INTISARI

TEMU KEMBALI INFORMASI BERBASIS SEMANTIK

STUDI KASUS : DOKUMEN PERUNDANG-UNDANGAN

Oleh
Eri Zuliarso

07/261373/SPA/162

Disertasi ini membahas model perolehan informasi berbasis semantik untuk dokumen perundangan. Dokumen perundangan mempunyai perbedaan dengan dokumen teks yang lain. Perbedaan itu antara lain dokumen perundangan mempunyai judul dan kata yang didefinisikan secara khusus untuk perundangan tersebut.

Penelitian ini dibagi menjadi empat bagian utama : membangun basisdata leksikal mengikuti format WordNet Princeton, mengkonstruksi ontologi yang mampu menyimpan informasi kamus leksikal, membangun model perolehan informasi berbasis semantik Bahasa Indonesia, dan membangun model perolehan informasi berbasis semantik lintas-bahasa Bahasa Indonesia dan Bahasa Inggris.

Untuk membangun basisdata leksikal mengikuti format Princeton WordNet dilakukan dengan cara 1. membangun thesaurus berdasarkan Kamus Besar Bahasa Indonesia, 2. membangun wordnet Bahasa Indonesia memanfaatkan Princeton WordNet, Wikipedia, Kamus Besar Bahasa Indonesia, 3. menambah leksikon dengan memanfaatkan dokumen perundangan yang tersedia dalam Bahasa Indonesia dan Bahasa Inggris.

Informasi tentang perundang-undangan dan informasi leksikal untuk suatu kata direpresentasikan dalam ontologi berdasar arsitektur WordNet dan disimpan dengan format Ontology Web Language (OWL). Komponen utama WordNet dan hirarki perundang-undangan ditransformasikan sebagai kelas dalam OWL. Relasi diantara himpunan sinonim, kata leksikal, perundang-undangan ditransformasikan sebagai properti OWL. Ontologi mempunyai hirarki kelas yang sederhana.

Struktur judul, definisi kata dan isi dalam perundangan digunakan untuk mencari kombinasi perluasan query yang menghasilkan presisi dan recall terbaik. Model Perolehan Informasi Berbasis Semantik *Bahasa Indonesia* dilakukan dengan memanfaatkan hubungan sinonim untuk memperluas query. Perluasan query pada judul dan kata di bagian ketentuan umum mendapatkan *presisi dan recall* terbaik. Query dengan kata yang mempunyai makna yang ambigu dalam kamus akan menyebabkan presisi dan recall yang rendah.

Model perolehan informasi berbasis semantik *cross-lingual* menterjemahkan term dalam query berdasarkan prioritas dengan urutan penterjemahan berdasarkan kata yang ada di judul perundang-undangan, kata di bagian ketentuan umum perundangan dan kamus bilingual. Kata yang dihasilkan dari hasil penterjemahan selanjutnya diperluas dengan memanfaatkan hubungan sinonim, hiperonim dan hiponim dari Princeton WordNet. Perluasan query pada judul dan kata di bagian ketentuan umum mendapatkan *presisi dan recall* terbaik. Query dengan kata yang mempunyai makna yang ambigu dalam kamus akan menyebabkan query hasil terjemahan yang ambigu. Query hasil terjemahan yang ambigu menyebabkan perolehan dokumen dengan relevansi yang rendah.

Kata kunci : Basisdata leksikal, WordNet, Ontology Web Language, perolehan informasi berbasis semantik

ABSTRACT

SEMANTIC BASED INFORMATION RETRIEVAL

CASE STUDY : LEGAL DOCUMENTS

By
Eri Zuliarso
07/261373/SPA/162

This dissertation addresses semantic-based information retrieval models for legal documents. Legal documents have differences with other text document since they have title and words which are defined specifically for that document.

This study comprises four main parts: building a database using Princeton WordNet architecture, constructing ontologies capable of storing information lexical dictionaries, build a model Indonesian semantic-based information retrieval, and build model Indonesian-English cross-language semantic based information retrieval.

Lexical database with Princeton WordNet architecture is constructed by 1. build a synonym by Indonesian Dictionary, 2. building Indonesian wordnet utilizing Princeton WordNet, Wikipedia, Indonesian Dictionary, 3. add to the lexicon by utilizing legal documents available in Indonesian and English. Every word stored in the lexical databases has links between languages, between words, the links between languages between words based on legislation, the links between the legislation.

Word lexical information and legal information represented in WordNet ontology-based architecture and saved in Web Ontology Language (OWL) format. The main components of WordNet and the hierarchy of regulation transformed as classes in OWL. The relation between set of synonyms, lexical words, and regulation transformed as OWL properties. Ontologies have a simple class hierarchy.

The structure of title, definition of the word and contents in the regulation used to find combinations of query expansion that produces the best precision and recall. Indonesian semantic based information retrieval use synonyms relationship to expand the query. Query expansion in title and general provisions get the best precision and recall. Query with words that have ambiguous meanings in the dictionary will lead to low precision and recall.

Cross-lingual semantic based information retrieval translate terms in a query based on priority. The priority order based on the existing translation in title of regulation, in general provisions of regulation and bilingual dictionaries. Result term from the translation expanded use synonyms, hyponym and hypernym relationship of Princeton WordNet. Query expansion in title and general provisions get the best precision and recall. Query with ambiguous words in the dictionary produce ambiguous query. Query results from ambiguous translation retrieve documents with low relevance.

Keywords: *Lexical database, WordNet, Web Ontology Language, semantic-based information retrieval*