

Text Clustering Using Universal Networking Language Representation

Bhoopesh Choudhary

Pushpak Bhattcharyya

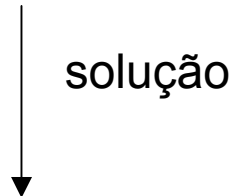
CSE Department
Indian Institute of Technology, Bombay
India

Apresentação
Alessandro Mueller
PPGEP

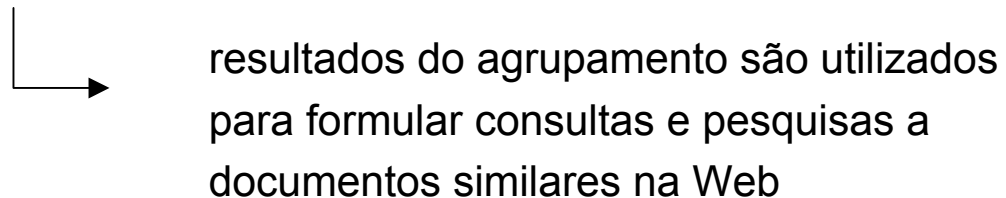
maio de 2002

mecanismos de pesquisa na Web

- localização de documentos
- URL
- pesquisas inconsistentes



*extração de características semânticas
das palavras ou da estrutura do documento
para classificação e categorização*



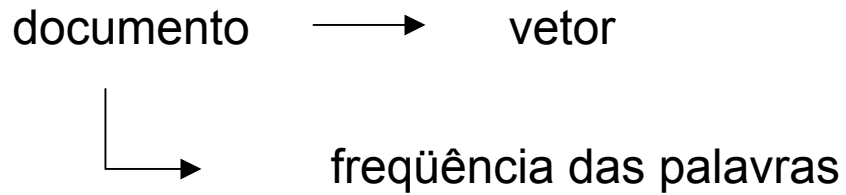
K Means Algorithm
Expectation Maximization
hierarchical clustering

...

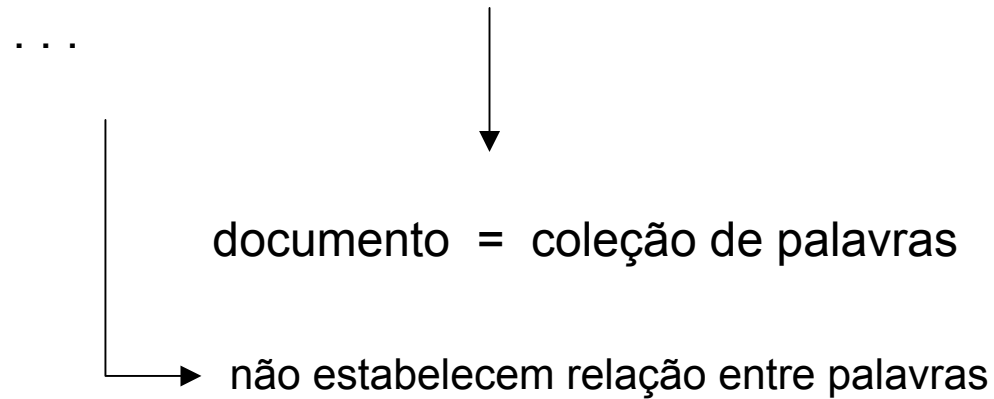
└─> algoritmos para agrupamento



utilizam vetores para formar os grupos



Inverse Document Frequency
Information Gain



problema

→ uma palavra com vários significados

He went to the bank to withdraw some money

The boat was beside the bank

*melhorar a acurácia do processo de
agrupamento pelo uso de informação
semântica das sentenças que
representam o documento*



UNL

métodos de freqüência de palavras

documento = coleção de palavras



freqüência palavras \longrightarrow vetor

deficiências dos métodos de frequência de palavras

- não consideram relações semânticas das palavras
 - a) sentenças com o mesmo conjunto de palavras e de significados distintos NÃO são categorizadas em grupos diferentes

Ram eats the apple beside the tree

The apple tree is beside Ram's house

- b) sentenças construídas com palavras distintas, mas com o mesmo significado, SÃO classificadas em grupos diferentes

Ram is an intelligent boy

Shyam is a brilliant student

- não necessariamente a palavra que apresenta a maior frequência é a que melhor descreve o documento

método de ligações UNL

sentença representada através de um hipergrafo

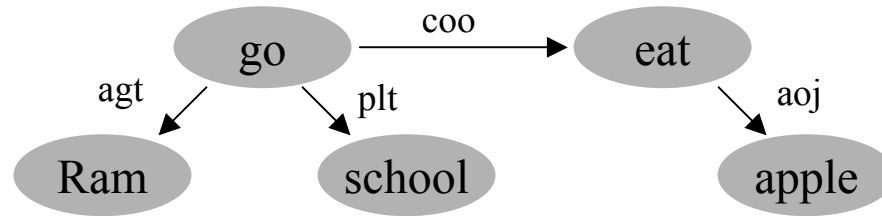
nodos = conceitos

ligações = relações entre conceitos

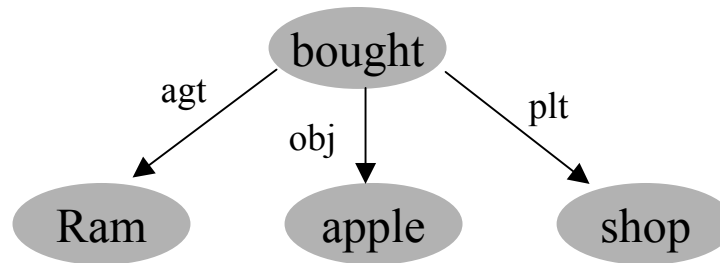
palavras do documento \longrightarrow UWs

cada componente do vetor corresponde ao número de ligações incidentes no nodo (considerando o grafo não direcionado)

*quanto maior o número de ligações **para** e **de** uma UW, maior a importância da palavra no documento*



Ram is going to the school eating an apple



Ram bought the apple from the shop

$X1 = \langle 1, 3, 1, 2, 1, 0, 0 \rangle$

$X2 = \langle 1, 0, 0, 0, 1, 3, 1 \rangle$

$\{ Ram, go, school, eat, apple, bought, shop \}$

A abordagem utilizando UNL não apenas considera a freqüência das palavras no documento, mas também adiciona informação sobre a estrutura da sentença, atribuindo maior peso para palavras importantes na sentença

Ram goes to the bank

Shyam goes to the market

{ Ram, goes, bank, Shyam, market }

método de ligações UNL

$X1 = \langle 1, 2, 1, 0, 0 \rangle$

$X2 = \langle 0, 2, 0, 1, 1 \rangle$

similaridade = 0.66

método de frequência de palavras

$X1 = \langle 1, 1, 1, 0, 0 \rangle$

$X2 = \langle 0, 1, 0, 1, 1 \rangle$

similaridade = 0.33

vantagens da abordagem utilizando ligações UNL

- implicitamente incorpora a frequência de UWs
- UWs capturam o significado da palavra de acordo com o contexto

crane (icl > bird)

crane (icl > machine)

The crane was eating fish

The crane lifted the load

processo de agrupamento

- vetor de frequência de palavras
- vetor de ligações UNL
- Self-organizing Maps (SOM)
*sinônimos e palavras estritamente relacionadas
freqüentemente mapeadas para o mesmo neurônio*

experimento

- 14 documentos, com 6 a 7 linhas cada
 - 6 documentos sobre um determinado tópico e
8 documentos sobre outro tópico
- tamanho do
 - vetor de frequência de palavras* = 534
 - vetor de ligações UNL* = 561
- SOM de tamanho 2 x 2
- treinamento considerando 100.000 iterações

8 + 1
9

0 + 2
2

0 + 0
0

0 + 3
3

freqüência de palavras

7 + 0
7

0 + 0
0

0 + 0
0

1 + 6
7

ligações UNL

→ método de freqüência de palavras = 3 grupos
método de ligações UNL = 2 grupos

número de documentos classificados de maneira
incorreta maior no método de freqüência de palavras