یازدهمین کنفرانس مهندسی برق ایران
۱۸-۱۶ اردیبهشت ۱۳۸۲

ساخت یک نمایه ساز خودکار برای متون فارسی

مسعود تشکری
Tashakori@noavar.com

محمدرضا میبدی
Meybodi@ce.aku.ac.ir

آزمایشگاه سیستمهای نرم‌افزاری
دانشکده مهندسی کامپیوتر و فن آوری اطلاعات
دانشگاه صنعتی امیر کبیر
تهران ایران

چکیده: در این مقاله یک نمایه ساز خودکار برای متون فارسی ساخته شده است. برای آزمایش این سیستم، ۴۵۰چکیده و ۳۲ پرس‌وجوی فارسی، که در زمینه‌های تخصصی کامپیوتر هستند، جمع آوری شده است. بر روی متون جمع آوری شده، ابتدا شیوه توزیع واژگان مطالعه می‌شود. سپس فهرستی از واژگانی که در زبان فارسی، عمومی به شمار می‌روند تهیه می‌شود. در این مقاله با مطالعه روش ارزیابی سیستمهای بازیابی متن، سیستم نمایه ساز خودکار ارائه شده، با استفاده از دو پارامتر "بازخوانی" و "دقت" ارزیابی می‌شود. همچنین با استفاده از ریشه یاب خودکار واژگان فارسی، تغییرات کیفی سیستم در هنگام استفاده از ریشه یابی واژگان نیز مورد بررسی قرار می‌گیرد.

واژگان کلیدی: نمایه سازی خودکار، واژگان عمومی زبان فارسی، بازیابی متون فارسی، ارزیابی نمایه ساز، ریشه یابی واژگان فارسی

۱- مقدمه

امروزه حجم زیاد و تقریباً نامحدود اطلاعات موجود باعث می‌گردد تا استفاده از اطلاعات و مدیریت آن با دشواریهایی روبرو باشد. با توجه به رشد و گسترش حجم اطلاعات و به موازات آن ضرورت به کار گیری و استفاده مؤثر از منابع اطلاعاتی، یکی از مهمترین و اساسی‌ترین نیازهای موجود قابلیت دستیابی به اطلاعات مورد نیاز در مدت زمان مناسب است. در واقع جستجوی سریع و یافتن اطلاعات مورد نظر جستجوگر از اهمیت فوق العاده‌ای برخوردار است .

با وجود آنکه اطلاعاتی که امروزه عرضه می‌شوند صورتهای مختلفی از قبیل تصویر، صوت، انیمیشن، و . . . به خود گرفته‌اند، هنوز هم پراستفاده‌ترین و حجیمترین اطلاعات موجود، متون غیر ساخت یافته هستند[۱]. به طور مثال در بایگانی سازمانها اطلاعات متنی زیادی در قالب نامه‌ها، جزوات، آئین نامه‌ها و دستورالعملها و قوانین(حقوقی، جزایی، مالیاتی)، و ... وجود دارد؛ و یا در روزنامه‌ها و مجلات، کتابخانه‌ها، و محیطهای وب و پست الکترونیک، که روزبروز نیز گسترش می‌یابند، غالب اطلاعات به صورت متن می‌باشد.

از زمان به کار گیری کامپیوتر با نگهداری متون به صورت الکترونیکی ، رشد قابل توجهی در جمع آوری اطلاعات و سازماندهی آنها به منظور بهره برداری هر چه بیشتر مشاهده می‌گردد. این نوع نگهداری باعث کاهش فضای ذخیره سازی و هزینه نگهداری و نیز انعطاف بیشتر در کاربرد متنها شده است. از این رو نگهداری الکترونیکی متون همواره مورد توجه می‌باشد و تلاشهای زیادی نیز برای بهبود سرعت و کیفیت روشهای بازیابی چنین اطلاعاتی صورت می‌پذیرد. منظور از بازیابی اطلاعات متنی، فرایند جستجو و یافتن این نوع اطلاعات در مجموعه بزرگی از متون است. سرعت و سهولت در جستجو و بازیابی اطلاعات نکته‌ای است که همواره باید مد نظر باشد تا از هزینه‌های غیر ضروری و اتلاف وقت اجتناب شود. تکنیکهای بازیابی اطلاعات به دلیل ناتوانی پایگاههای داده‌ای سنتی در مدیریت متون غیر ساخت یافته، به شدت مورد توجه قرار گرفته‌اند[۱].