Pesquisa e Publicação de Informação

Operações sobre Queries

Nuno D. Mendes

Licenciatura em Sistemas e Tecnologias de Informação

27 Abr 2012

ISEGI – UNL

Operações sobre Queries

Motivação

- A especificação de uma *query* de modo a traduzir uma **necessidade de informação** não é uma tarefa trivial

Operações sobre Queries

Motivação

- A especificação de uma *query* de modo a traduzir uma **necessidade de informação** não é uma tarefa trivial
- A *query* inicial deve ser vista como uma **especificação imperfeita/incompleta**, porque o utilizador não tem necessariamente consciência nem do **modelo** de Pesquisa de Informação usado, nem das características da **colecção de objectos informacionais** que está a interrogar

Operações sobre Queries

Motivação

- A especificação de uma *query* de modo a traduzir uma **necessidade de informação** não é uma tarefa trivial
- A *query* inicial deve ser vista como uma **especificação imperfeita/incompleta**, porque o utilizador não tem necessariamente consciência nem do **modelo** de Pesquisa de Informação usado, nem das características da **colecção de objectos informacionais** que está a interrogar
- Pretende-se optimizar a *query* de modo a identificar todos os documentos relevantes para a necessidade de informação

Operações sobre Queries

Motivação

- A especificação de uma *query* de modo a traduzir uma **necessidade de informação** não é uma tarefa trivial
- A *query* inicial deve ser vista como uma **especificação imperfeita/incompleta**, porque o utilizador não tem necessariamente consciência nem do **modelo** de Pesquisa de Informação usado, nem das características da **colecção de objectos informacionais** que está a interrogar
- Pretende-se optimizar a *query* de modo a identificar todos os documentos relevantes para a necessidade de informação

Tipos de operações

- Expansão da *query*: são introduzidos novos termos de pesquisa

Operações sobre Queries

Motivação

- A especificação de uma *query* de modo a traduzir uma **necessidade de informação** não é uma tarefa trivial
- A *query* inicial deve ser vista como uma **especificação imperfeita/incompleta**, porque o utilizador não tem necessariamente consciência nem do **modelo** de Pesquisa de Informação usado, nem das características da **colecção de objectos informacionais** que está a interrogar
- Pretende-se optimizar a *query* de modo a identificar todos os documentos relevantes para a necessidade de informação

Tipos de operações

- Expansão da *query*: são introduzidos novos termos de pesquisa
- Re-pesagem dos termos da *query*: são atribuídos novos pesos aos termos da *query*

Operações sobre Queries

Motivação

- A especificação de uma *query* de modo a traduzir uma **necessidade de informação** não é uma tarefa trivial
- A *query* inicial deve ser vista como uma **especificação imperfeita/incompleta**, porque o utilizador não tem necessariamente consciência nem do **modelo** de Pesquisa de Informação usado, nem das características da **colecção de objectos informacionais** que está a interrogar
- Pretende-se optimizar a *query* de modo a identificar todos os documentos relevantes para a necessidade de informação

Tipos de operações

- Expansão da *query*: são introduzidos novos termos de pesquisa
- Re-pesagem dos termos da *query*: são atribuídos novos pesos aos termos da *query*

Estratégias

- Feedback de Relevância

Operações sobre Queries

Motivação

- A especificação de uma *query* de modo a traduzir uma **necessidade de informação** não é uma tarefa trivial
- A *query* inicial deve ser vista como uma **especificação imperfeita/incompleta**, porque o utilizador não tem necessariamente consciência nem do **modelo** de Pesquisa de Informação usado, nem das características da **colecção de objectos informacionais** que está a interrogar
- Pretende-se optimizar a *query* de modo a identificar todos os documentos relevantes para a necessidade de informação

Tipos de operações

- Expansão da *query*: são introduzidos novos termos de pesquisa
- Re-pesagem dos termos da *query*: são atribuídos novos pesos aos termos da *query*

Estratégias

- Feedback de Relevância
- Análise Local

Operações sobre Queries

Motivação

- A especificação de uma *query* de modo a traduzir uma **necessidade de informação** não é uma tarefa trivial
- A *query* inicial deve ser vista como uma **especificação imperfeita/incompleta**, porque o utilizador não tem necessariamente consciência nem do **modelo** de Pesquisa de Informação usado, nem das características da **colecção de objectos informacionais** que está a interrogar
- Pretende-se optimizar a *query* de modo a identificar todos os documentos relevantes para a necessidade de informação

Tipos de operações

- Expansão da *query*: são introduzidos novos termos de pesquisa
- Re-pesagem dos termos da *query*: são atribuídos novos pesos aos termos da *query*

Estratégias

- Feedback de Relevância
- Análise Local
- Análise Global

Feedback de Relevância

- Recebe feedback do utilizador quanto à relevância dos documentos obtidos e modifica a *query* inicial de acordo com essa informação

Feedback de Relevância

- Recebe feedback do utilizador quanto à relevância dos documentos obtidos e modifica a *query* inicial de acordo com essa informação

No Modelo Vectorial

- A_q^+, conjunto de documentos relevantes identificados pelo utilizador dentre os documentos recuperados para a *query* inicial q
- A_q^-, conjunto respectivo de documentos identificados como não-relevantes
- $R_{q^*} \subseteq D$, conjunto de documentos relevantes para a query ideal q^* que satisfaz a necessidade de informação.

É possível demonstrar que

$$\vec{q^*} = \frac{1}{|R_{q^*}|} \sum_{\vec{d_j} \in R_{q^*}} \vec{d_j} - \frac{1}{N - |R_{q^*}|} \sum_{\vec{d_j} \notin R_{q^*}} \vec{d_j}$$

mas o conjunto R_{q^*} não é conhecido *a priori*.

Feedback de Relevância

- Recebe feedback do utilizador quanto à relevância dos documentos obtidos e modifica a *query* inicial de acordo com essa informação

No Modelo Vectorial

Estratégias clássicas

Rochio $\vec{q}_m = \alpha\ \vec{q} + \frac{\beta}{|A_q^+|}\ \sum_{\vec{d}_j \in A_q^+} \vec{d}_j - \frac{\gamma}{|A_q^-|}\ \sum_{\vec{d}_j \in A_q^-} \vec{d}_j$

IDE regular $\vec{q}_m = \alpha\ \vec{q} + \beta\ \sum_{\vec{d}_j \in A_q^+} \vec{d}_j - \gamma\ \sum_{\vec{d}_j \in A_q^-} \vec{d}_j$

IDE Dec-Hi $\vec{q}_m = \alpha\ \vec{q} + \beta\ \sum_{\vec{d}_j \in A_q^+} \vec{d}_j - \gamma\ \arg\max_{\vec{d}_j \in A_q^-} \sigma(d_j, q)$

Feedback de Relevância

- Recebe feedback do utilizador quanto à relevância dos documentos obtidos e modifica a *query* inicial de acordo com essa informação

No Modelo Vectorial

Estratégias clássicas

Rochio $\vec{q}_m = \alpha\ \vec{q} + \frac{\beta}{|A_q^+|}\ \sum_{\vec{d}_j \in A_q^+} \vec{d}_j - \frac{\gamma}{|A_q^-|}\ \sum_{\vec{d}_j \in A_q^-} \vec{d}_j$

IDE regular $\vec{q}_m = \alpha\ \vec{q} + \beta\ \sum_{\vec{d}_j \in A_q^+} \vec{d}_j - \gamma\ \sum_{\vec{d}_j \in A_q^-} \vec{d}_j$

IDE Dec-Hi $\vec{q}_m = \alpha\ \vec{q} + \beta\ \sum_{\vec{d}_j \in A_q^+} \vec{d}_j - \gamma\ \arg\max_{\vec{d}_j \in A_q^-} \sigma(d_j, q)$

Esta abordagem permite expansão da *query* (novos termos são introduzidos porque o seu peso deixa de ser nulo), e os termos originais têm um novo peso atribuído.

Feedback de Relevância

- Recebe feedback do utilizador quanto à relevância dos documentos obtidos e modifica a *query* inicial de acordo com essa informação

No Modelo Probabilístico

Ver aula anterior, mas em vez de considerar os r documentos mais relevantes segundo a estimativa inicial, considerar os r documentos indicados pelo utilizador.

Feedback de Relevância

- Recebe feedback do utilizador quanto à relevância dos documentos obtidos e modifica a *query* inicial de acordo com essa informação

Avaliação de Estratégias de Feedback de Relevância

As abordagens que pretendam **avaliar o desempenho** de estratégias diferentes de feedback de relevância devem considerar curvas de precisão/recall apenas em relação aos **documentos residuais**, ou seja, em relação aos documentos que se sabem relevantes, mas que ainda não tenham sido indicados pelo utilizador.

Análise Local

Nesta estratégia, os documentos extraídos por via da *query* inicial são analisados de modo a encontrar termos para expandir a *query* de forma automática.

Análise Local

Nesta estratégia, os documentos extraídos por via da *query* inicial são analisados de modo a encontrar termos para expandir a *query* de forma automática.

Expansão de Queries por Clustering Local

Para uma dada *query*, q, A_q designa o conjunto local de documentos extraído, e $\mathcal{L}_q$ um conjunto de todos os termos que ocorrem em A_q, designado de *vocabulário local*.

Clusters de Associação

$f_{i,j}$, a frequência do termo k_i no documento $d_j \in A_q$
$M = \{m_{ij}\}_{|\mathcal{L}_q| \times |A_q|}$ é a matriz de associação com $|\mathcal{L}_q|$ linhas e $|A_q|$ colunas , tal que $m_{ij} = f_{i,j}$
M^t a matriz transposta de M, então
$S = MM^t = \{s_{uv}\}$ é a matriz local de associação de termos.
Cada elemento s_{uv} exprime uma correlação c_{uv} entre o termo k_u e o termo k_v, i.e.

$$c_{uv} = \sum_{d_j \in A_q} f_{u,j} \times f_{v,j}$$

Se fizermos $s_{uv} = c_{uv}$ dizemos que as co-ocorrências de termos são não-normalizadas.
Se fizermos $s_{uv} = \frac{c_{uv}}{c_{uu} + c_{vv} - c_{uv}}$, dizemos que são normalizadas.

$S_u(n)$ define o cluster local de associação e é definido recursivamente como:

$$S_u(n) = \begin{cases} \emptyset & \text{if } n = 0 \\ \max\{s_{uv} \mid u \neq v,\ u, v \in K\} & \text{if } n = 1 \\ S_u(n-1) \cup \max\{s_{uv} \mid s_{uv} \notin S_u(n-1) \wedge u \neq v,\ u, v \in K\} & \text{if } n > 1 \end{cases}$$

Clusters de Associação

$f_{i,j}$, a frequência do termo k_i no documento $d_j \in A_q$
$M = \{m_{ij}\}_{|\mathcal{L}_q| \times |A_q|}$ é a matriz de associação com $|\mathcal{L}_q|$ linhas e $|A_q|$ colunas , tal que $m_{ij} = f_{i,j}$
M^t a matriz transposta de M, então
$S = MM^t = \{s_{uv}\}$ é a matriz local de associação de termos.
Cada elemento s_{uv} exprime uma correlação c_{uv} entre o termo k_u e o termo k_v, i.e.

$$c_{uv} = \sum_{d_j \in A_q} f_{u,j} \times f_{v,j}$$

Se fizermos $s_{uv} = c_{uv}$ dizemos que as co-ocorrências de termos são não-normalizadas.
Se fizermos $s_{uv} = \frac{c_{uv}}{c_{uu} + c_{vv} - c_{uv}}$, dizemos que são normalizadas.

$S_u(n)$ define o cluster local de associação e é definido recursivamente como:

$$S_u(n) = \begin{cases} \emptyset & \text{if } n = 0 \\ \max\{s_{uv} \mid u \neq v,\ u, v \in K\} & \text{if } n = 1 \\ S_u(n-1) \cup \max\{s_{uv} \mid s_{uv} \notin S_u(n-1) \wedge u \neq v,\ u, v \in K\} & \text{if } n > 1 \end{cases}$$

Em vez de usarmos os termos $k_i \in K$ directamente, podemos usar representantes dos

Clusters Métricos

Semelhantes aos clusters de associação, mas consideram as distâncias entre os termos no âmbito dos documentos locais.

r_{ij} é uma ponderação da distância dos termos k_i e k_j num mesmo documento, medida pelo número de termos que intermedeiam entre k_i e k_j, caso k_i e k_j nunca ocorram no mesmo documento então $r_{ij} = \infty$.
$[[k_i]]$ designa o conjunto de termos com o mesmo radical (stem) que k_i, definindo um fecho transitivo.

A matriz S passa a ser definida à custa de:

$$c_{uv} = \sum_{k_i \in [[k_u]]} \sum_{k_j \in [[k_v]]} \frac{1}{r_{ij}}$$

com $s_{uv} = c_{uv}$ no caso não-normalizado e
$s_{uv} = \frac{c_{uv}}{|[[k_u]]| \times |[[k_v]]|}$, no caso normalizado.

Os clusters $S_u(n)$ são definidos do mesmo modo.

Clusters Escalares

Definidos à custa da semelhança do vector de correlações de dois termos ou stems por via de uma medida escalar (e.g. produto interno normalizado)

c_{uv} a correlação entre os termos (ou stems) k_u e k_v.
$\vec{\alpha}_u = (c_{u1}, c_{u2}, \ldots, c_{un})$ o vector de correlações do termo (ou stem) k_u.

A matriz S passa a ser definida directamente como:

$$s_{uv} = \frac{\vec{\alpha}_u \cdot \vec{\alpha}_v}{|\vec{\alpha}_u| \times |\vec{\alpha}_v|}$$

Os clusters $S_u(n)$ são definidos do mesmo modo.

Análise Global

Nesta estratégia, a *query* inicial é expandida com termos semelhantes identificados através da análise de toda a colecção de documentos

Análise Global

Nesta estratégia, a *query* inicial é expandida com termos semelhantes identificados através da análise de toda a colecção de documentos

Expansão de Queries pela construção de um Tesouro de Semelhança

1. A *query* q passa a ser expressa num espaço de conceitos, semelhante ao modelo vectorial, mas as dimensões representam documentos, e as coordenadas são calculadas à custa da idf (inverse document frequency, semelhante à itf)
2. Com base na noção de distância no espaço de conceitos, é calculada a semelhança de cada termo com a *query* q
3. A *query* q é expandida com os r termos mais semelhantes

Análise Global

Nesta estratégia, a *query* inicial é expandida com termos semelhantes identificados através da análise de toda a colecção de documentos

Expansão de Queries pela construção de um Tesouro de Semelhança

1. A *query* q passa a ser expressa num espaço de conceitos, semelhante ao modelo vectorial, mas as dimensões representam documentos, e as coordenadas são calculadas à custa da idf (inverse document frequency, semelhante à itf)
2. Com base na noção de distância no espaço de conceitos, é calculada a semelhança de cada termo com a *query* q
3. A *query* q é expandida com os r termos mais semelhantes

Expansão de Queries pela construção de um Tesouro Estatístico

Os documentos da colecção são agrupados usando um algoritmo de clustering (complete link) e a noção de semelhança do modelo vectorial. As classes de termos do tesouro são calculadas a partir dos termos **raros** de cada cluster de documentos.