عنوان مقاله:
تشخیص سرقت علمی متون فارسی با رویکرد مبتنی بر بردار کلمات

محل انتشار:
نهمین کنفرانس فناوری اطلاعات و دانش (IKT 2017) (سال: 1396)

تعداد صفحات اصل مقاله: 9

نویسندگان:
محبوبه گلچین پور - *دانشجوی کارشناسی ارشد دانشگاه تهران*

هادی ویسی - *استادیار، عضو هییتعلمی دانشگاه تهران*

مصطفی صالحی - *استادیار، عضو هییت علمی دانشگاه تهران*

خلاصه مقاله:
گسترش اینترنت و دسترسی سریع و آسان به انبوه داده های متنی، سرقت علمی را به معضلی جدی و روبه رشد تبدیل کرده است. از این رو در این مقاله تابع فاصله جدیدی به نام فاصله برداری کلمات که مبتنی بر یادگیری عمیق است، برای تشابه یابی و تشخیص سرقت علمی متون فارسی پیشنهاد می گردد. این روش کلمات را به صورت بردارهایی در فضای N بعدی تعبیه و تشابه دو سند متنی را به صورت میانگین فاصله کسینوسی موردنیاز برای حرکت از کلمات تعبیه شده سند اول، برای رسیدن به کلمات مشابه شان در سند دوم تعریف میکند. روش فاصله برداری کلمات به آسانی می تواند تشابه اسناد متنی با کلمات مختلف ولی با مفهوم مشابه را تشخیص دهد. با استفاده از این روش دو سند متنی که حداکثر تشابه کسینوسی را نسبت به هم داشته باشند، مشابه نامیده و سرقت علمی تشخیص داده میشود. یکی از ضعف های روش ارایهشده عدم در نظر گرفتن طول رشته های متنی مورد مقایسه می باشد، از این رو با توجه به مزیت روش لونشتاین در بررسی تطابق کاراکتری رشته های متنی با طولهای مختلف، در این مقاله از روش لونشتاین به منظور کاهش خطای روش فاصله برداری کلمات استفاده شدهاست. نتایج استفاده از ترکیب این دو روش تشابه یابی، برای تشخیص سرقت علمی متون فارسی روی پیکره مبتنی PAN2015 دارای معیار F%97/9 می باشد.

کلمات کلیدی:
یادگیری عمیق، بازنمایی برداری کلمات، تشابه یابی، سرقت علمی، بردار کلمه