Analisis penerapan distribusi normal dan learning vector quantization dalam sistem deteksi plagiarisme makalah dwibahasa berbasis metode latent semantic analysis = Analysis of normal distribution and learning vector quantization implementation in latent semantic analysis based bilingual plagiarism detection system

Darien Jonathan, author

Abstrak

ABSTRAK
Distribusi normal adalah salah satu jenis persebaran kelompok data yang didefinisikan berdasarkan rata-rata dan standar deviasi dari sekelompok data, yang dapat digunakan untuk mengelompokkan data berdasarkan posisinya terhadap standar deviasi dari kelompok data tersebut. Learning Vector Quantization adalah salah satu jenis neural network yang bisa mempelajari sendiri masukan yang ia terima kemudian memberi keluaran sesuai dengan masukan tersebut, dengan metode supervised dan competitive learning. Skripsi ini membahas penerapan dan analisis dari kedua sistem tersebut untuk menguji hasil deteksi plagiarisme oleh sistem deteksi plagiarisme berbasis latent semantic analysis, yang berasal dari program Simple-O. Beberapa modifikasi dilakukan untuk meningkatkan akurasi pengujian, antara lain dengan melakukan variasi parameter-parameter dari metode distribusi normal, yakni dengan mengubah batas standar deviasi maupun dengan mengubah koefisien pengali batas nilai pada standar deviasi tertentu, dimana hasilnya adalah standar deviasi maupun koefisien pengalinya berbanding lurus dengan aspek relevansi program (recall) namun tidak pada akurasi (F-Measure). Modifikasi juga dilakukan pada parameter percepatan belajar dari algoritma learning vector quantization, dimana hasilnya adalah parameter percepatan belajar berbanding terbalik dengan relevansi program maupun akurasi. Kemudian variasi dan analisis dilakukan pada tujuh jenis besaran hasil keluaran sistem deteksi plagiarisme berbasis latent semantic analysis, yakni frobenius norm, slice, dan pad, beserta kombinasinya, dimana hasilnya keberadaan frobenius norm diwajibkan untuk melakukan evaluasi kemiripan antara dua teks. Kemudian hasil pengujian menggunakan kedua metode digabungkan menggunakan operasi AND yang memberikan hasil yang beragam, dengan catatan perlunya keseimbangan antara precision dan recall dari masing pengujian yang akan dilakukan operasi AND untuk memberikan hasil yang baik. Dengan menggunakan kombinasi metode dan parameter yang tepat, terdapat peningkatan akurasi sistem dari 35-46% pada penelitian sebelumnya hingga maksimal 65,98%.
<hr><i>ABSTRACT
Normal distribution is a type of data distributions which is defined from the average and standard deviation of the data cluster. It can be used to group datas based on its position from the standard deviation of the data cluster. Learning vector quantization is a type of neural networks that can learn from inputs it gets to give appropriate outputs, with supervised and competitive learning methods. This thesis discusses the implementation and analysis of both methods to verify the plagiarism detection results from detection plagiarism system based on latent semantic analysis, which is based on Simple-O program. Some modifications are made, such as by variating the parameters of normal distribution method, by changing the limits of standard deviation or by changing the factor of the number limit at a particular standard deviation. Both of them appear to be directly proportional to the relevance (recall), but not with accuracy (F-Measure).

Modifications are also made at the learning acceleration parameters from the learning vector quantization algorithm, which sees the parameters being inversely proportional to both the relevance and accuracy. Then, variations and analysis are done to seven types of magnitude from the results of the plagiarism detection system, which are frobenius norm, slice, and pad, and their combinations, which suggest that frobenius norm is the most verifiable results, and must be included to be evaluated when text similarity analysis are conducted. Then, verification results using both methods are combined using AND operation which gives diverse results. However, it is needed to have a balance between precision and recall from each verifications to produce good results. With correct combinations of methods and parameters, system accuracy are increased from 35-46% of last research to maximum accuracy of 65,98%.</i>