KLASIFIKASI HALAMAN WEB DENGAN MENGGUNAKAN ALGORITMA ANT COLONY WEB PAGE CLASSIFICATION WITH ANT COLONY ALGORITHM

Handisapto Harsiprasetio[1], -[2]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Salah satu aplikasi pada data mining yang sudah menerapkan Algoritma Ant Colony adalah pada klasifikasi. Tujuan tugas akhir ini adalah mengimplementasikan suatu pembangunan aturan data mining klasifikasi di lingkungan web content mining, dalam hal ini klasifikasi halaman web dengan menggunakan Algoritma Ant Colony.

Pada proses klasifikasi ini ada dua langkah yang utama yaitu langkah pertama membangun model, mendeskripsikan kelas dari data yang telah ditentukan yang disebut dengan data training. Tetapi, sebelum memperoleh data training perlu dilakukan suatu langkah yang disebut pre-processing. Pada proses pre-processing ini data halaman web diolah untuk menghasilkan data training. Selanjutnya, pada langkah kedua model tersebut digunakan untuk mengklasifikasikan data yang biasa disebut data testing. Dari performansi hasil ditunjukkan bahwa Algoritma Ant Colony dapat diterapkan pada klasifikasi data halaman web yang diambil dari situs bbc.co.uk dan abc.com dengan akurasi rule yang dihasilkan kompetitif jika dibandingkan dengan See5. Dalam tugas akhir ini, akan ditunjukkan juga suatu teknik pemrosesan teks HTML untuk mengurangi jumlah atribut yang sangat besar berkaitan dengan web content mining.

Kata Kunci : algoritma ant colony, data mining, klasifikasi, web content mining.

Abstract

One of applications in data mining that use and implement ant colony algorithm is classification technique. The purpose of this final project is to implement rule construction of data mining classification – in the field of web content mining, in this case web page classification done by implemented Ant Colony Algorithm.

The classification process consist of two main steps, the first step is to construct the model, describe data class which have been determined before called data training. However, before getting data training it need one step that called pre-processing. In this step web page data is processed to get data training.

The accuracy result show that ant colony algorithm can be implemented in web page classification by doing some experiment from bbc.co.uk and abc.com, where the result show rule accuracy competitive if compare with See5. In the second step, the model used for classify the data. This final project also show a technique based HTML text processing to reduce the large number of attribute in web content mining domain.

Keywords : ant colony algorithm, data mining, classification, web content mining.

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Seiring dengan pesatnya perkembangan jumlah informasi yang tersedia di web pada jaringan internet, maka kemudian bermunculan teknik-teknik *data mining* yang ditujukan untuk mempermudah pemakai dalam memperoleh data yang ada sesuai dengan kebutuhannya. Sebagai contoh, saat ini Google telah melakukan pencarian pada lebih dari 4,2 juta halaman web pada setiap proses *searching* yang dilakukan Maka seiring dengan jumlah web yang terus meningkat, kemampuan untuk mendapatkan informasi yang spesifik adalah sama pentingnya dengan web itu sendiri. *Data mining* tersusun atas pencarian pola yang menarik dari sekumpulan data, sehingga dapat ditemukan pola yang dapat digunakan untuk mengarahkan sebuah keputusan atau sebagai inputan ke tahap selanjutnya.

Klasifikasi Web yang dimaksud di sini merupakan bagian dari teknik *web mining* yaitu *web content mining*, dimana fokus pada analisa informasi-informasi pada teks yang tersimpan pada file yang ada di web. Web content mining tersebut menekankan pada hanya proses *mining* teks, jadi tidak termasuk media seperti *image*, suara dan film.

Tujuan dari tugas akhir ini adalah untuk menemukan kumpulan *rule* klasifikasi yang kemudian *rule* tersebut diuji pada *data testing* untuk mendapatkan akurasi prediksi berdasarkan kelas yang telah didefinisikan sebelumnya. Implementasi Algoritma *Ant Colony* dilakukan melalui penggunaan algoritma *Ant-Miner*[7] yang merupakan salah satu dari varian pertama algoritma *Ant Colony* yang digunakan untuk membangun aturan klasifikasi.

Sistem *Ant Colony* melibatkan agen yang sederhana (semut) yang bekerja sama antara satu dengan lainnya untuk mencapai tujuan bersama bagi sistem secara keseluruhan, menghasilkan sebuah sistem yang mampu menemukan solusi yang baik untuk masalah yang berskala besar.

1.2 Perumusan Masalah

Dalam tugas akhir ini, terdapat beberapa permasalahan yang timbul selama proses untuk menghasilkan aturan klasifikasi pada halaman web diantaranya:

1. Bagaimana mendapatkan data set yang diperoleh dari halaman web, mengingat halaman web berupa teks.
2. Jumlah atribut yang dalam hal ini *words* adalah jauh lebih banyak bila dibandingkan dengan aplikasi data mining lainnya.
3. Bagaimana menerapkan *Ant Colony Algorithm* dalam menghasilkan *rule* klasifikasi yang akurat di lingkungan *web content mining*.

1.3 Tujuan Pembahasan

Berdasarkan rumusan masalah di atas, diharapkan akan diperoleh hal-hal sebagai berikut:

1. Menghasilkan perangkat lunak yang menghasilkan suatu dataset yang diperoleh dari parsing halaman web situs bbc.co.uk maupun situs abc.com.
2. Dengan *rule* klasifikasi yang diperoleh pada tahap *training*, akan diuji prosentase akurasi prediksi dan simplisitas *rule*.
3. Menguji dan menganalisis hasil dari akurasi berdasarkan *meta tag* yang digunakan sebagai data training dan data testing.

1.4 Batasan Masalah

Untuk mencapai tujuan di atas, dilakukan pembatasan masalah sebagai berikut:

1. Implementasi yang dibuat dalam tugas akhir ini menekankan hanya pada proses mendapatkan dataset dari halaman web yang berupa teks (dalam hal ini akan memanfaatkan Meta Tag pada HTML), jadi tidak termasuk media seperti image, suara dan film.
2. Halaman Web yang diklasifikasikan hanya halaman web berbahasa Inggris dengan mengambil studi kasus pada halaman web situs bbc.co.uk dan abc.com serta pada halaman web tersebut mengandung meta tag HTML (*meta tag description* dan *meta tag keywords*).

1.5 Metode Penyelesaian Masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini adalah menggunakan metode studi pustaka atau studi literatur dan analisis dengan langkah kerja sebagai berikut :

1. Studi literatur, yang mempelajari literatur-literatur yang berhubungan dengan data mining, *web content mining*, rule asosiasi, klasifikasi, dokumen web, algoritma *ant colony.*
2. **Perumusan masalah dan pengumpulan data, yaitu data dokumen web berasal dari data dokumen web berekstensi .HTML.**
3. **Pengembangan klasifikasi halaman web dengan tahapan sebagai berikut:**
    a. Perencanaan

       Tahapan ini dilakukan untuk melakukan analisis dan perencanaan tentang bagaimana mengubah data dari halaman web menjadi data tabel yang nantinya digunakan untuk *data training* dan *data testing*. Pertama kali halaman web diambil dari suatu situs dengan menggunakan teleport. Kemudian diparser menjadi data tabel yang nantinya akan digunakan sebagai *data training* untuk pencarian *rule* klasifikasi menggunakan algoritma *ant colony*.

    b. Analisis

       Tahapan ini dilakukan untuk menentukan kebutuhan sistem, seperti identifikasi input, identifikasi output, identifikasi hardware maupun software, identifikasi user, identifikasi halaman web yang akan diteleport dan identifikasi atribut dalam pembangunan basis datanya.

    c. Perancangan

       Pada tahapan ini dijelaskan mengenai rancangan gambaran umum sistem, deskripsi umum perangkat lunak yang akan dibangun dengan tujuan memahami secara jelas proses yang dilakukan pada system tersebut dalam suatu diagram.

    d. Implementasi

Pengimplementasian apa yang telah dirancang dengan membuat suatu aplikasi dengan menggunakan suatu bahasa pemrograman yang ditentukan pada tahap sebelumnya ini yaitu PHP.

e. Testing

Tahap yang dilakukan untuk pengetesan aplikasi sekaligus menganalisis hasil uji coba perangkat lunak yang telah dibangun berdasarkan pemilihan dataset baik secara individu atau didampingi oleh pembimbing.

4. **Pengambilan kesimpulan**
5. **Penyusunan laporan dalam bentuk tertulis sebagai laporan tugas akhir.**

1.6 Sistematika Penulisan

Tugas akhir ini disusun berdasarkan sistematika penulisan sebagai berikut :

Bab I : Pendahuluan

Bab ini akan membahas kerangka penelitian atau percobaan dalam tugas akhir, meliputi latar belakang masalah, perumusan masalah, tujuan, batasan masalah, metode penyelesaian masalah, dan sistematika penulisan.

Bab II : Dasar Teori

Bab ini memuat dasar teori yang mendukung dan mendasari penulisan tugas akhir ini, yaitu mengenai pengertian *web content mining*, *data mining,* klasifikasi, konsep *ant colony.*

Bab III : Analisis dan Perancangan Sistem

Bab ini akan membahas mengenai analisis dan perancangan sistem klasifikasi halaman web dengan studi kasus data halaman web situs bbc.co.uk dan abc.com guna menghasilkan rule dengan menggunakan algoritma *ant colony*. Dalam tugas akhir ini analisis dan perancangan perangkat lunak dibangun dalam bentuk Diagram Aliran Data (DAD).

Bab IV : Implementasi dan Analisis Hasil

Bab ini menjelaskan tentang implementasi dari perancangan yang telah dilakukan, serta penganalisaan hasil dari sistem yang telah dibuat.

Bab V : Kesimpulan dan Saran

Bab terakhir ini menjelaskan kesimpulan secara umum dari seluruh rangkaian penelitian yang dilakukan dan saran untuk pengembangan selanjutnya.

BAB V
KESIMPULAN DAN SARAN

3.1 Kesimpulan

Kesimpulan yang dapat diambil setelah melakukan pengujian dan analisis adalah :

1. Algoritma *Ant Colony* dapat digunakan untuk klasifikasi halaman web.
2. *Rule* hasil klasifikasi akurasi prediksinya 66,47 persen. Besarnya kesalahan tergantung pada proses training, antara lain jumlah data training dan pemilihan data training.
3. Akurasi rata-rata terbaik ditunjukkan oleh training dengan memanfaatkan dataset hasil *pre-processing* dengan gabungan dari meta tag *description* dan meta tag *keywords*.
4. Jumlah atribut berpengaruh besar pada tahap *pre-processing*. Semakin banyak jumlah atribut semakin besar juga waktu *pre-processing* yang diperlukan. Oleh karena itu, untuk mengoptimalkan waktu pre-processing dapat digunakan teknik pengambilan meta tag dalam halaman web berbasis HTML.
5. Rule yang dihasilkan algoritma ant colony kompetitif atau dapat bersaing dengan Algoritma C 5.0 yang biasa digunakan pada teknik *data mining* klasifikasi, baik dalam hal akurasi prediksi maupun simplisitas aturan.
6. Distribusi data yang kurang merata pada data training mengakibatkan akurasinya kecil.
7. Penghilangan stopword tidak terlalu berpengaruh terhadap akurasi prediksi.
8. Nilai frekuensi untuk *feature selection* jika terlau besar akan mengakibatkan akurasi cenderung mengecil.

3.2 Saran

1. Untuk meningkatkan akurasi klasifikasi halaman dapat digunakan metode feature selection yang lain seperti *Chi-square, Information gain, Relevancy Score* dan data setidaknya representatif atau terdistribusi dengan merata.
2. Hasil dari aturan klasifikasi halaman web dapat diolah lagi untuk digunakan dalam suatu sistem yang lebih kompleks misalnya pembuatan kategori website berita.

3. Pengambilan halaman web lebih baik bila dilakukan secara *on-line*.
4. Untuk mengurangi jumlah atribut dapat dilakukan dengan teknik yang lebih baik seperti menggunakan teknik stemming, teknik *hypernym relationship* untuk menggeneralisasi *words* dalam halaman web.

DAFTAR PUSTAKA

[1] Abraham, A and Ramos, V , 2003. *Web Usage Mining Using Artificial Ant Colony Clustering and Genetic Programming*. Proc. Congress on Evolut. Comp. (CEC-2003). IEEE Press.

[2] Chakrabarti, S. *Mining the web: DiscoveringKnowledge from Hypertext Data*, Morgan Kaufmann, 2003.

[3] Cutler. M, S. S. Maniccam and W. Meng, 1999. *A New Study Using HTML Structures to Improve Retrieval*. Proc. 11th IEEE Int. Conf. on Tools with AI. IEEE.

[4] Daulani, Muhammad, 2005. *Analisis Data Mining Metode Klasifikasi dengan Algoritma ACO (Ant Colony Optimization) : Ant Miner3*. Bandung: STT Telkom.

[5] Holden, N, Freitas, A.A. *Web Page Classification with Ant Colony Algortihm.* University of Kent, 2004.

[6] Liu, Bo, Abbass, Hussein A, McKay, Bob, 2004. *Classification RuleDiscovery with Ant Colony Optimization*, IEEE Computational Intelligence Bulletin

[7] Parpinelli, Rafael S, Lopes, Heitor S, Freitas, A.A, 2003. *Data Mining with Ant Colony Optimization Algorithm*