نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

دسته بندی متون فارسی با استفاده از یادگیری نیمه نظارت شده

محسن طاهری نیا
دانشجوی کارشناسی ارشد دانشکده برق و کامپیوتر، دانشگاه صنعتی اصفهان m.taherinia@ec.iut.ac.ir

چکیده

امروزه با توجه به حجم و رشد روز افزون متون فارسی ، دسته بندی اتوماتیک اسناد و متون از ارزش بزرگ عملی برخوردار و به طور فزاینده ، زمینه‌ی مهمی برای تحقیق است . در این نوشتار به بررسی یکی از روش‌های یادگیری هوشمند به نام یادگیری نیمه نظارت شده در دسته بندی متون فارسی خواهیم پرداخت. بسیاری از روش‌های یادگیری هوشمندانه مانند یادگیری نظارت شده ، فقط بر روی داده‌های آموزشی بر چسب دار تکیه می‌کنند ، در شرایطی که بدست آوردن این داده‌های آموزشی دارای بر چسب بسیار پر هزینه است . حال آنکه حجم زیادی از داده‌های بدون برچسب به سرعت زیاد و با هزینه‌ی کم در دسترس هستند. در مقابل روش‌هایی مانند روش یادگیری بدون نظارت فقط بر روی داده‌های بدون برچسب تکیه می‌کنند. در ادامه به بررسی روش یادگیری نیمه نظارت شده که ما بین روش‌های یادگیری نظارت شده و یادگیری بدون نظارت قرار دارد و از ترکیبی از مثال‌های آموزشی برچسب دار و بدون برچسب برای یادگیری استفاده می‌کند پرداخته و از این تکنیک برای دسته بندی متون فارسی استفاده می‌کنیم.

کلمات کلیدی

دسته بندی متون فارسی - یادگیری نیمه نظارت شده - تئوری بیز - الگوریتم EM.

۱- مقدمه

یکی از کاربردهای موفق یادگیری نیمه نظارت شده[۱] [۱و۲و۳و۴] ، دسته بندی متون[۲] به صورت خودکار است [۵و۶و۷].در این نوشتار نشان داده خواهد شد که چگونه از طریق یادگیری نیمه نظارت شده، دقت یاد گرفته شده توسط یک دسته بندی کننده‌ی متن، به وسیله‌ی اضافه کردن مقدار کمی از متون برچسب دار[۳] به مقدار زیادی از متون بدون برچسب[۴] افزایش می‌یابد[۸].

استفاده از متون بدون برچسب بسیار حائز اهمیت است. چون مشکل بسیاری از الگوریتم‌های دسته بندی کننده‌ی متون بدست آوردن متون آموزشی بر چسب دار است که فراهم کردن این متون بسیار پرهزینه، زمان‌بر، خسته کننده، خطاپذیر و نیازمند افراد متخصص است در حالی که مقدار زیادی از متون بدون برچسب به آسانی و سرعت زیاد در دسترسی هستند[۹].

در زمینه دسته بندی متون به زبان انگلیسی تکنیک‌های مختلف مانند دسته بندی کننده بیزین[۵] [۱۰]، شبکه‌های عصبی مصنوعی[۶] [۱۱]، الگوریتم‌های نزدیک‌ترین همسایه[۷] [۱۲]، ماشین بردار پشتیبان[۸] [۱۳]، الگوریتم EM و تکنیک‌های مختلف دیگر ارائه شده است . در زمینه دسته بندی متون به زبان فارسی نیز تکنیک‌هایی مانند روش‌های بدون ناظر[۹] [۱۴] ، دسته بندی کننده بیزین[۱۵]، الگوریتم‌های نزدیک‌ترین همسایه[۱۶]، شاخص گذاری n-gram [۱۷] ، استفاده از دانش معنایی[۱۸] و... به کار گرفته شده‌اند ولی در زمینه روش نیمه نظارت شده فعالیت چشمگیری انجام نشده است.

در این نوشتار یک الگوریتم بر مبنای یادگیری نیمه نظارت شده برای یادگیری از متون بر چسب دار و بدون برچسب بر مبنای ترکیب دو الگوریتم Naïve Bayes و Expectation Maximization برای متون فارسی معرفی خواهد شد. این الگوریتم در ابتدا یک دسته بندی کننده را با استفاده از متون بر چسب دار موجود آموزش داده و بعد به صورت احتمالی متون بدون برچسب را بر چسب گذاری می‌کند. سپس یک دسته بندی کننده‌ی جدید را به وسیله برچسب گذاری برای تمام متون آموزش داده و این عمل آنقدر تکرار می‌شود تا همگرا شود. این رویه مبنایی از الگوریتم EM ، در جایی که داده‌ها مطابق مفروضات مدل ما هستند به خوبی کار می‌کند ، با این وجود فرضیاتمان در عمل نقض شده و باعث افت کارایی می‌شود.

۲- دسته بندی متون به وسیله‌ی روش بیز:

قبل از پرداختن به مقوله‌ی دسته بندی نیمه نظارت شده‌ی متون باید مقداری اطلاعات در مورد دسته بندی نظارت شده داشته باشیم بنابراین ابتدا به دسته بندی نظارت شده به وسیله روش بیز می‌پردازیم [۱۰].

این روش متداول‌ترین روش در دسته بندی متون است . در این روش متن به صورت مجموعه‌های از کلمات مستقل از یکدیگر و مستقل از محل قرار گرفتن در متن در نظر گرفته می‌شود [۱۵] .تعریف تابع احتمال هر متن از حاصل‌ضرب احتمال کلمات آن و احتمال رخداد متنی با آن طول بدست

1.Semi Supervised learning
2.Text classification
3. Labeled document
4. Unlabeled document
5. Navie Bayes classifier
6 Artificial neural network
7. K-nearest neighbour algorithms
8. Support vector machines (SVM)
9. Unsupervised learning