

INTISARI

DETEKSI KEMIRIPAN ARTIKEL BERTEMA ISLAM DENGAN METODE BOYER MOORE DAN COSINE SIMILARITY

Oleh

ADI ZANUWAR

17/409423/PA/17730

Artikel islami merupakan tulisan yang membahas seputar islam yang dapat ditemukan di berbagai platform. Masih banyak artikel islam di blog pribadi memiliki kemiripan dari satu situs dengan situs lain. Selain itu, *publisher* yang bermitra dengan *google adsense* harus menjaga kualitas artikel bebas plagiarisme. Dalam menghadapi permasalahan tersebut, dapat dibuat sistem dekteksi plagiarisme pada artikel bertema islam.

Penelitian ini mengimpelementasikan metode boyer moore sebagai filter data dan *cosine similarity* sebagai uji similaritas serta *euclidean distance* sebagai pembanding uji similaritas. Boyer moore dimanfaatkan sebagai filter data yang bertujuan untuk menemukan artikel dari koleksi data yang memiliki topik serupa dengan dokumen uji. Selain metode tersebut juga memanfaatkan pembobotan TF-IDF.

Pengujian dilakukan dari kinerja kebutuhan waktu dan nilai evaluasi *precision*, *recal*, *accuracy*, dan *f1-score*. Dengan filter mampu memangkas waktu yang dibutuhkan setiap tahapan pembobotan, dan uji similaritas. Sedangkan menggunakan evaluasi maka algoritma terbaik untuk uji similaritas adalah *cosine similarity* dengan mengacu *threshold* 60%, diperoleh nilai *precision* 86,7%, *recall* 86,7%, *accuracy* 86,7%, dan *f1-score* 86,7%. Akan tetapi sistem deteksi plagiarisme yang dibangun tidak cukup baik, karena deteksi plagiarisme yang dibangun membutuhkan waktu kerja rata-rata 3 menit 42. Karena pada tahap *preprocessing text* tidak melalui filter data, dimana memerlukan 3 menit 32 detik, yang artinya lama waktu deteksi plagiarisme di tahap *preprocessing text*.

Kata Kunci: Deteksi Kemiripan, Cosine Similarity, Boyer Moore, Atrikel Islam, Euclidean Distance

ABSTRACT

DETECTION OF SIMILARITY OF ISLAMIC THEME ARTICLES WITH THE BOYER MOORE METHOD AND COSINE SIMILARITY

Oleh

ADI ZANUWAR

17/409423/PA/17730

Islamic articles are writings that discuss various aspects of Islam and can be found on multiple platforms. Many Islamic articles on personal blogs often resemble those from other sites. Additionally, publishers partnering with Google AdSense must maintain high-quality, plagiarism-free content. To address this issue, a plagiarism detection system for Islamic-themed articles can be developed.

This research implements the Boyer-Moore method as a data filter and cosine similarity for similarity testing, with Euclidean distance as a comparative measure for similarity testing. Boyer-Moore is utilized as a data filter to identify articles from a data collection that have topics similar to the test document. Additionally, the method leverages TF-IDF weighting.

The system's performance was evaluated based on time efficiency and evaluation metrics such as precision, recall, accuracy, and F1-score. The filtering process successfully reduced the time required for each weighting stage and similarity test. The best algorithm for similarity testing was found to be cosine similarity, with a 60% threshold yielding precision, recall, accuracy, and F1-score values of 86.7%. However, the developed plagiarism detection system is not entirely effective, as it requires an average processing time of 3 minutes and 42 seconds. This inefficiency is primarily due to the preprocessing stage not incorporating data filtering, taking up 3 minutes and 32 seconds, indicating that the majority of the time spent in plagiarism detection is during the preprocessing text stage.

Keywords: Similarity Detection, Cosine Similarity, Boyer-Moore, Islamic Articles, Euclidean Distance