آموزشی باید با دقت طبقه‌بندی شوند، مدل نباید بیش برازش یا کم برازش شود. با اعمال الگوریتم های مختلف دقت مدل را با پارامترهای مختلف بررسی کنید تا دقت بالایی به‌دست آید. و در انتها تجسم داده‌ها برای استخراج دانش از داده‌های خام مانند نمودار میله‌ای، نمودار پراکندگی و غیره انجام می‌شود.



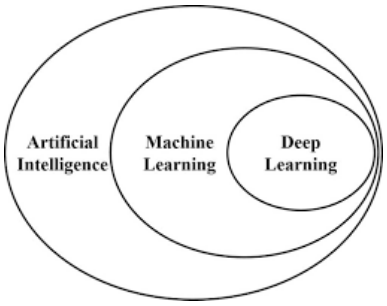
شکل ۷.۲: نمودار روند تحلیل داده بدون ساختار

۵.۲ هوش مصنوعی

هوش مصنوعی تکنیکی است که رایانه‌ها یا ماشین‌ها را قادر می‌سازد تا توانایی انسان برای انجام، رفتار یا عملکرد را تقلید کنند. هوش مصنوعی می تواند از داده‌های گذشته بیاموزد و قادر به تشخیص الگوها و رفتار است. این امکان را برای ماشین فراهم می کند که از تجربه درس بگیرد و رفتار خود را تحمل کند. اصطلاح هوش مصنوعی در سال ۱۹۵۶ توسط جان مک کارتی [۸] ابداع

[8] McCarthy

شد. زمینه های بین رشته‌ای زیادی تحت هوش مصنوعی وجود دارد. در شکل ۸.۲ موقعیت یادگیری ماشین [۹] و یادگیری عمیق [۱۰] را در ارتباط با هوش مصنوعی نشان می‌دهد.



شکل ۸.۲: ارتباط یادگیری ماشین و یادگیری عمیق

اصطلاح هوش مصنوعی بسیار گسترده است در حالی که یادگیری ماشین شاخه‌ای از هوش مصنوعی است. یادگیری ماشین اولین بار در سال ۱۹۵۹ توسط آرتور ساموئل ابداع شد که از روش های آماری استفاده می کند تا ماشین ها را قادر سازد عملکرد و رفتار خود را از طریق تجربه بهبود بخشند. با توجه به شکل ۸.۲، یادگیری عمیق زیر شاخه ای از یادگیری ماشین است. این اصطلاح در سال ۱۹۸۶ توسط رینا دچتر به جامعه یادگیری ماشین و در سال ۲۰۰۰ توسط ایگور آیزنبرگ به شبکه عصبی مصنوعی معرفی شد. آلاسکار و صبا (۲۰۲۱) در ادامه به مرور کلی بر مفاهیم اساسی یادگیری عمیق و یادگیری ماشین می‌پردازیم.

[9] Machine Learning
[10] Deep Learning

کتاب‌نامه

Abdullah, M. F. and Ahmad, K. (2013), "The mapping process of unstructured data to structured data," pp. 151–155.

Abidin, S. Z., Idris, N. M., and Husain, A. H. (2010), "Extraction and classification of unstructured data in WebPages for structured multimedia database via XML," pp. 44–49.

Alaskar, H. and Saba, T. (2021), "Machine learning and deep learning: a comparative review," *Proceedings of Integrated Intelligence Enable Networks and Computing: IIENC 2020*, 143–150.

Artiles, J., Penas, A., and Verdejo, F. (2004), "Word Sense Disambiguation based on term to term similarity in a context space," pp. 58–63.

Azizi, S. and Amini, Y. (2021), *Big Data and Internet of Things: Principles and Tools (in Persian)* : .

Gandhi, K. and Madia, N. (2016), "Information extraction from unstructured data using RDF," pp. 1–6.

Mishra, S. and Misra, A. (2017), "Structured and unstructured big data analytics," pp. 740–746.

Rusu, O., Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M., and Marinescu, V. (2013), "Converting unstructured and semi-structured data into knowledge," pp. 1–4.

Sambrekar, K., Rajpurohit, V. S., and Joshi, J. (2018), "A proposed technique for conversion of unstructured Agro-data to semi-structured or structured data," pp. 1–5.

Services, E. E. (2015), *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, Wiley.

Tekli, J. (2016), "An overview on xml semantic disambiguation from unstructured text to semistructured data: Background, applications, and ongoing challenges," *IEEE Transactions on Knowledge and Data Engineering*, 28, 1383–1407, special issue.

Verma, S., Jain, K., and Prakash, C. (2020), "An Unstructured to Structured Data Conversion using Machine Learning Algorithm in Internet of Things (IoT)," Proceedings of the International Conference on Innovative Computing & Communications (ICICC).

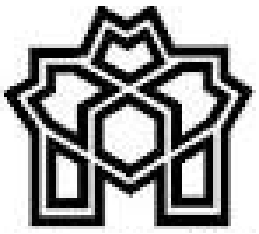
Abstract

In recent years, the Internet of Things (IoT) has received much attention from researchers. This technology receives information from its surroundings with the help of sensors and communicates with networks and other devices and sends the collected data to central systems. This technology is one of the important and powerful tools used to improve security, quality of life and productivity, and it is used in various industries such as health and medicine, agriculture, automotive, weather, etc. as a tool to improve efficiency and reduce costs. By using the Internet of Things, this information can be used to analyze and predict behaviors that may occur in the future, and as a result, it is possible to predict problems and solutions for them.

One of the challenges in the field of Internet of Things is data processing and its applications. This challenge has arisen due to the existence of unstructured data created through sensors, social network and web data collected by companies and organizations. Considering the high number of unstructured data available in different fields, if the unstructured data is converted into structured data, it gives us the possibility to use them in a better way, to identify the patterns that exist in them and to predict the behavior. Let's make better decisions in the future.

In this thesis, we investigate various processes of converting unstructured data into structured data and using machine learning methods to discover patterns from unstructured data of the Internet of Things.

Keywords: *Classification, IoT Sensors, Machine Learning Algorithm, Structed Data, Unstructed Data.*



Allameh Tabataba'i University
Faculty of Statistics, Mathematics and Computer
Department of Statistics

Thesis Submitted in Partial Fulfillment for the Degree of Master of Science (MSC) in the Data Science

Title

A Conversion of Unstructured to Structured Data Using Machine Learning Algorithm in Internet of Things (IoT)

Supervisor

Dr. Farzad Eskandari

Advisor

Dr. Mohammad Bameni Moghadam

Examiner

Dr. Reza Pourtaheri

By

Sajjad Hajizadeh

Fall 2024