Pemanfaatan dokumen unlabeled pada klasifikasi topik berbasis naive bayes dengan algoritma expectation maximization

Bayu Distiawan Trisedya, author

Abstrak

Klasifikasi dokumen teks adalah masalah sederhana namun sangat penting karena manfaatnya cukup besar mengingat jumlah dokumen yang ada setiap hari semakin bertambah. Namun, kebanyakan teknik klasifikasi dokumen yang ada memerlukan labeled documents dalam jumlah besar untuk melakukan tahap training. Dalam melakukan klasifikasi dokumen, pada tugas akhir ini digunakan algoritma Expectation Maximization yang dikombinasikan dengan algoritma Naïve Bayes untuk memanfaatkan unlabeled documents dengan tiga buah kumpulan data yaitu dokumen hukum, artikel media massa, dan 20Newsgroups dataset. Selain melihat pengaruh penggunaan unlabeled documents, percobaan pada tugas akhir ini juga menganalisis hasil klasifikasi dari beberapa aspek seperti pengaruh stopwords, penggunaan jumlah kategori, dan penggunaan empat buah jenis fitur yaitu presence, frequency, frequency normalized, dan pembobotan tf-idf. Secara umum, penggunaan unlabeled documents memberikan manfaat yang cukup berarti bagi peningkatan akurasi hasil klasifikasi. Dengan konfigurasi tertentu, rata-rata peningkatan akurasi yang diperoleh dapat mencapai angka 9,5%. Namun, penggunaan unlabeled documents ini harus didukung oleh penggunaan labeled documents dalam jumlah yang tepat. Dari percobaan yang telah dilakukan diperlukan sekitar 30 hingga 60 labeled documents tiap kategorinya untuk membangun initial classifier untuk dapat memanfaatkan unlabeled documents secara maksimal.

<hr>

Text documents classification is a simple problem but it is very important because the benefit is quite large considering the number of documents become more and more to handle each day. However, most of the document classification technique requires large numbers of labeled documents. In performing document classification on this final project, Expectation Maximization algorithm combined with Naïve Bayes algorithm is used to take advantage of unlabeled documents with the three set of data that is legal documents, news articles collection, and 20Newsgroups dataset. In addition to see the influence of unlabeled documents, we also analyze the classification results from several aspects such as the effect of stopwords, the number of categories, and the use of four types of features namely presence, frequency, frequency normalized, and TF-IDF. In general, the uses of unlabeled documents provide a significant benefit for increasing the classification accuracy. With a certain configuration, the average escalation in accuracy can be reached 9,5%. However, the use of unlabeled documents must be supported by the use of labeled documents in the appropriate amount. From the results obtained show that to get maximum benefit from unlabeled documents required 30 to 60 labeled documents per category.