PENGEMBANGAN FITUR DOKUMEN BERBASIS *BAG OF CONCEPTS* UNTUK MENDUKUNG PENGELOMPOKAN DOKUMEN

Yeti Nugraheni
UPT Komputer, Politeknik Negeri Bandung
Jl. Gegerkalong Hilir, Ds. Ciwaruga, Bandung, 40012
Telp : (022) 2013789, Fax : (022) 2013889
E-mail : yeti.nugraheni@polban.ac.id

Abstrak

Teknik pengelompokan dokumen (document clustering) standar umumnya menggunakan representasi bag of words, sedangkan teknik representasi bag of concepts belum banyak digunakan. Ekstraksi dan seleksi fitur merupakan tahap penting untuk merepresentasikan dokumen kedalam suatu bentuk data yang dapat mewakili informasi data teks. Fitur penting dan relevan yang ditemukan diharapkan dapat meningkatkan kualitas hasil pengelompokkan. Representasi teks dalam bentuk bag of concepts dapat diidentifikasi menggunakan wordnet. Pada penelitian ini dipilih representasi bag of concepts menggunakan wordnet yang secara konsep memiliki kemampuan dalam menjaga makna semantik dokumen. Tujuan dari penelitian ini adalah melakukan analisis terhadap representasi teks untuk meningkatkan akurasi pengelompokan dokumen dalam bentuk bag of concepts yang dihasilkan dari ekstraksi dan seleksi fitur menggunakan wordnet. Dokumen teks merupakan data yang tidak terstruktur, untuk itu diperlukan proses-proses pendukung meliputi preprocessing hingga diperoleh data yang dapat diolah dengan algoritma clustering K-Means untuk menghasilkan clusters.

Metode yang dilakukan yaitu dengan cara membandingkan kinerja representasi teks dalam bentuk bag of concepts yang dihasilkan dari wordnet terhadap representasi teks menggunakan sequential pattern maupun representasi teks dalam bentuk bag of word. Perbandingan kinerja dilakukan terhadap penjagaan makna semantik yang dapat dilakukan oleh masing-masing representasi, serta efisiensi waktu berdasarkan proses pembentukan representasi. Fitur yang didapat dari ekstraksi dan seleksi fitur menggunakan wordnet menghasilkan makna kata yang baik, karena makna dari setiap kata dalam dokumen tersimpan dalam basis data synset walaupun kebanyakan fitur masih dalam bentuk single word term. Sedangkan sequential pattern mampu mengurangi jumlah fitur, karena menghasilkan ruang dimensi pencarian yang lebih sedikit dan secara representasi mampu menghasilkan fitur dalam bentuk multi word term, sehingga secara efektif mampu meningkatkan kualitas pengelompokan.

Dengan menggunakan dataset jurnal dan paper terutama bersumber dari IEEE berbahasa Inggris dengan format file pdf serta data dari TNTD (Twenty News Group Text Data) sebagai pembanding, hasil pengujian yang diperoleh menunjukkan bahwa secara proses waktu yang dibutuhkan untuk menghasilkan fitur dalam bentuk bag of concepts menggunakan wordnet lebih sedikit membutuhkan waktu dibandingkan dengan representasi menggunakan sequential pattern maupun representasi dalam bentuk bag of word, namun rata-rata akurasi cluster paling tinggi adalah dengan representasi multi word terms menggunakan sequential pattern. Nilai akurasi clusters dokumen berdasarkan rumus F-Measure dipengaruhi oleh dataset yang dapat menghasilkan fitur yang tidak representatif terhadap topik dalam melakukan pengelompokan.

Untuk penelitian lebih lanjut perlu disempurnakan penemuan fitur menggunakan wordnet, tidak hanya menghasilkan frase dengan dua kata tetapi bisa lebih dari dua kata serta tidak hanya dengan memeriksa hubungan synonym tetapi juga dapat digunakan dengan antonym atau relasi lain. Selain itu perlu dilakukan pengembangan untuk melengkapi atau membuat database wordnet yang berisi frase ilmiah dan sesuai dengan bidang ilmu (ontology) tertentu sehingga seleksi fitur yang menggunakan wordnet dapat menghasilkan fitur yang lebih baik.

Kata Kunci: *document clustering, bag of concepts, wordnet, sequential pattern, bag of word.*

DAFTAR PUSTAKA

[1] Agrawal, Rakesh, Ramakrishnan Srikant. 1995. Mining Sequential Patterns. IBM Research Center.
[2] Tackstrom, Oscar., 2005, An Evaluation of Bag-of-Concepts Representations in Automatic Text Classification.
[3] Han, Jiawei., Micheline Kamber. 2006. Data Mining Concepts and Techniques. Morgan Kaufmann Publisher
[4] Alejandro Rodríguez, Ricardo Colomo, Juan Miguel Gómez, Giner Alor-Hernandez, Ruben Posada-Gomez, Ulises Juarez-Martinez,Jose Emilio Labra Gayo, Krishnamurthy Vidyasankar. 2009. A proposal for a Semantic Intelligent Document Repository Architecture. IEEE. Electronics, Robotics and Automotive Mechanics Conference.
[5] K. Zubrinic, D. Kalpic, and M. Milicevic, The automatic creation of concept maps from documents written using morphologically rich languages, Expert Systems with Applications, vol. 39, no. 16, pp. 12709–12718, 2012
[6] Miller, G. A. WordNet®: A Lexical Basis data for English. Communications of the ACM, November 1995 Vol.38 No.11 40-41. November 2014, dari WordNet®: A Lexical Basis data for English: http://WordNet®.princeton.edu/WordNet®/documentation/