ANALISA PENAMBANGAN DATA MENGGUNAKAN ALGORITMA ACO (ANT COLONY OPTIMIZATION) : ANT_MINER3

Jepronel Saragih
Program Studi Teknik Informatika
Fakultas Ilmu Komputer, Universitas Pelita Harapan
E-mail: jepronel@gmail.com

ABSTRACT

Nowadays, many companies already have a big amount of data, since the datas is important to the companies, not only for current importances, but for the future needs of the companies. In fact, the datas can be used to improve the company performance. It can be done with Data Mining.Data Mining have a several methods, one of them is classification. Classification main goal is to find a special pattern, that can be formed in tree, classification rule, or mathematics formula, within the datas. In order to find this, all we have to do is applying a certain algorithm. One of that is Ant Colony Optimization (ACO) or called Ant Miner. This Final Task analyzed the changes effect in Ant_Miner3 with the accuracy and the simplicity of rules that was found, and how system parameter effect to it's result. For that purpose, the Ant_Miner3 software was build, and compared it to another Data Mining tools See5 which uses a well known C5.0 algorithm in classifying the Breast Cancer, Tic-tac-toe and House Votes datasets. The result shows that Ant_Miner3 have a better accuracy than C5.0, and it also have a little differences in simplicity. The accuration rate can be improved by the use of pheromone and assigning the no_of_ants with a greater number. Besides, by assigning pheromone evaporation, max_uncovered_case and min_cases_per_rule with smaller number will increase the accuracy. While the simplicity can be improved with the use of pruning technique, and assigning max_uncovered_case and min_cases_per_rule with a greater number.
Keywords*: data mining, ant colony optimization, ant_miner, ant_miner3*

ABSTRAK

Pada saat ini, banyak perusahaan yang memiliki data dalam jumlah yang besar. Data dalam jumlah besar tersebut ternyata dapat dimanfaatkan untuk meningkatkan kinerja perusahaan. Untuk itu diperlukan proses Data Mining. Salah satu metode dalam Data Mining adalah klasifikasi. Klasifikasi bertujuan untuk memperoleh pola tertentu, dalam bentuk tree, aturan klasifikasi atau model matematis. Untuk memperoleh pola tersebut diperlukan algoritma tertentu. Salah satunya adalah dengan ACO (Ant Colony Optimization) atau dengan nama lain Ant Miner. Pada Skripsi ini dianalisa pengaruh perubahan yang dilakukan pada Ant_Miner3 terhadap tingkat akurasi dan simplisitas aturan yang dihasilkan, serta parameter sistem yang mempengaruhinya. Untuk itu dibangun perangkat lunak sebagai media pengujian algoritma Ant_Miner3, dan membandingkannya dengan hasil yang diperoleh dengan tools Data Mining See5 yang menggunakan algoritma yang sangat sering dipakai dalam Data Mining yaitu C5.0 pada dataset Breast Cancer, Tic-tac-toe, dan House Votes. Hasilnya tingkat akurasi Ant_Miner3 lebih baik daripada C5.0, sementara simplisitas aturan yang dihasilkan tidak jauh berbeda. Tingkat akurasi dapat ditingkatkan dengan menggunakan pheromone serta

dengan memperbesar nilai parameter no_of_ants dan no_rules converg. Selain itu, dengan memberikan nilai parameter pheromone evaporation, max_uncovered_case dan min_cases_per_rule yang kecil juga dapat meningkatkan akurasi pheromone. Sementara simplisitas aturan dapat ditingkatkan dengan menerapkan teknik pruning, dan memberikan nilai max_uncovered_cases dan min_cases_per_rule yang besar.
Kata Kunci: data mining, ant colony optimization, ant_miner, ant_miner3

PENDAHULUAN

Pada saat ini setiap perusahaan sudah memiliki data dalam jumlah yang besar. Data tersebut biasanya hanya disimpan di dalam tempat penyimpanan dan digunakan untuk keperluan tertentu. Data tersebut dapat digunakan lagi, terutama untuk mengevaluasi kinerja perusahaan atau bahkan membantu mengambil keputusan pihak manajemen perusahaan. Semua itu dapat dilakukan dengan menerapkan Data Mining.

Data mining adalah sebuah proses mengolah data yang didalamnya menerapkan suatu metode tertentu untuk mendapatkan suatu pola atau knowledge dari suatu data. Data mining terbagi ke dalam beberapa bagian yaitu klasifikasi, clustering, regresi, asosiasi dan lain-lain. Klasifikasi betujuan untuk mencari aturan klasifikasi dari suatu data. Aturan klasifikasi yang dihasilkan dari proses klasifikasi berupa pernyataan if-then seperti :

IF <kondisi> THEN <class>

Pada bagian IF (antecendent) terdiri atas kumpulan kondisi. Setiap kondisi disebut term. Setiap term memiliki bentuk <atribut,operator,nilai> misalnya <Gender=male>. Bagian THEN (consequent) menyatakan class yang ditujukan untuk semua case yang memenuhi semua kondisi pada bagian IF.

Banyak sekali algoritma yang dapat diterapkan didalam klasifikasi. Setiap algoritma dapat kita bandingkan dari beberapa hal, diantaranya akurasi prediktif dan simplisitas aturan. Algoritma semut sering dipakai dalam menyelesaikan masalah kombinatorial seperti TSP (Travelling Salesman Problem) dan penentuan jadwal kuliah. Namun algoritma semut masih tergolong baru dalam penggunaannya untuk data mining, khususnya klasifikasi. Alex F. Freitas dan kawan-kawannya merupakan yang pertama kali menerapkan algoritma semut di dalam data mining metode klasifikasi. Mereka menyebut algoritmanya dengan Ant_Miner. Algoritma Ant_Miner ini kemudian diperbaiki oleh Bo Liu dan kawan-kawan. Mereka mengubah rumus heuristic function menjadi lebih sederhana. Algoritma ini disebut dengan Ant_Miner2

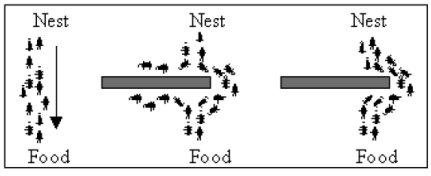
Untuk mengetahui apakah perubahan yang dilakukan mampu meningkatkan akurasi dan simplistas aturan yang dihasilkan, maka penulis merasakan perlu untuk menganalisa pengaruh perubahan yang terjadi didalam Ant_Miner3 terhadap tingkat akurasi dan simplisitas aturan yang dihasilkan. Selain itu, untuk melihat apakah Ant_Miner3 mampu bersaing dengan algoritma klasifikasi lain yang telah ada, maka dilakukan perbandingan dengan algoritma C5.0 yang sering dipakai di dalam Data Mining khususnya klasifikasi.

Koloni Semut

Semut di dalam kehidupan nyata, mampu menemukan jalur terpendek dari sumber makanan dan sarangnya (dapat berubah sesuai dengan perubahan pada lingkungan) tanpa informasi visual. Kemampuan semut yang buta akan keadaan sekitarnya telah dipelajari oleh

para ahli etholog. Mereka menemukan bahwa untuk bertukar informasi mengenai jalur yang harus dilalui, semut-semut saling berkomunikasi dengan penggunaan pheromone (zat kimia). Pada saat semut bergerak, sejumlah pheromone akan dijatuhkan, yang akan menandai jalur yang telah dilaluinya. Semakin banyak semut yang mengikuti jalur ini, akan membuat jejak ini semakin menarik untuk diikuti oleh semut lainnya.

Ketika jalur yang telah terbentuk antara sumber makanan dan sarang terhalang oleh suatu objek, semut-semut akan mencoba untuk mengelilingi rintangan tersebut. Pertama, setiap semut dapat memilih untuk mengelilingi objek tersebut lewat kiri atau kanan dengan kemungkinan 50%-50%. Semua semut bergerak kira-kira dengan kecepatan yang sama dan menjatuhkan pheromone dalam jumlah yang sama pula. Sehingga, semut yang mengelilingi rintangan dengan jarak terpendek akan mencapai jalur utama lebih cepat daripada semut yang melalui jalur yang lebih panjang. Hasilnya, pheromone akan terkumpul lebih cepat pada jalur terpendek. Karena semut-semut lain akan memilih jalur dengan pheromone yang lebih banyak, pada akhirnya semua semut akan beralih ke jalur terpendek (gambar 1).



Gambar 1. Jalur sarang semut dan sumber makanan

Ant Colony Optimization

Algoritma Ant Colony Optimization (ACO) ialah sebuah sistem yang berdasarkan agen-agen yang mensimulasikan perilaku dari semut-semut, termasuk mekanisme kerja sama dan adaptasi. Algoritma ACO dibuat berdasarkan gagasan berikut :

1. Setiap jalur yang dilalui oleh semut diasosiasikan dengan kandidat solusi dari suatu masalah.
2. Jika suatu semut mengikuti suatu jalur, jumlah pheromone pada jalur tersebut sebanding dengan kualitas kandidat solusi yang bersangkutan.
3. Jika suatu semut diharuskan untuk memilih antara dua jalur, jalur yang memiliki jumlah pheromone lebih banyak memiliki peluang lebih besar untuk dipilih semut tersebut.

Hasilnya, semut-semut akan memusat ke jalur yang pendek, dengan harapan menghasilkan solusi yang optimal atau mendekati optimal. Intinya, desain algoritma ACO berdasarkan hal-hal di bawah ini :

1. Dapat merepresentasikan masalah dengan baik, dimana semut-semut secara bertahap membangun dan mengubah solusi dengan menggunakan aturan transisi probabilistik, berdasarkan jumlah pheromone pada suatu jalur dan sebuah fungsi heuristik.
2. Sebuah fungsi heuristik yang sesuai dengan masalah ($\boldsymbol{\eta}$) yang menyatakan kualitas item yang akan ditambahkan ke solusi sementara.
3. Aturan dalam melakukan update jumlah pheromone, yang menjelaskan bagaimana mengubah jumlah pheromone ($\boldsymbol{\tau}$) suatu jalur.
4. Aturan transisi probabilistik berdasarkan nilai fungsi heuristik ($\boldsymbol{\eta}$) dan jumlah pheromone ($\boldsymbol{\tau}$) yang akan digunakan dalam membangun suatu solusi

Ant_Miner : Algoritma ACO untuk Data Mining

Algoritma ACO yang diterapkan dalam Data Mining disebut dengan Ant_Miner. Ant_Miner sampai saat ini telah

dikembangkan menjadi Ant_Miner2 dan Ant_Miner3.

Ant_Miner2

Berdasarkan fungsi heuristik yang digunakan pada Ant_Miner dapat disederhanakan. Hal ini disebabkan algoritma ACO tidak membutuhkan informasi yang akurat mengenai nilai heuristiknya, karena jumlah pheromone mampu mengimbangi kesalahan kecil dalam nilai heuristik. Dengan kata lain, dengan fungsi heuristik yang lebih sederhana mampu menghasilkan nilai yang setara dengan fungsi yang lebih kompleks. Fungsi heuristik yang lebih sederhana dapat dilihat di bawah ini :

$$\eta_{ij} = \frac{majority\_classT_{ij}}{|T_{ij}|} \quad (1)$$

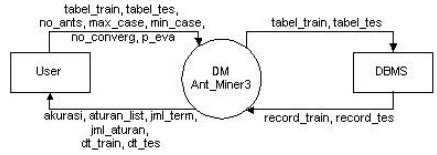
Ant_Miner3

Pada algoritma sebelumnya, semut-semut memilih term berdasarkan jumlah pheromone dan fungsi heuristik. Jumlah pheromone akan diupdate setelah aturan di-pruning, sementara nilai fungsi heuristik selalu sama, sehingga semut-semut berikutnya akan lebih memilih term yang digunakan pada aturan sebelumnya (karena jumlah pheromone yang lebih besar), dan mengabaikan term yang belum pernah digunakan (karena jumlah pheromone yang kecil). Sehingga semut-semut terlalu cepat berkumpul pada sebuah jalur. Hal ini akan membatasi penggunaan term lain yang mungkin lebih baik. Oleh karena itu, pada Ant_Miner3 dilakukan pengubahan pada saat pemilihan term pada aturan sementara dan untuk update jumlah pheromone dalam rangka meningkatkan akurasi aturan yang dihasilkan.

METODE PENELITIAN

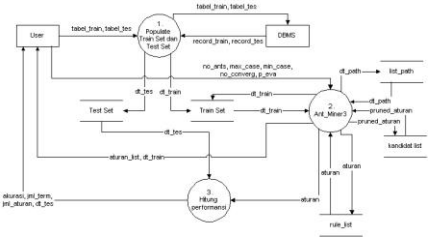
Perancangan sistem menggunakan metode terstruktur dengan *Data Flow Diagram* (Diagram Aliran Data). DFD menggambarkan aliran data dan proses yang terjadi di dalam sistem.

Terdapat dua entitas yang berperan dalam sistem ini, yaitu *User* dan RDBMS (dalam hal ini SQL Server). User berperan sebagai pihak yang mencoba mencari aturan klasifikasi dari dataset yang dipilihnya dan juga melihat performansi system. RDBMS berperan sebagai pihak yang menyediakan data Trainset dan Testset.



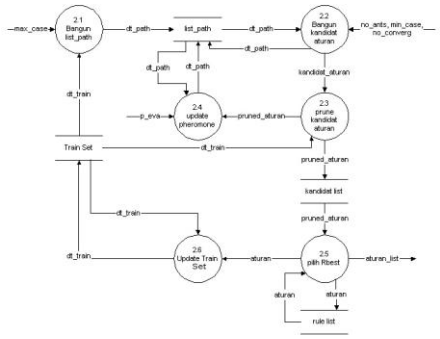
Gambar 2. Diagram Konteks

DFD Level 1 ini menggambarkan sistem secara lebih spesifik, dimana dijelaskan proses demi proses (secara deskriptif).



Gambar 3 Diagram Aliran Data Level 1

DFD Level 2 ini merupakan penjabaran dari proses 2. Pada Level ini dijelaskan bagaimana proses yang terjadi didalam Ant_Miner3, yang meliputi proses mebangun list path, membangun kandidat aturan, pruning kandidat aturan, update nilai pheromone, pemilihan Rbest, dan update data di dalam Train set.



Gambar 4. Diagram Aliran Data Level 2 proses 2 Ant_Miner3

HASIL DAN PEMBAHASAN

Perangkat lunak yang digunakan dalam perancangan sistem adalah sebagai berikut:

1. Sistem operasi Microsoft Windows 7 Professional
2. Microsoft SQL Server 2000
3. Borland Delphi 7.0

Sedangkan perangkat keras yang digunakan dalam perancangan sistem adalah sebagai berikut :

1. Prosesor Celleron 2000 Mhz
2. Memory DDRAM 512 MB
3. Harddisk 40 GB
4. VGA Card 256 MB
5. Keyboard dan mouse

Pada bagian ini akan dilakukan pengujian sistem yang telah dibuat. Dari hasil pengujian akan diperoleh aturan klasifikasi dari setiap Trainset serta parameter performansinya. Adapun parameter yang digunakan sebagai pengukur performansi adalah :

1. Akurasi prediktif
   Merupakan parameter untuk mengukur ketepatan aturan klasifikasi yang dihasilkan dalam mengklasifikasikan data di dalam Testset berdasarkan atribut yang ada ke dalam kelasnya. Dinyatakan dalam persentase (0-100 %). Semakin tinggi nilainya, maka semakin baik autran yang dihasilkan dalam mengklasifikasikan data baru ke dalam kelas yang benar.
2. Simplisitas aturan
   Merupakan parameter yang menyatakan kesederhanaan aturan yang dihasilkan. Parameter ini dinyatakan dalam jumlah rata-rata tiap term pada bagian antisenden setiap aturan. Semakin kecil nilainya maka semakin sederhana aturan yang dihasilkan, sehingga semakin mudah untuk dipahami.

Selain itu, akan dilakukan analisa juga mengenai parameter sistem yang digunakan didalam Ant_Miner3, yaitu :

1. No_of_ants : jumlah maksimal semut yang digunakan, berarti juga jumlah maksimal kandidat aturan yang dihasilkan
2. Min_cases_per_rule : jumlah case terkover yang harus dipenuhi oleh setiap kandidat aturan yang dihasilkan
3. Max_uncovered_cases : jumlah maksimal case yang tidak terkover pada Trainingset setelah dijalankan algoritma Ant_Miner3
4. No_rules_converg : jumlah aturan yang menyatakan bahwa semut-semut Ant_Miner3 telah konvergen ke satu jalur.
5. ρ (pheromone evaporation) : menyatakan tingkat penguapan pheromone.

Dataset yang digunakan disini adalah dataset dari repository UCI (University of California). Dataset yang dipilih adalah Breast Cancer, Tic-tac-toe, dan House Votes. Ketiga dataset ini dipilih karena atributnya tidak ada yang kontinu, dan tidak ada nilai yang hilang (*missing values*) atau null di dalamnya, serta data didalamnya acak. Karena aplikasi See5 yang digunakan sebagai pembanding masih berupa versi Demo, maka penggunaan dataset masih dibatasi yaitu hanya sebanyak 400 case.

Tabel 1. Karakteristik Dataset

Dataset	Jumlah kasus	Jumlah atribut	Jumlah kelas
Breast Cancer	390	10	2
Tic-tac-toe	400	10	2
House Votes	232	17	2

Pada bagian ini akan diperlihatkan tabel hasil pengujian secara keseluruhan dan analisa keseluruhan.

Tabel 2. Akurasi Prediktif Ant_Miner3 dan See5

	Akurasi Prediktif (%)	
Dataset	Ant_Miner3	See5
Breast Cancer	97.18	93.20
Tic-tac-toe	96.50	89.00
House Votes	98.69	96.10

Tabel 3. Perbandingan Akurasi Prediktif Ant_Miner3 dengan Pruning dan Tanpa Pruning

Dataset	Akurasi Prediktif (%) dengan pruning	Akurasi Prediktif (%) tanpa pruning
Breast Cancer	97.18	97.69
Tic-tac-toe	96.50	97.00
House Votes	98.69	97.39

Tabel 4. Perbandingan Akurasi Prediktif Ant_Miner3 dengan Pheromone dan Tanpa Pheromone

Dataset	Akurasi Prediktif (%) dengan pheromone	Akurasi Prediktif (%) tanpa pheromone
Breast Cancer	97.18	95.89
Tic-tac-toe	96.50	95.75
House Votes	98.69	96.52

Tabel 5. Akurasi Prediktif Ant_Miner3 untuk kombinasi no_of_ants dan no_rules_converg

Dataset Parameter	Breast Cancer	Tic-tac-toe	House Votes
No_of_ants=1000, no_ruls_converg=5	97.69	96.00	97.83
No_of_ants=1000, no_ruls_converg=10	96.92	92.75	97.39
No_of_ants=3000, no_ruls_converg=5	97.44	98.00	98.26
No_of_ants=3000, no_ruls_converg=10	97.69	94.25	97.39

Tabel 6. Akurasi Prediktif Ant_Miner3 untuk kombinasi max_uncovered_case dan min_cases_per_rule

Parameter Dataset	(max_uncovered_case, min_cases_per_rule)								
	5,5	5,10	5,15	10,5	10,10	10.15	15.5	15.10	15.15
Breast Cancer	95.13	95.38	93.33	97.18	95.64	92.82	93.08	95.13	93.59
Tic-tac-toe	93.75	95.25	91.00	92.00	90.50	88.25	83.50	89.25	89.25
House Votes	96.96	94.35	96.09	94.78	93.91	96.96	95.65	93.91	96.96

Tabel 7. Akurasi Prediktif Ant_Miner3 untuk kombinasi pheromone evaporation

Dataset	Akurasi Prediktif (%)				
	ρ=0.1	ρ=0.3	ρ=0.5	ρ=0.7	ρ=0.9
Breast Cancer	94.87	95.13	95.13	95.39	94.36
Tic-tac-toe	89.25	91.00	96.25	89.50	93.75
House Votes	94.35	93.91	93.91	93.48	93.91

Dengan melihat hasil yang diperoleh pada tahap analisa diatas, maka dapat ditarik kesimpulan bahwa akurasi prediktif Ant_Miner3 dapat meningkat jika dilakukan dengan menggunakan pheromone. Nilai no_of_ants yang semakin besar dan nilai ρ yang kecil juga dapat meningkatkan akurasi prediktifnya. Sementara untuk parameter max_uncovered_cases dan min_cases_per_rule, semakin kecil nilainya dapat meningkatkan akurasi prediktifnya.

Tabel 8. Simplisitas Aturan Ant_Miner3 dan See5

Simplisitas	Ant_Miner3		See5	
Dataset	Jml aturan	Jml term	Jml aturan	Jml term
reast Cancer	15.00	1.02	14.40	1.15
Tic-tac-toe	9.90	1.06	21.70	2.89
House Votes	5.70	1.25	2.40	1.15

Tabel 9. Perbandingan Simplisitas Aturan Ant_Miner3 dengan Pruning dan Tanpa Pruning

Simplisitas	Jml aturan		Jml term	
Dataset	Pruning	Tanpa pruning	pruning	Tanpa pruning
Breast Cancer	15.00	15.90	1.02	1.06
Tic-tac-toe	9.90	23.50	1.06	1.85
House Votes	5.70	7.90	1.25	1.87

Tabel 10. Perbandingan Simplisitas Aturan Ant_Miner3 dengan Pheromone dan Tanpa Pheromone

Simplisitas	Jml aturan		Jml term	
Dataset	Pheromone	Tanpa phero mone	Pheromone	Tanpa phero mone
Breast Cancer	15.00	15.00	1.02	1.02
Tic-tac-toe	9.90	8.50	1.06	1.08
House Votes	5.70	5.30	1.25	1.19

Tabel 11. Simplisitas Aturan Ant_Miner3 untuk kombinasi no_of_ants dan no_rules_converg

Simplisitas	Jml aturan				Jml term			
Dataset	(no_of_ants, no_rules_converg)				(no_of_ants, no_rules_converg)			
	1000,5	1000,10	3000,5	3000,10	1000,5	1000,10	3000,5	3000,10
Breast Cancer	15.00	14.90	15.40	12.40	1.01	1.02	1.01	1.03
Tic-tac-toe	10.30	8.30	9.50	9.30	1.06	1.02	1.03	1.07
House Votes	5.60	5.70	5.70	5.70	1.11	1.13	1.16	1.19

Tabel 11. Jumlah Aturan Ant_Miner3 untuk kombinasi max_uncovered_case dan min_cases_per_rule

Parameter Dataset	(max_uncovered_case, min_cases_per_rule)								
	5,5	5,10	5,10	0,5	0,10	0.10	5.5	5.10	5.10
Breast	1 1.	7. 5	5. 3	1 1.	7. 8	5. 4	9. 8	7. 7	5. 0

Cancer	30	0	0	10	0	0	0	0	0
Tic-tac-toe	8.7	8.6	7.7	9.0	8.0	7.4	7.6	7.9	7.6
House Votes	4.5	3.6	3.0	3.6	3.5	3.1	3.1	3.0	3.0

KESIMPULAN

Berdasarkan penelitian yang dilakukan peneliti, maka dapat ditarik kesimpulan :

1. Algoritma Ant_Miner3 dapat digunakan untuk menemukan aturan klasifikasi pada dataset Breast Cancer, Tic-tac-toe dan House Votes.
2. Algoritma Ant_Miner3 mengurangi kasus terhentinya algoritma Ant_Miner dengan penggunaan bilangan random.
3. Dalam hal akurasi prediktif dan simplistas aturan, Ant_Miner3 kompetitif dengan algoritma C5.0 yang banyak digunakan pada aplikasi data mining.
4. Akurasi Ant_Miner3 dapat ditingkatkan dengan beberapa hal, diantaranya menggunakan pheromone, memberikan nilai yang semakin besar untuk parameter no_of_ants dan no_rules_converg serta pheromone evaporation.
5. Simplisitas aturan Ant_Miner3 dapat ditingkatkan dengan beberapa hal, diantaranya menggunakan pruning, memberikan nilai yang semakin besar untuk parameter max_uncovered_case dan min_cases_per_rule.

DAFTAR PUSTAKA

[1] Han, Jiawei dan Micheline Kamber, Data Mining Concepts and Techniques. San Diego : Academic Press, 2001.

[2] Liu, Bo dan Hussein A. Abbas dan Bob McKay, Classification Rule Discovery with Ant Colony Optimization. IEEE Compitational Intellegence Bulletin, 2004.

[3] Parpinelli, Rafael S. dan Heitor S. Lopes dan Alex A. Freitas. Data Mining with an Ant Colony Optimization. Brazil.

[4] Maniezzo, Vittorio dan Luca Maria Gambardella dan Fabio de Luigi, Ant Colony Optimization. European Commission, 2001.

[5] Effendi, Arya Bima, Klasifikasi pada Data Mining Menggunakan Algoritma Ant Colony Optimization. Bandung : STT Telkom, 2004.

[6] Berry, Michael J.A. dan Gordon S. Linoff, Mastering Data Mining. USA : John Wiley & Sons, Inc, 2000.

[7] Middendorf, Martin dan Frank Reischle dan Hartmut Schmeck. Information Exchange in Multi Colony Ant Algorithm. Karslruhe, Germany.