International Journal of Web-Based Learning and Teaching Technologies
Volume 15 • Issue 2 • April-June 2020

ML-EC^2: An Algorithm for Multi-Label Email Classification Using Clustering

Aakanksha Sharaff, National Institute of Technology Raipur, Raipur, India
Naresh Kumar Nagwani, National Institute of Technology Raipur, Raipur, India

ABSTRACT

A multi-label variant of email classification named ML-EC^2 (multi-label email classification using clustering) has been proposed in this work. ML-EC^2 is a hybrid algorithm based on text clustering, text classification, frequent-term calculation (based on latent dirichlet allocation), and taxonomic term-mapping technique. It is an example of classification using text clustering technique. It studies the problem where each email cluster represents a single class label while it is associated with set of cluster labels. It is multi-label text-clustering-based classification algorithm in which an email cluster can be mapped to more than one email category when cluster label matches with more than one category term. The algorithm will be helpful when there is a vague idea of topic. The performance parameters Entropy and Davies-Bouldin Index are used to evaluate the designed algorithm.

KEYWORDS

Classification Using Clustering, Email Clustering, Latent Dirichlet Allocation, Multi Label Classification, Non-Negative Matrix Factorization, Taxonomic Terms

INTRODUCTION

As the number of incoming email messages increases, it becomes very difficult for the users to handle these emails. There are different tools for facilitating the management of incoming emails. e.g. use of threads and use of folders or labels for classifying incoming emails. Email categorization (classification) is a process of classifying emails to discrete set of predefined categories. Categorization of emails becomes difficult due to the enormous volume of emails (sent/received) as well as different topics may be discussed in an email. Hence, categorizing emails manually becomes a heavy burden for users. Categorizing emails by identifying categorical terms is an important issue. It adds semantics to email management. Multi label email classification is not explored in detail in literature.

The objective of this paper is to detect similar emails and categorize them in multi label classes as well as to identify (discover) categorical terms in a different way by adapting latent Dirichlet allocation (LDA) as topic modelling approach. Hence, to accomplish the objectives; an algorithm Multi Label Email Classification using Clustering (ML-EC^2) is proposed in this paper. It is a type of multiple classification of emails. Classifying emails into classes can be topic oriented or group oriented. Topic

DOI: 10.4018/IJWLTT.2020040102

oriented classification includes the emails belonging to such as "job opportunities", "entertainment" etc., whereas group oriented classification can be "place specific", "people specific", "course or project specific". In multi label classification of emails, each email file may belong to one or more number of categories. The algorithm ML-EC^2 has been designed for creating the categorized groups of similar emails using textual similarity of email attribute. It is a multi-label text clustering based classification algorithm, where one email cluster can be mapped to more than one email category. If in a single label there are 1500 emails on the same topic suppose on entertainment then it becomes very difficult and time consuming to find a desired email. Hence to overcome this problem a multi class categorization of email has been designed and implemented. A hierarchy is formed with a single label/class. For example, in entertainment class a hierarchy of music, videos, movies etc. can be formed and the email associated with concerned label (class) can be placed on these sub-hierarchies. The proposed technique of email categorization can reduce the problem of email overload.

LITERATURE REVIEW

Managing huge amount of emails received from users is a very challenging problem which needs to be solved in an effective and efficient way. Various researches have been done in the field of email mining. Some of the surveys done are as follows.

Park & An (2010) proposed an Email multicategory classification approach using semantic features and a dynamic category hierarchy reconstruction method in which the user reorganizes all e-mail messages into categories. Guan & Yuan (2013) reviews the existing work on mislabeled data detection techniques for pattern classification and classifies them into three types: Local learning-based, ensemble learning-based and single learning-based methods. The author Armentano & Amandi (2014) presented an approach to label the incoming emails based on user preference; a set of experiments using Google's webmail system, Gmail is performed to obtain a good rate of acceptance of the agent interactions. Alsmadi & Alhami (2015) introduced an algorithm for performing clustering and classification of email text corpus. They have proposed a model for classification of emails based on subject and folder using N-grams. Islam et al. (2009) proposed a new technique of e-mail classification based on the analysis of grey list (GL), which uses multi-classifier classification ensembles of statistical learning algorithms.

Sakurai & Suyama (2005) proposed a method to extract key concepts from e-mails and presents their statistical information which has been applied to three kinds of analysis tasks: a product analysis task, a contents analysis task, and an address analysis task in which acquired concept relation dictionaries gave high precision ratios in the classification. Koprinska et al. (2007) investigated the use of random forest for automatic e-mail filing into folders and spam e-mail filtering. Sappelli et al. (2005) presented an approach of categorizing emails that can alleviate the common problem of email overload. Sun et al. (2010) developed a clustering based algorithm for detecting duplicate emails by using hash function. Gomez et al. (2012) classified emails into spam and ham by reducing the dimensionality of email using Principal Component Analysis (PCA) and compare several feature selection methods with novel content-based statistical feature extraction techniques. Recently a Singular Value Decomposition (SVD) method has been proposed by Zareapoor et al. (2015) to classify email in order to compress sparse email data but retaining the most informative and discriminate features of email. Aloui & Neji (2010) developed a multi-agents system EQASTO (E-mails Question Answering System using Text-mining and Ontological techniques) to relieve the burden of e-mails processing by using a combination of text-mining and ontological techniques to classify semantically e-mails, fetch, generate, and send answers automatically to learners. Bekkarman et al. (2004) presented an email foldering scheme by using two large corpora i.e. Enron and SRI and point out the challenges arises by using email foldering scheme instead of traditional document classification. The author Beseiso et al. (2012) proposed an ontology based email knowledge extraction process which reduces the users time and resources to handle unstructured Email messages.