Filtro de Conteúdo para Sistemas SMS Baseado em Classificador Bayesiano e Agrupamento por Palavras

Dirceu Belém[1], Fátima Duarte-Figueiredo[1]

[1]Pontifícia Universidade Católica de Minas Gerais (PUC - Minas)
Rua Walter Ianni, 255, Belo Horizonte, Minas Gerais, 31980-110, MG
dirceu@fourtime.com, fatimafig@pucminas.br

Abstract. *There are many researches about e-mail spam filters. However, there are few researches that look at this issue for SMS (Short Message Service) systems. This is a result of the difficulty in having access to SMS platforms of mobile operators. Furthermore, the volume of spams to SMS systems has increased year after year. The main objective of this work is to propose the implementation of a content filter for SMS systems based on the Bayesian Classifier and word grouping. In order to evaluate the performance of this filter, 120 thousand messages sent from a content provider that services mobile operators were tested. The results demonstrated that the proposed filter reached a correct index spam detection close to 100%.*

Resumo. *Existem muitas pesquisas sobre filtro de spam para e-mails. No entanto, existem poucos trabalhos que abordam o assunto para sistemas SMS (Short Message Service). A quantidade de spams para sistemas SMS tem aumentado a cada ano. O principal objetivo deste trabalho é propor a implementação de um filtro de conteúdo para sistemas SMS baseado em classificador Bayesiano e agrupamento por palavras. Para avaliar o desempenho do filtro, foram utilizadas 120.000 mensagens de um provedor de conteúdo que presta serviço para operadoras de telefonia celular. Os resultados apresentados demonstraram que o filtro proposto obteve um índice de acerto na detecção de spams próximo de 100%.*

1. Introdução

Desde o lançamento da telefonia celular, até os dias de hoje, é possível perceber uma evolução dos dispositivos e dos serviços prestados pelas operadoras. Uma grande variedade de serviços passou a ser oferecida. O SMS (*Short Message Service*), conhecido popularmente como sistema de troca de mensagens de texto ou torpedos, tornou-se um dos mais importantes serviços, por ser de simples utilização e de baixo custo. A utilização dos serviços SMS representou, em 2007 12% da receita das operadoras de telefonia celular da Europa (Gartner, 2008). Na China, em 2000, foram enviadas um bilhão de mensagens SMS. Na Europa, em 2006, a receita das operadoras foi de 429,6 milhões de euros e aumentaria 30% em 2007, segundo He, Sun, Zheng, Wen (2008). As estimativas de faturamento global previstas até 2013, na utilização de serviços SMS, devem ser de US$ 177 milhões (Boas, 2008).

Considerando o alto percentual de mensagens indesejadas em sistemas SMS, este trabalho apresenta um filtro de conteúdo baseado em classificador Bayesiano, para

detectar spams previamente. Na bibliografia consultada, outros autores Deng e Peng (2006) implementaram filtros de conteúdo baseados em Classificadores Bayesianos. A proposta deste trabalho diferencia-se da deles por permitir o agrupamento de palavras das mensagens no filtro e na escalabilidade dos testes. O algoritmo desenvolvido na implementação do filtro, foi integrado a uma plataforma de envio de mensagens para sistemas SMS, de uma empresa provedora de conteúdo SMS, integrada a grandes operadoras de telefonia celular do país.

A implementação do filtro foi realizada em uma ESME (*External Short Message Entity*). Essa entidade é responsável por entregar as mensagens enviadas pelos provedores de conteúdo à SMSC (*Short Message Service Center*). Tanto a ESME quanto a SMSC são componentes de uma arquitetura da operadora de telefonia celular que será explicada detalhadamente no trabalho. O papel do filtro de conteúdo na ESME é classificar e bloquear as mensagens classificadas como spam.

Este trabalho está organizado da seguinte forma: a Seção 2 apresenta os principais conceitos, onde são apresentadas as tecnologias existentes sobre filtros de conteúdo. A Seção 3 apresenta os principais trabalhos relacionados. A Seção 4 apresenta a descrição do filtro baseado em Classificação Bayesiana e agrupamento por palavras. A Seção 5 apresenta os experimentos, os resultados e a análise. A Seção 6 apresenta as conclusões e os trabalhos futuros.

2. Principais Conceitos

2.1. Spam

Spam é definido por Sahami, Dumais, Heckerman e Horvitz (1998) como uma forma de bombardear caixas de mensagens com mensagens não solicitadas a respeito de tudo, desde artigos para venda e formas de como enriquecer rapidamente, a informações sobre como acessar sites pornográficos. Para Cranor e LaMacchia (1998), os principais fatores que contribuem para o crescimento do número de spam são a facilidade de enviá-lo para um grande número de destinatários e de se obter endereços de e-mails válidos, além do baixo custo de envio. Cormack e Lynam (2005), definem o spam como e-mail não solicitado, emitido de forma indiscriminada, direta ou indiretamente, por um remetente que não tem nenhum relacionamento com o destinatário. O spam pode ser considerado o equivalente eletrônico das correspondências indesejadas e dos telefonemas de telemarketing não solicitados.

2.2. Formas de Bloqueio e Detecção de Spam

As formas de bloqueio e detecção de um spam estão divididas em: listas de bloqueio, bloqueio temporário de mensagens (*greylisting*), autenticação da organização que está enviando aquele e-mail (*DomainKeys Identified Mail*) e filtros de conteúdo. As três primeiras técnicas se baseiam em bloquear o e-mail de acordo com o remetente ou quem está enviando o e-mail, podendo bloquear mensagens vindas de um remetente, servidor, ou domínio. A última técnica se baseia em analisar o conteúdo do e-mail e detectar se aquele e-mail é ou não spam. Algumas soluções podem utilizar mais de uma técnica ao mesmo tempo.

2.3. Filtros de Conteúdo propostos para E-mail

A filtragem de conteúdo consiste em analisar e identificar o conteúdo de acordo com uma classe, por exemplo, spam e não spam. Para que seja possível identificar as mensagens, o processo de filtragem precisa passar por um treinamento com uma amostra de mensagens das duas classes. Este processo de treinamento obtém os atributos necessários para a identificação e classificação das mensagens. Deng e Peng (2006), por exemplo, utilizam como atributos, quantidade de caracteres e o remetente, na identificação e classificação das mensagens spam e não spam.

Os primeiros filtros de conteúdo que surgiram foram para spams de e-mail, mas podem ser adaptados para sistemas SMS. Segundo Ming, Yunchun e Wei (2007), várias técnicas sobre filtragem de spam para e-mails foram criadas. Elas podem ser divididas em três técnicas: a primeira é baseada em palavra-chave, onde os algoritmos extraem características do corpo do e-mail e identifica as correspondências com essa palavra-chave. Essa técnica é considerada passiva e demorada. Por exemplo, quando os *spammers* alteram as palavras dos e-mails, o método se torna ineficaz por não encontrar as palavras e, quando há muitas palavras, as pesquisas são muito demoradas.

A segunda técnica filtra o conteúdo. Essa técnica pode ser vista como um caso particular de categorização de texto, onde apenas duas classes são possíveis: spam e não spam. As soluções que utilizam essa técnica estão entre os principais métodos: classificador Bayesiano, SVM (*Support Vector Machine*), K-NN (K-Vizinhos Mais Próximos), Redes Neurais Artificiais, Árvores *Boosting*, Aprendizado Baseado em Memória, Rocchio e Sistemas Imunológicos Artificiais. A filtragem de conteúdo possui algumas desvantagens. Primeiramente, ocorre um desperdício de largura de banda de rede, onde não é possível tomar decisão sobre o tipo do e-mail enquanto o mesmo não tenha sido baixado inteiramente no cliente. Em segundo lugar, é difícil garantir a atualidade da amostra, pela quantidade de e-mails e mudanças nos conteúdos de spam, sendo difícil garantir o efeito e a persistência do algoritmo anti-spam.

A terceira técnica é baseada na filtragem de spam. Essa técnica pode detectar spam e evitar as desvantagens da primeira e da segunda técnica. É uma técnica que reconhece spam através do comportamento dos usuários ao enviar novos e-mails, construindo o processo de decisão para receber e-mails. Essa técnica utiliza o comportamento dos *spammers* para detectar novos spams.

2.4. Classificador Bayesiano

A técnica filtragem de conteúdo baseada em classificador Bayesiano, segundo Silva (2009), é bastante utilizada em problemas de categorização de texto e em filtros anti-spam. A ideia básica é usar a probabilidade na estimativa de uma dada categoria presente em um documento ou texto. Assume-se que existe independência entre as palavras, tornando o filtro mais simples e rápido.

Na teoria da probabilidade, o teorema de Bayes é relacionado à probabilidade condicional de dois eventos aleatórios. Criado por Thomas Bayes, um famoso matemático britânico que viveu no século 18, o teorema de Bayes é utilizado para calcular probabilidades posteriores, a partir de informações coletadas no passado, e representa uma abordagem teórica estatística de inferência indutiva na resolução de problemas (Kantardzic 2003). Por exemplo, ao se observar alguns sintomas de um

paciente, é possível, utilizando-se o teorema, calcular a probabilidade de um diagnóstico (Sahami, Dumais, Heckerman e Horvitz, 1998).

De acordo com o teorema de Bayes, a probabilidade é dada por uma hipótese dos dados, considerada base posterior, que é proporcional ao produto da probabilidade vezes a probabilidade prévia. A probabilidade posterior representa o efeito dos dados atuais, enquanto a probabilidade prévia especifica a crença na hipótese, ou o que foi coletado anteriormente. O teorema nos permite combinar a probabilidade desses eventos independentes em um único resultado (Sahami, Dumais, Heckerman e Horvitz, 1998).

Kantardzic (2003) apresenta a classificação bayesiana da seguinte forma, seja X uma amostra de dados, cuja classe seja desconhecida, e seja H alguma hipótese, tal que os dados específicos da amostra X pertencem à classe C. Pode-se determinar $P(H \mid X)$, a probabilidade da hipótese H possui dados observados na amostra X, onde $P(H \mid X)$ é a probabilidade posterior que representa a confiança na hipótese. $P(H)$ é a probabilidade prévia de H, para qualquer amostra, independente de como a amostra dos dados aparecem. A probabilidade posterior $P(H \mid X)$ baseia-se em obter mais informações na probabilidade prévia $P(H)$. O teorema Bayesiano provê o cálculo da probabilidade posterior $P(H \mid X)$ utilizando as probabilidades $P(H)$, $P(X)$ e $P(X \mid H)$, conforme fórmula (1).

$$P(H \mid X) = \frac{P(X \mid H) \cdot P(H)}{P(X)} \qquad (1)$$

Suponha que existe um conjunto de m amostras $S = \{S_1, S_2, ..., S_m\}$ (conjunto de dados já treinados), onde cada amostra S_i é representada por um vetor de n dimensões $\{x_1, x_2, ..., x_n\}$. Valores de x_i correspondem aos atributos $A_1, A_2, ..., A_n$, respectivamente. Além disso, existem k classes $C_1, C_2, ..., C_k$ e todas as amostras pertence a uma dessas classes. Dada uma amostra de dados adicionais x (onde sua classe é desconhecida), é possível predizer a classe para x usando a probabilidade condicional mais elevada $P(C_i \mid X)$, onde $i = 1, ..., k$. As probabilidades são obtidas a partir do teorema de Bayes:

$$P(C_i \mid X) = \frac{P(X \mid C_i) \cdot P(C_i)}{P(X)} \qquad (2)$$

Em (2), $P(X)$ é constante para todas as classes, somente o produto $P(X \mid C_i) \cdot P(C_i)$ deve ser maximizado. $P(C_i)$ é o número de amostras treinadas para a classe C_i / m (m é o total de amostras treinadas). Tendo em vista a complexidade do cálculo de $P(X \mid C_i)$, especialmente para grandes conjuntos de dados, é realizado um pressuposto ingênuo de independência condicional entre os atributos. Utilizando esse pressuposto, é possível expressar $P(X \mid C_i)$ como um produto:

$$P(X \mid C_i) = \prod_{t=1}^{n} P(x_t \mid C_i) \qquad (3)$$

Em (3), x_t são valores para atributos na amostra X. As probabilidades $P(x_t \mid C_i)$ podem ser estimadas a partir do conjunto de dados treinados.

2.5. Filtros de Mensagens SMS

Dentre as técnicas conhecidas para filtragem de e-mail algumas delas também são utilizadas na filtragem de mensagens SMS. Listas de bloqueio e filtros de conteúdo, utilizando técnicas KNN, SVM e classificação Bayesiana, podem ser citados. Além dessas técnicas, foi encontrada também uma técnica chamada CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*) nos trabalhos de He, Wen e Zheng (2008), Shahreza (2008), Zhang, He, Sun, Zheng e Wen (2008) e Shahreza (2006), onde é possível realizar um teste cognitivo para identificar um ser humano e um computador. Nesse caso, o filtro solicita uma resposta ao remetente sobre uma pergunta que não pode ser realizada por um sistema, inibindo, assim, as mensagens indevidas.

2.6. O problema do falso positivo e do falso negativo

Um dos erros inaceitáveis na detecção de spam é não enviar uma determinada mensagem ao usuário por classificá-la incorretamente como spam. Esse erro é denominado falso positivo. Ou seja, ao classificar essa mensagem, o algoritmo classifica essa mensagem como spam e não a envia ao usuário. Em programas de e-mail, esse problema pode ser parcialmente resolvido configurando-os para enviar mensagens classificadas como spam para uma pasta específica. Sendo assim, o usuário poderá verificar essa pasta em busca de mensagens não spam classificadas incorretamente. Um outro tipo de erro, denominado falso negativo, não é tão grave. O único aborrecimento é que o usuário vai receber novas mensagens spam em sua caixa de entrada de e-mails. Quando isso ocorrer, o único trabalho que o usuário terá será o de excluir essa mensagem, ou denunciá-la como spam, caso o seu programa de e-mails tenha essa opção (Assis, 2006).

3. Principais Trabalhos Relacionados

O trabalho apresentado por Deng e Peng (2006) propõe um filtro de spam baseado em classificador Bayesiano para sistemas SMS. A proposta de Deng e Peng (2006) foi combinar a solução do filtro com o atributo palavra, com o filtro com os atributos, telefone, URL, tamanho da mensagem, valores monetários, lista negra (*blacklist*) de remetentes a serem bloqueados e lista branca (*whitelist*) de remetentes confiáveis.

A conclusão apresentada por Deng e Peng (2006), no trabalho realizado, foi que, ao se adicionar novos atributos propostos no cálculo, os resultados apresentaram uma precisão maior para se classificar as mensagens como spam e não spam. Ou seja, o filtro com os novos atributos obteve um resultado superior quando se compara com o algoritmo sem os novos atributos, reduzindo o número de falsos positivos e falsos negativos. Os resultados obtidos por Deng e Peng (2006) para os testes realizados adicionando todos os atributos foi de 99,1% de acerto.

4. Descrição do filtro baseado em classificação Bayesiana e agrupamento por palavras

A proposta deste trabalho é a implementação de um filtro de conteúdo baseado em classificação Bayesiana para sistemas SMS. A classificação Bayesiana foi escolhida devido à facilidade de implementação e conforme mencionado anteriormente, o filtro implementado neste trabalho se difere dos propostos por outros autores por agrupar palavras no treinamento do filtro e por estar implementando dentro de uma ESME. Este trabalho propõe a utilização de apenas palavras e três atributos para a classificação dos filtros.

4.1. Implementação do Filtro

Resumidamente, o trabalho engloba processos descritos a seguir: foi desenvolvida uma ESME e implantada em uma operadora de telefonia celular. Essa ESME recebe as mensagens enviadas pelos provedores de conteúdo e classifica as mensagens como spam e não spam antes de enviar à SMSC da operadora de telefonia celular. Caso a mensagem seja classificada como não spam, essa mensagem será encaminhada a SMSC da operadora que enviará essa mensagem ao dispositivo móvel do usuário. Foi desenvolvido também um aplicativo em Java para os dispositivos que suportam o sistema operacional Android, onde o usuário poderá classificar as mensagens como spam ou não spam, contribuindo com a base de dados do filtro.

O filtro implementado trabalha sobre cada mensagem para classificá-la como spam ou não spam, da seguinte forma: retira-se os caracteres especiais, acentos, e stopwords, transforma-se todos os caracteres em minúsculos. O conjunto de palavras restantes é armazenado em um vetor. Esse vetor é percorrido sequencialmente, pegando uma, duas, três, quatro, ou cinco palavras, de acordo com a opção feita antes de se iniciar o filtro. Todas as combinações de uma a cinco palavras são armazenadas no vetor. Cada elemento do vetor pode ser consequentemente uma, duas, três, quatro ou cinco palavras. Além disso, cala elemento pode conter um valor agregado, ou seja, um atributo.

Inicialmente, foram obtidas 120.000 mensagens de uma ESME implantada em uma operadora de telefonia celular. As 120.000 mensagens foram classificadas manualmente como spam e não spam de acordo com o caso. Todas as palavras de cada mensagem foram listadas e agrupadas até no máximo cinco palavras. Foi estabelecida uma probabilidade de cada uma dessas palavras e grupos de palavras aparecerem em uma mensagem spam e não spam, através da fórmula de Bayes. Por exemplo, ao se receber uma mensagem com a palavra “compre” a maioria dos usuários consideraria essa mensagem como spam, e, raramente, encontraria essa palavra em uma mensagem considerada não spam. Por isso, foi necessário treinar o filtro para que fosse possível identificar a probabilidade de uma mensagem com a palavra “compre” spam. O mesmo foi realizado para o grupo de duas, três e quatro palavras, como por exemplo “compre agora”, “compre agora carro” e “compre agora carro imperdível”. Esse treino foi feito com 25.000 mensagens spam e 95.000 mensagens não spam, totalizando 1.405.285 grupos de uma a cinco palavras. Durante o treino, o filtro ajustou as probabilidades de cada uma das palavras e grupos encontrados nas mensagens, de acordo com a categoria, spam e não spam.

Foi realizada a remoção de *stopwords*, que consiste em descartar as palavras que pouco refletem o conteúdo de um documento ou são tão comuns que não distinguem nenhuma categoria dos documentos, como por exemplo, artigos e preposições. Cada idioma possui uma lista de palavras consideradas *stopwords*. A lista para o idioma português foi obtida em Balinski (Balinski 2002). Para Silva (2009), nem todas as palavras são igualmente significativas para representar uma categoria. Algumas carregam mais significado que outras. Normalmente, os substantivos, seguidos dos adjetivos e verbos carregam mais representatividade do que outras classes gramaticais como os pronomes, as conjunções e os artigos. A remoção das *stopwords* consiste em descartar as palavras que pouco refletem no conteúdo de um documento ou são tão comuns que não distinguem nenhuma categoria dos documentos. Todas as palavras foram convertidas para minúsculas e acentos e caracteres especiais foram retirados.

Os atributos utilizados nessa pesquisa foram: grupos de uma a cinco palavras, telefones, URL's e valores monetários. Os demais atributos da pesquisa de Deng e Peng (2006) foram desconsiderados, como o tamanho da mensagem e o remetente. Esses atributos foram desconsiderados porque as mensagens utilizadas nessa pesquisa são mensagens enviadas por provedores de conteúdo. Essas mensagens possuem o tamanho maior quando se compara com mensagens enviadas de assinante para assinante, e os remetentes sempre são os mesmos. Cada um desses atributos selecionados contribui para o que se chama de probabilidade posterior no cálculo do teorema de Bayes. Em seguida, a probabilidade é calculada sobre todos os atributos de uma mensagem. O cálculo é realizado para ambos os casos, tanto para a mensagem ser spam, quanto para a mensagem não ser spam. A classificação é determinada a partir do maior valor das probabilidades de spam e não spam. Além disso, o filtro estará em constante atualização. A partir da chegada de novas mensagens, as probabilidades de cada uma das palavras são atualizadas, deixando, assim, cada vez mais precisa a detecção de mensagens spam e não spam.

4.2. Descrição do Algoritmo do Cliente implementado para o Sistema Operacional Android

O algoritmo do cliente foi desenvolvido para ser executado nos celulares que possuem o sistema operacional Android. Foram utilizados a linguagem Java SDK 1.5 do Android, e utilizando banco de dados SQLite. Nesta parte do projeto as pessoas poderão contribuir com as mensagens que recebem em seus celulares, indicando se são ou não spams. Quando a aplicativo é baixado e instalado, o mesmo lê todas as mensagens na caixa de entrada do celular do cliente e envia ao servidor os dados atualizados para serem incluídos na base de dados central do filtro. A partir desse momento, a cada nova mensagem que chegar ao celular, novas probabilidades são calculadas, e novos dados são enviados ao banco de dados do servidor.

4.3. Descrição do Algoritmo do Servidor

O algoritmo do servidor é a parte principal desse projeto. Ele contém toda a implementação do filtro de conteúdo. A cada nova mensagem que o algoritmo receber, são separados nos seguintes elementos: grupos de uma a cinco palavras e atributos. Após essa separação é realizado o cálculo da probabilidade dessa mensagem ser ou não spam. Para cada grupos de palavras e atributos obtidos em uma mensagem, é realizado

uma consulta na base de dados do filtro, identificando a incidência destes elementos como spam ou não spam. Posteriormente, é realizado o cálculo, utilizando a Fórmula (4) a seguir:

$$P(S \mid W) = \frac{P(W \mid S) \cdot P(S)}{P(W)} \qquad (4)$$

Após realizado o cálculo de cada um dos elementos da mensagem é realizado o cálculo da mensagem inteira ser spam, utilizando a Fórmula (5) abaixo.

$$P(S \mid W_i) = \prod_{t=1}^{n} P(s_t \mid W_i) \qquad (5)$$

O desenvolvimento do algoritmo ficou dividido em três módulos. No primeiro módulo, denominado Treinamento, foram inseridas uma lista de mensagens, spam e não spam. As probabilidades de cada um dos atributos e grupos de uma a cinco palavras são calculadas, classificando cada uma delas como spam ou não spam.

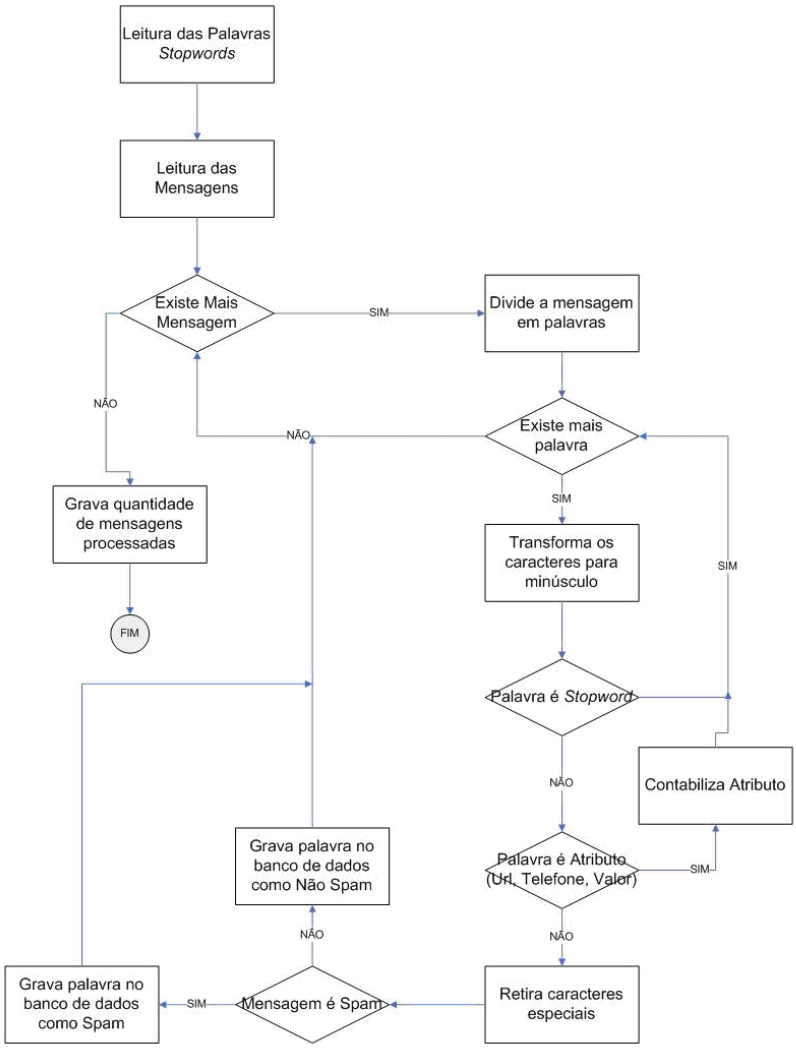

Figura 1. Fluxograma do Módulo Treinamento

No segundo módulo, denominado Classificação, o algoritmo recebe uma mensagem e a classifica como spam ou não spam. No terceiro módulo, denominado Aprendizado, os usuários que utilizarem o aplicativo no sistema operacional Android poderão contribuir com o filtro a partir das classificações realizadas por ele. O módulo Treinamento pode ser entendido melhor no fluxograma da Figura 1 e explicado melhor posteriormente.

O módulo Treinamento apresentado na Figura 1 apresenta o fluxograma de treinamento das mensagens utilizadas. Inicialmente foi realizado a carregamento dos *stopwords* na memória do servidor para utilização futura. Para cada mensagem lida foi realizado o processo de preparação da mensagem, onde todos os caracteres especiais, acentos e *stopwords* são retirados, e foi verificado se na mensagem existe algum atributo, como telefone, URL e valores monetários.

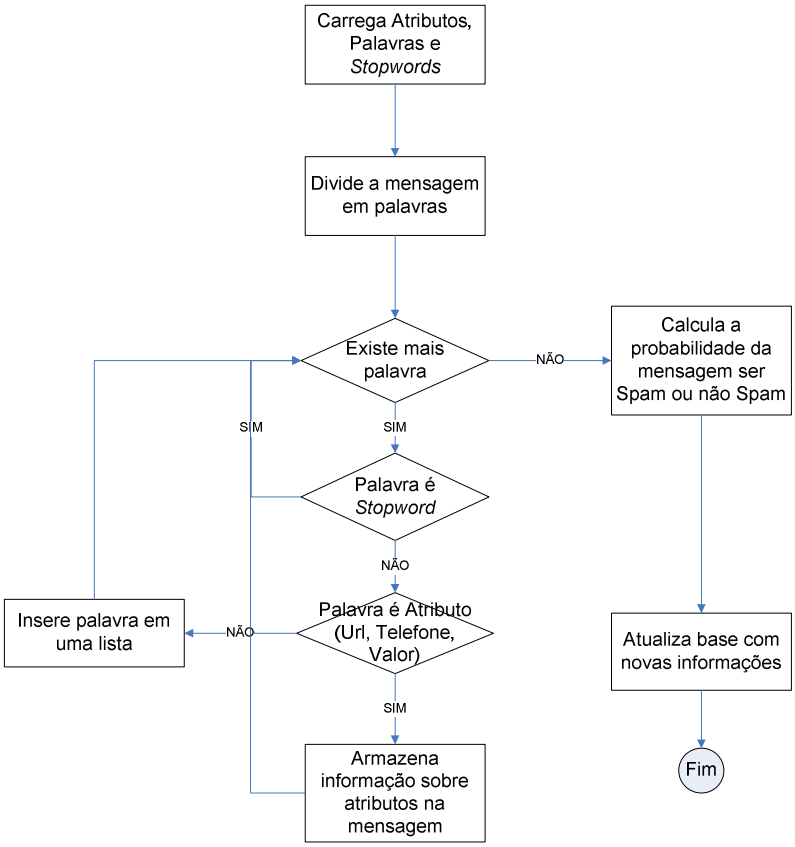

Figura 2. Fluxograma do Módulo Classificação e Aprendizado

O módulo Classificação e Aprendizado apresentado na Figura 2 apresenta o fluxograma de classificação da mensagem enviada. Inicialmente é realizado a carregamento dos *stopwords*, atributos e elementos na memória do servidor para utilização futura. A mensagem é preparada da mesma forma como demonstrado no processo da Figura 1, onde todos os caracteres especiais, acentos e stopwords são retirados, e é verificado se na mensagem existe algum atributo, como telefone, URL e valores monetários.

5. Experimentos e Resultados

Os experimentos foram realizados no filtro de conteúdo após a carga de 120.000 mensagens que foi realizado no processo de treinamento. Dessas 120.000 mensagens, 25.000 mensagens foram classificadas manualmente como spam e 95.000 mensagens foram classificadas manualmente como não spam. Essa carga realizada gerou 1.405.285 elementos, além dos atributos telefone, URL e valores monetários. Para realizar os experimentos foram obtidas mais 8.687 mensagens. Essas mensagens também foram classificadas manualmente, e dentre elas 8008 mensagens foram classificadas como não spam e 636 mensagens foram classificadas como spam. Essa classificação manual foi realizada para que após a obtenção dos resultados os mesmos possam ser conferidos e verificando assim o percentual de acerto do algoritmo. Os experimentos foram divididos em cinco partes. Na primeira parte dos testes, as 8.687 mensagens foram classificadas considerando apenas uma palavra e os atributos. Na segunda parte dos testes, as 8.687 mensagens foram classificadas considerando o grupo de uma a duas palavras, e os atributos. Na terceira parte dos testes, as 8.687 mensagens foram classificadas considerando o grupo de uma a três palavras, e os atributos. Na quarta parte dos testes, as 8.687 mensagens foram classificadas considerando o grupo de uma a quatro palavras, e os atributos. Na quinta parte dos testes, as 8.687 mensagens foram classificadas considerando o grupo de uma a cinco palavras, e os atributos.

Para realizar os experimentos do filtro de spam SMS, foi utilizado um equipamento com um processador Intel (R) Xeon (R), com dois processadores de 2.0 GHz, memória principal de 4 GB, memória secundária de 350 GB, sistema operacional Ubunbu 9.04 e banco de dados PostgreSQL 8.4. O tempo total gasto para este equipamento executar os testes chegou a 181,827 segundos na detecção de spams e não spams para 8.687 mensagens utilizando o agrupamento de uma a cinco palavras.

5.1. Resultados e Análises

A Tabela 1 apresenta os resultados para classificação de grupos de uma a cinco palavras e atributos. Os testes realizados para uma palavra e atributos apresentou o percentual de falsos positivos foi de 0,045% e o percentual de falsos negativos foi 0,044%. O desempenho total do algoritmo foi de 99,955% de acerto. Nessa parte do testes foram utilizados 69.515 grupos de uma palavra. O espaço utilizado para esses grupos foi de 0,49MB. Os testes realizados para o grupo de duas palavras e atributos apresentou o percentual de falsos positivos foi de 0,191% e o percentual de falsos negativos foi 0,007%. O desempenho total do algoritmo foi de 99,978% de acerto. Nessa parte do testes, foram utilizados 366.910 grupos de uma a duas palavras. O espaço utilizado para esses grupos foi de 4,38MB. Os testes realizados para o grupo de três palavras e atributos apresentou o percentual de falsos positivos foi de 0,209% e o percentual de falsos negativos foi 0,005%. O desempenho total do algoritmo foi de 99,9799% de acerto. Nessa parte do testes foram utilizados 717.550 grupos de uma a três palavras. O espaço utilizado para esses grupos foi de 11,05MB. Os testes realizados para o grupo de quatro palavras e atributos apresentou o percentual de falsos positivos foi de 0,218% e o percentual de falsos negativos foi 0,005%. O desempenho total do algoritmo foi de 99,9792% de acerto. Nessa parte do testes foram utilizados 1.073.356 grupos de uma a quatro palavras. O espaço utilizado para esses grupos foi de 19,95MB. Os testes realizados para o grupo de cinco palavras e atributos apresentou o percentual de falsos

positivos foi de 0,220% e o percentual de falsos negativos foi 0,005%. O desempenho total do algoritmo foi de 99,9790% de acerto. Nessa parte do testes foram utilizados 1.405.285 grupos de uma a quatro palavras. O espaço utilizado para esses grupos foi de 30,32MB.

Tabela 1. Resultados dos testes de grupos de uma, duas, três, quatro e cinco palavra e atributos

	Uma Palavra e Atributos	Duas Palavras e Atributos	Três Palavras e Atributos	Quatro Palavras e Atributos	Cinco Palavras e Atributos
Tempo em segundos do teste	53,815	99,041	131,117	148,856	181,827
Número de mensagens	8.687	8.687	8.687	8.687	8.687
Quantidade de mensagens spam	8.008	8.008	8.008	8.008	8.008
Quantidade de mensagens não spam	636	636	636	636	636
Falso positivo	29	122	133	139	140
Falso negativo	360	64	42	42	42
Grupos utilizados	69.515	366.910	717.550	1.073.356	1.405.285
Memória (MB)	0,49	4,38	11,05	19,95	30,32

Conforme demonstrado no gráfico da Figura 3, é possível perceber que o desempenho do filtro foi mais eficiente quando foi utilizado do grupo de até três palavras e atributos. A utilização do grupo de três palavras é mais eficiente por executar em um tempo menor que o grupo de quarto e cinco palavras, e por ter um índice de acerto melhor que o grupo de uma e duas palavras. Também vale destacar a diferença na utilização do grupo de uma palavra e o grupo de até duas palavras, onde o percentual de acerto aumentou significativamente. Quando foi utilizado o grupo de até quatro e cinco, o desempenho teve uma pequena queda.

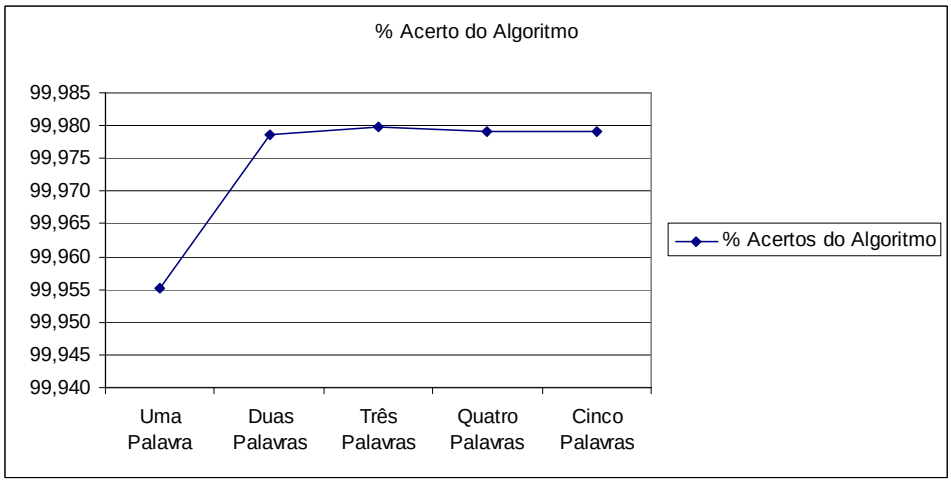

Figura 3. Percentual de acerto do algoritmo para o grupo de uma a cinco palavras

Além dos testes realizandos com 120.000 mensagens, foram realizados testes com 10.000 mensagens para uma comparação com a quantidade de mensagens utilizadas por Deng e Peng (2006). Dentre as 10.000 mensagens, 5.000 mensagens são

spam e 5.000 mensagens não spam. A Figura 4, apresenta os resultados obtidos com o treinamento com 10.000 mensagens e testes com 8.687 mensagens.

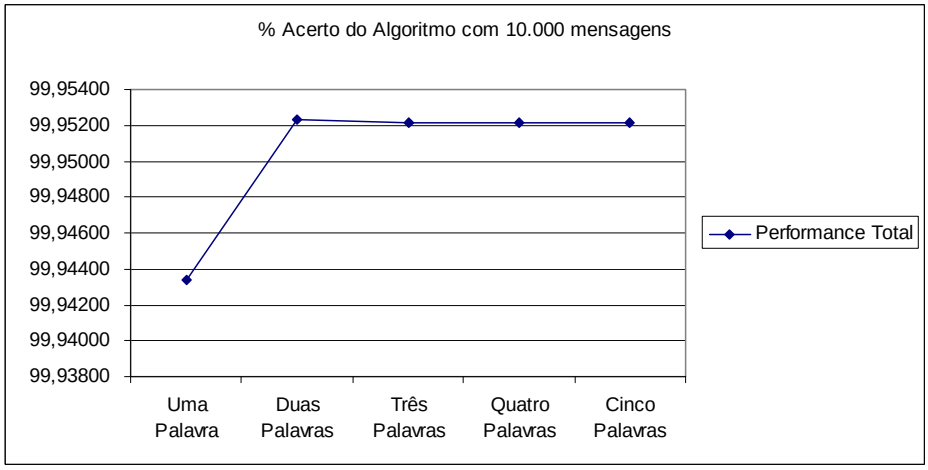

Figura 4. Percentual de acerto do algoritmo para o grupo de uma a cinco palavras com testes realizados com 10.000 mensagens

6. Conclusões e Trabalhos Futuros

A principal contribuição deste trabalho foi a implementação de um filtro de conteúdo para sistemas SMS baseado em classificador Bayesiano com agrupamento de palavras. Com este tipo de agrupamento de palavras, foi possível melhorar o percentual de acerto do algoritmo em comparação com a proposta de Deng e Peng (2006). Este trabalho foi motivado, principalmente, pela necessidade de se detectar e bloquear spams em ambientes reais de sistemas SMS de operadoras de telefonia celular. Como os testes foram realizados sobre mensagens reais, de uma operadora de telefonia celular, os resultados indicam que o filtro apresentado é altamente eficaz e consegue identificar e bloquear spams com quase 100% de acerto. Considera-se os resultados alcançados plenamente satisfatórios e o filtro totalmente possível de ser implementado em redes reais.

A diferença desta proposta para a de Deng e Peng (2006) está na implementação do algoritmo. Enquanto Deng e Peng (2006) usam apenas uma palavra, o filtro aqui proposto permite a combinação de uma a cinco palavras. Os testes para detecção de spams em um ambiente real foram feitos para 120.000 mensagens, enquanto Deng e Peng (2006) testaram apenas para 10.000 mensagens.

Os resultados mostraram que a utilização de grupos de palavras tornou o filtro mais eficiente na classificação de mensagens spam e não spam, principalmente para grupos de duas ou três palavras. A utilização de quatro e cinco palavras não obteve um resultado superior. O índice de acerto do filtro implementado chegou a 99,9799%, enquanto a proposta de Deng e Peng (2006) obteve 99,1% de acerto.

Quatro trabalhos futuros podem ser citados: (1) Extensão do filtro para todas as mensagens trafegadas na SMSC da operadora com o agrupamento de palavras, adicionando os atributos remetente e tamanho da mensagem, propostos por Deng e Peng (2006). (2) Implementação de um filtro de conteúdo com diversas classes. Dessa forma, o filtro ao invés de classificar as mensagens com apenas as classes spam e não spam

poderá classificar as mensagens por assunto, como por exemplo: promoção, esporte, financias, humor, políticas e religiosas. Dessa forma o usuário poderá liberar ou bloquear um determinado assunto. (3) Implementação de um filtro de conteúdo com classes por faixa etária. Dessa forma, o filtro, ao invés de classificar as mensagens com apenas as classes spam e não spam, poderá classificar as mensagens em faixas etárias dos usuários dos dispositivos móveis. Assim, os usuários dos dispositivos móveis que possuem idade abaixo dos 18 anos, por exemplo, não receberão mensagens com conteúdo ofensivo ou pornográfico. (4) Implementação de um filtro de conteúdo que faça uma tarifação diferenciada para o envio de mensagens promocionais ou propaganda. Nesse caso, a operadora, poderá tarifar dos provedores de conteúdo que enviam mensagens com propaganda de forma diferenciada.

Referências

ASSIS, J. M. C. Detecção de E-mails Spam Utilizando Redes Neurais Artificiais. Dissertação (Mestrado) — Universidade Federal de Itajubá - Programa de Pós-Graduação em Engenharia Elétrica, 2006.

BALINSKI, Ricardo. Filtragem de Informações no Ambiente do Direto. 2002. Dissertação (Mestrado em Informática) - Universidade Federal do Rio Grande do Sul, Porto Alegre.

BOAS, Roberta de Matos Vilas. Faturamento com SMS deve crescer globalmente, mas no Brasil preço alto dificulta uso. Disponível em < http://web.infomoney.com.br/templates/news/view.asp?codigo=1241084&path=/suas financas/estilo/tecnologia/>. Acesso em: 19 out. 2008.

CORMACK, G.; LYNAM, T. Spam corpus creation for TREC. In: Proceedings of the Second Conference on Email and Anti-Spam. Mountain View, CA, USA: CEAS, 2005. Disponível em: <http://www.ceas.cc/papers-2005/162.pdf>. Acesso em: 05 nov. 2009.

CRANOR, L. F. LAMACCHIA, B. A. Spam! In: Commun. ACM. New York, NY, USA: ACM, 1998. v. 41, p. 74–83. ISSN 0001-0782.

DENG, Wei-Wei., PENG, Hong. Research on a Naive Bayesian Based Short Message Filtering System. Machine Learning and Cybernetics, 2006 International Conference on, p. 1233-1237, Aug. 2006.

GARNER, Philip, MULLINS, Ian, EDWARDS, Reuben e COULTON, Paul. Mobile Terminated SMS Billing – Exploits and Security Analysis. Industrial Electronics and Applications, 2008. ICIEA 2008. 3rd IEEE Conference on, p. 1319-1324, jun. 2008.

HE, P. SUN, Y. ZHENG, W. WEN, X, Filtering Short Message Spam of Group Sending Using CAPTCHA, Knowledge Discovery and Data Mining, 2008. WKDD 2008. International Workshop on, p. 558 – 561, Jan. 2008.

HE, Peizhou, WEN, Xiangming e ZHENG Wei. A Novel Method for Filtering Group Sending Short Message Spam. Convergence and Hybrid Information Technology, 2008. ICHIT '08. International Conference on, p. 60–65, Aug. 2008.

KANTARDZIC, M, Data Minimg, Concepts, Models, Methods and Algorithms. J. B. Speed Scientific School University of Louisvile, IEEE Computer Society, Sponser 2003

MING L, YUNCHUN L. WEI L. Spam Filtering by Stages. Convergence Information Technology, 2007. International Conference on. p. 2209 - 2213, Nov. 2007

SAHAMI, M. DUMAIS, S. HECKERMAN, D. HORVITZ E. A Bayesian Approach to Filtering Junk E-mail, AAAI Workshop on Learning for Text Categorization, July 1998, Madison, Wisconsin. AAAI Technical Report WS-98-05

SHAHREZA, M. Verifyin Spam SMS y Arabic CAPTCHA, Information and Communication Technologies, 2006. ICTTA '06. 2nd, p. 78 – 83, 2006.

SHAHREZA, S. M.H., An Anti-SMS-Spam Using CAPTCHA. Computing, Communication, Control, and Management, 2008. CCCM '08. ISECS International Colloquium on, p. 318–321, Aug. 2008.

SILVA, Alission Marques Da. Utilização de Redeis Neurais Artificiais para Classificação de Spam. Dissertação (mestrado) – Centro Federal de Educação Tecnológica de Minas Gerais. 126f. Mar. 2009.

ZHANG, H. WEN, X; HE, P. ZHENG, W. Dealing with Telephone Fraud Using CAPTCHA, Computer and Information Science, 2009. ICIS 2009. Eighth IEEE/ACIS International Conference on, p. 1096 – 1099, June 2009.