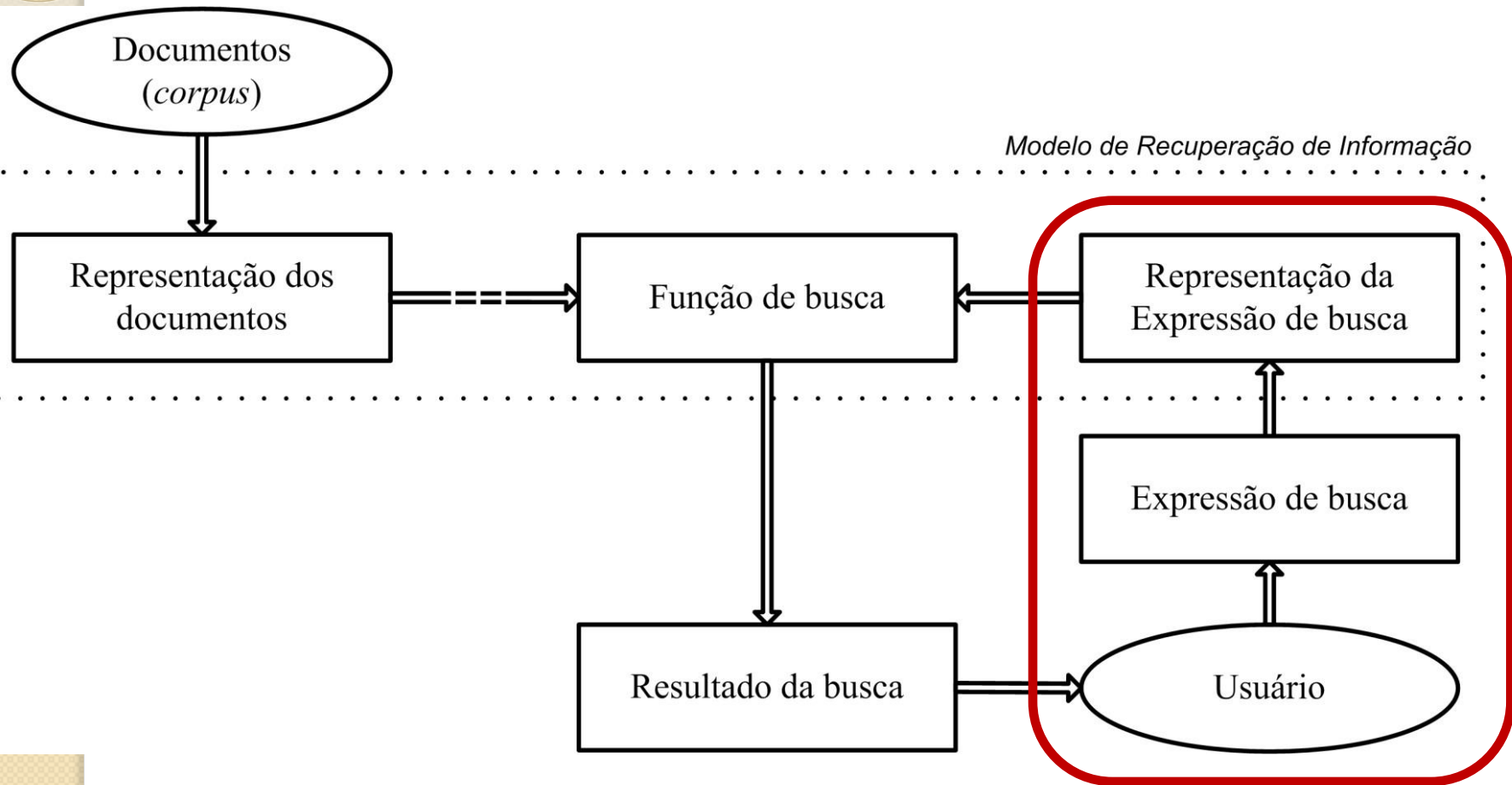
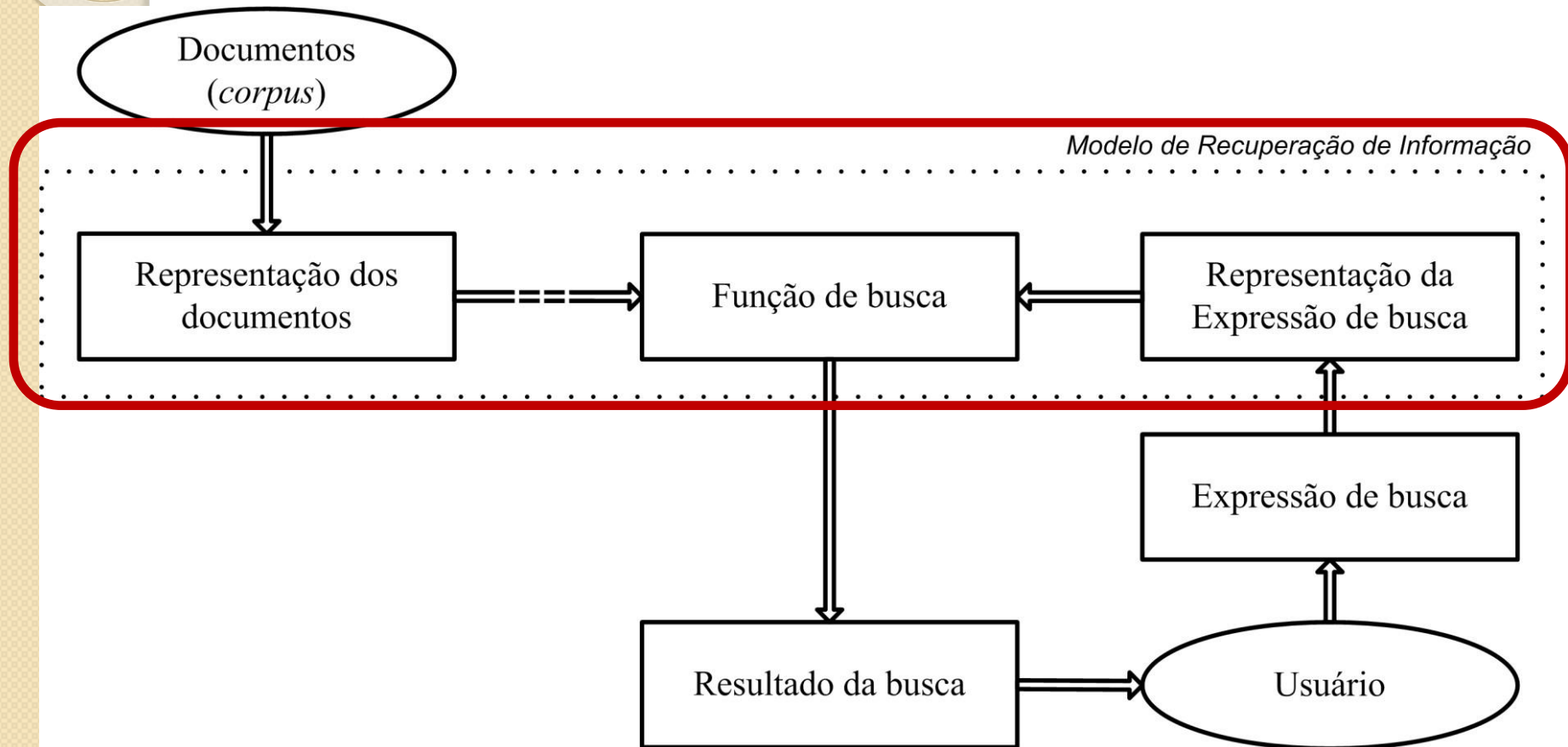




Onde estamos



Para onde vamos



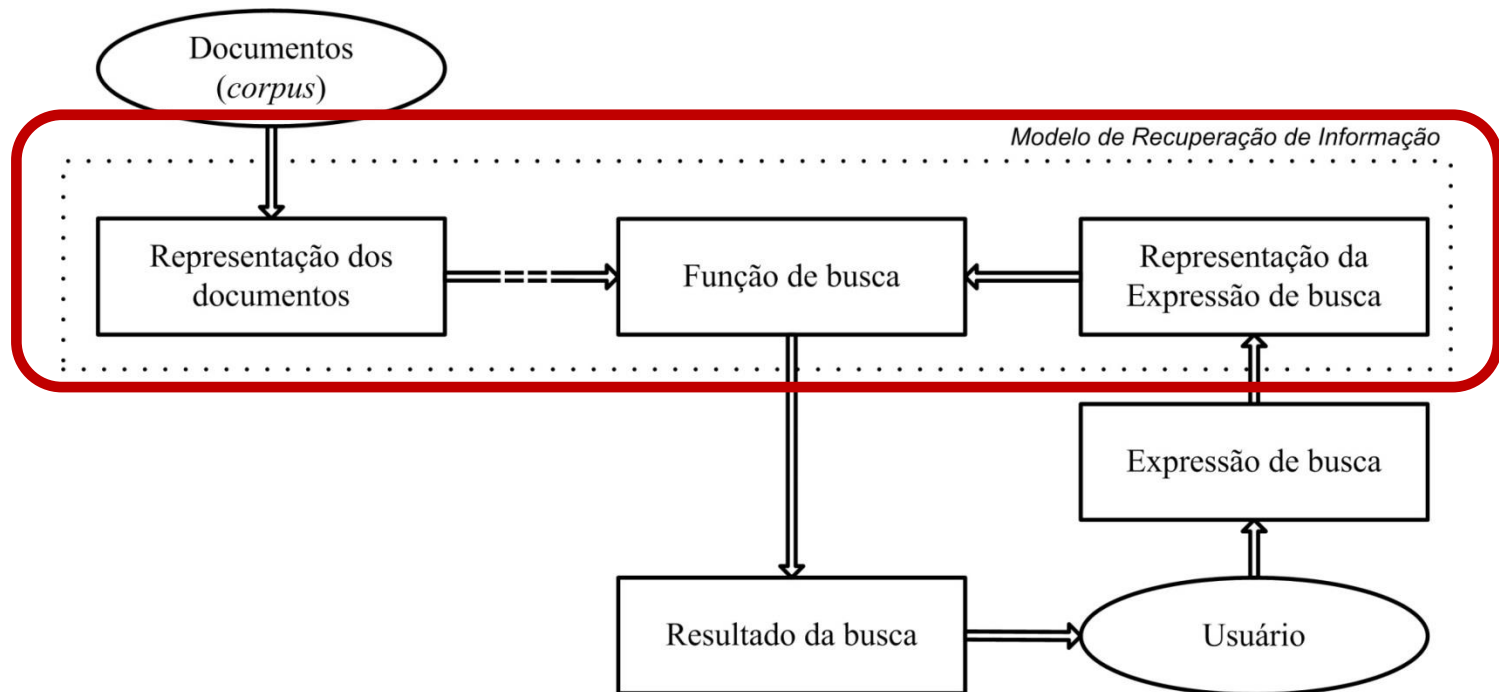
Sumário/Roteiro

- Modelo de Recuperação de Informação
- Modelos Clássicos
 - Modelo Booleano
 - A relevância no modelo booleano
 - Modelo Vetorial
 - O Sistema SMART
 - A relevância no modelo vetorial
 - Modelo Probabilístico
 - A relevância no modelo probabilístico
- Referências



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*.

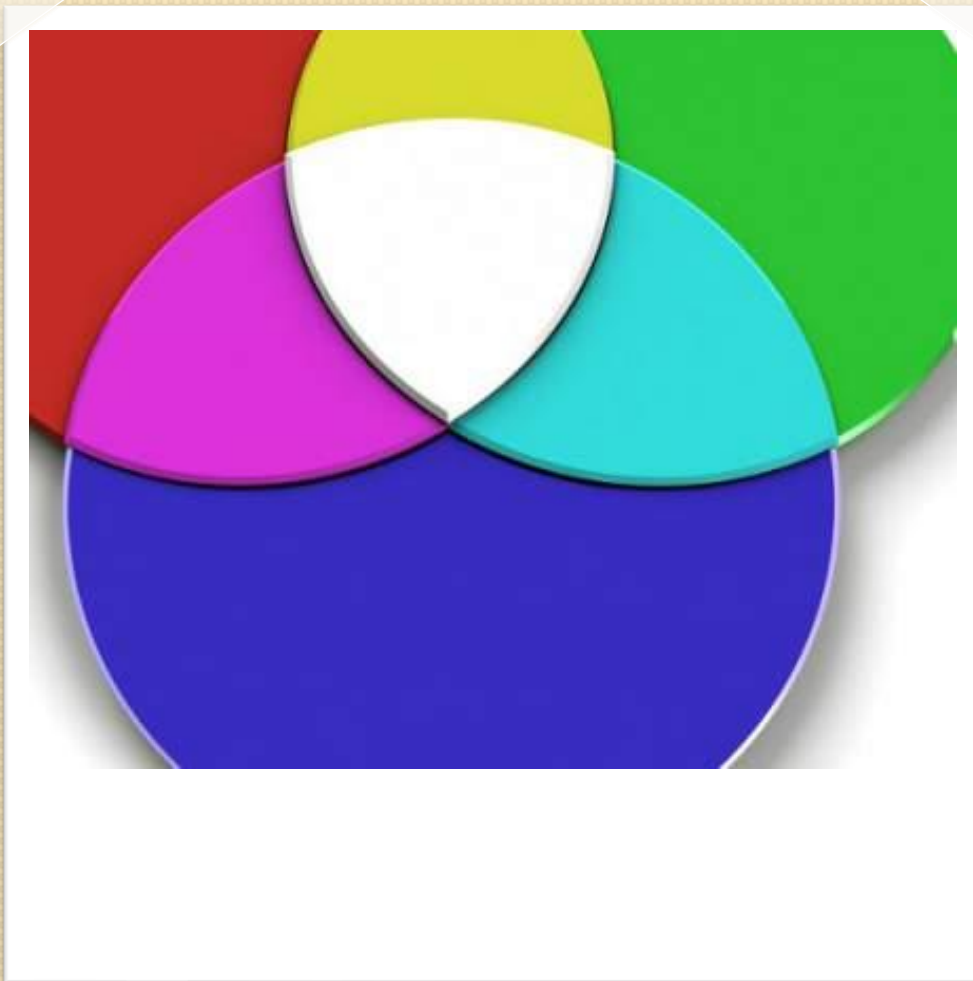
Modelo de Recuperação de Informação

- A eficiência e os recursos 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 Clásicos

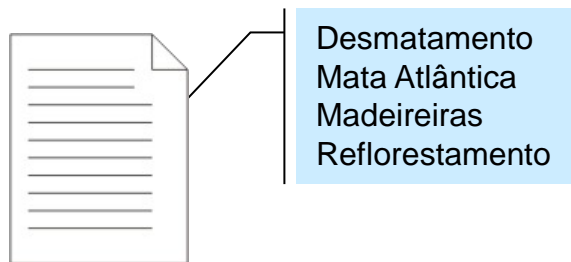
Modelos Clásicos



Modelo Booleano

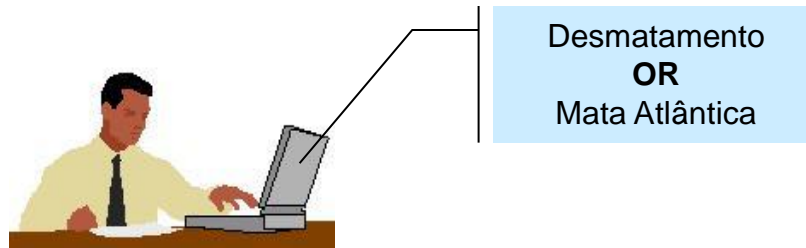
Modelo Booleano

- 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.



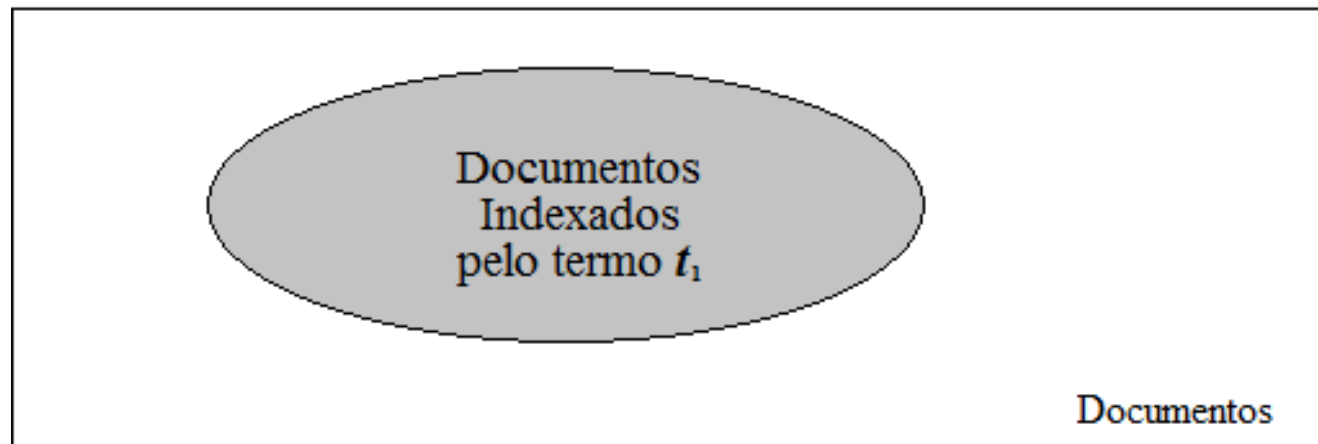
Modelo Booleano

- 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 t_1 terá como resultado o conjunto de documentos indexados por t_1 ;



Modelo Booleano



Desmatamento



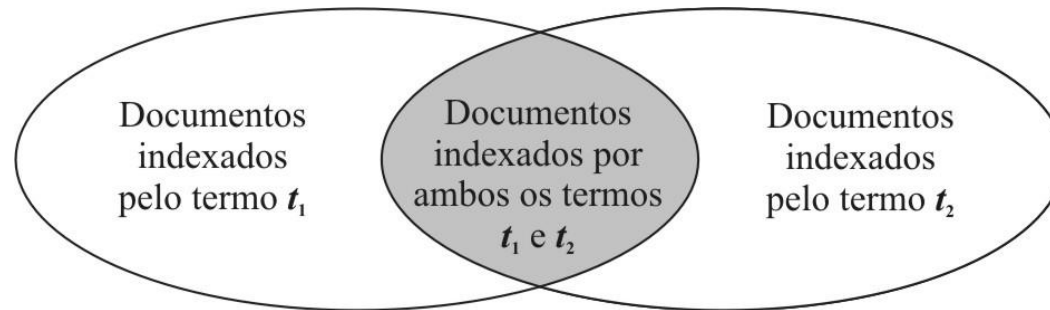
Desmatamento
Mata Atlântica
Madeiras
Reflorestamento



Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

Modelo Booleano

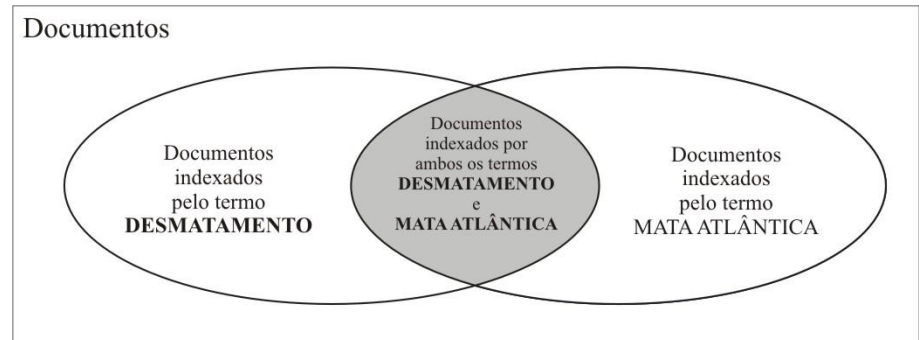
- Uma expressão conjuntiva de enunciado t_1 **AND** t_2 recuperará documentos indexados por ambos os termos (t_1 e t_2).
- Esta operação equivale à *interseção* do conjunto dos documentos indexados pelo termo t_1 com o conjunto dos documentos indexados pelo termo t_2 , representado pela área cinza na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



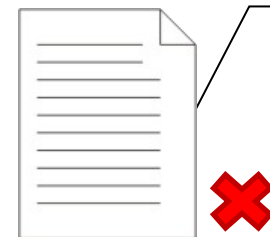
Modelo Booleano



Desmatamento
AND
Mata Atlântica



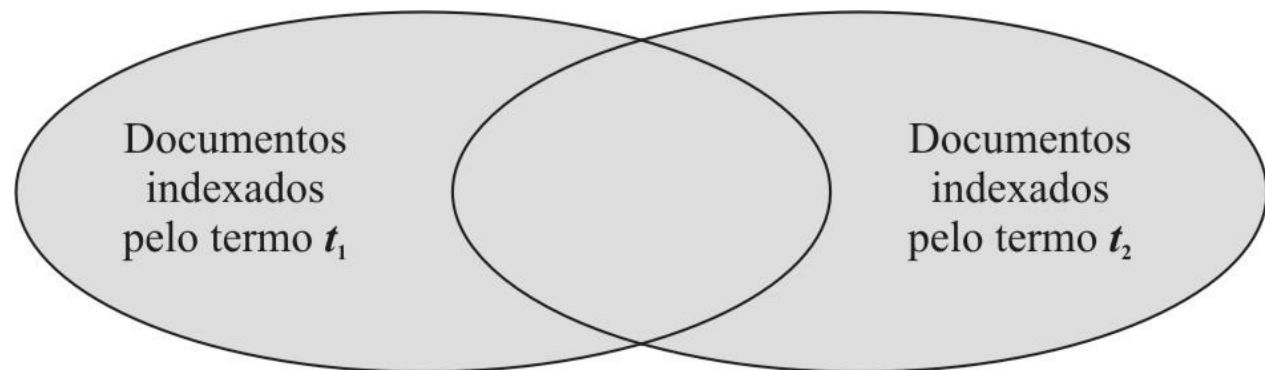
Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento



Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

Modelo Booleano

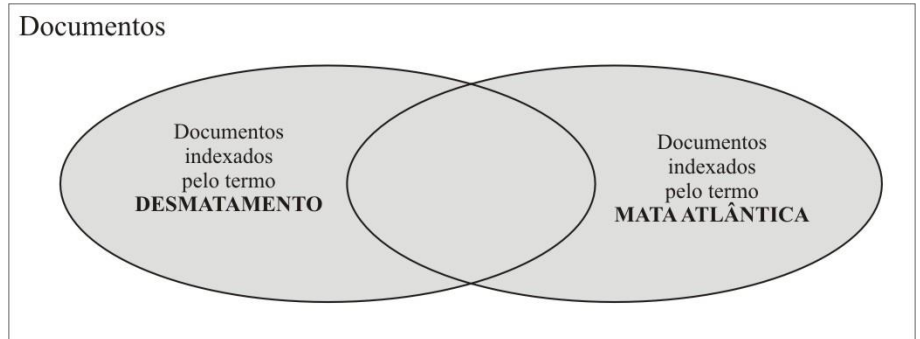
- Uma expressão disjuntiva t_1 **OR** t_2 recuperará o conjunto dos documentos indexados pelo termo t_1 ou pelo termo t_2 .
- Essa operação equivale à *união* entre o conjunto dos documentos indexados pelo termo t_1 e o conjunto dos documentos indexados pelo termo t_2 .



Modelo Booleano



Desmatamento
OR
Mata Atlântica



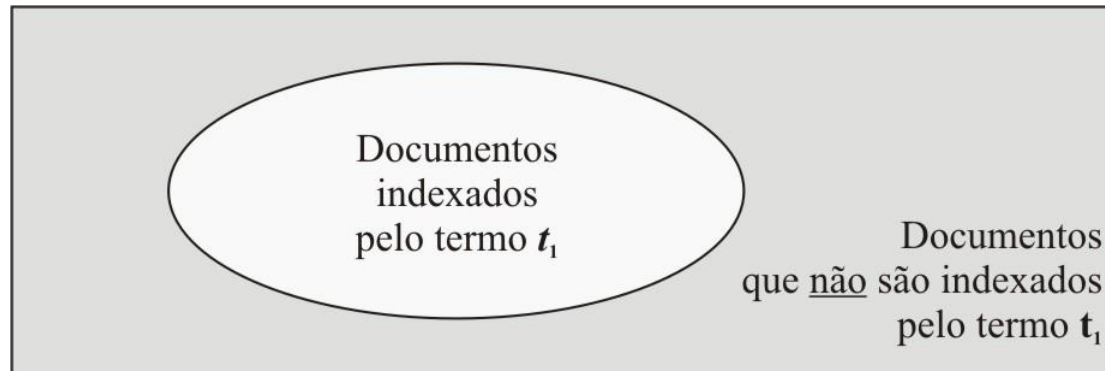
**Desmatamento
Mata Atlântica**
Madeireiras
Reflorestamento



**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



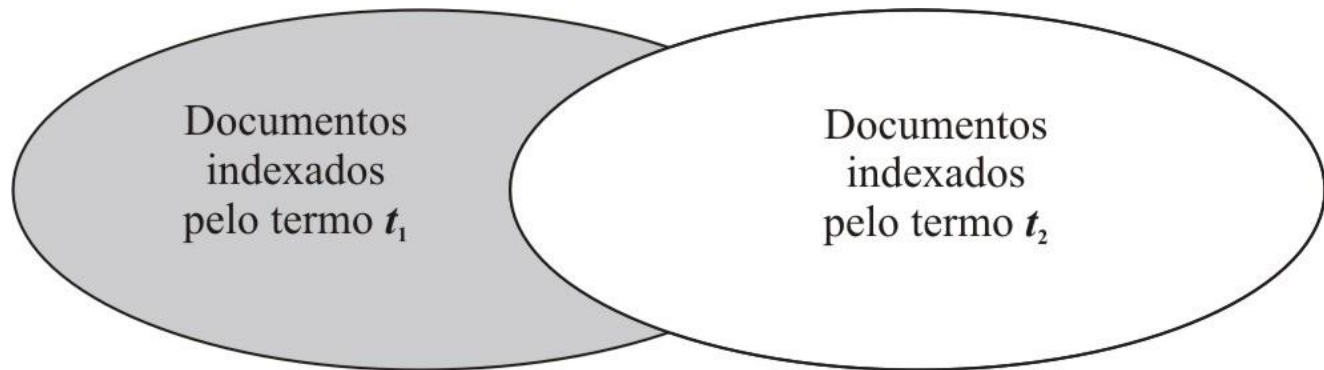
Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento



Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

Modelo Booleano

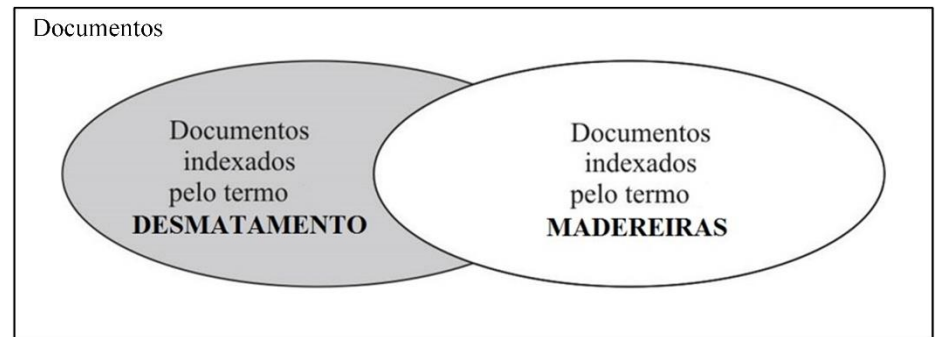
- As expressões t_1 **NOT** t_2 ou t_1 **AND NOT** t_2 terão o mesmo resultado: o conjunto dos documentos indexados por t_1 e que não são indexados por t_2 .
- Neste caso o operador NOT pode ser visto como um operador da diferença entre conjuntos.



Modelo Booleano



Desmatamento
AND NOT Madeireiras



Desmatamento
Mata Atlântica
Madeiras
Reflorestamento



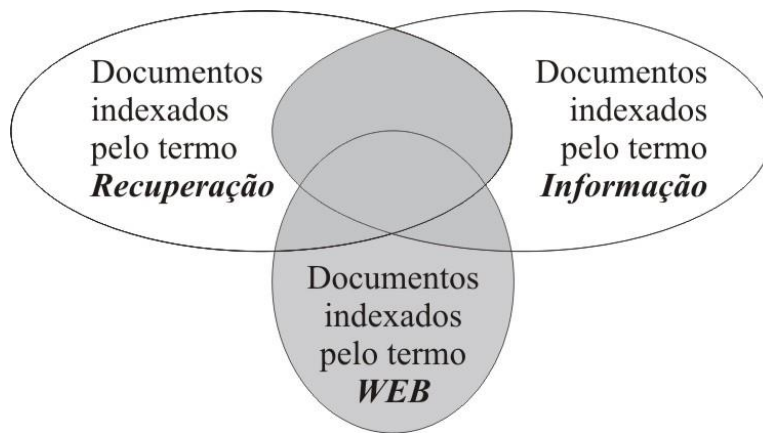
Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

Modelo Booleano

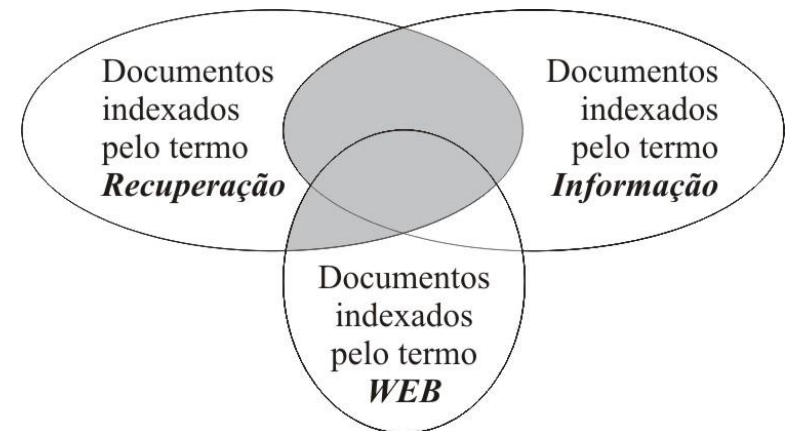
- Termos e operadores booleanos podem ser combinados para especificar buscas mais amplas ou restritivas.
- Como a ordem de execução das operações lógicas de uma expressão influencia no resultado da busca, muitas vezes é necessário explicitar essa ordem, delimitando partes da expressão por meio de parênteses.

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.



$(\text{Recuperação AND Informação}) \text{ OR WEB}$

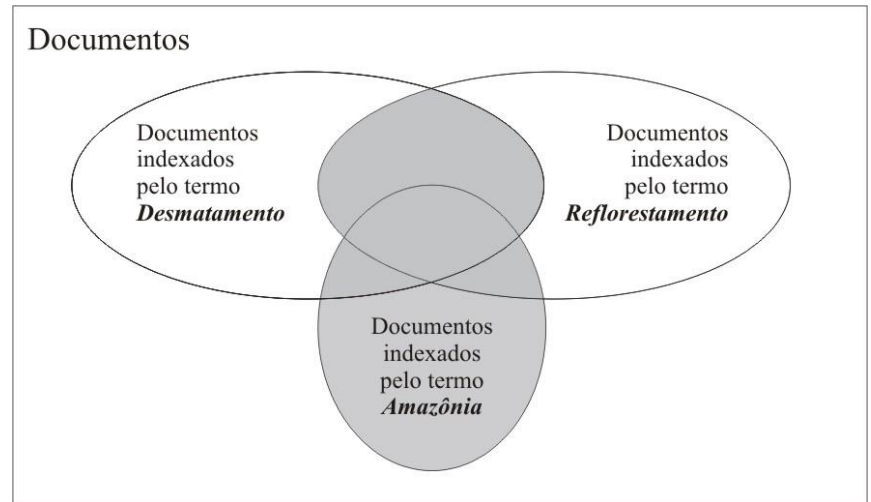


$\text{Recuperação AND (Informação OR WEB)}$

Modelo Booleano



(Desmatamento **AND** Reflorestamento)
OR
Amazônia



Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

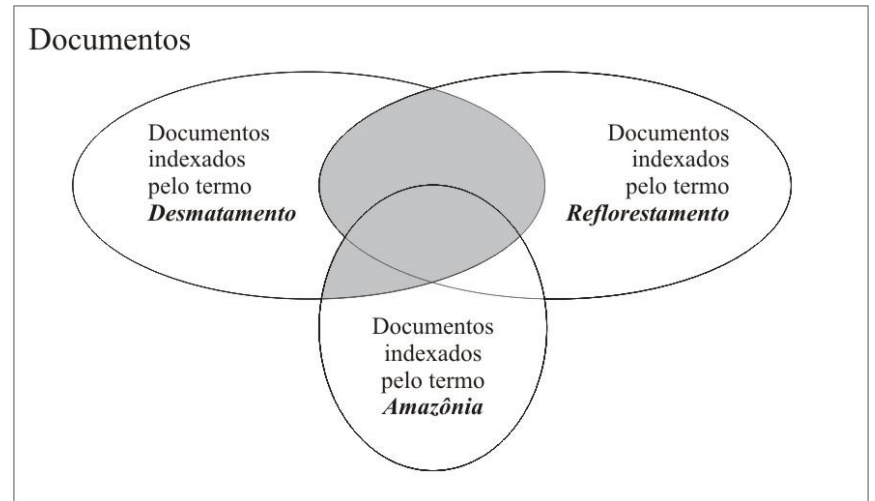


Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

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** entre o conjunto dos “gatos” e 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 incapacidade 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

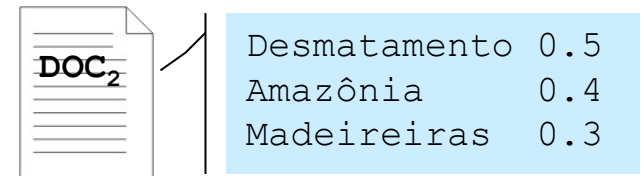
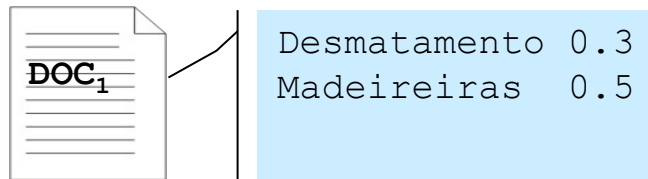
Modelo Vetorial

- 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 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.



Desmatamento	0.2
Amazônia	0.35
Madeireiras	0.1

Usuário e sua
necessidade de informação

Modelo Vetorial

cálculo da similaridade

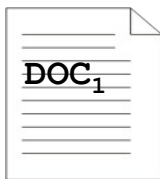
- A utilização de uma mesma forma de representação tanto para os documentos como para as expressões de busca permite calcular a **similaridade** entre uma expressão de busca e cada um dos documentos do *corpus*, ou ainda entre dois documentos;
- Em um espaço vetorial contendo ***N*** dimensões, a similaridade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

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

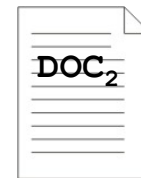
$w_{i,j}$ é o peso do i -ésimo termo do documento d_j e $w_{i,q}$ é o peso do i -ésimo termo da expressão de busca q .

Modelo Vetorial

cálculo da similaridade entre documentos



Desmatamento	0.3
Madeireiras	0.5



Desmatamento	0.5
Amazônia	0.4
Madeireiras	0.3

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

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

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

Modelo Vetorial

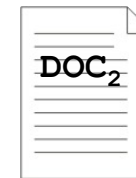
cálculo da similaridade entre busca e documentos



Desmatamento	0.2
Amazônia	0.35
Madeireiras	0.1



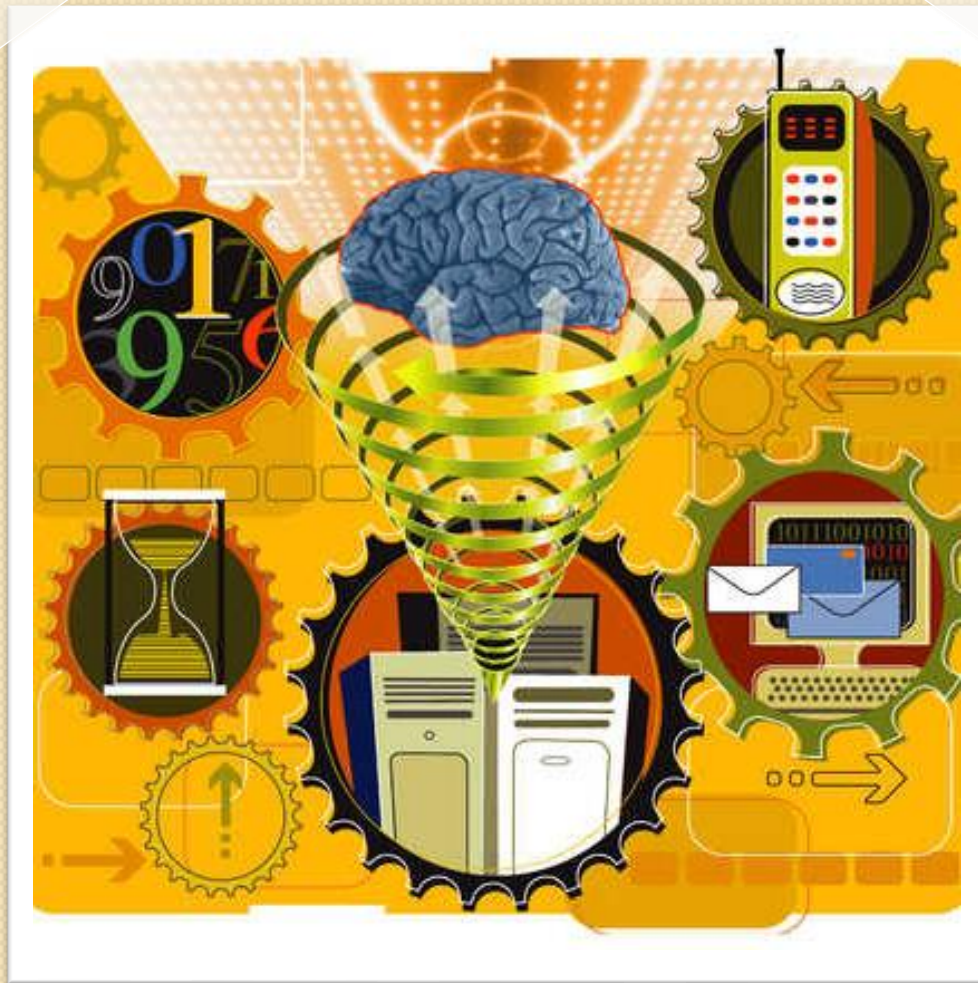
Desmatamento	0.3
Madeireiras	0.5



Desmatamento	0.5
Amazônia	0.4
Madeireiras	0.3

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

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



O sistema SMART

O sistema SMART

- O projeto SMART (*System for the **M**anipulation 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

- Processo de **Indexação**

1. Identificação dos termos/palavras de um texto (*tokenizer*)
2. Eliminação de *Stop Words*
 - As chamadas *Stop Words* são palavras semanticamente pobres para representar o conteúdo de um documento, por exemplo: artigos, preposições, etc.
3. Normalização de Termos - *Stemming*
 - Remover os sufixos e (possivelmente também os prefixos) para se chegar ao radical (ou uma forma normalizada) da palavra;
 - Reduzir a variabilidade lexical;
4. Identificação de termos compostos;
5. Cálculo do peso de cada termo de indexação (*tf-idf*)

O sistema SMART

- Cálculo do **peso dos termos de indexação**.
 - **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 $tf \times idf$ e possui a seguinte fórmula:

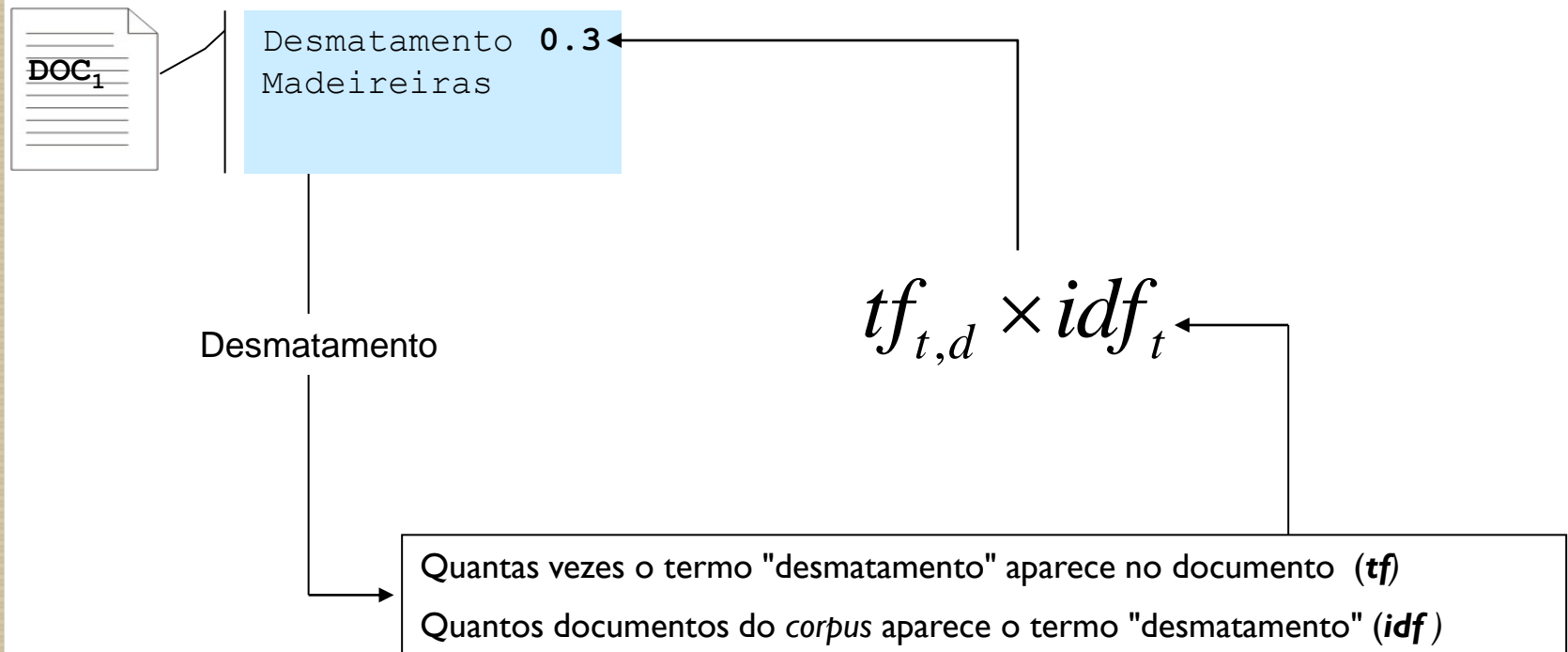
$$w_{t,d} = tf_{t,d} \times idf_t$$

- A medida $tf \times idf$ 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.

Salton e McGill (1983, p.204-207)

O sistema SMART

Calculo automático dos pesos dos termos de indexação



O sistema SMART

- Cada termo da expressão 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:

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

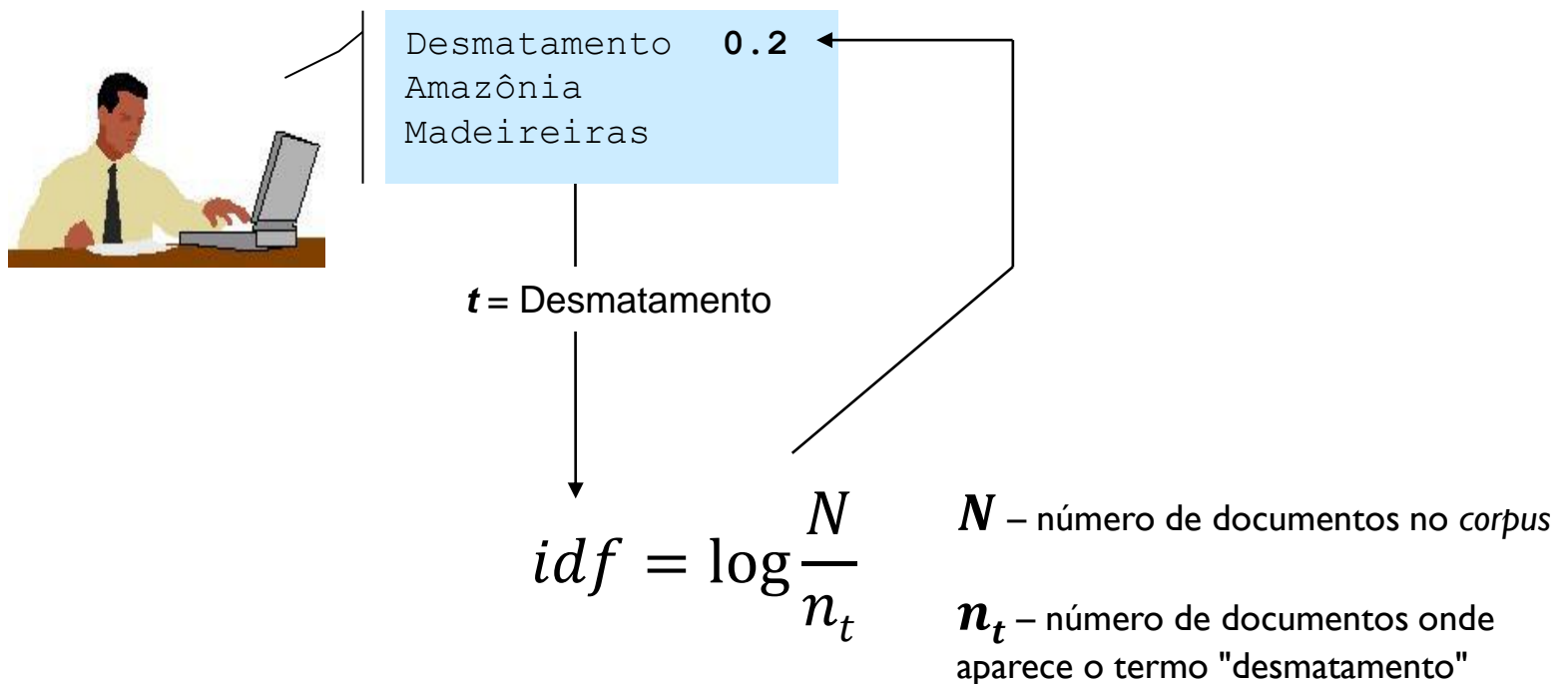
Inverse document frequency clássico

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

best weighted probabilistic weight

O sistema SMART

Calculo automático dos pesos da expressão de busca



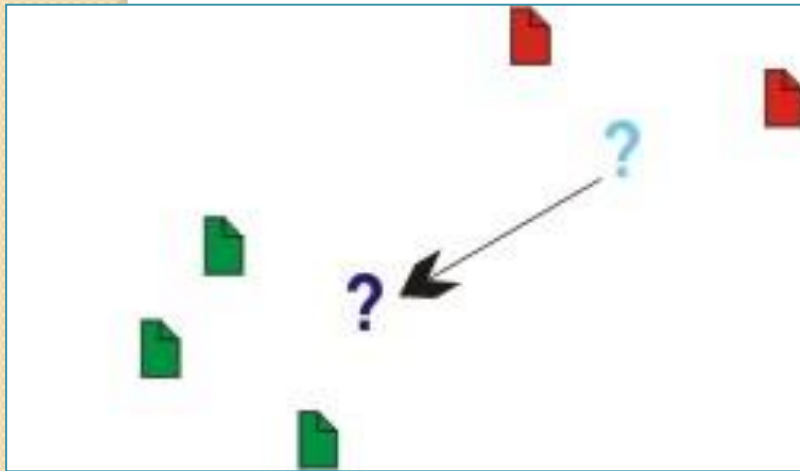
O sistema SMART

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

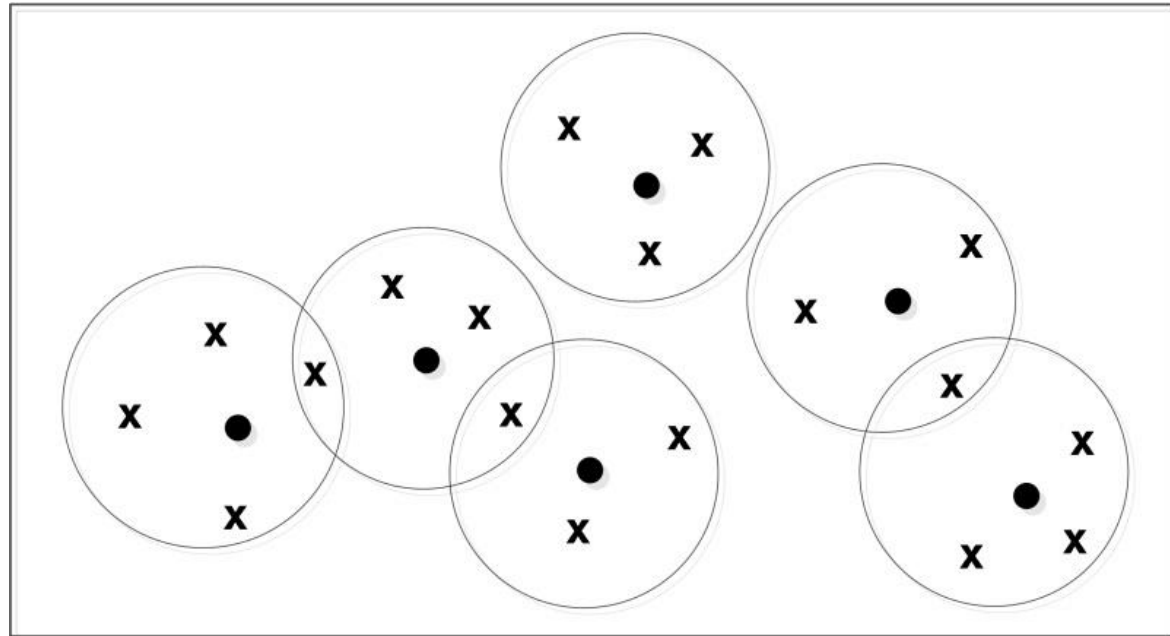
 Expressão de busca original

 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

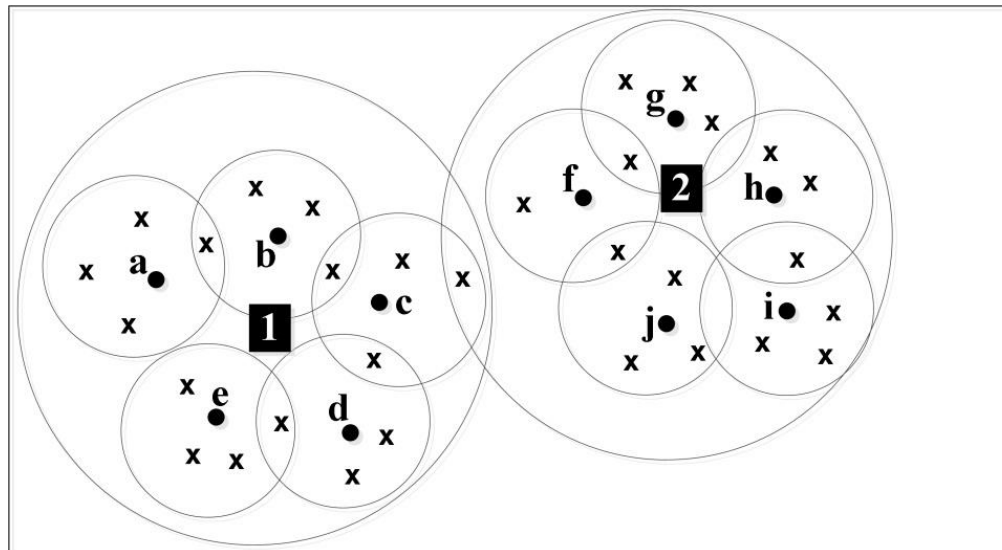
clustering



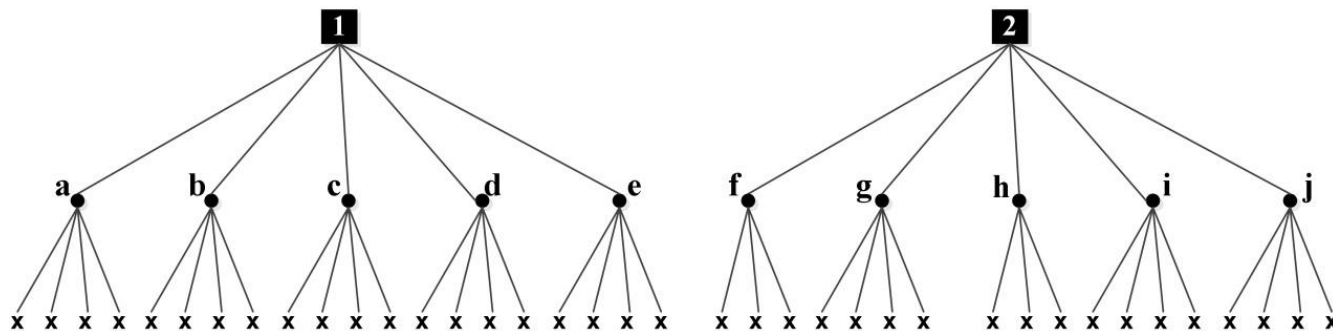
Clustering
(agrupamento)

- Cada x representa o vetor de um documento
- As circunferências representam os *clusters*
- Pode ser observado que alguns grupos se interseccionam, possuindo documentos em comum.
- 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*



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





**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

- 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 $\{2, 4, 6\}$.

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

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:

$$p(A | B) = \frac{p(A \text{ e } B)}{p(B)}$$

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;
- No modelo probabilístico os documentos e as buscas são representados por um conjunto de termos apenas, sem pesos.

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_i	$n_i - r_i$	n_i
documento que não contém t_i	$R - r_i$	$N - n_i - (R - r_i)$	$N - n_i$
	R	$N - R$	N

- **N** número de documentos no corpus
- n_i número de documentos que contém o termo t_i
- R número de documentos relevantes para uma dada consulta;
- r_i número de documentos relevantes que contém o termo t_i

Modelo Vetorial

cálculo da similaridade entre busca e documentos

1ª BUSCA

ainda não há informações sobre relevância

$$\sum_{i=1}^t \log \left(\frac{N + 0,5}{n_i + 0,5} \right)$$



Desmatamento
Amazônia

0,00	<input checked="" type="checkbox"/>	DOC ₁	Queimadas Madeireiras
0,73	<input checked="" type="checkbox"/>	DOC ₂	Desmatamento Amazônia Grilagem
0,26	<input checked="" type="checkbox"/>	DOC ₃	Desmatamento Mata Atlântica Madeireiras
0,00	<input type="checkbox"/>	DOC ₄	Reflorestamento Serrado

Modelo Vetorial

cálculo da similaridade entre busca e documentos

2ª BUSCA

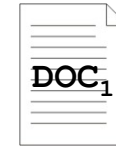
há informações sobre relevância

$$\sum_{i=1}^t \left(\log \frac{(r_i + 0,5)(N - n_i - R + r_i + 0,5)}{(R - r_i + 0,5)(n_i - r_i + 0,5)} \right)$$



Desmatamento
Amazônia

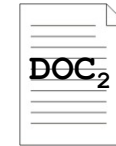
0,20



Queimadas
Madeireiras



0,80



Desmatamento
Amazônia
Grilagem



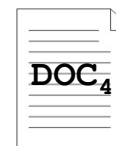
0,40



Desmatamento
Mata Atlântica
Madeireiras



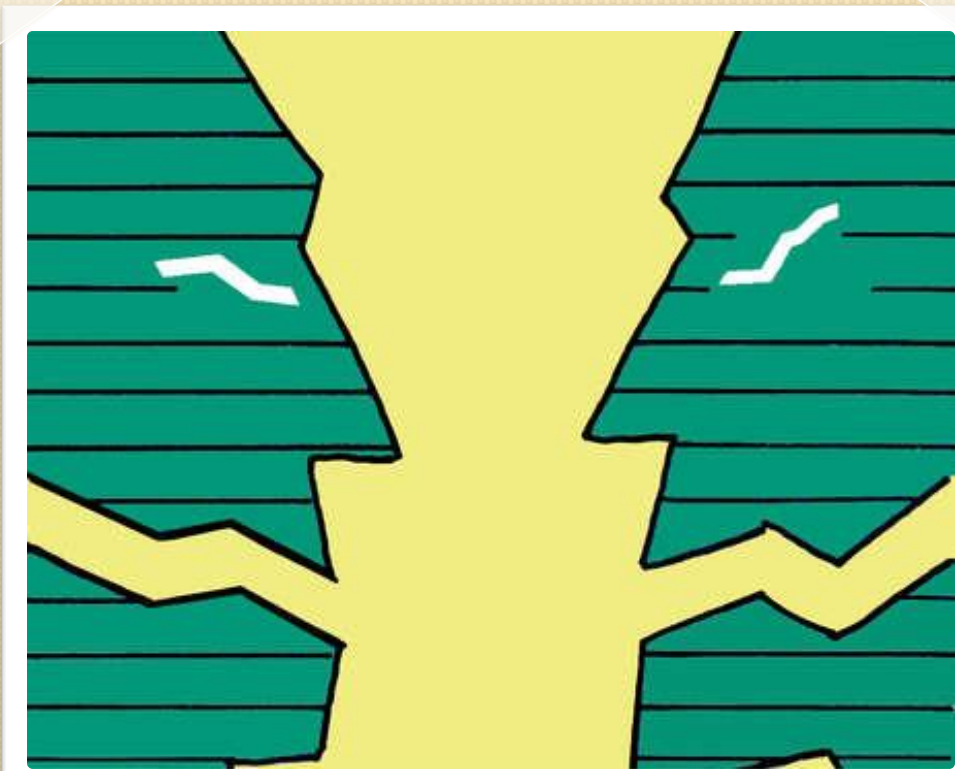
0,00



Reflorestamento
Serrado

O usuário pode novamente marcar os documentos que considera relevantes e submeter ao sistema.

Esse "diálogo" pode se prolongar até que o usuário fique satisfeito com os resultados



**A relevância no
modelo probabilístico**

A Relevância no Modelo Vetorial

- 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 Vetorial

- A sua complexidade desencoraja muitos desenvolvedores de sistema a abandonar os modelos booleano e vetorial (CHU, 2010, p.120; JONES; WALKER; ROBERTSON, 2000).



Referências

Referências

- 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.; KUHN, 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.