عنوان مقاله:
استخراج خودکار محتوای مفید صفحات وب با استفاده از آتوماهاتای یادگیر

محل انتشار:
کنفرانس بین المللی مهندسی کامپیوتر و فناوری اطلاعات (سال: 1395)

تعداد صفحات اصل مقاله: 15

نویسندگان:
زیبا جعفری - *دانشگاه آزاد اسلامی واحد کرمان،ایران*

محمد احمدی نیا - *دانشگاه آزاد اسلامی واحد کرمان،ایران*

خلاصه مقاله:
با توسعه سریع اینترنت، منابع اطلاعاتی متعددی به صورت صفحات HTML در شبکه جهانی وب منتشر شده اند. با این حال بسیاری از اطلاعات زائد و بی ربط در اینترنت وجود دارد از قبیل پانل ناوبری، جدول محتوا، تبلیغات، اظهارات حق انحصاری، کاتالوگ خدمات، سیاست حفظ حریم خصوصی و غیره. درنتیجه محتوای صفحات وب به دو صورت محتوای مفید (اصلی) و غیرمفید (غیر اصلی) درنظرگرفته شده اند. بیشتر دریافت کننده ها وکاربران نهایی فقط محتوای مفید را جستجو می کنند و نیاز به استخراج محتوای مفید از صفحات وب دارند که باید مشخص باشند. محتوای مفید، محتوای اصلی از صفحه وب است که بسیاری از اطلاعات مورد نیاز را به کاربر می دهد .در این مقاله، روشی جهت استخراج محتوای مفید صفحات وب پیشنهاد شده که ابتدا یک صفحه وب را دریافت می کند و بعد از استاندارد نمودن آن صفحه وب، درخت DOM را ایجاد می کند سپس مسیرهای درخت DOM از ریشه تا برگ استخراج می شوند. بعد از آن معادل با هر مسیر، یک اتوماتای یادگیر تصادفی تعیین می شود و به کمک آن وضعیت هربلوک جهت مفید بودن یا نبودن در یک فرآیند تکراری مشخص می شوند. درنهایت بلوک های حاوی محتوای مفید صفحات وب استخراج می شوند. این مدل می تواند نتایج موتورهای جستجو، تلخیص محتوای وب و برنامه های کاربردی داده کاوی را بالا ببرد. یک راه حل مفید خاص برای استخراج محتوای وب است. راه حل ارائه شده بر روی یک مجموعه داده ویکی اعمال گردیده است ونتایج حاصل ، بیانگر دقت و فراخوانی به ترتیب 97.2% و 98.1% می باشد.

کلمات کلیدی:
وب کاوی، استخراج محتوای مفید، آتوماتاهای یادگیر، مدل شئ سند

لینک ثابت مقاله در پایگاه سیویلیکا: