Analisis akurasi fuzzy C-means dengan reduki dimensi random projection pada pendeteksian topik = Accuracy analysis of fuzzy C-means with random projection dimensional reduction on topic detection

Muhammad Rifky Yusdiansyah, author

Abstrak

Pendeteksian topik (Topic detection) adalah suatu proses yang digunakan untuk menganalisis kata-kata pada suatu koleksi data tekstual untuk menentukan topik-topik yang ada pada koleksi tersebut, bagaimana hubungan topik-topik tersebut satu sama lainnya, dan bagaimana mereka berubah dari waktu ke waktu. Metode Fuzzy C-Means (FCM) merupakan metode clustering yang sering digunakan pada masalah pendeteksian topik. Fuzzy C-Means dapat mengelompokkan dataset ke beberapa cluster dengan baik pada dataset dengan dimensi yang rendah, namun gagal pada dataset yang berdimensi tinggi. Untuk mengatasi permasalahan tersebut, dilakukan reduksi dimensi pada dataset sebelum dilakukan pendeteksian topik menggunakan metode FCM. Pada penelitian ini digunakan data tweets akun berita nasional pada sosial media Twitter yang kemudian dilakukan pen-deteksian topik menggunakan metode Random space-based Fuzzy C-Means (RFCM) dan Kernelized Random space-based Fuzzy C-Means (KRFCM). Metode pembelajaran RFCM dan KRFCM terbagi menjadi dua langkah yaitu mereduksi dimensi dataset ke dimensi yang lebih rendah dengan menggunakan random projection dan melakukan metode pem-belajaran FCM pada RFCM dan metode pembelajaran KFCM pada KRFCM. Setelah didapatkan topik-topik, kemudian dilakukan evaluasi dengan menghitung nilai coher-ence pada topik. Nilai coherence yang digunakan pada penelitian ini menggunakan sa-tuan Pointwise Mutual Information (PMI). Penelitian dilakukan dengan membandingkan nilai rata-rata PMI dari RFCM dan KRFCM dengan Eigenspace-based Fuzzy C-Means (EFCM) dan Kernelized Eigenspace-based Fuzzy C-Means (KEFCM). Hasil yang didapatkan menggunakan data tweets akun berita nasional menunjukkan bahwa metode RFCMdan KRFCM menawarkan running time untuk reduksi dimensi yang lebih cepat namun memiliki rata-rata nilai PMI yang lebih kecil dibandingkan rata-rata nilai PMI yang di-hasilkan oleh metode pembelajaran EFCM dan KEFCM.

<hr>

Topic detection is a process that is used to analyze
words in a collection of textual data to determine which topics are in the collection, how the topics relate to each other, and how they change over time. Fuzzy C-Means (FCM) Method is a clustering method that is often used in topic detection problems. Fuzzy C-Means can group datasets into several clusters properly on dataset with low dimensions, but failed on the high dimension dataset. To overcome this problem, a dimension reduction is performed on the previous dataset Topic detection was performed using the FCM method. In this study used data on national news account tweets on Twitter social media which is then detected topics using the Randomspace-based Fuzzy C-Means (RFCM) method Kernelized Randomspace-based Fuzzy C-Means (KRFCM). RFCM learning methods and KRFCM is divided into two steps, namely reducing the dataset dimension to dimensions lower cost by using random projection and learning methods FCM on RFCM and KFCM learning methods on KRFCM. After obtained topics, then conducted an evaluation by calculating the value of coherence on the topic. The coherence value used in this study uses units Pointwise Mutual Information (PMI). Research carried out by comparing

the average PMI values of RFCM and KRFCM with Eigenspace-based Fuzzy C-Means (EFCM) and Kernelized Eigenspace-based Fuzzy C-Means (KEFCM). Results obtained using national news account tweets data shows that the RFCM method and KRFCM offers running time for faster dimension reduction however has an average PMI value that is smaller than the average PMI value produced by the EFCM and KEFCM learning methods.