A Method for Constructing Knowledgebase from Data in Emails using Clustering

Chellammal Surianarayanan
Ph.D. Scholar, School of Computer Science and Engineering,Bharathidasan University, Tiruchirappalli

Gopinath Ganapathy
School of Computer Science and Engineering
Bharathidasan University, Tiruchirappalli

ABSTRACT

Email is one of the most important mode of communication among various members involved in development of software. More specifically the knowledge in emails is very useful to developers who will be involved in the development of different modules as well as fixing bugs in various modules. As software developers tend to switch over different corporate, providing knowledge to new developers is really tough. As email is the chief medium of interaction among members of project, all important information will naturally be available in emails but in an unorganized form. In this work, an approach is proposed to construct knowledgebase using the contents of emails. It is proposed to use K-Means clustering to categorize the information. Here the value K will be number of different projects handled by the concern. The paper describes the need for knowledgebase from emails, related work, proposed methods and its benefits. At present efforts are towards implementing the method.

Index Terms

knowledgebase from emails, clustering of email data, email categorization

I. INTRODUCTION

In general each company will have one or more specialized domains such as banking and financial, tourism, education, entertainment, telecom, etc., and it provides services related its specialized domains to their clients. The company acquires projects from their clients in their specialized domains. At a given time a company will be dealing with more than one project. The management of the company will split the projects into different groups of employees based on the size of the project, availability of human resources, time required for development and release, skill set required to implement the project, etc. Each project will be assigned to a group of members. The members of a same project will generally will be scattered over different branches of the company. For example, in a project which is handled by say a group of 60 members, then out of 60 members 20 members may be working from Chennai, 10 members may be working from Mumbai, 15 may be working from Delhi and remaining 15 members may be working from USA. Besides their geographical separation, they are working in a collaborative manner for the same project with unified goal.

The global network infrastructure, Internet plays as the major backbone for business applications and their development. Nowadays, Internet services especially e-mail is used the most common method of communication in companies. The people who are working in the same project will be communicating to their group members though e-mail. Recent studies show that e-mails can make up to 75% of the company's communications and every employee spends around 90 minutes a day in organizing e-mail tasks. More specifically the developers will discuss about their assigned work with their technical managers and other group members through e-mail. One of the major issues faced by software industries is that the people hired for a particular project may not remain with the same company till end of the project. Further, some people may remain in the company but depending on the requirements of other projects and skill set of concerned people they be assigned with other projects. In such situations it is desirable to have knowledge base which gives details about the projects, various modules, databases, etc. This knowledgebase will serve as a valuable asset to the developers. At any instant, a developer (especially a new developer to the project) can find the required information related to his project from the knowledge base.

In software development communication through e-mail is the dominant form of interaction among various members of a project, developers, analysts, clients, technical managers, etc. Interactions through e-mail are essential especially to developers involved in a project. This is illustrated with an example. Let *A* be a newly joined programmer in a software company who is hired by the company for development and fixing bugs. Let *B* be the client manager of *A* located in USA. *B* assigns bugs to be fixed to *A* through e-mail. Let *A* be responsible for fixing the bugs related to say 'SNMP module' (in a router configuration). In such situation as a new programmer A needs to know more about the project as well as about SNMP module and its interaction with other modules of the project. He can interact with his client and other collaborating team members through emails. The members involved in a project (or even module) are located at different geographical locations. Most of the interactions among the members are through email. Besides this, a more frequent issue is that programmers who have been hired and trained for a particular project may resign his job or he may be assigned with some other projects where his expertise is very much needed. In such situations a content repository which contains information about the project and its various modules if available, will assist the concerned programmers well in fixing the bugs or developing further modules. From this, it is understood that the data present in e-mail becomes very important and such data should be organized in a repository to assist the developers involved in a project. Therefore it is proposed to construct a knowledgebase based on the contents of email. Such a knowledgebase will be more useful to developers. It improves the comprehension of developer while analyzing various modules. This understanding helps in fixing bugs quickly. Further, new tasks and complex logics can be added with reduced bugs. The quality of software being developed will be high. Further, the company can categorize its human resources based their expertise. For example, a person who is an expert in Simple Mail Transfer Protocol (SMTP) can be located.

2. RELATED WORK

Several research works have been proposed in e-mail mining including e-mail summarization, classification, spam detection, etc. Various e-mail mining tasks such as e-mail clustering, classification, summarization, spam filtering, automatic answering etc. have been described in [1]. Intelligent techniques for automated e-mail answering and e-mail summarization are presented in [2]. An approach for spam filtering based on noisy user feedback is presented in [3]. E-mail classification based on e-mail size and features using techniques such as neural network, Naïve Bayesian classifier and J48 classifier is discussed in [4]. A method for mapping *botnet* membership using traces of spam e-mail has been proposed in [5]. Methods such as [6], [7] focus identifying social networks/groups by mining the data available in e-mail communication. A method for finding similarity between objects (having attributes of different data types) has been carried out in research works such as [8], [9]. A weighted similarity based clustering model has been proposed to discover email groups in [10]. In contrast to the above approaches, the proposed work aims to construct a knowledge base that gives useful information about projects being taken up a company, various modules of projects, their bugs, methods of fixing bugs, etc using data contained in e-mails.

3. PROPOSED METHOD

The methodology used to implement the work is given in Fig. 1.

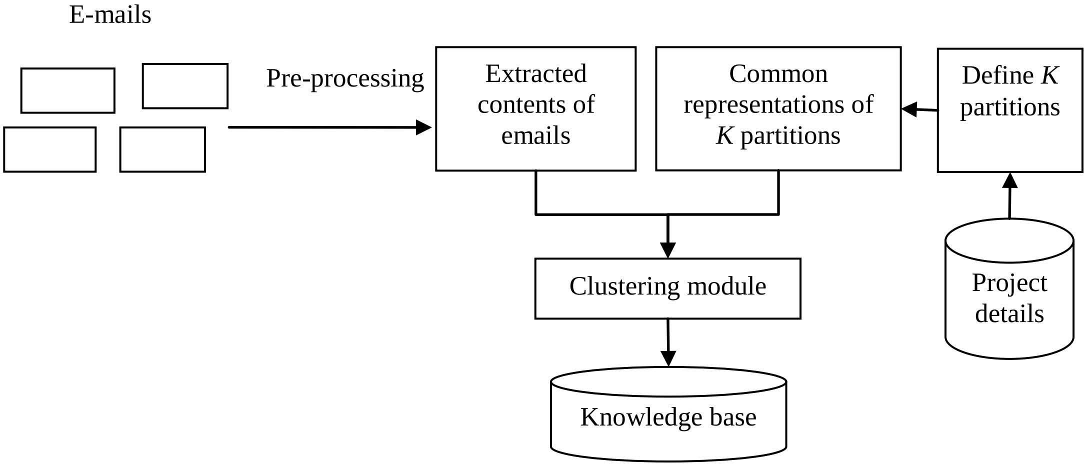

Fig.1 Proposed method

Basically an e-mail consists of a header, body and one or more attachments. The header of e-mail comprises of various fields namely from, to, subject, cc and bcc. In this project it is proposed to use contents from subject field and body of email are used for constructing the knowledgebase. The subject field and contents of the body are pre-processed for removal of all common words, stop words, etc. Depending on the number of projects, number of partitions is chosen. If a company has K number of projects, then K number of partitions is defined. Each partition is defined using a common representation. This representation includes the information about all the modules of a project. For example, the common representation may be a string consisting of all names of modules. Now, clustering module will compare the extracted contents of an e-mail with common representation of each partition. The partition which is most similar to the extracted contents of email is chosen as the relevant cluster of the e-mail and the e-mail is allotted to that partition. In this way all the emails are clustered. Further, it is proposed to test the approach with sufficient large data base.

4. CONCLUSION

In this paper, the importance of extracting contents from emails is analyzed from the perspective of software development. A simple clustering based method is suggested to extract and categorize the contents from emails. The knowledgebase we propose in this work will have the following benefits.

- It will improve the comprehension of programmers about their concerned modules and their interactions with other modules.
- The better understanding will help in fixing the assigned bugs more easily in a comparatively shorter time
- When new tasks are added to existing modules, the tasks can be added with reduced errors and it increases the quality of software being developed.
- Further, the proposed knowledgebase will assist the management in finding out categorization of employees based on their expertise and skill sets.

Presently, efforts are being taken to implement the proposed method with simulated dataset.

5. REFERENCES

[1] Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas, "Email Mining: Emerging Techniques for Email Management" Web Data Management Practices: Emerging Techniques and Technologies, Athena Vakali, George Pallis (Ed.), Idea Group Publishing, pp. 219-240, 2006.

[2] Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas, "Email Mining: Emerging Techniques for Email Management" Web Data Management Practices: Emerging Techniques and Technologies, Athena Vakali, George Pallis (Ed.), Idea Group Publishing, pp. 219-240, 2006.

[3] D. Sculley, Gordon V. Cormack, "Filtering Email Spam in the Presence of Noisy User Feedback", CEAS 2008: Proceedings of the Fifth Conference on Email and Anti-Spam. August, 2008.

[4] Seongwook Youn and Dennis McLeod, "A Comparative Study for Email Classification", Proceedings of International Joint Conferences on Computer, Information, System Sciences, and Engineering (CISSE'06), Bridgeport CT, December 2006.

[5] Li Zhuang, John Dunagan Daniel R. Simon Helen J. Wang, J. D.Tygar, Characterizing Botnets from Email Spam Records, Proceedings of the 1st Use nix Workshop on Large-Scale Exploits and Emergent Threats, Article no. 2, 2008.

[6] Vakali & G. Pallis, "Data Mining Email to Discover Social Networks and Emergent Communities", Web Data Management Practices: Emerging Techniques and Technologies, Ch. 10, Idea Group Inc., 2007.

[7] Christian Bird, Alex Gourley, Anand Swaminathan, Mining Email Social Networks, MSR’06, May 22–23, 2006, Shanghai, China.

[8] Naresh Kumar Nagwani, “OSIM: An Open Source Framework For Measuring Weighted Similarities Between Objects”, International Journal of Emerging Technologies And Applications in Engineering, Technology and Sciences (IJ-ETA-ETS), ISSN: 0974-3588, Jan 2010.

[9] Naresh Kumar Nagwani, Pradeep Singh, “Weight similarity measurement model based, object oriented approach for bug databases mining to detect similar and duplicate bugs", International Conference on Advances in Computing, Communication and Control archive, Proceedings of the International Conference on Advances in Computing, Communication and Control, pp. 202-207, 2009.

[10] Naresh Kumar Nagwani and Ashok Bhansali, “An object oriented email clustering model using weighted similarities between emails attributes”, International Journal of Research and Reviews in Computer Science , vol. 1, no.2, pp. 1-6, June 2010.