METODE HIBRIDASI ANT COLONY OPTIMIZATION DAN INFORMATION GAIN UNTUK SELEKSI FITUR PADA DOKUMEN TEKS ARAB

Heliza Rahmania Hatta[1,2)], Agus Zainal Arifin[1)] , Anny Yuniarti[1)]
[1)]Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
[2)]Ilmu Komputer, FMIPA, Universitas Mulawarman
heliza_rahmania@yahoo.com

Abstrak *- Kategorisasi teks telah membuat kemajuan pesat dan menjadi salah satu area penelitian di bidang pengolahan informasi. Tetapi kategorisasi teks memiliki masalah utama yaitu tingginya dimensi fitur sehingga dapat mengurangi kinerja klasifikasi. Karena itu dalam penelitian ini diusulkan sebuah metode untuk seleksi fitur menggunakan metode hibridasi Ant Colony Optimization (ACO) dan Information Gain (IG) pada dokumen teks Arab. Menggunakan dokumen teks Arab karena penelitian dalam bidang ini masih sedikit. Dokumen – dokumen teks Arab ini akan mengalami tahap preprocessing hingga menghasilkan fitur – fitur. Kemudian fitur – fitur tersebut akan diberi nilai IG dan akan digunakan untuk seleksi fitur menggunakan ACO. Informasi heuristik pada metode ACO menggunakan nilai IG yang telah dihitung sebelumnya. Pada percobaan ditunjukkan bahwa metode hibridasi ACOIG dapat mereduksi fitur sebanyak 89%, sedangkan metode ACO hanya 79%. Dan waktu performa yang dibutuhkan metode ACOIG lebih cepat dari metode ACO.*

Kata Kunci: Seleksi fitur, information gain, ant colony optimization, dokumen teks Arab.

Dewasa ini, kategorisasi teks telah membuat kemajuan pesat dan menjadi salah satu area penelitian di bidang pengolahan informasi seperti kategorisasi teks Arab. Kategorisasi teks banyak digunakan ketika mengorganisir dokumen dalam bentuk digital dan menjadi salah satu kunci teknologi pada pemrosesan dan mengorganisasi data dokumen yang berjumlah besar. Tetapi kategorisasi teks memiliki masalah utama yaitu tingginya dimensi fitur. Dimensi fitur terdiri dari puluhan atau ratusan ribu fitur unik yang diambil dari dokumen *input* yang dapat tidak saling berhubungan. Permasalahan yang muncul akibat dimensi fitur yang besar pada kategorisasi teks dapat mengurangi kinerja klasifikasi. Untuk mencegah situasi ini, fitur yang diekstrak harus di *filter* sebelum fase klasifikasi untuk menyeleksi fitur yang paling relevan dan yang terbaik untuk mewakili dokumen. Hal ini dilakukan dengan menghapus fitur *noninformative* dan membangun fitur set baru menggunakan metode seleksi fitur.

Karena itu diperlukan suatu metode untuk memilih fitur penting yang mewakili dokumen dan dapat mengurangi dimensi ruang fitur karena dapat meningkatkan kinerja klasifikasi. Seleksi fitur adalah proses memilih subset yang terbaik dari fitur originalnya berdasarkan pada beberapa kriteria [1] [2]. Banyak penelitian untuk menguji beberapa algoritma dan melakukan pengembangan pada algoritma tersebut untuk melakukan seleksi fitur. Beberapa metode seleksi fitur yang digunakan pada kategorisasi teks yaitu *Information Gain (IG), χ^2, Mutual Information (MI), Expected Cross Entropy, Weight of Evid, Odds Ratio (OR),* dan *Document Frequency (DF).*

IG adalah metode yang sering digunakan untuk seleksi fitur dan bekerja dengan baik dengan teks terutama untuk klasifikasi teks bahasa Inggris [3]. Hal ini dapat dikarenakan penelitian yang telah dilakukan dalam kategorisasi teks otomatis untuk dokumen lebih banyak dalam bahasa Inggris. Sehingga penelitian dalam kategorisasi teks Arab sangat terbatas karena bahasa Arab memiliki sifat struktur morfologi yang kompleks [4] [5]. Tetapi Mesleh [6] telah membuktikan bahwa IG dapat digunakan untuk seleksi fitur pada kategorisasi dokumen teks Arab. Metode IG dapat melihat setiap fitur untuk memprediksi label kelas yang benar karena memilih nilai

yang tertinggi dan lebih efektif untuk mengoptimalkan hasil klasifikasi [6].

Sedangkan ACO, menurut Zhou [8] dan Aghdam [7] dapat digunakan untuk seleksi fitur. Dengan kata lain, jika fitur diwakili sebagai graf, semut dapat menemukan kombinasi fitur terbaik saat mereka melintasi graf [9]. Dan ACO dapat mengatasi masalah saat beberapa dataset dengan lebih banyak fitur dan dapat menemukan yang lebih baik karena memiliki kemampuan eksplorasi yang kuat dalam menemukan solusi optimal [7] dan lebih cepat dalam hal waktu pemrosesan [6] [7].

Seleksi fitur merupakan proses memilih subset yang terbaik dari fitur originalnya berdasarkan pada beberapa kriteria untuk menghasilkan kategorisasi yang baik. Namun tingginya dimensi fitur menyebabkan berkurangnya kinerja klasifikasi dan lamanya waktu yang dibutuhkan saat proses berlangsung.

Oleh karena itu, dalam penelitian ini diusulkan sebuah metode hibridasi ACOIG untuk seleksi fitur dokumen teks Arab. Metode IG digunakan untuk menghitung informasi heuristik di ACO. Metode ACOIG ini diharapkan dapat mengoptimalkan seleksi fitur pada dokumen teks Arab dibandingkan dengan menggunakan metode yang lain.

METODOLOGI

IG merupakan pengurangan entropi klasifikasi berdasarkan pengamatan variabel tertentu dan digunakan dalam mesin pembelajaran dengan *decision tree* dalam menghitung pentingnya atribut. Setiap fitur dasar mendapatkan nilai informasi untuk menentukan apakah fitur tersebut harus dipilih atau dihapus.

Seperti disebutkan sebelumnya, IG adalah fungsi evaluasi yang sering digunakan untuk *term* atau seleksi fitur dalam bidang mesin pembelajaran. Mengukur jumlah bit informasi yang diperoleh untuk prediksi kategori dengan mengetahui ada atau tidak adanya *term* dalam dokumen. IG dapat didefinisikan sebagai berikut:

$$IG(t) = -\sum_{i=1}^{m} P_r(c_i) \log P_r(c_i) + P_r(t) \sum_{i=1}^{m} P_r(c_i|t) \log P_r(c_i|t) + P_r(\bar{t}) \sum_{i=1}^{m} P_r(c_i|\bar{t}) \log P_r(c_i|\bar{t}), \quad (1)$$

di mana $P_r(c_i)$ adalah probabilitas dari sebuah dokumen yang berada di label kelas c_i, $P_r(t)$ adalah probabilitas *term t* yang muncul di dokumen, $P_r(c_i|t)$ adalah probabilitas dari sebuah dokumen yang berada di label kelas c_i mengingat bahwa *term t* yang muncul di dalam dokumen dan $P_r(c_i|\bar{t})$ adalah probabilitas dokumen yang berada di label kelas c_i mengingat bahwa *term t* tidak muncul dalam dokumen. Nilai IG ini akan digunakan untuk menghitung informasi heuristik pada probabilitas di ACO.

Informasi heuristik berguna untuk mengarahkan probabilistik mencapai solusi terbaik di ACO. Semut buatan mengikuti aturan yang disebut aturan transisi probabilistik yang menentukan probabilitas sebuah fitur *k* semut memilih *i* fitur untuk dijadikan solusi pada *t* waktu. Aturan transisi probabilistik adalah membangun dua parameter yaitu informasi heuristik dan tingkat feromon sebagai berikut:

$$p_i^k(t) = \begin{cases} \frac{[\tau_i(t)]^\alpha \cdot [\eta_i]^\beta}{\sum_{u \epsilon J^k} [\tau_u(t)]^\alpha \cdot [\eta_u]^\beta} & jika\ i \epsilon J^k \\ 0 \end{cases}, \quad (2)$$

dimana J^k adalah himpunan fitur *k* semut yang belum dikunjungi, η_i adalah informasi heuristik untuk memilih fitur *i* untuk menjadi bagian dari solusi parsial, $\tau_i(t)$ adalah nilai feromon yang diletakkan di fitur *i*, sedangkan α dan β adalah dua parameter yang menentukan kepentingan relatif pada nilai feromon dan informasi heuristik.

Proses ACOIG ini akan dilakukan setelah dokumen teks arab mengalami proses *preprocessing* dan menghasilkan fitur – fitur asli. Fitur – fitur ini akan dihitung nilai IG terlebih dahulu lalu diseleksi menggunakan ACOIG. Proses metode hibridasi ACOIG untuk seleksi fitur dapat dilihat pada Gambar 1.

Pada Gambar 1 bahwa tingkat feromon pada setiap fitur adalah hasil dari feromon yang disimpan oleh semut selama perjalanan yang sebanding dengan kualitas dari solusi dan dianggap sebagai pengetahuan sebelumnya untuk semut lainnya. Sementara itu, keinginan heuristik dari semut yang melintas di antara fitur bisa ditafsirkan sebagai fungsi evaluasi. Dalam algoritma usulan, informasi heuristik untuk seleksi fitur dihitung dengan IG seperti dapat dilihat pada Persamaan 3.

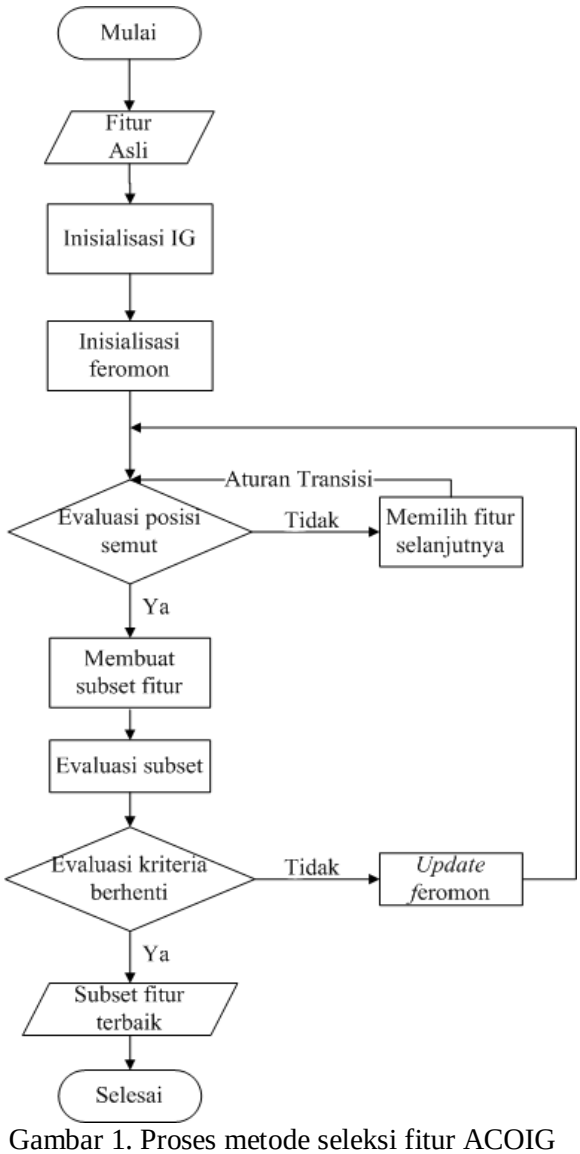
$$p_i^k(t) = \begin{cases} \frac{[\tau_i(t)]^\alpha \cdot [IG_i]^\beta}{\sum_{u \epsilon J^k} [\tau_u(t)]^\alpha \cdot [IG_u]^\beta} & jika\ i \epsilon J^k \\ 0 \end{cases}, \quad (3)$$

dimana α adalah konstanta yang menentukan

pengaruh dari jejak feromon (*exploitation*), β merupakan parameter yang menentukan pengaruh informasi heuristik (*exploration*).

Dari Persamaan ACOIG, dapat diihat bahwa informasi heuristik untuk calon fitur *i* dihitung dengan IG yang dikombinasikan dengan tingkat feromon τi untuk menentukan apakah akan memilih fitur *i* untuk menjadi bagian dari solusi parsial atau tidak yang dapat merepresentasikan dokumen.

HASIL DAN PEMBAHASAN

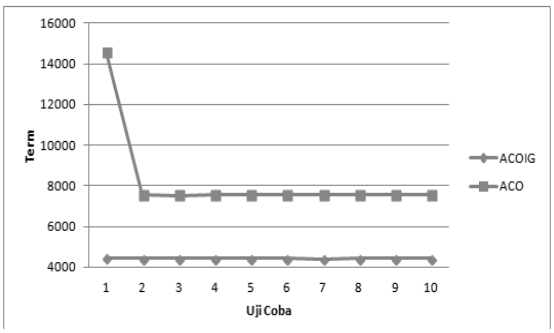


Gambar 1. Proses metode seleksi fitur ACOIG

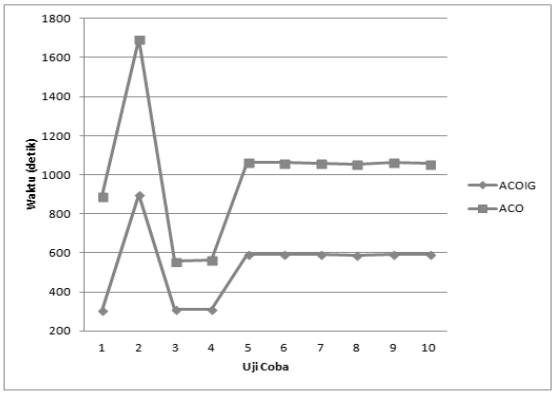
Pada penelitian ini akan digunakan dataset dengan bertipe .txt dari kumpulan dokumen surat kabar Arab, seperti Al-Jazirah (http://www.al-jazirah.com/) berjumlah 800 dokumen. Uji coba dilakukan dengan mengganti variasi nilai parameter $\alpha \in \{0, 0.5, 1, 2, 5, 6\}$, $\beta \in \{0, 0.5, 1, 2, 5, 6, 10, 20\}$ dan dengan $\rho = 0.2$ pada metode ACOIG. ρ adalah koefisien pengrusakan jejak feromon (*evaporation*). Semua nilai parameter yang ada dapat dikombinasikan. Sehingga dari semua kemungkinan kombinasi akhirnya didapat 160 buah kombinasi nilai parameter untuk metode ACOIG. Dengan masing – masing kombinasi menggunakan jumlah semut 10 dan jumlah iterasi 10 dengan fitur awal setelah *preprocessing* sebanyak 40231.

Hasil jumlah subset fitur yang ada pada Gambar 2 merupakan hasil dari rata – rata dengan menggunakan paramater α dan β. Misalkan pada data di uji coba 1 merupakan rata – rata hasil fitur dengan menggunakan $\beta = 0$ dan rata – rata dari $\alpha \in \{0, 0.5, 1, 2, 5, 6\}$ hal ini berlaku juga untuk Gambar 3. Berdasarkan Gambar 2, dapat disimpulkan bahwa seleksi fitur menggunakan metode ACOIG lebih bagus dari seleksi fitur menggunakan metode ACO. Metode ACOIG dapat mereduksi fitur rata – rata sebesar $\frac{fitur\ awal - rerata\ semua\ fitur\ yang\ dihasilkan}{fitur\ awal} \times 100\% = \frac{40231-4418}{40231} \times 100\% = 89\%$ sedangkan metode ACO sekitar $\frac{40231-8250}{40231} \times 100\% = 79\%$.

Berdasarkan Gambar 3 dapat disimpulkan waktu yang dibutuhkan metode ACOIG untuk seleksi fitur lebih cepat dari metode ACO. Hal ini didukung dengan subset fitur yang diperoleh ACOIG jauh lebih sedikit dari metode ACO. Waktu yang dibutuhkan untuk seleksi fitur metode ACOIG lebih cepat $\frac{\text{total waktu ACO} - \text{total waktu ACOIG}}{\text{total waktu ACO} + \text{total waktu ACOIG}} \times 100\% = \frac{1007-536}{1543} \times 100\% = 30.5\%$dari waktu metode ACO.



Gambar 2. Perbedaan jumlah subset fitur yang dihasilkan metode ACOIG dan ACO



Gambar 3. Perbedaan waktu metode ACOIG dan ACO

SIMPULAN

Seleksi fitur berguna untuk mengatasi masalah utama dalam kategorisasi dokumen yaitu masih tingginya dimensi fitur. Karena dimensi fitur terdiri dari puluhan atau ratusan ribu fitur unik yang dapat tidak saling berhubungan maka dibutuhkan sebuah metode yang mampu menyeleksi fitur yang paling relevan dan yang terbaik untuk mewakili dokumen. Hal ini dilakukan dengan menghapus fitur yang tidak penting dan membangun fitur set baru menggunakan metode seleksi fitur.

Dari hasil uji coba, dapat disimpulkan bahwa metode hibridasi ACOIG dapat digunakan untuk seleksi fitur pada dokumen teks Arab. Hal ini terbukti dengan metode hibridasi ACOIG mampu mereduksi fitur sebanyak 89% sedangkan metode ACO hanya 79%. Dan waktu proses metode hibridasi ACOIG lebih cepat dari metode ACO.

DAFTAR PUSTAKA

[1] Liu, L., Kang, J., Yu, J., dan Wang, Z., *"A comparative study on unsupervised feature selection methods for text clustering"*, IEEE, Proceeding of NLP-KE'O5, pp. 597–601. 2005.

[2] Uguz, Harun, *"A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm"*, Elsevier, Knowledge-Based Systems, pp. 1024-1032. 2011.

[3] Li, Y., Dai, G., dan Li, G., *"Feature Selection Method of Text Tendency Classification"*, IEEE, Fifth International Conference on Fuzzy Systems and Knowledge Discovery. 2008.

[4] Ameed, H., K., A., Ketbi, S., O., A., Kaabi, A., A., A., Shebli, K., S., A., Shamsi, N., F., A., Nuaimi, N., H., A., dan Muhairi, S., S., A., *"Arabic Light Stemmer: A new Enhanced Approach"*, The Second International Conference on Innovations in Information Technology. 2005.

[5] Harrag, F., El-Qawasmeh, E., dan Pichappan, P., *"Improving Arabic Text Categorization using Decision Trees"*, IEEE, 978-1-4244-4615-5/09. 2009.

[6] Mesleh, Abdelwadood Moh'd, *"Feature sub-set selection metrics for Arabic classification"*, Elsevier, Pattern Recognition Letters 32, pp. 1922-1929. 2011.

[7] Aghdam, M., H., Ghasem-Aghaee, N,. dan Basiri, M., E., *"Text feature selection using ant colony optimization"*, Elsevier, Expert Systems with Applications 36, pp. 6843-6853. 2009.

[8] Zhou, J., Ng., R., dan Li, X., *"Ant Colony Optimization and Mutual Information Hybrid Algorithms for Feature Subse Selection in Equipment Fault Diagnosis"*, IEEE, 2008 10th Intl. Conf. on Control, Automation, Robotics and Vision. 2008.

[9] Chen, Y.,, Miao, D., dan Wang, R., *"A Rough set approach to feature selection based on ant colony optimization"*, Elsevier, Pattern Recognition Letters 31, pp. 226-233. 2010.