A Review on Email Spam Filtering*

IN4325 Information Retrieval

Nivedita Prasad†
Student ID: 4712099 TU Delft
Delft, Netherlands
N.Prasad@student.tudelft.nl

ABSTRACT

Email Spam filters are present to prevent the delivery of Spam email. The field of Information retrieval has shown profound interest in this field. This paper aims to review Information Retrieval methods that can be used to build a Email Spam filter. This is a special case of Information Filtering. The objective is to review and summarize various Information Retrieval methods used by the research papers related to Information Retrieval based Email Spam filtering.

KEYWORDS

Information Filtering, Information Retrieval Methods, Spam Email Filtering

1 INTRODUCTION

A variety of Spam emails containing distorted images, audio, videos, plain text, url, word documents, pdfs attracts and convinces the receivers to believe and open the email sent. The characteristic of Email-Spam is *Unwanted*, *Indiscriminate*, *Disingenuous* and *backscatter*. The Email Spam Filter according to [4] is a *Passive Information Filtering system* and can be generalized or personalized based on the user preference. Finally the new incoming emails are classified as Spam and Ham (not Spam), with less false positives in the Spam folder and less False negatives in the Ham folder. The *core* of any Email Spam Filter is a classifier that classifies the Email as Spam or ham. The authors of [3], has stated that Spam Filtering is important to stop the delivery and action taken for unsolicited Spam emails. The actual information existence in the email will help the Spam filter in its purpose of identifying Spam. The popular *IR models* usually used for designing a email spam filter are Boolean model, Vector Spaced Model using term frequencies, Probabilistic modes, Axiomatic frame-work, Uni-gram model, Language model, Neural IR and ripper algorithm.

Fig 1, shows a basic Information Retrieval model which explains that the *information need* is converted into a query by the user and the relevant document will be fetched after ranking by the retrieval system. The tabulation section at the end of this survey will provide a compact summary of all the papers reviewed in this work. The rest of the survey are sections on Features of an Email, detailed documentation on Review on Email Spam Filtering and Fig 3 is a summary of all the research papers reviewed in this paper for clear understanding. The columns used are authors, title, aim , Spam Filter Type, Information Retrieval model, Dataset and finally performance of each research methodology based on the evaluation process adopted followed by Conclusion.

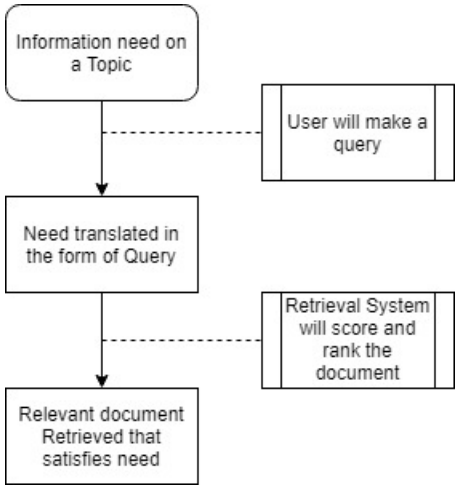

Figure 1: Information Retrieval Model

2 DEFINITION OF E-MAIL SPAM FILTERING

In this section, Fig 2 shows how a mail from a sender is being filtered by the email Spam filter and then classifies the email as Spam and Ham .The Ham which is classified as Spam is learned by the Spam filter and then the filter is refined. Cormack et al of [3] has explained that in the Spam filter, the incoming mail is processed one at a time and classified as ham (non-Spam) and Spam. The ham is classified into the Inbox which is regularly read by the user and the Spam email box is searched for any misclassified email and is reported to the Spam filter. The filter keeps an account of the misclassified emails and applies this information on the next incoming email. There are variants of the email Spam filtering deployment as cited in the same paper. They are Batch filtering, Batch Training, just-in-time, Deferred, receiver or sender engagement filters and Collaborative filtering.

*
†

,

,

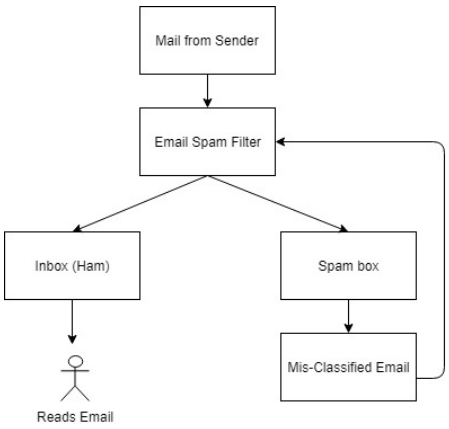

Figure 2: Email Spam Filter

2.1 Email Features contributing to IR models

The main features that the Information Retrieval models look for as features in the Email received, to retrieve relevant Spam email is explained in this section. The features for Email Spam are retrieved from the *header or body of the email* which in-turn is put into a vector called the feature vector. The process of extracting features after defining them is called *Feature Engineering*. This improves the Spam Filter's efficiency. We can form feature vectors using *bag of words* model where the distinct words are given numerical values and made into a feature vector, after which it is used for calculating the number of times a particular word appears in a *term frequency model* or to divide the number of documents that contain the term (td-idf model). The other way to indicate features are *Synthetic Words* where a particular word's presence in the head or body is checked. Word bi-grams, tri-grams and meta features selection is also possible .The next step is feature reduction for Information Retrieval models. The techniques are: *Stop word elimination* which is used in bag of words model where the common words are not given numerical value, then *Stemming* where the root word of the different variants of the word (dependent on the language) is obtained. Almost all Email Spam filtering techniques use variants of Bag of word model to identify a term and then apply it to their classifier. The other Email filtering techniques used are Human Classifier, Ad-Hoc, Rule Based Filtering, White lists, Black lists and Grey-listing.

3 REVIEW ON EMAIL SPAM FILTERING

The core idea of this review is to understand how each IR model uses the features it has been provided with to correctly classify Ham and Spam. There are many IR methods as mentioned in the section 1 to filter Email Spam which have been implemented in the below discussed research papers.

Probabilistic Models. An interesting and recent paper by Liu et al. [7] performs Email Spam filtering using On-line active multi-field Learning. The IR model used is a *Probabilistic Model using Bayesian conditional probability* that is making use of trec7 dataset. The Spam filter proposed by the authors has used lightweight String-Frequency Index Text Classification algorithm which uses the string-frequency index data structure. This stores the *feature string* that the authors use (overlapping word-level 4-gram model) and converts the on-line "training and classifying process" into an index increment *updating and retrieving/predicting* process. This solves the text feature space which is really large and the Spam score of the particular email can be easily calculated which is the arithmetical average of all the Spam scores. This was a good approach to classify email using probabilistic models .

Similar to [7], a Probabilistic IR Model by the authors Morteza Zi Hayat et al. of [5] have implemented Content-based concept drift detection for email Spam filtering Language Model. This detects concept drift based on the deviation of email contents. The concept based language model represents the difference between two probability distributions which is measured by KL divergence measure. If the difference is higher that the determined threshold then it is considered as a concept drift. If there is a concept drift then the Language models will have to adapt to classify the new incoming emails by updating itself. The updating is done for the term probabilities by combining them with the previous block when the older block has more influence on the new email, then the language models are updated if the old language model is influence the new model more and finally the prior probabilities of the Spam and ham classes are updated by combining the probabilities. The performance of the concept based concept drift model is evaluated by comparing with multi-nominal Naive Bayes classifier algorithm without updating the model by checking the performance of test mails correctly classified.

In comparison, to different IR models available, the Spam emails can be filtered conveniently using probabilistic Information Retrieval models.

Vector space model (VSM). This IR model uses the terms as feature vectors and finds the relevance to the text document retrieved in general. To this end, the authors Santos et al. of [8] explore usage of semantic Spam filtering with IR based *Semantic-Aware Enhanced Topic-based Vector Space Model (eTVSM)*. The Probabilistic Classifiers used are Bayesian, KNN and SVM. The TVSM model represents the topics as axes therefore the terms are related to how strongly they are related to the topic. The authors have introduced eTVSM which makes use of synonyms, homonyms, Metonymy, Homography, Word-groups. These linguistics are then are used in the *WordNet semantic ontology* (Semantic Lexicon Database), which will help in showing the relation between the terms to represent the vectors for the retrieval of related Spam email. A semantics-aware model is built by the authors which concentrates on synonyms as these kind of linguistics are considered because the Email Spam are using these kind of tactics to evade Spam Filters. The model uses all the information found in the body and subject of the email only to represent the messages in the eTVSM. The Retrieval process goes about as follows. The eTVSM eliminates the stop words (a,the,an..) which add more noise to the model and do not contribute to the term weight. The IR model is represented by 4-tuple $E, Q, F, R(q_i, e_j)$, where E is the representation set of the emails of the dataset, Q is the set of user representation queries, F is the framework consisting of the E, Q and the relationship between

them, $R(q_i, e_j)$ (Similarity function) is the ranking function that assigns a real number to the query and the email representation. The next step is to assign a weight $w_{i,j}$ (between zero to one) to the t terms in the email e_j. Hence email vector is a collection of weights now. eTVSM solves the problem of indepedence by using semantics and creating *interpretation vectors*. The machine learning concepts are applied on this model to classify the email as ham or Spam.

It is a good idea to consider the terms of an email as vectors and check how near or far is the other vector that is the incoming email to determine the Spam as it considers the weight of the terms to improve the model further.

Combined Models. A comparative study done on the "Learning rules that classify Email" by William W.Cohen et al of [2], has compared the method of TF-IDF weighting which is a *vector-spaced model* and the learning set of keyword-spotting rules that is based on *RIPPER rule algorithm* and ultimately checked the accuracy lost by using the using the *keyword-spotting rules*. They have demonstrated that both these methods obtain generalization and are efficient even with small dataset. The interesting part was the Keyword-Spotting rule, that primarily tests the condition of whether the word appears or does not appear in a certain field of a mail message. An *example* of keyword-splitting rule is that the mails containing keywords such as *"prize-money, lucky, password-required"* can be categorized as Spam. The authors made two extensions to the ripper algorithm, at first the first 100 words of the email were changed to boolean feature vectors and the vocabulary was further restricted by making the value of an attribute of email (from,to,cc,subject) be set of symbols. The second extension was to specify a loss ratio (misclassification cost) by changing the weights given to the errors. The trade off on recall for precision for TD-IDF is done by choosing a similarity threshold and the RIPPER algorithm choses an loss ratio. The email is parsed into a header and body. Hence as Ripper algorithm with extensions seemed to be a better IR model for classifying Spam.

Similar to the work done by [2], an interesting study was conducted on an email application *Ishmail* by Helfman et al of of [6]. The experiments conducted by the author concludes that a Email filtering system which has a combination of *user defined rules and keyword spotting rules* will be a favorable architecture. Here the Spam mail went into special mailboxes called misc which in-turn had to be manually moved to other user defined mailboxes and it additionally gave alert and archiving policies. The performance of ripper algorithm and td-idf was compared and very less error percentage was found in RIPPER. In *conclusion*, the Spam is classified well into a particular category if the rules are made of key-words from the mailing list that detect the presence of the term to make a decision and classify it as Spam or otherwise. Hence, the ripper and td-idf IR model in combination yields good performance results.

Neural Network in IR. A recent and growing field is the Neural IR. This field uses convolutional neural networks or artificial neural networks to improve the accuracy of the models after using a term frequency inverse document frequency models to retrieve the terms that is the features/nodes for the neural network. In accordance to this, the authors of [1], have extracted features from emails and have applied back propagation technique to classify email. The output of this neural model is user-defined email class. The input is the word of importance that is the term frequencies of in the email that is calculated after preprocessing steps. The preprocessing steps involves stop word elimination and punctuation removal. The neural nodes are binary in nature (0 or 1). This particular model developed by the authors yields 87% accuracy.

It is a really good idea to use neural networks for better accuracy of the email spam filter model as it back propagates for every error that is made during misclassification of the email.

4 CONCLUSION

Email Spam Filtering is a necessity now-a-days as the technology is improving .The easy availability of Bots at a low price increases the chances of receiving a tailor made Spam email. The other fields where Spam filtering is important and the information retrieval have shown good improvement are Web, SMS and Social media websites.

In conclusion, there are many Information Retrieval models that are very successful in predicting spam such as boolean model, BM25, Language modeling, models using pseudo relevance feedback or query optimization and recent fields like deep neural networks (Convolutional Neural Networks) that may not be addressed in this review. Although the above discussion has shown the most relevant works in the field of Information Retrieval methods to filter Spam emails, there are some problems that are not resolved yet. To this end, Email Spam Filtering is an integral part of any Email software it is necessary to develop Spam Filtering systems that will explicitly avoid Email Spam using Information Retrieval methods.

REFERENCES

[1] Taiwo Ayodele, Shikun Zhou, and Rinat Khusainov. 2010. Email classification using back propagation technique. *International Journal of Intelligent Computing Research (IJICR)* 1, 1/2 (2010), 1.

[2] William W Cohen et al. 1996. Learning rules that classify e-mail. In *AAAI spring symposium on machine learning in information access*, Vol. 18. California, 25.

[3] Gordon V Cormack et al. 2008. Email spam filtering: A systematic review. *Foundations and Trends® in Information Retrieval* 1, 4 (2008), 335–455.

[4] Uri Hanani, Bracha Shapira, and Peretz Shoval. 2001. Information filtering: Overview of issues, research and systems. *User modeling and user-adapted interaction* 11, 3 (2001), 203–259.

[5] Morteza Zi Hayat, Javad Basiri, Leila Seyedhossein, and Azadeh Shakery. 2010. Content-based concept drift detection for email spam filtering. In *Telecommunications (IST), 2010 5th International Symposium on*. IEEE, 531–536.

[6] Jonathan Isaac Helfman and Charles Lee Isbell. 1995. Ishmail: Immediate identification of important information. In *AT&T Labs*. Citeseer.

[7] Wuying Liu and Ting Wang. 2012. Online active multi-field learning for efficient email spam filtering. *Knowledge and Information Systems* 33, 1 (2012), 117–136.

[8] Igor Santos, Carlos Laorden, Borja Sanz, and Pablo G Bringas. 2012. Enhanced topic-based vector space model for semantics-aware spam filtering. *Expert Systems with applications* 39, 1 (2012), 437–444.

,

Authors	Title	Aim	Spam Filter Type	Information Retrieval model	Classification Method	Dataset	Performance
William W. Cohen	Learning Rules that Classify E-Mail	Rules to construct a personalized system for filtering and classification of email by comapring the TD-IDF with keyword spotting rule methods	Rule Based	Vector Spaced Model	TD-IDF weighting and Keyword spotting rules based on RIPPER rule learning algorithm	11 data sets created by collecting data from 3 different users of Ishmail mail application	10 fold cross validation - Ripper algorithm has less errors compared to TD-idf for all the three users
Liu, Wuying and Wang, Ting	Online active multi-field learning for efficient email spam filtering	Online active multi-flield Learning	Active Learning	Probabilistic Model	Incremental supervised binary streaming text classification	trec07p containing 75,419 total email messages (25,220 hams and 50,199 spams).	Evaluates the classifying confidence and has reduced label requirements and space-time and the application exceeds the performance of advanced induvidual text classification algorithm
Igor Santos, Carlos Laorden, Borja Sanz, Pablo G. Bringas	Enhanced Topic-based Vector Space Model for semantics-aware spam filtering	Exploring usage of semantic spam filtering with IR model	Semantic-Aware	Enhanced Topic-based Vector Space Model	Probailistic - Bayesian,Knn,SVM	Ling Spam dataset	10 fold cross validation - Bayesian network followed by random forest with performance above 92%
Hayat, Morteza Zi and Basiri, Javad and Seyedhossein , Leila and Shakery, Azadeh	Content-based concept drift detection for email spam filtering	Language Model that deyects concept drift based on the deviation of email contents	Probabilistic Model	Language Model	Language model combined with classifier	6 different datasets from Enron corpus - 3000 email message (Enron 1,2,3)	Compared with multinominal Naïve Bayes classifier algorith without updating the model by checking the % of test mails correctly classified

Figure 3: Tabulation of studies