21103 - Sistemas de Gestão de Bases de Dados
2014-2015
e-fólio C

Resolução e Critérios de Correção

PARA A RESOLUÇÃO DO E-FÓLIO, ACONSELHA-SE QUE LEIA ATENTAMENTE O SEGUINTE:

1) O e-fólio é constituído por 3 perguntas. A cotação global é de 3 valores.

2) O e-fólio deve ser entregue num único ficheiro PDF, não zipado, com fundo branco, com perguntas numeradas e sem necessidade de rodar o texto para o ler. Penalização de 1 a 3 valores.

3) Não são aceites e-fólios manuscritos, i.e. tem penalização de 100%.

4) O nome do ficheiro deve seguir a normal "eFolioC" + <nº estudante> + <nome estudante com o máximo de 3 palavras>. Penalização de 1 a 3 valores.

5) Na primeira página do e-fólio deve constar o nome completo do estudante bem como o seu número. Penalização de 1 a 3 valores.

6) Durante a realização do e-fólio, os estudantes devem concentrar-se na resolução do seu trabalho individual, não sendo permitida a colocação de perguntas ao professor ou entre colegas.

7) A interpretação das perguntas também faz parte da sua resolução, se encontrar alguma ambiguidade deve indicar claramente como foi resolvida.

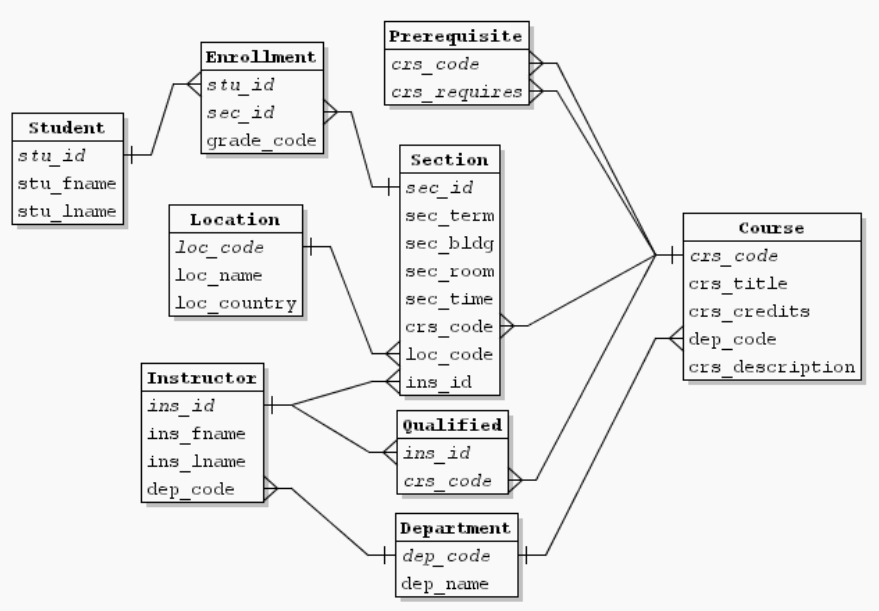
8**)** A legibilidade, a objectividade e a clareza nas respostas serão valorizadas, pelo que, a falta destas qualidades serão penalizadas.

A informação da avaliação do estudante está contida no vetor das cotações:

Questão:	1.a	1.b	2.1	2.2	2.3	3
Cotações:	5	5	3	4	3	10 décimas

1) (1 valor)
Na agregação de dados de uma base de dados transacional para um Data Warehouse existem 2 tipos de armadilhas no SQL ao utilizar junções (SQL traps):
- junções com múltiplos caminhos (“multiple access path problema”, “loop”)
- junções com agregações de dados de 2 tabelas (“connection traps”)

Exemplifique consultas que evidenciem os erros, com dados e resultados, para os seguintes casos:
1.a) junções com múltiplos caminhos



Resposta:

Vamos considerar os seguintes dados:

INSTRUCTOR			
INS_ID	INS_FNAME	INS_LNAME	DEP_CODE
1	JOAO	SANTOS	10
2	LUIS	OLIVEIRA	10
3	JOSÉ	MANUEL	20
4	PAULO	RODRIGUES	30

QUALIFIED	
INS_ID	CRS_CODE
1	1
2	3
3	4
4	2
1	2
2	4

SECTION		
SEC_ID	INS_ID	CRS_CODE
1	4	1
2	3	2
3	2	3
4	1	4
1	2	1
3	1	2

COURSE	
CRS_CODE	CRS_TITLE
1	Física
2	Informática
3	Matemática
4	Medicina

A junção de Instructor com Course, utilizando o caminho da tabela Section,

Instructor |><| Section |><| Course, será:

```
SELECT I.ins_id, I.ins_fname, I.ins_lname, C.crs_title
FROM instructor I, course C, section S
WHERE I.ins_id = S.ins_id
AND  S.crs_code= C.crs_code
ORDER BY I.ins_id, C.crs_title
```

Por outro lado, a junção de Instructor com Course, utilizando o caminho da tabela Qualified,

Instructor |><| Qualifed |><| Course, será:

```
SELECT I.ins_id, I.ins_fname, I.ins_lname, C.crs_title
FROM instructor I, course C, qualified Q
WHERE I.ins_id = Q.ins_id
AND Q.crscode = C.crs_code
ORDER BY I.ins_id, C.crs_title
```

Obtendo resultados diferentes:

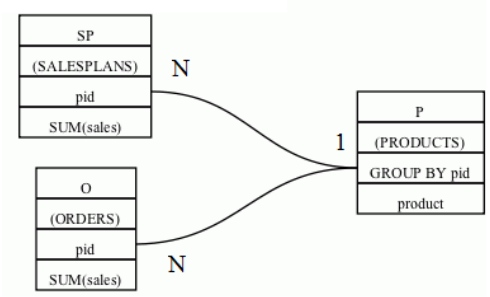
INS_ID	INS_FNAME	INS_LNAME	CRS_TITLE
1	JOAO	SANTOS	Informática
1	JOAO	SANTOS	Medicina
2	LUIS	OLIVEIRA	Física
2	LUIS	OLIVEIRA	Matemática
3	JOSÉ	MANUEL	Informática
4	PAULO	RODRIGUES	Física

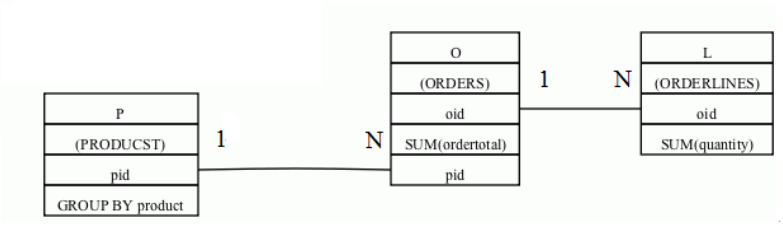
INS_ID	INS_FNAME	INS_LNAME	CRS_TITLE
1	JOAO	SANTOS	Física
1	JOAO	SANTOS	Informática
2	LUIS	OLIVEIRA	Matemática
2	LUIS	OLIVEIRA	Medicina
3	JOSÉ	MANUEL	Medicina
4	PAULO	RODRIGUES	Informática

Assim, confirmamos que junções com caminhos múltiplos, utilizando tabelas diferentes, podem originar resultados diferentes.

Critérios de correção (5 décimas):
- devem ser apresentadas as consultas, os dados e os resultados errados
- penalização de 1 a 2 décimas se faltarem as consultas, os dados ou os resultados

1.b) junções com agregações de dados de 2 tabelas





There are aggregates in more than one table: o.sales, sp.sales

There are aggregates in more than one table: o.ordertotal, l.quantity

Resposta:
Vamos considerar os seguintes dados:

PRODUCTS	
PID	PRODUCT
1	PRODUTO A
2	PRODUTO B
3	PRODUTO C

ORDERS	
PID	SALES
1	500
1	5.000
2	7.500
2	45
3	10.000
3	25

SALESPLANS	
PID	SALES
1	250
1	1.500
2	2.500
2	450
3	3.500
3	1.000

Ao juntar as três tabelas:

```
SELECT P.product, SUM(O.sales), SUM(SP.sales)
FROM products P, orders O, salesplans SP
WHERE P.pid = O.pid
AND P.pid = SP.pid
GROUP BY P.product
```

resultam os seguintes valores:

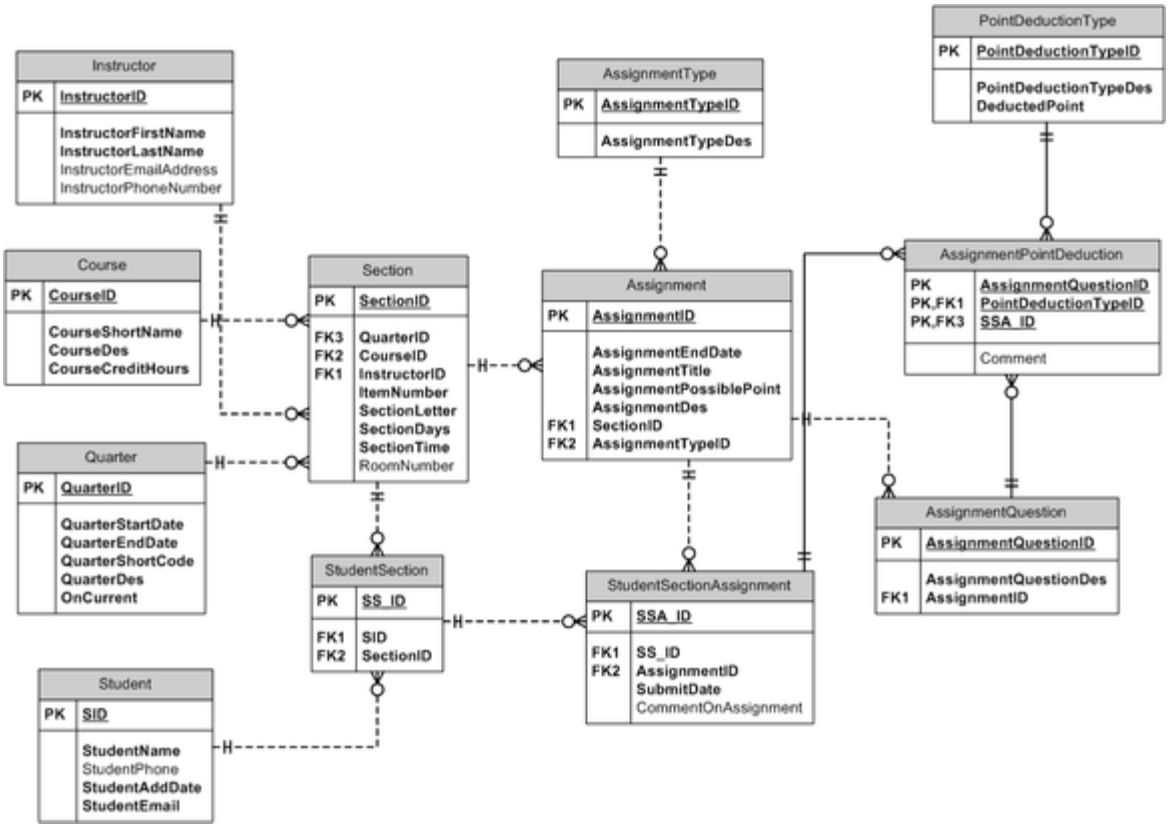
PRODUCT	SUM(O.SALES)	SUM(SP.SALES)
PRODUTO A	11.000	3.500
PRODUTO B	15.090	5.900
PRODUTO C	20.050	9.000

quando, na realidade se pretendiam os seguintes valores:

PRODUCT	SUM(O.SALES)	SUM(SP.SALES)
PRODUTO A	5.500	1.750
PRODUTO B	7.545	2.950
PRODUTO C	10.025	4.500

Critérios de correção (5 décimas):
- devem ser apresentadas as consultas, os dados e os resultados errados
- penalização de 1 a 2 décimas se faltarem as consultas, os dados ou os resultados

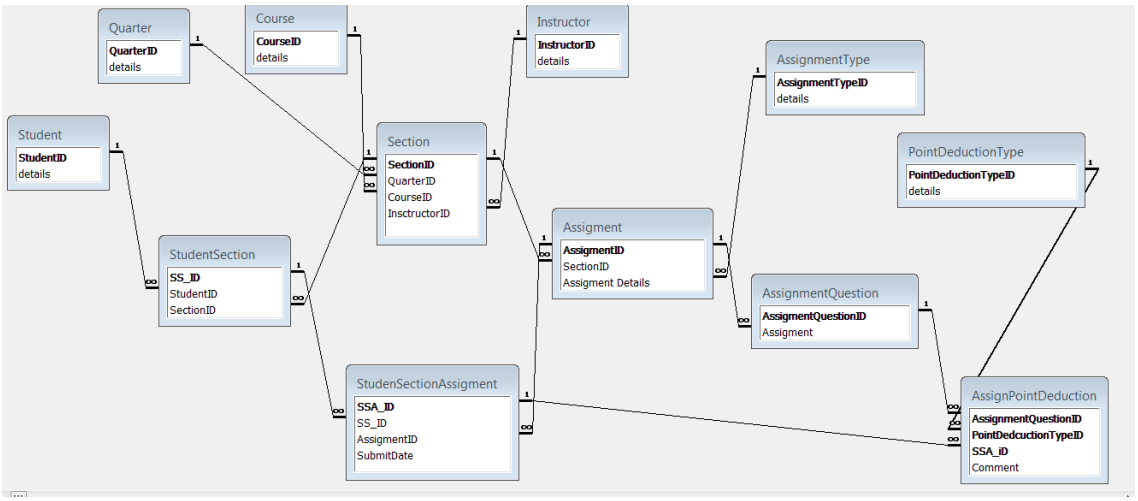
2) (1 valor) *Data Warehousing*
Considere o seguinte esquema relativo a uma escola:



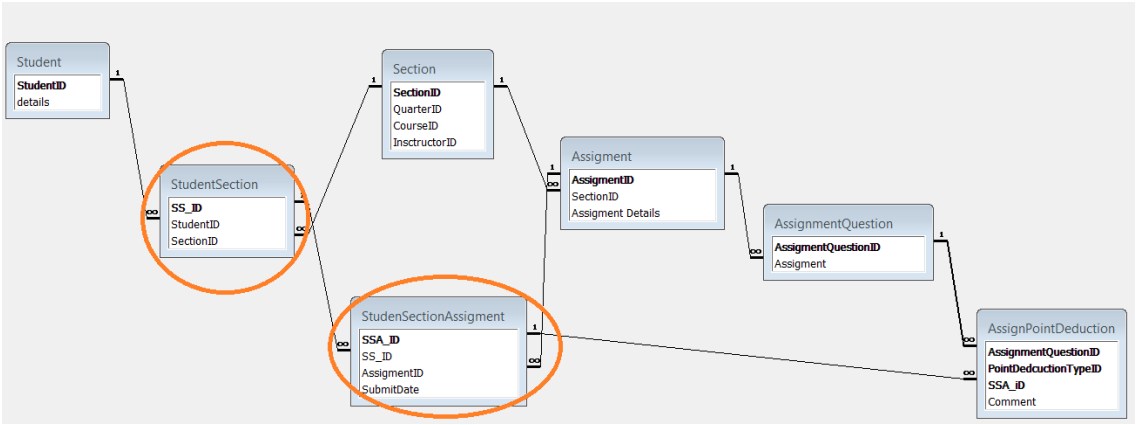
2.1- Desenhe uma base de dados transacional equivalente, na 3ª forma normal. De seguida remova a eventual transitividade que exista no esquema base de dados. Faça o carregamento de dados. Na representação gráfica das ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo.

Resposta:

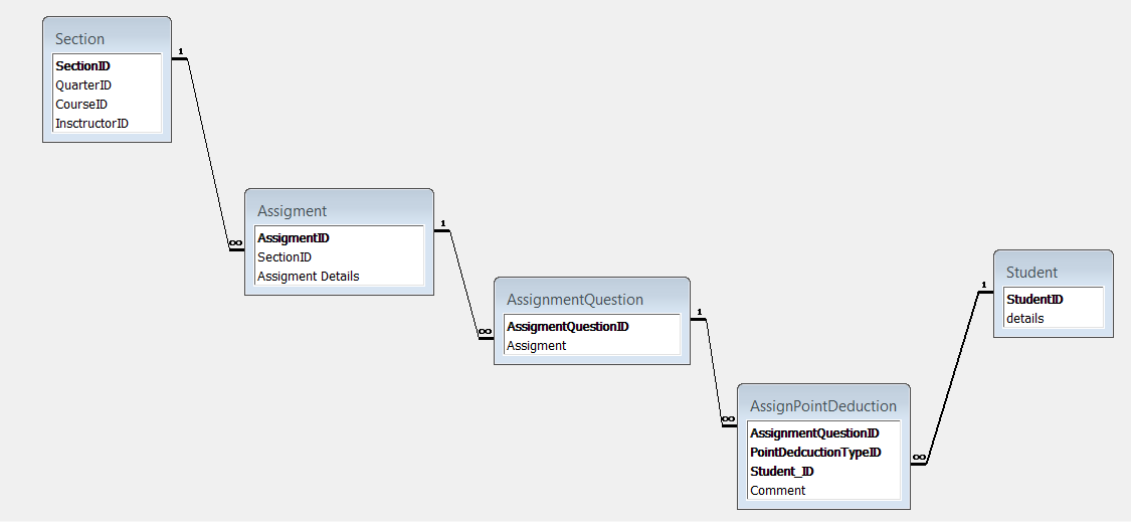
As tabelas originais com a representação gráfica das ligações de 1:N, a tabela com uma única linha é desenhada em cima e a tabela com várias linhas é desenhada por baixo.



Retirando parte das tabelas auxiliares teremos o seguinte. Visto que existem caminhos múltiplos, vamos tentar retirá-los. Podem ser retiradas relações ou tabelas. Neste caso visto que as tabelas StudentSection e StudentSectionAssignment são redundantes, irão ser retiradas.



Obtendo assim um esquema sem caminhos múltiplos.

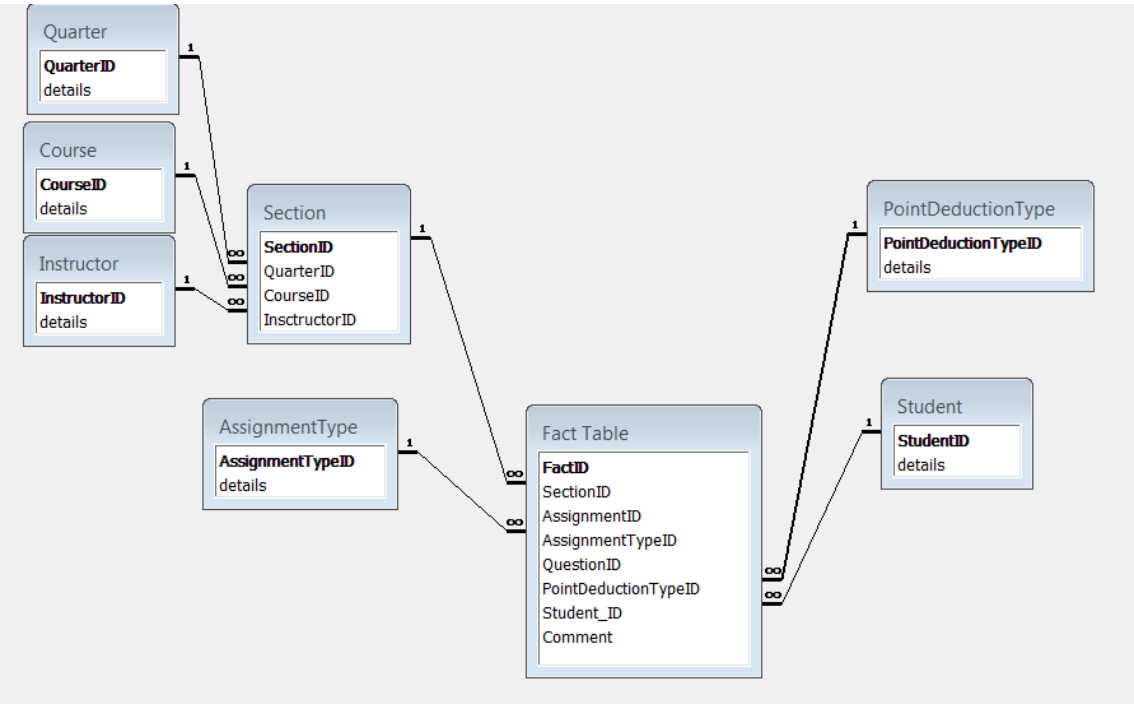


Critérios de correção (3 décimas):
- devem ser apresentada a base de dados sem caminhos múltiplos: 3 décimas
- penalização de 1 a 2 décimas para caminhos múltiplos

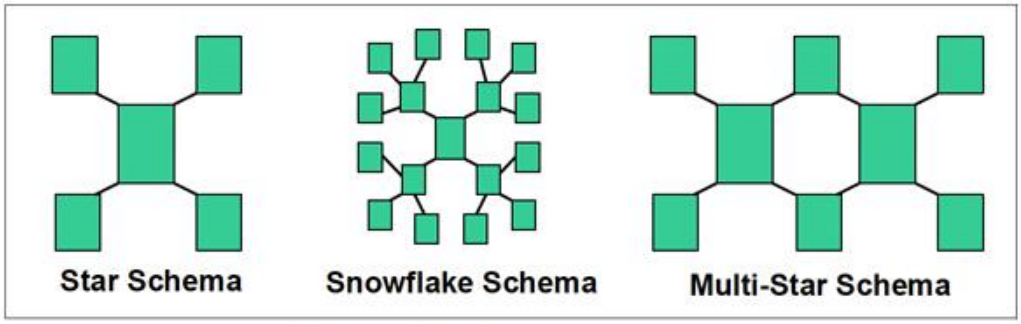
2.2- Pretendemos desenhar um “Data Warehouse” relacional em estrela ou em constelação, i.e. com duas ou mais estrelas. Defina a(s) tabela(s) de factos e mostre a tabela depois da desnormalização dos dados. Defina as dimensões com os níveis de agregação para o “Data Warehouse” relacional. Apresente a(s) tabela(s) de factos associada às dimensões. Ao juntar as tabelas transacionais tenha em consideração as eventuais armadilhas referidas na pergunta anterior.

Resposta:

A tabela de factos resulta da junção de e tabelas: Assignment, AssignmentQuestion e AssignmentPointDeduction.



Os *Data Warehouses* (DW) com uma única tabela de factos são conhecidas por DW com esquema em estrela; com dimensões que têm mais de uma tabela são conhecidas por DW com esquema em floco-de-neve (snowflake); e quando têm mais do que uma tabela de factos são conhecidos como DW com esquema em constelações (ou multi-estrela).



Types of dimensional models

No caso em estudo obtivemos uma esquema em floco-de-neve (snowflake).

Critérios de correção (4 décimas):
- devem ser apresentada o Data Warehouse com a tabela de factos
- penalização de 1 a 2 décimas para erros ou omissões

2.3- Crie duas perguntas e traduza para SQL utilizando pelo menos duas dimensões.

a) Contar os factos utilizando as dimensões Section e AssignmentType

	SectionID	Total de AssignmentID	A	B
	1	4	2	2
▶	2	2		2

```
TRANSFORM Count([Fact Table].AssignmentID) AS ContarDeAssignmentID
SELECT [Fact Table].SectionID, Count([Fact Table].AssignmentID) AS [Total de AssignmentID]
FROM [Fact Table]
GROUP BY [Fact Table].SectionID
PIVOT [Fact Table].AssignmentTypeID
```

b) Contar os factos utilizando as dimensões Student e AssignmentID

	Student_ID	Total de FactID	1	2	3	4
	1	1	1			
	2	1		1		
	3	1			1	
▶	4	1				1
	5	2	1	1		

```
TRANSFORM Count([Fact Table].FactID) AS ContarDeFactID
SELECT [Fact Table].Student_ID, Count([Fact Table].FactID) AS [Total de FactID]
FROM [Fact Table]
GROUP BY [Fact Table].Student_ID
PIVOT [Fact Table].AssignmentID
```

Critérios de correção (3 décimas):
- devem ser apresentados os resultados e as consultas SQL utilizador o Pivot

3) (1 valor) *Information Retrieval*
Escreva um texto, com pelo menos 500 palavras, onde descreva o que entende por *PageRank* e a sua importância nos SEO (Search Engine Optimization).

Resposta:

Vamos introduzir o conceito de motor de pesquisa, o algoritmo *PageRank* e a sua aplicação no SEO.

i) Motores de Pesquisa

Um motor de pesquisa organizam a sua informação em grandes matrizes com palavras e documentos (ou páginas).

Palavras\Documentos	1	2	3	4	5	6	7	8	9	10	11	12	13	14
universidade	x	x	x	x			x				x	x		
aberta	x		x	x			x							
educação	x				x	x		x		x		x		x
tecnologias	x					x				x				x
curso		x	x					x	x		x			x
informática		x	x					x	x		x			x
gestão					x			x	x			x		x
cultura					x		x		x			x	x	x
matemática				x				x	x	x		x	x	x

Os motores de pesquisa têm três subsistemas:
a) Subsistemas de indexação (Documento/Página) que insere novas colunas na matriz com base nas palavras do documento;
b) Subsistema Matriz (Documento, Palavra);
c) Subsistemas de Pesquisa (Palavra) que para cada palavra, ou linha, verifica os documentos que estão associados;

A pesquisa por palavra utiliza a álgebra booleana, selecionando os documentos mais relevantes.

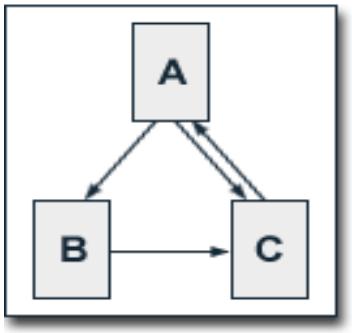
ii) PageRank

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 em 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

Dada a dimensão da Web, existem métodos iterativos que permitem calcular o *PageRank* sem recorrer aos sistemas de equações.

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

O *PageRank* é o algoritmo que permite calcular o "valor" de uma página na Web. O valor da página não depende apenas da quantidade de *links* apontados para ela, mas do "valor" das páginas que apontam para ela. No exemplo PR(C) depende de PR(A) e PR(B), PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B) /1).

iii) SEO (Search Engine Optimization)

O Marketing na Web fez crescer uma nova disciplina o SEO, que tem como objetivo melhorar o desempenho de uma dada página nos motores de pesquisa, para uma ou mais palavras.

Tal como referimos nos motores de pesquisa existem palavras e páginas, existem também duas formas de otimização em SEO:
1) otimização "on-page" está relacionada com a escolha das palavras ou "keywords"
2) otimização "off-page" está relacionada com o *PageRank*

Na otimização "on-page" distinguem-se ainda dois métodos:
1.1) métodos "white hat" ou métodos com ética
1.2) métodos " black hat" ou métodos sem ética

Dos métodos "black hat" podemos referir:
- repetir a utilização de uma palavra para aumentar a sua relevância na página;
- utilizar texto invisível ao utilizador mas captadas pelo motor de busca (utilizar palavras escondidas em letras da mesma cor do fundo ou formatar o tamanho da letra para zero);
- uso de redireccionamentos não autorizados ou de camuflagem do verdadeiro conteúdo da página;

A otimização "off-page" passa criar links externos que façam referência ao website, como por exemplo:
- escrever num blog sobre o website
- referir o website nas redes sociais
- submeter o website em motores de pesquisa (Google)
- submeter o website em diretórios (Yahoo)

Em conclusão: a importância do PageRank no SEO é evidente na otimização "off-page" do website.

Critérios de Correção (10 décimas):
- explicação e exemplifique o algoritmo PageRank (5 décimas)
- explicação SEO e a importância do PageRank (5 décimas)
- penalização de 1 a 2 décimas para erros ou omissões