Jurnal Informatika Universitas Padjadjaran ISSN : 2503-5258
Vol. 01 No. 01 Agustus 2016

KLASIFIKASI *NEWSGROUP* MENGGUNAKAN *VECTOR SPACE MODEL* DAN *NOVEL K NEAREST NEIGHBORS*

Mira Suryani[1)], , Ayi Muhammad Iqbal Nasuha[2)], Intan Nurma Yulita[3)], Erick Paulus[4)]

Email : [1)]mira.suryani@unpad.ac.id, [2)]m.iqbal.nasuha@email.unikom.ac.id, [3)]intan.nurma@unpad.ac.id, [4)]erick.paulus@unpad.ac.id,

[1,3,4)] Program Studi Teknik Informatika, Departemen Ilmu Komputer, FMIPA Universitas Padjadjaran
[2)] Departemen Teknik Informatika, Universitas Komputer Indonesia

ABSTRACT

One of the emerging study in the field of information retrieval is text classification. The text classification helps people to find the collection of information that relevant to the needs. This study explain about the process of newsgroup classification. The type of data selected due to the newsgroup itself is an application that have been used by many people for a long time to discuss in virtual world. So, the huge data that can be obtained and of course need to be classified. In this study, vector space model used as feature representation of the data after indexing and weighting process using term frequency. Represented feature then classified into three categories according to their target class. The experiment results shown the average precision of the classification is about 71% with 89 data classified correctly. The results obtained based on the setting of k value that gives the optimal classification with 30 in k nearest neighbors method.
Keywords—classification; knn; newsgroup; term-frequency; vector space model

ABSTRAK

Salah satu penelitian dalam bidang perolehan informasi yang hingga saat ini masih menjadi kajian adalah kategorisasi teks. Klasifikasi teks dapat membantu manusia untuk menemukan sekumpulan informasi yang relevan sesuai dengan kebutuhan secara cepat. Studi ini mengemukakan tentang proses mengkategorisasikan *newsgroup*. Data *newsgroup* dipilih sebagai dataset penelitian dikarenakan *newsgroup* sendiri merupakan aplikasi yang telah lama dan banyak digunakan oleh orang untuk berdiskusi di dunia maya, sehingga data *newsgroup* berada dalam jumlah besar dan perlu pengelolaan. Vector space model sebagai representasi fitur dari sebuah dokumen yang dihasilkan setelah melalui proses *indexing* dan pembobotan menggunakan *term frequency*. Representasi fitur kemudian diklasifikasikan ke dalam 3 kategori sesuai dengan kelas kategorinya. Dari hasil penelitian diperoleh nilai rata-rata precision sebesar 71% dengan jumlah data yang diklasifikasikan secara benar sebanyak 89 data. Hasil ini diperoleh dari penentuan jumlah k paling optimal yang berada pada nilai 30.
Kata Kunci—klasifikasi; knn; newsgroup; term-frequency; vector space model