GSI024 - Organização e recuperação de informação

Prof. Dr. Rodrigo Sanches Miani (FACOM/UFU)

Última atualização - Maio/2022

Modelos de RI e o Modelo Booleano

Tópicos

- Modelagem em RI;
- Caracterização de um modelo de RI;
- Recuperação de informação clássica;
- Modelo booleano.

Breve resumo da aula anterior

Aula anterior

- O objetivo da área de estudo conhecida como Recuperação de Informação é:
  - Prover aos usuários o acesso fácil as informações de seu interesse

- A diferença entre os objetivos iniciais da área e como esses objetivos mudaram com o advento da Web;

- Breve histórico.

Aula anterior

- O problema de RI está ligado basicamente a:
  - Como extrair as informações dos documentos?
  - Como utilizar tais informações para decidir sobre a sua relevância?

- Sistema de RI pode ser dividido em seis módulos:
  1. Obtenção e coleção de documentos;
  2. Indexação dos documentos;
  3. Consulta do usuário;
  4. Recuperação de documentos;
  5. Ranqueamento dos documentos;
  6. Apresentação para o usuário.

Modelagem em RI

Modelagem em RI

- Modelagem em RI é um processo complexo que tem o objetivo de produzir uma **função de ranqueamento**, ou seja, uma função que atribui valores a documentos em relação a uma consulta;

- Esse processo pode ser dividido em duas tarefas principais:
  1. A concepção de um **sistema lógico** para representar documentos e consultas (teoria de conjuntos, álgebra linear e probabilidades);
  2. A definição de uma **função de ranqueamento** que computa o grau de similaridade de cada documento em relação à consulta dada.

Caracterização de um modelo de RI

Caracterização de um modelo de RI

Um modelo de RI é uma quádrupla [D,Q,F,R(q_i,d_j)] onde:

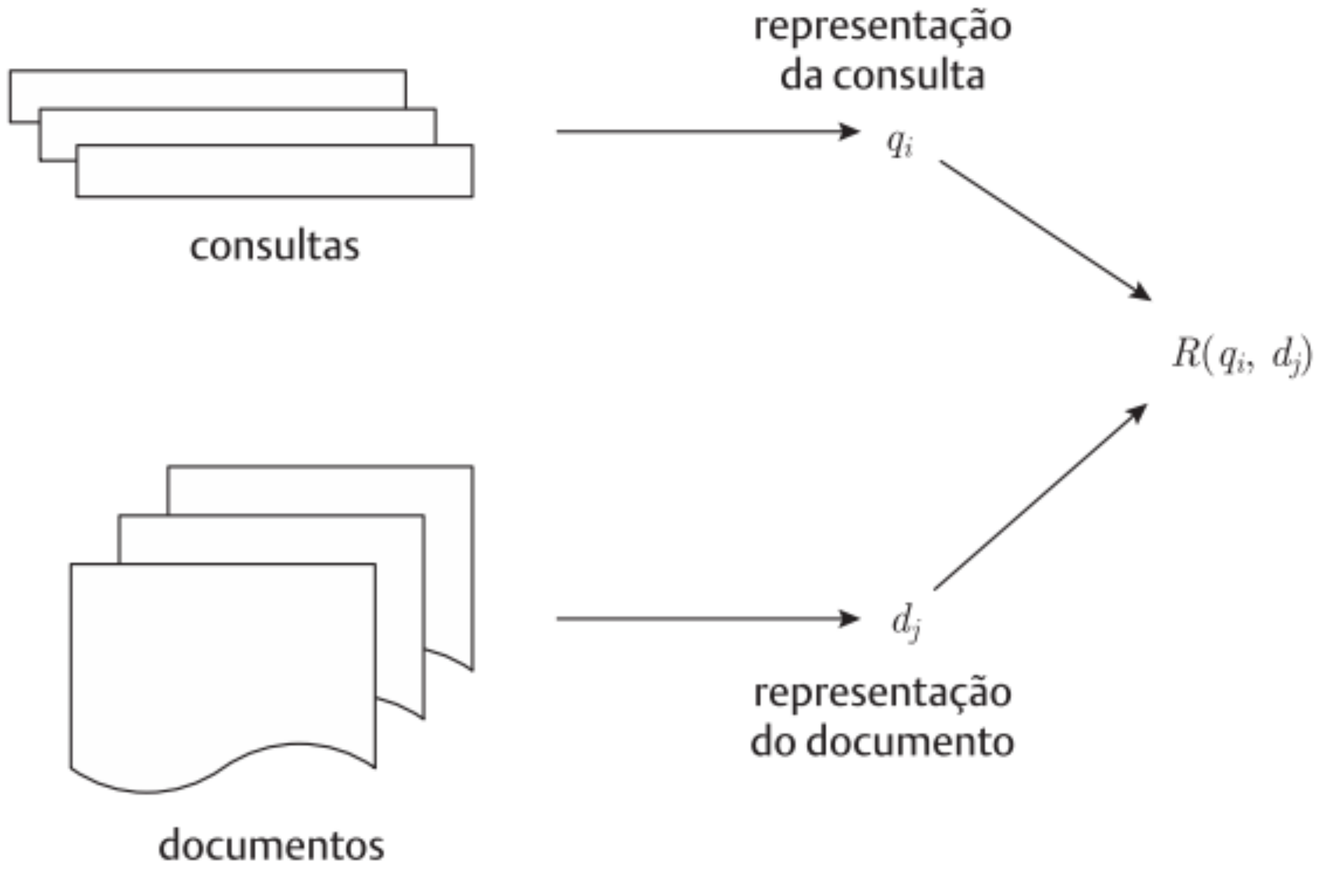
1. D é um conjunto composto por visões lógicas (ou representações) dos documentos da coleção.

2. Q é um conjunto composto por visões lógicas (ou representações) das necessidades de informação dos usuários. Essas representações são chamadas de consultas.

3. F é um sistema lógico usado para modelar as representações dos documentos, das consultas e de seus relacionamentos, como conjuntos e relações Booleanas;

4. R(q_i,d_j) é uma função de ranqueamento que associa um número real à representação de uma consulta $q_i \in Q$ e à representação de um documento $d_j \in D$. Esse ranking define um ordenamento entre os documentos em relação à consulta q_i.

Caracterização de um modelo de RI – Função de ranqueamento



Recuperação de informação clássica

Conceitos básicos

- Os modelos clássicos de RI consideram que cada documento é descrito por um conjunto de palavras-chave representativas, chamadas de **termos de indexação**:
- Um conjunto pré-selecionado de termos de indexação pode ser utilizado, por exemplo, para sumarizar o conteúdo dos documentos.
- Nesse caso, os termos são principalmente **substantivos** ou **grupos de substantivos**, uma vez que substantivos possuem significado próprio;
- Adjetivos, advérbios e conectores são menos úteis como termos de indexação, pois funcionam principalmente como complementos.

Conceitos básicos - Vocabulário

Definição:

Considere t como o número de termos de indexação na coleção de documentos e k_i como um termo de indexação genérico. V = {k_1, . . . , k_t} é o conjunto de todos os termos de indexação distintos na coleção e é comumente chamado de ***vocabulário*** *V da coleção. O tamanho do vocabulário é t.*

Dados os termos de um documento e de uma consulta como representá-los formalmente?

Conceitos básicos – Representação de documento e consulta

Definição:

Considere $V = \{k_1, k_2, \ldots, k_t\}$ como o vocabulário da coleção. Se três termos de indexação k_l, k_m e k_n ocorrem em um mesmo documento d_j, dizemos que o padrão $[k_l, k_m, k_n]$ de ***coocorrência*** *de termos foi observado. Cada um desses padrões de coocorrências de termos é chamado de componente* ***conjuntivo*** *de termo.*

- Exemplo: o padrão (1,0,...,0) indica a presença do termo k_1. O padrão (1,1,...,1) indica a presença de todos os termos no referido documento.

Conceitos básicos – Representação de documento e consulta

- Nesse caso, consultas e documentos são representados simplesmente pelos **componentes conjuntivos** de termo;

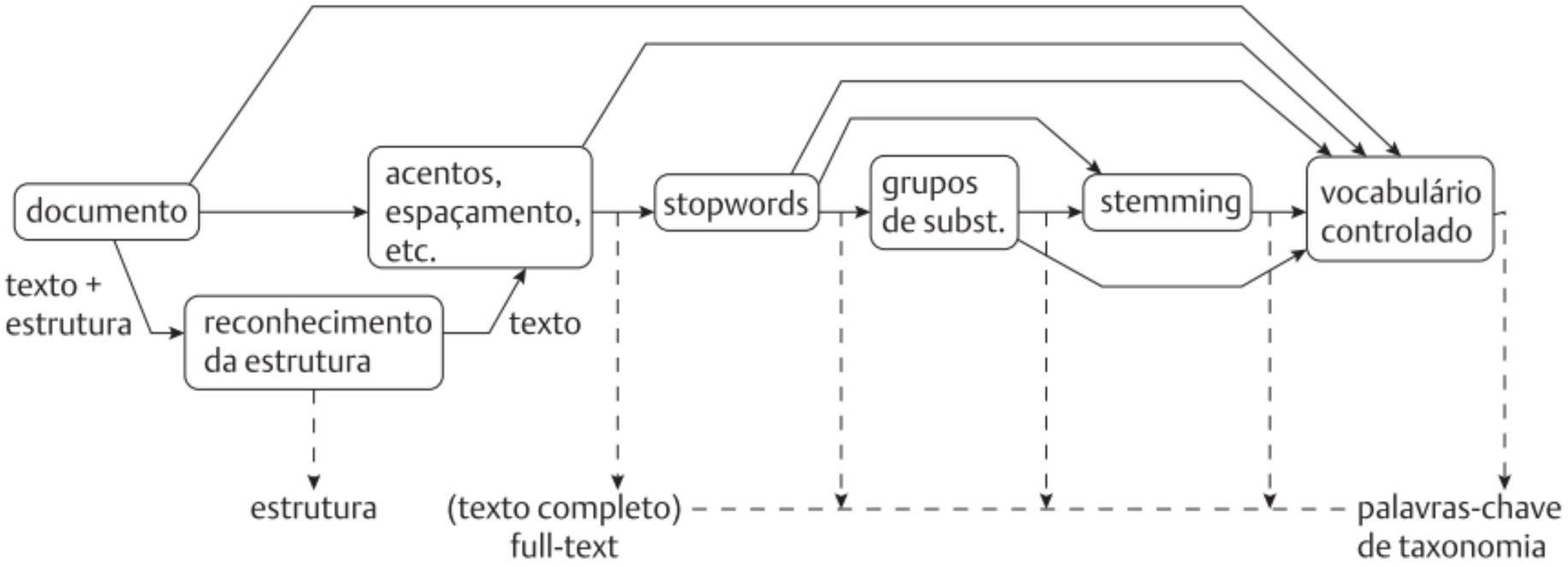
- Essa é a representação mais simples possível e é frequentemente conhecida como *bag of words* (saco de palavras).

Conceitos básicos – Representação de documento

Cada documento do conjunto pode ser representado por:

1. Um conjunto de termos indexados que melhor representem seus tópicos (existem diversas técnicas para construção desse conjunto)
2. Texto completo;
3. Texto completo + estrutura interna (capítulos, seções e etc).

Conceitos básicos – Representação de documento



Conceitos básicos – Matriz de termos e documentos

Abordagem simples para quantificar a relação entre a ocorrência de termos em determinados documentos:

$$\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array} \begin{array}{c} \begin{array}{cc} d_1 & d_2 \end{array} \\ \begin{bmatrix} f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ f_{3,1} & f_{3,2} \end{bmatrix} \end{array}$$

Onde k_n representa os termos de indexação, d_n os documentos e $f_{i,j}$ a frequência do termo k_i no documento d_j.

Matriz de termos e documentos - Exercício

Considere k1 = {liberdade} , k2 = {povo}, k3 = {forte} e k4 = {igualdade}. Seja d1 = {primeira estrofe do hino nacional} e d2 = {segunda estrofe do hino nacional}.

1) Monte a matriz de termos e documentos para esse sistema. Qual a ordem dessa matriz?

2) Considere que o vocabulário da coleção seja composto por todas as palavras com o número de letras maior ou igual a 3. Monte a matriz de termos e documentos nessas condições. Qual a ordem dessa matriz?

3) Sobre a matriz do ex. 2, qual a proporção de 0 (zero) e 1 (um)? Poderíamos pensar em uma representação mais eficiente para os termos e documentos?

Matriz de termos e documentos - Exercício

D1 = {Ouviram do Ipiranga as margens plácidas
De um povo heroico o brado retumbante,
E o sol da Liberdade, em raios fúlgidos,
Brilhou no céu da Pátria nesse instante.}

D2 = {Se o penhor dessa igualdade
Conseguimos conquistar com braço forte,
Em teu seio, ó Liberdade,
Desafia o nosso peito a própria morte!}

Modelo booleano

Modelo booleano

Lembrando que para cada um dos modelos de RI (booleano, vetorial e probabilístico) que serão estudados, veremos o funcionamento dos seguintes processos de RI:

1. Obtenção e coleção de documentos;
2. Indexação dos documentos;
3. **Consulta do usuário;**
4. **Recuperação de documentos;**
5. **Ranqueamento dos documentos;**
6. Apresentação para o usuário.

Modelo booleano

- O modelo Booleano é um modelo de recuperação de informação simples baseado na teoria de conjuntos e na álgebra Booleana;
- Como consequência, o modelo é bastante intuitivo e possui uma semântica precisa;
- Pela sua inerente simplicidade e formalismo, o modelo Booleano recebeu uma atenção considerável no passado e foi adotado por muitos dos primeiros sistemas bibliográficos comerciais.

Modelo booleano

- O modelo Booleano considera que os termos de indexação estão presentes ou ausentes nos documentos, ou seja, as frequências na matriz de termos por documentos são todas **binárias** (0 ou 1);
- Uma consulta q em um modelo booleano é composta por termos de indexação ligados por três conectivos Booleanos: **not**, **and** e **or**.
- Uma consulta é essencialmente uma expressão Booleana convencional sobre termos de indexação.

Modelo booleano - Definição

No modelo Booleano, uma consulta q é uma expressão Booleana convencional sobre termos de indexação. Considere c(q) como qualquer dos componentes conjuntivos da consulta. Dado um documento d_j, sendo $c(d_j)$ seu componente conjuntivo de documento correspondente, então a similaridade entre o documento e a consulta q é definida por:

$$sim(d_j, q) = \begin{cases} 1 & \text{se } \exists c(q) \mid c(q) = c(d_j) \\ 0 & \text{caso contrário} \end{cases}$$

Se sim(d_j, q) = 1, então d_j é relevante a consulta q.

Modelo booleano - Exemplo 1

- Suponha que o vocabulário da coleção seja dado por V = {k_a,k_b,k_c}. Seja d um documento que contém os termos k_a e k_c. Ou seja, d = {1,0,1}. Considere a consulta q = k_a AND k_b.

- Pergunta 1: o documento d satisfaz a consulta q? Ou seja, qual o grau de similaridade entre d e q (sim(d,q))?

- Pergunta 2: que documento satisfaz a consulta q? Nesse caso, qual seria o grau de similaridade entre esse documento e a consulta q?

Modelo booleano - Exemplo 2

- O que aconteceria se a consulta fosse “Brutus AND Casear AND NOT Calpurnia”, ou seja, quais documentos satisfazem essa consulta?

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Modelo booleano - Exemplo 2

- O que aconteceria se a consulta fosse “Brutus AND Caesear AND NOT Calpurnia”, ou seja, quais documentos satisfazem essa consulta?
- Intuitivamente, o que fizemos foi pegar os vetores Brutus, Caesar e (not) Calpurnia e fazer uma operação AND bit a bit:

110100 AND 110111 AND 101111
= 100100

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Modelo booleano - Relevância

- O modelo Booleano prevê que cada documento seja **relevante** ou **não relevante** perante a uma certa consulta;
- Não existe satisfação parcial das condições da consulta;
- Esse critério binário de decisão, sem nenhuma noção de grau, impede uma boa qualidade na recuperação de informação.

Uso das matrizes de termo e documento

- Por enquanto as coleções são pequenas…
- O que aconteceria se tivéssemos N = 1 milhão de documentos e que cada documento tivesse por volta de 1000 palavras? Nesse cenário seria comum ter por volta de M = 500.000 termos distintos!

1. Qual o tamanho da matriz de termos e documentos?
2. Qual seria uma característica dessa matriz?
3. Faz sentido usar essa representação?

Uso das matrizes de termo e documento

- Por enquanto as coleções são pequenas…
- O que aconteceria se tivéssemos N = 1 milhão de documentos e que cada documento tivesse por volta de 1000 palavras? Nesse cenário seria comum ter por volta de M = 500.000 termos distintos!

1. Qual o tamanho da matriz de termos e documentos? R = 500k x 1M
2. Qual seria uma característica dessa matriz? R = matriz esparsa! Muitos zeros…
3. Faz sentido usar essa representação? R = não! A ideia é registrar coisas que realmente ocorrem, por exemplo, as posições com valores 1. Um conceito central aqui para melhorar esse problema são os índices invertidos. Estudaremos ele nas próximas aulas.

Modelo booleano x Sistema de RI

- Consulta do usuário
  - Uso de expressões booleanas;
- Recuperação de documentos
  - Somente os documentos que satisfazem a consulta são recuperados;
- Ranqueamento dos documentos
  - Impossível!

Modelo booleano – Vantagens e desvantagens

- Vantagens
  - Formalismo claro;
  - Simplicidade;
  - Fácil de implementar;
  - Adoção de pesos binários para os termos de indexação.
- Desvantagens
  - Impossibilidade de realizar ranqueamento dos documentos;
  - Formulação de consultas booleanas pode ser incoveniente para os usuários.

Modelo booleano – Exercício

Considere três documentos D1, D2 e D3. D1 é a primeira estrofe do hino à bandeira, D2 é a primeira estrofe do hino da independência e D3 é a primeira estrofe do hino nacional. Considere somente os termos em destaque de cada documento.

a) Encontre o vocabulário dessa coleção.

b) Monte a matriz de termos e documentos.

c) Encontre a similaridade entre os documentos e cada uma das consultas a seguir: q1 = {liberdade AND brasil}, q2 = {patria} e q3 = {nobre OR heroico NOT liberdade}.

Modelo booleano – Exercício

D1 = {Salve, lindo pendão da **esperança**,
Salve, **símbolo** augusto da paz!
Tua **nobre** presença à lembrança
A grandeza da **Pátria** nos traz. }

D2 = {Já podeis, da **Pátria** filhos,
Ver contente a mãe gentil;
Já raiou a **liberdade**
No horizonte do **Brasil**.}

D3 = {Ouviram do **Ipiranga** as margens plácidas
De um povo **heroico** o brado retumbante,
E o sol da **Liberdade**, em raios fúlgidos,
Brilhou no céu da **Pátria** nesse instante.}

Modelo booleano - Passos

1) Encontrar o vocabulário da coleção;

2) Representar os documentos usando os componentes conjuntivos de termo;

3) Representar as consultas usando a forma normal disjuntiva (ou simplesmente encontrar o padrão de ocorrência);

4) Comparar os documentos com as consultas.

Comentários

No decorrer da aula vimos…

- Modelagem de sistemas de RI passa pela criação de uma função de ranqueamento (probabilidade de relevância para o usuário);

- A caracterização formal de um sistema de RI passa pela definição de quatro elementos: D (documentos da coleção), Q (consultas), F (sistema lógico para modelar as representações dos documentos e consultas) e R (função de ranqueamento).

No decorrer da aula vimos…

- Conceitos básicos da recuperação de informação clássica:
  - Vocabulário;
  - Componente conjuntivo;
  - Matriz de termos e documentos.

- Modelo booleano
  - Funcionamento;
  - Vantagens e desvantagens.

Próximas aulas

- Índices invertidos;
- Melhoria do modelo booleano com o auxílio da ponderação de termos;
- Modelo vetorial;
- Modelo probabilístico.

Roteiro de estudos

- Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca
  - Capítulo 2 – 2.1, 2.2.1, 2.2.2