News Clipping e Prevenção à Lavagem de Dinheiro

Aluno: Willian Gigliotti Orientador: Alair Pereira do Lago
Instituto de Matemática e Estatística - USP

Vovó do Pó

"Vovó do Pó foi como ficou conhecida Maria Auxiliadora Benaqui Cortat, mãe do traficante Denilson Benaqui Cortat. Ambos lideravam uma quadrilha de 16 pessoas na cidade de Rezende (RJ). Após a quebra do sigilo bancário e rastreamento dos bens dos integrantes da quadrilha, eles foram denunciados pelo Ministério Público do estado do Rio de Janeiro pelo crime de lavagem de dinheiro, formação de quadrilha e tráfico de drogas. A quadrilha utilizava lojas, duas auto-peças e uma drogaria, além de dez contas bancárias em diferentes bancos para lavar o dinheiro do tráfico e torná-lo legal."

Esse é o resumo de uma notícia publicada em março de 2011, pelo Jornal Extra (http://extra.globo.com).

Dois anos antes, em fevereiro de 2009, o traficante Denilson Benaqui Cortat foi preso junto com outras 17 pessoas por tráfico de drogas, no Vale do Paraíba (SP), durante uma operação do Ministério Público do Estado do Rio de Janeiro.

A prisão de Denilson em 2009 também teve notícias publicadas na mídia, no Jornal Vale de Aço (http://www.jvaonline.com.br/).

News Clipping e Prevenção à Lavagem de Dinheiro

O crime de lavagem de dinheiro, segundo o Conselho de Controle de Atividades Financeiras, COAF, "caracteriza-se por um conjunto de operações comerciais ou financeiras que buscam a incorporação na economia, de modo transitório ou permanente, de recursos, bens e valores de origem ilícita".

Resumidamente, *clipping* é a seleção de notícias ou outras informações de interesse, disponíveis em meios de comunicação. Esse trabalho pode ser feito, por exemplo, pela assessoria de imprensa de uma empresa com o objetivo de acompanhar sua reputação na mídia.

As instituições financeiras são obrigadas pelas Leis 9.613/98 e 12.683/12 a guardar informações de seus clientes com o objetivo de conhecer melhor a origem e a constituição do seu patrimônio.

No final de 2009, foi desenvolvido o Monitoração de Mídias, uma ferramenta elaborada para ajudar as instituições financeiras com a tarefa de conhecer melhor seus clientes.

O Monitoração de Mídias possui um banco de dados com cadastro de pessoas envolvidas em atividades criminosas. A principal fonte de informação do Monitoração de Mídias são notícias publicadas em território nacional. A seleção dessas notícias hoje é realizada manualmente por um técnico, que lê centenas de notícias diariamente em busca de notícias relevantes.

Objetivo do Projeto

A proposta deste projeto é desenvolver um sistema que seja capaz de processar as notícias, agrupando as repetidas em tópicos (*clusters*) e pré-selecionando as relevantes (*classificação*), melhorando a qualidade da seleção de notícias feita posteriormente por um técnico.

As funcionalidades do sistema foram divididas em dois módulos:

O primeiro módulo é automatizado. Ele deve periodicamente consultar novas notícias numa base de dados externa e processá-las. Em seguida, ele deve identificar as notícias repetidas, utilizando um algoritmo de *clustering*, então cada grupo de notícias será classificado, para decidir se é ou não relevante.

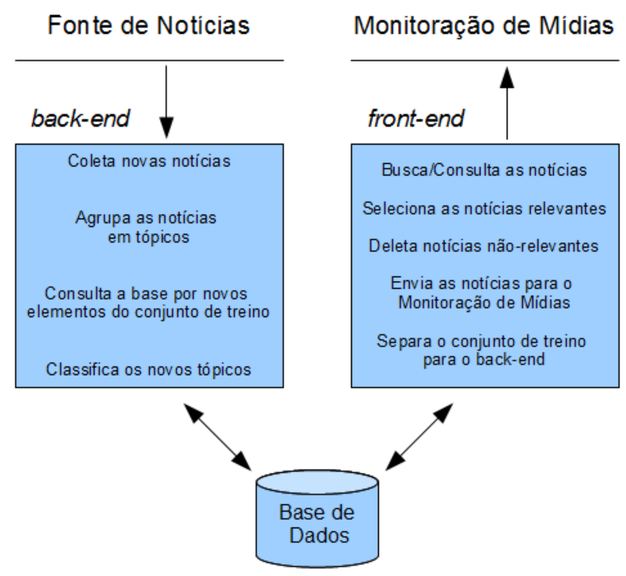
O segundo módulo é a interface com o usuário, onde é possível consultar as notícias pré-classificadas anteriormente, escolher as notícias que entram no Monitoração de Mídias e descartar as notícias não-relevantes.

Funcionalidades e Arquitetura

Técnicas de processamento de texto e classificação de documentos permitiram a criação de um sistema com aprendizado supervisionado capaz de agrupar e pré-selecionar documentos relevantes para o trabalho de *clipping*, resultando em ganho de tempo e de qualidade da seleção.

As principais funcionalidades do sistema de classificação e agrupamento para o *clipping* do Monitoração de Mídias são:

1. Periodicamente buscar por novas notícias numa base de dados externa, fornecida por um fornecedor terceirizado.
2. Processar os textos de cada nova notícia para determinar sua representação no espaço vetorial.
3. Agrupar as notícias em tópicos usando um algoritmo de *clustering*, para identificação das notícias repetidas.
4. Identificar as notícias relevantes utilizando um algoritmo de classificação.
5. Permitir que o usuário selecione as notícias de interesse e apague as não-relevantes.
6. Reservar as notícias marcadas como relevantes e não-relevantes, de acordo com as ações do usuário, integrando-as ao conjunto de treino do classificador.



Arquitetura do Sistema de classificação de notícias

Modelo Vetorial e Similaridade

Para realizar o agrupamento e a classificação das notícias é necessária uma representação matemática para esses elementos. A representação utilizada é o Modelo Vetorial, que atribui pesos para cada termo em seus documentos, com base em sua frequência.

$$\begin{aligned} tfi\text{-}df_{t,d} &= tf_{t,d}.idf_t \\ &= tf_{t,d}.log\left(\frac{N}{df_t}\right) \end{aligned}$$

Onde $tf_{t,d}$ é a frequência em que o termo t ocorre no documento d, df_t é a quantidade de documentos na coleção em que o termo t ocorre e N é o tamanho da coleção.

A partir do modelo vetorial descrito acima, é possível calcular a similaridade entre dois documentos calculando o cosseno do ângulo formado entre seus vetores:

$$sim(d_1, d_2) = \frac{V_{d_1}.V_{d_2}}{|V_{d_1}| \cdot |V_{d_2}|}$$

Clustering

Para realizar o agrupamento das notícias, identificando as repetidas, foi utilizado o algoritmo hierárquico associativo, que inicialmente considera que cada documento é um *cluster*, e a cada iteração do algoritmo é formado um novo *cluster* a partir da junção dos dois *clusters* com maior similaridade entre si.

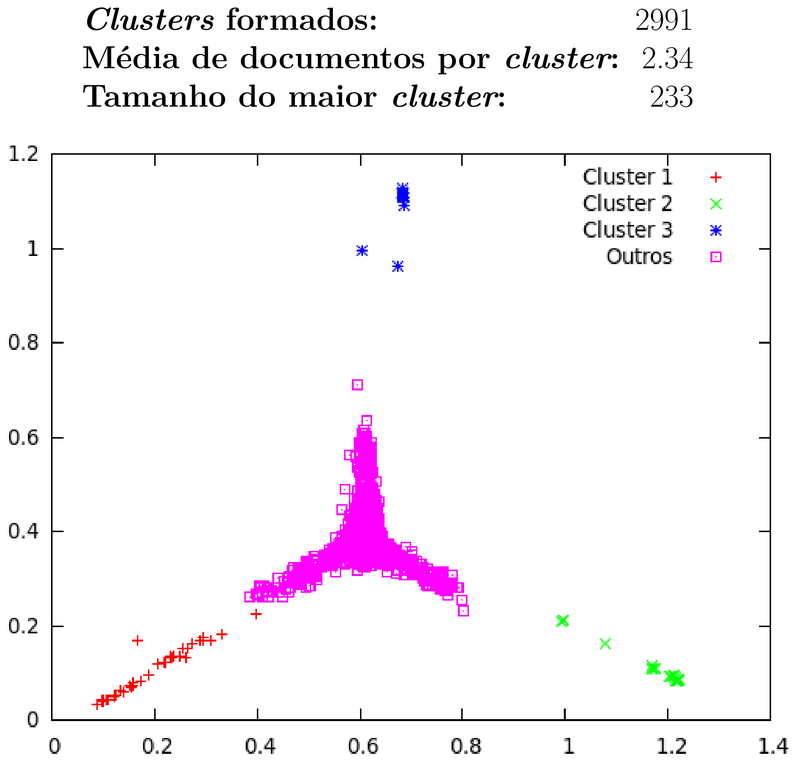
Classificação

A classificação das notícias foi feita a partir do algoritmo kNN. É necessário um conjunto de treino com documentos previamente classificados; a partir desse conjunto de treino, quando é necessário determinar a classificação de um novo documento, o kNN procura dentro do conjunto de treino pelos k documentos mais similares (ou mais próximos) e, baseado na quantidade de documentos de cada classificação encontrada, é determinada a classificação do novo documento.

Resultados de *clustering*

Foi utilizado nos testes de *clusters* um conjunto de 7000 notícias, publicadas entre os dias 14 e 15 de setembro de 2012. Para a realização dos testes, foi utilizado o algoritmo HAC, onde só foram agrupados os *clusters* com similaridades maiores que 0.5 entre si. (Similaridade obtida através do cálculo do cosseno dos ângulos formados entre os vetores centroides de cada *cluster*. Os resultados foram:

Documentos processados:	7000
***Clusters* formados:**	2991
Média de documentos por *cluster*:	2.34
Tamanho do maior *cluster*:	233



Vetores de documentos projetados no plano formado pelos centroides de 3 clusters

Resultados de classificação

Para testar a classificação foi utilizado o *10-fold cross-validation*, onde o conjunto de testes é dividido em 10 partes, e para testar uma dessas partes foram utilizadas como conjunto de treino as outras 9 partes. Os algoritmos testados foram o Rocchio e o kNN com $k = 31$.

Notícias relevantes (Sobre crimes):	1000
Notícias não-relevantes:	991
Total:	1991

Matriz de confusão acumulada para kNN:

	relevantes	não-relevantes	
recuperadas	995	45	1040
não-recuperadas	5	946	951
totais	1000	991	1991

Matriz de confusão acumulada para Rocchio:

	relevantes	não-relevantes	
recuperadas	998	67	1065
não-recuperadas	2	924	926
totais	1000	991	1991

Valores calculados da cobertura, precisão e medida F1 para os algoritmos:

Algoritmo	Coberturtura	Precisão	F1
kNN	0.995	0.956	0.975
Rocchio	0.998	0.937	0.966

Referências

[1] Soumen Chakrabarti. "**Mining the Web: Discovering Knowledge from Hypertext Data.**" Morgan Kaufmann, Amsterdam.

[2] COAF. "**O que é Lavagem de dinheiro?**" https://www.coaf.fazenda.gov.br/conteudo/sobre-lavagem-de-dinheiro-1. Último acesso em 30/9/2012.

[3] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. "**Introduction to Information Retrieval.**" Cambridge.