ABSTRAK

Nur Fajar, NIM 105841101121. Penerapan Sistem Pencarian Dokumen berdasarkan Frasa di Abstrak Perpustakaan Digital menggunakan Algoritma BM25 dan Word2Vec. Program Studi Informatika, Fakultas Teknik Universitas Muhammadiyah Makassar. Di bimbing oleh Fahrim Irhamna Rachman dan Ida.

Perkembangan perpustakaan digital menyebabkan meningkatnya volume abstrak dokumen sehingga menuntut metode pencarian yang akurat untuk menemukan buku relevan. Penelitian ini mengusulkan penerapan sistem pencarian berbasis frasa pada abstrak dengan menggabungkan algoritma BM25 dan Word2Vec untuk meningkatkan relevansi hasil. Dataset terdiri dari 500 abstrak skripsi yang dipreproses (lowercasing, tokenisasi, stopword removal); model Word2Vec dilatih dengan arsitektur skip-gram (vector_size=100, window=5, epochs=50) dan BM25 diinisialisasi pada representasi token dokumen. Skor BM25, Word2Vec (cosine similarity) dan TF-IDF dinormalisasi lalu digabungkan (rata-rata) untuk pemeringkatan akhir. Evaluasi dilakukan menggunakan metrik Precision, Recall dan F1-Score pada beberapa query uji. Hasil menunjukkan peningkatan performa pada banyak query (rata-rata F1 ≈ 0.80) dengan beberapa kasus mencapai nilai sempurna (1.00), meskipun ada variabilitas antar tipe query. Temuan ini menegaskan bahwa penggabungan pencocokan lesikal BM25 dan representasi semantik Word2Vec dapat meningkatkan relevansi pencarian; pengembangan lanjutan direkomendasikan pada metode penggabungan skor dan perluasan korpus.

Kata Kunci: Pencarian informasi, bm25, word2vec, tf-idf

ABSTRACT

Nur Fajar, NIM 105841101121. Implementation of a Phrase-Based Document Search System in Digital Library Abstracts Using the BM25 and Word2Vec Algorithms. Informatics Study Program, Faculty of Engineering, Muhammadiyah University of Makassar. Supervised by Fahrim Irhamna Rachman and Ida.

The development of digital libraries has led to an increase in the volume of document abstracts, thus demanding accurate search methods to find relevant books. This study proposes the implementation of a phrase-based search system on abstracts by combining the BM25 and Word2Vec algorithms to improve the relevance of the results. The dataset consists of 500 thesis abstracts that were preprocessed (lowercasing, tokenization, stopword removal); the Word2Vec model was trained with a skip-gram architecture (vector_size=100, window=5, epochs=50) and BM25 was initialized on the document token representation. The BM25, Word2Vec (cosine similarity) and TF-IDF scores were normalized and then combined (averaged) for the final ranking. Evaluation was performed using Precision, Recall and F1-Score metrics on several test queries. The results show improved performance on many queries (average F1 ≈ 0.80) with some cases achieving a perfect score (1.00), although there is variability between query types. These findings confirm that combining BM25 lexical matching and Word2Vec semantic representation can improve search relevance; further development of the score fusion method and corpus expansion is recommended.

Keywords: Information retrieval, bm25, word2vec, tf-idf