استخراج کلمات کلیدی جهت طبقه‌بندی متون فارسی

سمیه عربی نرئی[۱] ، مجتبی وحیدی اصل[۲] ، بهروز مینایی بیدگلی[۳].

چکیده

با رشد روز افزون اسناد و متون الکترونیکی به زبان فارسی، به کارگیری روش‌هایی سریع و ارزان برای دسترسی به متون مورد نظر از میان مجموعه وسیع این مستندات، اهمیت بیشتری می‌یابد. برای رسیدن به این هدف، استخراج کلمات کلیدی که بیانگر مضمون اصلی متن باشند، روشی بسیار موثر است. هدف ما در این مقاله، استخراج کلمات کلیدی موجود در مستندات فارسی، بر اساس معماری پیشنهادی، به منظور طبقه‌بندی کارآمد آنها در موتورهای جستجو است. روش ارائه شده شامل دو مرحله اصلی است: مراحل پیش‌پردازش و عملیات استخراج کلمات کلیدی. بدین‌منظور از ترکیبی از تکنیک‌های الهام گرفته از Wordnet و الگوریتم Porter، تطبیق یافته با زبان فارسی، و تکنیک Luhn[۴]، بهبود یافته، استفاده شده است. برای تسریع عملیاتِ استخراج کلمات‌کلیدی، از ساختمان داده‌ای مانند جداول درهم‌سازی و ساختار Trie استفاده می‌کنیم. یکی از مهمترین مسائلی که در این فرآیند، مورد توجه قرار گرفته، پوشش کلیه حالات دستوری کلمات و صورت‌های نگارشی مختلف آنها در زبان فارسی است. بر اساس بررسی‌های انجام شده بر روی یکصد متن فارسی و مقایسه نتایجِ بدست آمده با روش‌های دیگر، این روش می‌تواند کلمات کلیدی موجود در متون را با دقت و سرعت بیشتری استخراج نماید به گونه‌ای که این کلمات کلیدی، بیانگر مضمون اصلی متن باشند.

کلمات کلیدی

استخراج کلمات کلیدی، پیش‌پردازش، ساختار Trie.

---

[۱] دانشجوی کارشناسی‌ارشد نرم‌افزار- دانشگاه علم و صنعت ایران- دانشکده مهندسی کامپیوتر- atarabi@comp.iust.ac.ir
[۲] دانشجوی کارشناسی‌ارشد نرم‌افزار- دانشگاه علم و صنعت ایران- دانشکده مهندسی کامپیوتر- mojtabavahidi@comp.iust.ac.ir
[۳] استادیار کامپیوتر- دانشگاه علم و صنعت ایران- دانشکده مهندسی کامپیوتر- minaeibi@cse.msu.edu

[4] روشی برای خلاصه‌سازی متن. در این روش به هر جمله یک فاکتور اهمیت داده می‌شود، و جملات با بیشترین فاکتور اهمیت برای ایجاد خلاصه استفاده می‌شوند.

PDF created with FinePrint pdfFactory Pro trial version http://www.pdffactory.com