# ABSTRACT

Question answering (QA) system is a system designed to answer questions given by users on the document source provided. The challenge for QA systems in the era of big data today is that the number of documents is very large and stored in an unstructured format, which can become a nuisance that causes the answer source document not found. The aim of this research is to improve the accuracy of the document retrieval phase in Indonesian QA system in unstructured document collections.

The method used is query expansion, namely adding a list of defined keywords into the query based on the class of the query. Moreover, we employed boolean retrieval technique with "OR" boolean. In this research, there are three general stages, the first is preprocessing to create inverted index and TFIDF vector of document collection. The second stage is question analysis to extract the list of keywords and the class of the query, and query expansion. The third stage is document retrieval to retrieve list of relevant documents to the query using the “OR” boolean and calculate the similarity weight between all expansion queries and relevant documents. The data used were 66097 news articles and 90 questions. Each question was tested using all datasets, and the evaluation metrics used are recall, which indicates the existence of the document, and MRR which shows the position of the document being searched for on the list of documents returned by the system.

After conducting experiment on several thresholds for the number of documents returned, the best results were obtained at the threshold of 20 documents, with a 76.67% recall and 38.69% MRR. The analysis shows that 43.3% of the expanded queries successfully increase the similarity score to the answer source documents. Furthermore, the use of the boolean query "OR" significantly increase the recall score from 36.67% to 76.67%.

**Keywords** - question answering system, query expansion, information retrieval, search engine, natural language processing.

# INTISARI

*Question Answering* (QA) *system* adalah sistem yang dirancang untuk menjawab pertanyaan yang diberikan pada sumber dokumen yang disediakan. Tantangan untuk QA *system* pada *era big* data saat ini adalah jumlah dokumen sangat banyak dan tersimpan dalam format tidak terstruktur, yang dapat menjadi gangguan yang menyebabkan dokumen sumber jawaban tidak ditemukan. Fokus dari penelitian ini adalah meningkatkan akurasi pemilihan dokumen QA *system* bahasa Indonesia pada koleksi dokumen tidak terstruktur.

Metode yang digunakan adalah metode ekspansi kueri, yaitu menambahkan sejumlah kata kunci yang sudah didefinisikan ke dalam kueri berdasarkan kelas yang dimiliki. Sistem juga menggunakan *boolean retrieval* dengan *boolean query* "OR". Terdapat tiga tahap umum, pertama adalah *preprocessing* untuk membuat *inverted index* dan vektor TFIDF dari dokumen. Tahap kedua yaitu *question analysis* untuk mengekstrak daftar kata kunci dan kelas dari kueri, dan ekspansi kueri. Tahap ketiga yaitu *document retrieval* untuk mengambil daftar dokumen relevan dengan *boolean* "OR" dan menghitung bobot similaritas antara semua kueri ekspansi dan dokumen relevan. Data yang digunakan adalah 66097 artikel berita dan 90 pertanyaan. Masing-masing pertanyaan diuji menggunakan semua dataset yang tersedia, dan metrik evaluasi yang digunakan adalah *recall* yang menunjukkan keberadaan dokumen, serta MRR yang menunjukkan posisi dokumen yang dicari pada daftar dokumen yang ditemu kembalikan.

Setelah melakukan uji coba pada beberapa *threshold* jumlah dokumen yang ditemu kembalikan, hasil terbaik didapatkan pada *threshold* 20 dokumen, dengan *recall* 76,67% dan MRR 38,69%. Hasil analisis juga menunjukkan sebanyak 43,3% kueri yang telah diekspansi berhasil meningkatkan bobot similaritas dengan dokumen sumber jawaban. Selain itu, penggunaan *boolean query* "OR" berhasil meningkatkan bobot *recall* secara signifikan dari 36,67% menjadi 76,67%.

**Kata kunci –** sistem tanya jawab, ekspansi kueri, sistem temu kembali, mesin pencari, pemrosesan bahasa