



Recuperação de Informação

Edleno Silva de Moura





Modelagem em RI

Modelos em Recuperação de Informação

- Núcleo de qualquer sistema de recuperação de informação.
- Utilizados para representar características semânticas dos elementos envolvidos nos sistemas.
- Modelos clássicos: booleano, vetorial e probabilístico
- Modelos bastante utilizados: vetorial, language models (que são probabilísticos) e BM25 (que é probabilístico)

Modelo Vetorial

- Proposto em 1968 e continua sendo muito empregado hoje em dia.
- Proposto originalmente para resolver problemas de busca.
- Sucesso reside na eficiência e nos bons resultados obtidos.
- Todos os componentes do sistema são vistos como conjuntos de palavras.

Modelo Vetorial

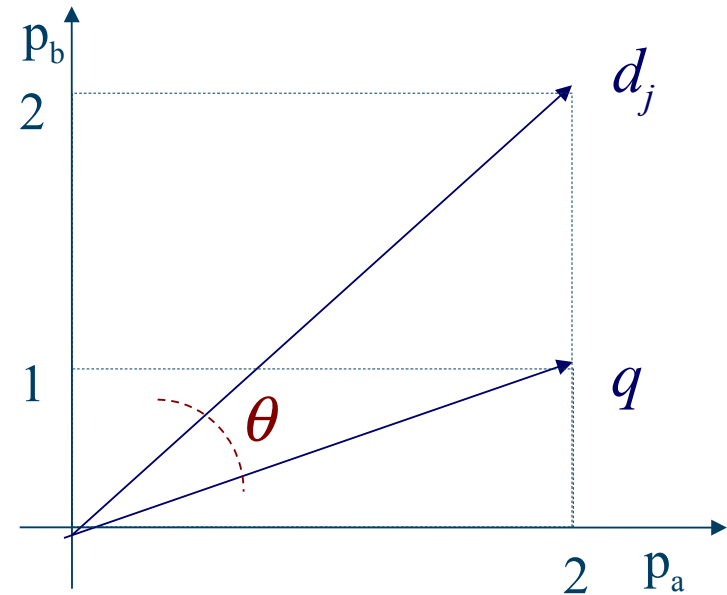
- Elementos a serem modelados são representados como vetores dentro de um espaço vetorial.
- Dimensão do espaço é dada pelo número de palavras distintas.

Modelo Vetorial

- Componentes do sistema são vistos como vetores cujas coordenadas são determinadas pelas palavras que os descrevem.

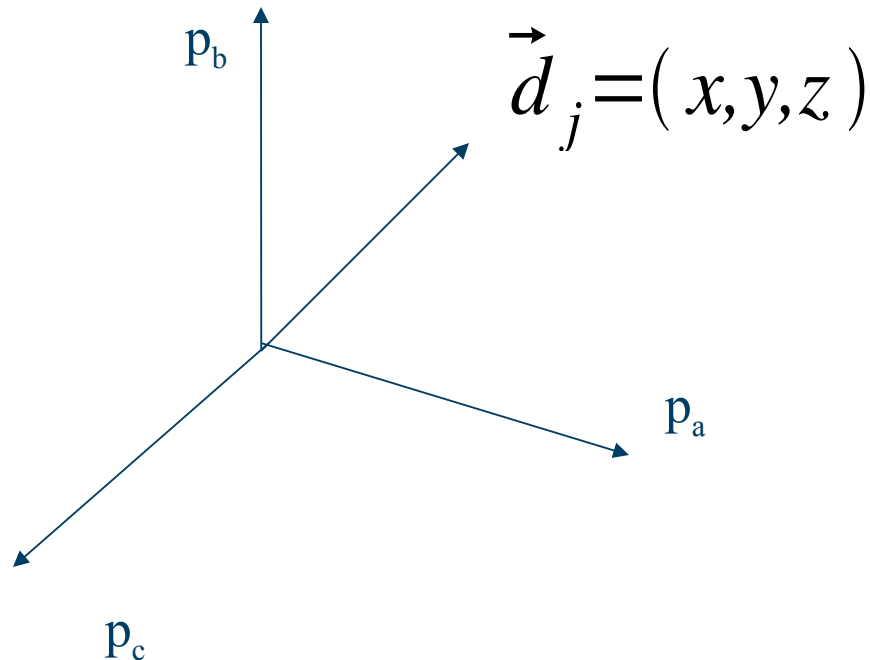
$$\vec{d}_j = (2, 2)$$

$$\vec{q} = (2, 1)$$



Modelo Vetorial

- Número de palavras distintas da coleção determina dimensão do espaço onde os documentos e consultas serão representados



Como determinar as coordenadas dos elementos ?

Medidas de Tf e Idf

- Idf tenta expressar a importância de uma palavra dentro de uma coleção.

N : número total de documentos de uma coleção

n_t : número de documentos onde a palavra t ocorreu

$$Idf(t) = \log \left(\frac{N}{n_t} \right)$$

Quanto mais rara a palavra, maior seu idf !

Determinação das Coordenadas

$$Idf(t) = \log\left(\frac{N}{n_t}\right)$$

Coordenada do
doc d no eixo t

$$w(d, t) = tf(d, t) \times idf(t)$$

Frequência da
palavra t no
documento d

Importância de t
na coleção

Exemplo

D1	A A A B
D2	A A C
D3	A A
D4	B B

$$idf(A) = \log\left(\frac{4}{3}\right) = 0,28$$

$$idf(B) = \log\left(\frac{4}{2}\right) = 0,69$$

Exemplo

D1	A A A B
D2	A A C
D3	A A
D4	B B

$$w(D1, A) = idf(A) \times tf(D1, A) = 0,28 \times 3 = 0,84$$

$$w(D1, B) = idf(B) \times tf(D1, B) = 0,69 \times 1 = 0,69$$

$$w(D1, C) = idf(C) \times tf(D1, C) = 1,38 \times 0 = 0$$

$$\vec{D1} = (0,84 ; 0,69 ; 0)$$

Exemplo

Para uma consulta Q composta por A e B :

$$w(Q, A) = idf(A) \times tf(Q, A) = 0,28 \times 1 = 0,28$$

$$w(Q, B) = idf(B) \times tf(Q, B) = 0,69 \times 1 = 0,69$$

$$w(Q, C) = idf(C) \times tf(Q, C) = 1,38 \times 0 = 0$$

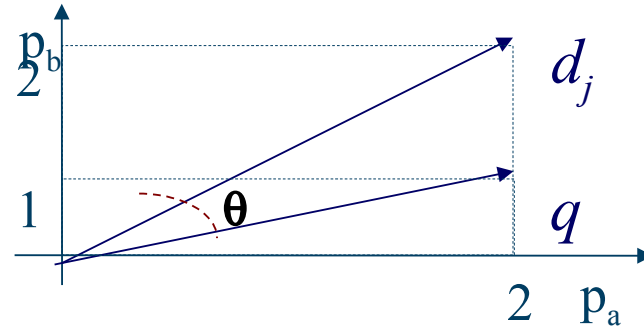
$$\vec{Q}1 = (0,28, 0,69, 0)$$

Similaridade

- Correlação entre dois vetores é utilizada para medir a proximidade entre os elementos reais modelados.

$$\vec{d}_j = (2, 2)$$

$$\vec{q} = (2, 1)$$



$$\text{sim}(d, q) = \cos \theta = \frac{\sum_{i=1}^t w(i, d) \times w(i, q)}{\sqrt{\sum (w(i, d))^2} \times \sqrt{\sum (w(i, q))^2}}$$

Norma de d

Norma de q



Implementação do Modelo Vetorial

Estrutura de Dados: Arquivo Invertido

- Composto de:
 - Vocabulário - contém cada termo (t) distinto da coleção;
 - Listas Invertidas - Para cada termo da coleção há uma lista invertida que indica a frequência com que o termo ocorre em cada documento da coleção;

Estruturas de Dados: Arquivo Invertido

D1	A A A B
D2	A A C
D3	A A
D4	B B

Vocabulário

A
B
C

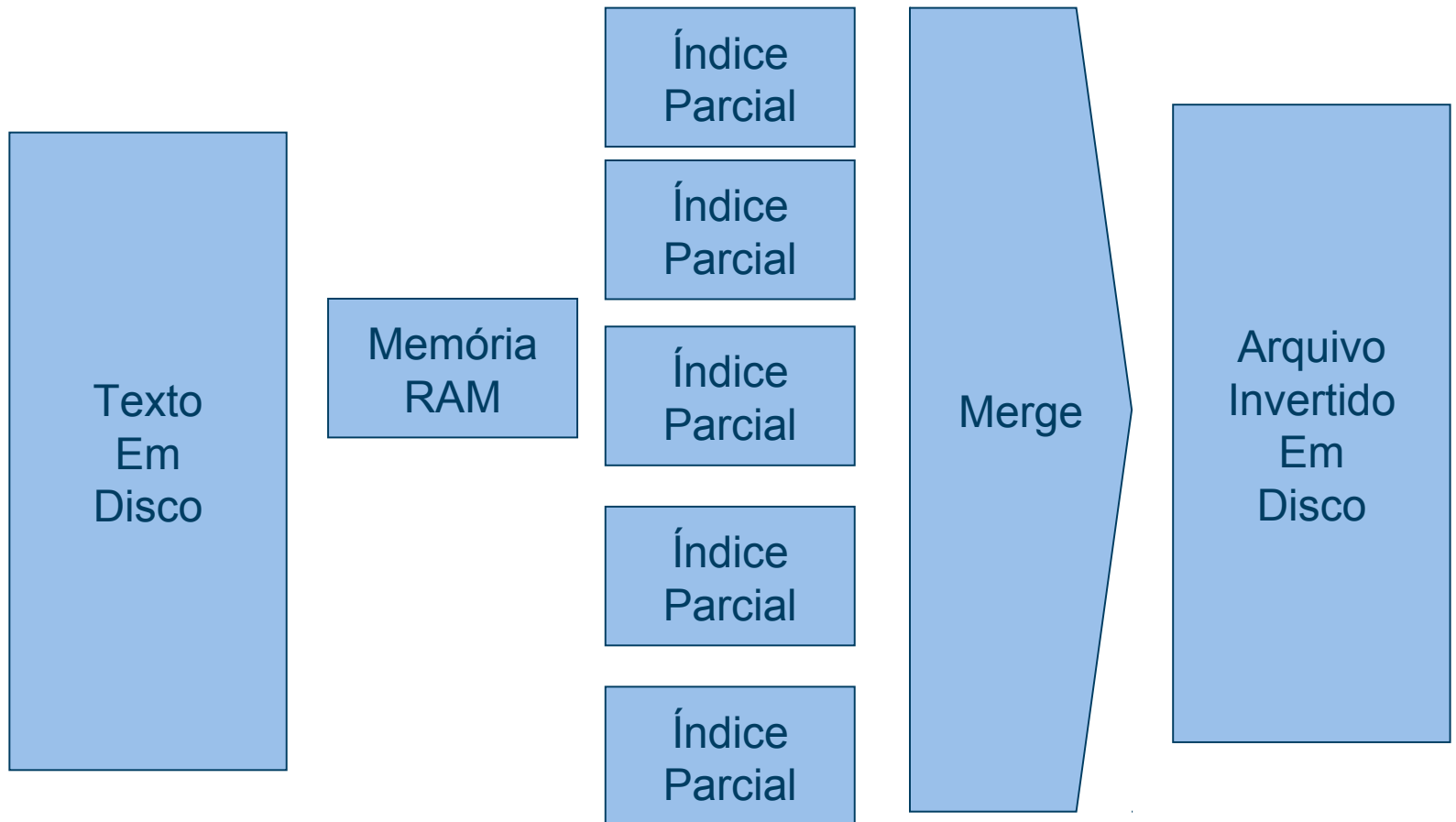
Listas invertidas

→ (1,3) (2,2) (3,2)
→ (1,1) (4,2)
→ (2,1)

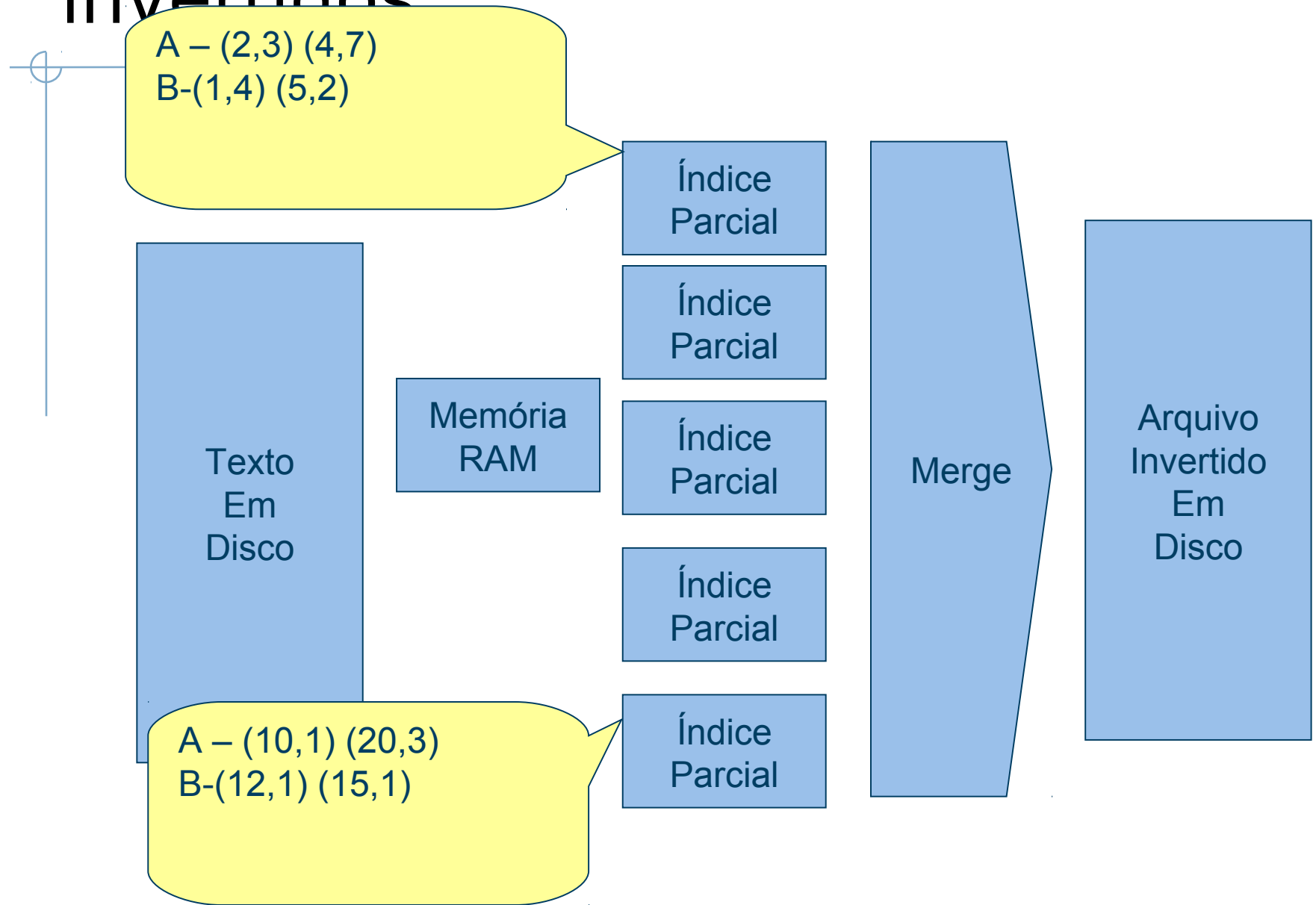
Arquivos Invertidos

- Construção:
 - Fácil de construir quando há memória RAM suficiente;
 - Algoritmos mais sofisticados são necessários quando a base de dados é muito grande;

Construção de arquivos invertidos



Construção de arquivos invertidos



Construção de arquivos invertidos

A – (2,3) (4,7)
B – (1,4) (5,2)

Texto
Em
Disco

Memória
RAM

Índice
Parcial

Índice
Parcial

Índice
Parcial

Índice
Parcial

Índice
Parcial

Merge

Arquivo
Invertido
Em
Disco

A – (10,1) (20,3)
B – (12,1) (15,1)

A – (2,3) (4,7) (10,1) (20,3) ...
B – (1,4) (5,2) (12,1) (15,1) ...

Merge

- Merge utiliza memória como buffer para acelerar o processo de indexação
- Compressão de dados pode ser usada para reduzir o tamanho dos índices e assim aumentar a velocidade de indexação

Processamento de Consultas

- Vocabulário e listas invertidas em disco;
- Idf dos termos e normas dos documentos são pré-computados;
- Palavras e listas freqüentes podem ter dados guardados em um cache

Processamento de Consultas

- Vocabulário fica em memória principal e listas invertidas em disco;
- Idf dos termos e normas dos documentos são pré-computadas e ficam em memória principal;
- Vocabulário pode ser guardado em um cache

Processamento de Consultas

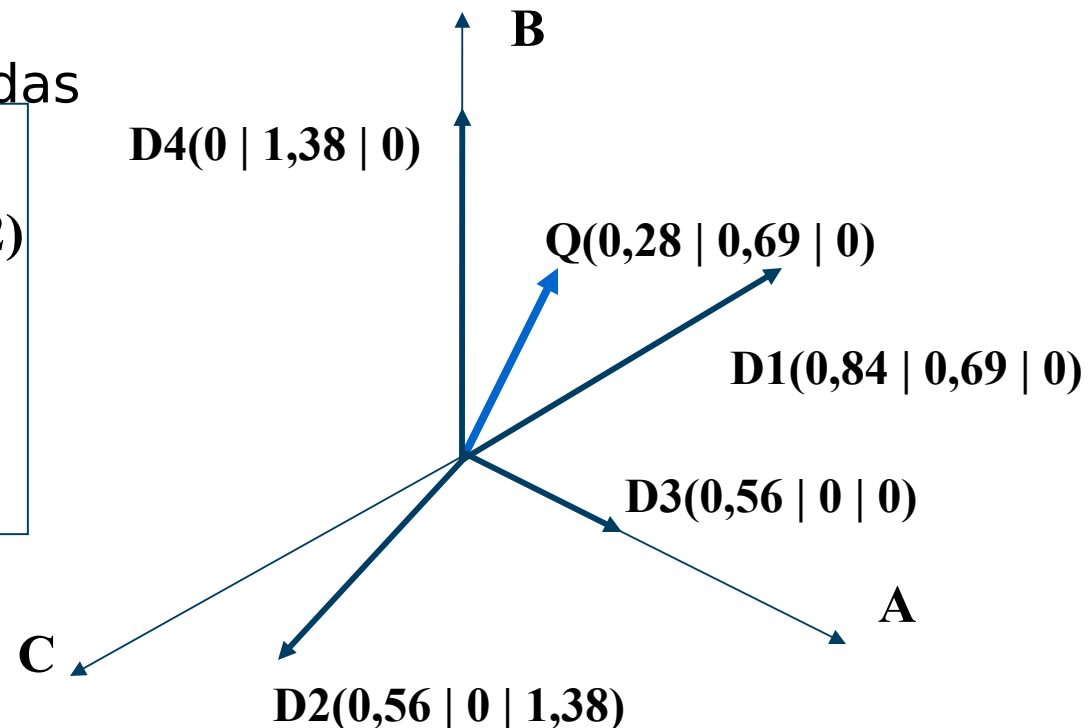
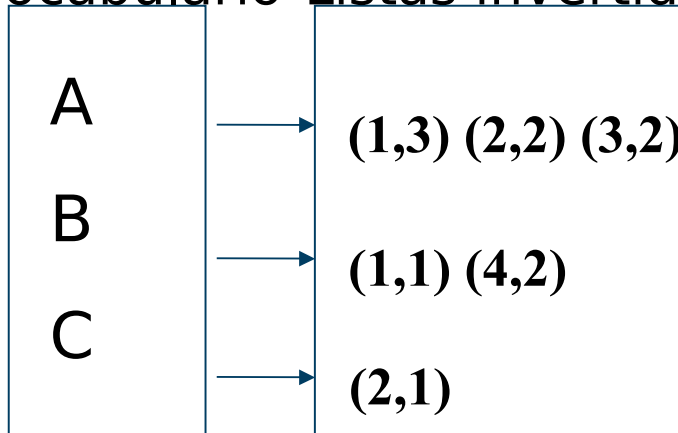
- Simplificações que aceleram o processamento:
 - Documentos que não possuem os termos das consultas têm similaridade igual a zero;
 - O cosseno é calculado de maneira que se possa ler as listas invertidas seqüencialmente durante o cálculo;

Estruturas de Dados: Arquivo Invertido

Q = "A B"

D1	A A A B
D2	A A C
D3	A A
D4	B B

Vocabulário Listas invertidas



Acumuladores para cálculo da similaridade parcial

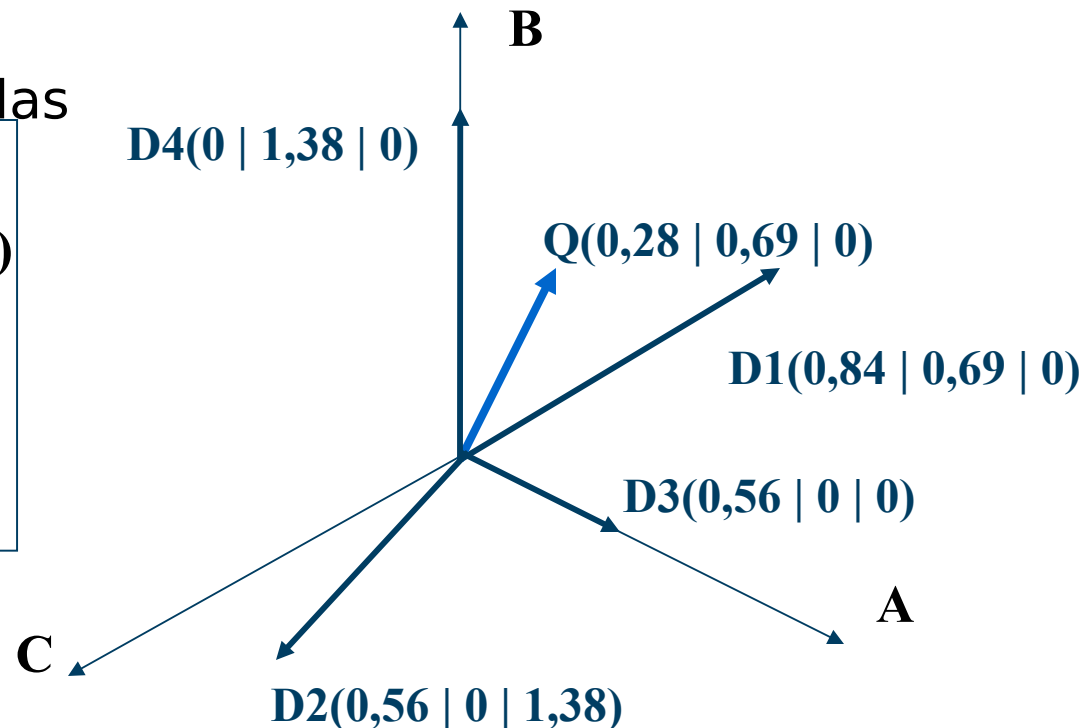
Acumuladores:

D1	D2	D3	D4
0	0	0	0

$$sim_{parc}(q, d, t) = w(d, t) \times w(q, t)$$

Vocabulário Listas invertidas

A	→	(1,3) (2,2) (3,2)
B	→	(1,1) (4,2)
C	→	(2,1)



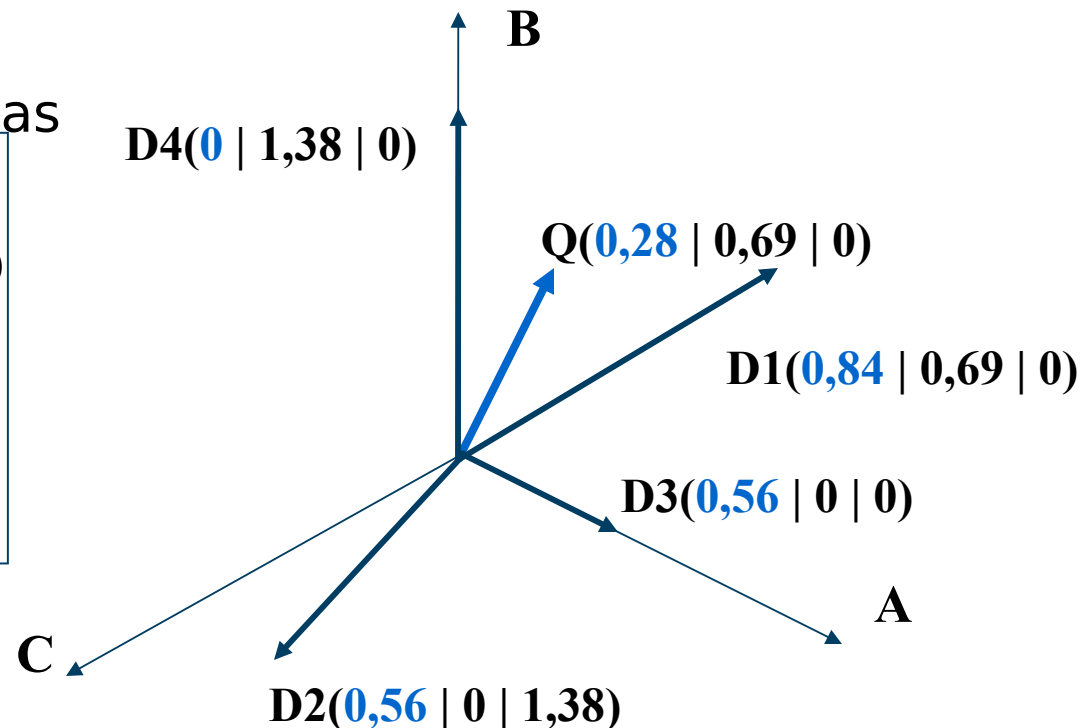
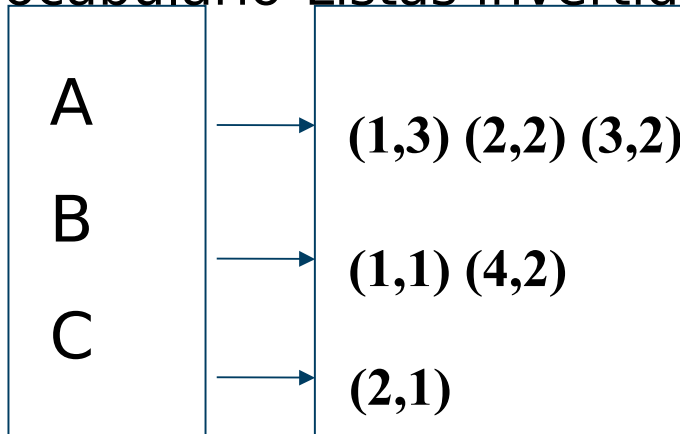
Acumuladores para cálculo da similaridade parcial

Acumuladores:

D1	D2	D3	D4
0,24	0,16	0,16	0

$$sim_{parc}(Q, D1, A) = w(D1, A) \times w(Q, A) = 0,84 \times 0,28 = 0,24$$

Vocabulário Listas invertidas



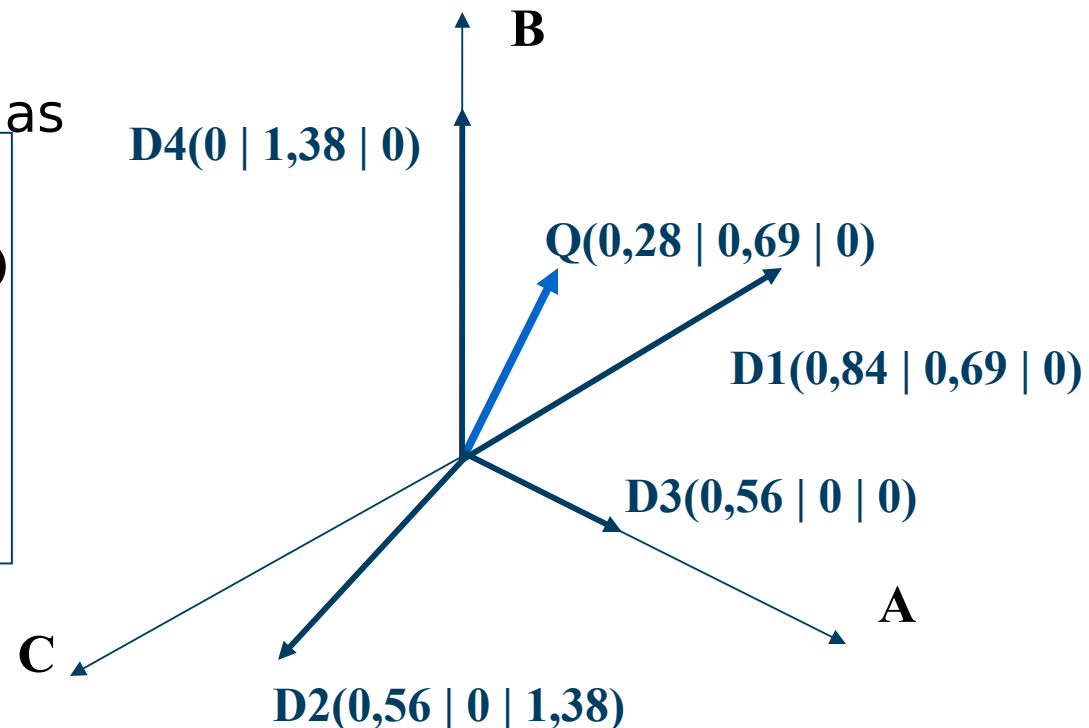
Acumuladores para cálculo da similaridade parcial

Acumuladores:

D1	D2	D3	D4
0,71	0,16	0,16	0,95

Vocabulário Listas invertidas

A	→	(1,3) (2,2) (3,2)
B	→	(1,1) (4,2)
C	→	(2,1)



Acumuladores para cálculo da similaridade parcial

Acumuladores:

D1	D2	D3	D4
0,71	0,16	0,16	0,95

$$\text{sim}(d, q) = \cos \theta = \frac{\sum_{i=1}^t w(i, d) \times w(i, q)}{\sqrt{\sum (w(i, d))^2} \times \sqrt{\sum (w(i, q))^2}}$$

Norma do documento

Norma da consulta

Acumuladores para cálculo da similaridade parcial

$$norma(d1) = \sqrt{\sum (w(i, d1))^2} = \sqrt{(0,84)^2 + (0,69)^2} = 1,08$$

$$norma(d2) = 1,49$$

$$norma(d3) = 0,56$$

$$norma(d4) = 1,38$$

Acumuladores para cálculo da similaridade parcial

Acumuladores:

D1	D2	D3	D4
0,71	0,16	0,16	0,95

$$\text{sim}(d1, q) = \frac{\text{Acum}(d1)}{\|\vec{d1}\| \times \|\vec{q}\|} = \frac{0,71}{1,08 \times \|\vec{q}\|} = \frac{0,66}{\|\vec{q}\|}$$

$$\text{norma}(d1) = 1,08$$

$$\text{sim}(d2, q) = \frac{0,16}{1,49 \times \|\vec{q}\|} = \frac{0,17}{\|\vec{q}\|}$$

$$\text{norma}(d2) = 1,49$$

$$\text{sim}(d3, q) = \frac{0,16}{0,56 \times \|\vec{q}\|} = \frac{0,28}{\|\vec{q}\|}$$

$$\text{norma}(d3) = 0,56$$

$$\text{sim}(d4, q) = \frac{0,95}{1,38 \times \|\vec{q}\|} = \frac{0,69}{\|\vec{q}\|}$$

$$\text{norma}(d4) = 1,38$$

Algoritmo

1. Para cada documento d na coleção, criar $A\{d\} = 0$.
2. Para cada termo t na consulta,
 - (a) Recuperar a lista invertida para o termo t do disco.
 - (b) Para cada entrada $\langle d, f(d, t) \rangle$ na lista invertida,
$$A\{d\} = A\{d\} + \text{sim}(q, d, t) \cdot f(d, t).$$
3. Dividir cada acumulador $A\{d\} \neq 0$ pela norma do documento $\|\vec{d}\|$.
4. Identificar os k valores mais altos de acumuladores, onde k é o número de documentos retornados para o usuário.

Figura 3: Algoritmo para o cálculo da similaridade no modelo vetorial.



Dúvidas?



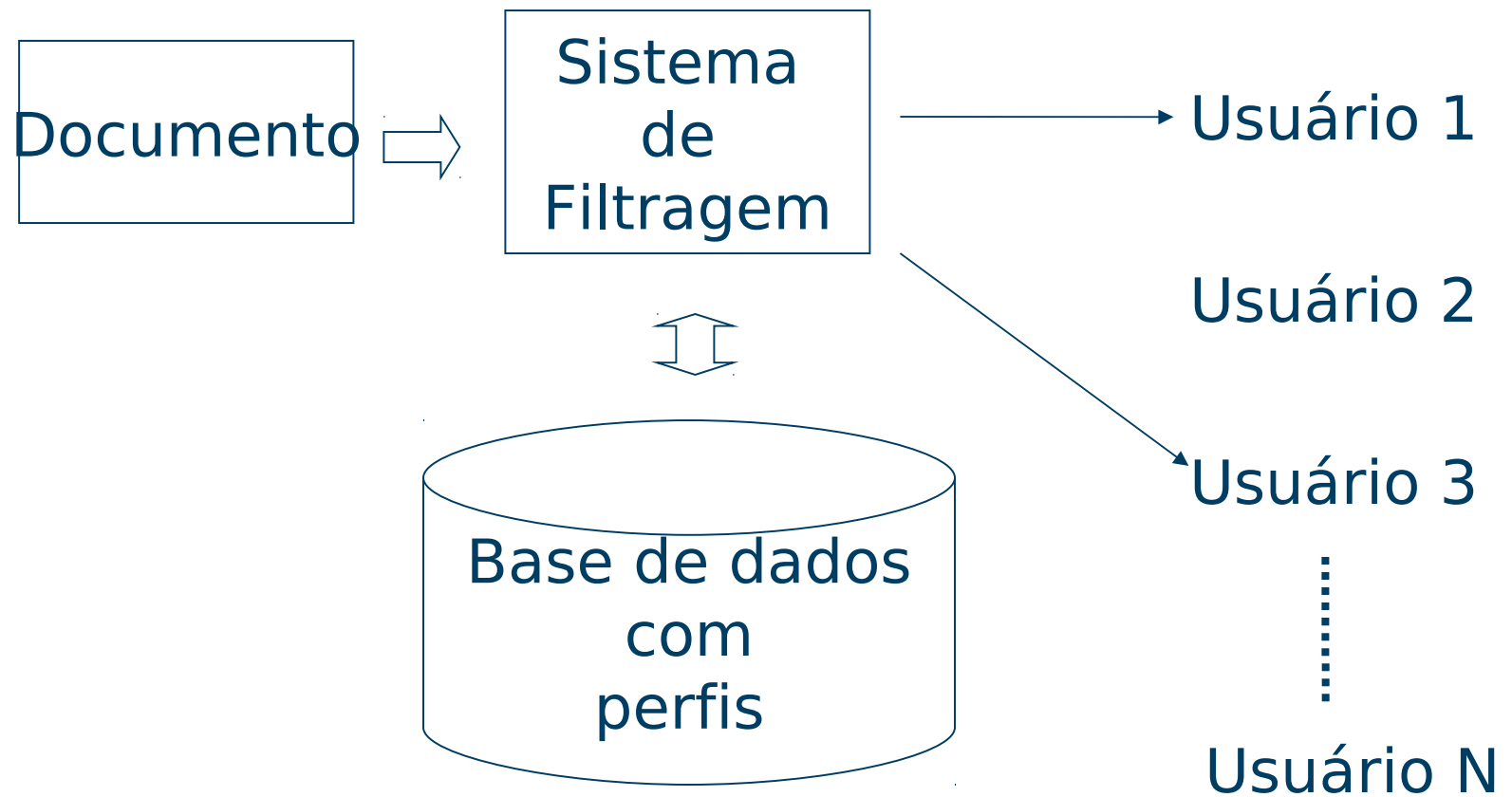
Busca com Modelo Vetorial

- Aplicação direta do modelo.

Filtragem com Modelo Vetorial

- Bases de dados contêm perfis no lugar de documentos.
- Perfis são conjuntos de termos que descrevem os interesses dos usuários.
- Documentos que chegam para o sistema são tratado como consultas.

Filtragem com Modelo Vetorial



Perfil

- Perfil pode ser um conjunto de palavras
- Exemplo:
 - Senado federal: senado, senador, votação, lei, nomes de senadores em geral e etc...
- Note que o perfil pode ser visto como um conjunto de palavras

Filtragem

- Guarda-se estatísticas sobre todas as palavras encontradas nos perfis e nos documentos processados pelo sistema para que se tenha o idf dessas palavras
- Similaridade entre perfil e documento é computada da mesma forma que similaridade com consulta é computada no modelo vetorial
- $Idf = \log ((qt \text{ perfil} + qt \text{ docs}) / (qt \text{ palavra perfil} + docs))$

Atualização Automática de Perfil

- Os perfis de um sistema de filtragem podem ser aperfeiçoados automaticamente com o tempo
- Para isso é necessário que os usuários realimentem o sistema com informação sobre os documentos recebidos
- Ex.: Quais documentos são relevantes ou quais não são relevantes
- Esta informação pode ser usada para melhorar automaticamente o perfil do usuário

Atualização Automática de Perfil

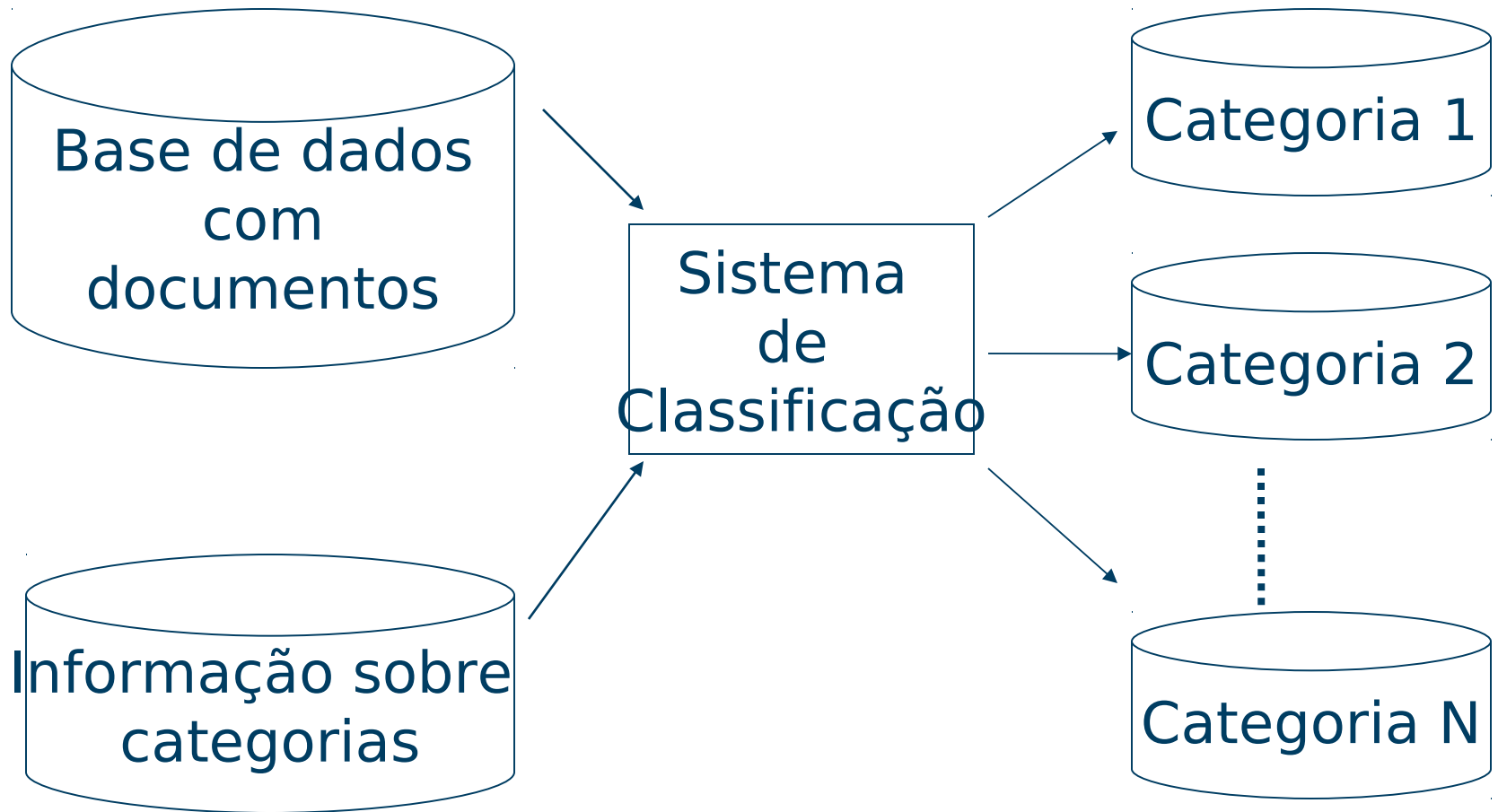
- Rocchio: novo perfil é calculado em função do perfil anterior e de informações dadas pelos usuários.
- Usuário indica se um documento recebido é relevante ou não.
- Palavras dos perfis podem ter pesos positivos ou negativos com o passar do tempo.
- Objetivo é fazer com que o perfil melhore com o tempo.

Fórmula de Rocchio

$$\vec{P}' = \alpha \times \vec{P} + \frac{\beta}{m_r} \sum_{i=1}^{m_r} \vec{D}_i^r - \frac{\gamma}{m_n} \sum_{i=1}^{m_n} \vec{D}_i^n$$

- α , β e γ são constantes definidas em experimentos que determinam a importância de cada termo da equação.
- M_r é o número de docs relevantes.
- M_n é o número de docs não relevantes.
- P é o perfil anterior, D^r são os vetores dos docs relevantes encontrados e D^n dos não relevantes encontrados.

Classificação de Documentos



Classificação com Modelo Vetorial

- Pequena coleção de documentos previamente classificada para treinar o programa (informação sobre categorias).
- As palavras muito freqüentes em uma categoria e menos freqüentes na coleção são selecionadas como descritores da categoria.
- Descritor é um conjunto de palavras que descreve uma categoria.

Classificação com Modelo Vetorial

- Calcula-se a similaridade entre cada documento da coleção completa e cada descritor.
- Se a similaridade de um documento ultrapassar um determinado limiar, então o documento é classificado como pertencente a categoria.

Exemplo

- Classificação da base de dados do sistema TodoBR (10 milhões de páginas).
- Categorias relacionadas à Educação:
 - Biologia;
 - Física;
 - Geografia;
 - História e outras.

Base de Dados para Treinamento

Categoria	N# de documentos (*)
Biologia	104
Física	121
Geografia	130
História	112

Resultados

Categoria	N# Docs. Classif.	% estimado de Acertos
Biologia	1477	97% (+/- 3.7%)
Física	1487	99% (+/- 2.5%)
Geografia	14667	95% (+/- 4.2%)
História	3295	93% (+/- 4.6%)

Exemplo de descritor:

- Biologia:

- Mitocôndria, Fagocitose, DNA, Genética...



Language models

Statistical Language Models

- Language models: Mecanismos probabilísticos para modelar fontes textuais
- Aplicação inicial em compressão de dados
- Muito aplicado em reconhecimento de voz (desde a década de 70)
- Aplicado em tradutores na década de 80
- Aplicado em RI a partir da década de 90

Statistical Language Models

- São modelos probabilísticos
- Ao invés de calcular a probabilidade de um documento ser relevante dada uma consulta
- Calcula-se a probabilidade de uma consulta ser gerada por um modelo extraído a partir de um documento.

Exemplo

- Espera-se que:
 - $P(\text{"departamento de ciência da computação"} | \text{HOME DO DCC}) >$
 - $P(\text{"departamento de ciência da computação"} | \text{HOME DA UFAM})$
- A fonte de informação HOME DO DCC tem mais chance de gerar a seqüência do que a fonte de informação HOME DA UFAM
- Desafio é aprender o modelo que representa cada documento D para calcular $P(Q|D)$.

Idéia similar à compressão de dados

- A idéia foi inspirada em um modelo proposto em 48 para compressão de textos

Modelo Mais simples

- Considerar cada palavra como sendo independente das demais
- $p(w_1 w_2 \dots w_n) = p(w_1)p(w_2)\dots p(w_n)$
- Probabilidade de uma dada consulta ser gerada por um documento:

$$P(Q = \{k_1, \dots, k_m\} | D) = P(K_1 | D) \times \dots \times P(K_m | D)$$

- Onde K_i é um termo de Q e m é o número de termos

Como estimar $P(K|D)$?

- Pode-se contar número de ocorrências de K em D e dividir pelo total de palavras em D :
- Ex: K aparece 10 vezes de um total de 100 palavras em D : $P(K|D)=0,1$
- Problemas com essa abordagem ???

Suavização (smoothing)

- Se uma palavra não aparece no documento, sua probabilidade não pode ser 0!
- Se uma palavra da consulta não aparece em D então $P(Q|D) = 0$?
- Solução é fazer uma suavização no modelo

Additive smoothing [Relatório Técnico: Chen & Goodman 98]:

- Adicionar uma constante à contagem de cada palavra:

$$P(K|D) = \frac{f(K,D) + C}{\sum f(k_i, D) + C \times V}$$

- Todas as palavras desconhecidas têm a mesma probabilidade
- Resolve o problema da multiplicação, mas continua sendo ruim!

Exercício

- Calcule o ranking para a coleção de exemplo com a consulta K1,K3 utilizando o modelo proposto. Considere $C = 1$;

	d1	d2	d3	d4	d5	d6
K1	1	0	1	0	1	0
K2	0	1	1	1	1	1
K3	0	0	0	1	0	0

Exercício

- $P(K1,d1) = (1+1)/1+3 = 0,5$

$$P(K|D) = \frac{f(K,D) + C}{\sum f(k_i,D) + C \times V}$$

Outras alternativas

- Há muitas alternativas de suavização
- Resultados acabam incluindo conceitos similares aos de tf e idf.
- Melhor resultado para busca obtido pelo método conhecido como Dirichlet (Zhai and J. Lafferty, SIGIR 2001 e SIGIR 2002)
- Testes com várias coleções(incluindo web trec 8), consultas curtas e longas

Dirichlet Prior/Bayesian

$$P(k|D) = \frac{f(k,D) + \mu P(k|REF)}{|D| + \mu} = \frac{|D|}{|D| + \mu} \frac{f(k,D)}{|D|} + \frac{\mu}{|D| + \mu} P(k, REF)$$

- $P(w|REF)$ é uma probabilidade do termo na coleção
- Problema de Dirichlet: vem da física e da matemática e consiste em encontrar uma função contínua que está encerrada dentro de determinados limites.

Suavização & TF-IDF

[Zhai & Lafferty 01a]

- Fórmula final proposta:

$$\log p(q|d) = \sum_{\substack{w \in V, c(w,d) > 0 \\ c(w,q) > 0}} \text{TF weighting} \cdot c(w,q) \log \frac{p_{DML}(w|d)}{\alpha_d p(w|REF)} + |q| \log \alpha_d + \sum_{w \in V} \text{Ignore for ranking} \cdot c(w,q) p(w|REF)$$

Doc length normalization
(long doc is expected to have a smaller α_d)

Words in both query and doc

IDF-like weighting

Ignore for ranking

$$\alpha_d = \frac{1 - \sum_{w \text{ is seen}} p_{DML}(w|d)}{\sum_{w \text{ is unseen}} p(w|REF)}$$

$$P_{DML}(w|d) = \frac{f(w,d) + \mu P(w|REF)}{d + \mu}$$

$$P(w, REF) = \frac{f(w, Col)}{\sum f(w_i, Col)}$$

Melhor valor de u

- O valor de u varia de coleção para coleção, mas nos experimentos ficou próximo a 2000.
- Método é bom para consultas curtas. Para consultas longas devem ser usadas outras formas de suavização

Modelos mais sofisticados

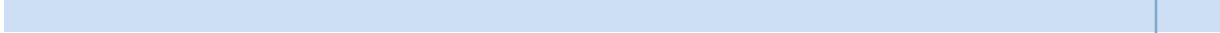
- Quando modelamos ocorrência de palavras como independentes estamos simplificando o modelo da linguagem
- A probabilidade de uma dada palavra ocorrer depende bastante das palavras que ocorreram anteriormente (dependência entre palavras)

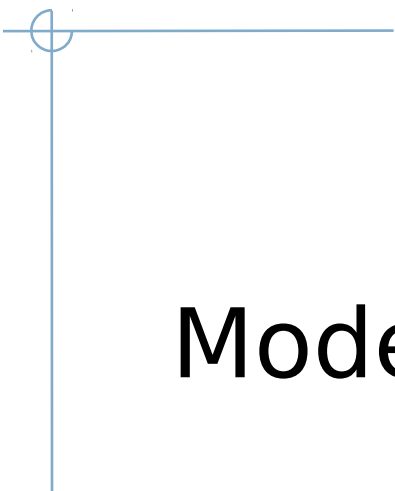
O que fazer para melhorar ?

- Modelos utilizando N-gramas (n-gram models):
 - Bigramas e trigramas tem sido estudados
 - Melhorias insignificantes (podemos mudar?)
- Parsimonius LM:
 - Equivalente à remoção de stopwords.

O que fazer para melhorar

- Como modelar relacionamentos entre palavras ???





Modelo Baseado em Conjuntos

SET-BASED MODEL

Modelo Baseado em Conjuntos (Set Based Model)

- Modela correlações entre termos das consultas utilizando teorias de conjuntos e técnicas de mineração de dados

Conceitos Básicos

- Termset
- Termset frequente
- Termset proximais
- Regras de associação entre termsets
- Fechamento

Termsets

- Um n-termset é um conjunto de n termos
- Vocabulário de conjuntos com 2^t elementos, onde t é o número de termos:
- $V = \{S_1, S_2, \dots, S_{2^t}\}$
- Cada termset S_i tem uma lista invertida IS_i (computada dinamicamente)

Termset Frequente

- Um termset é freqüente se sua lista invertida é maior do que um determinado limiar (conhecido como suporte em mineração de dados e como freqüência mínima em RI)
- Um n-termset é freqüente sse todos os seus subconjuntos também são freqüentes.

Exemplo

■ Coleção

A C
A C
E

C D
E D
E

A C
A C
A C

D E

A B C
D C D
E

B C
D F

- Vocabulário: {a,b,c,d,e,f}
- $Q = \{a,b,c,d,f\}$
- 32 possíveis termsets
- 23 ocorrem na coleção
- Quantos com frequência mínima 2 ?

Resposta

- 5 de 1 termo A,B,C,D,E
- 7 de 2 termos
- 3 de 3 termos
- 15 no total

Termsets proximais

- Termsets proximais são compostos por termos que ocorrem a uma distância máxima dentro do texto
- Considerar relações apenas entre termos que ocorrem dentro de um mesmo contexto
- Serve também para reduzir o número de termsets

Regras de associação

- Cada termset carrega informação sobre a associação entre termos
- Contudo, utilizar todas as associações pode não ser uma boa idéia (por exemplo, sobreposição)
- Para selecionar boas associações podemos utilizar regras de associação

Exemplo

- Na coleção de exemplo, os conjuntos S_{ac} (3 vezes) S_{abc} são freqüentes
- Devemos utilizar os dois ?

Regras de associação

- Uma regra é uma implicação $X \rightarrow Y$, onde X e Y são termsets.
- Regras são caracterizadas pelo grau de confiança, que indica a probabilidade de Y aparecer em um documento, dado que apareceu X

Exemplos

A C A C E	C D E D E	A C A C A C
D E	A B C D C D E	B C D F

REGRA	CONFIANÇA
$S_a \rightarrow S_{ab}$	33
$S_{ab} \rightarrow S_{abc}$	100
$S_{ac} \rightarrow S_{ab}$	33
$S_{abcd} \rightarrow S_{bcdf}$	0

- Associações com 100% de confiança implicam no descarte do conjunto menor

Fechamento

- O fechamento de um termset S é o conjunto de todos os termsets que co-ocorrem nos mesmos documentos de S
- Um termset é fechado se ele é o “maior” dentre os termsets de um fechamento
- Termsets fechados aparecem como conseqüências dentro das regras de associação com 100% de confiança.

Set-Based Vector Model

- Documentos e consultas são representados em um espaço determinado por todos os conjuntos possíveis
- Pesos são determinados por tf e idf:
 - Tf = número de vezes que conjunto ocorreu no documento
 - Idf = calculado em função do tamanho da lista invertida do conjunto
 - Utiliza apenas termsets fechados no

Resultados

- Ganhos de 10.66% sobre o vetorial na WBR99 (com proximidade)
- Ganho de 2.79% sem proximidade
- Ganhos de 30% na TREC8
- Tempo de execução próximo ao do vetorial

Dúvidas

- Qual seria o ganho quando combinado a outras evidências ?
- Qual seria a perda de eficiência em um sistema real ?
- Como selecionar os parâmetros para cada coleção (frequência de corte e distância) ?
- Uso de passagens melhoraria os resultados ?

Exercício

- Execute a consulta da nossa coleção de exemplo para o set-based model. Considere que não há cortes nos conjuntos nesse exemplo.



Avaliação de Sistemas de RI

Avaliação de Sistemas de Busca

- N – conjunto de documentos relevantes identificados pelos especialistas
- R – conjunto de documentos respondidos pelo sistema que foram examinados.

Precisão e Revocação

$$\text{Precisão} = \frac{|N \cap R|}{|R|}$$

$$\text{Revocação} = \frac{|N \cap R|}{|N|}$$

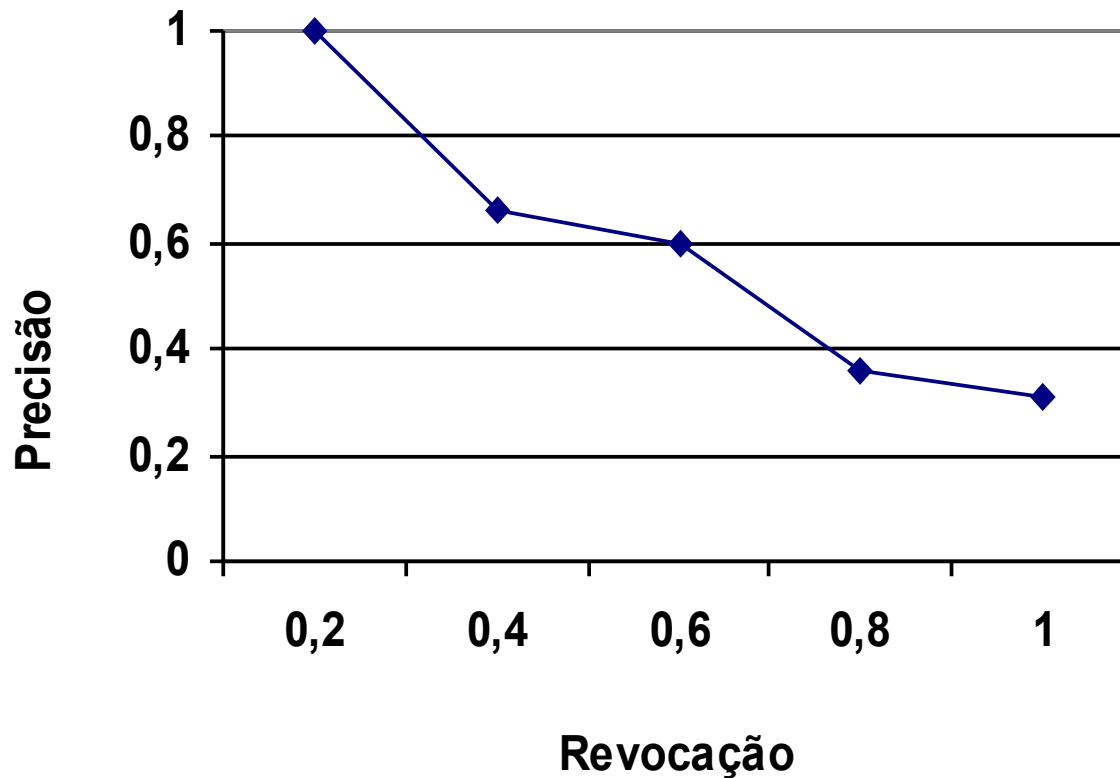
Curva de Precisão e Revocação

- Para facilitar a avaliação dos resultados é traçado um gráfico que mostra a evolução da precisão em função da revocação.
- Gráfico é conhecido como curva de precisão e revocação.

Exemplo

- Documentos relevantes: {1, 4, 8, 44, 72}.
- Um sistema recupera o vetor resultado: $\langle 8, 22, 72, 3, 1, 2, 24, 6, 33, 45, 4, 48, 55, 32, 11, 44 \rangle$.
- O nível de revocação 20% é atingido quando encontramos o primeiro documento relevante (8), a precisão é de $1/1 = 100\%$.
- Para revocação de 40% a precisão é igual a $2/3 = 66\%$.

Curva de Precisão e Revocação

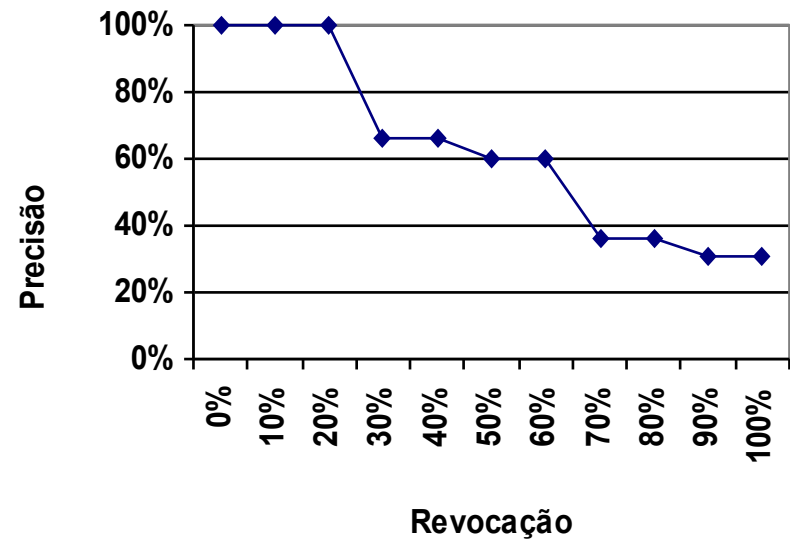
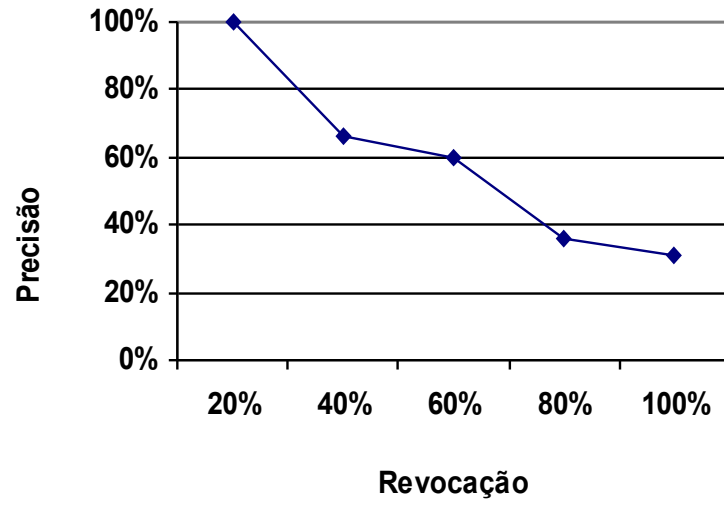


Precisão média

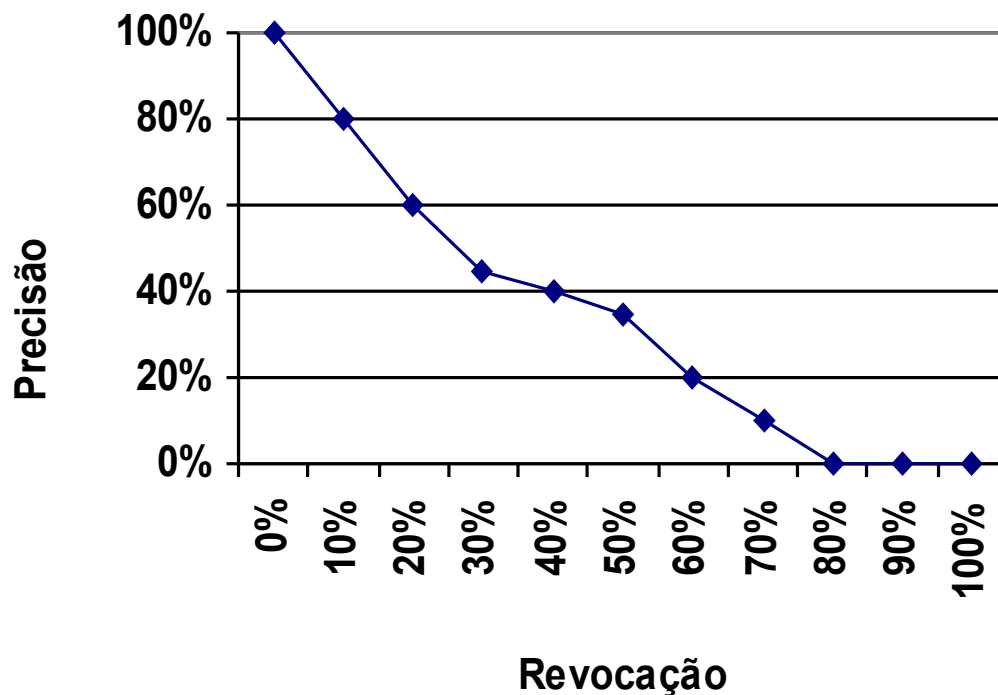
- Normalmente é interessante que se faça a avaliação do sistema utilizando-se uma média das precisões obtidas em várias consultas
- Pontos de precisão conhecidos são diferentes para cada consulta
 - Solução é criar uma forma de se ter valores conhecidos nos mesmos pontos em todas as consultas.

Precisão nos 11 pontos

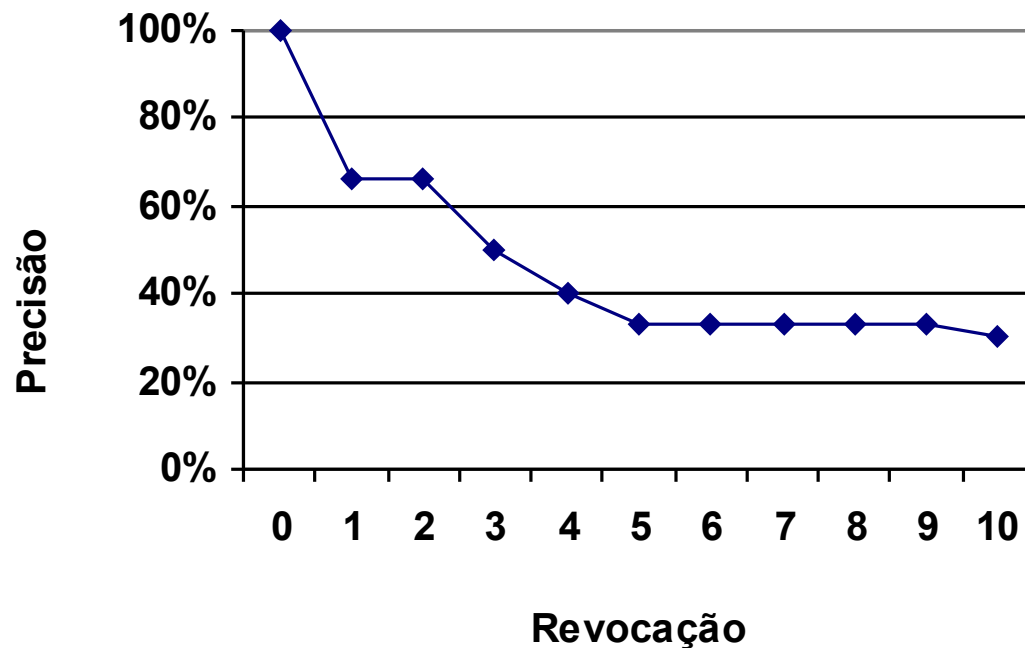
- Utiliza-se um método de interpolação para se obter a precisão em 11 pontos de revocação (0%, 10%, 20%, ..., 100%)
- Interpolação é feita tomando-se a precisão máxima conhecida entre o ponto atual e o próximo. Se não houver resultado, busca-se os próximos pontos até que se tenha uma definição



Se não recupera todos os relevantes precisão cai a zero



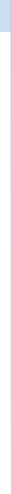
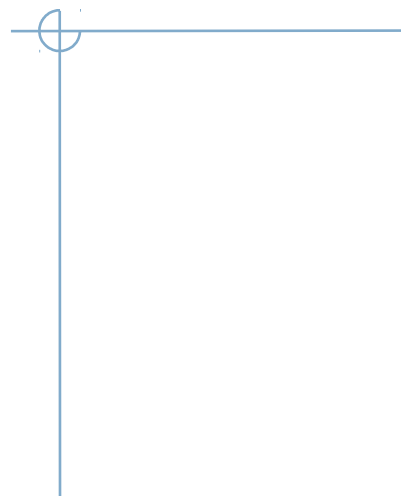
Precisão também pode ser computada em função do número de documentos vistos



Neste caso também utiliza-se interpolação

Exercício

- Documentos relevantes: {1, 4, 8, 25, 44, 53, 72}.
- Um sistema recupera o vetor resultado: <8, 22, 72, 1, 3, 2, 25, 6, 33, 45, 4, 48, 55, 32, 11, 44>.
- Mostre a curva de precisão e revocação utilizando os 11 pontos padronizados
- Mostre a curva de precisão nos 10 primeiros elementos da resposta



Calculando valores únicos

- Em alguns casos é interessante que se tenha um único valor de precisão para cada consulta
- Com este valor é possível comparar diretamente dois sistemas e determinar qual o melhor

Medidas de precisão escalares

- Precisão média nos relevantes encontrados
 - Tira-se a média das precisões nos pontos de revocação onde apareceram documentos relevantes (MAP não interpolado)
 - Há também o MAP Interpolado (média nos 11 pontos)
 - Note que sistemas que não encontraram todos os relevantes podem ser beneficiados pelo MAP não interpolado

Exercício

- Calcule o MAP para o exercício anterior

Medidas de precisão escalares

- Precisão-R

- Calcula-se a precisão na R-ésima posição do ranking
- Muito utilizada quando assume-se que usuário está interessado nos R primeiros.
- Por exemplo, nas máquinas costuma-se assumir que usuário está interessado em respostas apenas entre os 10 primeiros itens

Medidas Escalares (Vorhees, SIGIR, 2004)

$$bpref_{10} = \frac{1}{R} \sum_{r=1}^R 1 - \frac{Irrelevant_R(r)}{R+10}$$

- R é o número de relevantes
- $Irrelevant_R(r)$ é o número de documentos irrelevantes acima de r, entre os R+10 documentos do topo (valor maximo é R+10).
- Falha quando número de relevantes é pequeno
- Ex: $\frac{1}{7} \times ((1-0/17) + (1-1/17)(1-1/17) + (1-3/17) + (1-6/17) + (1-10/17) + 0) = 0.67$

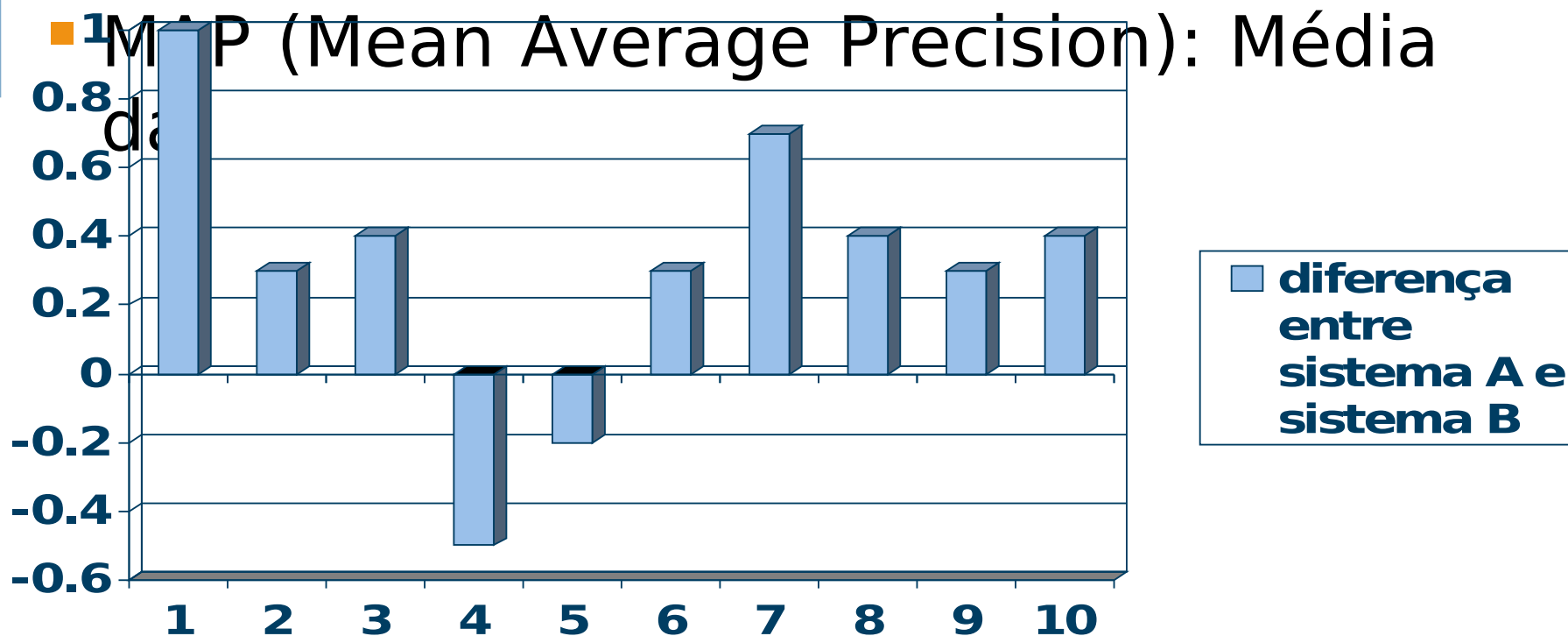
Exercício



Calcule o $B_{\text{pref-10}}$ para o exercício anterior!

Histograma de precisão

- Pode-se montar um histograma com as diferenças entre dois sistemas em vários pontos de precisão-R



Resumos comparativos

- Pode-se ainda montar resumos comparativos sobre os sistemas que estão sendo experimentados
 - Exemplo: número de consultas usadas no experimento, número médio de docs recuperados por consulta, número médio de relevantes por consulta, precisão-10 de cada sistema e assim por diante.

Medidas alternativas

- Medida F1(Média Harmônica)

$$F1(j) = \frac{2}{\frac{1}{R(j)} + \frac{1}{P(j)}}$$

Onde $R(j)$ é a revocação em um dado ponto j e $P(j)$ é a precisão neste ponto

A medida F é útil para combinar a precisão e a revocação em um único número

Medidas Alternativas

- Medida F

$$F(j) = \frac{1 + b^2}{\frac{b^2}{R(j)} + \frac{1}{P(j)}}$$

- Valores de b maiores que 1 indicam que o usuário está mais interessado na precisão
- Valores menores que 1 indicam que usuário está mais interessado na revocação

Exercício

- Calcule a medida F para o exercício anterior

MRR(Mean Reciprocal Ranking)

$$MRR(S, Q) = \frac{\sum_{q \in Q} \frac{1}{PosRel(S(q))}}{|Q|}$$

- Onde Q é um conjunto de consultas
- S é um sistema de ranking
- $PosRel(R(q))$ é a posição da primeira resposta relevante no ranking do sistema S para a consulta q
- $|Q|$ é o número de consultas avaliadas

Exercício

- Calcule o MRR para o exercício anterior

Métricas baseadas em Ganho Cumulativo (CG)

- Quando examinamos uma resposta de um sistema, fica claro que:
 - Alguns documentos relevantes atendem melhor às necessidades dos usuários que outros
 - Quanto mais longe do topo um documento relevante está, menor a sua utilidade na resposta
- Métricas baseadas em ganho cumulativo tentam incorporar estes dois fatos na avaliação

Ganho Cumulativo (CG)

- Documentos da resposta são substituídos pelos seus graus de relevância:

$\langle 3, 0, 1, 2, 3, 0, 0, 1 \rangle$

- Ganho cumulativo (CG) é igual a soma dos valores de ganho obtidos até cada posição:

$\langle 3, 3, 4, 6, 9, 9, 9, 10 \rangle$

- CG leva em consideração a relevância, mas não a posição

Calcule do CG para o exemplo abaixo ate o 6o documento

- Documentos relevantes: $\{1(3), 4(1), 8(2), 25(1), 44(3), 53(3), 72(2)\}$.
- Um sistema recupera o vetor resultado: $\langle 8, 22, 72, 1, 3, 2, 25, 6, 33, 45, 4, 48, 55, 32, 11, 44 \rangle$.
- Um sistema recupera o vetor resultado: $\langle 25, 22, 4, 3, 72, 2, 8, 6, 33, 45, 4, 48, 55, 32, 11, 44 \rangle$.

Ganho Cumulativo Descontado (DCG)

- Inclui a noção de que documentos relevantes têm a utilidade reduzida na medida em que são apresentados mais longe do topo da resposta
- A proposta é incluir um fator de desconto no ganho de acordo com a posição na qual os documentos são apresentados
- Uma proposta é dividir pelo log da posição no ranking
- A base do logaritmo ajusta o fator de desconto e o log não é aplicado para a primeira posição do ranking

DCG

- Usando log na base 2:
 $\langle 3 ; 3 ; 2,51 ; 3 \dots \rangle$
- Valores médios de DCG podem ser computados para avaliar o desempenho de um sistema e gráficos de DCG podem ser criados para facilitar a visualização dos resultados da avaliação

Calcule do DCG para o exemplo abaixo ate o 6o documento

- Documentos relevantes: $\{1(3), 4(1), 8(2), 25(1), 44(3), 53(3), 72(2)\}$.
- Um sistema recupera o vetor resultado: $\langle 8, 22, 72, 1, 3, 2, 25, 6, 33, 45, 4, 48, 55, 32, 11, 44 \rangle$.
- Um sistema recupera o vetor resultado: $\langle 25, 22, 4, 3, 72, 2, 8, 6, 33, 45, 4, 48, 55, 32, 11, 44 \rangle$.

NDCG (Jarvelin et al, TOIS, 2002)

- Normalized Cumulative Discount Gain
- Ganho Cumulativo Descontado Normalizado
- Valor ótimo de DCG poderia ser obtido com sistema que coloca os documentos ordenados de forma decrescente por valor de relevância:
<3,3,3,3,2,2,2,2,1,1,1,0,0,0,0>
- NDCG divide o DCG de cada sistema pelo DCG de um sistema ideal, obtendo valores entre 0 e 1 para cada posição do ranking.
- NDCG@K: NDCG obtido na k-ésima posição do ranking.

Calcule do NDCG para o exemplo abaixo ate o 6o documento

- Documentos relevantes: $\{1(3), 4(1), 8(2), 25(1), 44(3), 53(3), 72(2)\}$.
- Um sistema recupera o vetor resultado: $\langle 8, 22, 72, 1, 3, 2, 25, 6, 33, 45, 4, 48, 55, 32, 11, 44 \rangle$.
- Um sistema recupera o vetor resultado: $\langle 25, 22, 4, 3, 72, 2, 8, 6, 33, 45, 4, 48, 55, 32, 11, 44 \rangle$.

Para discutir...

- Como comparar Google x Altavista ??
- Problemas com métricas de avaliação
 - Subjetividade e contexto
 - Uso de logs x uso de consultas específicas
 - Níveis de relevância

Testes estatísticos

- Hipótese nula: Hipótese de que dois resultados comparados são iguais
- Hipótese alternativa: Há diferenças entre os dois resultados
- Testes estatísticos servem para dizer se a hipótese nula está descartada com um certo grau de certeza (p-value), normalmente entre 95% e 99%

Testes estatísticos

- Se o teste de significância falha, isso não quer dizer que os dois sistemas produzem resultados iguais. Isso quer dizer que seu experimento não foi conclusivo!
- Pode-se ampliar número de amostras para tornar experimento mais conclusivo
- Diferença ser considerada estatisticamente significativa não significa que diferença é necessariamente grande!!!!

Testes estatísticos

- Tentam estimar se os resultados de comparação entre dois sistemas foi obtido ao acaso (por sorte) ou se há uma diferença consistente
- Podem ser aplicados sobre as diversas métricas, tais como MAP, bpref, NDCG@K...

Há críticas sobre estes testes...

- Alguns autores da área de estatística criticam o uso deste tipo de testes e sugerem o uso de intervalos de confiança ao invés de testes estatísticos

T-test

- Proposto em 1908
- Autor (William S. Gosset) usou um codinome (student) porque não podia ser identificado. Por isso, o teste é conhecido como “t-student test”
- Uso em RI:
- Utilizado para verificar se a hipótese “a média de valores de duas populações é igual” é válida ou não.

T-test

- Média das diferenças sobre desvio padrão multiplicado pela raiz do número de amostras



■ Funcionamento de Máquinas de Busca

- Evidências utilizadas no processamento de consultas
- Coletores para a Web
- Indexação e remoção de ruído dos índices

Tipos de consultas

- Informacionais
- Navegacionais
- Transacionais
- Outros tipos....

Exemplos de fontes de informação para a Web

- Reputação de páginas
- Concatenação de âncoras
- Texto
- Nível das Urls
- Texto das Urls
- Click nas respostas
- Outras possibilidades...

Reputação das páginas

- Links podem ser considerados como indicadores de qualidade para uma página
- Reputação da página na Web pode ser estimada através da estrutura de links

Alguns exemplos de métodos

- Indegree (96)
- Pagerank (97)
- HITS (98)
- Diversas variações....
- Modelo de Hipergrafos
(HiperPagerank, HiperIndegree)

Indegree

- Conta o número de apontadores para uma dada página
- Possíveis variações incluem número de domínios, hosts e etc...

Pagerank

- Tenta estimar a probabilidade de um usuário chegar em uma página durante um caminhamento aleatório

Pagerank

- *A importância de uma página P é dada pela seguinte equação:*

- $$PR(p) = (1-d) + d (PR(t_1)/c_1 + \dots + PR(t_n)/c_n)$$

d – dump factor (geralmente entre 0.1 e 0.9)

T_i – página que aponta para P

c_i – quantidade de links em T_i

- *Pagerank procura expressar a probabilidade de uma página P ser acessada.*

Pagerank

- Menos suscetível a ataques do que Indegree
- Dá mais importância a apontadores oriundos de páginas com boa reputação
- Evita que páginas com muitos apontadores tenham influencia muito alta nos resultados

HITS

- *Utiliza valores de hub e autoridade para definir a reputação de uma página P .*
 - **hub** de uma página " P "– é dado em função dos valores de autoridade das páginas para onde ela aponta.
 - **autoridade** de uma página " P "– é dada em função dos valores de hub das páginas que apontam para P .
- *Um bom hub é uma página que aponta para boas autoridades e uma boa autoridade é uma página apontada por bons hubs.*

Autoridade de P = soma dos valores de hub das que apontam para P

Hub de P = soma dos valores de autoridade das páginas que apontam para P

HITS

- Utilizado em conjunto com a informação textual
- Monta-se um grafo de vizinhança envolvendo os K documentos do topo do ranking e:
 - Seus pais, filhos, demais pais dos seus filhos e filhos dos seus pais
- Valores de HUB e autoridade são calculados sobre este grafo

HITS

- Grafo de contexto ajuda a encontrar autoridades relacionadas ao assunto da página
- Custo computacional para calcular grafo durante o processamento da consulta é alto, o que dificulta a implementação do método

Modelo de Hipergrafos

- Trabalho de mestrado da UFAM
- Modifica forma de modelar a Web
- Permite a redefinição de métodos de análise de links para este novo modelo

Modelo de hipergrafos

- Um hipergrafo é representado por um conjunto de vértices e um conjunto de hiperarestas
- Cada hiperaresta conecta um subconjunto dos vértices do grafo a um outro subconjunto
- O hipergrafo pode ser dirigido

Modelo de Hipergrafos

- Web é particionada visando agrupar páginas com alto grau de relacionamento
- Arestas entre páginas da representação tradicional da Web são substituídas por hiperarestas que conectam partições a páginas
- Uma partição aponta para uma página P se e somente se contiver pelo menos um elemento com link para P .

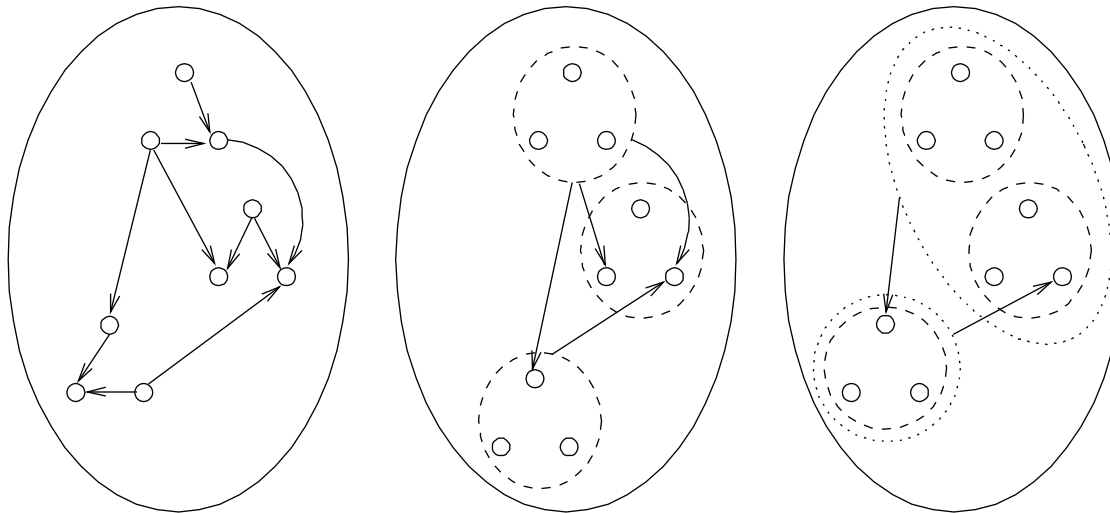
Modelo de Hipergrafos

- Particionamento visa encontrar um bom balanceamento entre número de hiperarestas e qualidade de cada hiperarestas
- Consideramos uma hiperaresta boa se a mesma conecta conjuntos de páginas “independentes”
- Estimamos a dependência através do grau de relacionamento entre as páginas na estrutura da Web.

Particionamento

- Inicialmente utilizamos heurística simples: host e domínio
- Estamos estudando heurísticas baseadas em métricas de conectividade

Modelo de Hipergrafo



- Page-based partitioning
- - - - Host-based partitioning
- Domain-based partitioning

Experimentos

- Criamos uma versão de Pagerank e de Indegree para hipergrafos
- Realizamos experimentos na Base WBR03 com consultas informacionais e navegacionais.
- Cada tipo de consulta foi dividido em popular e selecionadas aleatoriamente
- Resultados indicam ganhos para consultas navegacionais, sem perdas em informacionais

Concatenação de âncoras

- Representa cada página P com a concatenação de todos os textos utilizados em apontadores para P na Web
- Fonte de evidência textual que pode ser usada com modelos de RI similares aos de texto

Questões importantes

- Apontadores internos devem ser levados em consideração ?
- Pode ser facilmente burlado?
- Como tornar esta fonte de evidência segura ?

Propriedades interessantes

- Captura as diversas formas de descrever o conteúdo de uma página
- Serve para diminuir Gap entre vocabulário de quem formula a consulta e de quem escreve o texto
- Permite a descrição de conteúdos em páginas que não possuem texto

Nível da URL

- Atribui a cada página um nível de acordo com sua profundidade na estrutura de diretórios de um site
- Tem boa correlação com métodos para estimar reputação de páginas
- Serve como fonte de informação para distinguir páginas que não possuem reputação

Questões importantes

- O que fazer com sites que não utilizam estrutura de diretórios ?
- Agrupa páginas em poucas categorias
- Poderíamos utilizar um nível virtual em função do número de apontadores seguidos a partir da homepage do site ?

Texto das URLs

- Texto das URLs pode ser usado como fonte de informação
- Casamento com a consulta deve permitir aproximações ?
- Como utilizar esta fonte ?

Click nas páginas

- Páginas mais seguidas em respostas podem ter seu ranking melhorado
- Não há modelo padrão para o uso dessa informação
- Pode ser usado para pequenas modificações no ranking
- Deve ser usado com cuidado

Como combinar fontes ?

- Costumava-se normalizar resultados e assumir independência entre fontes
- Exemplos: Combinações lineares ou tratamento probabilístico
- Fala-se em percentual de impacto para cada fonte (não há artigos descrevendo idéia)
- Problema é mais complicado...

Problemas com combinação

- Fontes nem sempre são independentes
- Diferenças entre pesos pode não ser linear
- Combinação pode mudar de acordo com tipo de consulta
- Afeta métodos de poda
- Como utilizar evidências independentes de consulta (Pagerank, Nivel de URLs e etc...) ?

Alternativa de solução

- Utilizar algum método de combinação adhoc
- Utilização de métodos de aprendizagem automática
- Exemplos: GP, SVM, Técnicas de mineração de regras de associação

Métodos Adhoc

- Combinação linear dos graus de relevância dos documentos a uma consulta

$$SU(q,d) = P1 * A(q,d) + P2 * B(q,d) + P3 * C(q,d)$$

- Modelar cada fonte de evidência como uma probabilidade independente de relevância dos documentos

$$SU(q,d) = 1 - ((1 - A(q,d)) * (1 - B(q,d)))$$

Métodos Adhoc

- Vantagem

- Não precisam de treino

- Desvantagem

- ◆ Não utilizam propriedades das fontes de evidências para descobrir a melhor maneira de combinar
 - Dependências entre diferentes fontes de evidência
 - Variações na distribuição dos escores de cada evidência

Métodos de Aprendizagem de máquina

- Utilizam uma base de treino para “aprender” a melhor forma de combinar evidências
- Necessidade de treino é principal desvantagem
- Exemplo: GP

Combinação Utilizando GP

- Programação Genética pode ser usada para descobrir funções de combinação que:
 - ◆ aumentem a qualidade das respostas das máquinas de busca
 - ◆ utilizem as propriedades contidas nas evidências para gerar boas funções
 - ◆ generalizar a tarefa de combinação para um número qualquer de evidências de relevância
 - ◆ especificar a combinação de evidências para diferentes tipos de consultas existentes

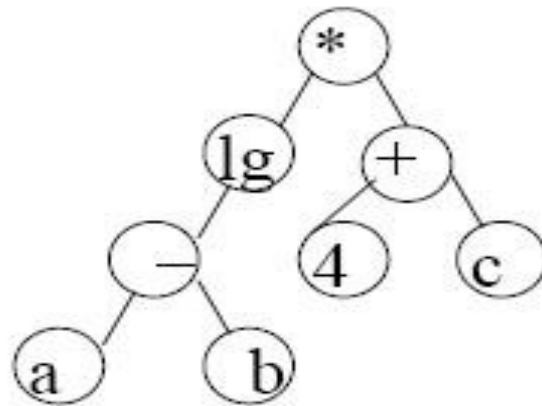
Programação Genética (PG)

- Baseado no conceito de *Seleção Natural*
- Técnica de aprendizagem automática (*machine learning*)
- Evolução de indivíduos (soluções para determinado problema)

Programação Genética

- Por que usar PG?
 - Tem sido usada com sucesso em várias aplicações, inclusive na derivação de funções de ranking.

Programação Genética



$$\text{Arvore}(a,b,c)=(\lg(a-b))*(4+c)$$

- Terminais: a,b,c e 4
- Funções: multiplicação, subtração, adição e logarítmo

PG para Combinação de Evidências

■ Entrada

■ Consultas pré-avaliadas

- ◆ Processa consulta
- ◆ Avalia as 50 primeiras respostas do topo de cada um dos ranks gerados por cada evidência

■ Fases da combinação

- Treinamento
- Validação

Configuração dos experimentos

- Evidências Utilizadas
 - Texto presente nas páginas
 - ◆ Cosseno vetorial aplicado sobre os textos
 - Textos de âncora dos documentos
 - ◆ Cosseno vetorial aplicado sobre os textos de âncora
 - Estruturas de ligação entre as páginas da Web
 - ◆ Pagerank

Configuração dos experimentos

- Classes de consultas
 - As consultas foram divididas segundo tipo de informação procurada e popularidade dos documentos
 - ◆ Consultas por tópico de informação (informacionais)
 - Populares e não-populares
 - ◆ Consultas por sites específicos (navegacionais)
 - Populares e não-populares
- Classificadas manualmente

Configuração dos experimentos

- Função de Fitness
 - Consultas informacionais: BPref
 - Consultas Navegacionais - MRR

Configuração dos experimentos

- Coleção
 - Base de dados do *TodoBr*
 - Mais de 12 milhões de páginas
- Fases dos experimentos
 - Treinamento
 - Validação
 - Teste

Configuração dos experimentos

- Informacionais

- ◆ 62 consultas extraídas do log do *TodoBr* composto por mais de 11 milhões de consultas
 - 70% utilizadas para o treinamento
 - 30% utilizadas para validação
 - Executamos 20 evoluções e escolhemos a que obteve melhor resultado na validação
 - 30 consultas para fase de teste em que foram avaliadas os 10 documentos do topo de cada evidência, inclusive as funções de combinação estudadas
- ◆ Os resultados apresentados se referem ao conjunto de teste

Configuração dos experimentos

- Navegacionais
 - 90 consultas extraídas do log
 - ◆ 50% para treino
 - ◆ 20% para validação
 - ◆ 30% para teste
- Executamos 20 evoluções e escolhemos a função que obteve melhor resultado quando aplicado a função de fitness para passar a fase de testes

Configuração dos experimentos

- Comparação de desempenho
 - Modelo de combinação de evidências por redes Bayesianas proposto por Silva
 - Modelo proposto por Craswell (baseado em treino)
 - Melhor combinação Linear (combinação linear utilizando treino)

Experimentos(1)

■ Consultas Informacionais

Query Type	Method	RankEff	Bpref-10	MAP
Informational Popular	GP	0.744	0.562	0.485
	BN	0.406	0.393	0.361
	BLC	0.731	0.563	0.312
	SIGM	0.704	0.520	0.462
Informational Non-Popular	GP	0.724	0.538	0.401
	BN	0.476	0.245	0.237
	BLC	0.648	0.365	0.312
	SIGM	0.632	0.336	0.300

Query Type	Evidence	RankEff	Bpref-10	MAP
Informational Popular	Text	0.715	0.414	0.336
	Anchor	0.632	0.348	0.259
	Pagerank	0.320	0.117	0.097
Informational Non-popular	Text	0.726	0.549	0.444
	Anchor	0.511	0.203	0.159
	Pagerank	0.317	0.190	0.134

Experimentos(2)

■ Consultas Informacionais

Query Type	Method	RankEff	Bpref-10	MAP
Informational Popular	GP	0.744	0.562	0.485
	BN	0.406	0.393	0.361
	BLC	0.731	0.563	0.312
	SIGM	0.704	0.520	0.462
Informational Non-Popular	GP	0.724	0.538	0.401
	BN	0.476	0.245	0.237
	BLC	0.648	0.365	0.312
	SIGM	0.632	0.336	0.300

Query Type	Evidence	RankEff	Bpref-10	MAP
Informational Popular	Text	0.715	0.414	0.336
	Anchor	0.632	0.348	0.259
	Pagerank	0.320	0.117	0.097
Informational Non-popular	Text	0.726	0.549	0.444
	Anchor	0.511	0.203	0.159
	Pagerank	0.317	0.190	0.134

Experimentos(3)

■ Consultas Navegacionais

Query Type	Method	MRR
Navigational Popular	GP	0.920
	BN	0.405
	BLC	0.581
	SIGM	0.479
Navigational Non-Popular	GP	0.803
	BN	0.408
	BLC	0.367
	SIGM	0.325

Query Type	Evidence	MRR
Navigational Popular	Text	0.153
	Anchor	0.178
	Pagerank	0.266
Navigational Non-popular	Text	0.209
	Anchor	0.364
	Pagerank	0.178

Experimentos(4)

- Consultas navegacionais

Query Type	Method	MRR
Navigational Popular	GP	0.920
	BN	0.405
	BLC	0.581
	SIGM	0.479
Navigational Non-Popular	GP	0.803
	BN	0.408
	BLC	0.367
	SIGM	0.325

Query Type	Evidence	MRR
Navigational Popular	Text	0.153
	Anchor	0.178
	Pagerank	0.266
Navigational Non-popular	Text	0.209
	Anchor	0.364
	Pagerank	0.178

Análise das Fórmulas

t-texto; a-âncora; p-pagerank

■ Informacionais Populares

$$Comb(a, p, t) = \begin{cases} t(3t + 2tp^2 + 2ap^2 + 6a + 4p) + 8ap + \\ p \ln(p)(6t^2p^2 + 2tap^2 + 9t + 3ta + 12tp + 4ap) \end{cases}$$

Informacionais não-populares

$$Comb(a, p, t) = t^5(2p + 6t^2p + 4tp + 3tp + 2p^2 + 6t^3 + 6t^5 + 4t^4 + 3t^4p + 2t^3p)$$

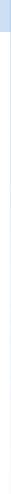
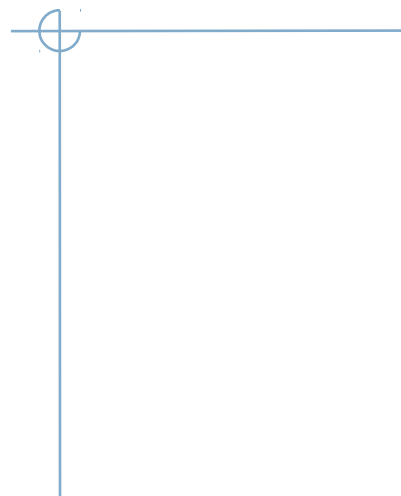
Análise das Fórmulas

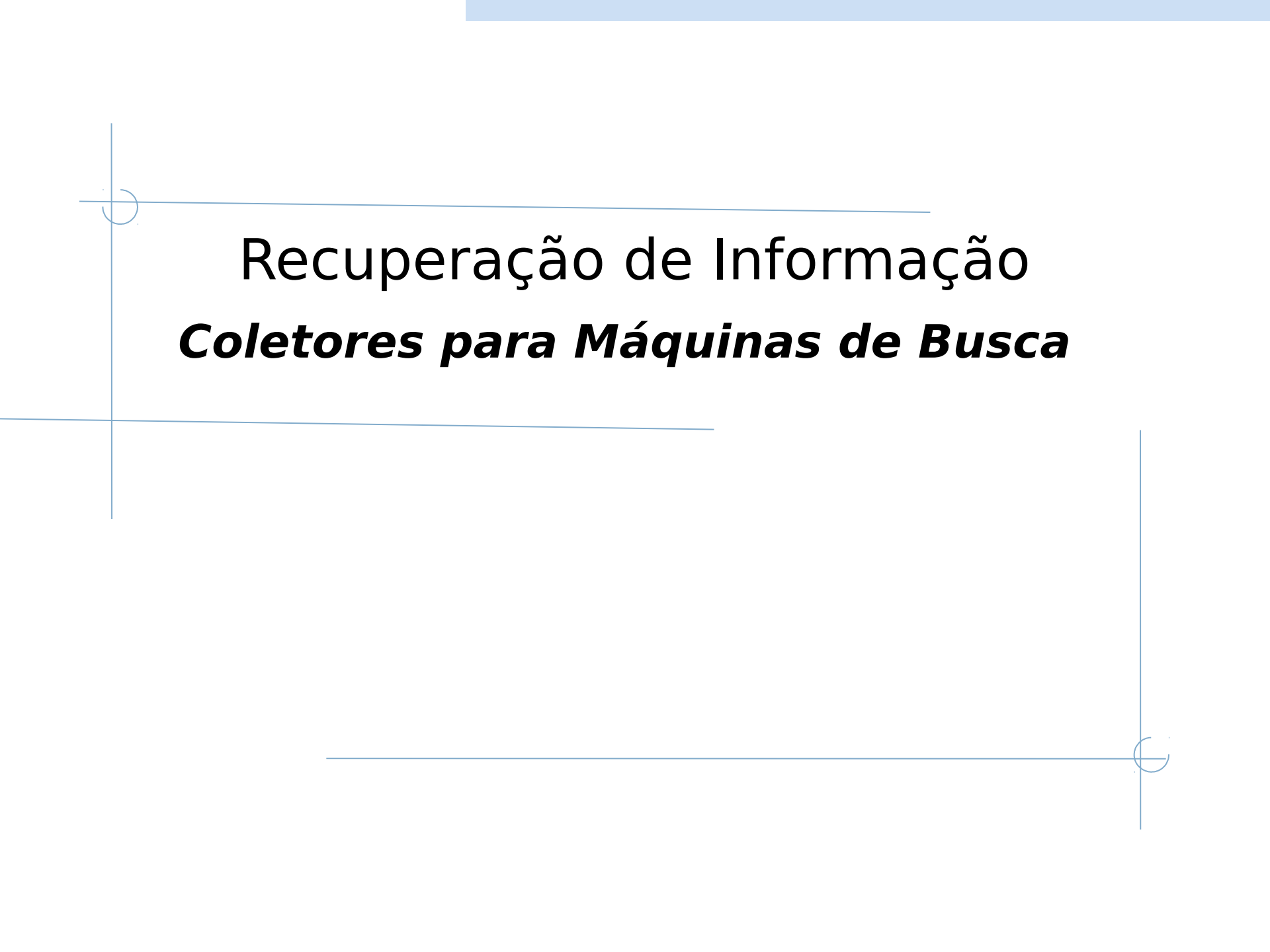
- Navegacionais populares

$$Comb(a, p, t) = \begin{cases} ap^2(8 + 7p - p^2 - a - \\ ap \ln(a)(8p - p^2 - \ln(p - 8 \ln(p - 8) + a \ln(p - 8)) - ap + 6)) \\ + ap(6 - \ln(p - 8 \ln(p - 8) + a \ln(p - 8)))) \end{cases}$$

- Navegacionais não-populares

$$Comb(a, p, t) = a^3 p(2t + p)$$





Recuperação de Informação

Coletores para Máquinas de Busca

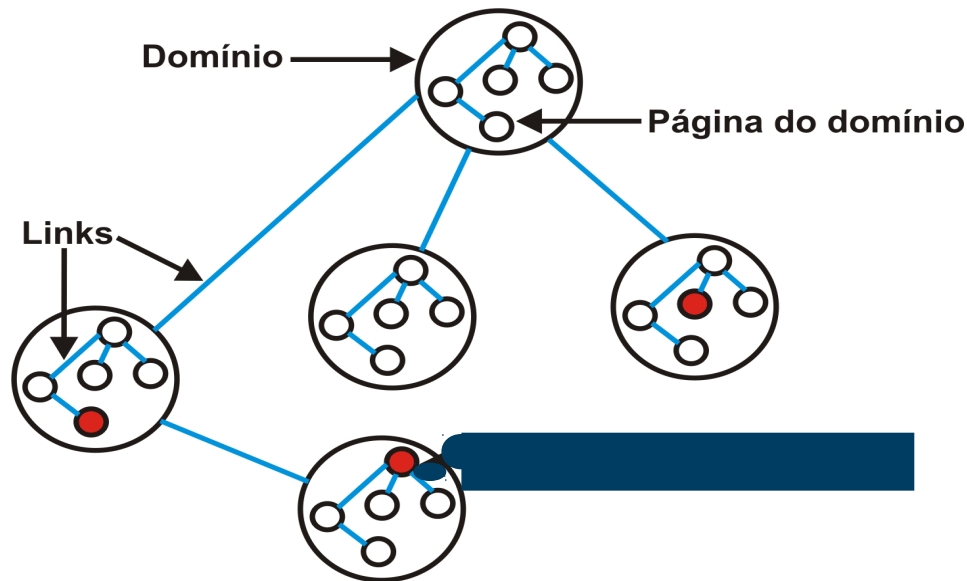
Roteiro



- Coletores
- Métricas de Importância
- Desafios de Projeto

Coletores

- Processo de se navegar entre páginas www usando estrutura de hyperlinks.
- A estrutura de hyperlinks da web pode ser modelada como um grafo.



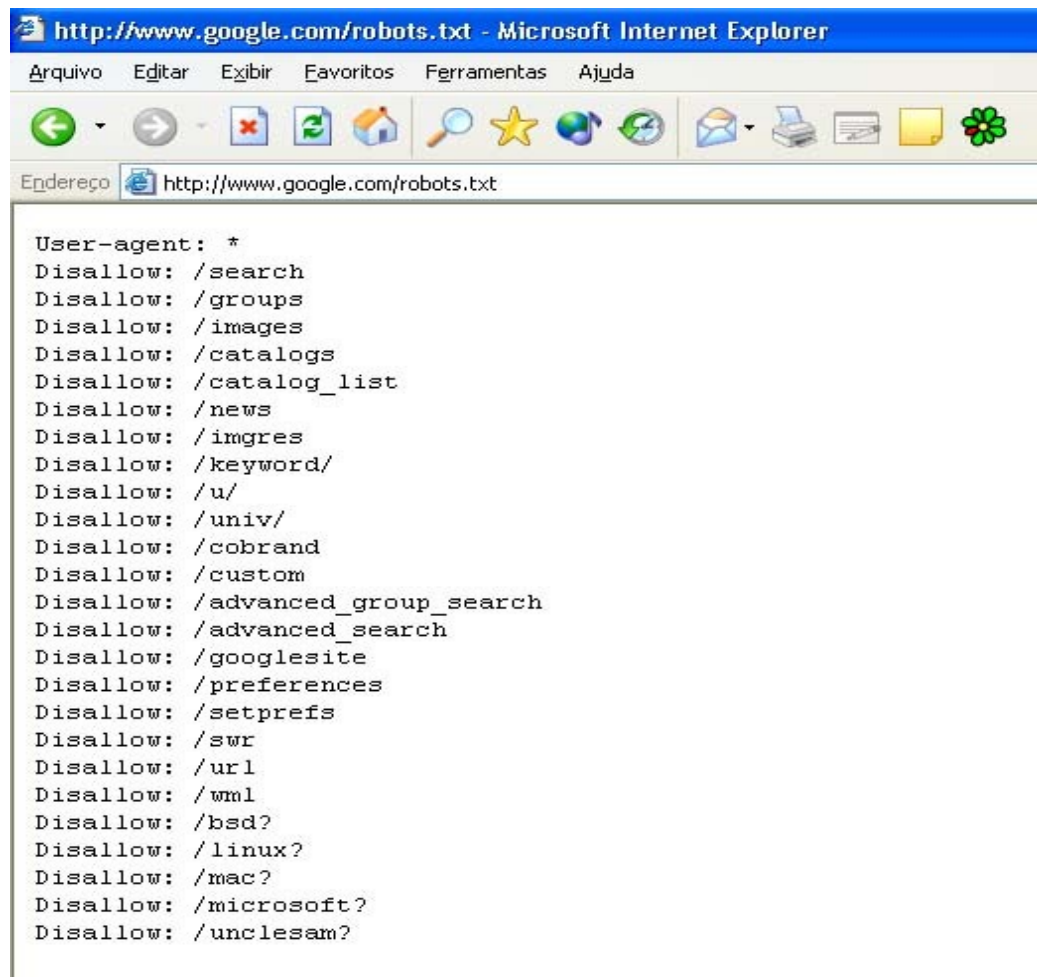
Coletores

- Navegadores automáticos entre páginas web que visam armazenar uma cópia local das páginas encontradas
- Devem obedecer algumas restrições ao visitar sites.
- O Protocolo de Exclusão de Robôs (Robot Exclusion Protocol) especifica algumas regras de acesso.
 - Principal regra é não deixar um intervalo de tempo entre acessos a cada servidor.

Coletores

- O Protocolo de Exclusão de Robôs:
 - Padrão definido em 30 de junho de 1994.
 - Define as permissões do *coletor* em um determinado site.
 - As diretrizes são descritas em um arquivo chamado “robots.txt”, localizado no servidor web coletado.

Exemplo de robots.txt

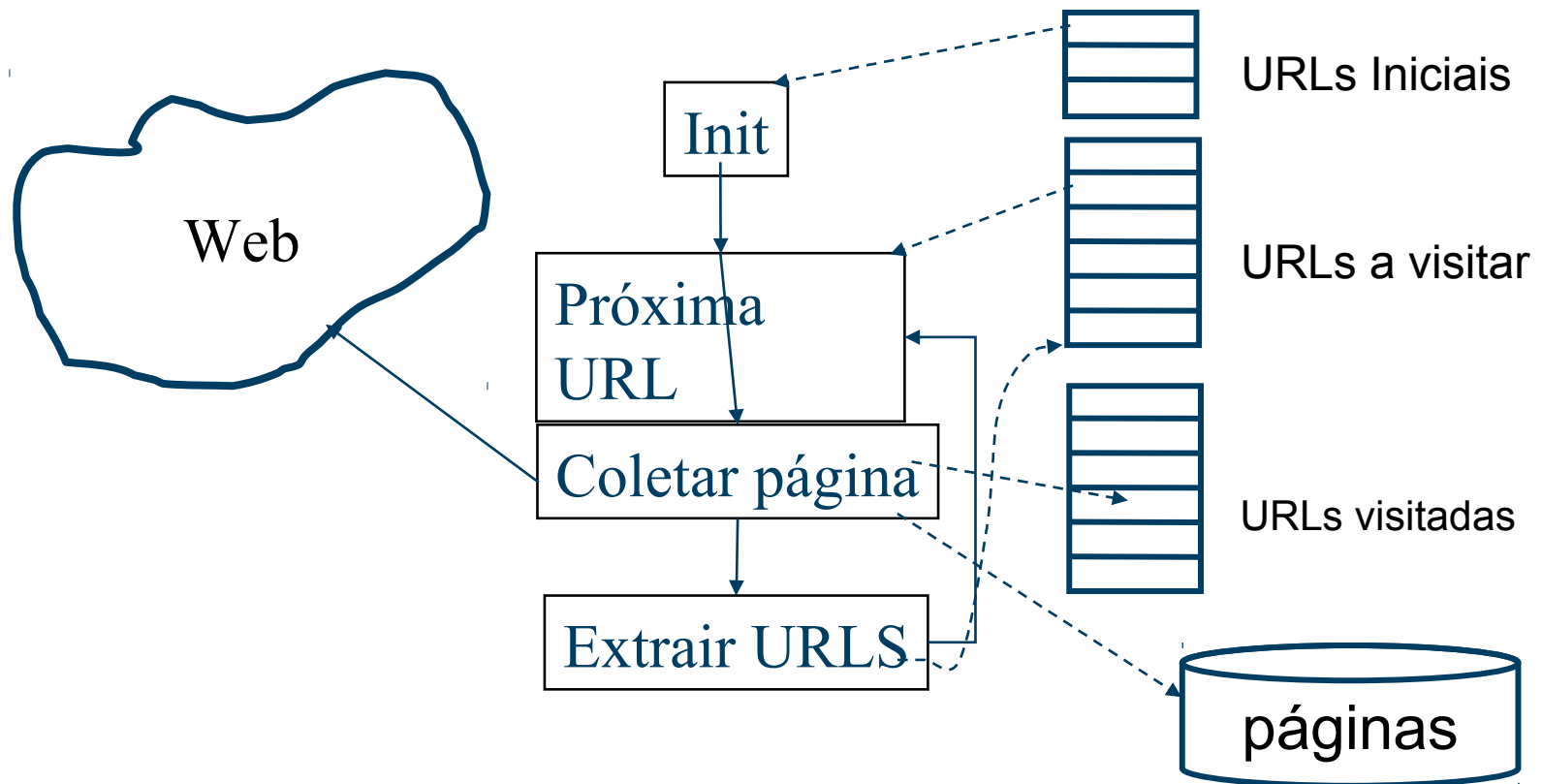


The screenshot shows a Microsoft Internet Explorer browser window. The title bar reads "http://www.google.com/robots.txt - Microsoft Internet Explorer". The menu bar includes "Arquivo", "Editar", "Exibir", "Favoritos", "Ferramentas", and "Ajuda". The toolbar contains various icons for navigation and actions. The address bar shows "Endereço http://www.google.com/robots.txt". The main content area displays the text of the robots.txt file.

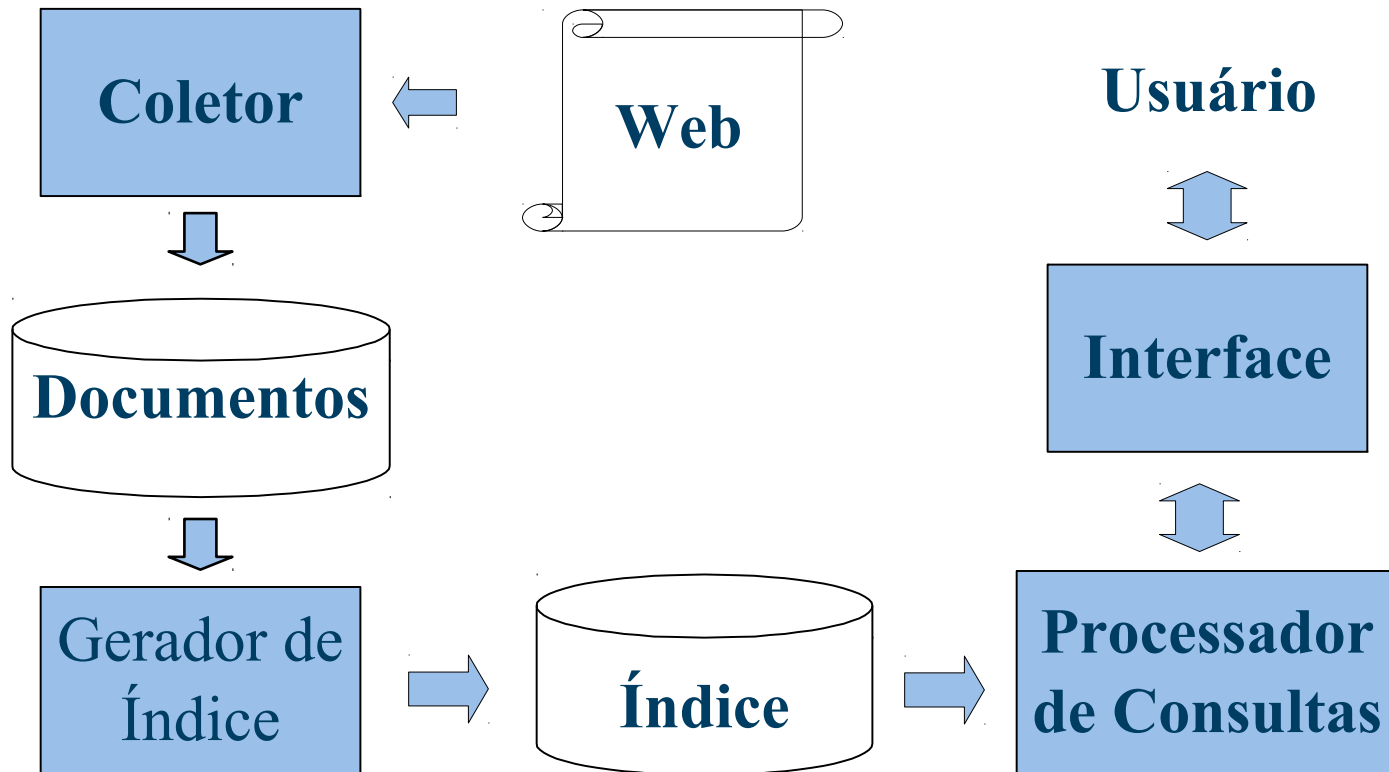
```
User-agent: *
Disallow: /search
Disallow: /groups
Disallow: /images
Disallow: /catalogs
Disallow: /catalog_list
Disallow: /news
Disallow: /imgres
Disallow: /keyword/
Disallow: /u/
Disallow: /univ/
Disallow: /cobrand
Disallow: /custom
Disallow: /advanced_group_search
Disallow: /advanced_search
Disallow: /googlesite
Disallow: /preferences
Disallow: /setprefs
Disallow: /swr
Disallow: /url
Disallow: /wml
Disallow: /bsd?
Disallow: /linux?
Disallow: /mac?
Disallow: /microsoft?
Disallow: /unclesam?
```

Coletor

- Esquema gráfico do funcionamento de um coletor.



Arquitetura da MB



Escalonamento

- Páginas devem ser coletadas seguindo métricas de importância
- Métricas devem definir prioridade de coleta e refrescamento de página
- Objetivo final deve ser maximizar a qualidade das respostas providas pela máquina de busca

Métricas de Importância

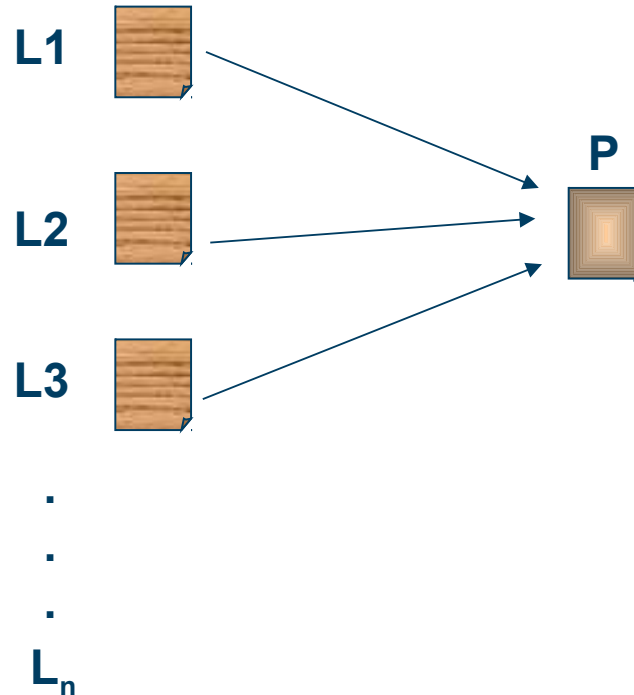
- Similaridade
- Backlink Count
- Page Rank
- HITS
- Foward Link Count

Similaridade

- Consultas são utilizadas para determinar importância das páginas.
- Algum modelo de RI é utilizado para computar a similaridade entre uma dada consulta Q e um determinado documento P (página web).

Backlink Count

- A importância de uma página “P” é definida pela quantidade de links que apontam para a mesma.



PageRank

- *A importância de uma página P é dada pela seguinte equação:*

- $$IR(P) = (1-d) + d (IR(T1)/c1 + ... + IR(Tn)/cn)$$

d – dump factor (geralmente entre 0.1 e 0.9)

T_i – página que aponta para P

c_i – quantidade de links em T_i

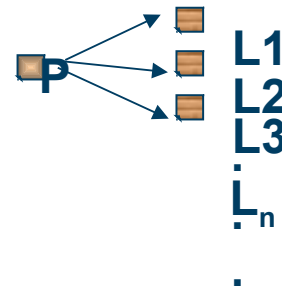
- *Page Rank procura expressar a probabilidade de uma página P ser acessada.*

HITS

- *Utiliza valores de hub e autoridade para definir a importância de uma página P.*
 - **hub** de uma página “P”– é dado em função dos valores de autoridade das páginas para onde ela aponta.
 - **autoridade** de uma página “P”– é dada em função dos valores de hub das páginas que apontam para P.
- *Um bom hub é uma página que aponta para boas autoridades e uma boa autoridade é uma página apontada por bons hubs.*

Forward Link Count

- A importância de uma página P é determinada pela quantidade de links existentes na mesma.
- Todos os links possuem peso 1, mas poderiam ser considerados pesos diferentes, de acordo com a importância do link.



Tipos de escalonamento

- Offline: Fila de prioridades para coleta é ordenada periodicamente fora do sistema de coleta
- Online: Fila de prioridades é reordenada continuamente

Desafios de Projeto

- Desafios ao se projetar um coletor:
 - Definir a periodicidade de atualização das páginas (Freshness) X Encontrar novas url's.
 - Usar o máximo de largura de banda sem sobrecarregar os *sites* visitados.
 - Identificar páginas redundantes (Mirror).
 - Coletar boas páginas.

Problemas práticos

- Sobrecarga do DNS
- Erros de acesso repetido a servidores
- Extração de links pode gerar URLs falsas ou links infinitos
- Coleta de páginas dinâmicas
- Normalização de URLs
- Diferença de velocidade entre servidores pode afetar velocidade
- ... (há muitos outros problemas)

Sobrecarga de DNS

- Coletores geram número muito alto de requisições de DNS
- Normalmente os servidores de DNS viram gargalos para o coletor
- Solução é manter um cache com DNSs previamente resolvidos

Erros de acesso repetido (falsos ataques)

- O coletor pode criar um falso ataque a um servidor Web por problemas como:
 - Uso de diferentes nomes para um servidor
 - Vários servidores em um mesmo local

Links infinitos

- Problemas na extração de links podem gerar erros que levam a links infinitos que são validados pelo servidor Web
- www.aa.bb.com/musica
- www.aa.bb.com/musica/musica

Páginas dinâmicas

- Alguns sites podem gerar número infinitos de páginas válidas
- Exemplo, um site que fornece um HTML com o dia da semana de qualquer data, onde a data entra na URL

Normalização de URLs

- URLs devem ser normalizadas para evitar repetições:
- <http://www.dcc.ufam.edu.br/~edleno>
- <http://www.dcc.ufam.edu.br/~edleno/index.htm>
- <http://www.dcc.ufam.edu.br/~edleno/>

Diferenças de velocidade

- Servidores mais lentos podem prejudicar processo de coleta por travar robôs

INDEXAÇÃO

O que fazer além das listas
invertidas ?

Alguns índices utilizados...

- Arquivo invertido de textos
- Arquivo invertido de concatenação de âncora
- Índice de inlinks/outlinks
- Pagerank (ou hiperpagerank 😊)
- Índice das URLs

Como gerar índice de links?

- Extrair links de cada página coletada para obter links de saída
- Inverter índice para computar links de entrada
- Fazer o mesmo com os textos de âncora
- Processo muito caro e de difícil paralelização

The slide features a minimalist design with light blue horizontal and vertical lines. A small blue circle is positioned at the top-left intersection of the lines, and another is at the bottom-right intersection. The title is centered in a large, black, sans-serif font.

Ruído em máquinas de Busca

Ruído

- Informação que atrapalha o funcionamento de Máquinas de busca
- Pode ser tanto intencional quanto involuntário.

Alguns Tipos de Ruído

- Páginas duplicadas
- Sites duplicados
- Links Tendenciosos (ou nepotísticos)
- SPAM

Efeitos prejudiciais

- Causam desperdício: aumento no custo de processamento, armazenamento e coleta de informação
- Degradam qualidade das respostas

Replicação de Sites

Paper Aceito no DKE

Paper publicado no SBBD 06

Alunos: André e Klessius

Professores: Altigran e Edleno

Um exemplo prático



The image is a screenshot of a Google search results page for the query "banco do brasil". At the top, the Google logo is on the left, and the search bar contains the text "banco do brasil". To the right of the search bar is a "Pesquisar" button and links for "Pesquisa avançada" and "Preferências". Below the search bar, there are radio buttons for "a web" (selected), "páginas em português", and "páginas do Brasil".

The search results are listed under the heading "Web". The first result is from bb.com.br, titled "... Promoção Prêmios Mil. O Banco do Brasil tem motivos a mais para você investir em Poupança. Veja os números sorteados. Oferta Pública de Ações Grendene. ...". It includes the URL www.bb.com.br/, a size of 45k, and a date of 4 nov. 2004. Below this result is another snippet from the same source, but with a different URL: www.bb.com.br/appbb/portal/index.jsp.

The second result is from www.bcb.gov.br, titled "Banco Central do Brasil". It provides information about the central bank and includes the URL www.bcb.gov.br/, a size of 61k, and a date of 4 nov. 2004.

The third result is from www.bancodobrasil.com.br, titled "[bb.com.br]". It includes the URL www.bancodobrasil.com.br/, a size of 46k, and a date of 4 nov. 2004.

The fourth result is from www.bnb.gov.br, titled "[Banco do Nordeste] - O Nosso Negócio é o Desenvolvimento". It includes the URL www.bnb.gov.br/, a size of 77k, and a date of 4 nov. 2004.

The fifth result is from <https://www2.bancobrasil.com.br>, titled "Mantenha sua senha em sigilo". It includes the URL <https://www2.bancobrasil.com.br/aapf/aai/login.pbk>, a size of 32k, and a date of 4 nov. 2004. Below this result is another snippet from the same source, titled "Verifique um pequeno cadeado fechado na parte inferior do ...", with the URL <https://www2.bancobrasil.com.br/pbank/index1.asp>.

The sixth result is from www.banco.com.br, titled "Welcome to Banco do Brasil - JAPAN". It includes the URL www.banco.com.br/, a size of 45k, and a date of 4 nov. 2004.

Método NormPaths

- Utilizamos a estrutura de diretórios dos sítios e o conteúdo das páginas para detectar replicação
- Consideramos que sítios que contêm páginas de conteúdo textual e caminhos idênticos são candidatos a réplicas.

Método NormPaths (2)

- Como comparar o conteúdo textual é custoso, utilizamos a norma dos documentos.
- Na falta da norma, qualquer outra forma de assinatura poderia ser utilizada.

Experimentos

- Para os experimentos, foi utilizada uma base de dados do TodoBR (10.077.722 páginas da Web Brasileira).

Resultados

Method	# replicas	Precision(%)	Recall(%)	F(%)
NormPaths	10868	54.34	71.67	61.81
Paths	7381	36.90	48.68	41.98
IPs	1753	9.17	12.09	10.43

Trabalhos Futuros

- Encontrar réplicas parciais de sites
- Criar clusters de réplicas

Ruído em Links

Paper WWW 06

Alunos: André e Paul

Professores: Edleno, Pavel e
Wolfgang

Ruidos em Links

- A intuição por trás da análise de links é que os links entre as páginas representam uma espécie de voto de confiança.
- Links ruidosos não têm este propósito
- Vários fatores acabam favorecendo o aparecimento de vários Links ruidosos.
- Spam, Troca de Links, Links Nepotísticos

Nossa abordagem

- Propusemos várias técnicas de detecção de links ruidosos.
- Estudamos o impacto da remoção dos mesmos no PageRank.
- Principal diferencial foi a análise de links entre sites ao invés de páginas

Nossos Métodos

- Mutual Site Reinforcement – trocas de links entre pares de sites
- Site Level Abnormal Support – Excesso de links de um site a outro
- Site Level Link Alliances – dependência entre os sites que apontam para um outro site indica ruído

Experimentos

- Posição Média e MRR dos relevantes em consultas navegacionais
- Precisão e Revocação nas consultas informacionais

Resultados

- Ganhos de 26% de MRR em consultas navegacionais populares
- Ganhos de 20% de MRR em consultas navegacionais selecionadas aleatoriamente
- Ganhos marginais em consultas informacionais

Futuro

- Pesos ao invés de corte
- Experimentos em outras bases de dados.
- Uso de métodos para melhorar detecção de réplicas

Métodos de Poda

Paper WWW 05/TOIS 2008

Alunos: Bruno e Célia

Professores: Edleno, Pavel, Altigran e Mário

Poda baseada em localidade

- Uso de localidade de ocorrência de termos em documentos como heurística para guiar poda
- Métodos simples que permitem a geração direta do índice com poda
 - Ex: Tomar apenas as primeiras sentenças do texto

Resultados

- Redução no tempo de indexação e ganhos na qualidade da poda quando comparados a métodos de poda estática previamente propostos
- Redução em até 50% do índice sem perda na qualidade das respostas

Pesquisas para melhorar a eficiência

- Compressão de Dados:
 - Compressão de textos;
 - Compressão de índices;
 - Busca em textos comprimidos;
- Novos Modelos e Algoritmos:
 - Formas eficientes para calcular a similaridade entre elementos no modelo vetorial;
 - Novos modelos para melhorar qualidade sem aumentar significativamente os

Pesquisas para melhorar a eficiência

- Processamento paralelo e distribuído:
 - Construção de índices e processamento de consultas utilizando ambientes distribuídos;
- Cache no processamento de consultas.

Avaliação em sistemas de Filtragem

- Utiliza medidas de precisão e revocação
- O conjunto de relevantes é determinado pelos docs que deveriam ser enviados a cada usuário
- Respostas dão lugar aos documentos que efetivamente são mostrados

Exemplo: Compare A e B

- De 10 documentos recebidos pelo sistema, os documentos {1,2,3,8,10} deveriam ter sido enviados a um usuário.
 - O sistema de filtragem A enviou apenas {1,2}
 - O sistema B enviou {1,2,6,8,10}
 - Calcule a precisão, revocação e a medida F para este exemplo.

Avaliação em classificação

- Realizada de forma similar a da filtragem

Coleções de Referência

- As principais coleções de referência existentes na literatura são:
- TREC, CACM, ISI, CFC (Cystic Fibrosis Collection), MED e NLM
- Porém há muitas outras disponíveis

TREC

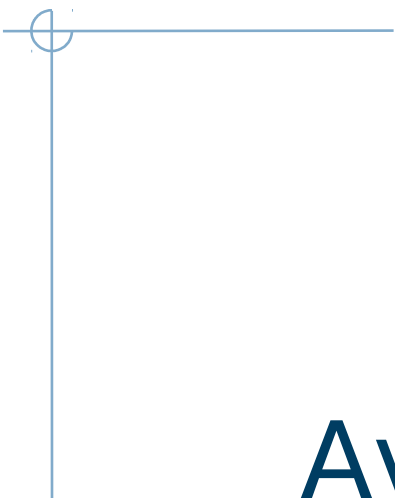


Pesquisas para melhorar a eficácia

- Extrair dados da estrutura das páginas:
 - Metatags, fontes, títulos, figuras e etc...
- Extrair dados da estrutura de apontadores (links):
 - Catálogos e Autoridades;
 - Textos contidos nos apontadores;

Pesquisas para melhorar a eficácia

- Extrair informação da coletividade:
 - Filtragem cooperativa;
 - Uso dos registros de acesso para alterar ordenação de respostas em sistemas de busca;



Avaliação de Sistemas de Filtragem de Informação

Avaliação de Sistemas de Filtragem

- Quase igual a avaliação realizada em sistema de busca.
- Somente alguns documentos são inseridos nas respostas e não há ordenação das respostas.

Avaliação de Sistemas de Filtragem

- Especialistas recebem documentos a serem filtrados.
- Conjunto de documentos filtrados pelo sistema (R) é comparado com o conjunto de documentos filtrados pelos especialistas (N).
- Obtém-se então um valor de precisão e revocação.
- Há outras formas de avaliação.



Avaliação de Sistemas de Classificação de Informação

Avaliação de Sistemas de Classificação

- Coleção de referência é classificada.
- Depois calcula-se o percentual dos documentos classificados corretamente e erroneamente.

GVSM(Exemplo de coleção)

	d1	d2	d3	d4	d5	d6
K1	1	0	1	0	1	0
K2	0	1	1	1	1	1
K3	0	0	0	1	0	0