Intent Based Filtering:
A Proactive Approach Towards Fighting Spam

By Priyanka Agrawal

Abstract-**The cyber security landscape has changed considerably over the past few years. Since 2003, while traditional malware threats such as Blaster and Welchia have continued to plague the networks and systems of enterprises around the world, new types of security risks such as spyware and adware have also emerged as significant IT threats. As the effects on both client systems and enterprise networks from spyware and adware have increased, IT administrators and managers are faced with a mounting challenge. The impact of spyware and adware can lead to decreased productivity, increased helpdesk calls, loss of privacy, and potential legal liability. However, there has been no broad agreement as to the definition and proper handling of applications referred to as spyware and adware, leading to significant confusion as organizations seek solutions to address this growing challenge.**

This paper discusses Intent Based Filtering of spam, where the intention of the mail is identified. Intent-Based Filtering employs artificial intelligence technology to recognize spam messages like a human reader would, and quickly eliminate them from the email stream.

I.INTRODUCTION

Intent-Based Filtering represents a true technological breakthrough in the proper identification of unwanted junk email, or "Spam". To date, spam-filtering techniques have fallen into several categories: blacklists, rules and heuristics, or more recently, Bayesian filters and signature technologies.

However, none of these existing techniques has done an adequate job of solving the Spam problem. For that reason, the adoption of these technologies has been limited and the results mixed. Strong predictions have been made for the future market growth of spam filtering, but still these predictions are based on the limited nature of existing solutions.

Existing techniques result in a constant battle of misclassification – either in the form of a legitimate message being mistaken for spam (false positive) or a spam message being mistaken as legitimate (false negative). But one thing remains the same: a human can identify a spam message almost immediately.

What is it about spam that allows a human to identify it so quickly and reliably? The answer is message intent. Regardless of who is sending a message, what specific words they use, or what servers those messages originate from, those would-be spammers cannot change the intent of the message they are sending. Identifying message intent is what Intent-Based Filtering (IBF) is all about, and it represents a shift in the way anti-spam solutions should be thought about and deployed, affecting the overall marketability of an anti-spam solution.

IBF represents an innovative proactive approach to dealing with the spam problem. Other approaches tend to be reactive, by creating sender blacklists, adding rules, creating signatures, or modifying statistics. By using a proactive system, IBF is able to identify spam even if that type of spam message or that sender has never been seen before.

II. THE VARIOUS APPROACHES TO SPAM FILTER AND THEIR FAILURES

Creating a system that understands the nature and subtleties of human language is not an easy task. It is these intricacies that make the spam problem so difficult to solve.

Consider the following statement from AT&T Bell Labs on natural language processing:

"The complex hidden structure of natural-language sentences is manifested in two different ways:
predictively, in that not every constituent (for example, word) is equally likely in every context,
and evidentially, in that the information carried by a sentence depends on the relationships among the constituents of the sentence."
-- Fernando Pereira, AT&T Bell Labs

Thus,it is the complex nature of understanding human language. In the context of spam, this means several things must be considered:

1. The words and phrases, which are used
2. The context in which those words and phrases are used
3. The relationships between those words and phrases.

A. *Early Approaches*
Approaches such as blacklisting and rules-based systems take a very simplistic view when dealing with spam. Blacklisting stems from the premise that the number of people sending spam is small. Therefore, creating a list of "known spammers" and then blocking messages from those senders will stop the spam problem. However, using a blacklist only stops a fraction of the spammers--those that you already know about--and worse yet, often stops legitimate messages that originate from locations that spammers used in the past. Spammers who use multiple addresses and systems to send spam, and in turn create a serious false positive problem – emails which are legitimate but have been classified as spam, easily fool these systems.

Likewise, rules-based systems take an approach that just doesn't work when dealing with actual language. Is all email that contains the word "mortgage" considered spam? No, although often it is. Without an understanding of the context surrounding the use of the word mortgage, such a simple approach fails to be effective.

Even the most complex rule structures reach a point where catching more spam results in increased false positives. This happens because the number of ways in which words and phrases can be used together with the number of potential relationships between those words and phrases simply cannot be modeled using a series of rules.

B. *Later Approaches*
More recent approaches to solving the spam problem have taken a more sophisticated approach, but still fail to effectively solve the problem. The most promising of these approaches has been Bayesian filtering, a technique whereby each word in an email is given a point value, positive or negative, and messages exceeding a certain value are considered spam. Sophistication is added in that Bayesian filters can be trained based on a set of known messages, helping them to identify the probability that certain words are spam or not-spam related. These probabilities are then translated back into point values as described above.

This solves the problem of simply filtering the word "mortgage" without looking at the rest of the email. "Mortgage" may have a high spam-value, but if enough low-value words are used, then the end result may avoid being misidentified.

Bayesian filtering focuses on the prediction aspect of dealing with natural language. However, because it is a simple point system associated with a dictionary of words, it lacks the intelligence required to interpret a message evidentially as defined by Pereira's statement on natural language processing. And, it is unable to analyze a message generally in its context.

As such, spammers use tricks such as replacing the word "mortgage" with "mort-gage" or "m0rtage". These variations keep the "point value" of the message low, and trick the Bayesian filter into believing that the message is legitimate. As these filters are tweaked to try and avoid such pitfalls, perhaps by raising the point value of other words, the end result is a higher rate of false positives.

Finally, signature-based techniques use a model similar to that used by many virus filtering applications. The basis is that you can take a "signature"--a digital fingerprint of each message--and then compare that signature against a list of known spam messages.

Unfortunately, this methodology breaks down when used against the volume of unique spam messages that are circulated and created every day.

Consider the virus world: When a new and previously unseen virus is unleashed, it often takes several days before it can be identified, a signature created, and that signature distributed to the filtering applications. During this time, numerous systems are affected, sometimes badly enough to make national headlines.

Spam seldom makes national headlines, but the principle is the same: New and previously unseen spam messages escape detection from these signature systems. However, unlike the virus world, thousands of new and unique spam messages are created and delivered every day. Attempting to create new signatures for each of these messages, in real time, is an intense, expensive and reactive effort that more often than not delivers too little too late.

III. WHY IBF WORKS

The technological breakthrough of IBF is its ability to understand the complex nature of human language, in particular, its ability to understand the words and phrases used, the context in which they are used and how they relate to each other.

The key to correctly interpreting the intent of an email message is to predictive and evidentially analyzes the language used in a message. In other words, to understand both the meaning of the individual words and phrases themselves, the meaning of the context and structure in which those words and phrases are used, and the relationships between the various words and phrases used in that context.

Understanding the contextual meaning of and relationships between words and phrases is not something that can be translated into rules or point values. The number of possible relationships between words is far too vast for those approaches to be practical.

The goal of IBF is to analyze these three things simultaneously: the words and phrases used in a message, the relationships between those words and phrases, and the context in which those words are used. Only by considering all of these things simultaneously can the intent of a message be reliably and correctly interpreted.

In other applications, AI has worked to varying degrees. In almost all cases, the success of AI is dependent on the level of known information versus the level of unknown information. In short, the more unknown variables and the more dynamic the situation, the more difficult it is for AI to succeed.

Take for example the card game, poker. In poker, you know what all the cards are and you know what the legal moves are. However, you can't see all the cards – only those in your hand and those that have already been played. This is an imperfect information environment.

In the AI world, a computer poker player will become better over the course of a deck. At first, only the cards held in your hand are known. This is when the AI is at its weakest position. As the game progresses, the AI learns the cards held by the other players, and over the course of a deck, discovers what cards are remaining on the deck versus what cards have already been played. The number of unknowns decreases over the course of the game and thus, the AI is able to perform better as the game goes on.

IV. UNDERSTANDING LANGUAGE INTENT –THE KILLER APP FOR AI?

For many years, there was great promise surrounding artificial intelligence. AI was to be
the savior for many of our daily tasks, from driving our cars for us to finding information
on the Internet before we even asked for it. But it often didn't work. What happened to this great promise? Why has AI failed to deliver on such practical applications?
Consider the following statement:
"One of the clearest results of artificial intelligence research so far is that solving even apparently simple problems requires lots of knowledge."-- Dr. Alison Cawsey, author of the book Essence of Artificial Intelligence. For example, we can find

amusement in the famous Chicago Daily Tribune headline" Dewey Defeats Truman". Understanding that headline requires a lot of prior knowledge: You must know that there was some kind of contest between those two people. You must have been expecting to hear about an outcome, enough so that justthose three words made sense even when taken away from the context of a full article.Then, to find the amusement in that headline, you need to know that Truman actually defeated Dewey, and you need to know that a newspaper would look very foolish for making such an incorrect statement. All that – just to understand three words: Dewey defeats Truman. Now, consider the complexity in driving a car, or firing the guns of a tank: new information arriving in real time, referenced against old knowledge, and forecasting the results in the blink of an eye. When making decisions as a human, we must consider that the information is far from perfect, conditions far from ideal, and decisions must be made from a combination of prior knowledge, personal instinct, and an educated guess.Translated into AI, the end result is this: the car drives correctly… most of the time. And the tank targets the enemy… most of the time. But in these applications, "most of the time" is not an acceptable outcome. Further, the more uncertain the inputs, the greaterthe chances are that the AI will break down.

A. Understanding where AI has succeeded

While AI has not yet delivered on the promise of self-driving cars, there are application where AI has worked very well and still others where it has worked reasonably well. AI has worked very well in situations where the environment is controlled, and all information about that environment can be known. These environments are known as perfect information environments. The best-known example of a perfect information application for AI is IBM's Deep Blue chess computer. While not able to defeat the world champion 100% of the time, Deep Blue can certainly outperform almost all chess players, and certainly plays at a championship level. But chess is an "easy"environment to understand. You know which moves are legal, you can see all the pieces at once, and you know the circumstances for winning and losing. "Solving" the chess problem becomes a matter of choosing moves, then predicting your opponents moves to find the best end-game solution. In other applications, AI has worked to varying degrees. In almost all cases, the success of AI is dependent on the level of known information versus the level of unknown information. In short, the more unknown variables and the more dynamic the situation, the more difficult it is for AI to succeed.

Take for example the card game, poker. In poker, you know what all the cards are and you know what the legal moves are. However, you can't see all the cards – only those in your hand and those that have already been played. This is an imperfect information environment. In the AI world, a computer poker player will become better over the course of a deck. At first, only the cards held in your hand are known. This is when the AI is at its weakest position. As the game progresses, the AI learns the cards held by the other players, and over the course of a deck, discovers what cards are remaining on the deck versus what cards have already been played. The number of unknowns decreases over the course of the game and thus, the AI is able to perform better as the game goes on.

B. Applying AI and Natural Language Procesing Principles to Spam

Using these principles, we must be able to take a unique AI approach to solving the spam problem. For years, the world's AI experts have been trying to get machines to understand human language. However, IBF simplifies this complex world into a manageable one: the AI does not need to understand all of human language. Instead, it need only recognize the intent of that language, and divide that intent into wide, manageable groups such as junk mail, bulk mail, and legitimate mail.

The key is to make the number of unknowns as small as possible for the AI. While natural language processing is far from a perfect information environment, such as chess, by breaking down the problem we are able to eliminate many of the unknowns.

First, we create an "end-game" set of circumstances. In defining our spam

problem, we create an environment in which all incoming messages must land into one of several groups spanning from legitimate email to junk email. The lines between those groups may not be clearly defined, however we do know that all messages will fall somewhere within that range.

Now, the IBF AI engine does three things: (1) it analyzes what words are used in a message, (2) it analyzes how those words are used both independently and in relationship with one another, and (3) it considers an end-game scenario for that message based on its analysis and knowledge of other messages. If necessary, it may eliminate certain data – further eliminating unknowns – and re-evaluate some or all of a message.

In this way, IBF is able to avoid the deception of spammers as they change words and spellings to trick other filters. As in the game of chess, you do not known exactly what move your opponent is going to make, but you can predict it with reasonable accuracy using AI. As Pereira explains:

Any language model or parser must include a generative mechanism or grammar that specifies how sentences are built from their parts, and how the information associated to the sentence derives from the information associated to its parts. Furthermore, to be able to cope with previously unseen sentences, any such system must involve generalization with respect to the data from which the language model or parser was developed.
-- Fernando Pereira, AT&T Bell Labs

Consider the sentence “Make m0n-ey fast from home!” Because of the spelling of the word “money” we are not familiar with that “word”. Therefore, this sentence becomes an imperfect information sentence, much like our poker game earlier. To IBF, this sentence becomes “Make ____ fast from home!” Next, IBF considers the other known words and phrases in the message and their relationship with this sentence. IBF can then predict the intent of that sentence and email as a whole, without having perfect information about the missing word specifically.

The prediction of intent does not rely on actually predicting the specific missing word but instead considers the other words and their relationship to the rest of the message. In this way, we are working with a set of defined circumstances, with limited unknowns, in relation to the whole of the message. This is the key to the robustness of IBF---it allows us to predict message intent generally, as Pereira suggests, rather than trying to predict a single word specifically.

V. CONCLUSION

IBF technology represents a significant breakthrough in spam filtering, and in contextual natural language processing as a whole. Two key things are accomplished by the application of IBF technology:

One, IBF takes into account the language used, the context of that language, and the relationship between parts of that language. It is able to accurately predict the intent of a message without having perfect information about the message itself. This provides for a system that is not easily fooled by the tricks of spammers.

In order to trick the system, the spammer must change both the language used and the context in which that language is used—essentially changing the meaning of the message itself—a scenario that defeats the very purpose of sending the spam in the first place.

Two, IBF can exist as a stand along technology, providing much greater flexibility in its deployment than traditional anti-spam techniques. Because it does not rely on blacklists, rules, or signatures, IBF does not require frequent updates. In fact, IBF can work for years without the need for updates at all.

This allows IBF to be embedded on or included in systems at any level of the network: on the mail server, on a separate filtering appliance, or even at the firewall or router level. It can be the core filtering technology for systems both large and small.

Most importantly, IBF technology delivers on the promise for a high level of filtering

effectiveness, very low false positives and no required updates or maintenance.

VI. REFERENCES

[1] Spam Assassin by Alan Schwartz