Técnicas para o Desenvolvimento de Recomendadores

Tássia Camões Araújo

Universidade de São Paulo

EXAME DE QUALIFICAÇÃO DE MESTRADO
Programa: Ciência da Computação
Orientador: Prof. Dr. Arnaldo Mandel

24 de fevereiro de 2011

K-NN

- Aprendizado de máquina supervisionado
- Proximidade entre objetos
- Vizinhança composta por k objetos
- A classe mais frequente em sua vizinha é atribuída ao objeto

K-NN

Medidas de distância e similaridade entre objetos

Distância euclidiana	$D(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$
Similaridade de cosseno	$sim(X,Y) = \frac{\vec{X} \cdot \vec{Y}}{\lvert\vec{X}\rvert \lvert\vec{Y}\rvert} = \frac{\sum_{1 \leq i \leq n} x_i y_i}{\sqrt{\sum_{1 \leq i \leq n} x_i^2} \sqrt{\sum_{1 \leq i \leq n} y_i^2}}$
Coeficiente de *Pearson*	$P(X,Y) = \frac{\sum_{1 \leq i \leq n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{1 \leq i \leq n} (x_i - \bar{x})^2} \sqrt{\sum_{1 \leq i \leq n} (y_i - \bar{y})^2}}$
Coeficiente de *Tanimoto*	$T(X,Y) = \frac{\vec{X} \cdot \vec{Y}}{\lvert\vec{X}\rvert^2 + \lvert\vec{Y}\rvert^2 - \vec{X} \cdot \vec{Y}}$

Bayes ingênuo

$$c_{MAP} = \arg\max_{c \in C} \hat{P}(c|x) \quad (1)$$

$$= \arg\max_{c \in C} \frac{\hat{P}(x|c)\hat{P}(c)}{\cancel{\hat{P}(x)}} \quad (2)$$

$$= \arg\max_{c \in C} \hat{P}(x|c)\hat{P}(c) \quad (3)$$

$$\hat{P}(x|c) = \hat{P}(x_1, x_2, ..., x_n|c) = \hat{P}(x_1|c)\hat{P}(x_2|c) \ ... \ \hat{P}(x_n|c) \quad (4)$$

Bayes Ingênuo

$$c_{MAP} = \arg\max_{c \in C} \hat{P}(c) \prod_{1 \leq i \leq n} \hat{P}(x_i|c) \tag{5}$$

$$\hat{P}(c) = \frac{N_c}{N} \tag{6}$$

$$\hat{P}(x|c) = \frac{T_{cx} + 1}{\sum_{x' \in V} T_{cx'} + 1} \tag{7}$$

Medida *tf-idf*

- Ordenação do resultado pela relevância dos documentos
- Stop words e normalização
- *Term frequency* ($tf_{t,d} = \#ocorrências$)
- *Inverse document frequency* ($idf_t = \log \frac{N}{df_t}$)

Medida *tf-idf*

- Peso composto: $tf\text{-}idf_{t,d} = tf_{t,d} \cdot idf_t$
  - É alto: t ocorre muitas vezes em d e em poucos documentos
  - Diminui: t ocorre menos vezes em d ou em muitos documentos
  - É muito baixo: t ocorre em quase todos os documentos
- Relevância de d para q
  - $R_{d,q} = \sum_{t \in q} tf\text{-}idf_{t,d}$
- Modelo de espaço vetorial
- Similaridade de cosseno
- *Queries* como documentos

Okapi *BM25*

- Princípio de Ordenação por Probabilidade
- Evento L: D é relevante para Q
- Bayes: $P(L|D) = \frac{P(D|L)P(L)}{P(D)}$
- Log da chance satisfaz o princípio

$$\log \frac{P(L|D)}{P(\bar{L}|D)} = \log \frac{P(D|L)P(L)}{P(D|\bar{L})P(\bar{L})}$$

$$= \log \frac{P(D|L)}{P(D|\bar{L})} + \log \cancel{\frac{P(L)}{P(\bar{L})}} \quad (8)$$

$$R\text{-}PRIM_D = \log \frac{P(D|L)}{P(D|\bar{L})} \quad (9)$$

Okapi *BM25*

- Suposição de independência de atributos

$$R\text{-}PRIM_D = \log \prod_i \frac{P(A_i = a_i|L)}{P(A_1 = a_1|\bar{L})} \quad (10)$$

$$= \sum_i \log \frac{P(A_i = a_i|L)}{P(A_1 = a_1|\bar{L})} \quad (11)$$

Okapi *BM25*

- Contabilando ausência como *zero*

$$R\text{-}BASIC_D = R\text{-}PRIM_D - \sum_i \log \frac{P(A_i = 0|L)}{P(A_1 = 0|\bar{L})} \quad (12)$$

$$= \sum_i \left(\log \frac{P(A_i = a_i|L)}{P(A_1 = a_1|\bar{L})} - \log \frac{P(A_i = 0|L)}{P(A_1 = 0|\bar{L})} \right) \quad (13)$$

$$= \sum_i \log \frac{P(A_i = a_i|L)P(A_1 = 0|\bar{L})}{P(A_1 = a_1|\bar{L})P(A_i = 0|L)} \quad (14)$$

Okapi *BM25*

- Peso W_i para cada termo do documento

$$W_i = log \frac{P(A_i{=}a_i|L)P(A_1{=}0|\overline{L})}{P(A_1{=}a_1|\overline{L})P(A_i{=}0|L)}$$

- $R\text{-}BASIC_D = \sum_i W_i$
- $p_i = P(t_i ocorre|L)$ e $\overline{p}_i = P(t_i ocorre|\overline{L})$
- $w_i = \log \frac{p_i(1-\overline{p_i})}{\overline{p_i}(1-p)}$

Okapi *BM25*

Tabela de contingência de incidência dos termos na coleção

	Relevante	Irrelevante	Incidência na coleção
t ocorre	r	$n-r$	n
t não ocorre	$R-r$	$N-n-R+r$	$N-n$
total de documentos	R	$N-R$	N

- $p = \frac{r}{R}$ e $\overline{p} = \frac{n-r}{N-R}$
- $w = \log \frac{r(N-n-R+r)}{(R-r)(n-r)}$

Okapi *BM25*

- Introduzindo fator de correção
$$rW = \log \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)}$$
- Considerando frequência dos termos
$$RD_{t,D} = \frac{tf_{t,D}(k_1+1)}{k_1((1-b)+b\frac{l_d}{l_{avg}})+tf_{t,D}}$$
- Consultas longas
$$RQ_{t,Q} = \frac{(k_3+1)qtf_{t,Q}}{k_3+qtf_{t,Q}}$$
- Estimativa de relevância
$$R_{D,Q} = \sum_{t \in Q} RW_t \cdot RD_{t,D} \cdot RQ_{t,Q}$$

Apriori

- Descoberta de correlações e padrões frequentes
  - Identificação de conjuntos frequentes
  - Geração de regras de associação
- Suporte e confiança
- Identificação de conjuntos frequentes sem analisar conjunto das partes

Apriori

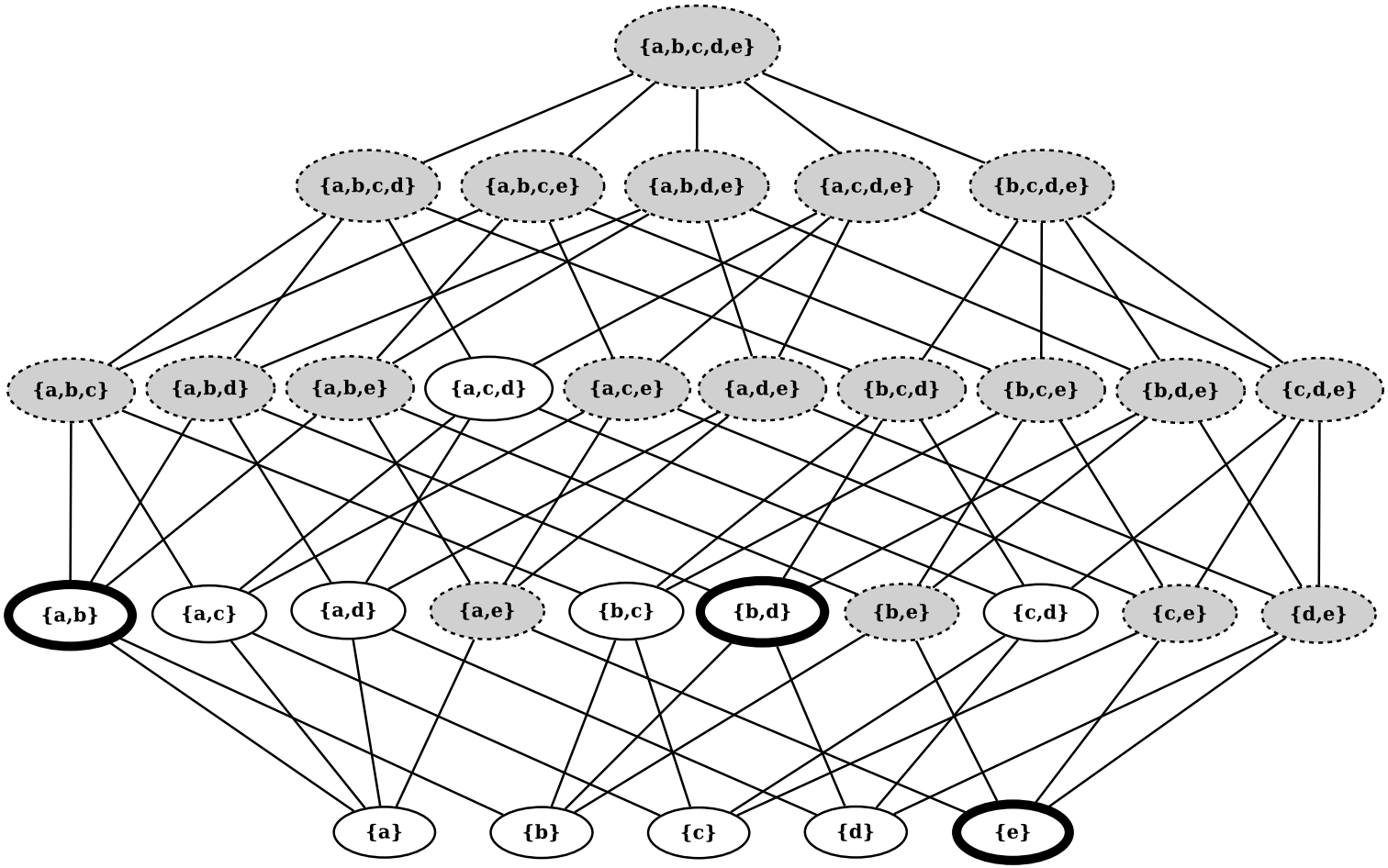

Figure: Geração de conjuntos candidatos pelo algoritmo Apriori