E-mail processing using data mining techniques

Augusto Villa Monte, César Estrebou, Laura Lanzarini

III-LIDI (Institute of Research in Computer Science LIDI)
Faculty of Computer Science, National University of La Plata
La Plata, Buenos Aires, Argentina
{avillamonte, cesarest, laural }@lidi.info.unlp.edu.ar

Abstract. A proposal to use data mining techniques to analyze e-mails corresponding to courses carried out through a distance education platform is made. The purpose of this type of analyses is determining which are the groups of relevant words that allow establishing communication topics of interest. Even though this new information can have various applications, they all involve an improvement in student service. The method proposed has been applied to the e-mails of the PACENI Project (Support Project for Improving First-Year Teaching in Courses of Studies in Exact and Natural Sciences, Economic Science and Computer Science) with satisfactory results.

Keywords : Information Retrieval, Text Mining, e-mails analysis, topic detection.

1 Introduction

Distance education platforms are a learning environment through which teachers and students interact by performing various types of activities.
In this context, electronic mail is the most commonly used mechanism, and it is therefore of interest for the study of techniques that allow analyzing and modeling the information shared through this medium. For example, it would be relevant knowing the topics most frequently enquired by students. This could have various applications:

- It would allow detecting shortcomings in the information provided, for instance, lack of information regarding exam dates or the need for reinforcement in any given topic because the theoretical material provided has not been clear enough.
- Automatically organizing e-mails to improve student service.
- Automatically identifying core discussion topics in order to improve decision-making.

An e-mail has a date, a set of addresses, a subject, and a body. The latter, even though it may contain various types of information, consists basically of text and can therefore be analyzed by means of text mining techniques.
Text mining is a branch of Data Mining, and its main purpose is the extraction of high-quality information from documents.
It has numerous applications in various areas:

– In Biomedicine, it has been used to automate the identification and extraction of information from the numerous papers published each year [1].

– In Molecular Biology, it has been used to automatically extract information about genes, proteins and their functional relations from large collections of texts [2].

– In Education, it has been used to facilitate resource searches by combining the documents from various Web sites from related organizations [3].

– In the commercial context, it has been used to analyze the information generated by a consumer complaint Web site in order to obtain word relations that allow understanding the data [4],

– In the hospitality industry, using information available on Internet about hotels and possible tourists, it has been used to develop competitive strategies by analyzing demographic features and browsing habits [5].

All these works are representative of the diversity of areas in which text mining techniques are applied. However, regardless of the type of problem at hand, in most of the cases the main purpose is determining the relevance of the document based on a previous query. This allows more efficient automatic classification and access.

However, the extraction of information from e-mails is based on some special considerations, since, in general, the texts are short and their wording is quite abbreviated. Thus, some of the metrics used are no longer relevant, such as text length or the frequency of any given word within it.

The method proposed in this paper was applied to the e-mails of the Tutors Program (PACENI). This program is promoted by the Ministry of Education and its purpose is reducing the number of students that drop out from their university courses of study during their first year. This program was implemented at UNLP in the 2009 school year. Through it, first-year students are accompanied by tutors, post-graduate students or advanced students, who help them overcome the initial difficulties of university life.

For the processing stage, a dictionary built automatically from the reduction of each word to its root (stemming) [6] and its subsequent selection was used. By using this dictionary, each e-mail was represented as a numerical vector and was then used to train a SOM (self organizing map) neural network. From the weights of each neuron in the trained network, the most frequent combinations of terms can be identified. Finally, association-rule metrics are used to establish the relevance of each combination.

This paper is organized as follows: in Section 2, some related works are mentioned; in Section 3, SOM networks and their training mechanism are briefly described; in Section 4, the method proposed is detailed; in Section 5, the results obtained are presented; and in Section 6, conclusions are drawn and future lines of work presented.

2 Related work

Obtaining information from e-mails is a relevant task whose main purpose is classification and interpretation.
In this sense, the identification of spam e-mails is a generalized problem and has therefore received a lot of attention [7–11].
There are also approached that seek to automatically identify the author of the e-mail or the core subject of the message. For example, [12] tries to identify the person writing the e-mail from features based on number of words, number of lines, and the frequency of significant key words. [13] proposed a method that assesses the words from e-mails based on their age. The age of a word is calculated based on the frequency with which e-mails including it are received. The problem of this approach is the number of different words that can be used to refer to the same concept.
There is a current approach that has become popular with the appearance of various social networks in work environments. Nowadays, the development of collaborative tasks and the use of e-mails as communication mechanism are common. This creates the need of solving some participation-related issues, which implies identifying project members and their categories, as well as central work topics [14].
The general objective of this paper is related to this latter approach—we try to obtain information from a group of e-mails generated by teacher-student relations during a course carried out through a distance-education platform.

3 SOM (Self-Organizing Maps)

The SOM (Self Organizing Maps) neural network was defined by Kohonen in 1982 [16]. Its main application is the clustering of available information. Its ability to preserve input data topology makes it a visualization tool that is widely used in various areas.
It can be represented as a two-layer structure: the input layer, whose function is only to allow information to enter the network, and the competitive layer, which is responsible for the clustering task. The neurons that form this second layer are connected and have the ability of identifying the number of "hops" or connections that separate them from each of the remaining neurons in this level. Figure 1 shows the structure of a SOM network where the input layer is formed by a D-dimensional vector and the competitive layer has 9x7=63 neurons. Each neuron in this second layer has 8 direct neighbors (immediate connections). This connection pattern can change depending on the problem to solve. Each competitive neuron is associated to a weight vector represented by the values of the arches that reach this neuron from the input layer. These values, for all the neurons in this layer, are represented in the figure by means of the W matrix.

Network weights, W values, are initially random, but they adapt with the successive presentations of input vectors.

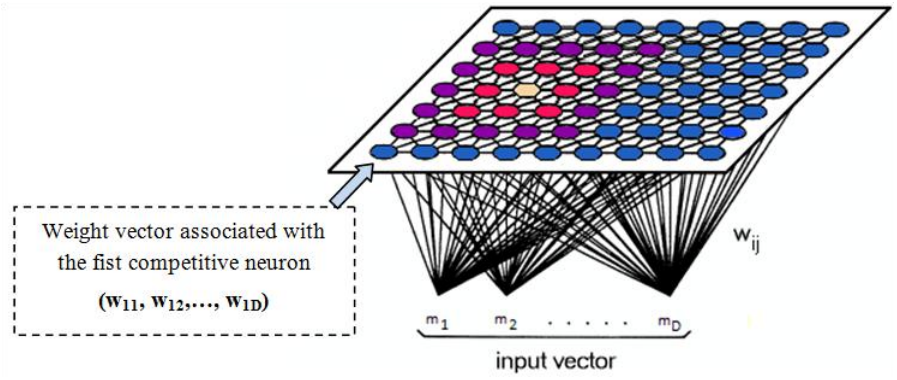

Fig. 1. Classic structure of a SOM network

Since this is a competitive structure, each input vector is considered to be represented by (or associated with) the competitive neuron that has the most similar weight vector based on a given similarity measure.
The final value of W is obtained by means of an iterative process that is repeated until the weight vectors do not present any significant changes or, in other words, until each input vector is represented by the same competitive neuron than in the previous iteration.
In each iteration, the neuron representing each input vector is determined. This neuron is called "winning neuron", since it is the one that "wins" the competition to represent the vector (is the most similar one so far). Then, the weight vector for that neuron and its neighborhood are updated following equation (1).

$$w_{ij} = w_{ij} + \alpha * (x_i - w_{ij}) \qquad i = 1..n \qquad (1)$$

where j is the competitive neuron whose vector is being updated and α is a value between 0 and 1 that represents a learning factor.
Equation (1) has variations that can be consulted in [15].
The concept of neighborhood is used to allow the network to adapt correctly. This implies that neighboring competitive neurons represent similar input patterns. For this reason, during the training process (obtaining W values) is started with a wide neighborhood that is then reduced as iterations occur.

Figure 2 shows the pseudo-code corresponding to the basic process for the adaptation of the SOM network.

4 Proposed method

To be able to operate with the network described above, e-mails have to be represented by numerical vectors. To this end, a dictionary of terms will be

```
W ← Random initial values.
Neighborhood ← set the size if the initial neighborhood
NoIteReduction ← set the number of iterations that must
                        occur to reduce the neighborhood
while termination criterion is not reached do
  for all each input vector do
    Input the vector to the network and calculate the winning neuron
    Update the winning neuron and its neighborhood
  end for
  Reduce the neighborhood {if applicable based on NoIteReduction }
end while
```

Fig. 2. Basic training pseudo-code for the SOM network

built by processing an only text formed by the concatenation of the subject and body of each e-mail. Each word in the text is reduced to its root by applying a stemming algorithm [17]. This process is important for processing text in Spanish due to the syntactic changes related to gender, number, and tense. For example, words such as 'trabajo' (work), 'trabajar' (to work), 'trabaja' (he/she works), 'trabajos' (the works), 'trabajoso'(laborious) are reduced to the common root 'trabaj' by applying the stemming algorithm.
Once the root of each word is obtained, its frequency of use in the entire text and its average length are calculated. By means of statistical analysis processes, terms that are less relevant are discarded; the dictionary to represent e-mails is then built with the remaining terms.
Then, each e-mail is represented by a fixed-length binary vector. The number of elements in the vector is determined by the number of words in the dictionary. Each position will have a value of 1 if the word appears in the e-mail or a value of 0 if it does not.
Be D the number of words in the dictionary and M the number of available e-mails, each e-mail will be represented as follows:

$$mail_i = [m_{i1}, m_{i2}, \ldots, m_{iD}] \qquad i = 1..M \tag{2}$$

$$m_{ij} = \begin{cases} 1 \; the \; word \; j \; is \; in \; e-mail \; i \\ 0 \; otherwise \end{cases} \qquad j = 1..D \tag{3}$$

Using the vectors defined in (2), the SOM network is trained by applying the algorithm shown in Figure 2.
Be N the number of competitive neurons that form the SOM network, W_k will be the weight vector of the k^{th} competitive neuron. When the training stage is complete, the weights of these neurons will have the following format

$$W_k = [w_{k1}, \ldots, w_{kD}] \qquad k = 1..N \tag{4}$$

where

$$no\_word_k = \{s \in 1..M / ||W_k - mail_s|| < ||W_j - mail_s||, \forall j, j \neq k\} \;\; k = 1..N \tag{5}$$

$$w_{kj} = \frac{\sum_{s \in no.\_word_k} m_{sj}}{\#no.\_word_k} \qquad j = 1..D \tag{6}$$

Therefore, if the subset of e-mails represented by the same competitive neuron includes the same word, the vector that is associated with that neuron will also have a value of 1 in the position corresponding to that word.
In other words, the positions that have high values (close to 1) in the vector associated to a competitive neuron represent words that appear repeatedly in the emails that have this neuron as the winning one.
The method proposed in this paper uses the self-organizing maps to achieve two objectives: in the first place, to discard the words that are less significant, and secondly, to determine the most relevant word associations. Both these tasks are of interest, since the former helps not having to make an *a priori* decision regarding the size of the dictionary, and the latter is the solution to the problem presented.
The less significant words will be those words whose own weight is not enough to be clearly represented by a limited subset of neurons. This is the case of words that are combined with many terms or that are infrequently used. In either case, these are terms that provide little information, since in the first case they do not determine the subject matter and in the second case are not sufficiently supported (number of occurrences) to be considered significant. The trained SOM network is able to detect these words because they do not go beyond a minimum threshold in any of the vectors associated with the competitive neurons (Equation 7). Therefore, the vectors of W (Equation 4) are converted to binary values by using this threshold.

$$Wbin_k = [wbin_{k1}, wbin_{k2}, \ldots, wbin_{kD}] \qquad k = 1..N \tag{7}$$

where

$$wbin_{kj} = \begin{cases} 1 \; if \; w_{kj} > threshold \\ 0 \; otherwise \end{cases} \qquad j = 1..D \tag{8}$$

and irrelevant words are obtained with equation (9).

$$IrrelevantWords = \{Word_j, j \in D \mid wbin_{kj} = 0, \forall k = 1..N\} \tag{9}$$

It should be mentioned that each element of $Wbin_k$ will have a value of 1 at the positions corresponding to relevant terms. This allows identifying the frequent terms in the e-mails represented by the kth neuron of the network, which can be used to form various association rules.
An association rule is an expression with the following format

IF (antecedent) THEN (consequence)

where both the antecedent and the consequence are logical expressions referring to the words present in the e-mail.
The following are examples of association rules:

– IF ('board' ∧ 'exten' ∧ 'certific') THEN ('transcr' ∧ 'present' ∧ 'academ' ∧ 'approve')
– IF ('pending') THEN ('academic subject' ∧ 'certificate')

The first rule indicates that each time an e-mail has the words 'board', 'exten' and 'certific', the words 'transcr', 'present', 'academ' and 'approve' are also present. This rule refers to the approval by the academic board of extensions to present school transcripts. The second rule shows the relationship between the word 'pending' and the words 'academic subject' and 'certificate'. It also refers to pending academic subject certificates.
Rules are formed by combining in all possible ways the terms that appear in any given weight vector defined as in equation (7).
There are various metrics that can be used to determine the importance of a rule. The most common ones are:

– Support: It is the proportion of examples (e-mails) that fulfill the rule. For example, if the words 'pending', 'academic subject' and 'certificate' are present in 300 e-mails from a total of 3,000 e-mails, the support of the rule

  IF ('pending') THEN ('academic subject' ∧ 'certificate')

  will be $300/3000 = 0.1$.
– Confidence: it is the quotient of the number of examples that fulfill the rule and the number of those that only fulfill the antecedent. Let us consider again the rule IF ('pending') THEN ('academic subject' ∧ 'certificate') verified by 300 of the 3,000 available e-mails. Let us assume that after revising the available e-los, it is observed that 350 of those contain the word 'pending,'; the confidence of this rule will be $300/350 = 0.85$

The importance of the rules obtained in this paper depends on the product of the two previously mentioned metrics. Therefore, the result that can be obtained is the interpretation of the most relevant rules.

5 Results

The method described in Section 4 was applied to the 2,995 e-mails from the Tutors Program (PACENI) between April and November 2009.
The initial dictionary was formed by 2,935 term roots; 287 of these were selected by statistical analysis. The selection criterion used had three stages:

i) First, words of atypical lengths were suppressed, considering as such all words whose value was more than 1.5 times the distance between the first and the third quartile (fourth dispersion). In the case of the PACENI e-mails, words that were longer than 18 characters or shorter than 3.5 characters were discarded. The average length was assessed based on all words that corresponded to the same root term.

ii) When analyzing the plot box corresponding to the occurrence frequency for each root term, it was observed that a large part of the population had a low value. That is, the most commonly used terms were the minority. Therefore, we decided to use those terms with extreme frequency. For the measured population, these were the terms with more than 49 occurrences.

iii) Finally, in the plot box of the reduced population, extreme values corresponding to the terms that are very frequently used in all e-mails are still observed. This reduces their importance. For this reason, those terms whose frequency was higher than 613—extreme value—were removed.

Figure 3 shows the plot box diagrams mentioned in ii) and iii).

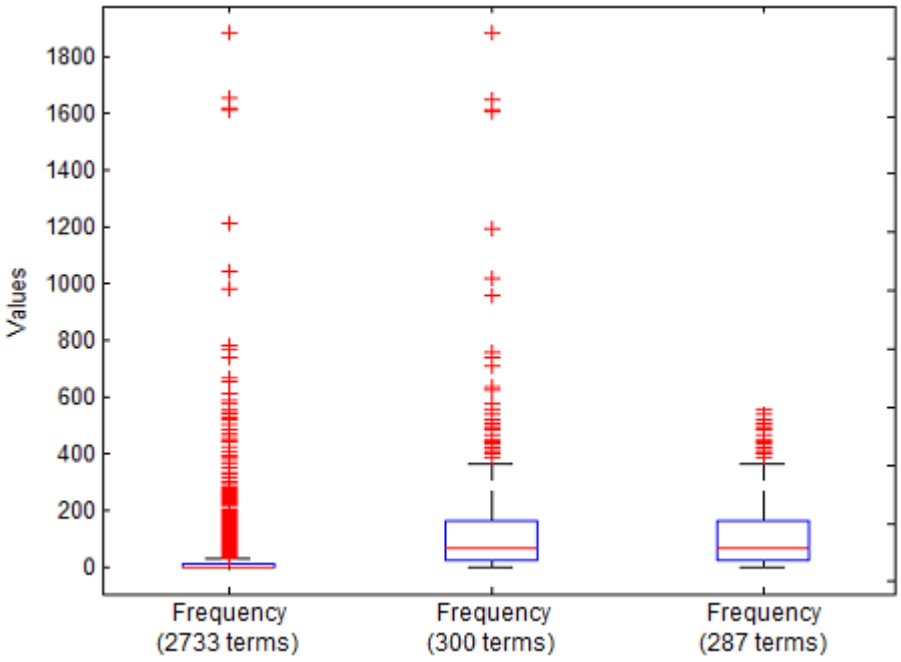

Fig. 3. Plot box diagrams corresponding to the successive reductions of dictionary terms as detailed in 5.ii) and 5.iii)

A SOM network with 13x13 competitive neurons with 4 neighbors per neuron was used. The initial size of the neighborhood was set as a third of the number of rows in the network, that is, 4 neurons. This value is high, since it is the radius (number of "hop") that determines the area around the winning neuron where weight vectors are modified. The reduction was carried out every 30 iterations, with a maximum of 180 iterations. This value ensures that successive reductions will be carried out until the adaptation only affects the winning neuron.
After training the network, all weights that were not significant were removed

from the matrix W; to this end, a threshold of 0.85 was used.
With the weight vectors of each competitive neuron, the combinations of terms that allow clustering the e-mails were determined.
To measure their relevance, they were used to form the corresponding association rules, considering all possible combinations. The combination was associated with the maximum value obtained by multiplying its support and confidence values.
After 50 independent training sessions of the neural network, the most commonly occurring terms are the following:

('transcript', 'academic', 'approved', 'board', 'extend', 'present', 'certific')
('beta', 'classroom', 'inscription', 'English)'
('included', 'scholarship', 'ministry', 'ICTs', 'find', 'http', 'inscription')
('alumns', 'inscription', 'Guaraní', 'segundo')
('situation', 'know', 'tutor', 'question', 'contact')

These combinations appear in various orders, but always within the 20 first best positioned ones. This determines their importance within the set of e-mails.
Another characteristic that was observed after the various tests is that the neural network allows discarding between 100 and 120 terms by means of the threshold function indicated in equation (8). This reduces considerably the time required to obtain the association rules to be measured.

6 Conclusions and future lines of work

An e-mail analysis mechanism based on data mining techniques has been presented. Even though the results obtained only refer to the 2009 Tutors Program of the PACENI, this analysis can be applied to other courses with no considerable changes.
Building the initial dictionary is essential to obtain good combinations of terms. The proposal presented in this paper included a statistical pre-processing so as to generate the dictionary as automatically as possible. This stage can be improved by manually entering additional information.
We are currently working with a dynamic SOM network so that adaptability is not limited. With this modification, we expect to solve the problem of neuron saturation. This is observed only in 0.2% of network neurons, but it may lead to the analysis of terms that are discarded with the current architecture.

References

1. Sophia Ananiadou, Douglas B. Kell, Jun-ichi Tsujii. Text mining and its potential applications in systems biology. Trends in Biotechnology. Elsevier Science London. Volume 24, Number 12, Pages 571-579. 2006.

2. Martin Krallinger, Alfonso Valencia. Text-mining and information-retrieval services for molecular biology. Genome Biology. BioMed Central. Volume 6, Number 7, Article 224. 2005.
3. Sophia Ananiadou, Paul Thompson, James Thomas, Tingting Mu, Sandy Oliver, Mark Rickinson, Yutaka Sasaki, Davy Weissenbacher, John McNaught. Supporting the education evidence portal via text mining. Philosophical Transactions of The Royal Society A. 368(1925): 3829-3844, 2010.
4. Kuan C. Chen. Text Mining e-Complaints Data From e-Auction Store With Implications For Internet Marketing Research. Journal of Business and Economics Research. The Clute Institute for Academic Research. Volume 7, Number 5, Pages 15-24. 2009.
5. Kin-Nam Lau, Kam-Hon Lee, Ying Ho. Text Mining for the Hotel Industry. Cornell Hotel and Restaurant Administration Quarterly. Cornell Hospitality Quarterly. Volume 46, Number 3, Pages, 344-362. 2005.
6. Carlos G. Figuerola, Raquel Gómez, Angel F. Zazo Rodríguez, José Luis Alonso Berrocal. Spanish Monolingual Track: The Impact of Stemming on Retrieval. Evaluation of Cross-Language Information Retrieval Systems. Springer Berlin/Heidelberg. Volume 2406, Pages 253-261. 2002.
7. Carlos M. Lorenzetti, Rocío L. Cecchini, Ana G. Maguitman, András A. Benczúr. Métodos para la Selección y el Ajuste de Características en el Problema de la Detección de Spam. XII Workshop de Investigadores en Ciencias de la Computación, Área Agentes y Sistemas Inteligentes. 2010.
8. Mehrnoush Famil Saeedian, Hamid Beigy. Dynamic classifier selection using clustering for spam detection. IEEE Symposium on Computational Intelligence and Data Mining. 2009.
9. Tsan-Ying Yu, Wei-Chih Hsu. E-mail Spam Filtering Using Support Vector Machines with Selection of Kernel Function Parameters. Fourth International Conference on Innovative Computing, Information and Control. 2009.
10. Wanli Ma, Dat Tran, Dharmendra Sharma. A Novel Spam Email Detection System Based on Negative Selection. Fourth International Conference on Computer Sciences and Convergence Information Technology. 2009.
11. Xiao Li, Junyong Luo, Meijuan Yin. E-Mail Filtering Based on Analysis of Structural Features and Text Classification. Second International Workshop on Intelligent Systems and Applications. 2010.
12. Olivier de Vel. Mining E-mail Authorship. In Proceedings of KDD 2000 Workshop on Text Mining. 2000.
13. Jason D. M. Rennie. ifile: An Application of Machine Learning to E-Mail Filtering. In Proceedings of KDD 2000 Workshop on Text Mining. 2000.
14. Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz, Anand Swaminathan. Mining Email Social Networks. In Proceedings of ICSE 2006 Workshop on Mining Software Repositories. 2006.
15. Teuvo Kohonen. Self-organizing Maps. 2nd Edition. Springer. 1997. ISBN 3-540-62017-6.
16. Teuvo Kohonen. Self-organized formation of topologically correct feature maps. Biological Cybernetics. Springer Berlin/Heidelberg. Volume 43, Number 1, Pages 59-69. 1982.
17. Dennis D. Perez Barrenechea. A Spanish Stemming Algorithm Implementation in PROLOG and C#. Accessed at www.ai.uga.edu/mc/pronto/perez.pdf. 2006.