بسم الله الرحمن الرحيم

دانشگاه پیام نور

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه

برای دریافت درجه کارشناسی ارشد

در رشته مهندسی نرم افزار کامپیوتر

عنوان پایان نامه:

دسته بندی موضوعی متون فارسی بر اساس روش قواعد انجمنی

استاد راهنما:

دکتر سید امیر حسن منجمی

استاد مشاور:

دکتر سید سعید آیت

نگارش:

سید محمد حسین احمدی

دی ماه ۱۳۹۰

دانشگاه پیام نور

بسمه تعالی

تصویب پایان نامه

پایان نامه تحت عنوان: دسته بندی موضوعی متون فارسی بر اساس روش قواعد انجمنی که در پیام نور مرکز شمیرانات تهیه و به هیأت داوران ارائه گردیده است مورد تأیید می باشد.

تاریخ دفاع : ۹۰/۱۱/۳۰ نمره : ۱۹,۲۵ درجه ارزشیابی:

اعضای هیأت داوران:

	نام و نام خانوادگی	هیات داوران	مرتبه علمی	امضاء
۱-	سید امیر حسن منجمی	استاد راهنما	دانشیار	
۲-	سید سعید آیت	استاد مشاور	استادیار	
۳-	داود کریم زادگان مقدم	استاد داور	استادیار	
۴-	محمد هادی معظم	نماینده گروه علمی	استادیار	
۵-	محمد هادی معظم	نماینده تحصیلات تکمیلی	استادیار	

ان الحسین مصباح الهدی و سفینه النجاه

تقدیم به سید و سالار شهیدان حضرت اباعبدالله الحسین
که با خون خود درخت اسلام را تا قیامت آبیاری کرد

سلام و صلوات بی پایان الهی بر وجود مقدسش باد

از زحمات صادقانه شادروان مرحوم مادرم (که رحمت و مغفرت الهی بر او باد)

پشتیبانی‌های دلسوزانه پدرم

و حمایت‌های صبورانه همسر عزیزم و فرزندان خوبم سید علی و سید محمد

و کمک‌ها و راهنمایی‌های اساتید محترم جناب آقای دکتر منجمنی و جناب آقای دکتر آیت

بی نهایت سپاسگذارم

چکیده:

برای دسته بندی متن از تکنیک‌های استخراج اطلاعات، پردازش زبان طبیعی و یادگیری ماشین به طور وسیع استفاده می‌شود. به طور کلی هدف یک دسته بند متون، دسته بندی اسناد در قالب تعداد معینی از دسته های از پیش تعیین شده می‌باشد. هر سند می‌تواند در یک، چند و یا هیچ دسته ای قرار بگیرد. در مورد هر سند به این سؤال پاسخ داده خواهد شد که این سند در کدام یک از دسته‌ها قرار می‌گیرد. این موضوع می‌تواند در قالب یک یادگیری خودکار قرار گیرد تا با استفاده از آن بتوان هر سند را به طور خودکار به دسته ای نسبت داد.

در این تحقیق، از روش دسته بندی بر مبنای قواعد انجمنی که از روی فرایند کاوش الگوهای مکرر مجموعه داده های آموزشی تولید شده‌اند، برای دسته بندی متون فارسی استفاده می‌شود. این فرآیند با فرآیندی که در داده کاوی داده های بزرگ پایگاه داده‌ها استفاده می‌شود یکسان می‌باشد. یکی از مهم‌ترین الگوریتم‌هایی که برای تولید قواعد انجمنی بکار می‌رود الگوریتم Apriori می‌باشد. در این تحقیق از الگوریتم CBA[1] که برای این کاربرد مناسب تشخیص داده شد، استفاده شده است. پیکره‌ی مورد استفاده برای انجام آزمایشات، پیکره متون فارسی همشهری ۲ می‌باشد، که مقالات آن کامل و حجیم بوده و به ۸ دسته خبری تقسیم شده‌اند. پس از انجام پیش پردازش‌های لازم بر روی پیکره همشهری ۲ و تبدیل آن به فرمت مناسب، کلمات کلیدی متون آموزشی با استفاده از تکنیک TFIDF[2] و نرم افزار قدرتمند Weka استخراج می‌شوند. سپس قواعد انجمنی دسته بندی از روی داده های آموزشی (کلمات کلیدی بدست آمده از مرحله قبل)، با استفاده از نرم افزار DMII CBA که الگوریتم CBA را پیاده سازی کرده است استخراج شده و دسته بند نهایی تولید و ذخیره می‌شود. در ادامه از این دسته بند برای دسته بندی متون آزمایشی استفاده می‌شود. آزمایشات انجام شده و ارزیابی آن‌ها نشان می‌دهد با افزایش تعداد متون آزمایشی و انتخاب مناسب کلمات کلیدی مرتبط با موضوع متن، دقت دسته بند به طور چشمگیری افزایش می‌یابد.

کلمات کلیدی:

طبقه بندی متون فارسی، دسته بندی متون فارسی، قواعد انجمنی، قواعد دسته بندی، استخراج کلمات کلیدی، داده کاوی متن

[1] Classification Based on Association

[2] Term Frequency - Inverse Document Frequency

فهرست مطالب

مقدمه **۱۲**

فصل ۱ : مفاهیم اولیه **۱۵**

۱-۱ مقدمه ۱۵

۱-۳ سابقه و ضرورت ۱۷

۱-٤ روش انجام تحقیق ۱۷

۱-٥ داده کاوي چیست؟ ۲۰

۱-٦ متن کاوي چیست؟ ۲۱

۱-۷ تعریف دسته بندي متن ۲۳

۱-۹ بعضي از روشهاي دسته بندي متن ۲٤

۱-۹-۱ درختهاي تصمیم گیري ۲٤

۱-۹-۲ ماشین بردار پشتیبان ۲٥

۱-۹-۳ k- نزدیکـترین همسایه ۲٦

۱-۹-٤ شبکه هاي عصبي ۲٦

۱-۹-٥ بیزین ساده ۳۲

۱-۹-٦ آنالیز معنایي پنهان (LSA) ۳۳

۱-۹-۷ آنالیز معنایي پنهان احتمالاتي (PLSA) ۳٦

فصل ۲ : بازیابی اطلاعات و استخراج کلمات کلیدی **۳۹**

۲-۱ مقدمه ۳۹

۲-۲ پیچیدگي‌هاي پردازش زبان فارسي ۳۹

۲-۳ انواع پردازش‌هاي زبان فارسي ٤۰

۲-۳-۱ پردازش لغوي ٤۰

۲-۳-۲ پردازش ساخت واژي ٤۰

۲-۳-۳ منابع زباني ٤۱

۲-٤ قالبهاي گوناگون پرونده هاي کامپیوتري ٤۲

۲-٤-۱ استاندارد خط در کامپیوتر ٤۳

۲-٥ بازیابي اطلاعات ٤٥

۲-٥-۱ تئوري لان ٤۷

۲-۵-۲ قانون ZIPF ٤٩
۲-٦ استخراج کلمات کليدي ٥١
۲-۷ تقسيم بندي روش‌ها ٥٢
۲-۷-۱ تقسيم بندي ابزاري ٥٢
۲-۷-۲ تقسيم بندي تکنيکي ٥٣
۲-۸ مراحل استخراج کلمات کليدي ٥٣
۲-۸-۱ حذف کلمات عمومي ٥٤
۲-۸-۲ ريشه يابي ٥٧
۲-۸-۳ وزن دهي به کلمات ٦٠
۲-۹ ارزيابي کلمات کليدي ٦٥
۲-۹-۱ روشهاي تشخيص نوع کلمات بکار رفته ٦٥
۲-۹-۲ نحوة ارزيابي کلمات کليدي ٦٦
۲-۹-۳ داوري مبتني بر سيستم‌هاي بازيابي اطلاعات ٦٦
۲-۱۰ بررسي کارهاي انجام شده ٦٨

فصل ۳ : دسته بندی براساس قواعد انجمنی ۷۲

۳-۱ مقدمه: ۷۲
۳-۲ مفاهيم اوليه قواعد انجمني ۷٤
۳-۳ معرفي الگوريتم Apriori ۷۷
۳-۳-۱ توليد مجموعه اقلام مکرر ۷۸
۳-۳-۲ توليد قواعد انجمني ۸۲
۳-٤ قالبهاي داده اي براي کاوش قواعد انجمني ۸۳
۳-٥ دسته بندي با استفاده از قواعد انجمني ۸٤
۳-٥-۱ دسته بند انجمني ۸٥
۳-٥-۲ شماي دسته بندي با مجموعه قواعد ۸۷
۳-٦ الگوريتم دسته بندي انجمني CBA ۸۹
۳-٦-۱ الگوريتم توليد کننده قواعد (CBA-RG) ۹۱
۳-٦-۲ الگوريتم سازنده دسته بند (CBA-CB) ۹٤
۳-٦-۳ رويکرد ساده دسته بند M۱ ..... ۹٥

فصل ۴ : دسته بندی متون فارسی بر اساس روش پیشنهادی ۹۷

٤-١ معماري كلي ٩٧
٤-٢ معرفي پيكره همشهري ٢ ..... ٩٧
٤-٣ استخراج كلمات كليدي متون ٩٩
٤-٣-١ كلمات كليدي ٩٩
٤-٣-٢ مراحل استخراج كلمات كليدي ٩٩
٤-٣-٣ حذف كلمات عمومي ١٠٠
٤-٣-٤ وزن دهي به كلمات با تكنيک TFIDF ١٠٠
٤-٣-٥ استفاده از نرم افزار Weka براي استخراج كلمات كليدي ١٠١
٤-٤ ايجاد دسته بند انجمني با استفاده از نرم افزار DMII CBA ١٠٢
فصل ٥ : ارزیابی و نتیجه گیری **١٠٤**
٥-١ محاسبه دقت دسته بند ١٠٤
٥-٢ ارزيابي و آزمايشات ١٠٥
٥-٣ نتيجه گيري و پيشنهادات ١٠٧
فهرست منابع **١٠٨**
پیوست الف: مقالات مستخرج از پایان نامه **١١٥**
پیوست ب: لغت نامه فارسی به انگلیسی **١١٦**
پیوست ج: لغت نامه انگلیسی به فارسی **١٢٧**

فهرست جداول

جدول ۲-۱ نتیجه عملی از قانون ZIPF روی پیکره Tom sawyer .....۵۰.....
جدول ۲-۲ لیست کلمات عمومی (حروف پر تکرار) .....۵۵.....
جدول ۲-۳ لیست کلمات عمومی (افعال پر تکرار) .....۵۵.....
جدول ۲-۴ معیار های مشابهت .....۶۴.....
جدول ۲-۵ جدول اشتراک اسناد مرتبط و بازیابی شده .....۶۷.....
جدول ۵-۱ ماتریس کانفیوژن .....۱۰۴.....
جدول ۵-۲ نتایج آزمایشهای انجام شده برای دسته بندی متون فارسی پیکره‌ی همشهری با استفاده از الگوریتمCBA .....۱۰۵.....

فهرست شکل‌ها

شکل ۱-۱ معماری کلی شبکه‌ی ۴CC .....۲۸.....
شکل ۱-۲ الگوریتم آموزش شبکه ۴CC .....۳۰.....
شکل ۲-۱ سهمی ارتباط بین فراوانی رتبه کلمات در متون .....۴۷.....
شکل ۲-۲ مراحل استخراج کلمات کلیدی .....۵۴.....
شکل ۲-۳ نتایج ارزیابی نمایه ساز سینا .....۷۱.....
شکل ۳-۱ الگوریتم Apriori .....۸۱.....
شکل ۳-۲ تابع Candidate-gen() .....۸۱.....
شکل ۳-۳ الگوریتم تولید قواعد انجمنی .....۸۳.....
شکل ۳-۴ الگوریتم CBA-RG .....۹۳.....
شکل ۳-۵ الگوریتم CBA-CB .....۹۴.....
شکل ۴-۱ معماری کلی دسته بندی انجمنی – آموزش و آزمایش .....۹۸.....
شکل ۴-۲ بلوک دیاگرام مراحل انجام دسته بندی متون فارسی با استفاده از قواعد انجمنی ....۱۰۳....

فهرست نمودارها

نمودار ۵-۱ رابطه دقت دسته بند با تعداد متون .....۱۰۶.....

نمودار ۵-۲ رابطه دقت دسته بند با تعداد کلمات کلیدی .....۱۰۶.....

مقدمه:

برای دسته بندی متن از تکنیک‌های استخراج اطلاعات، پردازش زبان طبیعی و یادگیری ماشین به طور وسیع استفاده می‌شود. به طور کلی هدف یک دسته بند متون، دسته بندی اسناد در قالب تعداد معینی از دسته های از پیش تعیین شده می‌باشد. هر سند می‌تواند در یک، چند و یا هیچ دسته ای قرار بگیرد. در مورد هر سند به این سؤال پاسخ داده خواهد شد که این سند در کدام یک از دسته‌ها قرار می‌گیرد. این موضوع می‌تواند در قالب یک یادگیری خودکار قرار گیرد تا با استفاده از آن بتوان هر سند را به طور خودکار به دسته ای نسبت داد.

در سالهای گذشته به طور وسیعی از روشهای آماری و یادگیری ماشین برای دسته بندی متون استفاده شده که از آن جمله می‌توان به روشهای درخت تصمیم گیری، K- نزدیک‌ترین همسایه، استنتاج قواعد، شبکه های عصبی و ماشین بردار پشتیبانی اشاره کرد [Markof۲۰۰۷,Sebastiani۲۰۰۲]. مدل جدیدی که اخیراً مطرح شده استفاده از قواعد انجمنی و ترکیب آن با قواعد دسته بندی[۱] و ایجاد مدل جدیدی با عنوان دسته بندی انجمنی[۲] [Yoon۲۰۰۷,Chen۲۰۰۵] و استفاده از آن برای دسته بندی متون می‌باشد. قواعد انجمنی از روی فرایند کاوش الگوهای مکرر مجموعه داده های آموزشی تولید می‌شوند. الگوهای مکرر در متون، همان کلمات کلیدی استخراج شده می‌باشند. از تکنیک‌های متعددی برای استخراج کلمات کلیدی استفاده می‌شود که یکی از مهم‌ترین آن‌ها که در این تحقیق از آن استفاده شده است، TFIDF می‌باشد [Sebastiani۲۰۰۲]. در تحقیقات نادری از روش دسته بندی انجمنی برای دسته بندی متون انگلیسی استفاده شده [Yoon۲۰۰۷,Chen۲۰۰۵,Kamruzzaman۲۰۰۴] ولی با توجه به بررسی‌های انجام شده، تاکنون از این روش برای دسته بندی متون فارسی استفاده نشده است.

دسته بند انجمنی[۳] دارای ویژگی‌های مطلوبی از جمله دقت دسته بندی مناسب و امکان تفسیر عالی می‌باشد. البته استفاده از این مدل در حوزه دسته بندی متون، دارای معایبی نیز می‌باشد که از آن جمله می‌توان به افزایش بعد فضای داده‌ها (مثل کلمات کلیدی) و تولید حجم زیادی از قواعد دسته بندی اشاره

[۱] Classification Rules

[۲] Associative Classification

[۳] Associative Classifier

کرد که باعث می‌شود فرایند آموزش بسیار طولانی شود [Yoon۲۰۰۷]. برای مقابله با این معایب نیز تکنیک‌های کارآمدی بکار رفته است [Sebastiani۲۰۰۲]. مدل دسته بندی انجمنی دارای سه فاز می‌باشد:

۱ - تولید تمام *قواعد انجمنی دسته‌ها*[۱] (CAR) با توجه به آستانه *حداقل پشتیبان*[۲] و با استفاده از الگوریتمی مثل Apriori. این قواعد به صورت " *برچسب دسته ←مشخصه‌ها* " می‌باشند.

۲ - ۲- ارزیابی کیفیت قواعد تولید شده در مرحله اول و انتخاب قواعد مفید با آستانه حداقل اطمینان[۳] و حذف قواعد زاید و کم تاثیر و در نهایت تشکیل دسته بند با قواعد انتخاب شده.

۳ - رتبه بندی دسته بند از تمام قواعد تشکیل دهنده داده جدید و انتخاب تمام یا تعدادی از قواعد مناسب برای پیش بینی کردن دستهٔ داده جدید و در نهایت نسبت دادن یک برچسب به آن [Liu۱۹۹۸].

در این تحقیق، از روش دسته بندی بر مبنای قواعد انجمنی که از روی فرایند کاوش الگوهای مکرر مجموعه داده های آموزشی تولید شده‌اند، برای دسته بندی متون فارسی استفاده می‌شود. این فرآیند با فرآیندی که در داده کاوی داده های بزرگ پایگاه داده‌ها استفاده می‌شود یکسان می‌باشد. یکی از مهم‌ترین الگوریتم‌هایی که برای تولید قواعد انجمنی بکار می‌رود الگوریتم Apriori می‌باشد. در این تحقیق از الگوریتم CBA[۴] که بر مبنای الگوریتم فوق بوده و برای این کاربرد مناسب تشخیص داده شد، استفاده شده است.

در فصل اول، ابتدا به مفاهیم اولیه اشاره شده و در ادامه به اختصار برخی از روشهای مشهور که برای دسته بندی متون انگلیسی و فارسی به کار رفته بررسی شده است.

در فصل دوم، در ابتدا مشکلات و پیچیدگی‌های پردازش زبان فارسی مطرح شده و بعضی از پیکره های موجود برای زبان فارسی معرفی شده است. در ادامه این فصل، به مفاهیم بازیابی اطلاعات و روش‌ها و تکنیک‌های مشهور برای استخراج و وزن دهی کلمات کلیدی متون اشاره شده است.

در فصل سوم، در بخشهای ۳-۲ الی ۳-۵، مفاهیم اصلی و اصطلاحاتی که در طول این تحقیق از آن‌ها استفاده می‌شود مطرح و تعریف شده‌اند. در بخش ۳-۶ ابتدا الگویتم اصلی CBA مطرح و بررسی

---

[۱] Class Association Rule

[۲] Minimum Support

[۳] Minimum Confidence

[۴] Classification Based on Association

اجمالی شده است. این الگوریتم شامل دو زیر الگوریتم می‌باشد که اولین آن‌ها قواعد انجمنی را تولید می‌کند و دیگری با اولویت بندی قواعد، دسته بند را ایجاد می‌کند. در ادامه همین بخش، در ابتدا به معماری کلی طرح اشاره شده، سپس نحوه آماده کردن متون پایگاه داده همشهری ۲ به عنوان پیکره‌ی مورد استفاده در این تحقیق، بیان شده است. کلمات کلیدی متون همشهری ۲ نیز با استفاده از تکنیک TFIDF و نرم افزار قدرتمند Weka [Witten۲۰۰۵] استخراج شده است. پس از انجام پردازش‌های لازم بر روی داده های بدست آمده (کلمات کلیدی استخراج شده از متون همشهری ۲) از مرحله قبل، قواعد دسته بندی با استفاده از نرم افزار DMII CBA(که از الگوریتم CBA استفاده می‌کند) استخراج شده و دسته بند بدست آمده از داده های آموزشی، تولید و ذخیره می‌شوند. در انتهای بخش، به مطالعه و ارزیابی دقت دسته بند با انجام آزمایش بر روی داده های آزمایشی پرداخته شده و در انتها نتیجه گیری می‌شود. آزمایشات انجام شده نشان می‌دهد دقت دسته بند با افزایش تعداد متون آزمایشی و انتخاب مناسب کلمات کلیدی به طور چشمگیری افزایش می‌یابد.

فصل ۱ : مفاهیم اولیه

۱-۱ مقدمه

روشهای زیادی برای دسته بندی متون وجود دارد که هر کدام مزایا و معایب خاص خود را دارند. بیشتر این روش‌ها مبتنی بر شیوه های شناخته شده در یادگیری خودکار هستند که در مسائل مختلفی به کار می‌روند، اما برخی به طور خاص برای حل مسئله دسته بندی و یا موارد مشابه بکار می‌روند. دسته بندی متن[۱] در واقع زیر شاخه ای از *متن کاوی*[۲] می‌باشد و متن کاوی هم زیر شاخه ای از علم *داده کاوی*[۳] می‌باشد. برای دسته بندی متن از تکنیک‌های *استخراج اطلاعات*[۴]، *پردازش زبان طبیعی*[۵] و *یادگیری ماشین*[۶] به طور وسیع استفاده می‌شود [Xia۲۰۰۷].

در این فصل مفاهیمی مانند داده کاوی، متن کاوی، بازیابی اطلاعات[۷]، استخراج اطلاعات، پردازش زبان طبیعی و دسته بندی متن تعریف شده و در ادامه به اختصار تعدادی از روشهای مشهور شامل درختهای تصمیم گیری[۸]، ماشینهای بردار پشتیبان[۹]، شبکه های عصبی[۱۰]، بیزین ساده[۱۱] و آنالیز معنایی پنهان[۱۲] که برای دسته بندی متون انگلیسی و فارسی به کار رفته است بررسی می‌شوند.

---

[۱] Text Classification

[۲] Text Mining

[۳] Data Mining

[۴] Information Extraction

[۵] Natural Language Processing

[۶] Machine Learning

[۷] Information Retrieval

[۸] Decision Trees

[۹] Support Vector Machines

[۱۰] Neural Networks

[۱۱] Naïve Bayesian

[۱۲] Latent Semantic Analysis

۱-۲ تعریف مسأله

با گسترش اینترنت و رشد سریع و روز افزون متون الکترونیکی و اطلاعات برخط، دسته بندی متون به یکی از ابزارهای کلیدی و مهم برای سازماندهی و مدیریت داده های متنی تبدیل شده است، که در کاربردهایی چون موتورهای جستجوی وب، ایجاد فهرست موضوعی و دسته بندی اخبار، نامه های الکترونیکی، اطلاعات برخط، کتاب‌ها و مقالات الکترونیکی، فیلتر کردن متون، و همچنین یافتن اطلاعات مورد علاقه کاربران در وب و تسهیل و هدایت جستجوی کاربران در میان ابرمتنها مورد استفاده قرار می‌گیرد [Sebastiani۲۰۰۲]. دسته بندی متون به صورت دستی به نیروی متخصص، هزینه و زمان زیاد نیاز دارد، علاوه بر اینکه رویه ای تکراری برای متون جدید خواهد بود. از اینرو دسته بندی خودکار متون بر اساس روش‌ها و تکنیک‌های بهینه، ضمن اینکه مناسب و کم هزینه خواهد بود، تدریجاً به یک مسئله ضروری تبدیل گشته است.

دسته بندی متن در واقع زیر شاخه ای از *متن کاوی* می‌باشد و متن کاوی هم زیر شاخه ای از علم *داده کاوی* می‌باشد. برای دسته بندی متن از تکنیک‌های *استخراج اطلاعات*، *پردازش زبان طبیعی* و *یادگیری ماشین* به طور وسیع استفاده می‌شود [Xia۲۰۰۷] .به طور کلی هدف یک دسته بند متن، دسته بندی اسناد در قالب تعداد معینی از دسته های از پیش تعیین شده می‌باشد. هر سند می‌تواند در یک، چند و یا هیچ دسته ای قرار بگیرد. در مورد هر سند به این سؤال پاسخ داده خواهد شد که این سند در کدام دسته (یا دسته‌ها) قرار می‌گیرد. این موضوع می‌تواند در قالب یک یادگیری خودکار حل شود تا با استفاده از آن بتوان هر سند را به طور خودکار به دسته ای نسبت داد [Chen۲۰۰۵].

در این پروژه، از روش دسته بندی بر مبنای قواعد انجمنی[۱] که از روی فرایند کاوش الگوهای مکرر[۲] مجموعه داده های آموزشی تولید شده‌اند، استفاده می‌شود. استفاده از قواعد انجمنی در زمینه دسته بندی، نسبت به سایر روشهای دسته بندی مثل بیزین ساده، ماشینهای بردار پشتیبان و k- نزدیک‌ترین همسایه[۳] دارای سابقه و تاریخچه کوتاه تری می‌باشد و روش جدیدتری است [Yoon۲۰۰۷]. سایر مزایا و معایب این روش در قسمت روش انجام تحقیق مورد اشاره واقع شده است.

---

[۱] Association Rules

[۲] Frequent Pattern Mining

[۳] K-Nearest Neighbor

۱-۳ سابقه و ضرورت

سابقه دسته بندی متون الکترونیکی به دهۀ ۱۹۶۰ میلادی باز می‌گردد ولی با گسترش اینترنت در دهۀ ۱۹۹۰ میلادی و به تبع آن گسترش سریع متون الکترونیکی با ساختارها و زبانهای متفاوت، توجه بسیاری از دانشمندان و محققان علوم کامپیوتر به استفاده از روش‌ها و تکنیک‌های بهینه و سریع جلب شد و هم اکنون نیز تحقیق در این زمینه در راستای افزایش سرعت و دقت روش‌ها همچنان ادامه دارد [Sebastiani۲۰۰۲].

تحقیق در زمینه های مختلف متن کاوی بر روی متون فارسی، خصوصاً دسته بندی متن، به طور جدی از اواخر دهۀ ۷۰ و اوایل دهۀ ۸۰ شمسی آغاز شده است و تاکنون نیز کارهای نسبتاً خوبی در این زمینه خصوصاً در دانشگاههای تهران، صنعتی امیر کبیر، علم و صنعت و صنعتی شریف انجام شده است [عرب سرخی ۱۳۸۵، امامی ۱۳۸۵، نقیبی ۱۳۸۷، شمس فرد ۱۳۸۵، محمدی ۱۳۸۸، پور حسن ۱۳۸۷]. با این وجود کار در زمینه متن کاوی فارسی و بکارگیری و تطبیق الگوریتم‌های جدید و کار آمد با زبان فارسی همچنان ادامه دارد. طبق بررسی‌های به عمل آمده تاکنون از روش قواعد انجمنی برای دسته بندی متون فارسی استفاده نشده و در این پروژه برای اولین بار از این روش استفاده خواهد شد. البته در زبان انگلیسی از این روش استفاده شده که نتایج بسیار خوب و قابل قبولی به همراه داشته است [Yoon۲۰۰۷,Chen۲۰۰۵].

۱-۴ روش انجام تحقیق

روشهای زیادی برای دسته بندی متون وجود دارد که هر کدام مزایا و معایب خاص خود را دارند. بیشتر این روش‌ها مبتنی بر شیوه های شناخته شده در یادگیری خودکار هستند که در مسائل مختلفی به کار می‌روند، اما برخی به طور خاص برای حل مسئله دسته بندی و یا موارد مشابه بکار می‌روند. تعدادی از روشهای مشهور شامل: درختهای تصمیم گیری[۱]، دسته بندی بر مبنای قواعد استنتاجی[۲]، ماشینهای بردار پشتیبان، k- نزدیک‌ترین همسایه، شبکه های عصبی[۳] و بیزین ساده می‌باشند [Sebastiani۲۰۰۲].

---

[۱] Decision Trees

[۲] Inductive Rules

[۳] Neural Networks

دسته بندی خودکار متون به طور کلی شامل شش فاز اصلی می‌باشد [Xia۲۰۰۷]:

۱ –تنظیم مجموعه داده‌ها، شامل متون آموزشی[۱] و متون آزمایشی[۲]

۲ –اندیس گذاری[۳] محتویات متن

۳ –استخراج و انتخاب ویژگی‌ها (کلمات یا عبارات کلیدی[۴])

۴ –طراحی یک دسته بند با استفاده از یادگیری ماشین

۵ –آزمایش دسته بندی طراحی شده با مجموعه داده های آزمایشی

۶ –ارزیابی روش دسته بندی

یکی از مسائل مهم در دسته بندی خودکار، بیان متون به صورتی است که برای الگوریتم دسته بندی مناسب باشد. یکی از پرکاربردترین روش‌ها، *مدل فضای برداری*[۵] می‌باشد که در آن هر متن به صورت برداری از *کلمه – مقدار*[۶] نمایش داده می‌شود. این روش شامل مراحلی می‌باشد که مهم‌ترین آن‌ها انجام پیش پردازشهای لازم بر روی متون، حذف کلمات عمومی[۷] و نشانه‌ها، ریشه یابی[۸] و استخراج کلمات یا عبارات کلیدی[۹] هستند [Sebastiani۲۰۰۲]. مسئله مهم دیگر، استفاده از الگوریتم دسته بندی مناسب می‌باشد. در بیشتر الگوریتم‌های دسته بندی، انتخاب نحوهٔ نمایش متون در افزایش دقت الگوریتم دسته بندی مؤثر بوده است.

روشی که در این پروژه برای دسته بندی متون استفاده شده است، دسته بندی بر مبنای قواعد انجمنی می‌باشد. دسته بندی بر مبنای قواعد انجمنی در واقع از قواعدی استفاده می‌کند که از روی فرایند کاوش الگوهای مکرر مجموعه داده های آموزشی تولید شده‌اند. این فرآیند با فرآیندی که در داده کاوی داده پایگاه داده های بزرگ استفاده می‌شود یکسان می‌باشد. همانطور که گفته شد استفاده از قواعد انجمنی در

---

[۱] Training Documents

[۲] Test Documents

[۳] Indexing

[۴] Keywords

[۵] Vector Space Model

[۶] Term-Value Vector

[۷] Stop Words

[۸] Stemming

[۹] Keywords Extraction

زمینه دسته بندی، نسبت به سایر روشهای دسته بندی مثل بیزین ساده، ماشینهای بردار پشتیبان، k-نزدیک‌ترین همسایه جدیدتر است [Yoon۲۰۰۷].

مزایای دسته بندی بر مبنای قواعد انجمنی چیست؟

دسته بندی بوسیله قواعدی مثل " کلاس ← ویژگی[۱]" انجام می‌شود و دارای مزایای متعددی می‌باشد که از آن جمله می‌توان به قابلیت تفسیر ساده، درک آسان قواعد توسط انسان و ویرایش مستقیم قواعدی که توسط فرآیند یادگیری استنتاجی[۲] تولید شده‌اند، اشاره کرد. حذف قواعد ضعیف و اضافه کردن قواعدی که با دقت و به صورت دستی توسط افراد خبره تعریف شده‌اند می‌تواند تا حد فوق‌العاده ای دقت دسته بندی را افزایش دهد. مزیت دیگر روش فوق به روز رسانی تدریجی قواعد توسط فرایندهای بعدی یادگیری ماشین می‌باشد. سایر تکنیک‌های دسته بندی اگرچه ممکن است دارای دقت قابل قبولی باشند ولی مزایای ذکر شدهٔ روشهای قاعده گرا را ندارند [Yoon۲۰۰۷].

یکی دیگر از مزایای مهم روش دسته بندی انجمنی این است که ویژگی‌ها هم می‌توانند منفرد باشند و هم چندگانه، یعنی می‌توان از اطلاعات ترکیبی ویژگیهای چندگانه استفاده کرد، در حالیکه روشهای دسته بندی دیگر مثل ماشینهای بردار پشتیبان و k- نزدیک‌ترین همسایه فقط از ویژگی منفرد استفاده می‌کنند. این بدین معنی است که در روش دسته بندی انجمنی امکان استفاده از اطلاعات *اصطلاح* یا *عبارت* همانند اطلاعات *لغت* وجود دارد.

دسته بندی بر مبنای قواعد انجمنی دارای معایبی هم هست که از آن جمله افزایش بعد فضای برداری ویژگی‌ها می‌باشد که برای رفع این مشکل از تکنیک‌های کاهش بعد فضای ویژگی‌ها استفاده می‌شود، و همچنین افزایش تعداد قواعدی که در فاز آموزش تولید شده‌اند و باعث افزایش بیهوده زمان محاسبات و کاهش تاثیر در دسته بندی انجمنی می‌شوند. برای رفع این مشکل هم از تکنیک *هرس کردن قواعد*[۳] استفاده می‌شود. در این تکنیک فقط قواعدی که دارای کیفیت و تأثیر بالایی هستند انتخاب می‌شوند [Sebastiani۲۰۰۲].

بنا بر مطالب فوق، دسته بندی بر مبنای قواعد انجمنی دارای سه فاز اصلی می‌باشد [Chen۲۰۰۵]:

۱ -تولید تمام قواعد انجمنی دسته (CAR)

---

۱ Feature → Class

۲ Inductive Learning

۳ Pruning Rules

۲ –ارزیابی کیفیت قواعد تولید شده در فاز قبلی و هرس کردن قواعد زاید، تکراری و کم تأثیر.
۳ –نسبت دادن یک برچسب دسته به داده های جدید.

۱–۵ داده کاوی چیست؟

داده کاوی، کشف دانش در پایگاه داده‌ها[۱] نامیده می‌شود. به بیان دیگر داده کاوی به روند کشف الگوهای مفید یا تعریف دانش از منابع داده‌ها مثل پایگاه داده‌ها، متون، تصاویر و وب اطلاق می‌شود. الگوها باید معتبر و به طور بالقوه مفید و قابل درک باشند. داده کاوی عرصه ای است که شامل زمینه های گوناگون پایگاه داده‌ها، یادگیری ماشین، آمار، هوش مصنوعی، بازیابی اطلاعات، و تجسم می‌باشد.

پایگاه داده‌ها به منظور تجزیه و تحلیل موثر مقادیر زیاد داده، ضروری می‌باشد. در این ارتباط، پایگاه داده، نه تنها به عنوان وسیله ای برای ذخیره سازی و دسترسی اطلاعات می‌باشد، بلکه امکان تجزیه و تحلیل داده را بوسیله الگوریتم‌های داده کاوی پشتیبانی می‌کند. به همین دلیل استفاده از تکنولوژی پایگاه داده در فرآیند داده کاوی می‌تواند بسیار مفید باشد.

یادگیری ماشین یکی از زمینه های هوش مصنوعی است که با توسعه تکنیک‌های تحلیل داده، امکان یادگیری را به کامپیوتر می‌دهد. بیشتر تمرکز روشهای یادگیری ماشین بر روی داده های نمادین می‌باشد. مسئله مهم در یادگیری ماشین، پیچیدگی الگوریتم‌ها در پیاده سازیهای محاسباتی می‌باشد.

آمار دارای زمینه‌هایی در ریاضیات می‌باشد و برای تجزیه و تحلیل داده های تجربی کاربرد دارد. آمار بر مبنای نظریه آمار می‌باشد که شاخه ای از ریاضیات کاربردی می‌باشد. در نظریه آمار، تصادف و عدم قطعیت توسط نظریه احتمال مدل سازی شده‌اند. امروزه روشهای آماری بسیاری در زمینه داده کاوی استفاده می‌شود.

بعضی از مفاهیمی که در داده کاوی وجود دارد شامل یادگیری با ناظر[۲] (یا دسته بندی)، یادگیری بدون ناظر[۳] (یا خوشه بندی[۴])، کاوش قواعد انجمنی و کاوش الگوهای مکرر می‌باشد.

---

۱ Knowledge Discovery in Databases

۲ Supervised Learning

۳ Unsupervised Learning

۴ Clustering