

# Spam Filtering

*Proseminar Data Mining*

Johannes Neutze  
Fakultät für Informatik  
Technische Universität München  
Email: neutze@in.tum.de

**Kurzfassung**—Das Internet ist in den letzten Jahren drastisch gewachsen. Nicht nur seine Infrastruktur hat stark zugenommen, sondern auch die Menge an versendeten Bytes ist deutlich angestiegen. Das Internet ist heute eines der wichtigsten Medien zur Kommunikation, jedoch werden nicht nur nützliche Informationen auf diesem Weg verbreitet, denn auch dubiose Unternehmer haben die Mail für sich entdeckt. So kommt es, dass über die Jahre ein immer größerer Teil des Mailverkehrs aus ungewollten Angeboten besteht. Legitime von unbrauchbaren Mails zu unterscheiden kostet jedoch Zeit und Zeit kostet Geld. In dieser Arbeit wird ein Ansatz behandelt, der unerwünschte Mails auf Basis der Entscheidungstheorie automatisch sortiert. Diese Methode, der Bayes Klassifikator, soll, nach einmaligem Einrichten durch den Benutzer, selbstständig immer exakter werden und Neuerungen erlernen. Zusätzlich werden für eine größere Genauigkeit domänen spezifische Attribute und händische Regeln hinzugefügt.

**Schlüsselworte**—Bayes Theorem, Bayes Filter, Bayes Klassifikator, Spamfilter

## I. EINLEITUNG

Der erste Vorgänger der Mail wurde 1971 zwischen zwei Computern in einem Netzwerk versendet. Als später aus lokalen Netzwerken regionale und schließlich globale Netzwerke wurden, gewann die Mail an Bedeutung. Je mehr Nutzer am Internet angeschlossen waren, desto mehr Nachrichten wurden verschickt. Der Umfang des Mailverkehrs explodierte geradezu und trotz des jetzt schon riesigen Anstiegs ist noch keine Ende des Wachstums in Sicht.

Aufgrund der Wirtschaftlichkeit (schnelle und billige Interaktion), ist die Mail einer der meist genutzten Kommunikationswege der heutigen Zeit. Auch Onlinevermarkter haben die Mail für ihre Zwecke entdeckt. Sie nutzen die Mail, da sie mit wenig Aufwand und Kosten sehr schnell sehr viele Leser erreichen können. Spammails werden ohne genaues Ziel an möglichst viele Empfänger versendet und sie versprechen meist in ihren Nachrichten Wunderprodukte, schnellen und einfach erreichbaren Reichtum oder werben für pornographische Internetseiten in ihren Nachrichten. Deswegen wächst mit der Zahl an das Internet angeschlossener Nutzer auch die Zahl der sogenannten Spammails. Spam, oder auch Junk, sind unerwünschte Nachrichten, die dem Empfänger ohne sein Verlangen zugeschickt werden. Laut dem Antivirusherstellers Symantec waren im Jahre 2010 beachtliche 89,1% des gesamten Mailverkehrs Nachrichten, die als Spam eingestuft werden. [1] Spam belästigt mit seinem Inhalt, beansprucht riesige Mengen an Speicher auf den Servern der Mailanbieter und vor allem kostet er Zeit. Würden eine Millionen Empfänger

auch nur eine Sekunde zum Löschen einer Junknachricht aufwenden, würde sich dies zu einem Schaden von 11.5 Tagen Arbeitszeit aufsummieren. Aus diesem Grund ist eine automatische Filterung von Spam dringend nötig.

Zunächst wurde das Problem mittels logischen Filtern angegangen, jedoch erfordert das Erstellen von Regeln des Filters erfahrene Benutzer, da die Regeln oft erneuert werden müssen, weil Spam sich sehr schnell weiterentwickelt. Würde zum Beispiel die Phrase "Click" als Indikator für Spam benutzt werden, würden 79,7% des Spams als solcher erkannt und es gäbe nur 1,2% false positives. [2] False positives sind legitime Nachrichten, die jedoch als Spam markiert werden. Sie sind der worst case jedes Filters und müssen unbedingt vermieden werden, denn sie stellen Informationsverlust dar, der mit hohen Kosten verbunden ist. Aus diesem Grund ist ein einfacher, regelbasierter Filter nicht gut genug und es muss nach einer anderen Methode gesucht werden, den Nachrichtenverkehr zu trennen.

Heutzutage werden adaptive wahrscheinlichkeitstheoretische Filter, welche auf dem Bayes Klassifikator basieren, bevorzugt. Sie haben den Vorteil, dass sie besser mit den Veränderungen von Spam mithalten können, da sie aus neuen Nachrichten lernen und sich durch die Informationen weiterentwickeln. Durch das Lernen aus persönlichen Mails des Nutzers, kann er sich auch den individuellen Präferenzen anpassen, was eine höhere Genauigkeit im Einsatz bei dem jeweiligen Benutzer zur Folge hat. Ein weiterer Vorteil ist, dass er sich bei der Sortierung nicht auf binäre Entscheidungen stützt, wie Regeln, die nur 0 oder 1 ergeben können, sondern mehrere Faktoren, in Form von Wahrscheinlichkeiten, in seine Entscheidung mit einbezieht. Dies kann zwar zu Konflikten führen, in dem er die Mail in Folge der Entscheidung als legitim einstuft. Jedoch ist dies nicht so schwerwiegend wie wenn er ein false positive erzeugt, was bei Regeln schneller der Fall ist.

So kommt die mathematische Methode, welche auf dem Bayes Theorem basiert, mit dem Problem der Klassifizierung zusammen, was gemeinsam den Bayes Klassifikator ergibt, welcher genauer im zweiten Kapitel behandelt wird. Das dritte Kapitel befasst sich mit der Verbesserung des Filters, die durch Erweiterung mit Domänen spezifischen Attributen und händischen zu erstellenden Regeln verwirklicht werden. So besteht der Filter nicht nur aus dem Textkörper, sondern bezieht auch weitere Faktoren in den Aufbau mit ein, welche sich auf diese Weise mit weiterentwickeln können. Im folgenden wird der Aufbau des Filters im Detail beschrieben.

## II. GRUNDLAGEN

Dieser Abschnitt beschäftigt sich mit den mathematischen Grundlagen des Bayes Klassifikators. Zuerst wird die Basis, das Bayes Theorem, behandelt, welches das Bayes Netzwerk aufbaut. Mittels diesem Theorem wird der Klassifikator zusammengesetzt.

### A. Das Bayes Theorem

Das Bayes Theorem dient zur Berechnung von bedingten Wahrscheinlichkeiten.

$$P(A|B)$$

Als bedingte Wahrscheinlichkeiten wird die Wahrscheinlichkeit eines Ereignisses A unter der Bedingung, dass bereits Ereignis B eingetreten ist, verstanden. Das Bayes Theorem lautet:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

Auf diese Grundlage baut das Bayes Netzwerk auf.

### B. Das Bayes Netzwerk

Das Bayes Netzwerk besteht aus einer Wurzel, Blätter und gerichteten, azyklischen Graphen, die von einer Wurzel zu den Blättern, welche die Variablen darstellen, zeigen. Dabei stellen die Kanten die bedingten Abhängigkeit dar, Knoten die nicht verbunden sind, sind unabhängig voneinander. Jedes Blatt besitzt somit eine von der Wurzel abhängige Wahrscheinlichkeit.

Durch diese Aufteilung kann für jedes Wort  $x_i$  die bedingte

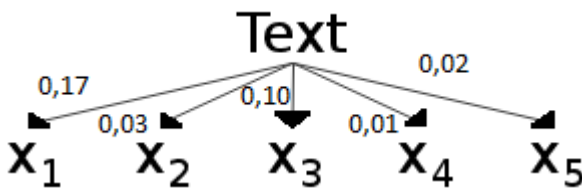


Abbildung 1. Ein Netzwerk, welches einem naiven Bayes Klassifikator entspricht

Wahrscheinlichkeit berechnet werden, mit der  $x_i$  auftritt, wenn zum Beispiel  $x_1$  aufgetreten ist. Weitert man dies jetzt auf mehrere bedingte Abhängigkeiten aus, kann zum Beispiel berechnet werden, welche Wahrscheinlichkeit vorliegt, wenn ein weiterer Faktor vorliegt. In diesem Beispiel ist die Wahrscheinlichkeit, dass eine Spammail das Wort "Sex" enthält, und die Wahrscheinlichkeiten, zu der das Wort Viagra vorkommt, wenn Sex auch vorkommt, und zu welcher Wahrscheinlichkeit Viagra in dem Fall in Spammnachrichten vorkommt. Außerdem ist gegeben, zu welcher Wahrscheinlichkeit eine Mail Spam ist, wenn keins, eines oder beide Worte enthalten sind. Daraus lässt sich dann die Wahrscheinlichkeit berechnen, zu der die Mail Spam ist, wenn beide Worte vorhanden sind:

$$\begin{aligned} &P(\text{Spam} = \text{true}, \text{Viagra} = \text{true}, \text{Sex} = \text{true}) = \\ &= P(\text{Sex} = \text{true}) \cdot P(\text{Viagra} = \text{true}, \text{Sex} = \text{true}) \\ &\cdot P(\text{Spam} = \text{true}, \text{Viagra} = \text{true}, \text{Sex} = \text{true}) = \end{aligned}$$

$$= 0,8 \cdot 0,99 \cdot 0,99 = 0,78$$

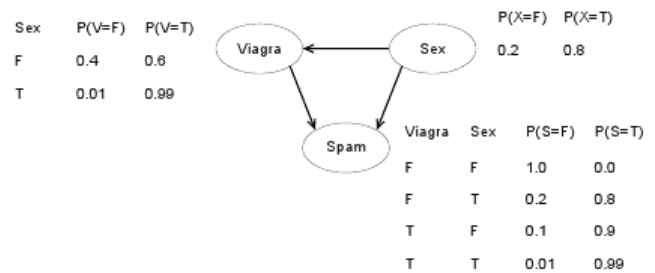


Abbildung 2. Beispiel für ein Bayes Netzwerk

## III. DER BAYES FILTER

Der Bayes Filter besteht aus einem, durch Domänen spezifischen Attributen und händischen Regeln, erweiterten Bayes Netzwerk. Der folgende Abschnitt geht genauer auf die Bestandteile des Filters ein.

### A. Der Bayes Klassifikator

Auf Basis des Bayes Netzwerkes wird der naive Bayes Klassifikator transformiert. Er ist im Grunde ein Bayes Netz, in dem die Geschwisterknoten unabhängig von einander sind, das heißt jedem Objekt, der Phrase, wird mit genau einem Attribut, der Wahrscheinlichkeit, die Klasse, legitime Mail oder Spam, zugeteilt. Um ihn zu Erstellen, werden zwei vorsortierte Sets an Mails benötigt. Sie werden zuerst in zwei Bayes Netzwerke für legitime Mails und Spam aufgeteilt, was in Form von zwei Tabellen realisiert wird. Die Tabellen enthalten die Phrase mit der Anzahl ihres Vorkommens. In einer dritten Tabelle werden die Wahrscheinlichkeiten zur Berechnung der Mail mit folgender Methode berechnet. [2]

$w$  : betrachtetes Wort

$legit(w)/spam(w)$  : Anzahl des Wortes in jeweiliger Klasse

$L/S$  : Anzahl der Mails

$P_{l/s}(w)$  : Wahrscheinlichkeit des Wortes in jeweiliger Klasse

$$P_l(w) = \min(1, 2 \cdot \frac{legit(w)}{L})$$

$$P_s(w) = \min(1, \frac{legit(w)}{S})$$

$$P(\text{Spam}|w) = \max(0,01, \min(0,99 \frac{P(s)}{P(l) + P(s)}))$$

$P(l)$ , beziehungsweise  $P(s)$ , ist die Wahrscheinlichkeit, zu welcher das Wort  $w$ , im Falle von  $P(l)$ , in einer legitimen Mail vorkommt. Zur Berechnung wird die Anzahl der Mails und nicht die der Wörter im jeweiligen Set genommen, um die Gewichtung nicht durch oftmals kurzen Spammnachrichten

zu verfälschen und einer daraus resultierenden Falscheinordnung vorzubeugen. Außerdem wird das Vorkommen eines Wortes in einer erwünschten Mail verdoppelt, damit zwischen häufigen und seltenen Wörtern besser unterschieden werden kann. Daraus kann dann  $P(\text{Spam}|w)$  berechnet werden, was die Wahrscheinlichkeit angibt, zu der ein Wort mit Spam in Verbindung gebracht wird. Dazu werden die bedingten Wahrscheinlichkeiten der Wörter aufsummiert. Somit erhält man eine Wahrscheinlichkeit, zu der die Nachricht Spam ist. [3]

$$P(x_1, \dots, x_i) = \prod_{j=1}^i (x_j | pa(x_j))$$

### B. Domänen spezifische Eigenschaften

Um den Einsatz des Filters noch zu verbessern, werden Domänen spezifische Eigenschaften ausgenutzt. Es gibt noch viele weitere Indizien für Spam, als nur auffällige Wörter. Zwar können Junknachrichten schon an charakteristischen Phrasen wie "FREE" oder vielen Ausrufezeichen erkannt werden, aber auch die Absenderdomain, .de oder .edu, sagen viel über die Nachricht aus. So können bereits im Vorfeld dem Filter spezielle Informationen hinzugefügt werden. Einschlägige Informationen, falls nicht im Kalibrierungsset vorhanden, werden dem Filter so ohne vorhandene Mails erlernt. Dazu wird eine händisch erstellte Liste von bekannten Spamphrasen, wie zum Beispiel "TEENS", "ONLY \$0.00" oder "be over 21", dem Filter hinzugefügt.// Um weitere Informationen aus einer Mail zu erhalten, wird nicht nur der Textkörper an sich, sondern auch der HTML-Code verwendet. Auf diese Weise werden HTML-spezifische Ausdrücke, wie zum Beispiel "ff0000", was für rote Schrift benötigt wird, welche in Spammails zur Erzeugung von Aufmerksamkeit eingesetzt wird, erkannt und dem Filter hinzugefügt. Auch kann die Absenderdomain extrahiert und hinzugefügt werden, was aufgrund ihres Informationsgehalts, da zum Beispiel kaum Spam von .edu Adressen verschickt wird, von großem Nutzen ist. Zusätzlich kann die Information, ob es eine Mail an eine Mailingliste oder nur an den Empfänger gesandt wurde, dem Filter erlernt werden. Des Weiteren können bekannte Spammer so direkt mit einer sogenannten Blacklist aussortieren werden. Adressen auf einer solchen Liste werden direkt als Spam markiert. Sie wurden davor aus einer alten Spammail entnommen und der Blacklist hinzugefügt. Diese Technik kann auch für eine Whitelist, eine Liste von Adressen, die ohne zu Filtern direkt in den Mailingang gehen, genutzt werden. Als Basis gilt in diesem Fall das Adressbuch des Nutzers. Die beiden Methoden haben aber nur einen geringen Nutzen, denn zum einen hat das Filtern einer Mail kaum Kosten, zum anderen schlägt das Verfahren nicht an, wenn ein Freund von einer noch unbekannten Adresse, zum Beispiel der Universitätsadresse, schreibt. Auch ist der Filter bei vertrauensvollen Absendern von einem gekaperten Mailservern, welchen Spammer in ihre Gewalt gebracht haben, machtlos. Manche Domänen spezifischen Eigenschaften sind nur in Verbindung mit einem Wahrscheinlichkeitstheoretischen Filter

effektiv, da sie bei einem regelbasierten Filter zu ungenau wären. Ein Beispiel ist die Auswertung des Anhangs. Spam hat fast immer die Information im Textkörper und nur selten einen Anhang. Probabilistische Filter können diese Eigenschaft nutzen, da er nicht nach strikten Regeln arbeitet, sondern nach Wahrscheinlichkeiten. Malwarenachrichten haben oft einen Anhang, den ein regelbasierter Filter nicht heraus filtern würde, da er davon ausgeht, dass Spam nur Text versendet und der Anhang ein starkes Argument zur Identifizierung von Spam wäre. Im Gegensatz dazu ist bei dem Wahrscheinlichkeitsfilter der Anhang nur eine weitere Variable zur Berechnung und er kann eventuelle Hinweise aus dem Inhalt der Mail ziehen und sie somit enttarnen. Ein weiteres Beispiel hierfür ist das Verhältnis zwischen Zeichen und Buchstaben. Zwar enthalten Junkmails oft viele Zeichen, wie Währungssymbole und Ausrufezeichen, jedoch kann hier keine klare Regel festgelegt werden, ab wie viel % eine Mail Spam ist. [3] Domänen spezifische Ergänzungen sind keine hundertprozentigen Entscheidungsträger, deswegen fließen sie nur als weiteres Objekt in den Filter ein.

### C. Benutzerdefinierte Einstellungen

Um undefinierte Fälle des Filters zu beseitigen, müssen noch vordefinierte Regeln erstellt werden, wie mit einem unbekannten Wort verfahren werden soll oder bei welchem Wert die Entscheidungsschwelle, ab welcher Wahrscheinlichkeit eine Mail als Spam eingestuft wird, liegen soll. Im Falle eines neuen Wortes, das während der Analyse einer neuen Mail auftaucht, muss eine Standardwahrscheinlichkeit eingeführt werden. Spam benutzt häufig den selben Wortschatz, deshalb sind neue Worte mit hoher Wahrscheinlichkeit neutral, wenn nicht sogar leicht positiv. Um dies zu berücksichtigen, wird diesen Worten die Wahrscheinlichkeit 0,4 zugeordnet. Auch der Wert, ab wann eine Nachricht als Junk definiert wird, muss festgelegt werden. Das Wichtigste ist, dass sich der Filter in der kommerziellen Nutzung keinen Fehler erlauben darf. Zwar liegen die Wahrscheinlichkeiten der meisten Mails mit 0,5 in der neutralen Zone, da sich gute und schlechte Phrasen gegenseitig aufheben. Jedoch gibt es auch Mails, die als false positives falsch eingeordnet werden, zum Beispiel zur Analyse weitergeleiteter Spam oder ein Artikel mit Neuigkeiten über Spam der viele ungute Wörter benutzt. Um dies auszuschließen muss die Schranke bei über 99% sein. [3]

### D. Einbeziehung der zusätzlichen Informationen

Um einen funktionsfähigen Filter mit einer guten Performance zu erhalten, müssen wir dem naiven Bayes Klassifikator die beiden anderen Methoden hinzufügen. Dazu werden die Erkenntnisse aus der Domänen spezifischen Analyse als weitere Objekte dem Netzwerk, beziehungsweise den Tabellen, hinzugefügt und die benutzerdefinierten Regeln implementiert. Um eine möglichst hohe Effektivität zu erhalten, werden nur die extremsten Werte in den Filter mit einbezogen.

## IV. ECHTZEIT SZENARIO

Dieser Teil der Ausarbeitung beschäftigt sich mit dem Ablauf des Filters in einem echtzeit Szenario. Dafür wird ein,

aus den oben definierten Punkten zusammengesetzter, Bayes Filter zuerst erstellt und dann an einer Mail getestet.

#### A. Erstellen des Filters

Zur Erstellung des Filters wird zunächst eine möglichst große Zahl an Mails benötigt, die als erste Anhaltspunkte für den Filter fungieren. Optimal ist es, den gesamten Mailverkehr des Nutzers zu bekommen, den er bereits in legitime und Spammessages unterteilt hat, um eine möglichst genaue Klassifizierung sicherzustellen, weshalb davor keine Mails gelöscht werden sollten. Um diesem Informationsverlust in Zukunft auch vorzubeugen, gibt es den Ansatz von zwei verschiedenen Löschvorgängen. Es gibt dazu zwei Löschmethoden für Nachrichten, "Löschen" und "Löschen als Spam", um falschem Entfernen von Nachrichten vorzubeugen. Auf diese Weise wird zwischen uninteressantem Informationsaustausch, welcher nicht des Speicherns würdig ist, und falsch einsortiertem Spam unterschieden. Die durch den Vorgang gewonnen Informationen fließen direkt in den Filter ein, um ihn für die Zukunft zu verbessern. Auf diese Weise geht keine Information verloren.

Das Initiieren des Filters durch den persönlichen Mailstrom hat gegenüber dem Erstellen aus einer vordefinierten Liste von Phrasen und deren Wahrscheinlichkeiten viele Vorteile. Zum Einen lernt der durch die persönlichen Mails basierte Filter die privaten Wortvorlieben, die dadurch als positiv in den Filter eingehen, sowie ab welcher Grenze der Benutzer Mail als Spam ansieht, was den Filter individuell effektiver macht. Zum anderen können Spammer sich so nicht auf spezielle Muster, zum Beispiel die immer gleichen Kriterien aus der vorgefertigten Liste, welche womöglich weit verbreitet sind, einstellen, was ihnen die Arbeit erschwert, da sie nicht wissen, welches Wort welche Gewichtung besitzt. Außerdem ist ein auf privaten Mails basierender Filter stärker gegen individuellen Spam, da er von einem Update und regionalen Spamunterschieden, von denen der Ersteller der Liste eventuell gar nichts mitbekommt, unabhängig ist.

Wird der Filter nur gestartet, teilt er die Phrasen zuerst auf zwei Tabellen auf und berechnet zuerst die einzelnen Wahrscheinlichkeiten  $P(l)$  und  $P(s)$ . Dann berechnet er mit Hilfe dieser beiden Wahrscheinlichkeiten die Spamwahrscheinlichkeiten  $P(\text{Spam}|w)$  für die dritte Tabelle. So erhält man die Basis des Filters. Kommen nun neue Mails, werden zuerst analysiert und die Worte je nach Entscheidung den einer der beiden Tabellen hinzugefügt. Dadurch entwickelt sich der Filter ständig weiter. Das selbe gilt auch für Wörter aus Mails, die nachträglich per "Löschen als Spam" entfernt werden.

#### B. Verarbeitung einer Mail

Im folgenden wird ein möglicher Aufbau des Filters dargestellt und das Verhalten nach dem Erhalt einer neuen Mail beschrieben.

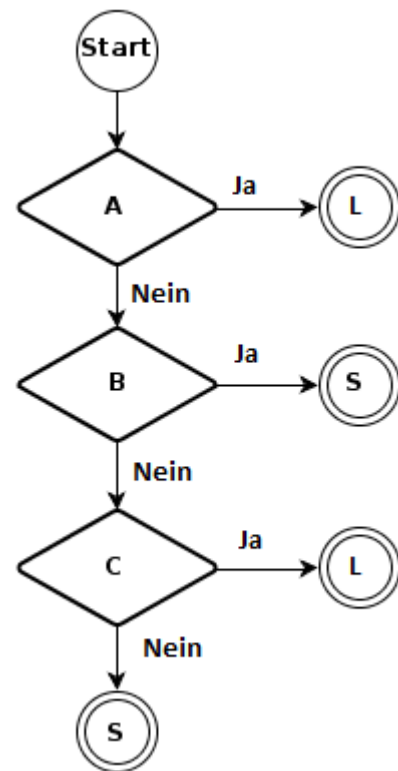


Abbildung 3. Ablauf des Filtervorganges

Zuerst wird in A geprüft, ob die Absenderadresse auf der Whitelist steht, wenn sie enthalten ist, wird die Nachricht direkt in den Posteingang L verschoben. Steht sie nicht auf der Whitelist, wird auf der Blacklist gesucht, ob sie dort referenziert ist. Ist sie darauf enthalten, wird die Mail als Spam klassifiziert und in den Spamordner S verschoben. Ist sie auf keiner der beiden Listen, wird mit dem Filter berechnet, ob die Wahrscheinlichkeit, dass die Mail Spam ist, niedriger als 99% ist, je nach Ergebnis wird die Nachricht daraufhin in den entsprechenden Ordner verschoben.

#### V. ZUSAMMENFASSUNG UND AUSBLICK

In dieser Arbeit wurden die Grundlagen und der Aufbau des Bayes Filters dargelegt. Er hat zwar Nachteile, doch die Vorteile überwiegen. Die Zeit, die zum Sortieren der Mails hätte aufgewendet werden müssen wäre gigantisch. Rechnet man die eingesparte Zeit hoch, so ergibt sich ein riesiger wirtschaftlicher Gewinn. Auch die Zeit, die durch nicht vorhandene Wartung frei ist, ist nicht zu vernachlässigen. Ein weiterer Vorteil dieses Filters ist seine lange Lebensdauer, die er durch die ständige Weiterentwicklung erreicht, welche durch die Verarbeitung der gesammelten Informationen aus den erhaltenen Mails vorangetrieben wird. Jedoch ist Weiterentwicklung auch gleichzeitig sein größter Feind. Das Internet lebt von dem Rennen zwischen Programmierern. Auf der einen Seite die Entwickler des Filters, auf der anderen Seite die Spammer. Es ist ein ständiges Ringen um den besten Algorithmus, der den anderen wieder in den Schatten stellt, genauso wie bei

der Entwicklung von Viren. Sobald ein neuer Virus entdeckt wird, setzen die Antiviren-Firmen alle daran, einen effektiven Schutz dagegen zu entwickeln, jedoch sobald dieser fertig ist, haben die Hacker bereits einen mächtigeren, besseren Virus erstellt. Deswegen kann der Filter auch per eingespielten Listen auf dem neusten Stand gehalten werden, denn ein wenig investierte Zeit ist besser als ein false positive. Das Hauptproblem bleibt dennoch, da Spam nicht genau definiert ist. Junk ist subjektiv und wird von jedem Nutzer anders wahrgenommen, für die einen ist erst aufdringliche Werbung oder graphische Pornographie Spam, für die anderen schon ein uninteressanter Newsletter. Deshalb ist es nicht möglich, den perfekten Spamfilter zu erstellen. Zwar versucht man mit selbst erstellten Listen und der persönlichen Kalibrierung dem entgegen zu wirken, jedoch kann das Empfinden von Tag zu Tag schwanken und somit ist der Filter nicht immer gut genug. Im Großen und Ganzen kann jedoch gesagt werden, dass die gewonnene Zeit fast alle Nachteile aufwiegt. Aber es darf auf keinen Fall ausser Acht gelassen werden, dass selbst eine 99,9% genaue Erkennung keinen kompletten Schutz vor false positives gibt. Denn wenn er nur eine Mail in der gesamten Laufzeit, zum Beispiel einen Lottogewinn, als false positive eingeordnet wird, hat der Filter mehr Schaden als Nutzen angerichtet.

Deswegen sollte weiterhin Zeit darauf verwendet werden, Spam Filter weiterzuentwickeln. Mit weiteren Verfahren kann die Grenze, ab wann eine Nachricht als Junk einzustufen ist, immer klarer definiert werden, um so die Effizienz und das Optimum immer weiter zu steigern.

#### LITERATUR

- [1] Symantec Corporation, MMessageLab Intelligence: 2010 Annual Security Report", 2011.
- [2] [1] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, A Bayesian approach to filtering junk e-mail," in Learning for Text Categorization: Papers from the 1998 workshop, 1998, vol. 62, no. Cohen, pp. 98-105.
- [3] P. Graham, "Hackers & Painters: big ideas from the computer age", O'Reilly, 2010, pp. 121-129.