

“ALGORITMA IMPROVED SQRT-COSINE SIMILARITY PADA SISTEM TEMU KEMBALI DOKUMEN PENELITIAN”

Wahid Arinanto Nugroho
(15.04.1.1.1.00121)

Pembimbing I : Eka Mala Sari Rochman, S.Kom., M.Kom.
Pembimbing II : Husni, S.Kom., M.T.

ABSTRAK

Banyaknya dokumen penelitian yang terkumpul dalam bentuk *soft file* mengakibatkan proses pencarian secara manual menjadi sulit. Pencarian pada *database* biasanya hanya mencari judul penelitian yang sesuai dengan kata kunci (*keyword*) yang diberikan, namun sistem tidak dapat mengukur dokumen yang paling mirip. Sistem temu kembali informasi merupakan salah satu solusi yang dapat membantu dalam mendapatkan informasi yang diinginkan. Dengan menerapkan sistem temu kembali informasi, diharapkan permasalahan pencarian dokumen penelitian dapat memberikan hasil yang relevan dan sesuai dengan kebutuhan pengguna. Salah satu masalah yang telah diselesaikan adalah sistem temu kembali dokumen menggunakan algoritma *Cosine Similarity* (CS). Perhitungan pada algoritma CS mengacu pada jarak *Euclidean*. Secara umum jarak *Euclidean* bukan matrik yang efektif dalam menangani probabilitas, dimana sering digunakan dalam analisis teks. Selain algoritma CS, algoritma lain untuk menghitung kemiripan dokumen adalah algoritma *Sqrt-Cosine Similarity* (SCS). Namun dalam beberapa kasus, algoritma SCS bertentangan dengan definisi dari pengukuran kemiripan itu sendiri. Penelitian ini menggunakan metode pembobotan *TF-IDF* dan algoritma yang digunakan untuk menghitung kemiripan antara dokumen dengan *keyword* adalah algoritma *Improved SQRT-Cosine Similarity* (ISC). Berdasarkan hasil pengujian yang telah dilakukan, diperoleh kesimpulan bahwa tingkat relevansi suatu dokumen yang dicari pada dasarnya tergantung pada pengguna sendiri. Penggunaan algoritma ISC yang telah diteliti ini sudah dapat mencari dokumen yang dianggap relevan dan sesuai dengan *keyword* yang diberikan, selisih nilai presisi yang didapatkan antara penggunaan algoritma ISC dan algoritma CS yaitu sebesar 5%. Nilai relevansi yang diperoleh setelah menggunakan algoritma ISC berdasarkan pengujian nilai rata-rata presisi yaitu sebesar 83%. Sedangkan penggunaan algoritma sebelumnya yaitu algoritma CS didapatkan nilai rata-rata presisi sebesar 78%.

Kata Kunci: Pencarian Dokumen, Temu Kembali Informasi, Kemiripan Dokumen, *Tf-Idf*, *Improved Sqrt-Cosine Similarity*.

“IMPROVED SQRT-COSINE SIMILARITY ALGORITHM ON RESEARCH DOCUMENT INFORMATION RETRIEVAL SYSTEMS”

Wahid Arinanto Nugroho
(15.04.1.1.1.00121)

Pembimbing I : Eka Mala Sari Rochman, S.Kom., M.Kom.
Pembimbing II : Husni, S.Kom., M.T.

ABSTRACT

The large number of research documents collected in softfiles has made the search process manually becomes difficult. Searching on database are usually just looking the title of the research titles that match the given keywords, but the system cannot measure the most similar document. Information Retrieval is one solution that can help in getting the desired information. By implementing an information Retrieval system, expected research document search problems can provide relevant results and in accordance with user needs. One problem that has been resolved is the document retrieval system using the Cosine Similarity (CS) algorithm. CS algorithm based on Euclidean distance. However, Euclidean distance is generally not an effective metric for dealing with probabilities, which are often used in text analytics. Besides CS algorithm, another algorithm to calculate document similarity is Sqrt-Cosine Similarity (SCS) algorithm. But in some cases, the manner of SCS is in conflict with the definition of similarity measurement. This research uses the TF-IDF weighting method and the algorithm that used to calculate the similarity between documents and keywords is the Improved Sqrt-Cosine Similarity (ISC) algorithm. Based on the results of the test that has been done, it is concluded that the relevance of a document sought is basically dependent on the user himself. The use of this researched ISC algorithm is already able to find a documents that are considered relevant and in accordance with the given keywords, the difference in the precision value obtained between the use of ISC algorithm and CS algorithm is 5%. The relevance value obtained after using the ISC algorithm based on testing the average precision is 83%. While the use of the previous CS algorithm, obtained an average precision value of 78%.

Keyword: Document Search, Information Retrieval, Document Similarity, TF-IDF, Improved Sqrt-Cosine Similarity.