UM AGENTE DE FILTRAGEM DE CORREIO ELECTRÓNICO INDESEJADO

José Campos
Departamento de Informática
Escola Superior de Tecnologia de Viseu
3500 VISEU
Tel: +351-32-480500, Fax: +351-32-424651
E-mail: jcampos@di.estv.ipv.pt

Edmundo Monteiro
Departamento de Engenharia Informática
Universidade de Coimbra - Pólo II
3030 COIMBRA
Tel.: +351-39-790000, Fax: +351-39-701266
E-mail: edmundo@dei.uc.pt

Sumário

Apresentação de um agente semi-inteligente para filtragem de *email* indesejado pelo seu conteúdo textual. O algoritmo central baseia-se nos métodos de extracção de padrões em textos e na medida *term frequency*. O agente apresenta um comportamento evolutivo em termos de aprendizagem.
São apresentados um protótipo e alguns resultados obtidos.

1. INTRODUÇÃO

A proliferação de mensagens de correio electrónico indesejadas e não solicitadas, comumente designadas por *spam*, tem vindo a aumentar. Estas mensagens, muitas vezes fraudulentas, são enviadas por entidades designadas por *spammers* com o intuito de promover serviços, produtos ou eventos. Testes realizados revelam os conteúdos destas mensagens [1]: 35% - oportunidades de fazer dinheiro fácil; 11% - entretenimento para adultos ou produtos e serviços pornográficos; 10% - marketing directo; 9% - guias informativos; 7% - serviços na Internet, promoção de hardware e *software* e outros produtos para escritório; e 25% - outros produtos e serviços.
Estas mensagens, quando em grande quantidade, provocam uma sobrecarga de informação que levam o utilizador a desperdiçar uma grande parte do seu tempo a filtrar e eliminar as mensagens indesejadas.
Os mecanismos mais comuns para a eliminação das mensagens indesejadas consistem em o utilizador estabelecer regras segundo as quais as mensagens devem ser consideradas *spam*. Estas regras incluem, tipicamente, a verificação de alguns dos *headers* das mensagens de correio electrónico (por exemplo, o *header subject*, o *header from*, etc). Muitos programas *anti-spam* disponibilizam uma lista extensa de *spammers* conhecidos que pode ser imediatamente referida pelo utilizador. Contudo, os *spammers* têm vindo alterar o *header from* por forma a conter um remetente falso, tornando a sua identificação virtualmente impossível.
Todos este mecanismos conseguem uma filtragem relativamente eficaz das mensagens *spam* conhecidas. No entanto, acabam por requerer demasiado empenho por parte do utilizador na sua configuração, o qual deve indicar as características mais peculiares dessas mensagens. Esta tarefa de análise das mensagens e configuração da aplicação *anti-spam* acaba por tornar a mensagem *spam* duplamente desagradável: depois do incómodo de a ter recebido, o utilizador tem que lidar com um sem número de menus de configuração da aplicação *anti-spam*, podendo vir a desperdiçar mais tempo na sua configuração do que propriamente a eliminar as novas mensagens *spam* manualmente.
O mercado encontra-se repleto de programas *anti-spam* que implementam estes mecanismos. Uns são completamente automáticos (eliminam imediatamente qualquer mensagem suspeita de ser *spam*) e outros semi-automáticos (que requerem que o utilizador examine as mensagens bloqueadas antes de serem eliminadas) [1].

2. ABORDAGEM AUTOMÁTICA

A abordagem apresentada neste documento elimina qualquer tipo de configuração ou criação de regras de filtragem por parte do utilizador. Baseia-se no conceito da existência de uma entidade semi-inteligente e autónoma, denominada neste documento por **Agente Anti-Spam** (AAS), capaz de analisar as mensagens de correio electrónico e extrair padrões que caracterizem as mensagens *spam*. O objectivo do AAS é construir um modelo ou perfil das mensagens *spam* para o utilizador e, baseado nesse perfil, eliminar ou marcar as mensagens que analiticamente são consideradas *spam*, antes que estas cheguem ao utilizador.
Com esta abordagem, o utilizador não necessita de configurar o AAS nem de definir as regras de reconhecimento de mensagens *spam*. É o próprio AAS que se "auto-configura".
Inicialmente, o AAS não possui um perfil das mensagens *spam* suficientemente preciso. As primeiras mensagens *spam* não são detectadas. O utilizador deverá indicar ao agente (por exemplo, via *email*) que mensagens devem ser filtradas. Esta abordagem tem, desta forma, a vantagem de não haver necessidade de conhecimento *a priori*, ou de pré-configuração.
Todas as mensagens são analisadas pelo AAS. Da análise resulta uma classificação (ou nota) para essa mensagem. As mensagens boas e as mensagens não desejadas deverão ter uma classificação suficientemente diferente para que seja possível discriminar umas em relação às outras. Por exemplo, as mensagens pessoais do utilizador poderão ter uma nota baixa (a tender para os 0%) enquanto as mensagens *spam* uma nota mais elevada, acima de um determinado limiar.

3. ARQUITECTURA DO SISTEMA

A arquitectura do sistema coloca o AAS na máquina (ou na LAN) onde se encontra o servidor POP3[1]. Deste modo, o AAS pode vir a gerir as *mailboxes* dos vários utilizadores com conta de correio electrónico no servidor. Para tal, deverá manter não um, mas vários perfis de mensagens *spam*, um perfil por cada utilizador. Isto deve-se ao facto de cada utilizador ter gostos e preferências diferentes. Muitas mensagens indesejadas têm origem na forma como o utilizador interage com os serviços de comunicação na rede: as *mailing list* que subscreve, os *sites* WWW que visita, os amigos com quem troca *email*, as *news* que frequenta. Cada utilizador tenderá a receber mensagens *spam* com conteúdos e proveniências diferentes. Mensagens diferentes geram perfis diferentes. A arquitectura do sistema segundo esta perspectiva encontra-se representada esquematicamente na figura 1.

[1] Post Office Protocol [6].

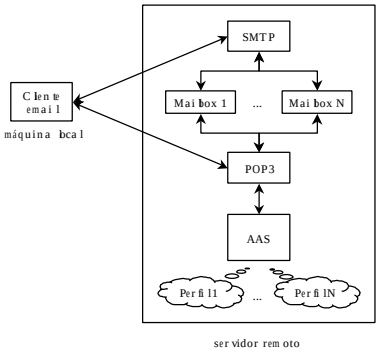

Fig. 1: O AAS residente no servidor

O AAS dialoga com o POP3 por forma a obter as mensagens existentes na *mailbox* do utilizador.
As mensagens não detectadas deverão ser enviadas pelo utilizador para o AAS. Esta operação poderá consistir em reencaminhar por correio electrónico toda a mensagem *spam* para o endereço *email* do AAS (que deverá ser criado pelo administrador do sistema para este efeito). O agente, ao receber essa mensagem, identifica o remetente (o utilizador cujo perfil deverá ser actualizado) e inicia o processo de aprendizagem do conteúdo da mensagem por forma a actualizar o respectivo perfil.
Esta abordagem apresenta, como maior desvantagem, o facto da interacção entre o utilizador e o agente se tornar algo complexa, pelo menos a nível cognitivo. O utilizador não vê o agente, nem pode interagir com ele directamente no seu computador. Terá que comunicar com o agente que se encontra numa máquina remota via *email*. A falta de visibilidade do agente poderá levar à construção de um falso modelo conceptual acerca do AAS [5].
Em contrapartida, esta arquitectura apresenta várias vantagens técnicas. O AAS analisa as mensagens antes de estas saírem da máquina servidora. Deste modo, só as mensagens boas são transmitidas ao respectivo utilizador, diminuindo assim os custos de ligação no caso do utilizador estar ligado ao servidor, por exemplo, via modem. Em alternativa, as mensagens poderão ser marcadas como *spam* em vez de serem eliminadas, para posterior análise do utilizador. Além disso, esta arquitectura permite a realização de filtragem de *mailbox*es para vários utilizadores.

4. REPRESENTAÇÃO DO CONHECIMENTO

O objectivo do sistema é poder confrontar cada mensagem recebida com o perfil das mensagens *spam* de um utilizador por forma decidir se a mensagem é ou não *spam*. Para tal, é necessário representar o conhecimento extraído das mensagens e estabelecer um meio de comparação entre a mensagens e o perfil. Uma forma simples e eficaz de o conseguir consiste em transformar ambos em vectores para depois medir o grau de proximidade entre eles [2].
A figura 2 representa uma mensagem e o perfil num espaço com duas dimensões. No entanto, para um sistema real, as mensagens e o perfil serão mapeados em vectores num hiper-espaço com N dimensões, em que N é o número de palavras relevantes a considerar nos cálculos.

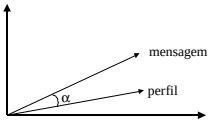

Fig. 2: Representação vectorial da mensagem e do perfil.

A proximidade da mensagem ao perfil pode ser medida através do ângulo α. Quanto menor for α, maior é o grau de semelhança entre a mensagem e o perfil.

5. TRANSFORMAR MENSAGENS EM VECTORES

As mensagens *email* são compostas por duas partes: a partes dos cabeçalhos (*headers*) e a parte do texto da mensagem propriamente dito.
Cada palavra do texto é extraída e testada. Se estiver incluída numa lista de palavras irrelevantes (por exemplo, artigos, advérbios, conjunções, etc), será desprezada. As palavras relevantes denominam-se **termos**.
Os termos não têm todos a mesma importância no que toca à caracterização do texto. Por isso, a cada termo é associado um **peso** que terá um valor proporcional à sua importância na caracterização do conteúdo do texto. Deste modo, um texto é representado por um vector de tuplos, cada tuplo contendo um termo e o respectivo peso:

$$T=\{<t1,p1>,<t2,p2>,...,<tN,pN>\}$$

em que *t1* ... *tN* são os termos relevantes do texto *T* da mensagem e *p1*...*pN* os respectivos pesos.
Qualquer *header* que forneça elemento identificativos do conteúdo e origem da mensagem pode ser utilizado para caracterizar a mensagem sob a forma de um vector. De cada *header* são extraídos os termos que melhor caracterizam a mensagem. A cada termo é atribuído um peso para reflectir a sua importância na caracterização do documento:

$$H=\{<t1,p1>,<t2,p2>,...,<tN,pN>\}$$

Uma mensagem consiste num conjunto de vectores:

$$M=\{V_i^m\}$$

em que V_i^m é o vector *i* da mensagem *M* (*i* pode assumir os valores "*texto*", "*subject*", "*from*", etc). O índice *m* indica que o vector pertence à mensagem, ao contrário de *p* que indica perfil.

6. REPRESENTAR O PERFIL

A representação do perfil é similar à das mensagens. Um perfil consiste num conjunto de vectores de termos. Cada vector representa um atributo relevante das mensagens, tais como "texto", "*subject*", "*from*", etc. Cada termo num vector tem associado um peso proporcional à importância do termo para efeitos de identificação:

$$P=\{V_i^p\}$$

em que V_i^p é o vector *i* no perfil *P* (*p* indica que o vector pertence ao perfil).
Os vectores de um perfil não têm todos a mesma importância na caracterização de uma mensagem (por exemplo, pode-se considerar que o *subject* é mais relevante do que o texto da mensagem). Assim, a cada vector do perfil é atribuído um peso adicional que indica a importância desse vector no perfil.

7. CÁLCULO DOS PESOS

Os pesos associados a cada termo dependem da frequência de ocorrência do termo na mensagem. Esta medida é dada pela conhecida equação *term frequency measure*, que assume que a importância de um termo é directamente proporcional à sua frequência no documento [3]:

$$peso(t) = \frac{frequência(t)}{NúmeroTermos(D)}$$

em que *frequência(t)* é o número de vezes que o termo *termo* ocorre no documento e *NúmeroTermos(D)* é o número de termos que o documento contém.

8. FILTRAGEM DE MENSAGENS

A filtragem de mensagens é o mecanismo que, baseado na representação das mensagens e do perfil, permite identificar as mensagens indesejadas. Uma forma de o conseguir é calcular a semelhança pesada entre os vectores correspondentes na mensagem e no perfil separadamente, e posteriormente, calcular a soma dessas semelhanças:

$$S(m,p) = \sum_i S(V_i^m, V_i^p) \cdot peso_i$$

em que $peso_i$ é o peso do vector i no perfil.
O cálculo da semelhança entre dois vectores é facilmente conseguido através do produto escalar de vectores [2]:

$$S(V_i^m, V_i^p) = \sum_k peso_{i,k}^m \times peso_{i,k}^p$$

em que $peso_{i,k}^m$ é o peso do termo k no vector i da mensagem, e $peso_{i,k}^p$ é o peso do termo k no vector i do perfil.
Para o problema presente, todos os vectores devem ser normalizados antes de se proceder ao seu produto escalar. Os vectores normalizados permitem estabelecer uma comparação entre os diversos pesos dentro de um documento ou perfil pelo facto de estarem todos sob a mesma escala. Além disso, quando os vectores são normalizados, o produto escalar permite-nos concluir que:

$$\vec{u} \cdot \vec{v} = \|\vec{u}\| \cdot \|\vec{v}\| \cdot \cos a = \cos a$$

pois $\|\vec{v}\| = \|\vec{u}\| = 1$, o que facilita o cálculo do ângulo α e consequentemente do grau de semelhança entre os documentos.

9. SELECÇÃO DE MENSAGENS INDESEJADAS

Do cálculo da semelhança *S(mensagem, perfil)* resulta um valor que é utilizado para determinar se uma mensagem é ou não desejada: cos α. Uma mensagem pessoal deverá ser representada por um vector quase perpendicular ao perfil, i.e., α deverá tender para os 90º. As mensagens *spam* deverão ter um α a tender para os 0º. Quanto maior for o valor de *S(mensagem, perfil)*, menor será o α entre os respectivos vectores. Para decidir se uma mensagem é *spam*, o agente utiliza um *threshold* denominado "*do-it*". O valor deste limiar é definido pelo utilizador, representando a **Confiança** que o AAS deverá ter na sua decisão. O utilizador é responsável pelo valor atribuído à Confiança, que deverá manter-se a um nível confortável para o utilizador. Se, por exemplo, o utilizador demonstrar uma grande apreenssão por o AAS eliminar as mensagens *spam* automaticamente, deverá colocar a Confiança no seu valor máximo [4]. O resultado de *S(mensagem, perfil)* é comparado com o *threshold "do-it"*. Se o seu valor for superior a mensagem é considerada indesejada.

10. APRENDIZAGEM

O agente não necessita de conhecimento *a priori*. De facto, partindo de um perfil vazio, o AAS é capaz de aprender, progressivamente, a reconhecer as mensagens indesejadas. Para tal, o utilizador deve enviar-lhe todas as mensagens que considera *spam*. O AAS analisa estas mensagens utilizando o mesmo mecanismo de transformação de mensagens em vectores (extracção de palavras relevantes e atribuição de pesos aos termos) e constrói ou actualiza o respectivo perfil.
A esta contribuição do utilizador dá-se o nome de *feedback*. O *feedback* pode ser positivo (quando o utilizador pretende que o AAS aprenda o conteúdo de uma nova mensagem ou quando pretende aprovar uma decisão correctamente tomada pelo AAS), ou negativo (para indicar que uma mensagem foi incorrectamente classificada). Todos os vectores do perfil são modificados em consequência do *feedback*. O ajuste do perfil é dado pela seguinte equação [2]:

$$P := P + b \times f \times M$$

em que P é o perfil, f é o *feedback* do utilizador e M a mensagem que vai contribuir para a actualização do perfil. O factor b indica a sensibilidade do processo de aprendizagem e pode ser definido pelo utilizador. Um b baixo corresponde a uma aprendizagem conservadora, isto é, que dá mais importância ao conhecimento adquirido no passado. Um b elevado implica uma aprendizagem mais agressiva em que se dá mais importância às mensagens analisadas mais recentemente.

11. IMPLEMENTAÇÃO

Para testar os algoritmos apresentados neste documento, foi implementado um protótipo em PERL. O sistema estabelece periodicamente uma ligação com o servidor POP3. Cada mensagem recebida é transformada em vectores, utilizando-se para tal o *subject* e o texto da mensagem. Ao *subject* foi atribuido um peso de 75% e ao texto um peso de 25%. Nesta transformação são retiradas as palavras irrelevantes e calculados os pesos das restantes palavras. A representação vectorial da mensagem é comparada com o perfil. As mensagens cuja classificação é superior ao *threshold "do-it"* são marcadas como sendo *spam*.
A Figura 5 apresenta um extracto do ficheiro que contém o perfil. A primeira linha indica que o AAS já utilizou 31 mensagens no seu processo de aprendizagem. As duas linhas seguintes indicam os pesos atribuídos a cada vector do perfil. As restantes linhas dividem-se em 4 colunas. A primeira coluna indica se se trata de um termo pertencente ao *subject* ou ao texto, a segunda contém o termo propriamente dito, a terceira o número de mensagens das 31 aprendidas em que ocorreu o termo e, por fim, a quarta coluna o seu peso.

```
31
subject weight:     0.75
text weight:        0.25
subject  day        29    0.538016310611867
subject  games      29    0.348351276087596
subject  net        29    0.348351276087596
subject  download   29    0.348351276087596
...
text     www        31    0.406224865436195
text     http       31    0.405584698804058
text     tipworld   29    0.400972719510715
text     cgi        31    0.273619548339064
text     html       31    0.19202305516097
text     world      31    0.172899458099404  ...
```

Fig. 3: Extracto do perfil.

12. EXPERIÊNCIAS

O procedimento de teste teve como objectivo reproduzir uma situação o mais realista possível. Para tal, foram realizadas algumas subscrições a *mailing lists* a fim de obter mensagens similares com alguma frequência. As mensagens recebidas enquadram-se em quatro categorias:

Categoria	Tamanho	Língua	Spam
1) PC WORLD'S Windows 95 Shareware Pick of the day	> 600 palavras	Inglês	SIM
2) PC WORLD'S Game Picks of the Day	> 600 palavras	Inglês	SIM
3) The GeoCities World Report	> 600 palavras	Inglês	NÃO
4) Emails pessoais	< 250 palavras	Português	NÃO

Foram consideradas mensagens indesejadas as mensagens da "*PC WORLD*", que tanto promove *shareware* para *Windows*ô (categoria 1) como jogos (categoria 2), cuja a língua é o inglês. Para tornar o teste mais realista, foram utilizadas mensagens da "*GeoCities*" (categoria 3), também em inglês. Estas mensagens não foram consideradas indesejadas. Por fim, as mensagens pessoais (categoria 4), na sua maioria escritas em português, não foram consideradas indesejadas.
O sistema foi testado com 381 mensagens nas seguintes proporções: 291 mensagens pessoais, 9 mensagens da "*GeoCities*" e 81 mensagens da "*PC-WORLD*".

TipWorld - http://www.tipworld.com
The Internet's #1 Source for Computer Tips, News, and Gossip

Proudly presents:
PC WORLD's Windows 95 Shareware Picks of the Day

Made possible today by

ThirdAge.com "Money Matters"
9 Safe Places to Put Your Money

Got the stock market jitters? Where else can you
put your money and keep peace of mind? Find out
nine safe places to stash your cash for 1998!
Visit ThirdAge.com-- the Web for GrownUps!
http://www.tipworld.com/arts.cgi?thirdage05

And now for today's shareware picks...

Fig. 4: Excerto de uma mensagem considerada *spam*.

O *threshold* "*do-it*" foi colocado a 20%, o que corresponde a um ângulo α entre os vectores de 72º. Isto significa que mensagens cuja classificação se situou abaixo de 20% ($\alpha \in [72º, 90º[$) foram consideradas não *spam* e mensagens cuja classificação se situou acima de 20% ($\alpha \in [0º, 72º[$) foram consideradas indesejadas. Do processo de aprendizagem resultou um perfil constituído por 2226 termos, dos quais a Figura 5 apresenta os 20 mais relevantes para o texto.

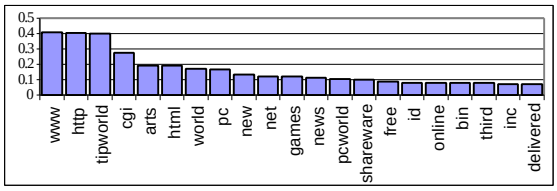

Fig. 5: Ranking dos 20 termos mais relevantes (perfil)

O sistema foi treinado positivamente com 29 mensagens indesejadas de treino. Quando o sistema tentou avaliar as mensagens da "*GeoCities*", o resultado da classificação colocou-as acima do *threshold* definido por duas vezes. Estas duas mensagens foram aprendidas negativamente pelo AAS e a partir de então, as mensagens da "*GeoCities*" passaram a ter uma classificação semelhante às mensagens pessoais. Nenhuma mensagem pessoal foi erradamente classificada pelo AAS ao longo do teste.
A Figura 6 apresenta a classificação das 150 primeira mensagens *não-spam* durante e após o treino. A Figura 7 apresenta as classificações atribuídas às 91 mensagens indesejadas.

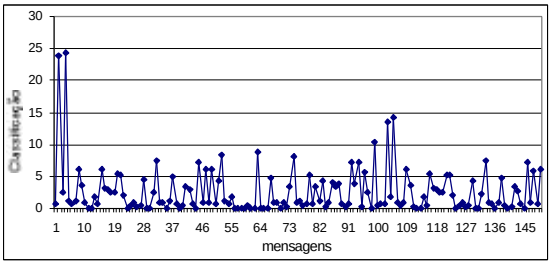

Fig. 6: Classificação para as 150 primeiras mensagens *não-spam*

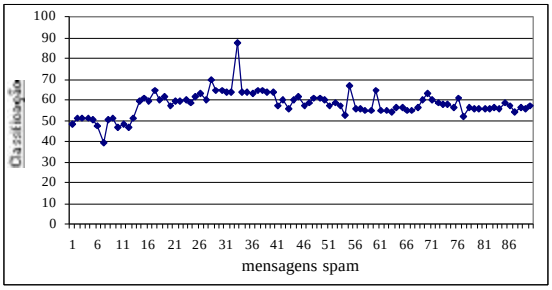

Fig. 7: Classificação das 91 mensagens *spam*.

13. CONCLUSÕES E DIRECÇÕES FUTURAS

Para o caso estudado o sistema demonstrou comportar-se como previsto. O valor do *threshold* "*do-it*" a 20% mostrou-se adequado. No entanto, é de prever que, em situações reais, o utilizador tenha que afinar o *threshold* ao longo do tempo.
Os algoritmos de transformação de mensagens em vectores e cálculo de pesos mostraram ser eficazes na representação das mensagens e na sua comparação com o perfil, evidenciando ser uma abordagem bem sucedida para resolver problemas deste tipo.
O algoritmo de aprendizagem, apesar da sua extrema simplicidade, pareceu atingir os seus objectivos. Outros mecanismos de extracção de padrões em textos poderão vir a demonstrar-se interessantes, tais como as medidas "*term frequency/inverse document frequency*" e "*term relevance*"

[3]. Do mesmo modo, técnicas mais elaboradas de aprendizagem automática conhecidas em Inteligência Artificial deverão ser experimentadas, em particular o C5.0, o CN2, o IBPL1 [3] e redes neuronais.
O sistema poderá ser reforçado com um segundo *threshold* denominado em Maes [4] por "*tell-me*". Este *threshold* deverá situar-se acima do *threshold* "*do-it*" e permitir ao AAS oferecer uma sugestão ou perguntar ao utilizador se uma mensagem é ou não desejada, por forma a desambiguar situações menos claras.

14. REFERÊNCIAS

[1] Cranor, Lorrie F., LaMacchia, Brian A. "Spam!". AT&T Labs-Research Technical Report TR 98.2.1, 1998.

[2] Sheth, Beerud. "A Learning Approach to Personalised Information Filtering". Master's Thesis, Department of Electrical Engineering and Computer Science, MIT, 1994.
ftp://ftp.media.mit.edu/pub/agents/interface-agents/news-filter.ps

[3] Edwards, Peter; Bayer, David; Green, Claire; Payne, Terrry. "Experience with Learning Agents which Manage Internet-Based Information". AAAI Spring Symposium on Machine Learning in Information Access, 31-40. Menk Park, CA:AAAI, 1996.
http://www.parc.xerox.com/istl/projects/mlia/papers/edwards.ps

[4] Maes, Pattie. "Agents that Reduce Work and Information Overload". Communications of the ACM 37(7): 30-40, 1994.

[5] Norman, Donald. "How Might People Interact with Agents". Software Agents, ed. J. Bradshaw, AAAI: 49-55, 1997.

[6] RFC 1939 – POP3 – Post Office Protocol - version 3.