

Realimentação de Relevância

Wendel Melo

Faculdade de Computação
Universidade Federal de Uberlândia

Recuperação da Informação

Realimentação de Relevância

- Ciclo de realimentação onde uma consulta q recebida do usuário é transformada em uma consulta modificada q_m :

Realimentação de Relevância

- Ciclo de realimentação onde uma consulta q recebida do usuário é transformada em uma consulta modificada q_m :
 - A expectativa é que q_m possa atender melhor a necessidade de informação do usuário;

Realimentação de Relevância

- Ciclo de realimentação onde uma consulta q recebida do usuário é transformada em uma consulta modificada q_m :
 - A expectativa é que q_m possa atender melhor a necessidade de informação do usuário;
 - Essa transformação pode ser feita, por exemplo, através de informações obtidas por meio de avaliação do resultado por parte do usuário, ou análise automática do topo do ranking.

Realimentação de Relevância

- Ciclo de realimentação onde uma consulta q recebida do usuário é transformada em uma consulta modificada q_m :
 - A expectativa é que q_m possa atender melhor a necessidade de informação do usuário;
 - Essa transformação pode ser feita, por exemplo, através de informações obtidas por meio de avaliação do resultado por parte do usuário, ou análise automática do topo do ranking.
- Essa filosofia já é incorporada, de certo modo, pelo modelo probabilístico. Todavia, ela também pode ser usada de modo mais genérico em qualquer modelo, incluindo o próprio probabilístico.

Tipos de Realimentação

Podemos identificar dois tipos de abordagem:

- **Realimentação explícita:** o usuário fornece diretamente as informações para a reformulação da consulta, por exemplo, classificando os docs no topo do ranking da consulta original.
 - Tende a ser dispendioso para o usuário.
- **Realimentação implícita:** o próprio sistema de RI produz as informações para a reformulação da consulta, por exemplo, analisando características em comum presentes nos docs no topo do ranking da consulta original, ou analisando fontes de informação externas.

Realimentação de Relevância

- O processo de formulação de uma consulta modificada que incorpore novos termos em relação à consulta original é denominado **expansão de consulta**.
- A expansão da consulta pode ser realizada tanto através de métodos de realimentação explícita quanto implícita.

Realimentação de Relevância

A modelagem de um ciclo de realimentação se constitui em duas etapas:

- 1) Determinar as informações de realimentação que estariam relacionadas à consulta q . Essas informações poderiam ser obtidas de modo explícito do usuário, ou implícito a partir de informações do sistema;
- 2) Determinar como usar as informações da etapa 1 para transformar a consulta original na expectativa de melhorá-la (fornecer resultados mais satisfatórios ao usuário).

Realimentação de Relevância Explícita

- Um exemplo clássico de realimentação de relevância explícita para o modelo vetorial é o **método de Rocchio**:
- O método de Rocchio parte dos pressupostos:
 - 1) Os documentos relevantes terão vetores de representação com certas semelhanças entre si.
 - 2) Os documentos não relevantes terão vetores de representação diferentes dos relevantes.
- A ideia básica é reformular a consulta, a partir da classificação do usuário, de modo que seu vetor de representação se aproxime dos docs relevantes e se afaste dos não relevantes.

Método de Rocchio

Sejam:

- D_r : conjunto de docs relevantes recuperados (avaliação do usuário);
- D_n : conjunto de docs não relevantes recuperados (avaliação do usuário);
- α, β, γ : constantes de ajuste (não negativas)
- O método de Rocchio calcula o vetor da consulta modificada q_m , a partir do vetor da consulta original q , segundo a expressão:

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{d_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{d_j \in D_n} \vec{d}_j$$

Método de Rocchio

- Note que o termo $\frac{\sum_{d_j \in D_r} \vec{d}_j}{|D_r|}$ representa o vetor médio dos docs relevantes
- Note que o termo $\frac{\sum_{d_j \in D_n} \vec{d}_j}{|D_n|}$ representa o vetor médio dos docs não relevantes

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{d_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{d_j \in D_n} \vec{d}_j$$

Método de Rocchio

- Observe que a expressão que calcula a consulta modificada pode incorporar pesos não nulos referentes a termos que não estavam na consulta original;
- Assim, na prática, é como se a consulta modificada pudesse incorporar novos termos.
- Através do ajuste dos parâmetros α , β , γ , pode-se ponderar a importância do vetor da consulta original e dos vetores dos docs em D_r e D_n no vetor da consulta modificada.

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{d_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{d_j \in D_n} \vec{d}_j$$

Realimentação de Relevância Explícita

- Os métodos de realimentação de relevância explícita possuem a vantagem de serem mais sensíveis à captação da subjetividade de cada usuário para melhorar a resposta, pois os próprios usuários avaliam diretamente os resultados.

Realimentação de Relevância Explícita

- Os métodos de realimentação de relevância explícita possuem a vantagem de serem mais sensíveis à captação da subjetividade de cada usuário para melhorar a resposta, pois os próprios usuários avaliam diretamente os resultados.
- A avaliação da resposta por parte de um determinado usuário trará uma carga de sua subjetividade que pode ajudar o sistema a chegar a resultados que lhe sejam mais satisfatórios.

Realimentação de Relevância Explícita

- Os métodos de realimentação de relevância explícita possuem a vantagem de serem mais sensíveis à captação da subjetividade de cada usuário para melhorar a resposta, pois os próprios usuários avaliam diretamente os resultados.
- A avaliação da resposta por parte de um determinado usuário trará uma carga de sua subjetividade que pode ajudar o sistema a chegar a resultados que lhe sejam mais satisfatórios.
- Por outro lado, o processo de avaliar as respostas pode ser muito dispendioso; Usuários podem não estar dispostos a avaliar resultados, especialmente em sistemas WEB.

Realimentação de Relevância Implícita

- Por serem menos incômodos aos usuários, há uma concentração maior de pesquisa e aplicação dos métodos de realimentação implícita, que podem ser subdivididos em:
- **Métodos de análise local:** usam informações referentes a resposta gerada para a consulta inicial, por exemplo, analisando o topo do ranqueamento.

Realimentação de Relevância Implícita

- Por serem menos incômodos aos usuários, há uma concentração maior de pesquisa e aplicação dos métodos de realimentação implícita, que podem ser subdivididos em:
- **Métodos de análise local:** usam informações referentes a resposta gerada para a consulta inicial, por exemplo, analisando o topo do ranqueamento.
- **Métodos de análise global:** usam fontes externas de informação, como tesouros (documento que relaciona termos de significado semelhante) e relações entre termos extraídas da coleção de documentos.

Realimentação Implícita por Análise Local

- Pode ser realizada através de **técnicas de agrupamento (clustering) local**:
 - A ideia principal consiste em gerar agrupamentos (cluster) de termos supostamente relacionados.
 - Esses agrupamentos podem então ser utilizados para expandir a consulta com novos termos presentes nos mesmos agrupamentos dos termos da consulta original.

Realimentação Implícita por Análise Local

- Por exemplo, suponha que, para uma consulta q , recupera-se uma lista inicial de documentos.
- Suponha que a consulta q engloba o termo A , e que, ao analisar os documentos no topo do ranqueamento, foi detectado que, frequentemente, A aparece com os termos B e C , embora B e C não estejam na consulta original.
- Nesse caso, temos um agrupamento local envolvendo os termos A , B e C , pois os mesmos aparecem juntos com frequência no contexto **local** da consulta q .
- Desse modo, podemos expandir a consulta adicionando à esta os termos B e/ou C .

Realimentação Implícita por Análise Local

- No exemplo anterior, o agrupamento envolvendo A, B e C é dito **local** porque foi construído apenas no contexto da consulta q .
- Uma outra consulta q' , tal que $q' \neq q$, que também envolva o termo A poderia gerar um agrupamento local diferente, com outros termos no lugar de B e C;
- Por sua vez, uma técnica de análise global produz agrupamentos observando a base de documentos como um todo, sem a consideração de nenhuma consulta em particular.

Realimentação Implícita por Análise Local

- Assim, técnicas de análise global podem ser aplicadas antes do sistema entrar em operação, já na etapa de indexação;
- Em contrapartida, as técnicas de análise local dependem da consulta recebida. Por isso, são aplicadas no processamento da resposta ao usuário;
- Por essa razão, a análise global pode utilizar técnicas computacionalmente mais pesadas, pois é realizada antes do usuário utilizar o sistema;
- A análise local, por sua vez, possui uma preocupação maior com o tempo de execução das técnicas adotadas, pois o usuário está esperando uma resposta;

Realimentação Implícita por Análise Local

- As técnicas de agrupamento se baseiam, em geral, em uma matriz de correlação de termos C de T linhas e T colunas, onde T é o número de termos do vocabulário.

$$C = \begin{matrix} & \begin{matrix} k_1 & k_2 & k_3 & \dots & k_T \end{matrix} \\ \begin{matrix} k_1 \\ k_2 \\ k_3 \\ \vdots \\ k_T \end{matrix} & \begin{bmatrix} & & & & \\ & & & & \\ & & c_{23} & & \\ & c_{32} & & & \\ & & & & \end{bmatrix} \end{matrix}$$

O elemento c_{ij} representa a correlação entre os termos k_i e k_j .

Na linha (coluna) i , temos as correlações de todos os termos com k_i .

Realimentação Implícita por Análise Local

$$C = \begin{matrix} & \begin{matrix} k_1 & k_2 & k_3 & \dots & k_T \end{matrix} \\ \begin{matrix} k_1 \\ k_2 \\ k_3 \\ \vdots \\ k_T \end{matrix} & \begin{bmatrix} & & & & \\ & & & & \\ & & c_{23} & & \\ & c_{32} & & & \\ & & & & \end{bmatrix} \end{matrix}$$

O elemento c_{ij} representa a correlação entre os termos k_i e k_j .

- A matriz $C \geq 0$ é quadrada e simétrica. Valores altos para c_{ij} indicam que k_i e k_j estão fortemente relacionados no contexto em questão. Valores próximos a zero indicam baixa correlação.

Exemplo de uso de matriz de correlação

- Seja C' a matriz de correlação local (isto é, construída no contexto de uma consulta q), e c'_{uv} o coeficiente em C' relativo aos termos k_u e k_v (isto é, o valor da correlação local entre k_u e k_v).
- Exemplo:** Suponha a seguinte matriz de correlação com vocabulário de 7 termos e a consulta $q = k_2 \text{ AND } k_4$:

$$C' = \begin{matrix} & \begin{matrix} k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & k_7 \end{matrix} \\ \begin{bmatrix} 0 & 25 & 30 & 28 & 0 & 38 & 10 \\ 40 & 28 & 10 & 61 & 0 & 150 & 6 \end{bmatrix} & \begin{matrix} k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \\ k_7 \end{matrix} \end{matrix}$$

Exemplo de uso de matriz de correlação

- Seja C^l a matriz de correlação local (isto é, construída no contexto de uma consulta q), e c'_{uv} o coeficiente em C^l relativo aos termos k_u e k_v (isto é, o valor da correlação local entre k_u e k_v).
- Exemplo:** Suponha a seguinte matriz de correlação com vocabulário de 7 termos e a consulta $q = k_2 \text{ AND } k_4$:

$$C^l = \begin{matrix} & \begin{matrix} k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & k_7 \end{matrix} \\ \begin{bmatrix} 0 & 25 & 30 & 28 & 0 & 38 & 10 \\ 40 & 28 & 10 & 61 & 0 & 150 & 6 \end{bmatrix} & \begin{matrix} k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \\ k_7 \end{matrix} \end{matrix}$$

Só é preciso calcular as linhas da matriz referentes aos termos da consulta.

Exemplo de uso de matriz de correlação

- Seja C' a matriz de correlação local (isto é, construída no contexto de uma consulta q), e c'_{uv} o coeficiente em C' relativo aos termos k_u e k_v (isto é, o valor da correlação local entre k_u e k_v).
- Exemplo:** Suponha a seguinte matriz de correlação com vocabulário de 7 termos e a consulta $q = k_2 \text{ AND } k_4$:

$$C' = \begin{matrix} & \begin{matrix} k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & k_7 \end{matrix} \\ \begin{bmatrix} 0 & 25 & 30 & 28 & 0 & 38 & 10 \\ 40 & 28 & 10 & 61 & 0 & 150 & 6 \end{bmatrix} & \begin{matrix} k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \\ k_7 \end{matrix} \end{matrix}$$

Só é preciso calcular as linhas da matriz referentes aos termos da consulta.

Os agrupamentos são montados observando os termos com maior correlação com os termos da consulta.

Exemplo de uso de matriz de correlação

- Exemplo:** Suponha a seguinte matriz de correlação com vocabulário de 7 termos e a consulta $q = k_2 \text{ AND } k_4$:

$$C^l = \begin{matrix} & \begin{matrix} k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & k_7 \end{matrix} \\ \begin{matrix} k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \\ k_7 \end{matrix} & \begin{bmatrix} 0 & 25 & \underline{30} & 28 & 0 & \underline{38} & 10 \\ \underline{40} & 28 & 10 & 61 & 0 & \underline{150} & 6 \end{bmatrix} \end{matrix}$$

Só é preciso calcular as linhas da matriz referentes aos termos da consulta.

Os agrupamentos são montados observando os termos com maior correlação com os termos da consulta.

- Assim, em relação a k_2 , monta-se o agrupamento com k_2 , k_3 e k_6
- Em relação a k_4 , monta-se o agrupamento com k_4 , k_1 e k_6 .

Exemplo de uso de matriz de correlação

$$C^l = \begin{matrix} & \begin{matrix} k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & k_7 \end{matrix} \\ \begin{bmatrix} 0 & 25 & 30 & 28 & 0 & 38 & 10 \\ 40 & 28 & 10 & 61 & 0 & 150 & 6 \end{bmatrix} & \begin{matrix} k_1 \\ k_2 \\ k_3 \\ k_4 \\ k_5 \\ k_6 \\ k_7 \end{matrix} \end{matrix}$$

Só é preciso calcular as linhas da matriz referentes aos termos da consulta.

Os agrupamentos são montados observando os termos com maior correlação com os termos da consulta.

- Assim, em relação a k_2 , monta-se o agrupamento com k_2 , k_3 e k_6
- Em relação a k_4 , monta-se o agrupamento com k_4 , k_1 e k_6 .
- A quantidade de termos nos agrupamentos é arbitrária, mas, em geral, deseja-se manter os agrupamentos pequenos.

Exemplo de uso de matriz de correlação

- Assim, a consulta modificada gerada será:

$$q_m = k_2 \text{ AND } k_4 \text{ AND } k_1 \text{ AND } k_3 \text{ AND } k_6$$

- A quantidade de termos nos agrupamentos é arbitrária, mas, em geral, deseja-se manter os agrupamentos pequenos.

Realimentação Implícita por Análise Local

Para a determinação de agrupamentos locais, três técnicas são comumente utilizadas:

- Agrupamentos de associação;
- Agrupamentos métricos;
- Agrupamentos escalares.

Cada uma dessas técnicas calculará a matriz de correlação C' de uma forma diferente. A partir da matriz de correlação, determina-se os agrupamentos observando os termos de maior correlação entre si.

Agrupamentos de Associação

- **Não normalizado:** define cada elemento c'_{uv} de C' da seguinte forma:

$$c'_{uv} = \sum_{d_j \in D_l} (f_{uj} \times f_{vj})$$

Onde:

- f_{ij} : frequência do termo k_i no documento d_j ;
- D_l : conjunto de docs recuperados pela consulta q , denominado *conjunto de documentos locais* (lembre-se de que o resultado da consulta está sendo usado para melhorá-la).

Agrupamentos de Associação

- **Normalizado:** Seja \hat{C}^l a matriz de correlação normalizada. Calcule-se cada elemento \hat{c}_{uv}^l como:

$$\hat{c}_{uv}^l = \frac{c_{uv}^l}{c_{uu}^l + c_{vv}^l - c_{uv}^l}$$

Onde:

$$c_{uv}^l = \sum_{d_j \in D_l} (f_{uj} \times f_{vj})$$

Agrupamentos de Associação

- O método de agrupamento de associação possui a vantagem de calcular a matriz de correlação de modo simples e intuitivo;
- No entanto, a matriz de correlação acaba não levando em conta a distância em que os termos aparecem no documento, o que pode ser um fator importante;

Agrupamentos Métricos

- O método de agrupamento métrico, por sua vez, parte da ideia de que dois termos que estejam próximos em um documento tendem a ter maior correlação do que dois termos que estejam distantes.
 - Dois termos que estejam na mesma frase tendem a ter maior correlação do que dois termos em parágrafos distantes.
- Assim, a correlação c'_{uv} entre os termos k_u e k_v é calculada em função das suas distâncias nos documentos.

Agrupamentos Métricos

- Cada elemento c'_{uv} de C' é calculado como:

$$c'_{uv} = \sum_{d_j \in \bar{D}_l(k_u, k_v)} \sum_{p=1}^{f_{uj}} \sum_{q=1}^{f_{vj}} \frac{1}{r(\bar{k}_u(p, j), \bar{k}_v(q, j))}$$

Onde:

- f_{ij} : frequência do termo k_i no documento d_j ;
- $\bar{k}_u(p, j)$: função que retorna a posição da p-ésima aparição do termo k_u no doc d_j (ex: posição referente aos bytes);
- $r(\bar{k}_u(p, j), \bar{k}_v(q, j))$: função que calcula a distância entre a p-ésima aparição de k_u e a q-ésima aparição de k_v no doc d_j (ex: nº de palavras);
- \bar{D}_l : docs locais (retornados pela consulta) que contém ambos k_u e k_v .

Agrupamentos Métricos

- Cada elemento c'_{uv} de C' é calculado como:

$$c'_{uv} = \sum_{d_j \in \bar{D}_l(k_u, k_v)} \sum_{p=1}^{f_{uj}} \sum_{q=1}^{f_{vj}} \frac{1}{r(\bar{k}_u(p, j), \bar{k}_v(q, j))}$$

Onde:

- f_{ij} : frequência do termo k_i no documento d_j ;
- $\bar{k}_u(p, j)$: função que retorna a posição da p-ésima aparição do termo k_u no doc d_j (ex: posição referente aos bytes);
- $r(\bar{k}_u(p, j), \bar{k}_v(q, j))$: função que calcula a distância entre a p-ésima aparição de k_u e a q-ésima aparição de k_v no doc d_j (ex: nº de palavras);
- \bar{D}_l : docs locais (retornados pela consulta) que contém ambos k_u e k_v .

A fórmula considera a distância entre cada aparição de k_u e todas as aparições de k_v .

Agrupamentos Métricos

- Cada elemento c'_{uv} de C' é calculado como:

$$c'_{uv} = \sum_{d_j \in \bar{D}_l(k_u, k_v)} \sum_{p=1}^{f_{uj}} \sum_{q=1}^{f_{vj}} \frac{1}{r(\bar{k}_u(p, j), \bar{k}_v(q, j))}$$

A fórmula considera a distância entre cada aparição de k_u e todas as aparições de k_v .

- Considerar cada par de aparição entre k_u e k_v é uma forma de se lidar com o fato de que o nº de aparições de k_u e k_v pode ser diferente.

Agrupamentos Escalares

- Método adicional para encontrar agrupamentos que usa o conceito da similaridade entre vizinhanças de termos;
- Parte-se da ideia de que termos com vizinhanças semelhantes possuem alguma relação de sinonímia;
- Assim, a relação entre os termos é dita indireta ou induzida pela vizinhança;

Agrupamentos Escalares

- Primeiramente, calcula-se uma matriz C' inicial com os coeficientes de correlação de alguma forma;
- Seja c'_u a linha (vetor) de C' referente ao termo k_u e c'_v a linha (vetor) referente a k_v ;
- Calcula-se então uma nova matriz de correlação local \bar{C}' onde cada coeficiente \bar{c}'_{uv} quantifica uma similaridade entre os vetores c'_u e c'_v obtidos com as correlações iniciais;
- É comum quantificar essa similaridade através do cosseno entre c'_u e c'_v .

Agrupamentos Escalares

- Assim:

$$\bar{c}_{uv}^l = \cos(c_u^l, c_v^l) = \frac{\sum_{i=1}^N c_{ui}^l \times c_{vi}^l}{\sqrt{\sum_{i=1}^N (c_{ui}^l)^2} \times \sqrt{\sum_{i=1}^N (c_{vi}^l)^2}}$$

- Desse modo, analisando a linha \bar{c}_u^l , obtemos os termos com maior correlação a k_u para fazer a expansão da consulta.

Expansão de consultas

- A expansão de consultas é um procedimento importante, pois tende a melhorar a revocação;
- Todavia, se não houver cuidado, a precisão pode cair. São necessários testes para que a expansão fique bem ajustada.