A New Algorithm to Categorize E-mail Messages to Folders with Social Networks Analysis

Urszula Boryczka, Barbara Probierz(✉), and Jan Kozak

Institute of Computer Science, University of Silesia,
Będzińska 39, 41–200 Sosnowiec, Poland
{urszula.boryczka,barbara.probierz,jan.kozak}@us.edu.pl

Abstract. This paper presents a new approach to an automatic categorization of email messages into folders. The aim of this paper is to create a new algorithm that will allow one to improve the accuracy with which emails are assigned to folders (Email Foldering Problem) by using solutions that have been applied in Ant Colony Optimization algorithms (ACO) and Social Networks Analysis (SNA). The new algorithm that is proposed here has been tested on the publicly available Enron email data set. The obtained results confirm that this approach allows one to better organize new emails into folders based on an analysis of previous correspondence.

Keywords: E-mail Foldering Problem · Ant Colony Optimization · Enron E-mail · Social Network Analysis

1 Introduction

In today's world, the media and social communication are among the most important factors in and methods of shaping contemporary social and economic reality. Communication via the Internet is a very simple yet convenient way of transmitting information. However, the biggest problem for the users is how to properly organize electronic mail and assign email messages to particular folders, especially when these are to be categorized automatically.

Email foldering is a complex problem because an automatic classification method can work for one user while for another it can lead to errors, which is often because certain users create new catalogs, but they also stop using some of the folders that were created earlier. At the same time folders do not always correspond to email subjects and some of them can refer to the tasks to do, project groups and certain recipients, while others only make sense in relation to previous email messages. Moreover, information items come at different times, which creates additional difficulties. A similar problem was dealt with in a paper [3] that presented a case study of benchmark email foldering based on the example of the Enron email data set. R. Bekkerman et al. classified emails from seven email boxes into thematic folders based on four classifiers: Maximum Entropy, Naive Bayes, Support Vector Machine and Wide-Margin Winnow.

Based on an analysis of this problem, the aim of the study was established, i.e. to create a new algorithm which would make it possible to better match new

M. Núñez et al. (Eds.): ICCCI 2015, Part II, LNCS 9330, pp. 89–98, 2015.
DOI: 10.1007/978-3-319-24306-1_9

emails to particular folders. The proposed method is based on the Ant Colony Optimization algorithm methodology and social networks. Social network analysis concerns the contacts between the sender and the recipients of email messages and it deals with studying the structure of folders that have been created in particular inboxes, which here was additionally marked with the pheromone trail.

This algorithm was used in a previously prepared data set of emails which had been obtained from the public Enron email data set especially for this purpose. The experiments that were conducted have confirmed that the proposed algorithm makes it possible to improve the accuracy with which email messages are assigned to folders.

This article is organized as follows. Section 1 comprises an introduction to the subject of this article. Section 2 describes Enron E-mail Dataset. Section 3 Social Network Analysis is presented. In section 4, describes Ant Colony Optimization. Section 5 focuses on the presented, new version approach based on ACO algorithm and SNA. Section 6 presents the experimental study that has been conducted to evaluate the performance of the proposed algorithm, taking into consideration Enron E-mail Dataset. Finally, we conclude with general remarks on this work and a few directions for future research are pointed out.

2 Enron E-mail Dataset

The Enron email data set constitutes a set of data which were collected and prepared as part of the CALO Project (a Cognitive Assistant that Learns and Organizes). It contains more than 600,000 email messages which were sent or received by 158 senior employees of the Enron Corporation. A copy of the database was purchased by Leslie Kaelbling with the Massachusetts Institute of Technology (MIT), and then it turned out that there were serious problems associated with data integrity. As a result of work carried out by a team from SRI International, especially by Melinda Gervasio, the data were corrected and made available to other scientists for research purposes.

This database consists of real email messages that are available to the public, which is usually problematic as far as other data sets are concerned due to data privacy. These emails are assigned to personal email accounts and divided into folders. There are no email attachments in the data set and certain messages were deleted because their duplicate copies could be found in other folders. The missing information was reconstructed, as far as possible, based on other information items; if it was not possible to identify the recipient the phrase no_address@enron.com was used. The Enron data set is commonly used for the purpose of studies that deal with social network analysis, natural language processing and machine learning.

The Enron email data set is often used to create a network of connections that represent the informal structure of an organization [13] and the flow of information within a company [7]. J. Shetty and J. Adibi constructed a database scheme on the basis of the Enron data set [15] and they also carried out research by using the entropy model to identify the most interesting and the most important nodes

in the graph [16]. Studies on the dynamic surveillance of suspicious individuals were conducted by Heesung Do, Peter Choi and Heejo Lee and presented in a paper [9] by using a method based on the Enron data set.

3 Social Network Analysis

A Social Network is a multidimensional structure that consists of a set of social entities and the connections between them. Social entities are individuals who function within a given network, whereas connections reflect various social relations between particular individuals. The first studies of social networks were conducted in 1923 by Jacob L. Moreno, who is regarded as one of the founders of social network analysis. SNA is a branch of sociology which deals with the quantitative assessment of the individual's role in a group or community by analyzing the network of connections between individuals. Moreno's 1934 book that is titled Who Shall Survive presents the first graphical representations of social networks as well as definitions of key terms that are used in an analysis of social networks and sociometric networks [14].

A social network is usually represented as a graph. According to the mathematical definition, a graph is an ordered pair $G = (V, E)$, where V denotes a finite set of a graph's vertices, and E denotes a finite set of all two-element subsets of set V that are called edges, which link particular vertices such that:

$$E \subseteq \big\{\{u, v\} : u, v \in V, u \neq v\big\}. \tag{1}$$

vertices represent objects in a graph whereas edges represent the relations between these objects. Depending on whether this relation is symmetrical, a graph which is used to describe a network can be directed or undirected.

Edges in a social network represent interactions, the flow of information and goods, similarity, affiliation or social relationships. Therefore, the strength of a connection is measured based on the frequency, reciprocity and type of interactions or information flow, but this strength also depends on the attributes of the nodes that are connected to each other and the structure of their neighborhood.

The degrees of vertices and the degree centrality of vertices are additional indicators that characterize a given social network. The degree of a vertex (indegree and outdegree) denotes the number of head endpoints or tail endpoints adjacent to a given node. Degree centrality is useful in determining which nodes are critical as far as the dissemination of information or the influence exerted on immediate neighbors is concerned. Centrality is often a measure of these nodes' popularity or influence. The probability that the immediate neighbors of vertex v are also each other's immediate neighbors is described by the clustering coefficient gc_v of vertex v such that:

$$gc_v = \frac{2E_v}{k_v(k_v - 1)}, \qquad k_v > 1, \tag{2}$$

where E_v is the number of edges k_v between the neighbors of vertex v.

Many studies that were carried out as part of SNA were aimed at finding correlation between a network's social structure and efficiency [12]. At the beginning,

social network analysis was conducted based on questionnaires that were filled out by hand by the participants [8]. However, research carried out by using email messages has become popular over time [1]. Some of the studies found that research teams were more creative if they had more social capital [11]. Social networks are also associated with discovering communication networks. The database which was used in the experiments that are presented in this article can be used to analyze this problem. G. C. Wilson and W. Banzhaf, among others, discussed such an approach, which they described in their article [19].

4 Ant Colony Optimization – Ant Colony Decision Trees

The inspiration for creating Ant Colony Optimization algorithms came from the desire to learn how animals that are almost blind, such as ants, are able to find the shortest path from the nest to the food source. The studies and experiments that were conducted by S. Goss and J.L. Deneubourg as well as those that were presented in papers [2,17,18] and aimed to explain how nature accomplishes this task constituted the first step toward implementing this solution in algorithms. However, only the attempts made by M. Dorigo [10], to create an artificial ant colony system and use it to find the shortest path between the vertices of a given graph were the key step toward creating ACO algorithms.

While searching for food, agent-ants create paths, on which they lay a pheromone trail. This allows them to quickly return to the nest and communicate information about the location of food to other ants. Based on feedback, the shortest paths, i.e. paths that are assigned large pheromone trail values, between the nest and the food source are created.

An agent-ant makes every decision concerning another step in accordance with the formula:

$$j = \begin{cases} \arg\max\limits_{r \in J_i^k} \{[\tau_{ir}(t)] \cdot [\eta_{ir}]^{\beta}\}, \text{ if } q \leq q_0 \\ p_{ij}^k(t), \qquad\qquad \text{otherwise,} \end{cases} \tag{3}$$

where τ_{ij} is the value of the reward, i.e. the degree of usefulness of the decision option that is being considered; η_{ir} is the value of the quality of a transition from state i to state r which was estimated heuristically; β is the parameter describing the importance of values η_{ir}; $p_{ij}^k(t)$ is the next step (decision) which was randomly selected by using the probabilities:

$$p_{ij}^k(t) = \begin{cases} \frac{\tau_{ij}(t) \cdot [\eta_{ij}]^{\beta}}{\sum\limits_{r \in J_i^k} \tau_{ir}(t) \cdot [\eta_{ir}]^{\beta}}, \text{ if } j \in J_i^k \\ 0, \qquad\qquad \text{otherwise,} \end{cases} \tag{4}$$

where J_i^k denotes the set of decisions that ant k can make while being in state i.

After a single ant-agent has covered the whole distance, pheromone trail is laid out on each edge it has visited (i, j). Let us assume that τ_{ij} denotes the pheromone trail value on the edge (i, j):

$$\tau_{ij}(t+1) = (1-\rho) \cdot \tau_{ij}(t) + \rho \cdot \tau_0, \tag{5}$$

where ρ denotes a coefficient such that $0 \leq \rho \leq 1$ represents the residue of the pheromone trail, whereas τ_0 value equal to the initial pheromone value.

Ant Colony Decision Trees (ACDT) algorithm [5] employs Ant Colony Optimization techniques for constructing decision trees. As mentioned before, the value of the heuristic function is determined according to the splitting rule employed in CART approach [6]. These algorithms were described in [4].

The evaluation function for decision trees will be calculated according to the following equation:

$$Q(T) = \phi \cdot w(T) + \psi \cdot a(T, P), \tag{6}$$

where $w(T)$ is the size (number of nodes) of the decision tree T; $a(T, P)$ is the accuracy of the classification object from a training set P by the tree T; ϕ and ψ are constants determining the relative importance of $w(T)$ and $a(T, P)$.

5 Proposed Algorithm

The previously conducted studies that are described in paper [4] confirmed the validity of using decision tables and ACO algorithms, whereas the creation of the social network of electronic mail users as well as an analysis of this network provided inspiration for creating a new algorithm that would make it possible to properly assign email messages to folders. The way in which this algorithm works is presented in Fig. 1.

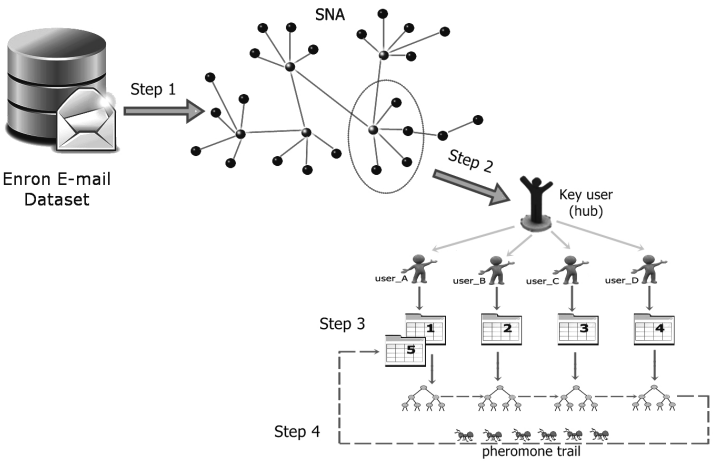

Fig. 1. Diagram of the proposed algorithm

The first step (Fig. 1 - step 1, Alg. 1 - line 1) that was taken to use this method for improving the accuracy with which email messages are assigned to folders was to create a social network based on the contacts between the sender and the recipients of emails obtained from the Enron data set. All users of the mailboxes obtained from the Enron email data set constitute the vertices,

whereas the connections for which the frequency of interactions, i.e. the number of emails sent between particular individuals, was more than 10 constitute the edges. Then, during the analysis of this network, key users (hubs) and their immediate neighbors were selected (Fig. 1 - step 2, Alg. 1 - line 2), thus creating groups of users within the network, which are presented in Tab. 1.

Table 1. Selected groups of users

Name of groups	Key user (hub)	Hub's closest neighbors
Group 1	lokay-m	hyatt-k, mcconnell-m, schoolcraft-d, scott-s, watson-k
Group 2	sanders-r	cash-m, dasovich-j, haedicke-m, kean-s, sager-e, steffes-j
Group 3	shackleton-s	jones-t, mann-k, stclair-c, taylor-m, ward-k, williams-j
Group 4	steffes-j	dasovich-j, gilbertsmith-d, presto-k, sanders-r, shapiro-r
Group 5	symes-k	scholtes-d, semperger-c, williams-w3
Group 6	williams-w3	mann-k, semperger-c, solberg-g, symes-k
Group 7	farmer-d	bass-e, beck-s, griffith-j, nemec-g, perlingiere-d, smith-m
Group 8	beck-s	buy-r, delainey-d, hayslett-r, kaminski-v, kitchen-l, may-l, mcconnell-m, shankman-j, white-s
Group 9	rogers-b	baughman-d, davis-d, griffith-j, kitchen-l, lay-k

Key users constitute the network's nodes with the highest vertex degree. The users who have no connections within the network are those individuals who received fewer than the predetermined number of emails. The parameters of selected groups are presented in Tab. 2.

Table 2. Parameters of selected groups of users.

Name of groups	Key user (hub) in group	N. of classes (folders) in group	N. of objects (emails) in group	N. of classes (folders) for the key user	N. of objects (emails) for the key user
Group 1	lokay-m	87	4963	11	2493
Group 2	sanders-r	230	8024	29	1181
Group 3	shackleton-s	126	4131	39	886
Group 4	steffes-j	154	4247	21	617
Group 5	symes-k	35	3598	11	767
Group 6	williams-w3	58	5138	18	2767
Group 7	farmer-d	156	5998	24	3538
Group 8	beck-s	185	6522	84	1703
Group 9	rogers-b	81	2963	14	1395

The next step that is to be taken in order to implement the proposed algorithm is to transform the set of data from the Enron email data set into a decision table, separately for each mailbox within a given group (Fig. 1 - step 3, Alg. 1 - line 6). The prepared decision table consists of six conditional attributes and one decision attribute *category*, which defines the folder to which the email is assigned. Conditional attributes were selected to define the most important information about each message. They consist of the information from the sender

field, the first three words from the email subject (with the exception of basic words and copulas) where additionally, words which belong to the set of decision classes are supported. The length of the message, and the information about the person who received the message was added to a courtesy copy (CC) (as the Boolean value) is also checked by the conditional attributes. If the person is not in a courtesy copy - the person is the recipient.

After the ACO algorithm has been run, a classifier is repeatedly constructed based on training data (Alg. 1 - line 14). Each classifier is tested on test data (Alg. 1 - line 15) and a pheromone trail is laid depending on the obtained results (Alg. 1 - line 20). While a classifier is being developed for each user, the communication network of the group that is assigned to this user is being analyzed. A classifier is built for a selected user via the algorithm, and then a classifier is constructed for each subsequent individual in the group by using the same pheromone trail matrix. After the classifiers have been built and the pheromone trail matrices have stabilized, the final classifier for a given user is developed (Alg. 1 - lines 21-24), in accordance with the diagram in Fig. 1 - step 4, which makes it possible to retain information related to the decisions made by the other members of the group (via the pheromone trail).

Algorithm 1. Pseudo code of the proposed algorithm

```
1  set_of_hubs = social_network_analysis(enron_e-mail_dataset);
2  user's_group = SNA_group(set_of_hubs); //For the chosen hub //eq(2)
3  //Hub is the first
4  for person=1 to number_of_users_in_the_group do
5    dataset[person]=prepare_decision_tables(person)
6  endFor
7  pheromone = initialization_pheromone_trail(); //Common to all users
8  //The first and last iteration for hub
9  for person=1 to (number_of_users_in_the_group+1) do
10   for i=1 to (number_of_iterations / (number_of_users+1)) do
11     best_classifier = NULL;
12     for j=1 to number_of_ants do
13      new_classifier = build_classifier_aACDT(pheromone, dataset[person]);//eq(3)
14       assessment_of_the_usefulness_classifier(new_classifier);
15       if new_classifier is_higher_usefulness_than best_classifier then//eq(6)
16         best_classifier = new_classifier;
17       endIf
18     endFor
19     update_pheromone_trail(best_classifier, pheromone); //eq(5)
20     //Only in last iteration - for hub
21     if person == (number_of_users_in_the_group+1) then
22       if best_classifier is_higher_usefulness_than best_built_class then//eq(6)
23         best_built_class = best_classifier;
24       endIf
25     endIf
26   endFor
27 endFor
28 result = best_built_class;
```

Information about the pheromone trail that has been laid is then passed on as feedback to the classifier that is being built for a given user, which has an influence on which folder will be chosen by this classifier. After the algorithm has finished running, the best classifier is obtained, which was developed based on test and training data (Alg. 1 - line 29). Validation data are regarded as future data, for which the accuracy with which email messages are assigned to folders is computed. The way in which this algorithm works is presented in Alg. 1.

6 Experiments

In order to check the usefulness of the proposed solution, experiments were conducted, the results of which are presented in Tab. 3. The proposed algorithm was implemented in C++. All computations were carried out on a computer with an Intel Core i5 2.27 GHz processor, 2.9 GB RAM, running on the Debian GNU/Linux operating system.

The construction of the social network and the selection of the group of users based on an analysis of this social network were carried out in a deterministic manner, as described in Section 5. As for the ACO algorithm, experiments were conducted and repeated 30 times for each of the data sets. Each experiment was carried out for 250 generations consisting of 25 agent-ants. The algorithm parameters were as follows: $q_0 = 0.3$; $\beta = 3$; $\rho = 0.1$. The results shown in Tab. 3 represent the results from the average of 30 executions.

Table 3. Comparison of all results with regard to the usefulness of the decision with which emails are assigned to folders

dataset	Classical algorithms presented in [3]				Previous algorithm	Proposed
	MaxEnt	Naive Bayes	SVM	WMW	presented in[4]	algorithm
beck-s	0.558	0.320	0.564	0.499	0.583	**0.600**
farmer-d	0.766	0.648	0.775	0.746	0.811	**0.834**
lokay-m	0.836	0.750	0.827	0.818	0.888	**0.891**
sanders-r	0.716	0.568	0.730	0.721	0.829	**0.871**
williams-w3	0.944	0.922	0.946	0.945	**0.962**	0.960
rogers-b	–	0.772	–	–	**0.911**	0.900
shackleton-s	–	0.667	–	–	0.709	**0.751**
steffes-j	–	0.755	–	–	0.841	**0.863**
symes-k	–	0.789	–	–	0.930	**0.937**

The obtained results, which are presented in Tab. 3 and Fig. 2, indicate that there was a significant improvement in the classification of emails when the proposed algorithm was used. The results for the other algorithms are cited based on papers [3,4].

The proposed algorithm each time generated better results, as compared to classical algorithms, for nine data sets that had been created based on nine groups of users. For some of the sets (sanders-r, symes-k), the accuracy with which email messages were assigned to folders increased by as much as 15%.

For two of the sets (rogers-b, williams-w3), the accuracy with which email messages are assigned to folders is very high (93-96%), but the use of social network analysis did not lead to the improvement of the results.

In addition, the comparison between the results yielded by the proposed algorithm and those produced by the algorithm that is described in paper [4] shows that the accuracy with which folders are assigned to emails improved by 1-3% (beck-s, farmer-d, steffes-j, symes-k) , whereas for sets sanders-r and shackleton-s this accuracy is even 5% higher, which confirms the validity of using social networks and analyzing mailboxes for a group of users, rather than individuals.

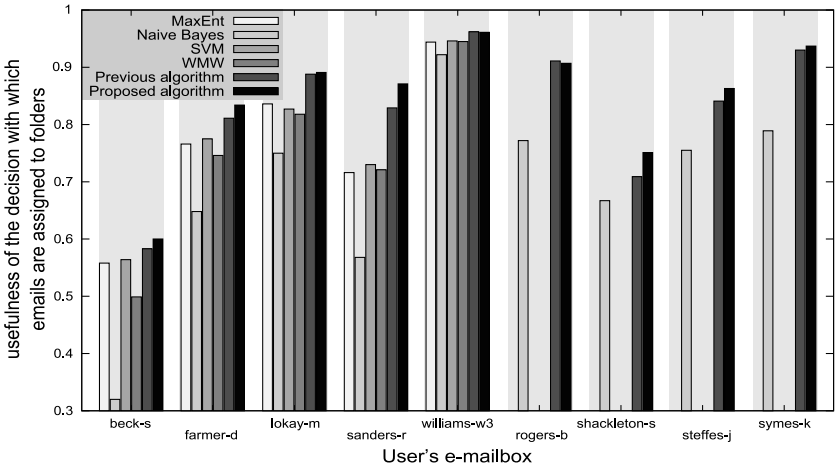

Fig. 2. The correctness of the proposed categorization method

7 Conclusions

Based on the conducted experiments, it has been confirmed that the accuracy with which email messages are automatically classified into folders improves when Ant Colony Optimization algorithms and social network analysis are used. The results improved significantly after creating a social network, based on which communication between the users had been analyzed.

The aim of this study has been achieved and the results that have been obtained are highly satisfactory, given that the tests were carried out on uncleaned data sets. However, the observations that were made during the experiments allow one to assume that this algorithm can be successfully used to suggest that new folders be created based on an analysis of the members of a group. If a rigid structure of folders is imposed on the users in a company, this should even further improve the accuracy with which email messages are assigned to proper folders, which we intend to study in the future.

References

1. Aral, S., Van Alstyne, M.: Network structure & information advantage (2007)
2. Beckers, R., Goss, S., Deneubourg, J., Pasteels, J.M.: Colony size, communication and ant foraging strategy. Psyche **96**, 239–256 (1989)
3. Bekkerman, R., McCallum, A., Huang, G.: Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora. Center for Intelligent Information Retrieval, Technical Report IR (2004)
4. Boryczka, U., Probierz, B., Kozak, J.: An ant colony optimization algorithm for an automatic categorization of emails. In: Hwang, D., Jung, J.J., Nguyen, N.-T. (eds.) ICCCI 2014. LNCS, vol. 8733, pp. 583–592. Springer, Heidelberg (2014)
5. Boryczka, U., Kozak, J.: Enhancing the effectiveness of ant colony decision tree algorithms by co-learning. Applied Soft Computing **30**, 166–178 (2015). http://www.sciencedirect.com/science/article/pii/S1568494615000575
6. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York (1984)
7. Chapanond, A., Krishnamoorthy, M., Yener, B.: Graph theoretic and spectral analysis of enron email data. Computational and Mathematical Organization Theory **11**(3), 265–281 (2005)
8. Cummings, J.N., Cross, R.: Structural properties of work groups and their consequences for performance. Social Networks **25**, 197–210 (2003)
9. Do, H., Choi, P., Lee, H.: Dynamic surveillance: a case study with enron email data set. In: Kim, Y., Lee, H., Perrig, A. (eds.) WISA 2013. LNCS, vol. 8267, pp. 81–99. Springer, Heidelberg (2014)
10. Dorigo, M., Caro, G.D., Gambardella, L.: Ant algorithms for distributed discrete optimization. Artif. Life **5**(2), 137–172 (1999)
11. Gloor, P., Grippa, F., Putzke, J., Lassenius, C., Fuehres, H., Fischbach, K., Schoder, D.: Measuring social capital in creative teams through sociometric sensors. International Journal of Organisational Design and Engineering (2012)
12. Gloor, P.A.: Swarm Creativity: Competitive Advantage through Collaborative Innovation Networks. Oxford University Press, USA (2006)
13. Keila, P., Skillicorn, D.: Structure in the enron email dataset. Computational and Mathematical Organization Theory **11**(3), 183–199 (2005)
14. Moreno, J.L.: Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama. Beacon House, Beacon (1953, 1978)
15. Shetty, J., Adibi, J.: The enron email dataset database schema and brief statistical report. Tech. rep. (2004)
16. Shetty, J., Adibi, J.: Discovering important nodes through graph entropy the case of enron email database. In: Proceedings of the 3rd International Workshop on Link Discovery, LinkKDD 2005, pp. 74–81. ACM, New York (2005)
17. Theraulaz, G., Goss, S., Gervet, J., Deneubourg, J.: Swarm intelligence in wasps colonies: an example of task assignment in multiagents systems. In: Proceedings of the 1990 IEEE International Symposium on Intelligent Control, pp. 135–143 (1990)
18. Verhaeghe, J., Deneubourg, J.: Experimental study and modelling of food recruitment in the ant tetramorium impurum. Insectes Sociaux **30**, 347–360 (1983)
19. Wilson, G.C., Banzhaf, W.: Discovery of email communication networks from the enron corpus with a genetic algorithm using social network analysis (2009)