ISSN:2249-5789

IMPLEMENTATION OF PERSONALIZED EMAIL PRIORITIZATION- A CONTENT BASED SOCIAL NETWORK ANALYSIS

K.HARINI[1], UPPE.NANAJI[2]

[1]Department of C.S.E, Avanthis St.Therissa institute of engineering technology, Garividi, Andra Pradesh, India, email: harini.kolluru1@gmail.com.

[2]Asst.professor, Department of C.S.E, Avanthis St.Therissa institute of engineering technology, Garividi, Andra Pradesh, India

Abstract— **Now a Day's, Email is one of the most prevalent personal and business communication tool, but it exhibits some significant drawbacks. One of the drawback of this is the portion of e-mail is that spam, has increased dramatically in the past few years. A recent study showed that 52 percent of e-mail users say spam has made them less trusting of e-mail, and 25 percent say that the volume of spam has reduced their e-mail use. This crisis has prompted proposals for a broad spectrum of potential solutions, ranging from more efficient antispam software tools to antispam laws at both the federal and state levels. But Still there is a necessity, to overcome this, in this paper We propose a novel approach that combines unsupervised clustering, social network analysis, semi supervised feature induction, and supervised classification to model user priorities among incoming email messages. This article presents the first study on PEP with a fully personalized methodology, where only each user's personal email data. Finally results show that this method somehow better compared to other methods.**

Index Terms— **WWW, Email, Containment, social network analysis.**

I. INTRODUCTION

Email is one of the most prevalent personal and business communication tools today; however, it is not without significant drawbacks. In contrast to telephone conversations or face-to-face meetings, communication through email is asynchronous in the sense that we receive messages (after some spam filtering) in the same way regardless of our level of interest, and a single sender can flood multiple receivers (unlike telephone or instant messaging). Users are left with the burden of having to process a large volume of email messages of differing importance. This tedious task has been shown to cause significant negative effects on both personal and organization performance [1, 2]. There is an urgent need to solve this information overload problem; i.e., we need to develop systems that automatically learn personal priorities for each user, and that identify personally interesting and important messages for user's attention.

Many statistical learning techniques have been studied in support of email-based prediction tasks, including supervised, unsupervised and semi-supervised methods for spam identification [3,4], folder recommendation [5], recipient reminding [6], action-item identification [7], social group analysis [8], etc. In spite of the wide variety of efforts and significant accomplishments, personalized email prioritization (PEP) remains an under-explored problem. Thorough investigations and conclusive solutions have been rare, mainly due to privacy issues in collecting personal data for training and testing. Unlike spam filtering where people are less concerned with sharing individually labeled spam messages, PEP requires personal judgments of the importance levels of non-spam email messages. Few are willing to share this data due to privacy concerns. Companies who have access to customers' email messages (like Google, Yahoo! and Microsoft) cannot share such data with academic institutes for the same reason. Personal importance judgments are also missing from the Enron corpus, which has been used as a benchmark dataset in email research and evaluations. A message important for an Enron employee might not be equally important for a high-level manager. In short, there is no publicly available dataset that contains personal importance judgments by real users and on personal messages, leaving researchers no choice but to go through a process of collecting private data under strict IRB (Institutional Review Board) guidelines. Such data collection processes are costly, time consuming, tedious, and difficult to scale to a large number of users with diverse criteria in judging the importance of email messages. As a result, PEP remains an area which has not been well studied thus far. This paper presents the first study with several statistical classifications and clustering methods (including our new approach) addressing the PEP problem based on personal importance judgments by multiple users.

II. LITERATURE SURVEY

Eric Horvitz[9] and his colleagues built an email alerting system that used support vector machines to classify newly arrived email messages into two categories— that is, high or low in terms of utility. However, their task did not consider personalization or investigate social network analysis.

Joshua Tyler[10] and his colleagues used the Newman Clustering algorithm to discover social structures from email messages. They found that the automatically discovered social structures (such as social leaders) are consistent with human interpretation of organizational structures. However, they did not focus on the email prioritization problem.

Carman Neustaedter[11] and her colleagues defined metrics for measuring the social importance of individuals based on the From, To, and CC fields in email messages and recorded user actions in replying and reading email. They used these metrics for retrieving old email messages rather than prioritization of new messages.

Lisa Johansen[12] and her colleagues used social clustering to predict the importance of email messages. The major difference between their method and ours is that their clusters were induced from a community social network, not based on personal social networks or the content information in email messages.

Lastly, Fei-Yue[13]Wang and his colleagues discussed the theoretical, methodological, and technological underpinnings of social computing in general and reviewed the major application areas. With this article, we leverage the good ideas in these previous works and develop new techniques for personalized email prioritization.

III. PROPOSED SYSTEM ARCHITECTURE

A. Existing System:

Email is one of the most prevalent personal and business communication tools today, but it exhibits some significant drawbacks. Unlike telephone conversations or face-to-face meetings, email messages are received (after some spam filtering) in the same way regardless of a user's level of interest, and a single sender can flood multiple receivers. As a result, users must process a large volume of email messages of different importance

B. Proposed system

Here we propose a mechanism called Personalized email prioritization (PEP) with a fully personalized methodology, where only each user's personal email data (textual content and social network information) is available for the system during the system's training and testing. This is an important assumption for the generality of PEP methods—that is, we cannot rely on the availability of centralized access to customer private data in the development cycle or evaluation phase, and we cannot take the liberty of using a particular user's private data to build models for other users because of the potential leak of private information. Such strictly separate data makes our work fundamentally different from research in spam filtering and other previous work on email-based prediction. We propose a novel approach that combines unsupervised clustering, social network analysis, semi-supervised feature induction, and supervised classification to model user priorities among incoming email messages.

We treat the priority prediction task as a supervised classification problem and use standard support vector machines (SVMs) as the classifiers. The novel part of our approach is the enriched representations of email messages and users, with automatically extracted features.

C. Deficiency in Existing approach:

The users must process a large volume of email messages of different senders. user required an accurate spam filter system. lack personal importance judgments agents the private data. There is a system that but to collect private data under strict Institutional Review Board (IRB)guide lines. Such data-collection processes are costly, time consuming, and tedious, making it difficult to acquire a large number of users with diverse criteria in judging the importance of email messages, But Rely on the availability of centralized access to customer private data in the development cycle or evaluation phase, and we cannot take the liberty of using a particular user's private data to build models for other users because of the potential leak of private information.

The importance of each sender group can be automatically learned by SVM classifiers. We chose the Newman Clustering (NC) algorithm, which researchers have used to successfully find social structures in large organizations. It defines the edge-betweenness (which we discuss in detail later) as a measure of the shortest path(s) going through a specific link among all-pairs shortest paths. A link with a high edge betweenness score is crucial for connecting two highly connected component clusters. By deleting links with high edge-betweenness scores and removing those edges from the graph, we obtain disconnected component clusters.

One way to control the granularity level of clusters is to prespecify the number of desired clusters, which might be based on domain knowledge about the social networks in email or automatically determined by algorithms with a certain optimization criterion or heuristic measure. For example, the NC method can pick the number that yields the largest decrease in the sum of edgebetweenness per cluster. We use this method in our work. We propose the Level-Sensitive PageRank (LSPR) approach to propagate labeled importance of the training examples to other messages and connected users.

IV. EXPERIMENTAL RESULTS

We recruited a set of subjects, mostly from the Language Technologies Each subject was asked to

label at least 400 nonspam messages during a one-month period using a five-level scale. Only seven users actually labeled more than 200 messages, which we used as the collected data for our experiments. In each personal data collection, we sorted the email messages temporally and split the sorted list into 70 and 30 percent portions. We used the 70 percent portion for training and parameter tuning and the remaining 30 percent for testing. Figure 2 shows the performance conditioned on varying training-set sizes of 30 to 150 labeled messages. Adding the social network based features (SI, NC, and LSPR) significantly reduced the importance prediction errors in both micro- and macro averaged MAE. We conducted Wilcoxon signed rank tests to compare the results of SVMs using only BF features versus using the additional features. The p values in these conditions are below 1 percent except in one case, when the training- set size is 60 and the p-value is 5 percent. These results strongly support the advantage of leveraging the social-network features in combination with content-based features over the baseline approach.

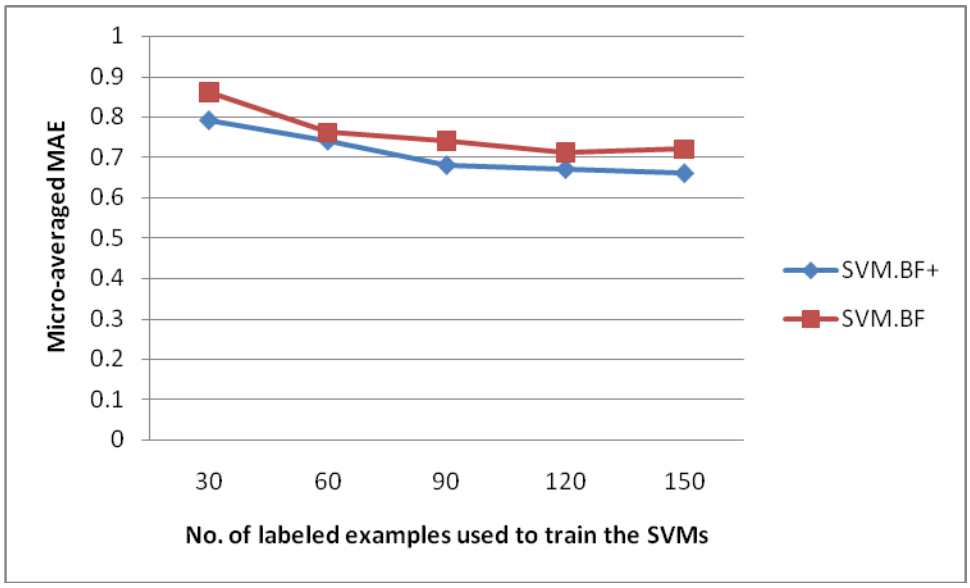

Fig 2 Performance of support vector machines (SVMs) in micro-averaged mean absolute error (MAE)

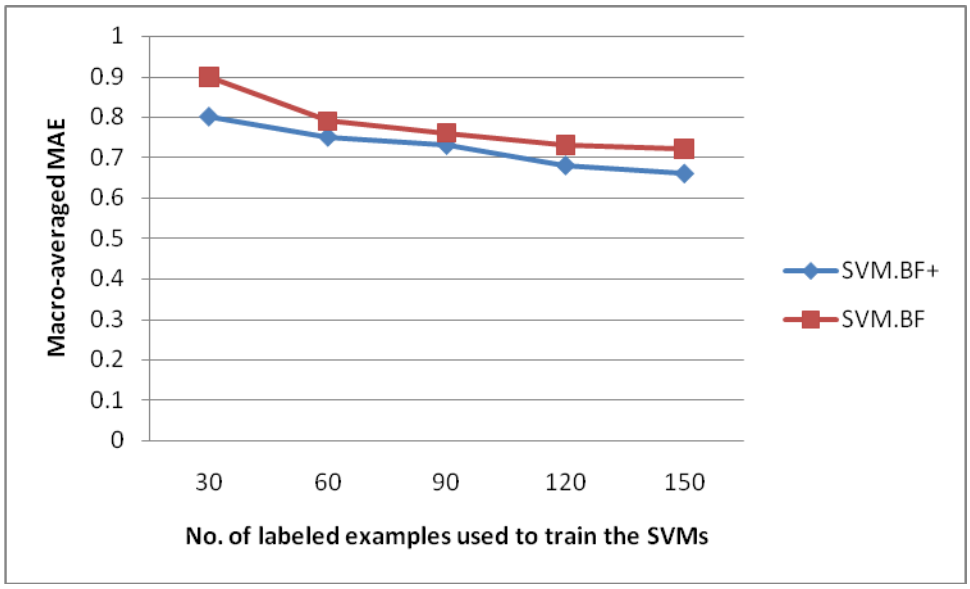

Fig 2 Performance of support vector machines (SVMs) in Macroaveraged MAE

V. CONCLUSION

Email is one of the most prevalent communication tools today, and solving the email overload problem is pressingly urgent. A good way to alleviate email overload is to automatically prioritize received messages according to the priorities of each user. However, research on statistical learning methods for fully personalized email prioritization (PEP) has been sparse due to privacy issues, since people are reluctant to share personal messages and importance judgments with the research community. This paper presents the first study (to the best of our knowledge) under such an assumption. Specifically, we focus on analysis of personal social networks to capture user groups and to obtain rich features that represent the social roles from the viewpoint of a particular user.

References

[1]. L. A. Dabbish and R. E. Kraut. Email overload at work: an analysis of factors associated with email strain. In P. J. Hinds and D. Martin, editors, Proceedings of the 2006 ACM Conference on Computer Supported Cooperative Work, CSCW 2006, Banff, Alberta, Canada, November 4-8, 2006, pages 431–440. ACM, 2006.

[2]. M. Wattenberg, Rohall, S. L., D. Gruen, and B. Kerr. E-mail research: Targeting the enterprise. Human-Computer Interaction, 20(1/2):139–162, 2005.

[3]. Joshua Goodman, Gordon V. Cormack, and David Heckerman. Spam and the ongoing battle for the inbox. Commununications of the ACM, 50(2):24–33, 2007.

[4]. M. Mojdeh and G. V. Cormack, Semi-supervised Spam Filtering: Does it Work?, SIGIR 2008.

[5]. B. Klimt and Y. Yang. The Enron Corpus: A New Dataset for Email Classification Research. ECML 2004.

[6]. R. Balasubramanyan, V. Carvalho and W. Cohen, CutOnce - Recipient Recommendation and Leak Detection in Action. In AAAI-2008, Workshop on Enhanced Messaging.

[7]. P.N. Bennett and J. Carbonell (2007). Combining Probability-Based Rankers for Action-Item Detection. In Proceedings of HLT-NAACL 2007.

[8]. A. McCallum, X. Wang and A. Corrada-Emmanuel. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email

[9]. E. Horvitz, A. Jacobs, and D. Hovel, "Attention-Sensitive Alerting," *Proc. Conf. Uncertainty and Artificial Intelligence*, Morgan Kaufmann, 1999, pp. 305–313.

[10]. J R. Tyler, D.M. Wilkinson, and B.A. Huberman, "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations," *Communities and Technologies*, M. Huysman,

[11]. E. Wenger, and V. Wulf, eds., Kluwer, 2003, pp. 81–96. 3. C. Neustaedter et al., "The Social Network and Relationship Finder: Social Sorting for Email Triage," *Proc. Conf. E-mail and Anti-Spam*, 2005;

[12]. L. Johansen, M. Rowell, and P. McDaniel, "Email Communities of Interest," *Proc. 4th Conf. E-mail and Anti-Spam*, 2007;

[13]. F.Y. Wang et al., "Social Computing: From Social Informatics to Social Intelligence," *IEEE Intelligent Systems*, vol. 22, no. 2, 2007, pp. 79–83.