عنوان مقاله:
پالایش صفحات وب بر اساس تحلیل هوشمند محتوا

محل انتشار:
چهاردهمین کنفرانس سالانه انجمن کامپیوتر ایران (سال: 1387)

تعداد صفحات اصل مقاله: 7

نویسندگان:
علی احمدی - *دانشکده برق و کامپیوتر دانشگاه صنعتی خواجه نصیر طوسی*

مهدی زمانیان - *دانشکده برق و کامپیوتر دانشگاه صنعتی خواجه نصیر طوسی*

هادی فرزین - *دانشکده برق و کامپیوتر دانشگاه صنعتی خواجه نصیر طوسی*

محمود خالقی - *مرکز تحقیقات مخابرات ایران*

خلاصه مقاله:
روش های موجود برای پالایش صفحات وب بیشتر مبتنی بر سد کردن نشانی های اینترنتی خاص از طریق جستجو در یک لیست مرجع از صفحات غیر مجاز و یا با استفاده از تحلیل ساده متن از طریق جستجوی کلمات کلیدی خاص در صفحات است. مشکل اصلی این روش ها نیاز برای به روزرسانی مداوم فهرست نشانی ها و نیز میزان قابل توجه اشتباه گرفتن صفحه های مجاز در آنهاست. در این مقاله یک روش پالایش هوشمند برای پالایش صفحات غیراخلاقی را پیشنهاد کرده ایم که با استفاده از هر سه نوع ویژگی ساختاری، متنی و تصویری و ترکیب سلسله مراتبی آنها یک دسته بندی هوشمند با دقت بالا (روی FN و FP هر دو) را به دست می دهد. الگوریتم روی 2600 صفحه وب شامل 1400 صفحه غیراخلاقی (دارای متن، تصویر، یا هر دو) انگلیسی و فارسی و 1200 صفحه مجاز شامل صفحات پزشکی، سلامت، ورزشی و غیره مورد آزمایش قرار گرفته و دقت دسته بندی بالای 95 % را به همراه داشته است.

کلمات کلیدی:
پالایش هوشمند، پالایش محتوا، شناسایی صفحات وب، صفحات غیر اخلاقی، پروفایل صفحات، رنگ پوست