

## Bayesian Filtering: the essentials

- A Must-take approach in any organization's Anti-Spam Strategy -

Whitepaper



- ✓ What is Bayesian Filtering
- ✓ How Bayesian Filtering works
- ✓ Benefits of Bayesian Filtering
- ✓ About Visendo Email Suite

---

ppedv AG - HQ – Burghausen, Germany

Tel: +49-8677-9889-110 Fax: +49-8677-9889-44

Info: [support@ppedv.de](mailto:support@ppedv.de) sales: [sales@ppedv.de](mailto:sales@ppedv.de) <http://www.visendo.com>

## Introduction

This white paper describes what Bayesian Filtering is, how it works and why it is among the best statistical intelligence methodology in SPAM filtering. It also explains why the Bayesian approach is the most effective way to tackle spam once and how it overcomes the obstacles faced by more static technologies such as blacklist checking, comparing to databases of known spam, keyword filtering

## What is Bayesian Filtering

Bayesian filtering is based on the principle that any event is dependent and that the probability of an event occurring in the future can be inferred from the history of occurrences of that event: history is always repeating (Various scientific researches have been made on Bayesian behavior [http://www-ccrma.stanford.edu/~jos/bayes/Bayesian\\_Parameter\\_Estimation.htm](http://www-ccrma.stanford.edu/~jos/bayes/Bayesian_Parameter_Estimation.htm)). Bayesian spam filtering has become a popular mechanism to distinguish illegitimate spam email from legitimate email (sometimes called "ham").

## How Bayesian Filtering Works

Before email can be filtered using this method, the administrator needs to generate a database with words, tokens, phrases, IP addresses, domains and so on collected from a sample of spam mail and valid mail (usually referred to as 'ham'). This has to be set up in the very beginning of the process. Then, a probability value is assigned to each word or token; the probability is based on calculations that take into account how often that word occurs in spam as opposed to legitimate mail (ham). This probability is influenced by the initial set-up of the database: analyzing the users' outbound mail and by analyzing known spam: all the tokens in both pools of email are analyzed to generate the probability that a particular word points to the email being spam.

Example of calculating these probabilities

If the word "antivirus" occurs in 100 of 1,000 spam mails and in 8 out of 100 legitimate emails, then its SPAM – PROBABILITY would be 11.1% (that is,  $[100/1000]$  divided by  $[8/100 + 100/1000]$ ).

## Creating the HAM database

It is important to note that the analysis of ham mail is performed on the organization's mail, and is therefore tailored to that particular organization. For example, a financial institution

---

ppedv AG - HQ – Burghausen, Germany

Tel: +49-8677-9889-110 Fax: +49-8677-9889-44

Info: [support@ppedv.de](mailto:support@ppedv.de) sales: [sales@ppedv.de](mailto:sales@ppedv.de) <http://www.visendo.com>

might use the word "debt" many times over and would get a lot of false positives if using a general anti-spam rule set. On the other hand, the Bayesian filter, if tailored to your company through an initial training period, takes note of the company's valid outbound mail (and recognizes "debt" as being frequently used in legitimate messages).

Note that some anti-spam software with simplistic Bayesian capabilities, such as the Outlook spam filter or the Internet Message Filter in Exchange Server. It does not create a customized ham data base / file for your company, but installs a standard ham data file. This method has 2 major drawbacks:

1. The ham data file is publicly available and can thus be hacked by professional spammers and therefore by passed.
2. Such a ham data file is a general one, and thus not customized to your company, it cannot be as effective and you will suffer from noticeably higher false positives.

### Creating the SPAM database

Besides valid mail, the Bayesian filter also relies on a spam data file/data base. This spam data file must include a large sample of known spam and must be constantly updated with the latest spam by the anti-spam software. This will ensure that the Bayesian filter is aware of the latest spam tricks, resulting in a high spam detection rate

(Note: this is achieved once the required initial SPAM - learning period is over).

### How the filtering is being processed

Once the ham and spam databases have been created, the word/token probabilities can be calculated and the filter is ready for use.

When a new mail arrives, it is broken down into words and the most relevant words – i.e., those that are most significant in identifying whether the mail is spam or not – are singled out. From these words, the Bayesian filter calculates the probability of the new message being spam or not. If the probability is greater than a certain level, say 0.75, then the message is classified as spam.

### Bayesian Filtering Advantages

1. The Bayesian method takes the whole message into account - It recognizes keywords that identify spam, but it also recognizes words that denote valid mail. For example: not every email that contains the word "free" and "cash" is spam. The advantage of the Bayesian method is that it considers the most interesting words (as defined by their deviation from the mean) and comes up with a probability that a message is spam. The Bayesian method would find the words "cash" and "free" interesting but it would also recognize the name of the business contact who sent the message and thus classify the message as legitimate, for instance;
2. A Bayesian filter is constantly self-adapting and self learning
3. The Bayesian technique is sensitive to the user. It learns the email habits of the company and understands that, for example, the word 'debt' might indicate spam if

- the company running the filter is, say, a restaurant, whereas it would not indicate it as spam if the company is a financial institution.
4. The Bayesian method is multi-lingual and international
  5. A Bayesian filter is difficult to fool, as opposed to any keyword filter

### Important observation

It is important to keep in mind that, when evaluating anti-spam software, if the product has advanced, customized Bayesian analysis, then it can only be truly evaluated after a few weeks. It is probable that basic anti-spam software might perform better initially, but after a few weeks the Bayesian filter catches up and well outperforms the conventional anti-spam filters once and for all.

### Innovation in Bayesian Filtering

Calculating the SPAM probabilities takes into account, as an essential factor, the number of incoming emails on a certain period of time. It is, hence, essential to make sure that you receive a big number of emails (i.e – 200 in 1 week) in order to build a comprehensive SPAM database.

Visendo has built its algorithm by taking into account a relatively small number of received emails with a low frequency. Visendo algorithm mixes already existing algorithms making it possible to build your SPAM database in a short period of time with a low inflow of emails. So, if you are a small or medium company, you don't have to wait for a long period of time to Make your Bayesian spam filter work effectively. Try out [Visendo Mail Checker Server](#) for further details.

## About Visendo Email Suite

popConnect 10 - is a POP3/IMAP-Connector for your mail server when sending and receiving your e-mail. It retrieves incoming e-mails from your POP3/IMAP-Account, transmits them to the correct recipient and vice versa. It offers support for APOP, GMAIL and it's main antiSPAM features are based on keyword filtering.

[Visendo Mail Checker](#) - Visendo Mailchecker is a complex, easy to use anti-spam and anti-phishing tool that works with any email client and, thanks to Multi-layered filters has a superior detection rate. It's filters are based on Bayesian Filtering.

[Visendo SMTP extender](#) - The idea behind Visendo SMTP extender is to connect and distribute the emails from the IIS SMTP server - Windows Server 2008, Vista, Windows 2008 SBS - with out using any POP3 connectors.

## About Visendo

---

ppedv AG - HQ – Burghausen, Germany

Tel: +49-8677-9889-110 Fax: +49-8677-9889-44

Info: [support@ppedv.de](mailto:support@ppedv.de) sales: [sales@ppedv.de](mailto:sales@ppedv.de) <http://www.visendo.com>

We are an Independent Software Vendor (ISV) specialized in internet systems integration on Microsoft technologies. We have always been one of the top 3 companies on the markets we entered. We were one of the first three Microsoft gold partners in e-commerce on the German market: this is the highest level for a Microsoft partner. Hence, you, as a potential customer, will benefit on our experience and quality of services that guarantees your investment.

For further information, please visit us on:website:

<http://www.visendo.com>, Product downloads

Blog: <http://www.blog.visendo.com>

Linkedin, Twitter , Facebook

Did you find this paper useful? Give us feedback NOW!

---

ppedv AG - HQ – Burghausen, Germany

Tel: +49-8677-9889-110 Fax: +49-8677-9889-44

Info: [support@ppedv.de](mailto:support@ppedv.de) sales: [sales@ppedv.de](mailto:sales@ppedv.de) <http://www.visendo.com>