Uma Introdução à Mineração de Informações

Fabrício J. Barth

Apontador

http://www.apontador.com.br

http://www.apontador.com.br/institucional/

fabricio.barth@lbslocal.com

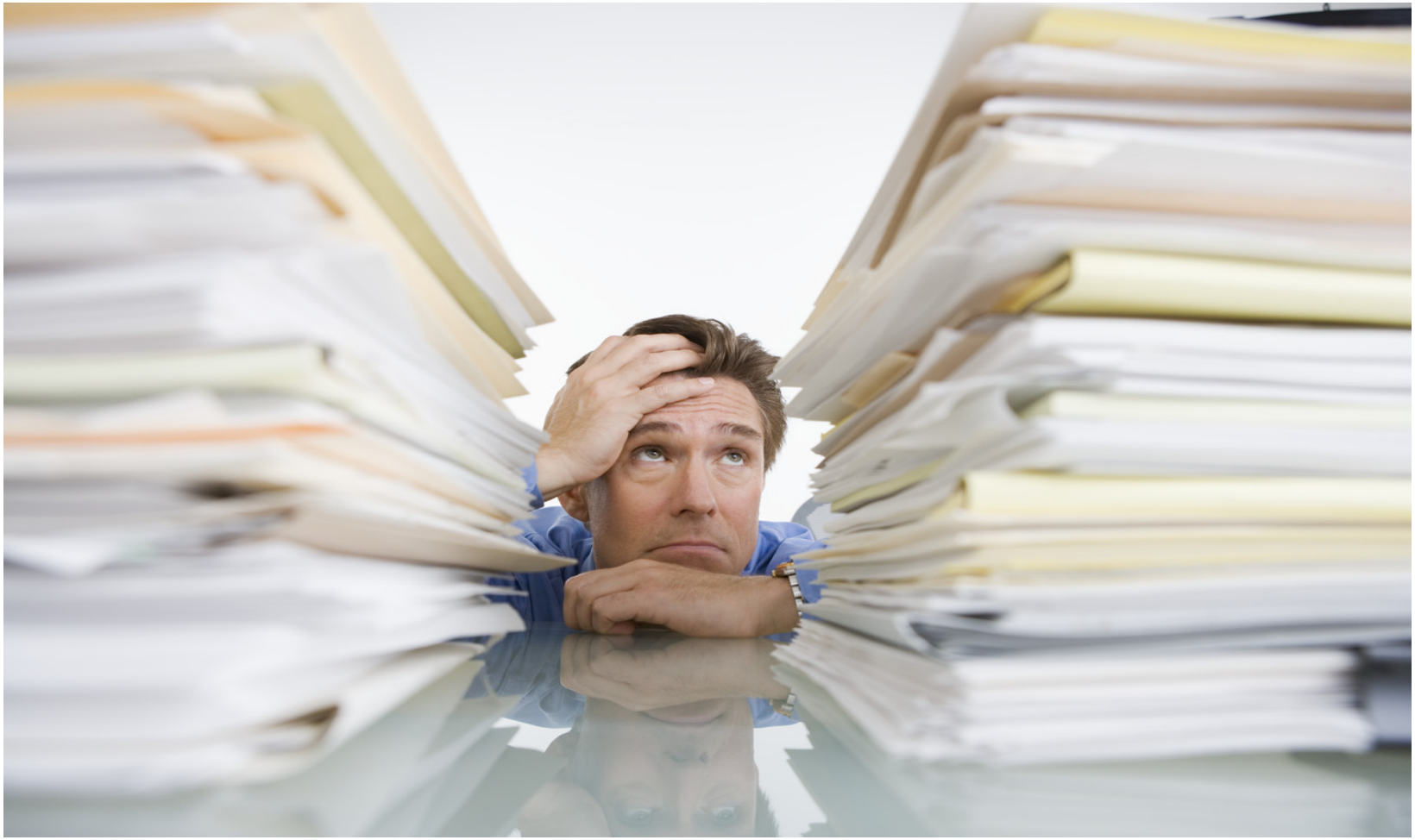
Outubro de 2010

Objetivo

Apresentar a importância do tema, os conceitos relacionados e alguns exemplos de aplicações.

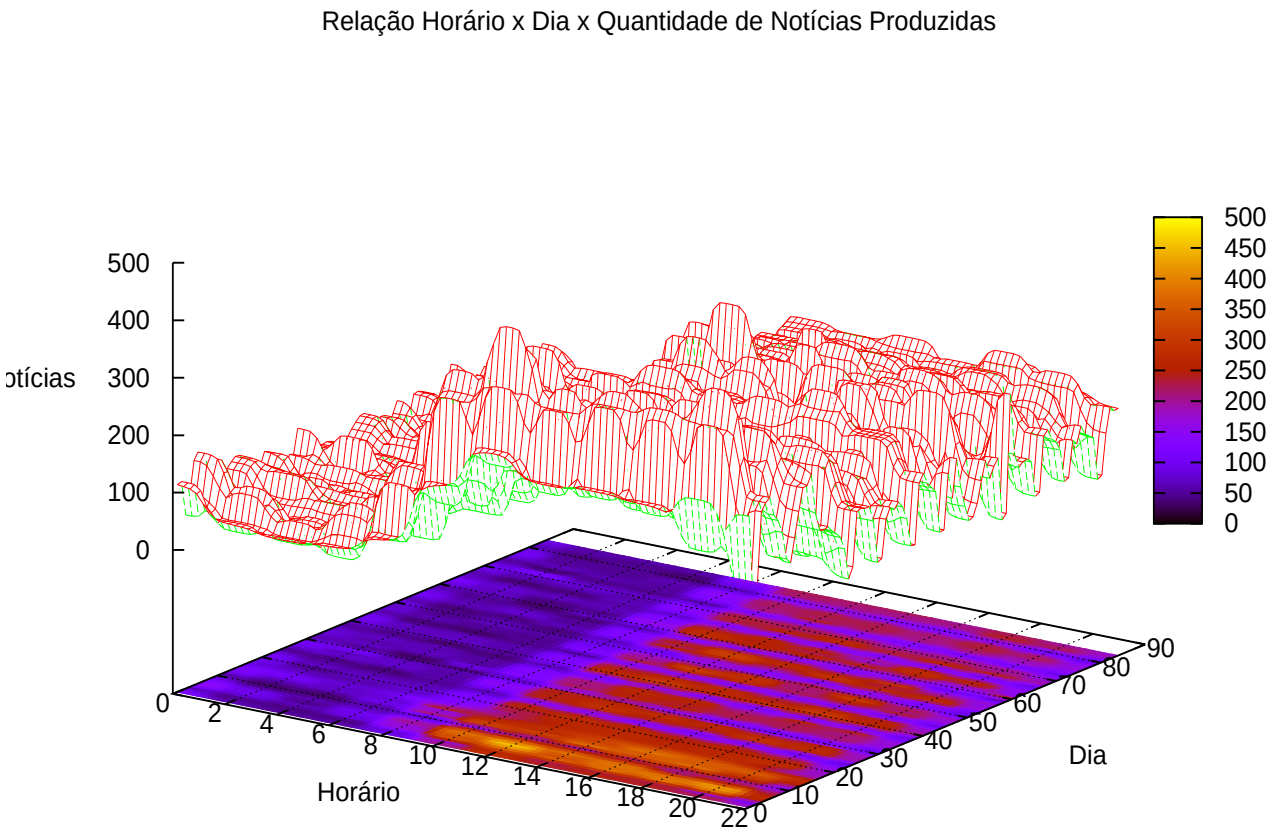
Importância do Tema

Problema



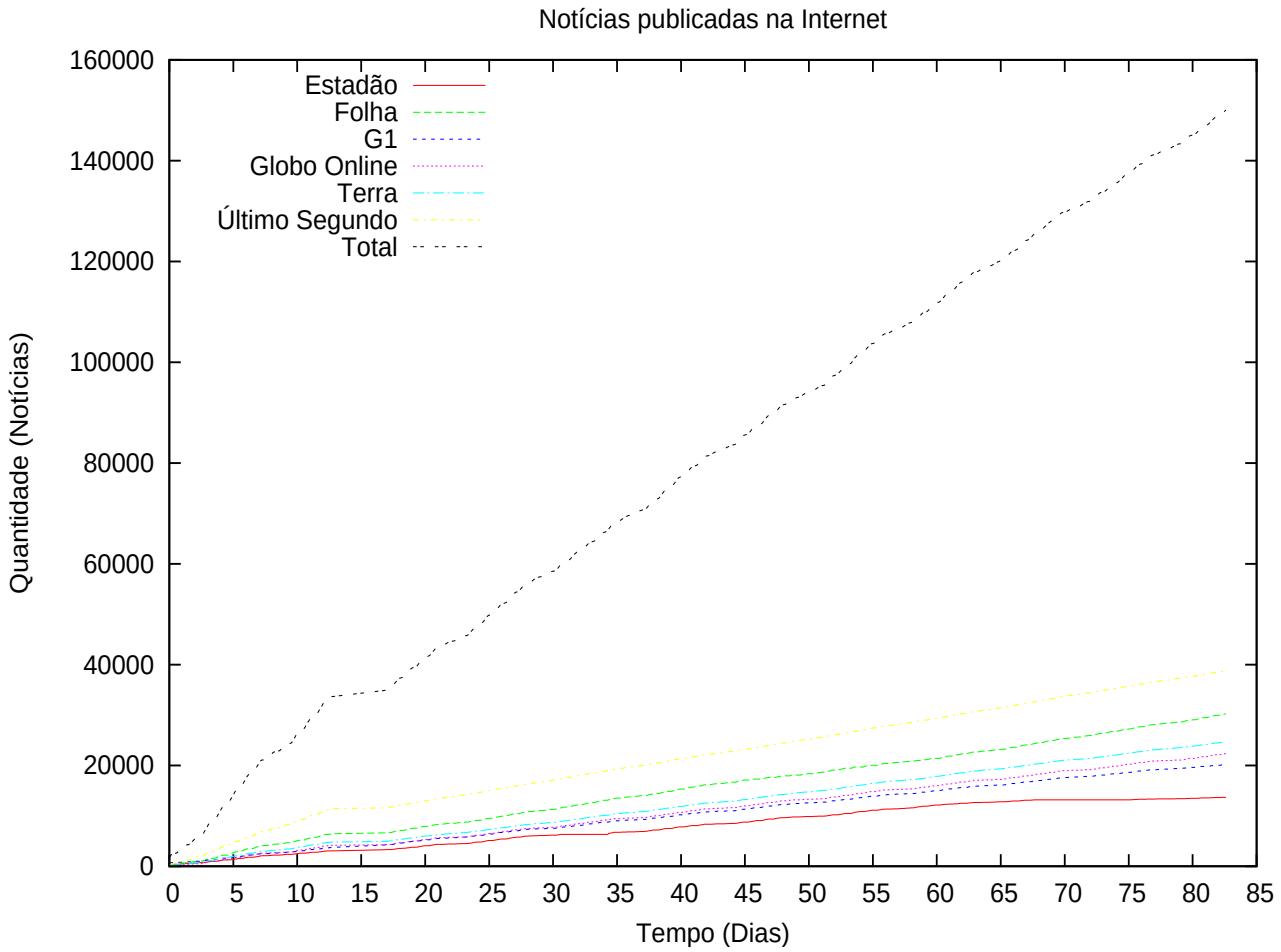
http://investingcaffeine.com/2010/01/07/tmi-the-age-of-information-overload/

Alguns dados...



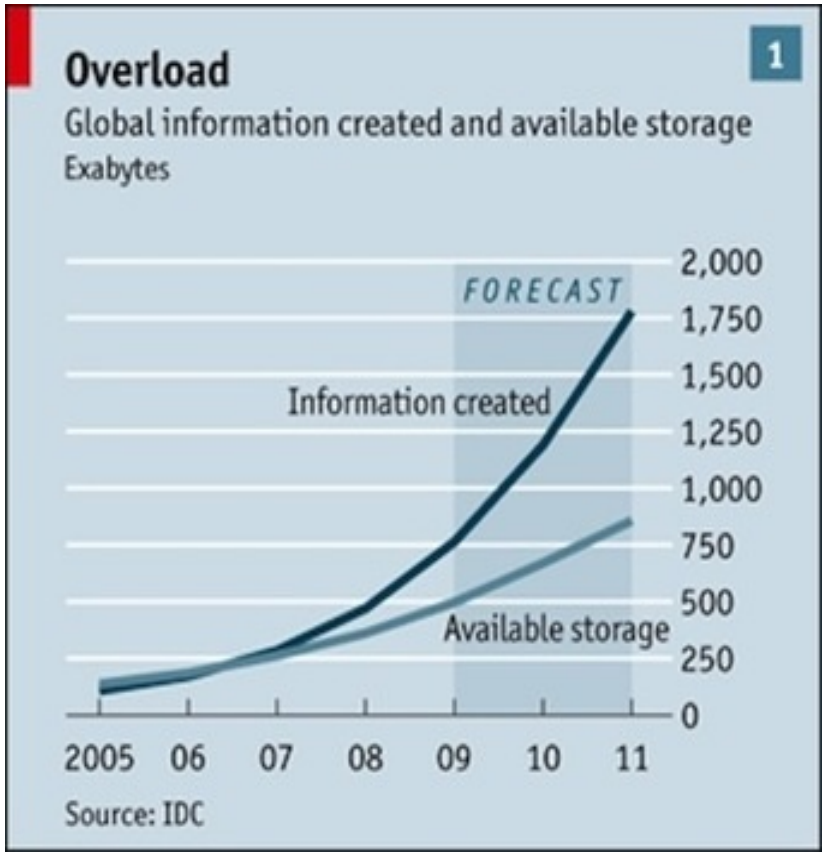
Quantidade de notícias publicadas na Web por apenas seis veículos de notícias (D0 = 17/07/2007)

Mais dados...



D0 = 17/07/2007

Big Data



"We collect an astonishing amount of digital information... ...we've long since surpassed our ability to store and process it all. Big data is here, and it's causing big problems..."[Data, data everywhere. A special report on managing information 2010]

Por que minerar informações?

- Explicitar conhecimento médico a partir de registros médicos.
- Sumarizar tendências de publicações de artigos e patentes sobre um determinado tema.
- Sumarizar e filtrar notícias relevantes.

- Sumarizar a opinião expressa na Web sobre a sua empresa.
- Identificar padrões de navegação em sites.
- Identificar conteúdo impróprio em sites.

Explicitar conhecimento médico a partir de registros médicos

Diagnóstico para o uso de lentes de contato

O setor de oftalmologia de um hospital da cidade de São Paulo possui, no seu banco de dados, um histórico de pacientes que procuraram o hospital queixando-se de problemas na visão.

A conduta, em alguns casos, realizada pelo corpo clínico de oftalmologistas do hospital é indicar o uso de lentes ao paciente.

Problema: Extrair do banco de dados do hospital uma hipótese que explica que paciente deve usar ou não lente de contatos.

Quais são os componentes deste sistema?

Componentes

- Que objetos são relevantes para a criação da hipótese?
- Como representá-los?
- Que linguagem de representação de conhecimento deve-se utilizar para representar a hipótese?
- Que algoritmo utilizar para gerar a hipótese?

- Que objetos são relevantes?
  - ⋆ Depois de um estudo detalhado do problema com especialistas da área...
  - ⋆ **Idade** do paciente.
  - ⋆ Se o paciente tem ou não **miopia**.
  - ⋆ Se o paciente tem ou não **astigmatismo**.
  - ⋆ Qual é a taxa de **lacrimejamento** dos olhos do paciente.
- Como representá-los? **Atributo/Valor**

Atributos

- idade (jovem, adulto, idoso)
- miopia (míope, hipermétrope)
- astigmatismo (não, sim)
- taxa de lacrimejamento (reduzido, normal)
- lentes de contato (forte, fraca, nenhuma)

Dados

Idade	Miopia	Astigmat.	Lacrimej.	Lentes
jovem	míope	não	reduzido	nenhuma
jovem	míope	não	normal	fraca
jovem	míope	sim	reduzido	nenhuma
jovem	míope	sim	normal	forte
jovem	hiper	não	reduzido	nenhuma
jovem	hiper	não	normal	fraca
jovem	hiper	sim	reduzido	nenhuma
jovem	hiper	sim	normal	forte
adulto	míope	não	reduzido	nenhuma

Idade	Miopia	Astigmat.	Lacrimej.	Lentes
adulto	míope	não	normal	fraca
adulto	míope	sim	reduzido	nenhuma
adulto	míope	sim	normal	forte
adulto	hiper	sim	reduzido	nenhuma
adulto	hiper	não	normal	fraca
adulto	hiper	sim	reduzido	nenhuma
adulto	hiper	sim	normal	nenhuma

Idade	Miopia	Astigmat.	Lacrimej.	Lentes
idoso	míope	não	reduzido	nenhuma
idoso	míope	não	normal	nenhuma
idoso	míope	sim	reduzido	nenhuma
idoso	míope	sim	normal	forte
idoso	hiper	não	reduzido	nenhuma
idoso	hiper	não	normal	fraca
idoso	hiper	sim	reduzido	nenhuma
idoso	hiper	sim	normal	nenhuma

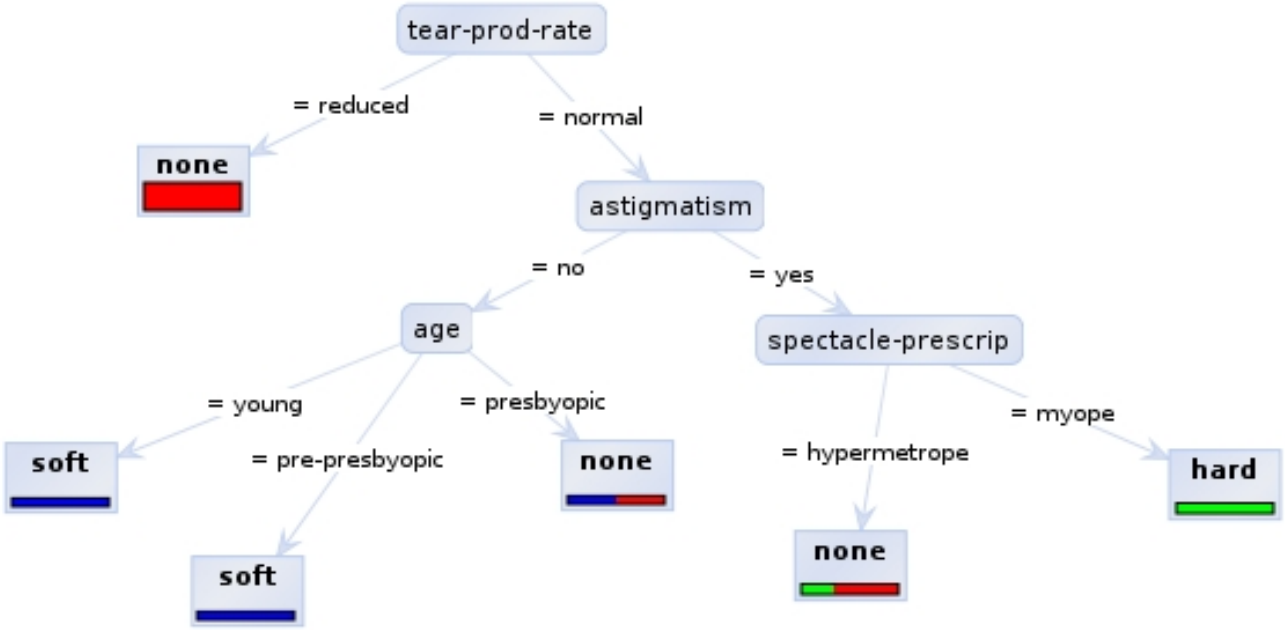
Extração de "conhecimento"

- O que foi apresentado nos slides anteriores pode ser considerado como conhecimento? **Não**
- Pode ser apresentado como uma informação que consegue explicar a tomada de decisão dos especialistas? **Não**
- **O que fazer?**

Extração de "conhecimento"

- Extrair a informação realmente relevante.
- Utilizar uma linguagem de representação **compreensível** ao ser humano.

Árvore de decisão

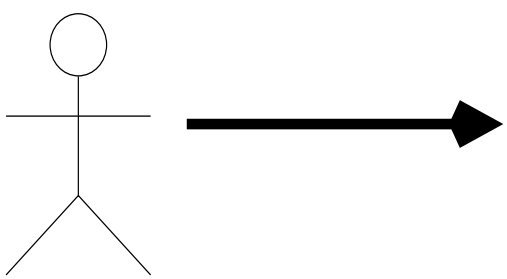


- Cada nodo interno testa um atributo.
- Cada aresta corresponde a um valor de atributo.
- Cada nodo folha retorna uma classificação.

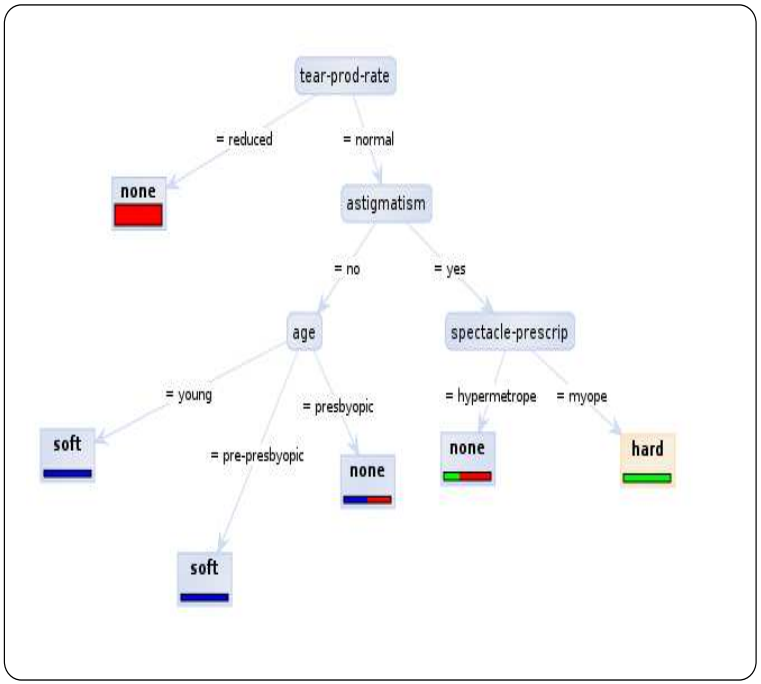
Algoritmos Indutores de Árvores de Decisão

- **Que algoritmo utilizar para gerar hipóteses na forma de árvores de decisão?**
- ID3, C4.5[Quinlan 1988]: são algoritmos indutores de árvore de decisão, **top-down**, **recursivos** e que fazem uso do conceito de **entropia** para identificar os melhores atributos que representam o conjunto de dados.

Resultado: Sistema Especialista



(none, soft, hard)



Sistema Especialista: Regras de Produção

- Baseado na premissa que o processo de tomada de decisão humano pode ser modelado por meio de regras do tipo **SE condições ENTÃO conclusões [FAÇA ações]**
- Convertendo uma árvore de decisão em regras de produção:

- SE lacrimejamento=reduzido ENTÃO lente=nenhuma
- SE lacrimejamento=normal E astigmatismo=não ENTÃO lente=fraca
- SE lacrimejamento=normal E astigmatismo=sim E miopia=miope ENTÃO lente=forte
- SE lacrimejamento=normal E astigmatismo=sim E miopia=hipermetrope ENTÃO lente=nenhuma

Um pouco de código...

- Gerar a árvore de decisão usando o RapidMiner[a].
- Codificar as regras de produção usando o *Drools Expert*[b].

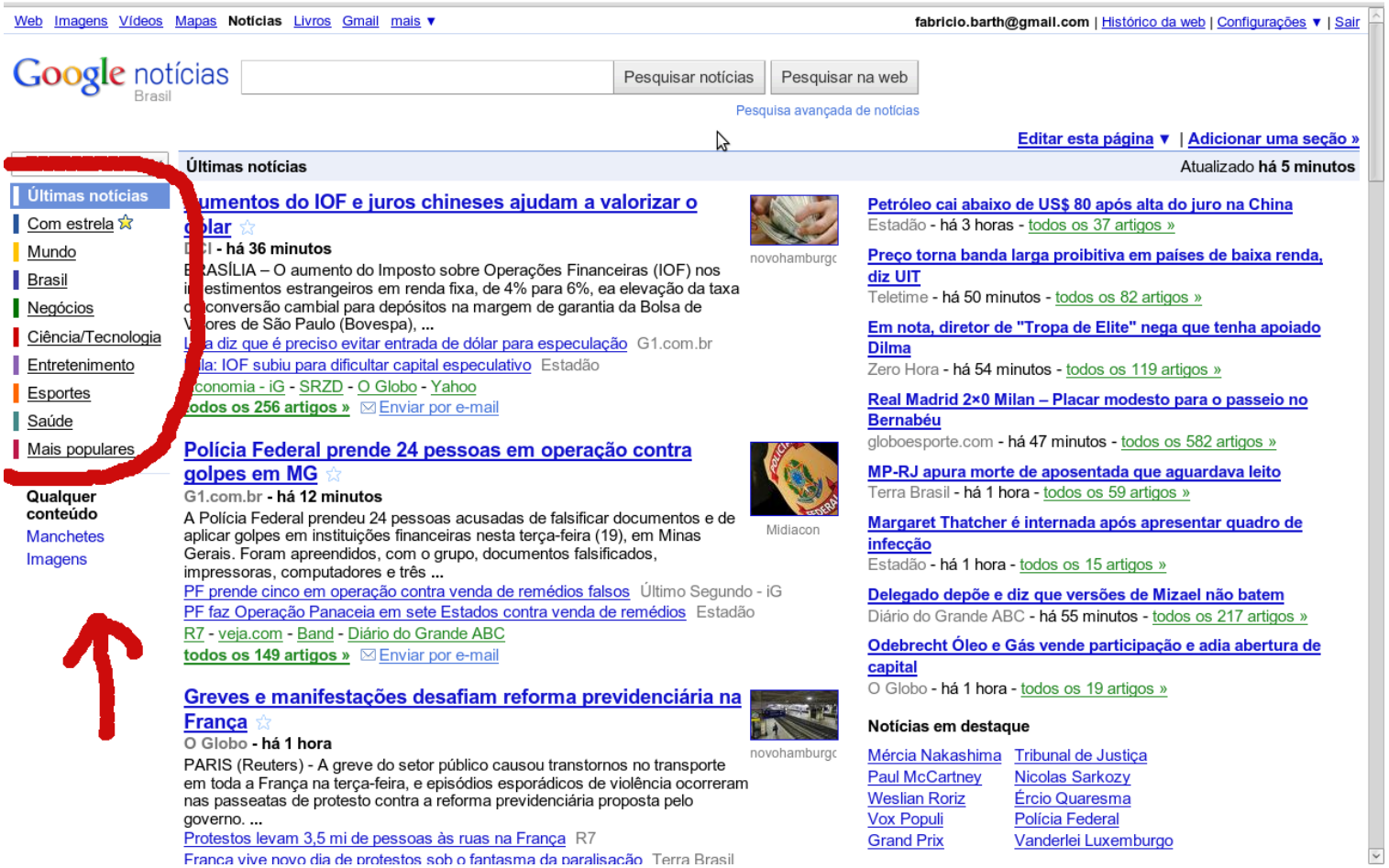
[a] http://www.rapidminer.com
[b] http://www.jboss.org/drools/drools-expert.html

Organizar documentos

O que fazer com grandes quantidades de documentos?

- Notícias, patentes, artigos, mensagens de twitter...
- **Para tirar proveito desta informação é necessário organizá-la de alguma forma**:
  - ⋆ Agrupamento de notícias, patentes, artigos e mensagens.
  - ⋆ Classificação, Recomendação e Filtragem de documentos (notícias, relatórios, mensagens do twitter, avaliação de itens).

Exemplo de classificação/agrupamento



Exemplo de classificação/agrupamento

- E quando não é possível fazer manualmente?
- Qual é o processo para classificar e agrupar documentos de forma automática?

Formato de um documento

... Esta disciplina tem como objetivo apresentar os principais conceitos da área de Inteligência Artificial, caracterizar as principais técnicas e métodos, e implementar alguns problemas clássicos desta área sob um ponto de vista introdutório.
A estratégia de trabalho, o conteúdo ministrado e a forma dependerão dos projetos selecionados pelos alunos.
Inicialmente, os alunos deverão trazer os seus Projetos de Conclusão de Curso, identificar intersecções entre o projeto e a disciplina, e propor atividades para a disciplina. ...

Atributo/Valor usando vetores

Como representar os documentos?

$$\overrightarrow{d_i} = (p_{i1}, p_{i2}, \cdots, pin) \tag{1}$$

- Os atributos são as palavras que aparecem nos documentos.
- Se todas as palavras que aparecem nos documentos forem utilizadas, o vetor não ficará muito grande?

Diminuindo a dimensionalidade do vetor

- Como filtrar as palavras que devem ser usadas como atributos?
- Em todos os idiomas existem átomos (palavras) que não significam muito. **Stop-words**

Esta disciplina **tem como** objetivo apresentar **os** principais conceitos **da** área **de** Inteligência Artificial**,** caracterizar **as** principais técnicas **e** métodos**, e** implementar alguns problemas clássicos **desta** área **sob um** ponto **de** vista introdutório**.**

. . .

Diminuindo ainda mais a dimensionalidade do vetor

- Algumas palavras podem aparecer no texto de diversas maneiras: **técnica**, **técnicas**, **implementar**, **implementação**...
- **Stemming** - encontrar o radical da palavra e usar apenas o radical.

Atributo/Valor usando vetores

- Já conhecemos os atributos.
- E os valores?
  - ⋆ **Booleana** - se a palavra aparece ou não no documento (1 ou 0)
  - ⋆ **Por freqüência do termo** - a freqüência com que a palavra aparece no documento (normalizada ou não)
  - ⋆ **Ponderação tf-idf** - o peso é proporcional ao número de ocorrências do termo no documento e inversamente proporcional ao número de documentos onde o termo aparece.

Por freqüência do termo

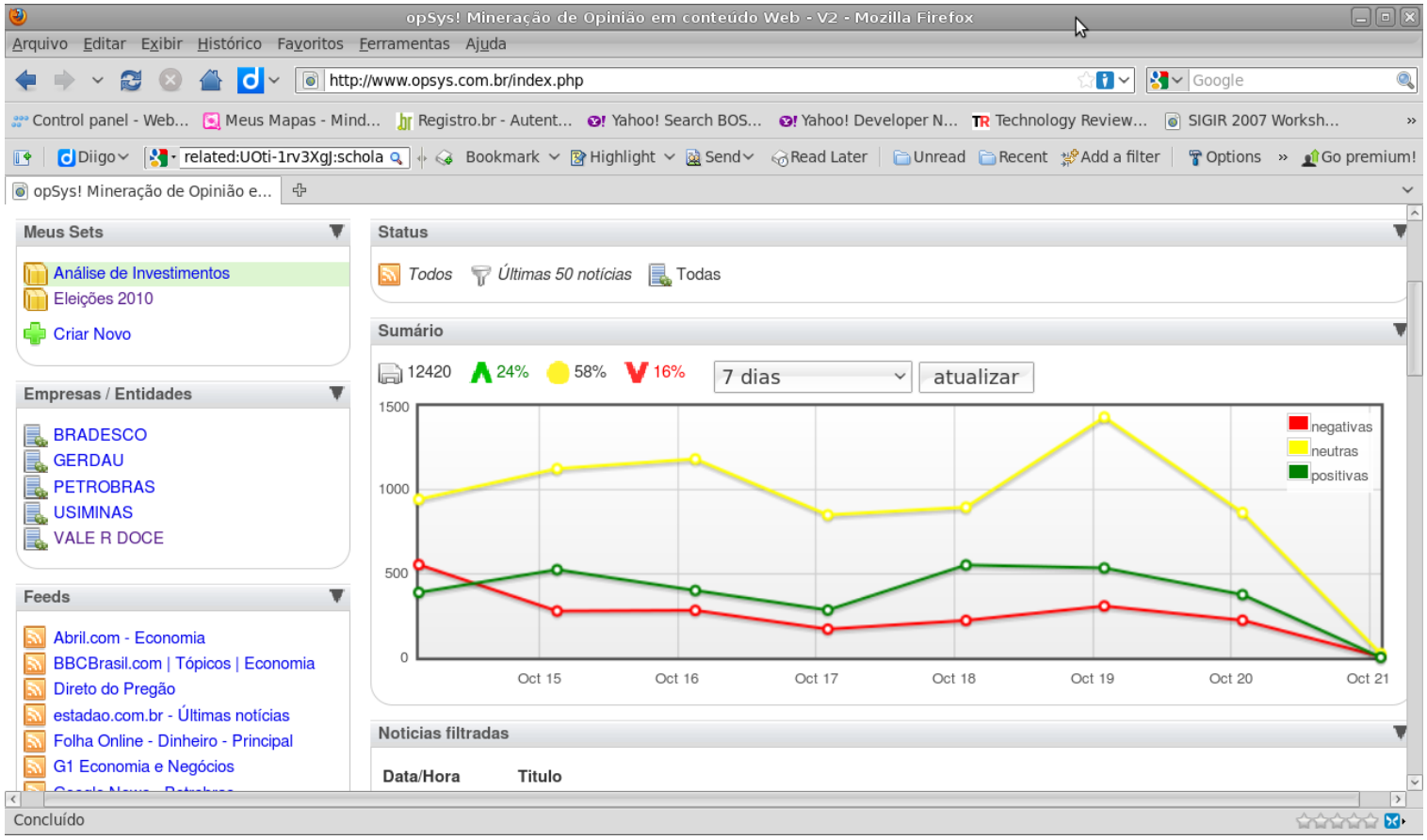
(apresent,0.33) (form,0.33) (tecnic,0.33) (caracteriz,0.33)
(projet,1.0) (introdutori,0.33) (objet,0.33) (inteligente,0.33)
(conclusa,0.33) (selecion,0.33) (intersecco,0.33) (classic,0.33)
(identific,0.33) (conceit,0.33) (trabalh,0.33) (disciplin,1.0)
(traz,0.33)

Conjunto de Exemplos - Atributo/Valor

Doc.	**apresent**	**form**	**tecnic**	**caracteriz**	$\cdots$
d_1	0.33	0.33	0.33	0.33	$\cdots$
d_2	0	0.5	0.2	0.33	$\cdots$
d_3	1	0.6	0	0	$\cdots$
d_4	0.4	0.3	0.33	0.4	$\cdots$
d_5	1	0.4	0.1	0.1	$\cdots$
d_n	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Classificação de documentos

Análise de Sentimento em mensagens no Twitter



Teor das mensagens sobre a Vale nos últimos sete dias.

Conjunto de Exemplos Rotulados

Doc.	Mensagem	Classe
d_1	A empresa X é uma empresa muito séria	Positivo
d_2	O produto Y é uma porcaria	Negativo
d_3	Gostei muito da palestra de fulano	Positivo
d_4	Aquela praia é muito bonita	Positivo
d_5	Gostei daquele restaurante	Positivo
d_n		...

Conjunto de Exemplos - Atributo/Valor e Classe

Doc.	**restaur**	**empres**	**bom**	**caracteriz**	$\cdots$	**Classe**
d_1	0.33	0.33	0.33	0.33	$\cdots$	Positivo
d_2	0	0.5	0.2	0.33	$\cdots$	Negativo
d_3	1	0.6	0	0	$\cdots$	Positivo
d_4	0.4	0.3	0.33	0.4	$\cdots$	Positivo
d_5	1	0.4	0.1	0.1	$\cdots$	Positivo
d_n	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Algoritmo Naïve Bayes

NaiveBayesLearn(exemplos): $P'(v_j)$ e $P'(a_i|v_j)$

for all valor alvo v_j **do**

$P'(v_j) \leftarrow$ estimar $P(v_j)$

for all valor de atributo a_i de cada atributo a **do**

$P'(a_i|v_j) \leftarrow$ estimar $P(a_i|v_j)$

end for

end for

ClassifyNewInstance(x): V_{NB}

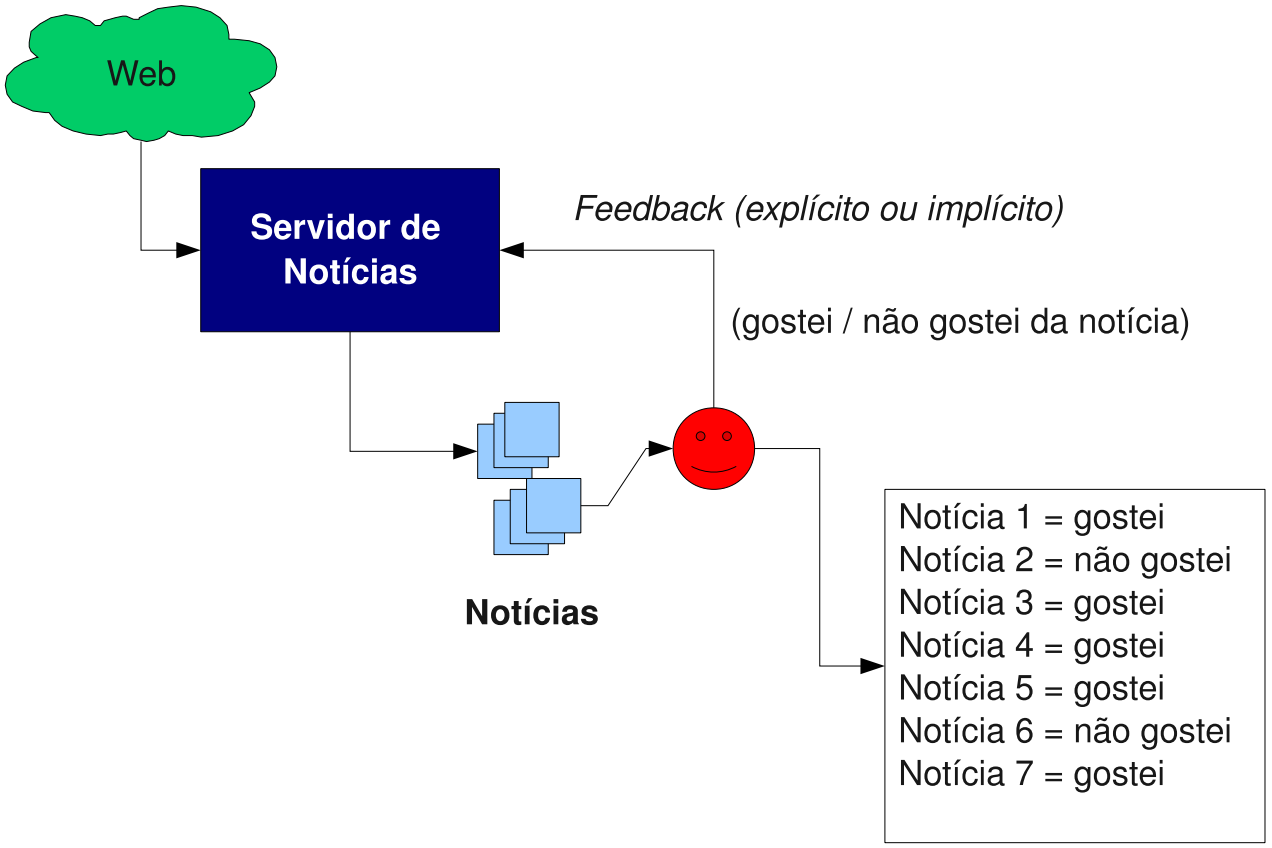
$V_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j)$

Exemplo

Execução de um demo usando estes conceitos...

```
Transformando o conjunto de treinamento em um vetor de palavras
Criando o modelo
Aplicando o modelo a novos casos
Mensagem "Meu voto e para X, com certeza!" e classificada como POSITIVA
Mensagem "Este produto e muito ruim" e classificada como NEGATIVA
Mensagem "Nunca mais compro naquela loja!" e classificada como NEGATIVA
Mensagem "Fulano e um mentiroso!" e classificada como NEGATIVA
Mensagem "X lidera intenÃ§Ã£o de voto" e classificada como POSITIVA
```

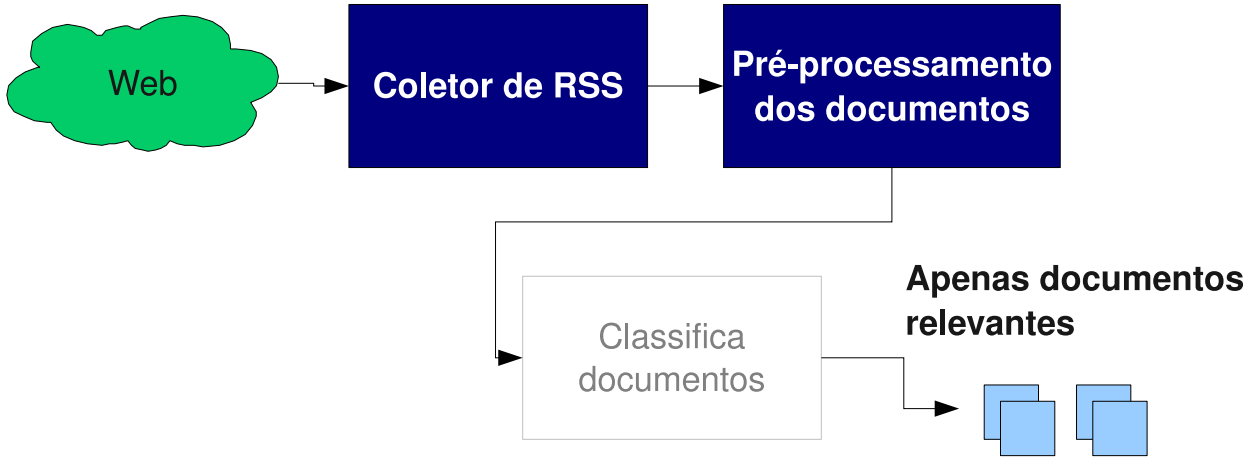
Classificação e Filtragem de Notícias



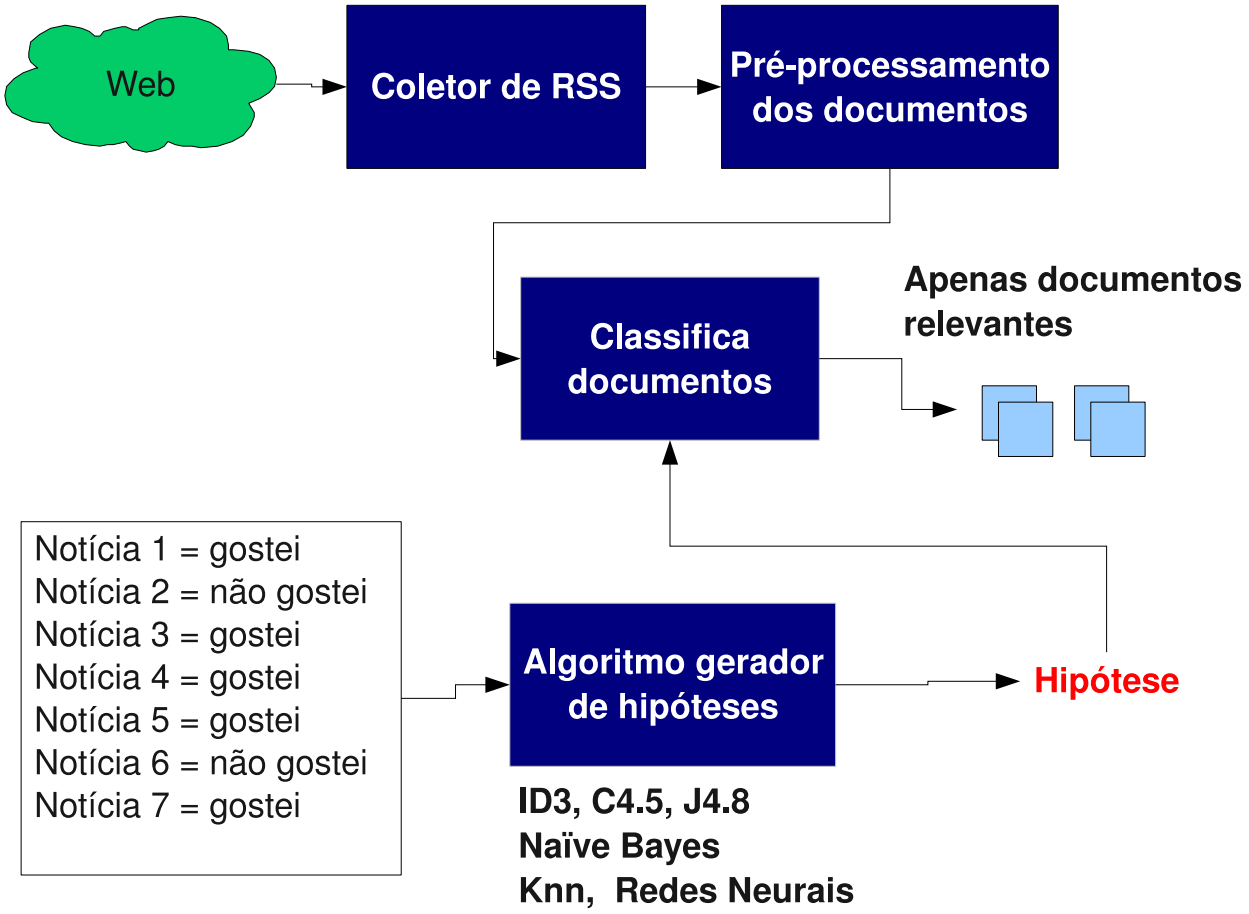
Conjunto de Exemplos - Atributo/Valor e Classe

Doc.	**apresent**	**form**	**tecnic**	**caracteriz**	$\cdots$	**Relevante**
d_1	0.33	0.33	0.33	0.33	$\cdots$	1
d_2	0	0.5	0.2	0.33	$\cdots$	0
d_3	1	0.6	0	0	$\cdots$	1
d_4	0.4	0.3	0.33	0.4	$\cdots$	1
d_5	1	0.4	0.1	0.1	$\cdots$	1
d_n	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

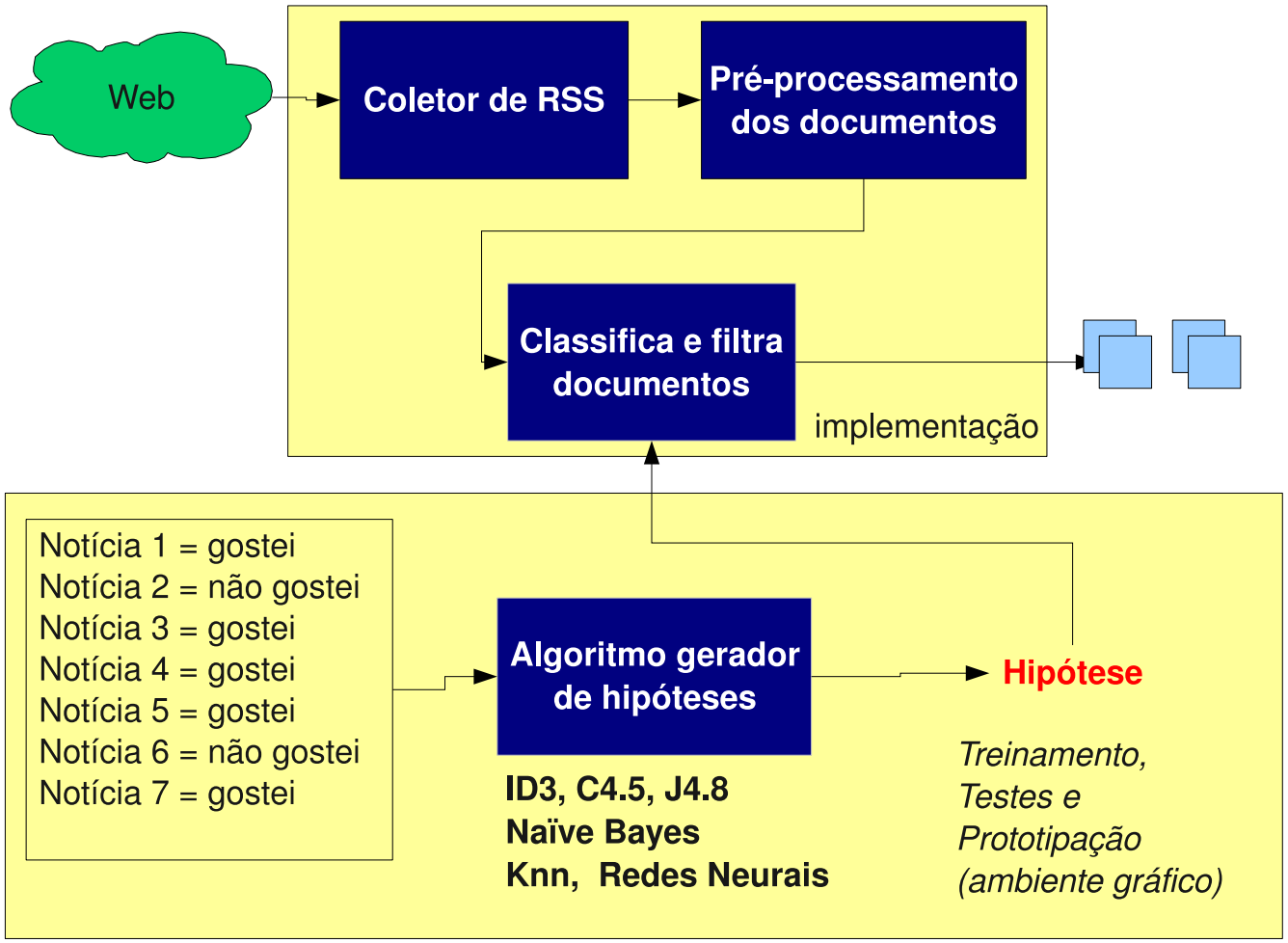
Qual é o problema?



Uma solução...


Web
Coletor de RSS
Pré-processamento dos documentos
Classifica documentos
Apenas documentos relevantes
Notícia 1 = gostei
Notícia 2 = não gostei
Notícia 3 = gostei
Notícia 4 = gostei
Notícia 5 = gostei
Notícia 6 = não gostei
Notícia 7 = gostei
Algoritmo gerador de hipóteses
Hipótese
ID3, C4.5, J4.8
Naïve Bayes
Knn, Redes Neurais

Processo de trabalho

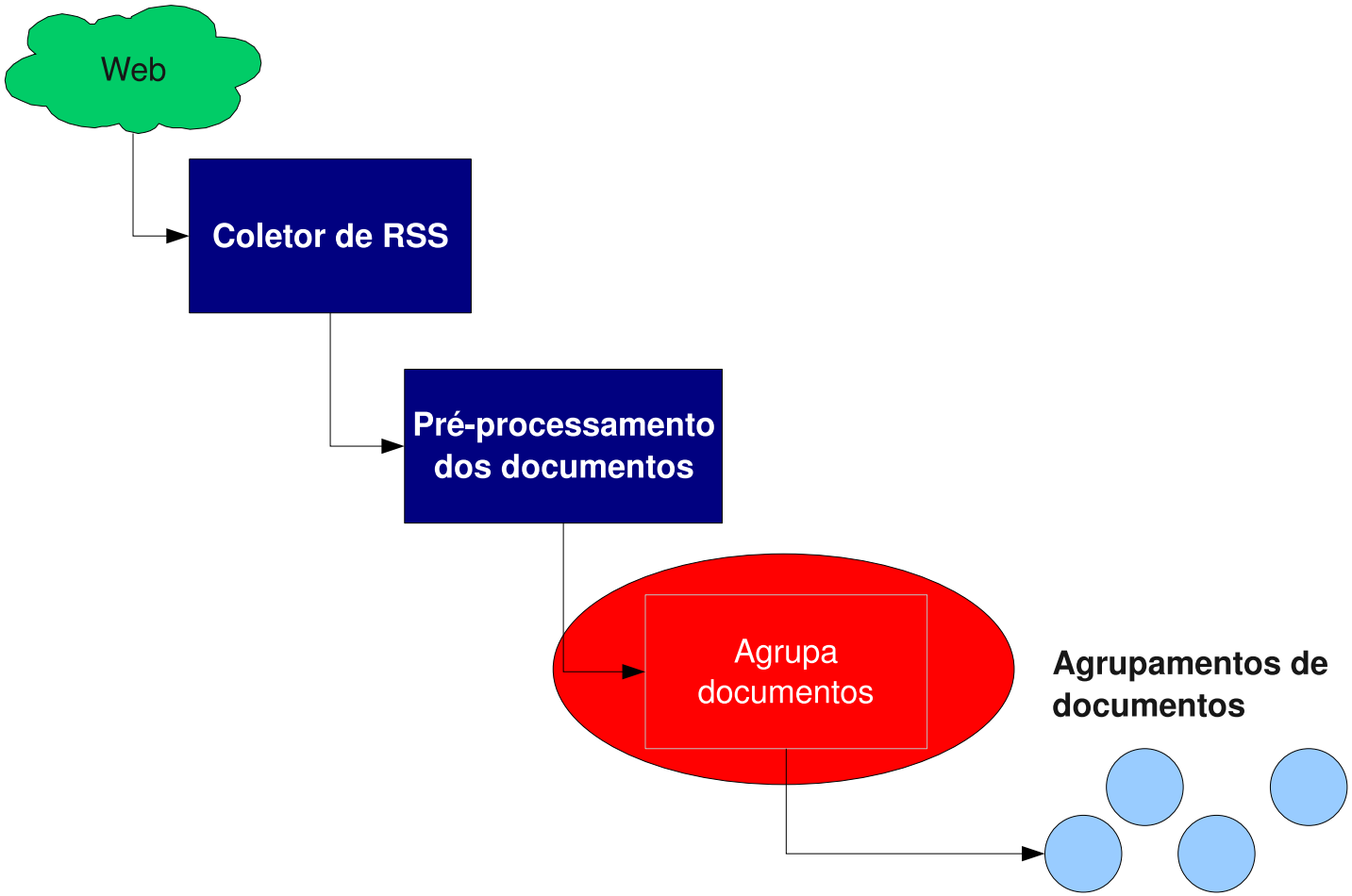


Agrupamento de documentos

Definições de Algoritmos de Agrupamento

- O objetivo dos algoritmos de agrupamento é colocar os objetos **similares** em um **mesmo grupo** e objetos **não similares** em **grupos diferentes**.
- Normalmente, objetos são descritos e agrupados usando um conjunto de **atributos e valores**.
- Não existe nenhuma informação sobre a classe ou categoria dos objetos.

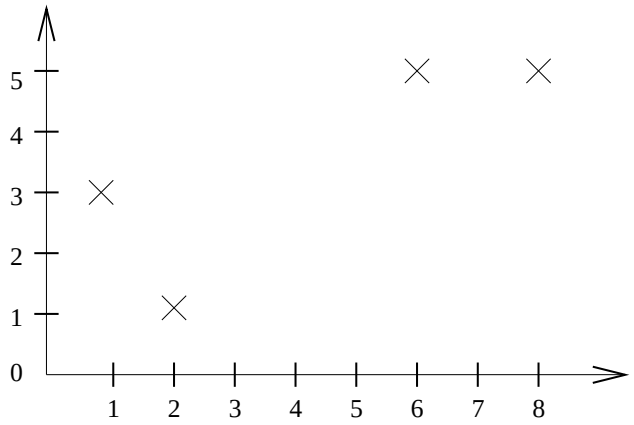
Componentes para uma solução...



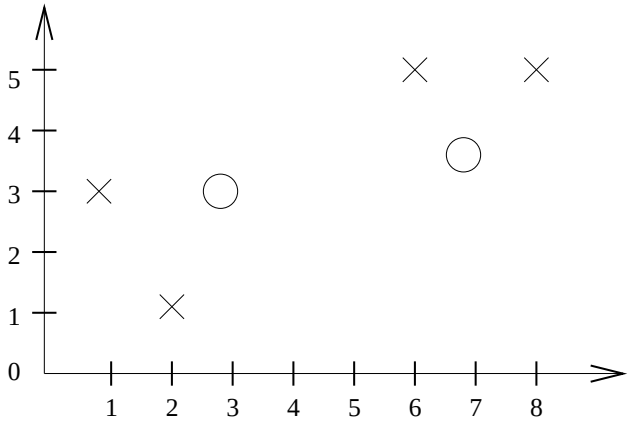
Algoritmos para Agrupamento - *K-means*

- **K** significa o número de agrupamentos (que deve ser informado à priori).
- Sequência de ações **iterativas**.
- A parada é baseada em algum critério de qualidade dos agrupamentos (por exemplo, similaridade média).

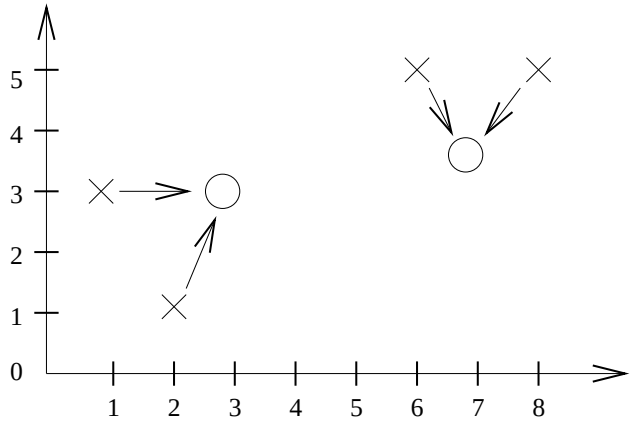
Algoritmo para Agrupamento - *K-means*



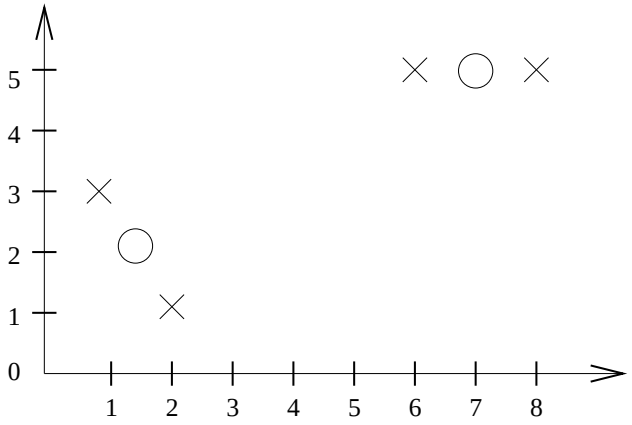
(1) Objetos que devem ser agrupados



(2) Sorteio dos pontos centrais dos agrupamentos



(3) Atribuição dos objetos aos agrupamentos



(4) Definição do centro do agrupamento

Algoritmos para agrupamento dos documentos - Weka

Execução do *K-means* no Weka[a].

[a] http://www.cs.waikato.ac.nz/ml/weka/

Algoritmo para agrupamento dos documentos - Resultados

```
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.1,0.2,0.3,0.1,0.8 está no cluster 1
A instância 0.3,0.4,0.5,0.8,0.9 está no cluster 0
A instância 0.3,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.3,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.8,0.7,0.8,0.8,0.8 está no cluster 0
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.6,0.5,0.6,0.6,0.6 está no cluster 0
A instância 0.6,0.5,0.6,0.6,0.6 está no cluster 0
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
A instância 0.2,0.8,0.8,0.7,0.9 está no cluster 0
A instância 0.1,0.1,0.1,0.1,0.1 está no cluster 1
```

Minerando o log de um servidor Web

Exemplo típico de log

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Pré-processamento do log: identificação de usuários

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

Pré-processamento do log: identificação das seções

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Matriz de transações

Pageviews

Sessions / users	A	B	C	D	E	F
user0	15	5	0	0	0	185
user1	0	0	32	4	0	0
user2	12	0	0	56	236	0
user3	9	47	0	0	0	134
user4	0	0	23	15	0	0
user5	17	0	0	157	69	0
user6	24	89	0	0	0	354
user7	0	0	78	27	0	0
user8	7	0	45	20	127	0
user9	0	38	57	0	0	15

Regras de Associação

- **Caso do supermercado** (fralda $\rightarrow$ cerveja)
- Quem acessa a página sobre futebol também acessa a página de volei em **90%** dos casos (futebol $\rightarrow$ volei).
- Quem acessa a página de ofertas e a página de material de construção também finaliza a compra em **83%** dos casos (ofertas $\wedge$ material_construção $\rightarrow$ compra)

Considerações Finais

Considerações Finais

- Foram vistos: problemas de classificação, agrupamento e análise de log. Tem muito mais de onde vieram estes...
- **Atenção para o processo!** Pré-processamento, criação dos modelos, avaliação e aplicação.
- Alguns algoritmos para mineração de informação são pesados. Talvez, parte da solução esteja na adoção de *cloud computing*.
- **Muitos dados... Muitas oportunidades...**

Outros exemplos

- Wiki2Group[a] - 2010
- Sistema Folkaliza[b] - 2009
- Sistema opSys[c] - 2008
- Sistema FaroFino - 2005
- Mais informações podem ser encontradas em **http://fbarth.net.br** e **http://fbarth.net.br/projetos/riInteligente.html**

[a] http://trac.fbarth.net.br/wikiAnalysis
[b] http://www.jessicacintra.com.br/jeh/folkaliza/Home/Default.aspx
[c] http://www.opsys.com.br

Referências

References

[Data, data everywhere. A special report on managing information 20
DATA, data everywhere. A special report on managing information. **The Economist**, p. 1–16, February 2010.

[Liu 2009]LIU, B. **Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)**. 1st ed. 2007. corr. 2nd printing. ed. Springer, 2009. Hardcover. ISBN 3540378812. Disponível em: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/3540378812>.

[Mitchell 1997]MITCHELL, T. M. **Machine Learning**. [S.l.]: McGraw-Hill, 1997.

[Quinlan 1988]QUINLAN, J. R. Knowledge acquisition for knowledge-based systems. In: _____. [S.l.]: Academic Press, 1988. cap. Simplifying Decision Trees.

[Russel e Norvig 2003]RUSSEL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. 2. ed. [S.l.]: Prentice-Hall, 2003. ISBN 0-13-790395-2.

[Witten e Frank 2005]WITTEN, I. H.; FRANK, E. **Data Mining: Practical Machine Learning Tools and Techniques**. Second. [S.l.]: Elsevier, 2005.