Volume 3, Issue 10, October 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Spam Detection using G-LDA

Shubham Saini* *Vellore Institute of Technology India*
Bhavesh Kasliwal *Vellore Institute of Technology India*
Shraey Bhatia *Vellore Institute of Technology India*

Abstract— Spam is any unwanted message, especially advertisement and fraud schemes. The average cost of spam per employee per year at 82 of the Fortune 500 companies is estimated to be $1934. The paper proposes a novel process called G-LDA, which takes concepts from Latent Dirichlet Allocation (LDA) and Genetic Evolution techniques. This involves framing a set of words having a high frequency of occurrence in any spam email. The method was tested on Enron spam corpus. The phrases that were evolved through the generations reflected significant improvement.

Keywords— Spam detection, Latent Dirichlet Allocation, Genetic Algorithm, Topic Modelling, Breeding, Fitness

I. INTRODUCTION

Spam is the use of electronic messaging systems to send unsolicited bulk messages, especially advertising and fraud schemes, indiscriminately. This paper proposes a novel method which combines Latent Dirichlet Allocation[1] (LDA) with genetic evolution techniques, hence the name G-LDA. LDA is a way of automatically discovering topics within a documents collection. Each word in the documents collection will belong to a particular topic. The technique can be applied to a spam email training data. The result will be a set of words having a high probability of occurrence within a spam email. Further, set of spam words generated earlier are subjected to genetic evolution techniques. This is done to get an improved set of words with a higher probability of occurrence within a spam email.

II. SET GENERATION

A set of 1000 spam emails from the Enron-Spam dataset is taken. Sets of words from these emails are generated using the JGibbLDA[4] package. JGibbLDA is a Java implementation of LDA using Gibbs Sampling technique. A total of 10 sets with 100 words each were generated. These sets were tested against 1000 ham emails and 1000 spam emails from the Enron dataset. Number of emails containing the words was found:

Number of emails matched (N_e) = (Number of spam emails matched) – (Number of ham emails matched)

Now, for each set of words, the sum of number of emails of the words in that set ($\sum N_e$) is calculated. Two sets having the highest sum are taken, rest are discarded. Genetic evolution techniques are applied on these words.

III. GENETIC EVOLUTION ALGORITHM

Genetic algorithm is applied as follows:

- Loop for n generations
  - Find Fitness Function for each keyword
  - Breeding using roulette wheel selection
  - Filtering to produce a new generation

A. Fitness Function

Based on the number of emails matched, a fitness value is assigned to each spam keyword:

$$\text{Fitness} = \text{Number of emails} * \left(1 + \frac{N - N_w}{N}\right)$$

Where
N = Threshold value
N_w = No. of words comprising the spam keyword

Threshold value is taken in order to provide a negative weight. Any keyword longer than N words will receive fitness penalty, and keywords less than N words receive a bonus.

B. Breeding

The process used created 1 child from 2 parents using the OR Method. Each survivor from previous generation breeds with a 2nd parent selected using Roulette Wheel Selection.

The OR Method – Take 2 parents, such as (Lottery|Cash) and (Cash|Win), breed them to produce the child (Lottery|Cash|Win). Here, $N_w = 3$ as the keyword comprises of 3 words.

Roulette Wheel Selection – It is a method for choosing a keyword from the breeding pool. Keywords with better fitness are more likely to be chosen.

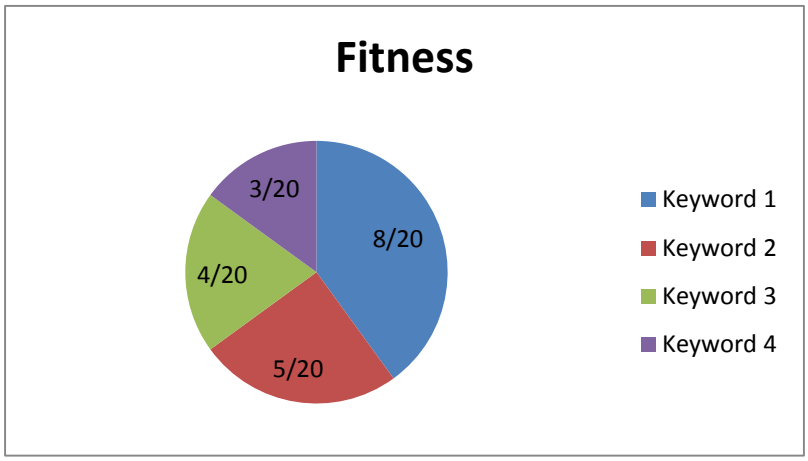

Fig. 1 Roulette Wheel Selection example

In Fig. 1, the breeding pool consists of 4 keywords. The probability of Keyword 1 being selected is 8/20, the highest among all the keywords. This is based on the fact that Keyword 1 has the highest fitness of 8.

Breading is carried out as follows to produce new children

- Mark all keywords as 'new'
- Loop till no keyword is marked 'new'
    - Take a keyword (P1) with highest fitness which is marked 'new'.
    - Mark P1 as 'old'
    - Take another keyword (P2) using Roulette Wheel Selection
    - Mark P2 as 'old'
    - Breed P1 and P2 to produce a child C = (P1|P2).
    - Mark C as 'old'

C. Filtering

Once breeding is done, fitness of the newly generated children is calculated and then filtering is performed in order to remove low fitness keywords. The keywords remaining after filtering constitute the new generation. Filtering criteria may vary with need and situation. For the purpose of spam detection, it was intended to keep only top 200 keywords. Hence, after each round of breeding, top 200 keywords were sent to the next generation and the rest of the keywords were discarded.

IV. PROOF OF CONCEPT RESULTS

In the Fitness function, as the threshold value (N) increases, the time required to find the number of emails for each phrase increases with each generation as phrases with more number of words tend to survive. After performing several tests, the threshold value was chosen to be 5. This is the optimum value which balances the time required for finding the number of emails for each phrase and the number of new generations required. The breeding is stopped when the increase in the fitness of the phrases become insignificant.

The top 10 phrases of the initial generation, 5th generation and the12th generation, with their fitness score is given below:

TABLE I: TOP 10 PHRASES OF THE INITIAL GENERATION

Phrase	**Fitness**
more	45.846
email	32.238
price	31.14
business	29.915998
over	29.862
money	26.37
address	24.335999
order	19.98
doll	19.421999
world	19.152

TABLE II: TOP 10 PHRASES AFTER THE FIFTH GENERATION

Phrase	**Fitness**
more	45.846
more|internationa	45.728004
more|software	44.528004
more|internationa|size	42.336002
business|address	41.472
more|internationa|remove|looking	40.584
more|internationa|size|bank	40.212
more|internationa|size|professional	39.456
business|link	38.72
more|las|client|pain	38.532

TABLE IIII: TOP 10 PHRASES AFTER THE TWELTH GENERATION

Phrase	**Fitness**
more|price|less	48.65
more|over|cheap	47.431995
more|email|money|statements	47.256
business|address|price	46.143997
more|over|order|hour	45.996002
more	45.846
more|internationa	45.728004
more|software	44.528004
more|software|order|investment	42.396
more|internationa|size	42.336002

The phrases generated through generations 1 to 12, with their number of emails matched and the fitness score are available in the CSV files at the GIT Repository. The Repository contains the Enron Email Corpus used for generating words using LDA, and for testing and breeding the phrases.

The collection enron_spam-1000 contains the collection of spam emails from which initial set of words has been generated. The enron_ham-1000 is used for initial filtering.

The enron_spam-6000 contains the collection of spam emails used for breeding and finding the fitness of the phrases. After each generation, the fittest 200 phrases are taken to the next round of breeding.
https://github.com/shubhamsaini/SpamDetection-G-LDA

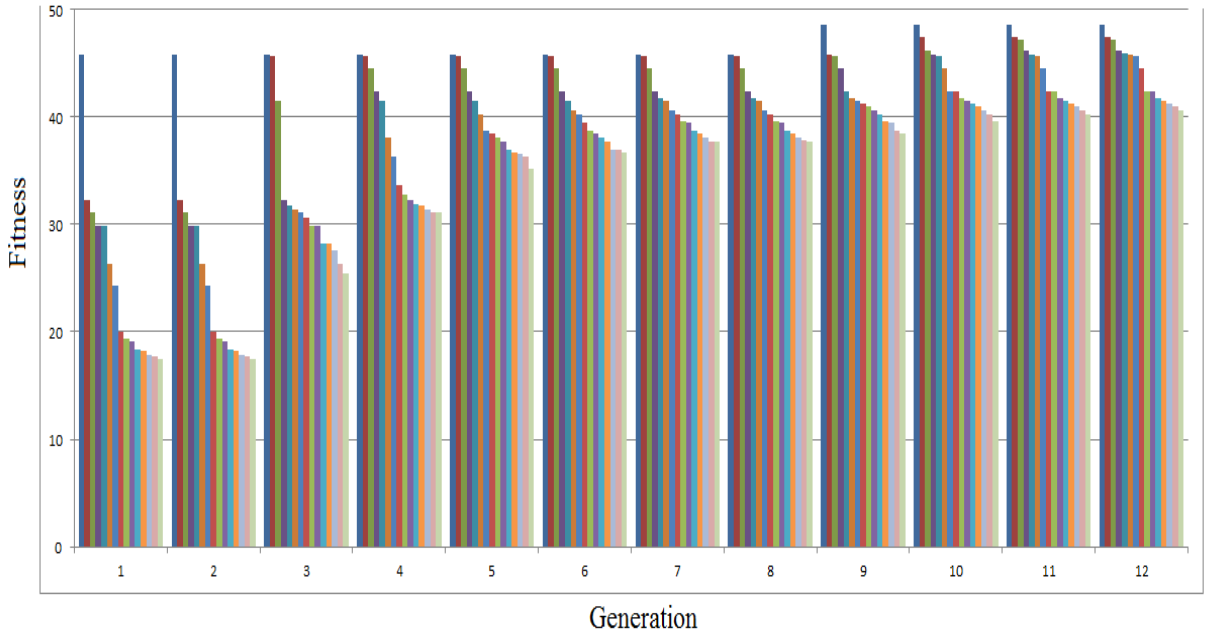

Fig. 2 Fitness of the top 15 phrases for each generation

It is evident from Fig. 2 that average fitness of the phrases increases gradually through the first 5 generations. This is because of the fact that the threshold value taken in the fitness function was 5. After the 5th generation, the increase in the average fitness of the phrases with each generation reduces. The increase in the fitness from 11th to the 12th generation is almost negligible, hence stopping the breeding process there.

V. CONCLUSIONS

Genetic algorithm simulates the processes of nature necessary for evolution, especially the principle laid down by Charles Darwin of 'Survival of the Fittest'. Combining this technique with topic modeling method, the set of phrases generated effectively match spam, and show significant improvement with each generation. The average of the fitness of the 12th generation of phrases was almost 4 times the average of the fitness of the initial generation.

VI. FUTURE WORK

Since the proposed model is based on LDA and Genetic Algorithm, there is still a probability of spam email going undetected. To take this into consideration a mechanism can be devised in which if a user flags a mail as spam, that mail is accommodated in the spam email set. So, after a fixed time quantum of time, there will be new additions in the spam set of emails. With this new set, the G-LDA model can be re-run and improved results can be obtained.

ACKOWLEDGEMENT

We take this opportunity to express our profound gratitude and deep regards to our guide Prof. Nisha V. M. for her exemplary guidance, monitoring and constant encouragement throughout the research.

REFERENCES

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[2] Bíró, István, Jácint Szabó, and András A. Benczúr. "Latent dirichlet allocation in web spam filtering." *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*. ACM, 2008.

[3] Conrad, Eric. "Detecting Spam with Genetic Regular Expressions." *SANS Institute InfoSec Reading Room* (2007).

[4] Phan, Xuan-Hieu, and Cam-Tu Nguyen. "Jgibblda: A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference." (2006).

[5] Ramage, D., and E. Rosen. "Stanford topic modeling toolbox." (2011).

[6] Wall, Matthew. "GAlib: A C++ library of genetic algorithm components."*Mechanical Engineering Department, Massachusetts Institute of Technology* 87 (1996): 54.

