# Pesquisa e Publicação de Informação

## Modelos de Pesquisa de Informação, Avaliação de Sistemas de PI

Nuno D. Mendes

Licenciatura em Sistemas e Tecnologias de Informação

20 Abr 2012

ISEGI – UNL

Parte I

# Modelo Probabilístico

# Modelo Probabilístico

## Definição

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

# Modelo Probabilístico

## Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é um conjunto de termos (*i.e.* $q \in 2^K$)

# Modelo Probabilístico

## Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é um conjunto de termos (*i.e.* $q \in 2^K$)
- $A_q \subseteq D$ é um conjunto de objectos informacionais tidos como relevantes para $q$ e $\bar{A}_q$ é o seu complemento ($A_q \cup \bar{A}_q = D$).

# Modelo Probabilístico

## Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é um conjunto de termos (*i.e.* $q \in 2^K$)
- $A_q \subseteq D$ é um conjunto de objectos informacionais tidos como relevantes para $q$ e $\bar{A}_q$ é o seu complemento ($A_q \cup \bar{A}_q = D$).
- Seja $P(A_q|\vec{d}_j)$ a probabilidade do objecto $d_j$ ser relevante para a query $q$ e $P(\bar{A}_q|\vec{d}_j)$ a probabilidade de que $d_j$ não é relevante para a query $q$.

# Modelo Probabilístico

## Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é um conjunto de termos (*i.e.* $q \in 2^K$)
- $A_q \subseteq D$ é um conjunto de objectos informacionais tidos como relevantes para $q$ e $\bar{A}_q$ é o seu complemento ($A_q \cup \bar{A}_q = D$).
- Seja $P(A_q|\vec{d}_j)$ a probabilidade do objecto $d_j$ ser relevante para a query $q$ e $P(\bar{A}_q|\vec{d}_j)$ a probabilidade de que $d_j$ não é relevante para a query $q$.
- A semelhança entre a query $q$ e o objecto $d_j$ é dada por

$$\sigma(d_j, q) = \frac{P(A_q|\vec{d}_j)}{P(\bar{A}_q|\vec{d}_j)}$$

# Modelo Probabilístico

## Definição

- $w_{ij} \in \{0, 1\}$
- $q \in Q$ é um conjunto de termos (*i.e.* $q \in 2^K$)
- $A_q \subseteq D$ é um conjunto de objectos informacionais tidos como relevantes para $q$ e $\bar{A}_q$ é o seu complemento ($A_q \cup \bar{A}_q = D$).
- Seja $P(A_q|\vec{d}_j)$ a probabilidade do objecto $d_j$ ser relevante para a query $q$ e $P(\bar{A}_q|\vec{d}_j)$ a probabilidade de que $d_j$ não é relevante para a query $q$.
- A semelhança entre a query $q$ e o objecto $d_j$ é dada por

$$\sigma(d_j, q) = \frac{P(A_q|\vec{d}_j)}{P(\bar{A}_q|\vec{d}_j)}$$

- Pela regra de Bayes

$$\sigma(d_j, q) = \frac{P(\vec{d}_j|A_q)P(A_q)}{P(\vec{d}_j|\bar{A}_q)P(\bar{A}_q)}$$

onde $P(\vec{d}_j|A_q)$ representa a probabilidade de escolher aleatoriamente um objecto $d_j$ de entre o conjunto de relevantes (resp. irrelevantes) e $P(A_q)$ designa a probabilidade de observar um documento relevante em $D$ (resp. irrelevante).

# Modelo Probabilístico

## Definição

- Temos que

$$\sigma(d_j, q) = \frac{P(\vec{d}_j|A_q)P(A_q)}{P(\vec{d}_j|\bar{A}_q)P(\bar{A}_q)}$$

# Modelo Probabilístico

## Definição

- Temos que

$$\sigma(d_j, q) = \frac{P(\vec{d}_j|A_q)P(A_q)}{P(\vec{d}_j|\bar{A}_q)P(\bar{A}_q)}$$

- Como $P(A_q)$ e $P(\bar{A}_q)$ são iguais para todos os documentos, escrevemos

$$\sigma(d_j, q) \propto \frac{P(\vec{d}_j|A_q)}{P(\vec{d}_j|\bar{A}_q)}$$

# Modelo Probabilístico

## Definição

- Temos que

$$\sigma(d_j, q) = \frac{P(\vec{d}_j|A_q)P(A_q)}{P(\vec{d}_j|\bar{A}_q)P(\bar{A}_q)}$$

- Como $P(A_q)$ e $P(\bar{A}_q)$ são iguais para todos os documentos, escrevemos

$$\sigma(d_j, q) \propto \frac{P(\vec{d}_j|A_q)}{P(\vec{d}_j|\bar{A}_q)}$$

- Assumindo a independência dos termos

$$\sigma(d_j, q) \propto \frac{\left(\prod_{g_i(\vec{d}_j)=1} P(k_i|A_q)\right)\left(\prod_{g_i(\vec{d}_j)=0} P(\bar{k}_i|A_q)\right)}{\left(\prod_{g_i(\vec{d}_j)=1} P(k_i|\bar{A}_q)\right)\left(\prod_{g_i(\vec{d}_j)=0} P(\bar{k}_i|\bar{A}_q)\right)}$$

onde $P(k_i|A_q)$ denota a probabilidade do termo $k_i$ estar presente num objecto aleatoriamente escolhido dentro os documentos do conjunto $A_q$ e $P(\bar{k}_i|A_q)$ a probabilidade do termo $k_i$ não estar presente num objecto escolhido do mesmo conjunto.

## Definição

- Lembrando que $P(k_i|A_q) + P(\bar{k}_i|A_q) = 1$ e ignorando factores constantes para todos os objectos e para a mesma query $q$, temos que

$$\sigma(d_j, q) \propto \sum_{i=1}^{t} \hat{w}_i w_{ij} \left( \log \frac{P(k_i|A_q)}{1 - P(k_i|A_q)} + \log \frac{1 - P(k_i|\bar{A}_q)}{P(k_i|\bar{A}_q)} \right)$$

## Definição

- Lembrando que $P(k_i|A_q) + P(\bar{k}_i|A_q) = 1$ e ignorando factores constantes para todos os objectos e para a mesma query $q$, temos que

$$\sigma(d_j, q) \propto \sum_{i=1}^{t} \hat{w}_i w_{ij} \left( \log \frac{P(k_i|A_q)}{1 - P(k_i|A_q)} + \log \frac{1 - P(k_i|\bar{A}_q)}{P(k_i|\bar{A}_q)} \right)$$

- Na primeira iteração do modelo, assume-se que
  - $P(k_i|A_q)$ é constante para todos os termos $k_i$ (tipicamente 0.5)
  - A distribuição de termos nos documentos irrelevantes pode ser aproximada pela distribuição de termos em $D$
  - Assim temos que:

$$P(k_i|A_q) = 0.5$$
$$P(k_i|\bar{A}_q) = \frac{n_i}{N}$$

onde $n_i$ é o número de documentos que contêm o termo $k_i$ e $N$ o número total de documentos

## Definição

- Para as iterações seguintes do modelo definimos dois conjuntos:
  - $V$ como os $r$ documentos mais relevantes segundo a estimativa da iteração anterior
  - $V_i \subseteq V$ o conjunto de documentos em $V$ que contêm o termo $k_i$
- Seja $v$ e $v_i$ o número de elementos de $V$ e $V_i$, respectivamente, então podemos fazer as seguintes aproximações na iteração corrente

$$P(k_i|A_q) = \frac{v_i}{v}$$

$$P(k_i|\bar{A}_q) = \frac{n_i - v_i}{N - v}$$

- Para evitar problemas nos casos degenerados (*e.g* $v = 1$ e $V_i = \emptyset$), introduzimos um factor de ajustamento tal que

$$P(k_i|A_q) = \frac{v_i + \frac{n_i}{N}}{v + 1}$$

$$P(k_i|\bar{A}_q) = \frac{n_i - v_i + \frac{n_i}{N}}{N - v + 1}$$

# Modelo Probabilístico

## Vantagens/Desvantagens

- Oferece uma função de *ranking* baseada na probabilidade do documento ser relevante, mas

# Modelo Probabilístico

## Vantagens/Desvantagens

- Oferece uma função de *ranking* baseada na probabilidade do documento ser relevante, mas
- Requer uma estimação inicial de documentos relevantes vs não-relevantes

# Modelo Probabilístico

## Vantagens/Desvantagens

- Oferece uma função de *ranking* baseada na probabilidade do documento ser relevante, mas
- Requer uma estimação inicial de documentos relevantes vs não-relevantes
- O modelo não leva em conta a frequência dos termos num dado documento (todos os pesos são binários)

# Modelo Probabilístico

## Vantagens/Desvantagens

- Oferece uma função de *ranking* baseada na probabilidade do documento ser relevante, mas
- Requer uma estimação inicial de documentos relevantes vs não-relevantes
- O modelo não leva em conta a frequência dos termos num dado documento (todos os pesos são binários)
- Os termos são tidos como independentes

Parte II

# Avaliação de Sistemas de Pesquisa de Informação

UNIVERSIDADE NOVA ISEGI Estatística e Gestão de Informação

## Noções Básicas

- Considere um conjunto $I$ de pedidos de informação e um conjunto indexado de conjuntos $\{R_q\}_{q \in I}$ de documentos relevantes

## Noções Básicas

- Considere um conjunto $I$ de pedidos de informação e um conjunto indexado de conjuntos $\{R_q\}_{q \in I}$ de documentos relevantes
- Para cada query $q \in I$ o sistema de PI em avaliação produz um conjunto resposta, $A_q$, de documentos relevantes (ou dos melhores documentos em termos do *ranking* de relevância)

## Noções Básicas

- Considere um conjunto $I$ de pedidos de informação e um conjunto indexado de conjuntos $\{R_q\}_{q \in I}$ de documentos relevantes
- Para cada query $q \in I$ o sistema de PI em avaliação produz um conjunto resposta, $A_q$, de documentos relevantes (ou dos melhores documentos em termos do *ranking* de relevância)
- Podemos definir duas medidas básicas de avaliação do sistema:

## Noções Básicas

- Considere um conjunto $I$ de pedidos de informação e um conjunto indexado de conjuntos $\{R_q\}_{q \in I}$ de documentos relevantes
- Para cada query $q \in I$ o sistema de PI em avaliação produz um conjunto resposta, $A_q$, de documentos relevantes (ou dos melhores documentos em termos do *ranking* de relevância)
- Podemos definir duas medidas básicas de avaliação do sistema:
  - ***Recall*** ou **Sensibilidade**, mede a fracção de documentos relevantes que foram recuperados

$$\text{Recall} = \frac{|R_q \cap A_q|}{|R_q|}$$

## Noções Básicas

- Considere um conjunto $I$ de pedidos de informação e um conjunto indexado de conjuntos $\{R_q\}_{q \in I}$ de documentos relevantes
- Para cada query $q \in I$ o sistema de PI em avaliação produz um conjunto resposta, $A_q$, de documentos relevantes (ou dos melhores documentos em termos do *ranking* de relevância)
- Podemos definir duas medidas básicas de avaliação do sistema:
  - ***Recall*** ou **Sensibilidade**, mede a fracção de documentos relevantes que foram recuperados

$$\text{Recall} = \frac{|R_q \cap A_q|}{|R_q|}$$

  - **Precisão**, mede a fracção de documentos recuperados que são relevantes

$$\text{Precision} = \frac{|R_q \cap A_q|}{|A_q|}$$

## Curvas de Precisão/Recall

- Tendo em conta que os documentos em $A_q$ estão ordenados por relevância, pode ser importante perceber a evolução das medidas de avaliação ao longo do *ranking*

## Curvas de Precisão/Recall

- Tendo em conta que os documentos em $A_q$ estão ordenados por relevância, pode ser importante perceber a evolução das medidas de avaliação ao longo do *ranking*
- Por outro lado, a avaliação do sistema deve ser feita em relação a um conjunto de queries $I$, captando o comportamento médio para as queries utilizadas

## Curvas de Precisão/Recall

- Tendo em conta que os documentos em $A_q$ estão ordenados por relevância, pode ser importante perceber a evolução das medidas de avaliação ao longo do *ranking*
- Por outro lado, a avaliação do sistema deve ser feita em relação a um conjunto de queries $I$, captando o comportamento médio para as queries utilizadas
- Tipicamente, assumem-se 11 níveis de recall $(0\%, 10\%, 20\%, \ldots, 100\%)$ e calcula-se a precisão média a cada um dos níveis

$$\hat{P}(r) = \sum_{q \in I} \frac{P_q(r)}{|I|}$$

em que $\hat{P}(r)$ é a precisão média ao nível de recall $r$, e $P_q(r)$ é a precisão para a query $q$ ao nível de recall $r$.

## Curvas de Precisão/Recall

- Tendo em conta que os documentos em $A_q$ estão ordenados por relevância, pode ser importante perceber a evolução das medidas de avaliação ao longo do *ranking*
- Por outro lado, a avaliação do sistema deve ser feita em relação a um conjunto de queries $I$, captando o comportamento médio para as queries utilizadas
- Tipicamente, assumem-se 11 níveis de recall $(0\%, 10\%, 20\%, \ldots, 100\%)$ e calcula-se a precisão média a cada um dos níveis

$$\hat{P}(r) = \sum_{q \in I} \frac{P_q(r)}{|I|}$$

em que $\hat{P}(r)$ é a precisão média ao nível de recall $r$, e $P_q(r)$ é a precisão para a query $q$ ao nível de recall $r$.

- Como o cálculo do recall ao longo de $A_q$ raramente corresponde aos níveis pré-definidos, usa-se um procedimento de interpolação.
  Seja $r_j$ com $j \in \{0, 10, 20, \ldots, 100\}$, cada um dos níveis de recall utilizados, e seja $r$ o nível de recall calculado em $A_q$, temos que:

$$P(r_j) = \max_{r_j \leq r \leq r_{j+10}} P(r)$$

## Medidas Unitárias

- Por vezes é útil ter apenas uma medida de avaliação de um sistema de PI

## Medidas Unitárias

- Por vezes é útil ter apenas uma medida de avaliação de um sistema de PI
- **R-precision** Corresponde, para uma query $q \in I$, a calcular a precisão de entre os primeiros $|R_q|$ elementos de $A_q$
- **Média Harmónica** (F-measure)

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

em que $r(j)$ (resp. $P(j)$) é a recall (resp. precisão) para o $j$-ésimo documento do *ranking*. Pode obter-se uma sumarização fazendo $F = \max_{j=1,\ldots,|A_q|} F(j)$.

## Medidas Unitárias

- Por vezes é útil ter apenas uma medida de avaliação de um sistema de PI
- **R-precision** Corresponde, para uma query $q \in I$, a calcular a precisão de entre os primeiros $|R_q|$ elementos de $A_q$
- **Média Harmónica** (F-measure)

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}}$$

em que $r(j)$ (resp. $P(j)$) é a recall (resp. precisão) para o $j$-ésimo documento do *ranking*. Pode obter-se uma sumarização fazendo $F = \max_{j=1,\ldots,|A_q|} F(j)$.

- **E-measure**

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}}$$

semelhante à F-measure, mas o parâmetro $b$ permite dar mais ou menos importância à recall/precisão. Para $b = 1$ a E-measure é o complemento da F-measure. Para $b > 1$, a precisão é mais importante do que a recall e com $b < 1$ a recall é mais importante.