Segmentasi dokumen bahasa indonesia menggunakan metode genetic algorithm

Vinky Halim, author

Abstrak

Segmentasi dokumen merupakan suatu proses untuk membagi dokumen menjadi bagian-bagian yang homogen atau memiliki keterkaitan yang tinggi. Pada tugas akhir ini digunakan genetic algorithm sebagai metode untuk melakukan segmentasi dokumen. Genetic algorithm merupakan suatu algoritma pencarian solusi terhadap permasalahan dengan search space yang besar dengan menggunakan pendekatan evolusi.

Penelitian tentang segmentasi dokumen menggunakan genetic algorithm telah dilakukan oleh Lamprier (Lamprier et al., 2007) terhadap dokumen bahasa Inggris dengan hasil yang memuaskan. Pada penelitian yang dilakukan Lamprier, proses segmentasi dilakukan dengan mengoptimisasi 2 fungsi objektif yaitu internal cohesion dan dissimilarity. Data yang digunakan pada percobaan ini terdiri dari dokumen artikel media massa Indonesia dan abstrak tulisan ilmiah dari Fakultas Ilmu Komputer Universitas Indonesia.

Percobaan ini dilakukan dan dianalisa dari beberapa aspek yaitu aspek fitness function, metode penghitungan similarity, jumlah iterasi, ukuran populasi, jumlah segmen, dan kemiripan antar dokumen penyusun. Selain itu dilakukan pula perbandingan hasil segmentasi antara metode genetic algorithm dengan metode Texttiling.

Hasil percobaan yang didapat adalah segmentasi dokumen menggunakan genetic algorithm dengan fitness function SPEA 2, metode penghitungan similarity menggunakan dice coefficient, jumlah iterasi 1000 iterasi, ukuran populasi 50 individu, tipe crossover two point crossover, dan probabilitas mutasi 0.09 memberikan hasil segmentasi terbaik. Pada percobaan untuk membandingkan 2 metode segmentasi yaitu genetic algorithm dan Texttiling diperoleh hasil precision 0.081 dan recall 0.46 untuk metode genetic algorithm dan precision 0.12 dan recall 0.58 untuk metode Texttiling.

Dari data hasil percobaan diperoleh kesimpulan bahwa hasil segmentasi dengan metode Texttiling lebih baik daripada hasil segmentasi dengan metode genetic algorithm. Hasil ini bertolak belakang dengan apa yang dilaporakan pada penelitian yang dilakukan Lamprier (Lamprier et al., 2007), hal tersebut dipengaruhi oleh data dan penggunaan genetic operator yang lebih kompleks.
<hr>
Document segmentation is a process to segments text into thematic homogeneous parts. The segmenting process uses genetic algorithm as a method to segment the text. Genetic algorithm is a searching algorithm for problem involving large search space by using evolution approach.

Research about document segmentation has been done by Lamprier (Lamprier et al., 2007) for English document and show satisfied results. The segmentation in Lamprier?s research uses internal cohesion and

dissimilarity as objective function to be optimized. This experiments use Indonesian mass media articles and abstracts of scientific paper from Lontar System of Faculty of Computer Science University of Indonesia.

Experiments have been done and analyzed towards several aspects such as fitness function, similarity calculating method, number of iteration, number of population, number of boundary, and similarity between appended documents. Furthermore the experiment to compare genetic algorithm and other segmentation method (Texttiling) is done in the last experiment.

The experiments shows that genetic algorithm using SPEA 2 as fitness function, dice coefficient as similarity calculating method, 1000 iteration, 50 individuals in population, two point crossover, and 0.09 mutation probability gives the best result. When comparing segmentation method between genetic algorithm and Texttiling, genetic algorithm gives precision 0.081 and recall 0.46 in other hand Texttiling gives precision 0.12 and recall 0.58.

The results show that Texttiling gives better segmentation than genetic algorithm, this conclusion is diffrent with the conclusion reported by Lamprier?s research (Lamprier et al., 2007). The diffrent is related with data and genetic operator used by Lamprier?s research.