The Mechanism to Predict Folders in Automatic Classification Email Messages to Folders in the Mailboxes

Barbara Probierz(✉)

Faculty of Informatics and Communication, Chair of Knowledge Engineering, University of Economics, Katowice, Poland
barbara.probierz@ue.katowice.pl

Abstract. This paper was proposed a new method for suggesting creating folders in users mailboxes by using Ant Colony Optimization algorithms and Social Networks Analysis. The aim of this paper is to create a mechanism to predict new folders in automatic classification email messages to folders in the mailboxes. The proposed algorithm uses the elements of Social Networks Analysis used to determine the groups of users who have a similar folder structure in mailboxes, on the basis of which the mechanism suggest new folders for the users. The operation of the proposed method has been tested on a public Enron E-mail Dataset.

Keywords: Enron E-mail Dataset · Ant Colony Optimization
Social network analysis · Predict folders

1 Introduction

One of the biggest problems the users, for whom e-mail is the basis of communication, is the correct ordering of email and assign messages to specific folders. Especially if this process is to be carried out in an automatic way. For this reason, there is an increasing interest in creating systems that automatically manage users' e-mails.

Unfortunately the problem automatically suggest to create folders and assign to them the email message is a very personalized, as it depends on the individual preferences of people interacting with each other. Mirrored these preferences can be represented in the form of social network, which allows the analysis to better understand the behavior of the users of mailboxes. However, through the use of Ant Colony Optimization algorithms, it is possible to search for a larger part of the solutions space exploration and exploitation.

The aim of the work is to create a social network based on contacts between senders and recipients of e-mail messages. Then during analysis and observation of social network users group are extracted with a similar structure of folders in mailboxes, on the basis of, which is a mechanism for suggesting the establishment of new folders in users mailboxes. The proposed method was applied to the

N. T. Nguyen et al. (Eds.): ICCCI 2018, LNAI 11056, pp. 355–365, 2018.
https://doi.org/10.1007/978-3-319-98446-9_33

collection of e-mail messages with the selected mailboxes from public Enron E-mail Dataset.

This article is organized as follows. Section 1 comprises an introduction to the subject of this article. Section 2 describes Social Network Analysis. Section 3 describes Ant Colony Decision Tree approach. In Sect. 4, Enron e-mail dataset is presented. Section 5 focuses on the presented a new method for suggesting creating folders in user's mailboxes by using Social Networks Analysis and Ant Colony Optimization algorithms. Section 6 presents the experimental study for the proposed method for suggesting creating folders for selected users mailboxes from public Enron E-mail Dataset. Finally, we conclude with general remarks on this work and a few directions for future research are pointed out.

2 Social Network Analysis

A network is a structure representation consisting of two elements: nodes and links that define the relationship between these nodes. On the other hand, social networks are networks in which the nodes are people or groups of people (e.g. teams, organizations), and links between these people, for example, relationships of knowledge, communication and dependence at work.

In 1923, Jacob L. Moreno conducted the first social network research and was recognized as one of the founders of the social network analysis discipline. 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 [10].

A social network is represented as a graph. According to the mathematical definition, a graph is an ordered pair:

$$G = (V, E), \tag{1}$$

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:

$$E \subseteq \big\{\{u, v\} : u, v \in V, u \neq v\big\}. \tag{2}$$

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 the flow of information, interactions, social relationships or similarity. The strength of a connection depends on the attributes of the nodes (e.g. the degree of relationship) that are connected to each other and the structure of their neighborhood (e.g. the number of common neighbors) but this strength also is measured based on the frequency, reciprocity and type of interactions or information flow.

The main indicators, that characterize a given social network, are degrees of vertices and the degree centrality of vertices. The degree of a vertex (indegree and outdegree) denotes the number of head endpoints or tail endpoints adjacent to a given node such that:

$$deg(v) = \sum_{u=1}^{n} k_{v,u}, \tag{3}$$

where k_{vu} is the edge between the vertex v and the vertex u.

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:

$$gc_v = \frac{2E_v}{k_v(k_v - 1)}, \qquad k_v > 1, \tag{4}$$

where E_v is the number of edges k_v between the neighbors of vertex v [12].

In order to find a correlation between the efficiency and social structure of the network, research was carried out by Gloor in [8]. At first, the social networks were examined through surveys filled in by hand by participants [7]. However, later, research was conducted via e-mail [1]. Some studies have shown that research teams were more creative if they had more social capital [9]. Social networks are also associated with the discovery of communication networks. The database used in the experiments G. C. Wilson and W. Banzhaf, can be used to the discovery of communication networks which they described in their article [11].

3 Ant Colony Decision Trees

One of the many popular Ant Colony Optimization algorithms (ACO) used in data mining is the Ant Colony Decision Tree algorithm (ACDT), which is based on using ACO algorithms in the process of optimizing the construction of decision trees. The conducted research has shown that the ACDT algorithm produces very good quality classifiers for many standard problems related to data mining [2]. The ACDT algorithm combines the idea of ACO algorithms with the idea of the CART algorithm. Performing the algorithm consists in selecting a test for each node, based on two factors. The first factor is the maximum value compliant with the criterion for the division of the CART algorithm, and the second factor is the additional information recorded in the form of the pheromone trail [2,3].

In each ACDT step an ant chooses an attribute and its value for splitting the objects in the current node of the constructed decision tree. The choice is made according to a heuristic function and pheromone values. The heuristic function is based on the Twoing criterion, which helps ants select an attribute-value pair which well divides the objects into two disjoint sets, i.e. with the intention that

objects belonging to the same decision class should be put in the same subset. The best splitting is observed when objects are partitioned into the left and right subtrees such that objects belonging to the same decision class are in the same subtree. Pheromone values indicate the best way (connection) from the superior to the subordinate nodes – all possible combinations are taken into account.

As mentioned before, the value of the heuristic function is determined according to the splitting rule employed in CART approach, that is, in the algorithm proposed by Breiman et al. in [6]. The probability of choosing the appropriate split in the node is calculated according to a classical probability used in ACO:

$$p_{i,j} = \frac{\tau_{m,m_{L(i,j)}}(t) \cdot \eta_{i,j}^{\beta}}{\sum_i^a \sum_j^{b_i} \tau_{m,m_{L(i,j)}}(t) \cdot \eta_{i,j}^{\beta}} \tag{5}$$

where:

$\eta_{i,j}$ is a heuristic value for the split using the attribute i and value j; t is a step of the algorithm;

$\tau_{m,m_{L(i,j)}}$ is an amount of pheromone currently available at step t on the connection between

nodes m and $m_{L(i,j)}$ (it concerns the attribute i and value j),

β is the relative importance of the heuristic value.

The value of the heuristic function is determined based on the splitting criteria used in the CART algorithm, i.e. in accordance with the following Eq. (6).

$$\underset{a_j \leq a_j^R, j=1,\ldots,M}{\arg\max} \left(\frac{P_l P_r}{4} \left[\sum_{k=1}^{K} |p(k|m_l) - p(k|m_r)| \right]^2 \right), \tag{6}$$

where:

$p(k|m_l)$ – probability of the occurrence of decision class k in node m_l (in the left subtree),

$p(k|m_r)$ – probability of the occurrence of decision class k in node m_r (in the right subtree),

P_l – probability of object transition to node m_l (in the left subtree),

P_r – probability of object transition to node m_r (in the right subtree),

K – decision classes.

The pheromone trail is updated by increasing pheromone levels on the edges connecting each tree node with its parent node (excepting the root):

$$\tau_{m,m_L}(t+1) = (1-\gamma) \cdot \tau_{m,m_L}(t) + Q(T) \tag{7}$$

where:

$Q(T)$ determines the evaluation function of decision tree (see Eq. (8)),

γ is a parameter representing the evaporation rate, equal to 0.1.

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

$$Q(T) = \phi \cdot w(T) + \psi \cdot a(T, P), \tag{8}$$

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 $a(T, P)$.

4 Enron E-mail Dataset

Enron E-mail Dataset is a collection of data collected and prepared by the CALO project (A Cognitive Assistant that Learns and Organizes). It contains over 600,000 e-mails sent or received by 150 senior employees from Enron Corporation. The data collection was taken over by the Federal Energy Regulatory Commission during the investigation after the collapse of the company, and then it was made public. A copy of the database was purchased by Leslie Kaelbling from the Massachusetts Institute of Technology (MIT), after which it turned out that there are big problems with data integrity in the collection. 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.

Each mailbox employees Enron Corporation is stored in a separate folder and the name of the staff member. In each of the folders are created automatically by the mail system (for example: sent mail, all documents, deleted items) and user-created folders. Within these folders are numbered consecutively e mail.

All messages in the collection of Enron Email Dataset may the same construction. These are text files that contain in the following lines the details i.e.: message ID, date sent, sender's postal address, email address of the recipient, subject, message, recipient, to which a copy of the messages sent, first and last name the sender of the message, the name of the recipient of the message, the name of the folder in which the message, the name of the mailbox, which is a message and the body of the message.

In Table 1 the parameters for the selected mailboxes from a Enron E-mail Dataset. There are data on the number of folders and the number of e-mail messages contained in the mailbox, as well as statistical data on the prevalence of messages in folders.

5 The Proposed Method for Suggesting Creating a New Folders in Mailboxes

The proposed mechanism allows you to suggest new folders for the users, based on the structure of the folders of other users designated by created a social network. The proposed method is based on an analysis of the matrix of pheromone trail created when classifying messages to folders.

Table 1. The parameters for the selected mailboxes from a Enron E-mail dataset

Dataset	Number of objects	Number of class	N. of messages in the folder		
			Avg	Min	Max
lokay-m	2493	11	226.64	6	1159
sanders-r	1188	30	39.6	4	420
shackleton-s	1001	53	18.89	3	259
steffes-j	625	23	27.17	3	317
symes-k	770	12	64.17	3	254
williams-w3	2769	18	153.83	3	1398
farmer-d	3672	25	146.88	5	1192
beck-s	1971	101	19.51	3	166

The same mechanisms suggest assigning messages to new folders are not new, as they are in practice used in some systems. However, it should be noted first of all their other classes of subject matter and the way it works. New folders are automatically generated messages or recognized by using a mail program, as news related to the forums, trade offers or social networking sites.

There is no way to find the algorithms with which it would be possible to suggest the more unusual folders for messages that are not automatically generated. Method not only is suggested by the new folders, but in addition is based on the possibilities afforded by the Social Networks and Ant Colony Optimization algorithms.

Functional diagram of the proposed algorithm with the prediction mechanism folders was shown in Fig. 1. An algorithm to automatically assign messages to folders along with suggesting users to create new folders in their mailboxes is to:

- an analysis not yet received e mail in terms of contacts users
- creation social network based on the contacts between the sender and the recipients of the message (Fig. 1 – step 1),
- extracting group of users who have similar social structure, the basis of the analysis and observation of social network Fig. 1 – step 2),
- processing data set to a decision within the Group (Fig. 1 – step 3),
- algorithm based solutions known from the Ant Colony Optimization algorithms (Fig. 1 – step 4),
- presentation folder prediction mechanism for users, based on the analysis of the classification matrix messages to folders (Fig. 1 – steps 5 and 6).

As a result of the work as designed the algorithm, as well as some form of memory (the decisions of other users) through the pheromone trail it is possible to specify the weight of the new suggestions. To this end, the proposed was the matrix of the pheromone trail, which is an analogy to the confusion matrix, which in this case is to visualize the resulting solution, and do not specify the classification error.

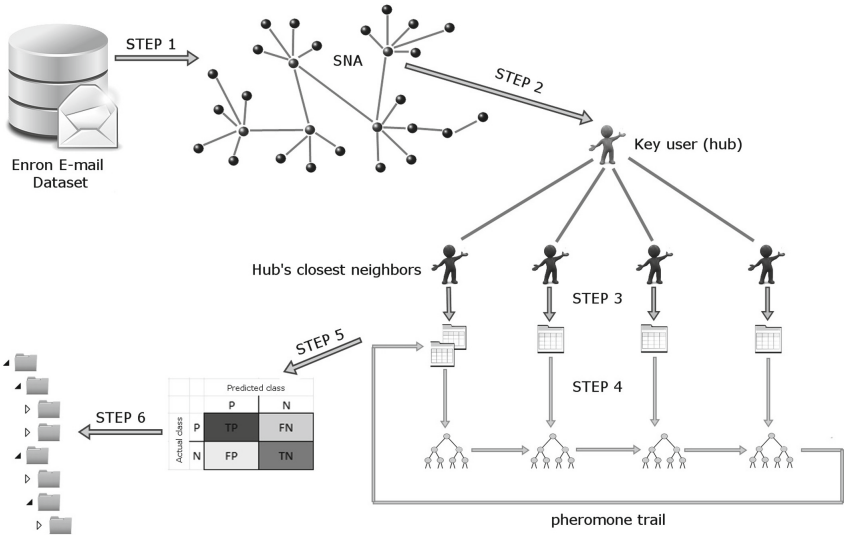

Fig. 1. Diagram of the proposed algorithm with the prediction mechanism folders

The Confusion Matrix is a tool used to assess the quality of classification models, which shows the dependence of the accuracy of the classification of each of the classes and errors that indicate properties classified in another class. Rows in this array correspond to the correct classes, and the columns to the decisions expected by the classifier. The accuracy of the classification of individual classes read based on the intersection of rows with the columns.

6 Experiments

Previous experiments (described in paper [4]) have confirmed the validity of using Ant Colony Optimization algorithms to classify e-mails to folders. After passing through steps 1–4 of this algorithm, the best constructed classifier is obtained, the operation of which is verified on the basis of test data. During the work of the algorithm, a pheromone trail matrix is created (Fig. 1 – step 5), the analysis of which allows the user to suggest new folders (Fig. 1 – step 6).

An essential aspect in this case is to extract the contact group for the user to whom you want to suggested new folders. To this end, in accordance with the established network of contacts, you need to determine the nearest neighbors of this user (treated as the user), and then on the basis of these preferences users make suggestions to create new folders. Create contact groups the user keyword with the immediate neighbours are presented in Table 2.

The main idea of the solution is based on an analysis of the common matrix of the pheromone trail for all users in the group. In the classic version of the proposed algorithm, described in paper [5], despite a group of users, as the available attribute values making only are allowed, which affects the user for

Table 2. 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

which performed is predictive. This is due to the fact that all messages that other users store in their own central folders are omitted.

Studies have been divided into three separate stages due to the breakdown of the data on a set of training and test. In each of the three stages of training set represent mailboxes people adjacent to the central user (key), while the test set is the user's mailbox. Depending on the stage of research test set can entirely be in the training set (stage I), is 50 % included in the training set (stage II), or is a completely new data for the classifier, as there is no common elements in relation to the collection of (stage III).

In this case, the acceptable attribute values making is the sum of values of attributes making all users in the Group (not only), in accordance with the formula:

$$D = D_1 \cup D_2 \cup \ldots \cup D_n, \tag{9}$$

where:

D_i is a set of decision values for the $i - th$ user,
n is the number of users in the group.

In a significant simplification, it can be concluded that if the news of similar characteristics (attributes), other users in the group will keep in a folder that the user does not have a central, it is suggested to create a new folder. As you can see, in this case, it is important to first processing data and customize the folder names to similar to differences arising for example. the folder name is not suggested the difference between folders.

For users of the key groups from Table 2 based on the social network based on the employees of Enron Corporation created and analysed matrices of the pheromone trail when classifying messages to folders. For each user created three matrices of the pheromone trail, in accordance with the three phases of research.

Table 3. The number of suggested folders for selected mailboxes

Dataset	Stage II		Stage III	
	Number of objects	Number of class	Number of objects	Number of class
lokay-m	7	124	3	40
sanders-r	4	849	14	894
shackleton-s	7	262	12	526
steffes-j	5	141	9	286
symes-k	3	550	8	750
williams-w3	2	1151	11	2250
farmer-d	27	564	59	2632
beck-s	12	268	25	727

Suggested the folders that should be created for a given user, are those to which in II or III stage has been classified a lot of messages, while they were not assigned to these folders during the stage I. Based on the analysis of the matrix of the pheromone trail during work the classifier, received the assignment results messages to folders for all users from Table 2 selected in accordance with the established social network. Folders that have been attributed to messages from the user keyword, all folders that contain in the mailboxes of all users in the group in accordance with the formula 9, regardless of the presence of these folders in the user's mailbox.

In the created matrix, on the diagonal as shown is the number of messages assigned correctly to folders, which indicate a classifier is identical to the user-specified folders in the mailbox. While the number of reported beyond the diagonal is a message whose assignment to folders based on folder structures of other users who are immediate neighbours a key user, who together form a group of users.

After the folders to which has not been assigned any message and those that have occurred in the user's mailbox key obtained is a list of the folders that are suggested to be created for that user. For users of the key of Table 2, after analyzing the received matrix of the pheromone trail, in Table 3 shows the results of the prediction mechanism folders, specifying that the minimum number of assigned messages to a folder is 10.

For stage II and stage III shows the number of the suggested folders to create the user's mailbox key due to the assignment of messages to such folders by other users with groups defined by created a social network. Moreover, the number of messages that have been assigned to the newly created folders. Figure 2 shows the number of the suggested folders for the selected mailboxes, while Fig. 3 shows the number of messages classified to the newly created folders.

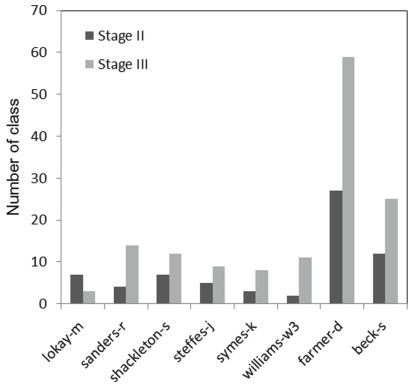

Fig. 2. The number of the suggested folders to create the user's mailbox.

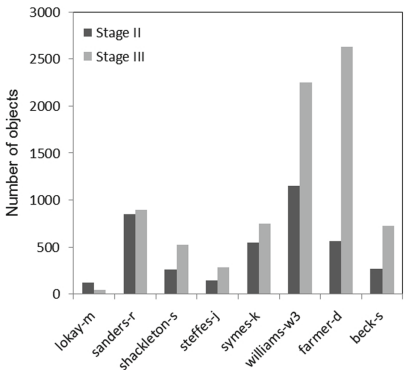

Fig. 3. The number of assigned messages to a new folders.

7 Conclusions

On the basis of experiments confirmed that the use of the social network can successfully contribute to the creation of a mechanism for predicting folders in user's mailboxes. The results depend not only on the high frequency of contacts between individuals, but above all from the subjectively created folder structures of other people. At the same time created arrays allow you to observe the real solutions – often a very large number of messages that are assigned to folders created by other users in relation to the number of remaining messages means that the proposed suggestion create a folder has big support in the case of a group of users. In contrast, the smaller the value, the weaker support suggestions.

The aim of this study has been achieved and the results that have been obtained are highly satisfactory. However, the observations that were made during the experiments allow one to assume that 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 and information advantage. In: Proceedings of the Academy of Management Conference, Philadelphia, PA, vol. 3 (2007)
2. Boryczka, U., Kozak, J.: Ant colony decision trees – a new method for constructing decision trees based on ant colony optimization. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010. LNCS (LNAI), vol. 6421, pp. 373–382. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16693-8_39

3. Boryczka, U., Kozak, J.: An adaptive discretization in the ACDT algorithm for continuous attributes. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011. LNCS (LNAI), vol. 6923, pp. 475–484. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23938-0_48
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 (LNAI), vol. 8733, pp. 583–592. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11289-3_59
5. Boryczka, U., Probierz, B., Kozak, J.: A new algorithm to categorize e-mail messages to folders with social networks analysis. In: Núñez, M., Nguyen, N.T., Camacho, D., Trawiński, B. (eds.) ICCCI 2015. LNCS (LNAI), vol. 9330, pp. 89–98. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24306-1_9
6. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Kluwer Academic Publishers, New York (1984)
7. Cummings, J.N., Cross, R.: Structural properties of work groups and their consequences for performance. Soc. Netw. **25**(3), 197–210 (2003)
8. Gloor, P.A.: Swarm Creativity: Competitive Advantage Through Collaborative Innovation Networks. Oxford University Press, Oxford (2006)
9. Gloor, P.A., Grippa, F., Putzke, J., et al.: Measuring social capital in creative teams through sociometric sensors. Int. J. Organisational Des. Eng. **2**(4), 380–401 (2012)
10. Moreno, J.L.: Who shall survive? Foundations of sociometry, group psychotherapy and socio-drama (1953)
11. Wilson, G., Banzhaf, W.: Discovery of email communication networks from the Enron corpus with a genetic algorithm using social network analysis. In: Evolutionary Computation, pp. 3256–3263. IEEE (2009)
12. Zhang, P., Wang, J., et al.: Clustering coefficient and community structure of bipartite networks. Phys. A: Stat. Mech. Appl. **387**(27), 6869–6875 (2008)