عنوان مقاله:

تشخیص موضوع در متون خبری با استفاده از گام برداری تصادفی تقویتی

محل انتشار:

بیست و دومین کنفرانس ملی سالانه انجمن کامپیوترایران (سال: 1395)

تعداد صفحات اصل مقاله: 7

نویسندگان:

سپهر آروین – دانشکده ی برق و کامپیوتر، پردیس دانشکده فنی، دانشگاه تهران، تهران

علی ورداسبی – دانشکده ی برق و کامپیوتر، پردیس دانشکده فنی، دانشگاه تهران، تهران

ح فیلی – دانشکده ی برق و کامپیوتر، پردیس دانشکده فنی، دانشگاه تهران، تهران

آزاده شاکری – دانشکده ی برق و کامپیوتر، پردیس دانشکده فنی، دانشگاه تهران، تهران

خلاصه مقاله:

تشخیص موضوع بر روی متون مختلف از جمله متون خبری یکی از مسایلی است که در سال های اخیر مورد توجه قرار گرفته و پژوهش های گوناگونی بر روی آن انجام شده است. برای حل این مسیله روش های مختلفی ارایه شده که در آن ها معمولا به تعیین فاصله میان متون و خوشه بندی آن ها می پردازند و یا در برخی از پژوهش ها از روش های مدل سازی موضوعی برای حل این مسیله استفاده می کنند. هدف این روش ها در نهایت تقسیم بندی این متون به خوشه های مختلف است به شکلی که هر خوشه شامل متونی باشد که از نظر موضوع به هم نزدیک باشند. از جمله روش های مورد استفاده برای خوشه بندی اسناد K-medoids است که این گونه از روش های خوشه بندی به انتخاب مراکز اولیه حساس بوده و با انتخاب مراکز اولیه مختلف نتیجه ی خوشه بندی تغییر می کند.در این مقاله یک روش تشخیص موضوع ارایه می شود که در این روش ابتدا برای تعیین فاصله میان اسناد از یکی از روش های مدل سازی موضوعی یعنی LDA (Latent Dirichlet Allocation)) استفاده می کنیم. با بهره گیری از توزیع LDA اسناد، فاصله میان اسناد محاسبه شده و از روی آن گراف اخبار که نشان دهنده ی میزان شباهت میان اخبار است تولید می شود. گراف حاصل توسط الگوریتم K-medoids خوشه بندی می شود. با توجه به حساس بودن این گونه از روش های خوشه بندی به مراکز اولیه، با استفاده از DivRank که یک روش گام برداری تصادفی تقویتی است مراکز اولیه مناسب مشخص می شوند و در اختیار الگوریتم K-medoids قرار می گیرند. آزمایش های ما بر روی مجموعه دادگان مختلف نشان می دهد که روش ما در نحوه ی تولید گراف و یافتن مراکز اولیه ی مناسب برای الگوریتم K-medoids در مجموع در روند تشخیص موضوع بهبود ایجاد می کند و در مقایسه با انتخاب تصادفی مراکز اولیه، با احتمالی بین ۷۰% تا ۹۲% (بسته به مجموعه دادگان متفاوت) به معیار F بالاتری می توان دست یافت.

کلمات کلیدی:

تشخیص موضوع، DivRank ،LDA (Latent Dirichlet Allocation)، خوشه بندی، معیار فاصله ،تعیین مراکز اولیه، K-medoids

لینک ثابت مقاله در پایگاه سیویلیکا: