**21.103 Sistemas de Gestão de Bases de Dados**

**Atividade Formativa**

Leia o capítulo 21 - Information Retrieval (6ª edição) do livro adotado e responda às seguintes questões:

1) Considere os seguintes documentos e as "stop words" como "o" e "a":

Doc1: solução para encontrar o óptimo

Doc2: antigo óptimo encontrado

Doc3: antiga abordagem para encontrar a solução óptima

Doc4: antiga experiência não encontra o óptimo

1.a) Represente a matriz dos Termos versus Documentos

1.b) Represente os índices invertidos com a posição dos termos

Resposta:

1.a) Represente a matriz dos Termos versus Documentos

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Doc1 | Doc2 | Doc3 | Doc4 |
| abordagem |  |  | 1 |  |
| antiga |  |  | 1 | 1 |
| antigo |  | 1 |  |  |
| encontra |  |  |  | 1 |
| encontrado |  | 1 |  |  |
| encontrar | 1 |  | 1 |  |
| experiência |  |  |  | 1 |
| não |  |  |  | 1 |
| óptima |  |  | 1 |  |
| óptimo | 1 | 1 |  | 1 |
| para | 1 |  | 1 |  |
| solução | 1 |  | 1 |  |

1.b) Represente os índices invertidos com a posição dos termos

|  |  |  |  |
| --- | --- | --- | --- |
| Termo |  |  |  |
| abordagem | doc3: posição2 |  |  |
| antiga | doc3: posição1 | doc4: posição1 |  |
| antigo | doc2: posição1 |  |  |
| encontra | doc4: posição4 |  |  |
| encontrado | doc2: posição3 |  |  |
| encontrar | doc1: posição3 | doc3: posição4 |  |
| experiência | doc4: posição2 |  |  |
| não | doc4: posição3 |  |  |
| óptima | doc3: posição6 |  |  |
| óptimo | doc1: posição4 | doc2: posição2 | doc4: posição5 |
| para | doc1: posição2 | doc3: posição3 |  |
| solução | doc1: posição1 | doc3: posição5 |  |

2) Quais os documentos devolvidos pela Consulta Booleana (que utiliza os operadores And, Or, Not): ((óptimo OR óptima) NOT experiência)

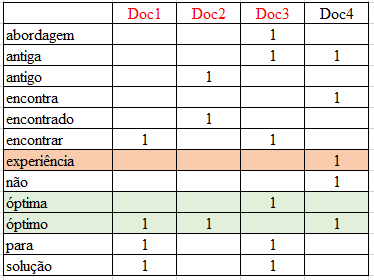
2.a) Na matriz Termos versus Documentos da alínea anterior

2.b) No primeiro motor de busca da Web, o Yahoo

Resposta:

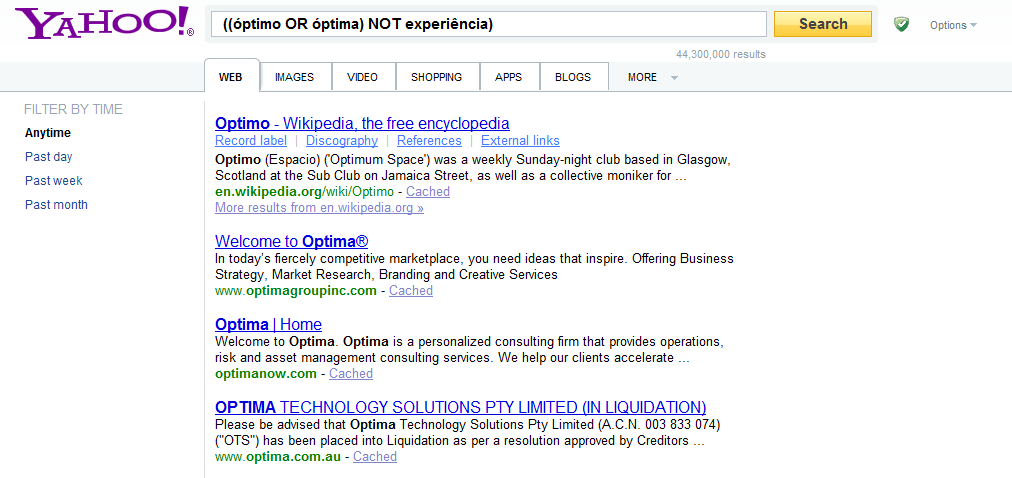
2.a) Na matriz Termos versus Documentos da alínea anterior

- Doc1, Doc2 e Doc3



2.b) No primeiro motor de busca da Web, o Yahoo

-



3) Seja n(d) o número de termos num documento "d" e n(d,t) o número de termos "t" num documento "d",

em que a Frequência de um Termo "t" num documento "d" é dado no manual por:



Seja n(t) o número de documentos que contêm o termo "t" e N o número total de documentos,

onde o *Inverse Document Frequency* (IDF) de Salton & Buckley 1988 é dado por:



Assim, a relevância de um termo "t" num documento "d" é dado por:

TF-IDF(d,t) = TF(d,t).IDF(t)

Considere um documento contendo 100 palavras em que a palavra "informática" aparece três vezes. Suponha ainda que temos 10 milhões de documentos e a palavra "informática" aparece em mil vezes. Qual o valor da relevância da palavra "informática" no documento em causa?

Resposta:

Como n(d,t)=3 e n(d)= 100, a frequência do termo TF = ln (1+3/100)=0,03

Como N=107 e n(t)=103, a frequência inversa de documentos é calculada da seguinte forma, IDF = log10 (107 / 103) = 4

Conclusão: A pontuação conjunta TF-IDF é o produto dessas quantidades: 0,03 x 4 = 0,12

4) O que entende por algoritmo PageRank? Como é que o pode ilustrar?

Resposta:

O algoritmo original de PageRank descrito por Lawrence Page and Sergey Brin em 1995 é dado por:

**PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))**

Onde:

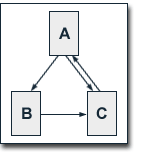
PR(A) é o PageRank da página A,

PR(Ti) é o PageRank das páginas Ti que estão ligadas (apontam) para a página A,

C(Ti) é o número de apontadores (“outbound links”) na página Ti

d é o fator de amortecimento que varia entre 0 e 1.

Exemplo:



Seja d=0.5,

PR(A) = 0.5 + 0.5 (PR(C) /1)

PR(B) = 0.5 + 0.5 (PR(A) /2)

PR(C) = 0.5 + 0.5 (PR(A) /2 + PR(B) /1)

Resolvendo o sistema de 3 equações e 3 incógnitas obtemos os seguintes PR:

PR(A) = 14/13 = 1.07692308

PR(B) = 10/13 = 0.76923077

PR(C) = 15/13 = 1.15384615

Dado que na pática existem milhões de incógnitas, o sistema de equações do PageRank é ser resolvido de forma iterativa, iniciando as variáveis com o valor 1:

|  |  |  |  |
| --- | --- | --- | --- |
| iterações | A | B | C |
| 0 | 1 | 1 | 1 |
| 1 | 1 | 0.75 | 1.25 |
| 2 | 1.125 | 0.75 | 1.125 |
| 3 | 1.0625 | 0.78125 | 1.15625 |
| 4 | 1.078125 | 0.765625 | 1.15625 |
| 5 | 1.078125 | 0.769531 | 1.152344 |
| 6 | 1.076172 | 0.769531 | 1.154297 |
| 7 | 1.077148 | 0.769043 | 1.153809 |
| 8 | 1.076904 | 0.769287 | 1.153809 |
| 9 | 1.076904 | 0.769226 | 1.15387 |
| 10 | 1.076935 | 0.769226 | 1.153839 |
| 11 | 1.07692 | 0.769234 | 1.153847 |
| 12 | 1.076923 | 0.76923 | 1.153847 |
| 13 | 1.076923 | 0.769231 | 1.153846 |
| 14 | 1.076923 | 0.769231 | 1.153846 |
| 15 | 1.076923 | 0.769231 | 1.153846 |
| 16 | 1.076923 | 0.769231 | 1.153846 |
| 17 | 1.076923 | 0.769231 | 1.153846 |
| 18 | 1.076923 | 0.769231 | 1.153846 |
| 19 | 1.076923 | 0.769231 | 1.153846 |
| 20 | 1.076923 | 0.769231 | 1.153846 |
| 21 | 1.076923 | 0.769231 | 1.153846 |
| PageRank | 2 | 3 | 1 |

Fonte: <http://pr.efactory.de/e-pagerank-algorithm.shtml>