

عنوان مقاله:

کشف سرقت ادبی در متون فارسی به کمک الگوریتمهای اثرانگشت

محل انتشار:

بیست و یکمین کنفرانس مهندسی برق ایران (سال: 1392)

تعداد صفحات اصل مقاله: 6

نویسندگان:

کبری کامران - دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران

علی احمدی - دانشگاه صنعتی خواجه نصیرالدین طوسی

فاطمه کاظمی ونهری

خلاصه مقاله:

سرقت ادبی یکی از معضلات بزرگ در عصر ارتباطات است. در خیلی از زبانها از جمله زبان انگلیسی این مساله به طور جدی مورد اهمیت بوده و ابزارهای قدرتمندی برای جلوگیری از این مساله تهیه شده است، اما متأسفانه تاکنون ابزاری برای کشف این مساله در زبان فارسی طراحی نشده است. این مقاله به کشف سرقت ادبی در متون فارسی براساس الگوریتمهای اثرانگشت میپردازد. الگوریتمهای اثرانگشت به خاطر ایجاد ویژگیهای فشرده و صرفاً مقایسه این ویژگیها بین اسناد اصل و مشکوک، سرعت بالایی در کشف سرقت ادبی دارند. به منظور افزایش دقت و سرعت کشف سرقت ادبی، میبایست قبل از بکارگیری الگوریتم عملیات پیشپردازش شامل: حذف تگهای HTML، نسخه اینترنتی مقاله، جداسازی کلمات، یکسانسازی، جایگزینی اعداد، حذف واژههای عمومی، و ریشهیابی روی متن صورت گیرد. در این مقاله چهار الگوریتم RareChunk، SimHash، Winnowing، SPEX بررسی و از نظر کارایی در زبان فارسی با یکدیگر مقایسه میشوند. پیادهسازی این الگوریتم-ها روی سیصد مقالات ثبت شده در پایگاه داده نور با موضوعات هنر، تاریخ و روانشناسی، نتایج قابل قبولی را نشان میدهد.

کلمات کلیدی:

سرقت ادبی، اثرانگشت، simHash، winnowing، SPEX، RareChunk

لینک ثابت مقاله در پایگاه سیویلیکا:

<https://civilica.com/doc/208842>

