Uma Introdução à Mineração de Informações na era do Big Data

Fabrício J. Barth

VAGAS Tecnologia e Faculdades BandTec

Setembro de 2012

Palestrante

- **Fabrício J. Barth**. Formado em Ciência da Computação pela **FURB**. Mestrado e Doutorado em Engenharia da Computação pela USP.
- Recuperação e mineração de informações para domínios de investigação; Identificação de temas emergentes e especialistas em bases de patentes e artigos; Desenvolvimento e gestão de uma plataforma de busca georeferenciada (www.apontador.com.br).
- Data Scientist na VAGAS Tecnologia (www.vagas.com.br). Professor da Faculdade BandTec (www.bandtec.com.br).

Objetivo

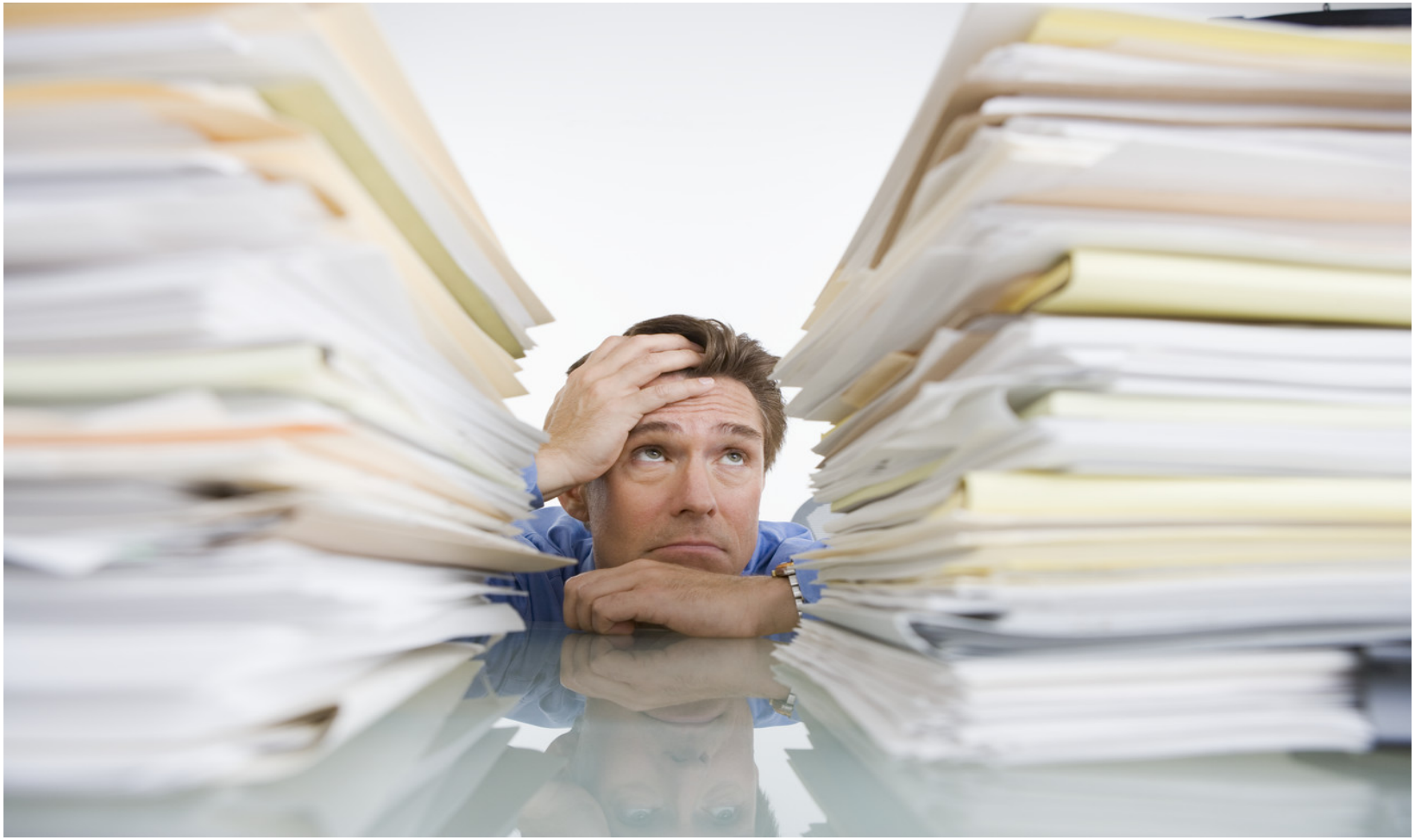
Apresentar a importância do tema, os conceitos relacionados e alguns exemplos de aplicações.

Sumário

- Importância do Tema
- Manipulando dados estruturados
- Manipulando dados não-estruturados (textos)
  - ⋆ Agrupamento
  - ⋆ Classificação
- Web Data Mining
- Considerações Finais
- Referências

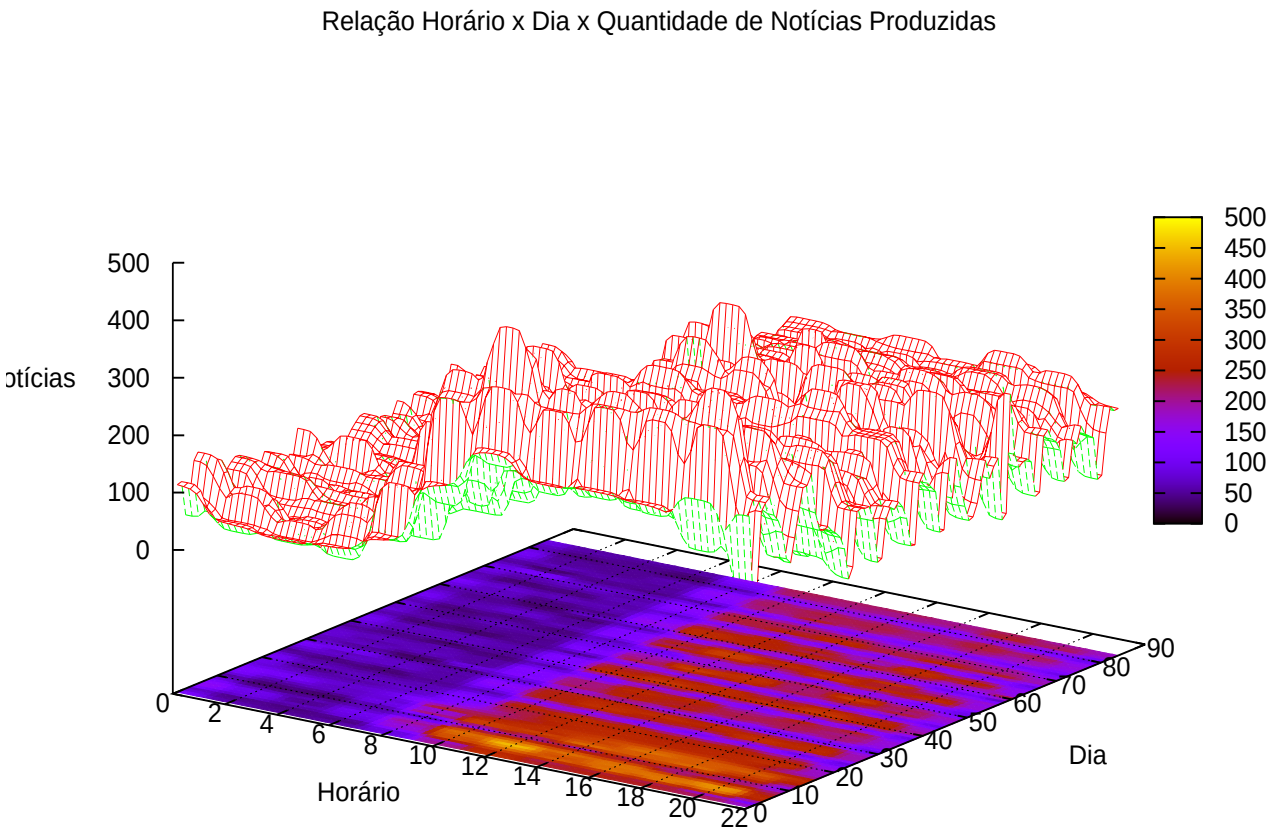
Importância do Tema

Problema



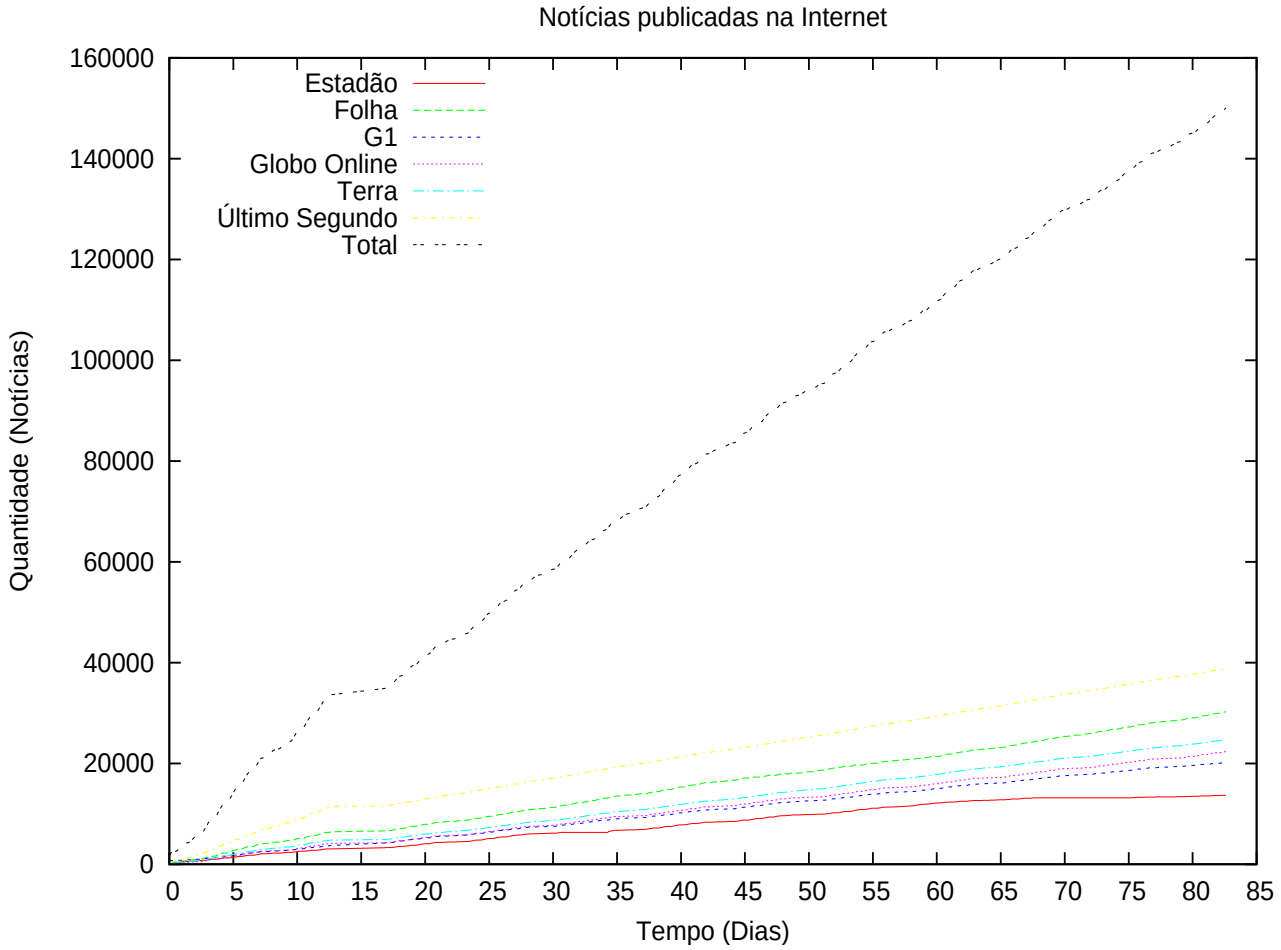
http://investingcaffeine.com/2010/01/07/tmi-the-age-of-information-overload/

Alguns dados...



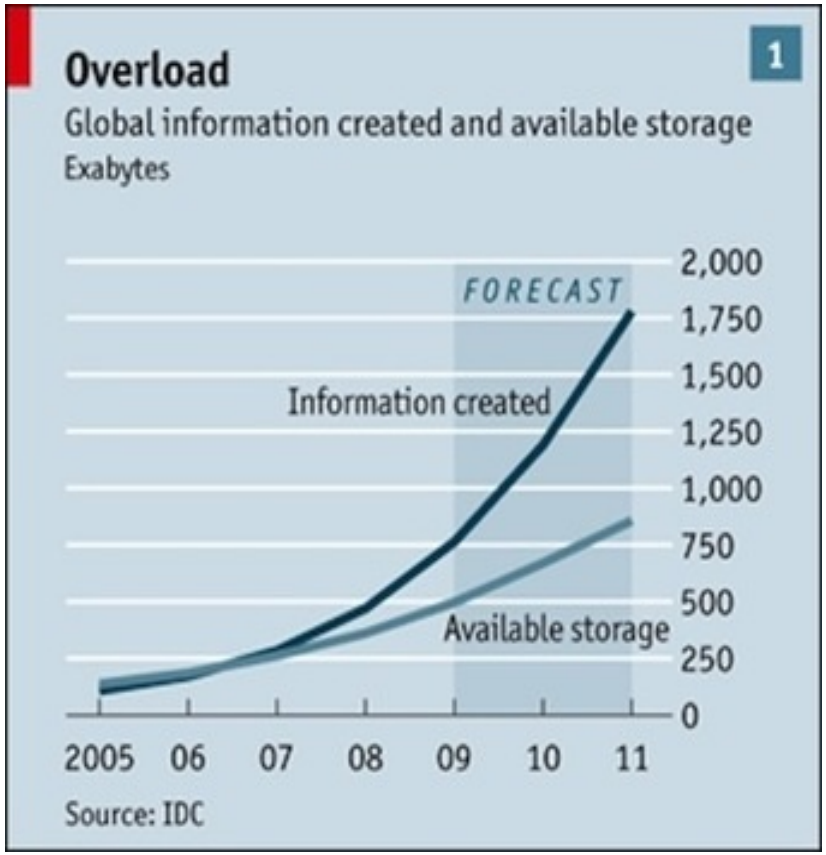
Quantidade de notícias publicadas na Web por apenas seis veículos de notícias ($D_0 = 17/07/2007$)

Mais dados...



$D_0 = 17/07/2007$

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..."[1]

Mais números

- A380: Heathrow $\rightarrow$ JFK: 640 TBs de log
- Twitter: 12+ TBs of tweet every day
- Facebook: 25+ TBs of log data every day
- Sistemas baseados em RFID
- Smartphones com GPS, acelerómetro, ...

http://www.ibmbigdatahub.com/
Mitchell. Mining our reality. Science. 2009

Por que minerar informações?

- Explicitar conhecimento médico a partir de registros médicos.
- Identificar comportamento anômalo (i.e., fraudes, falhas)
- 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.
- Recomendação de livros, filmes, restaurantes e **empregos**.

Explicitar conhecimento médico a partir de registros médicos

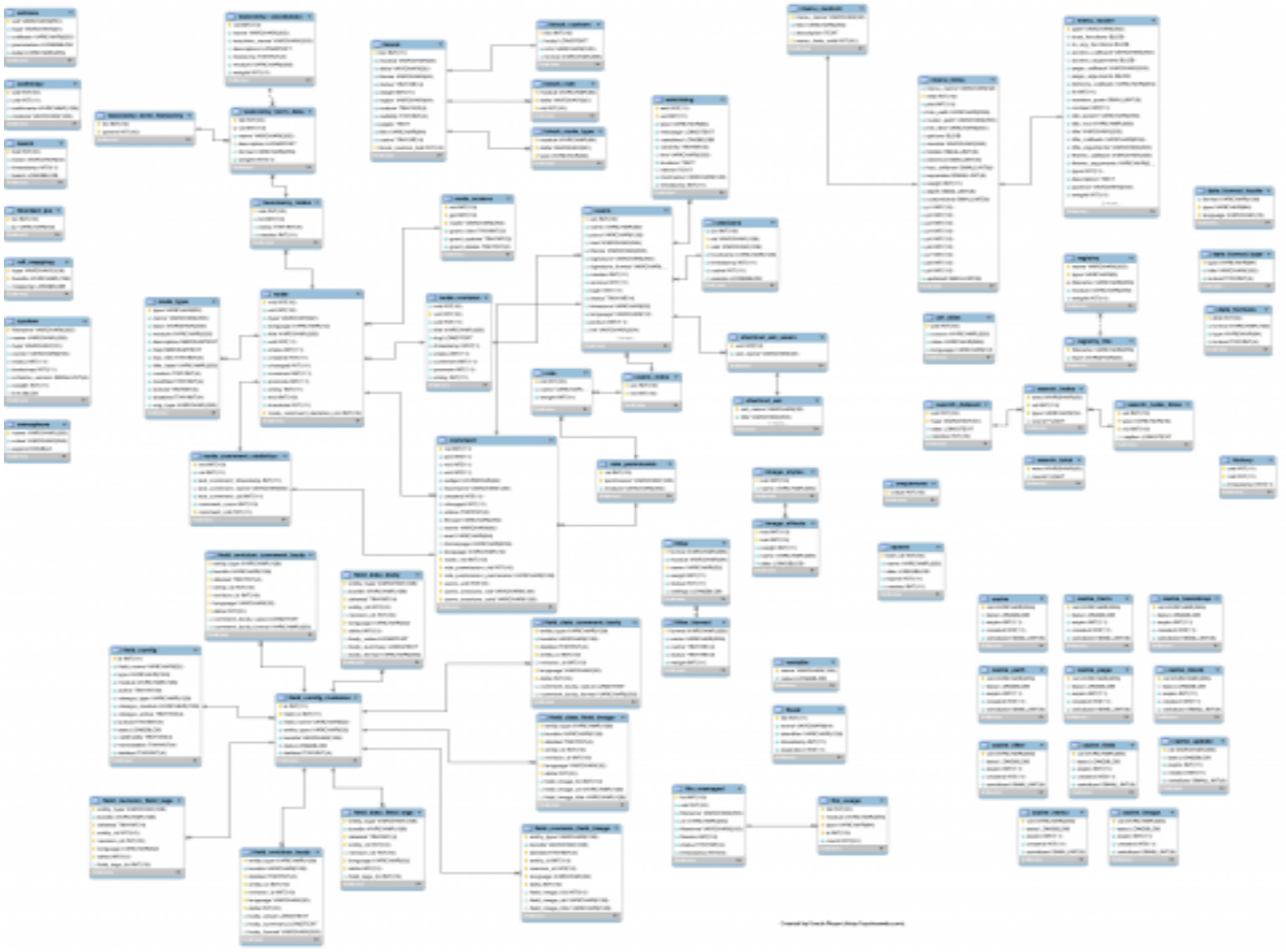
Diagnóstico para o uso de lentes de contato

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.

Banco de dados do ambiente de produção



Por onde começar?

Responder as seguintes perguntas:

- Que objetos/atributos 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”

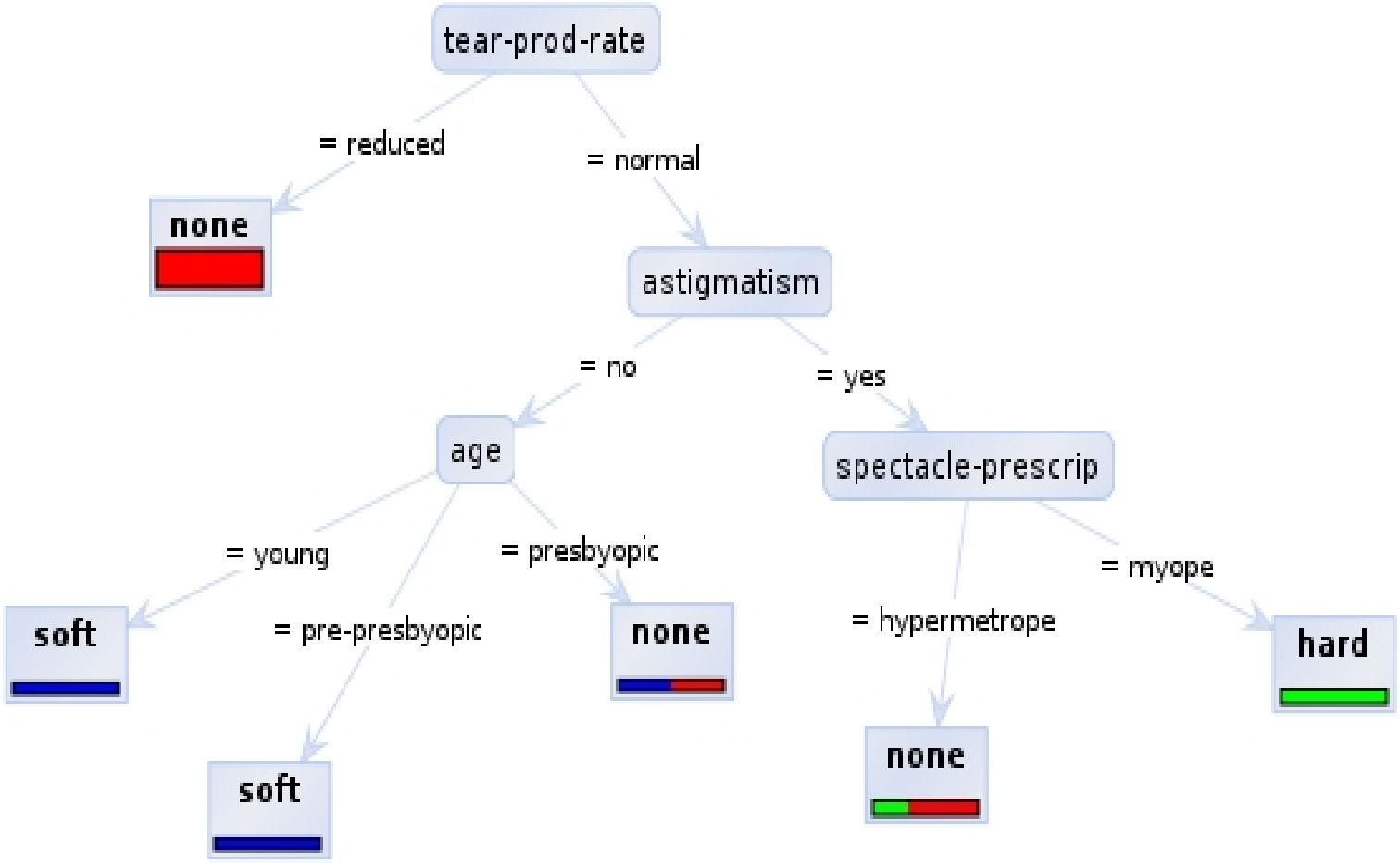
- O que foi apresentado nos slides anteriores pode ser considerado como conhecimento? **Não**
- Pode ser apresentado como uma informação que consegue explicar a tomada de decisão dos especialistas? **Não**
- **O que fazer?**

Extração de "conhecimento"

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

(mostrar exemplo no RapidMiner - www.rapid-i.com)

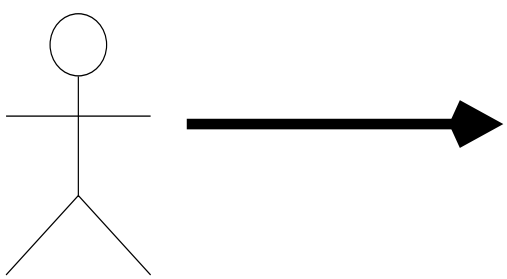
Árvore de decisão



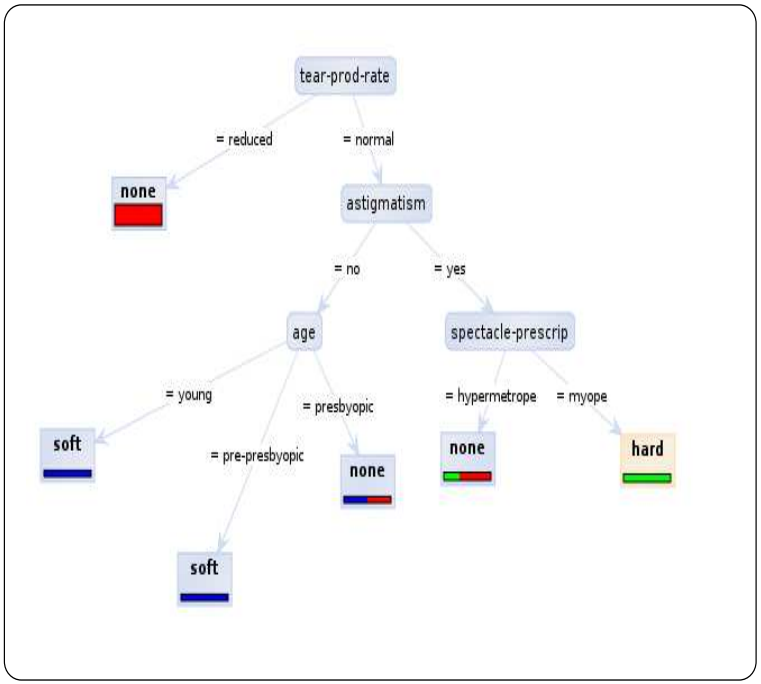
Algoritmos Indutores de Árvores de Decisão

- **Que algoritmo utilizar para gerar hipóteses na forma de árvores de decisão?**
- ID3, C4.5[7]: 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)

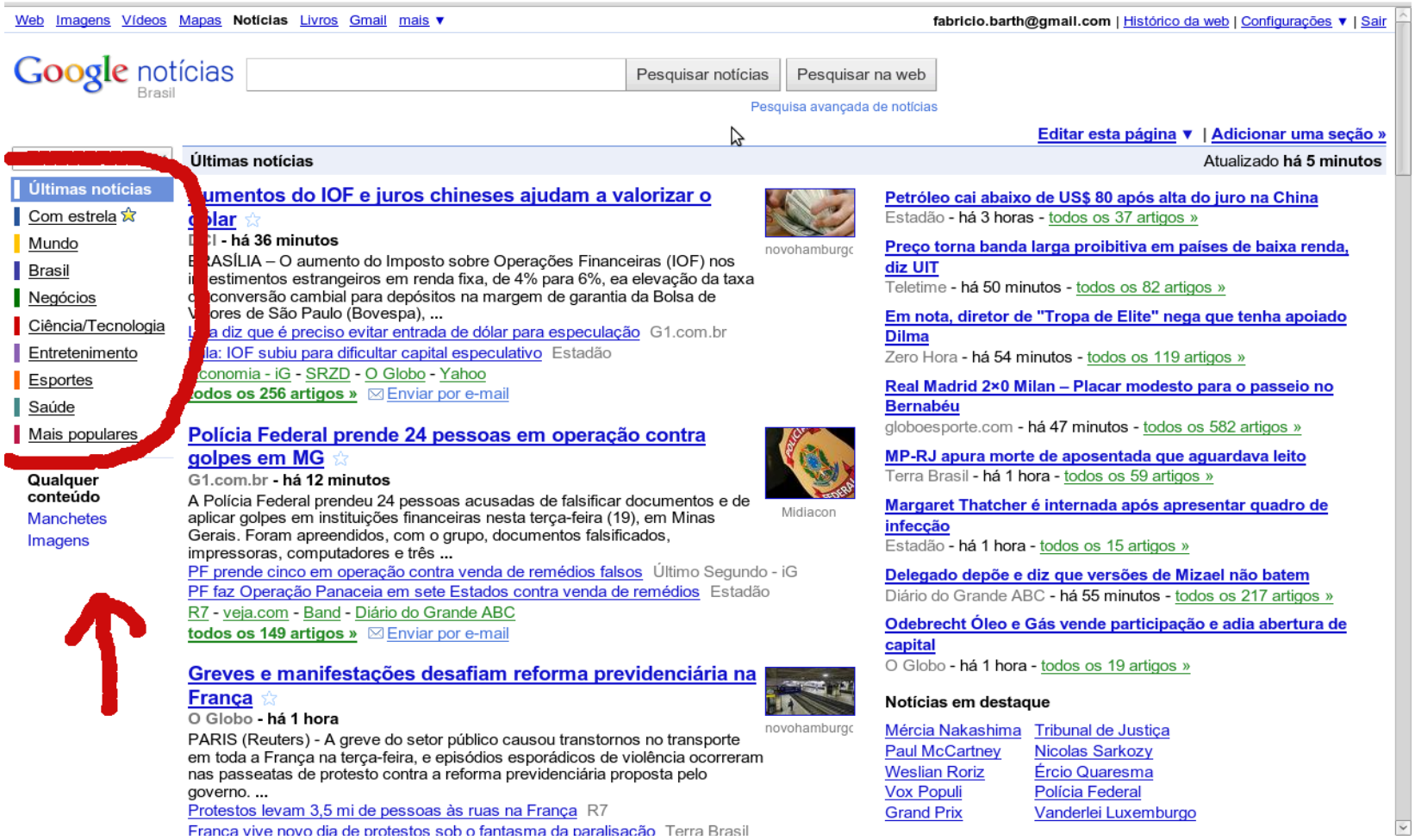


Organizar documentos

O que fazer com grandes quantidades de documentos?

- Notícias, patentes, artigos, mensagens de twitter, questões abertas de um questionário de pesquisa, ...
- **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



Etapas

- Pré-processamento dos dados.
- Modelagem (supervisionada ou não supervisionada).
- Avaliação do modelo.
- Utilização

Pré-processamento dos dados

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. ...

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	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

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

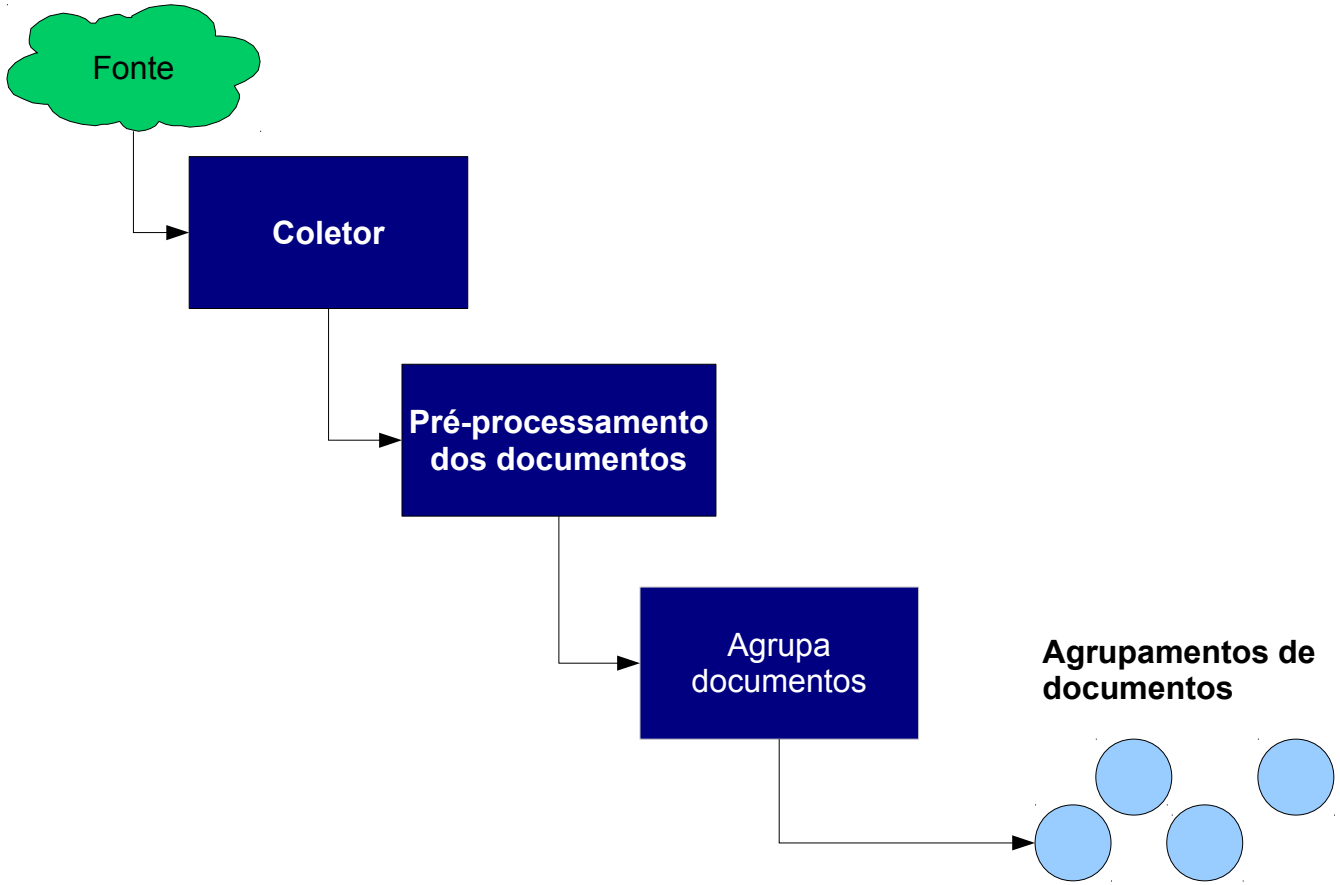
- Já conhecemos os atributos.
- E os valores?
  - ⋆ **Booleana** - se a palavra aparece ou não no documento (1 ou 0)
  - ⋆ **Por freqüência do termo** - a freqüência com que a palavra aparece no documento (normalizada ou não)
  - ⋆ **Ponderação tf-idf** - o peso é proporcional ao número de ocorrências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

Análise do conteúdo publicado no Twitter através de **algoritmos de agrupamento**

Componentes para uma solução...

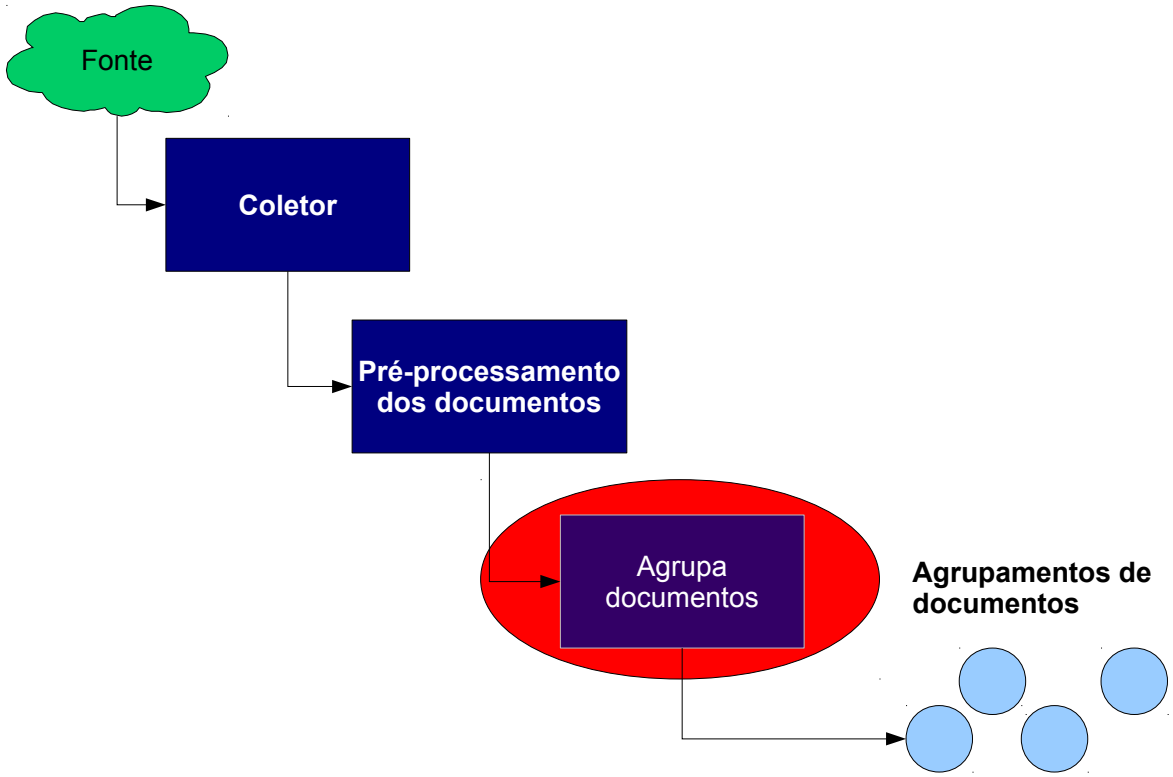


Coletando dados do twitter

```
$ curl -o aboutBrasil.txt https://stream.twitter.com/1/statuses/
filter.json?track=brasil -u user:passwd

% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 4549k    0 4549k    0     0   5986      0 --:--:--  0:12:58 --:--:--  6226

$ wc -l aboutBrasil.txt
1501 aboutBrasil.txt

$ date
Ter 28 Ago 2012 23:51:10 BRT
```

```ruby
#
# forma de uso: ruby twitter.rb > mensagens.csv
#

require 'rubygems'
require 'json'

content = File.open('aboutBrasil.txt')
content.each do |line|
        puts JSON.parse(line)['text']
end
```

Componentes para uma solução...



Wiki2Group - http://trac.fbarth.net.br/wikiAnalysis

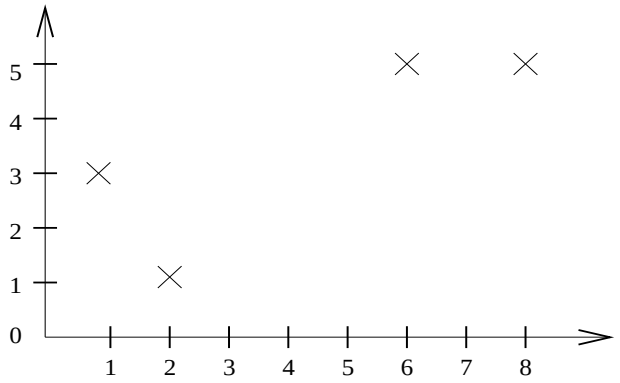
Definições de Algoritmos de Agrupamento

- 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.

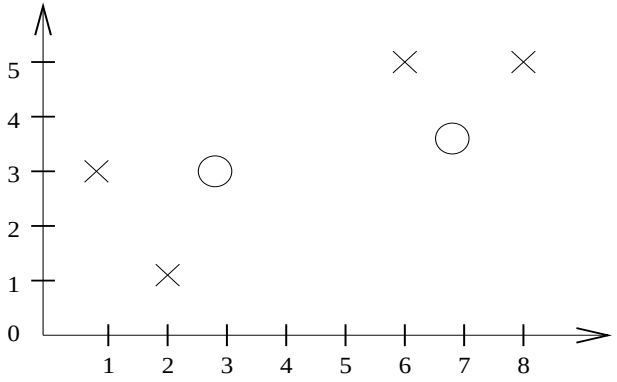
Algoritmos para Agrupamento - *K-means*

- **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).

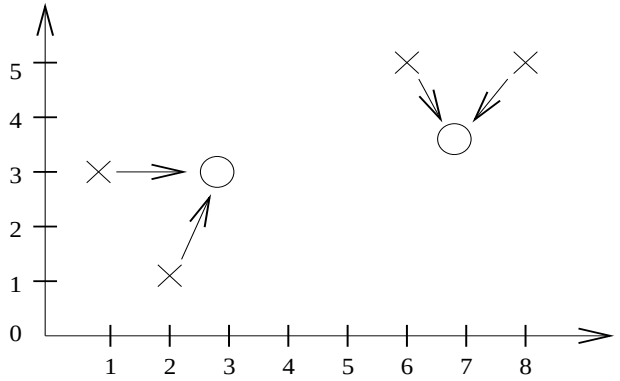
Algoritmo para Agrupamento - *K-means*



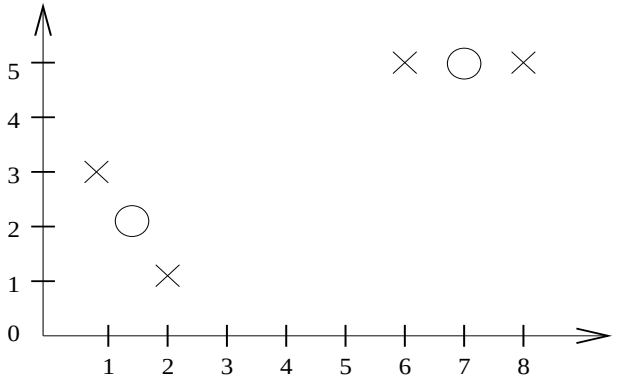
(1) Objetos que devem ser agrupados



(2) Sorteio dos pontos centrais dos agrupamentos



(3) Atribuição dos objetos aos agrupamentos



(4) Definição do centro do agrupamento

Algoritmo para agrupamento dos twittes - Resultados

Execução do processo no RapidMiner

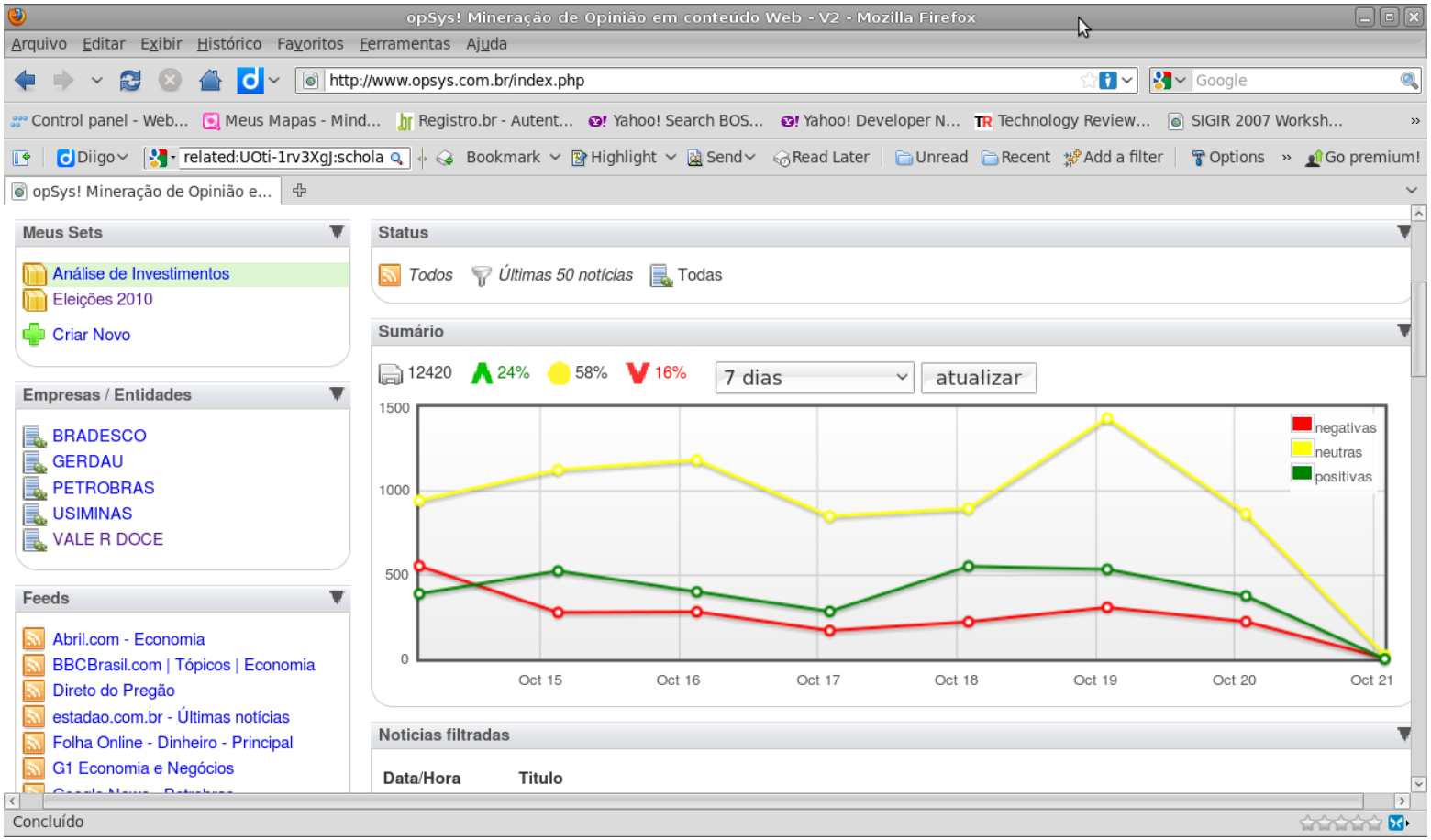
Análise dos agrupamentos (*clusters*)

```
RT @TWITTEI: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'tharara' '
RT @TWITTEI: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'tharara' '
RT @PiadaDePobre: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'thara
RT @soucrack: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'tharara'
RT @PiadaDePobre: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'thara
'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'tharara' 'parapapa' ach
RT @PiadaDePobre: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'thara
'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'tharara' 'parapapa' ach
RT @PiadaDePobre: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'thara
RT @PiadaDePobre: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'thara
RT @TWITTEI: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'tharara' '
RT @TWITTEI: 'tcherere tche tche' 'bara bere' 'tchu tcha' 'lelele' 'paragada' 'tharara' '
```


china
2012
setembro
video
awards
2021
liam
mim
bitch
ás
aniversario
sonha
planeja
ideb
6.0
gaga
@fofocas1d
lady
oglobo
show
to
futuroreprovado
brasil
@curiosidadesfas
2020
britairlandesa
tripulado
pron
music
ainda
ai
mtv
vôo
aqui
caros
21:00
passaram
ter
niver
dia
@choosekatybr
lua

Classificação de documentos

Análise de Sentimento em mensagens no Twitter



Teor das mensagens sobre a empresa Vale nos últimos sete dias. http://www.opsys.com.br/ - [4]

Conjunto de Exemplos Rotulados

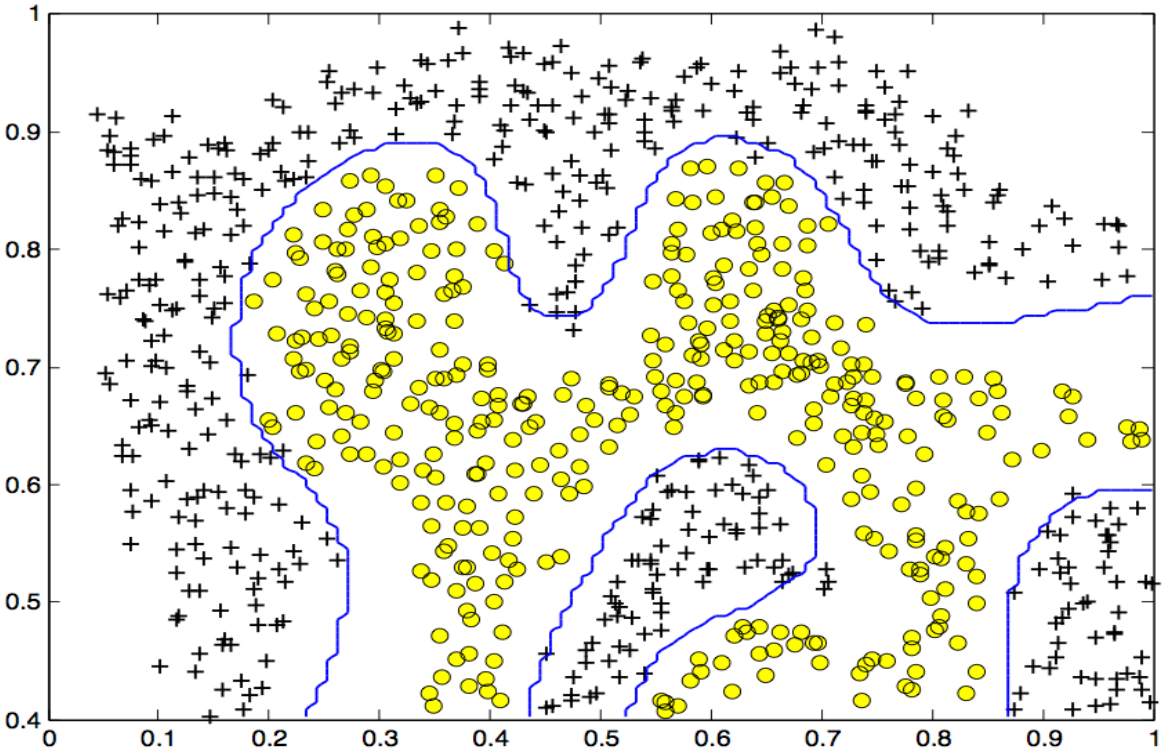
Doc.	Mensagem	Classe
d_1	A empresa X é uma empresa muito séria	Positivo
d_2	O produto Y é uma porcaria	Negativo
d_3	Gostei muito da palestra de fulano	Positivo
d_4	Aquela praia é muito bonita	Positivo
d_5	Gostei daquele restaurante	Positivo
d_n		...

- Rotular manualmente
- Utilizar emoticons :) :(

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	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Classificando objetos



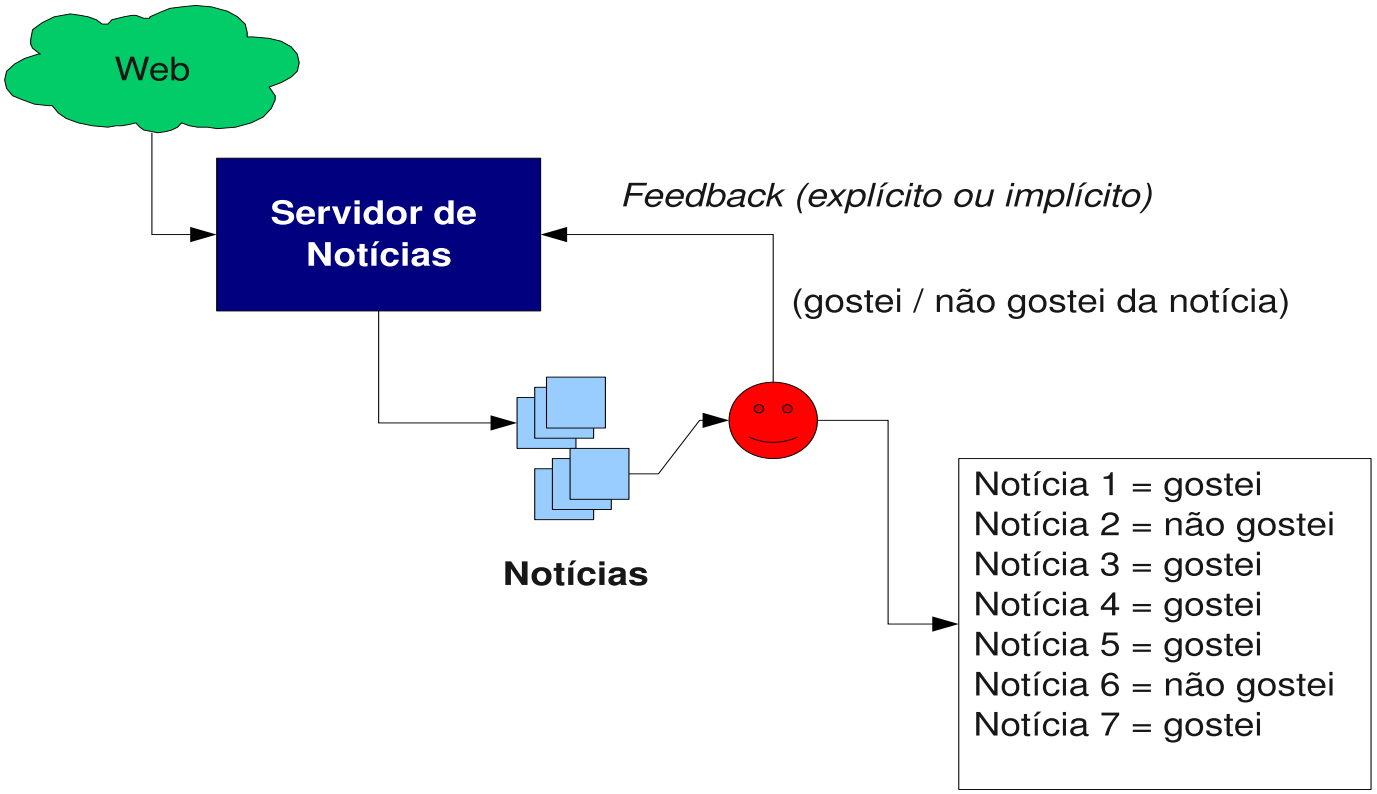
- Naïve Bayes [6]
- Support Vector Machines (SVM) [2]

Utilização

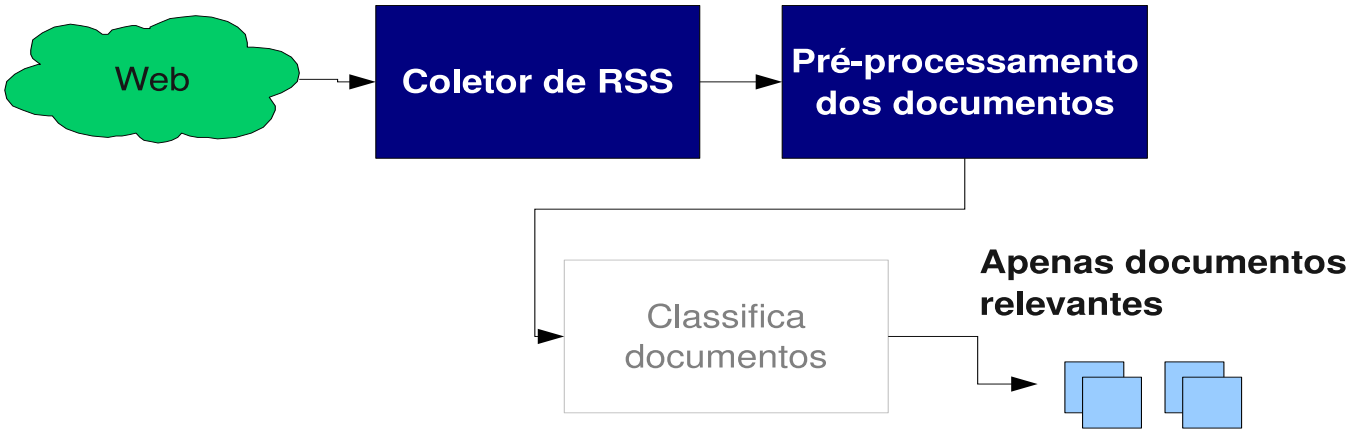
Utilizando o modelo criado é possível inferir se novos *twittes* possuem sentimento positivo ou negativo.

```
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
```

Outro Exemplo: Classificação e Filtragem de Notícias

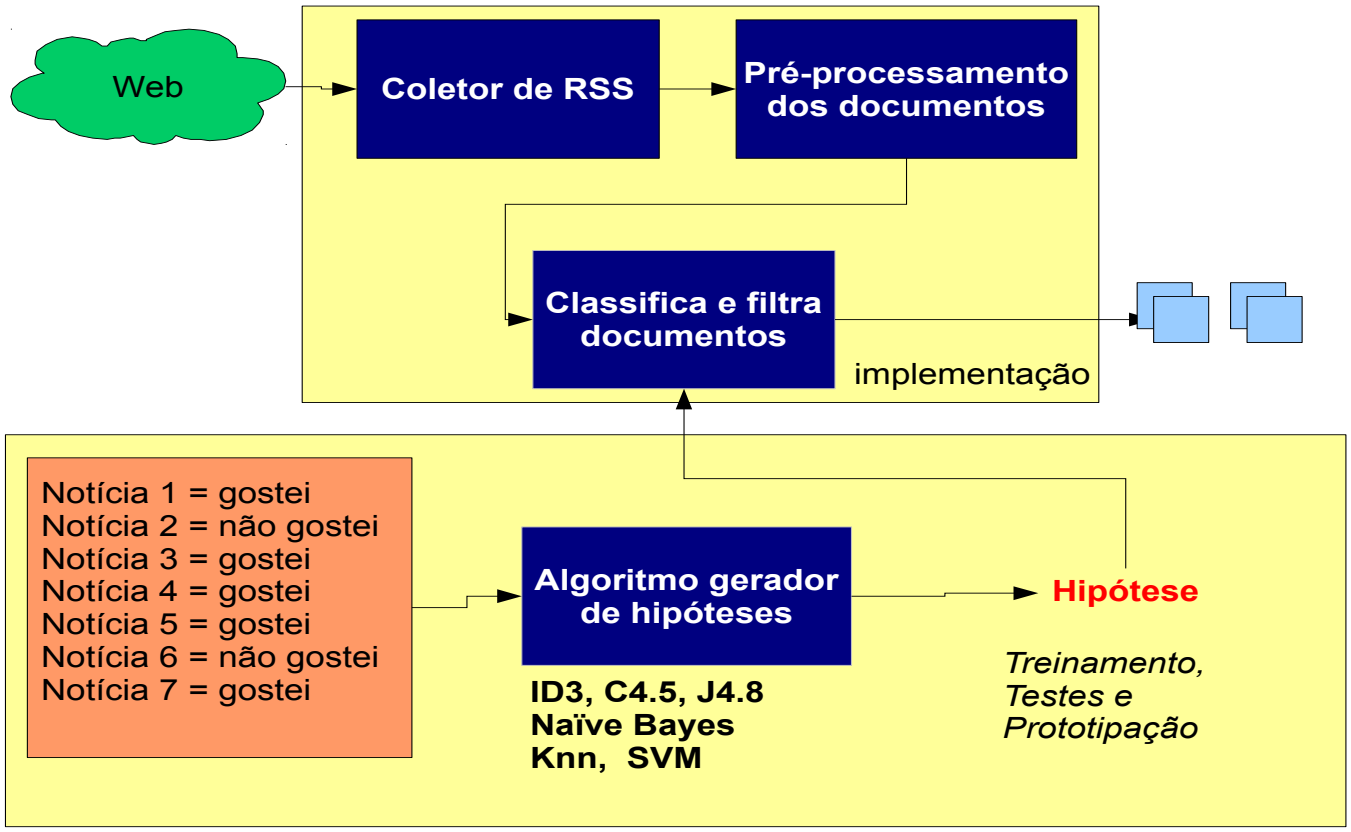


Qual é o problema?



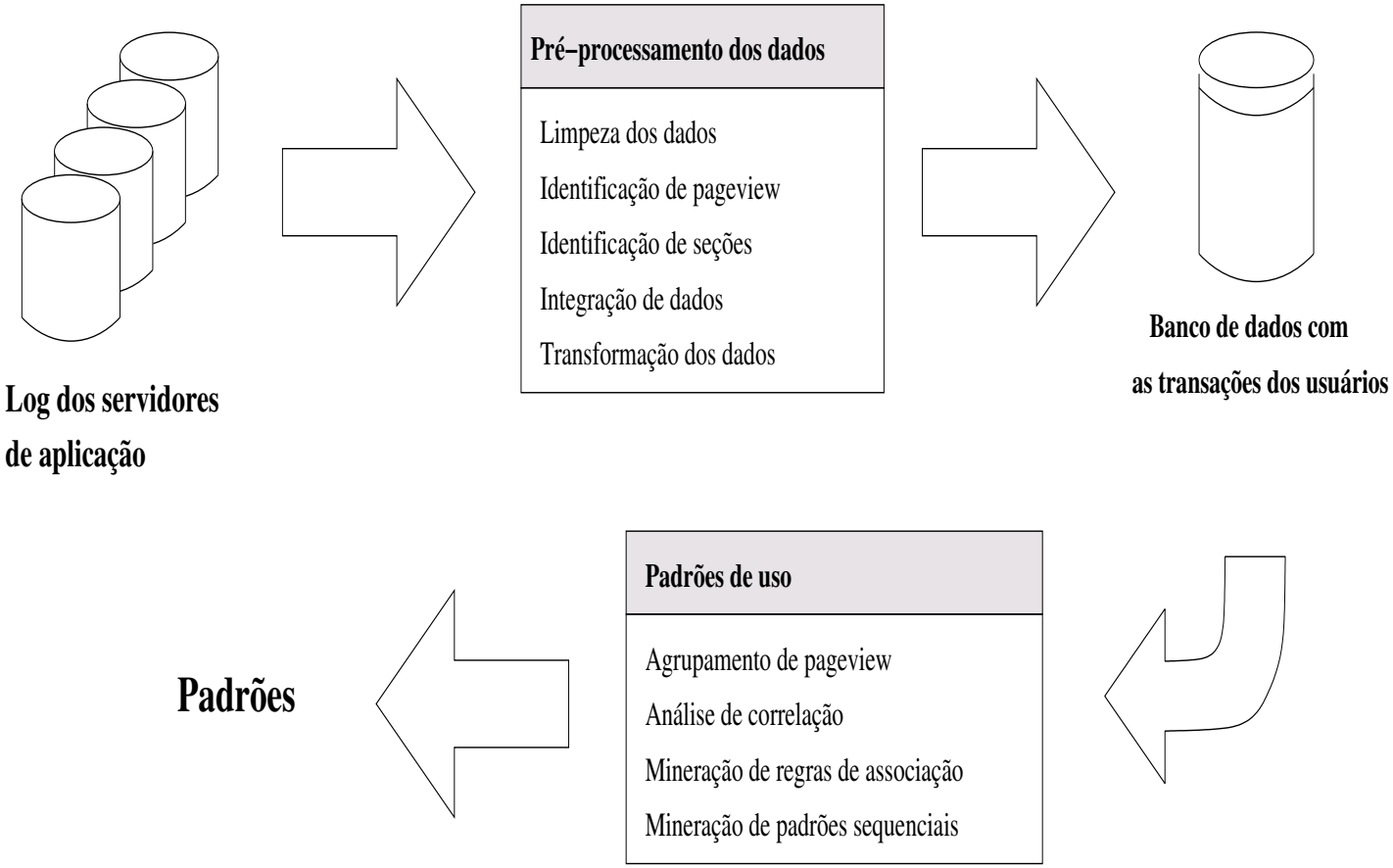
http://fbarth.net.br/projetos/riInteligente.html - Sistema FaroFino [5]

Processo de trabalho



Minerando o log de um servidor Web

Processo de mineração de padrões na 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

Matriz de transações com meta-informações sobre as páginas

usuário	$categoria_1$	$categoria_2$	$categoria_3$	$\cdots$	$categoria_m$
$user_1$	0	2	0	$\cdots$	1
$user_2$	1	1	0	$\cdots$	0
$user_3$	2	0	1	$\cdots$	0
$user_4$	0	1	0	$\cdots$	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$user_n$	1	1	0	$\cdots$	1

- Cada página pode pertencer a uma categoria (i.e., tipo de livro, tipo de estabelecimento comercial)
- Cada página pode estar associada a uma cidade (i.e., um estabelecimento, uma vaga de emprego)

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

Processo

1. Qual é a pergunta?
2. Aquisição e pré-processamento dos dados.
3. Análise Descritiva.
4. Modelagem: construção do modelo descritivo ou preditivo.
5. Avaliação do modelo.
6. Entrega: relatórios estáticos, dinâmicos, sistemas ou funcionalidade de sistemas.

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.
- *Mahout, Hadoop, Carrot2*. Antes de instalar ferramentas para a mineração das informações, tente entender os seus dados e os seus problemas! Faça uma análise descritiva dos dados.
- **Muitos dados... Muitas oportunidades...**

Obrigado!

http://fbarth.net.br

http://fbarth.net.br/materiais/palestras.html

fabricio.barth@gmail.com

Referências

References

[1] Data, data everywhere. a special report on managing information. *The Economist*, pages 1–16, February 2010.

[2] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 1st ed. 2007. corr. 2nd printing edition, January 2009.

[4] Thomas Jefferson P. Lopes, Gabriel Koji Lemos Hiratani, Fabrício J. Barth, Orlando Rodrigues, Jr., and Juliana Maraccini Pinto. Mineração de opiniões aplicada à análise de investimentos. In *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, WebMedia '08, pages 117–120, New York, NY, USA, 2008. ACM.

[5] João Carlos Medau, Maria Cristina Belderrain, and Fabrício J. Barth. Reordenação de resultados de busca na web conforme critério de relevância definido pelo usuário. In *Anais do XI Simpósio Brasileiro de Sistemas Multimídia e Web - WebMedia*, pages 220–222, 2005.

[6] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[7] J. R. Quinlan. *Knowledge Acquisition for Knowledge-Based Systems*, chapter Simplifying Decision Trees. Academic Press, 1988.

[8] Stuart J. Russel and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice-Hall, 2 edition, 2003.

[9] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, second edition, 2005.