Corporate Communication Network and Stock Price Movements Insights from Data Mining

SRI. CH. RATNA BABU[1], M. NAVEEN KUMAR[2], V. SRAVANI[3]
[1] *Associate Professor, Department of Computer Science and Engineering, RVR&JC College of Engineering, India*
[2, 3] *UG Student, Department of Computer Science and Engineering, RVR&JC College of Engineering, India*

Abstract- To detect communication patterns with in a company and find out how these patterns are related to performance of a company. In particular, the study concentrates on whether or not there exists any association relationships between the frequency of e-mail exchange of key employees in company and performance of company that reflects in stock prices. If there are such relationships, then they find out whether or not the company's stock price could be accurately predicted based on detected relationships.

Data Mining algorithm is proposed to mine the e-mail communication record and historical stock prices to detect the association relationships and based on the detected relationships, rules that can predict changes in stock prices can be constructed.

I. INTRODUCTION

Recent research the existence of interesting communication patterns among different participants of different social network platforms. These patterns have been shown to be useful in predicting product sales and stock prices. Compared to social network, which can be considered as representing connections among people in the public, a corporate network can be express opinions on any issues of interest, members of a corporate communication network are expected to mainly talk about company-specific business. In a company, employee communications can mean the success or failure of any major change program resulting from a merger, acquisition, new venture, new process improvement approach, or other management issues. In other words, employee communication can serve a critical "business function that drives performance and contributes to a company's financial success. Therefore, prediction of movements of stock price is the main focus of this paper.

Unlike social networks, in a corporate communication network, e-mails have long been used as a tool for interorganizational and information exchange. In the same way, a social network platform is able to capture participants' behaviour and their opinions about various issues and events. Thus, we argue that a corporate communication network in the form of an ecosystem also contains insightful information, such as organizational stability and robustness, about a company's development. We believe our argument is in line with corporate communication theory, which suggests that "employee communications can mean the success or failure of any major change program" resulting from a merger,acquisition,new venture, new process improvement approach, or other management issues.

In this paper, We propose that a company's performance, in terms of its stock price movements, can be predicted by internal communication patterns. To obtain early warnings signals, we believe that it is important for patterns in corporate communication networks to be detected earlier for the prediction of significant stock price movement to avoid possible adversities that a company may face in the stock market so that stakeholders' interests can be protected as much as possible. Despite the potential importance of such knowledge about corporate communication, little work has been done in this important direction.

II. PREVIOUS METHODS

- Support vector Machines:

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However,

it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes.

- Decision Trees:

The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data). As one of the traditional classifiers, a decision tree is a form of multiple variable analysis for classification and prediction. It has been applied in various fields including stock market prediction and stock trading rule generation

III. PROPOSED METHOD

- Modules to be implemented

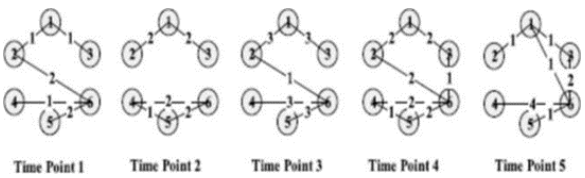
The modules to be implemented for prediction of movement in stock prices are:

- discretizing the weight matrices and the value of movement.
- discovering patterns for describing the relationship between communication frequency and stock price
- constructing predication rules based on patterns;
- predicting the movement of stock price using prediction rules.

- Module description

Notations and Definitions

Definition 1 (communication networks):Let G = {*G*1, *G*2 . . . *Gt* , . . . *Gp*} represents the communication networks(*p* is the total time points). The one of networks is represented as a graph with a 2-element tuple *Gt = (Vt ,Rt)*, with the following characteristics.



Vt= {*v*1, *v*2, . . . *vn*} represents a finite set of nodes (employees in the communication network), *n* is the total number of nodes in the network, and *vc* represents the central node in the network. In the Enron case, we choose the CEO as the central node and examined the communication network between the CEO with other employees.

Rt= $\{v^{i}_{j,t}\}$ denotes the set of edges representing all the links in the network. If there is an edge $R_t^{ij} \in R_t$, it means that the two vertices $v_i$and v_j are connected through the link of R_t at t^{th} time point.
1, $\tau \in Z+$ and $t = 1, 2, \ldots p$.

Let $S = \{s1, s2 \ldots, st-\tau \ldots st, \ldots sp\}$ represents a time series ofdata values collected over a period of time and *p* is the totaltime points. *St* represents the value of movement of stock price in the *t*th time point, where $st \in [Lt, Ut]$; *Lt* represents the lower bound and *Ut* represents the upper bound of the values,
$1 \leq \tau \leq t - 1$, $\tau \in Z+$ and $t = 1, 2, \ldots p$.

In order to better visualize the data we used, a synthetic data set shown in Fig. 3 is used to describe the format of input graph stream. Six nodes in the example represent six persons in the communication network, respectively. The value of the edge between two nodes represents communication frequency between two persons. The whole input data describes the changes of communication network during five time points. In this paper, the proposed algorithm can discover the patterns for describing the relationship between communication frequency network and the change of stock price. For testing the performance of the proposed algorithm, we predict the change of stock price of Enron using its e-mail corpus data set.

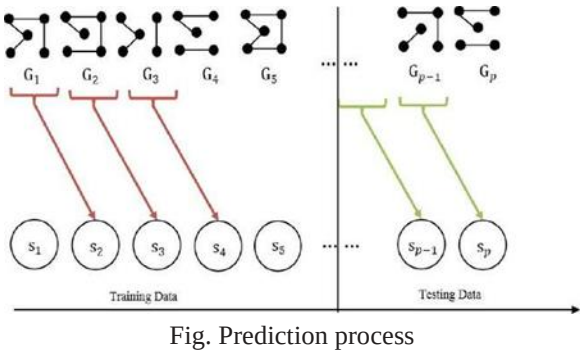


Fig. Prediction process

To better analyze the graphical data and link

information of *Gt*, we use a *n*×*n* matrix **D** = [*di j*](1 ≤ *i, j*≤ *n)* to represent the graph and the value of *di j* can capture the communication frequency that occurred between two nodes, as shown in the following:

dij ={ eij ,eij € E
0 ,other
}

By using the same synthetic data set, the graph set shown in Fig. 3 is transformed into a set of matrices as Fig. 5 demonstrated. In this case, we capture five matrices as shown in Fig. 5 to represent the communication network for all five time points.

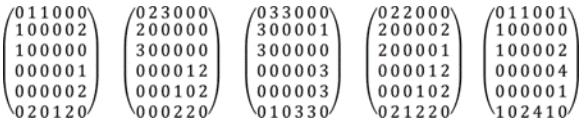


Fig. Weighted matrices for every time point

- Proposed Algorithm

Discretizing Communication Matrix and Stock Price: The value of edge (*dij)* between two different nodes for representing communication frequency between two nodes is numerous. In order to reduce and simplify the original data, numerous values are always replaced by a small number of interval label, which leads to a concise easy-to-use, knowledge-level representation of mining results. The existing and mature methods on unsupervised discretization are primarily equal frequency discretization and equal width discretization. The equal width method is typically used in every statistic program to produce regular histograms. However, equal width discretization can hardly handle the situation if outliers exist in the data set. Equal frequency can overcome the limitations of the equal width discretization by dividing the domain in intervals with the same distribution of data points. Hence, considering different people may have different habits to exchange e-mails with each other,we use equal frequency, the most represented algorithm, to discretize other nonzero value for the value of weight in the communication network.

Besides the differences between people, we also consider the differences of networks in different time points. Hence, we use GS*t* represent the total number of communications occurrences, denoted as $\sum^{n}j=1 \sum^{n}$ which can measure the overall graph sparsity in the *t*th time point.

Let a set of event-based values, denoted as *Eij*= {*Eij,1, Eij,2, . . .Eij,t . . ., Eij,p*}, *Eij,t€Eij* correspond to *vij,t€Rt*, where the domain of the value of *Ei*is denoted as dom(*Eij)*= [*Lij,Uij*], in which *Lij* represents the lower bound and *Uij* represents the upper bound of the values. Hence, the set of discrete states for *Eij,t*is denoted as *D(Eij,t)* = {dr1, dr2 . . . dr*k* . . . dr*n*, where dr*k* represents the discrete states for the value of *(*R*t /*GS*t), n* is the total number of states.

In terms of the Enron case, a secretary needs to send a great number of e-mails for coordination in course of the daily work. Therefore, the mean level, for instance 10 e-mails per day, is considered as the normal number for a secretary according to his/her communication pattern. In a relative sense, 10 e-mails per day might be already well above the mean level of the communication frequency for a front line employee. We standardize the communication level according to each person/node's mean of his/her communication frequency, and all continuous data are transformed into discrete data, such as dr1, dr2, dr3. Three levels are applied for Enron case which represent "no communication," "weak relationship," and "strong relationship".

Similarly, the value of stock price is discretized. First, the original value of "ups and downs" is represented using the price fluctuation ratio, which can be calculated by *(*CP*t*−1 − CP*t* /CP*t*−1*)* where CP*t*−1 represents "closing price" in the day of "t−1" and CP*t* represents "closing price" in the day of "t." Then the movement of stock price are classified into different states, such as ds1, ds2,which, respectively, represents "the ratio of increases of stock price is higher than 0%" and "the ratio of decreases of stock price is lower than 0%."

In summary, the original matrix is discretized as *Rt*={*Eij,t*|*Eij,t €D(Lij,t)*, *D(Lij,t)* = {dr*i*}, 1 ≤ *i, j* ≤ *n*} and the value of movement of stock price is discretized as *st*= *Et* ,where *Et €D(Lt)* and *D(Lt)* = {ds*i*}.

2-D TABLE WITH J ROWS AND I COLUMNS

o_{tk} (e_{tk})		***Stock Price*** (in the certain position later)					**Total**
		ds_1	...	ds_i	...	ds_I	
$E^c_{j,t-1}$	dr_1	o_{11} (e_{11})	...	o_{i1} (e_{i1})	...	o_{I1} (e_{I1})	o_{+1}
	. . .	. . .		. . .		. . .	. . .
	dr_j	o_{1j} (e_{1j})	...	o_{ij} (e_{ij})	...	o_{Ij} (e_{Ij})	o_{+j}
	. . .	. . .		. . .		. . .	. . .
	dr_J	o_{1J} (e_{1J})	...	o_{iJ} (e_{iJ})	...	o_{IJ} (e_{IJ})	o_{+J}
Total		o_{1+}	...	o_{i+}	...	o_{I+}	M'

FREQUENCY TABLE FOR $E^c_{j,t-1}$ AND STOCK PRICE

o_{ij} (e_{ij})		***Stock***			Total
		Up (ds_1)	Middle (ds_2)	Down (ds_3)	
$E^c_{j,t-1}$	None (dr_1)	0 (0)	0 (0)	0 (0)	0
	Strong (dr_2)	0 (0)	1 (0.5)	0 (0.5)	1
	Weak (dr_3)	0 (0)	1 (1.5)	2 (1.5)	3
Total		0	2	2	4

WEIGHT VALUES OF EVIDENCE FOR $E^c_{j,t-1}$ AND E_t

W_{ij}		***Stock***		
		Up	**Middle**	**Down**
$E^c_{j,t-1}$	**None**	N/A	N/A	N/A
	Strong	N/A	N/A	N/A
	Weak	N/A	-1	1

In a simple example, *zi j* (E^cj,*t*−1=strong→stock= middle*)* can be calculated as *zij*= 1 − 0.5*/*(0.5)1/2 =0.71 and statistical significances then tested based on the use of the following residual measure to determine the graph–stock pattern according to (3).*dij* = 0.71*/*((1 − 1/4)(1 − 2/4))1/2 = 1.154. And Table III shows all values of *dlk* to measure the correlations between $E^1$2,*t*−1 and *Et*. N/A in Table III represents nonexistence when *olk* is zero for counting thenumber of relationship takes on corresponded value.

The weight of evidence for or against a certain prediction of the attribute values of future objects by using the same information measure can be assessed quantitatively as follows.

Suppose that *vil* of stock at one time point later is dependent on *v jk* in the records stream. The weight of evidence measures the amount of positive or negative evidence that is provided by the value of relationship between central node and destination node supporting or refusing the stock price being observed together. Hence,*v jk* that provides positive evidence supporting stock at one time point later in the graph stream.

The value*vil* and *v jk* is the characteristics of weight for *R(Nc, NJ)* in the preceding time point. Otherwise, *v jk* provides negative evidence.

IV. IMPLEMENTATION

Input: D={$R_1, R_2, \ldots R_p$} (A set of matrices to represents communication network)
S = {$S_1, S_2, \ldots S_p$} (A set of stock prices),
v_c (central node)
sl (the number of states of stock prices)
rl (the number of states of communication frequency)
Output: *Sub-Res (List of Adjusted Residuals Value)*
Res (Itegrated all *Sub-Res*)
Discretize S into $E_t \subseteq (ds_1, ds_2, \ldots ds_i, \ldots ds_{sl})$
for *r=1: n, r≠c* (n is the total number of nodes)
Discretize $E^c_{j,t-1}$ *into* $D_r \subseteq (dr_1, dr_2, \ldots dr_j, \ldots dr_{rl})$
for i =1:*sl*
for j = 1:*rl*
calculate $d_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{\left(1-\frac{\sum_{u=1}^{I} o_{iu}}{M'}\right)\left(1-\frac{\sum_{u=1}^{I} o_{uj}}{M'}\right)} * e_{ij}}$
Add d_{ij} *into Sub-Res*
end
end
Add Result into Sub-Res
end

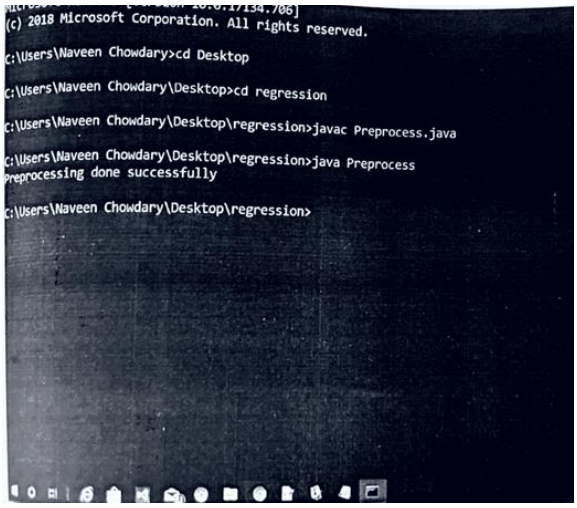
V. RESULTS

- Actual Results of the Work.

Firstly, the dataset containing tweets is pre-processed. Firstly, we test whether the dataset is being loaded or not and if is taken in the write format or not. If wrong format of dataset is given is given as input that the system is expected to return an error.

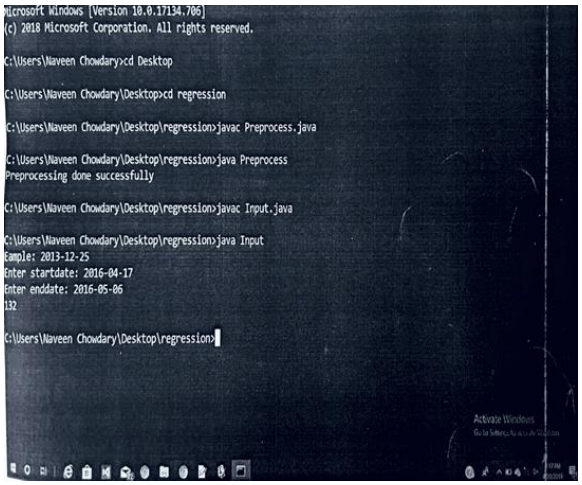
And here it is checked whether the system returns the error or not. If error is returned that the test case is passed else failed. If right file format is given, the system proceeds that the next module without any error. The probability of stock price moving upward or downward is obtained if done correctly, test case is passed. If there are no correct number of frequencies for calculating probabilities then the system should display no output to the user.

Next the input is entered in between the specified dates. Now the values entered are tested. The values of dates are entered as in the specified range. If the values entered tested. The values of dates are entered as in the specified range. If the values entered are out of the range then the specified range. If the values entered out of the range then the system should return an error else proceed to next step. The probability of stock price moving upward or downward is obtained if done correctly, test case is passed. If there are no correct number of frequencies for calculating probabilities then the system should display no output to the user. Input is entered in between the specified dates. Now the values entered are tested. The values of dates are entered as in the specified range.

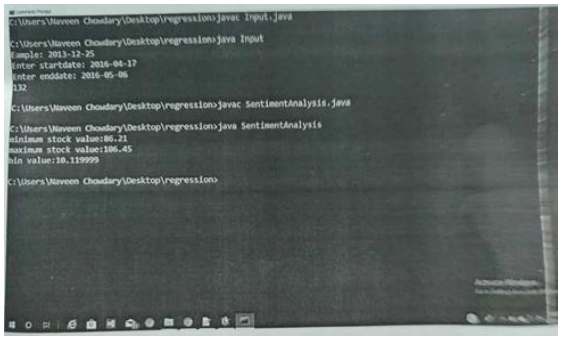


Screenshot of Importing the dataset

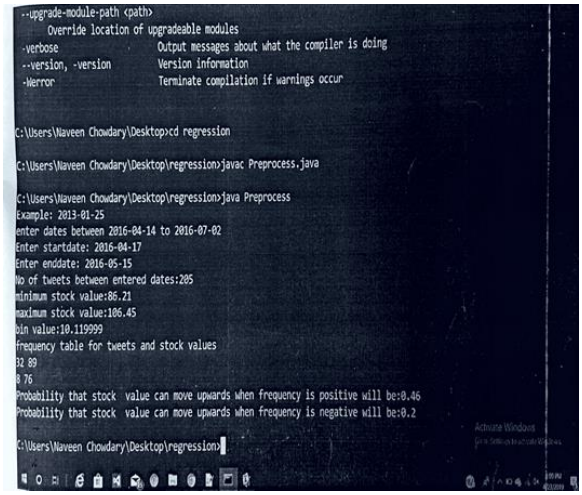
Next the input is entered in between the specified dates. Now values entered are tested. The values of dates are entered as in the specified range. If the values entered are out of range



Screenshot of entering the values



Screenshot of analysis



Screenshot of prediction

- Analysis of Results Obtained:

We analyze the complexity of proposed algorithm. Suppose that we have N nodes in a community network with the p total time points s represents states of stock price and r states of communication

frequency. After preprocessing and discretizing the stock price and communication between central nodes, the proposed algorithm needs to iterate in order to calculate a the adjusted residual to measure the correlations between communication frequency and stock prices.

CONCLUSION

The findings and theoretical implications from this paper are twofold. On one hand, we captured the communications among nodes in Enron's major corporate communication network and identified employees' communication patterns. This paper demonstrates that a corporate e-mail ecosystem contains meaningful information about employees' communication patterns. Even if we only focus on the communication frequency, a company (Enron in our case) has identifiable patterns of e-mail exchange. Such identifiable patterns can reveal important information about major corporate activities and organizational stability that may subsequently influence the focal company's performance in the stock market. Therefore, cooperate communication patterns can serve as a good proxy to predict a company's stock performance. Our experimental results demonstrated the existence of dependence between e-mail communication network and stock price for Enron.

This paper extended the existing communication theories to capture the patterns of corporate communication and the focal company's stock performance.

On the other hand, social networks have become a hot topic in the field of data mining recently. In this paper, we not only provided an innovative idea on using data-mining algorithms but also constructed the relationship between social network and finance. Hence, this paper demonstrated great potential to predict the amounts of increases and decreases of stock price based on the weighted rules.

ACKNOWLEDGMENT

We have taken a lot of effort into this project. However, completing this project would not have been possible without the support and guidance of our faculty of Computer Science and Engineering department at RVR&JC College Of Engineering and our fellow students. We would like to extend our sincere thanks to all of them.

We are highly indebted to Sri. Ch. Ratna Babu(Ass.Professor) for her guidance and supervision. We would like to thank her for providing the necessary information and resources for this project.

We would like to express our gratitude towards our parents & our friends for their kind co-operation and encouragement which help us a lot in completing this project. Our thanks and appreciations also go to our colleague in developing the project. Thank you to all the people who have willingly helped us out with their abilities.

REFERENCES

[1] B. Collingsworth, R. Menezes, and P. Martins, "Assessing organizational stability via network analysis," in Proc. IEEE Symp. Comput. Intell. Financial Eng., Mar. 2009, pp. 43–50.
[2] D. J. Barrett, "Change communication: Using strategic employee communication to facilitate major change," Corporate Commun., Int. J., vol. 7, no. 4, pp. 219–231, 2002.
[3] M.-C. Wu, S.-Y. Lin, and C.-H. Lin, "An effective application of decision tree to stock trading," Expert Syst. Appl., vol. 131, no. 2, pp. 270–274, 2006.
[4] C. W. Down, G. C. Phillip, and A. L. Pfeiffer, "Communication and organizational outcomes," in Handbook of Organizational Communication, G. Goldhaber and G. Barnett, Eds. Norwood, NJ, USA: Ablex, 1988.
[5] P. G. Clampitt and C. W. Downs, "Employee perceptions of the relationship between communication and productivity: A field study," J. Bus. Commun., vol. 30, no. 1, pp. 5– 28, 1993.
[6] G. S. Hansen and B. Wernerfelt, "Determinants of firm performance: The relative importance of economic and organizational factors," Strategic Manage. J., vol. 10, no. 5, pp. 399–411, 1989.
[7] W. W. Burke and G. H. Litwin, "A causal model of organizational performance and change," J. Manage., vol. 18, no. 3, pp. 523–545, 1992.

[8] R. Katz, “The effects of group longevity on project communication and performance,” Admin. Sci. Quarterly, vol. 27, no. 1, pp. 81–104, 1982.
[9] V. A. Zeithaml, L. L. Berry, and A. Parasuraman, “Communication and control processes in the delivery of service quality,” J. Marketing, vol. 52, no. 2, pp. 35–48, 1988.
[10] C. Hargreaves and Y. Hao, “Prediction of stock performance using analytical techniques,” J. Emerg. Technol. Web Intell., vol. 5, no. 2, pp. 136–142, 2013.
[11] R. B. Higgins and B. D. Bannister, “How corporate communication of strategy affects share price,” Long Range Planning, vol. 25, no. 3, pp. 27–35, 1992.