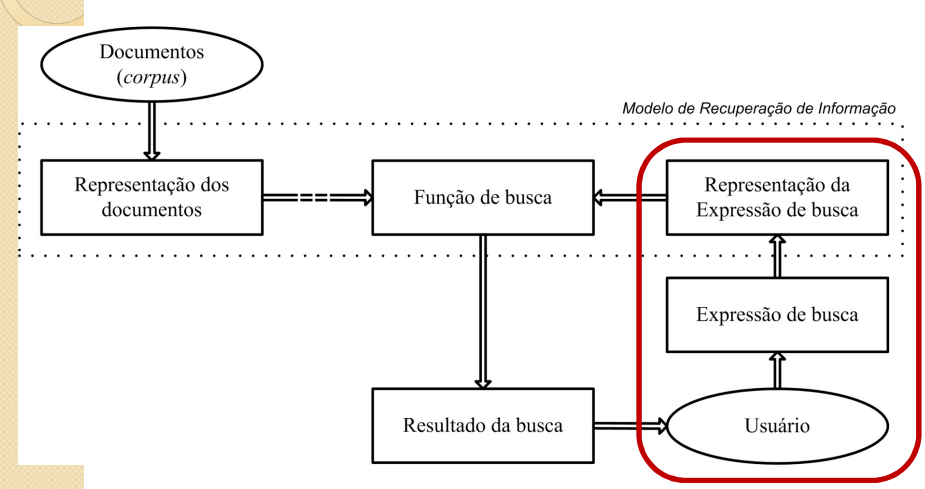
Onde estamos



Onde estamos

- As pesquisas em recuperação de informação:
  - Pesquisas centradas no sistema
    - Algoritmos de recuperação, ranqueamento, indexação,
    - Projetos de interface, etc.
  - Pesquisas centradas no usuário
    - Comportamento informacional
    - Métodos centrados no usuário ou Modelos cognitivos

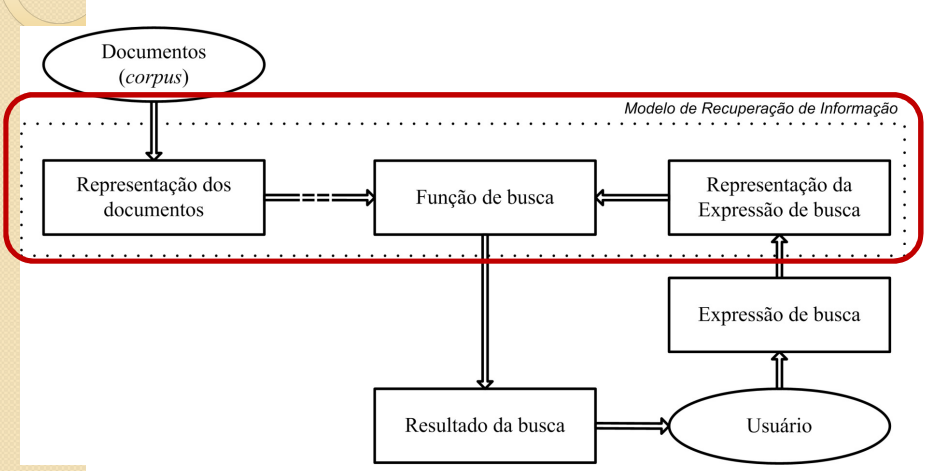
Onde estamos

- Modelos Centrados no Usuário
  - Modelo de Wilson
  - Modelo de Dervin
  - Modelo de Kuhlthau
  - Modelo de Ellis
  - Modelo Cognitivo de Ingwersen
- Técnicas e Tecnologias
  - *Relevance Feedback*
  - Análise de Logs
  - Expansão de Consulta

Onde estamos

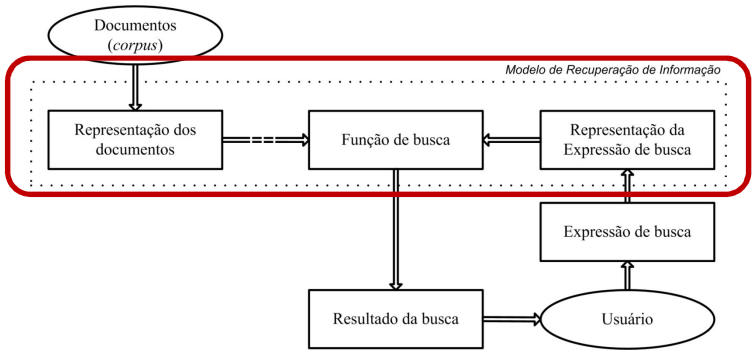
- Interfaces de Busca
  - Como as pessoas buscam informação?
    - Modelo Clássico/Linear;
    - Modelo Dinâmico;
  - Query
    - Envolve a comparação entre uma necessidade de informação, representada por uma expressão de busca, com a representação dos documentos de um *corpus*;
  - Browse
    - Permitem que o usuário percorra algum tipo de estrutura de informação na busca por documentos relevantes;

Para onde vamos



Modelo de Recuperação de Informação

Modelo de Recuperação de Informação



- Um modelo de recuperação de informação é a especificação formal de três elementos:
  - a *representação dos documentos;*
  - a *representação da necessidade* de informação por meio de uma *expressão de busca*
  - como estes dois elementos serão comparados, a *função de busca.*

Modelo de Recuperação de Informação

- Baeza-Yates e Ribeiro-Neto (2011, p.58) definem modelo de recuperação de informação como uma quadrupla:

$$[\ \mathbf{D}, \mathbf{Q}, F, R(q_i, d_j)\ ]$$

D é um conjunto composto por visões lógicas (representações) dos documentos no *corpus*;

Q é um conjunto composto de visões lógicas das necessidades de informação dos usuários;

F é um *framework* para a modelagem de representações dos documentos, consultas e seus relacionamentos;

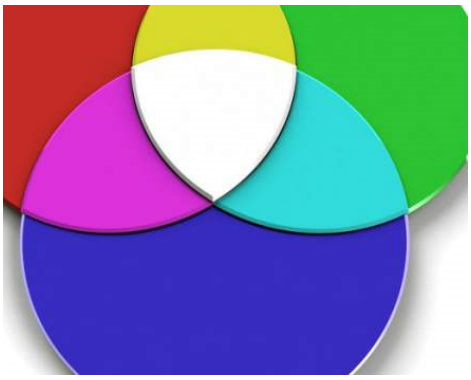
$\boldsymbol{R}(q_i, d_j)$ é uma função de ordenamento (*ranking*) que atribui um número real à relação entre uma representação da consulta q_i de **Q** e a representação de um documento d_j de **D**.

Modelo de Recuperação de Informação

- A eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que ele utiliza, influenciando diretamente em seu modo de operação.
- Apesar de alguns dos modelos de recuperação de informação terem sido criados nos anos 60 e 70 e aperfeiçoados nos anos 80, as suas principais ideias ainda estão presentes na maioria dos sistemas de recuperação atuais e nos mecanismos de busca da Web.

Modelos Clássicos

Modelos Clássicos de Recuperação de Informação



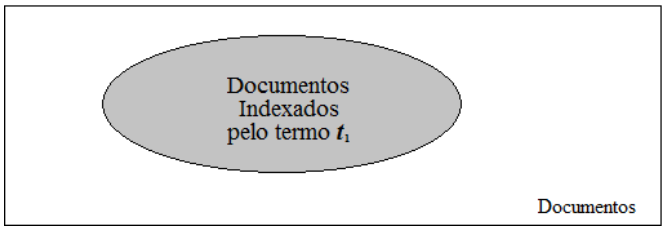
Modelo Booleano

Modelo Booleano

- No modelo booleano um **documento** é representado por um **conjunto de termos** de indexação que podem ser definidos de forma intelectual (manual) por profissionais especializados ou automaticamente, utilizando algoritmos computacionais.
- As **buscas** são formuladas por meio de uma **expressão booleana** composta por termos ligados por operadores lógicos AND, OR e NOT e apresentam como resultado os documentos cuja representação satisfazem às restrições lógicas da expressão de busca.

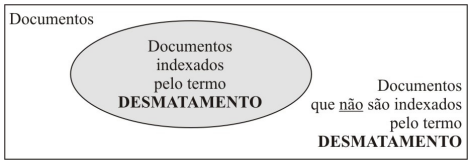
Modelo Booleano

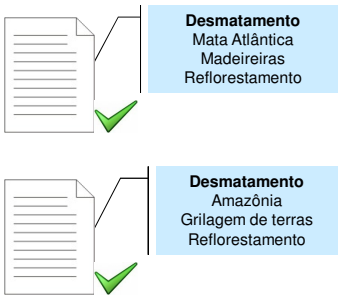
- Uma expressão de busca que utiliza apenas um termo $\mathbf{t_1}$ terá como resultado o conjunto de documentos indexados por $\mathbf{t_1}$;



Modelo Booleano

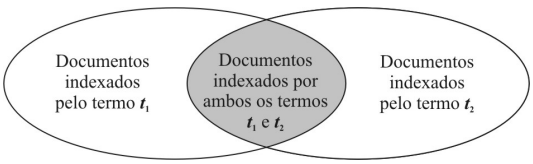
Desmatamento





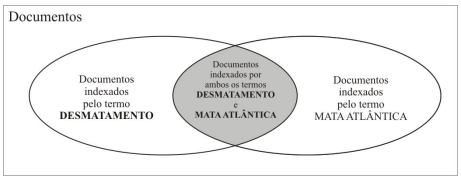
Modelo Booleano

- Uma expressão conjuntiva de enunciado t_1 **AND** t_2 recuperará documentos indexados por ambos os termos (t_1 e t_2).
- Esta operação equivale à *interseção* do conjunto dos documentos indexados pelo termo t_1 com o conjunto dos documentos indexados pelo termo t_2, representado pela área cinza na figura.



Modelo Booleano

Desmatamento
AND
Mata Atlântica

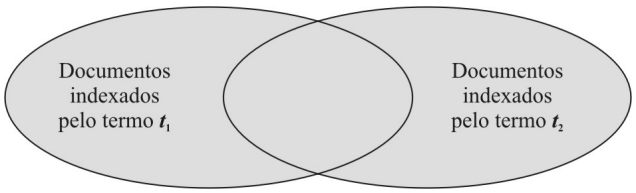


Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

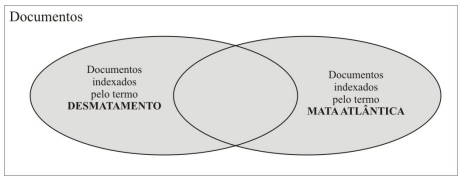
Modelo Booleano

- Uma expressão disjuntiva $\mathbf{t_1}$ **OR** $\mathbf{t_2}$ recuperará o conjunto dos documentos indexados pelo termo $\mathbf{t_1}$ ou pelo termo $\mathbf{t_2}$.
- Essa operação equivale à *união* entre o conjunto dos documentos indexados pelo termo $\mathbf{t_1}$ e o conjunto dos documentos indexados pelo termo $\mathbf{t_2}$.



Modelo Booleano

Desmatamento
OR
Mata Atlântica

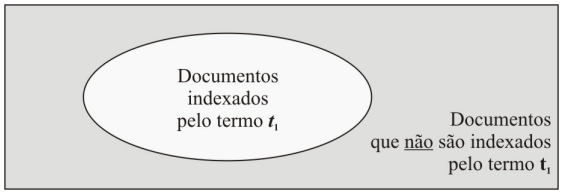


Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

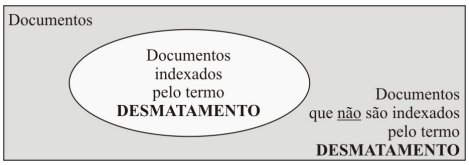
Modelo Booleano

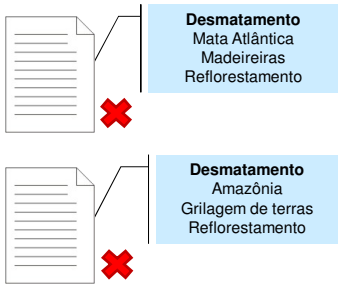
- A expressão **NOT t_1** recuperará os documentos que **não** são indexados pelo termo **t_1**, representados pela área cinza da figura.



Modelo Booleano

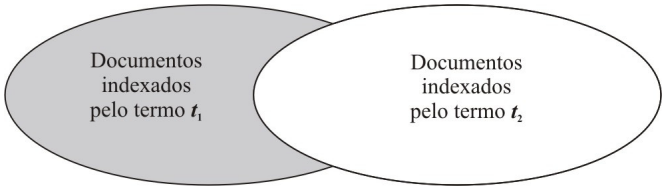
NOT Desmatamento





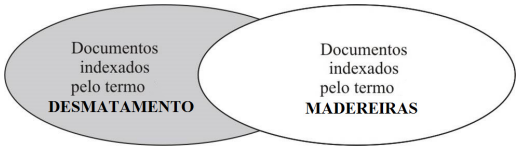
Modelo Booleano

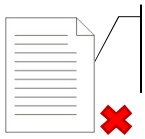
- As expressões t_1 **NOT** t_2 ou t_1 **AND NOT** t_2 terão o mesmo resultado: o conjunto dos documentos indexados por t_1 e que não são indexados por t_2.
- Neste caso o operador NOT pode ser visto como um operador da diferença entre conjuntos.



Modelo Booleano

Desmatamento
AND NOT Madeireiras





Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

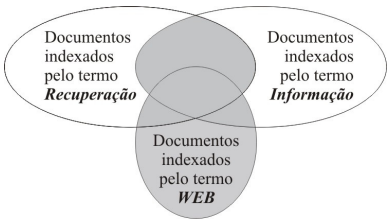
Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

Modelo Booleano

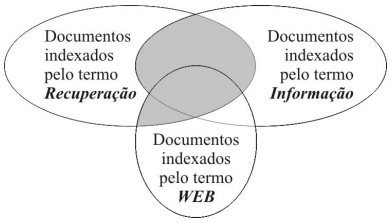
- Termos e operadores booleanos podem ser combinados para especificar buscas mais amplas ou restritivas.
- Como a ordem de execução das operações lógicas de uma expressão influencia no resultado da busca, muitas vezes é necessário explicitar essa ordem, delimitando partes da expressão por meio de parênteses.

Modelo Booleano

- As áreas cinza da figura representam o resultado de duas expressões de busca que utilizam os mesmos termos e os mesmos operadores, mas diferem na ordem de execução.



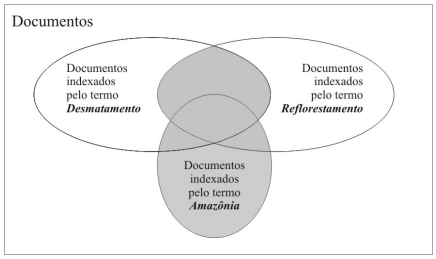
(*Recuperação* AND *Informação*) OR *WEB*



Recuperação AND (*Informação* OR *WEB*)

Modelo Booleano

(Desmatamento **AND** Reflorestamento)
OR
Amazônia

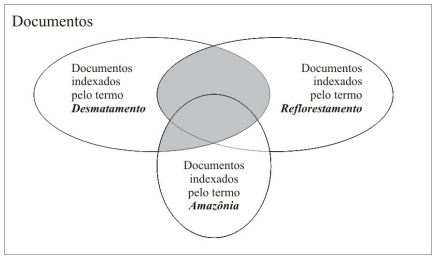


Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

Modelo Booleano

Desmatamento
AND
(Reflorestamento **OR** Amazônia)



Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

Modelo Booleano

- Operadores de Proximidade
  - Surgimento dos sistemas de texto completo
  - Operadores
    - Termos adjacentes
      - Desmatamento **ADJ** Amazônia
      - Desmatamento **NEAR/10** Amazônia
    - Sistema STAIRS
      - Desmatamento **WITH** Amazônia (mesmo parágrafo)
      - Desmatamento **SAME** Amazônia (mesma frase)
  - Frase Exata
    - "Recuperação de Informação";"Desmatamento na Amazônia"
  - Composição de Operadores
    - "Recuperação de" **ADJ** (informação OR documentos)

Modelo Booleano

- Operadores de Proximidade
  - Mesmo utilizando operadores de proximidade, o resultado de uma busca booleana será um conjunto de documentos que respondem verdadeiramente à expressão de busca e presumivelmente serão relevantes pelo usuário.
  - Apesar de os operadores de proximidade agregarem novos recursos aos sistemas de texto completo, tais operadores não alteram substancialmente as vantagens e limitações do modelo booleano

A relevância no modelo booleano

A Relevância no Modelo Booleano

- A lógica booleana difere da lógica natural;
  - Na linguagem cotidiana, quando falamos "gatos e cachorros", intuitivamente imagina-se uma **união** do conjunto dos "gatos" com o conjunto dos "cachorros".
  - Em um sistema de recuperação de informação a expressão t_1 AND t_2 resultará na **interseção** entre o conjunto dos documentos indexados pelo termo t_1 e o conjunto dos documentos indexados por t_2.
  - Na linguagem cotidiana, a expressão "café ou chá" expressa uma escolha ou seleção cujo resultado será apenas um dos elementos envolvidos.
  - Em um sistema de recuperação de informação, a expressão t_1 OR t_2 resultará uma **união** do conjunto de documentos indexados por t_1 com o conjunto de documentos indexados por t_2

(SMITH, 1993).

A Relevância no Modelo Booleano

- Não há nenhum mecanismo pelo qual os documentos resultantes de uma busca possam ser ordenados;
  - Os termos de indexação possuem a mesma importância (relevância) na representação do conteúdo dos documentos;
  - De forma similar, não é possível expressar que um termo de busca seja mais importante/relevante do que outro.
- O resultado de uma busca booleana é um conjunto de documentos que respondem verdadeiramente à expressão de busca;
  - O resultado se caracteriza por uma simples partição do corpus em dois subconjuntos: os documentos que atendem à expressão de busca e aqueles que não atendem;
  - Uma das maiores desvantagens do modelo booleano é a sua inabilidade em ordenar por relevância (ranquear) os documentos resultantes de uma busca.
- Para representar estratégias de busca mais complexas é necessário ter conhecimento da lógica booleana;

A Relevância no Modelo Booleano

- Apesar de suas limitações, o modelo booleano está presente em quase todos os sistemas de recuperação de informação e nos sistemas de banco de dados.
  - Facilidade de implementação;
  - Flexibilidade e poder, oferecendo certo controle sobre os resultados;
  - É fácil para o usuário entender porque um documento foi ou não recuperado

Gerard Salton
(1927-1995)

Modelo Vetorial

Modelo Vetorial

- O Modelo Espaço Vetorial (*Vector Space Model*) propõe um ambiente no qual é possível obter documentos que respondem parcialmente a uma expressão de busca.
- Isto é feito associando-se pesos tanto aos termos de indexação dos documentos como aos termos utilizados na expressão de busca.
- Como resultado, obtém-se um conjunto de documentos ordenado pelo grau de **similaridade** de cada documento em relação à expressão de busca.

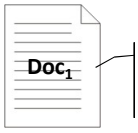
Modelo Vetorial

Representação dos documentos

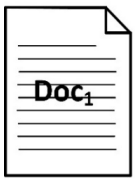
- Um documento é representado por um vetor onde cada elemento representa o peso, ou relevância, do respectivo termo de indexação para o documento.
- Cada vetor descreve a posição do documento em um espaço multidimensional, onde cada termo de indexação representa uma dimensão ou eixo.
- Cada elemento do vetor (peso) é normalizado de forma a assumir valores entre zero e um. Os pesos mais próximos de 1 indicam termos com maior importância para a descrição do documento.

Modelo Vetorial

Representação dos documentos



Mata Atlântica	0.7
Desmatamento	0.6
Reflorestamento	0.3

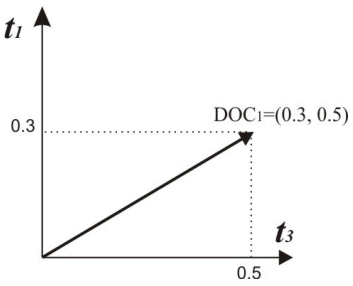


Mata Atlântica	Desmatamento	Reflorestamento
0,7	0,6	0,3

Modelo Vetorial

Representação dos documentos

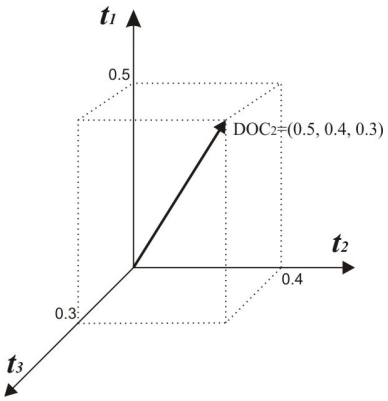
	t_1	t_3
DOC_1	0.3	0.5



Modelo Vetorial

Representação dos documentos

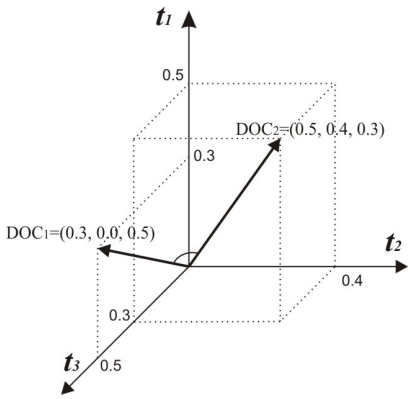
	t_1	t_2	t_3
DOC_2	0.5	0.4	0.3



Modelo Vetorial

- Os números positivos representam os pesos de seus respectivos termos. Termos que não estão presentes em um determinado documento possuem peso igual a zero.

	t_1	t_2	t_3
DOC_1	0.3	0.0	0.5
DOC_2	0.5	0.4	0.3



Modelo Vetorial

corpus documental

- Um *corpus* contendo ***n*** documentos e ***i*** termos de indexação pode ser representado da seguinte forma:

	t_1	t_2	t_3	...	t_i
DOC_1	$w_{1,1}$	$w_{2,1}$	$w_{3,1}$	...	$w_{i,1}$
DOC_2	$w_{1,2}$	$w_{2,2}$	$w_{3,2}$	...	$w_{i,2}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
DOC_n	$w_{1,n}$	$w_{2,n}$	$w_{3,n}$	...	$w_{i,n}$

onde $w_{i,n}$ representa o peso do *i*-ésimo termo do *n*-ésimo documento.

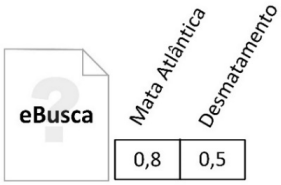
Modelo Vetorial

representação das buscas

- Uma expressão de busca também é representada por um vetor numérico onde cada elemento representa a importância (peso) do respectivo termo na representação da necessidade de informação do usuário, substanciada na expressão de busca.

Usuário e sua necessidade de informação

Mata Atlântica	0.8
Desmatamento	0.5



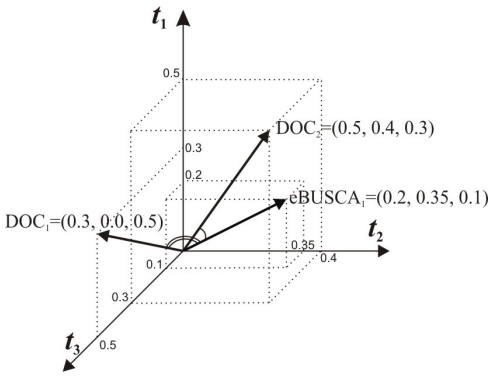
Modelo Vetorial

representação das buscas

- A figura mostra a representação da expressão de busca **eBUSCA$_1$** = (0.2, 0.35, 0.1) juntamente com os documentos **DOC$_1$** e **DOC$_2$** em um espaço vetorial formado pelos termos t_1, t_2 e t_3.

	t_1	t_2	t_3
eBUSCA$_1$	0.2	0.35	0.1

	t_1	t_2	t_3
DOC$_1$	0.3	0.0	0.5
DOC$_2$	0.5	0.4	0.3



Modelo Vetorial

cálculo da similaridade

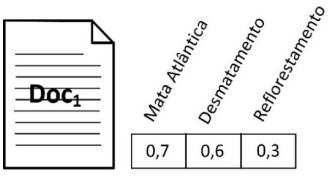
- A utilização de uma mesma forma de representação tanto para os documentos como para as expressões de busca permite calcular a **similaridade** entre uma expressão de busca e cada um dos documentos do *corpus*, ou ainda entre dois documentos;
- Em um espaço vetorial contendo ***N*** dimensões, a similaridade (***sim***) entre um documento d_j e uma expressão de busca ***q*** pode ser calculada utilizando a seguinte fórmula:

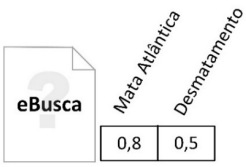
$$\text{sim}(d_j, q) = \frac{\sum_{i=1}^{N}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

$w_{i,j}$ é o peso do *i*-ésimo termo do documento d_j e $w_{i,q}$ é o peso do *i*-ésimo termo da expressão de busca *q*.

Modelo Vetorial

cálculo da similaridade



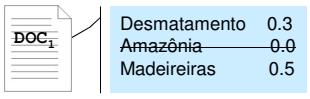


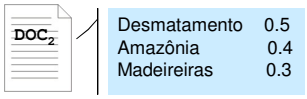
$$sim(Doc_1, \text{eBusca}) = \frac{\sum_{k=1}^{N}\left(w_{k,i} \times w_{k,q}\right)}{\sqrt{\sum_{k=1}^{N} w_{k,i}^2} \times \sqrt{\sum_{k=1}^{N} w_{k.q}^2}}$$

$$sim(Doc_1, eBusca) = \frac{(0{,}7 \times 0{,}8) + (0{,}6 \times 0{,}5)}{\sqrt{(0{,}7^2 + 0{,}6^2 + 0{,}3^2)} \times \sqrt{(0{,}8^2 + 0{,}5^2)}} \cong \mathbf{0,94}$$

Modelo Vetorial

cálculo da similaridade





$$\text{sim}(doc_1, doc_2) = \frac{\sum_{i=1}^{N}\left(w_{i,d_1} \times w_{i,d_2}\right)}{\sqrt{\sum_{i=1}^{N} w_{i,d_1}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,d_2}^2}}$$

$$\text{sim}(doc_1, doc_2) = \frac{(0.3 \times 0.5) + (0.0 \times 0.4) + (0.5 \times 0.3)}{\sqrt{0.3^2 + 0.0^2 + 0.5^2} \times \sqrt{0.5^2 + 0.4^2 + 0.3^2}}$$

$$\text{sim}(doc_1, doc_2) \cong \mathbf{0,728}$$

Modelo Vetorial

cálculo da similaridade

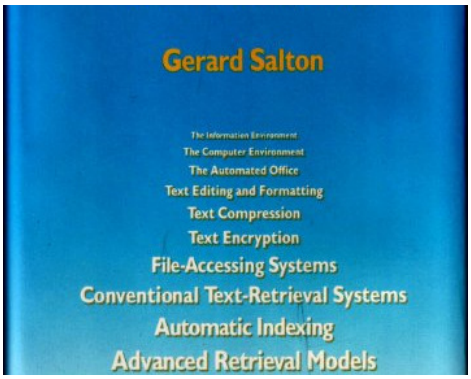
Desmatamento	0.2
Amazônia	0.35
Madeireiras	0.1

Desmatamento	0.3
Amazônia	0.0
Madeireiras	0.5

Desmatamento	0.5
Amazônia	0.4
Madeireiras	0.3

$$\text{sim}(q, doc_1) = \frac{(0.2\times0.3)+(0.35\times0.0)+(0.1\times0.5)}{\sqrt{0.2^2+0.35^2+0.1^2}\times\sqrt{0.3^2+0.0^2+0.5^2}} \cong \mathbf{0.457}$$

$$\text{sim}(q, doc_2) = \frac{(0.2\times0.5)+(0.35\times0.4)+(0.1\times0.3)}{\sqrt{0.2^2+0.35^2+0.1^2}\times\sqrt{0.5^2+0.4^2+0.3^2}} \cong \mathbf{0.92}$$



Automatic Text Processing:
The Transformation Analysis and Retrieval of Information by Computer
1988

O Sistema SMART

O Sistema SMART

- O projeto SMART (***S****ystem for the* ***Ma****nipulation and* ***R****etrieval of* ***T****ext*) teve início em 1961 na Universidade de Harvard.
- Mudou-se para a Universidade de Cornell após 1965.
- O sistema SMART é o resultado da vida de pesquisa de Gerard Salton e teve um papel significativo no desenvolvimento de toda a área da Recuperação de Informação.
- O SMART é uma implementação do modelo vetorial.

O Sistema SMART

- O sistema SMART fornece um método automático para o cálculo dos pesos não só dos vetores que representam os documentos, mas também para os vetores das expressões de busca.

O Sistema SMART

- Salton e McGill (1983, p.204-207)
  - ***tf*** (*term frequency*): número de vezes que um determinado termo **t** aparece no texto de um documento ***d***.

$$tf_{t,d} = freq_{t,d}$$

  - Um termo que aparece em todos os documentos terá provavelmente pouca utilidade em identificar. Portanto, para um cálculo preciso do peso de um determinado termo de indexação é preciso uma estatística global que caracterize o termo em relação a todo o corpus.
  - ***idf*** (*inverse document frequency):* mostra como o termo é distribuído pelo corpus;

$$idf_t = \log\left(\frac{N}{n_t}\right)$$

N → número de documentos no *corpus*
n_t → número de documentos que contém o termo **t**

O Sistema SMART

- O peso de um termo **t** em relação a um documento ***d*** ($w_{t,d}$) é calculado pela multiplicação da medida ***tf*** pela medida ***idf*** ;
- Essa medida é conhecida como tf x idf e possui a seguinte fórmula:

$$w_{t,d} = tf_{t,d} \times idf_t$$

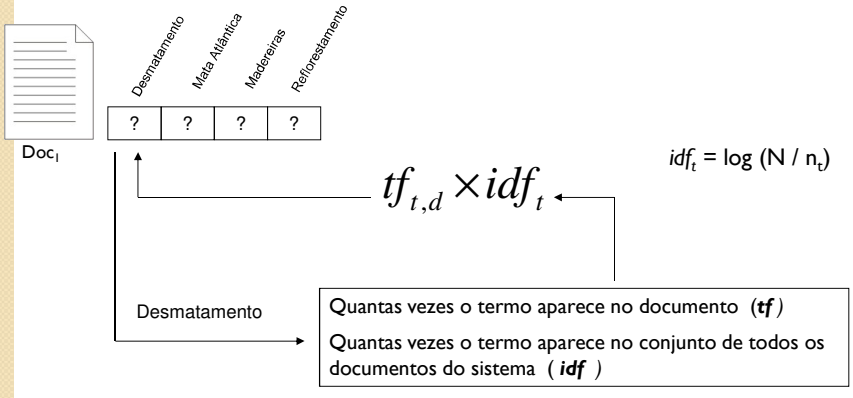
- A medida *tf x idf* é utilizada para atribuir peso a cada elemento dos vetores que representam os documentos do corpus;
- Os melhores termos de indexação (os que apresentarão maior peso) são aqueles que ocorrem com grande frequência em poucos documentos.

O Sistema SMART

- Processo de Indexação
  - Eliminação de *Stop Words*
    - São palavras semanticamente pobres para representar o conteúdo de um documento;
  - Normalização de Termos - *Stemming*
    - Remover os sufixos e (possivelmente também os prefixos) para se chegar ao radical da palavra;
    - Reduz a variabilidade lexical;
  - Calculo do peso de cada termo de indexação (tf-idf)

O Sistema SMART

Calculo automático dos pesos dos termos de indexação



O Sistema SMART

- Assim como os documentos, uma expressão de busca (consulta) também é representada por um vetor;
- Cada termo da busca recebe um número que expressa a importância relativa do termo para a necessidade de informação do usuário;
- Salton e Buckley (1988) descrevem algumas formas alternativas para calcular automaticamente os pesos não só para os termos de indexação, mas também dos termos de busca;
- O peso de cada termo **t** de uma expressão de busca (***q***) pode ser calculado utilizando uma das seguintes fórmulas:

$$idf_t = \log \frac{N}{n_t}$$

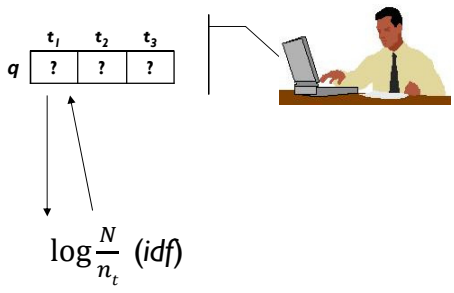
Inverse document frequency clássico

$$idf_t = \log \frac{N - n_t}{n_t}$$

best weighted probabilistic weight

O Sistema SMART

Calculo automático dos pesos da expressão de busca



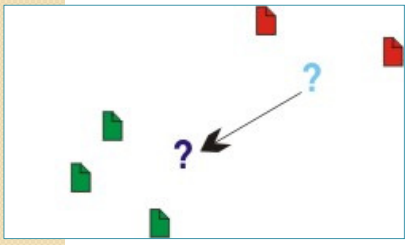
O Sistema SMART

relevance feedback / reformulação de consulta

- Outra técnica pioneira desenvolvida no sistema SMART é a reformulação da expressão de busca do usuário com o propósito de obter melhores resultados na recuperação;
- Essa reformulação pode ser feita automaticamente ou pela interação do usuário, em um processo conhecido como Relevance Feedback.
- Esse processo visa construir uma nova expressão de busca a partir dos documentos identificados como relevantes no conjunto de documentos resultantes de uma busca anterior;

O Sistema SMART

relevance feedback / reformulação de consulta



Documento considerado não-Relevante

Documento considerado Relevante

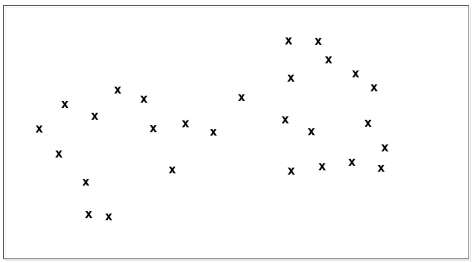
? Expressão de busca original

? Expressão de busca reformulada

1. Após uma busca, o usuário seleciona (marca) os documentos que considera relevantes e submete tal seleção aos sistema;
2. Os termos que ocorrem nos documentos identificados como relevantes são adicionados ao vetor da expressão de busca original, ou os pesos de tais termos são aumentados na construção de uma nova expressão de busca;
3. Termos que ocorrem em documentos identificados como não relevantes são excluídos da expressão de busca original, ou os pesos de tais termos são reduzidos;

O Sistema SMART

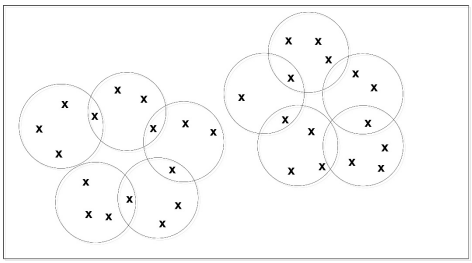
clustering



- Cada **x** representa o vetor de um documento

O Sistema SMART

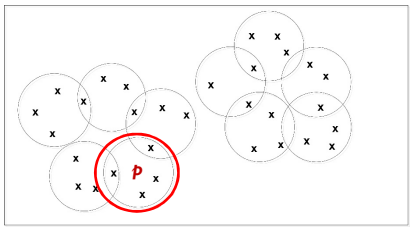
clustering



- As circunferências representam os *clusters*
- Pode ser observado que alguns grupos se intersecionam, possuindo documentos em comum.

O Sistema SMART

clustering

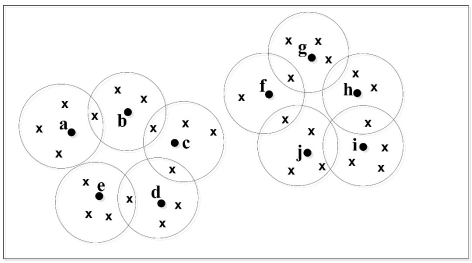


	t_1	t_2	t_3	t_4	t_5
Doc_1	0,5	0,3	0,2	0,8	0,35
Doc_2	0,7	0,2	0,4	0,6	0,2
Doc_3	0,4	0,0	0,3	0,7	0,4
Doc_4	0,6	0,1	0,3	0,85	0,28
C_p	**0,55**	**0,15**	**0,3**	**0,74**	**0,31**

- Considere um *cluster p* formado por quatro documentos (Doc_1, Doc_2, Doc_3, Doc_4), representados por seus vetores contendo cinco termos (t_1, t_2, t_3, t_4, t_5):
- O valor de cada elemento do vetor do centroide C*p* é calculado pela media dos valores dos elementos correspondentes dos documentos Doc_1, Doc_2, Doc_3 e Doc_4.

O Sistema SMART

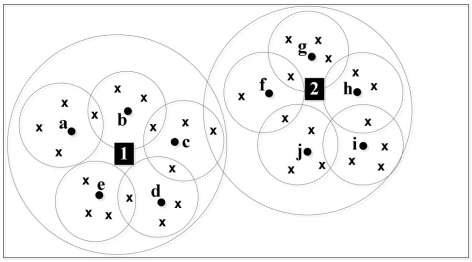
clustering



- Para que seja possível a manipulação de uma coleção de *clusters*, Salton e McGill (1983, p.125) propõem a criação de um tipo especial de vetor denominado “centroide”.
- Um centroide (●) é um vetor que não representa um documento, mas sim um *cluster*, podendo ser pensado como o seu “centro de gravidade”.

O Sistema SMART

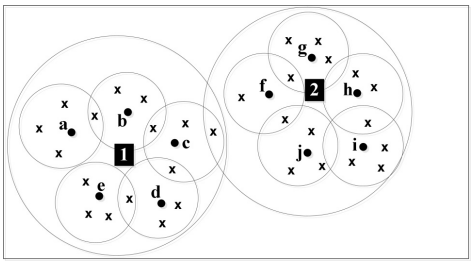
clustering



- Para um grande número de documentos será necessário um número excessivo de clusters com seus respectivos centroides, o que afetaria a eficiência do um sistema.
- Propõe-se então utilizar recursivamente a mesma metodologia, agora para criar superclasses compostas por agrupamentos de centroides, cada qual representada por um supercentroide.
- O corpus contendo 31 documentos (**x**) divididos em dez classes ou grupos, cada qual com o seu centroide (●) identificado por uma letra, e duas superclasses (*superclusters*) com seus respectivos supercentroides (■) identificados por números.

O Sistema SMART

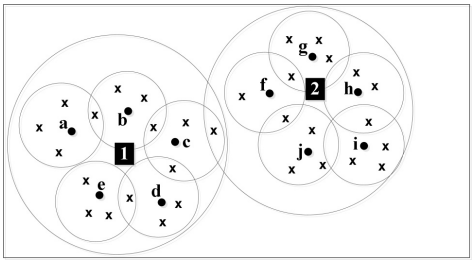
clustering



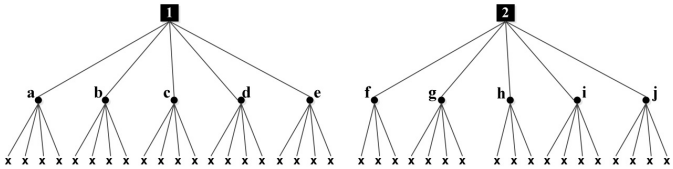
- Os dois círculos maiores formam duas superclasses (*superclusters*) com seus respectivos supercentroides (quadrados numerados)

O Sistema SMART

clustering



- À medida que o número de documentos aumenta, novas camadas de clusters podem ser criadas formando uma estrutura semelhante a uma árvore balanceada (B-tree)



A relevância no modelo vetorial

A Relevância no Modelo Vetorial

- Características do Modelo Vetorial
  - Utiliza pesos tanto para os termos de indexação quanto para os termos da expressão de busca. Esta característica permite o cálculo de um valor numérico que representa a relevância de cada documento em relação à busca;
  - O resultado de uma busca é um conjunto de documentos ordenados pelo grau de similaridade da expressão de busca e cada documento do *corpus*;
  - **Esse ordenamento permite restringir o resultado a um número máximo de documentos desejados. É possível também restringir a quantidade de documentos recuperados definindo um limite mínimo para o valor da similaridade;**

A Relevância no Modelo Vetorial

- Diferentemente do modelo booleano, o modelo vetorial utiliza pesos tanto para os termos de indexação quanto para os termos da expressão de busca.
- Essa homogeneidade é a característica fundamental que permite uma grande variedade de operações relacionadas à recuperação de informação, incluindo indexação, *clustering* (agrupamento), *relevance feedback*, classificação, reformulação da expressão de busca etc.

A Relevância no Modelo Vetorial

- O modelo de espaço vetorial assume que os termos são independentes;
- O fato de um termo ocorrer não diz nada sobre a ocorrência de outro termo;
- Isso é visto como uma limitação, mas as implicações dessa limitação ainda são debatidas;
- Uma limitação do modelo vetorial diz respeito à sua dificuldade em especificar relações frasais ou de sinonímia entre os termos das expressões de busca, pois não permite a utilização de operadores lógicos ou operadores de proximidade como no modelo booleano.

Modelo Probabilístico

Modelo Probabilístico

- Na matemática, a teoria das probabilidades estuda os experimentos aleatórios que, repetidos em condições idênticas, podem apresentar resultados diferentes e imprevisíveis.
- Isso ocorre, por exemplo, quando se observa a face superior de um dado após o seu lançamento, ou quando se verifica o naipe de uma carta retirada de um baralho.
- Por apresentarem resultados imprevisíveis, é possível apenas estimar a possibilidade ou a chance de um determinado evento ocorrer.

Modelo Probabilístico

- Espaço amostral (S) = conjunto dos possíveis resultados do experimento.
- No lançamento de um dado, por exemplo, o conjunto dos possíveis resultados é {1, 2, 3, 4, 5, 6}.
- Evento (E) = conjunto dos valores de interesse em um determinado experimento.
- No lançamento de um dado, por exemplo, pode-se estar interessado nos números pares {2, 4, 6}.

$$\mathrm{p}(E) = \frac{\mathrm{n}(E)}{\mathrm{n}(S)}$$

Modelo Probabilístico

- A probabilidade de um evento elementar *E* ocorrer em um espaço amostral *S* é a razão entre o número de elementos de *E*, simbolizado por **n(*E*)** e o número de elementos de *S* (**n(*S*)**).

$$p(E) = \frac{n(E)}{n(S)}$$

- No lançamento de um dado o espaço amostral é *S*={1, 2, 3, 4, 5, 6} e a probabilidade de sair um número par (*E*= { 2, 4, 6 }) é:

$$p(\{2,4,6\}) = \frac{n(E)}{n(S)} = \frac{3}{6} = 0.5$$

Modelo Probabilístico

- Quando dois eventos se mostram dependentes, o cálculo da probabilidade envolve as chamadas **Probabilidades Condicionais**. A probabilidade da ocorrência de um evento *A*, sabendo-se que o evento *B* ocorreu, é calculada como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Modelo Probabilístico

- O modelo probabilístico foi proposto inicialmente por Maron e Kuhns (1960) e posteriormente explorado por diversos outros pesquisadores;
- Utilização do processo de ***Relevance Feedback*** para a progressiva melhoria dos resultados de uma busca através de cálculos de probabilidade

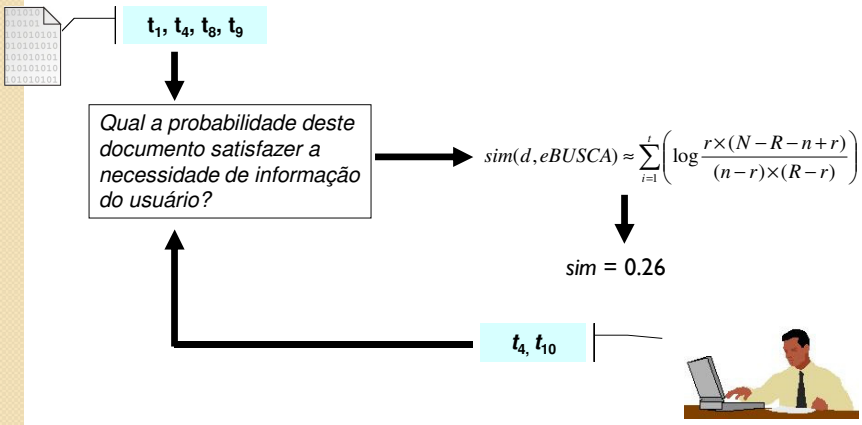
Modelo Probabilístico

- Todo cálculo de probabilidade resume-se a um problema de contagem. Portanto, para uma determinada expressão de busca, pode-se representar os documentos do corpus da seguinte forma:

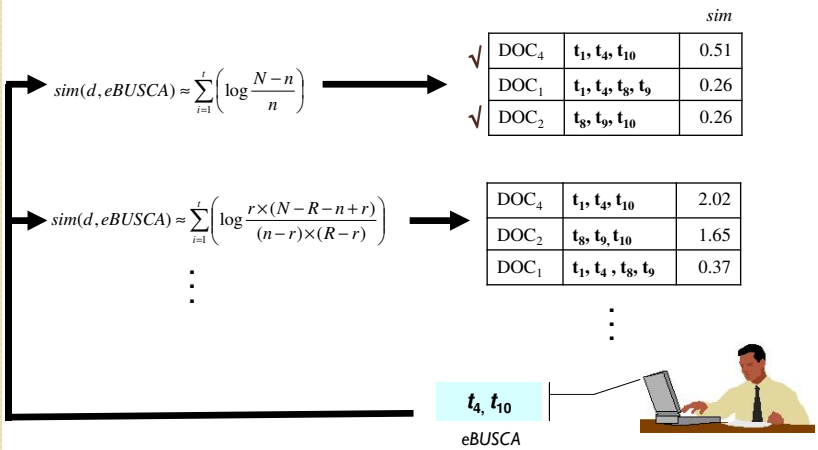
	Relevante	**não-Relevante**	
documento contendo t_i	r	n-r	n
documento que não contém t_i	R-r	N-R-n+r	N-n
	R	N-R	N

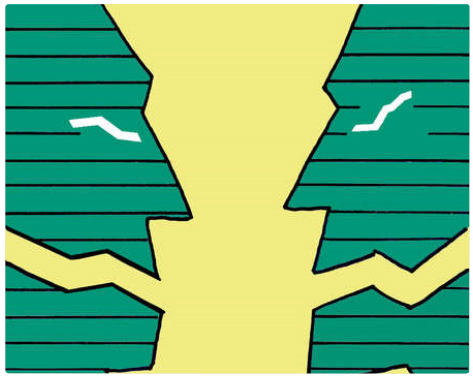
- Considerando um corpus com **N** documentos e um determinado termo $\mathbf{t_i}$, existe no *corpus* um total de **n** documentos indexados por $\mathbf{t_i}$. Desses **n** documentos apenas **r** são relevantes.

Modelo Probabilístico



Modelo Probabilístico





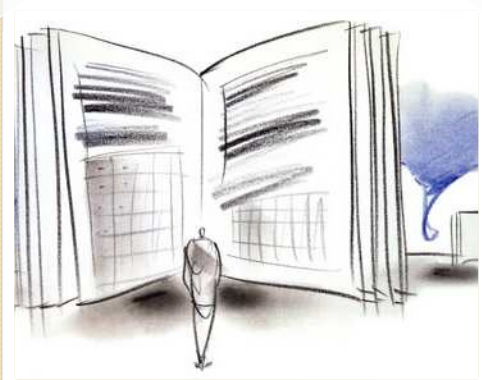
A relevância no modelo probabilístico

A Relevância no Modelo Probabilístico

- O processo de recuperação de informação é caracterizado por seu grau de incerteza no julgamento de relevância dos documentos em relação à expressão de busca;
- Portanto, é mais realístico pensar em uma **probabilidade de relevância** do que em uma pretensa relevância exata, como a utilizada nos modelos booleano e vetorial.
- O modelo probabilístico reconhece que a atribuição de relevância é uma tarefa do usuário. É o único modelo que incorpora explicitamente o processo de *relevance feedback* como base para a sua operacionalização.

A Relevância no Modelo Probabilístico

- Pode ser facilmente implementado por meio da estrutura proposta pelo modelo vetorial, permitindo integrar as vantagens desses dois modelos em um mesmo sistema de recuperação de informação.
- A sua complexidade desencoraja muitos desenvolvedores de sistema a abandonar os modelos booleano e vetorial (CHU, 2010, p.120; JONES; WALKER; ROBERTSON, 2000).



Referências bibliográficas

Referências bibliográficas

- BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. 2ª ed. Addison-Wesley, 2011.
- CHU, H. **Information Representation and Retrieval in the Digital Age**, Second Edition, Medford, N.J.: Information Today, 2010. (ASIST monograph series)
- JONES, K.S.; WALKER, S.; ROBERTSON, S.E. A probabilistic model of information retrieval: development and comparative experiments – Part 2. **Information Processing and Management**, v. 36, n. 6, 2000. p.809-840.
- MARON, M.E.; KUHNS, J.L. On relevance, probabilistic indexing and information retrieval. **Journal of the ACM**, v. 7, n. 3, 1960, p.216-244.
- SALTON, G.; McGILL, M. J. Introduction to Modern Information Retrieval. McGraw Hill, 1983.
- SALTON, G.; BUCKLEY, C. Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, v. 24, n. 5, 1988. p.513-523.
- SMITH, E.S. On the shoulders of giants: from Boole to Shannon to Taube: the origins and development of computerized information from the mid-19th century to the present. **Information Technology and Libraries**, n. 12, 1993 (june). p.217-226.