Pemotong imbuhan berdasarkan korpus untuk kata bahasa Indonesia

Muhammad Ichsan, author

Abstrak

Information Retrieval (IR) secara umum merupakan suatu teknik untuk menemukan informasi di dalam kumpulan-kumpulan dokumen atau di dalam media-media lainnya dengan memberikan query atau pertanyaan berupa teks, suara, gambar atau bentuk-bentuk lainnya. Penerapannya yang paling sering dijumpai adalah search engine atau mesin pencari. Untuk meningkatkan jumlah dokumen yang diperoleh salah satunya dilakukan dengan menggunakan pemotong kata berimbuhan (stemmer). Stemmer merupakan salah satu alat bantu paling sederhana dalam bidang Information Retrieval. Stemmer digunakan untuk mendapatkan kata dasar atau bentuk yang lebih umum dari suatu kata sehingga mengurangi variasi kata pada dokumen-dokumen. Dengan demikian dokumen yang diinginkan akan semakin banyak diperoleh. Contohnya dokumen yang mengandung kata-kata berimbuhan pendapat, pendapatan, didapat dan sebagainya akan dirujuk oleh kata dasar yang sama yaitu dapat. Namun beberapa kata berimbuhan yang mempunyai kata dasar yang sama, memiliki makna yang berbeda. Sehingga kurang tepat apabila menyamakan seluruh variasi kata tersebut kepada kata dasarnya dengan menggunakan stemmer. Misalnya kata pendapat dengan pendapatan. Meskipun keduanya memiliki kata dasar yang sama, tapi hakikatnya keduanya memiliki makna yang sangat berbeda. Selain masalah perbedaan makna di atas, juga ada masalah terkait dengan jenis korpus. Jenis korpus yang dapat mempengaruhi makna kata. Misalnya, kata membintangi dan bintang. Pada korpus astronomi kata membintangi tidak mempunyai makna yang sama dengan kata bintang. Sebaliknya pada korpus perfilman kedua kata ini bermakna sama yaitu pemain film. Sebuah penelitian mengenai stemmer yang berdasarkan pada korpus telah dilakukan untuk menghindari penyamarataan makna variasi kata. Stemmer yang telah diujikan pada bahasa Inggris dan Spanyol tersebut telah meningkatkan efektifitas sistem IR dalam mendapatkan informasi. Stemmer ini disebut stemmer corpus-based dengan menggunakan statistik co-occurrence dari variasi kata. Pada tulisan ini penulis mencoba untuk menggunakan teknik yang sama untuk menghindari penyamarataan makna variasi kata pada bahasa Indonesia. Karena pada bahasa Indonesia terdapat banyak variasi kata yang berakar pada kata dasar yang sama, namun memiliki perbedaan makna. Penulis mencoba memperbaiki efektifitas penggunaan stemmer Indonesia yang sudah ada dengan teknik stemmer corpus-based dengan menggunakan statistik co-occurrence dari variasi kata. Penulis tidak melakukan penelitian pada masalah yang terkait dengan korpus topik tertentu karena keterbatasan korpus pada bahasa Indonesia. Berdasarkan pembahasan dan uji coba yang telah dilakukan dengan menggunakan korpus yang berisi dokumen dari Tempo dan Republika, dapat disimpulkan bahwa penggunaan stemmer corpus-based dengan menggunakan statistik co-occurrence dari variasi kata (SVC) hanya menunjukkan sedikit perbaikan pada efektifitas sistem IR. Dibandingkan dengan perbaikan yang diperoleh dengan menggunakan stemmer masing-masing, dengan bantuan SVC, pada stemmer morfologi untuk bahasa Malaysia terjadi peningkatan