نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران
دانشکده مهندسی کامپیوتر، دانشگاه شهید بهشتی، تهران، ۱۵-۱۳ اسفندماه ۱۳۹۲

استخراج خودکار اطلاعات از متون فارسی در دامنه خاص

آرمان شریف زاده[۱]، مهرنوش شمس فرد[۲]

[۱] کارشناس ارشد نرم‌افزار، دانشکده مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی، قزوین، ایران
sharifzadeh.arman@gmail.com

[۲] آزمایشگاه پردازش زبان طبیعی، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، تهران، ایران
m-shams@sbu.ac.ir

چکیده

حجم انبوه متون قابل دسترس بخصوص در گستره جهانی اینترنت و اطلاعات موجود در این حجم انبوه، اهمیت استخراج خودکار اطلاعات از متن را بیشتر نشان می‌دهد. استخراج اطلاعات از متن شامل ارائه قالب ساخت‌یافته از اطلاعات دلخواه موجود در متن می‌باشد. در این مقاله به معرفی یک سامانه استخراج خودکار اطلاعات از متون فارسی در دامنه خاص می‌پردازیم. سامانه استخراج خودکار اطلاعات برای زبان فارسی در حوزه اخبار حوادث تروریستی بر اساس ترکیبی از روش‌های یادگیر مانند الگوریتم ماشین بردار پشتیبان و مدل میدان‌های تصادفی شرطی و روش‌های مبتنی بر الگوهای استخراج، معرفی و ارزیابی شده است. نتایج بدست آمده نشان می‌دهد که این سامانه در مقایسه با کارهای مشابه دارای دقت و بازخوانی قابل قبولی است.

کلمات کلیدی

استخراج اطلاعات، یادگیری ماشین، الگوریتم ماشین بردار پشتیبان، مدل میدان‌های تصادفی شرطی، زبان فارسی

۱- مقدمه

استخراج خودکار اطلاعات از موضوعات اصلی علم پردازش زبان طبیعی[۱] است که با موضوعات بسیاری مانند زبان‌شناسی[۲]، یادگیری ماشینی[۳]، بازیابی اطلاعات[۴]، پایگاه‌داده، وب جهان گستر و تجزیه و تحلیل اسناد در ارتباط می‌باشد. شناسایی خودکار نوع خاصی از موجودیت‌ها، روابط و یا وقایع از متن با عنوان استخراج اطلاعات از متن نامیده می‌شود[۱].

منابعی که در استخراج اطلاعات به عنوان ورودی مورد توجه هستند در سه نوع ساخت‌یافته[۵] (مثل جداول پایگاه‌داده یا هستان‌شناسی‌ها)، نیمه ساخت‌یافته[۶] (مثل فرهنگ‌های لغت و مستندات HTML) و غیرساخت‌یافته[۷] (مثل متون زبان‌طبیعی موجود در پیکره‌ها) معرفی شده‌اند.

جستجوی اطلاعات در ورودی غیرساخت‌یافته دشوارتر از جستجو در ورودی نیمه ساخت‌یافته و همچنین جستجو در ورودی نیمه ساخت‌یافته دشوارتر از جستجو در ورودی ساخت‌یافته است.

به طور کلی دو نگرش مختلف مهندسی دانش و یادگیری ماشین برای استخراج اطلاعات از متن وجود دارد که رویکردهای موجود برای استخراج اطلاعات بر اساس این دو نگرش در ۳ دسته طبقه‌بندی می‌شوند [۲]:

(۱) رویکرد دستی با استفاده از مهندسی دانش
(۲) رویکرد یادگیر با استفاده یادگیری ماشین
(۳) رویکرد ترکیبی

در حقیقت سامانه‌های استخراج اطلاعات از رویکرد مهندسی دانش به سمت مدل‌های آماری در حال حرکت هستند[۱].

روش‌های سنتی استخراج اطلاعات عمدتا مبتنی بر قانون و بر پایه استفاده از عبارات منظم[۸] بوده‌اند. اگرچه از روش‌های مبتنی بر گرامر[۹] هم استفاده شده است اما در مقیاس داده‌های حجیم و تعداد زیاد قوانین این روش دشوار به نظر می‌رسد. سامانه فاستاس[۱۰][۳] و همچنین کار هابز[۱۱] و همکاران[۴] از نمونه‌های سنتی است که برای استخراج اطلاعات از متن ارائه شده است. برای زبان فارسی نیز یک سامانه استخراج اطلاعات با رویکرد مبتنی بر قانون و با استفاده از الگوهای استخراج با نام "مرصاد" [۵] معرفی شده است که بر روی اخبار در حوزه نظامی کار می‌کند. البته این گونه سامانه‌ها به دلیل ابهام‌های گسترده بخصوص در جمله‌های طولانی تا حدی کند و مستعد خطا هستند.