



بسط دادن مخفف ها مبتنی بر دامنه با استفاده از مدل سازی موضوعی برای پاکسازی داده ها

علی عجبی آ* افسانه فاطمی آ* محمدعلی نعمتبخش آ

آگروه تحقیقاتی کلان داده، گروه مهندسی نرم افزار، دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران.

چکیده

اطلاعات مقاله

تاریخچه مقاله:

دریافت: 28 January 2024

اصلاح: 11 June 2024

پذیرش: 14 July 2024

انتشار آنلاین: 23 October 2024

کلمات کلیدی:

بسط دادن مخفف، ابهام زدایی داده ها، پاکسازی داده ها، مدل سازی موضوع، مدل LDA

پاکسازی داده ها یک فرآیند ضروری در تجزیه و تحلیل داده ها و مدیریت است و نقش اساسی در به دست آوردن داده های قابل اعتمادتر برای تجزیه و تحلیل و تجارت بیشتر دارد. مهم ترین چالش با داده های بدون ساختار متنی، نحوه سرمایه گذاری و فعال سازی پاکسازی داده ها است، جایی که حجم زیادی از داده های متنی به سرعت تولید می شوند که نیاز فوری به پاکسازی دارند تا بتوان از آن برای به دست آوردن دانش استفاده کرد. در داده های متنی، مخفف ها بیشتر و بیشتر در مجموعه داده های مختلف ظاهر می شوند و فرآیند ابهام زدایی اختصار را می توان به عنوان یک موضوع مهم در تجزیه و تحلیل داده ها و به دست آوردن داده های قابل اعتمادتر در نظر گرفت. بسیاری از رویکردهای بسط مخفف ها برای رسیدگی به این موضوع پیشنهاد شده است، اما این رویکردها هرگز به حوزه ای که این مخفف ها به آن تعلق دارند توجه نکرده اند. برای غلبه بر این اشکال، یک روش بسط دادن کخفف ها مبتنی بر دامنه با استفاده از مدل سازی موضوع در این مقاله پیشنهاد شده است. در این روش مدل سازی موضوع ابتدا بر روی متن اعمال می شود، سپس دامنه بر اساس سهم موضوعات تعیین می شود. این باعث کاهش فضای جستجو می شود. در نهایت، بسط به مخفف ها با توجه به دامنه آنها اعمال می شود. روش پیشنهادی با اعمال آن در مجموعه داده های توییتهای COVID-19 و استفاده از طبقه بندی کننده رگرسیون لجستیک تأیید شده است. دو نوع مقایسه انجام شده است، با خود مجموعه داده قبل از استفاده از روش پیشنهادی و با نتایج رویکرد CrowdCorrect. نتایج نشان دهنده بهبود واضحی در دقت و یادآوری و همچنین دقت طبقه بندی کننده است.

© 2024. تمامی حقوق محفوظ است.

* نویسنده مسئول.

آدرس رایانامه: a_fatemi@eng.ui.ac.ir (ا. فاطمی)

تمامی حقوق محفوظ است. ISSN: 2322-4460 © 2024

