

عنوان مقاله:

روشی جدید برای خوشه بندی اسناد HTML با استفاده از الگوریتم های تلفیقی

محل انتشار:

فصلنامه مطالعات مدیریت کسب و کار هوشمند، دوره 6، شماره 24 (سال: 1397)

تعداد صفحات اصل مقاله: 26

نویسندگان:

مریم شعار - استادیار، گروه مدیریت صنعتی، دانشکده مدیریت، دانشگاه آزاد اسلامی، واحد تهران شمال، تهران

علی اصغر سالارنژاد - کارشناسی ارشد، مدیریت فناوری اطلاعات، دانشکده مدیریت، دانشگاه آزاد اسلامی، واحد تهران شمال، تهران

خلاصه مقاله:

با عنایت به حجم بالای اطلاعات کنونی وب توجه به سیستم های خودکار استخراج اطلاعات بیشتر شده است. از مهم ترین روش های خودکار استخراج اطلاعات، خوشه بندی می باشد. روش های خوشه بندی زیادی تا به حال ارائه شده است که اکثراً مبتنی بر مدل برداری می باشند. در این مدل با هر سند مانند مجموعه ای از کلمات برخورد می گردد و توالی کلمات در جمله، نادیده گرفته می شود. از آنجایی که معانی در زبان طبیعی به طور کامل وابسته به توالی کلمات می باشند تقیصه بزرگی در این روش ها احساس می گردد. برای رفع این نقیصه در این مقاله روشی جدید در خوشه بندی اسناد HTML ارائه گردیده است که در آن الگوریتم Stc برای خوشه بندی Snippet ها لحاظ شده است. این روش که با عنوان خوشه بندی براساس جملات کلیدی Ks_Stc مطرح شده برای هر سند بردار وزن داری تهیه می کند و با استفاده از این بردار، جملات کلیدی هر متن از سند استخراج می گردد و نهایتاً این جملات کلیدی برای خوشه بندی به الگوریتم Stc داده می شود.

کلمات کلیدی:

افزونگی اطلاعات، خوشه بندی اسناد HTML، داده کاوی، سیستم های استخراج اطلاعات، کلاس بندی

لینک ثابت مقاله در پایگاه سیویلیکا:

<https://civilica.com/doc/804223>

