عنوان مقاله:
خوشه بندی متون فارسی به کمک الگوریتم K-means

محل انتشار:
دومین کنفرانس ملی توسعه کاربردهای صنعتی اطلاعات، ارتباطات و محاسبات (سال: 1392)

تعداد صفحات اصل مقاله: 7

نویسندگان:
پرویز کدخدایی - *کارشناس ارشد کامپیوتر-هوش مصنوعی-گروه کامپیوتر*

عرفان شمس - *کارشناس ارشد کامپیوتر نرم افزار- گروه کامپیوتر*

خلاصه مقاله:
بهره گیری از قدرت فرآیند داده کاوی جهت شناسایی الگوها و مدل ها و نیز ارتباط عناصر مختلف در پایگاه داده جهت کشف دانش نهفته در داده ها و نهایتا تبدیل داده به اطلاعات، روز به روز ضروری تر می شود. داده کاوی مجموعه روش هایی است کهبه کمک آن ها به صورت خودکار اطلاعات پیشگویانه از پایگاه داده های بزرگ استخراج می شود. سپس از این اطلاعات برای به وجود آوردن اطلاعات بهتر و در نتیجه اخذ تصمیمات مفیدتر استفاده می شود. در این مقاله سعی شده است از روش خوشه بندی توصیفی برای خوشه بندی و دسته بندی متون فارسی استفاده شود. برای نمونه مجموعه ای از متون فارسی که از روی سایت های خبریموجود در وب جمع آوری شده است، برای انجام این تحقیق بکار می رود. این متون در ابتدا بوسیله از بین بردن علائم نقطه گذاری وکلمات بی فایده، پیش پردازش می شوند. در خوشه بندی برای نمایش هر متن از یک بردار ویژگی استفاده می شود که شامل کلماتشاخص و میزان تکرار آن کمات در متن می باشد. اصول خوشه بندی بر پایه فرضیات آماری استوار است که متونی که در خوشه یکسانی قرار می گیرند، ویژگی های مشابهی دارند. برای خوشه بندی متن جدید، ابتدا بردار ویژگی آن متن ساخته شده، سپس با بردارهای ویژگی خوشه ها مقایسه می شود. در صورتی که خوشه جدید تشخیص داده شد به لیست خوشه ها اضافه میگردد و در غیر این صورت رشد خوشه متوقف می گردد

کلمات کلیدی:
خوشه بندی متن ، داده کاوی توصیفی ، زبان فارسی/K-means