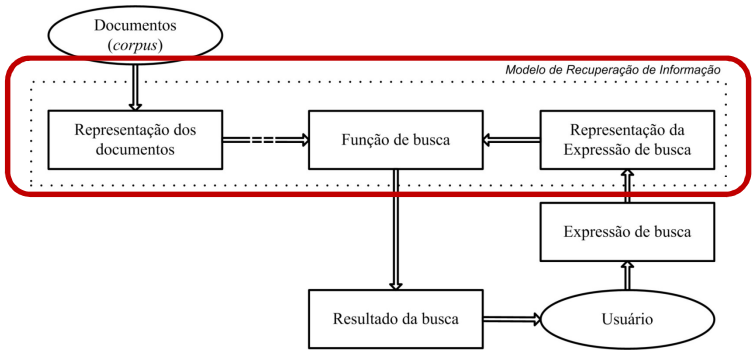
Modelo de Recuperação de Informação

Modelo de Recuperação de Informação



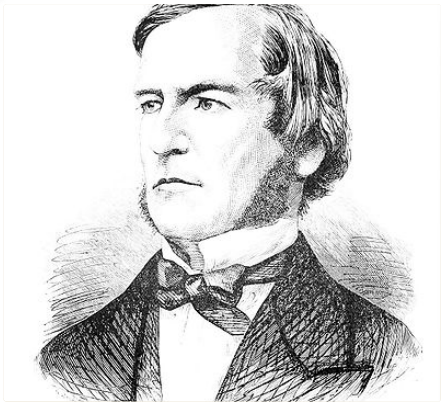
- Um modelo de recuperação de informação é a especificação formal de três elementos:
  - a *representação dos documentos;*
  - a *representação da necessidade* de informação por meio de uma *expressão de busca;* e
  - como estes dois elementos serão comparados, a *função de busca.*

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



George Boole
(1815 - 1864)

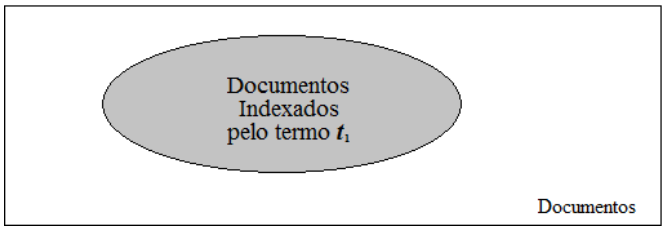
Modelo Booleano

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.

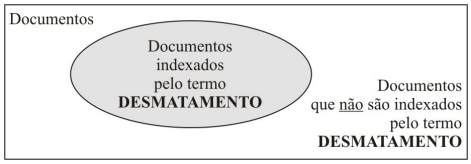
Modelo Booleano

- 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

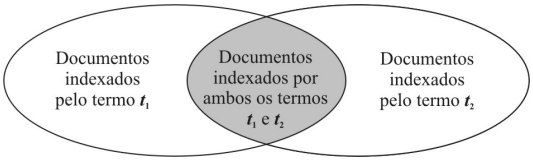


Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

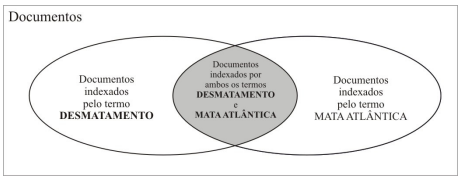
Modelo Booleano

- 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

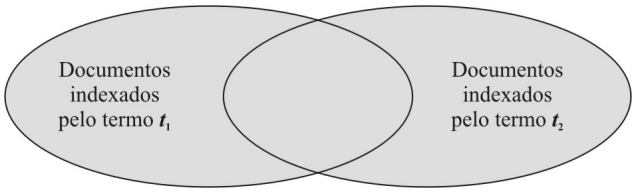


Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

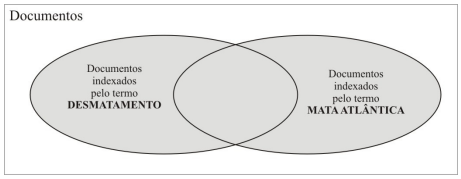
Modelo Booleano

- 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

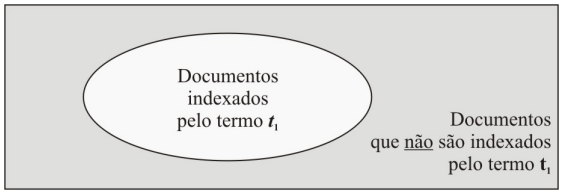


Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

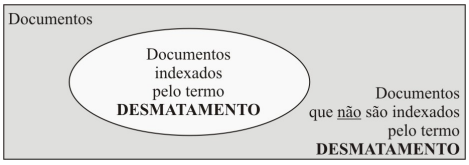
Modelo Booleano

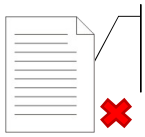
- A expressão **NOT t_1** recuperará os documentos que **não** são indexados pelo termo **t_1**, representados pela área cinza da figura.



Modelo Booleano

NOT Desmatamento



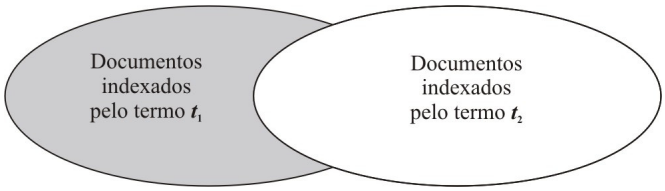


Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

Desmatamento
Amazônia
Grilagem de terras
Reflorestamento

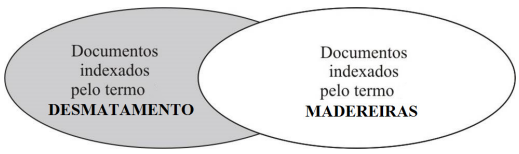
Modelo Booleano

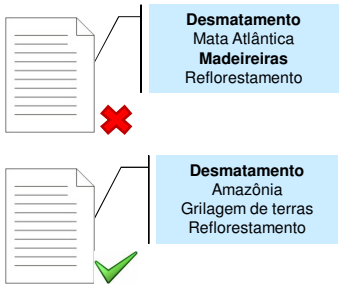
- As expressões t_1 **NOT** t_2 ou t_1 **AND NOT** t_2 terão o mesmo resultado: o conjunto dos documentos indexados por t_1 e que não são indexados por t_2.
- Neste caso o operador NOT pode ser visto como um operador da diferença entre conjuntos.



Modelo Booleano

Desmatamento **AND NOT** Madeireiras



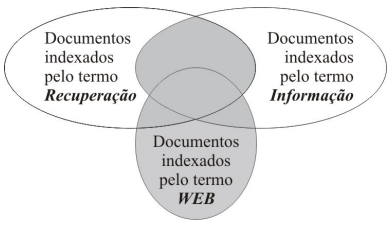


Modelo Booleano

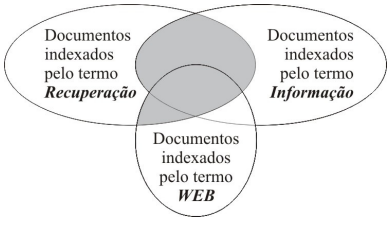
- Termos e operadores booleanos podem ser combinados para especificar buscas mais amplas ou restritivas.
- Como a ordem de execução das operações lógicas de uma expressão influencia no resultado da busca, muitas vezes é necessário explicitar essa ordem, delimitando partes da expressão por meio de parênteses.

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.



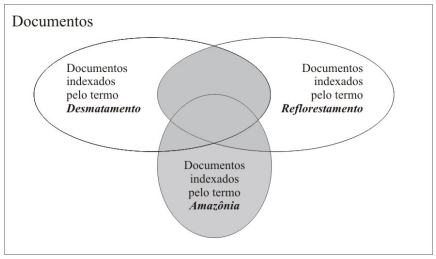
(*Recuperação* AND *Informação*) OR *WEB*



Recuperação AND (*Informação* OR *WEB*)

Modelo Booleano

(Desmatamento **AND** Reflorestamento)
OR
Amazônia

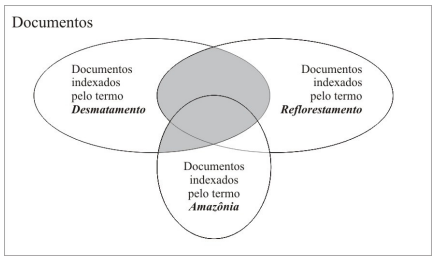


Desmatamento
Mata Atlântica
Madeireiras
Reflorestamento

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

Características do Modelo Booleano

Características do 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).

Características do 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;

Características do 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;

Gerard Salton
(1927-1995)

Modelo Vetorial

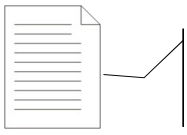
Modelo Vetorial

- 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:

- 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.

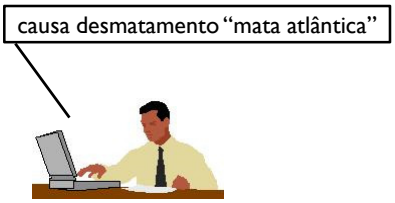
Modelo Vetorial



Desmatamento	0.7
Mata Atlântica	0.6
Madeireiras	0.3
Reflorestamento	0.2

Modelo Vetorial

- 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

Desmatamento	0.8
Mata Atlântica	0.5
Causa	0.7

Modelo Vetorial:

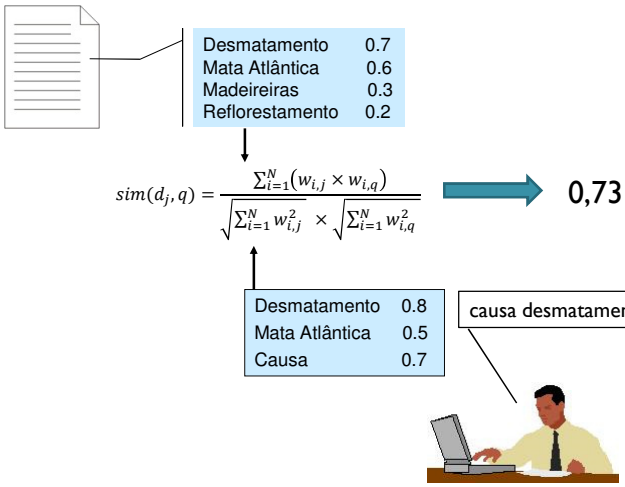
cálculo da similaridade

- 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 $\boldsymbol{d_j}$ e uma expressão de busca ***q*** pode ser calculada utilizando a seguinte fórmula:

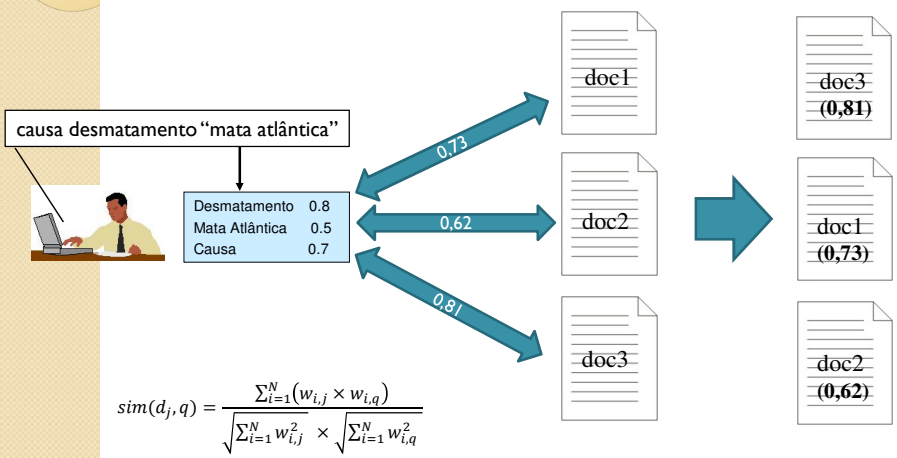
$$sim(d_j, q) = \frac{\sum_{i=1}^{N}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

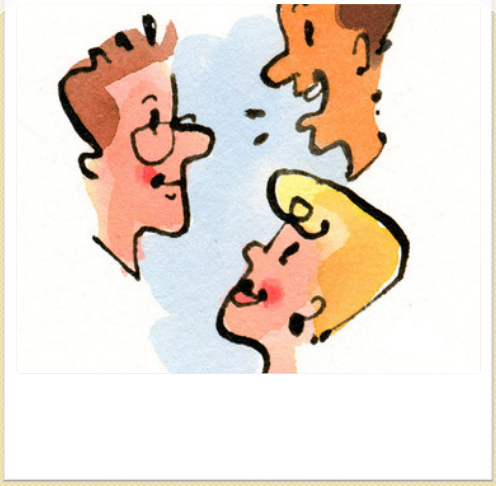
$w_{i,j}$ é o peso do *i*-ésimo termo do documento d_j e $w_{i,q}$ é o peso do *i*-ésimo termo da expressão de busca *q*.

Modelo Vetorial:
cálculo da similaridade



Modelo Vetorial:
cálculo da similaridade





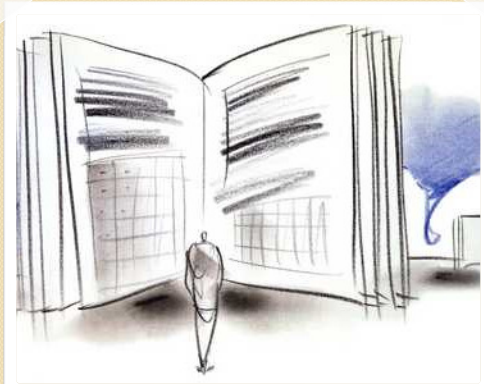
Características do modelo vetorial

Características do 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 (relevância) 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;

Características do 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.
- 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.



Referências bibliográficas

Referências bibliográficas

CHU, H. **Information Representation and Retrieval in the Digital Age**, Second Edition, Medford, N.J.: Information Today, 2010. (ASIST monograph series)

ROBERTSON, S.E.; JONES, K.S. Relevance weighting of search terms. **Journal of the Americam Society for Information Science**, v. 27, n. 3, 1976, p.129-146.

SALTON, G. Recent studies in automatic text analysis and document retrieval, Journal of the ACM, v. 20, n. 2, 1973. p.258-278

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.