1. How Many Folders Do You Really Need?: Classifying Email into a Handful of Categories

Aim

The authors propose an approach for automatically distinguishing between human and machine-generated emails as well as classifying emails without requiring the user to predefine any folder

Methodology

Feature Extraction

1. Content Extraction - Analyze the body and subject of the mail. They also eliminate top 400 words from the email (in at a the).

2. Address Extraction - Analyze the email's subdomain (.edu), subname (billing noreply), commercial words (flight, ticket, career, shipping, etc)

3. Behavior features - features extracted from recipient and sender actions over the message.

Classification (USES logistic regression)

Offline System - consist of two classifiers for training messages and sender information

1. Sender Table Classifier groups senders into tables
2. Message-level Classifier groups messages into folders

Online System - consist of three classifiers for categorizing incoming messages

Online lightweight Classifier - classifier uses a selection of rules based on the top 100 senders to quickly classify emails to deal with traffic

If this fails Online sender-based Classifier which uses all the entire sender table to try and classifier the mail

Should this also fail the final classifier online heavyweight classifier which uses all features to categorize the email

Results

After creating the systems, the authors tested the system, and it achieved precision and recall rates close to 90%.

2. Email classification and summarization: A machine learning approach

Aim

Classify email based on user activities

Generate a summary of the email

Classification

machine learning algorithm classifies the emails based on a set of rules and using the level of similarity in the email content and the subject.

Figure 4: **Classification Algorithm**

```
For every message M
Let FW = N most frequent words in the message
        Iterate over all activities and for each
activity AC
                Let AFW = common words in
        activity AC
                If (FW = AFW then
mark the activity AC
update the message activity as AC
create a rule that states // machine learning
        for each message received that has some
words like FW
        AC is the activity for this message
 Else create a new activity
End
```

The summarization algorithm selects sentences that have the most frequent word and arrange them in a logical order to make the email message summary.

Results

The authors achieved an accuracy of 90% for their classifier and “a recall score between 50% and 100% difference in the qualities of the email summaries while the precision score is between 20% and 70%”

3. Email classification for contact centers

Aim

The authors build a classifier which consists of two modules.

The first classifier categorizes message into single and root messages.

Single messages - do not require an require a response

root message - do not require an immediate response

The second module classifiers message into root, leaf and inner messages.

root - those that start a thread

leaf - end a thread

inner - message in between

Feature Extraction

The authors clean the data by initial cleaning the data by identifying deleting signature blocks and quotes passages from emails. They further use a tool to identify nouns and verb phrases in the data.

They also remove common words (the, in, a) from the messages.

Classification (uses support vector machines and Naive Bayes)

to group the message.

Results

The system achieved an accuracy of 90%.