**A Project Report**

**on**

**E-MAIL CLASSIFICATION USING DATA MINING**

Submitted in partial fulfillment of Degree of

Bachelor of Engineering

(COMPUTER ENGINEERING)

Academic Year: 2010-2011

**Submitted**

**BY**

**Mr. Archit Mehta** ID No. : 07CP619

**Mr. Raunakraj Patel**  ID No. : 07CP613

**Mr. Jatin Savaliya** ID No. : 07CP614

**Guided by**

**Ms. Hetal Gaudani**

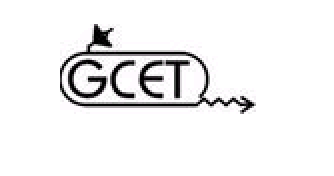
G. H. Patel Enginnering College,V V Nagar

Year 2010-2011

Department of Computer Engineering

### G. H. Patel College Of

### Engineering And Technology



G. H. Patel College of Engineering & Technology

Bakrol Road, Vallabh Vidyanagar-388120

Phone: (02692) 231651, 238490

[www.gcet.ac.in](http://www.gcet.ac.in)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Certificate**

We Archit Mehta (07 CP 619), Raunakraj Patel (07 CP 613), Jatin Savaliya (07 CP 614) , the student of Bachelor of Engineering (BE, Computer Engineering) program at G. H. Patel Engineering College, Vallabh Vidhyanagar, declare that the work presented in this report/thesis has not been copied from any other source.

The help, material, copy righted contents, contribution etc. from the other person/agencies/institution are acknowledged and given due credit by proper citation of reference wherever they appeared in this report.

Project Guide Head of the Department

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Ms. Hetal Gaudani Prof. Maulika S. Patel

\_\_\_\_\_\_\_\_\_\_\_\_\_

Date & Place.

**Acknowledgements**

A number of people have helped directly and indirectly with the work described in this thesis. In particular I would like to thank: Ms. Hetal Gaudani, our supervisor, for her guidance and patience, for many useful conversations and for always providing a different viewpoint.

Our special thanks to Mr. Deven Agravat, for his guidance for giving us direction throughout this project.

Our sincere thanks and gratitude to Prof. M. S. Patel, Head, Computer Engineering Dept, GCET Vallabh Vidyanagar.

I thankful to the entire staff member of COMPUTER DEPARTMENT of our college for providing us a helping arm during the project works by always showing us pros and cons of project.

I thankful to our college G. H. Patel Engineering College to allow us to carry out the project work at their esteemed organization and utilizing their provided facilities.

**Abstract**

Spam has serious negative on the usability of email and network resources. Spam is flooding the internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it. Negative effect of spam is 419 Scam and phishing are example of cyber crime. Because of IPv6 Spammers get freedom to send unsolicited bulk mail to millions of users. And despite the evolution of anti spam software, such as spam filters and spam blockers, the negative effects of spam are still being felt by individuals and businesses alike. . To prevent this advance techniques are necessary. Our application divides e-mails in spam class and non spam class according to different attribute values of spam. [2]

**Table of Contents**

1. **Introduction and objectives. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6**
   1. What is “spam”?
   2. Definitions
   3. The History of Spam
   4. Spammers
   5. Emerging spam for social networking attacks
   6. The Costs of Spam
   7. Costs to the Spammer
   8. Costs to the Recipient
   9. Spam and the Law
   10. Type of Spam
   11. Technique to send spam
   12. Spam filtering methods
2. **Understanding of algorithm. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21**
   1. Categories of algorithm
   2. Understanding C4.5
   3. How to deal with spam dataset
   4. Advantage of ID3:
   5. Disadvantage of ID3
   6. Advantage of C4.5
3. **Conclusion. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33**
4. **Future work. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .34**
5. **Reference. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .35**
6. **Web Reference. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .36**
7. **Introduction & Objectives**

Most e-mail readers spend a non-trivial amount of time regularly deleting junk e-mail (spam) messages, even as an expanding volume of such e-mail occupies server storage space and consumes network bandwidth. An ongoing challenge, therefore, rests within the development and refinement of automatic classifiers that can distinguish legitimate e-mail from spam. Some published studies have examined spam detectors using Naive Bayesian approaches and large feature sets of binary attributes that determine the existence of common keywords in spam, and many commercial applications also use Naive Bayesian techniques.

Spammers recognize these attempts to prevent their messages and have developed tactics to circumvent these filters, but these evasive tactics are themselves patterns that human readers can often identify quickly. This work had the objectives of developing an alternative approach using a neural network (NN) classifier brained on a corpus of e-mail messages from several users. The features selection used in this work is one of the major improvements, because the feature set uses descriptive characteristics of words and messages similar to those that a human reader would use to identify spam, and the model to select the best feature set, was based on forward feature selection. Another objective in this work was to improve the spam detection near 95% of accuracy using Artificial Neural Networks; actually nobody has reached more than 89% of accuracy using ANN.[R4]

* 1. **What is “Spam” ?**

Spam, in computing terms, means something unwanted. It has normally been used to refer to unwanted email or Usenet messages, and it is now also being used to refer to unwanted Instant Messenger (IM) and telephone Short Message Service (SMS) messages. Spam email is unwanted, uninvited, and inevitably promotes something for sale. Often the terms junk email, Unsolicited Bulk Email (UBE), or Unsolicited Commercial Email (UCE) are used to refer to spam email. Spam generally promotes Internet – based sales, but it also occasionally promotes telephone- based or other methods of sales too.

People who specialize in sending spam are called spammers. Companies pay spammers to send emails on their behalf, and the spammers have developed a range of computerized tools and techniques to send these messages. Spammers also run their own online businesses and market them using spam email.

The term “spam email” generally precludes email from known sources, regardless of however unwanted the content is. One example of this would be an endless list of jokes sent from acquaintances. Email virus, Trojan horses, and other malware (short for malicious software) are not normally categorized as spam either, although they share some common traits with spam. Emails that are not spam are often referred to as ham, particularly in the anti-spam community. Spam is subjective, and a message considered spam by one recipient may be welcomed by another.

Anti-spam tools can be partially effective in blocking malware; however, they are best at blocking spam. Special ant-virus software can and should be used to protect your inbox from other undesirable emails.

* 1. **Definitions**

The following definitions will be used throughout this work:

* Spam : Unsolicited commercial Email or UCE. It is any email that has not been requested and contains an advertisement of some kind.
* Ham : The opposite of spam email that is wanted
* False Negative : A spam email message that was not detected successfully.
* False Positive : A ham email message that was wrongly detected as spam.
  1. **The History of Spam**

Here are some important dates in the development of the internet:

* 1969: Two computers networked via a router
* 1971: First email sent using a rudimentary system
* 1979: Usenet (newsgroups) established
* 1990: The World Wide Web concept born
* 2004: The Internet is a major global network responsible for billions of dollars of commerce.

There is one omission from this time line:

1978: The first email spam sent.

Spam has been part of the Internet from a relatively early stage in its development. The first spam email was sent on May 3rd, 1978, when the U.S. Government funded Arpanet, as it was called then. The first spammer was a DEC engineer called Gary Thuerk who invited recipients of his email to attend a product presentation. This email was sent using the Arpanet, and caused an immediate response from the chief of the Arpanet, Major Raymond Czahor, at the violation of the non-commercial policy of the Arpanet.

Spam really took off in 1994 when an Arizona attorney, Laurence Carter, automated the posting of messages to many internet newsgroups (Usenet) to advertise his firm´s services. The resultant outcry from Usenet users included the coining of the term “spam”, when one respondent wrote “Send coconuts and cans of Spam to Cantor & Co.”. This sparked the beginning of spam as it is now experienced.

Spam email has increased in volume as the Internet has developed. In April 2009, PC Magazine reported that 98% of all email is spam.

* 1. **Spammers**

Typically, spammers are paid to advertise particular websites, products and companies, and are specialists in sending spam email. There are several well-known spammers who are responsible for a large proportion of spam and have evaded legal action. Individual managers of websites can send their own spams, but spammers have extensive mailing list and superior tools to bypass spam filters and avoid detection. Spammers have a niche in today´s marketing industry, and their clients capitalize on this.

Most spam emails are now sent from Trojanne ” computers, as reported in a press release by broadband specialist. The owners or users of trojanned computers have been tricked into running software that allows a spammer to send spam email from the computer without the knowledge of the user. The Trojan software often exploits security holes in the operating system, browser, or email client of a user. When a malicious website is visited, the Trojan software is installed on the computer. Unknown to users, their computer may become the source of thousands of spam email a day.

* 1. **Emerging spam for social networking attacks**

With individuals and businesses hooked on online social outlets, cybercriminals have taken notice and started using them for their gain. Beyond the common nuisances, such as wasted company time and bandwidth, malware and malicious data theft issues have presented serious problems to social networks and their users. Spam is now common on social networking sites, and social engineering—trying to trick users to reveal vital data, or persuading people to visit dangerous web links—is on the rise.

Social network logon credentials have become as valuable as email addresses, aiding the dissemination of social spam because these emails are more likely to be opened and trusted than standard messages. In many cases, spam and malware distribution are closely intertwined.

* 1. **The Costs of Spam**

Spam is very cheap to send, the cost are insignificant as compared to conventional marketing techniques, so marketing by spam is very cost-effective, despite very low rates of purchases in response. But it translates into major costs for the victim.

* 1. **Costs to the Spammer**

A report by Tom Galler, Executive Director of SpamCon Foundation, estimated that the cost to send a single spam email was as little as one thousandth of a cent, yet the cost to the recipient was around 10 cents.

The overhands in sending spam are low. The main costs are:

• **An Internet connection** : there are lots of flat-rate Internet Service Providers (ISPs) offering packages at around €20/month. A spammer doesn´t particularly need a Digital Subscriber Line (DSL) or cable modem service, a dial-up connection will also allow large quantities of spam to be sent. In fact, dial-up accounts are preferable, as spammer accounts are routinely shut down when complaints about spam are received. Dial-up accounts are easy to set up and can quickly activate within minutes, but DSL typically has a lead time of days.

• **Software** : specialist spam software is essential. A normal email client will restrict the number of spam messages that can be sent, and require the spammer to spend more time in front of the computer. Spammers usually write their own software, steal someone else´s, or buy one. A spammer with some technical knowledge and starting from scratch can have software ready after a week. To pay someone to develop that software would cost the spammer € 1000.

• **A mailing list** : most spammers will build up their own list of email addresses. For beginners, it is possible to buy a CD with 6 million email addresses on it for around €50. Ironically, these CDs are marketed via spam email. Email addresses that are guaranteed to be currently active sell for larger sums.

• **A web server** : this is an optional cost. It allows a spammer to deliver “web bug” images to validate their mailing list. Web bugs are discussed further in later chapters. Basic web hosting packages cost less than € 10 a month.

For less than € 1100, plus monthly cost of less than € 160, a spammer could have the software, internet connection, and a supply of addresses required to be operational.

A single computer can send thousands of emails an hour using dial-up. Spam varies, but a typical message size might be around 6,000 bytes. On a fast dial-up of around 50Kps, it would take one second to send this email to one recipient. It would take only a little longer to send it to 100 recipients. In other words, at least 3,600 emails can be sent in an hour. For smaller emails, the number sent per hour would be greater. The spammer needs to invest 15 minutes of their time, and the software will continue to send spam for many hours. With three phone lines, they could work for a total of an hour, and send approximately 10,000 emails an hour or 200,000 a day or more using DSL line.

* 1. **Costs to the Recipient**

The European Union performed a study into UCE in 2009. In the findings, it estimated the cost of receiving spam borne consumers and businesses to be around € 8 billon. These costs are partly incurred through lost productivity or time, partly in direct costs, and partly in direct costs incurred by suppliers, and passed on.

The cost of spam in a commercial environment is estimated to be as high as € 600 to 1000 per year, per employee. For a 50-person company, this cost could be as high as € 50.000 per year. Spam emails distract or take employees time and use disk space, processing power, and network bandwidth. Removing spam by hand is time consuming and laborious when there is a large amount of spam. In addition, there is a business risk, as genuine messages may be removed along with unwanted ones. Spam can also contain unsavory topics that some employee’s won´t tolerate.

* 1. **Spam and the Law**

In the USA, legislation proceeding on spam has been in progress since 1997. The latest legislation is the CAN-SPAM act (Bill number S.877) of 2009. This supersedes many state laws and is currently being used to prosecute persistent spammers. However, it is not proving a deterrent; the Coalition Against Unsolicited Commercial Email (CAUCE) reported in June 2008 that despite several high-profile lawsuits by the Federal Trade Commission (FTC) and ISPs, spam volumes were still increasing. The CAN-SPAM act is seen as weak on two counts: that consumers have to explicitly opt-out from commercial emails, and secondly, only IPSs can take action against spammers.

In Europe, legislation exists that makes spamming illegal. However, when Directive 2002/58/EC was passed in 2002, there were several problems with it. Business-to-business emails were excluded – a business could spam each and every account at any other business and stay within the law. Additionally, each individual member state has to pass its own laws and penalties for offenders. The law requires spammers to use opt-in emailing, where recipients have to explicitly request to receive commercial emails rather than the opt-out model proposed in the USA, where anyone can receive spam and has to request to be removed from mailing lists.

The Guardian, a UK newspaper, reported in June 2009 that gangs of spammers are moving their operations to the UK due to the leniency of the laws there. The maximum penalty they face in the UK is 5000 (pounds), while in Italy spammers face up to three years of imprisonment. Until June 2004, no one had been convicted under this act in the UK.

In Australia, the spam Act 2003 came into effect in April 2004. This makes spam illegal, using an opt-in model. Additionally, there have already been successful prosecutions for spamming in Australia using previous laws.

The internet is a multinational network and domestic legislation cannot reach to another country. A U.S-based spammer would be at risk of prosecution if it spammed U.S. citizens and advertised a product made and sold in the US. But the spammer from the Far East would be at very little risk of prosecution. Domestic legislation will not affect the volume of spam, but it may occasionally affect the types of products advertised via spam.

Spammers will often reroute spam via other nations, so spam is sent from the US to another country and relayed back to the US again. This makes it more difficult to trace the source of the email and to prosecute them. Many countries have no anti-spam laws and there is little or even no risk to the spammer. The blurring of geographical boundaries by the Internet does little to aid in tracing spam email to its source. Anti - spam in now moving towards tracking Spammers through other means. In May 2008, the New2008, the New York Times reported that the Direct Marketing Association is using paper trails in the real world to trace spammers in the virtual world with success.

* 1. **Type of Spam**
* Unsolicited bulk e-mail(UBE)

Unsolicited e-mail sent in large quantities.

* Unsolicited commercial e-mail(UCE)

This more restrictive definition is used by regulators whose mandate is to regulate commerce.

Spammers collect e-mail addresses form chat rooms, websites, customer lists, newsgroups, and viruses which harvest users address boos, and are sold to other spammers. They also use a practice as “e-mail appending” or “epending” in which they use known information about their target (such as a postal address) to search for the target’s email address.

* [Attachment spam](http://spamtrackers.eu/wiki/index.php/Attachment_spam)

This form of spam is distinctive as it entices victims to click on an attached file, such as a PDF file or electronic greeting cards.

* [Forum spam](http://spamtrackers.eu/wiki/index.php/Forum_spam)

The [Anti-Spam ACP](http://www.lithiumstudios.org/phpBB3/viewforum.php?f=25) module is highly effective

The KS BotDefender MOD was created for the [Kill Spammers Forums](http://www.thecarpcstore.com/phpbb2)

* [Geocities Spam](http://spamtrackers.eu/wiki/index.php/Geocities_Spam)

Geocities is a free web hosting service provided by Yahoo. It is predominantly used by hobbyist website creators to build fan sites, personal sites, photo galleries, etc. However starting in 2004 and continuing up until the present day, numerous spammers have started abusing Geocities websites in an automated fashion, building pages whose sole purpose is to redirect the user to the actual target website. This allows the spammer to get around numerous well-established spam filters, which would never block a geocities domain, since Yahoo is widely white listed on most block lists.

* [Greeting card spam](http://spamtrackers.eu/wiki/index.php/Greeting_card_spam) and [Image spam](http://spamtrackers.eu/wiki/index.php/Image_spam)

A type of [attachment spam](http://spamtrackers.eu/wiki/index.php/Attachment_spam), greeting card spam has victims click on what they believe to be an electronic greeting card from someone they know. The [URL](http://spamtrackers.eu/wiki/index.php/URL) is actually a means of transmission for computer contaminants.

In July 2007, Symantic reports that over 250 million customers were targeted with this type of spam.

* [Pump and dump](http://spamtrackers.eu/wiki/index.php/Pump_and_dump)

The term Pump and dump is a term used to define a particular type of stock fraud or stock market manipulation. It briefly means that a perpetrator buys cheap shares of a company, then makes somebody promote it. If there are enough stupid people to believe the hype and buy the stock, the price of the shares goes up, and the perpetrator sells them at a profit.

* 1. **Technique to send spam**
* Appending

Main article: E-mail appending

If a marketer has one database containing names, address, and telephone numbers of prospective customers, they can pay to have their database matched against an external database containing email addresses. The company then has the means to send email to persons who have not requested email address.

* Images Spam

Main article: Image spam

Image spam is an obfuscating method in which the text of the message is stored as a GIF or JPEG image and displayed in the email. This prevents text based spam filters form detecting and blocking spam messages. Image spam is currently used largely to advertise “pump and dump” stocks.

Often, image spam contains nonsensical, computer-generated text which simply annoys the reader. However, new technologies in some programs try to read the images by attempting to find text in these images. They are not very accurate, and sometimes filter out innocent images of products like a box that has words on it.

* Blank spam

Blank spam is spam lacking a payload advertisement. Often the message body is missing altogether, as well as the subject line. Still, it fits the definition of spam because of its nature as bulk and unsolicited email. Blank spam may be originated in different ways, either intentional or unintentionally.

* Blank spam can have been sent in a directory harvest attack, a form of dictionary attach for gathering valid addresses form an email service provider. Since the goal in such an attack is to use the bounes to separated invalid addresses form the valid ones, the spammer may with most elements of the header and entire message body, and still accomplish his or her goals.
* Blank spam may also occur when a spammer forgets or otherwise fails to add the payload when he or she sets up the spam run.
* Often blank spam headers appear truncated, suggesting that computer glitches may have contributed to this problem-from poorly-written spam software to the message body.
* Some spam may appear to be blank when in fact it is not. An example of this is the VBS. Davinia. B email worm which propagates through message that have no subject line and appears blank, when in fact it uses HTML code to download other files.
* Backscatter

Main article: Backscatter(e-mail)

Backscatter is a side-effect of e-mail spam, viruses and worms, where email servers receiving spam and other mail send bounce messages to an innocent party. This occurs because the original messsage’s envelope sender is forged to contain the e-mail address of the victim. A very large proportion of such e-mai is sent with a forged Form: header, matching the envelope sender. Since these messages were not solicited by the recipients, are substantially similar to each other, and are delivered in bulk quantities, they qualify as unsolicited bulk email or spam. As such, systems that generate e-mail backscatter can end up being listed on various DNSBLs and be in violation of internet service providers’’ Terms of Service.

Due to the alarming increase of the spam volume and its serious impact, providing vigilantly spam fighters has recently attracted considerable attention. In addition to regulations and legislations, several technical solutions including commercial and open-sources product have been proposed and deployed to alleviate this problem.

* 1. **Spam filtering methods**

Spam filtering methods fall into two broad categories:

1. Non-machine learning
2. Machine based learning
3. **Non machine learning based**

Most of the early anti-spam tools belonged to this category; for instance using a blacklist of known spammers, a white list of safe senders, or a human crafted list of keywords such as “Get Rich” either in the subject line or the message content. However, these static lists are simple to be goy around easily by spammers; for example by changing or spoofing the sender’s address or domain each time. Spammers have also learnt to deliberately avoid/misspell words or forge the content to bypass the spam filters. These methods also require periodic manual update and the likelihood to filter out a legitimate massage as spam is high which can be more serious than not filtering at all. According to the estimates by the British computer society(BCS), inaccurate anti-spam solutions may be responsible for wasting over five million working hours a year for users to check that legitimate massage were not mistakenly quarantined.

* **Heuristics**

**Benefits:**

The Heuristic mail filters are considered simple, highly accurate against the regular expression rules, and optimum in their speed of execution.

**Limitations:**

The Heuristic mail filters do not have intelligent learning capabilities (they do not adapt to emerging SPAM characteristics), they require administrator interference to 2 update the rule sets or rule sets need to be downloaded on regular basis, and they may also generate high rates of false positives as the sensitivity is increased.[wb3]

* **Signatures**

**Benefits:**

Based on high resistance of hash functions to collision, signature mail filters generates low rates of false positives.

**Limitations:**

The signature mail filters do not have intelligent learning capabilities (they cannot identify hashes for new SPAM emails), they require administrator’s interference to update the hash list of SPAM emails or the list needs to be fetched from a signature distribution server on a regular basis, and the filtering mechanism fails to detect pre-known SPAM emails if they are amended. Amendments to pre-known SPAM emails generate a different hash than the known one to the filtering system. Hence, the amended SPAM email will pass through the filtering system.[wb3]

* **Black Listing**

**Benefits:**

The black listing/white listing mail filters are considered simple, fast and easy to implement.

**Limitations:**

A main limitation of the black listing/white listing mail filters is that they can be easily fooled by spoofing the sender’s email address.[wb3]

* **Traffic Analysis**

**Benefits:**

The traffic Analysis mail filters are considered relatively complex. However, their mechanism proposes enhanced and fast mail filtering comparatively with actual analysis of email contents because they only analyze the SMTP logs.

**Limitations:**

The traffic Analysis mail filters do not have intelligent learning capabilities (they do not adapt to emerging SPAM characteristics). As of now, it is not possible for the filters to decide which of the characterizing attributes of email traffic are the most appropriate for a particular stream of email traffic.[wb3]

1. **Machine learning based**

Unlike traditional technique, machine learning based methods automatically analyze the content of received messages and build more robust models accordingly. Therefore, they can be more effective and dynamically updated to cope with spammers’ tactics. Several machine learning methods have been recently used for spam filtering, including

* **Bayesian**

**Benefits:**

The Bayesian mail filters have intelligent learning capabilities (ML filters) as well as enhanced filtering capacities based on content analysis. They allow email users to customize the filters based on the type of SPAM the users may receive. Additionally, the Bayesian mail filters are considered highly accurate.

**Limitations:**

The tokens in the Bayesian filters are formed of single words. Hence, combined words may pass from detection. Additionally, the filters lack the capability of analyzing consecutive words that form common SPAMming phrases in email contents, such as the phrase “special offer”. Instead, each word is analyzed separately. Certain consecutive words and combined words are frequently noticed in SPAM emails. Failing to recognize such words limits the filtering mechanism from detecting these types of SPAM emails. However, there are some algorithms that exist currently which help in analyzing permutations of single words, consecutive words, and words appearing within a distance of one another.[wb3]

* **Grey Listing**

**Benefits:**

Detecting SPAM emails using grey listing mail rejection is fast, simple and optimized. It utilizes the standard specification of the mail protocols. The blocking of emails is a feature of the mail servers (MTAs). Thus, it does not require the setup of additional hardware or software. It also provides a secure approach of combating SPAM emails by not receiving (rejecting) emails from SPAMming mail servers in the first place. The technique has proven to be effective in rejecting zombie machines that are used as SPAMming mail servers. The technique can be used as a premail filtering technique.

**Limitations:**

The grey listing mail rejection cannot be considered as a comprehensive anti-SPAM solution. Despite its effectiveness, it can cause major inconveniences for emails that require a prompt response. An example can be of websites that require user responses through email to complete their web registrations. The technique may hinder the users from completing their registration for some time. Another major inconvenience may occur if the period of waiting time given to the returned emails is not followed/not recognized by the source mail server. In such case, the mail will be sent after the expiration of the period. Thus, the mail will not reach the recipient. Consequently, the recipient mail server may consider the source mail server a SPAMmer and block it.[wb3]

1. **Understanding of algorithm**
   1. **Two categories:**

* Sequential covering algorithms
* Simultaneous covering algorithms.

Simultaneous covering algorithms like ID3 and C4.5 generate the entire rule set at once, while the sequential covering algorithms like AQ15 and CN2 learn the rule set in an incremental fashion.

* 1. **C4.5 Algorithm:**

C4.5 is a well known algorithm used for classifying datasets. It induces decision trees and rules from datasets, which could contain categorical and numerical attributes. The rules could be used to predict categorical values of attributes from new records. We discusses an overview of data classification and its techniques, the basic methods of C4.5 algorithm, the process and analysis of the results of an experiment, which utilizes C4.5 for classifying spam base dataset C4.5 performs well in classifying the dataset, but more data needs to be collected in order to gain useful rules.[R1]

C4.5 builds decision trees from a set of training data using the concept of [information entropy](http://en.wikipedia.org/wiki/Entropy_%28information_theory%29). The training data is a set S = s1,s2,... of already classified samples. Each sample si = x1,x2,... is a vector where x1,x2,... represent attributes or features of the sample. The training data is augmented with a vector C = c1,c2,... where c1,c2,... represent the class to which each sample belongs.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized [information gain](http://en.wikipedia.org/wiki/Information_gain) (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sublists.

This algorithm has a few base cases.

* All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
* None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
* Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.[wb4]

**Pseudo code:**

For generating decision tree from given training dataset

Input:

The training samples ; the set of candidate attributes, *attribute-list.*

Output:

A decision tree.

Method:

1. create a node N;
2. **if** *samples* are all of the same class, C **then**
3. return *N* as a leaf node labeled with the class C;
4. **if** *attribute-list* is empty **then**
5. return *N* as a leaf node labeled with the most common class in *samples*;//majority voting
6. select *test-attribute*, the attribute among *attribute-list* with the highest
7. label node N with *test-attribute*;
8. **for each** known value *ai* of *test-attribute;*
9. grow a branch from node N for the condition *test-attribute = ai;*
10. let *si* be the set of samples in *samples* for which *test-attribute = ai;* // a partition
11. **if** *si* is empty **then**
12. attach a leaf labeled with the most common class in *samples*;
13. **else** attach the node returned by Generate\_decision\_tree (*si, attribute-listtest-attribute);*

=====================================================================[R3]

**Basic Understanding:**

C4.5 follows a “divide-and-conquer” strategy to build a decision tree through recursive partitioning of a training dataset, and then translates the tree into an equivalent set of rules, one rule for each path from the root to a leaf of the tree. In order to build a decision tree, the following steps are performed recursively:

* If the training examples belong to the same class c or some stop criteria is met, then the decision tree for the training dataset is a leaf node labeled by the class c.
* Otherwise, select a test on a single attribute whose values best separate the training set into subsets according to a splitting criterion (the best separation is one where all examples in a subset belong to the same class), construct the decision tree by a node labeled with the corresponding attribute followed by the child nodes of the outcomes of the separation. 10
* Repeat the same procedure on each child node that contains a subset of the training examples, to build the branches of the decision tree.[R2]

**Example:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outlook** | **Temperature** | **Humidity** | **Windy** | **Play** |
| Sunny | Hot | High | False | Don’t Play |
| Sunny | Hot | High | True | Don’t Play |
| Overcast | Hot | High | False | Play |
| Rainy | Mild | High | False | Play |
| Rainy | Cool | Normal | False | Play |
| Rainy | Cool | Normal | True | Don’t Play |
| Overcast | Cool | Normal | True | Play |
| Sunny | Mild | High | False | Don’t Play |
| Sunny | Cool | Normal | False | Play |
| Rainy | Mild | Normal | False | Play |
| Sunny | Mild | Normal | True | Play |
| Overcast | Mild | High | True | Play |
| Overcast | Hot | Normal | False | Play |
| Rainy | Mild | High | True | Don’t Play |

Figure a[R2]

Above explains the decision tree building process by C4.5. The process begins by choosing an attribute to split the dataset into small subsets with the goal of having most of the examples in the subset belong to a single class. Heuristic splitting criteria are used to choose the best attribute to split the training examples into smaller subsets. Some of the popular splitting criteria include information gain, information gain ratio, Chi- square test, and gini index. The information gain of an attribute A relative to a set of examples S is defined as:

Entropy(S) is the entropy of S, which characterizes the purity of an arbitrary set of examples. If the class attributer has k different values, the Entropy(S) is defined as:

Entropy(S) = - ∑kc=1 nc/N log2 nc/N

Where :

Nc is the number of examples for the cth class

N=∑kc=1 nc is the total number of examples in the dataset.

Gain:

Gain(S,A) = Entropy(S) - ∑v Є Values(A)|Sv|/|S| Entropy(Sv),

Where:

Values(A) is the set of all possible values for A.

Sv is a subset of S for which attributer A has a value v

The higher the information gain, the more suitable the attributes is for choosing as the splitting attribute.

The higher the entropy, the k values are more evenly distributed across the dataset. Consider the weather dataset. The class attribute ‘play’ has two different values ‘play’ and ‘ Don’t play’ and ted dataset has 14 instance, out of which 9 instances have ‘play’ as their class attribute value, and 5 instances have ‘Don’t play’. The entropy for the entire dataset is

Entropy = -9/14 log2(9/14) -5/14log2(5/14)

= 0.940

Consider the first attribute ‘outlook’. The attribute has three different values in the dataset, ‘Sunny’,’Overcast’,’Rainy’

The Entropy of the subset with ‘outlook = sunny ’ is;

Entropy = -2/5 log2(2/5) -3/5log2(3/5)

= 0.970

The Entropy of the subset with ‘outlook = overcast ’ is;

Entropy = -4/4 log2(4/4) -0/4log2(0/4)

= 0

The Entropy of the subset with ‘outlook = rainy ’ is;

Entropy = -2/5 log2(2/5) -3/5log2(3/5)

= 0.970

So, the information gain for the attribute ‘outlook’ is;

Gain = 0.940-(5/14\*0.970+4/14\*0+5/14\*0.970)

= 0.247

Similarly, the information gain for the other three attributes is’

Gain(temperature) = 0.29

Gain(humidity) = 0.152

Gain(windy) = 0.048

Since the attribute ‘outlook’ has highest information gain, it is the first chosen to split the training set. The process is continued recursively until a stop criterion is met. Examples of stop criteria include the event that the depth of a number of examples in a subset for further partitioning is reached

.

Although information gain gives quite good results, it has a serious deficiency in that it favors attributes with many values over those with few values. An alternative criterion, information gain ratio which is robust and constantly provides a better choice of attributes than that provided by the information gain criterion is used in C4.5. The information gain ratio is defined as:

GainRatio(S,A) = Gain(S,A)/ SplittingInfo(S,A)

Where splitting information is used to penalize attributes with too many values, and is defined as:

SplittingInfo(S,A) = ∑ki=1 |Si| / |S| log2 |Si| / |S|,

Where i S is a subset of S (training dataset) that has the ith value of the attribute A. As an extreme example, an attribute may be ‘SSN’ which is unique for each example. The information gain for ‘SSN’ would be high, but the information gain ratio would be low.

The splitting information calculated for the weather dataset is as follows:

SplittingInfo(outlook) = -5/14 log2(5/14) - 4/14log2(4/14) - 5/14 log2(5/14)

= 1.577

SplittingInfo(temperature) = -4/14log2(4/14)-6/14log2(6/14)-4/14 log2(4/14)

= 1.576

SplittingInfo(humidity) = -7/14 log2(7/14) - 7/14log2(7/14)

= 1.0

SplittingInfo(windy) = -8/14 log2(8/14) - 6/14log2(6/14)

= 0.985

Therefore the gain ratio for the four attributes is:

gain ratio (outlook) = 0.247/1.577

= 0.156

gain ratio (temperature) = 0.029/1.556

= 0.018

gain ratio (humidity) = 0.152/1.0

= 0.152

gain ratio (windy) = 0.048/0.985

= 0.0487

For the weather dataset in, the attribute ‘outlook’ has the highest information gain, so, it is chosen to split the dataset.

Temperature and humidity could have been continuous attributes and would have been handled by C4.5 through a discretization scheme. In fact, Table shows the weather dataset with numerical attributes. C4.5 converts these numeric attributes into nominal attributes by sorting the training examples, producing a sequence of class values as shown in figure below

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 64 | 65 | 68 | 69 | 70 | 71 | 72 | 75 | 80 | 81 | 83 | 85 |
| Yes | No | Yes | Yes | Yes | No | No | Yes | No | Yes | Yes | No |

Numeric values for the temperature attribute in the weather dataset.

Weather dataset with numeric attributes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Outlook** | **Temperature** | **Humidity** | **Windy** | **Play** |
| Sunny | 85 | 85 | False | Don’t Play |
| Sunny | 80 | 90 | True | Don’t Play |
| Overcast | 83 | 78 | False | Play |
| Rainy | 70 | 96 | False | Play |
| Rainy | 68 | 80 | False | Play |
| Rainy | 65 | 70 | True | Don’t Play |
| Overcast | 64 | 65 | True | Play |
| Sunny | 72 | 95 | False | Don’t Play |
| Sunny | 69 | 70 | False | Play |
| Rainy | 75 | 80 | False | Play |
| Sunny | 75 | 70 | True | Play |
| Overcast | 72 | 90 | True | Play |
| Overcast | 81 | 75 | False | Play |
| Rainy | 71 | 80 | True | Don’t Play |

Figure b[R2]

C4.5 uses the mid-point between two numeric values as the branching value. For example, 71.5 is considered as a branching value, since it is the mid-point of 71 and 72. The information gain is calculated for each of the branching points (split points) in a similar manner to the nominal attributes and the branching point with the highest information gain is chosen to split the dataset. For example, in the numeric weather dataset in Table , if ‘71.5’ is chosen as a split point, the test ‘Temperature <= 71.5’ produces four ‘yes’ values and two ‘no’ values for the class attribute. The Test ‘Temperature > 71.5’ produces five ‘yes’ values and three ‘no’ values.

The Entropy of the subset with ‘Temperature <= 71.5’ is:

Entropy = -2/6 log2(2/6) -4/6log2(4/6)

= 0.918

The Entropy of the subset with ‘Temperature > 71.5’ is:

Entropy = -3/8 log2(3/8) -5/8log2(5/8)

= 0.954

So, the information gain for the attribute ‘Temperature’ is:

Gain = 0.940-(6/14\*0.918+8/14\*0.954)

= 0.0014

Figure b shows the decision tree generated by the C4.5 algorithm for the weather dataset with numerical attributes. This decision tree is very similar to the decision tree for the weather dataset with nominal attributes shown in Figure 2.3 except for the node labeled with the attribute ‘Humidity’. This node has two branches labeled ‘<= 75’ and ‘>75’ instead of ‘Normal’ and ‘High’.

* 1. **How to deal with spam dataset**

Information of Spambase.data file:

Number of total Instances = 4601

Number of total Spam Instances = 1813 (39.4% are Spam)

Number of total Non Spam Instances = 2788 (60.6% are Non Spam)

Information of Spambase.naem file:

Number of Attributes = 58 (57 continuous, 1 nominal class label) [wb2]

**Attribute Information**

The last column of 'spambase.data' denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail. Most of the attributes indicate whether a particular word or character was frequently occuring in the e-mail. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters.

48 continuous real [0,100] attributes of type word\_freq\_WORD = percentage of words in the e-mail that match WORD,

i.e. 100 \* (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non alphanumeric characters or end-of-string.

continuous real [0,100] attributes of type char\_freq\_CHAR = percentage of characters in the e-mail that match CHAR,

i.e. 100 \* (number of CHAR occurences) / total characters in e-mail

1 continuous real [1,...] attribute of type capital\_run\_length\_average = average length of uninterrupted sequences of capital letters

1 continuous integer [1,...] attribute of type capital\_run\_length\_longest = length of longest uninterrupted sequence of capital letters

1 continuous integer [1,...] attribute of type capital\_run\_length\_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail

1 nominal {0,1} class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0),

i.e. unsolicited commercial e-mail.

Missing Attribute Values: None

First we will pass the whole dataset, then we generate the random number that is 40% of spam and 40% of non spam. We write into another file that is called training dataset and other 60% we save it into another file called testing dataset. Then we apply C4.5 on training dataset and generate the rules. Then we test this rules through given process to check the accuracy of that generated rules.[wb2]

* 1. **Advantage of ID3:**
* Understandable prediction rules are created from the training data.
* Builds the fastest tree.
* Builds a short tree.
* Only need to test enough attributes until all data is classified.
* Finding leaf nodes enables test data to be pruned, reducing number of tests.
* Whole dataset is searched to create tree.
  1. **Disadvantage of ID3**
* Data may be over-fitted or over-classified, if a small sample is tested.
* Only one attribute at a time is tested for making a decision. Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.  
  1. **Advantage of C4.5**
* Avoiding over fitting the data

Determining how deeply to grow a decision tree.

* Reduced error pruning.
* Rule post-pruning.
* Handling continuous attributes.

e.g., temperature

* Choosing an appropriate attribute selection measure.
* Handling training data with missing attribute values.
* Handling attributes with differing costs.
* Improving computational efficiency.

1. **Conclusion**

|  |  |
| --- | --- |
| **Efficiency(in %)** | **Number Of Rules** |
| 0-35 | 8 rules |
| 36-50 | 10 rules |
| 51-75 | 12 rules |
| 76-90 | 10 rules |
| Above 90 | 68 rules |
| Not classified | 57 rules |

After testing all the rules we get average 90.35365% efficiency of C4.5 algorithm.

1. **Future work**

By doing pruning we can remove the node that don’t contribute and doesn’t make any change in efficiency.

We can speed up the process and reduce the complexity of indentifying e-mail as a spam

We can also implement CN2 for the same which may more accurate and more faster.

1. **Reference**

[R1] C4.5 machine learning - J Ross Quinlan

[R2] Rule induction using ant colony optimization for mixed variable attributes - Santhosh Swaminathan.

[R3] Towards the use of c4.5 algorithm for

classifying banking dataset - Veronica S. Moertini

[R4] Spam Classification Using Machine Learning Techniques - Sinespam

Jose Norte Sosa

1. **Web Reference**

[wb1] <http://www.techrepublic.com/blog/security/ipv6-blacklisting-will-it-be-harder-to-fight-spammers/5195?tag=nl.e036>.

[wb2] <http://archive.ics.uci.edu/ml/datasets.html>

[wb3] <http://www.spam.gov.sa/English/References/Articles%20-%20English/Article%205%20-%20Current%20Email%20Anti-SPAM%20Technical%20Solutions.pdf>

[wb4] <http://en.wikipedia.org/wiki/C4.5>