

یک روش crawling متمرکز جدید مبتنی بر الگوریتم‌های ژنتیک و مسیریابی AntNet

عباس شاهینی شمس آبادی و کامران زمانی‌فر

دانشگاه اصفهان، دانشکده فنی مهندسی، گروه کامپیوتر

E-mail: {shahini, zamanifar} @eng.ui.ac.ir

چکیده - Crawler های متمرکز برای جمع آوری صفحات وب در یک زمینه خاص استفاده می‌شوند. این crawler ها با چالش‌هایی از قبیل مشکل جستجوی محلی و چگونگی پیش بینی کیفیت صفحات وب قبل از بازیابی آنها، مواجه هستند. روش پیشنهاد شده در این مقاله از الگوریتم‌های ژنتیک و مسیریابی AntNet برای حل این مشکلات استفاده می‌نماید. نتایج آزمایشات انجام گرفته با در نظر گرفتن معیارهای harvest rate و میانگین کل امتیازات صفحات جمع آوری شده، کارایی بالای این روش را نشان می‌دهند.

کلید واژه - الگوریتم ژنتیک، الگوریتم مسیریابی AntNet crawler متمرکز.

۱- مقدمه

وب حاوی بیش از یازده و نیم بیلیون صفحه بوده و هر روز با یک رشد انفجاری، حدود هفت و نیم میلیون صفحه به آنها اضافه می‌شود. موتورهای جستجوی کلی نمی‌توانند این حجم عظیم از صفحات را با این سرعت رشد بالا پوشش دهند و کاربرانی که اطلاعاتی را درخواست می‌کنند، صفحات وب مهیا شده را نامناسب و تاریخ گذشته می‌یابند.

یک crawler با کارایی بالا حداقل باید یکی از دو خصوصیتی را که در ادامه ذکر می‌شود داشته باشد. اول اینکه از استراتژی‌های هوشمند در تصمیم‌گیری‌های خود استفاده نماید (مثلا انتخاب لینکی از میان لینک‌های دنبال نشده تا صفحه مربوط به آن بازیابی شود) و دوم اینکه ساختارهای سخت‌افزاری و نرم‌افزاری پشتیبانی کننده آن برای بازیابی مقادیر زیادی از صفحات در واحد زمان بهینه

وب بزرگ‌ترین منبع اطلاعاتی جهان بوده و موتورهای جستجو نقش مهمی را جهت فراهم نمودن صفحات مناسب برای کاربران ایفا می‌کنند. یک موتور جستجو می‌تواند به دو قسمت کلی on-line و off-line تقسیم شود. قسمت on-line، پرس و جو^۱ها را از کاربران گرفته و از میان صفحات ایندکس شده که قبلا ذخیره شده‌اند صفحات مناسب را به کاربر می‌فرستد. وظیفه بخش off-line، جمع-آوری صفحات وب و ایندکس‌گذاری آنها می‌باشد. موتورهای جستجو از crawlerها برای بازیابی و ذخیره صفحات وب استفاده می‌کنند.

طبق تحقیقات انجام شده در سال ۲۰۰۵ [۲۳] در آن زمان

^۱ Query.