Abstrak

Pada saat ini kategorisasi dokumen dilakukan dengan menggunakan pendekatan *machine learning* dimana pendekatan ini akan melakukan proses learning dari dokumen yang dijadikan sebagai contoh dan kemudian hasil dari learning tersebut akan digunakan sebagai acuan untuk mengkategorisasi dokumen yang lainnya. Terdapat suatu gejala dimana setiap orang dapat menggunakan kata yang berbeda untuk mengekspresikan maksud atau konsep yang sama. Oleh karena itu diperlukan suatu pendekatan yang membandingkan dokumen tidak hanya dari persamaan kata, tetapi juga mempertimbangkan persamaan konseptual dari kata tersebut.

Pada tugas akhir ini dibuat suatu perangkat lunak prototipe yang menerapkan metode *probabilistic latent semantic indexing* (PLSI) untuk proses kategorisasi dokumen. PLSI merupakan salah satu pengembangan dari metode *latent semantic lndexing* (LSI). Dari perangkat lunak yang dihasilkan, diukur performansi dari metode PLSI berupa efektifitas dan efesiensi.

Pada PLSI ditemukan bahwa dimensi yang digunakan pada dekomposisi memberikan pengaruh pada performansi yang dihasilkan, semakin besar dimensi yang digunakan maka *recall*, *precision*, dan waktu dekomposisi dari sistem akan meningkat dan *error* akan menurun. Jumlah data yang digunakan turut mempengaruhi performansi yang dihasilkan, namun term yang digunakan pada data pun turut mempengaruhi performansi pada jumlah data. Dari hasil penelitian pun ditemukan bahwa pembobotan TFDIF pada PLSI akan memperburuk performansi yang dihasilkan daripada pembobotan TF.

Kata kunci: kategorisasi dokumen, *probabilistic latent semantic indexing*