SHIV SHAKTI
International Journal in Multidisciplinary and Academic Research (SSIJMAR)
Vol. 7, No. 2, April 2018 (ISSN 2278 – 5973)

Spam Detection

*MR. NAVAKISHORE,**CH.SAI KUMAR REDDY
K.SRILEKHA,* K.RAHUL

* Assistant professor HITAM, Hyderabad, India, navakishor581@gmail.com
**Computer science and engineering, HITAM, Hyderabad, India,chiluverusaikumarreddy33@gmail.com
***Computer science and engineering, HITAM, Hyderabad, India, srilekhareddy5490@gmail.com
****Computer science and engineering, HITAM, Hyderabad, India,rahulyadav4796@gmail.com

Abstract

Spam is unsolicited mass email the it's the bulk nature of it which is so offensive. Spam is the another form of stealing. Spammers make people pay for their message without permission of the recipient.The increasing menace of spam is bringing down productivity. More than 68% of the email messages are spam, and it has become a challenge to separate such messages from the right ones. We have developed a spam identification engine which identifies spam. It not only identifies the spam but also tells the count of the spam. Huge amounts of mails can be scanned and we can identify the spam mails at a time. We are using Hadoop for this system which uses map reduce to know the count of spam messages. This is very useful for the research organisatations.

Keyword-Spam, Detection, Mail, Identifying Data,count.

1. INTRODUCTION

1.1 Email

Email (electronic mail) is the exchange ofComputer stored messages by telecommunication.(Some publications spell it email; we prefer theCurrently more established spelling of e-mail.) E-Mail messages are usually encoded in ASCII text.However, you can also send non-text files, such asgraphic images and sound files, as attachmentsSent in binary streams. E-mail was one of the firstuses of the Internet and is still the most popularuse.E-mail is a message that may contain text,files, images, or other attachments sent through a network to aSpecified individual or group of individuals.An email address is required to receive email, and that address is unique to the user. Some people useInternet-based applications and some use programs on their computer to access and store emails.Companies that are fully computerized make extensive use of e-mail because it is fast, flexible, andreliable.email communication is not only used in lieu of letter writing, it has also replaced telephone calls in manysocial situations and in professional environments.Some electronic mail systems are confined to a single computer system or network, but others havegateways to other computer systems, enabling users to send electronic mail anywhere in the world.

was one of the first methods of person-to-person communication made available through theInformation superhighway. In the early days of e-mail, simple text messages were sometimes difficult tomanage, and adding pictures or documents was

possible only if other software was available to maketransmission from e-mail to computer possible. Current e-mail software generally provides easy-to-useoptions for attaching photos, sounds, video clips, complete documents, and Hypertext Markup languages (HTML) code.Even with attachments, however, e-mail messages continue to be text messages -- we'll see why when weget to the section on attachments.

1.2. Hadoop:

Hadoop is a complete eco-system of open source projects that provide us the framework to deal with bigdata. Let's start by brainstorming the possible challenges of dealing with big data (on traditional systems)and then look at the capability of Hadoop solution.

Following are the challenges I can think of in dealing with big data :

I. .High capital investment in procuring a server with high processing capacity.
II. Enormous time taken
III. In case of long query, imagine an error happens on the last step. You will waste so much time making
IV. these iterations.
V. Difficulty in program query building

Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume.

Modules of Hadoop

- HDFS: Hadoop Distributed File System. Google published its paper GFS and on the basis of that
- HDFS was developed. It states that the files will be broken into blocks and stothe distributed architecture
- Yarn: Yet another Resource Negotiator is used for job scheduling and manage the cluster.
- Map Reduce: This is a framework which helps Java programs to do the parallel computation on data using key value pair.
- The Map task takes input data and converts it into a data set which can be computed in Key value pair
- . The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.
- Hadoop Common: These Java libraries are used to start Hadoop and are used by other Hadoop modules.

1.3 MapReduce:

Programming is not a good match for all problems. It's good for simple informationrequests and problems that can be divided into independent units, but it's not efficient for iterative andinteractive analytic tasks. Map Reduce is file-intensive. Because the nodes don't intercommunicate exceptthrough sorts and shuffles, iterative algorithms require multiple map-shuffle/sort-reduce phasesMapReduce is a core component of the apache hadoop software framework.Hadoop enables resilient, distributed processing of massive unstructured data sets across commodityComputer clusters in which each node of the cluster includes its own storage. MapReduce serves twoessential functions: which filters and parcels out work to various nodes within the cluster or map, a functionSometimes referred to as the mapper, and it organizes and reduces the results from each node into a Cohesive answer to a query, referred to as the reducer.The term MapReduce actually refers to two separate and distinct tasks that Hadoop programs perform.The first is the map job, which takes a set of data and converts it into another set of data, where individual
elements are broken down into tuples (key/value pairs).The reduce job takes the output from a map as input and combines those data tuples into a smaller set oftuples. As the sequence of the name MapReduce implies, the reduce job is always performed after themap job.

2. LITERARTURE SURVEY

IN THE SPAM MAIL FILTERING, A NEW APPROACH BASED ON THIS STRATEGY THAT HOW FREQUENTLY WORDS IN THE EVIDENCE ARE FOUND BY USAGE OF

The key sentences, those with the key words, of the incoming emails have to be tagged and

thereafter the grammatical roles of the entire words in the sentence need to be determined, finally they will be put together in a Vector in order to indicate the similarity between the received emails. So it takes advantage of an extraordinary algorithm called K-Mean algorithm to classify the received e-mails. It is Worthwhile to note that the so-called K-Mean algorithm follows some Simple and understandable rules which are too easy to work with it. This method is executed on 189 e-mails.142 of are e-mails are

non-spam e-mails and 49 spam e-mails were available among them. After repeating above steps for a repeated number of times the final precision for this method is obtained 84percent .In other thing they have integrated the content based spam detection using Bayesian Classifier and phishing URLs detection using Decision Tree C4.5. Thus they found that performance evaluated for combination approach of Bayesian classifier and Decision Tree are improved as compared to implementation using content based spam detection by Bayesian Classifier. In this spam there prime aim is to detect text as well as image based spam emails. To achieve the objective we applied three algorithms namely: KNN algorithm .

3.Existing System

The methods currently used by most anti-spamsoftware are static, mean that it is fairly easy to evade by tweaking the message little.To do this spammer simply examines the latest anti-spam techniques and find the ways how to dodge them. To effectively combat spam, an adaptive new technique is needed.This method must be familiar with spammer's tactics as they change over time. It must also able to adapt to the particular organization that it is protecting for the answer lies in Bayesian mathematics.

4. Proposed System

In this we are proposing a system which detects the spam messages and tells the no of spam messages in the mail by which we can know the spam count this is helpful for knowing the rate of spam whether it is increased or decreased.It is utmost necessary to stop all the unwanted messages as they contain viruses which harms the computer.it is reliable means it can be accessed in multiple system.The messages are separated using hadoop system. This helps the user to avoid the spam messages.

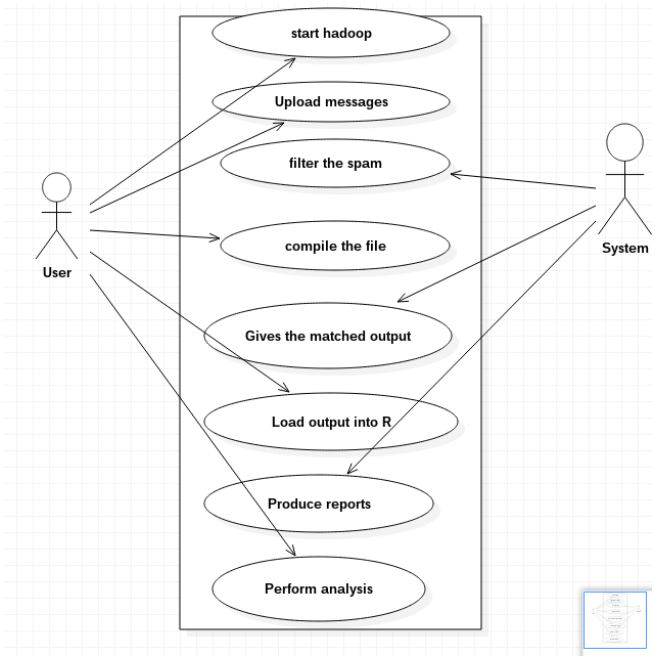

Fig 1 Use case Diagram

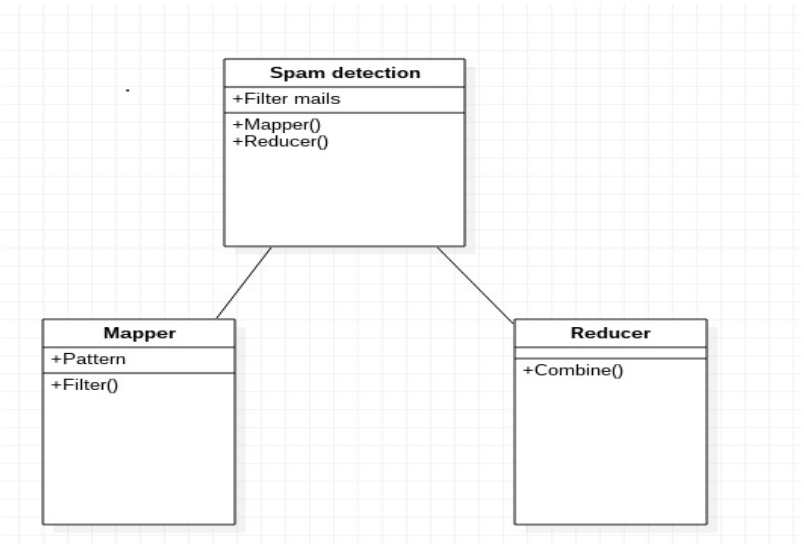

Fig 2 Class diagram

5.Advantages and drawbacks

ADVANTAGES:

Easy to identify spam message.
Spam count can be known very easily and fastly.
Huge data can be processed at a time without any interruption.

DRAWBACKS:
Need more ram size.
High configurations required.
User should have knowledge in coding and Hadoop.

6.SOFTWARE AND HARDWARE

SOFTWARE REQUIREMENTS:
Operating system: windows/Ubuntu
Software: hadoop
Language: java

HARDWARE COMPONENTS:
Personal Computers
6GB ram
1tb hard disk

7.Conclusion

We would like to extend our project by adding more spam filters to detect the spam easily and fast.Our project is more efficient as compared to the existing systems our system detects the spam messages and it tells the count of no of spam messages detected. Filtration of messages can be done very efficiently. We hope that it is accessible to every user further.

8.REFERENCES

[1] Ending spam- Book by Jonathan A. Zdziarski.

[2] spam filtering: a systematic review-Book by Gordon V. Cormack.

[3] S.Hinde (2002). Spam ,sacms,chains,hoaxes and other junk mails Computer &security vol21 pp592-606

[4] G. Cormack, "Email spam filtering: A systematic review," Foundations and Trends in Information Retrieval, vol. 1, no. 4, pp. 335–455, 2008.

[5] Al-Jarrah, 0., Khater, I., & Al-Duwairi, B. (2012). Identifying Potentially Useful Email Header Features for Email Spam Filtering. In ICDS 2012, The Sixth International Conference on Digital Society, pp. 140-145.

[6] <http://www.securelist.com/en/threats/spam?chapter=97>.

[7] Qaroush, A., Khater, 1. M., &Washaha, M. (2012). Identifying spam e-mail based-on statistical header features and sender behavior. In Proceedings of the CUBE international information Technology Conference, pp. 771-778.

[8] M. Basavaraju, R. Prabhakar, "A Novel Method of Spam Mail Detection using Text Based Clustering Approach", *International Journal of Computer Applications (0975–8887)*, vol. 5, no. 4, August 2010.

[9] https://pdfs.semanticscholar.org/9075/6e3a883234b6eb0ed8e7d7d7125d16c1569e.pdf

[10] http://www.ajer.org/papers/v2(10)/F02106373.pdf