Taxonomy for Email Classification and Summarization Techniques

Ahmed El Sharkasy Ahmed Kotb Amr Sharaf
Mohammad Kotb Moustafa Mahmoud

January 25, 2012

Contents

1 Abstract **3**

2 Introduction **3**

3 Email Classification Taxonomy **3**

4 Email Summarization Taxonomy **4**

5 Papers Summary **4**

5.1 Email Classification 4

5.1.1 Automatic Categorization of Email into Folders 4

5.1.2 Email Classifications For Contact Centers 7

5.1.3 Using GNUsmail to Compare Data Stream Mining Methods for On-line Email 8

5.1.4 E-Classifier: A Bi-Lingual Email Classification System 9

5.1.5 An Object Oriented Email Clustering Model Using Weighted Similarities between Emails Attributes 10

5.1.6 Content Based Email Classification System by applying Conceptual Maps 10

5.1.7 A new approach to Email classification using Concept Vector Space Model 10

5.2 Email Summarization 10

5.2.1 Detection of question-answer pairs in email conversations (2004, 41 citations) 10

5.2.2 Using Question-Answer Pairs in Extractive Summarization of Email Conversations 10

6 Results **10**

7 Conclusions **10**

1 Abstract

In this document we present a survey and taxonomy on recent research topics related to email classification and summarization. This document summarizes and organizes recent research results in the novel way that integrates and adds understanding to work in the field of email classification and summarization. It emphasizes the classification of the existing literature, developing a perspective on the area, and evaluating different trends.

Keywords Email, Classification, Summarization, Machine Learning.

2 Introduction

Email has been an efficient and popular communication mechanism as the number of Internet users increases. Therefore, email management has become an important and growing problem for individuals and organizations because it is prone to misuse. One of the problems that are most paramount is disordered email message, congested and unstructured emails in mail boxes. It may be very hard to find archived email message, search for previous emails with specified contents or features when the mails are not well structured and organized.

Many machine learning approaches have been applied in this field, the most State-of-the-Art algorithms in email classification include: support vector machines, neural network, nave bayes classifiers and entropy-based approach.

Email summarization is another important and challenging problem. We can think of automatic summarization as a type of information compression. To achieve such compression, better modelling and understanding of document structures and internal relations is required.

In this document we present a survey and taxonomy on recent research topics related to email classification and summarization.

3 Email Classification Taxonomy

The following table classifies some recent research papers in the field of email classification according to the different learning algorithms used in different papers

Learning Algorithm					
SVM	Nave Bayes	Neural Networks	Max. Entropy / Winnow	Nnge / Hoeffing Trees	Graph Mining
An Innovative Analyser for email classification Based on Grey List Analysis	Email Classification with Co-training	Email Classification: Solution with Back Propagation Technique	Automatic Categorization of Emails into Folders	Using GNUsmail to compare Data Stream Mining Methods for On-line Email Classification	A graph Based Approach for Multi-Folder Email Classification
Email Classification with Co-training	Automatic Categorization of Emails into Folders	Email Classification Using Semantic Feature Space			
Automatic Categorization of Emails into Folders					

4 Email Summarization Taxonomy

5 Papers Summary

5.1 Email Classification

5.1.1 Automatic Categorization of Email into Folders

Date 2004

Citations 112

Introduction

- users get alot of emails this days, not just spam but a large number of legitmate emails also that they need to process in a short time;
- the paper shows the results of an extensive benchmark on two large corpora (enron,sri) of 4 classification algorithms;
- the paper shows an enhancement to the exponential gradient method (winnow).

Related Work

- Clark and Niblet 1989: proposed a rule inductive algorithm CN2 and showed that it can outperform KNN;
- Cohen 1996: proposed the RIPPER classifier and showed that that it can outperfrom an tfidf classifier;
- Provost 1999: showed that Naive bayes can outperform RIPPER;
- Remmie 2000: achived a very high accuracy by classifying mails to 3 predefined folders ;
- Kiritchenko and Malwin 2001: showed that SVM can outperfom Naive Bayes.

Algorithms Benchmarked

- Maximum Entropy;
- Naive Bayes;
- SVM;
- Winnow (enhanced version).

Challenges in mail classification

- Email users often create folders and let it fall out of use (small no. of training data per folder);
- folders dont necessarily correspond to simple semantic topics (unfinished todos , project groups, certain recipient);
- differ drastically from one user to another;
- Email arrives in a stream over time which causes more difficulties , for example the topic of main folder can drift over time.

Data set pre-processing removing non topical folders (Inbox , sent , trash ,) removing small folders (folders that has a small number of emails)

Training/test set splits The paper shows a new way to split training data into training set and test sit , the new method takes time factor into considerations .. it works as follows sorting emails by time train the classifier for the first N emails then test it on the following N emails Train the classifier for the first 2N emails then test it for the following N emails and so on

Features Extraction traditional bag of words representation.

Datasets Enron : http://www.cs.cmu.edu/ enron/ 150 users with more than 500,000 Emails Applied to the following 7 employees folder only (the largest 7) beck-s,farmer-d, kaminski-v, kitchen-l, lokay-m, sanders-r and williams-w3. Removed the non topical folders like all documents, calendar, contacts, deleted items, discussion threads, inbox, notes inbox, sent, sent items and sent mail. Flatten all the folder hierarchies. Removed folders with less than 3 messages Removed the X-Folder field from email messages. (The X-Folder field contains the class label) SRI : http://www.ai.sri.com/project/CALO Applied to the following 7 folders only : acheyer, bmark, disrael, mgervasio, mgondek, rperrault ,and vchaudri Removed the non topical folders(inbox , draft, sent , trash). Flatten all the folder hierarchies. Removed folders with less than 3 messages.

Critique They didnt use Stemming in their preprocessing to the dataset not including precision , recall and f1 score for accuracy measures.

Conclusion Naive Bayes is inferior to other algorithms. SVM achieved the highest accuracies in most of the tests.

Future Work Different sections of each email can be treated differently. For example, the system could create distinct features for words appearing in the header, body, signature, attachments, etc Named entities may be highly relevant features. It would be desirable to incorporate a named entity extractor (such as MinorThird3 , see, e.g., Cohen and Sarawagi (2004)) into the foldering system.

5.1.2 Email Classifications For Contact Centers

Date 2003

Citations 14

Main Topic Proposing an automatic system to classify mail message for contact centers mails are categorized into 2 classes Single messages : messages that dont require a response Root messages : messages that require immediate response those message can be sub divided into 3 classes Root : the start of the communication (contains a problem or a question) inner : communication on a certain problem leaf : marks the end of this interaction (eg. the problem was solved)

Tools used Rainbow : an implementation for naive bayess algorithm svm-light : an implementation for SVM algorithm wordNet : used for parts of speach taging Ltchunk : used to identify noun phrases and count number of sentences in email.

Dataset Pine-info discussion list web archive http://www.washington.edu/pine/pine-info

Pre processing removing reply blocks (blocks from previous emails in the current mail) removing signature blocks

Features (for SVM algorithm) Non-infected words nouns,verbs,adjective,adverb using wordNet Noun phrases using Ltchunk Verb phrases Punctuation letters count assuming that single message will have different punctuation than letters count Length of email (number of sentences) using Ltchunk Dictionary 2 dictionaries were made one for the most common words in single messages and the other for the most common words in root message

Conclusion High accuracy was achieved on root vs leaf (92

5.1.3 Using GNUsmail to Compare Data Stream Mining Methods for On-line Email

Date 2011

Citations 0

Main Topic Introducing GNUsmail , an open source framework used for mail classification , focusing on online incremental learning. proposing new techniques for testing other than holdout and cross-validation like prequential measure.

Evaluation methods prequential measure. sliding and fading windows. McNemar test.

Dataset Enron a layer was added to feed the learning algorithm the new emails one by one, simulating new incoming emails. Algorithms: OzaBag over NNge, using DDM for concept drift detection. NNge. Hoeffding Trees. Majority class.

Tools GNUsmail http://code.google.com/p/gnusmail/

Result Improved GNUsmail by incorporating new different methods to evaluate data stream mining algorithms in the domain of email classification.

Future Work Current online learning algorithm implementations have an important limitation that affects the learning process: learning attributes have to be fixed before beginning the induction of the algorithm. They need to know all the attributes, values and classes before the learning itself, since it is not possible to start using a new attribute in the middle of the lifetime of a learning model. Future methods should support online addition of new features

5.1.4 E-Classifier: A Bi-Lingual Email Classification System

Date 2008

Citations 0

Problem classifying Arabic and English emails. implementing an outlook add-in e-classifier,

Related Work English Email Classifiers PopFile: http://popfile.sourceforge.net uses naive bayes algorithm only. SpamBayes http://spambayes.sourceforge.net Binary Classifier (Spam or not) Arabic Email Classifiers there are no Email classification work on arabic language, the related work are on arabic documents not emails El-Kourdiet

Dataset English : enron dataset Arabic : translated documents that have been converted to emails document obtained from http://www.comp.leeds.ac.uk/eric/latifa/rese

pre-processing English removing stop words removing punctuation marks converting all the letters to lowercase porter stemmer Arabic no root extraction technique was used due to the lack of non commercial product.

Results 85 percent of English emails were classified correctly. 60 percent of English emails were classified correctly.

Critique used only overall accuracy measure which might not good indicator in case of skewed data.

5.1.5 An Object Oriented Email Clustering Model Using Weighted Similarities between Emails Attributes

5.1.6 Content Based Email Classification System by applying Conceptual Maps

5.1.7 A new approach to Email classification using Concept Vector Space Model

5.2 Email Summarization

5.2.1 Detection of question-answer pairs in email conversations (2004, 41 citations)

5.2.2 Using Question-Answer Pairs in Extractive Summarization of Email Conversations

6 Results

In this section we describe the results.

7 Conclusions

We worked hard, and achieved very little.