King Saud University

Journal of King Saud University –
Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

Clustering and classification of email contents

Izzat Alsmadi [a,*], Ikdam Alhami [b]

[a] Department of Computer Science, Boise State University, USA
[b] Yarmouk University, Jordan

Received 3 September 2013; revised 12 February 2014; accepted 13 March 2014
Available online 7 January 2015

KEYWORDS

Emails classification;
Document similarity;
Document classification;
Feature extraction;
Subject classification;
Content classification

Abstract Information users depend heavily on emails' system as one of the major sources of communication. Its importance and usage are continuously growing despite the evolution of mobile applications, social networks, etc. Emails are used on both the personal and professional levels. They can be considered as official documents in communication among users. Emails' data mining and analysis can be conducted for several purposes such as: Spam detection and classification, subject classification, etc. In this paper, a large set of personal emails is used for the purpose of folder and subject classifications. Algorithms are developed to perform clustering and classification for this large text collection. Classification based on NGram is shown to be the best for such large text collection especially as text is Bi-language (i.e. with English and Arabic content).

1. Introduction

Emails are used by most humans on earth. It is estimated that there are more than 3 billion email accounts of almost half of the world population. They are expected to reach 4 billion by the year 2015 (Email Statistics Report, 2011). Even kids are allowed under certain conditions to have email accounts supervised by parents.

Spam in emails is one of the most complex problems in email services. Spam emails are those unwanted, unsolicited emails that are not intended for specific receiver and that are sent for either marketing purposes, or for scam, hoaxes, etc. It is estimated that in 2009 more than 97% of emails were classified as spam (Elements of Computer Security, 2010). This is why many research papers which studied or analyzed emails focused on this aspect (i.e. the classification of emails into spam or not). However, the struggle between spammers and spam detection tools is continuous where each side is trying to create new ways to overcome the techniques developed by the other.

Some local papers that conducted spam assessment (e.g. Abdullah Al-Kadhi, 2011 paper) showed that the problem is serious. Authors conducted surveys to assess the current status of Spam distribution in KSA. Authors tried also to summarize major reasons of spreading of spam messages and emails including: Sexual contents, commercials, phishing, religious reasons, etc. Of course major disadvantage of spam spread is the overconsumption and bandwidth and resources for no good purposes.

[*] Corresponding author.
E-mail address: izzatalsmadi@boisestate.edu (I. Alsmadi).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.03.014
1319-1578

In this aspect, an email spam-based classifier is not only expected to accurately classify spam emails as spams, but also expected to classify non-spam emails as non-spam or normal. This is since both are considered conditions for evaluating the quality of its classification or prediction. Four prediction metrics are used then to evaluate the quality of email prediction. True Positive (TP) indicates that the spam detection tool predicts that the email is spam and truly it was a spam. True Negative (TN) indicates that the tool or the email system predicts that the email is normal and not spam and correctly it was so. False Positive (FP) indicates that by mistake the tool predicts that a good email is spam (aka false alarms). Last, False Negative (FN) indicates also another mistake where it is predicted that a spam email is normal. As such, a perfect detection system should have the values: TP 100%, TN 100%, FP 0%, and FN 0%. In reality such perfect situation is impossible and impractical. TP and FP complement each other for 100% (i.e. their total should be 100%). Same thing is applied for TN and FN.

The challenge of some email detection systems is that if it is restricted through many spam-detection roles, TP may go high, but at the account of getting many false alarms. On the opposite very lean rules may get very high TN but at the account of FN.

Another challenge in emails' spam detection is speed. In security, speed or performance is always in a trade off with security where too many roles may slow down the system.

In addition to spam based classification, papers that conducted research in emails discussed other aspects such as: Automatic subject or folder classification, priority based filtering of email messages, emails and contacts clustering, etc. Some papers evaluated replies in emails to classify emails on different threads. Currently some email servers such as Gmail combine email together if they came as a reply.

Following are some of the focuses in the research of email analysis (Based on our review of papers related to research papers in data mining in emails' datasets):

1. Generally, email analysis can be classified under text categorization in its most activities. Algorithms such as: VSM, KNN, Ripper, Maximum Entropy (MaxEnt), Winnow, ANN are examples of algorithms used in email analysis.
2. A major research subject in email classification is to classify emails into spam or no spam emails. This can be further used for the real time prediction of spam emails.
3. Some email classification research papers tried to classify emails based on the gender of the sender given some of the common aspects that may distinguish emails from females or males.
4. Email classification can be also used to automatically assign emails to predefined folders.
5. Rather than spam and non spam emails, emails can be also classified into: Interesting and uninteresting emails.
6. Features are extracted from the email content or body, title or subject or some of the other Meta data that can be extracted from the emails such as: sender, receiver, BCC, date of sending, receiving, number of receivers, etc. The method to extract feature can be based on words, bags of words, etc.
7. Email clustering is also considered to cluster emails into different subjects or folders.
8. The time information in emails (e.g. when: sent, received, etc.) is used also in some research papers to classify emails.
9. Some research papers tried to classify emails based on similar threads or subjects. Some email systems such as Gmail connect emails related to each other (e.g. by reply or forward events) together.

In this paper, a personal email archive of more than 19,000 normal messages is used for analysis and evaluation. The focus is to study the email content and address and classify each email into one of three: Personal, professional and other based on sender, content and header.

The rest of the paper is organized as the following: Section two presents several research papers in email analysis. Section three presents goals and approaches. Section four presents experiment and analysis and paper is concluded with conclusion section.

2. Related work

As mentioned earlier, collecting an archive of emails for analysis can be done for several purposes. One of the major goals is spam detection. This sub section describes some research papers related to spam email classification.

2.1. Spam–non-spam email classification

We selected some papers, based on citation, related to spam detection or filtering. Those papers are: Zhuang et al., 2008; Blanzieri and Bryl, 2008; Webb et al., 2006; Mishne et al., 2005; Sculley and Wachman, 2007; Zhou et al., 2010; Pérez-Díaz et al., 2012; Xie et al., 2006; Katakis et al. 2007; Bogawar et al. 2012; Ozcaglar 2008. Different papers discussed the using of different algorithms and also applying the algorithms in different places between email senders and receivers.

Zhuang et al.'s (2008) paper focused on trying to find Botnets. Botnets are groups responsible for spreading spam emails. Methods are evaluated to detect such sources of spam campaigns that share some common features. Spammers however try to change spam emails through some intended mistakes or obfuscations especially in popular filtered keywords. Certain finger prints are defined where all emails that have those finger prints are then clustered together.

Blanzieri and Bryl (2008) presented a technical report in 2008 to survey learning algorithms for spam filtering. The paper discussed several aspects related to spam filtering such as the proposals to change or modify email transmission protocols to include techniques to eliminate or reduce spams. Some methods focused only on content while others combined header or subject with content. Some other email characters such as size, attachments, to, from, etc. were also considered in some cases. Feature extraction methods were also used for both email content, attached and embedded images.

Webb et al.'s (2006) paper talked about web spam and how to use email spam detection techniques to detect spam web pages. Similar to the approaches to detect spam in emails, web pages are scanned for specific features that may classify them as spam pages such as using irrelevant popular words, keywords stuffing, etc. Mishne et al.'s (2005) paper represents another example of web or link spam research paper. Blogs, social networks, news or even e-commerce websites now allow