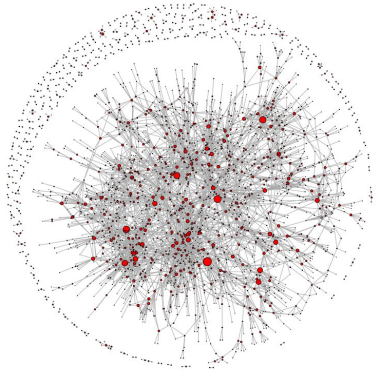
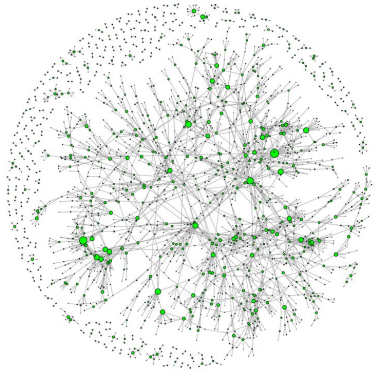
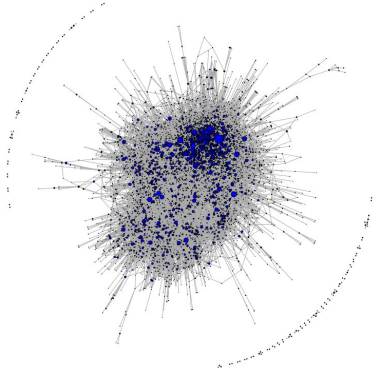






News Clipping

Como calcular ângulos entre notícias pode ajudar na prevenção à lavagem de dinheiro

Willian Gigliotti

wgigliotti@gmail.com

Departamento de Ciências da Computação

Instituto de Matemática e Estatística

Universidade de São Paulo

12 de Novembro de 2012

Agenda

Lavagem de Dinheiro

- Mensalão
- Banco Cruzeiro do Sul
- Banco Panamericano

O que é lavagem de dinheiro

"O crime de lavagem de dinheiro caracteriza-se por um conjunto de operações comerciais ou financeiras que buscam a incorporação na economia de cada país, de modo transitório ou permanente, de recursos, bens e valores de origem ilícita"

Conselho de Controle de Atividades Financeiras - COAF

Combate à lavagem de dinheiro

- Criação do COAF em 1998.

- Leis 9.613/98 e 12.683/12
  - Art. 9º - Responsáveis

    - Bolsa de Valores, Bancos, Seguradoras, Imobiliárias
  - Art. 10º - Responsabilidades

    - Guardar informações dos clientes. Cadastros e transações.

Monitoração de Mídias

- Cadastro negativo de pessoas envolvidas em atividades criminosas.
- Principal fonte: Notícias
- Outras fontes: Lista de trabalho escravo, Empresas incluídas no CEIS, ONGs incluídas no CEPIM ...

Monitoração de Mídias: Como é utilizado

- Durante o cadastro de novos clientes
  - Abertura de contas
  - Concessão de crédito ou financiamento
  - Contratação de fornecedores ou funcionários
- Cruzamento de base de dados
- Acompanhamento diário

Dificuldades encontradas

- Seleção das notícias é totalmente manual
- Quantidade de notícias muito grande, cerca de 7 mil diariamente
- Muitas notícias não são relevantes
- Existe muita repetição de notícias

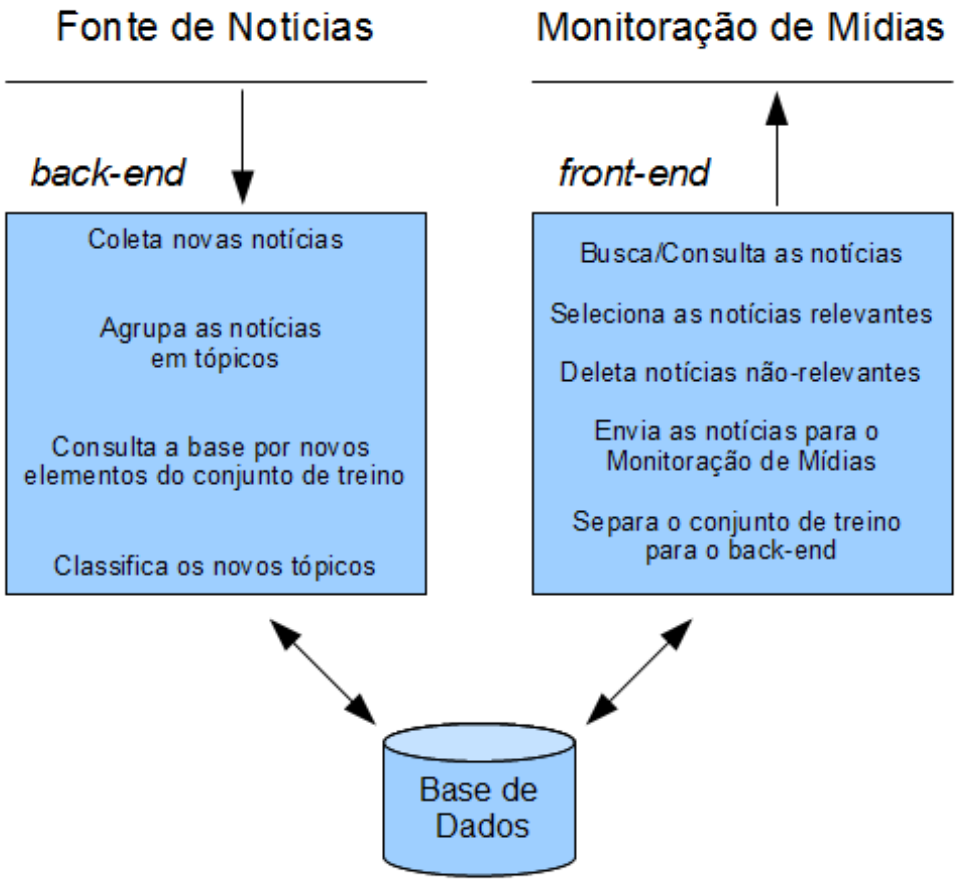
Proposta

- Identificar notícias repetidas
- Pré-selecionar notícias de interesse
- Permitir que o usuário faça a escolha final
- Utilizar as ações do usuário para melhorar as respostas futuras

Ganhos

- Economia de tempo na escolha das notícias
- Melhorar a cobertura
- Melhorar a qualidade da escolhas das notícias

Arquitetura da solução



Modelo Vetorial

Representação dos documentos no espaço vetorial, baseada no peso dos termos em cada documento.

- Quanto maior a frequência de um termo no documento, maior é a relevância do termo para o documento.
- Quanto menor a frequência de um termo no universo de documentos, maior é o peso desse termo para os documentos que o contêm.

Modelo Vetorial

Pesos $tf\text{-}idf_{t,d}$ para os termos de um documento

$$tf\text{-}idf_{t,d} = tf_{t,d}.log\left(\frac{N}{df_t}\right)$$

onde, $tf_{t,d}$ é a frequência do termo t no documento d

e df_t é a quantidade de documentos em que o termo t é encontrado.

Comparação entre documentos: Distância

Comparação por distância euclidiana.

$$dist(V_1, V_2) = |V_1 - V_2|$$

Onde, V_1 e V_2 são os vetores dos documentos d_1 e d_2, respectivamente.

Comparação entre documentos: Cosseno

Comparação pela similaridade do cosseno do ângulo formado entre os documentos.

$$sim(V_1, V_2) = \frac{V_1 . V_2}{|V_1| . |V_2|}$$

Ou apenas:

$$sim(v_1, v_2) = v_1 . v_2$$

se v_1 e v_2 forem os vetores V_1 e V_2 normalizados. ($|v_1| = |v_2| = 1$).

Clustering: HAC

Forma novos clusters a partir de um conjunto de clusters

Crie um grupo $\{d\}$ para cada documento d em D

Seja G o conjunto dos *clusters*

while $|G| > 1$ **do**

Escolha $c_1 \epsilon G$ e $c_2 \epsilon G$, segundo algum critério de similaridade $sim(c_1, c_2)$

Remova c_1 e c_2 de G

$c \Leftarrow c_1 \cup c_2$

Insira c em G

end while

Classificação de Documentos

Foram testados dois algoritmos diferentes:

- Rocchio
- *kNN*

Classificação de Documentos - Rocchio

- **Treino:** Considera os centroides dos vetores dos documentos de cada classe.

- **Teste:** Procura pelo centroide mais próximo do vetor do documento testado.

Classificação de Documentos - kNN

- **Treino:** Não precisa de treino. Precisa apenas dos vetores dos documentos de cada classe.

- **Teste:** Procura pelos k vetores mais próximos do documento testado, então toma uma decisão.

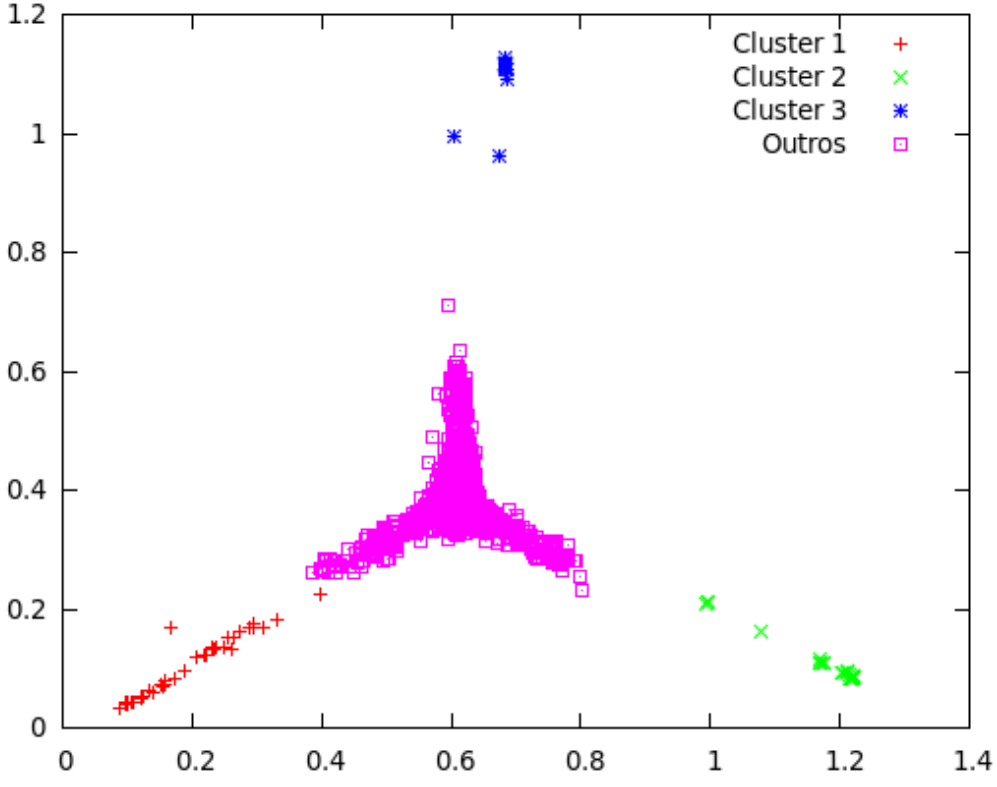
Testes - Agrupamento de Notícias

- Foram utilizadas 7000 notícias
- Notícias principalmente do dia 14 de setembro. (Algumas poucas do dia 15)
- Algoritmo HAC, com limite mínimo de similaridade de 0,5

Resultados - Agrupamento de Notícias

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

Projeção de clusters no plano



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

Testes - Classificação de Notícias

- **Método:** 10-fold cross-validation
- **Notícias relevantes (sobre crimes):** 1000
- **Notícias não-relevantes (outros assuntos):** 991

Resultados - Rocchio

Matriz de confusão acumulada - Rocchio:

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

Resultados - kNN

Matriz de confusão acumulada - kNN:

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

Métricas - Classificação de notícias

$$\text{cobertura} = \frac{\textit{relevantes} \cap \textit{recuperadas}}{\textit{relevantes}}$$

$$\text{precisão} = \frac{\textit{relevantes} \cap \textit{recuperadas}}{\textit{recuperadas}}$$

$$F_1 = 2.\frac{\text{precisão.cobertura}}{\text{precisão} + \text{cobertura}}$$

Resultados - Classificação de notícias

Resultados:

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

Duvidas?