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

پردازش عبارت در موتور جستجو

جواد پاک سیما[۱]، علی محمد زارع بیدکی[۲]، ولی درهمی[۳]

[۱] دانشجوی دکترای دانشکده برق و کامپیوتر- دانشگاه یزد
paksima@stu.yazd.ac.ir

[۲] استادیار دانشکده برق و کامپیوتر- دانشگاه یزد
alizareh@yazd.ac.ir

[۳] دانشیار دانشکده برق و کامپیوتر- دانشگاه یزد
vderhami@yazd.ac.ir

چکیده

تحقیقات زیاد روی موتورهای جستجو نشان می‌دهد که اکثر پرس و جوهای کاربران بیش از یک کلمه می‌باشد و ممکن است بطور مشخص با استفاده از علامت نقل قول به عنوان عبارت معرفی شده باشند یا از علامت نقل قول استفاده نشده باشد ولی در بیشتر مواقع منظور کاربر یک عبارت باشد. اکثر الگوریتم‌های رتبه بندی از فرکانس رخداد یک کلمه در سند(TF) برای امتیاز دهی به اسناد استفاده می‌کنند اما برای عبارت تعریف روشنی از این پارامتر وجود ندارد. از طرفی تعداد رخداد یک عبارت به تنهایی مفید نیست و باید فاصله بین کلمات عبارت محاسبه گردد. در این مقاله پارامترهای فاصله، فرکانس رخداد یک عبارت(PF) و IDF با توجه به فاصله تعریف می‌شود و الگوریتم‌هایی برای محاسبه آنها ارائه می‌گردد. همچنین نتایج الگوریتم پیشنهادی با الگوریتم پیاده سازی شده توسط نمایه ساز متن باز لوسین مقایسه گردیده است.

کلمات کلیدی

موتور جستجو، عبارت، فاصله، فرکانس عبارت(PF)

۱- مقدمه

پیدا کردن صفحات وب دارای کیفیت بالا یکی از مهمترین وظایف موتورهای جستجو می‌باشد. مفهوم میزان ارتباط اسناد پیدا شده با پرس و جو بیشتر وابسته به نظر کاربر می‌باشد و این موضوع باعث افزایش پیچیدگی الگوریتمهای رتبه بندی می‌شود. نکته‌ی دیگر این است که غالباً کاربران ۱۰ تا ۲۰ نتیجه اول را بررسی می‌کنند[۱] در حالیکه برای یک پرس وجو ممکن است میلیونها صفحه ی مرتبط وجود داشته باشد. بنابراین موتورهای جستجو باید برای یافتن مرتبطترین صفحات الگوریتم مناسب با کارایی بالا ارائه نمایند.

بخش رتبه بندی یکی از مهمترین قسمتهای موتور جستجو می‌باشد. رتبه بندی فرآیندی است که کیفیت صفحه توسط موتور جستجو تخمین زده می‌شود. در حال حاضر دو روش عمده برای رتبه بندی صفحات وب وجود دارد. در روش اول رتبه بندی براساس محتوای اسناد انجام می‌شود(رتبه بندی سنتی). مدلهایی مانند مدل بولی ، مدل احتمالی و مدل فضای برداری جهت رتبه بندی اسناد مبتنی بر محتوا ارائه شده اند[۹]. در روش دوم براساس گراف و اتصالات وب و میزان اهمیت صفحات رتبه بندی صورت می‌گیرد.

در رتبه بندی سنتی، موتور جستجو سعی می‌کند با پیدا کردن میزان ارتباط سند و پرس و جو رتبه بندی را انجام دهد. در این الگوریتمها برای هر پرس و جو، اسناد با محتوای شبیه تر به کلمات موجود در پرس و جو امتیاز بالاتری را دارا می‌باشند. بعنوان مثال الگوریتم های TF-IDF[۷]، BM25[۸] دو نمونه از رایج‌ترین الگوریتم‌های این نوع رتبه بندی هستند.

وب شامل تعداد زیادی اسناد غیر ساخت یافته می‌باشد که بهم متصل هستند و یک گراف خیلی بزرگ را ایجاد می‌کند. معمولاً تعداد کلمات پرس و جوها کم بوده (۲.۴ کلمه در هر پرس و جو[۱۰]) و مجموعه‌ی کل لغات خیلی زیاد می‌باشد. این در حالیست که در اکثر