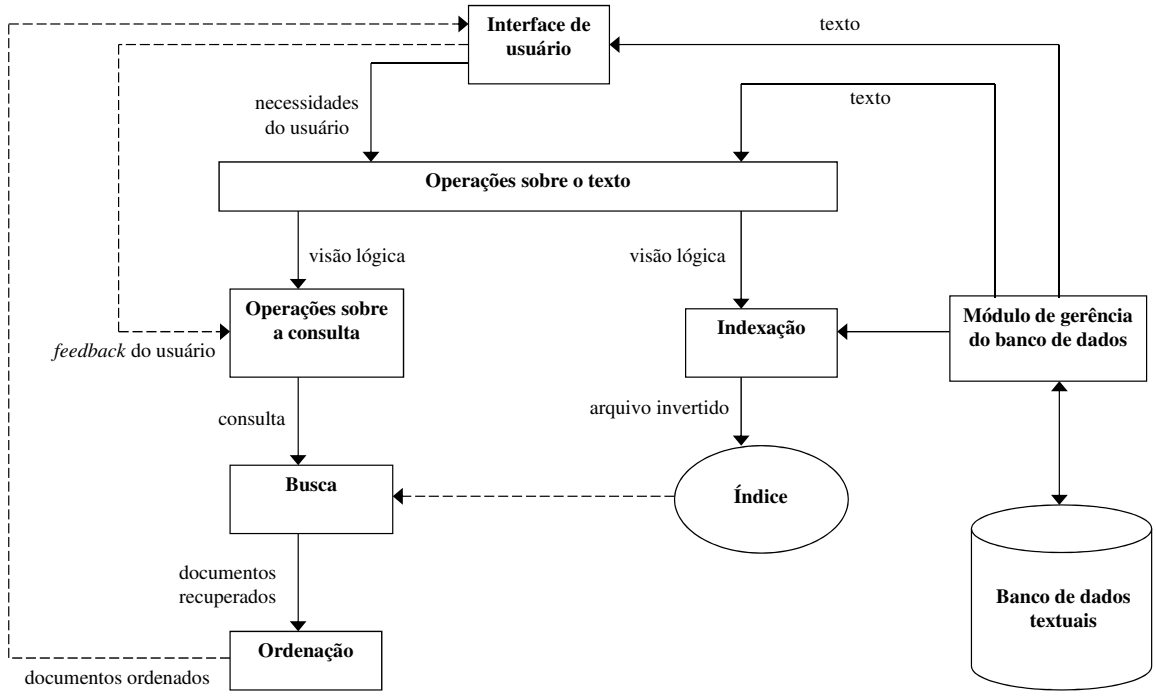
Modelos de Linguagem para Recuperação de Informação

Recuperação de Informação na *Web*

Prof. Guilherme Tavares de Assis

Universidade Federal de Ouro Preto – UFOP
Instituto de Ciências Exatas e Biológicas – ICEB
Departamento de Computação – DECOM

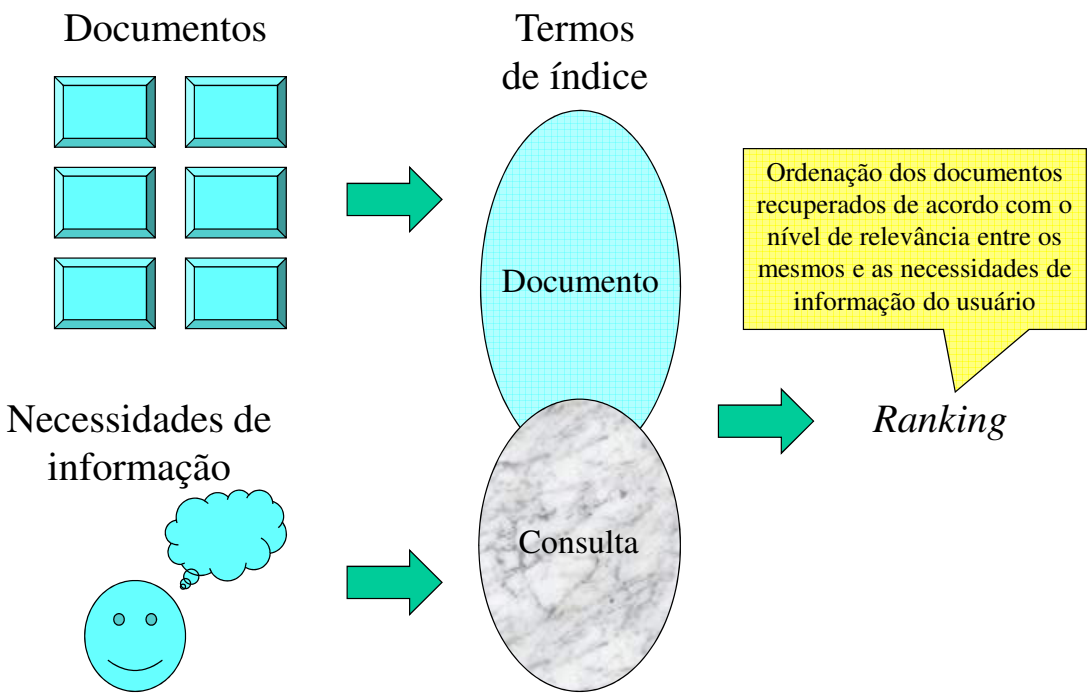
Processo de Recuperação



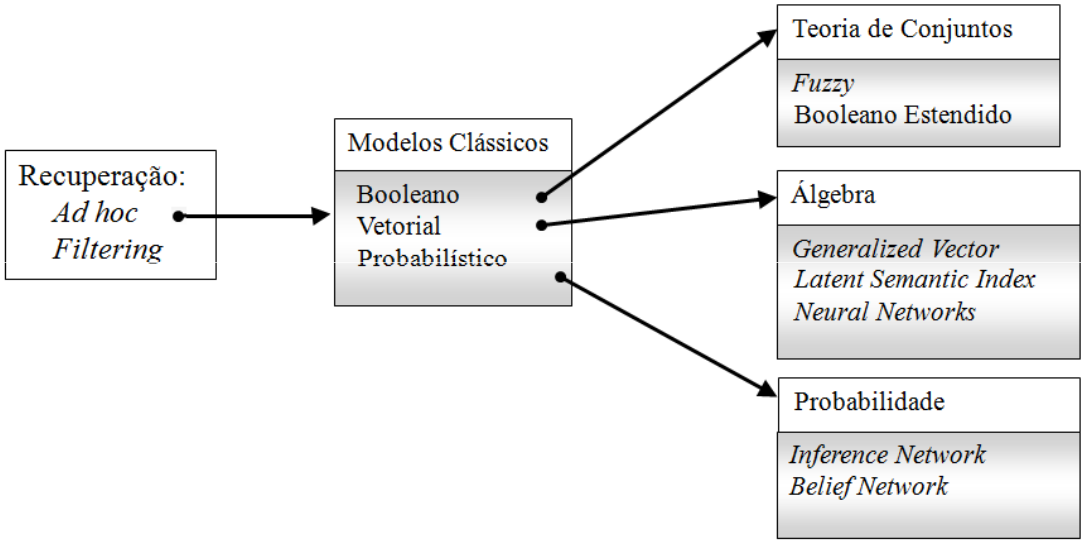
Modelagem

- Sistemas de RI geralmente adotam termos de índice para processar consultas.
  - Já que usuários não são treinados em "elaboração de consultas", mediante uma necessidade de informação, o resultado pode não ser satisfatório.
  - Visando escalabilidade, geralmente, um arquivo invertido é confeccionado para os termos de índice de uma coleção.
- A determinação da relevância entre uma consulta e os documentos de uma determinada coleção é uma questão crítica em sistemas de RI.
  - Os modelos de RI tentam determinar tal relevância.

Modelagem

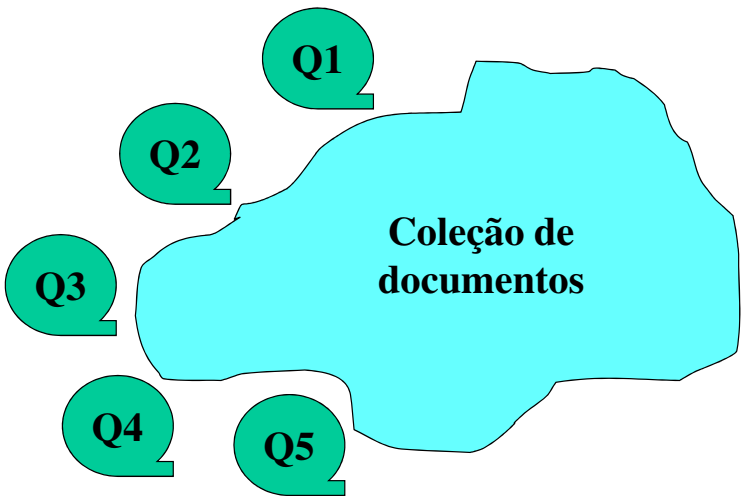


Modelagem



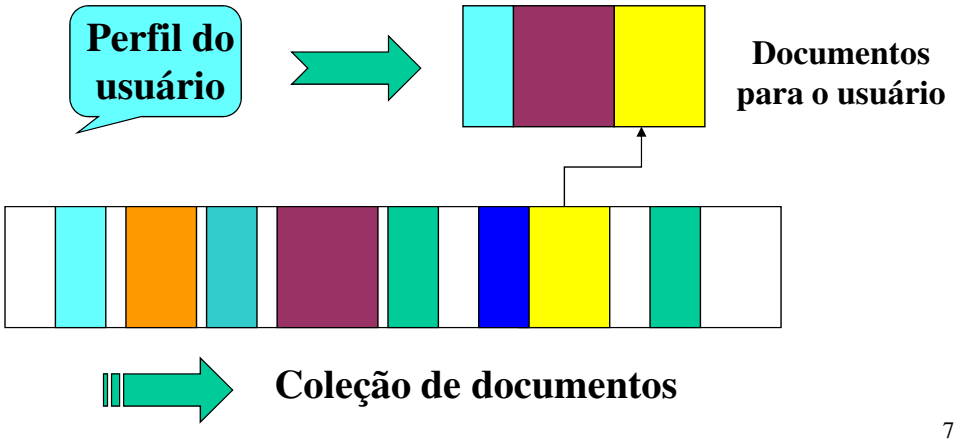
Modelagem

- Recuperação: modo operacional *ad hoc*.
  - Os documentos na coleção permanecem intactos enquanto consultas são submetidas por meio de um sistema de RI.



Modelagem

- Recuperação: modo operacional *filtering*.
  - Consultas permanecem relativamente intactas enquanto a coleção é alterada (entrada e saída de documentos).
  - Um "perfil do usuário" é criado descrevendo as necessidades do usuário.
  - O perfil é comparado com os documentos da coleção, na tentativa de encontrar aqueles que são de interesse para o usuário.



Modelagem

- Cada documento é representado por um conjunto de termos de índice ou palavras-chave representativas.
  - Termos de índice podem ser apenas substantivos já que possuem significado próprio.
  - Em uma representação *full text*, máquinas de busca assumem que todas as palavras são termos de índice.
  - Nem todos os termos são igualmente úteis para representar o conteúdo de um documento, já que os termos apresentam frequências distintas no documento.
- Em uma coleção, termos menos frequentes permitem identificar um conjunto mais restrito de documentos.
  - A importância dos termos de índice em uma coleção pode ser representada por pesos associados a eles (termos ponderados).

Modelagem

- Definições:
  - t é o número total de termos de índice da coleção.
  - k_i é o i-ésimo termo de índice da coleção.
  - $K = (k_1, k_2, ..., k_t)$ é o vetor de termos de índice da coleção.
  - d_j é o j-ésimo documento da coleção.
  - $w_{i,j}$ é o peso associado ao par (k_i, d_j).
    - O peso $w_{i,j}$ quantifica a importância do termo k_i na descrição do conteúdo do documento d_j, sendo sempre ≥ 0.
    - $w_{i,j} = 0$ indica que o termo k_i não pertence ao documento d_j.
  - $\text{vec}(d_j) = (w_{1,j}, w_{2,j}, ..., w_{t,j})$ é o vetor de pesos dos termos de índice associado ao documento d_j.

Modelo Booleano

- O Modelo Booleano é um modelo simples e fácil de implementar, baseado na teoria de conjuntos.
  - Os termos da consulta estão presentes ou não em um documento, sem distinção de importância; logo, $w_{i,j} \in \{0,1\}$.
  - A recuperação é baseada em decisão binária; logo, um documento é ou não relevante à consulta.
  - Não há *ranking*.
  - O modelo Booleano, frequentemente, retorna poucos ou muitos documentos em resposta à consulta do usuário.
- A necessidade de informação deve ser traduzida em uma expressão booleana.
  - Geralmente, as consultas booleanas formuladas são simples.

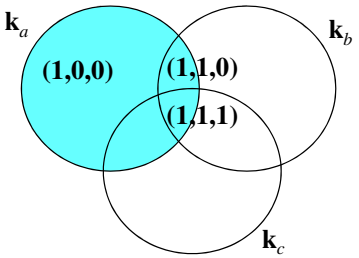
Modelo Booleano

- Exemplo de consulta:

$$q = k_a \wedge (k_b \vee \neg k_c)$$

  - Forma normal disjuntiva da consulta (FND), envolvendo três componentes conjuntivos:

$$(1,1,1) \vee (1,1,0) \vee (1,0,0)$$



- Um documento d_j é relevante à consulta q se os pesos dos termos do documento (vec(d_j)) forem iguais a de algum componente conjuntivo da consulta.

Modelo Booleano - Exemplo

- Coleção composta por 20 documentos relativos a curiosidades de Copas do Mundo de Futebol.

Doc.	Texto do documento
d_1	Em 1994, o Brasil sagrou-se campeão, porém o artilheiro da competição foi o búlgaro Hristo Stoichkov, com 6 gols.
d_2	O primeiro gol brasileiro em Copas do Mundo foi marcado por Preguinho, atacante do Fluminense, em 1930, no Uruguai.
d_3	Gols do Brasil na Copa de 1994: Brasil 2 x 0 Rússia (Romário, Raí); Brasil 3 x 0 Camarões (Romário, Márcio Santos, Bebeto); Brasil 1 x 1 Suécia (Romário); Brasil 1 x 0 EUA (Bebeto); Brasil 3 x 2 Holanda (Romário, Bebeto, Branco); Brasil 1 x 0 Suécia (Romário). Artilheiro da competição: Hristo Stoichkov (6 gols).
d_4	A Copa do Mundo surgiu com a intenção de ampliar, em termos mundiais, a chamada *Cup* britânica, instituída pela *The Football Association* em 1872.
d_5	O goleiro mexicano Antonio Carbajal foi o jogador que participou do maior número de Copas (1950, 1954, 1958, 1962 e 1966).
d_6	Leônidas da Silva e Ademir de Menezes, em 1938 e 1950, respectivamente, foram os únicos brasileiros que conseguiram se tornar o artilheiro de uma Copa do Mundo.
d_7	Na Copa de 1994, o artilheiro Romário, com sua genialidade e seus gols, contrabalançou o pobre futebol demonstrado pelo Brasil e pelos adversários.
d_8	A seleção da Alemanha foi a grande campeã da Copa de 1990, quando venceu a Argentina na final, com um gol de pênalti de Brehme, aos 40 minutos do 2º tempo.

Modelo Booleano - Exemplo

d_9	O número de países participantes da Copa do Mundo, passou de 13 (em 1930) para 24 (em 1994).
d_{10}	Com Passarela, Kempes e Fillol, a Argentina venceu a Holanda por 3 x 1 na final do mundial de 1978.
d_{11}	O maior artilheiro da história das Copas foi o francês Just Fontaine, que em 1958 marcou 13 gols.
d_{12}	Na final da Copa de 1974, o Carrossel Holandês, como era conhecida a seleção da Holanda, foi anulado pela anfitriã Alemanha Ocidental, que venceu por 2 x 1 e ficou com o título.
d_{13}	Em 1950, a seleção brasileira perdeu a chance de conquistar seu primeiro título jogando em casa, perdendo a final de forma inesperada para o Uruguai.
d_{14}	Desde seu início, a Copa de 1954, disputada na Suíça, parecia destinada àquele fantástico time da Hungria. Da estréia, massacrando a Coréia do Sul por 9 x 0, até a final, contra a Alemanha, seu ataque não deixou barato.
d_{15}	O artilheiro da seleção brasileira na Copa do Mundo de 1994 foi o jogador Romário, que marcou 5 gols.
d_{16}	A Copa da Suíça mantém até hoje a maior média de gols em mundiais. Foram 140 tentos em 26 jogos (média de 5,28 gols por jogo).
d_{17}	2000 jornalistas cobriram a Copa de 1958. Destes, 200 (10%) eram da Alemanha, a então campeã do mundo.
d_{18}	Eusébio, jogador de Portugal, foi o artilheiro da Copa do Mundo de 1966.
d_{19}	Menor média de público da história das Copas: 1938 (20.829 pessoas). Maior média de público da história das Copas: 1994 (68.991 pessoas).
d_{20}	Classificação final da Copa do Mundo de 1986: Argentina (1º), Alemanha (2º), França (3º) e Bélgica (4º).

UFOP – RIW – Prof. Guilherme Tavares de Assis

Modelo Booleano - Exemplo

- Necessidade de informação: "Que jogador foi o artilheiro do Brasil na Copa de 1994? Quantos gols ele marcou?".
  - Documentos relevantes definidos por um especialista da área: d_{15}, d_3, d_7 (nesta ordem).
  - Possível consulta: artilheiro $\wedge$ brasil $\wedge$ 1994 $\wedge$ gols.

- Como os termos de índice da consulta são ligados, simplesmente, pelo conectivo $\wedge$ (*and*), tem-se que:
  - FND = (1,1,1,1);
  - pelo modelo Booleano, são recuperados, como relevantes, os documentos d_1, d_3 e d_7;
  - O documento d_{15} (mais relevante pelos especialistas) foi ignorado e o documento irrelevante d_1 foi recuperado.

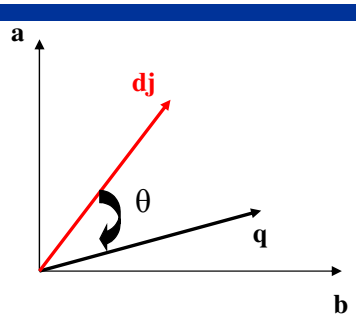
Modelo Vetorial

- O Modelo Vetorial permite o uso de pesos não binários, associados aos termos de índice, proporcionando combinação parcial entre a consulta e os documentos da coleção.
  - Tais pesos permitem o cálculo do grau de similaridade entre a consulta e um determinado documento.
  - Um documento é retornado se há combinação parcial entre os termos de índice do documento e da consulta.
  - Os documentos recuperados são ordenados de acordo com o grau de similaridade calculado, em ordem decrescente, permitindo um resultado mais preciso em relação ao Modelo Booleano.

Modelo Vetorial

- Definições:
  - Documentos (d_j) e consultas (q) são representados como vetores de pesos dos termos de índice.
  - $w_{i,j}$ é o peso associado ao par (k_i, d_j): $w_{i,j} > 0$ se $k_i \in d_j$.
  - $w_{i,q}$ é o peso associado ao par (k_i, q): $w_{i,q} > 0$ se $k_i \in q$.
  - $\text{vec}(d_j) = (w_{1,j}, w_{2,j}, ..., w_{t,j})$.
  - $\text{vec}(q) = (w_{1,q}, w_{2,q}, ..., w_{t,q})$.
  - Cada termo k_i está associado a um vetor unitário $\text{vec}(i)$.
    - Os t vetores unitários $\text{vec}(i)$ formam uma base ortogonal (ou seja, os termos de índice ocorrem nos documentos de forma independente) para o espaço t-dimensional.

Modelo Vetorial



- O grau de similaridade entre o documento d_j e a consulta q é dado pela correlação entre os vetores associados. Tal correlação pode ser quantificada, por exemplo, pelo coseno do ângulo θ entre tais vetores.

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|d_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2_{i,j}} \times \sqrt{\sum_{j=1}^{t} w^2_{j,q}}}$$

onde $|\vec{d_j}|$ e $|\vec{q}|$ são as normas dos vetores do documento e da consulta, respectivamente. A norma $|\vec{q}|$ não afeta o *ranking* porque é a mesma para todos os documentos. Já a norma $|\vec{d_j}|$ proporciona uma normalização no espaço dos documentos.

Modelo Vetorial

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{|d_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2_{i,j}} \times \sqrt{\sum_{j=1}^{t} w^2_{j,q}}}$$

- Para os documentos recuperados ($w_{i,j} > 0$ e $w_{i,q} > 0$ para algum termo k_i), tem- se que:

$$0 \leq \text{sim}\,(d_j, q) \leq 1.$$

- Problema: como calcular os pesos $w_{i,j}$ e $w_{i,q}$?

Modelo Vetorial

- Para calcular $w_{i,j}$, é utilizada uma estratégia de ponderação de peso, chamada de esquema *tf-idf*, que se baseia nos princípios básicos relativos às técnicas de agrupamento. Deve-se então quantificar:
  - a similaridade intra-agrupamento (*term frequency* - *tf*): frequência com que um termo incide no documento, determinando se o mesmo descreve bem ou não o conteúdo do documento;
  - a não-similaridade inter-agrupamento (*inverse document frequency* - *idf*): frequência inversa do termo nos documentos da coleção, determinando se o mesmo é considerado útil ou não, dentro da coleção, para descrever a relevância de um documento.

- Logo, balanceando os dois fatores, tem-se que:

$$w_{i,j} = f_{i,j} \times idf_i$$

Modelo Vetorial

- Como calcular, para cada termo k_i presente em um documento d_j, a frequência do termo *tf* ($f_{i,j}$)?

- Como calcular, para cada termo k_i presente na coleção, a frequência inversa do termo *idf* (idf_i)?

- Definições:
  - n é o número total de documentos na coleção.
  - n_i é o número de documentos que contêm o termo k_i.
  - $freq_{i,j}$ é o número de vezes que o termo k_i aparece no texto do documento d_j.

Modelo Vetorial

- O fator normalizado $f_{i,j}$, referente ao termo k_i presente no documento d_j, é dado por:

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}}$$

onde $max_l freq_{l,j}$ é a frequência máxima dentre as frequências de todos os termos k_i no documento d_j.

- O fator idf_i, referente ao termo k_i, é dado por:

$$\log \frac{N}{n_i} = idf_i$$

onde o *log* torna os valores de *tf* e *idf* comparáveis.

Modelo Vetorial

- Para os termos de índice presentes na consulta, tem-se que:

$$w_{i,q} = \left(0{,}5 + \frac{0{,}5\, freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log \frac{N}{n_i}$$

onde:

  - $freq_{i,q}$ é o número de vezes que o termo k_i é mencionado no texto da consulta q;
  - $max_l freq_{l,q}$ é a frequência máxima dentre as frequências de todos os termos k_i na consulta q.

- Como geralmente, em uma determinada consulta, os termos de índice não se repetem, tem-se que:

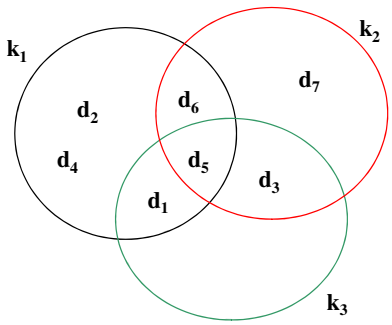
$$w_{i,q} = idf_i$$

Modelo Vetorial

- Vantagens:
  - É um modelo simples, rápido de computar e tão eficaz quanto qualquer outro modelo de *ranking* existente.
  - A fórmula de *ranking*, baseada no coseno, ordena os documentos recuperados de acordo com o grau de similaridade à consulta, permitindo a recuperação de documentos que não possuem todos os termos da consulta.
  - O esquema *tf-idf*, para ponderação dos termos, é eficiente e melhora a qualidade do conjunto resposta.
- Desvantagem:
  - Assume independência entre os termos de índice.

Modelo Vetorial

- Exemplo simples:



	k_1	k_2	k_3	$q \bullet d_j$
q	1	2	3	-----
d_1	2	0	1	**5**
d_2	1	0	0	**1**
d_3	0	1	3	**11**
d_4	2	0	0	**2**
d_5	1	2	4	**17**
d_6	1	2	0	**5**
d_7	0	5	0	**10**

Modelo Vetorial - Exemplo

- Necessidade de informação: "Que jogador foi o artilheiro do Brasil na Copa de 1994? Quantos gols ele marcou?".
  - Documentos relevantes definidos por um especialista da área: d_{15}, d_3, d_7 (nesta ordem).
  - Possível consulta: artilheiro $\wedge$ brasil $\wedge$ 1994 $\wedge$ gols.

Modelo Vetorial - Exemplo

- Como todos os termos de índice da consulta aparecem uma única vez, tem-se que $w_{i,q} = idf_i$.
- Logo:

Termo de índice	n_i	$idf_i = \log \frac{N}{n_i}$
artilheiro	6	0,523
brasil	3	0,824
1994	6	0,523
gols	6	0,523

Modelo Vetorial - Exemplo

Pesos $w_{i,j}$ referentes aos termos da consulta

Documentos da coleção em que pelo menos um termo da consulta aparece. Para os demais, tem-se $w_{i,j}$=0 para todos os termos da consulta.

Doc.	$w_{artilheiro,j}$	$w_{brasil,j}$	$w_{1994,j}$	$w_{gols,j}$
d_1	0,523	0,824	0,523	0,523
d_3	0,075	0,824	0,075	0,149
d_6	0,523	0	0	0
d_7	0,523	0,824	0,523	0,523
d_9	0	0	0,523	0
d_{11}	0,523	0	0	0,523
d_{15}	0,523	0	0,523	0,523
d_{16}	0	0	0	0,523
d_{18}	0,523	0	0	0
d_{19}	0	0	0,523	0

Modelo Vetorial - Exemplo

Retorno final do Modelo Vetorial

Doc.	$\sum_{i=1}^{t} w_{i,j} \times w_{i,q}$	$\sqrt{\sum_{i=1}^{t} w^2_{i,j}}$	$\sqrt{\sum_{j=1}^{t} w^2_{j,q}}$	$sim(d_j, q)$
d_1	1,500	1,225	1,225	1
d_7	1,500	1,225	1,225	1
d_3	0,780	0,884	1,225	0,754
d_{15}	0,821	0,906	1,225	0,740
d_{11}	0,547	0,740	1,225	0,603
d_6	0,274	0,523	1,225	0,428
d_9	0,274	0,523	1,225	0,428
d_{16}	0,274	0,523	1,225	0,428
d_{18}	0,274	0,523	1,225	0,428
d_{19}	0,274	0,523	1,225	0,428

Modelo Vetorial - Exemplo

- Pelos resultados obtidos, observa-se que:
  - entre os quatro documentos recuperados com o maior grau de similaridade, estão os três considerados relevantes;
  - como no Modelo Booleano, o documento d_1 continua no topo da classificação, mesmo sendo irrelevante.
- Analisando os documentos da coleção, verifica-se que a consulta poderia ter sido melhor formulada para se obter o resultado esperado.
  - Substituindo o termo "brasil" por "seleção brasileira", provavelmente o documento d_{15} apareceria no topo da classificação por apresentar tal termo e o documento d_1 cairia na classificação por não ter tal termo.

Modelo Probabilístico

- O Modelo Probabilístico fornece um arcabouço probabilístico para se resolver problemas de recuperação de informação, possibilitando estimar a probabilidade de um documento ser relevante a uma consulta do usuário.
  - Dada uma consulta, o modelo assume que existe um conjunto ideal de respostas que contém exatamente os documentos relevantes.

Modelo Probabilístico

- O processo de consulta pode ser visto como um processo de especificação das propriedades do conjunto ideal de respostas.
- Problema: como definir tais propriedades?
  - Já que as propriedades não são conhecidas, deve-se definir um conjunto inicial de respostas como sendo o ideal (suposição);
  - Iniciam-se interações com o usuário no intuito de melhorar a descrição probabilística do conjunto ideal de respostas.
- Logo, o modelo é descrito como uma série de interações com o usuário, no intuito de refinar o conjunto ideal de respostas.

Modelo Probabilístico

- Dada uma consulta q e um documento d_j, o modelo estima a probabilidade que o usuário ache tal documento relevante.
  - O modelo assume que tal probabilidade de relevância depende apenas das representações da consulta e do documento.
  - No modelo, são associados pesos binários aos termos de índice dos documentos e consultas.
- O conjunto ideal de respostas R deve maximizar a probabilidade de relevância. Documentos em R são considerados relevantes.
- Para computar a chance de um documento d_j ser relevante à consulta q, a similaridade entre os mesmos é dada por:

$$\text{sim}(d_j,q) = P(d_j \textit{ relevante-a } q) \,/\, P(d_j \textit{ não-relevante-a } q)$$

Modelo Probabilístico

- Aplicando as regras de Bayes e assumindo independência entre os termos de índice, tem-se que:

$$sim(d_j, q) = \sum_{i=1}^{t} w_{i,q} \times w_{i,j} \times \left(\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})} \right)$$

onde:

- $w_{i,j}$ é o peso do termo k_i no documento d_j;
- $w_{i,q}$ é o peso do termo k_i na consulta q;
- R é conjunto ideal de respostas (documentos relevantes);
- $\overline{R}$ é o complemento de R (documentos não relevantes);
- $P(k_i \mid R)$ é a probabilidade do termo k_i estar presente em um documento aleatoriamente selecionado de R;
- $P(k_i \mid \overline{R})$ é a probabilidade do termo k_i estar presente em um documento aleatoriamente selecionado de $\overline{R}$.

Modelo Probabilístico

- Considerações:
  - Se um termo k_i está presente no documento d_j, o peso $w_{i,j}$ é igual a 1; caso contrário, o peso $w_{i,j}$ é igual a 0.
  - Se um termo k_i é um termo de índice da consulta q, o peso $w_{i,q}$ é igual a 1; caso contrário, o peso $w_{i,q}$ é igual a 0.
  - Inicialmente, $P(k_i \mid R)$ é igual a 0.5 para qualquer termo de índice k_i.
  - Inicialmente, $P(k_i \mid \overline{R})$ é igual a (n_i/N) para qualquer termo de índice k_i, onde n_i é o número de documentos que contém k_i e N representa o número total de documentos da coleção.

Modelo Probabilístico

- Aplicando-se a fórmula de similaridade, obtém-se o conjunto inicial de documentos recuperados.
  - Assim, é possível definir um subconjunto V contendo os r documentos de maior similaridade, que é utilizado para refinar as seguintes fórmulas de probabilidade para as próximas interações:

$$P(k_i \mid R) = \frac{V_i + \frac{n_i}{N}}{V + 1} \qquad P(k_i \mid \overline{R}) = \frac{n_i - V_i + \frac{n_i}{V}}{N - V + 1}$$

Onde:

- V é o número de documentos do próprio subconjunto;
- V_i é o número de documentos de V que contém o termo k_i.

Modelo Probabilístico

- Vantagem:
  - Classifica os documentos, em ordem decrescente, pela probabilidade de relevância à consulta.
- Desvantagens:
  - Necessita definir as probabilidades iniciais e, a cada processo interativo, um conjunto ideal de respostas.
  - Não considera a frequência dos termos de índice nos documentos da coleção.
  - Assume independência entre os termos de índice.

Modelo Probabilístico - Exemplo

- Necessidade de informação: "Que jogador foi o artilheiro do Brasil na Copa de 1994? Quantos gols ele marcou?".
  - Possível consulta: artilheiro ∧ brasil ∧ 1994 ∧ gols.

- Probabilidades iniciais e cálculos relacionados, segundo o Modelo Probabilístico, para cada termo da consulta:

Termo de índice	n_i	$P(k_i \mid R)$	$P(k_i \mid \overline{R}) = \frac{n_i}{N}$	$\log \frac{P(k_i \mid R)}{1-P(k_i \mid R)} + \log \frac{1-P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})}$
artilheiro	6	0,500	0,300	0,368
brasil	3	0,500	0,150	0,753
1994	6	0,500	0,300	0,368
gols	6	0,500	0,300	0,368

Modelo Probabilístico - Exemplo

Resultados obtidos na 1ª interação do Modelo

Doc.	**Termos de índice**	$sim(d_j, q)$
d_1	artilheiro, brasil, 1994, gols	1,857
d_3	artilheiro, brasil, 1994, gols	1,857
d_7	artilheiro, brasil, 1994, gols	1,857
d_{15}	artilheiro, 1994, gols	1,104
d_{11}	artilheiro, gols	0,736
d_6	artilheiro	0,368
d_9	1994	0,368
d_{16}	gols	0,368
d_{18}	artilheiro	0,368
d_{19}	1994	0,368

Modelo Probabilístico - Exemplo

- Probabilidades e cálculos relacionados, na 2ª interação do Modelo, para cada termo da consulta, considerando que o subconjunto V é composto pelos 5 primeiros documentos do conjunto inicial recuperado:

Termo de índice	n_i	V	V_i	$P(k_i \mid R)$	$P(k_i \mid \overline{R})$	$\log\frac{P(k_i \mid R)}{1-P(k_i \mid R)} + \log\frac{1-P(k_i \mid \overline{R})}{P(k_i \mid \overline{R})}$
artilheiro	6	5	5	0,883	0,138	1,674
brasil	3	5	3	0,525	0,038	1,446
1994	6	5	4	0,717	0,200	1,006
gols	6	5	5	0,883	0,138	1,674

- Existem várias formas de se definir os documentos do subconjunto V. Uma delas é considerar a metade superior dos documentos do *ranking* gerado na interação inferior do Modelo Probabilístico.

Modelo Probabilístico - Exemplo

Resultados obtidos na 2ª interação do Modelo

Doc.	Termos de índice	$sim(d_j, q)$
d_1	artilheiro, brasil, 1994, gols	5,800
d_3	artilheiro, brasil, 1994, gols	5,800
d_7	artilheiro, brasil, 1994, gols	5,800
d_{15}	artilheiro, 1994, gols	4,354
d_{11}	artilheiro, gols	3,348
d_6	artilheiro	1,674
d_{16}	gols	1,674
d_{18}	artilheiro	1,674
d_9	1994	1,006
d_{19}	1994	1,006

Modelo Probabilístico - Exemplo

- Pelos resultados obtidos, observa-se que:
  - da 1ª para a 2ª interação, só houve diferença no *ranking* dos 5 últimos documentos;
  - em uma possível 3ª interação, como o subconjunto *V* seria o mesmo da 2ª, não haveria diferença no *ranking* gerado (ou seja, o resultado da aplicação do Modelo estabilizou-se na 2ª interação);
  - o Modelo Probabilístico teve um comportamento semelhante ao Modelo Vetorial.

- A ideia do Modelo Probabilístico é fazer com que o *ranking* dos documentos melhore, a cada interação, a partir dos documentos do conjunto *V*, até um ponto de estabilização.

- Da mesma forma que no Modelo Vetorial, verifica-se que a consulta poderia ter sido melhor formulada para se obter um resultado melhor.

Modelo Booleano Estendido

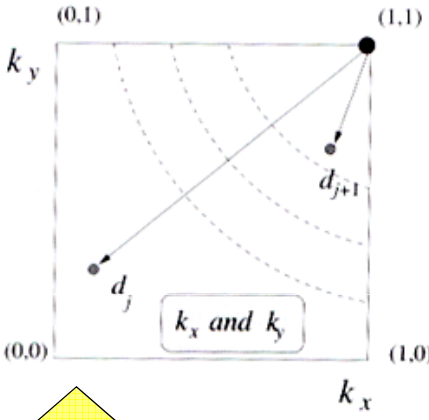
- O Modelo Booleano estendido surgiu em 1983, com o propósito de suprir a deficiência do Modelo Booleano clássico de não gerar um *ranking* dos documentos recuperados, retornando geralmente um número grande ou pequeno de documentos.

- A ideia é combinar as características dos Modelos Clássicos Booleano (expressões booleanas) e Vetorial (vetores de pesos).
  - Por exemplo, para um determinado documento d_j formado pelos termos k_x e k_y, o modelo assume pesos $w_{x,j}$ e $w_{y,j}$ que podem ser calculados da mesma forma que no Modelo Vetorial; ou seja

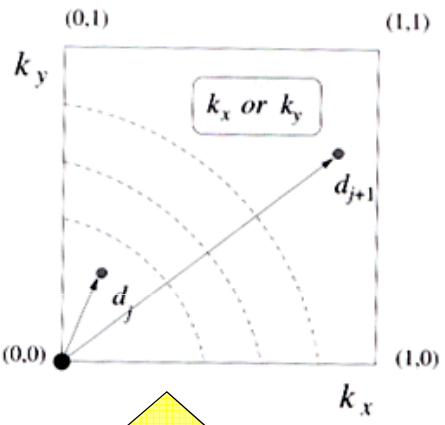
$$w_{i,j} = f_{i,j} \times idf_i \text{, onde:}$$

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \qquad \log \frac{N}{n_i} = idf_i$$

Modelo Booleano Estendido



Para a consulta conjuntiva "$q_{and} = k_x \wedge k_y$", já que o ponto (1,1) é o mais desejável, **sim**(q_{and}, d_j) pode ser calculada pelo complemento da distância entre o documento d_j e o ponto (1,1).



Para a consulta disjuntiva "$q_{or} = k_x \vee k_y$", já que o ponto (0,0) deve ser evitado, **sim**(q_{or}, d_j) pode ser calculada pela distância de d_j a partir do ponto (0,0).

Modelo Booleano Estendido

- Representando os peso $w_{x,j}$ e $w_{y,j}$ como sendo x e y, as similaridades de um documento d em relação às consultas q_{or} e q_{and} são dadas por:

$$sim(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}} \qquad sim(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

- Se todos os pesos são binários ($w_{i,j} \in \{0,1\}$), um documento encontra-se em um dos cantos: (0,0), (0,1), (1,0), (1,1).
  - Os valores de sim(q_{or}, d_j) podem ser $0,\ 1/\sqrt{2}$ e 1
  - Os valores de sim(q_{and}, d_j) podem ser $0,\ 1 - 1/\sqrt{2}$ e 1

  Contudo, pesos não binários são adotados.

Modelo Booleano Estendido

- Dado que o nº de termos de índice é t, o Modelo considera distâncias euclidianas em um espaço com t dimensões.

- Contudo, uma generalização mais compreensiva é adotar a teoria das normas de vetores.
  - O modelo *p-norm* generaliza a noção de distância para incluir não apenas distâncias euclidianas, mas também distâncias p, onde $1 \leq p \leq \infty$.
  - As formas generalizadas para consultas q_{or} e q_{and} tornam-se:

$$q_{or} = k_1 \vee^p k_2 \vee^p \ldots \vee^p k_m$$

$$q_{and} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m$$

Modelo Booleano Estendido

- Assim, as similaridades de um documento d_j em relação às consultas q_{or} e q_{and} são dadas por:

$$sim(q_{or}, d_j) = \left(\frac{x_1^p + x_2^p + \ldots + x_m^p}{m} \right)^{\frac{1}{p}}$$

$$sim(q_{and}, d_j) = 1 - \left(\frac{(1 - x_1)^p + (1 - x_2)^p + \ldots + (1 - x_m)^p}{m} \right)^{\frac{1}{p}}$$

  Onde x_i é o peso $w_{i,j}$ associado ao par $[k_i, d_j]$ e $1 \leq p \leq \infty$.

- Variando o valor do parâmetro p, pode-se variar o comportamento da classificação: característica poderosa do Modelo Booleano Estendido.

Modelo Booleano Estendido

- Para as consultas que envolvem tanto componentes disjuntivos quanto conjuntivos, os operadores são agrupados pela ordem de precedência.
- Para $q = (k_1 \wedge^p k_2) \vee^p k_3$, a similaridade é dada por:

$$sim(q, d_j) = \left(\frac{\left(1 - \left(\frac{(1-x_1)^p + (1-x_2)^p}{2}\right)^{\frac{1}{p}}\right)^p + x_3^p}{2} \right)^{\frac{1}{p}}$$

Modelo Booleano Estendido

- O modelo é bem poderoso e relaxa a álgebra booleana, interpretando os operadores como distâncias algébricas.
- Um ponto positivo é a possibilidade de variar a distância p entre 1 e infinito.
  - Para $p = 1$, $\text{sim}(q_{or},d_j) = \text{sim}(q_{and},d_j) = (\sum x_i)/m$ ($\approx$ vetorial).
  - Para $p = \infty$, $\text{sim}(q_{or},d_j) = \max(x_i)$ e $\text{sim}(q_{and},d_j) = \min(x_i)$ ($\approx$ Fuzzy).
- No entanto, o modelo não é muito utilizado, já que a computação é um pouco complexa.

Modelo Booleano Estendido - Exemplo

- Consulta conjuntiva: artilheiro ∧ brasil ∧ 1994 ∧ gols.

Os pesos $w_{i,j}$ são calculados da mesma forma que no Modelo Vetorial.

Doc.	$w_{artilheiro,j}$	$w_{brasil,j}$	$w_{1994,j}$	$w_{gols,j}$
d_1	0,523	0,824	0,523	0,523
d_3	0,075	0,824	0,075	0,149
d_6	0,523	0	0	0
d_7	0,523	0,824	0,523	0,523
d_9	0	0	0,523	0
d_{11}	0,523	0	0	0,523
d_{15}	0,523	0	0,523	0,523
d_{16}	0	0	0	0,523
d_{18}	0,523	0	0	0
d_{19}	0	0	0,523	0

Modelo Booleano Estendido - Exemplo

Retorno final do Modelo para $p = 3$.

O *ranking* é parecido com o do Vetorial, sendo diferente na troca dos docs d_3 e d_{15} (o mais relevante).

Doc.	$sim(d_j, q) = 1 - \left(\frac{(1 - x_{ARTILHEIRO})^3 + (1 - x_{BRASIL})^3 + (1 - x_{1994})^3 + (1 - x_{GOLS})^3}{4} \right)$
d_1	0,564
d_7	0,564
d_{15}	0,308
d_3	0,181
d_{11}	0,178
d_6	0,081
d_9	0,081
d_{16}	0,081
d_{18}	0,081
d_{19}	0,081

Modelo *Generalized Vector*

- Todos os modelos apresentados assumem independência entre os termos de índice, ou seja, cada par de termos k_i e k_j é ortogonal ($\vec{k}_i \bullet \vec{k}_j = 0$) .
- Já que a independência entre os termos pode restringir os modelos, em 1985, foi proposto o modelo *Generalized Vector*, assumindo que os vetores dos termos de índice são independentes, mas os pares de vetores não são necessariamente ortogonais.
- Considerando $w_{i,j}$ binário, é possível representar todas as possibilidades de ocorrência mútua dos termos nos documentos da coleção, por meio de um conjunto de 2^t mintermos, onde t é o número total de termos de índice.

Modelo *Generalized Vector*

Mintermo (t = 3)	**Descrição da ocorrência mútua**
$m_1 = (0,0,0)$	Indica os documentos que não contêm algum dos termos
$m_2 = (1,0,0)$	Indica os documentos que contêm apenas o termo k_1
$m_3 = (0,1,0)$	Indica os documentos que contêm apenas o termo k_2
$m_4 = (0,0,1)$	Indica os documentos que contêm apenas o termo k_3
$m_5 = (1,1,0)$	Indica os documentos que contêm os termos k_1 e k_2
$m_6 = (1,0,1)$	Indica os documentos que contêm os termos k_1 e k_3
$m_7 = (0,1,1)$	Indica os documentos que contêm os termos k_2 e k_3
$m_8 = (1,1,1)$	Indica os documentos que contêm os termos k_1, k_2 e k_3

- Considere $\boldsymbol{g_i(m_r)}$ uma função que retorna o peso binário do termo de índice k_i no mintermo m_r. Logo, por exemplo, $g_i(m_1) = 0$, para todo k_i.

Modelo *Generalized Vector*

- A ideia é criar vetores ortogonais $\vec{m}_i$ associados aos mintermos.
  - Para todo i ≠ j, $\vec{m}_i \bullet \vec{m}_j = 0$. No entanto, os pares de vetores $\vec{m}_i$ ortogonais não implicam na independência entre os termos de índice, pois estes estão correlacionados pelos mintermos.

Vetor $\vec{m}_i$	Mintermo associado
$\vec{m}_1 = (1, 0, 0, 0, 0, 0, 0, 0)$	m_1
$\vec{m}_2 = (0, 1, 0, 0, 0, 0, 0, 0)$	m_2
$\vec{m}_3 = (0, 0, 1, 0, 0, 0, 0, 0)$	m_3
$\vec{m}_4 = (0, 0, 0, 1, 0, 0, 0, 0)$	m_4
$\vec{m}_5 = (0, 0, 0, 0, 1, 0, 0, 0)$	m_5
$\vec{m}_6 = (0, 0, 0, 0, 0, 1, 0, 0)$	m_6
$\vec{m}_7 = (0, 0, 0, 0, 0, 0, 1, 0)$	m_7
$\vec{m}_8 = (0, 0, 0, 0, 0, 0, 0, 1)$	m_8

Modelo *Generalized Vector*

- O Modelo adota a ideia de que co-ocorrência de termos de índice em documentos da coleção induz dependência entre tais termos.

- A dependência entre termos de índice melhora a eficácia em um processo de recuperação?
  - Ainda há dúvidas.
  - Não é claro, por exemplo, que a estrutura do Modelo *Generalized Vector* proporciona uma vantagem em situações práticas, já que é mais complexo e mais custoso computacionalmente que o Modelo Vetorial.

Modelo *Generalized Vector*

- Deve-se determinar, para cada termo k_i, o vetor de termo $\vec{k}_i$ associado.
  - Deve-se, então, calcular a soma normalizada dos vetores dos mintermos m_r onde o termo k_i possui o valor 1:

$$\vec{k}_i = \frac{\sum_{\forall r,\ g_i(m_r)=1} c_{i,r}\vec{m}_r}{\sqrt{\sum_{\forall r,\ g_i(m_r)=1} c^2{}_{i,r}}} \qquad c_{i,r} = \sum_{d_j \mid g_1(\vec{d}_j) = g_l(m_r)\ para\ todo\ l} w_{i,j}$$

onde:

- r é o número do mintermo m_r ,que varia de 1 a 2^t;
- $g_i(m_r)$ é a função que retorna o peso binário do termo k_i no mintermo m_r;
- $\vec{m}_r$ é o vetor associado ao mintermo m_r;
- $c_{i,r}$ é o fator de correlação definido entre o vetor $\vec{m}_r$ e o termo de índice k_i; tal fator é calculado a partir dos documentos d_j, cuja ocorrência dos termos de índice coincide com o mintermo m_r;
- $w_{i,j}$ é o peso associado ao par (k_i, d_j), calculado como no Modelo Vetorial.

Modelo *Generalized Vector*

- Os vetores $\vec{k}_i$ servem de base para o cálculo dos vetores $\vec{d}_j$ e $\vec{q}$, que são utilizados para se obter a similaridade de um documento d_j em relação à consulta q, fornecida pela medida de coseno utilizada pelo Modelo Vetorial.

$$\vec{d}_j = \sum_{\forall i} w_{i,j}\vec{k}_i \qquad \vec{q} = \sum_{\forall i} w_{i,q}\vec{k}_i$$

$$sim(d_j, q) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times \left|\vec{q}\right|}$$

Modelo *Generalized Vector* - Exemplo

- Consulta conjuntiva: artilheiro ∧ brasil ∧ 1994 ∧ gols.

Relação de mintermos e vetores $\vec{m}_i$ associados para a consulta q.

Mintermo	Vetor $\vec{m}_i$ associado
$m_1 = (0, 0, 0, 0)$	$\vec{m}_1 = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$m_2 = (1, 0, 0, 0)$	$\vec{m}_2 = (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$m_3 = (0, 1, 0, 0)$	$\vec{m}_3 = (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
. . .	. . .
$m_{15} = (0, 1, 1, 1)$	$\vec{m}_{15} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0)$
$m_{16} = (1, 1, 1, 1)$	$\vec{m}_{16} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)$

Modelo *Generalized Vector* - Exemplo

Fatores de correlação $c_{i,r}$ para os termos de índice e mintermos (apenas aqueles que combinam com algum documento da coleção), considerando os pesos $w_{i,j}$ calculados no Modelo Vetorial.

Minitermo	$c_{artilheiro,r}$	$c_{brasil,r}$	$c_{1994,r}$	$c_{gols,r}$
m_2	1,046	0,0	0,0	0,0
m_4	0,0	0,0	1,046	0,0
m_5	0,0	0,0	0,0	0,523
m_8	0,523	0,0	0,0	0,523
m_{14}	0,523	0,0	0,523	0,523
m_{16}	1,121	2,472	1,121	1,195

Modelo *Generalized Vector* - Exemplo

Vetor $\vec{k}_i$ vinculado a cada termo k_i, uma vez calculados os fatores de correlação $c_{i,r}$.

Termo de índice	$\sqrt{\sum_{\forall r,\, g_i(m_r)=1} c^2_{i,r}}$	Vetor $\vec{k}_i$
artilheiro	1,702	(0, 0.615, 0, 0, 0, 0, 0, 0.307, 0, 0, 0, 0, 0, 0.307, 0, 0.659)
brasil	2,472	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)
1994	1,620	(0, 0, 0, 0.646, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.323, 0, 0.692)
gols	2,249	(0, 0, 0, 0, 0.233, 0, 0, 0.233, 0, 0, 0, 0, 0, 0.233, 0, 0.531)

Modelo *Generalized Vector* - Exemplo

- Como $w_{i,q} = idf_i$ para todos os termos da consulta, tem-se que:

$\vec{q}$ = (0, 0.322, 0, 0.338, 0.122, 0, 0, 0.283, 0, 0, 0, 0, 0, 0.452, 0, 1.809)

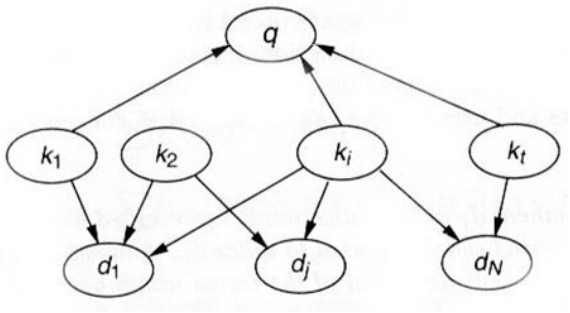
Doc.	Vetor $\vec{d}_j$	sim(d_j,q)
d_1	$\vec{d}_1$ = (0, 0.332, 0, 0.338, 0.122, 0, 0, 0.283, 0, 0, 0, 0, 0, 0.452, 0, 1.809)	1
d_7	$\vec{d}_7$ = (0, 0.332, 0, 0.338, 0.122, 0, 0, 0.283, 0, 0, 0, 0, 0, 0.452, 0, 1.809)	1
d_{15}	$\vec{d}_{15}$ = (0, 0.332, 0, 0.338, 0.122, 0, 0, 0.283, 0, 0, 0, 0, 0, 0.452, 0, 0.985)	0,968
d_3	$\vec{d}_3$ = (0, 0.046, 0, 0.048, 0.035, 0, 0, 0.058, 0, 0, 0, 0, 0, 0.082, 0, 1.004)	0,967
d_{11}	$\vec{d}_{11}$ = (0, 0.332, 0, 0, 0.122, 0, 0, 0.283, 0, 0, 0, 0, 0, 0.283, 0, 0.623)	0,914
d_{16}	$\vec{d}_{16}$ = (0, 0, 0, 0, 0.122, 0, 0, 0.122, 0, 0, 0, 0, 0, 0.122, 0, 0.278)	0,893
d_6	$\vec{d}_6$ = (0, 0.332, 0, 0, 0, 0, 0, 0.161, 0, 0, 0, 0, 0, 0.161, 0, 0.345)	0,829
d_{18}	$\vec{d}_{18}$ = (0, 0.332, 0, 0, 0, 0, 0, 0.161, 0, 0, 0, 0, 0, 0.161, 0, 0.345)	0,829
d_9	$\vec{d}_9$ = (0, 0, 0, 0.338, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.169, 0, 0.362)	0,828
d_{19}	$\vec{d}_{19}$ = (0, 0, 0, 0.338, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.169, 0, 0.362)	0,828

Modelo *Generalized Vector* - Exemplo

- Observa-se que:
  - o resultado é bem parecido com o do Modelo Vetorial;
  - como no Modelo Booleano Estendido, o documento d_{15} aparece melhor classificado em relação ao Modelo Vetorial;
  - o documento d_{16} melhorou no *ranking* em relação aos demais documentos que possuem apenas um termo de índice (isso se deve ao fato de que o termo *gols*, único presente no documento, possui um maior fator de correlação para o mintermo m_{16}, em comparação com os termos *artilheiro* e *1994*).

Modelo *Belief Network*

- O funcionamento do Modelo é descrito como a associação de variáveis aleatórias (termos de índice) relativas aos documentos e à consulta.
  - Assim, tantos os documentos quanto a consulta são modelados de tal forma a gerar a topologia de uma rede de crenças.



  - A classificação do documento d_j relativa à consulta q é interpretada como o quanto a consulta q cobre o documento d_j.

Modelo *Belief Network*

- Aplicando as regras de *Bayes* e instanciando as variáveis aleatórias dos termos de índice, o que os torna mutuamente independentes, a probabilidade do documento d_j ser relevante para a consulta q é estabelecida por:

$$P(d_j \mid q) \approx \sum_{\forall \vec{k}} P(d_j \mid \vec{k}) \times P(q \mid \vec{k}) \times P(\vec{k})$$

Modelo *Belief Network*

- A definição $P(d_j \mid q)$ serve para conceituar a rede de crenças.
- Para que o Modelo se torne aplicável, deve-se definir uma estratégia de classificação que, associada à rede de crenças, permita a recuperação ordenada dos documentos.
  - Pelo Modelo Vetorial, são estabelecidas as probabilidades:

$$P(d_j \mid \vec{k}) = \begin{cases} \dfrac{w_{i,j}}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,j}}} \; se \; \vec{k} = \vec{k}_i \wedge g_i(\vec{d}_j) = 1 \\ 0 \; caso \; contrário \end{cases}$$

$$P(\vec{k}) = \left(\frac{1}{2}\right)^t$$

$$P(q \mid \vec{k}) = \begin{cases} \dfrac{w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,q}}} \; se \; \vec{k} = \vec{k}_i \wedge g_i(q) = 1 \\ 0 \; caso \; contrário \end{cases}$$

Modelo *Belief Network* - Exemplo

- Consulta conjuntiva: artilheiro ∧ brasil ∧ 1994 ∧ gols.

Retorno do Modelo, considerando os pesos $w_{i,j}$ e $w_{i,q}$ calculados no Vetorial.

Doc.	$sim(d_j, q) = P(d_j \mid q)$
d_1	0,427*0,427*0,063 + 0,673*0,673*0,063 + 0,427*0,427*0,063 + 0,427*0,427*0,063 = **0,063**
d_7	0,427*0,427*0,063 + 0,673*0,673*0,063 + 0,427*0,427*0,063 + 0,427*0,427*0,063 = **0,063**
d_3	0,085*0,427*0,063 + 0,932*0,673*0,063 + 0,085*0,427*0,063 + 0,169*0,427*0,063 = **0,050**
d_{15}	0,577*0,427*0,063 + 0,577*0,427*0,063 + 0,577*0,427*0,063 = **0,047**
d_{11}	0,707*0,427*0,063 + 0,707*0,427*0,063 = **0,038**
d_6	1,000*0,427*0,063 = **0,027**
d_9	1,000*0,427*0,063 = **0,027**
d_{16}	1,000*0,427*0,063 = **0,027**
d_{18}	1,000*0,427*0,063 = **0,027**
d_{19}	1,000*0,427*0,063 = **0,027**

Modelo *Belief Network* - Exemplo

- Percebe-se, pelo resultado obtido, que a ordem dos documentos recuperados coincide com o *ranking* gerado pelo Modelo Vetorial.
  - Entre os quatro documentos recuperados com o maior grau de similaridade, estão os três considerados relevantes.
  - Não há, no caso, nenhuma conclusão adicional, além das observações já descritas na ilustração do Modelo Vetorial.

Modelo *Belief Network*

- Vantagens:
  - Modelo rápido de computar.
  - Possibilidade de adoção de outras estratégias para gerar o *ranking* dos documentos.
- Desvantagem:
  - Assume independência entre os termos de índice.