Tiago Agostinho de Almeida

SPAM: do Surgimento à Extinção

Tese de doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos exigidos para a obtenção do título de Doutor em Engenharia Elétrica.
Área de concentração: Automação.

Orientador: Akebo Yamakami

Campinas
2010

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

AL64s

Almeida, Tiago Agostinho de
SPAM: do surgimento à extinção / Tiago Agostinho de Almeida. --Campinas, SP: [s.n.], 2010.

Orientador: Akebo Yamakami.
Tese de Doutorado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Spam (Mensagens eletrônicas). 2. Inteligência artificial. 3. Classificação. 4. Categorização (Linguística). 5. Sistemas inteligentes. I. Yamakami, Akebo. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: SPAM: from the rise to the extinction
Palavras-chave em Inglês: Spam (Electronic mail), Artificial intelligence, Classification, Categorization (Linguistic), Intelligent systems
Área de concentração: Automação
Titulação: Doutor em Engenharia Elétrica
Banca examinadora: João Paulo Papa, Sahudy Montenegro González, Tatiane Regina Bonfim, Romis Ribeiro de Faissol Attux
Data da defesa: 10/09/2010
Programa de Pós Graduação: Engenharia Elétrica

Tiago Agostinho de Almeida
Mestre em Engenharia Elétrica – UNICAMP

SPAM: do Surgimento à Extinção

Tese de doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos exigidos para a obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: Automação. Aprovada pela banca examinadora no dia 10 de Setembro de 2010.

Banca Examinadora:

Prof. Dr. Akebo Yamakami
FEEC/UNICAMP

Prof. Dr. João Paulo Papa
FC/UNESP

Profa. Dra. Tatiane Regina Bonfim
FAC/Campinas

Prof. Dr. Romis Ribeiro de Faissol Attux
FEEC/UNICAMP

Profa. Dra. Sahudy Montenegro González
CMCC/UFABC

Campinas
2010

COMISSÃO JULGADORA - TESE DE DOUTORADO

Candidato: Tiago Agostinho de Almeida

Data da Defesa: 10 de setembro de 2010

Título da Tese: "SPAM: do Surgimento à Extinção"

Prof. Dr. Akebo Yamakami (Presidente): _____

Prof. Dr. João Paulo Papa: _____

Profa. Dra. Tatiane Regina Bonfim: _____

Profa. Dra. Sahudy Montenegro González: _____

Prof. Dr. Romis Ribeiro de Faissol Attux: _____

Dedico este trabalho a todo aquele que de alguma forma contribui para a evolução da ciência e da consciência.

Agradecimentos

Agradeço,

ao Prof. Akebo a oportunidade e aos seis anos de valorosa orientação cujos conselhos e sugestões foram muito valiosos tanto para este trabalho quanto para a minha vida.

aos meus pais, Jurandy e Sônia, e avós, Anésio (*in memorian*) e Virgínia (*in memorian*), pela formação como ser humano e por me ensinarem às coisas mais preciosas que a vida oferece.

à minha esposa, Ana Carolina, pela dedicação, companheirismo, paciência e carinho.

ao meu irmão Júnior aos inúmeros conselhos, correções e sugestões.

ao meu irmão Marcos a paciência e ao apoio.

aos colegas de trabalho mais próximos: Ricardo, Renato, Ricardo Coelho, Taís, André e Cristiano pela boa convivência e frutíferas discussões.

aos demais colegas do Departamento de Telemática a ótima convivência.

aos professores da FEEC: Von Zuben, Romis, Paulo Valente e Takaaki, os ótimos cursos oferecidos.

aos membros da banca examinadora os comentários, sugestões e contribuições, que ajudaram a melhorar a qualidade e a redação final deste manuscrito.

à agência CNPq o apoio financeiro concedido durante todo o período de doutoramento.

à FEEC/UNICAMP a ótima estrutura que oferece aos estudantes e pesquisadores.

à CAPES o portal de periódicos eletrônicos, que permite o acesso rápido e eficiente ao conhecimento científico.

ao Google o complemento que faz ao item anterior.

a todos que de alguma forma contribuíram com o meu progresso como aluno e como Ser.

A maravilhosa disposição e harmonia do universo só pode ter tido origem segundo o plano de um Ser que tudo sabe e tudo pode. Isto fica sendo a minha última e mais elevada descoberta.

Isaac Newton

Resumo

Nos últimos anos, *spams* têm se tornado um importante problema com enorme impacto na sociedade. A filtragem automática de tais mensagens impõem um desafio especial em categorização de textos, no qual a característica mais marcante é que os filtros enfrentam um adversário ativo, que constantemente procura evadir as técnicas de filtragem. Esta tese apresenta um estudo abrangente sobre o problema do *spamming*. Dentre as contribuições oferecidas, destacam-se: o levantamento histórico e estatístico do fenômeno do *spamming* e as suas consequências, o estudo sobre a legalidade do *spam* e os recursos jurídicos adotados por alguns países, a análise de medidas de desempenho utilizadas na avaliação dos filtros de *spams*, o estudo dos métodos mais empregados para realizar a filtragem de *spams*, a proposta de melhorias dos filtros Bayesianos através da adoção de técnicas de redução de dimensionalidade e, principalmente, a proposta de um novo método de classificação baseado no princípio da descrição mais simples auxiliado por fatores de confidência. Vários experimentos são apresentados e os resultados indicam que a técnica proposta é superior aos melhores filtros anti-*spams* presentes tanto comercialmente quanto na literatura.

Palavras-chave: Filtragem de *Spam*. Classificação. Categorização de texto. Propaganda enviada em massa. Recuperação de informação. Aprendizado de máquina. Princípio da descrição mais simples.

Abstract

Spam has become an increasingly important problem with a big economic impact in society. Spam filtering poses a special problem in text categorization, in which the defining characteristic is that filters face an active adversary, which constantly attempts to evade filtering. In this thesis, we present a comprehensive study of the spamming problem. Among many offered contributions we present: the statistical and historical survey of spamming and its consequences, a study regarding the legality of spams and the main juridic methods adopted by some countries, the study and proposal of new performance measures used for the evaluation of the spam classifiers, the proposals for improving the accuracy of Naive Bayes filters by using dimensionality reduction techniques and a novel approach to spam filtering based on the minimum description length principle and confidence factors. Furthermore, we have conducted an empirical experiments which indicate that the proposed classifier outperforms the state-of-the-art spam filters.

Keywords: Spam filtering. Classification. Text categorization. Junk mail. Information retrieval. Machine learning. Minimum description length.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos e Notação

ARPA	*Advanced Research Projects Agency* (Agência de Projetos de Pesquisa Avançada)
AUC	*Area Under the ROC Curve* (Área sobre a curva ROC)
AUTODIN	*Automatic Digital Network* (Rede Digital Automática)
CAN-SPAM	*Controlling the Assault of Non-Sollicited Pornography and Marketing Act*
CCM	Coeficiente de Correlação de Matthews
CEAS	*Conference on E-mail and Anti-Spam* (Congresso Internacional de E-mail e Anti-Spam)
CERN	*Conseil Européen pour la Recherche Nucléaire* (Centro Europeu de Pesquisas Nucleares)
CF	*Confidence Factors* (Fatores de confidência)
CGI	Comitê Gestor da Internet
CTSS	*Compatible Time-Sharing System* (Sistema Compatível de Divisão por Tempo)
DNS	*Domain Name System* (Sistema de Nome de Domínio)
FTC	*Federal Trade Commission* (Comissão Norte-Americana de Comércio)
HTML	*HyperText Markup Language* (Linguagem de Marcação de Hipertexto)
HTTP	*Hypertext Transfer Protocol* (Protocolo de Transferência de Hipertexto)
IMAP	*Internet Message Access Protocol* (Protocolo de Internet para Acesso de Mensagem)
IP	*Internet Protocol* (Protocolo da Internet)
ISP	*Internet Service Provider* (Provedor de Serviços da Internet)
LAM	*Logistic Average Misclassification* (Erro médio logístico)
MAAWG	*Messaging Anti-Abuse Working Group*
MDL	*Minimum Description Length* (Princípio da descrição mais simples)
MIME	*Multipurpose Internet Mail Extensions* (Extensões Multi-função para Mensagens de Internet)
MIT	*Massachusetts Institute of Technology* (Instituto Tecnológico de Massachusetts)
MMS	*Multimedia Messaging Service* (Serviço de mensagens multimídia)
NB	*Naive Bayes*
NCSA	*National Center for Supercomputing Applications* (Centro Nacional de Aplicações de Supercomputação)
NRPOL	Norma de Referência da Privacidade Online

POP	*Post Office Protocol* (Protocolo de Postagem)
RCT	Razão do Custo Total
RFC	*Request for Comments* (Requisição de Comentários)
ROC	*Receiver Operating Characteristic* (Características de Operação do Receptor)
SAGE	*Semi-Automatic Ground Environment* (Ambiente Base Semiautomático)
SMTP	*Simple Mail Transfer Protocol* (Protocolo de Transferência de Correspondência Simples)
SMS	*Short Message Service* (Serviço de Mensagens Curtas)
SVM	*Support Vector Machine* (Máquina de Vetores de Suporte)
TCP	*Transfer Control Protocol* (Protocolo de Controle e de Transfêrencia)
TCPA	*Telephone Consumer Protection Act*
TF	*Term Frequency* (Frequência do termo)
TREC	*Text REtrieval Conference* (Congresso de Recuperação de Texto)
TST	Técnica de Seleção de Termos
WWW	*World Wide Web* (Teia Mundial)

c_i	Classe da mensagem (c_s=*spam* e c_l=legítima)
$CF(t_k)$	Fator de confidência do termo k
$\mathcal{L}$	Conjunto de mensagens legítimas (*hams*)
$L_m(c)$	Aumento da descrição da classe c pela inserção da mensagem m
m	mensagem, e-mail, texto, conjunto de termos
$P(p)$	Probabilidade de ocorrência de p
$P(q\|p)$	Probabilidade de ocorrência de p dado q
$\mathcal{S}$	Conjunto de mensagens de *spams*
$\mathcal{T}$	Espaço/Conjunto de termos
T	Limiar de classificação
t_k	termo ou *token* k
Tr	Conjunto de treinamento
$\vec{x}$	representação vetorial de uma mensagem m
X_k	valor do atributo x_k de $\vec{x}$ associado ao termo t_k

Introdução Geral

E-mail é um dos meios de comunicação mais popular, mais rápido e mais barato. Ele tornou-se parte do cotidiano de milhares de pessoas, mudando a maneira como elas trabalham e colaboram. O e-mail não é mais empregado apenas para dar suporte à comunicação pessoal, mas também é utilizado como agenda, gerenciador de tarefas, organizador de contatos, sistema de envio e armazenamento de documentos, etc. A desvantagem desse enorme sucesso é o volume sempre crescente de *spams* (mensagens eletrônicas comerciais não-solicitadas) que recebemos. O problema do *spam* pode ser quantificado em termos econômicos, uma vez que muitas horas são desperdiçadas todos os dias por trabalhadores. Não se trata apenas do tempo que eles perdem com a leitura do *spam*, mas também o tempo que gastam para excluir essas mensagens. Isso, sem considerar os inúmeros problemas com fraudes, roubos e danos diretos.

De acordo com relatórios anuais divulgados por grandes corporações de segurança computacional, a quantidade de *spams* em circulação está aumentando de maneira assustadora. A média de *spams* enviados por dia passou de 2,4 bilhões em 2002[1] para 300 bilhões em 2010[2]. Em 2008, calculou-se que aproximadamente 96,5% de todo o e-mail recebido por empresas foram *spams*[3]. Estima-se que, atualmente, 90,4% de todo o tráfego de e-mail seja representado por *spams*[4].

O volume de *spams* enviado no Brasil é um problema crescente com proporções alarmantes e enorme impacto econômico para toda a sociedade. Apesar de tratar-se de um problema de escala global, tal fenômeno vem ganhando grande destaque internamente, fazendo com que o Brasil seja um dos países que mais enviam lixo virtual no mundo. Segundo fontes, o país foi o maior emissor de mensagens não-solicitadas, sendo origem de aproximadamente 20% dos *spams* enviados nos dois primeiros meses de 2010[5] e responsável por 7,7 trilhões de mensagens deste tipo somente em 2009[2].

A ausência de uma regulamentação jurídica aliada à alta vulnerabilidade dos sistemas computacionais e à falta de conhecimento dos usuários de e-mail são indicados como os principais

[1]Consulte `http://www.spamunit.com/spam-statistics/`

[2]Consulte `www.ciscosystems.cd/en/US/prod/collateral/vpndevc/cisco_2009_asr.pdf`

[3]Disponível em `http://www.sophos.com/articles/2008/dirtydozjul08.html`

[4]Consulte `http://www.messagelabs.com/MLIReport_2009_FINAL.pdf`

[5]Panda Security Report Q12010, disponível em:
`http://www.pandasecurity.com/img/enc/Quarterly_Report_Pandalabs_Q1_2010.pdf`

fatores que impulsionaram o Brasil a tornar-se o maior disseminador de *spams* do mundo.

Felizmente, diversos métodos vêm sendo propostos para evitar e minimizar o impacto desse problema, sendo que um dos recursos mais promissores é o emprego de técnicas de aprendizado de máquina para realizar a filtragem automática de e-mails (Cormack, 2008). Tais abordagens incluem métodos que são considerados os melhores em categorização de texto, tais como Rocchio (Joachims, 1997; Schapire *et al.*, 1998), Boosting (Carreras & Marquez, 2001), Máquina de Vetores de Suporte (*Support Vector Machines* – SVM) (Drucker *et al.*, 1999; Kolcz & Alspector, 2001; Hidalgo, 2002; Almeida *et al.*, 2010a), regressão logística (Goodman & Yih, 2006; Lynam *et al.*, 2006), filtros Bayesianos (Androutsopoulos *et al.*, 2004; Metsis *et al.*, 2006; Almeida *et al.*, 2009, 2010b), além de outros.

Contudo, apesar do avanço significativo dos métodos de filtragem disponíveis, ainda existem alguns pontos fracos bem conhecidos, como alta sensibilidade a ruídos, por exemplo, que são explorados pelos *spammers*, possibilitando o constante aumento no volume de *spams* em circulação (Cormack & Lynam, 2007; Cormack, 2008; Guzella & Caminhas, 2009). Além disso, a evolução e a adaptação dos termos das mensagens com o intuito de burlar as regras de classificação constitui um grande desafio para os métodos de filtragem (Zhang *et al.*, 2004).

Se por um lado os mecanismos jurídicos contra o fenômeno do *spamming* ainda estão em gestação no Brasil e engatinham no restante do mundo, os métodos computacionais de filtragem e classificação automática avançam com passos largos. A principal justificativa é que a ausência ou a ineficiência dos recursos legislativos fazem com que a sobrevivência do sistema atual de e-mail dependa diretamente da evolução dos métodos de classificação.

O fato mais preocupante é que existe o risco cada vez maior de que o e-mail, dentro de alguns anos, deixe de ser o meio de comunicação simples, eficiente e acessível que é hoje, caso a tendência ao crescimento desenfreado do volume de *spams* não seja revertida. Nesse sentido, uma correta avaliação e integração no plano internacional de combate ao *spam* não é somente um auxílio precioso – é uma necessidade de primeira ordem.

Objetivos e contribuições

O principal objetivo desta tese é oferecer um estudo abrangente sobre o fenômeno do *spamming*, abordando as principais alternativas para a solução desse problema e, dessa forma, contribuir no combate contra o *spam* e possibilitar que o fruto dessa pesquisa possa ser usufruída por toda a sociedade.

Dentre as contribuições oferecidas neste trabalho, destacam-se:

1. Introdução do problema do *spam* com apresentação de dados históricos, estatísticos e quadro atual no Brasil e no mundo;

2. Estudo relacionado ao combate do *spam* por meios jurídicos usando recursos legislativos que já se encontram em vigor no Brasil, além do estudo de ferramentas legislativas que

são utilizadas em outros países;

3. Análise das medidas de desempenho utilizadas para avaliar o desempenho dos filtros anti-*spam* e apresentação do coeficiente de correlação de Matthews como boa alternativa neste quesito;

4. Apresentação e avaliação dos principais métodos de classificação de *spams* disponíveis na literatura;

5. Proposta e avaliação do emprego de técnicas de redução de dimensionalidade do espaço de dados para aumentar o desempenho dos filtros de *spams*;

6. Apresentação, proposta e validação de um novo classificador automático de *spams* baseado no princípio da descrição mais simples auxiliado por fatores de confidência.

Organização

Este manuscrito está organizado da seguinte maneira:

- No Capítulo 1, é apresentada uma introdução sobre o fenômeno do *spamming*, bem como o seu surgimento e o contexto histórico em que ele surgiu. Dessa forma, é oferecida uma breve apresentação relacionada ao surgimento da Internet, do e-mail e do *spam*, destacando os principais fatos históricos que auxiliam na explicação do seu contexto atual.

- No Capítulo 2, são apresentadas informações sobre como a prática do *spamming* fere os direitos do cidadão brasileiro, oferecendo, assim, direções para o combate efetivo do *spam* por meio de intervenção jurídica. Além disso, é oferecida uma análise dos modelos jurídicos empregados por outros países.

- No Capítulo 3, são apresentados os principais conceitos básicos necessários para a compreensão deste manuscrito, tais como: representação adotada pelos métodos de classificação, bases de dados de e-mail comumente utilizadas para avaliar os filtros anti-*spam* e medidas de desempenho empregadas na comparação de resultados obtidos por diferentes classificadores. Posteriormente, são apresentados os principais métodos de classificação disponíveis tanto na literatura quanto comercialmente. Adicionalmente, é oferecido um estudo comparativo de desempenho entre os métodos mais populares empregados na filtragem de *spams*.

- No Capítulo 4, é apresentado um estudo comparativo de diversas técnicas de seleção de termos empregadas em conjunto com diversas variações dos filtros anti-*spam* Naive Bayes detalhados no Capítulo 3. O principal objetivo desse capítulo é examinar como os métodos seletores de termos podem ajudar a manter, ou até mesmo melhorar, o desempenho dos diferentes classificadores baseados na probabilidade Bayesiana.

- No Capítulo 5, é apresentada a proposta de um novo método de classificação baseado no princípio da descrição mais simples auxiliado por fatores de confidência. Para realizar uma avaliação rigorosa do método proposto, foram simulados os três grandes campeonatos de filtros anti-*spam* usando as mesmas funções, bases dados, medidas de desempenho e tarefas empregadas em cada um deles. Destaca-se que os resultados obtidos mostram que o filtro proposto é superior aos melhores classificadores disponíveis atualmente.

- Finalmente, no Capítulo 6, são apresentadas as conclusões finais sobre o estudo desenvolvido, bem como direções para trabalhos futuros.

Capítulo 1

O Surgimento

Antes de apresentar as características e o funcionamento dos *spams*, é importante conhecer um pouco da sua história. Fatos como o seu surgimento e o contexto histórico em que ele surgiu podem ajudar a compreender o seu estado atual.

Neste capítulo, são descritos de maneira resumida o surgimento da Internet, do e-mail e do *spam*, apresentando os principais fatos históricos que auxiliam na explicação do seu contexto atual. Para tal, este capítulo está estruturado da seguinte forma: a Seção 1.1 apresenta, resumidamente, a história da Internet, o seu funcionamento e a sua posição atual. O surgimento do e-mail e as suas características são detalhadas na Seção 1.2. Por fim, na Seção 1.3, são apresentados dados sobre a história do *spam*, sobre a sua expansão e sobre o quadro atual.

1.1 A Internet

A Internet é um sistema global de redes de computadores interconectadas que trocam dados por transferência de pacotes usando os protocolos padrões TCP/IP (*Transfer Control Protocol/Internet Protocol*). Trata-se de uma "rede de redes" que consiste de milhares de redes privadas e públicas, acadêmicas, comerciais e governamentais.

A Internet proporciona diversos recursos de informações e serviços, como e-mail, bate-papo em tempo real, transferência e compartilhamento de arquivos, jogos *on-line*, inter-conectividade de documentos hipertexto, além de outros recursos da *World Wide Web* (www).

A Internet surgiu por vias militares nos períodos de apogeu da Guerra Fria. Na década de 60, quando dois blocos antagônicos politicamente, socialmente e economicamente exerciam enorme controle e influência no mundo, qualquer mecanismo, qualquer inovação, qualquer ferramenta nova poderia contribuir nessa disputa liderada pela ex-União Soviética e pelos Estados Unidos. As duas superpotências compreendiam a eficácia e a necessidade absoluta dos meios de comunicação (Edwards, 1996).

Nessa perspectiva, o governo dos Estados Unidos temia um ataque às suas bases militares pela ex-URSS, pois informações importantes e sigilosas poderiam ser perdidas. Dessa forma, foi

idealizado um modelo de troca e compartilhamento de informações que permitisse a descentralização das mesmas. Neste contexto surgiu a ARPANET, criada pela ARPA (*Advanced Research Projects Agency*), uma subdivisão do departamento de defesa dos EUA (Edwards, 1996).

A ARPANET funcionava através de um sistema conhecido como chaveamento de pacotes, que é um esquema de transmissão de dados em rede de computadores no qual as informações são divididas em pequenos "pacotes" que, por sua vez, contém trecho dos dados, o endereço do destinatário e informações que permitem a remontagem da mensagem original (Abbate, 2000).

Na década de 70, a tensão entre ex-URSS e EUA diminuiu: as duas potências entraram definitivamente naquilo em que a história se encarregou de chamar de coexistência pacífica (Edwards, 1996). Não havendo mais a iminência de um ataque imediato, o governo dos EUA permitiu que pesquisadores desenvolvessem, nas suas respectivas universidades, estudos na área de defesa que pudessem, também, entrar na ARPANET. Com isso, a ARPA começou a ter dificuldades em administrar todo este sistema devido ao grande e crescente número de localidades universitárias contidas nela (Carvalho, 2006).

Assim, o sistema dividiu-se em dois grupos, a MILNET, que possuía as localidades militares, e a nova ARPANET, que possuía as localidades não militares. O desenvolvimento da rede, nesse ambiente mais livre, pôde então acontecer. Pesquisadores e alunos passaram a ter acesso aos estudos já empreendidos e somaram esforços para aperfeiçoá-los (Abbate, 2000).

Um recurso técnico denominado IP (*Internet Protocol*) permitia que o tráfego de informações fosse encaminhado de uma rede para outra. Todas as redes conectadas pelo endereço IP na Internet comunicavam-se para que todas pudessem trocar mensagens. Através da *National Science Foundation*, o governo norte-americano investiu na criação de *backbones*, computadores com alto poder de processamento conectados por linhas que têm a capacidade de dar vazão a grandes fluxos de dados (Abbate, 2000).

Na década de 80, o interesse pela rede ampliou e mais recursos foram investidos em pesquisas. Em 1989, o cientista Sir Tim Berners-Lee do CERN (*Conseil Européen pour la Recherche Nucléaire* - Centro Europeu de Pesquisas Nucleares) criou a *World Wide Web*, um sistema que, inicialmente, interligava sistemas de pesquisas científicas e acadêmicas. A rede coletiva ganhou uma maior divulgação pública a partir de 1990 (Abbate, 2000).

Em Agosto de 1991, Berners-Lee publicou seu novo projeto para a *World Wide Web*, dois anos depois de começar a criar o HTML, o HTTP e as poucas primeiras páginas *Web* no CERN, na Suíça. Em 1993, o Centro abriu mão do direito de propriedade dos códigos básicos do projeto de um sistema global de hipertexto (Abbate, 2000).

Em 1994, a NCSA (*National Center for Supercomputing Applications*) lançou o primeiro navegador para a *Web*, o *X Windows Mosaic* 1.0. O lançamento deste *browser* ajudou na popularização da Internet, que desta maneira passou a fazer parte do cotidiano das pessoas (Carvalho, 2006). Nesta época, o então senador Al Gore idealizou a *Superhighway of Information* (super-estrada da informação) que tinha como unidade básica de funcionamento a troca, compartilhamento e o fluxo contínuo de informações pelos quatro cantos do planeta, através

de uma rede mundial. O crescente interesse mundial aliado ao interesse comercial, que evidentemente observava o potencial financeiro e rentável da Internet, proporcionou o *boom* e a popularização da Internet na década de 90 (Abbate, 2000).

Por fim, em 1996, a palavra Internet já era de uso comum, principalmente nos países desenvolvidos, referindo-se na maioria das vezes à WWW. Esta confusão entre a nomenclatura Internet e *web* é frequente até hoje. Entretanto, vale lembrar que a *web* é só uma parte da Internet.

Segundo dados da *World Internet Usage Statistics*[1], no final do ano 2000, a Internet possuía aproximadamente 360 milhões de usuários. Ao fim de 2009, este número superou a marca de 1 bilhão e 800 milhões de usuários, representando um crescimento superior a 400% e correspondendo a aproximadamente 27% de toda a população mundial.

1.2 O E-mail

O e-mail (correio eletrônico) é um método que permite compor, enviar e receber mensagens através de sistemas eletrônicos de comunicação. O termo e-mail é aplicado tanto aos sistemas que utilizam a Internet e são baseados no protocolo SMTP (*Simple Mail Transfer Protocol*), como àqueles sistemas conhecidos como *intranets*, que permitem a troca de mensagens dentro de uma empresa ou organização e são, normalmente, baseados em protocolos proprietários.

Por mais curioso que pareça, o e-mail surgiu antes da Internet, e foi uma ferramenta crucial para a criação da rede internacional de computadores.

O primeiro sistema de troca de mensagens entre computadores de que se tem notícia foi criado em 1965[2], e possibilitava a comunicação entre os múltiplos usuários de um computador do tipo *mainframe* (computador de grande porte capaz de oferecer serviços de processamento a milhares de usuários através de terminais conectados diretamente ou através de uma rede). Acredita-se que os primeiros sistemas criados com tal funcionalidade foram o Q32 da *System Development Corporation* e o *Compatible Time-Sharing System* (CTSS) do MIT.

O sistema eletrônico de mensagens transformou-se rapidamente em um "e-mail em rede", permitindo que usuários situados em diferentes computadores trocassem mensagens. O sistema AUTODIN (*Automatic Digital Network*), de 1966, parece ter sido o primeiro a permitir que mensagens eletrônicas fossem transferidas entre computadores diferentes, mas é possível que o sistema SAGE (*Semi-Automatic Ground Environment*) tivesse a mesma funcionalidade algum tempo antes.

A rede de computadores ARPANET trouxe uma grande contribuição para a evolução do e-mail, aumentando significativamente a sua popularidade. Relatos indicam que houve transferência de mensagens eletrônicas entre diferentes sistemas situados nesta rede logo após a sua criação, em 1969. Em 1971, o programador Ray Tomlinson introduziu o uso do sinal @ para

[1]Fonte: `http://www.internetworldstats.com/stats.htm`.
[2]Fonte: `http://www.multicians.org/thvv/mail-history.html`.

separar os nomes do usuário e da máquina (domínio) no endereço de correio eletrônico[2].

O envio e recebimento de uma mensagem de e-mail são realizados através de um sistema de correio eletrônico. Esse sistema é composto de programas que suportam a funcionalidade de cliente de e-mail e de um ou mais servidores de e-mail que, através de um endereço, conseguem transferir uma mensagem de um usuário para outro. Estes sistemas utilizam protocolos de Internet que permitem o tráfego de mensagens de um remetente para um ou mais destinatários que possuem computadores conectados à Internet.

O formato na Internet para mensagens de e-mail é definido na RFC 2822[3] e uma série de outras RFCs (RFC 2045 até a RFC 2049) que são conhecidas como MIME (*Multipurpose Internet Mail Extensions*).

Mensagens de e-mail são compostas basicamente por duas partes principais:

- **Cabeçalho (*header*)** – é estruturado em campos que contém o remetente, destinatário e outras informações sobre a mensagem;
- **Corpo (*body*)** – contém o texto da mensagem separado do cabeçalho por uma linha em branco.

A Figura 1.1 apresenta a sequência típica de eventos que ocorre ao enviar um e-mail. A título de exemplo, consideremos que Alice escreveu uma mensagem para Bob. A partir do momento em que ela clica no botão "enviar" no seu editor de e-mails, as seguintes etapas são realizadas:

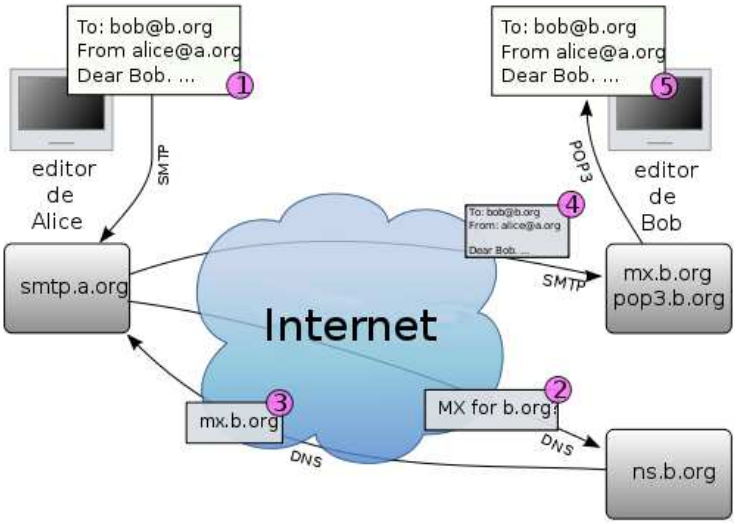


Figura 1.1: Sequência de eventos para envio de e-mails.

1. O editor de e-mails de Alice formata a mensagem deixando-a no formato padrão de e-mails da Internet. Utilizando o protocolo SMTP, ele envia a mensagem para o agente de

[3]RFC é um acrônimo para o inglês *Request for Comments*. Trata-se de um documento que descreve os padrões de cada protocolo da Internet.

transferência local de e-mail, neste caso `smtp.a.org`, realizado pelo provedor de serviços da Internet de Alice;

2. O agente de transferência analisa o endereço de destino fornecido pelo SMTP, neste caso `bob@b.org` e procura pelo nome do domínio de Bob (`b.org`) no Sistema de Nome de Domínio (DNS) para encontrar o servidor de troca de mensagens que aceite aquele domínio;

3. O servidor DNS do domínio `b.org`, `ns.b.org`, responde com o registro MX[4], neste caso `mx.b.org`, um servidor do provedor de serviços da Internet de Bob;

4. `smtp.a.org` envia a mensagem para `mx.b.org` usando SMTP, que entrega a mensagem à caixa de mensagens de Bob;

5. Finalmente, quando Bob clicar em "Ler mensagem" no seu editor de e-mails, este carregará a mensagem utilizando o protocolo POP3 (*Post Office Protocol*).

As aplicações do e-mail normalmente oferecem ao usuário uma série de facilidades. A maior parte delas fornece um editor de textos embutido e a possibilidade do envio de arquivos anexados à correspondência. Além disso, a maioria das aplicações permite o envio de correspondências para um único destinatário, para mais de uma pessoa ou para um grupo de pessoas.

Embora não tenha sido desenvolvida como uma ferramenta de trabalho cooperativo, os serviços de e-mail adaptaram-se muito bem ao ambiente de grupos de trabalho, tendo se tornado indispensáveis nas organizações, agilizando processos, democratizando o acesso às informações e diminuindo os custos. Esta é uma das formas mais utilizadas para o estabelecimento de comunicações através do computador.

De acordo com um estudo realizado pelo grupo de pesquisa *Radicati Group*, o volume de e-mails enviados diariamente passou de 210 bilhões registrados em 2007 para 247 bilhões em 2009. Isso significa que, em média, mais de 2,8 milhões de e-mails são enviados por segundo. No mesmo período, o número de usuários de e-mail também aumentou de 1,2 bilhão para 1,4 bilhão[5].

1.3 O *Spam*

Simultaneamente ao desenvolvimento e a popularização da Internet e do e-mail, ocorreu o crescimento de um fenômeno que, desde o seu surgimento, tornou-se um dos principais problemas da comunicação eletrônica em geral: o envio em massa de mensagens não-solicitadas. Esse fenômeno ficou conhecido como *spamming*, as mensagens em si como *spam* e os seus autores como *spammers* (Gyöngyi & Garcia-Molina, 2005).

[4] *Mail exchanger record*. Lista de servidores que deveriam receber e-mail para um determinado domínio e sua relação de prioridade.

[5] Fonte: `http://email.about.com/od/emailtrivia/f/emails_per_day.htm`.

Apesar da existência de mensagens não-eletrônicas que podem ser comparadas ao *spam*, como por exemplo correspondências, ligações telefônicas e folhetos promocionais não-solicitados, o termo é reservado aos meios eletrônicos devido a motivações que os tornam muito mais propícios ao crescimento do fenômeno.

O termo *spam*, abreviação em inglês de "*spiced ham*" (presunto condimentado), diz respeito a de uma mensagem eletrônica não-solicitada enviada em massa. Consequentemente, as mensagens legítimas passaram a ser chamadas de *hams*.

Na sua forma mais popular, um *spam* consiste numa mensagem de e-mail com fins publicitários. O termo *spam*, no entanto, pode ser aplicado a mensagens enviadas por outros meios e com outras finalidades. Geralmente, os *spams* têm caráter apelativo e, na grande maioria das vezes, são incômodos e inconvenientes.

O primeiro registro oficial de uma mensagem eletrônica não solicitada enviada em massa ocorreu no CTSS do MIT. Pouco tempo depois da sua criação, Tom Van Vleck e Noel Morris implementaram o programa CTSS MAIL que permitia que usuários se comunicassem através de mensagens. Em 1971, um administrador do CTSS chamado Peter Bos utilizou o CTSS MAIL para enviar a mensagem pacifista intitulada "*THERE IS NO WAY TO PEACE. PEACE IS THE WAY*", ou "Não há caminho para a paz. A paz é o caminho". Quando Tom Van Vleck considerou tal comportamento como inadequado, Bos se defendeu dizendo que aquilo era importante[2].

Em 1978, ocorreu o primeiro registro de uma mensagem comercial não solicitada enviada em massa utilizando mensagem eletrônica[6], reproduzida abaixo:

> Mail-from: DEC-MARLBORO rcvd at 3-May-78 0955-PDT
> Date: 1 May 1978 1233-EDT
> From: THUERK at DEC-MARLBORO
> Subject: ADRIAN@SRI-KL
>
> DIGITAL WILL BE GIVING A PRODUCT PRESENTATION OF THE NEWEST MEMBERS OF THE DECSYSTEM-20 FAMILY; THE DECSYSTEM-2020, 2020T, 2060, AND 2060T. THE DECSYSTEM-20 FAMILY OF COMPUTERS HAS EVOLVED FROM THE TENEX OPERATING SYSTEM AND THE DECSYSTEM-10 <PDP-10> COMPUTER ARCHITECTURE. BOTH THE DECSYSTEM-2060T AND 2020T OFFER FULL ARPANET SUPPORT UNDER THE TOPS-20 OPERATING SYSTEM. THE DECSYSTEM-2060 IS AN UPWARD EXTENSION OF THE CURRENT DECSYSTEM 2040 AND 2050 FAMILY. THE DECSYSTEM-2020 IS A NEW LOW END MEMBER OF THE DECSYSTEM-20 FAMILY AND FULLY SOFTWARE COMPATIBLE WITH ALL OF THE OTHER DECSYSTEM-20 MODELS.

[6]Fonte: `http://www.templetons.com/brad/spamterm.html`.

> WE INVITE YOU TO COME SEE THE 2020 AND HEAR ABOUT THE DECSYSTEM-20 FAMILY AT THE TWO PRODUCT PRESENTATIONS WE WILL BE GIVING IN CALIFORNIA THIS MONTH. THE LOCATIONS WILL BE:
>
> TUESDAY, MAY 9, 1978 - 2 PM
> HYATT HOUSE (NEAR THE L.A. AIRPORT)
> LOS ANGELES, CA THURSDAY, MAY 11, 1978 - 2 PM
> DUNFEY'S ROYAL COACH
> SAN MATEO, CA (4 MILES SOUTH OF S.F. AIRPORT AT BAYSHORE, RT 101 AND RT 92)
>
> A 2020 WILL BE THERE FOR YOU TO VIEW. ALSO TERMINALS ON-LINE TO OTHER DECSYSTEM-20 SYSTEMS THROUGH THE ARPANET. IF YOU ARE UNABLE TO ATTEND, PLEASE FEEL FREE TO CONTACT THE NEAREST DEC OFFICE FOR MORE INFORMATION ABOUT THE EXCITING DECSYSTEM-20 FAMILY.

Apesar do número considerável de reações indignadas e de subsequentes discussões a respeito, em parte devido ao fato de que a ARPANET era considerada como sendo de uso exclusivo para assuntos do governo norte-americano, a questão não foi considerada por todos como sendo de maior importância na época.

Embora os dois registros possam ser considerados como o início do *spamming*, o termo *spam* não foi associado ao envio de mensagens não-solicitadas até a década de 80. O início exato do uso da palavra é incerto e alvo de muita especulação, mas existe um consenso de que ele provavelmente originou-se nos *Multi-User Dungeons* (MUDs), ambientes virtuais onde múltiplos usuários conectados por uma rede podem interagir e conversar.

Alguns relatos descrevem que o ato de prejudicar o sistema através do envio excessivo de dados, e o ato de enviar múltiplas mensagens com o objetivo de deslocar as mensagens de outros usuários para fora da tela, passaram a ser conhecidos como *spamming*, quando alguns usuários começaram a comparar esse comportamento ao dos *vikings* presentes em um quadro do grupo humorístico britânico Monty Python. Nesse episódio, um casal vai a um restaurante onde todos os pratos são servidos com Spam, uma marca americana de carne enlatada da empresa *Hormel Foods Corporation* (Figura 1.2). A mulher não gosta do alimento, mas não consegue nenhuma opção sem ele. Ao longo do diálogo, a palavra é repetida insistentemente pelos protagonistas, principalmente por um grupo de *vikings* presentes no local, que começa a cantar: "*Spam, spam, spam, spam. Lovely spam! Wonderful spam!*" atrapalhando a conversa no restaurante da mesma maneira que as mensagens não-solicitadas atrapalhavam a comunicação nos MUDs.

O quadro foi escrito para ironizar o racionamento de comida ocorrido na Inglaterra durante a Segunda Guerra Mundial. Spam foi um dos poucos alimentos excluídos desse racionamento, o que eventualmente levou as pessoas a enjoarem do produto e motivou a criação do quadro.

Figura 1.2: Latas de carne suína da marca Spam.

Existem outras versões, menos populares, a respeito da etimologia que associam o termo *spam* a acrônimos, como: *Sending and Posting Advertisement in Mass* (enviar e postar publicidade em massa), *Shit Posing As Mail* (porcaria fingindo ser correspondência), ou ainda *Single Post to All Messageboards* (mensagem única para todos os fóruns de discussão), além de outras[6].

Na década de 90, o termo *spam* tornou-se popular quando passou a ser usado na rede USENET, o maior sistema de grupos de notícias e listas de discussão da época[7]. Entretanto, outras mensagens não-solicitadas já haviam sido enviadas anteriormente na USENET. Em 1988, Rob Noha enviou para diversas listas de discussão um pedido de doações para seu fundo de faculdade e David Rhodes iniciou no mesmo período a circulação de uma corrente eletrônica conhecida como "*Make Money Fast*". Contudo, o primeiro uso conhecido da palavra *spam* na USENET para designar esse tipo de comportamento foi feito por Joel Furr após um episódio em 1993 que ficou conhecido como "*ARMM Incident*". Nesse incidente, um software experimental chamado ARMM, criado por Furr para moderar mensagens inadequadas em listas de discussão, acidentalmente enviou recursivamente dezenas de mensagens para a lista `news.admin.policy` devido a uma falha de implementação. Joel Furr lamentou o ocorrido e declarou que "não era sua intenção enviar *spam* para as listas"[6].

Em 18 de Janeiro de 1994 foi enviado o primeiro e mais importante *spam*. Clarence Thomas, administrador do sistema da *Andrews University*, enviou a todos os grupos de notícias uma mensagem religiosa intitulada "*Global Alert for All: Jesus is Coming Soon*". E quatro meses depois, em 12 de Abril de 1994, Laurence Canter e Martha Siegel, dois advogados da cidade norte-americana de Phoenix, que trabalhavam em casos de imigração, enviaram uma mensagem anunciando serviços que teoricamente ajudavam as pessoas a ganharem vistos de permanência

[7]Fonte: `http://www.oreillynet.com/pub/a/network/2001/12/21/usenet.html`.

(*Green Card*) nos Estados Unidos. Por causa disso, a mensagem ficou conhecida como *Green Card Spam* e, já na época, imediatamente gerou as mesmas reações do *spam* atual, com questionamentos sobre ética e legalidade da prática. Apesar de não tratar-se de uma mensagem inédita, eles usaram uma tática inovadora: contrataram um programador para criar um *script* simples e enviar o anúncio da dupla para todos os milhares de grupos de notícias da USENET. O esquema deu certo e todos receberam o primeiro *spam* comercial em larga escala da história[8]:

> From: Laurence Canter (nike@indirect.com)
> Subject: Green Card Lottery- Final One?
> Date: 1994-04-12 00:40:42 PST
>
> Green Card Lottery 1994 May Be The Last One!
> THE DEADLINE HAS BEEN ANNOUNCED.
>
> The Green Card Lottery is a completely legal program giving away a certain annual allotment of Green Cards to persons born in certain countries. The lottery program was scheduled to continue on a permanent basis. However, recently, Senator Alan J Simpson introduced a bill into the U. S. Congress which could end any future lotteries. THE 1994 LOTTERY IS SCHEDULED TO TAKE PLACE SOON, BUT IT MAY BE THE VERY LAST ONE.
>
> PERSONS BORN IN MOST COUNTRIES QUALIFY, MANY FOR FIRST TIME.
>
> The only countries NOT qualifying are: Mexico; India; P.R. China; Taiwan, Philippines, North Korea, Canada, United Kingdom (except Northern Ireland), Jamaica, Domican Republic, El Salvador and Vietnam.
>
> Lottery registration will take place soon. 55,000 Green Cards will be given to those who register correctly. NO JOB IS REQUIRED.
>
> THERE IS A STRICT JUNE DEADLINE. THE TIME TO START IS NOW!!
>
> For FREE information via Email, send request to cslaw@indirect.com
>
> –
> **********
> Canter & Siegel, Immigration Attorneys
> 3333 E Camelback Road, Ste 250, Phoenix AZ 85018 USA
> cslaw@indirect.com telephone (602)661-3911 Fax (602) 451-7617

[8]Fonte: `http://www.computerweekly.com/Articles/2002/10/03/190015/spam-spam-spam-spam.htm`.

Diversas pessoas referiram-se à mensagem como *spam*, e o termo passou a ser utilizado em qualquer outra instância de comportamento análogo. Em poucos anos, com a popularização cada vez maior da Internet, o envio de mensagens não-solicitadas passou a crescer no meio do correio eletrônico, em parte estimulado pela existência dos programas para envio automático de mensagens, e se expandiu rapidamente para os outros meios disponíveis[6].

Dentre os meios de propagação de *spam* podem ser destacados: e-mails, mensagens instantâneas, grupos de discussão e fóruns, mensagens de texto em celulares, jogos *on-line*, sites de busca, *blogs*, *wikis*, *guestbooks* e sites de compartilhamento de arquivos e vídeos.

Atualmente, os assuntos mais abordados nos *spams* são: pornografia (25%), ganho financeiro (22%), produtos (13%), saúde (10%), Internet (9%), enganos (7%), lazer (6%), religioso (5%) e demais (5%)[9].

Segundo relatório divulgado no primeiro trimestre de 2010 pela Panda Security[10], 13,76% dos *spams* que são enviados diariamente partem do Brasil. A Figura 1.3 apresenta a origem geográfica dos *spams* em termos de porcentagem, conforme publicado no relatório.

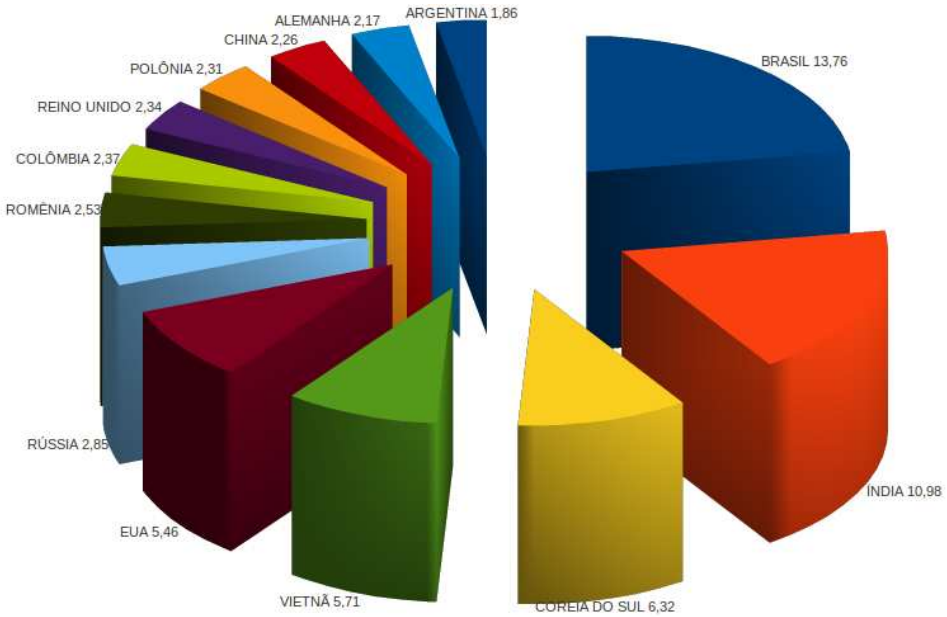


Figura 1.3: Origem geográfica dos *spams* (%).

A consequência direta do *spam* é o consumo de recursos de máquina e de rede dos provedores de Internet e dos usuários, além do custo do tempo e da atenção humana gastos para selecionar e apagar as mensagens não-solicitadas. Indiretamente, há o custo gerado pelas vítimas dos crimes que geralmente acompanham os *spams*, tais como roubo financeiro, roubo de identidade, roubo de dados e propriedade intelectual, vírus e outras infecções por *malwares*, pornografia infantil, fraudes e propagandas enganosas (Goodman *et al.*, 2005).

[9]Fonte: `http://spam-filter-review.toptenreviews.com/spam-statistics.html`.
[10]Fonte: `http://www.pandasecurity.com/homeusers/media/press-releases/viewnews?noticia=10111`.

Os *spammers* frequentemente utilizam falsos nomes, endereços, números de telefones, e outras informações para contato, necessários para criar e configurar contas de e-mail na maioria dos provedores de Internet. Em muitos casos, eles usam cartões de crédito falsificados ou roubados para pagar por essas contas. Dessa forma, quando o provedor detecta e remove uma conta eles podem reagir rapidamente criando outras novas (Goodman *et al.*, 2007).

O custo associado aos *spams* também inclui gastos colaterais causados pelo conflito constante entre os *spammers* e os administradores e usuários do meio atacado pelo *spamming*[11]. Além disso, o conteúdo do *spam* frequentemente é ofensivo, já que os produtos mais anunciados envolvem pornografia. Uma vez que os *spammers* enviam seus *spams* indiscriminadamente e em larga escala, anúncios pornográficos podem ser expostos em locais de trabalho ou até mesmo a crianças.

Os *spams* exemplificam uma "tragédia dos comuns" (Hardim, 1968): *spammers* usam recursos (físicos e humanos) sem se preocuparem com o custo total. Com isso, o custo acaba sendo dividido com cada usuário. Em alguns aspectos o *spam* pode ser visto como uma grande ameaça ao sistema de e-mail atual.

Uma vez que o custo de envio de e-mail é extremamente baixo, um grupo reduzido de *spammers* pode saturar a Internet com mensagens não-solicitadas. Apesar de somente uma pequena porcentagem dos seus alvos ser motivada a comprar os seus produtos, o baixo custo pode determinar um ganho suficiente para manter o *spamming* ativo.

De acordo com um estudo realizado por Kanich *et al.* (2008), os *spammers* conseguem obter lucros, mesmo com um índice de resposta de um para cada 12,5 milhões de e-mails enviados. Os autores conseguiram se infiltrar em uma rede que envia *spams* e estudar a sua estrutura econômica. A análise sugere que um baixo índice de respostas é suficiente para produzir milhões de dólares de lucro por ano. A pesquisa também indica que as pessoas que enviam *spams* estão mais suscetíveis a ataques de *hackers*, o que encarece o custo da operação. Os autores se infiltraram na rede Storm, que utiliza computadores de usuários domésticos "sequestrados" para enviar *spam* – ou seja, o computador do usuário envia *spams* sem que o seu dono perceba. No seu ápice, acredita-se que o Storm controlou até 1 milhão de computadores pelo mundo. Os autores acompanharam os *spams* enviados por 75.869 computadores sequestrados – apenas uma pequena fração da rede Storm. Um dos *spams* promovia o site de uma farmácia que oferecia remédios para aumento da libido. O site era falso e foi criado pelos próprios autores. A página da farmácia falsa retornava uma mensagem de erro cada vez que alguém tentava comprar remédios com cartão de crédito. Foram enviadas 469 milhões de mensagens de *spam*, a grande maioria promovendo a falsa farmácia. Depois de 26 dias, apenas 28 vendas foram realizadas. O índice de resposta da campanha foi inferior a 0,00001% – muito abaixo do índice de 2,15% prometido por serviços privados e legalizados de envio de mensagens. O lucro gerado foi de US$ 2.731,88 – um pouco mais de US$ 100 por dia, no período computado. Ao expandir esse número para

[11]Fonte: `http://linxnet.com/misc/spam/thank_spammers.html`.

a totalidade da rede Storm, os autores indicam que apenas essa rede de *spams* produz cerca de US$ 7.000 por dia, ou US$ 3,5 milhões por ano.

A quantidade de *spams* em circulação continua aumentando de forma preocupante e, por enquanto, não apresenta sinais de enfraquecimento. O número de *spams* que os usuários vêem em suas caixas de mensagens é apenas a ponta do *iceberg*, uma vez que as listas de e-mails dos *spammers* contém uma enorme porcentagem de endereços inválidos e muitos filtros simplesmente recusam ou apagam os *spams* óbvios.

Segundo uma estimativa divulgada no *Cisco 2009 Annual Security Report*[12], realizado pela Cisco Systems Inc., cerca de 300 bilhões de *spams* deverão ser enviados por dia em 2010, representando mais de 90% de todos os e-mails enviados diariamente. O estudo também revela que um dos principais meios utilizados para envio dessas mensagens são as redes de computadores "sequestrados". O computador é usado para disparar *spam*, promover ataques ou furtar informações pessoais. Ainda, de acordo com o estudo, as ameaças evoluem diariamente, à medida que criminosos descobrem novas formas de explorar pessoas, redes e a Internet.

Os *spammers* usam inclusive um método para identificar e "sequestrar" senhas consideradas fracas, facilmente detectáveis. Senhas ruins nesse caso são nomes de familiares, aniversários, endereços residenciais ou termos facilmente dedutíveis. A partir do momento em que eles conseguem acesso a essas contas de e-mail, começam a enviar milhares de mensagens usando o remetente do dono da conta.

De acordo com outro experimento realizado pela empresa de segurança McAfee em 2009[13], o Brasil é o segundo país que mais recebe *spam*. A empresa convidou 50 voluntários de dez países – inclusive do Brasil – para participar de uma pesquisa na qual os participantes teriam que passar um mês navegando na Internet sem filtro ou proteção contra *spam*. A empresa forneceu um computador limpo para cada voluntário e analisou a quantidade de *spams* recebidos. Os Estados Unidos ficaram em primeiro lugar na lista dos países que mais receberam e-mails indesejados durante o experimento. Os participantes norte-americanos receberam um total de 23.233 *spams* nos 30 dias de teste. Já os brasileiros ocuparam o segundo lugar da lista. Os cinco voluntários brasileiros receberam 15.856 e-mails indesejados durante o mês. Em terceiro lugar ficou a Itália, seguida pelo México e Reino Unido. Entre os países analisados, a Alemanha ficou com a última colocação, com cerca de 2 mil mensagens. Ao analisar os dados de todos os países, o relatório aponta que um usuário convencional da Internet recebe uma média de 70 e-mails indesejados por dia.

A Tabela 1.1 apresenta o aumento significativo do volume médio de *spams* enviados diariamente em escala global nos períodos aferidos[14].

O projeto Spamhaus (projeto voluntário, fundado por Steve Linford em 1998, com o intuito

[12]Fonte: `http://cisco.com/en/US/prod/vpndevc/annual_security_report.html`.

[13]Fonte: `http://news.softpedia.com/news/McAfee-Releases-Final-Report-On-the-SPAM-Experiment.shtml`.

[14]Fonte: `http://www.spamunit.com/spam-statistics/`.

Ano	Quantidade (bilhões)
2004	11
2005	30
2006	55
2007	100
2008	210
2009	250
2010	300

Tabela 1.1: Quantidade média de *spams* enviada por dia no mundo.

de rastrear *spammers*) estima que em 2008, aproximadamente 96,5% dos e-mails recebidos por empresas foram *spams*.

Já a fundação MAAWG (*Messaging Anti-Abuse Working Group* – grupo de pesquisa anti-*spam* formado por provedores de serviços da Internet e companhias de telecomunicações) estima que mais de 90% das mensagens recebidas no quarto semestre de 2009 foram *spams*. O conjunto de amostras utilizado para o estudo da MAAWG é composto por mais de 100 milhões de endereços de e-mail [15].

A título de curiosidade, segundo uma reportagem veiculada pela BBC News[16], de acordo com Steve Ballmer – chefe executivo da Microsoft – Bill Gates recebe, em média, 4 milhões de e-mails por ano, sendo que a maioria deles é *spam*. Por outro lado, Jef Poskanzer, dono do domínio `acme.com`, recebia até 2005, cerca de 1 milhão de *spams* por dia[17] – quase 100 vezes mais que Bill Gates.

Em relação ao custo gerado pelo *spam*, a Comissão de Comércio Interno da União Européia estimou em 2001 que o *spam* custou, somente naquele ano, cerca de 10 bilhões de euros aos usuários da Internet[18].

Segundo o *US Technology Readiness Survey*, produzido pelo Centro de Excelência em Serviço da Universidade de Maryland, o custo gerado pela perda de produtividade por causa do *spam*, em escala global, foi estimado em US$ 50 bilhões, somente em 2004[19]. Em 2009, estimou-se que mais de US$ 130 bilhões foram perdidos, em escala global, por causa do *spam*[20].

Destaca-se que, nesse contexto, o Brasil vem ganhando posição de destaque. Em 2008, o Brasil ocupava a 6^a posição dos países que mais enviaram *spams* e no primeiro bimestre de 2010, o país assumiu a indesejada liderança. A ausência de uma regulamentação jurídica aliada à alta vulnerabilidade dos sistemas computacionais e à falta de conhecimento dos usuários de e-mail são indicados como os principais fatores que impulsionaram o Brasil a tornar-se o líder do "seleto" grupo de países que mais enviam lixo virtual.

[15]Fonte: `http://www.maawg.org/sites/maawg/files/news/MAAWG_2009-Q3Q4_Metrics_Report_12.pdf`.
[16]Fonte: `http://news.bbc.co.uk/2/hi/business/4023667.stm`.
[17]Fonte: `http://www.acme.com/mail_filtering/`.
[18]Fonte: `http://europa.eu/rapid/pressReleasesAction.do?reference=IP/01/154`.
[19]Fonte: `http://www.smith.umd.edu/ntrs/NTRS_2004.pdf`.
[20]Fonte: `http://www.ferris.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/`.

Contudo, apesar do conhecimento e das calorosas discussões sobre o mal ocasionado pelo *spam*, nenhuma solução realmente eficiente foi adotada e, ao que tudo indica, não será, pelo menos em breve. Acredita-se que, a menos que o sistema convencional de e-mail passe por profundas reestruturações, o problema continuará oferecendo sérios riscos à existência e manutenção do próprio sistema.

Basicamente, existem duas formas de combater o *spamming* sem a necessidade de reestruturar o atual sistema de e-mail. Elas são dadas por:

1. Métodos diretos – tratam-se de medidas que atacam diretamente os *spammers* e, portanto, contribuem para a redução do volume de *spams* em circulação. Os principais representantes desses métodos são os meios legislativos aliados à fiscalização e à punição, bem como o aumento da segurança dos sistemas computacionais e a correta educação dos usuários do sistema de e-mail;

2. Métodos indiretos – são recursos empregados para evitar o problema do *spamming*. Basicamente, são representados por mecanismos computacionais que filtram e excluem automaticamente os *spams*.

Os métodos diretos são, obviamente, bem mais eficazes. Entretanto, eles são de difícil implementação e os resultados são percebidos apenas no longo prazo (após anos). Por outro lado, os métodos indiretos, apesar de não resolverem o problema do *spamming*, são simples de utilizar e apresentam resultados no curto prazo (após horas ou dias).

No capítulo seguinte são apresentadas com maior profundidade informações sobre os métodos diretos, principalmente no que tange aos recursos relacionados à legislação e regulamentação contra mensagens indesejadas.

Capítulo 2

Métodos Diretos – Combatendo o Problema

Atualmente, no Brasil, não existe regulamentação jurídica específica com relação ao *spam*. Entretanto, há projetos de lei que versam sobre esse tema. Este capítulo apresenta informações sobre como a prática do *spamming* fere os direitos do cidadão brasileiro, oferecendo, assim, direções para o combate do *spam* por meio de intervenção jurídica.

Baseado no trabalho de Lemos *et al.* (2007), uma breve introdução sobre o assunto está descrita na Seção 2.1. Na Seção 2.2 é abordado o tema relativo à tutela dos dados pessoais e como o comércio desses dados fere direitos constitucionais. Além disso, uma breve introdução dos modelos empregados por outros países é oferecida na Seção 2.3.

2.1 Tutela Legal do Spam

O envio de *spams* pode ser visto como uma forma de violação ao direito à privacidade do indivíduo, prerrogativa que é tutelada pela Constituição Federal Brasileira e pelo Código Civil (Lei n. 10.406, de 10 de janeiro de 2002). O envio dessas mensagens eletrônicas indesejadas somente é viabilizado em decorrência da obtenção dos endereços de e-mails e dados pessoais dos destinatários. O núcleo desse problema reside na legalidade da obtenção desses dados.

A Constituição Federal, em seu artigo 5°, incisos X e XII, garante a inviolabilidade da intimidade, da vida privada, da correspondência e dos meios de comunicações, conforme exposto a seguir:

> *"Artigo 5°, X: São invioláveis a intimidade, a vida privada, a honra e a imagem das pessoas, assegurado o direito a indenização pelo dano material ou moral decorrente de sua violação."*

> *"Artigo 5°, XII: É inviolável o sigilo da correspondência e das comunicações telegráficas, de dados e das comunicações telefônicas, salvo, no último caso, por ordem judicial, nas hipóteses e na forma que a lei estabelecer para fins de investigação criminal ou instrução processual penal."*

Em Bastos & Martins (1989), os autores oferecem o seguinte esclarecimento a respeito do artigo 5°, inciso X:

> *"O inciso oferece guarida ao direito à reserva da intimidade assim como ao da vida privada. Consiste na faculdade que tem cada indivíduo de obstar a intromissão de estranhos na sua vida privada e familiar, assim como impedir-lhes o acesso a informações sobre a privacidade de cada um, e também impedir que sejam divulgadas informações sobre esta área da manifestação existencial do ser humano."*

A Constituição Federal, ainda, dá guarida à privacidade do indivíduo em seu artigo 5°, LXXII, o qual proporciona-lhe a possibilidade de recorrer ao Judiciário para que tenha acesso as suas informações pessoais contidas nas bases de dados de entidades do governo ou as de caráter público, por meio da garantia dada pelo habeas-data[1] (Lemos *et al.*, 2007), assim prevista:

> *"Artigo 5°, LXXII - conceder-se-á habeas-data:*
>
> *a) para assegurar o conhecimento de informações relativas à pessoa do impetrante, constantes de registros ou bancos de dados de entidades governamentais ou de caráter público;*
>
> *b) para a retificação de dados, quando não se prefira fazê-lo por processo sigiloso, judicial ou administrativo."*

Segundo Ferraz Junior (2001), a tutela constitucional depara-se no processo de transmissão dos dados pessoais, fornecendo a quem interessar meios que possam impedir a manipulação intencional de dados (meios de comunicação violados ou grampeados), a divulgação de informações distorcidas (amparo ao direito à imagem) ou ainda que invadam a privacidade pessoal (dados pessoais que são coletados e armazenados em banco de dados).

A proteção ao direito à privacidade, no ordenamento jurídico brasileiro, não é esgotado no âmbito da Constituição, outras leis regulamentam a privacidade em áreas específicas. A Lei n. 9296/96, por exemplo, estabelece as condições necessárias para interceptação telefônica; por sua vez, a Lei n. 5250/67, mais conhecida como Lei da Imprensa, estabelece penalidades para a pessoa que, no exercício da atividade jornalística, revelar fatos que violem a privacidade e a intimidade alheias.

O Código Civil brasileiro, por sua vez, ampara a privacidade da pessoa natural, ou seja, da pessoa física, em seu artigo 21, conforme segue:

[1] *Habeas-data* é um remédio jurídico (facultativo) na formação de uma ação constitucional que pode, ou não, ser impetrada por pessoa física ou jurídica (sujeito ativo) para tomar conhecimento ou retificar as informações a seu respeito, constantes nos registros e bancos de dados de entidades governamentais ou de caráter público. Pode-se também entrar com ação de *habeas-data* com o intuito de adicionar, retirar ou retificar informações em cadastro existente. É remédio personalíssimo, só podendo ser impetrado por aquele que é o titular dos dados questionados.

> *"Art. 21. A vida privada da pessoa natural é inviolável, e o juiz, a requerimento do interessado, adotará as providências necessárias para impedir ou fazer cessar ato contrário a esta norma."*

De maneira complementar, em Doneda (2003), o autor expõe o seguinte:

> *"Ao clamar pela criatividade do magistrado para que tome as providências adequadas, o Código Civil dá mostras da necessidade de uma atuação específica de todo o ordenamento na proteção da privacidade da pessoa humana, que seja uma resposta eficaz aos riscos que hoje corre."*

Com relação à privacidade na Internet, é recorrente a venda de cadastros de consumidores, sem seu conhecimento prévio ou consentimento. Tal prática se desenvolveu amplamente tendo em vista a facilidade de comunicação dos fornecedores com os consumidores em potencial por meio do correio eletrônico, proporcionando o envio de *spams*, disseminando, assim, mensagens publicitárias sem autorização. Esse comércio de dados pessoais que se estabelece na Internet arrosta diretamente com o artigo 43, do Código do Consumidor, que estabelece, em sua redação, que o consumidor deverá ter acesso às suas informações registradas e ser comunicado da abertura de banco de dados que detenha suas informações de caráter pessoal.

> *"Art. 43. O consumidor, sem prejuízo do disposto no art. 86, terá acesso às informações existentes em cadastros, fichas, registros e dados pessoais e de consumo arquivados sobre ele, bem como sobre as suas respectivas fontes.*
>
> *§ 1° Os cadastros e dados de consumidores devem ser objetivos, claros, verdadeiros e em linguagem de fácil compreensão, não podendo conter informações negativas referentes a período superior a cinco anos.*
>
> *§ 2° A abertura de cadastro, ficha, registro e dados pessoais e de consumo deverá ser comunicada por escrito ao consumidor, quando não solicitada por ele.(...)"*

A seguir, oferecemos um estudo sobre os mecanismos de retenção e de propagação de informações pessoais, bem como a sua legalidade, que possibilitam que os *spammers* enviem mensagens em larga escala, atingindo um grande número de usuários de e-mail.

2.2 Dados Pessoais: a tutela na Internet

As ameaças ao direito à privacidade aumentam à medida a que o progresso tecnológico propicia o desenvolvimento de novas formas de violação às informações pessoais. A Internet constitui um ambiente fértil para práticas que afrontam a privacidade alheia, pois um número significativo dos usuários não conhece os meios pelos quais as informações pessoais são coletadas. Dessa forma, é importante ressaltar que o tratamento da informação por computadores permite

não apenas o seu processamento para fins idôneos, mas também para o uso indevido dos dados pessoais.

A legalidade do armazenamento de informações pessoais através de *cookies* tem-se apresentado como uma das questões mais controversas no que se refere à tutela dos direitos da personalidade na Internet.

Cookie é um grupo de dados trocados entre o navegador e o servidor de páginas, colocado num arquivo texto criado no computador do usuário. Eles podem ser utilizados para autenticação, rastreamento de sessões, armazenamento de preferências, lista de produtos colocados em carrinho de compras (comum em *sites* de vendas), identificador de sessão ou para guardar qualquer outra espécie de informação que possa ser armazenada em arquivo texto.

O emprego dos *cookies* desempenhou uma função de grande importância para o sucesso da Internet, na medida em que ele permite ao usuário obter uma navegação personalizada. Entretanto, inúmeras práticas ilícitas, que representam séria ameaça à privacidade, têm sido praticadas por intermédio da sua utilização.

A violação da privacidade do usuário pode ser analisada em três momentos diferentes da utilização dos *cookies*: a coleta, o armazenamento e a utilização dos dados pessoais (Lemos *et al.*, 2007).

Com relação ao armazenamento de dados, é importante destacar que o usuário deve ter noção de que algumas informações pessoais fornecidas podem ser coletadas quando este acessar um *site*. No Ordenamento Jurídico brasileiro, a questão está tutelada, no âmbito das relações de consumo. Destacamos que o Código de Defesa do Consumidor, em seu Capítulo V, Seção VI, contempla uma regulamentação especial relacionada aos bancos de dados e cadastros formados a partir de informações dos consumidores (Lemos *et al.*, 2007). Como previsto no artigo 43, várias obrigações são impostas aos administradores dos bancos de dados, como, por exemplo, demonstrar a cada consumidor a informação coletada a seu respeito, como vemos a seguir:

> *Art. 43. §3.º O consumidor, sempre que encontrar inexatidão nos seus dados e cadastros, poderá exigir sua imediata correção, devendo o arquivista, no prazo de cinco dias úteis, comunicar a alteração aos eventuais destinatários das informações incorretas."*

Assim sendo, o Direito brasileiro não possibilita que informações pessoais sejam armazenadas sem o consentimento do consumidor. Todavia, essa prática tem sido descumprida violando diretamente o Código de Defesa do Consumidor.

Em relação ao armazenamento das informações pessoais, o usuário deverá ter acesso aos seus dados uma vez constantes do banco de dados da empresa que explora o *site*, sendo-lhe permitido exigir a sua correção, caso encontre alguma inexatidão. O descumprimento da requisição realizada pelo usuário submete o infrator às disposições do disposto no art. 84 do Código de Defesa do Consumidor, podendo o mesmo, sob pena de multa, ser condenado a cumprir a sua obrigação de fazer ou, até mesmo, pagar indenização por perdas e danos que forem causados ao

usuário.

A utilização das informações armazenadas tem o objetivo de proteger a pessoa cujas informações foram coletadas contra a utilização indevida de seus dados pessoais. Entretanto, destaca-se nesse contexto a prática disseminada na Internet de venda de cadastros, sem que seja notificado o usuário que forneceu os dados.

A Fundação Vanzolini[2], em cooperação com a Universidade de São Paulo – USP, em junho de 2000, publicou uma Norma Padrão para os *sites* brasileiros adequarem-se aos níveis internacionais de políticas de privacidade. A elaboração da NRPOL – Norma de Referência da Privacidade Online[3] obteve o patrocínio de diversas empresas e, progressivamente, vem alcançando efeitos positivos no mercado virtual brasileiro.

Vários são os princípios éticos determinados pela NRPOL, com a finalidade de se conservar a privacidade do usuário da Internet, a serem utilizados pelos *sites* brasileiros. Tais como:

i) o acesso irrestrito do usuário às informações armazenadas ao seu respeito;

ii) a segurança de que a informação, uma vez recolhida, é apropriada e de que não será utilizada para propósitos diversos daqueles que ocasionaram o seu recolhimento; e

iii) o emprego, pela empresa recolhedora dos dados, de procedimentos que previnam danos e o uso, sem autorização, dessas informações.

Em síntese: o emprego dos *cookies* não representa em si uma violação ao direito da privacidade. Por outro lado, a forma como são realizados os procedimentos de coleta, armazenamento e utilização das informações pessoais é que pode determinar a ilegalidade da conduta do administrador do banco de dados.

2.3 Modelos Estrangeiros

Devido ao volume significativo de *spams* em circulação, alguns países passaram a propor e a testar diversas possibilidades de solução. Em geral, as medidas adotadas podem ser divididas em três categorias: a auto-regulação e o recurso a normas sociais; as medidas técnicas e a via jurídica (Sorkin, 2001). A via judicial e a via legislativa passaram a ser utilizadas com maior frequência nos últimos anos. Nesta seção, são apresentadas as linhas gerais de ação no combate ao *spam* empregadas nos Estados Unidos e na Europa, com ênfase na via legislativa e na sua eficácia (Cho & LaRose, 1999).

O fenômeno do *spam* na Europa se manifestou posteriormente aos Estados Unidos, e percebido de maneira um pouco diferenciada. Esta diferença justifica-se pelo fato de ter sido os Estados Unidos o país pioneiro em que o e-mail tornou-se uma ferramenta amplamente utilizada, como também foi ali que a Internet começou a ser empregada com fins comerciais e de

[2]Consulte em `http://www.vanzolini.org.br/`.
[3]Disponível em `http://www.privacidade-vanzolini.org.br/`.

forma massificada. Estes motivos justificam um certo atraso na percepção econômica e social do fenômeno do *spam* no território europeu.

Após ponderar determinadas diferenças básicas, hoje, pode-se dizer que é comum tanto aos Estados Unidos quanto à Europa, a percepção do *spam* como um grave problema a ser imediatamente confrontado, pois é capaz de comprometer a utilização das redes de comunicação em sua eficiência e confiabilidade.

2.3.1 O sistema norte-americano

Nos Estados Unidos a internet passou a apresentar grande importância e dimensão, isto somado a outros fatores como o fato do país possuir uma forte tradição de *marketing* direto, um tanto mais hostil que em outros países, geraram o acumulo do que é, possivelmente, a maior experiência judicial em matéria de *spam*.

Observando o âmbito legislativo norte-americano, desde uma fase que poderíamos atribuir como anterior ao enorme impacto proporcionado pela Internet nas comunicações eletrônicas, podem ser mencionadas algumas normas interessantes ao tema como a legislação federal existente desde 1991, sobre chamadas telefônicas e por fax comerciais com fins de propaganda, o TCPA (*Telephone Consumer Protection Act.*). (Lemos *et al.*, 2007).

Em termos de jurisprudência, os norte-americanos possuem uma experiência bastante razoável em relação ao *spam*. As demandas judiciais mais frequentes são: ações de operadoras de serviços de telecomunicações e provedores de Internet contra *spammers* pela utilização não autorizada de seus sistemas, ações contra operadores de *open relays*[4], ações contra *spammers* que utilizam falsas identidades para obter acesso a sistemas ou para enviar mensagens não autorizadas e, ações de provedores de acesso contra usuários que utilizam os seus serviços para enviar *spam* (Sorkin, 2001).

O primeiro estado americano a publicar uma lei contra *spam* foi Nevada, consecutivamente Califórnia, Washington e Virginia, todos entre 1997 e 1998[5]. Estas leis apresentam um elenco de medidas variadas de combate ao *spam*, que podem ser resumidas em (Lemos *et al.*, 2007): repressão à falsificação da identidade do remetente nos cabeçalhos do e-mail, requisição de uma indicação de que se trata de uma mensagem comercial no campo "Assunto:" do e-mail (*labelling*), reconhecimento formal das políticas anti-*spam* dos provedores de acesso, previsão de ressarcimento por cada mensagem recebida caracterizada como *spam* e até mesmo, no caso do estado de Delaware, na obrigatoriedade da preexistência de um relacionamento comercial entre remetente e destinatário para legitimar o envio do e-mail. Até o momento 38 estados norte-americanos possuem leis próprias sobre *spam*[6] (Lemos *et al.*, 2007).

Em 2003 foi finalmente aprovado o CAN-SPAM Act (*Controlling the Assault of Non-Sollicited*

[4]Servidores de e-mail que permitem sua utilização indiscriminada por usuários da Internet. Eles costumam ser utilizados com o intuito de disseminar *spam* ao mesmo tempo que dificultam a localização de sua origem.

[5]Disponível em `http://www.spamlaws.com` e `http://www.cauce.org/legislation`.

Pornography and Marketing Act)[6], essa normativa prescreve um sistema padrão para o envio de mensagens comerciais não solicitadas, como também fortalece o papel da *Federal Trade Commission* (FTC)[7] como entidade cuja função é combater o *spam* em nível nacional.

O CAN-SPAM Act institui a necessidade do e-mail possuir um endereço eletrônico válido ou outro mecanismo para que o destinatário possa requisitar não receber outras mensagens – conhecido como mecanismo de *opt-out*. Entre as medidas adotadas, estão mecanismos de inibição, bem como obrigações de ressarcimento (com penalidades estipuladas em até US$ 250,00 por e-mail, até o limite de dois milhões de dólares para o *spammer* que desrespeitar as medidas).

Segundo Lemos *et al.* (2007), além da adoção do sistema de *opt-out*, o CAN-SPAM Act realizou uma solicitação expressa à FTC para investigação sobre a viabilidade da formação de uma lista do gênero *Do-Not-Call List*[8], possuindo os endereços eletrônicos das pessoas que, expressamente, não desejassem receber mensagens comerciais não solicitadas, indesejadas. A FTC, após analisar a questão, declarou-se contra à criação de uma listagem do gênero, ao menos até que fossem implementados meios capazes de autenticar exatamente a origem do e-mail considerado como *spam*[9].

Em termos quantitativos, apesar das diversas prisões já ocorridas e a aplicação de severas multas, a eficácia da norma ainda está para ser demonstrada (Kraut *et al.*, 2005).

2.3.2 O sistema europeu

O tratamento do *spam* empregado na Europa recebeu influência direta do direito comunitário. A União Europeia não editou uma normativa exclusivamente para o problema do *spam*, porém o tema foi tratado em algumas importantes diretivas. O tratamento legislativo desse tema foi conduzido para o campo da proteção dos dados pessoais, matéria que já conta com um tempo de maturação razoável no espaço jurídico europeu (Lugaresi, 2004).

O enfoque dado pela legislação comunitária ao *spam* visa atingir dois objetivos: o primeiro, prático, de reduzir o volume de *spams*, e o segundo, ético, de procurar garantir o controle individual sobre o fluxo de informações (Lugaresi, 2004).

A primeira disposição legal no direito comunitário que produziu efeitos sobre a prática do *spam* é a Diretiva 95/46/CE, de 1995, a qual institui a disciplina geral de proteção dos dados pessoais. Ela especifica que, no tratamento informatizado de dados pessoais, estes devem ser recolhidos por meios lícitos e para finalidades determinadas e precisas, entre outras disposições.

[6]Para informações, consulte `http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=108_cong_public_laws&docid=f:publ187.108.pdf`.

[7]Disponível em `http://www.ftc.gov/spam/`.

[8]Lista mantida pela *Federal Trade Commission* com a finalidade de regular a utilização do marketing direto telefônico nos Estados Unidos.

[9]A FTC considerou basicamente as dificuldades técnicas devidas ao atual estado tecnológico da rede que tornam possível ao *spammer* dificultar, forjar ou mesmo impossibilitar a sua localização. Também, foram levados em conta outros fatores, entre eles o risco à privacidade dos próprios integrantes da lista. A íntegra do relatório elaborado pala FTC entitulado "*National do not email registry*" encontra-se disponível em `http://www.ftc.gov/reports/dneregistry/report.pdf`.

Esta diretiva não abordava diretamente o problema do *spam*, porém forneceu base jurídica para as futuras previsões normativas referentes ao tema (Lemos *et al.*, 2007).

Alguns países europeus, após a Diretiva 95/46/CE, procuraram tratar o problema do *spam* no âmbito do direito interno, prevendo o que viria a se tornar a solução padrão no espaço comunitário, ou seja, a introdução do princípio da autorização preliminar para o envio de mensagens comerciais – abordagem conhecida por *opt-in*.

Os primeiros sinais de uma normatização europeia que aborda diretamente o tema do *spam* podem ser observados na diretiva sobre contratos à distância, de 1997, que tratando da propaganda direta estabelecia a regra do *opt-in* para comunicações realizadas mediante fax e chamadas telefônicas automáticas, e um sistema de *opt-out* para as demais (que incluíam, portanto, o e-mail).

No âmbito do direito comunitário, a Diretiva sobre Comércio Eletrônico, de 2000, mencionou diretamente, pela primeira vez, o e-mail. Na diretiva era proposto uma técnica que previa um sistema que identifica a mensagem comercial não-solicitada no campo do assunto do e-mail, técnica conhecida como *labelling*. Nela também era prevista a possibilidade de medidas que obrigam aos remetentes de e-mails comerciais a consultarem listas de usuários que optaram por não receber tal mensagens. Entretanto, a aplicabilidade destas medidas foram prejudicadas em função de alguns problemas, sendo um deles o fato de que para eficácia da técnica de *labelling* é necessária elevada harmonização entre todas as partes envolvidas para que possa ser efetivada – aspecto com o qual a diretiva não se preocupou; sem contar que as próprias listas de *opt-out* não existiam e nem, sequer, as implantaram posteriormente. Quanto ao *opt-in*, a diretiva manteve a política de deixá-lo como uma opção a ser considerada por cada estado membro no seu direito interno, não o impondo nem o incentivando de modo especial (Lemos *et al.*, 2007).

Por sua vez, a Diretiva sobre Privacidade nas Comunicações Eletrônicas, de 2002, trouxe a abordagem definitiva do problema do *spam*, na esfera do direito comunitário. Esta diretiva foi editada no momento o qual os efeitos do *spam* já refletiam com bastante nitidez e a demanda por barreiras era crescente. Assim, este foi o primeiro instrumento legal que cuidou diretamente desta problemática no direito comunitário.

O método proposto na diretiva mencionada resume-se essencialmente à adoção de um sistema *opt-in*, conforme disposto em seu artigo 13(1):

> *"A utilização de sistemas de chamada automatizados sem intervenção humana (aparelhos de chamada automáticos), de aparelhos de fax ou de correio eletrônico para fins de comercialização direta apenas poderá ser autorizada em relação a assinantes que tenham dado o seu consentimento prévio.*

Dentre as motivações para a adoção do regime de *opt-in* estão o crescimento desenfreado do volume de *spam* na União Europeia que foi verificado durante o período de apreciação da proposta que culminou na diretiva, bem como a observação de que era o sistema de *opt-in* que poderia oferecer uma maior proteção contra o *spam*. Foi considerado, também, que um sistema

de *opt-in* poderia ser executado de forma mais fácil e a partir de um quadro normativo mais simples do que o de *opt-out* (Sorkin, 2001; Lugaresi, 2004).

Um dos países que enfrentaram o o problema do *spam* antes mesmo da Diretiva 2002/58/CE, foi a Itália que estabeleceu regras bastante próximas ao que posteriormente veio a sedimentar-se com a internalização da norma comum europeia, tendo como base legal, para tal construção, a norma sobre proteção de dados pessoais já existente (Lemos *et al.*, 2007).

Em outros países europeus, a solução empregada não foi diversa. A França, por exemplo, proibiu a utilização comercial do correio eletrônico "utilizando-se das coordenadas de uma pessoa física" caso esta pessoa física não expresse sua autorização prévia para receber essas mensagens[10]. Na Espanha, proibiu-se o envio de "comunicações comerciais não autorizadas" por meios eletrônicos[11].

2.3.3 Conclusões sobre os modelos apresentados

Após serem apontadas as principais abordagens empregadas pelos EUA e pela Europa para combater o problema do *spam*, podemos traçar algumas considerações gerais a respeito.

Na Tabela 2.1 são apresentadas as leis adotadas por diversos países para regulamentar o envio do *spam*[12].

As considerações europeias com relação ao *spam* foi, desde seu início, limitadas, tanto é que as discussões acabaram concentrando-se na adoção do *opt-in* ou do *opt-out*, sem que fossem consideradas com tanta reflexão outras opções e outras vias que não a legal. Em contrapartida, a estrutura do estatuto jurídico do *spam* no ordenamento jurídico comunitário está solidamente fundada, mesmo por estarem vinculados à proteção de dados pessoais, uma experiência que já é razoavelmente madura.

Nos Estados Unidos, as opções de embate contra o *spam*, aparentemente, foram exploradas com mais criatividade, tendo em vista a riqueza de abordagens e inúmeros ensaios de confronto ao problema, legalmente ou não. No entanto, alguns fatores importantes impediram, de maneira geral, que a solução norte-americana apresentasse efetivamente seus resultados mais interessantes do que a europeia.

Uma grande preocupação na tomada de medidas legais para combater o *spam* é que o fato de estabelecer regras para regulá-lo pode acabar tornando lícitos, determinados tipos de *spam* e, consequentemente, legitima-los incentivando, ainda mais, esta prática.

[10]Consulte `http://www.legifrance.gouv.fr/html/actualite/actualite_legislative/decrets_application/2004-575.htm`.

[11]Disponível em `http://www.mityc.es/dgdsi/lssi/normativa/Paginas/normativa.aspx`.

[12]As informações apresentadas foram coletadas nas seguintes fontes: `http://www.oecd-antispam.org/countrylaws.php3` e `http://www.itu.int/osg/spu/spam/legislation/Background_Paper_ITU_Bueti_Survey.pdf`.

País	Legislação
Alemanha	*Gesetz gegen Unlauteren Wettbewerb (UWG)*
Argentina	*Personal Data Protection Act (2000)*
Austrália	*Spam Act* 2003
Áustria	*Austrian Telecommunications Act (1997)*
Bélgica	*Loi du 11 mars 2003*
Canadá	*Personal Information Protection and Electronic Documents Act*
China	*Regulations on Internet E-Mail Services (2006)*
Chipre	*Regulation of Electronic Communications and Postal Services Law*
Cingapura	*Spam Control Act 2007*
Dinamarca	*Danish marketing practices act*
Espanha	*Act 34/2002 on Inf. Society Services and Electronic Commerce*
EUA	*CAN-SPAM Act (2003)*
Finlândia	*Act on Data Protection in Electronic Communications (516/2004)*
França	*Loi du 21 juin 2004 pour la confiance dans l'économie numérique*
Hungria	*Act CVIII of 2001 on Electronic Commerce*
Indonésia	*Undang-undang Informasi dan Transaksi Elektronic* (ITE)
Inglaterra	*Privacy and Electronic Communications Regulations 2003*
Israel	*Communications Law, 1982* (Emenda 2008)
Itália	*Data Protection Code (Legislative Decree no. 196/2003)*
Japão	*The Law on Regulation of Transmission of Specified Electronic Mail*
Malásia	*Communications and Multimedia Act 1998*
Malta	*Data Protection Act (CAP 440)*
Holanda	*Dutch Telecommunications Act*
Nova Zelândia	*Unsolicited Electronic Messages Act 2007*
Paquistão	*Prevention of Electronic Crimes Ordinance 2007*
Rep. Checa	*Act No. 480/2004 Coll., on Certain Information Society Services*
Suécia	*Marknadsföringslagen (1995:450)*
U. Europeia	*Directive on Privacy and Electronic Communications*

Tabela 2.1: Leis adotadas por países para regulamentar o envio do *spam*.

2.4 Análise do Spam no Ordenamento Jurídico Brasileiro

O crescimento da utilização e aplicação comercial da Internet no Brasil, a partir do final da década de 90, fez com que diversos projetos relacionados ao tratamento de mensagens eletrônicas não solicitadas fossem submetidos às Casas Civis. Dentre estes, destaca-se o projeto nº 2.186 de 2003, de autoria do Deputado Nelson Proença, aprovado com unanimidade pela Comissão de Ciência e Tecnologia da Câmara dos Deputados em 6 de dezembro de 2006[13].

Entre os projetos apresentados, destaque deve ser conferido ao projeto de Lei Substitutivo oferecido pelo mesmo Deputado Nelson Proença ao Projeto de Lei nº 2.186/2003 apresentado pelo Deputado Ronaldo Vasconcellos. Outros três projetos foram incorporados ao Projeto nº

[13]Consulte `http://www.camara.gov.br/sileg/integras/428816.htm`.

2.186/2003 e a seu substitutivo. São eles (Lemos *et al.*, 2007):

(i) Projeto de Lei nº 2.423, de 2003, do Deputado Chico da Princesa, que autoriza o envio, por uma única vez, do *spam* e que tipifica o crime de enviar mensagem com arquivo ou comando destinado a inserir ou a capturar dados, código executável ou informação do destinatário, punível com reclusão de até quatro anos e multa.

(ii) Projeto de Lei nº 3.731, de 2004, do Deputado Takayama, que permite o envio do *spam* por uma única vez e sujeita o emissor à detenção de seis meses a dois anos e multa de quinhentos reais por mensagem enviada.

(iii) Projeto de Lei nº 3.872, de 2004, do Deputado Eduardo Paes, que admite o envio, por uma única vez, do *spam* e sujeita o emissor à pena de multa de duzentos reais, bem como obriga o provedor de acesso a dispor de recurso para bloquear tais mensagens.

Além disso, a pedido do Comitê Gestor da Internet (CGI), em Lemos *et al.* (2007), os autores apresentam uma proposta de anteprojeto de lei para combater o envio de *spams*. O anteprojeto sugere a adoção do sistema *opt-in*, que prevê que e-mails comerciais só podem ser enviados depois da concordância por parte do destinatário. O sistema atual usual, o *opt-out*, exige que os remetentes ofereçam aos destinatários a opção de avisar que não desejam mais receber mensagens daquele endereço.

De acordo com os autores, o modelo atual, baseado na legislação norte-americana, "legitima" o *spam*, pois não pune o remetente que envia a mensagem não-desejada pela primeira vez. O anteprojeto também pretende evitar a criminalização da atividade de *spam*, mas desestimular seu envio como ferramenta de publicidade. A lei, se aprovada, também vai permitir que associações como o Procon, enviem reclamações coletivas de abusos à justiça (Lemos *et al.*, 2007).

Apesar da urgência da adoção de soluções por meios jurídicos, até o presente momento o anteprojeto não foi aprovado.

Finalmente, em 29 de outubro de 2009, a Secretaria de Assuntos Legislativos do Ministério da Justiça (SAL/MJ), em parceria com a Escola de Direito do Rio de Janeiro da Fundação Getúlio Vargas, lançou o projeto para a construção colaborativa de um Marco Civil da Internet no Brasil.

O marco a ser proposto tem o propósito de determinar de forma clara os direitos e as responsabilidades relativas à utilização dos meios digitais. O foco, portanto, é o estabelecimento de uma legislação que garanta direitos, e não uma norma que restrinja liberdades[14]. Um exemplo foi o caso da modelo Daniela Cicarelli. Após processar o site YouTube, pelo fato de usuários terem postado um vídeo íntimo da modelo, um juiz havia determinado que o site de vídeos fosse bloqueado no país.

A principal justificativa é que a ausência de um marco civil tem gerado incerteza jurídica quanto ao resultado de questões judiciais relacionadas ao tema. A falta de previsibilidade, por

[14]Para mais informações, consulte `http://culturadigital.br/marcocivil/`.

um lado, desincentiva investimentos na prestação de serviços por meio eletrônico, restringindo a inovação e o empreendedorismo. Por outro, dificulta o exercício de direitos fundamentais relacionados ao uso da rede, cujos limites permanecem difusos e cuja tutela parece carecer de instrumentos adequados para sua efetivação. Entretanto, o processo de elaboração normativa sobre o tema deve ter o cuidado de se ater ao essencial. A natureza aberta e transnacional da Internet, bem como a rápida velocidade de sua evolução tecnológica, podem ser fortemente prejudicados por uma legislação que tenha caráter restritivo. Qualquer iniciativa de regulamentação da Internet deve, portanto, observar princípios como a liberdade de expressão, a privacidade do indivíduo, o respeito aos direitos humanos e a preservação da dinâmica da Internet como espaço de colaboração.

Dentre os temas a serem abordados na discussão do marco civil, incluem-se regras de responsabilidade civil de provedores e usuários sobre o conteúdo postado na Internet e medidas para preservar e regulamentar direitos fundamentais do internauta, como a liberdade de expressão e a privacidade. Também poderão ser abordados princípios e diretrizes que visem a garantir algumas das premissas de funcionamento e operacionalidade da rede, como a neutralidade da Internet.

Destaca-se, no entanto, que a discussão não abrangerá de forma aprofundada temas que vêm sendo discutidos em outros foros e/ou que extrapolam a questão da Internet, como direitos autorais, crimes virtuais, comunicação eletrônica de massa e regulamentação de telecomunicações, dentre outros.

O debate sobre o Marco Civil da Internet encerrou na meia-noite do dia 30 de maio de 2010. Posteriormente, uma comissão do Ministério da Justiça reunirá as contribuições enviadas por internautas e empresas e consultarão os ministérios envolvidos (Ministérios da Cultura e da Casa Civil) para, em seguida, enviar o projeto de lei para votação na Câmara dos Deputados. Objetiva-se que a votação ocorra no segundo semestre de 2010.

Segundo dados divulgados pelo Comitê Gestor da Internet no Brasil (CGI)[15], a explicação para o Brasil ser um dos maiores emissores de *spams* reside no fato dele ser o país com maior número de máquinas comprometidas ou mal configuradas sendo abusadas por *spammers*, ou seja, o problema não é o número de *spammers* presentes no Brasil, mas um número muito grande, e crescente, de máquinas de usuários finais, conectadas via banda larga, que não possuem proteção, tornando-se alvos fáceis para *spammers* do mundo todo. Portanto, é importante salientar que o problema que está ocorrendo no Brasil não reduzirá com a definição de uma lei, pois para diminuir estes abusos é necessário um conjunto de ações que envolvem a adoção de políticas pelas Operadoras de Telecomunicações e, principalmente, a conscientização dos usuários sobre a necessidade de adotar uma postura mais pró-ativa na Internet.

O fato mais preocupante é que existe o risco cada vez maior de que o e-mail, dentro de alguns anos, deixe de ser o meio de comunicação simples, eficiente e acessível que é hoje, caso

[15]Consulte `http://www.cgi.br/publicacoes/documentacao/spam.htm`.

a tendência ao crescimento desenfreado do volume de *spams* não seja revertida. Neste sentido, uma correta avaliação e integração no plano internacional de combate ao *spam* não é somente um auxílio precioso – é uma necessidade de primeira ordem.

Se, por um lado, os métodos diretos ainda estão em gestação no Brasil e engatinham no restante do mundo, os métodos indiretos avançam com passos largos. Nos últimos anos, os métodos de classificação e filtragem automática de *spams* vêm ganhando cada vez maior destaque. A principal justificativa é que a ausência ou a ineficiência dos métodos diretos fazem com que a sobrevivência do sistema atual de e-mail dependa diretamente da evolução dos métodos indiretos.

Nos capítulos seguintes, são abordadas as principais técnicas de classificação de *spams*, bem como são oferecidas diversas contribuições nesse campo de pesquisa.

Capítulo 3

Métodos Indiretos – Parte I: Conceitos Básicos e Análise de Filtros

A ausência de métodos diretos, que poderiam ser empregados no combate contra o *spamming*, aliada ao crescente volume de lixo virtual em circulação têm impulsionado a criação e o desenvolvimento de métodos indiretos como possível alternativa capaz de sustentar a sobrevivência do atual sistema de e-mail.

Diversos métodos têm sido propostos para realizar a classificação automática de *spams*, tais como filtros baseados em regras, listas negras, filtros colaborativos, técnicas que levam em consideração o domínio do remetente, dentre muitos outros. Entretanto, dentre todas essas abordagens, algoritmos de aprendizado de máquina vêm obtendo maior sucesso (Androutsopoulos *et al.*, 2004; Cormack, 2008). Tais métodos incluem técnicas que são consideradas as melhores em categorização de textos, como algoritmos de indução de regras (Cohen, 1995, 1996), Rocchio (Joachims, 1997; Schapire *et al.*, 1998), Boosting (Carreras & Marquez, 2001), aprendizado baseado em memória (Androutsopoulos *et al.*, 2000c), máquinas de vetores de suporte (SVM) (Drucker *et al.*, 1999; Kolcz & Alspector, 2001; Hidalgo, 2002) e classificadores Bayesianos (Sahami *et al.*, 1998; Androutsopoulos *et al.*, 2000a). Dentre eles, os dois últimos vêm ganhando considerável destaque por serem muito populares em filtros comerciais ou abertos (*open-source*). Isso deve-se, provavelmente, a sua simplicidade de implementação, baixa complexidade computacional e acurácia, características que são comparáveis aos métodos de aprendizado mais elaborados (Metsis *et al.*, 2006).

Neste capítulo, inicialmente, são apresentados os conceitos básicos necessários para a compreensão do restante deste manuscrito, tais como: representação adotada, bases de dados de e-mail e medidas de desempenho. Em seguida, são apresentados os principais métodos de classificação disponíveis tanto na literatura quanto comercialmente. Nesse sentido, é oferecida uma comparação de desempenho entre os filtros mais populares: SVM e classificadores Bayesianos.

Este capítulo está estruturado da seguinte maneira: a Seção 3.1 apresenta a representação adotada pela maioria dos métodos de classificação baseados em técnicas de aprendizagem de máquina. As principais bases de dados de e-mail utilizadas na avaliação de filtros anti-*spam*

estão descritas na Seção 3.2. A Seção 3.3 apresenta uma análise das principais medidas de desempenho empregadas para avaliar os resultados obtidos por diferentes classificadores anti-*spam*. Os principais classificadores, bem como uma comparação de desempenho, são apresentados na Seção 3.4. Finalmente, conclusões são oferecidas na Seção 3.5.

3.1 Representação

Considerando que cada mensagem m é composta por um conjunto de termos (*tokens*) $m = t_1, \ldots, t_n$, sendo que cada termo t_k corresponde a uma palavra ("adulto"), um conjunto de palavras ("para ser removido") ou um único caractere ("$"), pode-se representar cada mensagem como um vetor $\vec{x} = \langle x_1, \ldots, x_n \rangle$, onde $x_1, \ldots, x_n$ são valores dos atributos $X_1, \ldots, X_n$ associados aos termos $t_1, \ldots, t_n$. No caso mais simples, cada termo representa uma única palavra e todos os atributos são Booleanos: $X_i = 1$ se a mensagem contém t_i ou $X_i = 0$, caso contrário.

Alternativamente, os atributos podem ser valores inteiros obtidos a partir da frequência do termo (*term frequency* – TF) representando quantas vezes cada termo ocorre na mensagem. Esse tipo de representação oferece mais informação que a Booleana (Metsis *et al.*, 2006). Uma terceira alternativa é associar cada atributo X_i a um TF normalizado, $x_i = \frac{\sharp t_i(m)}{|m|}$, onde $\sharp t_i(m)$ corresponde ao número de ocorrências do termo representado por X_i em m, e $|m|$ representa o comprimento de m mensurado pelas ocorrências dos termos. TF normalizado leva em consideração a repetição dos termos em relação ao tamanho da mensagem. Ele é similar à pontuação TF–IDF (*term frequency-inverse document frequency*), comumente empregada em recuperação de informação, sendo que o componente IDF poderia ser adicionado para denotar termos que são comumente encontrados nas mensagens de treinamento.

3.2 Bases de Dados

Existe um esforço significativo para gerar bases de dados públicas que possam ser aplicadas na filtragem de *spams*. Um dos maiores problemas está relacionado à proteção de informações privadas dos usuários, cujas mensagens legítimas são incluídas nas bases.

A primeira base pública que surgiu é a Ling-Spam (Androutsopoulos *et al.*, 2000a), que é composta por arquivos de um fórum de discussão sobre linguística. Ling-Spam tem a desvantagem de conter mensagens legítimas extremamente específicas em um único tópico. Dessa forma, a classificação torna-se mais fácil e os classificadores tendem a apresentar alto desempenho.

A coleção Spambase[1] evita o problema da privacidade substituindo as palavras das mensagens por suas frequências de aparecimento. Entretanto, essa metodologia resulta em uma base mais restritiva que a Ling-Spam, já que as suas mensagens não estão disponíveis na forma

[1]Spambase está disponível em `http://www.ics.uci.edu/~mlearn/mlrepository.html`

natural e, portanto, é impossível testar os classificadores com atributos diferentes daqueles escolhidos pelos seus criadores.

Outra base pública é o PU corpora (Androutsopoulos *et al.*, 2004), que consiste em mensagens recebidas por um usuário particular após a substituição de cada termo por um valor único. O problema dessa base é que a perda dos termos originais impõe uma série de restrições, como por exemplo, é impossível testar os classificadores com diferentes formas de tokenização (*tokenizers*)[2].

Posteriormente, Metsis *et al.* (2006) apresentou seis bases conhecidas como Enron (Klimt & Yang, 2004; Shetty & Adibi, 2004; Keila & Skillicorn, 2005). Cada coleção é composta por mensagens legítimas reais extraídas da caixa de mensagens de um ex-funcionário distinto da Enron Corporation e *spams* coletados de diferentes fontes. Dessa forma, as mesmas características de uma caixa de mensagens de um usuário real são mantidas.

Concomitantemente, outras bases maiores foram apresentadas. Em 2005, ocorreu o primeiro *Spam Challenge*, evento realizado em conjunto com a *Text REtrieval Conference* (TREC). Devido ao sucesso obtido, o evento foi realizado novamente no ano seguinte. Por fim, em 2008, ocorreu o *Spam Challenge* organizado pela CEAS (*Conference on E-mail and Anti-Spam*)[3].

Atualmente, as bases de dados e metodologias empregadas na *TREC Spam Track* e no *CEAS Spam Challenge* constituem o maior e mais realista "laboratório" para avaliação de filtros anti-*spam*. A base de dados pública TREC05 contém mensagens de e-mail de 150 executivos da Enron, coletadas e disponibilizadas para o domínio público. Essas mensagens foram previamente rotuladas e, posteriormente, foram adicionadas mensagens de *spams*, alteradas para aparentar terem sido recebidas no mesmo período das mensagens legítimas (Cormack & Lynam, 2005). Por outro lado, a base TREC06 é totalmente composta por mensagens extraídas da Internet, incrementada com mensagens de *spams* de domínio público (Cormack, 2006). A principal tarefa da *TREC Spam Track – immediate feedback* (resposta imediata) – simula o envio *on-line* das mensagens para a caixa de e-mails de um usuário ideal – aquele que fornece uma resposta imediata ao filtro (Cormack & Lynam, 2005; Cormack, 2006).

O *CEAS Spam Challenge* 2008 alterou a metodologia empregada na *TREC Spam Track* em vários aspectos[4] (Cormack, 2008):

- Uma grande quantidade de mensagens foi capturada e distribuída aos participantes em tempo real;
- Os filtros foram avaliados empregando várias coleções reais e privadas de mensagens;
- O fluxo de mensagens comunicava-se com os filtros utilizando protocolos padrões (SMTP, IMAP e POP), ao invés de uma interface de linha de comando;

[2]Tokenização é o processo de "quebrar" um fluxo de texto em partes menores chamadas *tokens*. Essa técnica é massivamente empregada em linguística (onde também é conhecida como segmentação de texto) e em computação, compondo parte da análise léxica.

[3]Consulte `http://ceas.cc/2010/main.shtml`.

[4]Consulte `(http://www.ceas.cc/2008/challenge/)`.

- Os filtros foram obrigados a retornar apenas a classificação de cada mensagem (*hard classification*), ao invés do valor de probabilidade de pertencer à uma dada classe (*soft classification*);

- Os filtros foram irrestritamente autorizados a acessar recursos externos, tais como servidores de nomes e listas negras.

O *CEAS Spam Challenge* 2008 avaliou o desempenho dos filtros empregando diferentes tarefas:

- *Live simulation task* (simulação ao vivo) – oferece resposta (*feedback*) parcial e atrasada, sendo que a resposta é retardada e disponibilizada apenas para mensagens endereçadas para um subconjunto de destinatários. O principal objetivo dessa tarefa é simular uma caixa de e-mail real, a fim de explorar o impacto causado no filtro quando o *feedback* do usuário é imperfeito ou limitado.

- *Active learning task* (aprendizagem ativa) – os filtros realizam o processo de classificação sem treinamento prévio. Durante todo o desafio, cada classificador possui uma quota n de consulta, ou seja, ele poderá consultar a classe correta de apenas n mensagens.

3.3 Medidas de Desempenho

Diversas medidas têm sido propostas com o intuito de avaliar e comparar o desempenho obtido por diferentes filtros anti-*spam*. Entretanto, ainda não está claro qual é o melhor sistema de comparação, nem quais medidas oferecem uma avaliação justa (Cormack, 2008).

Infelizmente, não existe uma padronização em relação aos sistema de avaliação do desempenho dos classificadores de e-mail. Consequentemente, a maioria dos resultados disponíveis na literatura é divulgada usando diferentes medidas de desempenho, dificultando a comparação dos resultados entre trabalhos diferentes. Nesse sentido, existe um esforço por parte dos congressos e campeonatos com o propósito de incentivar que os resultados publicados nessa área sejam reportados utilizando a mesma metodologia. Dessa forma, as informações divulgadas tornam-se mais reprodutíveis.

De acordo com Cormack (2008), os filtros deveriam ser julgados em quatro dimensões: autonomia, imediatismo, identificação de *spams* e identificação de *hams* (mensagens legítimas). Entretanto, não é obvio um procedimento capaz de mensurar cada uma dessas dimensões separadamente, nem como combinar essas medidas em uma única expressão cujo propósito seja comparar o desempenho dos filtros anti-*spam*.

Sejam $\mathcal{S}$ e $\mathcal{L}$ os conjuntos de mensagens de *spam* e legítimas, respectivamente, e todos os resultados possíveis de predição apresentados na Tabela 3.1. Algumas medidas de avaliação bastante populares na literatura de aprendizagem de máquina estão apresentadas na Tabela 3.2.

Tabela 3.1: Possíveis resultados de predição de um classificador de e-mails.

Notação	Denominação	Descrição
$\mathcal{VP}$	Verdadeiros positivos	*Spams* classificados corretamente
$\mathcal{VN}$	Verdadeiros negativos	*Hams* classificados corretamente
$\mathcal{FN}$	Falso negativos	*Spams* classificados como *hams*
$\mathcal{FP}$	Falso positivos	*Hams* classificados como *spams*

Tabela 3.2: Medidas de desempenho mais empregadas para avaliar métodos de classificação.

Medida	Equação (em %)
Taxa de verdadeiro positivo (*spam recall*)	$Tvp = \frac{\lvert\mathcal{VP}\rvert}{\lvert\mathcal{S}\rvert}$
Taxa de verdadeiro negativo (*ham recall*)	$Tvn = \frac{\lvert\mathcal{VN}\rvert}{\lvert\mathcal{L}\rvert}$
Taxa de falso positivo	$Tfp = \frac{\lvert\mathcal{FP}\rvert}{\lvert\mathcal{L}\rvert}$
Taxa de falso negativo	$Tfn = \frac{\lvert\mathcal{FN}\rvert}{\lvert\mathcal{S}\rvert}$
Precisão de *spams*	$Ps = \frac{\lvert\mathcal{VP}\rvert}{\lvert\mathcal{VP}\rvert+\lvert\mathcal{FP}\rvert}$
Precisão de legítimos	$Pl = \frac{\lvert\mathcal{VN}\rvert}{\lvert\mathcal{VN}\rvert+\lvert\mathcal{FN}\rvert}$
Taxa de acurácia	$Tac = \frac{\lvert\mathcal{VP}\rvert+\lvert\mathcal{VN}\rvert}{\lvert\mathcal{S}\rvert+\lvert\mathcal{L}\rvert}$
Taxa de erro	$Ter = \frac{\lvert\mathcal{FP}\rvert+\lvert\mathcal{FN}\rvert}{\lvert\mathcal{S}\rvert+\lvert\mathcal{L}\rvert}$

Todas as medidas apresentadas na Tabela 3.2 consideram um falso negativo tão prejudicial quanto um falso positivo. Todavia, as consequências materiais causadas por falhas na identificação de mensagens legítimas e *spams* são diferentes (Cormack, 2008; Cormack & Lynam, 2007). Classificar incorretamente um e-mail legítimo aumenta o risco de perda da informação contida na mensagem, ou ao menos, poderá causar atraso na leitura. A quantidade de risco e atraso incorridos são difíceis de mensurar, assim como as suas consequências, pois dependem do conteúdo da mensagem. Por outro lado, falhas na identificação de *spams* também variam em importância, mas geralmente são menos prejudiciais que as falhas na identificação de e-mails legítimos. Vírus, códigos maliciosos e mensagens fraudulentas podem ser exceções, pois apresentam alto risco ao usuário.

Independente da medida adotada, um aspecto importante que deve ser considerado é a

assimetria dos custos envolvidos nos erros de classificação. Geralmente, um *spam* classificado incorretamente como *ham* é um problema relativamente pequeno, pois o usuário apenas necessitará removê-lo. Por outro lado, uma mensagem legítima classificada como *spam* pode ser inaceitável, pois implica na perda de informações potencialmente importantes, particularmente em configurações nas quais os *spams* são apagados automaticamente.

A seguir são apresentadas as principais medidas de avaliação que são mais utilizadas para comparar o desempenho dos filtros anti-*spam*.

3.3.1 Curva ROC – Receiver Operating Characteristic

Uma curva ROC (do inglês *Receiver Operating Characteristic* – características de operação do receptor) trata-se de um gráfico de sensibilidade por (1 - especificidade) para um classificador binário conforme varia-se o limiar de classificação. A curva ROC também pode ser representada pelo gráfico das taxas de verdadeiro positivo (Tvp) por falso positivo (Tfp).

Um espaço ROC é definido por Tfp e Tvp como eixos x e y, respectivamente, representando a relação entre verdadeiros positivos (benefício) e falsos positivos (custo). Sendo Tvp equivalente a sensibilidade e Tfp igual a (1 - especificidade), a curva ROC é, algumas vezes, simplesmente chamada de gráfico de sensibilidade por (1 - especificidade). Cada resultado de predição representa um ponto no espaço ROC (Figura 3.1).

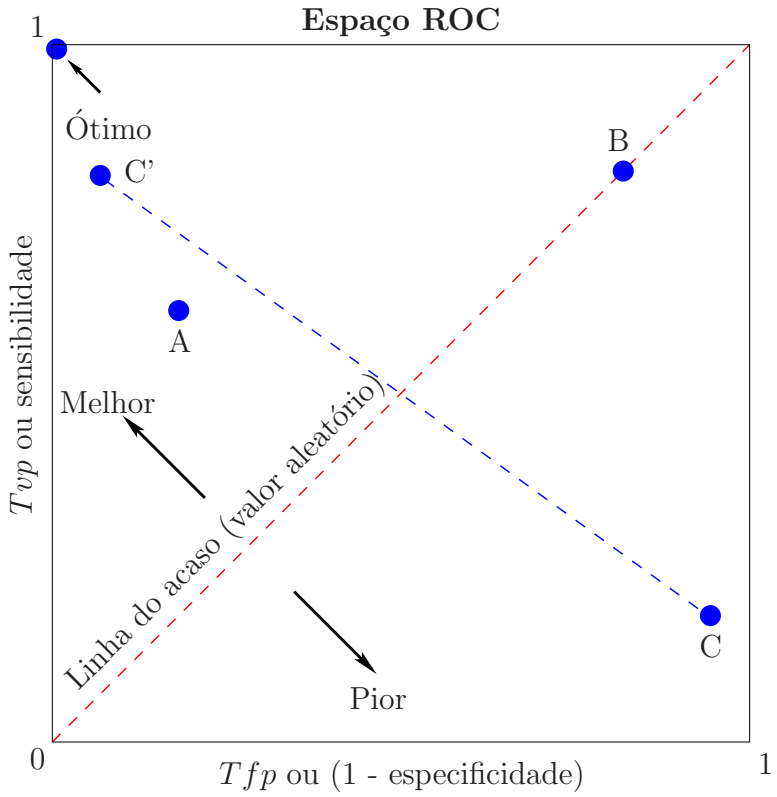


Figura 3.1: Exemplo de espaço ROC.

O melhor classificador possível produziria um ponto localizado no canto esquerdo superior,

coordenada $(0, 1)$ do espaço ROC, representando 100% de sensibilidade (nenhum falso negativo) e 100% de especificidade (nenhum falso positivo). O ponto $(0, 1)$ é também conhecido como ponto ótimo ou perfeito.

Na Figura 3.1, quatro pontos representam resultados de predição obtidos por métodos diferentes. O resultado do método A claramente ilustra que a sua capacidade de predição é superior às de B e C. O resultado obtido por B está localizado em cima da linha do acaso. Por outro lado, quando C é espelhado através do ponto central $(0,5; 0,5)$, o método resultante C' passa a ser superior ao classificador A. O método espelhado simplesmente inverte as predições realizadas pelo classificador C. É interessante notar que, apesar do método original C ter capacidade de predição negativa, o simples fato de inverter as suas decisões produz um novo classificador C' que possui o melhor desempenho dentre os métodos comparados. A distância entre a linha do acaso e um ponto em ambas as direções é a melhor indicação da capacidade de predição que um método possui.

Classificadores "discretos" (*hard classifiers*), como aqueles baseados em árvores de decisão ou conjunto de regras, produzem respostas binárias. Dessa forma, quando aplicados a um conjunto de instâncias, tais métodos produzem um único ponto no espaço ROC. Por outro lado, os filtros flexíveis (*soft classifiers*), como os filtros Bayesianos e máquina de vetores de suporte (SVM), retornam valores de probabilidade de uma amostra pertencer a uma determinada classe. Para esses métodos, cada configuração do limiar de classificação T produz um determinado ponto no espaço ROC. Portanto, é possível produzir uma curva dispondo-se graficamente cada ponto ROC para cada T possível.

Em resumo, a eficácia de um filtro flexível pode ser definida pelo conjunto de todos os pares distintos de (Tvp, Tfp) para diferentes valores de T. Esse conjunto de pontos caracteriza uma curva ROC de um filtro anti-*spam* (Cormack, 2008). Um filtro cuja curva ROC está estritamente acima de outra é superior em todas as situações, enquanto que um filtro cuja curva ROC cruza com outra é superior para algumas configurações de T e inferior para outras.

A área sob a curva ROC (*Area Under the ROC Curve – AUC*) fornece uma estimativa do desempenho de um classificador flexível para todas as configurações de T. AUC também possui uma interpretação probabilística: trata-se da probabilidade de um filtro atribuir pontuação maior para uma mensagem aleatória de *spam* do que para uma mensagem legítima (Cormack, 2008). O índice $(1 - AUC)\%$ é frequentemente empregado no lugar da AUC devido ao fato de que na filtragem de *spams* o valor AUC é muito próximo de 1.

Apesar das curvas ROC e do índice $(1 - AUC)\%$ serem medidas de desempenho intuitivas e eficazes para a avaliação de métodos de classificação, elas possuem pelo menos duas desvantagens: elas só podem ser empregadas para avaliar o desempenho de filtros flexíveis e é difícil comparar o desempenho de classificadores quando as suas curvas ROC cruzam-se entre si.

Atualmente, a medida $(1 - AUC)\%$ é o critério padrão de avaliação de filtros utilizada nos campeonatos anti-*spam*.

3.3.2 Logistic Average Misclassification

Outra medida de desempenho utilizada nos campeonatos é a *Logistic Average Misclassification* (Erro médio logístico – $LAM\%$). Trata-se de uma medida composta por uma única expressão, baseada somente no resultado da classificação binária (*ham* ou *spam*). Ela é calculada por

$$LAM\% = logit^{-1}\left(\frac{logit(Tfp\%) + logit(Tfn\%)}{2}\right),$$

sendo que

$$logit(p) = log\left(\frac{p}{100\% - p}\right).$$

$LAM\%$ não estabelece *a priori* nenhuma diferença na importância entre os erros de classificação de *hams* e *spams*, e estabelece igualmente um fator fixo na chance de ocorrência de ambas (Cormack, 2008).

Obviamente, $LAM\%$ sempre é maior ou igual a zero. Quanto menor o valor de $LAM\%$, melhor o desempenho do filtro. Entretanto, essa medida falha quando o filtro classifica todas as mensagens dentro de uma mesma classe.

3.3.3 Razão do Custo Total

Com o intuito de considerar a assimetria dos custos relacionados aos erros de classificação de e-mails, Androutsopoulos *et al.* (2000a) propôs um refinamento baseado na precisão e revocação de *spams* que permite avaliar o desempenho do classificador usando uma única medida. Tal proposta considera um falso positivo λ vezes mais custoso que um falso negativo, com λ igual a 1, 9 ou 999, dependendo do cenário escolhido pelo usuário. Dessa forma, cada falso positivo é contabilizado como λ erros, com acurácia ponderada (TAc_p) dada por

$$TAc_p = \frac{|\mathcal{VP}| + \lambda|\mathcal{VN}|}{|\mathcal{S}| + \lambda|\mathcal{L}|}.$$

Nesse caso, a Razão do Custo Total (RCT) pode ser calculada por

$$RCT = \frac{|\mathcal{S}|}{\lambda|\mathcal{FP}| + |\mathcal{FN}|}.$$

RCT oferece uma indicação da melhoria obtida pelo uso do filtro. Quanto maior o valor RCT, melhor o desempenho e, para $RCT < 1$, não utilizar o filtro é melhor.

3.3.4 Coeficiente de Correlação de Matthews

De acordo com Carpinter & Hunt (2006) e Cormack & Lynam (2007) o valor de λ é muito difícil de ser determinado, porque ele depende do conteúdo das mensagens legítimas, uma vez que algumas são mais importantes que outras (por exemplo, mensagens pessoais).

Além disso, o problema de usar a RCT é que ela não retorna um valor dentro de uma escala pré-definida. Considere, por exemplo, dois classificadores A e B empregados para classificar 600 mensagens (450 *spams* + 150 *hams*, $\lambda = 1$). Suponha que A obteve uma predição perfeita com $|\mathcal{FP}_A| = |\mathcal{FN}_A| = 0$ e que B classificou incorretamente apenas 3 *spams* como *hams*, então $|\mathcal{FP}_B| = 0$ e $|\mathcal{FN}_B| = 3$. Dessa forma, $RCT_A = +\infty$ e $RCT_B = 150$. Intuitivamente, pode-se observar que ambos os classificadores obtiveram desempenho similares, com uma pequena vantagem para A. Entretanto, analisando-se somente a RCT, conclui-se erroneamente que o classificador A é muito melhor do que B. Além disso, a RCT não é um valor representativo, com o qual podemos fazer análises importantes sobre o desempenho de um único classificador. Isso ocorre porque tal medida fornece apenas o ganho obtido pelo uso do filtro, mas não fornece informações sobre o quanto o classificador pode ser melhorado.

Com o intuito de evitar essas características, este trabalho propõe a utilização do coeficiente de correlação de Matthews (CCM) (Matthews, 1975) (Equação 3.1):

$$CCM = \frac{(|\mathcal{VP}|.|\mathcal{VN}|) - (|\mathcal{FP}|.|\mathcal{FN}|)}{\sqrt{(|\mathcal{VP}| + |\mathcal{FP}|).(|\mathcal{VP}| + |\mathcal{FN}|).(|\mathcal{VN}| + |\mathcal{FP}|).(|\mathcal{VN}| + |\mathcal{FN}|)}} \tag{3.1}$$

O CCM é empregado em aprendizado de máquina como medida de qualidade de classificadores binários e que fornece mais informações que a RCT e $LAM\%$. Ele retorna um valor real entre -1 e $+1$. Sendo que um coeficiente igual a $+1$, indica uma predição perfeita; 0, uma predição aleatória média; e -1, uma predição inversa.

CCM é uma medida que oferece uma avaliação mais balanceada de predição do que outras métricas, tais como a proporção de predições corretas, especialmente se as duas classes são de tamanho muito diferentes (Baldi *et al.*, 2000).

A principal vantagem desta medida é que ela fornece uma avaliação mais justa, pois, numa situação real, o número de *spams* recebidos é muito maior do que o número de mensagens legítimas. Dessa forma, CCM tende a ajustar automaticamente o quanto um falso positivo é pior que um falso negativo. À medida que a razão entre o número de *spams* e mensagens legítimas aumenta, um falso positivo tende a ser muito pior do que um falso negativo.

Usando o exemplo anterior, os classificadores teriam obtido $CCM_A = 1,000$ e $CCM_B = 0,987$. Dessa maneira, podemos tirar conclusões corretas tanto sobre as predições dos classificadores quanto sobre o desempenho obtido individualmente.

Provavelmente, a forma mais justa de avaliar o desempenho de filtros diferentes seja combinar várias medidas de avaliação, tais como $Tvp \times Tfp$, CCM e $(1 - AUC)\%$, por exemplo.

A Tabela 3.3 sumariza medidas de desempenho apresentadas nesta seção.

Tabela 3.3: Medidas de desempenho empregadas para avaliar os filtros anti-*spam*.

Medida	Denotação	Equação
Área sob a curva ROC	$(1 - AUC)\%$	–
Erro médio logístico	$LAM\%$	$logit^{-1}\left(\frac{logit(Tfp\%)+logit(Tfn\%)}{2}\right)$
Taxa de acurácia ponderada	TAc_p	$\frac{\lvert\mathcal{VP}\rvert+\lambda\lvert\mathcal{VN}\rvert}{\lvert\mathcal{S}\rvert+\lambda\lvert\mathcal{L}\rvert}$
Razão de custo total	RCT	$\frac{\lvert\mathcal{S}\rvert}{\lambda\lvert\mathcal{FP}\rvert+\lvert\mathcal{FN}\rvert}$
Coef. correlação Matthews	CCM	$\frac{(\lvert\mathcal{VP}\rvert.\lvert\mathcal{VN}\rvert)-(\lvert\mathcal{FP}\rvert.\lvert\mathcal{FN}\rvert)}{\sqrt{(\lvert\mathcal{VP}\rvert+\lvert\mathcal{FP}\rvert).(\lvert\mathcal{VP}\rvert+\lvert\mathcal{FN}\rvert).(\lvert\mathcal{VN}\rvert+\lvert\mathcal{FP}\rvert).(\lvert\mathcal{VN}\rvert+\lvert\mathcal{FN}\rvert)}}$

3.4 Classificadores

Dado um conjunto de mensagens $\mathcal{M} = \{m_1, m_2, \ldots, m_j, \ldots, m_{|\mathcal{M}|}\}$ e um conjunto de classes $\mathcal{C} = \{c_s = \texttt{spam}, c_l = \texttt{legitima}\}$, sendo que m_j é a j-ésima mensagem de $\mathcal{M}$ e $\mathcal{C}$ são as possíveis classes, a tarefa de classificação de *spams* consiste, basicamente, em uma função de categorização Booleana $\Phi(m_j, c_i) : \mathcal{M} \times \mathcal{C} \rightarrow \{\texttt{Verdadeiro}, \texttt{Falso}\}$. Quando $\Phi(m_j, c_i)$ é `Verdadeiro`, indica que a mensagem m_j pertence a categoria c_i; caso contrário, m_j não pertence a c_i.

Na classificação de *spams* existem apenas duas possíveis categorias: `spam` e `legitima`. Cada mensagem $m_j \in \mathcal{M}$ pode ser associada a apenas uma delas, mas nunca a ambas. Dessa forma, podemos utilizar uma função de categorização mais simplificada $\Phi_{\texttt{spam}}(m_j) : \mathcal{M} \rightarrow \{\texttt{Verdadeiro}, \texttt{Falso}\}$. Assim, uma mensagem é classificada como *spam* quando $\Phi_{\texttt{spam}}(m_j)$ for `Verdadeiro` e legítima, caso contrário.

A aplicação dos algoritmos de aprendizagem supervisionada com ênfase em filtragem de *spams* consiste em duas etapas (Vapnik, 1995):

1. *Treinamento*. Um conjunto de mensagens rotuladas (Tr) deve ser fornecido como fonte de treinamento. Elas são previamente convertidas para uma representação que o algoritmo consiga compreender. A representação mais empregada para filtragem de *spams* é o modelo de espaço vetorial, sendo que cada documento $m_j \in Tr$ é transformado em um vetor real $\vec{x}_j \in \Re^{|\mathcal{V}|}$, no qual $\mathcal{V}$ é o vocabulário (conjunto de características) e as coordenadas de $\vec{x}_j$ representam o peso de cada característica de $\mathcal{V}$. Assim, podemos utilizar o algoritmo de aprendizado e os dados de treinamento para criar um classificador $\Phi_{\texttt{spam}}(\vec{x_j}) \rightarrow \{\texttt{Verdadeiro}, \texttt{Falso}\}$.

2. *Classificação.* O classificador $\Phi_{\texttt{spam}}(\vec{x_j})$ é aplicado na representação vetorial de uma mensagem nova $\vec{x}$ para produzir uma predição se ela é *spam* ou não.

Atualmente, os métodos de classificação vêm recebendo destaque na filtragem de *spams*. Gigantes da computação, como Google, Yahoo, IBM, Microsoft, dentre outras, investem constantemente no avanço dessa área, por meio do desenvolvimento direto e do patrocínio de congressos e campeonatos.

Diversos métodos têm sido propostos para classificação automática de *spams*. De maneira geral, as técnicas existentes podem ser enquadradas em dois grupos: filtros baseados em regras e filtros baseados em aprendizagem de máquina (Cormack, 2008).

Os principais representantes do grupo de filtros baseados em regras são:

- métodos baseados em listas (negras, brancas e cinzas) (Ramachandran *et al.*, 2007);
- técnicas que levam em conta o domínio do remetente (Jung & Sit, 2004);
- filtros que realizam a predição por presença e/ou ausência de termos (Zhang *et al.*, 2004; Wang *et al.*, 2009).

Por outro lado, algoritmos de aprendizado de máquina vêm obtendo mais sucesso (Cormack, 2008; Guzella & Caminhas, 2009). Dentre esses métodos, destacam-se:

- Árvores de decisão (Pampapathi *et al.*, 2006; Meizhen *et al.*, 2009);
- Algoritmo indutivos (Cohen, 1995, 1996; Katakis *et al.*, 2009);
- Rocchio (Joachims, 1997; Schapire *et al.*, 1998)
- Boosting (Carreras & Marquez, 2001; Cournane & Hunt, 2004);
- Aprendizado baseado em memória (Androutsopoulos *et al.*, 2000c; Fdez-Riverola *et al.*, 2007a; Sakkis *et al.*, 2003; Luo & Chung, 2008);
- Redes neurais artificiais (Chuan *et al.*, 2005);
- Filtros colaborativos (Lemire, 2005; Boykin & Roychowdhury, 2005; Kong *et al.*, 2006);
- Heurísticas (Fdez-Riverola *et al.*, 2007b; Meizhen *et al.*, 2009; Guzella & Caminhas, 2009);
- Predição por correspondência parcial (Bratko *et al.*, 2006; Cormack & Lynam, 2005);
- Regressão logística (Perlich *et al.*, 2003; Goodman & Yih, 2006; Koprinska *et al.*, 2007);
- Compressão dinâmica Markoviana (Bratko *et al.*, 2006);
- Máquinas de vetores de suporte (*Support Vector Machine* – SVM) (Drucker *et al.*, 1999; Kolcz & Alspector, 2001; Hidalgo, 2002; Sculley *et al.*, 2006; Lynam *et al.*, 2006; Sculley *et al.*, 2006; Sculley & Wachman, 2007; Peng *et al.*, 2008; Liu & Cui, 2009); e
- Classificadores Bayesianos (Friedman *et al.*, 1997; Sahami *et al.*, 1998; Androutsopoulos *et al.*, 2000a,b; Metsis *et al.*, 2006; Seewald, 2007; Song *et al.*, 2009; Marsono *et al.*, 2009).

De acordo com vários estudos e resultados de campeonatos, dentre todas as técnicas mencionadas neste capítulo, os classificadores Bayesianos e os SVMs vêm apresentando os melhores desempenhos na tarefa de filtragem de *spams* (Zhang *et al.*, 2004; Cormack, 2008; Guzella & Caminhas, 2009). A seguir, são apresentados detalhes de funcionamento de cada um desses classificadores.

3.4.1 Filtros Anti-Spam Naive Bayes

Os filtros probabilísticos são historicamente os primeiros a terem sido propostos e até hoje são amplamente utilizados (Dunning, 1993). Tais métodos são bastante empregados na classificação automática de *spams* devido a sua simplicidade e alto desempenho (Zhang *et al.*, 2004).

O primeiro programa que utilizou recursos da probabilidade Bayesiana para filtragem de *spams* foi o iFile, desenvolvido por Jason Rennie, em 1996[5]. Todavia, o primeiro trabalho acadêmico a propor um filtro Bayesiano anti-*spam* foi apresentado por Sahami *et al.* (1998). A partir de então, diversas outras técnicas variantes têm sido propostas em uma grande quantidade de trabalhos de pesquisa e programas comercializados. Em 2002, os princípios da filtragem Bayesiana foram amplamente difundidos por Paul Graham[6].

Classificadores Bayesianos anti-*spam* tornaram-se mecanismos populares e diversos servidores modernos de e-mail passaram a utilizá-lo. Filtros locais (por usuário) consagrados, como OSBF-Lua[7], DSPAM[8], SpamAssassin[9], SpamBayes[10], Bogofilter[11] e ASSP[12] são baseados em técnicas de classificação Bayesiana.

O classificador Naive Bayes (NB) proposto por Duda & Hart (1973) é o filtro mais simples derivado da teoria da decisão Bayesiana (Bernardo & Smith, 1994). Do teorema de Bayes e da teoria da probabilidade total, sabe-se que a probabilidade de uma mensagem $\vec{x} = \langle x_1, \ldots, x_n \rangle$ pertencer a categoria $c_i \in \{c_s, c_l\}$ é:

$$P(c_i|\vec{x}) = \frac{P(c_i).P(\vec{x}|c_i)}{P(\vec{x})}.$$

Uma vez que o denominador não depende da classe, o filtro NB classifica cada mensagem na categoria que maximiza $P(c_i).P(\vec{x}|c_i)$. Isso é equivalente a classificar uma mensagem como *spam* (c_s) se

$$\frac{P(c_s).P(\vec{x}|c_s)}{P(c_s).P(\vec{x}|c_s) + P(c_l).P(\vec{x}|c_l)} > T,$$

com $T = 0,5$. Variando-se T, obtêm-se mais verdadeiros negativos e menos verdadeiros positivos ou vice-versa. A probabilidade $P(c_i)$ pode ser estimada pela frequência de documentos que pertencem a classe c_i no conjunto de treinamento Tr, enquanto que $P(\vec{x}|c_i)$ é praticamente impossível de ser estimada diretamente, pois isso requer que Tr contenha mensagens idênticas as que estão sendo classificadas. Entretanto, os classificadores NB fazem a suposição de que os termos de uma mensagem são condicionalmente independentes e que a ordem em que eles aparecem é irrelevante.

[5] Consulte `http://people.csail.mit.edu/jrennie/ifile/`

[6] Consulte `http://www.paulgraham.com/spam.html`

[7] Vencedor do *TREC's Spam Track* 2006 e do *CEAS 2008 Spam Filter Live Challenge*. Consulte `http://osbf-lua.luaforge.net/`

[8] Veja `http://dspam.nuclearelephant.com/index.shtml`

[9] Veja `http://spamassassin.apache.org/`

[10] Consulte `http://spambayes.sourceforge.net/`

[11] Consulte `http://bogofilter.sourceforge.net/`

[12] Veja `http://assp.sourceforge.net/`

Apesar dessa suposição ser um tanto quanto simplista, diversos estudos indicam que o classificador NB é surpreendentemente eficaz na filtragem de *spams* (Androutsopoulos *et al.*, 2004; Zhang *et al.*, 2004; Metsis *et al.*, 2006; Song *et al.*, 2009).

Embora esses filtros sejam bastante estudados e utilizados para a classificação de *spams*, um fato não muito contemplado pela literatura é a existência de várias formas distintas de estimar as probabilidades Bayesianas $P(\vec{x}|c_i)$, sendo que cada uma delas representa uma versão diferente de filtro anti-*spam* NB (Almeida *et al.*, 2009, 2010b; Almeida & Yamakami, 2010).

A seguir, são apresentadas sete versões distintas de classificadores anti-*spam* NB.

NB Básico

Conhece-se por NB básico o primeiro classificador anti-*spam* NB proposto por Sahami *et al.* (1998). Seja $\mathcal{T}' = \{t_1, \ldots, t_n\}$ o conjunto de termos (*tokens*) extraídos das mensagens, cada mensagem m é representada por uma vetor binário $\vec{x} = \langle x_1, \ldots, x_n \rangle$, sendo que o valor de cada x_k indica se o termo t_k ocorreu ou não em m. As probabilidades $P(\vec{x}|c_i)$ são calculadas por

$$P(\vec{x}|c_i) = \prod_{k=1}^{n} P(t_k|c_i),$$

e o critério para classificar a mensagem como *spam* é dado por:

$$\frac{P(c_s).\prod_{k=1}^{n} P(t_k|c_s)}{\sum_{c_i \in \{c_s, c_l\}} P(c_i).\prod_{k=1}^{n} P(t_k|c_i)} > T.$$

Assim, as probabilidades $P(t_k|c_i)$ são estimadas por

$$P(t_k|c_i) = \frac{|Tr_{t_k,c_i}|}{|Tr_{c_i}|},$$

sendo $|Tr_{t_k,c_i}|$ a quantidade de mensagens de treinamento da categoria c_i que contêm o termo t_k e $|Tr_{c_i}|$ o número total de mensagens de treinamento pertencentes a categoria c_i.

NB Multinomial com Frequência de Termos

No classificador NB multinomial com frequência de termos (NB MN FT), cada mensagem é representada como um conjunto de termos $m = \{t_1, \ldots, t_n\}$, no qual cada t_k informa a quantidade de vezes que ele aparece em m. Nesse caso, m pode ser representado por um vetor $\vec{x} = \langle x_1, \ldots, x_n \rangle$, tal que cada x_k corresponde ao número de ocorrências de t_k em m. Além disso, cada mensagem m de categoria c_i pode ser interpretada como o resultado da escolha independente de $|m|$ termos de $\mathcal{T}'$ como substitutos e probabilidade $P(t_k|c_i)$ para cada t_k (McCallum & Nigam, 1998). Portanto, $P(\vec{x}|c_i)$ é a distribuição multinomial:

$$P(\vec{x}|c_i) = P(|m|).|m|!.\prod_{k=1}^{n} \frac{P(t_k|c_i)^{x_k}}{x_k!}.$$

Dessa forma, o critério para classificar uma mensagem como *spam* torna-se:

$$\frac{P(c_s).\prod_{k=1}^{n} P(t_k|c_s)^{x_k}}{\sum_{c_i\in\{c_s,c_l\}} P(c_i).\prod_{k=1}^{n} P(t_k|c_i)^{x_k}} > T,$$

e as probabilidades $P(t_k|c_i)$ são calculadas pela estimativa Laplaciana:

$$P(t_k|c_i) = \frac{1 + N_{t_k,c_i}}{n + N_{c_i}},$$

sendo N_{t_k,c_i} correspondente ao número de ocorrências do termo t_k nas mensagens de treinamento da categoria c_i e $N_{c_i} = \sum_{k=i}^{n} N_{t_k,c_i}$.

NB Multinomial Booleano

O classificador NB multinomial Booleano (NB MN Bool) é similar ao multinomial com frequência de termos, incluindo a estimativa de $P(t_k|c_i)$, exceto pelo fato de que cada atributo x_k é Booleano. Note que esses métodos não levam em conta a ausência de termos ($x_k = 0$) nas mensagens. Schneider (2004) demonstrou que o classificador NB multinomial Booleano costuma ter desempenho superior ao multinomial com frequência de termos.

Isso deve-se ao fato de que o classificador NB multinomial com frequência de termos é equivalente a versão do classificador NB com atributos modelados seguindo uma distribuição de Poisson em cada categoria, assumindo que o comprimento de cada mensagem é independente da categoria. Dessa forma, o classificador NB multinomial pode obter desempenho melhor usando atributos Booleanos, caso as frequências dos termos não sigam uma distribuição de Poisson.

NB Multivariado Bernoulli

Seja $\mathcal{T}' = \{t_1, \ldots, t_n\}$ o conjunto de termos extraídos das mensagens, o classificador NB multivariado Bernoulli (NB MV Bern) representa cada mensagem m em termos de presença e ausência de cada termo. Dessa forma, m pode ser representada como um vetor binário $\vec{x} = \langle x_1, \ldots, x_n\rangle$, em que cada x_k informa se t_k está presente ou não em m. Além disso, cada mensagem m de categoria c_i é vista como o resultado de n tentativas de Bernoulli, no qual cada tentativa decide se t_k aparece ou não em m. A probabilidade de obter um resultado positivo em uma tentativa k é dada por $P(t_k|c_i)$. Dessa forma, as probabilidades $P(\vec{x}|c_i)$ são calculadas por

$$P(\vec{x}|c_i) = \prod_{k=1}^{n} P(t_k|c_i)^{x_k}.(1 - P(t_k|c_i))^{(1-x_k)}.$$

O critério para classificar uma mensagem como *spam* torna-se:

$$\frac{P(c_s).\prod_{k=1}^{n} P(t_k|c_s)^{x_k}.(1 - P(t_k|c_s))^{(1-x_k)}}{\sum_{c_i\in\{c_s,c_l\}} P(c_i).\prod_{k=1}^{n} P(t_k|c_i)^{x_k}.(1 - P(t_k|c_i))^{(1-x_k)}} > T,$$

e as probabilidades $P(t_k|c_i)$ são calculadas pela estimativa Laplaciana:

$$P(t_k|c_i) = \frac{1 + |Tr_{t_k,c_i}|}{2 + |Tr_{c_i}|},$$

sendo $|Tr_{t_k,c_i}|$ a quantidade de mensagens de treinamento que pertencem a categoria c_i e que possuem o termo t_k e $|Tr_{c_i}|$ o número total de mensagens de treinamento da categoria c_i. Para maiores informações a respeito desse classificador, consulte Losada & Azzopardi (2008).

NB Booleano

Denomina-se NB Booleano o classificador similar ao NB multivariado Bernoulli com a diferença de que ele não leva em consideração a ausência de termos. Portanto, as probabilidades $P(\vec{x}|c_i)$ são calculadas somente por

$$P(\vec{x}|c_i) = \prod_{k=1}^{n} P(t_k|c_i),$$

e o critério para classificar uma mensagem como *spam* torna-se:

$$\frac{P(c_s).\prod_{k=1}^{n} P(t_k|c_s)}{\sum_{c_i \in \{c_s,c_l\}} P(c_i).\prod_{k=1}^{n} P(t_k|c_i)} > T.$$

As probabilidades $P(t_k|c_i)$ são estimadas da mesma forma do classificador NB multivariado Bernoulli.

NB Gauss Multivariado

O classificador NB Gauss multivariado (NB Gauss MV) utiliza atributos com valores reais, assumindo que cada atributo segue uma distribuição Gaussiana $g(x_k; \mu_{k,c_i}, \sigma_{k,c_i})$ para cada categoria c_i, tal que os valores de μ_{k,c_i} e σ_{k,c_i} de cada distribuição são estimados do conjunto de treinamento Tr.

As probabilidades $P(\vec{x}|c_i)$ são calculadas por

$$P(\vec{x}|c_i) = \prod_{k=1}^{n} g(x_k; \mu_{k,c_i}, \sigma_{k,c_i}),$$

e o critério para classificar uma mensagem como *spam* torna-se:

$$\frac{P(c_s).\prod_{k=1}^{n} g(x_k; \mu_{k,c_s}, \sigma_{k,c_s})}{\sum_{c_i \in \{c_s,c_l\}} P(c_i).\prod_{k=1}^{n} g(x_k; \mu_{k,c_i}, \sigma_{k,c_i})} > T.$$

Bayes Flexível

O classificador Bayes flexível é similar ao NB Gauss multivariado. Entretanto, ao invés de usar uma única distribuição normal para cada atributo X_k por categoria c_i, o filtro Bayes flexível

representa as probabilidades $P(\vec{x}|c_i)$ como médias de L_{k,c_i} distribuições normais com diferentes valores para μ_{k,c_i}, mas com o mesmo valor para σ_{k,c_i}:

$$P(x_k|c_i) = \frac{1}{L_{k,c_i}} \sum_{l=1}^{L_{k,c_i}} g(x_k; \mu_{k,c_i,l}, \sigma_{c_i}),$$

sendo que L_{k,c_i} representa a quantidade de valores diferentes que o atributo X_k assume nas mensagens de categoria c_i do conjunto de treinamento Tr. Cada um desses valores é usado como $\mu_{k,c_i,l}$ de uma distribuição normal da categoria c_i. No entanto, todas as distribuições de uma categoria c_i são calculadas com o mesmo $\sigma_{c_i} = \frac{1}{\sqrt{|Tr_{c_i}|}}$.

Assim, a distribuição de cada categoria torna-se cada vez mais "limitada" conforme as mensagens de treinamento de cada categoria são acumuladas. Pela média de muitas distribuições normais, o filtro Bayes flexível consegue aproximar as distribuições verdadeiras dos atributos melhor que o classificador NB Gauss multivariado, quando a suposição de que eles seguem uma distribuição normal é violada. Para obter maiores detalhes sobre esse método de classificação, consulte Androutsopoulos *et al.* (2004) e John & Langley (1995).

A Tabela 3.4 sumariza as sete versões de classificadores anti-*spam* NB apresentadas nesta seção[13].

Tabela 3.4: Diferentes versões de filtros anti-*spam* Naive Bayes.

Classificador NB	$P(\vec{x}\|c_i)$	Complexidade Treinamento	Complexidade Classificação
NB Básico	$\prod_{k=1}^{n} P(t_k\|c_i)$	$O(n.\|Tr\|)$	$O(n)$
NB MN FT	$\prod_{k=1}^{n} P(t_k\|c_i)^{x_k}$	$O(n.\|Tr\|)$	$O(n)$
NB MN Booleano	$\prod_{k=1}^{n} P(t_k\|c_i)^{x_k}$	$O(n.\|Tr\|)$	$O(n)$
NB MV Bernoulli	$\prod_{k=1}^{n} P(t_k\|c_i)^{x_k}.(1 - P(t_k\|c_i))^{(1-x_k)}$	$O(n.\|Tr\|)$	$O(n)$
NB Booleano	$\prod_{k=1}^{n} P(t_k\|c_i)$	$O(n.\|Tr\|)$	$O(n)$
NB Gauss MV	$\prod_{k=1}^{n} g(x_k; \mu_{k,c_i}, \sigma_{k,c_i})$	$O(n.\|Tr\|)$	$O(n)$
Bayes Flexível	$\prod_{k=1}^{n} \frac{1}{L_{k,c_i}} \sum_{l=1}^{L_{k,c_i}} g(x_k; \mu_{k,c_i,l}, \sigma_{c_i})$	$O(n.\|Tr\|)$	$O(n.\|Tr\|)$

[13] As complexidades computacionais estão de acordo com Metsis *et al.* (2006). Em relação ao custo de classificação, a complexidade do Bayes flexível é $O(n.|Tr|)$ porque ele precisa somar L_k distribuições.

3.4.2 Máquinas de Vetores de Suporte

Máquina de vetores de suporte (do inglês, *Support vector machine* – SVM) é uma das técnicas mais recentes empregadas com grande sucesso em categorização de textos e filtragem de *spams* (Cormack, 2008). Basicamente, cada mensagem é representada por um ponto em um espaço p-dimensional. O principal objetivo do método é proporcionar a separação entre os pontos usando um hiperplano $(p-1)$-dimensional. Nesse caso, tal método é chamado de classificador linear. Entretanto, existem infinitos hiperplanos que podem realizar essa separação dos dados. Assim, o SVM procura pelo hiperplano ótimo que apresenta a maior separação (margem) entre as duas classes. Se tal hiperplano existir, ele é chamado hiperplano com margem máxima e o classificador linear é chamado de classificador de margem máxima (Figure 3.2).

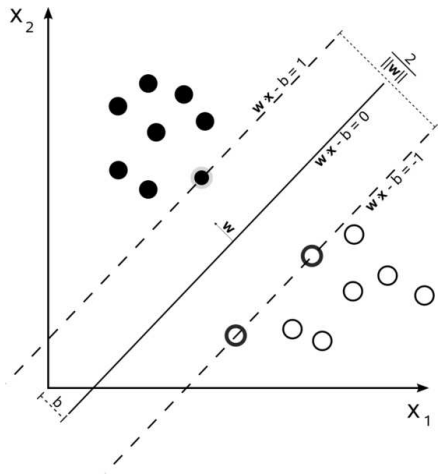


Figura 3.2: Hiperplano com margem máxima e as respectivas margens obtidas para um SVM treinado com amostras de duas classes. Os pontos situados nas margens são chamados de vetores de suporte.

Os SVMs pertencem a uma família de classificadores embasada na teoria de aprendizado estatístico. Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. Uma propriedade especial dos SVMs é que eles simultaneamente minimizam os erros empíricos de classificação e maximizam a margem geométrica. Existem vários estudos na literatura que demonstram que os SVMs lineares são bastante adequados e eficientes na tarefa de classificação de *spams*.

No caso geral de problema de classificação binária, os vetores de entradas possuem p atributos ($\vec{x_i} \in \mathbb{R}^p$) e as saídas são dadas por $c_i \in \{-1, 1\}$. Os atributos podem seguir a representação Booleana, frequência de termos ou TF normalizada.

O caso linear e separável de SVM consiste em achar um hiperplano de dimensão $p-1$, $\mathcal{H}_{p-1}$, dado pela igualdade (Equação 3.2), que separa as entradas de exemplos positivos dos negativos (Vapnik, 1995).

$$\vec{w} \cdot \vec{x} - b = 0 \tag{3.2}$$

Neste caso, $\vec{w}$ é um vetor normal (perpendicular ao hiperplano) e $b \in \mathbb{R}$. Assim, pode-se calcular $\frac{b}{||\vec{w}||}$ para obter a distância do hiperplano em relação à origem. Dessa forma, o método SVM procura por $\vec{w}$ e b que maximizam a margem, ou que a distância entre os hiperplanos paralelos seja a maior possível, enquanto ainda separar corretamente os dados. Estes hiperplanos podem ser descritos por $\vec{w} \cdot \vec{x} - b = 1$ e $\vec{w} \cdot \vec{x} - b = -1$. Caso os dados de treinamento sejam linearmente separáveis, é possível selecionar os dois hiperplanos da margem, de maneira que não haja pontos entre eles e, em seguida, tentar maximizar as suas distâncias (Vapnik, 1995).

Como a distância entre os dois hiperplanos pode ser calculada por $\frac{2}{||\vec{w}||}$, a maximização da margem de separação dos dados em relação a $\vec{w} \cdot \vec{x} - b = 0$ pode ser obtida pela minimização de $||\vec{w}||$. Dessa forma, recorre-se ao seguinte problema de otimização (Equação 3.3) (Vapnik, 1995):

$$\begin{aligned} \underset{\vec{w},b}{\text{Minimizar}} \quad & \frac{1}{2}||w||^2 \\ \text{sujeito a} \quad & c_i(\vec{w} \cdot \vec{x}_i - b) - 1 \geq 0, \forall i = 1, \ldots, n. \end{aligned} \tag{3.3}$$

As restrições são impostas para garantir que não haja dados de treinamento entre as margens de separação das classes. O problema de otimização obtido é quadrático, cuja solução é amparada pela teoria matemática. Como a função objetivo é convexa e todos os pontos que satisfazem as restrições formam um conjunto convexo, esse problema possui um único mínimo global. Problemas dessa natureza podem ser solucionados com a introdução de uma função Lagrangiana, que engloba as restrições à função objetivo, associadas a parâmetros denominados multiplicadores de Lagrange α_i (Equação 3.4) (Vapnik, 1995):

$$\begin{aligned} \underset{\vec{\alpha}}{\text{Maximizar}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j c_i c_j (\vec{x}_i \cdot \vec{x}_j) \\ \text{sujeito a} \quad & \alpha_i \geq 0, \forall i = 1, \ldots, n \\ & \sum_{i=1}^{n} \alpha_i c_i = 0. \end{aligned} \tag{3.4}$$

Essa formulação é chamada forma dual, enquanto o problema original é denominado forma primal. Note que o problema dual é formulado utilizando apenas os dados de treinamento e os seus rótulos.

Seja $\vec{\alpha}^*$ a solução do problema dual, os dados que possuem $\alpha_i^* \geq 0$ são denominados vetores de suporte e podem ser considerados os dados mais informativos do conjunto de treinamento, pois somente eles participam na determinação da equação do hiperplano separador.

Em situações reais, é difícil encontrar aplicações cujos dados sejam linearmente separáveis. Isso se deve a diversos fatores, entre eles a presença de ruídos e *outliers* nos dados ou à própria natureza do problema, que pode ser não linear. Para lidar com problemas dessa natureza, foram propostos SVMs com margens suaves que permitem que alguns dados possam violar as restrições da Equação 3.3. Isso é feito com a introdução de variáveis de folga $\xi_i, \forall i = 1, \ldots, n$. Essas variáveis relaxam as restrições impostas ao problema de otimização primal, que

se tornam (Forman *et al.*, 2009)

$$c_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i = 1, \ldots, n.$$

A aplicação desse procedimento suaviza as margens do classificador linear, permitindo que alguns dados permaneçam entre os hiperplanos e também a ocorrência de alguns erros de classificação.

Os SVMs lineares são eficazes na classificação de conjuntos de dados linearmente separáveis ou que possuam uma distribuição aproximadamente linear, sendo que a versão de margens suaves tolera a presença de alguns ruídos e *outliers*. Porém, há muitos casos em que não é possível dividir satisfatoriamente os dados de treinamento por um hiperplano.

As SVMs lidam com problemas não lineares mapeando o conjunto de treinamento de seu espaço original, referenciado como de entradas, para um novo espaço de maior dimensão, denominado espaço de características. Seja $\Phi : X \rightarrow \Im$ um mapeamento, em que X é o espaço de entradas e denota o espaço de características, um *kernel* K é uma função que recebe dois pontos $\vec{x}_i$ e $\vec{x}_j$ do espaço de entradas e calcula o produto escalar desses dados no espaço de características (Forman *et al.*, 2009):

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j).$$

Para realizar o mapeamento, aplica-se Φ aos exemplos presentes no problema de otimização representado na Equação 3.4, conforme ilustrado a seguir (Forman *et al.*, 2009):

$$\underset{\vec{\alpha}}{\text{Maximizar}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j c_i c_j (\Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j)) \tag{3.5}$$

Para garantir a convexidade do problema de otimização formulado na Equação 3.5 e também que o *kernel* represente mapeamentos nos quais seja possível o cálculo de produtos escalares, utiliza-se funções que seguem as condições estabelecidas pelo teorema de Mercer. Alguns dos *kernels* mais utilizados na prática são os Polinomiais, os Gaussianos ou RBF (*Radial-Basis Function*) e os Sigmoidais (Forman *et al.*, 2009).

Maiores detalhes sobre a teoria geral dos SVMs podem ser encontrados no trabalho original apresentado em Vapnik (1995) e as técnicas mais recentes divulgadas em Forman *et al.* (2009). Entretanto, mais informações sobre o desempenho dos SVMs na classificação de *spams* podem ser encontrados em Drucker *et al.* (1999); Kolcz & Alspector (2001); Hidalgo (2002); Sculley *et al.* (2006); Sculley & Wachman (2007); Cormack (2008) e Liu & Cui (2009).

3.4.3 Filtros Naives Bayes *vs.* Support Vector Machines

Após apresentar sete modelos diferentes de filtros anti-*spam* Naive Bayes e introduzir os SVMs lineares, faz-se necessário comparar o desempenho de cada um deles, verificar qual possui

maior capacidade de predição para classificação automática de *spams* e citar as principais vantagens e desvantagens encontradas. Nesse ponto, vale destacar que a comparação de desempenho foi realizada utilizando como único critério a capacidade de predição dos métodos, uma vez que existem outros trabalhos, como Androutsopoulos *et al.* (2004); Cormack (2008) e Zhang *et al.* (2004), que oferecem análises de desempenho por tempo e complexidade computacional. Em resumo, tais trabalhos indicam que tanto os filtros Bayesianos quanto os SVMs são equivalentes e muito eficientes na etapa de classificação de novas mensagens. Entretanto, é sabido que os SVMs são muito mais lentos que os filtros NB nos processos de treinamento e de aprendizado de novas mensagens, pois ele necessita resolver problemas de otimização quadrática para cada nova mensagem a ser inserida na base de dados.

A seguir, é apresentada a metodologia empregada nessa fase experimental, bem como os principais resultados obtidos.

Metodologia

Esta seção apresenta o protocolo experimental utilizado para a avaliação de desempenho das sete diferentes versões de filtros anti-*spam* Bayesiano e do SVM linear.

Os experimentos realizados nessa seção foram conduzidos nas seis bases de dados Enron (Metsis *et al.*, 2006). A grande variedade de tópicos (assuntos), a alta dimensionalidade da base e o fato de as mensagens nelas contidas serem exatamente iguais às originais foram as principais características que motivaram a escolha dessa base. Diversos detalhes estatísticos sobre base Enron podem ser encontrados em Shetty & Adibi (2004). Com o intuito de simular uma caixa de e-mails real, as mensagens foram ordenadas de acordo com a data de recebimento.

O treinamento dos filtros foi realizado utilizando 10% das mensagens mais antigas recebidas por cada um dos seis usuários da Enron. A seguir, o conjunto de teste (90% das mensagens restantes) foi classificado por cada uma das sete diferentes versões do filtro anti-*spam* Naive Bayes (NB) e pelo filtro SVM linear.

Na etapa de classificação, os filtros NB foram configurados com um limiar T igual 0,5 ($\lambda = 1$). Para o filtro SVM, foi adotado o mesmo conjunto de parâmetros e detalhes de implementação apresentados por Sculley & Wachman (2007).

Conforme apresentado na Seção 3.3, apesar de existirem várias medidas de avaliação, optou-se por empregar as taxas de revocação de *spams* (Tvp) e *hams* (Tvn) e o coeficiente de correlação de Matthews (CCM) como medidas mais relevantes para comparar os resultados obtidos pelos filtros. Adicionalmente, são apresentadas as taxas de precisão (Ps e Pl), taxa de acurácia ponderada (TAc_p) e Razão de Custo Total (RCT). Nesse caso, foi utilizado $\lambda = 1$ para calcular os valores TAc_p e RCT.

Resultados

No restante desta seção, considere as seguintes abreviações: NB Básico como Bas, NB Booleano como Bool, NB Booleano MN como Bool MN, NB MN com frequência de termos como

MN FT, NB MV Bernoulli como Bern, NB Gauss MV como Gauss e Bayes flexível como BF.

As Tabelas 3.5, 3.6, 3.7, 3.8, 3.9 e 3.10 apresentam os resultados obtidos para cada um dos classificadores avaliados em cada base Enron. Valores em negrito indicam a maior pontuação obtida em uma dada medida de desempenho. No entanto, considere o valor CCM como mais relevante na avaliação dos filtros testados, pois ele combina as taxas de precisão e revocação em uma única medida, facilitando a análise dos resultados.

Tabela 3.5: Resultados obtidos para diferentes filtros usando a coleção Enron 1.

Medidas	Bas	Bool	MN TF	Bool MN	Bern	Gauss	BF	SVM
$Tvp(\%)$	91,33	**96,00**	82,00	82,67	72,00	78,67	87,33	83,33
$Ps(\%)$	85,09	51,61	75,00	62,00	61,71	**87,41**	86,18	**87,41**
$Tvn(\%)$	93,48	63,32	88,86	79,35	81,79	**95,38**	94,29	95,11
$Pl(\%)$	96,36	**97,49**	92,37	91,82	87,76	91,64	94,81	93,33
$TAc_p(\%)$	**92,86**	72,78	86,87	80,31	78,96	90,54	92,28	91,70
RCT	**4,054**	1,064	2,206	1,471	1,376	3,061	3,750	3,488
CCM	**0,831**	0,540	0,691	0,578	0,516	0,765	0,813	0,796

Tabela 3.6: Resultados obtidos para diferentes filtros usando a coleção Enron 2.

Medidas	Bas	Bool	MN TF	Bool MN	Bern	Gauss	BF	SVM
$Tvp(\%)$	80,00	**95,33**	75,33	74,00	65,33	62,67	68,67	90,67
$Ps(\%)$	97,57	81,25	96,58	**98,23**	81,67	94,95	98,10	90,67
$Tvn(\%)$	99,31	92,45	99,08	**99,54**	94,97	98,86	**99,54**	96,80
$Pl(\%)$	93,53	**98,30**	92,13	91,77	88,87	88,52	90,25	96,80
$TAc_p(\%)$	94,38	93,19	93,02	93,02	87,39	89,61	91,65	**95,23**
RCT	4,545	3,750	3,659	3,659	2,027	2,459	3,061	**5,357**
CCM	0,850	0,836	0,812	0,814	0,652	0,717	0,776	**0,875**

Tabela 3.7: Resultados obtidos para diferentes filtros usando a coleção Enron 3.

Medidas	Bas	Bool	MN TF	Bool MN	Bern	Gauss	BF	SVM
$Tvp(\%)$	57,33	99,33	57,33	62,00	**100,00**	52,67	52,00	91,33
$Ps(\%)$	**100,00**	99,33	**100,00**	**100,00**	84,75	89,77	96,30	96,48
$Tvn(\%)$	**100,00**	99,75	**100,00**	**100,00**	93,28	97,76	99,25	98,76
$Pl(\%)$	86,27	99,75	86,27	87,58	**100,00**	84,70	84,71	96,83
$TAc_p(\%)$	88,41	**99,64**	88,41	89,67	95,11	85,51	86,41	96,74
RCT	2,344	**75,000**	2,344	2,632	5,556	1,875	2,000	8,333
CCM	0,703	**0,991**	0,703	0,737	0,889	0,613	0,644	0,917

Tabela 3.8: Resultados obtidos para diferentes filtros usando a coleção Enron 4.

Medidas	Bas	Bool	MN TF	Bool MN	Bern	Gauss	BF	SVM
$Tvp(\%)$	94,67	98,00	93,78	96,89	98,22	94,44	94,89	**98,89**
$Ps(\%)$	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
$Tvn(\%)$	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**	**100,00**
$Pl(\%)$	86,21	94,34	84,27	91,46	94,94	85,71	86,71	**96,77**
$TAc_p(\%)$	96,00	98,50	95,33	97,67	98,67	95,83	96,17	**99,17**
RCT	18,750	50,000	16,071	32,143	56,250	18,000	19,565	**90,00**
CCM	0,903	0,962	0,889	0,941	0,966	0,900	0,907	**0,978**

Tabela 3.9: Resultados obtidos para diferentes filtros usando a coleção Enron 5.

Medidas	Bas	Bool	MN TF	Bool MN	Bern	Gauss	BF	SVM
$Tvp(\%)$	89,67	87,23	88,86	94,29	**98,10**	86,68	88,86	89,40
$Ps(\%)$	98,80	**100,00**	**100,00**	**100,00**	92,56	96,37	98,79	99,70
$Tvn(\%)$	97,33	**100,00**	**100,00**	**100,00**	80,67	92,00	97,33	99,33
$Pl(\%)$	79,35	76,14	78,53	87,72	**94,53**	73,80	78,07	79,26
$TAc_p(\%)$	91,89	90,93	92,08	**95,95**	93,05	88,22	91,31	92,28
RCT	8,762	7,830	8,976	**17,524**	10,222	6,033	8,178	9,200
CCM	0,825	0,815	0,835	**0,909**	0,828	0,743	0,814	0,837

Tabela 3.10: Resultados obtidos para diferentes filtros usando a coleção Enron 6.

Medidas	Bas	Bool	MN TF	Bool MN	Bern	Gauss	BF	SVM
$Tvp(\%)$	86,00	66,89	76,67	92,89	**96,22**	92,00	89,78	89,78
$Ps(\%)$	98,98	**99,67**	99,42	97,21	92,32	94,95	98,30	95,28
$Tvn(\%)$	97,33	**99,33**	98,67	92,00	76,00	85,33	95,33	86,67
$Pl(\%)$	69,86	50,00	58,50	81,18	**87,02**	78,05	75,66	73,86
$TAc_p(\%)$	88,33	75,00	82,17	**92,67**	91,17	90,33	91,17	90,05
RCT	6,716	3,000	4,206	**10,227**	8,491	7,759	8,491	6,818
CCM	0,757	0,574	0,661	**0,816**	0,757	0,751	0,793	0,727

Em relação aos filtros avaliados, a análise dos resultados obtidos indica que {SVM, NB MN Booleano, NB Básico, NB Booleano} > {Bayes Flexível, NB MV Bernoulli} >> {NB MN com frequência de termos, NB Gauss MV}, sendo que "$>$" significa "oferece melhor desempenho que". Uma análise geral aponta que o classificador SVM linear apresenta o melhor desempenho médio para todas as bases testadas. É importante destacar que o SVM foi o único filtro a obter taxa de acurácia superior a 90% para todas as coleções analisadas. Entretanto, os filtros NB Booleano, NB MN Booleano e NB Básico merecem destaque, pois apresentaram desempenho comparável ao SVM na maioria das instâncias. Note que, uma vez que não houve treinamento durante a etapa de classificação, o desempenho obtido por esses filtros foi bastante satisfatório, em linha com os melhores resultados presentes na literatura.

Adicionalmente, em acordo com os resultados divulgados por Schneider (2003, 2004), esses experimentos comprovaram que os filtros que utilizam atributos Booleanos são mais eficientes que aqueles que empregam frequência de termos, apesar do fato de que a frequência de termos oferece mais informações sobre as mensagens.

Apesar do classificador Bayes flexível não ter apresentado um bom desempenho, Metsis *et al.* (2006) mostrou que tal filtro é menos sensível a variação do parâmetro T que os demais filtros. Isso significa que o Bayes flexível é capaz de obter um bom índice de reconhecimento de *spams*, mesmo em situações nas quais requer-se um alto reconhecimento de mensagens legítimas (altos valores para T). Os demais classificadores tendem a obter melhor reconhecimento de *spams* conforme T diminui.

Finalmente, é importante destacar que a RCT não é realmente uma medida muito informativa. Por exemplo, para a base Enron 4 (Tabela 3.8), os filtros SVM e NB MV Bernoulli obtiveram resultados similares ($SVM_{CCM} = 0,978$ e $Bern_{CCM} = 0,966$), fato que pode ser confirmado pela proximidade das demais taxas Tvp, Tvn e TAc_p . Entretanto, os valores RCT indicam uma grande diferença de desempenho ($SVM_{RCT} = 90,000$ e $Bern_{RCT} = 56,250$).

3.5 Conclusões

Este capítulo apresentou os principais conceitos básicos necessários para a compreensão deste manuscrito e de quaisquer outras pesquisas divulgadas na área de filtragem de *spams*, tais como: representações empregadas pelos principais métodos de classificação, bases de dados utilizadas para testar filtros anti-*spam* e medidas de desempenho empregadas na avaliação e comparação desses filtros. Nesse último item, a medida CCM demonstrou ser bastante adequada para avaliar e comparar resultados obtidos por diferentes filtros anti-*spam*. Vale ressaltar que o CCM oferece uma avaliação mais balanceada do que a RCT, especialmente se as duas classes possuirem tamanhos diferentes. Outro ponto a favor do CCM é o fato dele retornar um valor dentro de um conjunto pré-definido, provendo mais informações sobre o desempenho do classificador.

Adicionalmente, foram apresentados os principais classificadores estudados na literatura e largamente utilizados por *softwares* comerciais e *open-sources*. Nesse quesito, contribuiu-se

detalhando sete modelos diferentes de filtros Bayesianos, fato pouco contemplado na literatura.

Finalmente, foi realizado um estudo comparativo entre os métodos Bayesianos e o classificador linear SVM aplicados na tarefa de filtragem automática de *spams*. Resultados experimentais foram conduzidos utilizando seis bases de dados bastante conhecidas, reais, grandes e públicas. Em resumo, os resultados indicam que os filtros SVM linear, NB Booleano, NB MN Booleano e NB Básico são as escolhas mais adequadas para a realização de tal tarefa. Entretanto, é importante destacar que o classificador SVM obteve o melhor desempenho médio. Ele foi o único método que obteve taxa de acurácia ponderada superior a 90% para todas as bases utilizadas.

Capítulo 4

Métodos Indiretos – Parte II: Redução de Dimensionalidade

Conforme apresentado no capítulo anterior, os classificadores Bayesianos vêm ganhando considerável destaque por serem muito populares em filtros comerciais e abertos (*open-sources*). Diversos servidores modernos de e-mail, como Google, Yahoo, HotMail, dentre outros, passaram a utilizá-lo. Além disso, filtros bastante consagrados, como OSBF-Lua (vencedor dos dois últimos campeonatos de filtros anti-*spams*), DSPAM, SpamAssassin, SpamBayes, Bogofilter, ASSP, dentre muitos outros, utilizam técnicas de classificação Bayesiana.

Apesar de todas as vantagens apresentadas por esses métodos, eles possuem um ponto fraco bastante conhecido: a queda de desempenho conforme a dimensão da base de dados aumenta (Androutsopoulos *et al.*, 2004; Zhang *et al.*, 2004; Cormack, 2008). Nesse sentido, este capítulo apresenta um estudo comparativo de oito técnicas de seleção de termos empregadas em conjunto com as sete variações do filtro anti-*spam* Naive Bayes estudados no Capítulo 3.

Este capítulo visa examinar como os métodos seletores de termos podem ajudar a manter ou até mesmo melhorar o desempenho dos diferentes classificadores baseados na probabilidade Bayesiana. A escolha de tais métodos deve-se ao fato de que os filtros Bayesianos são os mais empregados para realizar a classificação automática de e-mails (Zhang *et al.*, 2004; Cormack, 2008; Guzella & Caminhas, 2009). Dessa forma, a principal contribuição desse estudo é proporcionar melhorias aos filtros que já se encontram em atividade.

O restante deste capítulo está estruturado da seguinte maneira: a Seção 4.1 apresenta detalhes sobre as técnicas mais conhecidas empregadas para seleção de termos. A Seção 4.2 descreve a metodologia empregada nos experimentos realizados. Os resultados obtidos são apresentados na Seção 4.3. Finalmente, a Seção 4.4 oferece conclusões a respeito do estudo realizado.

4.1 Redução de Dimensionalidade

Ao contrário do que ocorre em recuperação de textos, em categorização de texto, a alta dimensionalidade do espaço de termos ($\mathcal{T}$) pode ser problemática. Os algoritmos de recuperação

de texto podem manipular altos valores de $|\mathcal{T}|$, o que não ocorre com a maioria dos métodos de aprendizado empregados para classificação. Consequentemente, antes da etapa de classificação, geralmente, é aplicado um passo de redução de dimensionalidade, cujo efeito é reduzir o tamanho do vetor de espaço de $|\mathcal{T}|$ para $|\mathcal{T}'| \ll |\mathcal{T}|$. O conjunto $\mathcal{T}'$ é chamado de conjunto reduzido de termos (Sebastiani, 2002).

A redução da dimensionalidade também é benéfica uma vez que ela tende a reduzir o super-ajustamento (*overfitting*). Classificadores que super-ajustam os dados são bons na reclassificação dos dados usados no treinamento, porém eles tendem a classificar incorretamente dados que ainda não foram vistos. Além disso, muitos classificadores apresentam baixo desempenho quando manipulam uma grande quantidade de atributos. Dessa forma, é recomendável um procedimento para reduzir o número de termos utilizados para representar as mensagens (Forman, 2003; Forman & Kirshenbaum, 2008).

Métodos empregados para reduzir a dimensionalidade do espaço de termos podem ser separados em duas classes distintas dependendo se tarefa é realizada localmente (por exemplo, para cada categoria individualmente) ou globalmente. Além disso, tais técnicas são compostas por seletores de termos ($\mathcal{T}' \subset \mathcal{T}$) ou extratores de termos (os termos de $\mathcal{T}'$ não são iguais aos termos de $\mathcal{T}$ e são obtidos por combinações ou transformações dos termos originais). O motivo de utilizar termos sintéticos é evitar problemas com polissemia, homônimos e sinônimos. Devido a essas características, técnicas de seleção de termos são geralmente aplicadas para reduzir a dimensionalidade do espaço de termos. De acordo com Yang & Pedersen (1997), algumas dessas técnicas podem reduzir a dimensionalidade em até 100 vezes sem causar perda (ou até mesmo com um pequeno aumento) de eficácia.

A seguir, são apresentados os oito métodos mais utilizados para Redução do Espaço de Termos (RET) presentes na literatura. Probabilidades são interpretadas como eventos no espaço das mensagens (por exemplo, $P(\bar{t}_k, c_i)$ corresponde a probabilidade de, para uma mensagem aleatória m, o termo t_k não ocorrer em m e m pertencer a classe c_i) e são estimadas pela contagem das ocorrências no conjunto de treinamento Tr.

Como existem apenas duas categorias na filtragem de *spams* ($c = \{\texttt{spam}(\texttt{c}_\texttt{s}), \texttt{legitimo}(\texttt{c}_\texttt{l})\}$), algumas funções são especificamente "locais" em relação a uma dada categoria c_i. Para calcular o valor de um termo t_k num senso "global" e independente de categoria, tanto a soma $f_{sum}(t_k) = \sum_{i=1}^{|c|} f(t_k, c_i)$, a soma ponderada $f_{wsum}(t_k) = \sum_{i=1}^{|c|} P(c_i).f(t_k, c_i)$ ou o máximo $f_{max}(t_k) = max_{i=1}^{|c|} f(t_k, c_i)$ dos valores relativos a cada categoria específica $f(t_k, c_i)$ são comumente calculados. Essas funções tentam capturar a intuição de que os melhores termos para c_i são aqueles melhor distribuídos nos conjuntos de amostras positivas e negativas de c_i.

4.1.1 Frequência de Documentos (FD)

Trata-se de uma função simples, global e eficiente para RET. Ela é obtida pela frequência de mensagens com o termo t_k na base de treinamento, ou seja, somente os termos que aparecem

em maior número de mensagens são selecionados. Nesse caso, supõem-se que os termos que aparecem raramente não são informativos para a predição da categoria e nem influenciam no desempenho global. Dessa forma, a remoção dos termos raros reduz a dimensionalidade do espaço de termos e, possivelmente, melhora o desempenho do classificador, principalmente, quando a maioria desses termos são ruídos inseridos pelos *spammers*. O valor FD de um termo t_k é calculado por

$$FD(t_k) = \frac{|Tr_{t_k}|}{|Tr|},$$

sendo $|Tr_{t_k}|$ a quantidade de mensagens que contém o termo t_k na base de treinamento Tr e $|Tr|$ a quantidade total de mensagens processadas (Yang & Pedersen, 1997).

4.1.2 Fator de Associação DIA (DIA)

O fator DIA de um termo t_k em uma classe c_i mede a probabilidade de encontrar mensagens da classe c_i dado o termo t_k. Essas probabilidades são obtidas pela frequência dos termos na base de treinamento Tr (Fuhr & Buckley, 1991; John *et al.*, 1994),

$$DIA(t_k, c_i) = P(c_i|t_k).$$

As pontuações específicas de cada classe podem ser combinadas usando as funções f_{sum} ou f_{max} para calcular a representatividade do termo em escala global.

4.1.3 Ganho de Informação (GI)

GI é frequentemente empregado em aprendizagem de máquina como critério de representatividade de termos (Mitchell, 1997). Ele calcula o número de *bits* de informação obtido pela predição de categoria através do conhecimento da presença ou ausência de um termo em uma mensagem (Yang & Pedersen, 1997). O GI de um termo t_k é calculado por

$$GI(t_k) = \sum_{c \in [c_i, \bar{c}_i]} \sum_{t \in [t_k, \bar{t}_k]} P(t, c).log \frac{P(t, c)}{P(t).P(c)}.$$

4.1.4 Informação Mútua (IM)

IM é um critério comumente utilizado em linguagem estatística para modelar associações entre palavras e aplicações relacionadas (Yang & Pedersen, 1997). A informação mútua entre t_k e c_i é definida por

$$IM(t_k, c_i) = log \frac{P(t_k, c_i)}{P(t_k).P(c_i)}.$$

$IM(t_k, c_i)$ tem um valor natural igual a zero se t_k e c_i são independentes. Para medir a representatividade de um termo em um senso global, as pontuações específicas de cada categoria podem ser combinadas usando as funções f_{sum}, f_{wsum} ou f_{max}, como apresentado anteriormente.

Em alguns trabalhos, GI também é chamado de IM, causando uma certa confusão. Provavelmente, isso deve-se ao fato de GI ser a média ponderada de $IM(t_k, c_i)$ e $IM(\bar{t}_k, c_i)$, na qual os pesos são dados pelas probabilidades conjuntas $P(t_k, c_i)$ e $P(\bar{t}_k, c_i)$, respectivamente. Consequentemente, GI é também conhecido como IM média (Sebastiani, 2002).

Existem duas diferenças fundamentais entre GI e IM: primeiro, GI usa informação relativa à ausência de termos, enquanto que IM a ignora e, segundo, GI normaliza a pontuação IM usando as probabilidades conjuntas, enquanto que IM utiliza uma pontuação não normalizada.

4.1.5 Estatística χ^2

A estatística χ^2 calcula a falta de independência entre o termo t_k e a classe c_i. Ela tem valor natural igual a zero se t_k e c_i são independentes. Pode-se calcular a estatística χ^2 de um termo t_k em uma classe c_i por

$$\chi^2(t_k) = \frac{|Tr|.[P(t_k, c_i).P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i).P(\bar{t}_k, c_i)]^2}{P(t_k).P(\bar{t}_k).P(c_i).P(\bar{c}_i)}.$$

4.1.6 Pontuação de Relevância (PR)

PR mede a relação entre a presença do termo t_k na classe c_i e a ausência do mesmo na classe oposta $\bar{c}_i$ (Kira & Rendell, 1992),

$$PR(t_k, c_i) = log\frac{P(t_k, c_i) + d}{P(\bar{t}_k, \bar{c}_i) + d},$$

sendo d um fator constante de amortecimento.

As funções f_{sum}, f_{wsum} ou f_{max} podem ser utilizadas para combinar as pontuações específicas de cada categoria.

4.1.7 Razão de Chances (RC)

A RC foi proposta por Van Rijsbergen (1979) para selecionar termos por realimentação de relevância. Trata-se de uma medida particularmente importante em estatística Bayesiana e regressão logística. Ela calcula a razão entre as chances do termo aparecer em uma mensagem relevante em relação as chances dele aparacer em uma mensagem não relevante. Em outras palavras, a RC é capaz de encontrar termos comumente presentes em mensagens que pertencem a uma certa classe (Cunningham *et al.*, 2003). A razão de chances entre t_k e c_i é dada por

$$RC(t_k, c_i) = \frac{P(t_k, c_i).(1 - P(t_k, \bar{c}_i))}{(1 - P(t_k, c_i)).P(t_k, \bar{c}_i)}.$$

Um RC igual a 1 indica que o termo t_k é representativo em ambas as classes c_i e $\bar{c}_i$. RC maior que 1 indica que t_k é mais representativo na classe c_i. Por outro lado, RC menor que 1 indica que t_k é menos representativo na classe c_i. Entretanto, o valor de RC sempre deve ser maior ou igual a zero. Quanto mais as chances de c_i aproximarem-se de zero, o valor de RC

também será próximo de zero. Quando as chances de $\bar{c}_i$ aproximarem-se de zero, o valor de RC tenderá ao infinito.

Assim como em IM e PR, para calcular a representatividade de um termo em um senso global, podemos utilizar as funções f_{sum}, f_{wsum} ou f_{max}.

4.1.8 Coeficiente GSS (GSS)

O coeficiente GSS pode ser visto como uma variação simplificada da estatística χ^2, definido por (Galavotti *et al.*, 2000):

$$GSS(t_k) = P(t_k, c_i).P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i).P(\bar{t}_k, c_i).$$

Valores positivos correspondem a termos com indícios de pertinência, enquanto que valores negativos representam indícios de não pertinência. Isso significa que quanto maior (menor) for o valor positivo (negativo), maior será o indício de pertinência (não pertinência) de t_k na classe c_i.

Para melhor conveniência, as equações matemáticas de todas as medidas apresentadas nesta seção estão sintetizadas na Tabela 4.1[1].

Tabela 4.1: Técnicas mais populares usadas para seleção de termos.

Técnica	Denotação	Equação
Frequência de documento	$FD(t_k)$	$\frac{\lvert Tr_{t_k}\rvert}{\lvert Tr\rvert}$
Fator de associação DIA	$DIA(t_k, c_i)$	$P(c_i\lvert t_k)$
Ganho de informação	$GI(t_k)$	$\sum_{c\in[c_i,\bar{c}_i]}\sum_{t\in[t_k,\bar{t}_k]} P(t,c).log\frac{P(t,c)}{P(t).P(c)}$
Informação mútua	$IM(t_k, c_i)$	$log\frac{P(t_k,c_i)}{P(t_k).P(c_i)}$
Estatística χ^2	$\chi^2(t_k)$	$\frac{\lvert Tr\rvert.[P(t_k,c_i).P(\bar{t}_k,\bar{c}_i)-P(t_k,\bar{c}_i).P(\bar{t}_k,c_i)]^2}{P(t_k).P(\bar{t}_k).P(c_i).P(\bar{c}_i)}$
Pontuação de relevância	$PR(t_k, c_i)$	$log\frac{P(t_k,c_i)+d}{P(\bar{t}_k,\bar{c}_i)+d}$
Razão de chances	$RC(t_k, c_i)$	$\frac{P(t_k,c_i).(1-P(t_k,\bar{c}_i))}{(1-P(t_k,c_i)).P(t_k,\bar{c}_i)}$
Coeficiente GSS	$GSS(t_k)$	$P(t_k, c_i).P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i).P(\bar{t}_k, c_i)$

[1]A Tabela 4.1 apresenta todas as funções em termos de probabilidade subjetiva. Em alguns casos, como $\chi^2(t_k)$, isso é um pouco artificial, uma vez que essa função não é vista habitualmente em termos de probabilidade. As equações referem-se a forma "local" (específica por categoria) das funções.

4.2 Metodologia

Nesta seção, é apresentado o protocolo experimental utilizado para a avaliação dos diferentes métodos de seleção de termos introduzidos na Seção 4.1. Eles foram aplicados para reduzir a dimensionalidade do espaço de termos antes da etapa de classificação realizada pelos classificadores anti-*spam* Naive Bayes apresentados no Capítulo 3.

Assim como nos experimentos discutidos no Capítulo 3, em todos os testes apresentados nesta seção, foram utilizadas as seis bases de dados Enron. A grande variedade de mensagens aliada à alta dimensionalidade da base foram as principais características que motivaram a sua escolha.

Com o intuito de simular uma caixa de e-mails real, as mensagens foram ordenadas de acordo com a data de recebimento. Para realizar uma redução de termos agressiva (significativo número de atributos), foram utilizadas 90% das mensagens mais antigas como fonte de treinamento (conjunto de treinamento). As demais foram separadas para etapa de classificação (conjunto de teste).

Após coletar os atributos do conjunto de treinamento, todos os termos irrelevantes foram eliminados. Nesse caso, foram considerados irrelevantes os termos que apareceram em menos de cinco mensagens durante o processo de treinamento. A remoção dos termos irrelevantes é uma estratégia comumente aplicada em categorização de textos para reduzir o espaço de termos sem afetar a capacidade de predição do classificador (Sebastiani, 2002).

Uma vez que a etapa de treinamento dos filtros é concluída, as técnicas de seleção de termos são aplicadas para reduzir a dimensionalidade do espaço de dados. Todos os métodos apresentados na Seção 4.1[2] foram analisados.

Para produzir uma avaliação completa, variou-se o número de termos selecionados pelos métodos seletores de 10% a 100% de todos os atributos obtidos na etapa de treinamento. A seguir, o conjunto de teste foi classificado por cada uma das sete diferentes versões do filtro anti-*spam* Bayesiano.

Foram realizadas todas as combinações possíveis entre os filtros anti-*spam* Bayesianos, os métodos de seleção de termos e a variação do número de termos selecionados.

Na etapa de classificação foi utilizado um limiar T igual 0,5 ($\lambda = 1$) como sugerido por Metsis *et al.* (2006) e previamente comentado no Capítulo 3. Novamente, apesar das diversas medidas de desempenho existentes, optamos por empregar o coeficiente de correlação de Matthews (CCM) para comparar os resultados obtidos pelos filtros.

4.3 Resultados Experimentais

A seguir, são apresentados os resultados experimentais obtidos em cada base de dados Enron. No restante desta seção, considere as seguintes abreviações: NB Básico como Bas, NB Booleano

[2]Para a técnica PR, foi utilizado um fator de amortecimento $d = 0,1$.

como Bool, NB Booleano MN como Bool MN, NB MN com frequência de termos como NB MN FT, NB MV Bernoulli como Bern, NB Gauss MV como Gauss e Bayes flexível como BF.

Devido ao elevado número de resultados obtidos, optou-se por apresentar os melhores resultados e as melhores predições médias obtidas por cada um dos classificadores NB usando diferentes técnicas de seleção de termos (TST) e variando o número de atributos selecionados[3]. Considere como "melhor resultado" a combinação que obteve o CCM mais alto.

As Tabelas 4.2, 4.5, 4.8, 4.11, 4.14, e 4.17 apresentam os cinco maiores CCM obtidos para cada base Enron. Nas Tabelas 4.3, 4.6, 4.9, 4.12, 4.15, e 4.18 é apresentado o melhor desempenho obtido para cada classificador. Nas Tabelas 4.4, 4.7, 4.10, 4.13, 4.16, e 4.19 é detalhado o conjunto completo de resultados obtido pelos dois classificadores que atingiram os melhores desempenhos.

A Figura 4.1 apresenta os TSTs que obtiveram as melhores predições médias para cada classificador NB variando o número de termos selecionados na etapa de treinamento.

Tabela 4.2: Cinco melhores desempenhos obtidos para a coleção Enron 1.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB MV Bernoulli	DIA_{max}	50	0,885
NB MV Bernoulli	DIA_{max}	60	0,885
NB MV Bernoulli	RC_{wsum}	40	0,872
NB Básico	PR_{wsum}	80	0,872
NB MV Bernoulli	DIA_{max}	40	0,871

Tabela 4.3: Melhor CCM obtido para cada classificador quando aplicado com a base Enron 1.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB MV Bernoulli	DIA_{max}	50	0,885
NB Básico	PR_{wsum}	80	0,872
NB Booleano	DIA_{max}	50	0,867
NB Booleano MN	RC_{wsum}	50	0,861
NB MN FT	RC_{wsum}	50	0,844
NB Gauss MV	GI	70	0,839
Bayes flexível	GI	70	0,833

[3]O conjunto completo de resultados está disponível em `http://www.dt.fee.unicamp.br/~tiago/Research/Spam/spam.htm`.

Tabela 4.4: Dois classificadores que obtiveram os melhores desempenhos para a base Enron 1.

Medida	Bern & DIA_{max}	Bas & RC_{wsum}
$\|\mathcal{T}'\|$(% de $\|\mathcal{T}\|$)	50	80
$Tvp(\%)$	99,33	88,00
$Ps(\%)$	85,14	93,62
$Tvn(\%)$	92,93	97,55
$Pl(\%)$	99,71	95,23
$TAc_p(\%)$	94,79	94,79
RCT	5,556	5,556
CCM	0,885	0,872

Tabela 4.5: Cinco melhores desempenhos obtidos para a coleção Enron 2.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB MV Bernoulli	RC_{max}	40	0,952
NB MV Bernoulli	RC_{sum}	40	0,952
NB MV Bernoulli	RC_{wsum}	50	0,944
NB MV Bernoulli	RC_{max}	50	0,943
NB MV Bernoulli	RC_{sum}	50	0,943

Tabela 4.6: Melhor CCM obtido para cada classificador quando aplicado com a base Enron 2.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB MV Bernoulli	RC_{max}/RC_{sum}	40	0,952
NB Booleano	DIA_{sum}	50	0,915
NB Básico	PR_{max}	30	0,909
NB Gauss MV	χ^2	20	0,896
NB MN FT	RC_{max}/RC_{sum}	40	0,874
NB Booleano MN	RC_{max}/RC_{sum}	40	0,861
Bayes flexível	GI	10	0,855

Tabela 4.7: Dois classificadores que obtiveram os melhores desempenhos para a base Enron 2.

Medida	Bern & RC_{max}/RC_{sum}	Bool & DIA_{sum}
$\|\mathcal{T}'\|$(% of $\|\mathcal{T}\|$)	40	50
$Tvp(\%)$	99,33	88,00
$Ps(\%)$	93,71	99,25
$Tvn(\%)$	97,71	99,77
$Pl(\%)$	99,77	96,04
$TAc_p(\%)$	98,13	96,76
RCT	13,636	7,895
CCM	0,952	0,915

Tabela 4.8: Cinco melhores desempenhos obtidos para a coleção Enron 3.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB Booleano	GI	60	0,991
NB Booleano	GI	30	0,986
NB Booleano	GI	50	0,986
NB Booleano	GI	70	0,986
NB Booleano	GI	20	0,982

Tabela 4.9: Melhor CCM obtido para cada classificador quando aplicado com a base Enron 3.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB Booleano	GI	60	0,991
NB MV Bernoulli	GI	30	0,973
NB Básico	GI	10	0,950
NB Booleano MN	GI	10	0,936
NB Gauss MV	χ^2	10	0,917
NB MN FT	GI	10	0,884
Bayes flexível	IM_{max}	20	0,880

Tabela 4.10: Dois classificadores que obtiveram os melhores desempenhos para a base Enron 3.

Medida	Bool & GI	Bern & GI
$\|\mathcal{T}'\|$(% of $\|\mathcal{T}\|$)	60	30
$Tvp(\%)$	98,67	99,33
$Ps(\%)$	100,00	96,75
$Tvn(\%)$	100,00	98,76
$Pl(\%)$	99,50	99,75
$TAc_p(\%)$	99,64	98,91
RCT	75,000	25,000
CCM	0,991	0,973

Tabela 4.11: Cinco melhores desempenhos obtidos para a coleção Enron 4.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB MV Bernoulli	GI	10	1,000
NB Booleano	DIA_{max}	40	1,000
NB Booleano	RC_{max}	40	1,000
NB Booleano	RC_{sum}	40	1,000
NB Booleano	RC_{wsum}	40	1,000

Tabela 4.12: Melhor CCM obtido para cada classificador quando aplicado com a base Enron 4.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB MV Bernoulli	GI	10 / 20	1,000
NB Booleano	DIA_{max}/RC	40	1,000
NB Booleano MN	DIA_{max}/RC	40	0,996
NB MN FT	DIA_{max}/RC	40	0,996
NB Básico	DIA_{sum}	40	0,978
Bayes flexível	DIA_{max}/RC	40	0,974
NB Gauss MV	DIA_{max}/RC	40	0,970

Tabela 4.13: Dois classificadores que obtiveram os melhores desempenhos para a base Enron 4.

Medida	Bern & GI	Bool & DIA_{max}/RC
$\|\mathcal{T}'\|$(% of $\|\mathcal{T}\|$)	10 / 20	40
$Tvp(\%)$	100,00	100,00
$Ps(\%)$	100,00	100,00
$Tvn(\%)$	100,00	100,00
$Pl(\%)$	100,00	100,00
$TAc_p(\%)$	100,00	100,00
RCT	$+\infty$	$+\infty$
CCM	1,000	1,000

Tabela 4.14: Cinco melhores desempenhos obtidos para a coleção Enron 5.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB MV Bernoulli	RC_{max}	50	0,972
NB MV Bernoulli	RC_{sum}	50	0,972
NB Booleano MN	RC_{wsum}	50	0,967
NB Booleano MN	RC_{max}	40	0,963
NB Booleano MN	RC_{sum}	40	0,963

Tabela 4.15: Melhor CCM obtido para cada classificador quando aplicado com a base Enron 5.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB MV Bernoulli	OR_{max}/OR_{sum}	50	0,972
NB Booleano MN	OR_{wsum}	50	0,967
NB Booleano	OR_{max}/OR_{sum}	60	0,955
NB MN FT	OR_{max}/OR_{sum}	50	0,954
Bayes flexível	χ^2	10	0,931
NB Básico	DF	10	0,924
NB Gauss MV	GSS	20	0,895

Tabela 4.16: Dois classificadores que obtiveram os melhores desempenhos para a base Enron 5.

Medida	Bern & RC_{max}/RC_{sum}	Bool MN & RC_{wsum}
$\|\mathcal{T}'\|$(% of $\|\mathcal{T}\|$)	50	50
$Tvp(\%)$	99,46	99,18
$Ps(\%)$	98,92	98,92
$Tvn(\%)$	97,33	97,33
$Pl(\%)$	98,65	97,99
$TAc_p(\%)$	98,84	98,65
RCT	61,333	52,571
CCM	0,972	0,967

Tabela 4.17: Cinco melhores desempenhos obtidos para a coleção Enron 6.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB Booleano	RC_{wsum}	60	0,929
NB Booleano	RC_{max}	50	0,927
NB Booleano	RC_{sum}	50	0,927
NB MV Bernoulli	RC_{max}	50	0,923
NB MV Bernoulli	RC_{sum}	50	0,923

Tabela 4.18: Melhor CCM obtido para cada classificador quando aplicado com a base Enron 6.

Classificador	TST	% de $\|\mathcal{T}\|$	CCM
NB Booleano	RC_{wsum}	60	0,929
NB MV Bernoulli	RC_{max}/RC_{sum}	50	0,923
NB Booleano MN	RC_{max}/RC_{sum}	60	0,897
Bayes flexível	GI	10	0,873
NB Básico	FD	10	0,866
NB MN FT	RC_{max}/RC_{sum}	50	0,829
NB Gauss MV	RC_{max}/RC_{sum}	50	0,819

Tabela 4.19: Dois classificadores que obtiveram os melhores desempenhos para a base Enron 6.

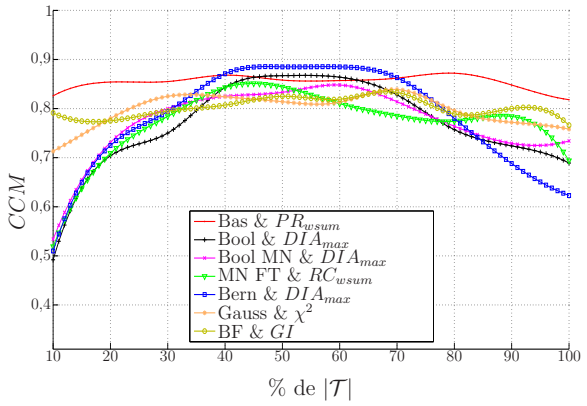
Medida	Bool & RC_{wsum}	Bern & RC_{max}/RC_{sum}
$\|\mathcal{T}'\|$(% of $\|\mathcal{T}\|$)	60	50
$Tvp(\%)$	98,44	96,89
$Ps(\%)$	98,01	99,09
$Tvn(\%)$	94,00	97,33
$Pl(\%)$	95,27	91,25
$TAc_p(\%)$	97,33	97,00
RCT	28,125	25,000
CCM	0,929	0,923

No processo de análise dos resultados pode-se verificar um ponto fraco da medida RCT. O filtro NB MV Bernoulli com GI selecionando 10% de $|\mathcal{T}|$ realizou uma predição perfeita ($|\mathcal{FP}| = |\mathcal{FN}| = 0$), obtendo $CCM = 1,000$ e $RCT = +\infty$ para a base Enron 4. Por outro lado, o classificador NB Booleano MN, com a técnica DIA_{max} usando 40% de $|\mathcal{T}|$, classificou apenas um *spam* incorretamente como legítimo ($|\mathcal{FP}| = 0, |\mathcal{FN}| = 1$), obtendo $CCM = 0,996$ e $RCT = 450$. Se o valor RCT for analisado isoladamente, pode-se concluir erroneamente que a primeira combinação obteve resultado muito superior à segunda.

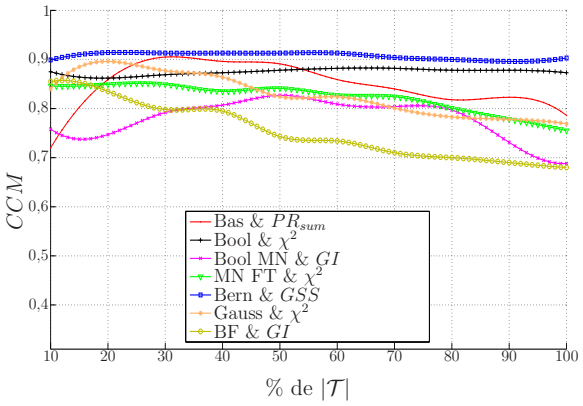
É importante notar que, geralmente, os classificadores pioram o desempenho quando o conjunto completo de termos $|\mathcal{T}|$ é utilizado, com exceção da técnica IM. A Figura 4.2 ilustra como o emprego do método IM afeta o desempenho dos diferentes classificadores variando-se o número de termos selecionados para a base Enron 1. Nota-se também que IM consegue obter melhores resultados quando o conjunto completo de termos é utilizado. Portanto, não empregar IM para reduzir a dimensão do espaço de termos é mais favorável à classificação. Essa mesma característica foi observada em todas as bases utilizadas neste experimento[4]. Entretanto, existe um intervalo entre 30% – 60% de $|\mathcal{T}|$ que oferece melhor desempenho para os demais métodos de seleção. Mesmo um conjunto de termos reduzido, com apenas 10% – 30% de $|\mathcal{T}|$, oferece resultados superiores ao conjunto completo. Dessa forma, empiricamente é possível demonstrar que o emprego de técnicas de seleção de termos na filtragem de *spams*, além de reduzir a dimensionalidade, aumenta a capacidade de predição dos classificadores Naive Bayesianos.

Com relação as TSTs, os resultados indicam que $\{RC, DIA, GI\} > \{PR, \chi^2, GSS, FD\} >> IM$, sendo que "$>$" significa "oferece melhor desempenho que". No entanto, ao levar-se em conta o desempenho médio obtido por todos os classificadores, conclui-se que GI e χ^2 oferecem melhores resultados que RC e DIA, pois eles são menos sensíveis a variação de $|\mathcal{T}'|$. Por outro lado, IM apresentou os piores resultados tanto médios quanto individuais.

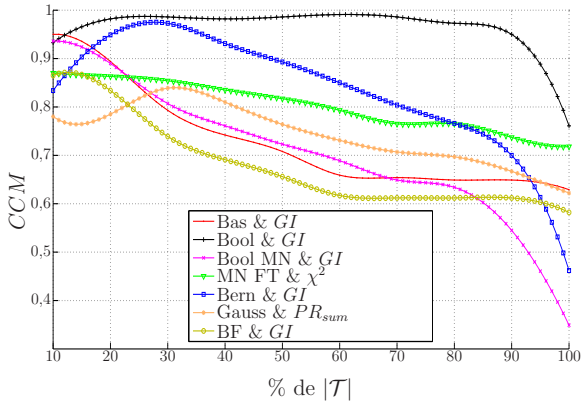
[4]Consulte a página do projeto disponível em `http://www.dt.fee.unicamp.br/~tiago/Research/Spam/spam.htm`



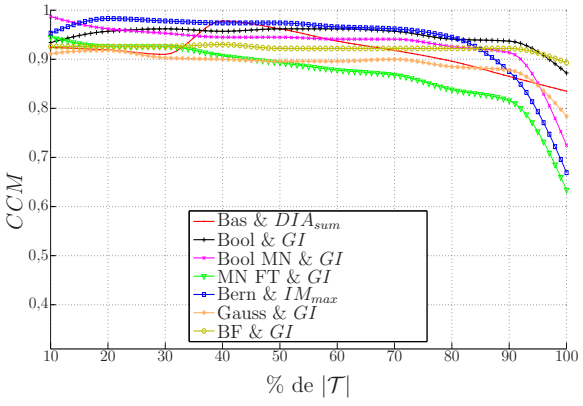
(a) Enron 1



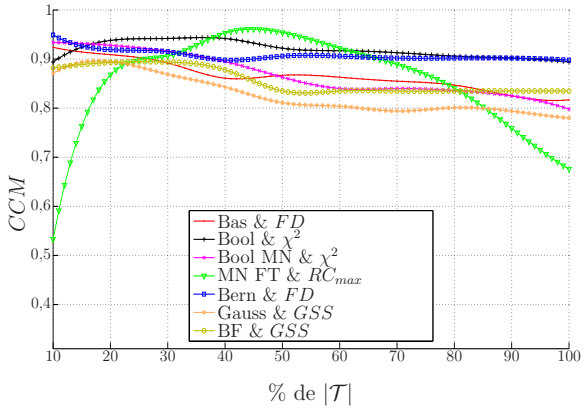
(b) Enron 2



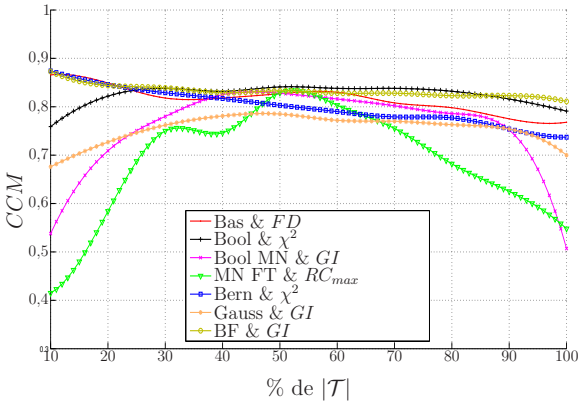
(c) Enron 3



(d) Enron 4



(e) Enron 5



(f) Enron 6

Figura 4.1: TSTs que obtiveram as melhores predições médias para cada classificador NB variando o número de termos selecionados.

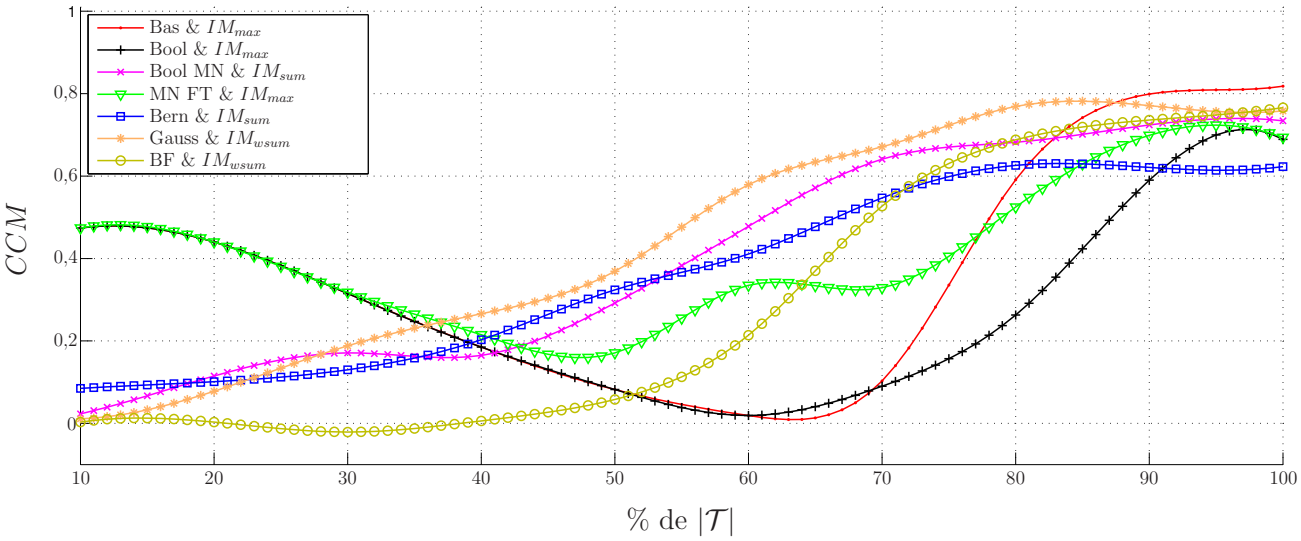


Figura 4.2: TSTs que obtiveram as piores predições médias por classificador NB variando o número de termos para a coleção Enron 1.

Foi também verificado que o desempenho dos classificadores Bayesianos é altamente sensível à qualidade dos termos selecionados pelos TSTs e ao número de termos selecionados $|\mathcal{T}'|$. Por exemplo, para a base Enron 4, o classificador NB MV Bernoulli obteve uma predição perfeita ($CCM = 1,000$) usando 10% de $|\mathcal{T}|$ selecionados por GI, enquanto que o mesmo classificador obteve $CCM = -0,082$ quando empregado com IM_{sum} e com mesma quantidade de termos.

Com respeito ao desempenho dos classificadores, os resultados médios e individuais indicam que os filtros NB MV Bernoulli e NB Booleano são superiores aos demais quando a dimensão do espaço de termos é reduzida adequadamente. Vale a pena lembrar que NB MV Bernoulli é o único método que leva em consideração a ausência de termos nas mensagens. Essa característica oferece mais informações que auxiliam a predição do classificador em casos nos quais os usuários geralmente recebem mensagens com termos específicos, por exemplo, assinaturas. Entretanto, na presença de ruídos (termos inseridos propositalmente pelos *spammers*) esse método tende a apresentar maior perda de desempenho que os demais. Isso explica porque tal método não se destacou nos experimentos discutidos no Capítulo 3.

Novamente, os experimentos comprovaram que os filtros que utilizam atributos Booleanos, mesmo com o uso de técnicas de seleção de termos, ainda são mais eficientes que aqueles que empregam uma representação inteira ou real.

4.4 Conclusões

Neste capítulo foi apresentada uma avaliação experimental de diversos métodos seletores de termos empregados para reduzir a dimensionalidade do espaço de dados na tarefa de filtragem automática de *spams* realizada por classificadores Bayesianos.

Em relação às técnicas empregadas para seleção de termos, os resultados empíricos apontaram RC, fator de associação DIA e GI, respectivamente, como os métodos mais agressivos na realização dessa tarefa. PR, estatística χ^2, coeficiente GSS e FD também ofereceram ganho de desempenho. Por outro lado, o uso do IM obteve resultados ruins e frequentemente piorou a performance dos classificadores. Os resultados também indicam que GI e χ^2 apresentam os melhores desempenhos médios porque são menos sensíveis a variação de $|\mathcal{T}'|$.

Após a etapa de redução de dimensionalidade, dentre todos os classificadores examinados, NB MV Bernoulli e NB Booleano obtiveram os melhores desempenhos médios e individuais para a maioria das bases avaliadas. A análise dos resultados também é conclusiva no sentido de que o emprego de atributos Booleanos é melhor que os baseados em frequência de termos. Também confirmou-se que o desempenho dos classificadores Bayesianos é altamente sensível à qualidade e ao número de termos selecionados na etapa de treinamento.

Capítulo 5

Uma Nova Proposta: O Princípio da Descrição mais Simples

A filtragem automática de *spams* impõe um desafio especial em categorização de textos, no qual a característica mais marcante é que os filtros enfrentam um adversário ativo, que constantemente procura evadir as técnicas de filtragem. Os métodos de classificação de *spams* disponíveis apresentam pelo menos uma das seguintes deficiências:

1. Alta dependência de aprendizado inicial;
2. Alta sensibilidade a ruídos;
3. Queda de desempenho por dimensionalidade; e
4. Alto custo de treinamento e/ou classificação.

Neste capítulo é apresentada a proposta de um novo método de classificação (MDL-CF) baseado no princípio da descrição mais simples auxiliado por fatores de confidência. O principal objetivo dessa proposta era a criação de um método que não tivesse nenhuma das deficiências apresentadas, ou que, pelo menos, minimizasse significativamente cada uma delas.

O princípio da descrição mais simples (do inglês, *Minimum Description Length* – MDL) é um método relativamente recente de inferência indutiva que ainda é muito pouco explorado em tarefas de categorização de textos. Ele parte do pressuposto que a melhor explicação para um conjunto limitado de dados é aquela que permite a sua maior compressão. Os métodos baseados em MDL são particularmente adequados para resolver problemas de seleção de modelos, predição e estimação, em situações nas quais eles podem ser arbitrariamente complexos e a sobreposição dos dados é um grave problema (Rissanen, 1978; Barron *et al.*, 1998; Grünwald, 2005).

Fatores de confidência (do inglês, *Confidence Factors* – CF) são uma técnica utilizada para reduzir o impacto dos ruídos introduzidos pelos termos com baixa incidência e/ou baixa significância (Assis *et al.*, 2006). Trata-se de uma tentativa de imitar o que é feito intuitivamente

ao inspecionar-se uma mensagem e dizer se ela é ou não um *spam*. Em geral, ela é empregada para reduzir a influência da probabilidade local de um determinado atributo, tal que essa redução ocorre de maneira inversamente proporcional ao poder de separação das classes que aquele atributo possui.

Para realizar uma avaliação rigorosa do método proposto, foram simulados três grandes campeonatos de filtros anti-*spam* usando as mesmas funções, bases de dados, medidas de desempenho e tarefas empregadas em cada um deles. Os resultados obtidos mostram que o filtro desenvolvido é superior aos melhores classificadores atualmente disponíveis.

Este capítulo está estruturado da seguinte forma: a Seção 5.1 apresenta detalhes do novo método proposto. A Seção 5.2 descreve a metodologia empregada no processo experimental. Os resultados experimentais são apresentados na Seção 5.3. Finalmente, a Seção 5.4 oferece conclusões e direções para trabalhos futuros.

5.1 MDL-CF: Um Novo Filtro Anti-Spam

Nesta seção são apresentados os conceitos mais importantes relacionados as principais técnicas empregadas no filtro proposto: o princípio da descrição mais simples e os fatores de confidência.

5.1.1 Princípio da Descrição mais Simples

O principal objetivo da modelagem estatística é a descoberta de regularidades nos dados observados. O sucesso em encontrar tais regularidades pode ser medido pela quantidade de informação empregada para descrevê-las. Essa é a base do princípio da descrição mais simples (Rissanen, 1978). A ideia fundamental é que toda a regularidade de um determinado conjunto de dados pode ser usada para comprimí-los.

O princípio da MDL é uma formalização da navalha de Occam, que afirma que a melhor hipótese para um determinado conjunto de dados é aquela que produz representações mais compactas (simples). Assim, de acordo com esse pressuposto, o melhor modelo é aquele que produz a descrição mais simples para o modelo e para os dados. Em outras palavras, o modelo selecionado é aquele que melhor comprime os dados. Esse critério de seleção naturalmente equilibra a complexidade do modelo e do grau com que ele se ajusta aos dados.

Seja $\mathcal{Z}$ um conjunto finito ou contável e P a distribuição de probabilidade de $\mathcal{Z}$. Existe um código prefixo C para $\mathcal{Z}$, tal que, para todo $z \in \mathcal{Z}$, existe um comprimento de código $L_C(z) = \lceil -log_2 P(z) \rceil$. C é chamado de código correspondente a P. Semelhantemente, seja C o código prefixo de $\mathcal{Z}$. Então existe uma distribuição de probabilidade P (possibilidade imperfeita), tal que, para todo $z \in \mathcal{Z}$, $-log_2 P'(z) = L_{C'}(z)$. P' é chamada de distribuição de probabilidade correspondente a C'. Dessa forma, quanto maior a probabilidade de acordo com P implica em um comprimento de código menor de acordo com o código correspondente a P, e

vice-versa (Rissanen, 1978; Barron *et al.*, 1998; Grünwald, 2005).

O objetivo da inferência estatística pode ser moldado como a tentativa de encontrar a regularidade dos dados. Regularidade pode ser definida como a capacidade de compactação. Assim, o princípio da MDL combina essas duas perspectivas e, portanto, dela temos que, para um determinado conjunto de hipóteses $\mathcal{H}$ e um conjunto de dados $\mathcal{D}$, deve-se tentar encontrar a hipótese ou a combinação de hipóteses em $\mathcal{H}$ que melhor comprime $\mathcal{D}$ (Rissanen, 1978; Barron *et al.*, 1998; Grünwald, 2005).

O princípio da MDL pode ser aplicado a todos os tipos de problemas de inferência indutiva, mas acaba por ser mais frutífero em problemas de seleção de modelos. De acordo com Grünwald (2005), uma propriedade importante dos métodos de MDL é que eles fornecem automaticamente e inerentemente proteção contra a sobreposição dos dados (*overfitting*), podendo ser aplicados para estimar os parâmetros e a estrutura de um modelo. Em contrapartida, para evitar sobreposição na estimativa da estrutura de um modelo, os métodos tradicionais, tais como o método da máxima verossimilhança (do inglês, *maximum likelihood*), devem ser alterados e estendidos, tipicamente com princípios *ad hoc*.

Essencialmente, algoritmos de compressão de dados podem ser aplicados na categorização de textos pela construção de um modelo de compressão obtido através do treinamento de mensagens de cada classe e, posteriormente, pelo emprego desses modelos para avaliar a mensagem-alvo.

Seja Tr um conjunto de treinamento composto por mensagens pré-classificadas, a tarefa de filtragem consiste em associar uma mensagem alvo m, com um rótulo desconhecido, a uma das classes $c \in \{spam, ham\}$. Dessa maneira, o método calcula o aumento no comprimento da descrição do conjunto de dados como resultado da adição da mensagem alvo. Finalmente, ele escolhe a classe cujo aumento no comprimento da descrição é mínimo (Almeida *et al.*, 2010a).

Nessa proposta, cada classe (modelo) c é definida como uma sequência de termos (*tokens*) extraídos das mensagens e inseridos no conjunto de treinamento. Cada termo t da mensagem m tem um comprimento de código L_t, baseado na sequência de termos presente nas mensagens de treinamento da classe c. O comprimento de m, quando associada a classe c, corresponde a soma de todos os comprimentos de código associados a cada termo de m, $Lm = \sum_{i=1}^{|m|} L_{t_i}$. Calcula-se $L_{t_i} = \lceil -log_2 P_{t_i} \rceil$, sendo P a distribuição de probabilidade relativa aos termos das classes. Seja $n_c(t_i)$ o número de vezes que t_i aparece em mensagens da classe c, a probabilidade de qualquer termo pertencer a c é dada pela estimativa da máxima verossimilhança:

$$P_{t_i} = \frac{n_c(t_i) + \frac{1}{|\Omega|}}{n_c + 1}$$

na qual, n_c corresponde a soma de $n_c(t_i)$ para todos os termos que aparecem nas mensagens que pertencem a c e $|\Omega|$ corresponde ao tamanho do vocabulário. Neste trabalho, assumiu-se que $|\Omega| = 2^{32}$, ou seja, cada termo em um modo não comprimido corresponderá a um símbolo de 32 bits(Braga & Ladeira, 2008). Essa estimativa reserva uma "porção" de probabilidade para termos que nunca foram apresentados ao classificador (Almeida *et al.*, 2010a).

5.1.2 Fatores de confidência

Os fatores de confidência (*Confidence Factors* – CF) foram inicialmente propostos por Assis *et al.* (2006) com o intuito de reduzir o impacto dos ruídos introduzidos pelos termos com baixa incidência e/ou baixa significância. Para esse propósito, eles tentam imitar o que é feito intuitivamente ao inspecionar-se uma mensagem e dizer se ela é ou não um *spam*. Intuitivamente são considerados apenas poucos símbolos que carregam fortes indícios, de acordo com o aprendizado, e são descartados os termos que aparecem (aproximadamente) de maneira igual em ambas as classes.

O fator de confidência ($0 \leq CF < 1$) de um termo (t_i) é calculado levando-se em conta o peso e as suas frequências máxima e mínima sobre as classes, usando a seguinte expressão (Assis *et al.*, 2006):

$$CF(t_i) = \frac{\left(\frac{(H_{max} - H_{min})^2 + (H_{max} \times H_{min}) - \frac{K_1}{SH}}{SH^2} \right)^{K_2}}{1 + (\frac{K_3}{SH})},$$

sendo que:

- H_{max} corresponde ao número de documentos que possuem o termo t_i e que pertencem a classe com maior probabilidade local;
- H_{min} corresponde ao número de documentos que possuem o termo t_i e que pertencem a classe com menor probabilidade local;
- SH corresponde a soma de H_{max} e H_{min};
- K_1, K_2, K_3 são constantes usadas para melhorar o desempenho do filtro. Elas ajustam a velocidade de decaimento do fator de confidência conforme reduz a diferença das contagens dos termos.

H_{max} e H_{min} são normalizados pelo número máximo de aprendizados das duas classes envolvidas. Além de modular a estimativa do classificador e reduzir o impacto dos ruídos, os fatores de confidência também podem ser usados para restringir o intervalo de probabilidade, evitando a certeza implicada falsamente por uma contagem de zero para uma determinada classe.

5.1.3 Pseudocódigo

Resumidamente, o filtro MDL-CF classifica uma mensagem seguindo as seguintes etapas:

1. *Tokenização*: o classificador extrai todos os termos de uma nova mensagem $m = \{t_1, \ldots, t_{|m|}\}$;
2. Calcula-se o aumento no comprimento da descrição de cada classe quando m é inserida em cada classe $c \in \{spam, ham\}$:

$$L_m(spam) = \sum_{i=1}^{|m|} \left\lceil -log_2 \left(\frac{n_{spam}(t_i) + \frac{1}{|\Omega|}}{n_{spam} + 1} \right) \right\rceil \times \frac{1}{1 - CF(t_i)}$$

$$L_m(ham) = \sum_{i=1}^{|m|} \left\lceil -log_2 \left(\frac{n_{ham}(t_i) + \frac{1}{|\Omega|}}{n_{ham} + 1} \right) \right\rceil \times \frac{1}{1 - CF(t_i)}$$

3. Se $L_m(spam) > L_m(ham)$, então m é classificada como *spam*; caso contrário, m é rotulada como *ham*.

4. Método de treinamento: atualização da frequência dos termos a partir da inserção da nova mensagem.

A seguir, são apresentados mais detalhes sobre os passos 1 e 4 do pseudocódigo.

5.1.4 Pré-processamento e Tokenização

O método proposto não emprega nenhum tipo de técnica de pré-processamento de texto, tais como *stemming* (reduzir a palavra a sua raiz), *stop-words* (remoção de palavras "vazias", como pronomes, preposições, etc) ou *case folding* (todos os caracteres são convertidos para maiúsculas ou minúsculas). Diversos trabalhos publicados na literatura demonstram que o emprego de tais práticas na filtragem de *spams* tende a afetar a acurácia do classificador (Androutsopoulos *et al.*, 2004; Zhang *et al.*, 2004; Metsis *et al.*, 2006). Entretanto, antes da etapa de classificação, o filtro MDL-CF utiliza o método *normalizemime*[1], desenvolvido por Jaakko Hyvattis, para realizar o pré-processamento específico de e-mails. Esse programa converte o conjunto de caracteres para a codificação UTF-8, traduz a codificação de endereços de Internet e adiciona sinais de aviso em caso de erro nessas análises. Ele também acrescenta uma cópia do corpo das mensagens que estão em HTML/XML com a maioria dos rótulos de programação removida, decodifica as entradas HTML e limita o tamanho de arquivos binários anexados.

Tokenização é o primeiro estágio executado dentro do processo de classificação. Ele envolve a quebra da sequência do texto em termos ("*tokens*"), geralmente realizada por meio de uma expressão regular. Nesta etapa, foram considerados todos os termos que iniciam com um caractere imprimível (*printable character*), seguido por qualquer número de caracteres alfanuméricos, exceto pontos, vírgulas e dois pontos. Seguindo-se esse padrão, nomes de domínios e endereços de e-mail são quebrados nos pontos e, assim, o classificador será capaz de reconhecer um domínio mesmo se o seu sub-domínio variar (Siefkes *et al.*, 2004). Assim como proposto por Drucker *et al.* (1999) e Metsis *et al.* (2006), não foram considerados o número de vezes que cada termo aparece em cada mensagem. Nesse sentido, cada *token* é computado apenas uma vez para cada mensagem em que ele aparece.

[1]Disponível em `http://hyvatti.iki.fi/jaakko/spam`

5.1.5 Método de Treinamento

Os filtros de *spams*, assim como muitos outros classificadores, geralmente constrõem os seus modelos de predição por meio do aprendizado com exemplos. Um método de treinamento básico inicia-se com um modelo vazio, classifica cada nova amostra e treina o filtro informando a classe correta quando ele errar a classificação. Esse processo é chamado de treinamento por erro (*Train On Error* – TOE). Uma melhoria desse método consiste no treinamento do filtro quando ele também acertar a classificação e a pontuação (saída) resultante ficar entre limitantes, ou seja, treinar por erro ou perto do erro (*Train On or Near Error* – TONE) (Siefkes *et al.*, 2004).

A vantagem do método TONE sobre o TOE é que ele acelera o processo de aprendizado pela exposição do filtro a exemplos difíceis de classificar. Por essa razão, o treinamento realizado pelo filtro MDL-CF usa o princípio TONE.

Uma grande vantagem do classificador MDL-CF é que ele pode ser inicializado com um conjunto vazio de treinamento e, de acordo com a resposta do usuário, ele vai construindo os modelos de cada classe durante a sua execução. Além disso, não é necessário manter as mensagens usadas no treinamento já que os modelos são incrementalmente construídos pela frequência de aparecimento dos termos. Dessa forma, a complexidade computacional é linear em relação ao tamanho da base.

5.2 Configurações e Metodologia

Conforme visto no Capítulo 3, o problema de classificação de *spams* difere das tarefas de categorização de textos em alguns aspectos, que podem ser brevemente resumidos a seguir:

- O custo gerado por erro de classificação é altamente desbalanceado. Falhas na identificação de mensagens legítimas e *spams* têm consequências materiais diferentes. Classificar incorretamente um e-mail legítimo aumenta o risco de perda da informação contida na mensagem, ou ao menos, poderá causar atraso na leitura. Por outro lado, falhas na identificação de *spams* também variam em importância, mas geralmente são menos prejudiciais que as falhas na identificação de e-mails legítimos (Cormack, 2008).

- Mensagens de e-mail são recebidas em ordem cronológica e devem ser classificadas no momento da entrega. Além disso, em situações reais, é muito comum os filtros entrarem em execução sem receberem dados prévios de treinamento. Outro ponto a ser considerado é que, embora estudos presentes na literatura costumam apresentar experimentos realizando validação cruzada, a adequação desse tipo de validação é questionável nesse cenário (Bratko *et al.*, 2006).

- Diversas informações úteis para o classificador podem ser coletadas nos cabeçalhos, formatos, codificações, padrões de pontuação e estrutura da mensagem. Consequentemente, na etapa de validação, é altamente recomendável utilizar mensagens reais e intactas que

não passaram por nenhum tipo ofuscação ou codificação (Bratko *et al.*, 2006; Cormack, 2008).

Todas essas características refletem o processo de validação adotado para avaliar o desempenho do filtro proposto. A seguir, são especificadas as bases de dados e a metodologia de avaliação empregadas nos experimentos realizados.

5.2.1 Base de Dados e Tarefas

Com o intuito de realizar um processo rigoroso de validação, foram simulados três grandes campeonatos de filtros de *spams* usando os mesmos programas, funções e bases de dados fornecidos pelos organizadores de cada evento[2]. Na Tabela 5.1 são apresentados os dados estatísticos de cada base de e-mail utilizada.

Tabela 5.1: Bases de dados utilizadas na avaliação do filtro MDL-CF.

Base	Nº de *Hams*	Nº de *Spams*	Total	Proporção de *Spams*
TREC05	39.399	52.790	92.189	57,3%
TREC06	12.910	24.912	37.822	65,9%
CEAS08	27.129	110.576	137.705	80,3%
Total	**79.438**	**188.278**	**267.716**	-

Conforme apresentado no Capítulo 3, as bases de dados e metodologias empregadas na *TREC Spam Track* e no *CEAS Spam Challenge* constituem o maior e mais realista "laboratório" para avaliação de filtros anti-*spam* (Cormack, 2008).

A principal tarefa do desafio *TREC Spam Track* é o *immediate feedback* (resposta imediata) que simula o envio em tempo real das mensagens para a caixa de e-mails de um usuário ideal, isto é, aquele que fornece uma resposta imediata ao filtro (Cormack & Lynam, 2005; Cormack, 2006).

Por outro lado, o *CEAS Spam Challenge* 2008, além de alterar a metodologia empregada (ver Capítulo 3), também avaliou o desempenho dos filtros utilizando diferentes tarefas:

- *Live simulation task* (simulação ao vivo) – oferece uma resposta (*feedback*) parcial e atrasada, que é retardada e disponibilizada apenas para mensagens endereçadas à um subconjunto de destinatários. O principal objetivo é simular uma caixa de e-mail real, a

[2]O filtro MDL-CF também foi avaliado usando os mesmos moldes do experimento apresentado no Capítulo 3. Os resultados obtidos, apresentados em Almeida *et al.* (2010a), mostram que o filtro proposto obteve o melhor desempenho quando comparado com os classificadores Bayesianos e o SVM linear.

fim de explorar o impacto causado no filtro quando a resposta do usuário é imperfeita ou limitada.

- *Active learning task* (aprendizagem ativa) – os filtros realizam o processo de classificação sem treinamento prévio. Durante todo o desafio, cada classificador possui uma quota n de consulta, ou seja, ele poderá consultar a classe correta de apenas n mensagens.

5.2.2 Medidas de Desempenho

Conforme detalhado no Capítulo 3, existem diversas medidas de desempenho que podem ser empregadas para avaliar os resultados obtidos pelos filtros anti-*spam*, dentre as quais o Coeficiente de Correlação de Matthews (CCM) demonstrou-se bastante adequado. Entretanto, para manter o padrão dos campeonatos e dar total credibilidade aos resultados obtidos, foram adotadas exatamente as mesmas medidas de avaliação que foram utilizadas por eles: $(1 - AUC)\%$, porcentagem de *spams* bloqueados ($SB\%$ – benefício), porcentagem de *hams* bloqueados ($HB\%$ – custo) e $LAM\%$.

5.2.3 Paramêtros

Nesta seção são detalhados os parâmetros do classificador MDL-CF usados nos experimentos. Todos eles foram empiricamente calibrados para obter os melhores resultados. Na descrição desses parâmetros, o filtro MDL-CF foi dividido em quatro partes: *tokenizador*, fatores de confidência (redução de ruídos), princípio da descrição mais simples (classificador) e treinamento.

- *Tokenizador*:

  1. Foram considerados somente os 3000 primeiros caracteres de cada mensagem, incluindo o cabeçalho;
  2. São preservados somente os *tokens* que possuem tamanho maior ou igual a 1 caractere e menor ou igual a 20 caracteres;

- Fatores de confidência (redução de ruídos):

  1. Constantes: $K1 = 0,25$, $K2 = 10,0$, $K3 = 8,0$.

- Princípio da descrição mais simples (classificador):

  1. A pontuação final (*score*) de cada mensagem é calculada por

$$score = 1,0 - \frac{L_m(spam)}{L_m(ham)},$$

  se $L_m(spam) < L_m(ham)$. Caso contrário,

$$score = -1,0 \times \left(1,0 - \frac{L_m(ham)}{L_m(spam)}\right).$$

Portanto, a pontuação final de cada mensagem é um valor real que varia de -1 (maior probabilidade da mensagem ser *ham*) a $+1$ (maior probabilidade da mensagem ser *spam*).

- Treinamento:

  1. O filtro solicita o método de treinamento em caso de detecção de erro na classificação ou quando alguma mensagem recebe uma pontuação final entre $-0,1$ e $0,15$ (TONE).

5.3 Experimentos e Resultados

Os resultados obtidos pelo filtro MDL-CF foram coletados após um extensivo processo de avaliação, no qual o desempenho do método proposto foi comparado com os resultados atingidos por filtros consolidados e outros divulgados em publicações de grande relevância. A Tabela 5.2 lista todos os sistemas utilizados nos experimentos e comparados com o classificador proposto.

O método MDL-CF foi implementado em linguagem C++ e todas as simulações ocorreram em uma máquina AMD Athlon Dual Core 2.4Ghz com 2Gb de memória RAM, usando o sistema operacional Linux Ubuntu 9.04.

Todos os programas, funções, bases de dados e instruções necessárias para a realização das simulações dos campeonatos estão disponíveis em `http://plg.uwaterloo.ca/~gvcormac/jig/` para os desafios *TREC Spam Track* e `http://plg.uwaterloo.ca/~gvcormac/ceascorpus/` para o campeonato *CEAS Spam Challenge 2008*.

As Tabelas 5.3 e 5.4 apresentam os resultados obtidos na tarefa *immediate feedback* para as bases de dados TREC05 e TREC06, respectivamente. As Tabelas 5.5 e 5.6 apresentam os resultados obtidos pelos filtros para a base CEAS08 nas tarefas *live simulation* e *active learning*, respectivamente. Os seis melhores resultados de cada simulação foram selecionados e ordenados em ordem crescente de valor $(1 - AUC)\%$, uma vez que tal medida é considerada de maior importância nos campeonatos. Outras medidas analisadas são: *Spams* Bloqueados (SB) %, *Hams* Bloqueados (HB) % e $LAM\%$, como descrito anteriormente.

Como pode ser visto nos resultados apresentados, o filtro MDL-CF obteve excelente desempenho nas simulações dos três campeonatos, independente da tarefa realizada. Ao considerar-se que a medida $(1 - AUC)\%$ é a medida padrão e mais importante utilizada nos campeonatos, certamente pode-se concluir que o método proposto foi superior aos melhores filtros da atualidade. Note que, se por um lado, o MDL-CF não conseguiu atingir as melhores taxas de *spams* bloqueados ($SB\%$), por outro, ele obteve ótimas taxas de *hams* bloqueados ($HB\%$) e $LAM\%$ para todas as bases empregadas. Essa característica é muito adequada na tarefa de filtragem de *spams*, já que os falsos positivos são considerados muito mais problemáticos que os falsos

Tabela 5.2: Filtros e resultados reproduzidos para comparação com o MDL-CF.

Filtro	Descrição
Bogofilter	Versão 0.94.0, parâmetros padrões. (`http://www.bogofilter.org`)
PPM	*Prediction by Partial Matching* (Bratko *et al.*, 2006; Cormack & Lynam, 2005).
dbacl	Versão personalizada do filtro de código-aberto dbacl (Breyer, 2005).
CRM-114	Versão 20041231, parâmetros padrões (Cormack & Lynam, 2005). (`http://crm114.sourceforge.net`)
SpamProbe	Versão 1.0a, parâmetros padrões (Cormack & Lynam, 2005). (`http://spamprobe.sourceforge.net`)
SpamAssassin	Versão 3.0.2, parâmetros padrões (Cormack & Lynam, 2005). (`http://spamassassin.apache.org`)
SVM	*Relaxed Linear Support Vector Machine* (Cormack, 2006; Sculley *et al.*, 2006).
OSBF-Lua	*Orthogonal Sparse Bigrams with confidence Factors* (Assis *et al.*, 2006). (`http://osbf-lua.luaforge.net/`)
DMC	*Dynamic Markov Compression* (Bratko *et al.*, 2006).
LR	*Logistic Regression* (LR) com caractere 4-*grams* (Goodman & Yih, 2006).
LR + DMC	Fusão *On-line* do *Dynamic Markov Compression* (DMC) e *Logistic Regression* (Cormack & Lynam, 2007).
IBM Kassel	Filtro anti-*spam* comercial da IBM. (`www-935.ibm.com/services/us/index.wss/offering/iss/a1027071`
Solido Systems	Filtro anti-*spam* comercial da Solido Systems. (`http://solidosystems.com/spamfilter/`)

negativos. Apesar de o filtro Bogofilter ter obtido excelentes taxas $HB\%$, ele obteve as piores taxas $SB\%$, ou seja, o custo é baixo somente em situações em que o benefício também é baixo.

É importante notar que o filtro MDL-CF também foi superior aos filtros comerciais mais consagrados, tais como *Solido Systems Spam Filter* e *IBM Kassel* (*IBM Proventia Network Mail Security System*), tanto nas tarefas *immediate feedback* quanto em *live simulation*. Mesmo o atual campeão, OSBF-Lua, considerado por muitos como o melhor filtro anti-*spam* disponível,

Tabela 5.3: Melhores resultados obtidos na tarefa *immediate feedback* para a coleção TREC05.

Filtros	$(1 - AUC)\%$	SB%	HB%	$LAM\%$
MDL-CF	0,019	96,76	0,05	0,42
PPM	0,020	99,05	0,23	0,50
dbacl	0,037	99,07	0,51	0,69
CRM-114	0,042	99,85	2,56	0,63
Bogofilter	0,048	89,53	0,01	0,30
SpamAssassin	0,059	98,71	0,06	0,57

Tabela 5.4: Melhores resultados obtidos na tarefa *immediate feedback* para a coleção TREC06.

Filtros	$(1 - AUC)\%$	SB%	HB%	$LAM\%$
MDL-CF	0,048	97,58	0,10	0,50
OSBF-Lua	0,054	99,41	0,42	0,50
SVM	0,060	99,51	0,71	0,62
PPM	0,061	97,98	0,11	0,47
Bogofilter	0,084	91,77	0,00	0,00
CRM-114	0,113	99,20	0,27	0,47

Tabela 5.5: Melhores resultados obtidos na tarefa *live simulation* para a base CEAS08.

Filtros	$(1 - AUC)\%$	SB%	HB%	$LAM\%$
MDL-CF	0,0119	96,22	0,01	0,17
OSBF-Lua	0,0127	98,97	0,03	0,19
Solido Systems	0,0131	99,88	0,68	0,28
Bogofilter	0,0157	89,04	0,00	0,00
CRM114	0,0359	99,68	0,75	0,49
LR + DMC	0,0384	99,29	0,25	0,42

vencedor do *TREC Spam Track 2006* e do *CEAS Spam Challenge 2008 – live simulation*, apresentou desempenho inferior ao método proposto neste trabalho.

Outra característica marcante é que o filtro MDL-CF obteve um bom desempenho mesmo quando o seu aprendizado foi realizado com um número bastante limitado de mensagens. Para a base CEAS 2008, na tarefa *active learning* (Tabela 5.6), o método MDL-CF contou com apenas 1000 mensagens enquanto que o número de e-mails da base é próximo de 138.000. Mesmo assim, ele atingiu um resultado expressivo, mostrando-se superior aos demais filtros avaliados.

Tabela 5.6: Melhores resultados obtidos na tarefa *active learning*, com quota igual a 1000, para a coleção CEAS08.

Filtros	$(1 - AUC)\%$	SB%	HB%	$LAM\%$
MDL-CF	0,0071	99,23	0,01	0,11
LR + DMC	0,0129	99,92	0,12	0,23
Online SVM	0,0257	98,65	0,07	0,31
PPM	0,1031	94,28	0,01	0,18
DMC	0,2988	98,11	0,34	0,80
LR	0,3669	96,92	0,31	0,91

5.4 Conclusões

Neste capítulo, foi proposto e apresentado um classificador anti-*spam* baseado no princípio da descrição mais simples e auxiliado por fatores de confidência. Destaca-se que o novo método possui várias características apropriadas para a filtragem automática de *spams*: consome poucos recursos computacionais, é fácil de ser implementado, muito rápido na classificação de novas mensagens e seu treinamento e aprendizado são muito simples.

Para verificar o desempenho do filtro proposto, foram conduzidos experimentos por meio da simulação de três campeonatos renomados. Empiricamente, foi possível demonstrar que o método MDL-CF é bastante eficaz e eficiente na tarefa de classificação de *spams*, superando os resultados obtidos por métodos consagrados e por outros filtros previamente publicados na literatura.

Projetos em andamento envolvem a extensão e o emprego efetivo desse filtro por meio do desenvolvimento de *plug-ins*, a fim de possibilitar uma maneira simples e eficaz de empregar esse classificador em conjunto com os principais gerenciadores de e-mails, tais como Microsoft Outlook e Mozilla Thunderbird e, assim, possibilitar que o fruto desta pesquisa possa ser usufruído por toda a sociedade.

Outro ponto interessante é que o método proposto pode ser estendido e aplicado na solução de problemas de diversas áreas que envolvam classificação e categorização de textos, tais como: *web spamming*, *blog spamming*, *social network spamming*, *mobile spamming*, além de qualquer outro tipo de *spam* disseminado por texto.

Capítulo 6

Sobre a Extinção e Conclusões Finais

Um fato pode ser previsto: o *spam* irá desaparecer um dia! A maior preocupação é que não se sabe se ele irá deixar de existir por causa da avaliação e integração de um plano de combate ao *spamming* ou por causa da extinção do sistema de e-mail causado pelo próprio *spam*.

Apesar dos avanços significativos apresentados pelos métodos indiretos, o volume de *spams* em circulação continua aumentando, pois, além de ser um negócio lucrativo, ele ainda não é considerado ilegal. Para piorar a situação, a inviabilidade na alteração do atual sistema de e-mail faz com que todas as esperanças se concentrem nas propostas e na aplicação dos métodos diretos. Nesse sentido, uma correta avaliação e integração no plano internacional de combate ao *spam* não é somente um auxílio precioso – é uma necessidade de primeira ordem.

Neste trabalho foi apresentado um estudo abrangente sobre o problema do *spamming*, abordando temas que vão desde o seu surgimento até a proposta de um método de classificação automática de *spams*. Pode-se resumir as principais contribuições desta tese em:

1. *Introdução do problema do spam, dados históricos, estatísticas atualizadas e quadro atual do spamming no Brasil e no mundo.* Neste ponto, foi mostrado que o volume de *spams* vem aumentando de forma surpreendente e que o Brasil é um dos países que mais enviam esse tipo de mensagem. Também, foram apresentadas as consequências e os custos causados pelo *spamming*, além das formas mais comuns de operações utilizadas pelos *spammers*.

2. *Estudo relacionado ao combate do spam por meios jurídicos usando recursos legislativos que já se encontram em vigor no Brasil, além do estudo de ferramentas legislativas utilizadas em outros países.* Foi apresentado que o Brasil está atrasado em relação aos seus pares, pois ainda não possui nenhuma legislação no que se refere ao envio de *spams*. Também, foram apresentados e analisados os modelos aplicados atualmente pelos EUA e pela Europa.

3. *Análise das medidas de desempenho utilizadas para avaliar os filtros anti-spam.* Foram apresentadas as informações básicas necessárias para a compreensão deste manuscrito e quaisquer publicações na área de classificação de e-mails. Dentre elas, foram discutidas as

bases de dados de e-mail mais utilizadas na avaliação de filtros de *spams*, a representação comumente adotada pelos métodos de classificação e o estudo referente as medidas de desempenho empregadas na avaliação e comparação de desempenho obtido pelos filtros. Destaca-se a apresentação do coeficiente de correlação de Matthews como boa alternativa neste quesito.

4. *Apresentação e avaliação dos principais classificadores de spams disponíveis na literatura.* Os principais métodos de classificação de *spams* foram apresentados, sendo que sete diferentes variações do filtro anti-*spam* Naive Bayes foram analisados. Os desempenhos obtidos foram comparados com o classificador SVM linear e a análise dos resultados indica que o filtro SVM e os classificadores Bayesianos Booleano multinomial, Básico e Booleano são mais adequados para a filtragem automática de *spams*.

5. *Apresentação e avaliação do emprego de técnicas de redução de dimensionalidade do espaço de dados.* Neste ponto, foram analisadas oito técnicas diferentes empregadas para reduzir a dimensão do espaço de termos com o intuito de melhorar o desempenho dos filtros Bayesianos. Foi demonstrado que tal redução é benéfica para os classificadores avaliados e que as técnicas IG e χ^2 apresentaram o melhor desempenho médio, pois foram menos sensíveis a variação da quantidade de atributos selecionados.

6. *Proposta, apresentação e validação de um novo filtro de spams.* Um classificador baseado no princípio da descrição mais simples auxiliado por fatores de confidência foi proposto. Para realizar uma avaliação rigorosa, foram simulados três campeonatos de classificação de *spams* usando as mesmas funções, bases de dados, medidas de desempenho e tarefas empregadas em cada um deles. Os resultados obtidos mostram que o filtro proposto é superior aos melhores classificadores disponíveis.

Perspectivas

Projetos em andamento envolvem a extensão e o emprego efetivo do filtro MDL-CF por meio do desenvolvimento de *plug-ins*, a fim de possibilitar uma maneira simples e eficaz de empregar o classificador em conjunto com os principais gerenciadores de e-mails, tais como Microsoft Outlook e Mozilla Thunderbird.

Outro destaque, é que o método proposto pode ser estendido e aplicado na solução de problemas de diversas áreas que envolvam classificação e categorização de textos. A seguir, são descritos alguns exemplos que podem ser abordados:

- De acordo com um relatório recentemente divulgado pela empresa de segurança Websense[1], durante a segunda metade de 2009, 95% do conteúdo gerado no espaço para comentários

[1]Disponível em `http://investor.websense.com/releasedetail.cfm?ReleaseID=409218`

em *blogs*, salas de bate-papo e fóruns era *spam* ou continha algum tipo de código malicioso. Nesse sentido, o filtro MDL-CF pode ser estendido para classificar mensagens antes delas serem divulgadas em tais *Web* sites.

- Uma pesquisa realizada recentemente nos Estados Unidos pela empresa de segurança computacional Norton aponta que 44% dos usuários de redes sociais são vítimas de cibercrimes[2]. O relatório indica que os crimes vão desde vírus, fraudes de cartão de crédito, pornografia não solicitada, *spams* e *phishing* (mensagens falsas enviadas por criminosos se passando por bancos e outras empresas de varejo). De maneira similar ao item anterior, o filtro pode ser aplicado na classificação de mensagens postadas em redes sociais.

- Estudos recentes indicam que a atividade de *web spamming* (ações destinadas a induzir os sistemas de busca a errar fazendo com que uma página seja classificada como mais relevante do que ela realmente é) aumentou de 13,8% para 22,1% somente nos últimos 12 meses, causando a degradação dos resultados fornecidos pelos mecanismos de buscas (Gyöngyi & Garcia-Molina, 2005; Webb *et al.*, 2007). A extensão e aplicação do filtro nesse contexto é um desafio científico altamente relevante, pois se a taxa de propagação de *web spams* atingir valores alarmantes como no caso do *spam* por e-mail, os resultados obtidos pelos sistemas de buscas poderão ser fortemente impactados.

- Os telefones celulares estão se tornando o próximo alvo dos *spammers*. Apesar das mensagens de texto com anúncios não serem tão comuns entre os celulares, o problema pode piorar gradativamente, assim como aconteceu com o e-mail. A aplicação do filtro neste escopo poderia antecipar a um problema em eminência.

Trabalhos futuros nesta linha de pesquisa devem levar em conta que a filtragem automática de *spams* é um problema co-evolutivo, pois enquanto os filtros tentam evoluir as suas capacidades de predição, os *spammers* tentam evoluir as suas mensagens para que estas não sejam detectadas pelos classificadores. Dessa forma, uma abordagem eficiente deve possuir uma maneira eficaz de ajustar as regras de classificação para detectar rapidamente as mudanças nas caraterísticas dos *spams*. Nesse sentido, estratégias empregadas em filtros colaborativos poderiam ser usadas para auxiliar o classificador, acelerando a adaptação das regras e aumentando o desempenho do filtro.

Publicações

A seguir, destacam-se os trabalhos submetidos e publicados que apresentam os resultados das pesquisas relatadas neste manuscrito:

[2]Disponível em `http://nortononlineliving.com/documents/NOLR_studyreport031609.pdf`

Revistas

1. T.A. ALMEIDA & A. YAMAKAMI. Facing the Spammers: A Very Effective Approach to Avoid Junk E-mails. *Information Sciences*, Elsevier. Submetido em 2010.

2. T.A. ALMEIDA & A. YAMAKAMI. Occam's Razor Based Spam Filter. *Knowledge and Information Systems*, Springer. Submetido em 2010.

3. T.A. ALMEIDA, J. ALMEIDA & A. YAMAKAMI. Spam Filtering: How the Dimensionality Reduction Impacts on the Accuracy of Naive-Bayes Classifiers? *Journal of Internet Services and Applications*, Springer. Submetido em 2010.

4. T.A. ALMEIDA, J. ALMEIDA & A. YAMAKAMI. A Tutorial on Dimensionality Reduction Approaches in Spam Filtering. *Brazilian Journal of Theoretical and Applied Computation*. Submetido em 2009.

Congressos

1. T.A. ALMEIDA & A. YAMAKAMI. Content-Based Spam Filtering. *Proceedings of the 23rd IEEE International Joint Conference on Neural Networks (IJCNN'10)*, 1–7, Barcelona, Espanha, Julho, 2010.

2. T.A. ALMEIDA, J. ALMEIDA & A. YAMAKAMI. Filtering Spams using the Minimum Description Length Principle. *Proceedings of the 25th ACM Symposium On Applied Computing (SAC'10)*, 1854-1858, Sierre, Suíça, Março, 2010. (Taxa de aceitação = 17%)

3. T.A. ALMEIDA, J. ALMEIDA & A. YAMAKAMI. Probabilistic Anti-Spam Filtering with Dimensionality Reduction. *Proceedings of the 25th ACM Symposium On Applied Computing (SAC'10)*, 1802-1806, Sierre, Suíça, Março, 2010. (Taxa de aceitação = 17%)

4. T.A. ALMEIDA, J. ALMEIDA & A. YAMAKAMI. Evaluation of Approaches for Dimensionality Reduction Applied with Naive Bayes Anti-Spam Filters. *Proceedings of the 8th IEEE International Conference on Machine Learning and Applications (ICMLA'09)*, 517-522, Miami, FL, EUA, Dezembro, 2009.

Filtro bloqueia entrada de spams

Ferramenta desenvolvida na FEEC classifica automaticamente mensagens de e-mail

ISABEL GARDENAL
bel@unicamp.br

Foto: Antoninho Perri

O pesquisador Tiago Agostinho de Almeida: "Não existe no Brasil uma regulamentação jurídica específica com relação ao spam"

O e-mail é um dos meios de comunicação mais popular, mais rápido e mais barato. Ele tornou-se parte do cotidiano de milhares de pessoas, mudando a maneira como trabalhar. Não é somente empregado para dar suporte à comunicação pessoal, mas também é utilizado como agenda, gerenciador de tarefas, organizador de contatos, dentre muitas outras funções. A desvantagem desse enorme sucesso é o volume crescente de spams, mensagens eletrônicas não-solicitadas e que chegam às pessoas constantemente. O problema causado por ele pode ser quantificado em termos econômicos, em razão do desperdício de tempo que ele traz todos os dias aos usuários de computador. Isso sem considerar os inúmeros problemas com fraudes, roubos e danos diretos.

Em 2008, internautas brasileiros enviaram 2,7 trilhões de spams. No primeiro bimestre deste ano, o país passou à primeira posição no ranking mundial, após ter sido responsável por 7,7 trilhões somente em 2009. "Isso é lamentável, porque não dispomos de respaldo jurídico contra esse tipo de fraude. Com isso, a situação tende a se agravar", prevê o pesquisador Tiago Agostinho de Almeida. Tecnólogo em computação, Almeida há pouco defendeu doutorado na Faculdade de Engenharia Elétrica e de Computação (FEEC), abordando a problemática gerada pelos spams, orientado pelo professor do Departamento de Telemática Akebo Yamakami. Idealizou um filtro para classificar automaticamente mensagens de e-mail: o MDL-CF Spam Filter.

O novo filtro, método computacional que classifica os e-mails como spam ou como e-mail legítimo, deriva de duas técnicas: MDL (Minimum Description Length – Princípio da Descrição mais Simples) e CF (Confidence Factors – Fatores de Confidência). O objetivo era oferecer uma classificação balanceada proporcionando uma alta taxa de bloqueio de spams e, simultaneamente, tomando os cuidados necessários para evitar classificação incorreta de um e-mail legítimo. "A nossa técnica mostrou-se simples, eficiente e rápida. Seus resultados indicam que ela é superior aos melhores filtros anti-spam que existem."

Na maioria dos casos, os próprios servidores de e-mail oferecem filtros anti-spam, como o GMail, o Hotmail e o Yahoo. Entretanto, a sua eficácia depende diretamente dos seus usuários. "É preciso saber usar corretamente a ferramenta oferecida pelo gerenciador de e-mails. Se souberem, a eficácia pode chegar a 95%", garante o tecnólogo.

O pesquisador explica que o maior desafio dos filtros é não classificar um e-mail legítimo como spam. Isso é considerado um erro grave, pois a mensagem acaba sendo enviada para a caixa de spams. "Os prejuízos podem ser enormes, pois o usuário pode não tomar conhecimento de uma informação muito importante."

Simulações

O tecnólogo simulou campeonatos de spams com base em modelos existentes. Ele explica que grandes corporações como Google e Microsoft, com frequência, financiam estes eventos para avaliar os filtros anti-spam. No estudo de Almeida, foram simuladas três competições em que concorreram o filtro proposto contra 13 métodos consagrados. "O MDL-CF obteve o melhor desempenho. Em uma situação real, ele teria sido tricampeão", conta.

O resultado foi comemorado pelo autor do projeto. "O nosso filtro teve uma melhor performance em relação aos métodos comparados, mesmo aqueles que partem de grandes corporações, que recebem um alto investimento e que têm uma grande equipe dando-lhes suporte", sinaliza.

O pesquisador pretende oferecer plug-ins para possibilitar uma maneira simples e eficaz de empregar o filtro MDL-CF em conjunto com os principais gerenciadores de e-mail disponíveis, como o *Microsoft Outlook* e o *Mozilla Thunderbird*. "Vamos tentar desenvolver os plug-ins e oferecê-los gratuitamente como uma forma de fazer com que o fruto dessa pesquisa seja usufruído pela sociedade", almeja.

Um único spam pode danificar o equipamento, pelo fato de muitas vezes vir acompanhado de vírus. A dica de Almeida é sempre verificar os e-mails de maneira consciente. "É preciso ter um filtro anti-spam instalado ou ainda usar os recursos anti-spam oferecidos pelo provedor. Além do filtro, existem outras ferramentas que devem ser empregadas para aumentar a segurança dos usuários, como um antivírus e um firewall, um programa que bloqueia acessos", aconselha.

Sistemática

Os spams são enviados pelos spammers, pessoas que geralmente estão interessadas na comercialização de produtos. Existem também aqueles cuja intenção é prejudicar os usuários mediante o roubo de senhas, seja de alguma informação pessoal e até mesmo do uso do seu computador para enviar spams, uma prática que vem aumentando no Brasil.

Uma das consequências dos spams é que eles envolvem um custo elevado, difícil de ser calculado. Relaciona-se inclusive à perda da produtividade de funcionários nas empresas, que gastam o seu tempo lendo e apagando mensagens não solicitadas. O conteúdo é bastante abrangente. Fora os casos de transmissão de vírus, eles permeiam propagandas, pornografia, boatos, esquemas de enriquecimento ilícito e mensagens religiosas.

Existem duas formas mais corriqueiras de se apropriar dos dados pessoais dos usuários. Numa delas, o usuário se cadastra numa empresa idônea, o qual comercializa as informações de seus consumidores para os spammers. Outra é realizada através de um arquivo chamado *cookie*, que armazena as informações trocadas entre o navegador e o servidor.

Ao informar o endereço de e-mail mediante esses *cookies*, o site poderá usar aquela informação e vendê-la. "Existe uma máfia de compra e venda nesta direção", constata Almeida, e é muito difícil chegar à fraude. "O e-mail é uma presa fácil de ser burlada. Quem envia o spam não necessariamente é aquela pessoa. Os endereços que eles enviam em geral são inválidos."

Em alguns países há legislações que criminalizam o envio de spams. Os exemplos mais significativos são os Estados Unidos e países da Comunidade Europeia. Ambos têm meios legais e diferentes de condenar o envio de spams. Nos Estados Unidos, o método adotado permite que o spam seja enviado desde que ele respeite algumas regras. Dentre elas, o endereço de e-mail do remetente deve ser verdadeiro e a mensagem deve conter um mecanismo explícito para que o usuário solicite nunca mais recebê-lo. "O que acontecer fora disso, é considerado ilegal. Se conseguirem pegar o spammer, ele poderá ser multado ou, até mesmo, preso", revela o tecnólogo. Já na Europa, o método adotado não permite que o spammer envie qualquer tipo de mensagem sem o consentimento prévio do usuário. "Se enviar um primeiro spam que seja, ele já estará cometendo uma infração."

No Brasil, a realidade é outra. "Apesar de a situação ser alarmante, não existe uma regulamentação jurídica específica com relação ao spam", relata Almeida, enfatizando a urgência de uma medida que impeça essa prática.

O volume de spams vem crescendo de forma surpreendente no Brasil, diz o tecnólogo, a priori porque o número de usuários está aumentando e muitos utilizam o computador de maneira desprotegida. Portanto, apesar de o país ser um grande emissor de lixo virtual, não significa que ele tenha o maior número de spammers, e sim que os usuários, por falta de conhecimento, acabam sendo alvos de spammers de outros países. O pesquisador também destaca que uma das estratégias mais utilizadas pelos spammers é "sequestrar" computadores de usuários comuns e utilizá-los para enviar mais spams.

Carne enlatada originou termo

Existem diversas versões a respeito da origem da palavra spam. A mais aceita afirma que o termo originou-se da marca SPAM, um tipo de carne enlatada da Hormel Foods Corporation, e foi associado ao envio de mensagens não solicitadas devido a um quadro do grupo humorístico inglês Monty Phython.

O episódio ironiza o racionamento de comida ocorrido na Inglaterra durante a Segunda Guerra Mundial. SPAM foi um dos poucos alimentos excluídos desse racionamento, o que eventualmente levou as pessoas a enjoarem do produto.

Esse quadro envolve um casal discutindo com uma garçonete em um restaurante a respeito da quantidade de SPAM presente nos pratos. Enquanto o casal pergunta por um prato que não contenha a carne enlatada, a garçonete repete constantemente a palavra SPAM. Eventualmente, a discussão faz com que um grupo bizarro de vikings, presente no local, comece a cantar "SPAM, amado SPAM, glorioso SPAM, maravilhoso SPAM!", "impossibilitando qualquer tipo de conversa, assim como o spam atrapalha a comunicação por e-mail", compara Almeida.

Publicação
Tese de Doutorado "SPAM: do surgimento à extinção"
Autor: Tiago Agostinho de Almeida
Orientador: Akebo Yamakami
Unidade: Faculdade de Engenharia Elétrica e de Computação (FEEC)
Financiamento: CNPq

Notoriedade é porta aberta para invasores

As maiores vítimas dos spams são pessoas que possuem endereços de e-mail bastante divulgados, como presidentes de grandes corporações. Alguns exemplos são o dono do domínio acme.com, Jef Poskanzer, e o empresário co-fundador da empresa de tecnologia Apple Inc, Steve Jobs.

De acordo com o presidente-executivo da Microsoft, Steve Ballmer, em matéria veiculada na BBC News, o fundador da Microsoft – Bill Gates – recebe em média 4 milhões de e-mails por ano, sendo a grande maioria spams. Curiosamente, Poskanzer recebia já em 2005 cerca de 1 milhão de spams por dia, quase 100 vezes mais que Bill Gates. "Sem os filtros anti-spam, provavelmente o sistema de e-mail não seria mais suportável", comenta Almeida.

Fato é que os spams são um negócio lucrativo, isso porque o e-mail é uma forma de comunicação muito barata e que atinge muitas pessoas simultaneamente. Desta forma, os spammers conseguem obter lucro mesmo que a taxa de resposta dos usuários seja baixa. Um estudo recente mostrou que mesmo um índice de resposta de 1 para cada 12,5 milhões de e-mails enviados ainda consegue gerar bons lucros aos spammers.

Em 2002, o mundo todo recebia em média 860 milhões de spams por dia. Para 2010, a Cisco Systems, empresa de segurança computacional, estimou que, em média, cerca de 300 bilhões de spams serão enviados diariamente, representando mais de 90% dos e-mails em circulação.

Durante o período de doutoramento, os seguintes trabalhos foram produzidos em colaboração com outros pesquisadores ou de maneira independente:

Revistas

1. T.A. ALMEIDA, A. YAMAKAMI & M.T. TAKAHASHI. Artificial Immune System to Solve the Minimum Spanning Tree Problem with Fuzzy Parameters. *Pesquisa Operacional*, v. 27, n. 1, 131-154, ISSN 0101-7438, 2007.

Capítulos de livros

1. J. ALMEIDA, R. MINETTO, T.A. ALMEIDA, R. TORRES & N.J. LEITE. Robust Estimation of Camera Motion Using Optical Flow Models. *Advances in Visual Computing*. Book Series: Lecture Notes in Computer Science, Springer Berlin/Heidelberg, v. 5879/2009, 435-446, ISBN 978-3-642-10330-8, 2009.

2. P. BERBERT, L. FREITAS, T.A. ALMEIDA, M. CARVALHO & A. YAMAKAMI. Artificial Immune System to Find a Set of k-Spanning Trees with Low Costs and Distinct Topologies. *Artificial Immune Systems*. Book Series: Lecture Notes in Computer Science, Springer Berlin/Heidelberg, v. 4628/2007, 396-406, ISBN 978-3-540-73921-0, 2007.

Congressos

1. J. ALMEIDA, R. MINETTO, T.A. ALMEIDA, R. TORRES & N.J. LEITE. Estimation of Camera Parameters in Videos Sequences with a Large Amount of Scene Motion. *Proceedings of the 17th International Conference on Systems, Signals and Image Processing (IWSSIP'10)*, Rio de Janeiro, Brasil, Junho, 2010.

2. T.A. ALMEIDA, C.H. DIAS, J. ALMEIDA, R.C. SILVA & A. YAMAKAMI. Sistema Imunológico Artificial Aplicado ao Problema Generalizado de Atribuição. *Proceedings of the 41st Brazilian Symposium on Operational Research (SBPO'09)*, 1-12, Porto Seguro, Brasil, Setembro, 2009.

3. T.A. ALMEIDA & R.C. SILVA. Programação Multi-objetivo Irrestrita com Incertezas. *Proceedings of the 41st Brazilian Symposium on Operational Research (SBPO'09)*, 2994-3005, Porto Seguro, Brasil, Setembro, 2009.

4. T.A. ALMEIDA & A. YAMAKAMI. Immune-Inspired Algorithm to Find the Set of k-Spanning Trees with Lowest Costs in Graphs with Fuzzy Parameters. *Proceedings of the 27th North American Fuzzy Information Processing Society's Conference (NAFIPS'08)*, 1-6, New York, NY, EUA, Maio, 2008.

5. T.A. ALMEIDA & A. YAMAKAMI. An Evolutionary Approach to Find the k-Spanning Trees with Lowest Costs in Graphs with Fuzzy Parameters. *Proceedings of the 10th International Joint Conference on Information Sciences (JCIS'07)*. World Scientific Publishing, v. 1, 1292-1298, Salt Lake City, UT, EUA, Julho, 2007.

Relatórios Técnicos

1. J. ALMEIDA, R. MINETTO, T.A. ALMEIDA, R. TORRES & N.J. LEITE. Robust Estimation of Camera Motion using Optical Flow Models. *Technical Report IC-09-24*, Julho, 2009.

2. J. ALMEIDA, R. MINETTO, T.A. ALMEIDA, R. TORRES & N.J. LEITE. Robust Estimation of Camera Motion using Local Invariant Features. *Technical Report IC-09-12*, Abril, 2009.

Bibliografia

Abbate, J. (2000). *Inventing the Internet*. MIT Press, Cambridge, MA, USA.

Almeida, T. & Yamakami, A. (2010). Content-Based Spam Filtering. In *Proceedings of the 23rd IEEE International Joint Conference on Neural Networks*, pages 1–7, Barcelona, Spain.

Almeida, T., Yamakami, A., & Almeida, J. (2009). Evaluation of Approaches for Dimensionality Reduction Applied with Naive Bayes Anti-Spam Filters. In *Proceedings of the 8th IEEE International Conference on Machine Learning and Applications*, pages 517–522, Miami, FL, USA.

Almeida, T., Yamakami, A., & Almeida, J. (2010a). Filtering Spams using the Minimum Description Length Principle. In *Proceedings of the 25th ACM Symposium On Applied Computing*, pages 1856–1860, Sierre, Switzerland.

Almeida, T., Yamakami, A., & Almeida, J. (2010b). Probabilistic Anti-Spam Filtering with Dimensionality Reduction. In *Proceedings of the 25th ACM Symposium On Applied Computing*, pages 1804–1808, Sierre, Switzerland.

Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., & Spyropoulos, C. (2000a). An Evalutation of Naive Bayesian Anti-Spam Filtering. In *Proceedings of the 11st European Conference on Machine Learning*, pages 9–17, Barcelona, Spain.

Androutsopoulos, I., Koutsias, J., Chandrinos, K., & Spyropoulos, C. (2000b). Experimental Comparison of Naive Bayesian and Keyword-based Anti-spam Filtering with Personal E-mail Messages. In *Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Infornation Retrieval*, pages 160–167, Athens, Greece.

Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C., & Stamatopoulos, P. (2000c). Learning to Filter Spam E-mail: A Comparison of a Naive Bayesian and a Memory-Based Approach. In *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 1–13, Lyon, France.

Androutsopoulos, I., Paliouras, G., & Michelakis, E. (2004). Learning to Filter Unsolicited Commercial E-Mail. Technical Report 2004/2, National Centre for Scientific Research "Demokritos", Athens, Greece.

Assis, F., Yerazunis, W., Siefkes, C., & Chhabra, S. (2006). Exponential Differential Document Count – A Feature Selection Factor for Improving Bayesian Filters Accuracy. In *Proceedings of 4th the MIT Spam Conference*, pages 1–6, Cambridge, MA, USA.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., & Nielsen, H. (2000). Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics*, **16**(5), 412–424.

Barron, A., Rissanen, J., & Yu, B. (1998). The Minimum Description Length Principle in Coding and Modeling. *IEEE Transactions on Information Theory*, **44**(6), 2743–2760.

Bastos, C. & Martins, I. (1989). *Comentários à Constituição do Brasil*, volume 2. Saraiva, São Paulo, Brasil.

Bernardo, J. & Smith, A. (1994). *Bayesian Theory*. Wiley & Sons.

Boykin, P. & Roychowdhury, V. (2005). Leveraging Social Networks to Fight Spam. *Computer*, **38**(4), 61–68.

Braga, I. & Ladeira, M. (2008). Filtragem Adaptativa de Spam com o Principio Minimum Description Length. In *Proceedings of the 28th Brazilian Computer Society Conference*, pages 11–20, Belem, Brazil.

Bratko, A., Cormack, G., Filipic, B., Lynam, T., & Zupan, B. (2006). Spam Filtering Using Statistical Data Compression Models. *Journal of Machine Learning Research*, **7**, 2673–2698.

Breyer, L. (2005). DBACL at the TREC 2005. In *Proceedings of the 14th Text REtrieval Conference*, Gaithersburg, MD, USA.

Carpinter, J. & Hunt, R. (2006). Tightening the Net: A Review of Current and Next Generation Spam Filtering Tools. *Computers and Security*, **25**(8), 566–578.

Carreras, X. & Marquez, L. (2001). Boosting Trees for Anti-Spam Email Filtering. In *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing*, pages 58–64, Tzigov Chark, Bulgaria.

Carvalho, M. (2006). *A Trajetória da Internet no Brasil: do Surgimento das Redes de Computadores à Instituição dos Mecanismos de Governança*. Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil.

Cho, H. & LaRose, R. (1999). Privacy Issues in Internet Surveys. *Social Science Computer Review*, **17**(4), 421–434.

Chuan, Z., Xianliang, L., Mengshu, H., & Xu, Z. (2005). A LVQ-based Neural Network Anti-spam E-mail Approach. *ACM SIGOPS Operating System Review*, **39**(1), 34–39.

Cohen, W. (1995). Fast Effective Rule Induction. In *Proceedings of 12nd International Conference on Machine Learning*, pages 115–123, Tahoe City, CA, USA.

Cohen, W. (1996). Learning Rules that Classify E-mail. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, pages 18–25, Stanford, CA, USA.

Cormack, G. (2006). TREC 2006 Spam Track Overview.

Cormack, G. (2008). Email Spam Filtering: A Systematic Review. *Foundations and Trends in Information Retrieval*, **1**(4), 335–455.

Cormack, G. & Lynam, T. (2005). TREC 2005 Spam Track Overview.

Cormack, G. & Lynam, T. (2007). Online Supervised Spam Filter Evaluation. *ACM Transactions on Information Systems*, **25**(3), 1–11.

Cournane, A. & Hunt, R. (2004). An Analysis of the Tools used for the Generation and Prevention of Spam. *Computers and Security*, **23**(2), 154–166.

Cunningham, P., Nowlan, N., Delany, S., & Haahr, M. (2003). A Case-Based Approach to Spam Filtering than Can Track Concept Drift. In *Proceedings of the 5th International Conference on Case Based Reasoning*, pages 115–123, Trondheim, Norway.

Doneda, D. (2003). *Os Direitos da Personalidade no Novo Código Civil*. A Parte Geral do Código Civil. Renovar, Rio de Janeiro, RJ, Brasil.

Drucker, H., Wu, D., & Vapnik, V. (1999). Support Vector Machines for Spam Categorization. *IEEE Transactions on Neural Networks*, **10**(5), 1048–1054.

Duda, R. & Hart, P. (1973). *Bayes Decision Theory*, chapter 2, pages 10–43. John Wiley & Sons.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, **19**(1), 61–74.

Edwards, P. (1996). *The Closed World*. MIT Press, Cambridge, MA, USA.

Fdez-Riverola, F., Iglesias, E., Diaz, F., Mendez, J., & Corchado, J. (2007a). Applying Lazy Learning Algorithms to Tackle Concept Drift in Spam Filtering. *Expert Systems with Applications*, **33**(1), 36–48.

Fdez-Riverola, F., Iglesias, E., Diaz, F., Mendez, J., & Corchado, J. (2007b). SpamHunting: An Instance-based Reasoning System for Spam Labelling and Filtering. *Decision Support Systems*, **43**(3), 722–736.

Ferraz Junior, T. (2001). A Liberdade como Autonomia Recíproca no Acesso à Informação. *Revista dos Tribunais*, page 247.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, **3**, 1289–1305.

Forman, G. & Kirshenbaum, E. (2008). Extremely Fast Text Feature Extraction for Classification and Indexing. In *Proceedings of 17th ACM Conference on Information and Knowledge Management*, pages 1221–1230, Napa Valley, CA, USA.

Forman, G., Scholz, M., & Rajaram, S. (2009). Feature Shaping for Linear SVM Classifiers. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 299–308, Paris, France.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, **29**(3), 131–163.

Fuhr, N. & Buckley, C. (1991). A Probabilistic Learning Approach for Document Indexing. *ACM Transactions on Information Systems*, **9**(3), 223–248.

Galavotti, L., Sebastiani, F., & Simi, M. (2000). Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization. In *Proceedings of 4th European Conference on Research and Advanced Technology for Digital Libraries*, pages 59–68, Lisbon, Portugal.

Goodman, J. & Yih, W.-T. (2006). Online Discriminative Spam Filter Training. In *Proceedings of the 3rd Conference on Email and Anti-Spam*, pages 1–3, Mountain View, CA, USA.

Goodman, J., Heckerman, D., & Rounthwaite, R. (2005). Stopping SPAM. *Scientific American*, **292**(4), 42–49.

Goodman, J., Cormack, G., & Heckerman, D. (2007). Spam and the Ongoing Battle for the Inbox. *Communications of the ACM*, **50**(2), 25–33.

Grünwald, P. (2005). A Tutorial Introduction to the Minimum Description Length Principle. In P. Grünwald, I. Myung, & M. Pitt, editors, *Advances in Minimum Description Length: Theory and Applications*, pages 3–81. MIT Press.

Guzella, T. & Caminhas, W. (2009). A Review of Machine Learning Approaches to Spam Filtering. *Expert Systems with Applications*. In press.

Gyöngyi, Z. & Garcia-Molina, H. (2005). Web Spam Taxonomy. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web*, pages 1–9, Chiba, Japan.

Hardim, G. (1968). The Tragedy of the Commons. *Science*, **162**(3859), 1243–1248.

Hidalgo, J. (2002). Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization. In *Proceedings of the 17th ACM Symposium on Applied Computing*, pages 615–620, Madrid, Spain.

Joachims, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Proceedings of 14th International Conference on Machine Learning*, pages 143–151, Nashville, TN, USA.

John, G. & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the 11st International Conference on Uncertainty in Artificial Intelligence*, pages 338–345, Montreal, Canada.

John, G., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. In *Proceedings of 11st International Conference on Machine Learning*, pages 121–129, New Brunswick, NJ, USA.

Jung, J. & Sit, E. (2004). An empirical study of spam traffic and the use of dns black lists. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, pages 370–375, Taormina, Italy.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G., Paxson, V., & Savage, S. (2008). Spamalytics: An Empirical Analysis of Spam Marketing Conversion. In *Proceedings of the 15th ACM conference on Computer and Communications Security*, pages 3–14, Alexandria, VA, USA.

Katakis, I., Tsoumakas, G., & Vlahavas, I. (2009). Tracking Recurring Contexts using Ensemble Classifiers: An Application to Email Filtering. *Knowledge and Information Systems*.

Keila, P. & Skillicorn, D. (2005). Structure in the Enron Email Dataset. *Computational & Mathematical Organization Theory*, **11**(3), 183–199.

Kira, K. & Rendell, L. (1992). A Practical Approach to Feature Selection. In *Proceedings of the 9th International Workshop on Machine Learning*, pages 249–256, Aberdeen, Scotland, UK.

Klimt, B. & Yang, Y. (2004). Introducing the Enron Corpus. In *Proceedings of the 1st Conference on Email and Anti-Spam*, pages 1–2, Mountain View, CA, USA.

Kolcz, A. & Alspector, J. (2001). SVM-based Filtering of E-mail Spam with Content-Specific Misclassification Costs. In *Proceedings of the 1st International Conference on Data Mining*, pages 1–14, San Jose, CA, USA.

Kong, J., Rezaei, B., Sarshar, N., Roychowdhury, V., & Boykin, P. (2006). Collaborative Spam Filtering using E-mail Networks. *Computer*, **39**(8), 67–73.

Koprinska, I., Poon, J., Clark, J., & Chan, J. (2007). Learning to Classify E-mail. *Information Sciences*, **177**(10), 2167–2187.

Kraut, R., Sunder, S., Telang, R., & Morris, J. (2005). Pricing Electronic Mail to Solve the Problem of Spam. *Human-Computer Interaction*, **20**(1), 195–223.

Lemire, D. (2005). Scale and Translation Invariant Collaborative Filtering Systems. *Information Retrieval*, **8**(1), 129–150.

Lemos, R., Doneda, D., Souza, C., & Rossini, C. (2007). Estudo sobre a Regulamentação Jurídica do Spam no Brasil. Trabalho comissionado pelo Comitê Gestor da Internet no Brasil ao Centro de Tecnologia e Sociedade (CTS), da Escola de Direito do Rio de Janeiro – Fundação Getúlio Vargas.

Liu, S. & Cui, K. (2009). Applications of Support Vector Machine Based on Boolean Kernel to Spam Filtering. *Modern Applied Science*, **3**(10), 27–31.

Losada, D. & Azzopardi, L. (2008). Assessing Multivariate Bernoulli Models for Information Retrieval. *ACM Transactions on Information Systems*, **26**(3), 1–46.

Lugaresi, N. (2004). European Union vs. Spam: A Legal Response. In *Proceedings of the 1st Conference on Email and Anti-Spam*, pages 145–152, Mountain View, CA, USA.

Luo, C. & Chung, S. (2008). A Scalable Algorithm for Mining Maximal Frequent Sequences Using a Sample. *Knowledge and Information Systems*, **15**(2), 149–179.

Lynam, T., Cormack, G., & Cheriton, D. (2006). On-Line Spam Filter Fusion. In *Proceedings of the 29th International ACM SIGIR Conference*, pages 123–130, Seattle, Washington, USA.

Marsono, M., El-Kharashi, N., & Gebali, F. (2009). Targeting Spam Control on Middleboxes: Spam Detection Based on Layer-3 E-mail Content Classification. *Computer Networks*, **53**(6), 835–848.

Matthews, B. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta*, **405**(2), 442–451.

McCallum, A. & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classication. In *Proceedings of the 15th AAAI Workshop on Learning for Text Categorization*, pages 41–48, Menlo Park, CA, USA.

Meizhen, W., Zhitang, L., & Sheng, Z. (2009). Fuzzy Decision Tree Based Inference Technology for Spam Behavior Recognition. In *Proceedings of the 7th International Symposium on Parallel and Distributed Processing with Applications*, pages 463–468, Chengdu, China.

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam Filtering with Naive Bayes - Which Naive Bayes? In *Proceedings of the 3rd International Conference on Email and Anti-Spam*, pages 1–5, Mountain View, CA, USA.

Mitchell, T. (1997). *Machine Learning*. McCraw-Hill.

Pampapathi, R., Mirkin, B., & Levene, M. (2006). A Suffix Tree Approach to Anti-spam Email Filtering. *Machine Learning*, **65**(1), 309–338.

Peng, T., Zuo, W., & He, F. (2008). SVM Based Adaptive Learning Method for Text Classification from Positive and Unlabeled Documents. *Knowledge and Information Systems*, **16**(3), 281–301.

Perlich, C., Provost, F., & Simonoff, J. (2003). Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *Journal of Machine Learning Research*, **4**, 211–255.

Ramachandran, A., Feamster, N., & Vempala, S. (2007). Filtering Spam with Behavioral Blacklisting. In *Proceedings of the 14th ACM conference on Computer and Communications Security*, pages 342–351, Alexandria, Virginia, USA.

Rissanen, J. (1978). Modeling by Shortest Data Description. *Automatica*, **14**, 465–471.

Sahami, M., Dumais, S., Hecherman, D., & Horvitz, E. (1998). A Bayesian Approach to Filtering Junk E-mail. In *Proceedings of the 15th National Conference on Artificial Intelligence*, pages 55–62, Madison, WI, USA.

Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C., & Stamatopoulos, P. (2003). A Memory-based Approach to Anti-spam Filtering for Mailing Lists. *Information Retrieval*, **6**(1), 49–73.

Schapire, R., Singer, Y., & Singhal, A. (1998). Boosting and Rocchio Applied to Text Filtering. In *Proceedings of the 21st Annual International Conference on Information Retrieval*, pages 215–223, Melbourne, Australia.

Schneider, K. (2003). A Comparison of Event Models for Naive Bayes Anti-spam E-mail Filteriring. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 307–314, Budapest, Hungary.

Schneider, K. (2004). On Word Frequency Information and Negative Evidence in Naive Bayes Text Classification. In *Proceedings of the 4th International Conference on Advances in Natural Language Processing*, pages 474–485, Alicante, Spain.

Sculley, D. & Wachman, G. (2007). Relaxed Online SVMs for Spam Filtering. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 415–422, Amsterdam, The Netherlands.

Sculley, D., Wachman, G., & Brodley, C. (2006). Spam Filtering using Inexact String Matching in Explicit Feature Space with On-Line Linear Classifiers. In *Proceedings of the 15th Text REtrieval Conference*, pages 1–10, Gaithersburg, MD, USA.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, **34**(1), 1–47.

Seewald, A. (2007). An Evaluation of Naive Bayes Variants in Content-based Learning for Spam Filtering. *Intelligent Data Analysis*, **11**(5), 497–524.

Shetty, J. & Adibi, J. (2004). The Enron Email Dataset – Database Schema and Brief Statistical Report. Technical report, University of Southern California, Los Angeles, CA, USA.

Siefkes, C., Assis, F., Chhabra, S., & Yerazunis, W. (2004). Combining Winnow and Orthogonal Sparse Bigrams for Incremental Spam Filtering. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 410–421, Pisa, Italy.

Song, Y., Kolcz, A., & Gilez, C. (2009). Better Naive Bayes Classification for High-precision Spam Detection. *Software – Practice and Experience*, **39**(11), 1003–1024.

Sorkin, D. (2001). Technical and Legal Approaches to Unsolicited Electronic Mail. *University of San Francisco Law Review*, **325**, 357–367.

Van Rijsbergen, C. (1979). *Information Retrieval*. Butterworths, London, 2 edition.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA.

Wang, J., Gao, K., Jiao, Y., & Li, G. (2009). Study on Ensemble Classification Methods Towards Spam Filtering. In *Advanced Data Mining and Applications*, Lecture Notes in Computer Science, pages 314–325. Springer.

Webb, S., Caverlee, J., & Pu, C. (2007). Characterizing Web Spam Using Content and HTTP Session Analysis. In *Proceedings of the 4th International Conference on Email and Anti-Spam*, pages 1–9, Mountain View, CA, USA.

Yang, Y. & Pedersen, J. (1997). A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420, Nashville, TN, USA.

Zhang, L., Zhu, J., & Yao, T. (2004). An Evaluation of Statistical Spam Filtering Techniques. *ACM Transactions on Asian Language Information Processing*, **3**(4), 243–269.