

خوشه بندی توزیع شده و فازی متون مبتنی بر آنتولوژی

سید محمد جعفر راشدی، حسن رشیدی، امید سجودی

۱- دانشجوی کارشناسی ارشد نرم افزار، دانشگاه آزاد اسلامی واحد علوم تحقیقات قزوین

۲- هیأت علمی گروه رایانه دانشگاه علامه طباطبائی

۳- مدیر گروه مهندسی نرم افزار دانشگاه آزاد اسلامی واحد علوم تحقیقات قزوین

jafar.rashedi@gmail.com

نام ارائه دهنده: سید محمد جعفر راشدی

خلاصه

سراسر وب از انواع و تعداد زیادی اسناد متنی تشکیل یافته است و قالب عمومی نگهداری داده ها در بانک های اطلاعاتی بصورت متن است. خوشه بندی اسناد تکنولوژی قدرتمندی است که به طور گسترده می تواند در ساماندهی داده ها در هسته های کوچکتر و مدیریت پذیر استفاده گردد. برای حل مساله روشهای متعدد پیشنهادی از مشکلات متعددی چون ابعاد بالا، مترادف بودن، ابهام و فقدان محتوای توصیفی رنج می برند. با استفاده از یک دانش پس زمینه مانند شبکه معنایی کلمات فارسی (فارس نت) می تواند استاندارد الگوریتم های خوشه بندی را افزایش داد. در این مقاله تلفیقی از «فارس نت» و الگوریتم «خوشه بندی کلمات تکرار شونده» با نام PFMDC ارائه گردیده که به صورت رویکردی فازی و موازی پیاده سازی شده است. نتایج بدست آمده نشان می دهد که این رویکرد بهبود قابل توجهی در کیفیت خوشه بندی متون و بهبود حدود پنجاه درصدی در زمان اجرا داشته است.

کلمات کلیدی: خوشه بندی فازی متون، خوشه بندی کلمات تکرار شونده، فارس نت، آنتولوژی.

۱- مقدمه

سیل صفحات و مقالات متنی در سراسر وب افزایش روزانه فوق العاده دارد. این مسأله کاملاً غیر ممکن است که یک فرد بتواند از میان همه آنها داده های مرتبط را تشخیص دهد در حالی که این مقدار از منابع اطلاعاتی به صورت آنلاین و به سرعت در حال افزایش هستند. یکی از مهم ترین روشها برای سازماندهی این مقدار عظیم از اطلاعات تکنیک خوشه بندی است. خوشه بندی، فرآیند پارتیشن مجموعه ای از اشیاء به چندین زیر مجموعه بر اساس معیار شباهت می باشد. در مورد خوشه بندی کاربر هر زیر مجموعه خوشه کاربر نامیده شده و شامل درخواستهای مشخص مشابه کاربر از متن انتخابی در مجموعه اسناد می باشد. ثابت شده که خوشه بندی یک روش مفید برای بازیابی اطلاعات بوسیله یافتن اطلاعات مورد علاقه مرکزی و توزیع در داده های زیرین است. به طور کلی خوشه بندی در ساخت و ساماندهی پارتیشن های با معنی از مجموعه اشیاء گسترده کمک می نماید. بعنوان مثال در موارد زیر کاربرد دارد.

الف) ساماندهی نتیجه جستجوها

ب) ساماندهی گروه های موضوعی با استفاده از سایر روشهای بازیابی اطلاعات مانند خلاصه سازی

ج) در حوزه سیستم های توصیه کننده که با افزایش کارایی کاربران نهایی را به بخشهای مرتبط هدایت می کند.

د) در حوزه یادگیری ماشین مانند کاوش سری های زمانی که از آیت های اخیرا استفاده شده برای یافتن قوانین وابستگی در یک پایگاه داده مبتنی

بر تراکنش استفاده می کند. [۱]

خوشه بندی اسناد می تواند یک سیستم بازیابی اطلاعات را در برابر حجم عظیمی از اطلاعات برای کاربر فراهم می آورد که به اطلاعات مورد

نیازش سریعاً دستیابی پیدا کند. [۲]