Data Mining in Personal Email Management

Gunjan Soni

E-mail is still a popular mode of Internet communication and contains a large percentage of every-day information. Hence, email overload has grown over the past years becoming a problem for personal information management for users and a financial issue for companies. This survey reviews research on how a Machine Learning and Data Mining technique, such as classification, clustering can contribute to the solution to the problem by constructing intelligent techniques which automate email managing tasks. This survey contains annotations of research publications describing approaches used to aid in better understanding of the research for personal email information management. Some email mining applications such as automatic folder creation, mail summarization, automatic answering and spam filtering will be also presented.

Categories. and Subject Descriptors: H.2.8 [Database Applications]: Data Mining

General. Terms: machine learning, text mining

Additional. Key Words and Phrases: email, feature selection, clustering, topic detection.

Contents

1. INTRODUCTION

This survey is about research on different email management techniques, such as classification of emails in different folders, spam detection, email summarization etc for the purpose of reducing email overload. The papers used to write this survey can be easily found using the Google Scholar with keywords such as "Email Management", "Email Mining" etc; also the use of web portal of University of Windsor Leddy library's with various digital libraries such as IEEE Xplore, ACM, Lecture Notes in Computer Science etc. was very helpful for having the relevant papers.

Papers that were published in conference proceeding and journal, included Whittaker et al. [1996], Mock et al.[2001], Stolfo et al. [2003], Aery et al. [2004], Berry et al. [2005], Kulkarni et al. [2005], Kushmerick et al.[2005], Tang et al. [2005], Li et al. [2006], Schuff et al. [2006], Appavu et al. [2007], Haider et al. [2007], Yang et al. [2007], Mojdeh et al. [2008], Haider et al. [2009], Li et al. [2009], Yoo et al. [2009], Nagwani et al. [2010], Yang et al. [2010], Cernain et al. [2011].

The rest of the survey is organized as follows. Next subsection gives introduction to the different technique applied for email data representation and the feature selection for further processing. Section 2 presents different approaches for email management mainly focusing on classification and clustering of emails and describe research on the application of email management. Later sections 3 and 4 concludes the survey with proposing some future work and annotations of different research papers respectively.

1.1 Email Preprocessing and Representation

This section describes how the emails are preprocessed after parsing for the tokenization and data representation like use of vector space model which can further be used in clustering or classification.

The first step for email management is the data extraction and processing. Here short lexical entities such as single words or word pairs are extracted from the e-mail data set. For data extraction there exist lot of APIs like javamail based on JSP to extract email data from email servers.

After data extraction from the server, we need to remove stop words such as 'the', 'for'', of' etc from the corpus. Next, stemming algorithm is used to stem the data. For example, words 'connection', 'connecting', 'connected' will be converted to 'connect'.

After stemming we need to prepare those data for clustering. In this phase data are represented with data cluster features such as unigrams, bigrams and co-occurrences.

Unigram: unigrams are just significant individual words. For example, in a data such as "hello dear, how are you?" the unigrams are 'hello', 'dear', 'how', 'are' and 'you'.

Bigram: bigrams are pair of two adjacent words. For example, in a data such as "hello dear, how are you?" the bigrams are 'hello dear', 'dear how', 'how are' and 'are you'. Besides, in bigrams word sequence is important. 'hello dear' and 'dear hello' are two different units.

Co-occurrence: Co- occurrences are same as bigrams but the only difference is here word sequence is not important. For example, 'hello dear' and 'dear hello' are treated as single unit as order does not matter.

There is another feature called target co-occurrence which is same as co-occurrence with one target word inside each pair.

2. SURVEY OF RESEARCH

2.1 Management by clustering

Email Clustering goes one step further. Subject-based folders can be automatically constructed starting from a set of incoming messages in this case, the goal is to build automatic organization systems which will analyze an inbox recognize clusters of messages with the same concept, give an appropriate name to each cluster and then put all messages into their corresponding folders. Research papers that are presented in this section use clustering methods to manage all emails.

2.1.1 *A multi-attribute, multi-weight clustering approach to manage e-mail overload.* According to Schuff et al. [2006] there is no efficient automated process exists to manage the e-mail overload, which will help users to manage hundreds of email automatically based on the content of a message. An efficient email management system can reduce the information overload and mental workload of a certain user.

Schuff et al. [2006] do not refer to any of my selected papers as their related work.

Schuff et al. [2006] propose a new multi weight, multi attribute clustering system that will automatically create folder structure in user's inbox based on the combination of email subject, sender, and receiver and text body. In their proposed system the user can set their desired weight to a particular attribute.

Schuff et al. [2006] state that for evaluation their experimental subjects were daily emails of 65 students from an introductory computer literacy class. The data used analyzed using both multivariate and uni-variate analysis of variance models. To verify the appropriateness of multivariate, it is also verified that the assumptions of normality and homogeneity of error variance across groups were upheld.

According to Schuff et al. [2006] the results of this research are potentially important for both academics and practitioners. For academics, this study integrates the concepts of semantic network theory and research on human memory chunking from cognitive psychology with prior information- science studies on textual document clustering. We extended this research to include clustering on key attributes of a textual document (in this case, attributes of an e-mail message). The ACEMS experiment has two implications for theory. First, while the application of a semantic network to an e-mail collection resulted in a nearly 41% improvement in task effectiveness, the additional increase from customizing the structure of the network was only marginally.

Schuff et al. [2006] claim that their proposed multi-weighted, multi-attribute method increase retrieval effectiveness reduces perceived effort and increase intension to use. They also claim that their system offers a general contribution in extending the application of semantic network theory.

The work of Schuff et al. [2006] is cited by Yang et al. [2007].

2.1.2 *Adding semantic to email clustering.* According to Li et al. [2006] email classification is a ways to manage emails but supervised classification needs a predefined taxonomy which requires user involvement and also after the development of clustering technique, it was also not possible to have satisfactory performance.

No previous work is mentioned by Li et al. [2006]

Li et al. [2006] propose a model to automatically mine the semantic knowledge from the subject line of an email and create a cluster according to the similarity. In this method, each subject line is treated as a sentence and parsed through natural language processing techniques. The algorithm consists of four levels: 1. Generalization of terms in email subject line, the subject line parsing is done to create a syntactic tree using Microsoft NLPWin tool; 2. Mine Generalized Sentence Pattern (GSP), patterns are generated from the generalized terms; 3. GSPs grouping and selection, GSPs in the same group will represent the same cluster; 4. GSP-PCL: GSP as pseudo class label.

The GSP-PCL clustering algorithm was experimented on two datasets: the open dataset Enron email dataset and a private email dataset collected by the Li et al. [2006]. In Enron email dataset, the minimum support threshold (min_sup) was set to 4 and the minimum length of GSPs was restricted to 2.

When Li et al. [2006] compared GSP-means and K-means clustering on Enron email dataset and personal email dataset, the result showed that the readability is improved by 68.5%.

Li et al. [2006] states that model suggested automatically extract embedded knowledge from the email subjects to help improve email clustering and GSP-PCL obtains significant improvement both on the clustering quality and cluster name readability compared with the basic K-means algorithm.

The work of Li et al. [2006] is cited by Yang et al. [2010].

2.1.3 *Managing email overload with an automatic nonparametric clustering approach.* According to Yang et al. [2007] the email overload is a problem which user faces to process the large number of emails received/sent. As result it affects the usage or purpose of emails as effective knowledge management tool for communication.

Yang et al. [2007] mentioned the previous work of Schuff et al. [2006].

According to Yang et al. [2007] the work of Schuff et al. [2006] relies on the user involvement, i.e. they used techniques which is semi-supervised by user.

Yang et al. [2007] present an automatic email clustering system for automatic categorization of email into different meaningful groups by proposing a new automatic nonparametric clustering approach to manage email overload. The method works as: firstly, read the email messages from email client's data file, then it converts email texts into vector matrix and generate similarity matrix. Now once matrices are generated they are input into to the nonparametric text clustering algorithm. Then, the algorithm produces email clusters

Yang et al. [2007] email data sets are from real life email collections. The comparison is made with the results of the authors approach to the results of the k-mean algorithm and the hierarchical agglomerative algorithm. The quality is measured by Hubert's G statistic, simple matching coefficient, and Jaccard coefficient.

Yang et al. [2007] result shows that for computational time analysis, hierarchical agglomerative algorithm takes 808% time more from the proposed algorithm to perform the clustering, and k-means algorithm takes 342% time more from the proposed algorithm to perform the clustering. For Hubert's G statistic is always higher than 0.764 when using the proposed algorithm which is mostly higher than Hubert's G statistic for other two algorithm. The Jaccard coefficient is found to be more than 0.821 for all data sets.

Yang et al. [2007] claim that email users get clustered emails easily without any input. The experiments shows that proposed algorithm has high efficiency and high clustering quality in terms of computation time and clustering quality.

There are no specific references to the work of Yang et al. [2007] by other researchers in this survey.

2.1.4 *Bayesian clustering for email campaign detection.* According to Haider et al. [2009] there exist problems in clustering elements according to the sources that have generated them. For the independent binary attributes, a closed form of Bayesian solution exist but for dependent attributes that is based on a transformation of the instance was proposed by the authors.

The author refer previous work by Haider et al. [2007].

According to author the work of Haider et al. [2007] is not workable in practical where the effort of partitioning the data is much higher than the effort of labeling the labeling data for classification.

Haider et al. [2009] discussed the clustering of emails according to the sources that have generated using the Bayesian clustering algorithm. There are three main parts of algorithm: Firstly, they developed a model that produces a cluster of binary features vectors, based on a transformation of the input vectors. Secondly, generate an optimization problem and algorithm that produce the features transformation.

Haider et al. [2009] presented a large-scale case study that analyzes Bayesian clustering solution for email campaign detection.

Haider et al. [2009] found a small fraction of spam messages, a total of 139,250 spam messages in correct chronological order. In order to maintain the users' privacy, authors blend the stream of spam messages with an additional stream of 41,016 non-spam messages from public sources. The non-spam portion contains newsletters and mailing lists in correct chronological order as well as Enron emails and personal mails from public corpora which are not necessarily in chronological order. Every email is represented by a binary vector of 1,911,517 attributes that indicate the presence or absence of a word. The feature transformation technique introduces an additional 101,147 attributes.

Haider et al. [2009] claimed that they devised a model for Bayesian clustering of binary features vectors based on Bayesian solution of the data likelihood in which the model parameters.

There are no specific references to the work of Haider et al. [2009] by other researchers in this survey.

2.1.5 *An object oriented email clustering model using weighted similarities between emails attributes.* According to Nagwani and Bhansali [2010] it is possible to discover useful patterns from emails dataset which can further be used to manage the emails.

The authors refer to previous work by Bird et al. [2006].

The problem with the previous work was that it was not accurate.

Nagwani and Bhansali [2010] propose an automatic organization system which analyzes an inbox to recognize cluster of messages and put them in their corresponding folders. This system measures the weighted email attribute similarity between a pair of email objects like from-mail-id, to-mail-id, subject, message, sending time etc. using OSim (Object Similarity) distance function. The proposed method has three stages – 1. Pre-processing, it includes parsing, stemming and email representation technique for parsed information; 2. Weighted attributes similarity of Emails, it includes fetching the email attributes from processed database, then calculating the pair-wise attribute similarity of email document and finally assigning weights to the similarity measured for attribute pair-wise to calculate the overall similarity between a pair email document; and 3. Applying clustering technique over the measured similarity information to create email clusters.

Nagwani and Bhansali [2010] tested their algorithm by experimenting with an inbox folder of "bass-e" from Enron email corpus datasets with Java as programming language and Simmetric & Weka as the other open source API's to support some functionality. Nagwani and Bhansali [2010] also evaluated the accuracy of the proposed model by the 10-fold cross validation technique.

Nagwani and Bhansali [2010] state that the selected inbox folder consists of around 310 emails and total of eight clusters were generated from the given dataset by implementing this model and gives the similarity thresholds for the cluster as 0.05%. The evaluation of accuracy results around 78%.

Finally, Nagwani and Bhansali [2010] claim that the proposed model is implemented for discovering the email groups with good accuracy.

There are no specific references to the work of Nagwani and Bhansali et al. [2010] by other researchers in this survey.

2.1.6 *Automatically detecting personal topics by clustering emails.* According to Yang et al. [2010] there are three problems in detecting topics by clustering. Firstly, choosing the method for text feature selection, Secondly, the way to combine the email subject and body features and lastly, since Yang et al. [2010] use the k-mean clustering algorithm to cluster email therefore there is a problem in finding the value of k automatically and selecting the appropriate initial k kernels.

The authors refer to previous work by Li et al. [2006].

No shortcomings of previous work were mentioned by the Yang et al. [2010].

Yang et al. [2010] propose a model to automatically detect the personal topic from the email inbox using a clustering algorithm. The approach is divided into three steps 1. Email representation with the EVSM (Email Vector Space Model); 2. Kernel selection algorithm based on lowest similarity; and 3. Email topic detection algorithm. The email representation with the EVSM is again split into three stages – Selection of body and subject features by selecting the n top-ranked high frequency words, Combine the body and subject of the email and Construction of the EVSM by applying the standard vector space model approaches.

Yang et al. [2010] did three experiments with four folders of the mini_newsgoups which is part of the data source 20NewsGroup. Experiment 1 consisted of implementing the standard k-mean algorithm. Secondly, implementing the proposed algorithm and lastly, clustering email by combining the body and subject fields with the proposed approach.

The results of implementation of the clustering algorithm are measured by Yang et al. [2010] is in terms of F-value which 0.8584 for standard K-means and 0.9163 for improved K-means.

Yang et al. [2010] claimed that the automatic detection of personal topic by clustering emails is successfully implemented and also they did some improvement on the construction of the EVSM and the kernel selection of the k-mean algorithm including the criteria of space and time complexity of the large-scale data processing.

There are no specific references to the work of Yang et al. [2010] by other researchers in this survey.

2.1.7 *The design and validation of an automatic email clustering system based on semantics.* According to Cernian et al. [2011] a user sends and receives hundreds of emails every day and hence managing the emails is time consuming and annoying when done manually, even searching the email is also difficult if it have many messages in the respective folder.

No previous work is mentioned by Cernian et al. [2011].

Cernian et al. [2011] propose a novel approach for managing email using email clustering based on semantic criteria, by using the subject line and body of the messages in the inbox or from another folder of a Gmail server. The application can only work for the English and Romanian languages. .

After preprocessing the email the massages are sent to the clustering engine for generates the clustering, based on its interpretation of the distance matrix. Clustering engine consist of: The text processing engine, the BZIP2 compression algorithm and the UPGMA clustering algorithm. The non-spam portion contains newsletters

and mailing lists in correct chronological order as well as Enron emails and personal mails from public corpora which are not necessarily in chronological orderFor each group a folder is created on the local computer in a predefined location.

For the validation Cernian et al. [2011] purpose 2 datasets were used: a set of 50 emails written in Romanian and a set of 50 emails written in English and FScore was calculated to assess the quality and robustness of the classification process.

The first set of experiment by Cernian et al. [2011] experiments with Romanian dataset the FScore for: Unprocessed – 0.70, Stop words – 0.71, Stemming - 0.69 and complete – 0.93. Secondly, from the second set of experiment with English dataset the FScore for: Unprocessed – 0.79, Stop words – 0.82, Stemming - 0.84 and complete – 0.95.

According to Cernian et al. [2011] whole system was successfully implemented and the FScore values obtained proved the impending of the clustering by grouping to correctly interpret the informational content of the data.

There are no specific references to the work of Cernian et al. [2011] by other researchers in this survey.

2.1.8 *A novel approach for clustering e-mail users using pattern matching.* According to Shazmeen et al. [2011] emails have been considered as a useful resource for research in fields like link analysis, social network analysis and textual analysis and discovering useful patterns from emails can be useful for reducing the overload problem with today email inbox.

No previous work is referred by Shazmeen et al. [2011].

Shazmeen et al. [2011] propose a novel approach for clustering e-mail users using pattern matching email attributes such as sender email-id, receiver email-id Subject, message, sending-time, and attachments etc. clustering is used to discover email groups. The whole process is divided into different stages: In pre-processing phase all the email data is prepared for clustering. All the important email attributes are retrieved by parsing the email documents. Most of the attributes are of text data type, so stemming techniques are used to eliminate the unwanted texts from the parsed attribute information, after preprocessing clustering the users who are showing similarity in discussing the same context is clustered and graphically representing the cluster.

Shazmeen et al. [2011] algorithm is tested with Enron Email dataset.

Experiment conducted by Shazmeen et al. [2011] showed that there are two clusters formed, “announcement” with cluster size 6 with threshold 1 and “conference” with cluster size 8 with threshold 1.

In this paper Shazmeen et al. [2011] claimed that an email clustering approach is proposed and implemented to show text similarities and that the proposed technique shows the email attributes and how the text similarities are used to cluster the users.

There are no specific references to the work of Cernian et al. [2011] by other researchers in this survey.

Year	Authors	Title	Contribution
2006	Scuff, Turetken and Arcy	A multi-attribute, multi-weight clustering approaches to managing " email-overload"	They first introduced the multi attribute, multi-weight clustering approaches which has increased the retrieval effectiveness
2006	Li, Shen, Zhang and Yang	Adding semantic to email clustering	novel algorithm to mine the semantic knowledge from subject line and then to cluster similar emails accordingly.
2007	Yang	Managing email overload with an automatic nonparametric clustering approach.	Automated email categorization algorithm
2009	Haider and Scheffer	Bayesian clustering for email campaign detection	they devised a model for Bayesian clustering of binary features vectors based on Bayesian solution of the data likelihood in which the model parameters.
2010	Nagwani and Bhansali	An object oriented email clustering model using weighted similarities between emails attributes	proposed object oriented email clustering process
2010	Yang, Luo, Yin and Liu	Automatically detecting personal topics by clustering emails.	Automatically detecting personal topics by clustering emails Cernian et al. [2011] - The design and validation of an automatic email clustering system based on semantics. – proposed and integrated system for automated clustering
2011	Shazmeen and Gyani	A Novel Approach for Clustering E-mail Users Using Pattern Matching	novel approach for clustering e-mail users using pattern matching according to email attributes
2011	Cernian, Florea, Carstoiu and Sgarciu	The design and validation of an automatic email clustering system based on semantics.	novel approach for managing email using email clustering based on semantic criteria

Table I.

2.2 Management by classification

Most email mining tasks are accomplished by using email classification at some point. In general, what email classification confronts is the assignment of an email message to one from a pre-defined set of categories. Automatic email classification aims at building a model (typically by using machine learning techniques), which will undertake this task on behalf of the user. Research paper that is presented in this section use classification method to manage all emails.

2.2.1 *Supervised clustering of streaming data for email batch detection.* According to Haider et al. [2007] filtering spam more efficiently by exploiting the collective information about entire batched or group of jointly generated message is one of the important parts of email management. Haider et al. [2007] addressed the problem of detecting batches in an email streaming that have been created according to the same template.

No previous work is mentioned by Haider et al. [2007]

Haider et al. [2007] generates a model for detecting batches of emails into a well-defined problem setting of supervised learning. For the whole process Haider et al. [2007] first derived a compact optimization problem based on the LP approximation to correlation clustering to learn the weights of the similarity measure, then they devised an efficient clustering algorithm with computational complexity linear in the number of emails and in-turn to complete this task they integrated method for learning the weight vector.

Haider et al. [2007] evaluated the performance and benefit of batch detection on an emails collection and also evaluated the method of identification of email batched for spam or non-spam email detection. These experiments are done on Enron corpus datasets. Firstly, Haider et al. [2007] created an email corpus that reflects the features of an email stream. Secondly, the comparison is done of four strategies for clusters/batches identification: LP decoding, sequential decoding, agglomerative decoding and decoding strategies, using the similarity matrix obtained from pairwise learning. Thirdly, the evaluation of the classification of email as spam or non-spam is done.

Haider et al. [2007] found that the final corpus contains 2,000 spam messages, 500 Enron messages, and 500 newsletters. Secondly, while finding the ideal batch information, the risk of misclassification is reduced by 43.8%, while with non-ideal batch information obtained through approximation clustering still 41.4% reduction are achieved.

Haider et al. [2007] devised a sequential clustering algorithm and two integrated formulations for learning a similarity measure to be used with correlation clustering. Haider et al. [2007] also claimed that a sequential clustering algorithm can efficiently make supervised batch detection at enterprise-level scale. The work of Haider et al. [2007] is cited by Haider et al. [2009].

year	Authors	Title	Contribution
2007	Haider, Brefeld and Scheffer	Supervised clustering of streaming data for email batch detection.	derived a compact optimization problem based on the LP approximation to correlation clustering to learn the weights of the similarity measure, then they devised an efficient clustering algorithm with computational complexity linear in the number of emails and in-turn to complete this task they integrated method for learning the weight vector.

Table II.

2.3 Management by statistical classification and clustering

This section presents a study with statistical classification and clustering methods.

2.3.1 *Mining social networks for personalized email prioritization.* According to Yoo et al [2009] email overload creates problems for personal information management since it is a burden for user to process a large volume of email messages of differing importance; in turn it causes much negative effect on both personal and organization performance. The email overload can reduced by automatically prioritize received messages according to the priorities of each user called personalized email prioritization (PEP).

No previous work is by Yoo et al [2009].

Yoo et al [2009] present a study with statistical classification and clustering methods addressing the PEP problem based on personal importance judgements by multiple users and also developed a novel transductive learning algorithm that propagates importance labels from training dataset to test dataset via message and user nodes in a personal email network. Firstly, a user as a member of a group is chosen based on unsupervised clustering, and then inference is made on the importance of that particular user from other group members. Later these clusters can be used by SVM classifier as input features to each message.

Yoo et al [2009] engaged 25 experimental subjects where each subject was requested to label at least 400 non-spam messages during a one-month period. The five importance levels are: absolutely non-important, relatively non-important, neutral, important, and most important.

Yoo et al [2009] found the following results: Firstly, below are the performance curves of SVM runs with different representation schemes for email messages. Secondly, the author claims to have obtained significant performance improvement over the baseline system (without induced social features) in our experiments on a multiuser data collection: the relative error reduction in MAE was 31% in micro-averaging, and 14% in macro-averaging.

Year	Authors	Title	Contribution
2009	Yoo, Yang, Lin and Moon	Mining social networks for personalized email prioritization.	present a study with statistical classification and clustering methods addressing the email overload problem

Table III.

Yoo et al [2009] claim that the algorithm they designed is successfully implemented for fulfilling the purpose of the system with less error rate. There are no specific references to the work of Yoo et al. [2009] by other researchers in this survey.

3. CONCLUDING COMMENTS

Email is very important for interpersonal communication and professional life. Therefore its problems demand immediate attention and efficient solutions. Email categorization into folders, email answering and summarization, spam filtering, are only a few examples. All of these applications have been explored repeatedly in the literature with very promising results.

In this survey it is found that different researchers have work with different approaches such as multi weight approach, object oriented approach, clustering approach and many more to solve the email overload problem. It is observed that most of the researchers have used unigrams and TF-IDF methods to process and represent their email dataset before clustering or classification. Only Kulkarni and Pedersen [2005] have used bigrams for representing their data. Schuff et al [2005], Yang et al [2010], Xiang et al [2007], Cselle et al [2007], Manco et al [2008], Kushmerick and Lau [2005], Guan et al [2011], Surendran et al [2005] and Ayodele et al [2009] used unigrams and TF-IDF method. In this survey it is also found that unsupervised clustering methods are more efficient than the supervised classification methods. Moreover, it is found that most of the researchers have used the hierarchical and K-means clustering algorithms for performing the clustering to the email dataset.

It is observed that many future works can be done to solve the email overload problem. Haider et al. [2007], Yang et al. [2010] have referred their future work as to improve the space and time complexity of their proposed algorithms. Nagwani et al. [2010] want to consider the email attachments to do the mining in future. Yoo et al. [2009] want to use graph mining techniques in future to prioritize the user emails.

New solutions had to be proposed in already discussed areas due to email data peculiarity. Additionally, domain specific problems provoked the development of new applications like spam filtering, email answering and thread summarization. While effective solutions have been proposed to most email problems, not all of them have been implemented in popular email clients.

4. ANNOTATIONS

4.1 Cernian et al. 2011

Citation: Cernian, A., Florea, I., Carstoiu, D., and Sgarciu, V. 2011. The design and validation of an automatic email clustering system based on semantics. *In Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2011 IEEE 6th International Conference on. Vol. 2. 629-632.*

Problem: According to authors a user sends and receives hundreds of emails every day and hence managing the emails is time consuming and annoying when done manually, even searching the email is also difficult if it have many messages in the respective folder.

Previous work: No previous work is mentioned by authors.

Shortcoming of previous work: Authors propose a novel approach for managing email using email clustering based on semantic criteria, by using the subject line and body of the messages in the inbox or from another folder of a Gmail server. The application can only work for the English and Romanian languages.

New Idea/Algorithm/Architecture: After preprocessing the email the massages are sent to the clustering engine for generates the clustering, based on its interpretation of the distance matrix. Clustering engine consist of: The text processing engine, the BZIP2 compression algorithm and the UPGMA clustering algorithm. For each group a folder is created on the local computer in a predefined location. Experiments and analysis conducted: For the validation, authors purpose 2 datasets were used: a set of 50 emails written in Romanian and a set of 50 emails written in English and FScore was calculated to assess the quality and robustness of the classification process.

Results: The first set of experiment by author experiments with Romanian dataset the FScore for: Unprocessed – 0.70, Stop words – 0.71, Stemming - 0.69 and complete – 0.93. Secondly, from the second set of experiment with English dataset the FScore for: Unprocessed – 0.79, Stop words – 0.82, Stemming - 0.84 and complete – 0.95

Claims: According to author whole system was successfully implemented and the FScore values obtained proved the impending of the clustering by grouping to correctly interpret the informational content of the data.

Citation by other: There are no specific references to the work of Cernian et al. [2011] by other researchers in this survey.

4.2 Haider et al. 2007

Citation: Haider, P., Brefeld, U., and Scheffer, T. 2007. Supervised clustering of streaming data for email batch detection. *In Proceedings of the 24th International Conference on Machine Learning.* ICML '07. ACM, New York, NY, USA, 345-352.

Problem: According to author filtering spam more efficiently by exploiting the collective information about entire batched or group of jointly generated message

is one of the important parts of email management. Author addressed the problem of detecting batches in an email streaming that have been created according to the same template.

Previous Work: No previous work is mentioned by Haider et al. [2007]

Shortcoming of previous work: Not applicable

New Idea/Algorithm/Architecture: Author generates a model for detecting batches of emails into a well-defined problem setting of supervised learning. For the whole process author first derived a compact optimization problem based on the LP approximation to correlation clustering to learn the weights of the similarity measure, then they devised an efficient clustering algorithm with computational complexity linear in the number of emails and in-turn to complete this task they integrated method for learning the weight vector.

Experiments and analysis conducted: Author evaluated the performance and benefit of batch detection on an emails collection and also evaluated the method of identification of email batched for spam or non-spam email detection. These experiments are done on Enron corpus datasets. Firstly, author created an email corpus that reflects the features of an email stream. Secondly, the comparison is done of four strategies for clusters/batches identification: LP decoding, sequential decoding, agglomerative decoding and decoding strategies, using the similarity matrix obtained from pair wise learning. Thirdly, the evaluation of the classification of email as spam or non-spam is done.

Results: Author found that the final corpus contains 2,000 spam messages, 500 Enron messages, and 500 newsletters. Secondly, while finding the ideal batch information, the risk of misclassification is reduced by 43.8%, while with non-ideal batch information obtained through approximation clustering still 41.4% reduction are achieved.

Claims: Author devised a sequential clustering algorithm and two integrated formulations for learning a similarity measure to be used with correlation clustering. Author also claimed that a sequential clustering algorithm can efficiently make supervised batch detection at enterprise-level scale.

Citation by others: This work is referred by Haider et al. [2009]

4.3 Haider et al. 2009

Citation: Haider, P. and Scheffer, T 2009. Bayesian clustering for email campaign detection. *In Proceedings of the 26th Annual International Conference on Machine Learning.* ICML '09. ACM, New York, NY, USA, 385-392.

Problem: According to author there exist problems in clustering elements according to the sources that have generated them. For the independent binary attributes, a closed form of Bayesian solution exist but for dependent attributes that is based on a transformation of the instance was proposed by the authors.

Previous work: The author refer previous work by Haider et al. [2007]

Shortcoming of previous work: According to author the work of Haider et al. [2007] is not workable in practical where the effort of partitioning the data is much higher than the effort of labeling the labeling data for classification.

New Idea/Algorithm/Architecture: Author discussed the clustering of emails according to the sources that have generated using the Bayesian clustering algorithm. There are three main parts of algorithm: Firstly, they developed a model that produces a cluster of binary features vectors, based on a transformation of the input vectors. Secondly, generate an optimization problem and algorithm that produce the features transformation.

Experiments and analysis conducted: Author presented a large-scale case study that analyzes Bayesian clustering solution for email campaign detection.

Results: Author found a small fraction of spam messages, a total of 139,250 spam messages in correct chronological order. In order to maintain the users' privacy, authors blend the stream of spam messages with an additional stream of 41,016 non-spam messages from public sources. The non-spam portion contains newsletters and mailing lists in correct chronological order as well as Enron emails and personal mails from public corpora which are not necessarily in chronological order. Every email is represented by a binary vector of 1,911,517 attributes that indicate the presence or absence of a word. The feature transformation technique introduces an additional 101,147 attributes.

Claims: Author claimed that they devised a model for Bayesian clustering of binary features vectors based on Bayesian solution of the data likelihood in which the model parameters.

Citation by other: There are no specific references to the work of Haider et al. [2009] by other researchers in this survey.

4.4 Li et al. 2006

Citation: Li, H., Shen, D., Zhang, B., Chen, Z., and Yang, Q. 2006. Adding semantics to email clustering. *In Sixth International Conference on Data Mining.* ICDM '06. 938-942.

Problem: According to author email classification is a ways to manage emails but supervised classification needs a predefined taxonomy which requires user involvement and also after the development of clustering technique, it was also not possible to have satisfactory performance.

Previous Work: No previous work is mentioned by the author.

Shortcoming of previous work: Not applicable

New Idea/Algorithm/Architecture: Author proposes a model to automatically mine the semantic knowledge from the subject line of an email and create a cluster according to the similarity. In this method, each subject line is treated as a sentence and parsed through natural language processing techniques. The algorithm consists of four levels: 1. Generalization of terms in email subject line, the subject line parsing is done to create a syntactic tree using Microsoft NLPWin tool; 2. Mine

Generalized Sentence Pattern (GSP), patterns are generated from the generalized terms; 3. GSPs grouping and selection, GSPs in the same group will represent the same cluster; 4. GSP-PCL: GSP as pseudo class label.

Experiments and analysis conducted: The GSP-PCL clustering algorithm was experimented on two datasets: the open dataset Enron email dataset and a private email dataset collected by the author. In Enron email dataset, the minimum support threshold (min_sup) was set to 4 and the minimum length of GSPs was restricted to 2.

Results: When author compared GSP-means and K-means clustering on Enron email dataset and personal email dataset, the result showed that the readability is improved by 68.5%.

Claims: Author states that model suggested automatically extract embedded knowledge from the email subjects to help improve email clustering and GSP-PCL obtains significant improvement both on the clustering quality and cluster name readability compared with the basic K-means algorithm.

Citation by other: The work of Li et al. [2006] is cited by Yang et al. [2010].

4.5 Nagwani et al. 2010

Citation: Nagwani, N. and Bhansali, A. 2010. An object oriented email clustering model using weighted similarities between emails attributes. *International Journal of Research and Reviews in Computer Science* (IJRRCS) 1, 2, 1-6.

Problem: According to author it is possible to discover useful patterns from emails dataset which can further be used to manage the emails.

Previous work: The authors refer to previous work by Bird et al. [2006].

Shortcoming of previous work: The problem with the previous work was that it was not so accurate.

New Idea/Algorithm/Architecture: Author propose an automatic organization system which analyzes an inbox to recognize cluster of messages and put them in their corresponding folders. This system measures the weighted email attribute similarity between a pair of email objects like from-mail-id, to-mail-id, subject, message, sending time etc. using OSim (Object Similarity) distance function. The proposed method has three stages – 1. Pre-processing, it includes parsing, stemming and email representation technique for parsed information; 2. Weighted attributes similarity of Emails, it includes fetching the email attributes from processed database, then calculating the pair-wise attribute similarity of email document and finally assigning weights to the similarity measured for attribute pair-wise to calculate the overall similarity between a pair email document; and 3. Applying clustering technique over the measured similarity information to create email clusters.

Experiments and analysis conducted: Author tested their algorithm by experimenting with an inbox folder of "bass-e" from Enron email corpus datasets with Java as programming language and Simmetric & Weka as the other open source API's to support some functionality. Author also evaluated the accuracy of the proposed model by the 10-fold cross validation technique.

Results: Author state that the selected inbox folder consists of around 310 emails and total of eight clusters were generated from the given dataset by implementing this model and gives the similarity thresholds for the cluster as 0.05%. The evaluation of accuracy results around 78%.

Claims: Finally, author claim that the proposed model is implemented for discovering the email groups with good accuracy.

Citation by other: There are no specific references to the work of Nagwani and Bhansali [2010] by other researchers in this survey.

4.6 Schuff et al. 2006

Citation: Schuff, D., Turetken, O., and D'Arcy, J. 2006. A multi-attribute, multi-weight clustering approach to managing e-mail overload. *Decision Support Systems* 42, 3, 1350-1365.

Problem: The authors state that there is no efficient automated process exists to manage the e-mail overload, which will help users to manage hundreds of email automatically based on the content of a message. An efficient email management system can reduce the information overload and mental workload of a certain user.

Previous Work: The authors do not refer to any of my selected papers as their related work.

Shortcomings of previous work: No shortcomings of previous work were mentioned by the authors.

New Idea/Algorithm/Architecture: The authors propose a new multi weight, multi attribute clustering system that will automatically create folder structure in user's inbox based on the combination of email subject, sender, and receiver and text body. In their proposed system the user can set their desired weight to a particular attribute.

Experiments Conducted: The authors state that for evaluation their experimental subjects were daily emails of 65 students from an introductory computer literacy class. The data used analyzed using both multivariate and uni-variate analysis of variance models. To verify the appropriateness of multivariate, it is also verified that the assumptions of normality and homogeneity of error variance across groups were upheld.

Results: According to author the results of this research are potentially important for both academics and practitioners. For academics, this study integrates the concepts of semantic network theory and research on human memory chunking from cognitive psychology with prior information- science studies on textual document clustering. We extended this research to include clustering on key attributes of a textual document (in this case, attributes of an e-mail message). The ACEMS experiment has two implications for theory. First, while the application of a semantic network to an e-mail collection resulted in a nearly 41% improvement in task effectiveness, the additional increase from customizing the structure of the network was only marginally.

Claims: The authors claim that their proposed multi-weighted, multi-attribute method increase retrieval effectiveness reduces perceived effort and increase intension to use. They also claim that their system offers a general contribution in extending the application of semantic network theory.

Citation by other: The work of Schuff et al. [2006] is cited by Yang et al. [2007].

4.7 Shazmeen et al. 2011

Citation: Shazmeen, S. and Gyani, J. 2011. A novel approach for clustering e-mail users using pattern matching. *In Electronics Computer Technology (ICECT), 2011 3rd International Conference on.* Vol. 6. 205-209.

Problem: According to author emails have been considered as a useful resource for research in fields like link analysis, social network analysis and textual analysis and discovering useful patterns from emails can be useful for reducing the overload problem with today email inbox.

Previous Work: No previous work is referred by Shazmeen et al. [2011].

Shortcoming of previous work: Not applicable.

New Idea/Algorithm/Architecture: Shazmeen et al. [2011] propose a novel approach for clustering e-mail users using pattern matching email attributes such as sender email-id, receiver email-id Subject, message, sending-time, and attachments etc. clustering is used to discover email groups. The whole process is divided into different stages: In pre-processing phase all the email data is prepared for clustering. All the important email attributes are retrieved by parsing the email documents. Most of the attributes are of text data type, so stemming techniques are used to eliminate the unwanted texts from the parsed attribute information, after preprocessing clustering the users who are showing similarity in discussing the same context is clustered and graphically representing the cluster.

Experiments and analysis conducted: Shazmeen et al. [2011] algorithm is tested with Enron Email dataset.

Results: Experiment conducted by Shazmeen et al. [2011] showed that there are two clusters formed, "announcement" with cluster size 6 with threshold 1 and "conference" with cluster size 8 with threshold 1.

Claim: In this paper Shazmeen et al. [2011] claimed that an email clustering approach is proposed and implemented to show text similarities and that the proposed technique shows the email attributes and how the text similarities are used to cluster the users.

Citation by other: There are no specific references to the work of Shazmeen et al. [2011] by other researchers in this survey.

4.8 Xiang et al. 2007

Citation: Xiang, Y., Zhou, W., and Chen, J. 2007. Managing email overload with an automatic nonparametric clustering approach. In Network and Parallel Computing, K. Li, C. Jesshope, H. Jin, and J.-L. Gaudiot, Eds. Lecture Notes in Computer Science, vol. 4672. Springer Berlin / Heidelberg, 81-90.

Problem: According to author the email overload is a problem which user faces to process the large number of emails received/sent. As result it affects the usage or purpose of emails as effective knowledge management tool for communication.

Previous Work: Author mentioned the previous work of Schuff et al. [2006].

Shortcoming of previous work: According to the author work of Schuff et al. [2006] relies on the user involvement, i.e. they used techniques which is semi-supervised by user.

New Idea/Algorithm/Architecture: Author present an automatic email clustering system for automatic categorization of email into different meaningful groups by proposing a new automatic nonparametric clustering approach to manage email overload. The method works as: firstly, read the email messages from email client's data file, then it converts email texts into vector matrix and generate similarity matrix. Now once matrices are generated they are input into to the nonparametric text clustering algorithm. Then, the algorithm produces email clusters.

Experiments and analysis conducted: Author used email data sets are from real life email collections. The comparison is made with the results of the authors approach to the results of the k-mean algorithm and the hierarchical agglomerative algorithm. The quality is measured by Hubert's G statistic, simple matching coefficient, and Jaccard coefficient.

Results: Author result shows that for computational time analysis, hierarchical agglomerative algorithm takes 808% time more from the proposed algorithm to perform the clustering, and k-means algorithm takes 342% time more from the proposed algorithm to perform the clustering. For Hubert's G statistic is always higher than 0.764 when using the proposed algorithm which is mostly higher than Hubert's G statistic for other two algorithm. The Jaccard coefficient is found to be more than 0.821 for all data sets.

Claims: Author claim that email users get clustered emails easily without any input. The experiments shows that proposed algorithm has high efficiency and high clustering quality in terms of computation time and clustering quality.

Citation by other: There are no specific references to the work of Yang et al. [2007] by other researchers in this survey.

4.9 Yang et al. 2010

Citation: Yang, H., Luo, J., Yin, M., and Liu, Y. 2010. Automatically detecting personal topics by clustering emails. *In Second International Workshop on Education Technology and Computer Science*. Vol. 3. 91-94.

Problem: According to the author there are three problems in detecting topics by clustering. Firstly, choosing the method for text feature selection, Secondly, the way to combine the email subject and body features and lastly, since author use the k-mean clustering algorithm to cluster email therefore there is a problem in finding the value of k automatically and selecting the appropriate initial k kernels.

Previous Work: The authors refer to previous work by Li et al. [2006].

Shortcomings of previous work: No shortcomings of previous work were mentioned by the author.

New Idea/Algorithm/Architecture: Author propose a model to automatically detect the personal topic from the email inbox using a clustering algorithm. The approach is divided into three steps 1. Email representation with the EVSM (Email Vector Space Model); 2. Kernel selection algorithm based on lowest similarity; and 3. Email topic detection algorithm. The email representation with the EVSM is again split into three stages – Selection of body and subject features by selecting the n top-ranked high frequency words, Combine the body and subject of the email and Construction of the EVSM by applying the standard vector space model approaches.

Experiments and analysis conducted: Author did three experiments with four folders of the mini_newsgoups which is part of the data source 20NewsGroup. Experiment 1 consisted of implementing the standard k-mean algorithm. Secondly, implementing the proposed algorithm and lastly, clustering email by combining the body and subject fields with the proposed approach. Results: The results of implementation of the clustering algorithm are measured by author is in terms of F-value which 0.8584 for standard K-means and 0.9163 for improved K-means.

Claims: Authors claimed that the automatic detection of personal topic by clustering emails is successfully implemented and also they did some improvement on the construction of the EVSM and the kernel selection of the k-mean algorithm including the criteria of space and time complexity of the large-scale data processing.

Citation by other: There are no specific references to the work of Yang et al. [2010] by other researchers in this survey.

4.10 Yoo et al 2009

Citation: Yoo, S., Yang, Y., Lin, F., and Moon, I. 2009. Mining social networks for personalized email prioritization. *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '09. ACM, New York, NY, USA, 967-976.

Problem: According to author email overload creates problems for personal information management since it is a burden for user to process a large volume of email messages of differing importance; in turn it causes much negative effect on both personal and organization performance. The email overload can reduced by automatically prioritize received messages according to the priorities of each user called personalized email prioritization (PEP).

Previous work: No previous work is by author.

Shortcoming of previous work: Not applicable.

New Idea/Algorithm/Architecture: Author present a study with statistical classification and clustering methods addressing the PEP problem based on personal importance judgments by multiple users and also developed a novel transductive learning algorithm that propagates importance labels from training dataset to test dataset via message and user nodes in a personal email network. Firstly, a user as a

member of a group is chosen based on unsupervised clustering, and then inference is made on the importance of that particular user from other group members. Later these clusters can be used by SVM classifier as input features to each message.

Experiments and analysis conducted: Author engaged 25 experimental subjects where each subject was requested to label at least 400 non-spam messages during a one-month period. The five importance levels are: absolutely non-important, relatively non-important, neutral, important, and most important.

Results: Author found the following results: Firstly, below are the performance curves of SVM runs with different representation schemes for email messages. Secondly, the author claims to have obtained significant performance improvement over the baseline system (without induced social features) in our experiments on a multiuser data collection: the relative error reduction in MAE was 31% in micro-averaging, and 14% in macro-averaging.

Claims: Author claim that the algorithm they designed is successfully implemented for fulfilling the purpose of the system with less error rate.

Citation by other: There are no specific references to the work of Yoo et al. [2009] by other researchers in this survey.

5. REFERENCES

Aery, M. and Chakravarthy, S. 2004. eMailSift: mining-based approaches to email classification. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '04. ACM, New York, USA, 580-581.

Bird, C., Gourley, A., Devanbu, P., Gertz, M. and Swaminathan, A. 2006. Mining email social networks. In *Proceedings of the 2006 International Workshop on Mining software Repositories.* MSR '06. ACM, New York, USA, 137-143.

Cernian, A., Florea, I., Carstoiu, D., and Sgarciu, V. 2011. The design and validation of an automatic email clustering system based on semantics. *In Intelligent Data Acquisition and Advanced Computing Systems (IDAACS).* 2011 IEEE 6th International Conference on. Vol. 2. 629-632.

Cselle, G., Albrecht, K. and Wattenhofer, R. 2007. BuzzTrack: topic detection and tracking in email. In *Proceedings of the 12th International Conference on Intelligent User Interfaces.* IUI '07. ACM, New York, USA, 190-197.

Haider, P., Brefeld, U., and Scheffer, T. 2007. Supervised clustering of streaming data for email batch detection. *In Proceedings of the 24th International Conference on Machine Learning.* ICML '07. ACM, New York, NY, USA, 345-352.

Haider, P. and Scheffer, T 2009. Bayesian clustering for email campaign detection. *In Proceedings of the 26th Annual International Conference on Machine Learning.* ICML '09. ACM, New York, NY, USA, 385-392.

Ho, V., Wobcke, W. and Compton, P. 2003. EMMA: an e-mail management assistant. In *Intelligent Agent Technology, 2003. IAT 2003. IEEE/WIC International Conference on.* 67-74.

Kushmerick, N. and Lau, T. 2005. Automated email activity management: an unsupervised learning approach. In *Proceedings of the 10th International Conference on Intelligent User Interfaces.* IUI '05. ACM, New York, USA, 67-74.

Li, H., Shen, D., Zhang, B., Chen, Z., and Yang, Q. 2006. Adding semantics to email clustering. *In Sixth International Conference on Data Mining.* ICDM '06. 938-942.

Li, W., Zhong, N., Yao, Y. and Liu, J. 2009. An Operable Email Based Intelligent Personal Assistant. In *World Wide Web 12.* 125-147.

Mock, K. 2001. An experimental framework for email categorization and management. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '01. ACM, New York, USA, 392-393.

Nagwani, N. and Bhansali, A. 2010. An object oriented email clustering model using weighted similarities between emails attributes. *International Journal of Research and Reviews in Computer Science* (IJRRCS) 1, 2, 1-6.

Schuff, D., Turetken, O., and D'Arcy, J. 2006. A multi-attribute, multi-weight clustering approach to managing e-mail overload. *Decision Support Systems* 42, 3, 1350-1365.

Shazmeen, S. and Gyani, J. 2011. A novel approach for clustering e-mail users using pattern matching. *In Electronics Computer Technology (ICECT), 2011 3rd International Conference on.* Vol. 6. 205-209.

Stolfo, S. J., Hershkop, S., Wang, K., Nimeskern, O. and Hu, C. 2003. Behavior profiling of email. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics.* ISI'03. 74-90.

Tang, J., Li, H., Cao, Y. and Tang, Z. 2005. Email data cleaning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining.* KDD '05. ACM, New York, USA, 489-498.

Whittaker, S. and Sidner, C. 1996. Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground.* CHI '96. ACM, New York, USA, 276-283.

Xiang, Y., Zhou, W., and Chen, J. 2007. Managing email overload with an automatic nonparametric clustering approach. In Network and Parallel Computing, K. Li, C. Jesshope, H. Jin, and J.-L. Gaudiot, Eds. Lecture Notes in Computer Science, vol. 4672. Springer Berlin / Heidelberg, 81-90.

Yang, H., Luo, J., Yin, M., and Liu, Y. 2010. Automatically detecting personal topics by clustering emails. *In Second International Workshop on Education Technology and Computer Science.* Vol. 3. 91-94.

Yoo, S., Yang, Y., Lin, F., and Moon, I. 2009. Mining social networks for personalized email prioritization. *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* KDD '09. ACM, New York, NY, USA, 967-976.