Knowledge dictionary for information extraction on the Arabic text data = Knowledge dictionary untuk ekstraksi informasi pada data teks Arab

---

Abstrak

Ekstraksi informasi merupakan sebuah tahap awal dari proses analisis data tekstual. Ekstraksi informasi diperlukan untuk mendapatkan informasi dari data tekstual sehingga dapat digunakan untuk proses analisis seperti misalnya klasifikasi dan kategorisasi. Data tekstual
sangat dipengaruhi oleh bahasa, jika sebuah data tekstual berbahasa Arab maka karakter yang digunakan adalah karakter arab.
Knowledge dictionary merupakan sebuah kamus yang dapat digunakan untuk mengekstraksi informasi dari data tekstual. Informasi yang diekstraksi menggunakan knowledge dictionary adalah konsep.
Knowledge dictionary biasanya dibangun secara manual oleh seorang pakar yang tentunya membutuhkan waktu yang lama dan spesifik untuk
setiap masalah. Pada penelitian ini diusulkan sebuah metode untuk membangun knowledge dictionary secara otomatis. Pembentukan
knowledge dictionary dilakukan dengan cara mengelompokkan kalimat yang memiliki konsep yang sama, dengan asumsi kalimat yang memiliki konsep yang sama akan memiliki nilai simi laritas yang tinggi.
Konsep yang telah diekstraksi dapat digunakan sebagai fitur untuk proses komputasi berikutnya misalnya klasifikasi ataupun kategorisasi.
Dataset yang digunakan dalam penelitian ini adalah dataset teks Arab. Hasil ekstraksi diuji dengan menggunakan mesin klasifikasi
decision tree dan didapatkan nilai presisi tertinggi 71,0% dan nilai recall tertinggi 75,0%.
<hr>
Abstract
Information extraction is an early stage of a process of textual data analysis. Information extraction is required to get information from textual data that can be used for process analysis, such as classification and categorization. A textual data is strongly influenced by the language. Arabic is gaining a significant attention in
many studies because Arabic language is very different from others, and in contrast to other languages, tools and research on the Arabic language is still lacking. The information extracted using the knowledge dictionary is a concept of expression. A knowledge dictionary is usually constructed manually by an expert and this would take a long time and is specific to a problem only. This paper proposed a method for automatically building a knowledge dictionary. Dictionary knowledge is formed by classifying sentences having the same concept, assuming that they will have a high similarity value. The concept that has been extracted can be used as features for subsequent computational process such as classification or categorization. Dataset used in this paper was the Arabic text dataset. Extraction result was tested by using a decision tree classification engine and the highest precision value obtained was 71.0% while the highest recall value was 75.0%.