

دسته‌بندی وب سایت‌های عمیق با استفاده از نشانه‌های رابط جستجو

حسین شریفی‌پناه، علی موقر

دانشکده‌ی مهندسی کامپیوتر، دانشگاه صنعتی شریف

E-mail: {sharifipناه, movaghar}@ce.sharif.edu

چکیده - امروزه حجم بسیار زیادی از اطلاعات در دنیای وب وجود دارند که تنها راه دسترسی به آن‌ها از طریق رابط‌های جستجو است. همین مسأله بسیاری از خزش‌گرهای موجود را با مشکلات اساسی روبرو کرده است. شناخت رابط جستجو و گزینش یک پرس‌وجوی مناسب از یک طرف، و بازیابی مستنداتی که به عنوان نتیجه، توسط وب سایت برگردانده می‌شوند از طرف دیگر، مسائلی هستند که چالش‌های موجود بر سر راه خزش‌گرهای این نوع وب سایت‌ها را دو چندان نموده است. در این مقاله سعی شده است الگوریتمی ارائه شود که فارغ از مسائل مطرح شده، بتواند به دسته‌بندی وب سایت‌های عمیق بپردازد. این دسته‌کننده که SIAC نام دارد، با استفاده از نشانه‌های موجود در رابط جستجو به این مهم می‌پردازد. نتایج ارزیابی کارایی این الگوریتم، نشان‌دهنده کارایی قابل قبول روش پیشنهادی است.

کلید واژه- خزش‌گر، دسته‌کننده، موتور جستجو، وب عمیق.

۱- مقدمه

صفحات خود را تولید کرده و در اختیار کاربران می‌گذارند [۳]. همچنین سرعت رشد نرم‌افزارهایی که تولید این صفحات را تسهیل می‌کند، رو به افزایش است [۴-۶]. اطلاعات این سایت‌ها بر روی پایگاه‌داده‌هایی است که تنها راه دستیابی به این اطلاعات عبور از رابط موجود در سایت می‌باشد. به مجموعه‌ی این سایت‌ها، وب عمیق و یا وب مخفی می‌گویند [۷].

براساس بررسی‌های صورت گرفته، اندازه‌ی اطلاعات عمومی در وب عمیق، در حدود ۴۰۰ تا ۵۵۰ برابر اندازه‌ی وب سطحی (مجموعه سایت‌هایی که عمیق نیستند) می‌باشد. از طرف دیگر، وب عمیق در حدود ۷۵۰۰ ترابایت اطلاعات دارد. در حالیکه اطلاعات در وب سطحی ۹۰ ترابایت است. یکی دیگر از مسائلی که باعث پر رنگ‌تر شدن اهمیت وب عمیق می‌گردد میزان ترافیک موجود در این وب‌سایت‌ها است. به طور میانگین ترافیک وب سایت‌های عمیق در طی یک ماه، پنجاه

گسترش روزافزون وب از یک‌سو و افزایش نیاز کاربران به اطلاعات موجود بر روی وب از سوی دیگر، لزوم حضور موتورهای جستجو را برای تسهیل امر دستیابی به منابع مورد نیاز را بیش از پیش روشن می‌سازد. طبق تحقیقات، ۸۵٪ کاربران وب از موتورهای جستجو برای پیدا کردن اطلاعات مورد نیازشان در وب استفاده می‌کنند [۱]. بنابراین یکی از مهم‌ترین مسائل، قدرت موتورهای جستجو در بدست آوردن صفحات وب و دسته‌بندی آن‌ها می‌باشد. اما آنچه که امروزه به عنوان چالشی بر سر راه پیشبرد اهداف موتورهای جستجو قرار گرفته است، تغییر ماهیت ساختاری وب است. تحقیقی تحت نظر NPD نیز با نشان دادن این موضوع که نرخ شکست در جستجو از سال ۱۹۹۷ افزایش یافته است، مؤید این مهم است [۲]. با توجه به تحقیقات، میزان قابل توجهی از وب‌سایت‌ها، به صورت پویا و براساس درخواست کاربر،