

Re-Ranker Berbasis Fitur untuk Information Retrieval pada Domain Legal = Feature Based Re-Ranker for Legal Domain Retrieval

Kenneth Jonathan, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920567386&lokasi=lokal>

Abstrak

Terdapat beberapa masalah yang muncul seiring dengan bertambahnya peraturan. Hal tersebut menyebabkan proses pengumpulan dan evaluasi peraturan memakan waktu yang relatif lebih lama. Oleh karena itu, diperlukan suatu sistem yang dapat mengotomatiskan kebutuhan tersebut, salah satunya adalah Information Retrieval. Penelitian ini bertujuan untuk meningkatkan efektivitas sistem Information Retrieval melalui pendekatan re-ranker berbasis fitur dengan memanfaatkan beberapa jenis fitur, seperti atribut kuantitatif sederhana, skor text matching, dan document embeddings. Ditemukan bahwa skor kesamaan Jaccard, nilai relevansi BM25 dan nilai relevansi LemurTF_IDF merupakan karakteristik yang dapat membantu peningkatan efektivitas re-ranking secara konsisten dalam domain legal. Sementara itu, fitur yang memanfaatkan embeddings dari BERT maupun T5 didapatkan bermanfaat, namun memiliki kontribusi yang lebih kecil dari fitur perhitungan sederhana seperti kesamaan Jaccard. Selain itu, didapatkan bahwa pemanfaatan seluruh fitur sebagai masukan dari re-ranker LambdaMART dapat meningkatkan seluruh metrik sistem sekitar 4,17% secara signifikan dengan nilai metrik utama, recall@3, tertinggi diperoleh DLH13 (Reranker) dengan nilai 0,6632 dan peningkatan sebesar 5,64%. Namun, saat dilakukan percobaan menggunakan hanya ketiga fitur tersebut, didapatkan peningkatan sebesar 3,739% yang tidak signifikan.There are several issues that arise with the increasing number of regulations. This causes the process of collecting and evaluating regulations to take relatively longer. Therefore, a system is needed to automate these needs, one of which is Information Retrieval. This research aims to improve the effectiveness of the Information Retrieval system through a feature-based re-ranker approach by utilizing several types of features, such as simple quantitative attributes, text matching scores, and document embeddings. It was found that Jaccard similarity scores, BM25 relevance values, and LemurTF_IDF relevance values are characteristics that can consistently help improve re-ranking effectiveness in the legal domain. Meanwhile, features that utilize BERT and T5 embeddings were found to be beneficial but contributed less than simple calculation features like Jaccard similarity. Additionally, it was found that using all the features as input for the LambdaMART re-ranker can significantly improve all system metrics by about 4,17%, with the highest main metric value, recall@3, achieved by DLH13 (Reranker) with a value of 0,6632 and an increase of 5,64%. However, when experiments were conducted using only the three features mentioned, an insignificant increase of 3,739% was obtained.