عنوان مقاله:
استخراج اطلاعات از صفحات وب بر اساس ساختار آن ها

محل انتشار:
اولین کنفرانس داده کاوی ایران (سال: 1386)

تعداد صفحات اصل مقاله: 11

نویسندگان:
میثم قادریان - *دانشجوی کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر*

احسان درویشی - *دانشجوی کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر*

حسن ابوالحسنی - *استادیار دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف*

خلاصه مقاله:
در این مقاله روشی برای استخراج اطلاعات ساخت یافته از صفحات وب مانند صفحات ویژگی های محصولات ارایه شده است. اکثر روش های موجود برای استخراج اطلاعات بر پایه استنتاج لفافه (wrapper) می باشند. بر خلاق روش استنتاج لفافه که به مجموعه اولیه ای از صفحات برچسب گذاری شده نیاز دارد، این روش یک روش یادگیری بدون ناظر است، هنگامی که یک صفحه جدید با هیچ کدام از صفحات برچسب گذاری شده مطابقت نداشته باشد آن صفحه را برچسب گذاری شده بیشتر گشته که به این ترتیب صفحات جدید بیشتری با صفحات برچسب گذاری شده قبلی مطابقت پیدا می کنند، بنابراین برچسب های آن ها به راحتی انتخاب می گردد. این روش بر خلاف روش استنتاج لفافه، با اجتناب از برچسب گذاری صفحاتی که دارای قالب یکسان هستند، مشکل اساسی یادگیری استنتاجی را حل می کند. مجموعه صفحات برچسب گذاری صفحاتی که دارای قالب یکسان هستند، مشکل اساسی یادگیری استنتاجی را حل می کند. مجموعه صفحات بر چسب دار ممکن است قالب تمام صفحات را پوشش ندهد، چرا که داده های ساخت یافته بر روی وب معمولا در چند قالب ثابت قرار می گیرند و صفحاتی که از یک قالب استفاده می کنند، می توانند با استفاده از یک نمونه صفحه برچسب دار، استخراج شوند. معیارهای موجود بر مبنای فاصله اقلیدسی یا شباهت متنی، به علت تفاوت در موارد استخراج شده از صفحات مختلف به راحتی قابل اجرا نمی باشد. برای رفع مشکل مذکور این مقاله یک معیار شباهت جدید مبتنی بر ساختار صفحات وب را ارایه می دهد که بر روی صفحات وب قالب دار به راحتی اجرا می گردد. نتایج آزمایش این روش در استخراج اطلاعات ، نشان می دهد که با دقت بالاتری نسبت به روش استنتاج لفافه اطلاعات خواسته شده را استخراج می کند.

کلمات کلیدی:
استخراج اطلاعات ، داده کاوی