Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Realimentação de relevância e Expansão de consulta

Profa. Giseli Rabello Lopes

Roteiro

- Introdução
- Um framework para métodos de realimentação
  - Realimentação de relevância explícita
  - Realimentação explícita através de cliques
  - Realimentação implícita através de análise local
  - Realimentação implícita através de análise global
- Referências

Introdução

- A maioria dos usuários encontra dificuldades para formular consultas bem projetadas para fins de recuperação
- No entanto, usuários muitas vezes precisam reformular suas consultas para obter os resultados que lhes interessam
  - Assim, a primeira formulação da consulta deve ser tratada como uma tentativa inicial de recuperar informações relevantes
  - Os documentos inicialmente recuperados poderiam ser analisados por relevância e usados para melhorar a formulação da consulta inicial

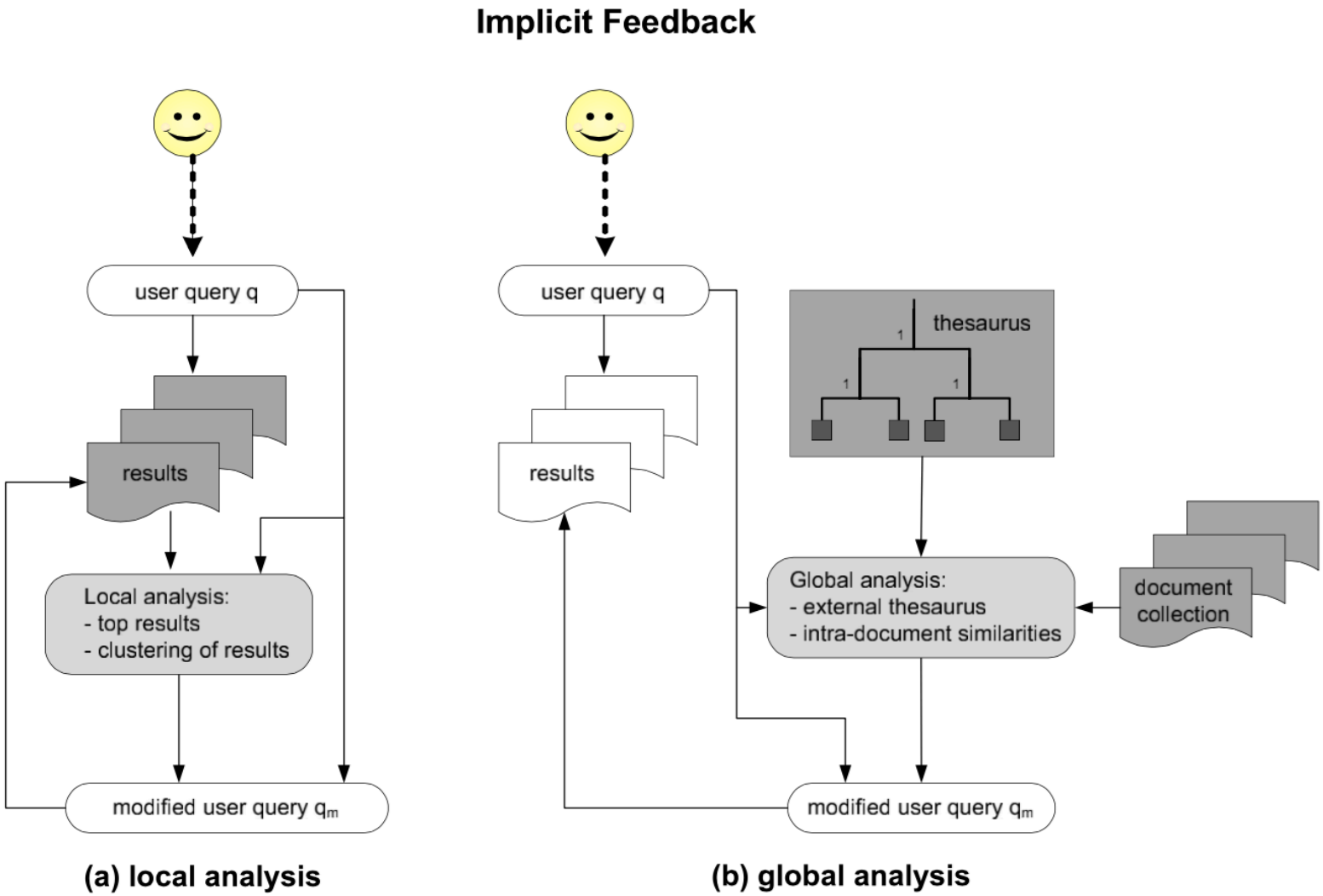
Introdução

- Duas abordagens básicas de métodos de realimentação:
  - **Realimentação explícita**, em que a informação para reformulação da consulta é fornecida diretamente pelos usuários, e
  - **Realimentação implícita**, em que a informação para a reformulação da consulta é derivada implicitamente pelo sistema

Um *framework* para métodos de *realimentação*

- Em um ciclo de **realimentação de relevância implícita**, a informação de realimentação é derivada implicitamente pelo sistema
- Existem duas abordagens básicas para a coleta de informações implícitas de realimentação:
  - **Análise local**, que deriva a informação de realimentação dos documentos ranqueados no topo do conjunto de resultados
  - **Análise global**, que deriva a informação de realimentação de fontes externas tais como um tesauro

Informação implícita de realimentação



Análise local

- Análise local consiste em derivar informação de realimentação de documentos recuperados para uma dada consulta q
- Isso é semelhante a um ciclo de realimentação de relevância, mas feito sem o envolvimento do usuário
- Será discutida a seguinte estratégia local: **clustering local**

Clustering local

- A adoção de técnicas de clustering para expansão de consulta tem sido uma abordagem básica em RI
- O procedimento padrão é quantificar correlações entre termos e então usar os termos correlacionados para a expansão da consulta
- Correlações entre termos podem ser quantificadas pelo uso de estruturas globais, tais como **matrizes de associação**
- Entretanto, estruturas globais podem não se adaptar bem para o contexto local definido pela consulta atual
- Para lidar com esse problema, **clustering local** pode ser usado, como será discutido agora

Matriz de correlação entre termos para uma coleção de exemplo [Baeza-Yates & Ribeiro-Neto, 2013]

$$
\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array}
\begin{array}{c}
\begin{array}{cc} d_1 & d_2 \end{array} \\
\left[\begin{array}{cc} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \\ w_{3,1} & w_{3,2} \end{array}\right]
\end{array}
\qquad \mathbf{M}
$$

$$
\times
$$

$$
\begin{array}{c} \\ d_1 \\ d_2 \end{array}
\begin{array}{c}
\begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\
\left[\begin{array}{ccc} w_{1,1} & w_{2,1} & w_{3,1} \\ w_{1,2} & w_{2,2} & w_{3,2} \end{array}\right]
\end{array}
\qquad \mathbf{M}^T
$$

$$
\Downarrow
$$

$$
\begin{array}{c} \\ k_1 \\ k_2 \\ k_3 \end{array}
\begin{array}{c}
\begin{array}{ccc} k_1 & k_2 & k_3 \end{array} \\
\left[\begin{array}{ccc}
w_{1,1}w_{1,1} + w_{1,2}w_{1,2} & w_{1,1}w_{2,1} + w_{1,2}w_{2,2} & w_{1,1}w_{3,1} + w_{1,2}w_{3,2} \\
w_{2,1}w_{1,1} + w_{2,2}w_{1,2} & w_{2,1}w_{2,1} + w_{2,2}w_{2,2} & w_{2,1}w_{3,1} + w_{2,2}w_{3,2} \\
w_{3,1}w_{1,1} + w_{3,2}w_{1,2} & w_{3,1}w_{2,1} + w_{3,2}w_{2,2} & w_{3,1}w_{3,1} + w_{3,2}w_{3,2}
\end{array}\right]
\end{array}
$$

Clustering de associação

- Para uma dada consulta q, sejam
  - D_l: **conjunto de documentos locais**, ou seja, conjunto de documentos recuperados por q
  - N_l: número de documentos em D_l
  - V_l: vocabulário local, ou seja, conjunto de todas as palavras distintas em D_l
  - $f_{i,j}$: frequência de ocorrência do termo k_i em um documento $d_j \in D_l$
  - M_l=[m_{ij}]: matriz de termos por documentos com V_l linhas e N_l colunas
  - $m_{ij} = f_{i,j}$: um elemento da matriz M_l
  - M^T_l : transposta de M_l
- A matriz $\mathbf{C}_\ell = \mathbf{M}_\ell \mathbf{M}_\ell^T$
  - é a matriz de correlação local entre termos

Clustering de associação

- Cada elemento $c_{u,v} \in C_l$ expressa uma correlação entre termos k_u e k_v
- Esse relacionamento entre termos é baseado em suas coocorrências dentro de documentos da coleção
- Quanto maior o número de documentos nos quais dois termos coocorrem, mais forte sua correlação
- A força de correlação pode ser usada para definir clusters locais de termos próximos
- Termos no mesmo cluster podem ser então usados para expansão de consulta

Clustering de associação

- Um clustering de associação é computado a partir da matriz de correlação local C_l
- Para isso, redefinimos os fatores de correlação $c_{u,v}$ entre qualquer par de termos k_u e k_v, como segue:

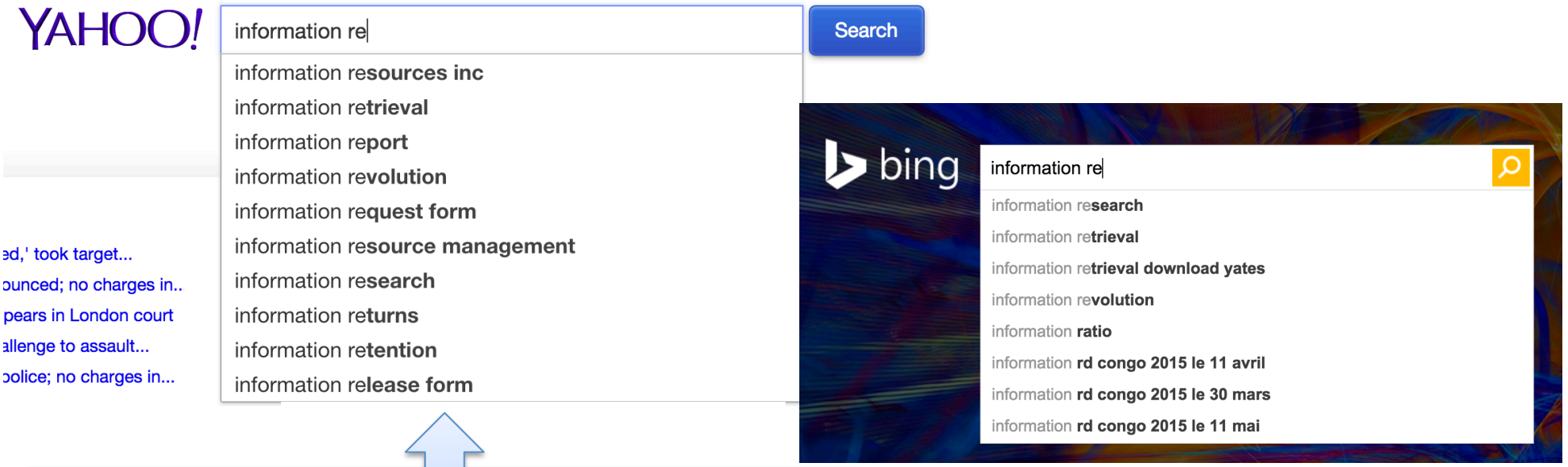
$$c_{u,v} = \sum_{d_j \in D_l} f_{u,j} \times f_{v,j}$$

- Nesse caso a matriz de correlação é referenciada como uma **matriz de associação local**
- A motivação é que termos que coocorrem frequentemente dentro de documentos têm uma associação de sinonímia

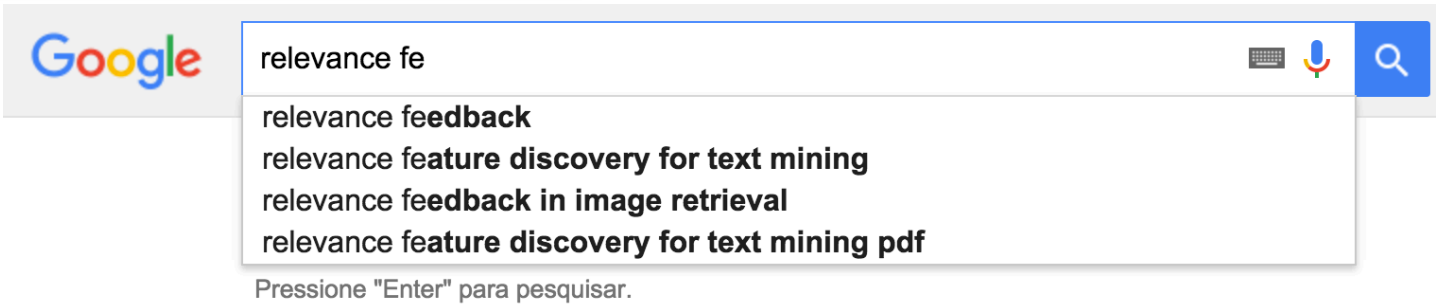
Clustering de associação

- Dada uma matriz de associação local C_l, podemos usá-la para construir clusterings de associação locais como segue
- Seja $C_u(n)$ uma função que retorna os n maiores fatores $c_{u,v} \in C_l$, onde v varia sobre o conjunto de termos locais e $v \neq u$
- Então, $C_u(n)$ define um cluster de associação local, uma vizinhança, em torno do termo k_u
- Dada uma consulta q, estamos normalmente interessados em encontrar clusters apenas para os $|q|$ termos da consulta
- Isso significa que tais clusters podem ser computados eficientemente em tempo de consulta

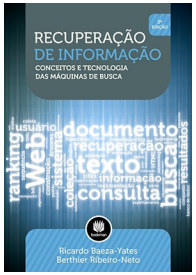
Assistência na consulta (expansão assistida)
[Manning et al. 2008]



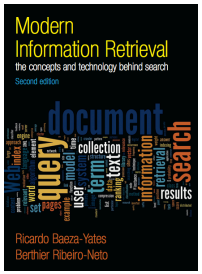
Would you expect such a feature to increase the query volume at a search engine?



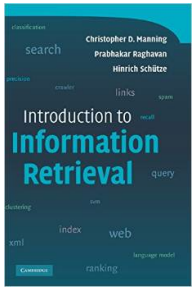
Referências



- Baeza-Yates, R.; Ribeiro-Neto, B. Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca. 2 ed. Bookman, 2013.



- Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Wokingham, UK: Addison-Wesley, 2 ed., 2011.



- Manning, C. D.; Raghavan, P.; Schütze, H. Introduction to Information Retrieval. Cambridge University Press, 2008.

  Online edition 2009: http://nlp.stanford.edu/IR-book/

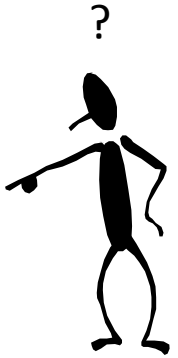
Universidade Federal do Rio de Janeiro (UFRJ)
Departamento de Ciência da Computação (DCC)

DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO - UFRJ -

Recuperação da Informação (MAB605)

Dúvidas?

Profa. Giseli Rabello Lopes
giseli@dcc.ufrj.br
CCMN - DCC - Sala E-2012



Material selecionado e traduzido de slides do capítulo 4 do livro [Baeza-Yates & Ribeiro-Neto, 2013]