Filtering Spam by Using Factors Hyperbolic Trees

Hailong Hou*, Yan Chen, Raheem Beyah, Yan-Qing Zhang
Department of Computer science
Georgia State University
P.O. Box 3994
Atlanta, GA 30302-3994, USA
*Contact author: hhou2@student.gsu.edu, ychen44@student.gsu.edu, rbeyah@cs.gsu.edu, yzhang@cs.gsu.edu

Abstract—Most of current Anti-spam techniques, like the Bayesian anti-spam algorithm, primarily use lexical matching for filtering unsolicited bulk E-mails (UBE) and unsolicited commercial E-mails (UCE). However, precision of spam filtering is usually low when the lexical matching algorithms are used in real dynamic environments. For example, an E-mail of refrigerator advertisements is useful for most families, but it is useless for Eskimos. The lexical matching anti-spam algorithms cannot distinguish such processed E-mails that are junk to most people but are really useful for others. We propose a Factors Hyperbolic Tree (FHT) based algorithm that, unlike the lexical matching algorithms, handles spam filtering in a dynamic environment by considering various relevant factors. The new Ranked Term Frequency (RTF) algorithm is proposed to extract indicators from E-mails that are related to environmental factors. Type-I and Type-2 fuzzy logic systems are used to evaluate the indicators and determine whether E-mails are spam based on the environmental factors. Additionally, weights of factors in a FHT database are continuously updated according to dynamic conditional factors in a real environment. Simulation results show that the FHT algorithm filters out spams with high precision. Furthermore, the FHT algorithm is more efficient than other methods when it filters E-mails with complex influencing factors. The main contribution of this paper is that the Factors hyperbolic tree based algorithm can filter E-mails based on influencing factors instead of matched words to allow dynamic filtering of spam E-mails.

Index Terms—spam; Bayesian algorithm; Ranked Term Frequency; fuzzy logic; factors hyperbolic trees.

I. INTRODUCTION

E-mail spam, also known as "bulk E-mail" or "junk E-mail", has existed since the beginning of the Internet. The basic idea is that nearly identical messages are sent to numerous recipients by E-mail [1]. Spam can be described as unsolicited bulk E-mail (UBE) where unsolicited means that the recipient has not granted verifiable permission for the message to be sent and bulk means that the message is sent as part of a larger collection of messages, all having substantially identical content [2]. Unsolicited commercial E-mail (UCE) is the most common type of spam. UCE seeks to engage a potential consumer in order to exchange goods or services for money. Spam has become a significant problem and has grown to about 90 billion messages a day. Botnets and virus infected computers, account for about 80% of spam. Symantec [3] recently reported that they detected a 30% increase in phishing attempts from Jan 2006 to the end of the year.

Statistics from the Distributed Checksum Clearinghouse (DCC) project [4] shows that 51% of the E-mail messages checked by the DCC network in 2007 were likely to be bulk E-mail. About 85.65% threats came from spammers in Jan 2008, which were checked by MX Logic [5]. Laws to prevent certain types of spam have been enacted in many countries. For example, in the United States, spam is legally permissible according to the CAN-SPAM Act of 2003 provided it follows certain criteria; the European Union (EU) Directive on Privacy and Electronic Communications (2002/58/EC) provides that the EU member states shall take appropriate measures to ensure that unsolicited communications are prevented; also, in Australia, the relevant legislation is the Spam Act 2003 which covers certain types of E-mail and phone spam, which took effect on 11 April 2004. However, good-faith compliance with anti-spam laws is not always enough to keep a legitimate Internet or wireless marketer out of trouble because of the considerable cost to analyze the relevance between the spam messages and the law. This is heightened by the lack of concern of the laws by malicious spammers. We propose the use of a FHT and RTF algorithm that allows spam to be filtered dynamically as some bulk E-mail may be of interest to certain users. This technique assists the users as they will receive E-mail that matches their interests and unsolicited (or unwanted) E-mail will be filtered.

The rest of this paper is organized as follows. Section II shows the related work. Section III introduces Factors hyperbolic tree and shows how it works. Section IV describes mining data by ranked term frequency algorithm and how to process features of E-mail. Section V depicts computation model based on FHT and fuzzy logic. Section VI discusses that personal interests may affect spam identification. Section VII evaluates the performance of our model by comparing simulation results of Bayesian algorithm and FHT algorithm. Section VIII concludes our work and presents future work.

II. RELATED WORK

Some popular methods for filtering spam have been deployed by Internet Service Providers (ISP), like DNS-based blackhole lists (DNSBL), greylisting, spamtraps, enforcing technical requirements, checksumming systems to detect bulk E-mail, and by putting some sort of cost on the sender via a Proof-of-work system or a micropayment. However, each method has strengths and weaknesses and each is controversial due to its weaknesses. Thus, new methods have been
developed to replace many of the aforementioned techniques for handling E-mail spam. Among these methods, Bayesian filtering is probably one of the most widely used to identify spam E-mail, and is therefore integrated in many popular E-mail clients [6, 7]. Algorithms based on Naive Bayes extract keywords and other indicators from E-mail messages, and determine whether the messages are spam using statistical or heuristic schemes. However, spammers nowadays are using sophisticated techniques to trick content based filters by clever manipulation of the spam content [8]. A learning approach to spam sender detection based on features extracted from social networks constructed from E-mail exchange logs was proposed in [9]. The approach extracts several features from E-mail social networks for each sender. Based on these features, a supervised model is used to learn the behaviors of spammers and legitimate senders and then assign a legitimacy score to each sender. Scores are made available in a database where online mitigation methods can query for the score of a particular sender. In [10], the method Spam Filtering Model Based on Support Vector Machine (SVM) is proposed. A SVM is a new learning algorithm which has some attractive features such as eliminating the need for feature selections, which enables easier spam classification. Producing a high dimensional feature space is vital for getting better performance from SVM for filtering spam to produce a high dimensional feature space. However, neither Social network nor SVM considers dynamic factors in a real environment. The changes of factors indubitably alter E-mail’s property so that a useful E-mail may become useless.

In this paper, we propose the FHT based fuzzy decision algorithm for spam detection, which may deals with the dynamic factors without maintenance of blacklists and white lists. Specifically, the proposed framework extracts several dynamic features from E-mail. A FHT model is used to learn and periodically update the factors’ values of features relied on these dynamic features. The fuzzy logic is then used to compute the score of the E-mail. The score is compared to a criterion score $\lambda$. If the score greater than $\lambda$, it is useful E-mail; otherwise, it is spam.

### III. FACTORS HYPERBOLIC TREE

A decision is made normally by considering a certain amount of factors along with specific reasoning. For example, determining whether air conditioning (AC) is useful depends on two factors: temperature and zone. For another example, the distance required and potential terrain would determine what kind of vehicle needed for an expedition. Thus, an objects’ (e.g., AC use, vehicle type) value can be influenced by certain types of factors (e.g., temperature, distance, etc.), with the factors relevant at specific times. To express the relationship among decisions, factors and objects, we propose a FHT, which can describe every object of interest with related factors, and the values of factors can be constantly updated adapting over time.

To implement a Factors hyperbolic tree based system, a large database needs be formed with objects, factors and the relationships between them. The records in the database include all the information around us such as products, weather, human, earth, society, country, culture, zone etc. The following simple example can present the relationship between one object and several related factors.

**Example 1:**

**What factors are related to heater?**

**Weather**  
In summer, a heater may be useless; in winter, it is possibly very useful.

**Zone**  
In Florida, heater is practically useless; in New York, it is useful.

**People**  
For old people, a heater may be needed; for young people, it may not be as necessary.

**Popularity**  
If people would like to buy a heater in winter degree instead of AC, it has an increased popularity relative to an AC; otherwise, it does not.

Example 1 shows the relationship among heater and several factors, each explanation of the factor describes how it works. In fact, the description is very easily digitized by fuzzy logic systems so that they may be computed by other methods. Moreover, because a database will be used to depict the FHT system, the data structure and obtaining these data are important issues that will be discussed further in the rest of this paper.

#### A. Classification of factors

The factors can be classified into two categories: normal and abnormal.

**Normal factors**: these factors change with time naturally or periodically, such as weather, age, etc.

**Abnormal factors**: factors that happen suddenly and affect the object immediately, such as disasters and popularity degree, etc.

#### B. Expression of normal factors by fuzzy logic system

To simplify conditional factors, normal factors are standardized to a number ranging between 0 and 1 by a type-2 fuzzy set, which achieves a more accurate expression. Type-2 fuzzy sets are an extension of type-1 fuzzy sets in which uncertainty is represented by an additional dimension. The Type-2 fuzzy sets are used to normalize factors because: (1) the variation of normal factors obey natural rules, which means the value of normal factor at certain time must fluctuate in a range. And (2) a range of values will be generated if periodic values of normal factors are selected.
Figure 1 (a) shows the temperature curve of one year. The value of every month fluctuates in a range (the blue dots indicate lowest temperature, orange dots indicate highest temperature and the red dots indicate average value). Thus, the type-2 membership function of a specific month’s temperature is given in figure 1(b) where each uncertainty temperature value is represented by an additional dimension (the red line indicates the defuzzied membership function).

Overall, in FHT based system, all the initial weights between factors and object are set to 0; type-2 fuzzy logic is used to compute the weights between normal factors and objects $w_{ni}$ (i is the index normal factor).

### C. Data structure of FHT

Every object has its own attributes, which are the influencing factors in the FHT algorithm. The database plays a crucial role in the FHT algorithm and stores all the relationships and objects, which should be represented precisely by the data structure of records in the database. In this paper, Factors hyperbolic tree (FHT) is proposed to express our algorithm. Figure 2(a) depicts the data structure of the FHT. In the figure, the decision node is the root and it will organize sub trees; every other node is composed of objects and their related factors. The relationship between object and factors is shown in figure 2(b), specifically, objects can share more than one factors in order to avoid factor redundancy. In the FHT, the object is represented by a factor set, and the values of factor in the set will be updated corresponding to environmental factors changing.

*Figure 1. Type-2 fuzzy logic membership function.*

*Figure 2. Structure of Factors hyperbolic tree.*
D. Data mining of influencing factors

To avoid confusion of the relationship between object and factors, relationships must be well predefined and initialized in the database. Once an event triggers input, the weights of the FHT will be updated. However, obtaining data corresponding to the factors is challenging because the database is large.

Data mining for such a huge database is very difficult; fortunately, the Internet is a good way to obtain information although not all websites provide correct information. Accordingly, authority websites are the best options for extracting information to populate the database. For instance, weather information can be obtained from weather.com; information about the law can be mined from government websites; sports information can be got from TNT or ESPN websites, etc. Additionally, for more conveniently selecting authority websites, an efficient classification algorithm for distinguishing authority and non-authority websites must be elaborately designed.

IV. RANKED TERM FREQUENCY AND APPLICATION

Key words are used to extract the features of E-mails because words are basic units of languages. Furthermore, nouns are the primary words extracted by the Ranked Term Frequency (RTF) algorithm since the spammers often trap recipients by using a high frequency of nouns.

A. Ranked Term Frequency

The Term Frequency Inverse Document Frequency (TF-IDF) weight is an algorithm often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is for a document in a collection or corpus. The importance increases proportionally to the number of times a word appearing in the document but is offset by the frequency of the word in the corpus. We propose a RTF algorithm to mine noun words in the document while the IDF concept is ignored because more accurate features are expected to be obtained only from the single E-mail, but not the entire corpus.

The RTF in the given document is simply the ranked number of times that a given term appears in that document. This count is usually normalized to prevent bias towards longer documents (which may have a higher term frequency regardless of the actual importance of that term in the document) to give a measure of the importance of the term ti within the particular document dj.

\[
tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}
\]

Where \(n_{i,j}\) is the number of occurrences of the considered term in document \(d_j\), and the denominator is the number of occurrences of all terms in document \(d_j\).

The RTF result will be shown after all \(tf_{i,j}\) have been ranked, a high weight in RTF is reached by a high term frequency (in the given document); hence the weights tend to filter out important noun terms in \(d_j\). However, we must primarily clarify what kinds of noun are associated to factors; and then it is possible to use RTF results to obtain meaningful information from E-mails. Following this rule, the first m (3 ≤ m ≤ 8) highest weights nouns are selected to be processed in the FHT algorithm to avoid a potential misclassification of the E-mail.

B. E-mail Processing

Real spam is generally E-mail advertising for some product or service sent to a mailing list or newsgroup. In addition to wasting people's time with unwanted E-mail, spam also consumes a lot of network bandwidth.

In order to filter E-mail by conditional factors, original E-mail must be processed by following steps: (1) extract features of E-mail; (2) search factors related to the features in FHT; (3) represent E-mail by features with associated factors. After these three steps, E-mail is possible to be computed in our FHT based computation model. The difference between original E-mail and processed E-mail are listed in table 1. From the table, processed E-mail expresses its special feature which related with certain factors.

<table>
<thead>
<tr>
<th>Description</th>
<th>Original E-mail</th>
<th>Processed E-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standardized into format</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Could be legal message even it is spam</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>now, vice versa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words related to other affection factors</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Can be digitized and computed</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

V. DECISION COMPUTATION MODEL

Decision is the result of analyzing activities of conditional factors, which will be obtained by searching in FHT. To make a decision, a computation model based on FHT and fuzzy logic is shown in figure 3. The algorithm is described as:

**FHT based Algorithm**

**Start:**

**Step 1.** Extract features by using RTF and input to FHT Processing Unit.

**Step 2.** Search related factors to the features in FHT and output factors to Fuzzy Logic Unit in two types: normal factors and abnormal factors.

**Step 3.** Fuzzy Logic Unit will compute normal factors to obtain an incomplete result; then abnormal factors will be processed to obtain another incomplete result.

**Step 4.** Weighted Computation Unit will calculate two incomplete result with predefined weights to obtain final decision.

**End**

During the step 3, factorial analysis is applied for optimally setting up a rulebase. That is, factorial analysis algorithm can be used to obtain affection degree of normal factors by...
analyzing historical data. The values of these degrees are represented by 1*n matrix:

\[ \text{Degree} = \{d_1, d_2, \ldots, d_n\} \]

Then the rules can be computed by equation (2):

\[ \text{Rule}_i = 1 - (1 - a_1 * d_1)(1 - a_2 * d_2) \ldots \ldots (1 - a_n * d_n) \]  
(2)

where \( \alpha \) is the parameter computed from membership function.

Moreover, for the abnormal factors, type-1 fuzzy system is applied to compute an incomplete output. Meanwhile, factorial analysis can be also applied to adjust rulebase. Finally, equation (3) is used in step 4 to compute the complete output.

\[ \lambda = w_1 * \text{output}_{\text{normal}} + w_2 * \text{output}_{\text{abnormal}} \]  
(3)

where \( w_1 \) and \( w_2 \) are predefined weights.

**VI. PERSONALIZED DECISION**

Although several influencing factors in the FHT based system partly determine whether incoming E-mail is spam, personal preference must also be considered in the procedure. For an object, user preference is very useful in adjusting the value of natural factors’ weights after factorial analysis. Frequency value of the factors which related to user preference can be normalized to \([0,1]\) and represented by a 1*n matrix:

\[ F_{\text{value}} = \{f_1, f_2, \ldots, f_n\} \]

This matrix is applied to equation (2) to get new formula:

\[ \text{Rule}_i = 1 - (1 - a_1 * d_1)(1 - a_2 * d_2) \ldots \ldots (1 - a_n * d_n)^{f_i} \]  
(2)

Furthermore, the preference frequency also can be applied to the final step which computes the complete result to adjust the predefined weights.

To define the spam E-mail by FHT, we test the final output \( \lambda \) compared to prior known spam and get the criterion value by averaging \( \lambda \)s. Then E-mail is distinguished by FHT - if the output is smaller than \( \lambda \), it is spam; otherwise, it is not.

**VII. PERFORMANCE AND EVALUATION**

In this section, we discuss experiments set up to evaluate the FHT based algorithm and compare it to Bayesian algorithm.

**A. Experiment setup**

Experiment environment: Windows XP professional


Fig 3. Computation model base on Factors Hyperbolic Tree and fuzzy logic.

The FHT based system and Bayesian system developed in this experiment was written in Visual C# 2005. The plain text files ‘*.txt’ are used to emulate real E-mails and an Access database stores all the objects, factors and relationships and key words for Bayesian algorithm.

**B. Results and evaluation**

The E-mail filtering result of FHT depends on the weights of normal and abnormal factors. Different ratios of these two weights yield different crisp outputs \( \lambda \). In our experiment, the ratio is set to a fixed value 1 so that we can test several groups of E-mails conveniently. Furthermore, the criterion \( \lambda = 0.55 \) is obtained by averaging all outputs after 100 spam were inputted into FHT based software. Therefore, if the final result computed by FHT system is greater than 0.55, it should be normal E-mail; otherwise, it should be spam.

Fig 4. Incorrectly filtering comparison.

Moreover, to compare FHT based algorithm with Bayesian algorithm, two different groups of E-mails will be analyzed: 1000 spam and 1000 normal E-mails. For every group, the 1000 E-mails were divided into 10 sub groups with same size and were filtered by the FHT algorithm and Bayesian algorithm.
Figure 4 shows the incorrectly filtered result of FHT algorithm and Bayesian algorithm where E-mails are 100% white E-mails; The FHT algorithm expresses a lower incorrect filtering percentage than Bayesian algorithm, which indicates that more white E-mails were classified into spam by Bayesian than by FHT.

In figure 6, identifying percentages of spam are compared between these two algorithms. FHT algorithm also presents higher filtering percentage than Bayesian algorithm, which indicates that FHT algorithm is more efficient for filtering spam than Bayesian algorithm in our experiment. More spam from standard set will be added into our experiment for better comparing the filtering efficiency of FHT and Bayesian algorithm.

VIII. CONCLUSION

In this paper, we propose the FHT algorithm which filters E-mail by conditional factors instead of word matching. The results of the experiments show this technique can efficiently filter E-mail, and also improve the filtering degree of precision. Experimentally, FHT is a feasible technique to apply to anti-spam filters, especially in the case when information related to current environment factors is not accurately represented by lexical matching algorithms.

In our future work, we will complete factors hyperbolic trees to build a novel and complete system. Then apply it to other fields, such as semantic analysis and decision supporting field. Moreover, a more efficient method to find factors of objects will be investigated.

Acknowledgement: We will thank Yuchun Thang for his help to provide the spam set to test our algorithm.

REFERENCES

[9] HoYu Lam, DitYan Yeung, “a learning approach to spam sender detection based on social networks”, CEAS 2007 Fourth Conference on E-mail and AntiSpam, August 23, 2007,