Pesquisa e Publicação de Informação

Modelos de Pesquisa de Informação

Nuno D. Mendes

Licenciatura em Sistemas e Tecnologias de Informação

13 Abr 2012

ISEGI – UNL

Sistemas de Pesquisa de Informação

- Frequentemente baseados em **termos de pesquisa**
- Problemas?

Sistemas de Pesquisa de Informação

- Frequentemente baseados em **termos de pesquisa**
- Problemas?
  - Semântica dos objectos informacionais é perdida
  - Termos de pesquisa são mera aproximação
  - Pedido de informação (também baseado em termos)
  - **Pedido de informação** não é o mesmo que a **Necessidade de Informação**
  - Utilizador nem sempre sabe traduzir a sua NI num PI
- Baseados em *a prioris* sobre relevância dos objectos

Taxonomia de modelos de PI

- Pesquisa
  - Modelos Clássicos
    - Booleano
      Extensões (modelos baseados em teoria de conjuntos)
      Conjuntos *Fuzzy*
      Booleano estendido
    - Vectorial
      Extensões (modelos algébricos)
      Vectorial generalizado
      Indexação de Semântica Latente
      Redes Neuronais
    - Probabilístico
      Extensões (modelos probabilísticos)
      Redes de inferência
      Redes de crenças
  - Modelos Estruturados
    - Segmentado
    - Hierárquico
- Navegação
  - Simples
  - Orientada pela estrutura
  - Hipertexto

Modelos de PI quanto à tarefa do utilizador e abstracção dos objectos informacionais

	Termos de Pesquisa	Texto completo	Texto completo + Estrutura
Pesquisa	Clássicos Teoria de conjuntos Algébricos Probabilísticos	Clássicos Teoria de conjuntos Algébricos Probabilísticos	Estruturados
Navegação	Simples	Simples Hipertexto	Orientada pela estrutura Hipertexto

Definição Formal de Sistema de Pesquisa de Informação

Sistema de Pesquisa de Informação

Um sistema de Pesquisa de Informação para uma colecção de objectos informacionais, $\mathcal{C}$, é um tuplo $(D, Q, \mathcal{F}, R)$, onde:

- D é um conjunto de abstracções para os objectos informacionais da colecção $\mathcal{C}$
- Q é um conjunto de abstracções para as necessidades de informação do utilizador (**queries**)
- $\mathcal{F}$ representa a modelação das relações entre as abstracções dos objectos e as das queries
- $R : D \times Q \mapsto \Re$ é uma função que associa uma query $q_i \in Q$ e um objecto $d_j \in D$ a um número real $R(q_i, d_j)$, que representa o *ranking* de d_j em relação à query q_i de entre todos os documentos em D

Conceitos gerais

Seja t o número de *termos de pesquisa* no sistema e seja k_i o i-ésimo termo.

$K = \{k_1, \ldots, k_t\}$ é o conjunto de termos de pesquisa.

A cada termo k_i do objecto d_j está associado um peso $w_{ij} > 0$. Para todo o termo que não ocorre em d_j, $w_{ij} = 0$.

Assim, ao objecto d_j está associado um vector de pesos $\vec{d_j} = (w_{1j}, \ldots, w_{tj})$.

Adicionalmente, consideramos uma função g_i que devolve o peso associado ao termo k_i em qualquer vector t-dimensional (*i.e.* $g_i\left(\vec{d_j}\right) = w_{ij}$).

Conceitos gerais

Seja t o número de *termos de pesquisa* no sistema e seja k_i o i-ésimo termo.

$K = \{k_1, \ldots, k_t\}$ é o conjunto de termos de pesquisa.

A cada termo k_i do objecto d_j está associado um peso $w_{ij} > 0$. Para todo o termo que não ocorre em d_j, $w_{ij} = 0$.

Assim, ao objecto d_j está associado um vector de pesos $\vec{d_j} = (w_{1j}, \ldots, w_{tj})$.

Adicionalmente, consideramos uma função g_i que devolve o peso associado ao termo k_i em qualquer vector t-dimensional (*i.e.* $g_i\left(\vec{d_j}\right) = w_{ij}$).

- Assumimos que o conhecimento de w_{ij} não diz nada acerca de $w_{i'j}$, *i.e.* não há correlação entre ocorrências de k_i e $k_{i'}$ em d_j
- É uma simplificação do problema porque frequentemente há correlações (*e.g.* "pesquisa da informação")
- Não é trivial incorporar a correlação entre termos nos modelos

Modelo Booleano

Definição

- $w_{ij} \in \{0, 1\}$

Modelo Booleano

Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é uma expressão Booleana de termos (*e.g.* $k_a \wedge (k_b \vee \neg k_c)$)

Modelo Booleano

Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é uma expressão Booleana de termos (*e.g.* $k_a \wedge (k_b \vee \neg k_c)$)
- Expressa na forma normal disjuntiva (DNF) (*e.g.* $\vec{q}_{\text{dnf}} = (1, 0, 0) \vee (1, 1, 0) \vee (1, 1, 1)$)

Modelo Booleano

Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é uma expressão Booleana de termos
  (*e.g.* $k_a \wedge (k_b \vee \neg k_c)$)
- Expressa na forma normal disjuntiva (DNF)
  (*e.g.* $\vec{q}_{\text{dnf}} = (1, 0, 0) \vee (1, 1, 0) \vee (1, 1, 1)$)
- Seja $\vec{q}_{\text{cc}}$ cada uma das conjunções de $\vec{q}_{\text{dnf}}$, então, a semelhança entre uma query q e um documento d_j é expressa pela seguinte expressão:

Modelo Booleano

Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é uma expressão Booleana de termos
  (*e.g.* $k_a \wedge (k_b \vee \neg k_c)$)
- Expressa na forma normal disjuntiva (DNF)
  (*e.g.* $\vec{q}_{\text{dnf}} = (1, 0, 0) \vee (1, 1, 0) \vee (1, 1, 1)$)
- Seja $\vec{q}_{\text{cc}}$ cada uma das conjunções de $\vec{q}_{\text{dnf}}$, então, a semelhança entre uma query q e um documento d_j é expressa pela seguinte expressão:

$$\sigma(d_j, q) = \begin{cases} 1 & \text{se} \quad \exists \vec{q}_{\text{cc}} \in \vec{q}_{\text{dnf}} \quad \forall k_i \quad g_i(\vec{d}_j) = g_i(\vec{q}_{\text{cc}}) \\ 0 & \text{c.c.} \end{cases}$$

Modelo Booleano

Exemplo

d_1

Isto é um exemplo para um modelo Booleano.

q_1

$[\text{exemplo} \wedge \text{Booleano}]$

d_2

Este é outro exemplo.

q_2

$[\text{Isto} \vee \neg\text{Booleano}]$

Modelo Booleano

Exemplo

d_1

Isto é um exemplo para um modelo Booleano.

q_1

$[\text{exemplo} \wedge \text{Booleano}]$

d_2

Este é outro exemplo.

q_2

$[\text{Isto} \vee \neg\text{Booleano}]$

$$\sigma(d_1, q_1) = 1$$

Exemplo

d_1

Isto é um exemplo para um modelo Booleano.

q_1

$[\text{exemplo} \wedge \text{Booleano}]$

d_2

Este é outro exemplo.

q_2

$[\text{Isto} \vee \neg\text{Booleano}]$

$$\sigma(d_1, q_1) = 1$$
$$\sigma(d_1, q_2) = 1$$

Exemplo

d_1

Isto é um exemplo para um modelo Booleano.

q_1

$[\text{exemplo} \wedge \text{Booleano}]$

d_2

Este é outro exemplo.

q_2

$[\text{Isto} \vee \neg\text{Booleano}]$

$$\sigma(d_1, q_1) = 1$$
$$\sigma(d_1, q_2) = 1$$
$$\sigma(d_2, q_1) = 0$$

Modelo Booleano

Exemplo

d_1

Isto é um exemplo para um modelo Booleano.

q_1

$[\text{exemplo} \wedge \text{Booleano}]$

d_2

Este é outro exemplo.

q_2

$[\text{Isto} \vee \neg\text{Booleano}]$

$$\sigma(d_1, q_1) = 1$$
$$\sigma(d_1, q_2) = 1$$
$$\sigma(d_2, q_1) = 0$$
$$\sigma(d_2, q_2) = 1$$

ISEGI Estatística e Gestão de Informação

Modelo Booleano

Vantagens/Desvantagens

- Oferece uma **semântica precisa**

Modelo Booleano

Vantagens/Desvantagens

- Oferece uma **semântica precisa**
- É **simples** e permite uma avaliação rápida, mas

Modelo Booleano

Vantagens/Desvantagens

- Oferece uma **semântica precisa**
- É **simples** e permite uma avaliação rápida, mas
- Considera apenas um **critério de decisão binário**

Modelo Booleano

Vantagens/Desvantagens

- Oferece uma **semântica precisa**
- É **simples** e permite uma avaliação rápida, mas
- Considera apenas um **critério de decisão binário**
- Um documento é apenas considerado relevante ou não-relevante, não existe *ranking*

Modelo Booleano

Vantagens/Desvantagens

- Oferece uma **semântica precisa**
- É **simples** e permite uma avaliação rápida, mas
- Considera apenas um **critério de decisão binário**
- Um documento é apenas considerado relevante ou não-relevante, não existe *ranking*
- Assemelha-se mais a um modelo de recuperação de dados

Definição

- $w_{ij} > 0$

Definição

- $w_{ij} > 0$
- $\hat{w}_i$, peso associado ao i-ésimo termo da query q, $\hat{w}_i > 0$

Definição

- $w_{ij} > 0$
- $\hat{w}_i$, peso associado ao i-ésimo termo da query q, $\hat{w}_i > 0$
- $\vec{q} = (\hat{w}_1, \ldots, \hat{w}_t)$

Definição

- $w_{ij} > 0$
- $\hat{w}_i$, peso associado ao i-ésimo termo da query q, $\hat{w}_i > 0$
- $\vec{q} = (\hat{w}_1, \ldots, \hat{w}_t)$
- $\vec{d}_j = (w_{1j}, \ldots, w_{tj})$

Definição

- $w_{ij} > 0$
- $\hat{w}_i$, peso associado ao i-ésimo termo da query q, $\hat{w}_i > 0$
- $\vec{q} = (\hat{w}_1, \ldots, \hat{w}_t)$
- $\vec{d_j} = (w_{1j}, \ldots, w_{tj})$
- A semelhança entre uma query q e um documento d_j é expressa pela seguinte expressão:

Definição

- $w_{ij} > 0$
- $\hat{w}_i$, peso associado ao i-ésimo termo da query q, $\hat{w}_i > 0$
- $\vec{q} = (\hat{w}_1, \ldots, \hat{w}_t)$
- $\vec{d_j} = (w_{1j}, \ldots, w_{tj})$
- A semelhança entre uma query q e um documento d_j é expressa pela seguinte expressão:

$$\sigma(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times \hat{w}_i}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} \hat{w}_i^2}}$$

Definição

- $w_{ij} > 0$
- $\hat{w}_i$, peso associado ao i-ésimo termo da query q, $\hat{w}_i > 0$
- $\vec{q} = (\hat{w}_1, \ldots, \hat{w}_t)$
- $\vec{d}_j = (w_{1j}, \ldots, w_{tj})$
- A semelhança entre uma query q e um documento d_j é expressa pela seguinte expressão:

$$\sigma(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times \hat{w}_i}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} \times \sqrt{\sum_{i=1}^{t} \hat{w}_i^2}}$$

- O que corresponde a calcular o coseno do ângulo entre $\vec{d}_j$ e $\vec{q}$ no espaço t-dimensional dos termos

Atribuição de pesos

- Baseados na noção de *clustering*

Atribuição de pesos

- Baseados na noção de *clustering*
- A query q é usada como referência para determinar duas classes ou *clusters* de documentos (documentos relevantes e não-relevantes)

Atribuição de pesos

- Baseados na noção de *clustering*
- A query q é usada como referência para determinar duas classes ou *clusters* de documentos (documentos relevantes e não-relevantes)
- Para adaptar o problema de pesquisa de informação a um problema de *clustering* é necessário resolver dois problemas
  - Maximizar a semelhaną dos documentos contidos no mesmo *cluster* (semelhança *intra-cluster*)

Atribuição de pesos

- Baseados na noção de *clustering*
- A query q é usada como referência para determinar duas classes ou *clusters* de documentos (documentos relevantes e não-relevantes)
- Para adaptar o problema de pesquisa de informação a um problema de *clustering* é necessário resolver dois problemas
  - Maximizar a semelhaną dos documentos contidos no mesmo *cluster* (semelhança *intra-cluster*)
  - Maximizar a diferença entre documentos contidos em *clusters* diferentes (diferença *inter-cluster*)

Atribuição de pesos

- Baseados na noção de *clustering*
- A query *q* é usada como referência para determinar duas classes ou *clusters* de documentos (documentos relevantes e não-relevantes)
- Para adaptar o problema de pesquisa de informação a um problema de *clustering* é necessário resolver dois problemas
  - Maximizar a semelhaṇa dos documentos contidos no mesmo *cluster* (semelhança *intra-cluster*)

    É usado frequentemente o **factor tf** (*term frequency*) como forma de medir a poder descritivo de um termo para um dado documento e corresponde à frequência absoluta do termo no documento (*i.e.* o número de ocorrências do termo)

  - Maximizar a diferença entre documentos contidos em *clusters* diferentes (diferença *inter-cluster*)

Atribuição de pesos

- Baseados na noção de *clustering*
- A query q é usada como referência para determinar duas classes ou *clusters* de documentos (documentos relevantes e não-relevantes)
- Para adaptar o problema de pesquisa de informação a um problema de *clustering* é necessário resolver dois problemas
  - Maximizar a semelhaną dos documentos contidos no mesmo *cluster* (semelhança *intra-cluster*)
  - Maximizar a diferença entre documentos contidos em *clusters* diferentes (diferença *inter-cluster*)

    É usado frequentemente o **factor idf** (*inverse document frequency*) que capta a ideia que são os termos raros que mais permitem distinguir entre documentos diferentes (*i.e.* termos que aparecem em quase todos os documento não são bons discriminantes)

    $$\text{idf}_i = \log \frac{N}{n_i}$$

    onde N é o número total de documentos, e n_i é o número de documentos com ocorrências do termo k_i.

Atribuição de pesos

- Frequência normalizada de um termo k_i no documento d_j

$$f_{i,j} = \frac{\text{freq}_{i,j}}{\max_{\xi=1,\ldots,t} \text{freq}_{\xi,j}}$$

- Peso do termo k_i no documento d_j (método **tf-idf**)

$$w_{ij} = f_{i,j} \log \frac{N}{n_i}$$

- Peso do termo k_i na query q (Salton & Buckley, 1988)

$$\hat{w}_i = \left(\frac{1}{2} + \frac{\frac{1}{2}\hat{\text{freq}}_i}{\max_{\xi=1,\ldots,t} \hat{\text{freq}}_\xi} \right) \log \frac{N}{n_i}$$

Exemplo

d_1

Isto é um exemplo para um modelo Booleano.

q_1

$[exemplo, Booleano]$

d_2

Este é outro exemplo.

q_1

$[Isto]$

Vantagens/Desvantagens

- Simples e eficaz

Vantagens/Desvantagens

- Simples e eficaz
- A utilização de pesos melhora bastante a performance, mas

Vantagens/Desvantagens

- Simples e eficaz
- A utilização de pesos melhora bastante a performance, mas
- Assume independência entre os termos, o que não é sempre verificado

Vantagens/Desvantagens

- Simples e eficaz
- A utilização de pesos melhora bastante a performance, mas
- Assume independência entre os termos, o que não é sempre verificado
- Embora a captação de dependência seja difícil e por vezes, contraproducente (*e.g.* correlações localizadas que não se verificam sistematicamente em toda a colecção de documentos)