

بهبود دسته‌بندی اسناد علمی فارسی با استفاده از اصطلاح‌نامه‌های تخصصی فارسی

بهاره داودآبادی فراهانی^۱، سید امید فاطمی^۲

^۱ دانشکده مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی، دانشگاه تهران، تهران
b.davoodabadi@ut.ac.ir

^۲ دانشکده مهندسی برق و کامپیوتر، پردیس دانشکده‌های فنی، دانشگاه تهران، تهران
ofatemi@ut.ac.ir

چکیده

به منظور بهبود دسته‌بندی اسناد چندین رویکرد مختلف وجود دارد که یکی از مهم‌ترین آن‌ها استفاده از بردار ویژگی مناسب است. در دسته‌بندی اسناد معمولاً کلمات، اجزای بردار ویژگی را تشکیل می‌دهند. از آن جایی که در هر مجموعه‌ای اسناد تعداد بسیار زیادی کلمه وجود دارد؛ در این مقاله روشی نوین برای تولید بردار ویژگی مناسب، به منظور بهبود دسته‌بندی اسناد علمی فارسی ارائه می‌شود. روش پیشنهادی ارتباطات معنایی بین کلمات هر سند را با استفاده از اصطلاح‌نامه استخراج می‌کند. سپس به منظور انتخاب ویژگی‌های دسته‌بند، روش نوین و کارآمد گروه‌بندی کلمات سند از لحاظ ارتباطات معنایی را، بکار می‌گیرد. برای ارزیابی و مقایسه روش پیشنهادی از مجموعه پایان‌نامه‌های ایران‌داک (پژوهشگاه علوم و فناوری اطلاعات ایران) استفاده می‌شود. نتایج آزمایشات نشان می‌دهد، عملکرد دسته‌بندی با بکارگیری این روش نسبت به انتخاب ویژگی بدون استفاده از اصطلاح‌نامه به صورت قابل توجهی افزایش می‌یابد.

کلمات کلیدی

دسته‌بندی اسناد علمی فارسی، بردار ویژگی، اصطلاح‌نامه، ارتباطات معنایی، انتخاب ویژگی، بهبود وزن‌دهی، دسته‌بند k -NN (k Nearest Neighbors)، بازیابی اطلاعات

۱- مقدمه

تمامی این روش‌ها از پردازش زبان طبیعی، داده‌کاوی و یادگیری ماشین برای تشخیص الگوهای اسناد و دسته‌بندی خودکار استفاده می‌کنند. بنابراین عملکرد دسته‌بندها را می‌توان با استفاده از روش‌های مختلفی مانند نمایه‌سازی مناسب متون، نمایش صحیح اسناد و انتخاب ویژگی مناسب و یافتن تابعی مطلوب برای دسته‌بندی بهبود داد [3]. در این میان، نمایش سند و انتخاب ویژگی، یکی از اساسی‌ترین مراحل مورد نیاز پیش از هر نوع دسته‌بندی است؛ که در آن‌ها سعی می‌شود، کلماتی به عنوان ویژگی انتخاب شوند، که نسبت به کلمات دیگر وابستگی بیشتری به توزیع کلاس داشته‌باشند [2]. تا به حال مطالعات زیادی با هدف بهبود دسته‌بندی انجام شده‌است؛ که بسیاری از آن‌ها به دنبال یافتن روشی کارآمد برای نمایش سند و انتخاب ویژگی هستند. یکی از روش‌های ابتدایی، BOW (Bag Of Words) نام دارد. این الگوریتم سند را به صورت برداری از ترم‌های مستقل از هم نشان می‌دهد که وزن هر ترم برحسب فرکانس وقوع آن محاسبه می‌شود [4]. یانگ [5] پنج روش انتخاب ویژگی MI^3 ، CHI^3 ، IG^2 ، DF^1 ، MI^4 ،

امروزه با پیشرفت فناوری اطلاعات، حجم زیادی از اسناد علمی به صورت الکترونیکی بر روی اینترنت منتشر شده‌است. از آن جایی که اسناد علمی شامل تمامی انتشارات علمی مانند پایان‌نامه‌ها، کتاب‌ها، مقالات مجلات و کنفرانس‌ها، گزارشات فنی و غیره است؛ حجم زیاد و نیاز به دسترسی سریع به این اسناد، ضرورت سازماندهی مؤثر و کارآمد آن‌ها را بیش از پیش کرده‌است.

دسته‌بندی یک روش مفید برای سازماندهی مجموعه‌ای داده‌ها در تعدادی گروه است. موتورهای جستجویی که از دسته‌بندی استفاده می‌کنند، اسناد مرتبط را سریع‌تر و آسان‌تر در اختیار کاربران قرار می‌دهند [2]. تاکنون الگوریتم‌های بسیاری از جمله k امین نزدیک‌ترین همسایه، درخت تصمیم، SVM (Support Vector Machine)، شبکه عصبی، و مدل بیزی، برای دسته‌بندی اسناد ارائه شده‌است [2].