UBAYA
UNIVERSITAS SURABAYA

snastia 2013

SEMINAR NASIONAL
TEKNOLOGI INFORMASI DAN MULTIMEDIA

PROCEEDINGS

"Pemanfaatan Teknologi Informasi, Komunikasi dan Multimedia untuk Meningkatkan Kualitas Kehidupan Masyarakat"

21 September 2013

Teknik Informatika - Sistem Informasi - Multimedia - IT Dual Degree

PROSIDING

SNASTIA

Seminar Nasional

Teknologi Informasi dan Multimedia

UBAYA

UNIVERSITAS SURABAYA

Vol. 4 Tahun 2013

ISSN: 1979-3960

21 September 2013

UNIVERSITAS SURABAYA

SURABAYA

Reviewer

Prof. Dr. Ir. Arif Djunaidy, M.Sc.
Prof. Ir. Handayani Tjandra, M.Sc. Ph.D.
Prof. Ir. Hening Widi Oetomo, M.M., Ph.D.
Prof. Ir. Joniarto Parung, Ph.D.
Prof. Drs. Nur Iriawan, M.Sc., Ph.D.
Prof. Ir. Supeno Djanali, M.Sc., Ph.D.
Djuwari, Ph.D.
Nemuel Daniel Pah, S.T., M.Eng., Ph.D.
Daniel Hari Prasetyo, S.Kom., M.Sc.
Stephanus Eko Wahyudi, M.M.M.

Daftar Isi

KATEGORISASI UNBALANCED TEXT MENGGUNAKAN COMPLETE GINI INDEX DAN RELATIVE WEIGHT K-NEAREST NEIGHBOR

Monica Widiasri, S. Kom., M. Kom.[1], Army Justitia, S. Kom., M. Kom.[2]
Universitas Surabaya[1], Universitas Airlangga[2]
monica@ubaya.ac.id[1], army.justitia@yahoo.com[2]

Abstract

Feature selection is necessary to reduce a large feature space in text categorization. Especially in unbalanced text, where the number of training documents between each category is unbalanced, it needs proper feature selection method that features selected are appropriate and can distinguish between categories. Also, a proper categorization method is needed to categorize unbalanced text, because unsuitable categorization method for unbalanced text categorization can lead poor results. This research used Complete Gini index (CGI) for feature selection and Relative Weight K-Nearest Neighbor (RWKNN) for unbalanced text categorization method. CGI can select representative features for each category in the unbalanced text. RWKNN can overcome the problems of unbalanced text categorization. The experiment shows that the accuracy of CGI-RWKNN is better than CGI-KNN, at least 5% improved. CGI-RWKNN can select representative features for each category, show better results and stable unbalanced text categorization.

Keywords: feature selection, unbalanced text, Complete Gini Index, Relative Weight K-Nearest Neighbor.

1. Pendahuluan

Pada kategorisasi teks dimensi fitur yang besar dapat mengurangi efisiensi dan presisi dari hasil kategorisasi, sehingga diperlukan seleksi fitur untuk mengurangi ukuran dimensi fitur dan memilih fitur yang sesuai untuk tiap kategori. Pada kondisi nyata kategorisasi teks, sering dijumpai data kategorisasi bersifat *unbalanced*, yaitu jumlah dokumen antar kategori tidak seimbang [Zheng, Wu, dan Srihari, 2004]. Sifat *unbalanced text* juga dapat mengakibatkan hasil kategorisasi menurun. Oleh karena itu diperlukan suatu metode seleksi fitur serta metode kategorisasi yang tepat pada *unbalanced text* supaya menghasilkan kategorisasi teks yang baik.

Terdapat beberapa metode seleksi fitur yang sering digunakan pada kategorisasi teks multi kategori yaitu *Information Gain, χ^2, Mutual Information, Expected Cross Entropy, Weight of Evid dan Odds Ratio* [Zheng, Wu, dan Srihari, 2004]. Pada 2007, diusulkan algoritma Gini Index yang biasa digunakan pada pembuatan *Decision Tree* untuk memilih fitur penting pada kategorisasi *unbalanced text* [Shang, Huang, dan Zhu, 2007]. Namun metode ini masih terdapat kelemahan, yaitu beberapa fitur penting dapat tidak terpilih karena nilai gini indeks yang dihasilkan sangat kecil (mendekati nol). Untuk mengatasi masalah tersebut dan meningkatkan kinerja kategorisasi dikembangkan metode *Complete Gini Index* (CGI) [Park, Kwon, dan Kwon, 2010]. Metode CGI terbukti dapat memilih fitur penting dengan lebih baik pada kategorisasi *unbalanced text*. Namun pada penelitian tersebut digunakan KNN sebagai metode pengkategorisasi. Padahal kinerja kategorisasi KNN terbukti menurun pada data *unbalanced text* [Liu, Ren, dan Yuan, 2010]. Oleh karena itu, diperlukan metode kategorisasi yang tepat untuk *unbalanced text* yang menggunakan seleksi fitur *Complete Gini Index*. Metode Relative Weight K-Nearest Neighbor (RWKNN) yang mengembangkan KNN terbukti dapat digunakan sebagai metode pengkategorisasi pada *unbalanced text* [Liu, Ren, dan Yuan, 2010]. Pada penelitian ini, metode CGI yang digabungkan dengan RWKNN digunakan untuk mengkategorisasikan *unbalanced text*, diharapkan menghasilkan kinerja kategorisasi teks yang lebih baik. Uji coba dilakukan dengan membandingkan nilai $F_{1\ macro}$ hasil kategorisasi teks antara CGI-RWKNN dengan CGI-KNN.

Paper ini disusun dengan struktur sebagai berikut : bagian 2 membahas model vector dokumen. Bagian 3 membahas mengenai teori seleksi fitur CGI. Bagian 4 membahas RWKNN. Sedangkan metode CGI-RWKNN dijelaskan pada bagian 5. Bagian 6 menjelaskan tentang uji coba yang dilakukan. Kesimpulan dituliskan pada bagian terakhir paper ini.

2. Model Vektor Dokumen

Sebuah dokumen *d* digambarkan sebagai model vektor yang direpresentasikan sebagai sebuah vektor dengan *n* fitur. Sebelum *indexing* dan mendapatkan representasi dokumen dalam model vektor, dilakukan *preprocessing* dokumen. *Preprocessing* terdiri dari segmentasi kata, penghilangan stop word dan *stemming*, untuk mendapatkan daftar fitur keseluruhan dokumen corpus.

Representasi vektor untuk sebuah dokumen *d*, sebagai berikut :

$$\vec{d} = \langle t_1, w_1; t_2, w_2; t_3, w_3; \ldots; t_n, w_n \rangle, \qquad (1)$$

dimana t_i adalah fitur ke-*i* dari dokumen dan w_i adalah bobot TF-IDF dari sebuah fitur.