INTISARI

ANOTASI OTOMATIS BERBASIS TESAURUS DAN ONTOLOGI PADA DOKUMEN PDF

Diah Sekarsari Rahmaniati

13/356418/PPA/04399

Penyimpanan dokumen secara digital telah menjadi hal yang umum dilakukan. Berbagai perangkat lunak dikembangkan untuk memudahkan pengelolaan dokumen. Salah satu format digital yang dapat digunakan untuk menyimpan dokumen adalah *Portable Document Format* (PDF).

Penelitian ini bertujuan untuk memudahkan pencarian kembali dokumen PDF yang disimpan secara digital. Anotasi secara otomatis diberikan pada dokumen PDF dengan tujuan untuk mengoptimalkan hasil pencarian dokumen. Ontologi, WordNet dan DBpedia digunakan sebagai basis pengetahuan untuk mendapatkan kata yang valid. Dokumen PDF di ekstraksi untuk mendapatkan isi di dalam dokumen. Hasil ekstraksi isi dokumen ini kemudian dijadikan kumpulan token dan ditelusuri dengan menggunakan basis pengetahuan. Penelusuran dengan menggunakan basis pengetahuan ini dapat memberikan hasil kata yang ambigu. Kata-kata yang ambigu ini kemudian di saring dengan menggunakan metode *specification* marks, *hypernym/hyponym* heuristic, *definition* heuristic dan *gloss hypernym/hyponym heuristic*. Hasil anotasi dari proses penelusuran melalui metode-metode *word sense disambiguation* (WSD) di atas kemudian disaring kembali dengan memanfaatkan kata kunci. Hasil anotasi yang memiliki hubungan *hypernym* atau domain dari kata kunci ini dijadikan sebagai hasil anotasi akhir.

Anotasi yang telah didapat kemudian disematkan ke dalam dokumen PDF. Hasil pencarian berdasarkan anotasi otomatis menunjukkan *recall* sebesar 0,75 dan presisi sebesar 0,16 serta *recall* sebesar 0,45 dan presisi 0,05 untuk pencarian dengan menggunakan pendekatan *vector space model*. Waktu rata-rata pencarian dengan anotasi otomatis adalah 44,78 s dan waktu rata-rata pencarian dengan menggunakan metode *vector space model* adalah 35,67 s.

Kata Kunci: Anotasi Otomatis, Pengelolaan Dokumen, Perangkat Lunak, Temu Balik Dokumen

ABSTRACT

AUTOMATIC ANNOTATION BASED ON THESAURUS AND ONTOLOGY FOR PDF DOCUMENT

Diah Sekarsari Rahmaniati

13/356418/PPA/04399

Storing document in digital format is a common activity nowadays. There are a lot of software developed to ease document management. One of the common digital document extention is Portable Document Format (PDF).

The purpose of this research is to facilitate PDF document retrieval. Annotation was embedded in the PDF document to optimize the document retrieval results. Ontology, WordNet dan DBpedia are the knowledge-based resources being used to validate the words. PDF document extraction was performed to extract document's text. This text extraction result will be stored as token. After the tokenization process, comparation between the knowledge-based resources and tokenization results will be performed to validate each words inside the tokenization results. This comparation may giving an ambiguouse result from a single word. Specification marks method, hypernym/hyponym heuristic, definition heuristic and gloss hypernym/hyponym heuristic methods are being used to solved the issue. Results from the word sense disambiguation (WSD) methods will be filtered using keywords. The annotation results that have relation with the keywords will be the annotation final results.

Automatic annotation results will be stored inside the document as document information properties. The retrieval result showed that annotation based recall was 0,75 and annotation based precision was 0,16. Recall for vector space model approach was 0,45 and its precision was 0,05. Average time needed to proceed annotation based document retrieval is 44,78 s and 35,67 s for vector space model.

Keywords: Automatic Annotation, Document Management, Software, Document Retrieval