Data, Text and Web Mining

Fabrício J. Barth

TerraForum Consultores

Junho de 2010

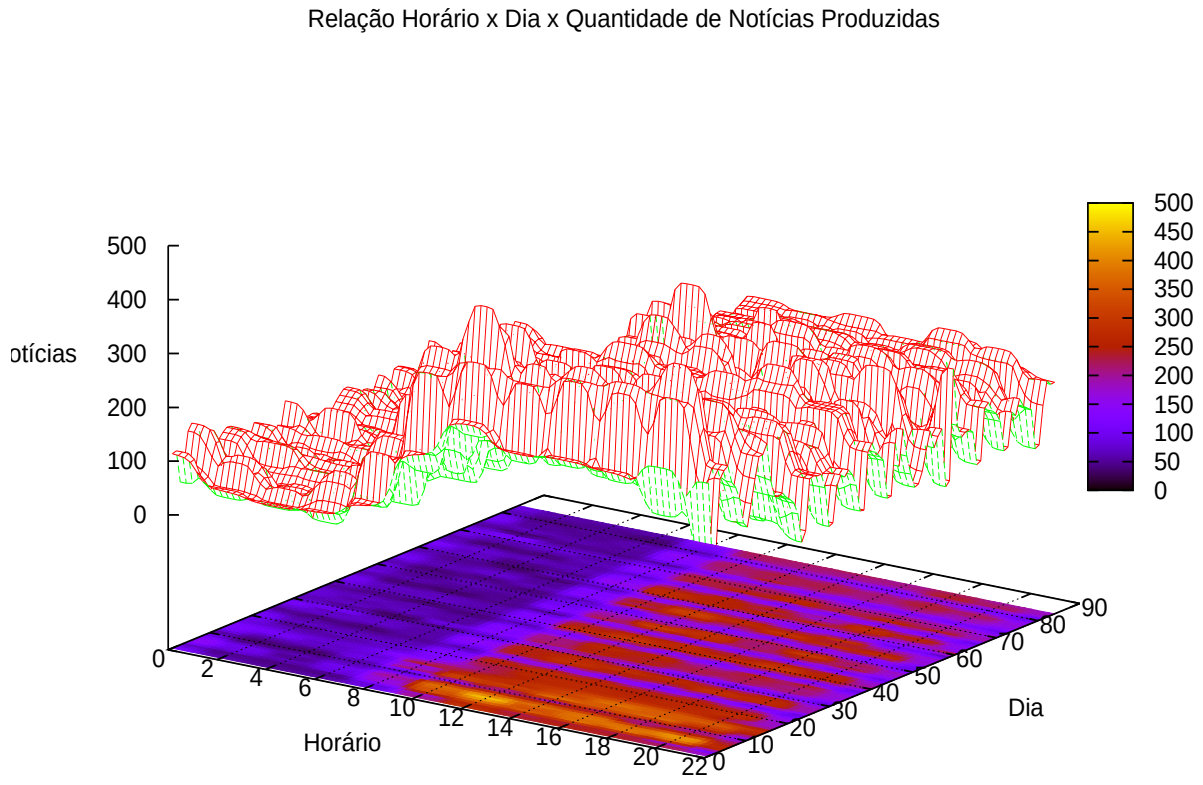
Objetivo

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

Importância do Tema

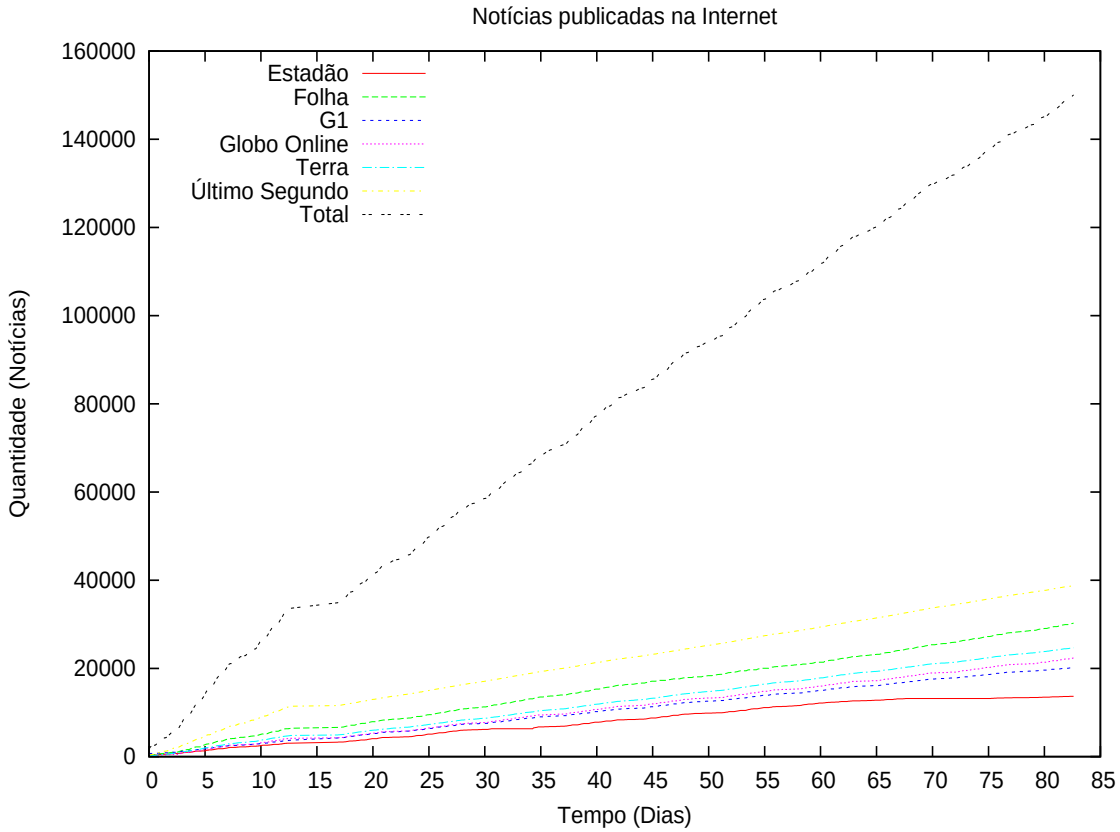
Problema

Alguns dados...



Quantidade de notícias publicadas na Web por apenas seis veículos de notícias.

Alguns dados...



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 grupos de usuários com perfil similar em ambientes de escrita colaborativa.

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

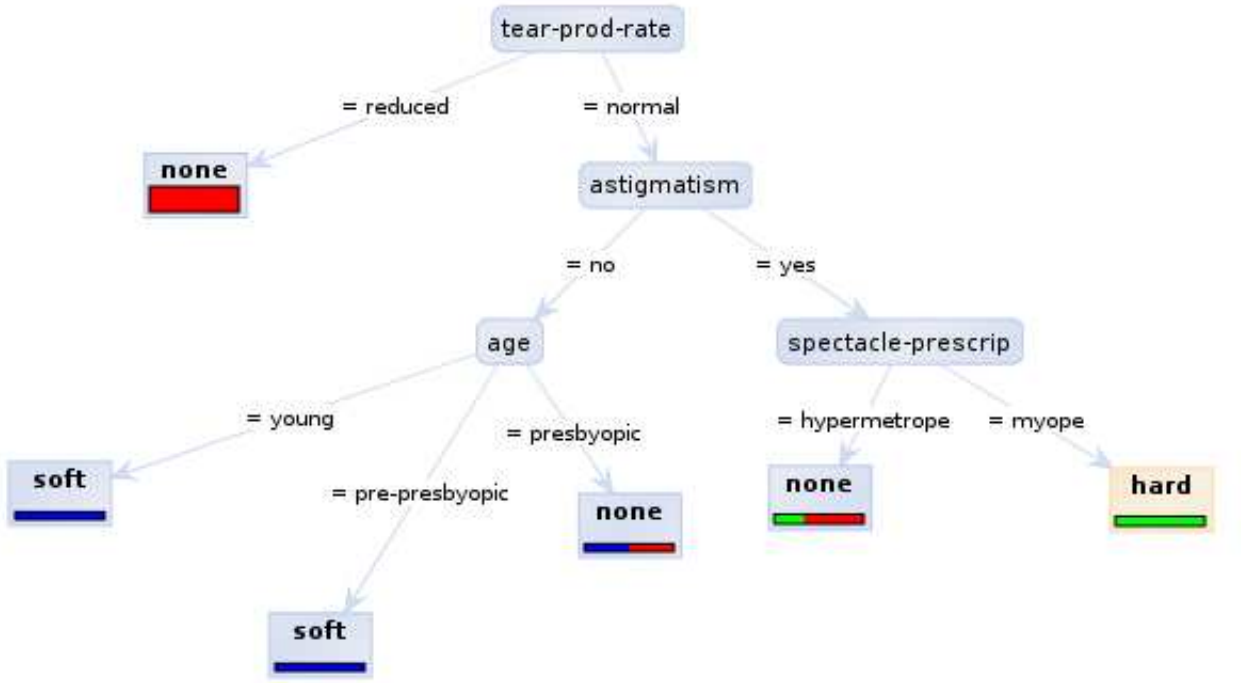
Extração de "conhecimento"

- 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

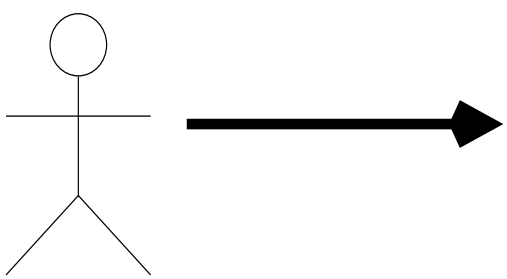


- Cada nodo interno testa um atributo.
- Cada aresta corresponde a um valor de atributo.
- Cada nodo folha retorna uma classificação.

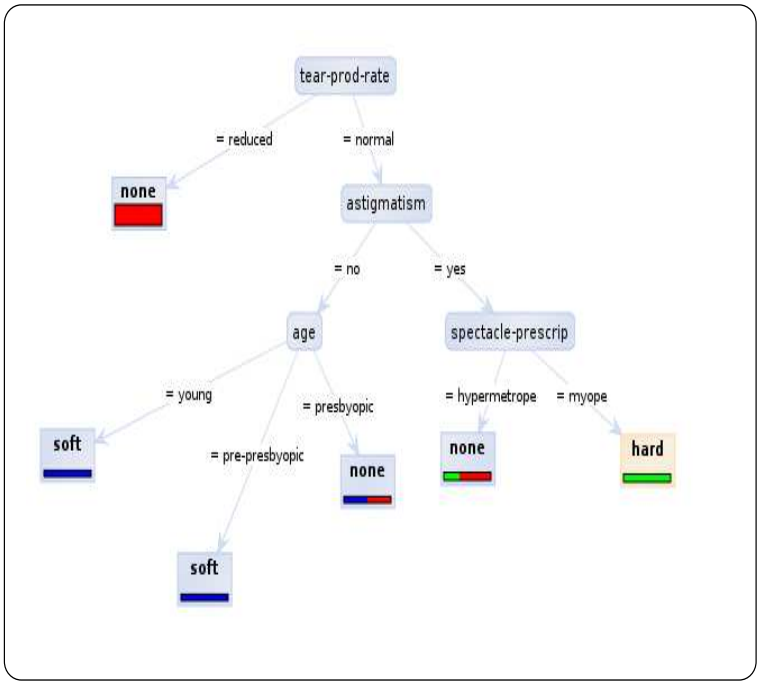
Algoritmos Indutores de Árvores de Decisão

- **Que algoritmo utilizar para gerar hipóteses na forma de árvores de decisão?**
- ID3, C4.5[2]: 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...
- **Para tirar proveito desta informação é necessário organizá-la de alguma forma**:
  - ⋆ Agrupamento de notícias, patentes, artigos.
  - ⋆ Classificação, Recomendação e Filtragem de Notícias.

Exemplo de agrupamento

E quando não é possível fazer manualmente?

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.

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.

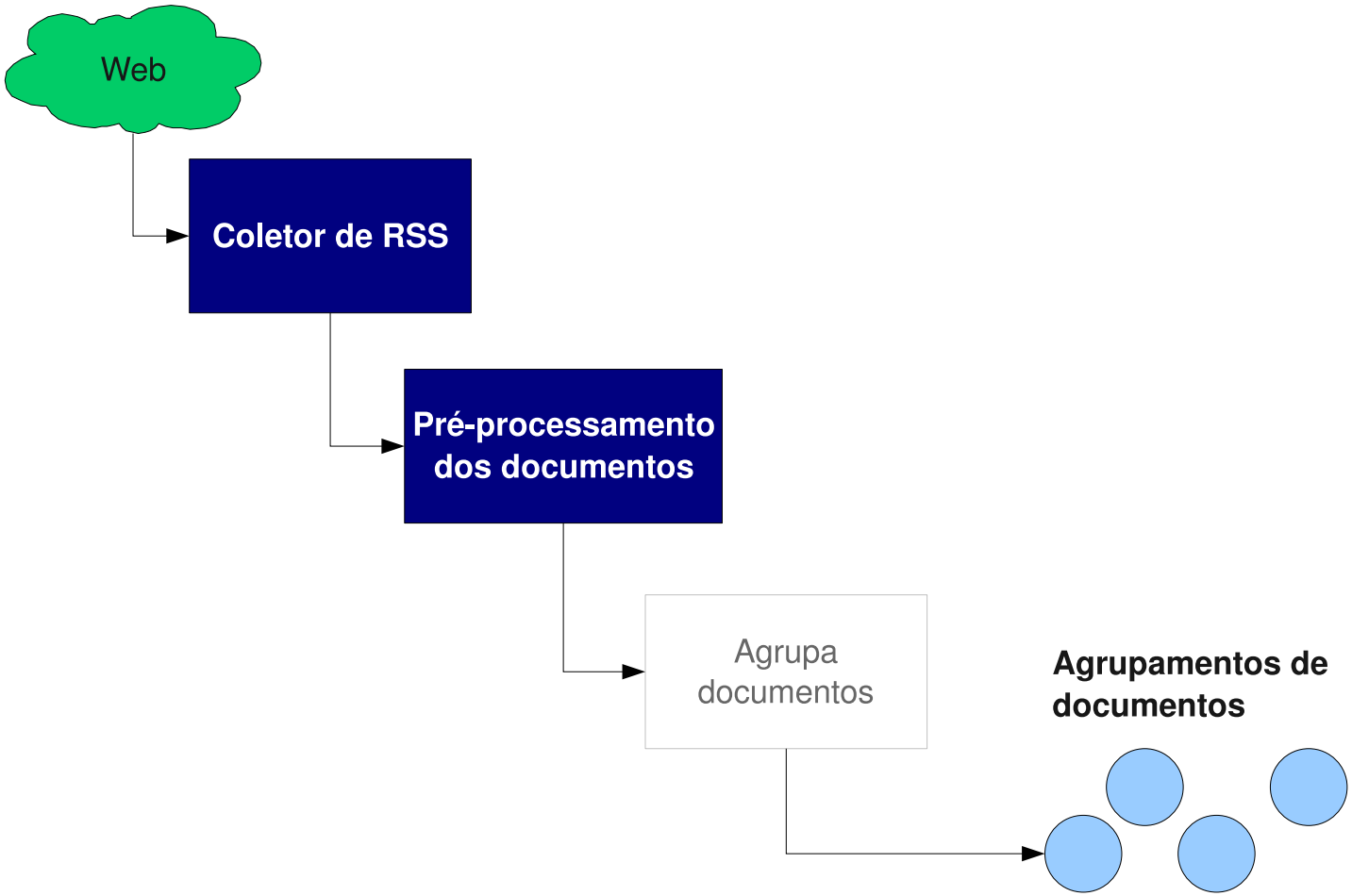
Por freqüência do termo

(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	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

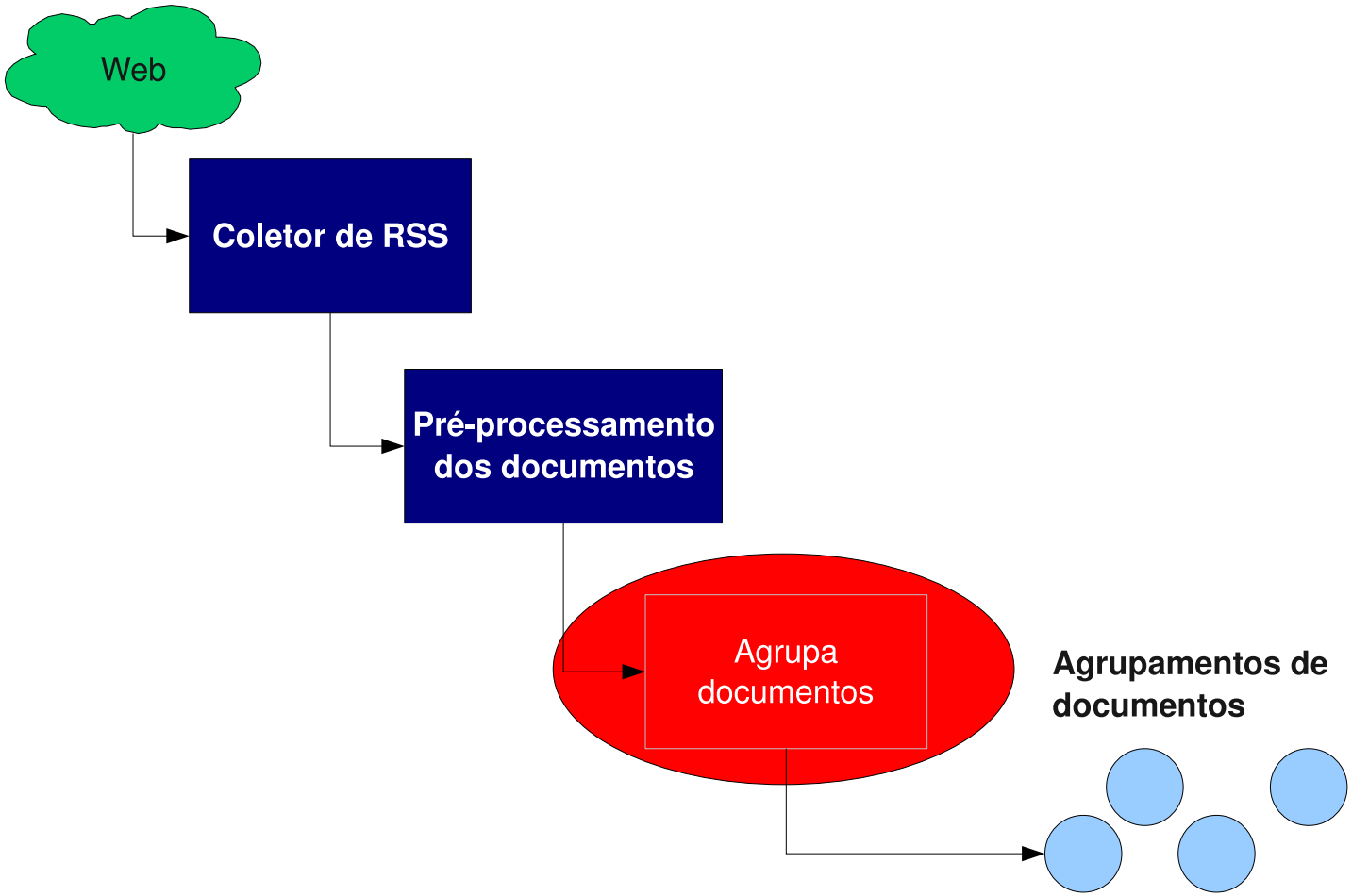
Componentes para uma solução...



Pré-processamento dos documentos: um pouco de código...

Converter texto em TF-IDF usando o RapidMiner.

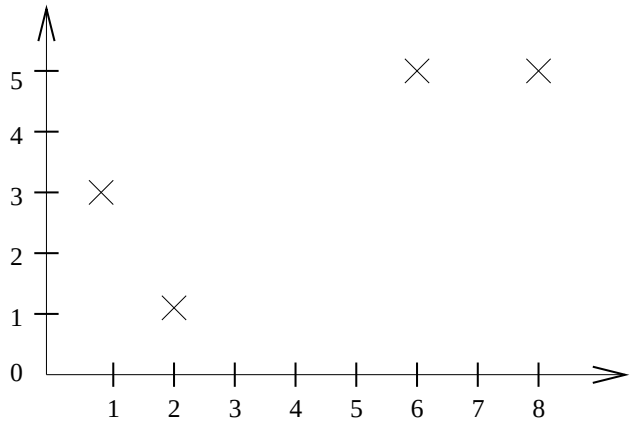
Que algoritmo de agrupamento utilizar?



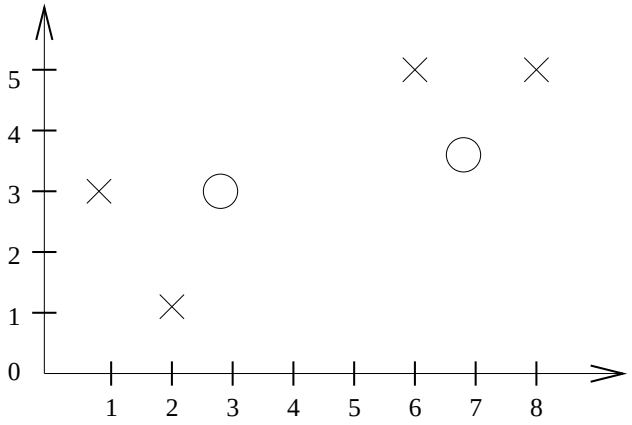
Algoritmos para Agrupamento - *K-means*

- **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).

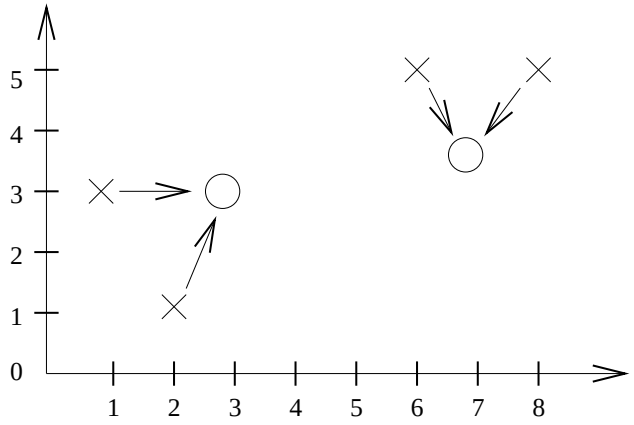
Algoritmo para Agrupamento - *K-means*



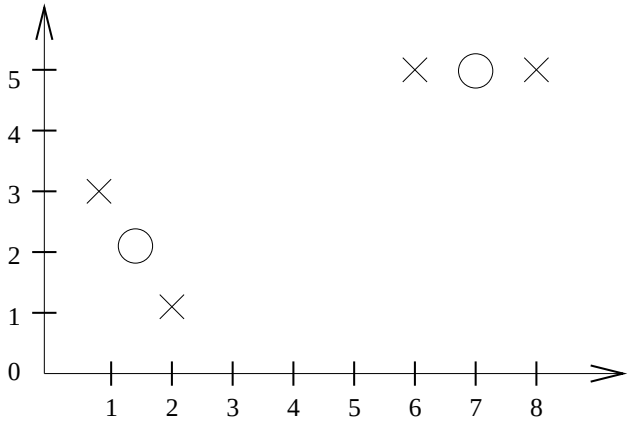
(1) Objetos que devem ser agrupados



(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].

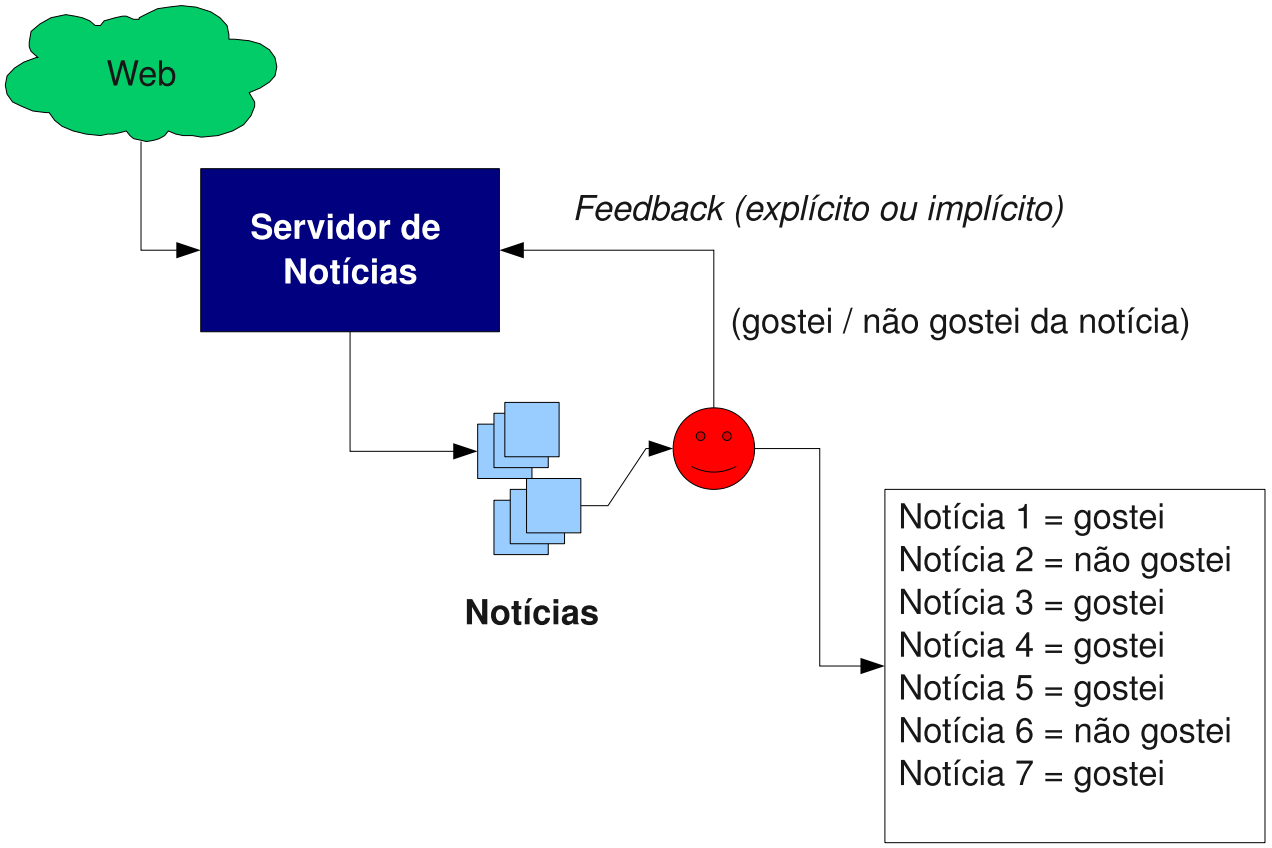
[a] http://www.cs.waikato.ac.nz/ml/weka/

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
```

Classificação de documentos

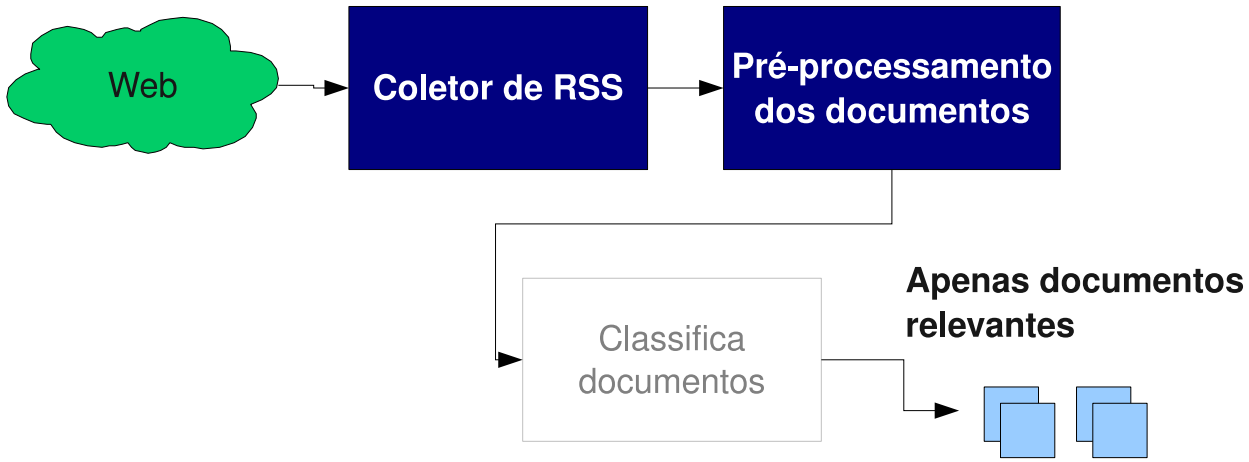
Classificação e Filtragem de Notícias



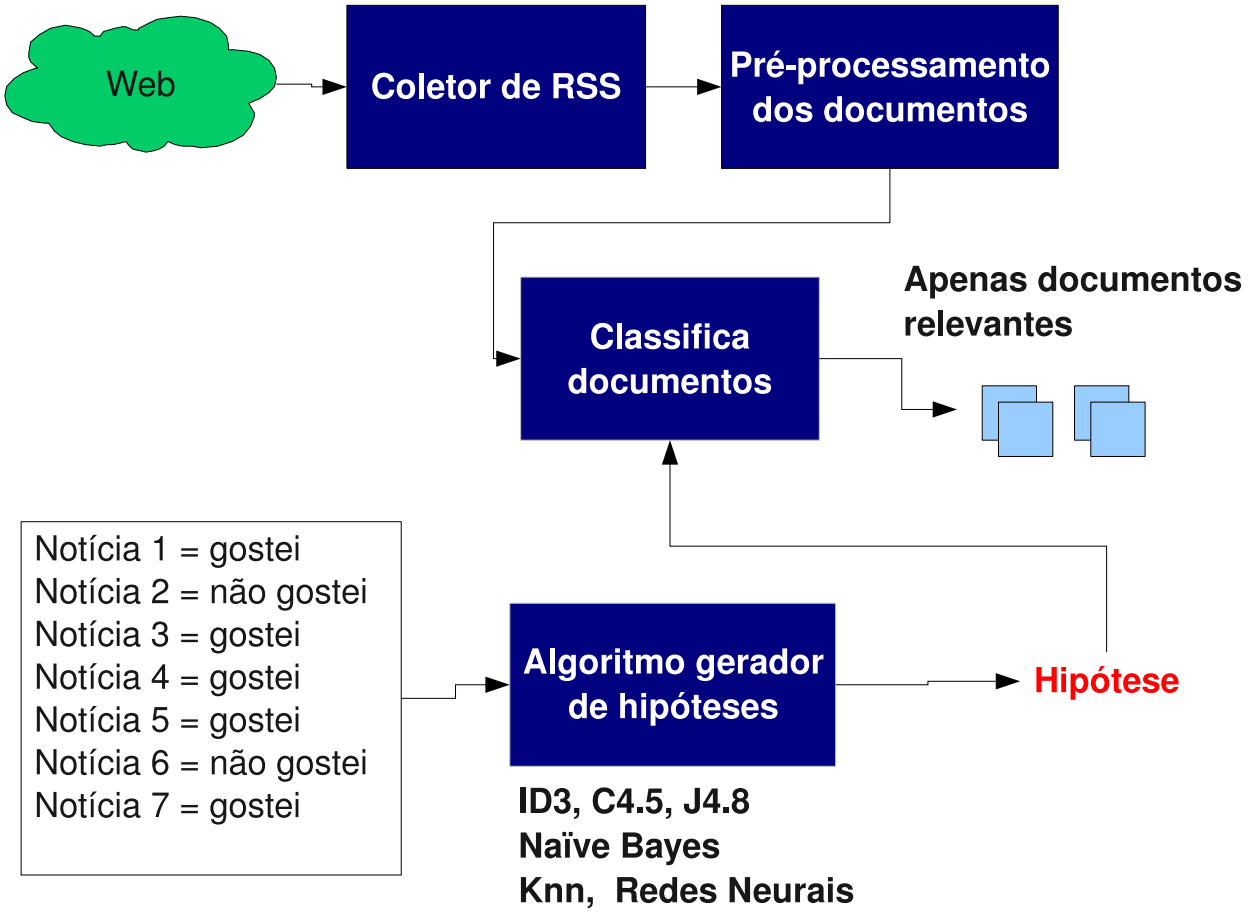
Conjunto de Exemplos - Atributo/Valor e Classe

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

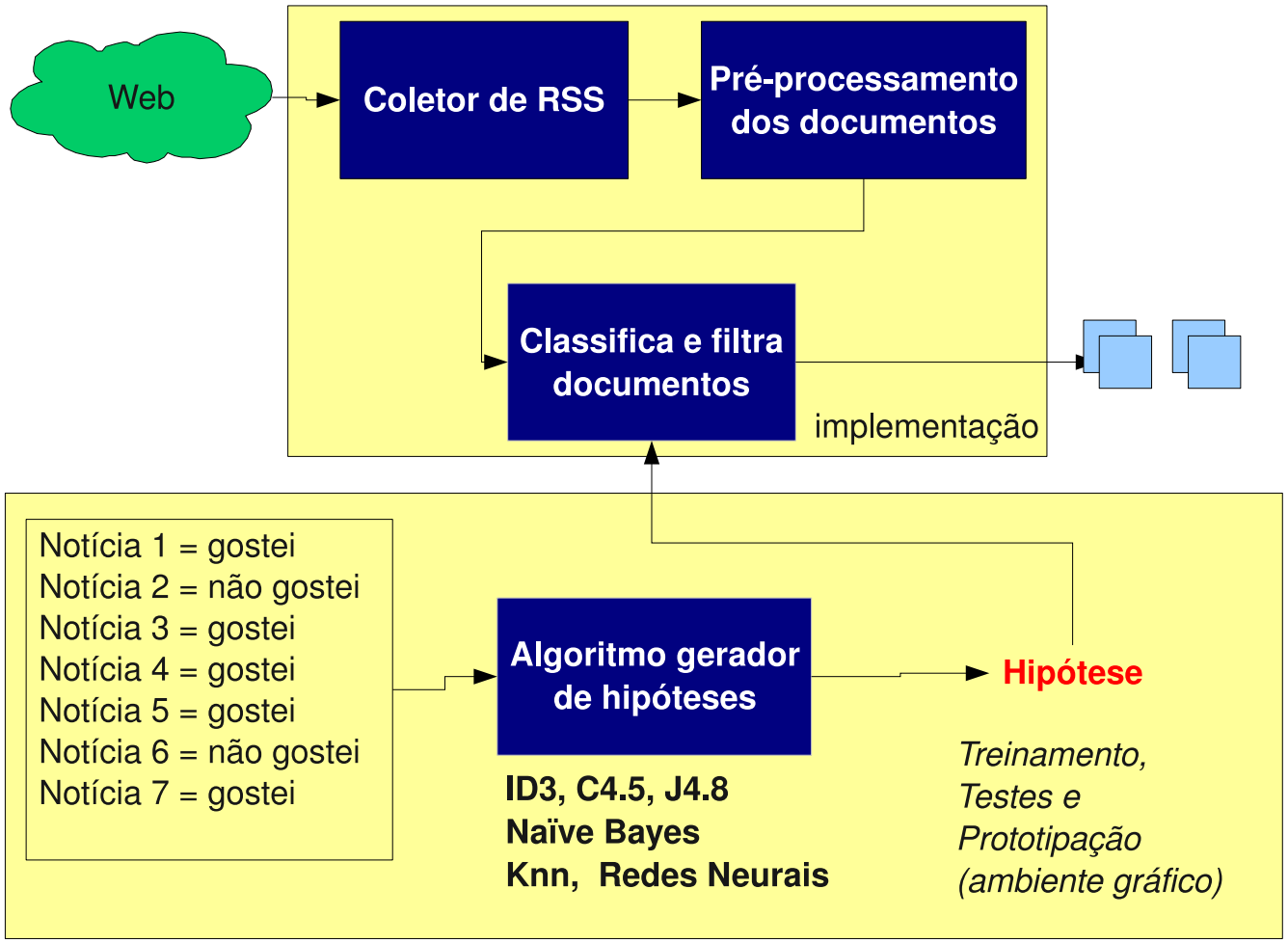
Qual é o problema?



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



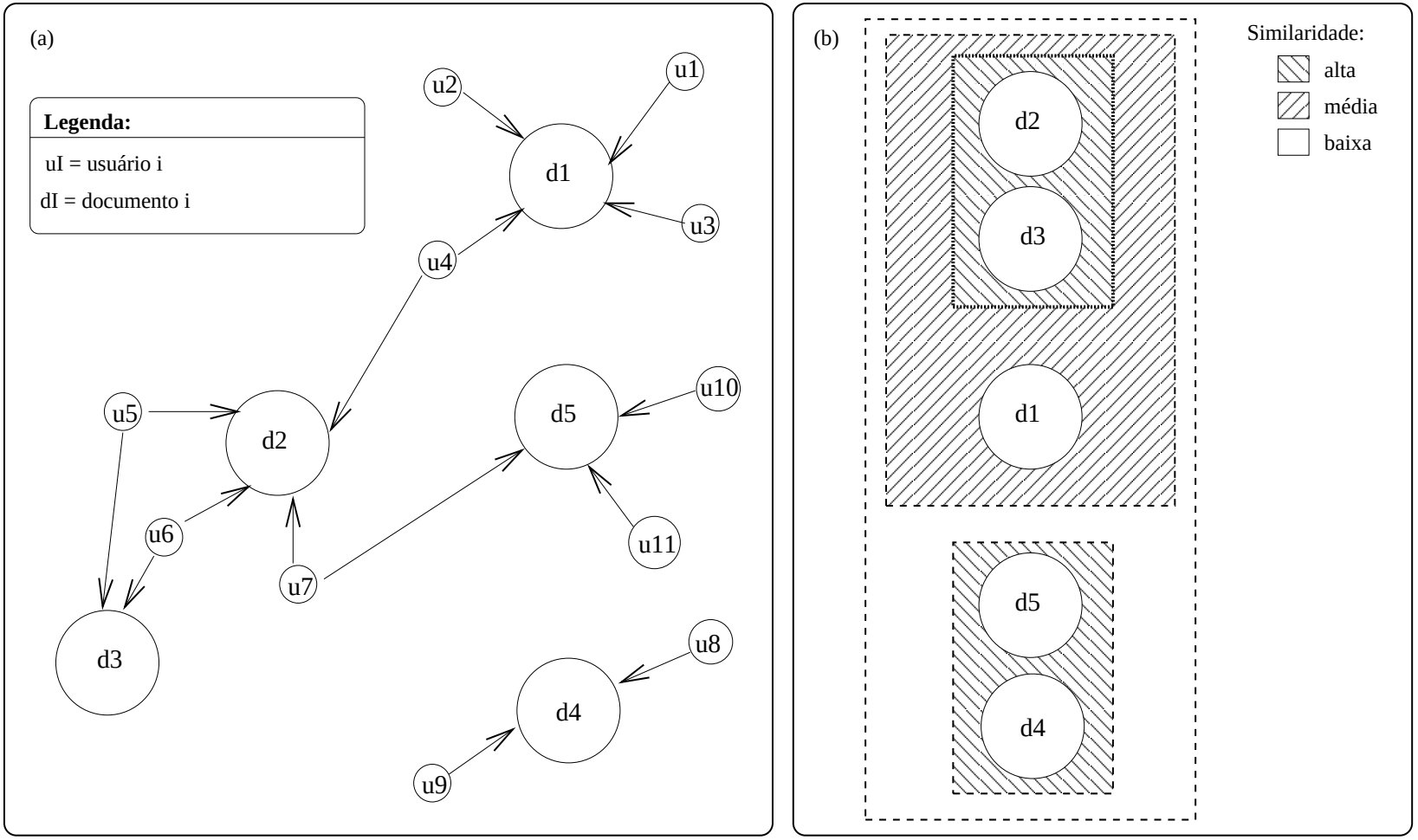
Minerando ambientes colaborativos de escrita (Wiki)

Que informações um Wiki tem?

Documento	Versão	Editor	Data	Documento	Versão	Editor	Data
d_1	1	u_1	$\cdots$	d_2	4	u_7	$\cdots$
d_1	2	u_2	$\cdots$	d_3	1	u_5	$\cdots$
d_1	3	u_2	$\cdots$	d_3	2	u_6	$\cdots$
d_1	4	u_3	$\cdots$	d_3	3	u_6	$\cdots$
d_1	5	u_4	$\cdots$	d_4	1	u_8	$\cdots$
d_2	1	u_4	$\cdots$	d_4	2	u_9	$\cdots$
d_2	2	u_5	$\cdots$	d_5	1	u_{10}	$\cdots$
d_2	3	u_6	$\cdots$	d_5	2	u_{11}	$\cdots$

Exemplo de histórico de criação e alteração de páginas em um WIKI.

O que eu posso fazer com isto?



Exemplo

Execução de uma aplicação usando estes conceitos.

Considerações Finais

Considerações Finais

- Caso sobre conhecimento médico: *Data Mining*.
- Agrupamento, classificação e filtragem de documentos: *Text Mining*.
- Caso WIKI: *Web Mining*
- O que existe em comum nestes casos?

Referências

References

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

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

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

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