
Introdução à Hipermídia

Vanessa de Paula Braganholo
{braganholo@dcc.ufrj.br}

Roteiro

- **Introdução à Hipermídia**
- Introdução a XML
- Esquemas para XML
- Apresentação e transformação em XML
- Linguagens de Consulta a XML
- Processamento de dados XML: APIS SAX e DOM
- Navegação em XML
- Armazenamento de dados XML
- Web Semântica

Material sobre a aula de hoje

1. **State of the Art Review on Hypermedia Issues And Applications**, V. Balasubramanian

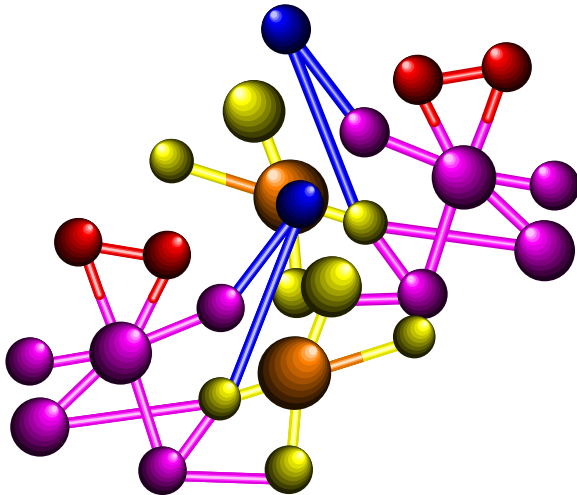
- Disponível em

<http://citeseer.ist.psu.edu/cache/papers/cs/15808/http:zSzzSzwww.dsi.unive.itzSz~simmzSzdocszSzbalasubramanian94.pdf/balasubramanian94state.pdf>

2. **A Web Odyssey: From Codd to XML**, V. Vianu, Proc. 20th ACM Symp. on Principles of Database Systems (PODS), 2001.

- Disponível em <http://www.db.ucsd.edu/CSE291S01/invited.pdf>

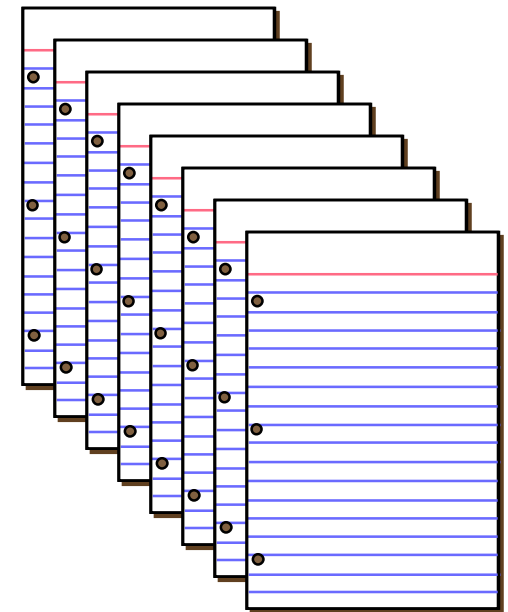
Origens



O pensamento é
multidimensional

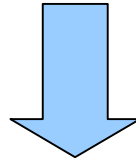
mas

o texto convencional é
linear



Sistemas de Hipermídia

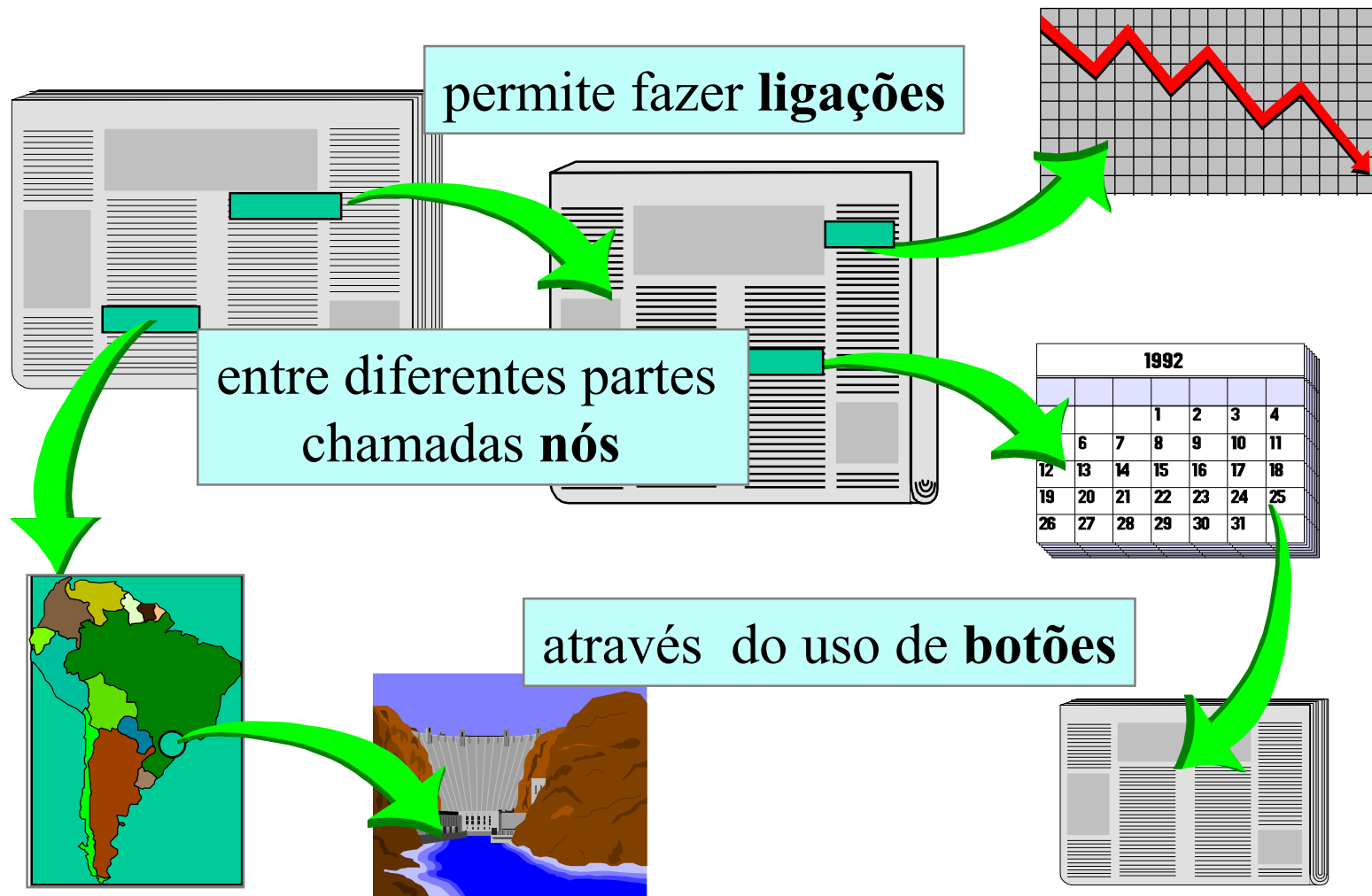
Idéia de textos não lineares ou não sequenciais



Documentos consistem de uma coleção de nós conectados por ligações direcionadas

É um sistema que manipula um conjunto de informações, pertencendo a vários tipos de mídia (texto, som, imagem, etc.), podendo estas informações serem lidas de forma não-linear através dos diversos caminhos de acesso disponíveis.

O Conceito de Hipermídia



O Conceito de Hipermídia

- É uma maneira de gerenciamento e acesso a informações, onde a **navegação** é a principal característica
- É também um esquema de representação, uma forma de rede semântica

Principais Características de um Sistema de Hipermídia

- Modularidade e associação baseada no *conteúdo*
- Flexibilidade de expansão
- Interface amigável
- Apresentação dinâmica
- **O leitor pode escolher diferentes sucessores para uma idéia**

Hipertextos e seus princípios (Lévy)

- Princípio de Metamorfose
 - ✓ a rede hipertextual está em constante construção e renegociação
- Princípio de Heterogeneidade
 - ✓ os nós e conexões de uma rede hipertextual são heterogêneos
- Princípio de Multiplicidade e de Encaixe das Escalas
 - ✓ organização “fractal”: cada nó, quando analisado, pode revelar-se como sendo composto por toda uma rede

Hipertextos e seus princípios (Lévy)

- **Princípio de Exterioridade**
 - ✓ a rede não possui unidade orgânica, nem motor interno: seu crescimento e composição dependem de um exterior indeterminado
- **Princípio de Topologia**
 - ✓ nos hipertextos, tudo funciona por proximidade, por vizinhança: O curso dos acontecimentos é uma questão de caminhos
- **Princípio de Mobilidade dos Centros**
 - ✓ a rede não tem centro: possui diversos centros, pontas móveis com ramificações

Funcionalidades

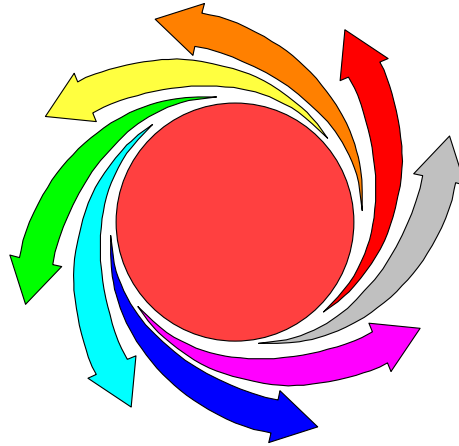
- **Trilhas:** são seqüências de nós que foram percorridas pelo usuário durante uma sessão de navegação aleatória no hiperdocumento
- **Excursões:** são trilhas pré-definidas
- **Mapas:** provê a visualização do conjunto de nós e os relacionamentos existentes entre eles
- **Visões:** permitem estabelecer o contexto sob o qual o leitor irá "ver" o hiperdocumento
- **Versões:** capacidade de preservar as diversas edições históricas de criação do hiperdocumento
- **Marcadores:** (*bookmarks*)
- **Anotações**

Sistemas Comerciais

- Hypercard (Bill Atkinson, Apple)
- Guide (Owl)
- Folio Views (Folio Corporation)
- Toolbook (Asymetrix)
- Lotus Notes
- Ferramentas da Web

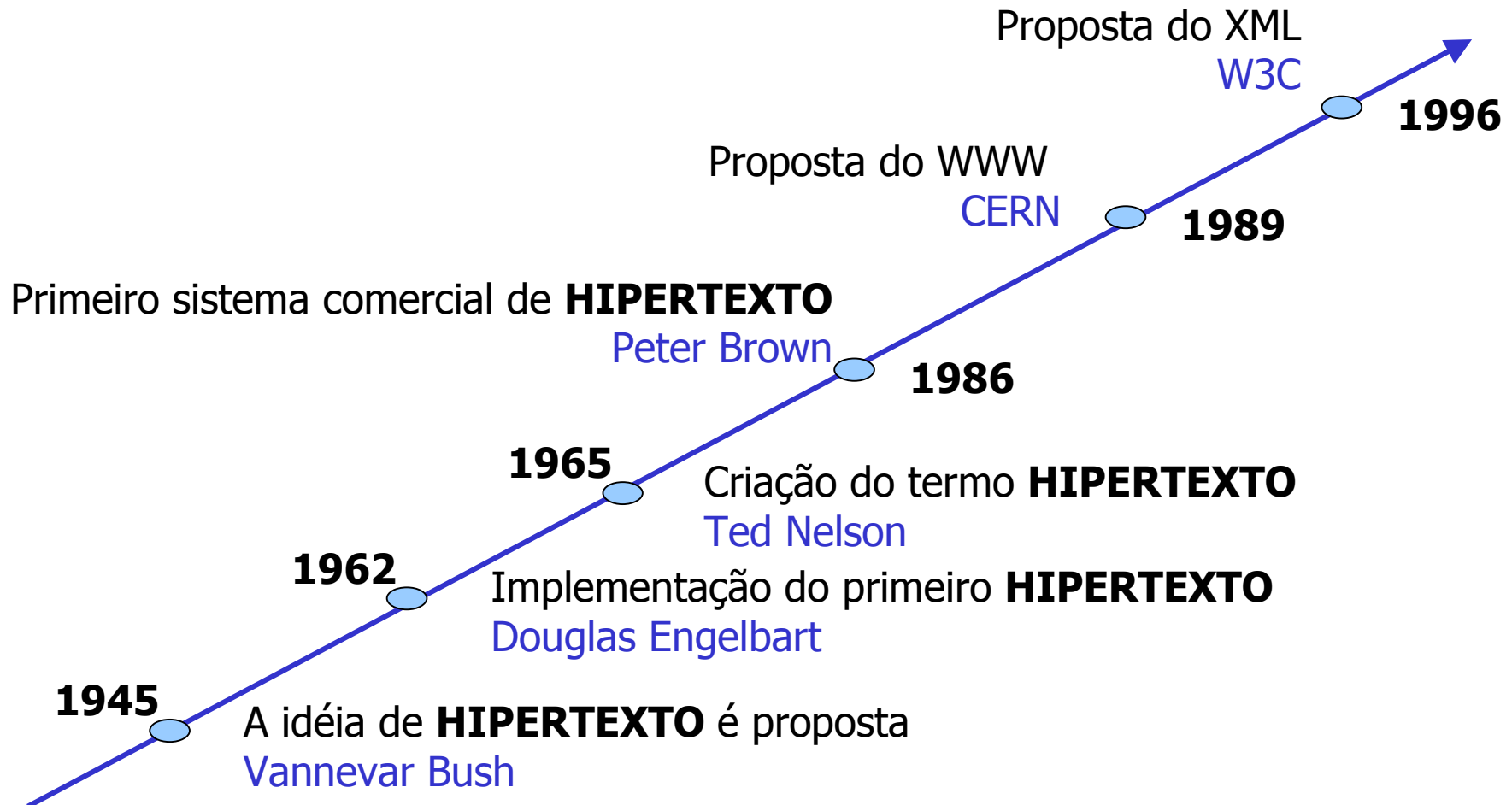
Situação até a chegada da Web

- Dezenas de Sistemas de Hipertextos Comerciais
 - ✓ ambientes diversos
 - ✓ funcionalidade variada
- Produção artesanal
 - ✓ quase nenhum método
 - ✓ processo demorado
- Milhares de hiperdocumentos gerados
 - ✓ uso específico
 - ✓ aplicações construídas do “zero”

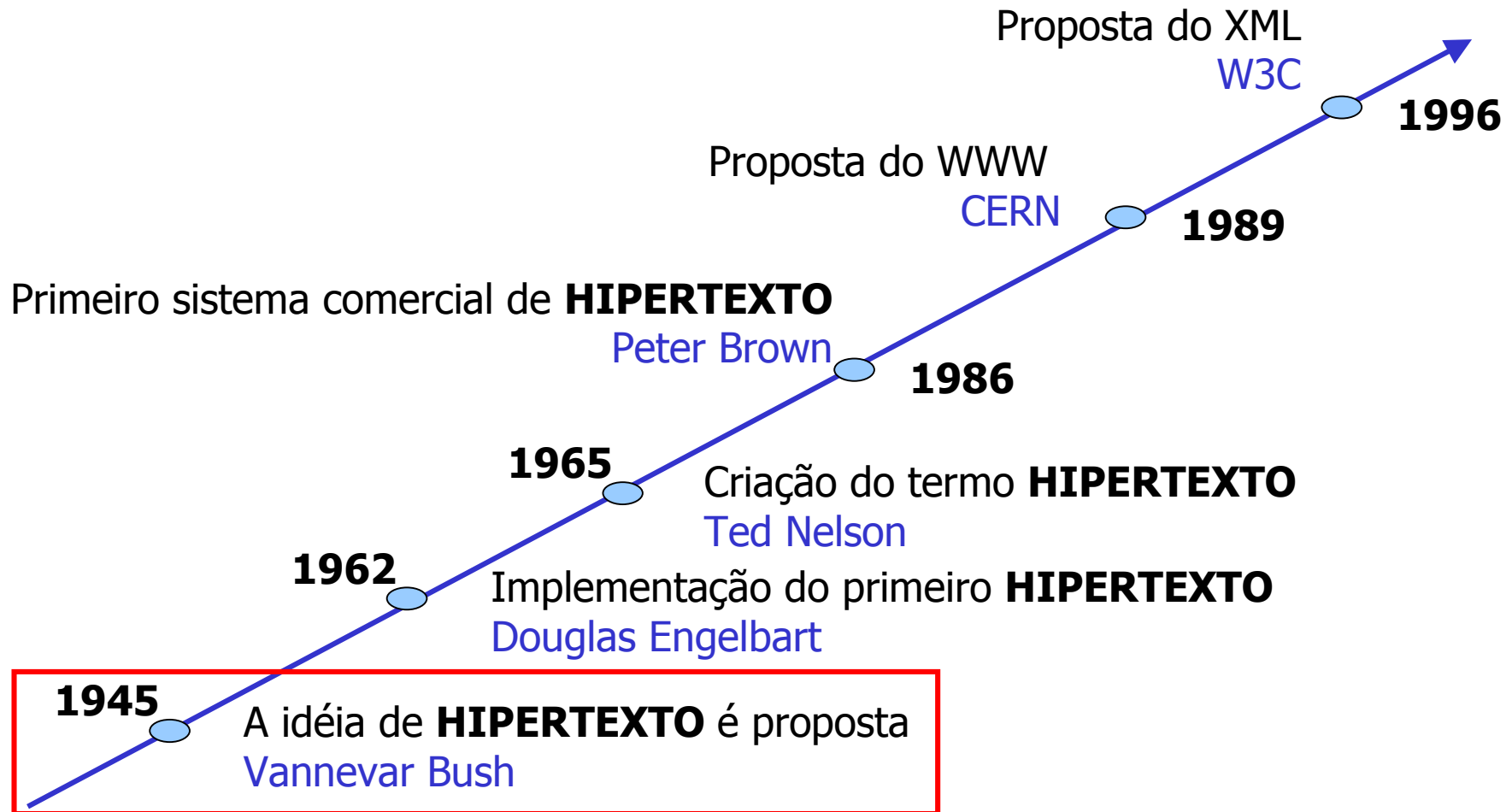


PERSPECTIVA HISTÓRICA

Uma Breve História



Uma Breve História



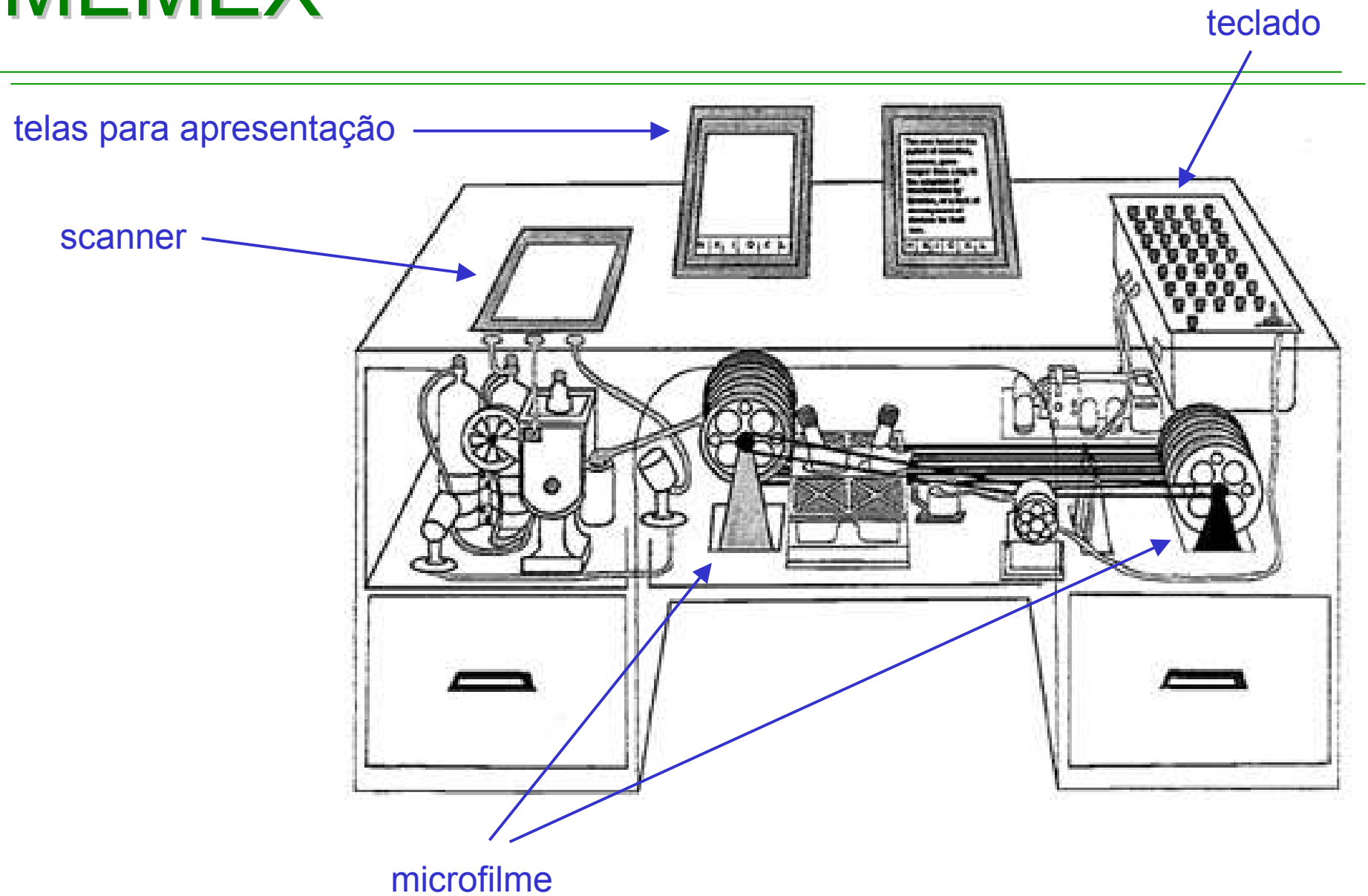
Bush - O Conceito de Hipertexto

- Cenário:
 - ✓ Pós-Guerra (1945)
 - ✓ Muitos documentos para gerenciar e analisar
 - ✓ Decisões estratégicas precisavam ser tomadas de forma rápida, mas era impossível devido ao volume de documentos (mapas, fotos, etc.)
- Proposta
 - ✓ “As We May Think” - 1945
 - ✓ Biblioteca e arquivo mecanizado privado
 - ✓ MEMEX (“memory extension”) - para funcionar como um suplemento de memória
 - ✓ Objetivo: armazenar registros, livros, comunicações, etc.

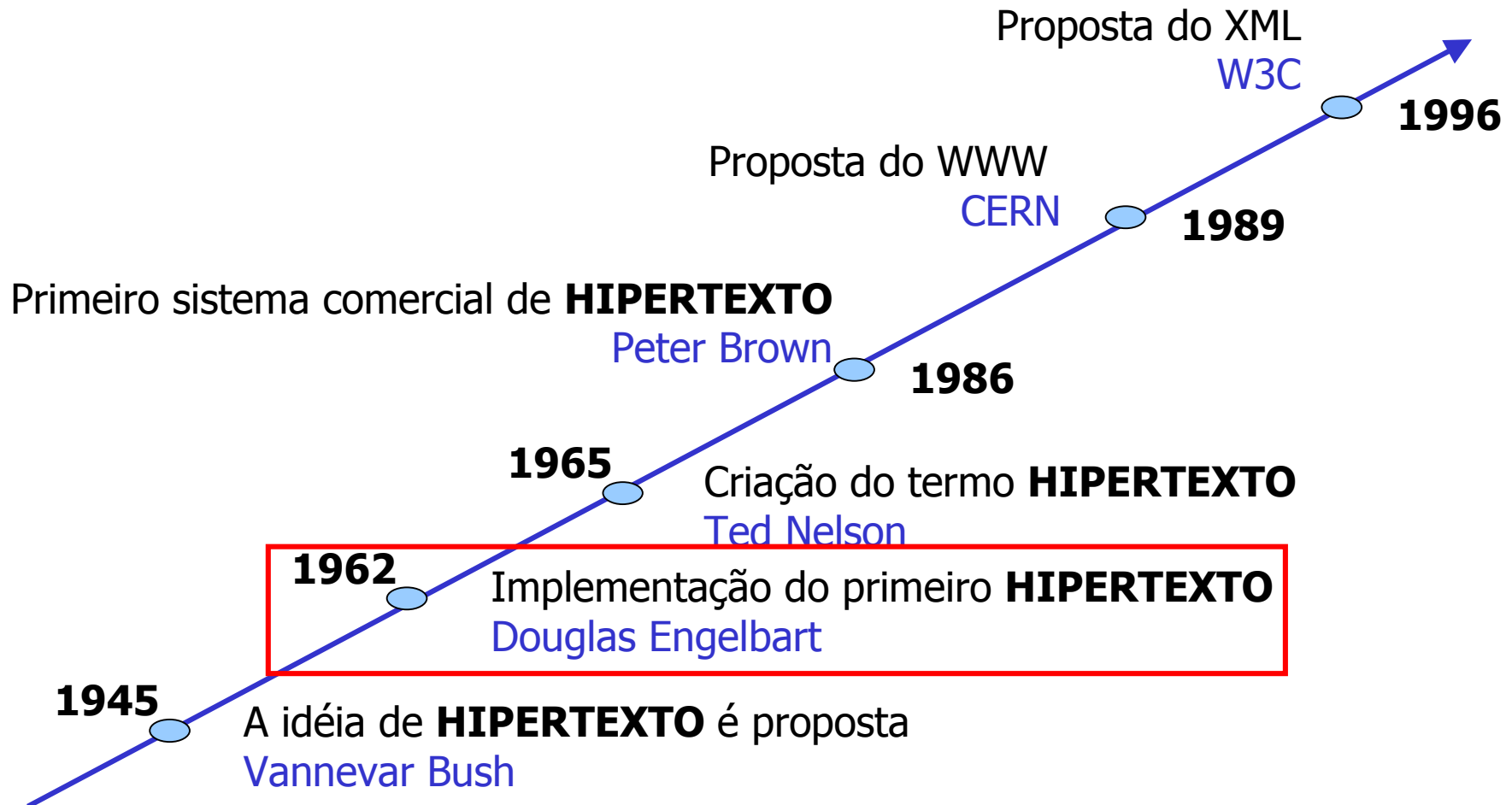
MEMEX - Características

- Era como uma mesa de trabalho comum, contendo mais um teclado e 2 telas para apresentação
- Entrada de dados: scanning (fotografia)
- Armazenamento: microfilme
- **Adição de ligações individualizadas**
- “browsing” rápido, indexação
- **usuário pode criar excursões**

MEMEX



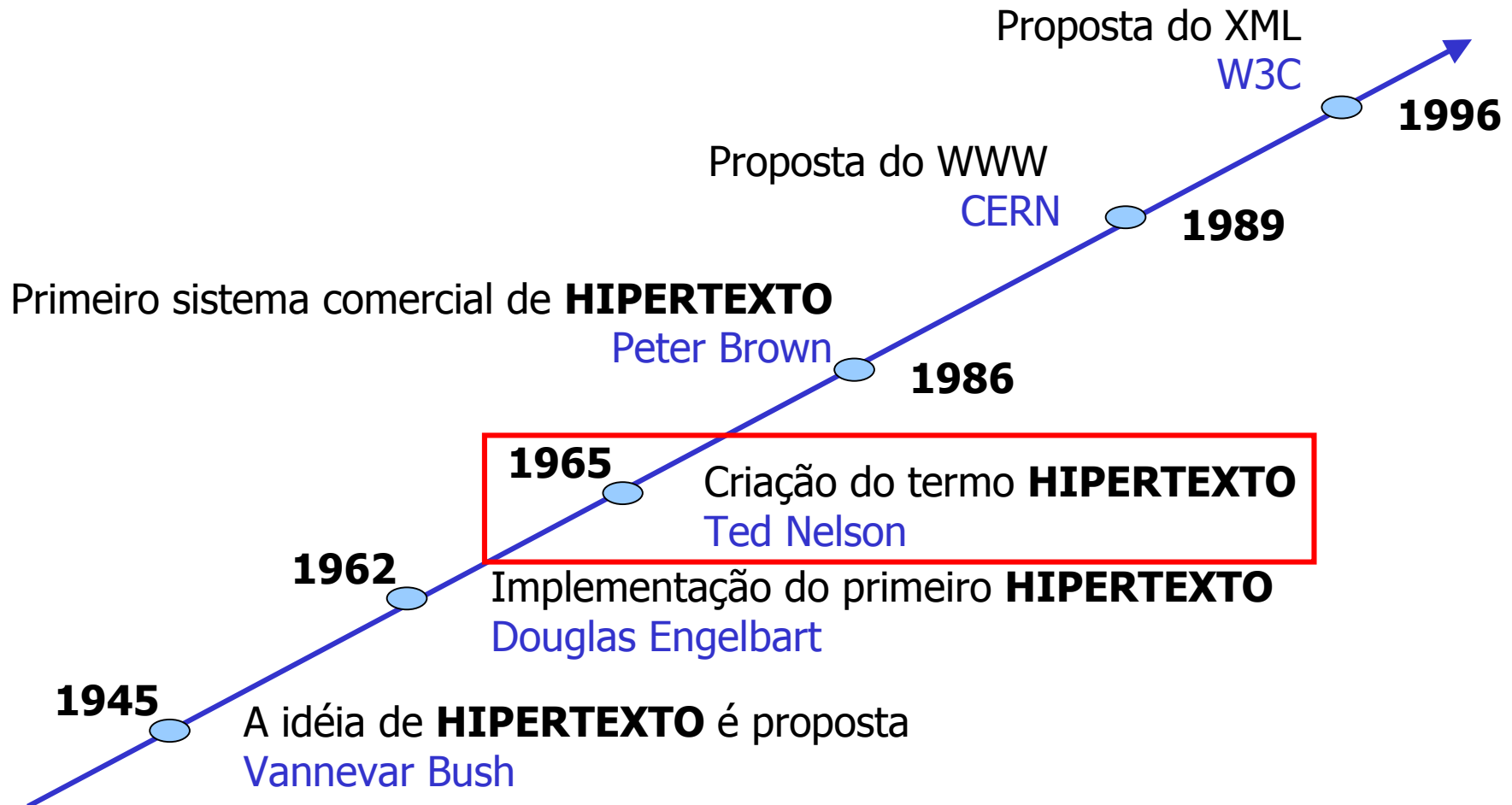
Uma Breve História



Engelbart - Primeiro Hipertexto

- AUGMENT - primeiro sistema de hipertexto operacional (1962 – 1975)
- “precursor” dos computadores pessoais e do trabalho cooperativo
- inovações:
 - ✓ mouse
 - ✓ múltiplas janelas na tela
 - ✓ correio eletrônico
 - ✓ processador de textos
 - ✓ sistema de ajuda
 - ✓ teleconferência
 - ✓ ligações e nós hipertextuais

Uma Breve História



Nelson - O Termo HIPERTEXTO

- Projeto **XANADU** - iniciado em 1965
- HIPERTEXTO: “escrita não sequencial”
- criação de uma biblioteca eletrônica distribuída global no formato hipermídia: Docuverse
- continha a maioria dos conceitos dos sistemas de hipermídia atuais (inclusive WWW!)
- “royalty” automático: usando o copyright e *transpublicação*

Nelson - XANADU

“Universal or grand hypertext, then, means a new publishing system - an accessible great universe of linked documents and graphics (and audio recordings and video and movies).”

“Imagine making your own notes and connections any way you choose in this great interconnected corpus; so that any time you want to reopen this great hypertext world at any of these private annotations that make it your own, it will be like opening a book to a bookmark.”

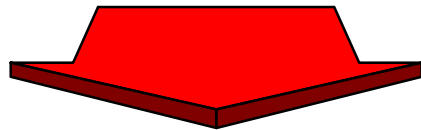
Sistemas de Hipertexto - Outros marcos importantes

- van Dam (Brown Univ.) - hipertexto no ensino (1968)
- ZOG (Carnegie-Mellon Univ.) - comercializado depois como KMS (1972)
- GUIDE (Univ.of Kent) - primeiro sistema comercial para computadores pessoais (1986)
- HYPERCARD - primeiro grande “hit” comercial (1987)
- ◆ WWW e a INTERNET - larga difusão da tecnologia de hipermídia
- ◆ Surgimento do XML

Hipertexto: Uma área de grandes profetas!

Futuro de Hipertexto (Nielsen – 1990)

- aparecimento de um mercado de massa para os hipertextos
- integração dos hipertextos a outras facilidades computacionais

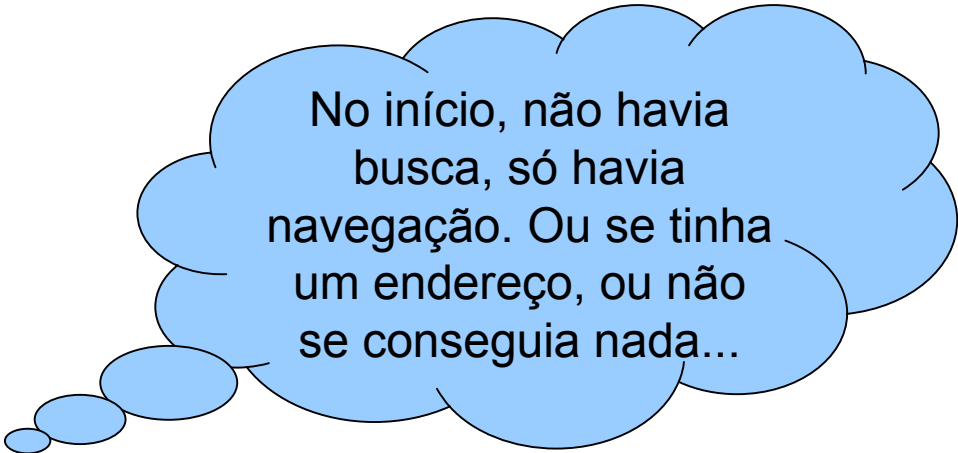


PADRONIZAÇÃO

- linguagens de marcação
- linguagens script
- formatos de armazenamento multimídia

Hipermídia: a Era Web

- Reconhecimento das vantagens da navegação
- Agora:
 - ✓ Não mais preso no universo de um hiperdocumento autocontido
 - ✓ Autonomia na criação de ligações entre hiperdocumentos
 - ✓ Descentralização



No início, não havia busca, só havia navegação. Ou se tinha um endereço, ou não se conseguia nada...

WWW - Objetivos Iniciais

- Ambiente distribuído, Cliente-servidor
- Múltiplas plataformas
- Facilidades:
 - Histórico, Anotação, Mecanismos de busca
- Flexibilidade de formato dos documentos
- Notificação de novo material (*)
- Versões(*)
- Ligações tipadas(*)
- Nomes lógicos para os nós(*)

(*) não disponíveis na era HTML!

Melhorias nos Navegadores

- “Frames”: múltiplas seções, com conteúdo dinâmico
- Objetos “vivos” com “viewers” inline
- Programas de autoria associados
- Outras ferramentas auxiliares
 - ✓ desenvolvimento de aplicações
 - ✓ verificadores de ligações
 - ✓ fácil acesso a BD relacionais
- Maior integração com correio eletrônico e outros
- ...

Web X Hipermídia “stand-alone”

- Popularidade - Grátis !
- Grande portabilidade de plataformas e sistemas
- Abundância de ferramentas de domínio público
- Expansão espantosa de funcionalidades (adições ou “helpers”, etc.)
- Multimídia ainda deve ser usado com cautela
- Controle do “todo” fora do alcance do desenvolvedor (em geral)

Comparando Abordagens de Tratamento da Informação

Sistemas de Banco de Dados

X

Sistemas Hipermídia

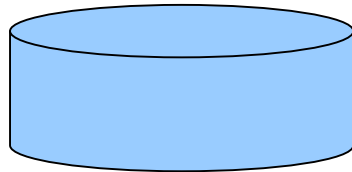
X

Sistemas de Recuperação de Informação

Sistemas de Banco de Dados

Abordagem de Banco de Dados

Funcionalidades de um SGBD:



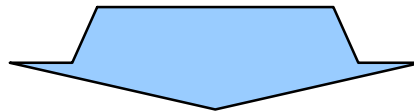
- Compartilhamento de dados
- Controle de redundância
- Controle de acesso
- Múltiplas interfaces
- Representação de relacionamentos
- Restrições de integridade
- Capacidade de “backup” e restauração

Aplicações de Banco de Dados

CUST_NAME	ORDER_ID	DATE
Hill's Hardware	245067	8/10/96
Denton Wood	403140	9/12/96
Philly's Products	394010	8/15/96
Power Townsend	497678	8/18/96
Keith's Controls	568922	9/22/96
George's Kitchen	60895	9/24/96



Característica Principal: Estruturação da Informação



- Poder de consulta “semântica” ao banco de dados
 - Facilidade de Manutenção

Abordagem de Banco de Dados

- Fornece armazenamento *eficiente* para uma grande quantidade de dados
- *Estrutura* fornece suporte para mecanismos de recuperação mais elaborados: consultas

```
SELECT DATA_NASC, ENDEREÇO  
FROM EMPREGADO  
WHERE NOME = 'JOSÉ DA SILVA'
```

- Diferentes níveis de *abstração*
- *Flexibilidade* de mudanças
- Potencial para o estabelecimento e o cumprimento de *padrões*

Modelo, Esquema, Instância

- Modelo
 - ✓ Coleção de conceitos para descrever um conjunto de dados e as operações que os manipulam
- Esquema
 - ✓ Representação de uma porção específica da realidade, a partir de um modelo de dados particular
- Instância
 - ✓ Coleção dinâmica de dados que se conforma à estrutura definida pelo esquema

Ex: BD Relacional

- Modelo: Relacional
 - ✓ Quais são os conceitos e operações?

Modelo, Esquema, Instância

Vendor

vendorId	vendorName	url	state	country
01	Amazon	www.amazon.com	WA	USA
02	Barnes and Noble	www.barnesandnoble.com	NY	USA

SellBook

vendorId	isbn	price
01	1111	38
01	2222	29
02	1111	38
02	2222	38

Warehouse

wId	vendorId	address	city	state	country
D1	01	1245, Bourbom Street	Seattle	WA	USA
D2	02	1478, 25th Avenue	New York	NY	USA
D3	01	4545, 15th Avenue	Seattle	WA	USA

Book

isbn	title	publisher	year
1111	Unix Network Programming	Prentice Hall	1998
2222	Computer Networks	Prentice Hall	1996

Dvd

asin	title	genre	nrDisks
D1111	Friends	Comedy	4

SellDvd

vendorId	asin	price
01	D1111	29

Constraints:

On table Vendor:

- primary key(vendorId)

On table Warehouse

- primary key(wId)
- foreign key(vendorId) references Vendor

On table Book

- primary key(isbn)

On table Dvd

- primary key(asin)

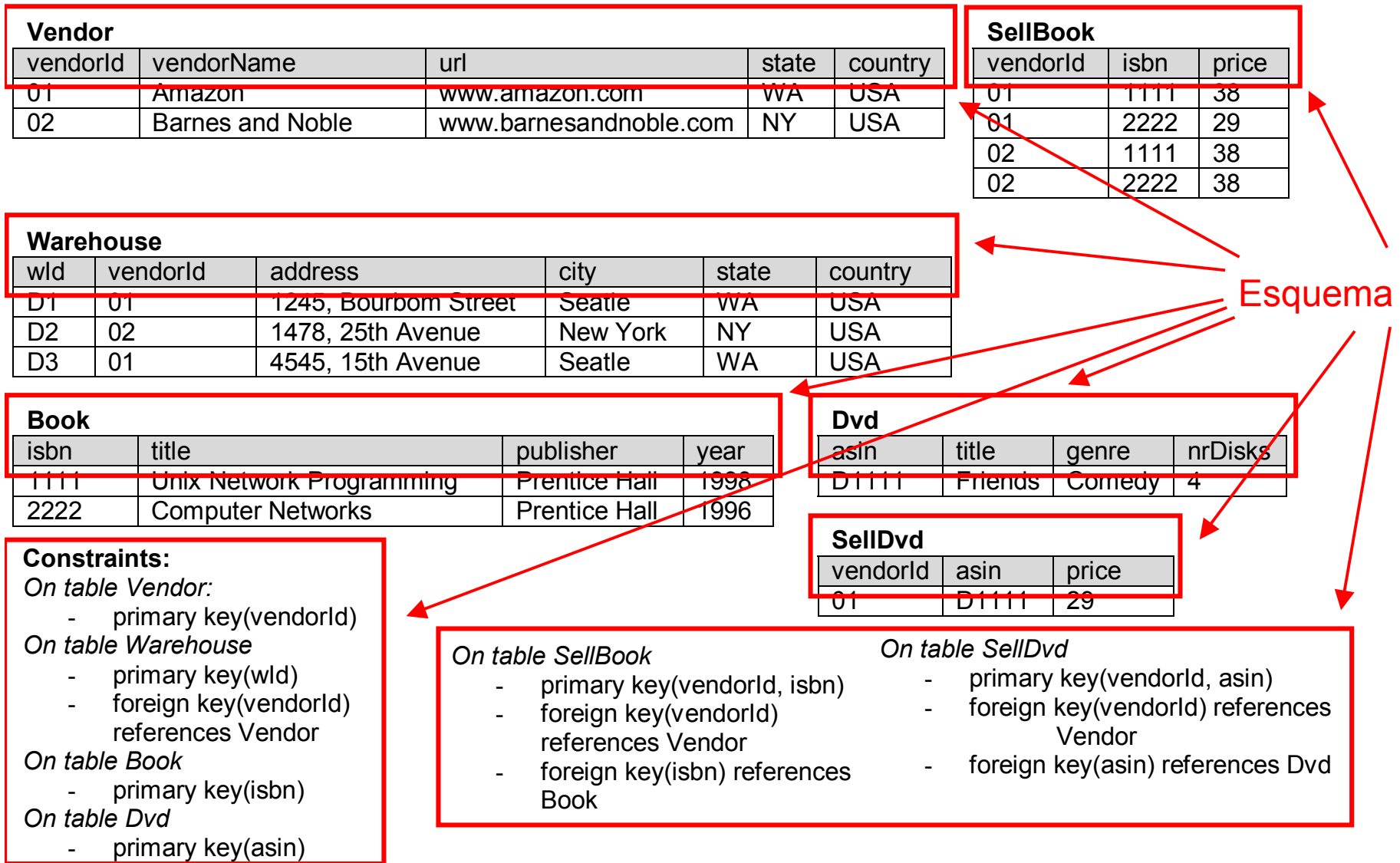
On table SellBook

- primary key(vendorId, isbn)
- foreign key(vendorId) references Vendor
- foreign key(isbn) references Book

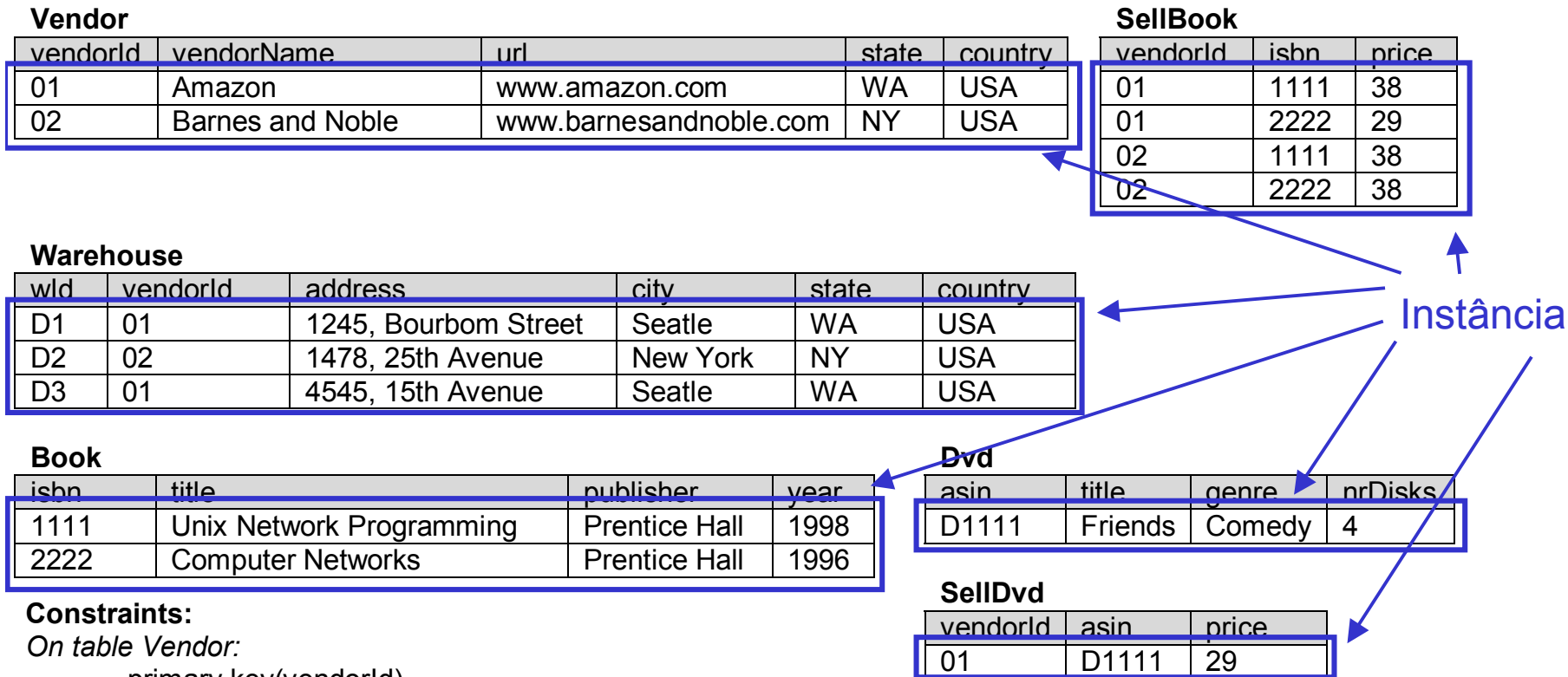
On table SellDvd

- primary key(vendorId, asin)
- foreign key(vendorId) references Vendor
- foreign key(asin) references Dvd

Modelo, Esquema, Instância



Modelo, Esquema, Instância



Constraints:

On table Vendor:

- primary key(vendorId)

On table Warehouse

- primary key(wId)
- foreign key(vendorId) references Vendor

On table Book

- primary key(isbn)

On table Dvd

- primary key(asin)

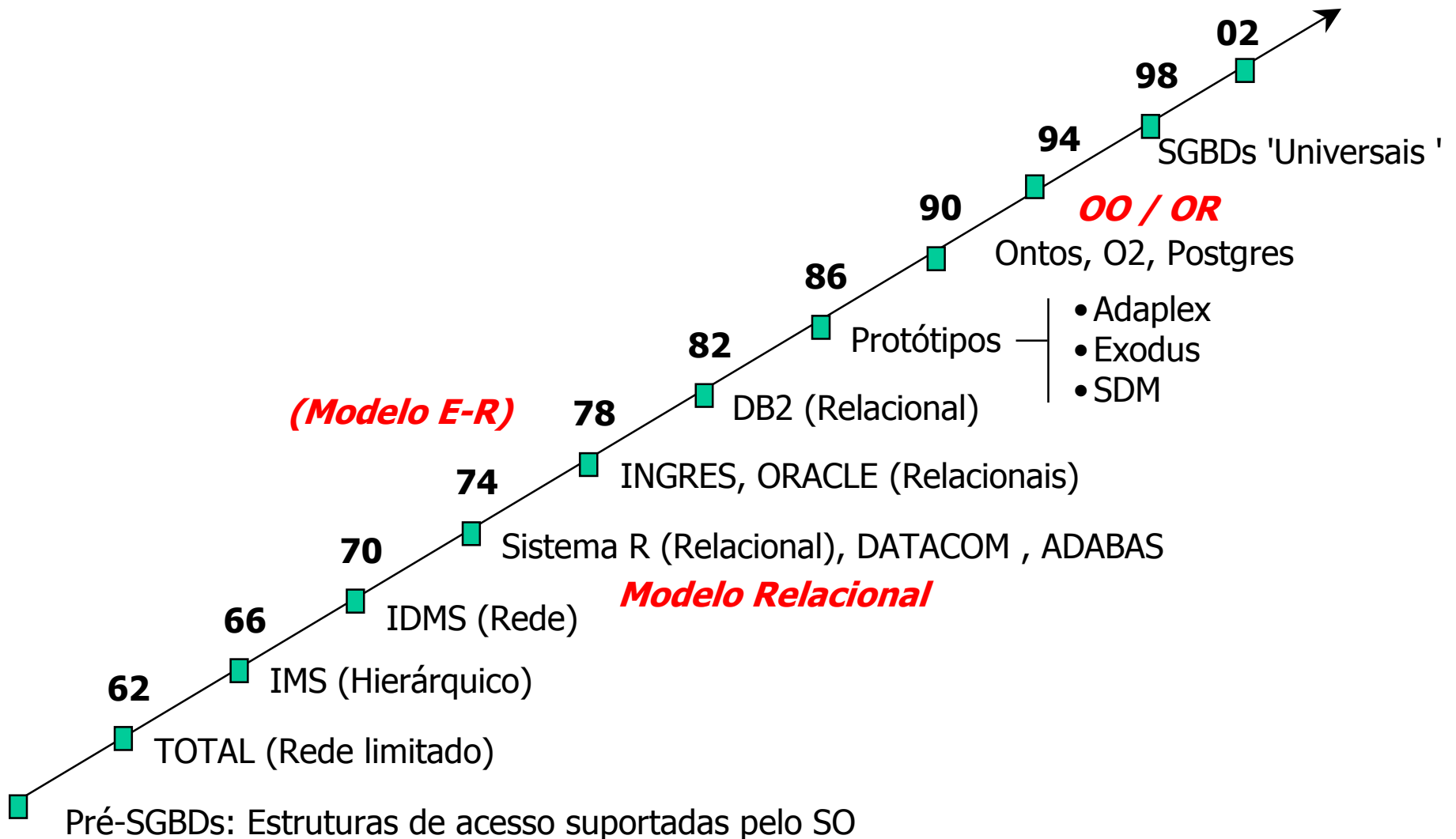
On table SellBook

- primary key(vendorId, isbn)
- foreign key(vendorId) references Vendor
- foreign key(isbn) references Book

On table SellDvd

- primary key(vendorId, asin)
- foreign key(vendorId) references Vendor
- foreign key(asin) references Dvd

Evolução dos Modelos e SGBDs



Perspectiva Histórica dos SGBDs

- Evolução no caminho de níveis mais altos de abstração

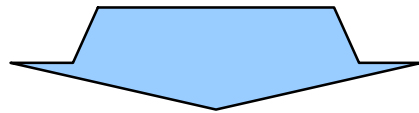
Modelos de Dados mais próximos dos conceitos da realidade sendo representada
e
menos ligados à forma de representação na máquina

Sistemas Hipermídia

Aplicações de Hipermídia

Característica Principal:

Flexibilidade de tratar informações não estruturadas



Facilidade de Navegação – interface amigável

Abordagem Hipermídia

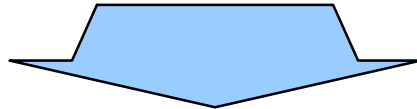
- Associações no nível da instância
- Não pressupõe a existência de uma esquema
- Flexibilidade !
- Hipermídia “tradicional”
 - ✓ problemas na manutenção
 - ✓ “lost in space”
- ✓ Capacidade de Consultas?

Sistemas de Recuperação de Informação

Aplicações de Recuperação da Informação (*Information Retrieval*)

Característica Principal:

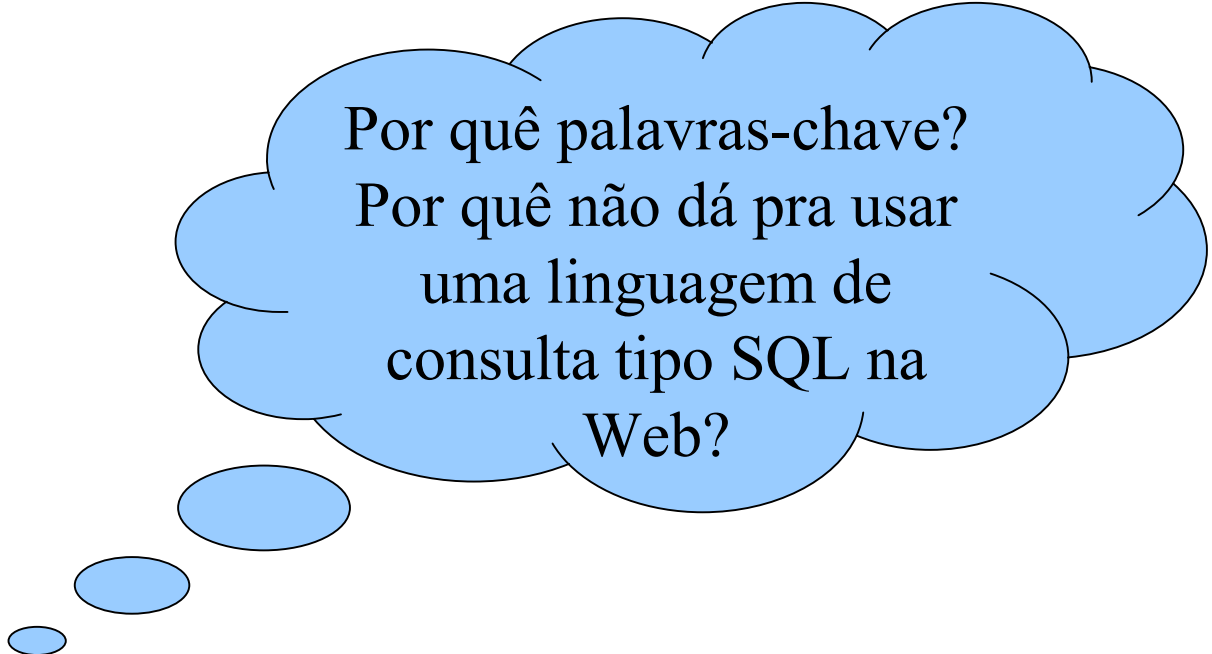
Indexação e classificação de documentos para posterior recuperação por comparação de palavras-chaves



Capacidade de organizar e consultar acervos de documentos

Abordagem de Recuperação da Informação

- Busca automática baseada em palavra-chave
- Técnicas de Indexação
- Técnicas de Classificação



Por quê palavras-chave?
Por quê não dá pra usar
uma linguagem de
consulta tipo SQL na
Web?

Indexação

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

Vocabulário

A
B
C

Listas invertidas

(D1,3) (D2,2) (D3,2)
(D1,1) (D4,2)
(D2,1)

Critérios utilizados em Recuperação da Informação

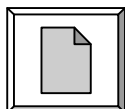
- Revocação (*recall*)
 - ✓ Grau de sucesso: número de documentos relevantes recuperados em relação ao total de documentos relevantes existentes
- Precisão (*precision*)
 - ✓ Mede o sucesso da filtragem: número de documentos relevantes recuperados em relação ao total de documentos recuperados
- Esforço do usuário
- Cobertura: volume de informações disponíveis
- Formato de saída
- Tempo de resposta
- Atualidade

Técnicas de Recuperação da Informação

- Relacionamento entre termos
 - ✓ Operadores booleanos
 - ✓ Proximidade entre termos
 - ✓ Linguagem natural
 - ✓ Através de vocabulário
- Interpretação de uma única palavra
 - ✓ Truncagem
 - ✓ Distinção entre maiúsculas e minúsculas
 - ✓ Delimitação por campo
 - ✓ Eliminação de palavras não significativas
 - ✓ Atribuição de peso a termos
 - ✓ Incorporação automática de sinônimos

Desafio

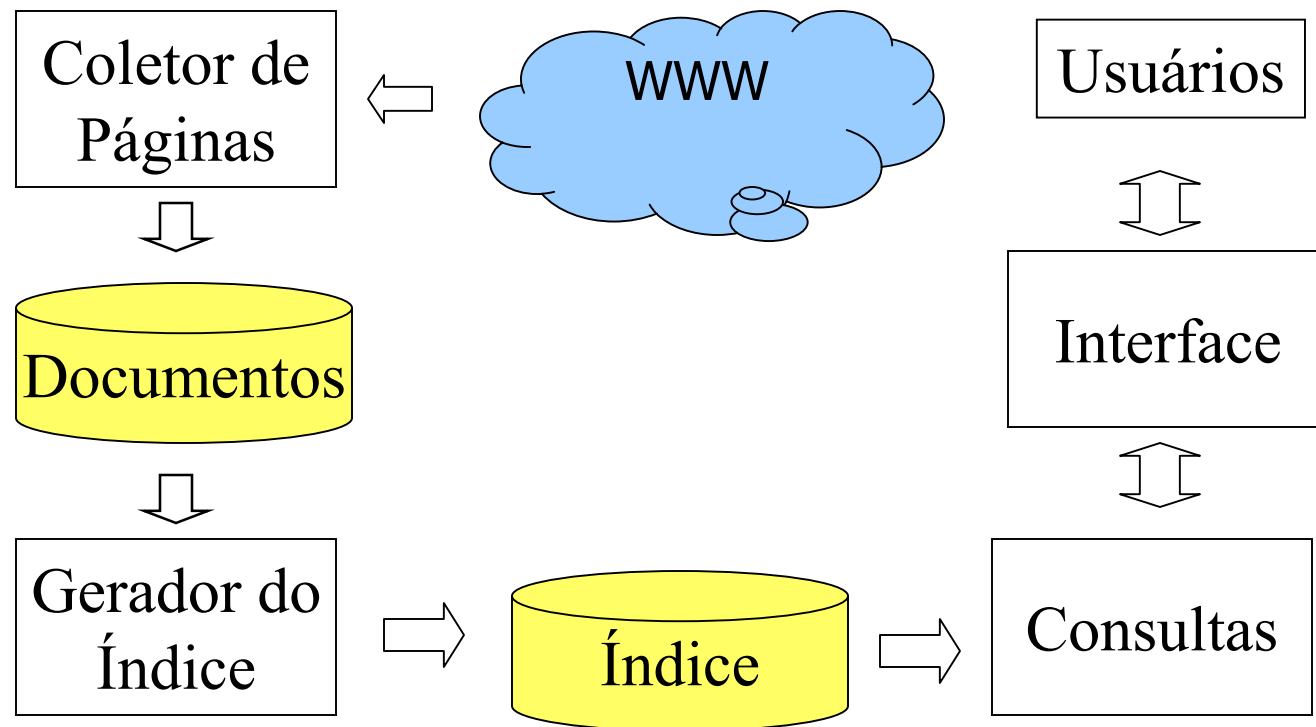
- Vamos descobrir que outras técnicas o Google utiliza para descobrir que documentos são relevantes para uma busca de um usuário?
- Reúnam-se em grupos e pesquisem na Web sobre o Google
- Depois vamos discutir o que vcs descobriram...



Sistemas de Busca na WWW

- Desenvolvidos como aplicações em separado, não são parte dos “browsers”
 - ✓ Necessidade de complementar capacidade de navegação com facilidades de busca
- Ferramentas de busca consistem de três componentes:
 - ✓ coleta
 - ✓ busca
 - ✓ ordenação

Máquinas de busca para Web

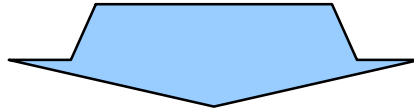


Aplicações na Web

- Surgiram como típicas aplicações hipermídia
- Mas com características próprias:
 - ✓ grande volume de informações
 - ✓ descentralização
- Necessidade de:
 - ✓ facilitar localização e recuperação de informações
 - ✓ maior semântica
 - ✓ facilitar manutenção
 - ✓ tratamento uniforme para qualquer tipo de recurso
- Mudança nas aplicações corporativas: INTRANETS e EXTRANETS
- Evoluíram para uma integração de abordagens:
 - ✓ Banco de Dados, Hipermídia, Recuperação de Informação

Tendência

Necessidade de integrar cada vez mais estas abordagens



Fronteira entre o tipo de informação tratada por cada abordagem tende a desaparecer

Abordagens apresentam funcionalidades **complementares**

Mas e o XML?

- Tenta agir como a integração destas abordagens
 - ✓ Dados semi-estruturados

Curiosidade: Google

Material de Referência:

- ✓ Artigo: The anatomy of a large- scale hypertextual web seach engine. Sergey Brin e Lawrance Page, WWW 1998

Cenário da época (1997):

- ✓ Usuários faziam buscas seguindo links a partir de índices mantidos por humanos, como o Yahoo;
- ✓ Ou faziam buscas automáticas (altavista, lycos, etc.), que retornavam respostas de baixíssima qualidade
- ✓ Páginas comerciais tentavam “burlar” os sistemas automáticos, fazendo com que suas páginas sempre aparecessem nas primeiras posições do resultado

Em 1994

- ✓ Acreditava-se que um índice que fosse completo seria capaz de encontrar tudo facilmente
 - Mas isso não é o suficiente – qualidade da resposta (índices cresceram, mas a capacidade das pessoas de examinar milhares de respostas, não)
 - Pessoas querem examinar apenas alguns documentos, e que eles sejam relevantes
- ✓ Das 4 melhores máquinas de busca comerciais disponíveis em 97, apenas **uma** delas conseguia encontrar a si mesma! (testem com o Google...)

Google

- ✓ O Google é uma proposta para melhorar este cenário
- ✓ Nome **Google**: é como se pronuncia o número 10^{100} (**googol**), o que reflete o objetivo de construir uma máquina de busca de larga escala para a Web

Google em 1997

- ✓ Índice de 24 milhões de páginas
- ✓ Tamanho do repositório comprimido: 53.5Gb (repositório sem compressão = 148Gb)
- ✓ Tamanho do arquivo de índices: 37.2Gb

Google em 2000

- ✓ Indexava aprox. 1,3 bilhões de páginas
- ✓ Respondia à mais de 100 milhões de consultas por dia
- ✓ Respostas em menos de meio segundo

Histórico do Google

Material de Referência:

✓ Artigo: Google History,
www.google.com/corporate/history.html

Histórico...

✓ Fundadores:

- Larry Page e Sergey Brin
- Se conheceram em 1995 em Stanford, quando ambos eram alunos de doutorado
- Larry (24 anos) estudava em Michigan e foi visitar Stanford no fim de semana
- Sergey (23 anos) era um dos responsáveis por ciceronear Larry pelo campus
- Eles não gostaram um do outro logo de cara...
 - Tinham opiniões muito fortes, brigavam sempre e por tudo

1996

- ✓ Larry e Sergey começaram a colaborar num projeto de uma máquina de buscas chamada BackRub
- ✓ O dinheiro era curto, e eles saíam à caça de computadores que pudessem ser emprestados para sua rede
- ✓ O BackRub começou a ficar famoso no campus...

Até metade de 1998

- ✓ Larry e Sergey continuaram trabalhando para melhorar a tecnologia da máquina de busca
- ✓ Compraram 1Tb de disco bem baratinho
- ✓ Construíram sua própria sede para os computadores no dormitório do Larry na faculdade – o primeiro CPD do Google

Enquanto isso...

- ✓ Sergey montou um escritório e começou a procurar por parceiros potenciais
- ✓ Apesar da febre .com da época, poucos se interessaram
- ✓ O fundador do Yahoo, David Filo, aconselhou os dois a criar eles mesmo uma empresa, e amadurecer a tecnologia, e quando estivesse mais sólida, voltariam a conversar

Então...

- ✓ Eles decidiram tocar o negócio eles mesmos, e colocaram os planos de doutorado de lado...
- ✓ Mas precisavam de dinheiro
- ✓ Procuraram então um “anjo” que os ajudasse
- ✓ Foram falar com um dos fundadores da Sun, Andy Bechtolsheim, que era amigo de um dos professores da universidade

Então...



- ✓ Andy viu um demo, achou que o projeto tinha potencial, mas estava com pressa pra ir a uma reunião
- ✓ Decidiu então não entrar em detalhes, e escreveu um cheque de US\$ 100.000,00 em nome de **Google Inc.**
- ✓ Mas não existia nenhuma empresa chamada **Google Inc.**, então não era possível depositar o cheque

Então...

- ✓ O cheque ficou na mesa de Larry enquanto ele organizava a empresa e arrumava mais alguns investidores entre família e amigos
- ✓ O total arrecadado ficou próximo de 1 milhão de dólares



Setembro de 1998

- ✓ Google abre as portas em Menlo Park, California
- ✓ As portas tinham um controle remoto...
 - A empresa foi estabelecida na garagem de um amigo...
 - Tinha várias vantagens: uma secadora e lavadora, e uma banheira
 - Também tinha lugar para o primeiro funcionário contratado estacionar (Craig Silverstein, hoje diretor de tecnologia do Google)



1998

- ✓ Já era Google.com, versão beta
- ✓ Respondia uma média de 10 mil consultas por dia
- ✓ Começou a chamar a atenção... Artigos sobre ele foram publicados na USA TODAY e no Le Monde
- ✓ Em Dezembro, a PC Magazine colocou o Google na lista de Top 100 Sites Web e Máquinas de Busca de 1998

Fevereiro de 1999

- ✓ Se mudaram para um escritório em Palo Alto
- ✓ Tinha 8 empregados
- ✓ Respondia mais de 500.000 consultas por dia
- ✓ O interesse na companhia cresceu

Junho de 1997

- ✓ Investidores interessados
- ✓ US\$ 25 milhões de duas companhias (competidoras) de capital do Silicon Valley
- ✓ Google continua a expandir, contratando mais pessoal chave
- ✓ Novo escritório: Googleplex, em Mountain View, Califórnia

Setembro de 1999

- ✓ O label “beta” desapareceu
- ✓ Enquanto o Google continuava a crescer, uma cultura única no escritório evoluía
 - Sem “cubículos”
 - Móveis confortáveis
 - Jogos
 - Cachorros no escritório
 - etc...

1999

- ✓ Essa atmosfera informal acelerou a troca de idéias
- ✓ Muitas melhorias ao Google
 - Diretório Google
 - Versões em 10 linguas diferentes

Junho de 2000

- ✓ Google se torna a maior máquina de busca da Web
- ✓ Google e Yahoo anunciam parceria – aumento da reputação do Google
- ✓ 18 milhões de consultas por dia
- ✓ Final de 2000: mais de 100 milhões de consultas por dia

Até 2003

- ✓ Google lança uma série de inovações: busca por imagens, Google News, Froogle (busca por produtos e preços), etc...
- ✓ Em 2003, Google compra o Pyra Labs, e passa a ser o dono do Blogger

2004

- ✓ Local Search (para quando se quer algo na vizinhança (comida, lavanderia, etc.))
- ✓ GMail
- ✓ ...

Características do Google

Material de Referência:

- ✓ Artigo: The anatomy of a large- scale hypertextual web seach engine. Sergey Brin e Lawrance Page, WWW 1998
- ✓ Artigo: Web Search for a planet: The Google cluster architecture. Luiz André Barroso, Jeffrey Dean, Urs Hölzle. IEEE Micro, Março/Abril de 2003

Características

- ✓ Page rank
- ✓ Anchor Text

Page Rank

- ✓ Aproveita o grafo de links da Web
- ✓ Links têm pesos diferentes para calcular o **page rank** de uma página
- ✓ $PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$
- ✓ d normalmente é 0,85
- ✓ $C(T1)$ = número de links que saem de T1

Exemplo

Página	Aponta para
Y1	P1,P2
Y2	P2
P1	P2

Y1 e Y2 não são apontadas por ninguém

- $PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$
- $PR(Y1) = (1-0,85) = 0,15$
- $PR(Y2) = (1-0,85) = 0,15$
- $PR(P1) = (1-0,85) + 0,85(PR(Y1)/C(Y1)) = 0,15 + 0,85(0,15/2) = 0,213$
- $PR(P2) = (1-0,85) + 0,85(PR(Y1)/C(Y1) + PR(Y2)/C(Y2) + PR(P1)/C(P1)) = 0,15 + 0,85(0,15/2 + 0,15/1 + 0,213/1) = 0,52$

Texto Âncora

- ✓ Texto Âncora = **texto dos links**
- ✓ O texto do link é associado com
 - A página em que o link está
 - A página para qual o link aponta

Consulta

- ✓ Quando o usuário faz uma consulta, a primeira coisa que é feita é decidir que servidor (cluster) atenderá o pedido
- ✓ Cada cluster tem alguns milhares de máquinas
- ✓ É escolhido o cluster geograficamente mais próximo
- ✓ Clusters estão espalhados geograficamente para evitar que o sistema seja afetado por catástrofes (terremotos, grandes faltas de energia, etc.)

Servidor de consultas

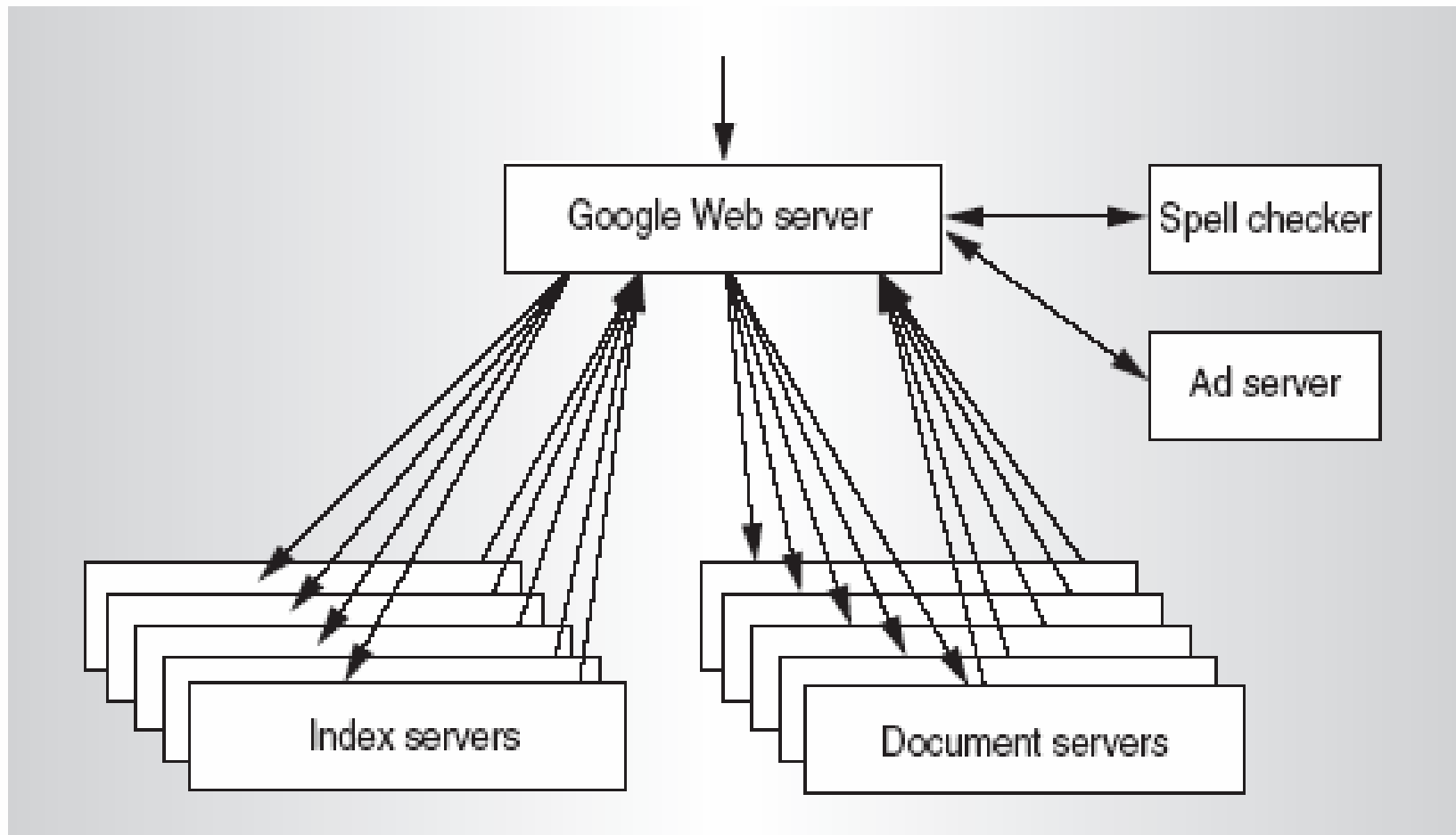


Figure 1. Google query-serving architecture.

Fonte: Web Search for a planet – the google cluster architecture

Execução das consultas

- Duas fases: Fase 1
 - ✓ **Servidores de índice** (*Index servers*) consultam um **índice** (arquivo invertido) procurando por cada palavra da consulta, e encontram a lista de documentos que contêm tais palavras
 - ✓ Eles fazem a intersecção da lista de documentos para encontrar os documentos que contêm **todas** as palavras
 - ✓ Calculam a relevância de cada documento

Execução das consultas

- Desafio: quantidade de dados
 - ✓ Base de documentos: dezenas de terabytes (em 2003)
 - ✓ Índice: vários terabytes (2003)
 - ✓ Consulta é paralelizada: o índice é dividido em pedaços, cada computador processa um pedaço
 - ✓ Resultado da primeira fase: uma lista de docIds ordenada

Execução das consultas

- Duas fases: Fase 2
 - ✓ Pegar a lista de docIds e computar o título e resumo de cada documento (contexto da consulta)
 - ✓ Isso é feito pelos **servidores de documentos** (*document servers*)

Replicação

- ✓ Base de documentos e índices são replicados em vários servidores
- ✓ Resultado: Google armazena várias “cópias” da Web em seus servidores

Tarefas auxiliares

- Durante o processamento da consulta:
 - ✓ Corretor ortográfico para sugerir novas consultas
 - ✓ Servidor de anúncios vê se existe algum patrocinador relacionado com a consulta do usuário, e mostra o(s) anúncio(s) em caso afirmativo

Outros detalhes...

- O segredo eles não revelam...