Spam Filtering using Naïve Bayesian Classification

Presented by: Samer Younes
Outline

► What is spam anyway?
► Some statistics
► Why is Spam a Problem
► Major Techniques for Classifying Spam
  ▪ Transport Level Filtering
  ▪ Fingerprint Analysis
  ▪ Lexical Analysis
  ▪ Artificial Intelligence
  ▪ Statistical Analysis
  ▪ Heuristics
► Naïve Bayesian Classification
  ▪ Some Definitions
  ▪ Equations
► Evaluation Measures
► Conclusions
So What is Spam?

- **Spam = Unsolicited Bulk e-mail**

- **Used to promote:**
  - mass advertising & marketing
  - Scams and Fraud
  - Other less than desirable content

- **Reasons for using e-mail**
  - Cheap and cost effective
  - No widely spread legislation banning its use for such purposes
Some Statistics

Current spam traffic accounts for:

- 60% of all e-mail traffic
- Approximately 10 billion messages / day
- Approximately 2.6 trillion messages / year
- Equivalent to 2000TB / day
Some Statistics (ctnd.)

Loss of Productivity adds up to more than $10 billion in the US alone!

Source: Brightmail Incorporated

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>64%</td>
<td>April 2004</td>
</tr>
<tr>
<td>63%</td>
<td>March 2004</td>
</tr>
<tr>
<td>62%</td>
<td>February 2004</td>
</tr>
<tr>
<td>60%</td>
<td>January 2004</td>
</tr>
<tr>
<td>58%</td>
<td>December 2003</td>
</tr>
<tr>
<td>56%</td>
<td>November 2003</td>
</tr>
<tr>
<td>52%</td>
<td>October 2003</td>
</tr>
<tr>
<td>54%</td>
<td>September 2003</td>
</tr>
<tr>
<td>50%</td>
<td>August 2003</td>
</tr>
<tr>
<td>50%</td>
<td>July 2003</td>
</tr>
<tr>
<td>49%</td>
<td>June 2003</td>
</tr>
<tr>
<td>48%</td>
<td>May 2003</td>
</tr>
</tbody>
</table>

Percentages of Total Internet Email Identified as Spam

Number of Fraudulent Emails Filtered by Brightmail

Over 96 Billion Email Messages Filtered by Brightmail in April 2004

Over 3.1 Billion Fraudulent Emails Filtered by Brightmail in April 2004.
Why is Spam a Problem

- Financial Costs

- Resource Abuse
  - Storage
  - Employee Productivity
  - Over-Burdened Mail Servers

- Fraud

- Social Costs
So how to Stop Spam

Several Techniques Exist:

- Transport Level Filtering
- Fingerprint Analysis
- Lexical Analysis
- Artificial Intelligence
- Statistical Analysis
- Heuristics
Transport Level Filtering

Transport Level Filtering:
- Secure SMTP
- Realtime Blackhole Lists (RBLs)
Fingerprint Analysis

- Similar to Anti-Virus Software
- Requires large fingerprint database
- Not a viable solution considering subtle variations in messages
Lexical Analysis

► Context examination of words or phrases

► Uses rule based discrimination and weighted words to filter
Artificial Intelligence

Uses Neural Networks techniques to filter out spam
Statistical Analysis

- Includes Bayesian classifiers
- Similar classification rates to NN
- Associates probabilities to words
- Will be covered in subsequent slides
Heuristics

- Is a combination of the above listed methods
- Used in large enterprises architectures
Naïve Bayesian Classification

► Some Definitions

- Corpus = Body of words used in statistical classification and training.

- Vector $\mathbf{v} = (x_1, x_2, \ldots, x_n)$ is an array of values with associated attributes.
  (in spam filters attributes are words)

- Lemmatizer = Engine that returns the root of a word (i.e. $\text{Lem(“going”) = “go”}$)

- Stop-List = List of frequently used wording
  (i.e. “the”, “is”, etc...)
What is Bayesian Classification

Given a document $d$ create a vector $v$ of values and compute the mutual information (MI) of each candidate using the following formula:

$$MI(X; C) = \sum_{x \in \{0,1\}, c \in \{spam, legitimate\}} P(X = x, C = c) \cdot \log \frac{P(X = x, C = c)}{P(X = x) \cdot P(C = c)}$$

Taking the attributes with highest MI’s they are compared to frequency ratios from the training corpus.
What is Bayesian Classification (ctnd.)

► Having computed the MI's a new vector $\vec{v}'$ is constructed and fed to the following formula:

$$P(C = c | \vec{X} = \vec{x}) = \frac{P(C = c) \prod_{i=1}^{n} P(X_i = x_i | C = c)}{\sum_{k \in \{spam, legitimate\}} P(C = k) \prod_{i=1}^{n} P(X_i = x_i | C = k)}$$

► With the assumption that the various attributes of vector $\vec{v}'$ are conditionally independent
What is Bayesian Classification (ctnd.)

- The classification is committed if the following criteria is satisfied:

\[
\frac{P(C = \text{spam} \mid \vec{X} = \vec{x})}{P(C = \text{legitimate} \mid \vec{X} = \vec{x})} > \lambda
\]

- The threshold \( \lambda \) is usually set to 999 in order to avoid classifying legitimate e-mail as spam.
How it all fits together

1. Stop-list
2. Lemmatizer
3. Construct Token Vectors
4. Compute MI
5. Compute Probabilities
6. Compute Threshold
7. Bayesian Classifier

Spam
NOT Spam
Results and Measures

Results based on the testing of the Bayesian Classifier were as follows:

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Total Messages</th>
<th>Testing Messages</th>
<th>% Spam</th>
<th>Spam Precision</th>
<th>Spam Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>words only</td>
<td>1789</td>
<td>251</td>
<td>88.2%</td>
<td>97.1%</td>
<td>94.3%</td>
</tr>
<tr>
<td>words + phrases</td>
<td>1789</td>
<td>251</td>
<td>88.2%</td>
<td>97.6%</td>
<td>94.3%</td>
</tr>
<tr>
<td>words + phrases + non-textual</td>
<td>1789</td>
<td>251</td>
<td>88.2%</td>
<td>100.0%</td>
<td>98.3%</td>
</tr>
<tr>
<td>words + phrases + non-textual</td>
<td>2815</td>
<td>222</td>
<td>~20%</td>
<td>92.3%</td>
<td>80.0%</td>
</tr>
</tbody>
</table>
Evaluation was also performed to assess the usefulness of lemmatizers and stop-lists in increasing the accuracy of the filter with a threshold of 999.
Results and Measures (ctnd.)

The following table shows precision, recall and accuracy of the filter with various configurations.

<table>
<thead>
<tr>
<th>Filter Configuration</th>
<th>( \lambda )</th>
<th>No. of attrib.</th>
<th>Spam Recall</th>
<th>Spam Precision</th>
<th>Weighted Accuracy</th>
<th>Baseline W. Acc.</th>
<th>TCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) bare</td>
<td>1</td>
<td>50</td>
<td>81.10%</td>
<td>96.85%</td>
<td>96.408%</td>
<td>83.374%</td>
<td>4.63</td>
</tr>
<tr>
<td>(b) stop-list</td>
<td>1</td>
<td>50</td>
<td>82.35%</td>
<td>97.13%</td>
<td>96.649%</td>
<td>83.374%</td>
<td>4.96</td>
</tr>
<tr>
<td>(c) lemmatizer</td>
<td>1</td>
<td>100</td>
<td>82.35%</td>
<td>99.02%</td>
<td>96.926%</td>
<td>83.374%</td>
<td>5.41</td>
</tr>
<tr>
<td>(d) lemmatizer + stop-list</td>
<td>1</td>
<td>100</td>
<td>82.78%</td>
<td>99.49%</td>
<td>97.064%</td>
<td>83.374%</td>
<td>5.66</td>
</tr>
<tr>
<td>(a) bare</td>
<td>9</td>
<td>200</td>
<td>76.94%</td>
<td>99.46%</td>
<td>99.419%</td>
<td>97.832%</td>
<td>3.73</td>
</tr>
<tr>
<td>(b) stop-list</td>
<td>9</td>
<td>200</td>
<td>76.11%</td>
<td>99.47%</td>
<td>99.401%</td>
<td>97.832%</td>
<td>3.62</td>
</tr>
<tr>
<td>(c) lemmatizer</td>
<td>9</td>
<td>100</td>
<td>77.57%</td>
<td>99.45%</td>
<td>99.432%</td>
<td>97.832%</td>
<td>3.82</td>
</tr>
<tr>
<td>(d) lemmatizer + stop-list</td>
<td>9</td>
<td>100</td>
<td>78.41%</td>
<td>99.47%</td>
<td>99.450%</td>
<td>97.832%</td>
<td>3.94</td>
</tr>
<tr>
<td>(a) bare</td>
<td>999</td>
<td>200</td>
<td>73.82%</td>
<td>99.43%</td>
<td>99.912%</td>
<td>99.980%</td>
<td>0.23</td>
</tr>
<tr>
<td>(b) stop-list</td>
<td>999</td>
<td>200</td>
<td>73.40%</td>
<td>99.43%</td>
<td>99.912%</td>
<td>99.980%</td>
<td>0.23</td>
</tr>
<tr>
<td>(c) lemmatizer</td>
<td>999</td>
<td>300</td>
<td>63.67%</td>
<td>100.00%</td>
<td>99.993%</td>
<td>99.980%</td>
<td>2.86</td>
</tr>
<tr>
<td>(d) lemmatizer + stop-list</td>
<td>999</td>
<td>300</td>
<td>63.05%</td>
<td>100.00%</td>
<td>99.993%</td>
<td>99.980%</td>
<td>2.86</td>
</tr>
</tbody>
</table>

TCR = Total Cost Ratio \( \rightarrow \) Higher TCR = Better Results
Conclusions

► Authors believe naïve bayesian classification is not viable.

► In vivo results of the technique have shown a high degree of accuracy in catching spam.

► This technique is used in various commercial products targeted mainly to the single user.