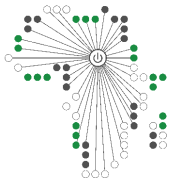


Universidade Virtual Africana

INFORMÁTICA APLICADA: ITI 4301

RECUPERAÇÃO DE INFORMAÇÃO

Marcelo Francisco de Barros Correia

Prefácio

A Universidade Virtual Africana (AVU) orgulha-se de participar do aumento do acesso à educação nos países africanos através da produção de materiais de aprendizagem de qualidade. Também estamos orgulhosos de contribuir com o conhecimento global, pois nossos Recursos Educacionais Abertos são acessados principalmente de fora do continente africano.

Este módulo foi desenvolvido como parte de um diploma e programa de graduação em Ciências da Computação Aplicada, em colaboração com 18 instituições parceiras africanas de 16 países. Um total de 156 módulos foram desenvolvidos ou traduzidos para garantir disponibilidade em inglês, francês e português. Esses módulos também foram disponibilizados como recursos de educação aberta (OER) em oer.avu.org.

Em nome da Universidade Virtual Africana e nosso patrono, nossas instituições parceiras, o Banco Africano de Desenvolvimento, convido você a usar este módulo em sua instituição, para sua própria educação, compartilhá-lo o mais amplamente possível e participar ativamente da AVU Comunidades de prática de seu interesse. Estamos empenhados em estar na linha de frente do desenvolvimento e compartilhamento de recursos educacionais abertos.

A Universidade Virtual Africana (UVA) é uma Organização Pan-Africana Intergovernamental criada por carta com o mandato de aumentar significativamente o acesso a educação e treinamento superior de qualidade através do uso inovador de tecnologias de comunicação de informação. Uma Carta, que estabelece a UVA como Organização Intergovernamental, foi assinada até agora por dezenove (19) Governos Africanos - Quênia, Senegal, Mauritânia, Mali, Costa do Marfim, Tanzânia, Moçambique, República Democrática do Congo, Benin, Gana, República da Guiné, Burkina Faso, Níger, Sudão do Sul, Sudão, Gâmbia, Guiné-Bissau, Etiópia e Cabo Verde.

As seguintes instituições participaram do Programa de Informática Aplicada: (1) Université d'Abomey Calavi em Benin; (2) Université de Ougagadougou em Burkina Faso; (3) Université Lumière de Bujumbura no Burundi; (4) Universidade de Douala nos Camarões; (5) Universidade de Nouakchott na Mauritânia; (6) Université Gaston Berger no Senegal; (7) Universidade das Ciências, Técnicas e Tecnologias de Bamako no Mali (8) Instituto de Administração e Administração Pública do Gana; (9) Universidade de Ciência e Tecnologia Kwame Nkrumah em Gana; (10) Universidade Kenyatta no Quênia; (11) Universidade Egerton no Quênia; (12) Universidade de Addis Abeba na Etiópia (13) Universidade do Ruanda; (14) Universidade de Dar es Salaam na Tanzânia; (15) Universite Abdou Moumouni de Niamey no Níger; (16) Université Cheikh Anta Diop no Senegal; (17) Universidade Pedagógica em Moçambique; E (18) A Universidade da Gâmbia na Gâmbia.

Bakary Diallo

O Reitor

Universidade Virtual Africana

Créditos de Produção

Autor

Marcelo Francisco de Barros Correia

Par revisor(a)

Martina Barros

UVA - Coordenação Académica

Dr. Marilena Cabral

Coordenador Geral Programa de Informática Aplicada

Prof Tim Mwololo Waema

Coordenador do módulo

Robert Oboko

Designers Instrucionais

Elizabeth Mbasu

Benta Ochola

Diana Tuel

Equipa Multimédia

Sidney McGregor

Barry Savala

Edwin Kiprono

Kelvin Muriithi

Victor Oluoch Otieno

Michal Abigael Koyier

Mercy Tabi Ojwang

Josiah Mutsogu

Kefa Murimi

Gerisson Mulongo

Direitos de Autor

Apoiado por

Projeto Multinacional II da UVA financiado pelo Banco Africano de Desenvolvimento.

Índice

Descrição Geral do Curso

Bem-vindo(a) a Informática Aplicada- Recuperação de Informação

Este Curso pretende levar aos profissionais da área a capacidade de gerir e recuperar informações. Com o avanço das novas tecnologias de informação passa-se a dispor de grande volumes de informações,o que está a transformar se em Caos,em vez de se transformar numa mais-valia para a organização. O mundo vive "afogado"no mar de informações,o que consiste numa ansiedade descontrolada de encontrar o melhor e o melhor de todos, isso faz com que os utilizadores passem mais tempo á procura e armazenar grande quantidade de informação e produzem pouco conhecimento divido ao caos informacional. Depois essas informações dificilmente vão ser revisadas como forma de transforma-la em conhecimento ou em valor informacional,de forma a ser utilizado em tempo útil e opurtuno, porque não conseguem ter acesso à informação de maneira satisfatória.

Alguns aspectos a como rapidez, localização e abrangência de conteúdos, surgem outras facetas, como relevância, especificidade e utilidade dos elementos recuperados, determinados de acordo com a necessidade informacional do indivíduo. O expressivo e crescente volume de conteúdos disponibilizados na web contribui para dificultar a ampliação do grau de especificidade na recuperação da informação, elemento-chave para garantir sua relevância informacional e sua utilidade no desenvolvimento da produção científica.

Com este módulo pretende se mostrar aos alunos e profissionais da área as técnicas e forma de gerir, recuperar, acessar e disponibilizar em momentos opurtuno respeitando o valor da informação e dar mais impulso a organização institucional.

Pré-requisitos

1. Para este curso é suposto que os alunos despõem de conhecimento básico de sistemas de informação e do seu funcionamento
2. Noções básicas de matemática tais como a lógica e domínio das ferramentas de internet.

Volume horário/Tempos

1. A duração deste módulo é de 120 horas repartidas entre leituras, actividades práticas, trabalhos dirigidos e avaliações formativas e sumativas.
2. Para a leitura são 20 horas para as 4 unidades. Para as actividades práticas, 20 horas. Para as consultas dos links e recursos, 20 horas. Para os trabalhos dirigidos são 20 horas e, Para as avaliações formativas e sumativa, 40 horas.

Materiais

Os materiais necessários para completar este curso incluem:

1. CD-Rom
2. Livros
3. E-books
4. Tutoriais
5. Computadores
6. Internet
7. Vídeo aulas

Para além destes os alunos (as) podem recorrer a outros materiais ou softwares suplementares como forma de reforçar a compreensão e realizar simulações.

Objectivos do Curso

1. Identificar os componentes de um sistema de informação;
2. Conhecer os fundamentos da tecnologia da informação;
3. Diferenciar os diferentes modelos dos sistemas de recuperação da informação;
4. Caracterizar os sistemas de pesquisa Web;
5. Apresentar os modelos de banco de dados e as etapas de projeto e implementação.
6. Explicar as diferentes formas de recuperar informções via web
7. Compreender conceitos fundamentais de recuperação e extracção de informação
8. Avaliar o desempenho do sistema de IR,as técnicas e algoritmos utilizados na recuperação e extracção de informação baseada em texto (IR e IE).

Unidades

Unidade 0: Conceitos Recuperação de Informação

Com o avanço das novas tecnologias de informação as informações passaram a ser armazenadas em diferentes meios e tornaram se mais volumosos e heterogéneos, de forma a produzir um caos informacional. Com isso tornou se imprescindível o papel de Recuperação de informação-RI.

Unidade 1: Modelos de RI e Métrica de avaliação da eficiência de recuperação

Os Modelos são responsáveis por recuperar os documentos de que precisa, e avaliar o grau de satisfação do utilizador final em função do resultado da consulta formulada.

Unidade 2:Recuperação de informação na Web /pesquisa web

Com a revolução tecnológica a maior fonte de informação a nível mundial, encontra-se armazenado em base de dados distribuidos, que se encontram remotamente localizados em servidores dispersos no universo utilizando como canal Internet.

Unidade 3: Sistema de Gestão da base de dados

Para que as informações possam ser recuperadas, tem de ser armazenadas numa base de dados, sendo assim é de extrema importância conhecer o papel da base de dados dentro do SRI.

Avaliação

Em cada unidade encontram-se incluídos instrumentos de avaliação formativa a fim de verificar o progresso do (a)s aluno(a)s.

No final de cada módulo são apresentados instrumentos de avaliação sumativa, tais como testes e trabalhos finais, que compreendem os conhecimentos e as competências estudadas no módulo.

A implementação dos instrumentos de avaliação sumativa fica ao critério da instituição que oferece o curso. A estratégia de avaliação sugerida é a seguinte:

1	Teste	35%
2	Teste 2	30%
3	Fichas	35%

Calendarização

Unidade	Temas e Actividades	Estimativa do tempo
Conceitos de recuperação de informação	Técnicas básicas e avançadas de informação textual: Introdução ao Armazenamento e Recuperação de Informação (definição, componentes, tipos de sistemas de recuperação de informação, o processo de recuperação, etc)	10 h
Modelos de Recuperação de Informação	Modelos de recuperação de espaço booleanos,vectoriais e probabilistico Métrica de avaliação da eficiência de recuperação	4h
Recuperação de informação na Web	Interface Pesquisa Web e rastreamento Algoritmos baseados em ligações e meta-dados da web	4h
Sistema de Gestão da base de dados	DBMS versus sistemas de recuperação de informação Consultas de textos em bases de dados e questões atuais sobre recuperação de informação •Indexação baseada em texto •índices invertidos •Eficiência na indexação baseada em texto •Categorização e classificação de texto/ documentos e classificação; mineração de texto	10h

Leituras e outros Recursos

1. CD-Rom
2. Livros
3. E-books
4. Tutoriais
5. Computadores
6. Internet

Os alunos podem recorrer a outros materiais ou softwares suplementares como forma de reforçar a compreensão e realizar simulacões.

Unidade 0

Leituras e outros recursos obrigatórios:

- "Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)", R. Baeza-Yates, B. Ribeiro-Neto B., 2011, Addison Wesley Professional.
- Recuperação de Informação: estudo sobre a contribuição da

Ciência da Computação para a Ciência da Informação FERNEDA, E.. São Paulo, 2003.

Leituras e outros recursos opcionais:

- Managing Gigabytes , Ian Witten, Alistair Moffat, and Timothy Bell, Morgan-Kaufmann, 1999, ISBN: 1558605703
- Recuperação de Informação - Conceitos e Tecnologia Das Máquinas de Busca - 2ª Ed. 2013 Baeza-Yates, RicardoRibeiro-Neto, Berthier

Unidade 1

Leituras e outros recursos obrigatórios:

- Modern Information Retrieval BAEZA-YATES, R.; RIBEIRO-NETO.
- Introdução Aos Modelos Computacionais de Recuperação de Informação,Ferneda, Edberto

Leituras e outros recursos opcionais:

- Os Fundamentos Da Lógica Aplicada À Recuperação Da Informação, Ariadne Chloe Furnival
- Precisão no processo de pesquisa e recuperação da informação, ARAÚJO JÚNIOR, Rogério Henrique, 2007.

- [http://comum.rcaap.pt/handle/123456789/172?mode=full&submit_simple=Mostrar+registo+em+formato+completo
- http://libros.metabiblioteca.org/bitstream/001/227/8/84-932537-7-4.PDF

Unidade 2

Leituras e outros recursos obrigatórios:

- Sistemas de gerenciamento de banco de dados - 3.ed. Raghu Ramakrishnan,Johannes Gehrke

Leituras e outros recursos opcionais:

- http://minhateca.com.br/bibliotecadeti/Banco+de+dados/Sistema_de_Banco_de_Dados+Abraham+Silberschatz,78506107.pdf, acessado em Novembro de 2014.
- http://minhateca.com.br/bibliotecadeti/Banco+de+dados/Banco_Dados,78497529.pdf, acessado em Novembro 2014.
- Projecto de banco de dados, Carlos Alberto Heuser, http://pt.slideshare.net/laercionunesdacosta/carlos-alberto-heuser-projeto-de-banco-de-dados acessado em Novembro 2014

Unidade 3

Leituras e outros recursos obrigatórios:

- GARMAN, Nancy, Meta search engines. Online, v. 23, n.3, p. 75-78,May/June 1999.
- WWW search tools in reference services. The reference librarian,. KIMMEL, Stacey.
- METODOS DE PESQUISA PARA INTERNET, AMARAL, ADRIANA, FRAGOSO, SUELY,RECUERO, RAQUELeditora: SULINA,2011
- Ricardo Baeza-Yates and Berthier Ribeiro-Neto: Modern Information Retrieval, Addison Wesley, 2011, 913 pages, Second edition.
- Ian H. Witten, Alistair Moffat e Timothy C. Bell: Managing Gigabytes: Compressing and Indexing Documents and Images, Morgan Kaufmann Publishers, 1999, second edition.

Leituras e outros recursos opcionais:

- http://unefazuliasistemas.files.wordpress.com/2011/04/fundamentos-de-bases-de-datos-silberschatz-korth-sudarshan.pdf
- Visualization for Information Retrieval ZHANG, J.. New York, NY: Springer. 2008

Unidade 0. Conceitos de Recuperação de Informação

Introdução à Unidade

O propósito desta unidade é verificar o conhecimento que possui relacionado com este curso.

Os alunos devem saber usar, entender o significado, parafrasear os conceitos relacionados com o mundo da recuperação da informação, passando pela sua história, os processos, e das etapas da recuperação de informação. Qualquer profissional que lida com a recuperação de informação tem de entender esses conceitos de forma que lhe ajude na análise, planificação e implementação dos projectos de recuperação de informação.

Objectivos da Unidade

Após a conclusão desta unidade, deverá ser capaz de:

1. Definir Dados
2. Definir Informação
3. Diferenciar Dados de Informação
4. Avaliar sistema RI
5. Interpretar e descrever as etapas Principais de Recuperação de informação
6. Distinguir e seleccionar modelos de Recuperação de Informação
7. Analisar e interpretar o processo de recuperação de informação
8. Distinguir a recuperação de informação da recuperação de dados
9. Enumerar alguns sistemas correntes de recuperação de informação
10. Definir os conceitos de documentos,coleção, necessidade de informação, relevância, resposta

Termos-chave

Recuperação de Informação: localização e recuperação de documentos que podem ser relevantes a uma pesquisa. É necessário um sistema para filtrar esses documentos especificados pelo utilizador e indexar as palavras-chave encontradas.

Extração de Informação: os termos considerados relevantes nos documentos são extraídos e convertidos em dados afim de que possam ser utilizados durante o processo de mineração.

Mineração da Informação: assim que a informação é armazenada de forma estruturada, a descoberta de informação é feita através da mineração sobre o banco de dados criado

Termo: string de caracteres alfanumérico,unidade de informação

Pesquisas: ("queries") são feitas por um único termo ou por composição de termos utilizando-se conectores lógicos (and, or, not), operadores relacionais (>, <, =) e meta-caracteres (*, ?)

Robô (ou spider): programas que percorrem links na web, geralmente com objetivo de indexá-la

Corpus: conjunto de documentos etiquetados

Routing: faz a mesma coisa que filtragem, a medida que os documentos vão sendo adicionados ao Corpus

Arquivo invertido: termos (índices) mapeando os documentos em que aparecem

Stop List: lista de palavras comuns, irrelevantes ex:Artigos: a, os, ...

Stemming e n-grams: redução de termos. Ex:CONNECTCONNECTED CONNECTING CONNECTION CONNECTIONS

Base de Índice: banco de dados de um sistema de índices

Similaridade: o grau de quanto 2 documentos são semelhantes

Thesaurus: identifica o relacionamento entre termos

Classificação da informação: Processo que permite agrupar as informações com as características e propriedades idênticas, facilitando assim o seu tratamento e uso.

Digitalização: Conversão de informação analógica (som, imagem, papel e vídeo)em valores digitais correspondentes, manipuláveis por computador.

Recuperação de Dado**s: A recuperação de itens(tuplos,** documentos, páginas Web, etc...) cujo conteúdo satisfaz exactamente as condições especificadas na interrogação (tipo expressão regular)

Recuperação distribuída de informação: A utilização de técnicas de computação distribuída para resolver o problema da recuperação da informação

Recuperação de Informação Multimedia: Sistema de recuperação de informação que manipula documentos multimedia.

Recuperação de Informação (Information Retrieval): Área das ciências da computação que estuda a recuperação de informação (não de dados) duma colecção de documentos. Os documentos devolvidos têm como objectivo satisfazer uma necessidade de informação do utilizador expressa normalmente em linguagem naturalFicheiro invertidoResultado da inversão das tabelas de ocorrência de todos os documentos que constituem a colecção.

Filtragem de informação: Sistema de RI que indexa perfis de informação que correspondem a necessidades de informação e compara com os documentos dum fluxo fazendo chegar aos utilizadores os documentos considerados relevantes pelo respectivo perfil

Indexador de Documentos: Componente do sistema de RI que processa os documentos extraindo a informação considerada útil para construir o ficheiro invertido e os registos de meta-informação.

Lista de ocorrência: Conjunto de registos de ocorrência relativo a um termo, isto é, informação a respeito da ocorrência do termo em cada documento. A terminologia usada em inglês é Posting List.

Termo de índice:Uma palavra usada para identificar o conteúdo dum documento. Normalmente é um nome ou uma frase (2 ou mais nomes).

Thesaurus:Uma estrutura de dados composta de uma lista de palavras importantes dum dado domínio de conhecimento e para palavra da lista, um lista de palavras relacionadas (sinónimos, etc...).

Recuperação de Informações e Seus Objectivos

1. Ultimamente tem crescido muito a necessidade de se recuperar informações, e o primeiro objectivo é a classificação dos documentos úteis no conjunto de informações.
2. Segundo Baeza Yates e Ribeiro Neto(1999), apud (TAKAO, 2001), recuperação de Informação incorpora os conceitos de modelagem, classificação de documentos, filtragem, interfaces, linguagem, etc.
3. A Internet revolucionou o mundo de informação tornando um depósito universal de conhecimento humano e cultura, no qual tem dado uma grande colaboração de ideias e informações numa escala nunca vista antes.
4. Sistema de Recuperação de Informação definição apresentada por Robertson(Robertson apud [BEJ 92]). "Sistema de Recuperação é assim considerado se o mesmo tiver a funcionalidade de conduzir o utilizador para aqueles documentos que melhor irão habilitá-lo a satisfazer as suas necessidades de informação."
5. Para você melhor entender e analisar Sistemas de Recuperação de Informação tem que se debruçar sobre alguns conceitos de extrema importância, tais como:

Dados

Numa primeira abordagem, dado pode ser definido como INFORMAÇÃO BRUTA.

Definimos dado como uma sequência de símbolos quantificados ou quantificáveis. Portanto, um texto é um dado. De fato, as letras são símbolos quantificados, já que o alfabeto por si só constitui uma base numérica. Também são dados imagens, sons e animação, pois todos podem ser quantificados a ponto de alguém que entra em contacto com eles ter eventualmente dificuldade de distinguir a sua reprodução, a partir da representação quantificada, com o original. É muito importante notar-se que qualquer texto constitui um dado ou uma sequência de dados, mesmo que ele seja ininteligível para o leitor. Como são símbolos quantificáveis, dados podem obviamente ser armazenados em um computador e processados por ele.

Em nossa definição, um dado é necessariamente uma entidade matemática e, desta forma, puramente sintática. Isto significa que os dados podem ser totalmente descritos através de representações formais, estruturais.

Tipos de dados

Existem diferentes tipos de dados como nos é ilustrada na Figura 1:

1. Quais os tipos de dados que temos hoje?
2. Dados Estruturados
3. Dados Semi-Estruturados
4. Dados não-estruturados

Dados estruturados

Dados organizados em blocos semânticos (relações),numa estrutura plana (tabelas)

Dados de um mesmo grupo possuem as mesmas descrições (atributos)

Descrições para todas as classes de um grupo possuem o mesmo formato (esquema)

Dados mantidos em um SGBD são chamados de Dados Estruturados por manterem a mesma estrutura de representação (rígida), previamente projetada (esquema)

Dados semi estruturados

Divido a heterogeneidade dos dados, muitos dados não são mantidos no SGBD

1. Dados Web, por exemplo, apresentam uma organização bastante heterogênea.
2. A alta heterogeneidade dificulta as consultas a estes dados
3. Assim, estes dados são classificados como semi-estruturados

- Não são estritamente tipados
- Não são complementarmente não-estruturados

FIGURA 1
Dados conforme sua estrutura

Figura 1:Estrutura de Dados

A tabela que se segue ilustra as principais diferenças entre os diferentes tipos de dados

Dados Estruturados	Dados Semiestruturados	Dados Não Estruturados
Esquema predefinido	Nem sempre há esquemas	Não há esquemas
Estrutura regular	Estrutura irregular	Estrutura irregular
Estrutura independente dos dados	Estrutura embutidado dados	Pode não ter estrutura alguma
Francamente evolutiva	Fortemente evolutiva (estrutura modifica com frequência)	Fortemente evolutiva (estrutura modifica com frequência)
Prescritivas (esquemas fechados e restrições de integridade	Estrutura descritiva	Estrutura descritiva
Distinção entre estrutura e dados é clara	Distinção entre estrutura e dados não é clara	Distinção entre estrutura e dados não é clara
Etrutura reduzida	Estrutura extensa (particularidade de cada dado, visto que cada um pode ter uma organização própria)	Estrutura extensa (particularidade de cada dado, visto que cada um pode ter uma organização própria)

Processo de Recuperação de Informação

Depois de conhecer os diferentes tipos de dados, nesse documento concentra se sobretudo na recuperação de documentos textual.

Um sistema de recuperação de informações textuais é um sistema desenvolvido para indexar e recuperar documentos do tipo textual, ou seja, documentos cujas informações estão descritas através da linguagem natural. São sistemas que tratam basicamente informações do tipo texto (ASCII), mas, que através de filtros adequados podem analisar outros formatos que contenham textos, figuras, tabelas e imagens, mas que possuam um aspecto de documento textual, como o PDF, o PS ou DOC).

Conceitos Básicos

Um sistema automático para RI pode ser visto como a parte do sistema de informação responsável pelo armazenamento ordenado dos documentos em um banco de dados, e sua posterior recuperação para responder a consulta do utilizador.

Um sistema de recuperação de informação pode ser representado por três componentes: entrada, processador e saída (Van RIJSBERGEN, 1979). Analisando as entradas (documentos e consultas), o principal desafio é obter uma representação de cada documento e consulta.

Na figura 2 estão exibidas todos os passos que precisa para a recuperar informação.

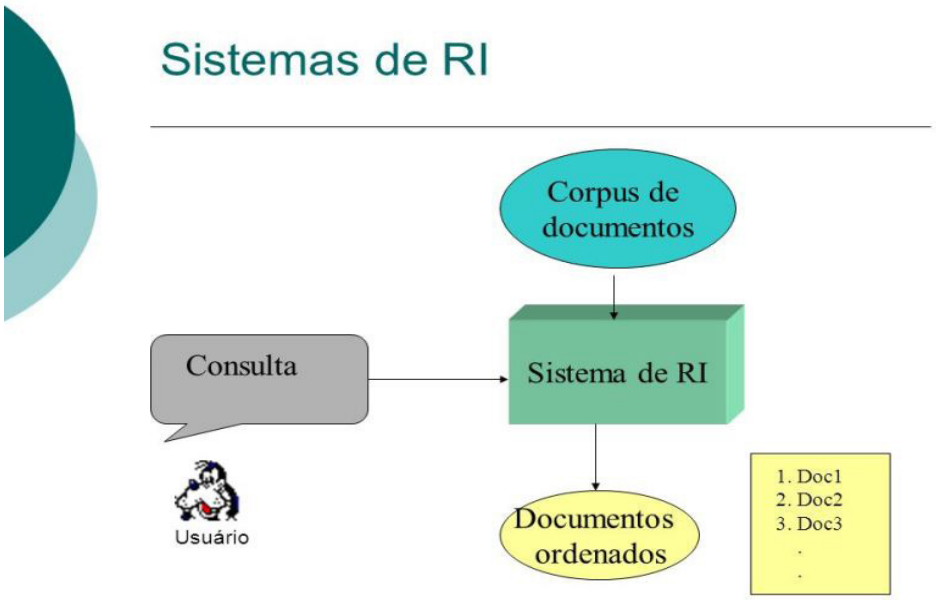


Figura 2:Processo de Recuperação de Informação

Definição da informação

Numa primeira abordagem pode ser definido a informação como sendo dados contextualizados.

Hoje conta-se com um grande número de informações, segundo (CRESTANI,1991 apud TAKAO,2001) elas podem ser textuais, visuais ou auditivas. Como consequência os bancos que armazenam tais informações estão se tornando cada vez maiores. A IR trata em desenvolver recursos para recuperar informações, independente do tamanho do banco de dados. Normalmente um sistema de IR conduz o utilizador aos documentos que irão melhor possibilitar a satisfação de suas necessidades.

Um utilizador que necessita de uma informação, devido à grandeza do banco de informações, precisa do apoio de uma ferramenta chamada SRI. Ao contrário de um sistema tradicional de base de dados, o SRI não fornece uma resposta exacta, mas um ranking de documentos com informações relevantes, os bons SRI apresentam primeiramente os documentos com maior similaridade com a pergunta, caso estes documentos sejam irrelevantes, o utilizador reformula a pergunta para um novo ranking de documentos, conforma a figura abaixo, descrita por (CRESTANI, 1998,apud TAKAO, 2001).

Informação é uma abstracção informal (isto é, não pode ser formalizada através de uma teoria lógica ou matemática), que representa algo significativo para alguém através de textos, imagens, sons ou animação. Note que, isto não é uma definição - isto é uma caracterização, porque «algo», «significativo» e «alguém» não está bem definido; assumimos aqui um entendimento intuitivo desses termos. Não é possível processar informação directamente em um computador. Para isso é necessário reduzi-la a dados. A representação da informação pode eventualmente ser feita por meio de dados. Nesse caso, pode ser armazenada em um computador. Mas, atenção, o que é armazenado na máquina não é a informação, mas a sua representação em forma de dados. Essa representação pode ser transformada pela máquina - como na formatação de um texto - mas não o seu significado, já que este depende de quem está entrando em contato com a informação. Por outro lado, dados, desde que inteligíveis, são sempre incorporados por alguém como informação, porque os seres humanos (adultos) buscam constantemente por significação e entendimento.

Uma distinção fundamental entre dado e informação é que o primeiro é puramente sintático e o segundo contém necessariamente semântica (implícita na palavra "significado" usada em sua caracterização). É interessante notar que é impossível introduzir semântica em um computador, porque a máquina mesma é puramente sintática (assim como a totalidade da matemática). Se examinássemos, por exemplo, o campo da assim chamada "semântica formal" das "linguagens" de programação, notaríamos que, de fato, trata-se apenas de sintaxe expressa através de uma teoria axiomática ou de associações matemáticas de seus elementos com operações realizadas por um computador (eventualmente abstrato).

	Recuperação de dados	Recuperação de informação
Adequação	Exata	Adequação parcial, melhor adequação
Inferência	Dedução	Indução
Modelo	Determinística	Probabilistica
Classificação	Monotética	Politética
Linguagem de consulta	Artificial	Natural
Especificação de questão (*query*)	Completa	Incompleta
Itens desejados	Adequação	Relevante
Resposta ao erro	Sensível	Insensível

Etapas Para Recuperação de Informação

Etapas

- Aquisição (seleção) dos documentos
- Preparação dos documentos
- Indexação dos documentos
- Armazenamento

- Recuperação
- Pesquisa (casamento com a consulta do utilizador)
- Ordenação dos documentos recuperados
- Aquisição (seleção) de Documentos
- Manual para sistemas gerais de RI
- Por exemplo, sistemas de bibliotecas

Automática para sistemas na Web

Uso de crawlers, spiders ou robots, Programas que navegam pela Web e fazem download das páginas para um servidor.

Preparação dos Documentos

Objectivo

Criar uma representação computacional do documento seguindo alguma visão lógica do documento, que consiste na realização operações sobre o texto que permite a sua representa e facilite a sua recuperação.

Aquisição (Seleção) dos Documentos

Na primeira Etapa é realizada a aquisição dos documentos para que estes sejam submetidos à extração automática da informação, sua classificação e armazenamento.

Na aquisição dos documentos, o utilizador pode escolher documentos armazenados em disco rígido, CD, na Web ou transformando os documentos em formato de papel por meio de digitalização. Para a extração da informação, o utilizador seleciona os ficheiros de seu interesse. Logo na etapa seguinte inicia o processo de preparação do documento (ficheiro) seleccionado.

ParaTransformar os documentos em formato de papel em digital as seguintes tarefas devem ser realizadas:

- Escaneamento dos documentos originais e reconhecimentos de textos mediantes software específico (ocr optical character reconhecimento)
- Processamento manual do resultado de reconhecimento para corrigir os erros para reconhecimento.
- Inserção do documento no sistema, tendo em consideração os diferente formatos suportados pelo sistema (pdf,doc,html,xml,etc)

Preparação dos Documentos

Nesta etapa os documentos selecionados são analisados com o objetivo de identificar quais termos serão definidos como palavras-chave. Como resultado, é obtido um conjunto de palavras-chave (termos) que identificam o conteúdo do documento. Durante a etapa de preparação dos documentos são realizadas os seguintes processos:

- Filtragem e Padronização do Texto;
- Elaboração de stoplist - remoção de stop-words ;
- Stemming: algoritmo que reduz as palavras na sua forma raiz;
- Determinação das probabilidades de relevância dos termos de acordo com o modelo probabilístico e com o modelo probabilístico exponencial, e os pesos de acordo com o modelo vetorial.

Todos esses processos estão visualizados na Figura 2. O resultado desta etapa será armazenado numa base de dados, para ser utilizado na etapa seguinte.

Filtragem

Este algoritmo tem como entrada o texto e como saída uma versão processada ou filtrada desse texto. Isto é feito para reduzir o tamanho do texto ou padronizá-lo, tornando a pesquisa mais eficiente. As operações de filtragem mais comuns são:

- As palavras comuns são removidas, usando uma lista de stopwords (palavras da stoplist) gerando os termos(TOKEN) da entrada, que depois de um processamento adicional é candidato a ser uma entrada na indice;
- Letras maiúsculas são transformadas em minúsculas;
- Símbolos especiais são removidos e sequências de múltiplos espaços reduzidos para apenas um espaço;
- Números e datas são transformados para o formato padrão;
- Palavras são truncadas, removendo sufixos e/ou prefixos;
- Extração automática de palavras chaves;
- Ranqueamento das palavras.

Desvantagens: Qualquer query, antes de consultar o banco de dados, deve ser filtrada como é o texto; e não é possível pesquisar por palavras comuns, símbolos especiais ou letras maiúsculas, nem distinguir fragmentos de textos que tenham sido mapeados para a mesma forma interna.

Stoplist

1. A maioria das palavras mais frequentemente usadas em línguas são palavras sem valor de índice, isto é:

- Classificação gramatical (artigos, determinantes, pronomes, numerais, advérbios, preposições)
- Palavras mais frequentes

Entre estas, por exemplo, temos “o”, “a”, “de”, “para”, entre muitas outras.

2. Uma consulta feita com um desses termos retornaria quase todos os itens do banco de dados sem considerar a relevância destes.

3. Estas palavras compõem uma grande fração da maioria dos textos de cada documento.
4. Por conta disso, a eliminação de tais palavras no processo de indexação salva uma enorme quantidade de espaços em índices, e não prejudica a eficácia da recuperação.
5. A lista das palavras filtradas durante a indexação automática, em virtude delas gerarem índices pobres, é chamada de stoplist ou um dicionário negativo.
6. Uma maneira de melhorar a performance do sistema de recuperação de informação, então, é eliminar as stopword – palavras que fazem parte da stoplist – durante o processo de indexação automática.
7. A geração da stoplist depende das características do banco de dados, do utilizador e do processo de indexação.

Ex: TERMOS (Token)

8. Entrada: "Friends, Romans and Countrymen"
9. Saída: TERMOS (Tokens)

- Friends
- Romans
- Countrymen
- Como implementar uma StopList?

Existem duas maneiras de filtrar as stopwords de uma sequência deTERMOS- tokens de entrada. O analisador léxico examinar a saída e remover qualquer stopword, ou remover as stopwords como parte da análise léxica.

O primeiro método, filtrar stopwords da saída do analisador léxico, causa um problema: todo TERMO-token deve ser examinado na stoplist e removido da análise se encontrado. A solução mais rápida para isso é a utilização de hashing.

Stemming ou Truncagem

Uma outra técnica para melhorar a performance e eficiência e, também, facilitar as pesquisas para o utilizador, é prover aos pesquisadores do sistema meios de encontrar variantes morfológicas dos termos de pesquisa, ou seja, múltiplas representações da palavras como um único stem (redical). Se, por exemplo, um pesquisador entrar com o termo "comput" como parte da query, é provável que ele ou ela também estará interessado em tais variantes como "computable" , "computation", "computational" , etc.

Stemming é muito usado nos sistemas de recuperação de informação para reduzir o tamanho dos arquivos de índices. Desde que um único termo truncado (stem) tipicamente corresponde à vários termos completos, armazenar stems ao invés de termos, o fator de compressão chega a 50%.

A vantagem de stemming é a eficiência no tempo de indexação e a compressão do arquivo de índice. A desvantagem de indexar stemming é que a informação sobre os termos completos serão perdidas, ou adicional armazenamento será necessário para armazenar ambas as formas truncadas e não-truncadas.

Uma desvantagem é que o uso da truncagem causa um aumento da imprecisão da recuperação, pois o valor da precisão não está baseado em encontrar todos os itens relevantes, mas justamente minimizar a recuperação de termos irrelevantes.

Informação sobre os termos completos serão perdidas, ou adicional armazenamento será necessário para armazenar ambas as formas truncadas e não-truncadas.

Existem diversos métodos de truncagem. Entre eles, o mais comum é remover sufixos e/ou prefixos dos termos, deixando apenas a raiz do termo. Um outro método, é gerar uma tabela de todos os termos indexados e seus stems.

Por exemplo:

Termo	Stem
Engenheiro	Engenh
Engenharia	Engenh
Engenhenho	Engenh

Os termos da query e índices poderiam ser, então, truncados via pesquisa na tabela. Usando B-Tree ou hash, tais pesquisas seriam mais rápidas.

Questão: Aonde cortar?

Técnicas de limiar de similaridade e truncagem no agrupamento de texto

Mesmo com a utilização dessas medidas de similaridade, conforme Wives (2004),apud MARCUS VINICIUS CARVALHO GUELPELI) ainda é necessário o uso de um limiar (threshold) e de uma técnica de truncamento. A técnica de limiar estabelece um critério de corte, um limiar, ou threshold. O corte pode ser feito a partir da importância da palavra no texto, estipulando-se um valor mínimo ou máximo. A técnica de truncamento é estabelecida pelo limite máximo de palavras representativas do texto. De acordo com Wives (1999),apud MARCUS VINICIUS CARVALHO GUELPELI na maioria dos casos, 50 palavras são suficientes. Então, pelo método da truncagem, esse número pré-estabelecido é utilizado para formar o índice de cada texto, desconsiderando-se as demais palavras. Para as duas técnicas, são necessários o cálculo do peso das palavras e o seu ordenamento decrescente.

- Ponto de corte fixo
- Dependendo do número de variedades de sucessores
- Regras de remoção de sufixos e prefixos

Ex.: gênero, número, -mente, -ologia, -ional, -zação, -ismo, hiper-, meta- , re-, super-.

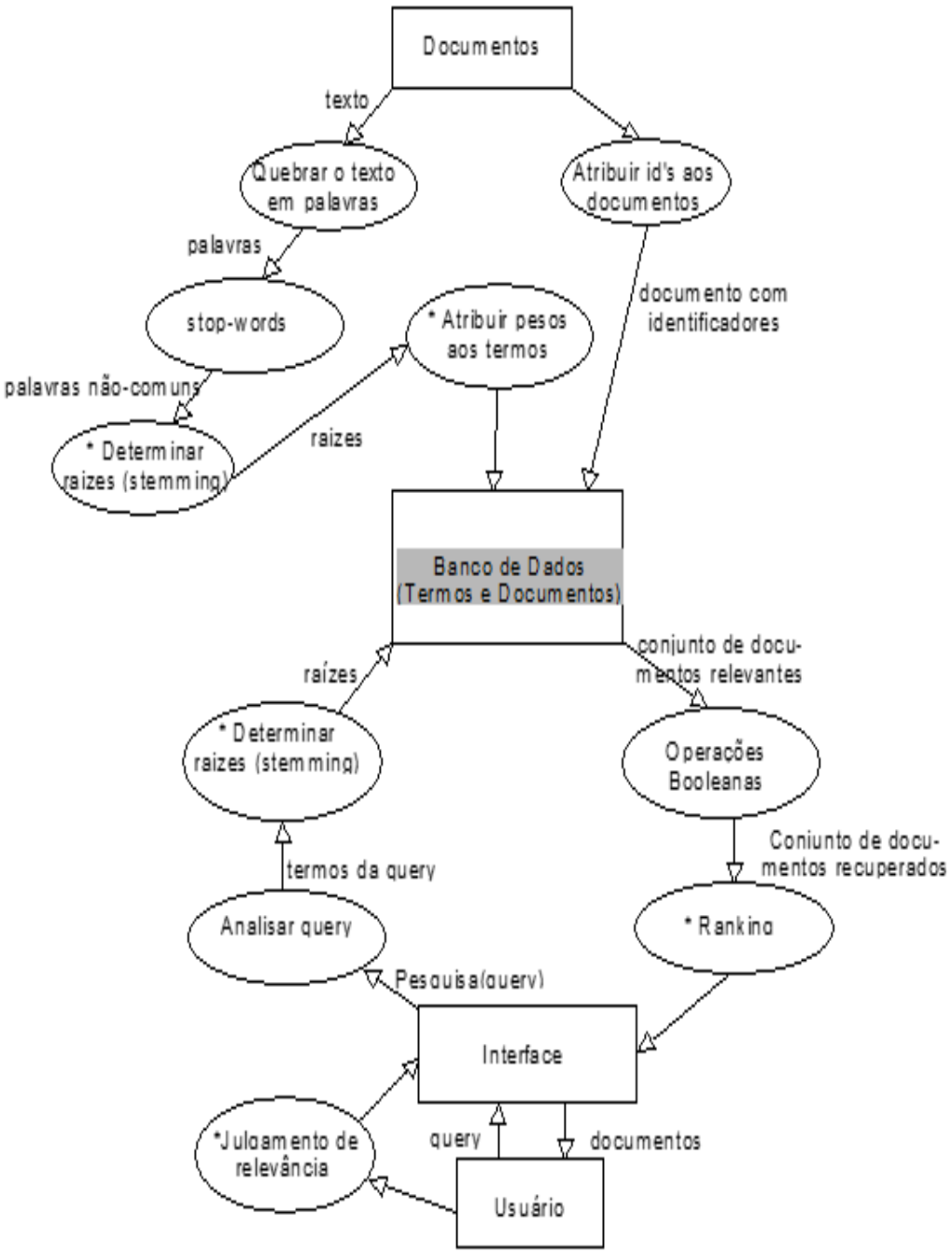


Figura 3: Processo de Recuperação de Informação

Indexação

O processo de indexação é responsável por converter os textos em um formato que proporcione buscas rápidas a elementos do texto, eliminado o processo seqüencial e lento da exploração por elementos do texto. Esse processo gera como saída um índice, que pode ser uma estrutura de dados que proporcione o acesso rápido às palavras armazenadas no documento, funcionando como o índice de um livro que possibilitam ao leitor a localização rápida das páginas procuradas por meios do termo indexado.

Indexar significa incluir um documento num repositório de informações, selecionando termos,em função do assunto para representar o indice,que facilite o acesso ao documento.

Esta estrutura guarda em si dois problemas: a indexação é trabalhosa e se realizada manualmente muitas vezes depende da subjetividade do classificador. Se automática, necessita de um vocabulário de referência que traduza efetivamente o contexto da busca.

Mesmo a indexação automática demanda tempo e capacidade de processamento, daí a importância de uma estrutura lógica para facilitar generalizações e melhorar o desempenho. A forma mais usual é a construção de um tesauros específico para a área de domínio ensejada. Neste caso, optamos pelo tesauros desenvolvido no projeto CDCON (AMORIM, 2006)

A partir desta base, o processo de indexação automática compara e identifica os termos relevantes (descritores) nos documentos de uma coleção e os insere em uma estrutura de índice. As fases normalmente encontradas nesse processo são a identificação de termos (simples ou compostos), a remoção de stopwords (palavras irrelevantes), a normalização morfológica (stemming) e a seleção de termos (Krug 2004).

Este processo é bem descrito nesta literatura técnica, assim como a discussão sobre a ponderação que cada documento deve receber ao ser inserido no índice. Existem três formas para esta análise: a Freqüência absoluta de termos (term frequency ou absolute frequency) Freqüência relativa de termos (relative frequency), Freqüência inversa de documentos (Inverse document frequency) A análise de relevância é realizada através da função de similaridade, mas esta comparação entre termos consultados e documentos em geral traz documentos irrelevantes. Para melhorar estes resultados foram propostos modelos conceituais de recuperação, posteriormente adaptados às ferramentas de busca: o modelo booleano, o espaço-vetorial, o probabilístico, o de busca direta, o de aglomerados (clusters) e o contextual ou conceitual.

Ainda assim, o desempenho dos sistemas de recuperação de informação depende primordialmente da organização e estrutura da base de dados de referência. Se ela refletir melhor o contexto do universo pesquisado, ela resulta em melhor precisão e revocação do sistema.

A metodologia de indexação consiste na identificação de termos (simples ou compostos), a remoção de stopwords (palavras irrelevantes), a normalização morfológica (lemetizaçao e stemming) e a seleção de termos baseada o no sistema de classificação pela medição da importância de cada palavra, identificando seu "peso" ou "força" de representatividade (term weight) considerando-se a freqüência das palavras nos documentos.

Processo de indexação consiste em três passos:

1. Definição do vocabulário de indexação

É o processo de atribuir termos ou códigos de indexação a um registo ou documento, termos ou códigosesses que serão úteis na recuperação do documento ou do registo.

2. Atribuição de termos de indexação a cada documento

   - A atribuição dos termos de indexação pode ser feito automaticamente ou manualmente
   - Os termos de indexação podem ser extraído de uma lista padrão (vocabulário controlado) ou um tesauro instalado no computador com base na ocorrência de palavras contidas no documento.
   - Uma outra alternativa será a indexação feita manual (Homem) com base no julgamento subjectivo que faz acerca do conteúdo do documento.

3. Construção da estrutura de dados do índice

Definição do vocabulário de indexação

O vocabulário empregado em um sistema de recuperação deve ser um vocabulário controlado, caracterizado por um conjunto limitado de termos, os quais se encontram organizados em alguma forma de estrutura que permita controlar sinônimos e remissivos que indiquem relações entre os termos (LANCASTER, 1985 apud AMORIM, Sergio et al 2007). Deste modo, um sistema de recuperação possui duas bases de dados distintas: uma armazena o conjunto de documentos, dos quais se deseja obter informações, e a outra contém as entradas que representam os documentos do sistema. Estas entradas são os descritores obtidos no processo de indexação, podendo ser considerado como um índice da outra base de dados (LANCASTER, 1985, apud AMORIM, Sergio R. Leusin ,2007)

Um Vocabulário é composto por conjunto de termos de indexação utilizáveis para representar o conteúdo temático dos documentos, que podem ser cabeçalhos de assuntos (listas de cabeçalhos de assunto), descritores (tesauros) ou símbolos de classificação (sistemas de classificação bibliográfica).

Esses termos podem ser definidos:

1. Manualmente por um especialista humano - vocabulário de indexação manual

   - Vocabulário controlado, tesauro,...
   - Refletem diretamente os assuntos dos documentos

2. Pode ser definido automaticamente pelo sistema de RI (após operadores de texto)-

Vocabulário de indexação automática

Consiste basicamente no conjunto de termos que aparecem nos documentos da

base após a preparação dos documentos

Vantagens:

- Maior cobertura de termos
- Maior velocidade no processo de indexação

Desvantagens

Pode gerar baixa precisão

Vocabulário de Indexação Manual

Definido pelo vocabulário controladoadotado:

1. Lista de cabeçalho de assuntos

   Lista simples de termos sem hierarquia

2. Taxonomia

   Lista de termos organizados com hierarquia

3. Tesauro

   Hierarquia de termos com relações associativas

4. Ontologia

   Hierarquia de assuntos organizados em classes e com relações associativas complexas

Vantagens:

- É possível ter uma visão lógica dos documentos que compõem a base
- É possível direcionar melhor a busca realizada pelo utilizador
- Aumenta a precisão na busca

Desvantagens:

- Cada documento é indexado por um humano (processo lento)
- Nem sempre é possível construir uma boa estrutura de assuntos
- O utilizador pode realizar buscas com termos que não aparecem no vocabulário controlado

Normalização do vocabulário

A normalização tem como objectivo melhorar a análise e classificação do conjunto de documentos.

Normalmente os documentos estão escritos utilizando linguagem natural, baixo nível, sendo assim o documento é descrito por um conjunto de palavras. O primeiro passo da normalização é remover as palavras que aparecem em excesso no corpo do texto e que não possuem grande importância (preposições, artigos, conjunções etc). Assim, pode-se dizer que depois dessa primeira etapa teremos as palavras-chave. A próxima etapa é a classificação automática das classes de palavra-chave (Van RIJSBERGEN, 1979 apud MARCELLO ERICK BONFIM,2006).

O objectivo é reduzir as variações a única base ,de modo a fazerem parte do Vocabulário

- Tematização (Lemmatization)
- Termos em forma nominal
- Evitar preposições, excepto as muito usadas
- Limitar o uso de adjectivos
- Eliminar diferenças: maiúsculas/minúsculas, gênero, número, abreviações, iniciais, acrônimos, pontuação

Construção: Manual

Automática (métodos estatísticos, métodos linguísiticos)

semiautomática (técnicas de I. A.)

mesclarem de thesauri existentes

Ex:

Quer casar U.S.A. com USA

1. Um termo é uma palavra (normalizada), que é uma entrada para o dicionário de um sistema de recuperação de informação
2. Classes de equivalência são definidas para os termos, por exemplo:

- Remover os pontos
- U.S.A.,USA ↳ USA
- Remover os hífens
- Para-choques, parachoques ↳ parachoques
- Fui, era,sou → ser
- Carro, carrinho, carrão, carros→carro

Letras Maiúsculas - É prática comum converter todas as letras maiúsculas em minúsculas. Esta conversão permite que palavras começadas por letra maiúscula não sejam compreendidas pelo software como palavras diferentes quando estão escritas completamente em minúsculas.

Stemming e Lemmatization - O objectivo das duas técnicas é precisamente o mesmo, ou seja, reduzir palavras que se encontram em formas derivadas para a sua forma base. Exemplo disso é a transformação das formas "carro", "carrinho" para a sua forma base "carro". A diferença reside no facto do stemming se tratar de um processo heurístico que simplesmente corta as extremidades das palavras na tentativa de alcançar na maioria das vezes o objectivo pretendido. Quanto à Lemmatization, este procura atingir o objectivo com o uso de vocabulário e análise morfológica daspalavras.

Construção de thesaurus

"Tesauro é uma lista estruturada de termos associada empregada por analistas de informação e indexadores, para descrever um documento com a desejada especificidade, em nível de entrada, e para permitir aos pesquisadores a recuperação da informação que procura", (CAVALCANTI, 1978. Apud JEROCIR BOTELHO MARQUES DE JESUS,2002)

Enquanto a formação de frases junta termos muito genéricos para obter expressões mais específicas, a construção de thesaurus procura expressões mais genéricas para termos muito específicos.

Para a construção de um Tesauro, é necessário examinar seus elementos e seleccionar aquele que produzirá um adequado desempenho para um sistema específico. Assim, para assegurar a recuperação de um número desejável de documentos relevantes (revocação) e garantir uma selecção mais precisa (precisão), devemos fazer um controle da terminologia, que delimite os meios pelos quais se poderá expressar idéias, não necessariamente estabelecer limites, mas sim, regras que permitam a expansão e efetividade do sistema, através de bom controle vocabular, que garanta efetividade nas relações entre perguntas e respostas.

O objectivo principal do tesauro é dar assistência ao utilizador (pesquisador ou indexador) de maneira que ele consiga encontrar o termo que represente um determinado significado para o que se procura, ou seja, com a ajuda do tesauro, o utilizador no momento da busca poderá identificar termos alternativos, que permitirá descrever a informação contida no document de forma mais adequada

Relações Hierarquicas – é-um

Subclasse (Automóvel- Veículo)

Instância (João da Silva - Pessoa)

Relação Hierarquicas – parte-de (meronímia)

Agregado (Motor - Automóvel)

Conjunto (Estudante - Turma)

Relação Horizontais

Sinonímia (Carro - Automóvel)

Antonímia (Empregado - Desempregado)

Relacionado (Carro - motorista - estrada)

Ranqueamento

A recuperação de informação em textos precisa de ser classificado, em função da relevância dos documentos.

Esta classificação gera um índice que reflete a relevância do documento no contexto da busca. Este processo é descrito por LANCASTER (1985 1987 apud Amorin). O

Ordenação (ranking)

- Atribuir um valor (score) no intervalo [0,1] que representa o nível de "match" entre uma interrogação e um dado documento
- Como calcular este valor?

Cálculo do grau de semelhança entre documentos "score":

Medida de Jaccard:

jaccard(A,B)= |A ∩ B| / |A ∪ B|, onde A e B são osconjuntos de termos dos respectivos documentos.

- Não tem em conta a frequência dos termos;
- Cada termo/palavra tem a mesma importância.

Durante vários anos os sistemas de banco de dados existiam para buscar dados estruturados, e a recuperação de informação se restringia à busca de documentos simples e previamente ordenados. Atualmente, com crescimento rápido na quantidade e diversificação dos tipos de informação, as páginas na web se proliferam sem nenhum controle de qualidade ou custos de publicação. Algoritmos de ranqueamento são utilizados para melhorar os resultados da busca capturando as informações relevantes, desempenhando um papel fundamental nos engenhos de busca.

Problemas com recuperação sem classificação

1. Os utilizadores querem analisar poucos resultados – e não milhares.
2. É muito difícil escrever consultas que resultem em poucos resultados.

- Classificação é importante porque ela reduz um grande conjunto de resultados para um conjunto muito menor

3. "Utilizadores querem analisar poucos resultados"
4. Na verdade, na grande maioria dos casos, os utilizadors só examinam 1, 2 ou 3 resultados.

5. Leitura dos resumos: Os utilizadors estão a procura de documento mas não estão com paciencia para consultar todos o s resumos sendo assim preferem ler os resumos dos resultados mais bem classificados (1, 2, 3, 4) do que os resumos dos resultados com classificação inferior (7, 8, 9, 10).

6. Mesmo que o resultado número 1 não seja relevante, 30% dos utilizadors clicaram nele.

- Classificar de forma correcta é importante.
- Classificar corretamente os melhores resultados é ainda mais importante

Arquivos Invertidos

Um arquivo invertido é um arquivo de índices, que contém duas partes: um Vocabulário ou Dicionário, contendo um conjunto de palavras distintas do texto, e uma Lista de Ocorrências, indicando para cada termo do vocabulário, em quais documentos de uma determinada colecção este termo (palavra) ocorre, indicando, ainda, qual a frequência do termo em cada um destes documentos. Para um melhor entendimento do que vem a ser um arquivo invertido, será utilizada, como exemplo, a colecção de documentos textuais apresentada a seguir:

Documento	Texto
DOC1	A casa da mãe
DOC2	Casa da mãe
DOC3	Mãe da casa
DOC4	A casa da mãe
DOC5	A casa mãe da casa da mãe

Sendo a Tabela em cima, a representação de uma coleção de documentos com seus respectivos conteúdos, pode-se criar índices para uma eficiente recuperação de informação desses documentos, usando para a indexação, a técnica de arquivos invertidos. A Figura 2.1 representa esta técnica, destacando cada palavra existente na coleção de documentos da Tabela 2.1, seguida de seu respectivo número de ocorrência em cada documento.

Termos

TERMO 1	CASA
TERMO 2	MÃE

Forma original

	TERMO 1	TERMO 2
DOC1	1	1
DOC 2	1	1
DOC 3	1	1
DOC 4	1	1
DOC 5	2	2

Forma invertida

	DOC1	DOC 2	DOC3	DOC4	DOC5
Termo 1	1	1	1	1	2
Termo 2	1	1	1	1	2

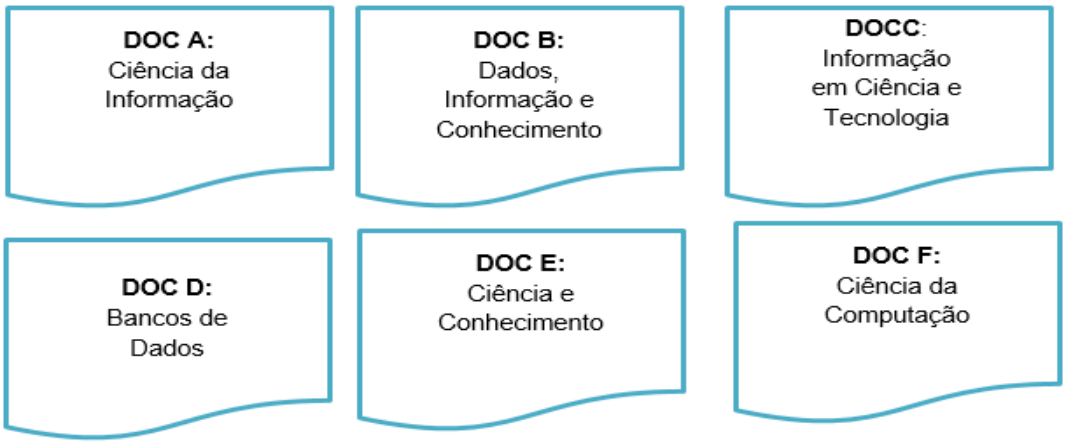
DOCUMENTOS	ID	TERMOS	POSIÇÃO DO DOCUMENTO
DOC 1	1	MÃE	1,

Um arquivo invertido é um índice ordenado de palavras-chaves, com cada palavra-chave contendo encadeamentos para os documentos que as contém.

Extensões

- Distâncias entre termos
- Pesos de termos
- Sinônimos
- Termos truncados (sufixos, prefixos ou infixos)

Exemplo



Elimine as stopwords e construa o arquivo invertido para os documentos

DOCUMENTOS	TEXTOS
DOC A	Ciência da Informação
DOC B	Dados,Informação e Conhecimento
DOC C	Informação em Ciência e Tecnologia
DOC D	Bancos de Dados
DOC E	Ciência e Conhecimento
DOC F	Ciência da Computação

Stoplist

Da | E | Em | de |

DOCUMENTOS	TEXTOS
DOC A	Ciência da Informação
DOC B	Dados,Informação e Conhecimento
DOC C	Informação em Ciência e Tecnologia
DOC D	Bancos de Dados
DOC E	Ciência e Conhecimento
DOC F	Ciência da Computação

Arquivo Invertido

Nº	TERMO	OCORRÊNCIAS (Nº de vezes)	DOCS(LISTA DOS DOCUMENTOS ONDE O TERMO APARECE)
1	Bancos	1	D
2	Ciência	4	A,C,E,F
3	Conhecimento	2	B,E
4	Dados	2	B,D
5	Informação	2	A,B,C
6	Tecnologia	1	C

Arquivos de assinatura

É uma forma extremamente compacta de caracterizar um texto por meio de uma "assinatura".

Assinatura = um bitstring que caracteriza uma palavra-chave -> um bloco -> um document
X uma consulta

Trará uma alternativa aos arquivos invertidos para recuperação de informações, pois requerem menos espaço de armazenamento (em torno de 10-20 % do arquivo principal, comparado com os 50-300% dos arquivos invertidos)

- Contém as "assinaturas" dos registos armazenados no arquivo principal.

Preliminares

Atributos Binários

Em muitas situações, um único bit é suficiente para representar uma informação.

Ex: Arquivo com atributos de Pessoas

Ex: Um arquivo de recuperação de documentos em que os documentos aparecem em um eixo e as palavras-chave no outro

Ex: Os ingredientes de receitas de sobremesa como atributos binários dos registos

O processo de recuperação dos registos desejados consiste de 4 etapas:

1. Construir o vetor de sobreposição para a query desejada
2. Fazer o casamento desse vetor com as assinaturas de todos os registos no conjunto e identificar aqueles cujas assinaturas contém os bits do vetor da query
3. Recuperar os registos correspondentes no arquivo principal.
4. Checar os registos recuperados para garantir que elesrealmente atendem a query.
5. Obs.: Esta última etapa é necessária para evitar as ocorrências falsas (false hits, false drops). Lembrar que quando a informação é condensada, alguma coisa é perdida.

Abaixo estão as assinaturas de todas as receitas doexemplo.

Observe-se que caso desejemos recuperar as receitas que contêm chocolate aparecerá a receita de glazed pound cake que não contém aquele ingrediente no vetor original (false drop).

Recuperação de Infromação

Com base na definição de informação, que se baseia no conceito de dados,pode se concluir que existem diferentes tipos de informações,o que consequentemente exigem técnicas diferentes como nos mostra a tabela seguinte, para a recuperação da mesma.

Definição

"Recuperação da Informação se preocupa em encontrar objectos (normalmente documentos) de natureza desestruturada (tipicamente texto) que satisfazem necessidades de informação em grandes coleções de documentos (geralmente armazenados em computador)." (Manning et al. 2009 apud Leandro Balby Marinho http//www.dsc.ufcg.edu.br/~lbmarinho)

"Recuperação da Informação se preocupa em representar, buscar e manipular grandes coleções de texto e outros dados da linguagem Humana." (Büttcher et al. 2010 apud Leandro Balby Marinho http//www.dsc.ufcg.edu.br/~lbmarinho)

Recuperação de Informação ou Information Retrieval (RI ou IR) lida com a representação, armazenamento, organização e acesso a itens de informação (documentos). A representação e a organização da informação deve dar ao utilizador de um Sistema de Recuperação de Informação (SRI) um acesso fácil a informação de seu interesse.

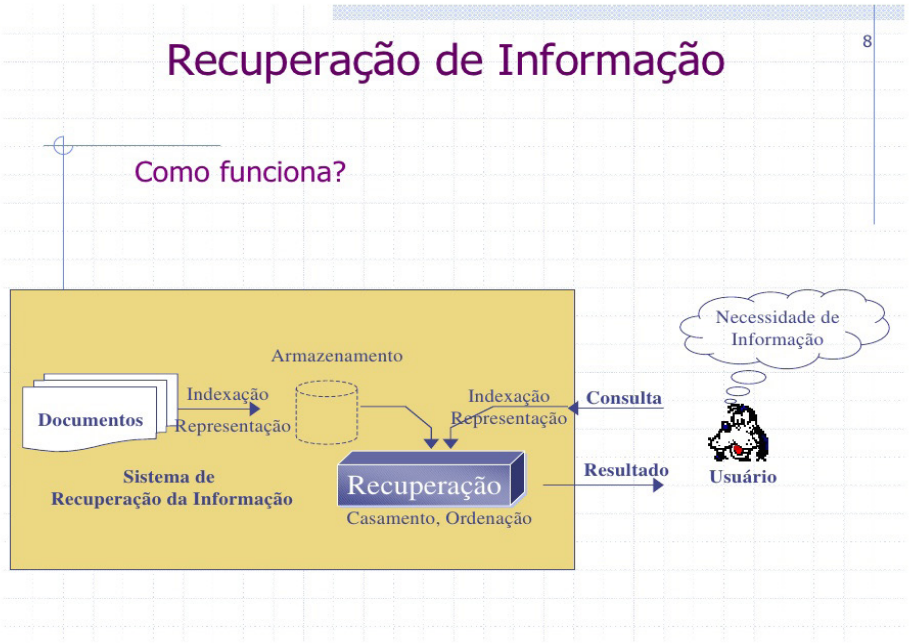


Figura 4: Recuperação de Informação

1° problema: Como caracterizar as necessidades de informação do utilizador?

Primeiro o utilizador deve traduzir o que deseja numa forma de consulta (query), que possa ser processada por um SRI. Na sua forma mais comum, esta consulta é escrita utilizando palavras-chave (ou termos de indexação) que resumem a descrição da necessidade de informação do utilizador.

Como um utilizador pode ter a certeza de que termos escolher?

Dada uma consulta, o principal objectivo do SRI é retornar informações úteis (relevantes) ao utilizador. A ênfase é na recuperação de informação e não na recuperação de dados.

Recuperação de dados num SRI consiste apenas em determinar que documentos de uma coleção contêm as palavras-chave que aparecem na consulta de um utilizador, e isto não é o suficiente para satisfazer as suas necessidades de informação, na maioria das vezes. O utilizador de um SRI prefere que informações sejam recuperadas sobre determinado assunto, que contenha os dados que aparecem na consulta. O SRI deve de alguma forma "interpretar" o conteúdo das informações encontradas nos documentos de uma coleção e ordená-los de acordo com um grau de relevância para o utilizador.

Relevância é a palavra central de um SRI. É objetivo do SRI recuperar todos os documentos que são relevantes a uma consulta de um utilizador e o menor número possível de documentos não relevantes.

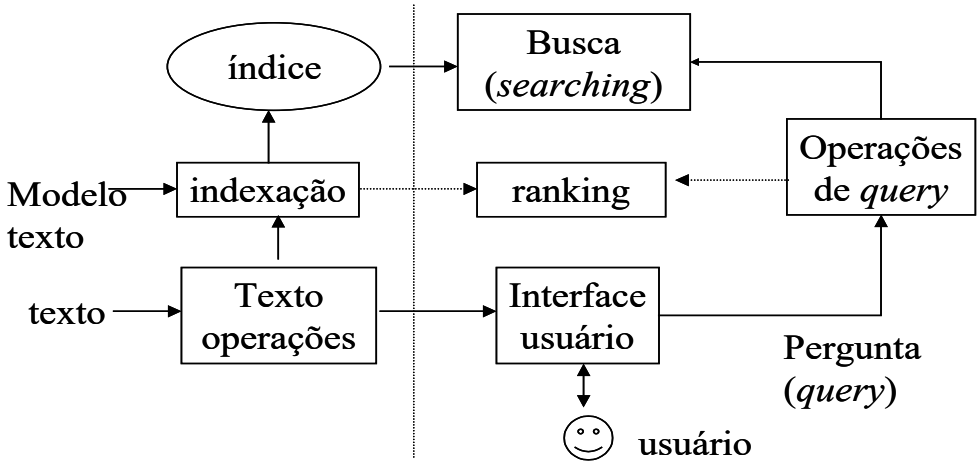


Figura 5: Arquitetura de um SRI

A recuperação eficaz de informações relevantes está directamente ligada tanto pela tarefa do utilizador, quanto pela visão lógica dos documentos.

O utilizador de um SRI tem que traduzir sua necessidade de informações em uma consulta, escrita na linguagem fornecida pelo sistema. Geralmente isto implica em especificar um conjunto de palavras que conduzam a semântica de sua necessidade. Neste caso, o utilizador está pesquisando por informações úteis executando uma tarefa de recuperação.

Existe uma distinção entre duas tarefas que podem ser executadas pelo utilizador de um SRI: a recuperação de informação ou a navegação entre documentos. SRI clássicos normalmente permitem recuperação de informação rápida. Sistemas de hipertexto são geralmente criados para permitir navegação rápida. Bibliotecas digitais modernas e interfaces para a web devem tentar combinar estas duas tarefas, entretanto esta ainda não é uma abordagem estabelecida e não é o principal paradigma desta área.

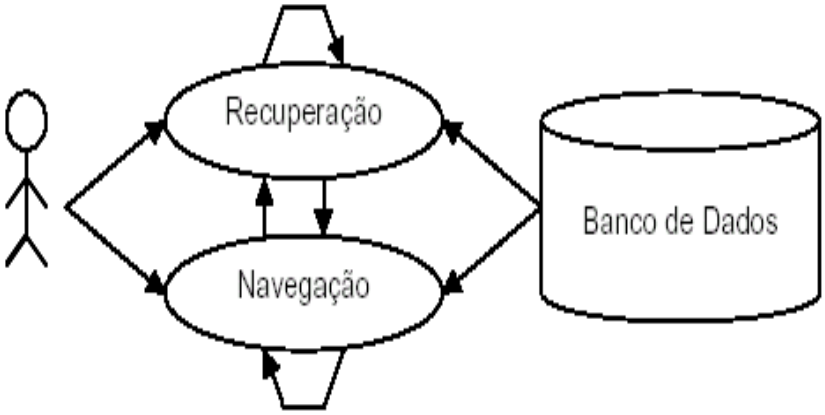


Figura 6:Tarefas dos SRI

O objectivo geral de um sistema de informação é minimizar o overhead para localização da informação para o utilizador. Overhead pode ser expresso como o tempo que o utilizador gasta em todas as etapas até encontrar a informação que procura.

Inclua nesse overhead:

- Tempo para geração da pergunta
- Tempo para execução da pergunta
- Tempo para pesquisar os resultados
- Tempo para organização da resposta para ser mostrada ao utilizador
- Tempo perdido para encontrar o resultado com leituras de documentos não relevantes

A tabela seguinte nos visualiza os sistemas que estão envolvidos na recuperação de informação, assim como os objectos de busca e as operações realizadas sobre os mesmos.

	OBJECTOS	OPERAÇÃO	TAMANHO
SRI	Documento	Recuperação (probabilistica)	Grande
SGDB	Registo	Recuperação (detrministica)	Grande
SBC	Regra	Inferência	Pequena

Mineração de texto (Text minning)

O processo de Mineração de Textos tem como objectivo recuperar informações textuais relevantes não estruturado.

Este processo envolve algum grau de dificuldades, porque as informações normalmente estão disponíveis em linguagem natural sem a preocupação com a sua estruturação de dados.

Essa área de sistemas de recuperação de informação surgiu com a finalidade de tratar os dados e as informações não-estruturadas considerando o alto nível de complexidade envolvida neste tipo de representação de informação. A mineração de texto utiliza técnicas de análise e extracção de dados a partir de textos, frases ou apenas palavras. Envolve a aplicação de algoritmos computacionais que processam textos e identificam informações úteis e implícitas, que normalmente não poderiam ser recuperadas utilizando métodos tradicionais de consulta, pois a informação contida nestes textos não pode ser obtida de forma directa, uma vez que, em geral, estão armazenadas em formato não estruturado.

A maior parte da informação disponível no mundo não está de forma estruturada e nem padronizada, armazenada em tabelas de bancos de dados relacionais. Ao invés disso, se encontra disponibilizada digitalmente como texto: livros, jornais, revistas, páginas Web, blogs, perfis de redes sociais, e-mails, arquivos PDF, documentos XML, arquivos JSON, etc. No final dos anos 90 esta situação foi percebida tanto por pesquisadores como pelas empresas. Mais ou menos nesta época surgiu a seguinte ideia: "que tal analisarmos estas 'montanhas de texto digital' para que novas informações sobre nossos clientes, fornecedores, produtos e serviços possam ser reveladas e, assim, utilizadas de forma estratégica em processos de tomada de decisões?".

Para começar, quando trabalha com dados textuais precisa lidar com informações que, na maioria das vezes, não possuem um esquema para descrever a sua estrutura. Ou seja, ao contrário do que acontece com os "bem-comportados" dados estruturados em tabelas relacionais, os dados textuais normalmente não estão organizados em campos, cada qual com seu tipo, tamanho e faixas de valores possíveis. Sendo assim, comparada com a informação gravada em SGBDs relacionais, a informação em formato texto é bem mais difícil de coletar, tratar, analisar e sumarizar.

FIGURA 2
Posição da mineração de textos no contexto da indexação manual

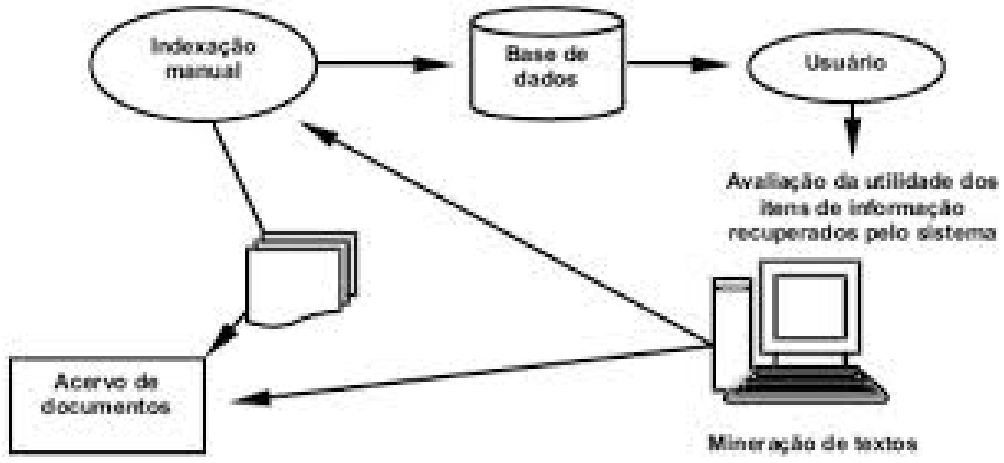


Figura 7:Mineração de textos no contexto de indexação Manual

Segundo o Text Mining Research Group, "Mineração de textos é a procura por padrões em um texto em linguagem natural e pode ser definido como o processo deanálise do texto para extrair informação dele para um propósito em particular".

A figura 8 nos exibe os dois mundos relacionados com sistemas de recuperação de informações

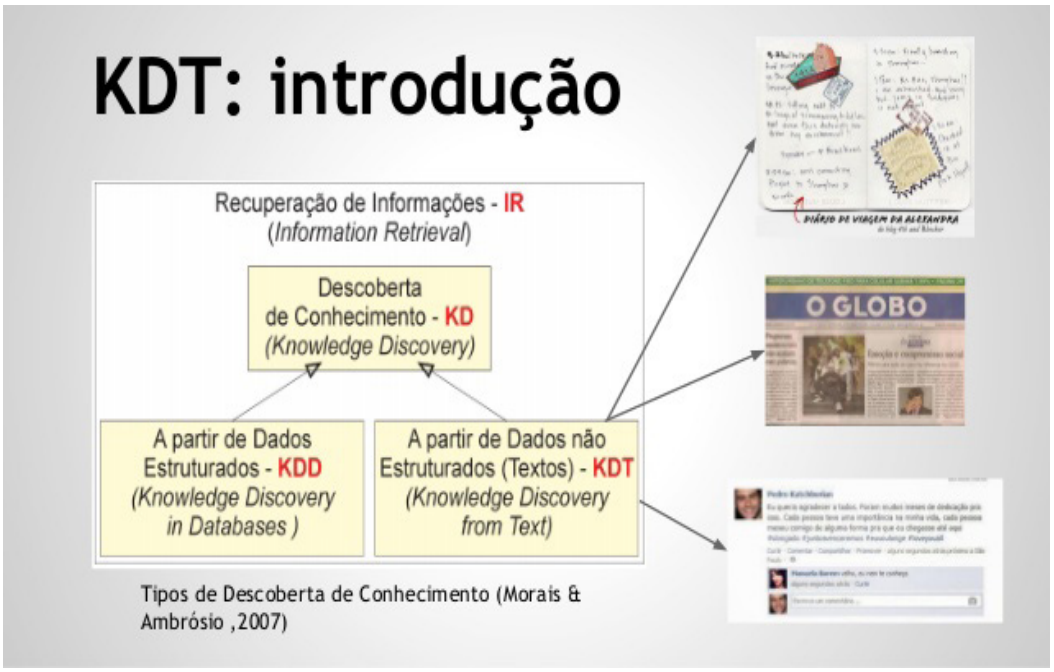


Figura 8: Tipos de descoberta de conhecimento

Etapas de mineração

Pré-processamento:

A etapa de pré-processamento diz respeito à limpeza dos dados para facilitar as análises da etapa seguinte. Esta etapa consiste na remoção do que for desnecessário para o entendimento do texto, o documento gerado é utilizado como base para a fase seguinte.

A etapa de pré-processamento pode ser dividida em três grandes etapas que são: a correcção Ortográfica, a remoção de Stopwords e o Stemming, cada uma dessas etapas deve ser aplicada, nesta ordem, no texto a ser pré-processsado, porém, vale ressaltar que nem todas as etapas são obrigatoriamente executadas

- Exclusão de palavras e números, baseada no tamanho, nas letras inicial e final ou outros critérios.
- Manutenção ou exclusão de palavras baseada em uma lista previamente definida.
- Identificação de sinónimos e antónimos.
- Determinação de radicais

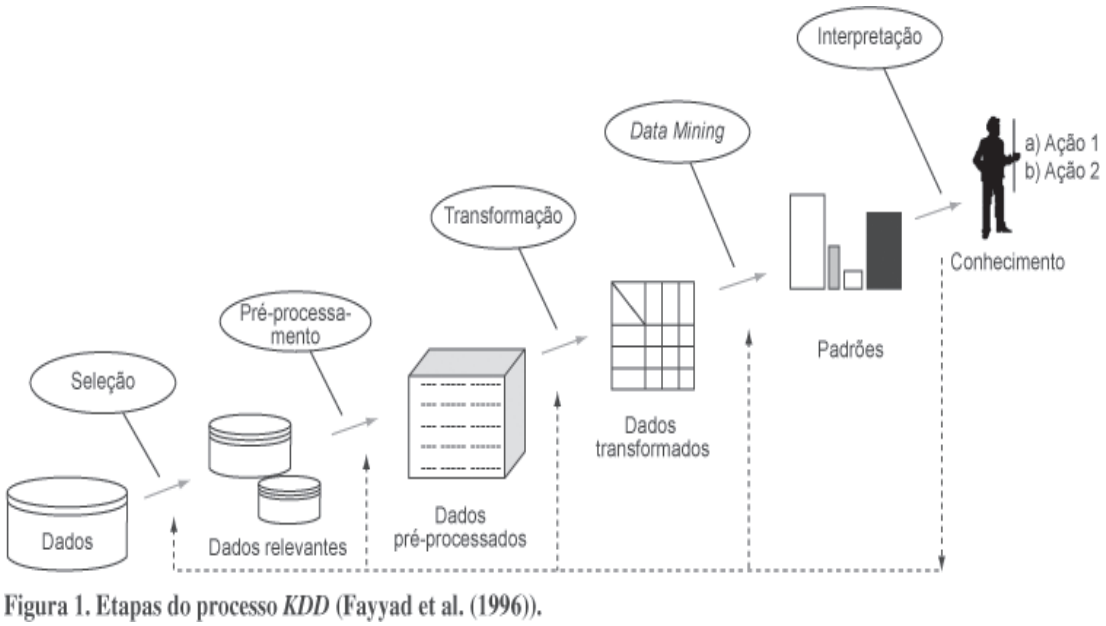


Figura 1. Etapas do processo *KDD* (Fayyad et al. (1996)).

Figura 9: Etapas do processo KDD (Fayyad et al.1996)

Avaliação da Unidade

Verifique a sua compreensão!

a. Relacione com um exemplo conceitos de dados, informação e conhecimento

b. Num índice invertido, qual a importância da ordenação dos documentos na lista invertida? Dê um exemplo.

c. Explique a importância de cada um dos seguintes procedimentos de indexação, determinando o efeito de cada método para aumentar o desempenho do sistema em termos de precisão e cobertura.

- Radicalização das palavras
- Uso dum dicionário de sinónimos
- Utilização da informação de localização das palavras
- Uso de peso dos termos

d. Normalize a frase utilizando o processo da filtragem de stopwors

Um sistema automático para RI pode ser visto como a parte do sistema de informação responsável pelo armazenamento ordenado dos documentos em um banco de dados,

e sua posterior recuperação para responder a consulta do utilizador.

Resposta

Um sistema automático para RI pode ser visto como a parte do sistema de informação responsável pelo armazenamento ordenado dos documentos em um banco de dados,

esua posterior recuperação para responder a consulta do utilizador.

Critérios:

Classificação gramatical (artigos, determinantes, pronomes, numerais, advérbios, preposições)

Palavras mais frequentes

Arquivo invertido

e. Dada a lista de documentos seguinte

DocID	Texto
d1	Cabo Verde é um pais Verde
d2	Verde, mesmo verde das chuvas
d3	Chuvas dão cabo de cabo verde num Cabo
d4	Chuvas e Verde Tempo da chuvas
d5	Cabo Verde de origem Verde das chuvas do outro Cabo de chuvas

Assumindo que a lista de palavras comuns (stopwords) é {or, and, for}, crie um índice invertido para o conjunto de documentos acima. O índice deve consistir das seguintes três peças: o dicionário, o conjunto de listas invertidas e o método de armazenamento do índice no disco. Cada registo de ocorrência no índice deve conter a identificação do documento (docID) e a frequência do termo nesse documento. Não se esqueça de indicar os pressupostos que utilizar.

Utilize por favor as tabelas apresentadas a seguir para as suas respostas.

Dicionário			Listas Invertidas
Palavra	Frequência	Localização	

f. Calcule o peso dos documentos

um novo documento d2 contém as palavras

- 'Petróleo' 18 vezes
- 'Refinaria' 8 vezes

na base de 2048 documentos ocorrem:

- 'Petróleo' em 128 deles
- 'Brasil' em 16 deles
- 'Refinaria' em 1024 deles

Teríamos os cálculos:

Vetor com tf : <'petróleo': 18, ' refinaria': 8>

Cálculo com tfitf ('petróleo')= 18*log(2048/128) = 18*1,2 = 21,6

tfitf ('Brasil')= 0

tfitf ('refinaria')= 8*log(2048/1024) = 8*0,3 = 2,4

Logo temos o vetor d2 = <21.6, 0, 2.4>

Um novo documento d3 contém as palavras

- 'Petróleo' 10 vezes
- 'Brasil' 10 vezes

Teríamos os cálculos:

Vetor com tf : <'petróleo': 10, 'Brasil': 10>

Cálculo com tfitf ('petróleo')= 10*log(2048/128) = 10*1,2 = 12

tfitf ('Brasil')= 10*log(2048/16) = 10*2,1 = 21

Logo temos o vetor d3 = <12, 21, 0>

Seja a consulta com

tf : <'petróleo': 1, 'Brasil': 1, ' refinaria': 1>

c = <1.2, 2.1, 0.3>

i=1..n pki X pci

sim(c,d1) = _____

i=1..n pki2 x i=1..n pci2

(1.2x4.8 + 2.1x16.87+0.3x3)

= ---------- = 0.97

(23.04+284.6+9) x(1.44+4.41+0.09)

Fazendo os cálculos para D2 e D3 temos: 0,50 e 0,99 resp.

Assim, a resposta à consulta seria ranqueada da seguinte forma:

1) D3, 0,99
2) D1, 097
3) D2, 0,50

Problemas com o Modelo Vetorial:

- Mudanças na base
- Os termos são independentes entre si
- Se um documento discute petróleo no Brasil e futebol na Argentina irá atender a uma consulta sobre petróleo no futebol.

g. Qual o objetivo de um sistema de recuperação de informação?

h. Para um sistema de recuperação de informação, o que significa overhead? O que pode gerar o overhead?

i. Esboce um plano para a recuperação da informação, (gerando assim Stopwords,Stemming,llatezation) como forma de normalizar o Vocabulario para o

processo de indexação

j. O sistema de Recuperação de Informação trousse á informática e a Tecnologia de Informação vento Informacional.

k. Esboce um fluxograma para indexar um documento?

R-1. Qual a diferença entre dado, informação e conhecimento?

Dados: são fatos e valores que isoladamente não tem significado.

Informação: consiste nos dados interpretados num dado contexto.

Conhecimento: Gerado ou adquirido a partir da informação, permite a tomada de decisão (MARCHI, 2009).

Leituras e Outros Recursos

Introduction to Information Retrieval, by C. Manning, P. Raghavan, and H. Schütze. Cambridge University Press. (Disponível gratuitamente online) ACESSADO NOVEMBRO 2014

http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf Acessado Novembro2014

Managing Gigabytes, by I. Witten, A. Moffat, and T. Bell.

Information Retrieval: Algorithms and Heuristics by D. Grossman and O. Frieder.

Modern Information Retrieval, by R. Baeza-Yates and B. Ribeiro-Neto.

Finding Out About, by R. Belew.

Mining the Web, by S. Chakrabarti.

Unidade 1: Modelos de Recuperação de Informação

Introdução à Unidade

O propósito desta unidade é verificaros conhecimentos que você possui relacionados com este curso.

É de extrema importância ter em mente o algoritmo de funcionamento dos diferentes modelos de recuperação de informação existente. Isso permite aos utilizadores equacionarem os seus problemas em função das suas necessidades e dos resultados desejados.

Objectivos da Unidade

Após a conclusão desta unidade, deverá ser capaz de:

1. Definir modelo booleano
2. Definir modelo vectorial
3. Definir modelo probabilístico
4. Definir os modelos extendidos
5. Diferenciar os diferentes tipos de modelos
6. Identificar as vantagens e desvantagens de cada modelo
7. Diferenciar modelos de Recuperação de Informação
8. Analisar cada um dos modelos

Termos-chave

Modelo BooleanoModelo clássico de recuperação de informação baseado na teoria dos conjuntos

Modelo Booleano Extendido: Um modelo de recuperação de informação basado numa extensão do modelo booleano clássico. A ideia é a interpretação das unificações parciais como distâncias euclideanas representadas num espaço vectorial de termos de índice.

Modelo do Espaço Vectorial: Modelo clássico de recuperação de informação baseado na representação de documentos e interrogações como vectores de termos. O modelo pressupõe a independência entre os termos.

Modelo Probabilístico: Modelo clássico de recuperação de informação baseado na interpretação probabilística da relevância dum documento para uma dada interrogação.

Modelo de Listas não sobrepostas: Um modelo de recuperação de documentos estruturados através de estruturas de índices concretizados como listas não sobrepostas.

Modelo de Recuperação de Informação: Um conjunto de premissas e um algoritmo para pontuar documentos para uma determminada interrogação do utilizador.

Navegação (Browsing): Processo interactivo em que o utilizador está mais interassado em explorar e conhecer os documentos do que satisfazer uma necessidade específica de informação.

Necessidade de informação do utilizador: Uma frase em linguagem natural que especifique a necessidade de informação do utilizador.

Precisão: Medida de eficácia usada em RI que corresponde à fracção de documentos relevantes no total de devolvidos.

Relevância: Qualidade dos documentos que satisfazem a necessidade de informação do utilizador, explicitada na interrogação.

Unidade I Modelos RI

Modelos de Recuperação de Informação

Os modelos de recuperação de informação consideram que cada documento é descrito por palavras-chave chamadas de termos de indexação. Um termo de indexação é uma palavra cuja semântica ajuda a localizar os temas principais de um documento. Adjectivos, advérbios, conjunções são menos úteis como termos de indexação.

Segundo Baeza e Ribeiro (1999), dado um conjunto de termos de indexação para um documento, nota-se que nem todos os termos podem ser usados para descrever o conteúdo do documento. Não é uma tarefa fácil determinar a importância de um termo de indexação em um documento. Considerando uma colecção com cem mil documentos, uma palavra que aparece em cada um dos cem mil documentos é completamente inútil como um termo de indexação porque ela não trás somente documentos de interesse do utilizador. Por outro lado, uma palavra que aparece em cinco documentos é completamente útil porque se estreita o espaço dos documentos que interessam na pesquisa.

Segundo Baeza e Ribeiro (1999), os três modelos clássicos de recuperação de informação são: Booleano, Vetorial e Probabilístico; estes são responsáveis por recuperar os documentos relevantes utilizando um mecanismo de comparação entre a consulta e os documentos armazenados.

O modelo Booleano é um modelo que utiliza a expressão booleana (álgebra booleana) para conjugar ou formular a expressão da consulta por meios da operação entre as palavras chaves utilizando a álgebra booleana, AND (E),OR (OU) e NOT (NÃO) para junção das palavras chaves a serem formuladas na consulta como forma de comparar com a informação armazenada.

Para formular a consulta utiliza-se:

1. AND (E) - Usado quando se deseja a junção dos conjuntos. Ex. Computador e Disco
2. OR (OU) - Usado quando se deseja as duas ocorrências. Ex. Computado ou Disco.
3. NOT (NÃO) - Usado quando se deseja excluir uma ocorrência. Ex. Computador, mas não disco.

As informações recuperadas são identificadas como fruto da consulta formulada, de acordo com elementos relacionados e com a elaboração lógica da consulta.

Mas para documentos (corpus) com tamanho que contem um volume muito grande de informações de forma não estruturadas tem de recorrer uma outra técnica de recuperação. O que dificulta bastante o desenvolvimento de algoritmos eficientes nesses documentos.

A técnica mas ideal para recuperar informação neste caso seria o uso de Matriz Termo - Documento:

Matriz Termo - Documento:

Considere que o utilizador deseja realizar a seguinte consulta nas obras de Shakespeare:

- "Quais as obras contem os personagens Brutus, Caesar e não Calpurnia". Uma maneira de realizar essa consulta e utilizando o conceito Matriz Termo – Documento.

Suponha a matriz abaixo:

Considere que o utilizador deseja realizar a seguinte consulta nas obras de Shakespeare:

- "Quais as obras contem os personagens Brutus, Caesar e não Calpurnia". Uma maneira de realizar essa consulta e utilizando o conceito Matriz Termo – Documento.

Suponha a matriz abaixo:

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Propriedades:

Cada posição da matriz pode assumir valor de 0 (Ausência no documento) ou 1 (presença no documento);

1. As linhas representam os termos;
2. As colunas representam os documentos.

Para realizar a consulta "Brutus AND Caesar AND NOT Calpurnia" primeiramente recupera-se as linhas dos temos:

Exemplo:

(Brutus) 110100

(Caesar)110111

(Calpurnia) 101111 Complemento pois na expressão o utilizador deseja NOT

O resultado e a operação AND entre as linhas, o resultado obtido e 100100.

Observando o resultado obtém-se as obras Antony and Cleópatra (primeira coluna) e Hamlet (quarta coluna),isto é cada um bit representa uma coluna.

Vantagens do modelo booleano:

1. Expressividade completa se o utilizador souber exactamente o que quer.
2. É facilmente programável e exacto.
3. Fácil de entender e implementar
4. Modelo simples baseado em teoria bem definido

Desvantagens do modelo booleano:

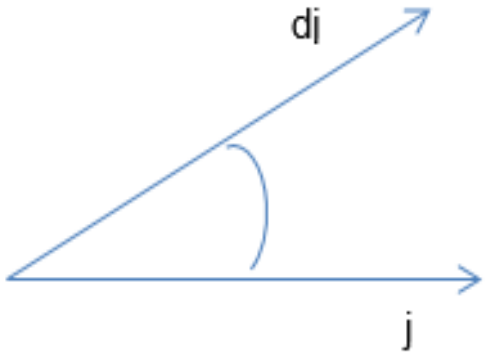
1. Pessoas lidam com conhecimento parcial.
2. Saída pode ser nula, ou haver overload (problema que dá um falso senso de segurança).
3. Não permite ordenação de documentos recuperados,Saída desordenada.
4. Nem todo o utilizador é capaz de formular a consulta
5. O modelo retorna ou poucos documentos ou demasiados documentos, dependendo da consulta

Modelo Vectorial:

O modelo espaço-vetorial (ou simplesmente vetorial) foi desenvolvido por Gerard Salton, para ser utilizado num SRI chamado SMART. No modelo vetorial cada documento é representado como um vetor de termos e cada termo possui um valor associado que indica o grau de importância (peso) para o documento. A consulta do utilizador também é representada por um vetor. Desta forma, os vetores dos documentos podem ser comparados com o vetor da consulta e o grau de similaridade entre cada um deles pode ser identificado. Os documentos mais similares (maispróximos no espaço) à consulta são considerados relevantes para o utilizador e retornados como resposta para ela.

O resultado da consulta é um vector construído através do calculo de similaridade baseado no ângulo (co.seno) entre o vector que representa o documento e o vector que representa a consulta. . (BAEZA e RIBEIRO, 1999 apud).

Neste modelo, o documento (dj) e a query (j) do utilizador são representados como um vetor t-dimentional.



Fonte: BAEZA YATES, R.; RIBEIRO NETO, B. A. Modern information retrieval. New York: ACM Press ; Harlow: Addison-Wesley, 1999. 513p.

Existe correlações entre os vetores dj e j, como mostra a figura x. Esta correlação pode ser quantificada pelo cosseno do ângulo entre os dois vectores.

onde (dj) são os documentos, (j) são as queries, (wi,j) é o peso de cada documento e (wi,j) é o peso de cada query.

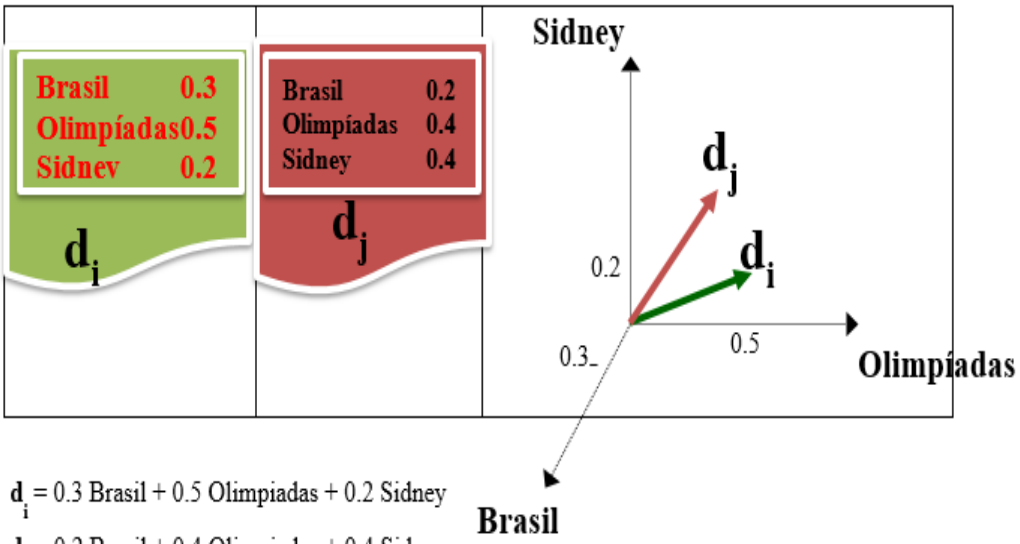
As principais vantagens deste modelo são: o esquema de pesagem de termo melhora o desempenho da recuperação; sua estratégia permite recuperação de documentos que aproximam às condições das queries, e o grau do cosseno permite ordenar os documentos de acordo com a semelhança com a questão.

No modelo vetorial cada documento é representado como um vetor de termos e cada termo possui um valor associado que indica o grau de importância (peso - weight) deste no documento. Ou em outras palavras, cada documento possui um vetor associado que é constituído por pares de elementos na forma {(palavra_1, peso_1), (palavra_2, peso_2),..., (palavra_n, peso_n)} .

D: um vetor

f : espaço vetorial t-dimensional e operações de álgebra linear sobre vetores

- As dimensões do espaço vetorial são os termos do documento
- Os termos recebem pesos de relevância no documento (negrito, título, etc)
- Esses pesos são usados como índices do vetor
- Modelo mais utilizado em IR



d_i = 0.3 Brasil + 0.5 Olimpiadas + 0.2 Sidney

d_j = 0.2 Brasil + 0.4 Olimpiadas + 0.4 Sidney

- **Sim: produto interno / produto das normas**

$$Sim = \frac{d_i * d_j}{|d_i| * |d_j|} = \frac{0,3*0,2+0,5*0,4+0,2*0,4}{(0,09+0,25+0,04)^{1/2}*(0,04+0,16+0,16)^{1/2}} = 0.28$$

Observações:

Documentos são representados como vetores no espaço de termos.

- Termos são ocorrências únicas nos documentos.
- Documentos são representados pela presença/ausência de um termo.
- Todos os termos combinados podem definir cada documento.

As consultas são representadas da mesma forma.

Aos termos das consultas e documentos são acrescentados pesos. Os pesos especificam o tamanho e a direção de sua especificação como vetor.

A distância entre vetores pode medir a sua "relação" com uma consulta.

O peso de um termo em um documento pode ser calculado de diversas formas. Os pesos são usados para calcular a similaridade entre cada documento armazenado e uma consulta feita pelo utilizador. Estes métodos de cálculo de peso geralmente se baseiam no número de ocorrências do termo no documento (freqüência).

Cada elemento do vetor de termos é considerado uma coordenada dimensional. Assim, os documentos podem ser colocados em um espaço euclidiano de n dimensões (onde n é o número de termos) e a posição do documento em cada dimensão é dada pelo seu peso.

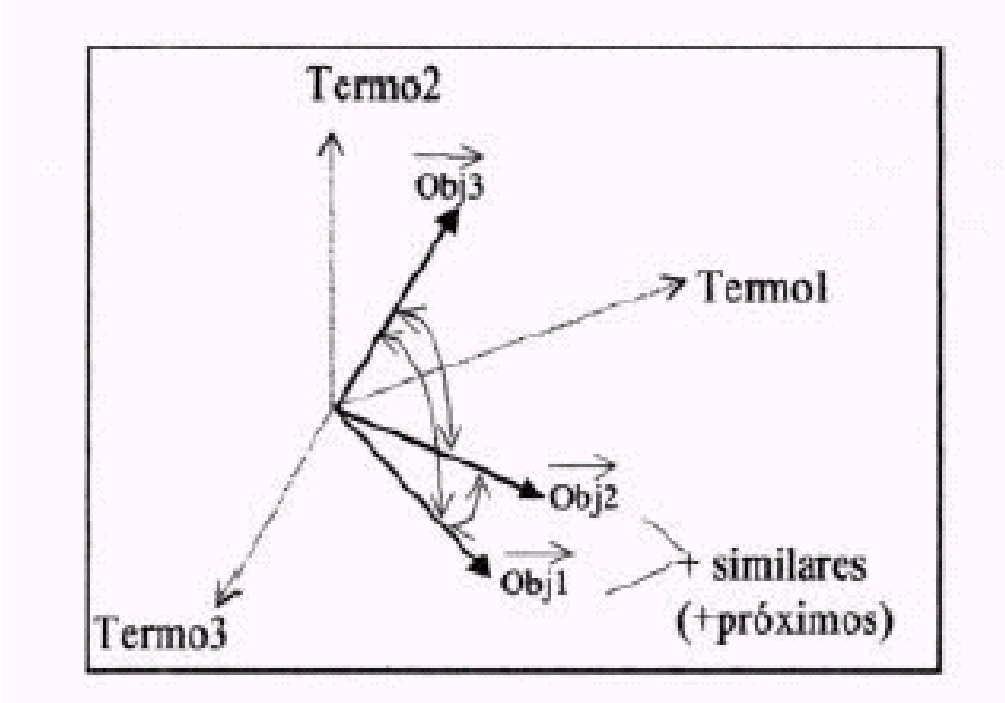


Figura 10:Modelo vetorial

As distâncias entre um documento e outro indicam seu grau de similaridade, ou seja, documentos que possuem os mesmos termos acabam sendo colocados em uma mesma região do espaço e, em teoria, tratam de assuntos similares.

A consulta do utilizador também é representada por um vetor. Desta forma, os vetores dos documentos podem ser comparados com o vetor da consulta e o grau de similaridade entre cada um deles pode ser identificado.

Os documentos mais similares (mais próximos no espaço) à consulta são considerados relevantes para o utilizador e retornados como resposta para ela.

Uma das formas de calcular a proximidade entre os vetores é testar o ângulo entre estes vetores. No modelo original, é utiliza uma função batizada de cosine vector similarity que calcula o produto dos vetores de documentos através da fórmula:

$$\text{similaridade}\,(Q,D) = \frac{\sum_{k=1}^{n} w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^{n} (w_{qk})^2 \cdot \sum_{k=1}^{n} (w_{dk})^2}}$$

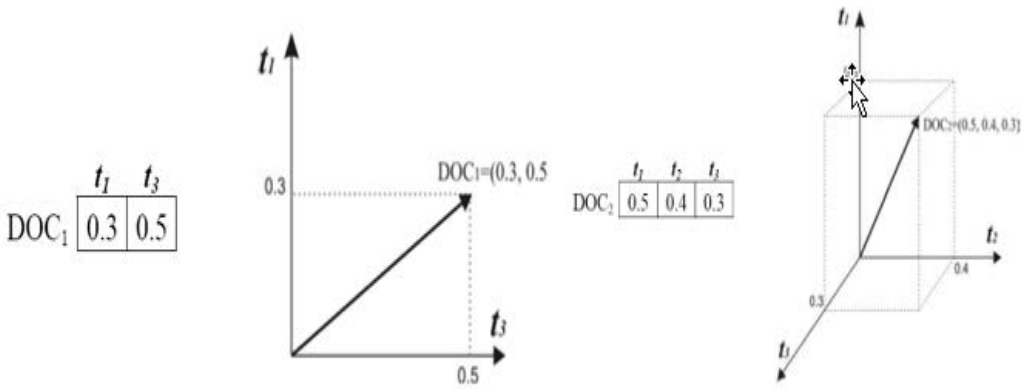
Onde,

- Q representa o vetor de termos da consulta;
- D é o vetor de termos do documento;
- Wqk são os pesos dos termos da consulta; e
- Wdk são os pesos dos termos do documento.

Depois dos graus de similaridade terem sido calculados, é possível montar uma lista ordenada (ranking) de todos os documentos e seus respectivos graus de relevância à consulta, da maior para a menor relevância.

Ex:

A representação gráfica de dois documentos: DOC1, com termos de indexação t1 e t3, com pesos 0.3 e 0.5, e DOC2 com termos de indexação t1, t2 e t3, com pesos 0.5, 0.4 e 0.3, dá-se:



Vantagens do modelo vetorial:

- Atribuir pesos aos termos melhora o desempenho.
- É uma estratégia de encontro parcial (função de similaridade), que é melhor que a exatidão do modelo booleano.
- Co-senoordena os documentos são ordenados de acordo seu grau de similaridade com a consulta.

Desvantagens do modelo vetorial:

- Ausência de ortogonalidade entre os termos, isto poderia encontrar relações entre termos que aparentemente não têm nada em comum.
- É um modelo generalizado.
- Um documento relevante pode não conter termos da consulta.

Cálculo dos Pesos:

O Método utilizado para calcular o pesoTF-IDF

Term Frequency (TF)- Frequência do termo no documento

Quanto maior, mais relevante é o termo para descrever o documento

Inverse Document Frequency (IDF)- Inverso da frequência do termo entre os documentos da coleção

Termo que aparece em muitos documentos não é útil para distinguir relevância

Peso associado ao termo tenta balancear os dois fatores

Term-frequency (tf) é Uma medida que utiliza o número de ocorrências do termo tj no documento di. Porém quando termos com alta freqüência aparecem na maioria dos documentos da coleção eles não fornecem informação útil para diferenciar documentos.

A medida inverse document frequency (idf) favorece termos que aparecem em poucos documentos da coleção. Tal medida varia inversamente ao número x dedocumentos que contem o termo tj em uma coleção de documentos e é definida como log (n/x). Baseado nessas duas medidas de freqüência pode-se definir a medida tfidf, combinando-as, como mostra a tabela a seguir:

Medida	Fórmula	Comentário
tf	$\#(t_j, d_i)$	$\#(t_j, d_i)$ é o número de vezes que o termo t_j ocorre no documento d_i.
$tfidf$	$\#(t_j, d_i).\log\frac{n}{x}$	as medidas tf e idf são combinadas, na qual x representa o número de documentos em D em que o termo t_j ocorre pelo menos uma vez.
$tfidfn$	$\frac{tfidf(t_j,d_i)}{\sqrt{\sum_{s=1}^{m}(tfidf(t_s,d_i))^2}}$	um fator de normalização é utilizado na equação $tfidf$ para que documentos de tamanhos diversos sejam tratados com a mesma importância.

Modelo Probabilístico:

Possui esta denominação justamente por trabalhar com conceitos provenientes da área de probabilidade e estatística. Neste modelo os termos indexados dos documentos e das consultas não possuem pesos pré-definidos. A ordenação dos documentos é calculada pesando dinamicamente os termos da consulta relativamente aos documentos. É baseado no princípio da ordenação probabilística (ProbabilityRanking Principle). Neste modelo, pesquisa-se saber a probabilidade de um documento ser ou não relevante para uma 8 consulta. Tal informação pode ser obtida assumindo-se que a

Segundo Baeza Yates e Ribeiro Neto (1999), o modelo probabilístico tenta

recuperar informações usando a teoria da probabilidade. Dado uma query do utilizador, há um conjunto de documentos que contém exactamente os documentos relevantes, e nenhum outro. Dado uma query, o retorno seria um conjunto de resposta ideal. Um grande problema, é saber as propriedades significativas dessa resposta ideal. Esta suposição permite gerar uma descrição probabilística preliminar do conjunto de resposta ideal, que é usada para recobrar o primeiro conjunto de documentos.

- Princípio da hipótese probabilística, segundo Baeza Yates e Ribeiro Neto (1999), revelam que: dado ao utilizador uma query (q) e um documento (dj) na coleção, o modelo probabilístico estima a probabilidade do utilizador achar o documento (dj) relevante. O modelo determina que esta probabilidade de relevância depende somente da query e da representação do documento. Além disso, o modelo determina que existe um sub-conjunto de todos os documentos que o utilizador pretende como resultado de sua busca para a query (q). O resultado ideal, que é denominado por (R), deve maximizar toda probabilidade relevante para o utilizador. Documentos no conjunto (R) são previstos ser relevante para a query. Documentos fora deste conjunto são previstos ser não relevantes.

  Esta hipótese é bastante preocupante porque não declara explicitamente como calcular a relevância das probabilidades. Na verdade, nem mesmo o espaço da amostra que é usado para definir tal probabilidade é dado. Dado uma query (q), o modelo probabilístico determina que para cada documento (dj), como uma medida desimilaridade para a query, a relação que calcula a probabilidade do documento (dj) ser relevante para a query (q).

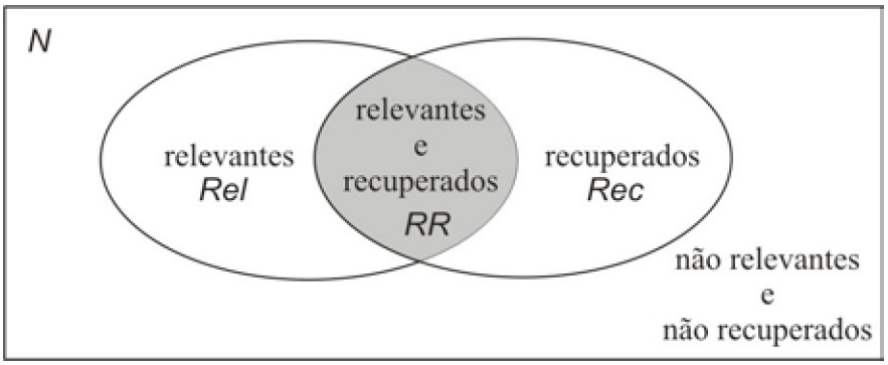
- Objetivo deste modelo é estimar Pq(dk/R), a probabilidade do documento dk, dentro do conjunto de resposta ideal (R), ser relevante para a questão q. Para calcular P q(dk/R), os SRI's deve de alguma forma, representar e armazenar os documentos. Os SRI's freqüentemente representam o documento por um conjunto de palavras conhecidas, os termos índices. No geral, os termos índices são aquelas palavras do documento que ficam no topo da listas (lista das palavras comuns). Estas palavras são obtidas freqüentemente, removendo-se os prefixos e sufixos.

  Inicia então, uma interação com o utilizador, com o propósito de melhorar a descrição probabilística do conjunto de resposta ideal. O utilizador analisa os documentos recobrados e decide os relevantes ou não. O sistema usa esta informação para refinar a descrição do conjunto de resposta ideal. Repetindo-se muitas vezes este processo, existiria então uma evolução e conseqüentemente uma aproximação do conjunto de resposta ideal. Assim, o utilizador deverá ter em mente no princípio, a necessidade, para prever o conjunto de resposta ideal.

- Modelo Probabilísticorepresenta o processo de recuperação de informação sob um ponto de vista probabilístico, ou seja, calcula a probabilidade de que o documento seja relevante para a consulta

Dada uma expressão de busca, podem-se dividir os N documentos de um corpus em quatro subconjuntos:

- Conjunto dos documentos relevantes (Rel)
- Conjunto dos documentos recuperados (Rec)
- Conjunto dos documentos relevantes e recuperados (RR) e
- Conjunto dos documentos não relevantes e não recuperados.



O resultado ideal de uma busca é o conjunto que contenha todos e apenas osdocumentos relevantes para o utilizador, isto é, todo o conjunto Rel.

Após obter os resultados da primeira busca, é possível melhorar os resultados a partir de interações com o utilizador.

Seja Rel o conjunto de documentos relevantes, e o complemento de Rel, a probabilidade de um documento d ser relevante em relação à expressão de busca é designada por p(Relld).

A similaridade (sim) de um documento d em relação à expressão de busca eBUSCA é definida como:

$$sim(d, eBUSCA) = \frac{\mathrm{p}(Rel \mid d)}{\mathrm{p}(\overline{Rel} \mid d)}$$

Vantagens:

Medida de similaridade: os documentos são retornados em ordem decrescente do seu grau de semelhança

Desvantagens

Necessidade de separar os documentos relevantes a priori

Métricas de Avaliação de Desempenho

Estudos na área de busca e recuperação da informação apontam para a necessidade de avaliar e comparar os diversos mecanismos de pesquisa para a Web, devido às características específicas de cada abordagem usada para criar esses mecanismos (MANDL, 2008 apud Barros Oliveira Ricardo) (AMENTO, TERVEEN et al., 2000).

Lancaster (1986, p. 131 apud Viera e Garrido) entende que, para avaliar a efectividade de um sistema de recuperação, devemos determinar o quão bem ele encontra as necessidades dos utilizadores, levando em consideração também critérios de performance como: qualidade (medida pela revocação e precisão), esforço (usabilidade do sistema ou até mesmo seu custo de uso) e tempo de resposta. Para o autor, todos os utilizadores de um sistema de recuperação de informação têm pré-requisito em comum: esperam que o sistema seja capaz de recuperar documentos relevantes que contribuam a satisfação de suas necessidades informacionais.

A avaliação de um mecanismo de RI é determinada pelo grau de satisfação da resposta á consulta lançada pelo utilizador, ou seja, se ele está satisfeito com o conjunto de resposta, o que significa que ele encontrou a informação que procurava. Para um mecanismo de recuperação de informação, existem duas medidas específicas mais comuns que devem ser utilizadas para avaliar o seu desempenho.

Estas medidas são conhecidas como revocação (recall) e precisão.

Medidas

Revocação mede a proporção de documentos relevantes que formam retornados como resultado a uma consulta do utilizador.

Precisão mede quantos documentos relevantes foram recuperados

Figura 11:: Revocação e precisão

Gráfico Precision-Recall

Esses dois parâmetros estão inversamente relacionados, significando que a melhoria de um implica na piora do outro. O gráfico abaixo, ilustra essa relação:

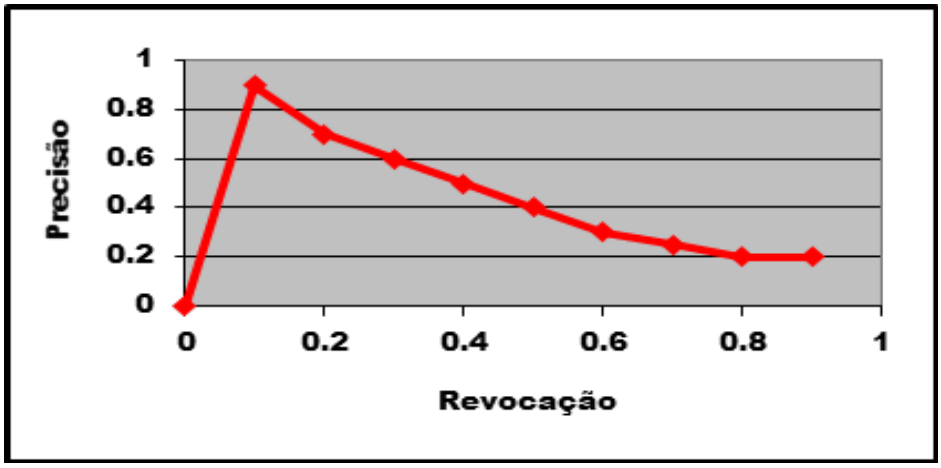


Figura 12: Precisão x Revocação

Tradicionalmente, existem dois factores muito importantes que governam a eficiência da indexação de um documento e consequentemente, influenciam na revocação e na precisão . São a exaustividade e a especificidade. Suas definições variam entre diversos autores, daí escolhemos uma para defini-los:

- Exaustividade: Define o número de diferentes conceitos (tópicos) que estão indexados
- Especificidade: Define o grau de precisão da linguagem de indexação em descrever um dado documento.

Precisão: a precisão é definida como sendo a razão entre o número de objectos relevantes que foram retornados pelo número total de objectos retornados.

$$\text{Precisão} = \frac{D_r}{D_t}$$

Revocação / Abrangência: é a razão entre o número de objectos relevantes que foram retornados pelo número de objectos relevantes que estão presentes na

base.

$$\text{Revocação} = \frac{D_r}{N_r}$$

Média-F – também conhecida como F-mean ou Média Harmônica, é a combinação entre Abrangência e Precisão (figura 5.3). Esta função retorna um valor no intervalo entre zero e um. Quanto mais próximo de zero, menos relevantes são os documentos e quanto mais próximo de um, mais relevantes são os documentos da base.

$$\text{Média-F} = \frac{2}{\frac{1}{abrangencia} + \frac{1}{precisão}}$$

É interessante ressaltar que de acordo com o objetivo de cada processo de descoberta de conhecimento, métricas de avaliação de desempenho diferente das citadas devem ser utilizadas. Por exemplo, uma tarefa de sumarização não será bem avaliada por medidas como abrangência, precisão ou média-F.

Avaliação da Unidade

Modelos de recuperação de informação

1. O que é o modelo de recuperação booleano?
2. Formule 6 interrogações booleanas. Tenha o cuidado de usar toda a variedade de operadores booleanos e de utilizar diferentes combinações destes elementos.
3. Para cada uma dessas 6 interrogações, indique quais são os documentos devolvidos em cada uma delas.
4. O modelo booleano é um modelo de recuperação com ordenação? Porquê? Indique um exemplo de um sistema de recuperação com ordenação.
5. Verifique como é que pode utilizar os operadores booleanos no Google. Por exemplo, escolha uma palavra, como informação, e submeta as seguintes
6. Interrogações

   (i) informação,

   (ii) informação AND informação,

   (iii) informação OR informação. Olhe para o número de resultados.
7. Fazem sentido em termos de lógica booleana? Consegue perceber a razão? E se tentar com palavras diferentes? Por exemplo, introduza as seguintes interro -gações (i) recuperação, (ii) informação, (iii) recuperação OR informação. Qual o limite que as duas primeiras interrogações colocam à terceira? Este limite é observado?
8. Dada a matriz de termos-documentos

	LIVRO DA RI	UVA
ACÇÃO	1	0
SONHO	1	1
SONÂMBOLO	0	0

Interrogação:

- Sonho AND acção AND NOT sonâmbolo
- 11 AND 10 AND NOT 00
- 11 AND 10 AND 11
- 10

Resposta: "Livro da RI"

4. Dada a expressão de busca com os seguintes pesos eBUSCA=(0.2,0.35,0.1), juntamente com os documentos DOC1 e DOC2, em um espaço vetorial formado pelos termos t1, t2 e t3, com os seguintes pesos: DOC1=(0.3,0.0,0.1) DOC2 (0.5,0.4,0.3)

1. Faça a representação gráfica do modelo vectorial?

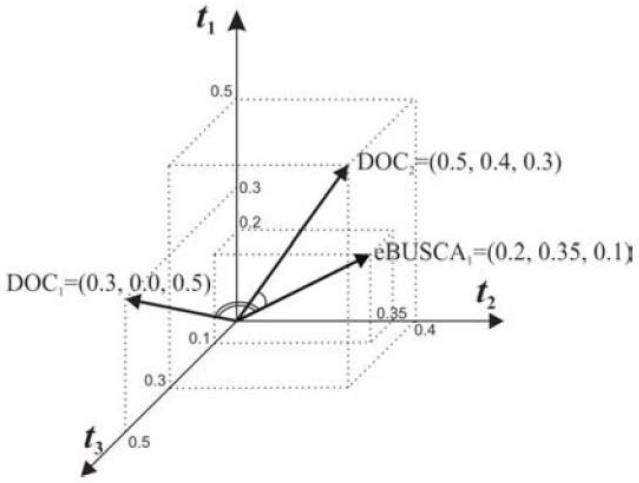
2. Calcula o grau de semilaridade?

Resposta:

1.

	t_1	t_2	t_3
$eBUSCA_1$	0.2	0.35	0.1

	t_1	t_2	t_3
DOC_1	0.3	0.0	0.5
DOC_2	0.5	0.4	0.3



2 .

$$sim(DOC_1, DOC_2) = \frac{(0.3 \times 0.5) + (0.0 \times 0.4) + (0.5 \times 0.3)}{\sqrt{0.3^2 + 0.0^2 + 0.5^2} \times \sqrt{0.5^2 + 0.4^2 + 0.3^2}} =$$

$$sim(\mathrm{DOC}_1, \mathrm{DOC}_2) = \frac{0.15 + 0.0 + 0.15}{\sqrt{0.34} \times \sqrt{0.5}} = 0.73$$

$$sim(\mathrm{DOC}_1, \mathrm{eBUSCA}_1) = 0.45 \quad (45\%)$$

$$sim(\mathrm{DOC}_2, \mathrm{eBUSCA}_1) = 0.92 \quad (92\%)$$

5. Construa o ficheiro invertido da seguinte colecção de documentos

$$d_1 = (t_1, t_2, t_6, t_7)\ d_2 = (t_1, t_2, t_3)\ d_3 = (t_1, t_2, t_5, t_6)$$
$$d_4 = (t_2, t_3, t_7, t_9)\ d_5 = (t_5, t_6, t_9)\ d_6 = (t_8, t_9)$$
$$d_7 = (t_7, t_8, t_9)\ d_8 = (t_4, t_6)\ d_9 = (t_2, t_3, t_4, t_5, t_6)$$

6. Usando os dados do exercício anterior identifique os documentos associados com cada uma das seguintes expressões booleanas

$$q_1 = t_1 \wedge t_2 \qquad q_2 = t_1 \vee t_3 \qquad q_3 = t_1 \vee (t_2 \wedge t_3)$$
$$q_4 = t_2 \wedge (\neg t_3) \qquad q_5 = t_2 \wedge (t_1 \vee t_3)$$

7. Considere a seguinte amostra de colecção de documentos

$$d_1 = (1,0,1,0,0,0)\ d_2 = (3,0,2,1,0,0)\ d_3 = (1,2,3,0,1,0)$$
$$d_4 = (0,0,0,2,1,2)\ d_5 = (1,1,1,4,2,1)\ d_6 = (1,1,0,0,3,2)$$

Gere a matriz de similaridade entre documentos usando a seguinte fórmula de similaridade

$$Sim(d_1, d_2) = \frac{\sum w_1 * w_2}{\sqrt{\sum w_1^2 * \sum w_2^2}}$$

Escolha um valor de corte para a similaridade e agrupe os documentos em aglomerados (clusters)

8. Considere as seguintes interrogações $i_1 = (2,0,2,0,0,0)$ e $i_2 = (0,0,0,2,0,2)$

Calcule o RSV (Retrieval Status Value) de cada par documento-interrogação usando para esse efeito a fórmula do coseno do Modelo do Espaço Vectorial e apresente os documentos por ordem decrescente da relevância estimada.

A fórmula do Coseno é usada na pergunta anterior para sim(D,I) em que $w_k = f_k * idf_k$, com $idf_k = \log \frac{N}{n_k}$

9. Tente pensar um pouco nos seguintes aspectos

Reclama-se que a resposta pontuada aumenta a eficácia da busca e a satisfação do utilizador.

- Justifique esta afirmação
- Embora os sistemas baseados no modelo booleano não produzam respostas ranked, porquê que algumas pessoas continuam a dizer que estes sistemas são bons?
- A formulação de interrogações booleanas e a tecnologias dos ficheiros invertidos não produzem directamente respostas ranked. Que métodos existem para produzir respostas pontuadas neste tipo de sistemas.

10. Que problema surge num sistema que usa a realimentação de relevância quando uma dada interrogação não devolve nenhum item relevante?

11. Explique os parâmetros Revocação e precisão - para que eles servem e como eles são calculados?

12. Cite vantagens e desvantagens do modelo booleano.

13. Cite vantagens e desvantagens do modelo vetorial.

14. Cite vantagens e desvantagens do modelo probabilístico.

Modelo booleano (VERDADEIRO OU FALSO)

– Peso igual a todos os termos VERD

O PESO DOS TERMOS DEVEM SER CALCULADOS FALS

PESO DOS TERMOS É IRRELEVANTE VER

– Não permite ordenação (ranking) VER

Leituras e Outros Recursos

Christopher D. Manning

Prabhakar Raghavan

Hinrich Schütze

Online edition (c) 2009 Cambridge UP

http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf Acessado Novembro2014

Unidade II- Base de Dados

1. Banco de dados
2. Definição;
3. Componentes dos sistemas de gerenciamento de banco de dados;
4. Modelos de banco de dados;
5. Projeto conceitual de banco de dados;
6. Projeto lógico de banco de dados;
7. Projeto físico de banco de dados;
8. Manipulação de dados

Objectivos :

1. Explicar os conceitos fundamentais relacionados aos sistemas de banco de dados
2. Listar os modelos de dados.
3. Projetar um modelo de recuperação de dados na base de dados
4. Linguagem de definição e manipulação de dados (SQL – DDL e DML);
5. Explicar as diferenças entre banco de dados e o processamento tradicional de arquivos
6. Compreender as diferentes arquiteturas de Banco de Dados existentes;
7. Aplicar a técnica de transição do Modelo Conceitual para uma arquitetura Relacional de BD;

Termos-chave

dados - fatos isolados que podem ser armazenados, ou valor do campo quando é armazenado no Banco de Dados ex: ex:nomes,telefones, endereços

Base de dados - coleção de dados interelacionados logicamente, ex: agenda de telefones

Sistema de Gerência de Bases de Dados (SGBD) - coleção de programas que permite a criação e gerência de bases de dados ou Sistema de Banco de Dados

SGBD (Sistema Gerenciador de Banco de Dados): é um software com recursos específicos para facilitar a manipulação das informações de um BD e o desenvolvimento de programas aplicativos. Exemplos: Oracle, Paradox, MySQL,

Access, Interbase, Sybase.

Entidade: A entidade é a representação de um objecto do mundo

Atributo: Os atributos são as características da entidade que desejamos guardar.

Domínio: podem ser numérico, texto, data e boleano. Podemos chamar também de tipo do valor que o atributo irá receber.

Campo: É a menor unidade de informação existente em um arquivo de banco de dados.

Registo: Conjunto de campos ou conjunto de atributos pertencente a uma entidade que identifica entrada única num banco de dados.

Chave primária: Um atributo (campo/ coluna) que permite identificar a entidade perante o seu micromundo,ou identicar os restantes campos pertencente a um atributo.

Tabelas: Representam as estruturas de armazenamento de dados dos sistemas Gestão da Base de Dados,compostos por colunas (Atribuots) e linhas (Registo).

SBD (Sistema de Banco de Dados): é um sistema de manutenção de dados envolvendo quatro componentes principais: dados, hardware, software e utilizadores.

GERENCIADORES DE BANCO DE DADOS; Conjunto de programas (software) para gerenciar (criando, modificando e usando) um banco de dados e garantir a integridade e segurança dos dados. São exemplos de "Sistemas Gerenciadores de Banco de Dados": MYSQL, SQL Server, Oracle, DB2, ADABAS etc.

Dado: factos em "bruto", ou seja, a mais pequena p qp porção de dados que o computador reconhece ex:pode ser um número, um caracter,...;

Campo: um caracter ou conjunto de caracteres com um significado específico como por exemplo um nome; significado específico, como por exemplo um nome;

Abstração:O processo de escolher as características mais importantes e reduzir um sistema a elas, deixando as menos importantes de lado (abstraindo-as)

ANSI (Instituto Nacional Americano de Padronização):

Organização americana que desenvolve padrões amplamente usados pela indústria da informática.

Arquivo:Conjunto de dados ou instruções armazenado em meio digital e identificado por nome.

Software: É uma sentença (Instruções) escrita em linguagem de computador, legível e interpretável por máquina. A sentença (software) é composta por uma seqüência de instruções (comandos) e declarações de dados, armazenável em meio digital.

SQL (Structured Query Language): Linguagem de Consulta Estruturada. É uma linguagem padrão de comunicação com sistemas gerenciadores de banco de dados (SGBD), utilizada para acessar sistemas de banco de dados relacional.

Hardware:É a parte física dos recursos computacionais, ou seja, o conjunto de componentes eletrônicos, circuitos integrados, placas etc. São exemplos de hardware o monitor, o mouse, o pendrive, as mídias etc.

Introdução

Banco de Dados

Um banco de dados é uma colecção de dados relacionados sobre uma determinada entidade (objecto) do mundo real ou sobre um domínio. Ex.: lista telefonica, Dicionários, Acervo bibliográfico. Os dados normalmente estão estruturados respeitando os princípios e a estrutura de SGDB de modo a produzir informação.

O conceito de Banco de Dados está intimamente relacionado com o mundoda recuperação de Informação.

Recuperação de Informação (RI) é a actividade de recuperar objetos informacionais armazenados em um meio acessível por computador. Um objecto informacional é geralmente constituído de texto, tais como documentos diversos, páginas Web e livros, embora possa conter outros tipos de conteúdo, tais como imagens, áudios, gráficos e figuras. A representação e organização desses objectos devem permitir às pessoas o acesso à informação relevante a partir da formulação da consulta em função de uma necessidade de informação.

Para que se possa representar esses objectos de forma a permitir a recuperação das informações armazenadas relacionada com esses objectos tem de se recorrer ao processo de modelação que permite estruturar os dados de forma a permitir o seu armazenamento e manipulação pelos sistemas de recuperação de informação e Sistemas de Gestão da base de dados.

Numa base de dados os dados estão estruturados de forma a facilitar operações como inserção, manipular e remoção.

Software que se utiliza para gestão da colecção dos dados chama se Sistema Gerenciador de Banco de Dados (SGBD). Normalmente um SGBD adota um modelo de dados, de forma pura, reduzida ou extendida. No nosso dia-a-dia,o termo banco de dados é usado como sinônimo de SGDB.

O modelo de dados mais adotado hoje em dia ó o modelo relacional, onde as estruturas têm a forma de tabelas, compostas por linhas e colunas.

A primeira área de interesse de SGDB foca na estrutura e armazenamento de dados em forma de tabelas relacionadas,mas em texto sem formatos,de modo que esse dados são manipulados por meio de consultas utilizando linguagem SQL. Mas como existem outros tipos de dados,semi- estruturado (XML).Pode se constatar que SQL está mais virado para recuperação de informações em forma de textos,que obdece uma determinada estrutura. E isso não é o suficiente para recuperação de informação,uma vez que existe documentos Semi-estruturados.

1. O que se pode concluir é que o modelo relacional da base de dados relacional permite recuperar documentos
2. Textual
3. Classificação dos documentos
4. Consultas com limitações semânticas (metodos de recuperação classico (palavras de buscas)
5. Que os resultados não apresentam a formatação

O que se precisa é a recuperação de informação e com base na definição de dados pode se concluir que existe diferentes tipos e formatos de informação.

Para que se possa recuperar informações precisa se:

- Classificação automatica de documentos
- Linguagem mais poderosa
- Metodos de recuperação de informação
- Apresentação dos resultados classificados em função da relevancia-

Razões:

- A qualidade de informação depende em grande medida da forma como está catalogada
- O esforço a ser realizado é demasiado e não pode ser exaustivo
- A linguagem da utilizada para consulta não é apropriado

As informações recuperadas utilizando linguagem SQL tem algumas limitações tais como:

- Classificação Manual dos documentos
- Consultas com semanticas limitadas
- Metodos classicos de recuperação (busca por palavras)
- Apresentação dos resultados sem processamento

E como estáperante o processo recuperação de informação não se pode limitar a documentos em formato de texto. O que precisa é um sistema de recuperação de informação,isto é:

- Classificação automatica dos documentos
- Linguagem da consulta mais poderosa
- Métodos de recuperação de informação
- Apresentação dos resultados da consulta formatada

Base de Dados Relacional

É uma base de dados estruturado unicamente na forma de tabelas múltiplas, que podem ser reunidas através da utilização de um campo em comum.

Com o avanço das tecnologias web e o aumento de volume das informações sobretudo com a digitalização dos documentos nas instituições a base de dados passou armazenar, classificar, organizar grande quantidade de informação que é de extrema importância para a vida das empresas. Mas para que isso aconteça tem de se recorrer a um Sistema de Gestão da Base de Dados (SGDB) que é um software que permite a gestão e a manipulação dos dados como forma de poder utilizar as operadores para formular a consulta que traduz o desejo do utilizador final e recuperar a informação necessária. Logo, um BD é uma fonte de onde se podem extrair informações derivadas, que possui um patamar de interacção com eventos como a globalização e tecnologia que representa (HENRY, 2006 apud Albuquerque).

Para a recuperação de informação em bases de dados, o uso dos recursos como ferramentas de busca e operadores booleanos, são instrumentos indispensáveis no momento da pesquisa, tanto na pesquisa livre, como na estruturada, seja ela realizada em bases de dados ou na Internet. Para o utilizador ter acesso às informações oferecidas pela base de dados é necessário que ele planeia a extretegia de busca utilizando diferentes ferramentas e operadores para relacionar termos ou palavras-chave numa expressão de busca.

Sem a presença de um sofware (DBMS) a gestão e a recuperação de informação numa base de dados é quase impossível para não dizer impossível.E no mundo concorrencial e das competitividades em que se vive hoje, em que a informação é o sistema nervosos de uma organização, sobretudo no que toca ao apoio á decisão.

Modelo de Dados

Um modelo de dados é caracterizado como um grupo de ferramentas conceituais que são utilizadas para descrever os dados; o relacionamento entre esses dados, e as normas de consistência. A sua função principal é buscar a organização de dados de estratégia de processamento que otimizem a performance de acesso e métodos de acesso que sejam independentes na ação (SILBERSCHATZ, 2006 apud Marluce Nunes Albuquerque,2010).

De acordo com a classificação de Silberschatz (2006 apud merluce 2010), os modelos de dados podem ser divididos em três categorias: modelo relacional, modelo de dados baseado em objeto, modelo de dadossemi-estruturado.

Sistema de Base de Dados

Um SBD possui como objectivos isolar o utilizador de detalhes mais internos do BD (abstração de dados) e prover independência de dados às aplicações (estrutura física de armazenamento e estratégia de acesso).

Pode ser conceituado um sistema de base de dados como o conjunto de quatro componentes básicos: dados, hardware, software e utilizadores. A Figura 13 ilustra os componentes de um sistema de banco de dados.

As vantagens da utilização de um SBD em relação aos sistemas tradicionais degerenciamento de arquivos são:

- Rapidez na manipulação e no acesso à informação;
- Redução do esforço humano no desenvolvimento e utilização das aplicações;
- Disponibilização da informação no tempo necessário;
- Controle integrado de informações distribuídas fisicamente;
- Redução da redundância e de inconsistência de informações;
- Compartilhamento de dados;
- Aplicação automática de restrições de segurança;
- Redução de problemas de integridade.

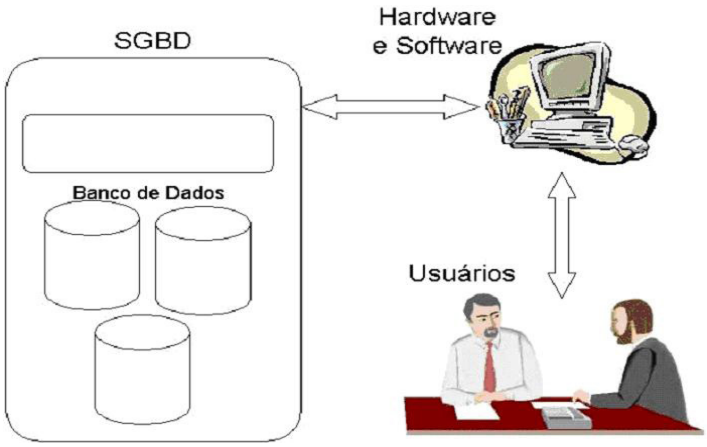


Figura 13: Componentes SBD (fontes Dvmedia)

Sistema de Banco de dados é um sistema de manutenção de registos utilizando computador envolvendo quatro componentes principais, sendo eles dados, hardware, software e utilizadores. O sistema de banco de dados pode ser considerado como uma sala de arquivos eletrônica. Existe uma série de métodos, técnicas e ferramentas que visam sistematizar o desenvolvimento de banco de dados.

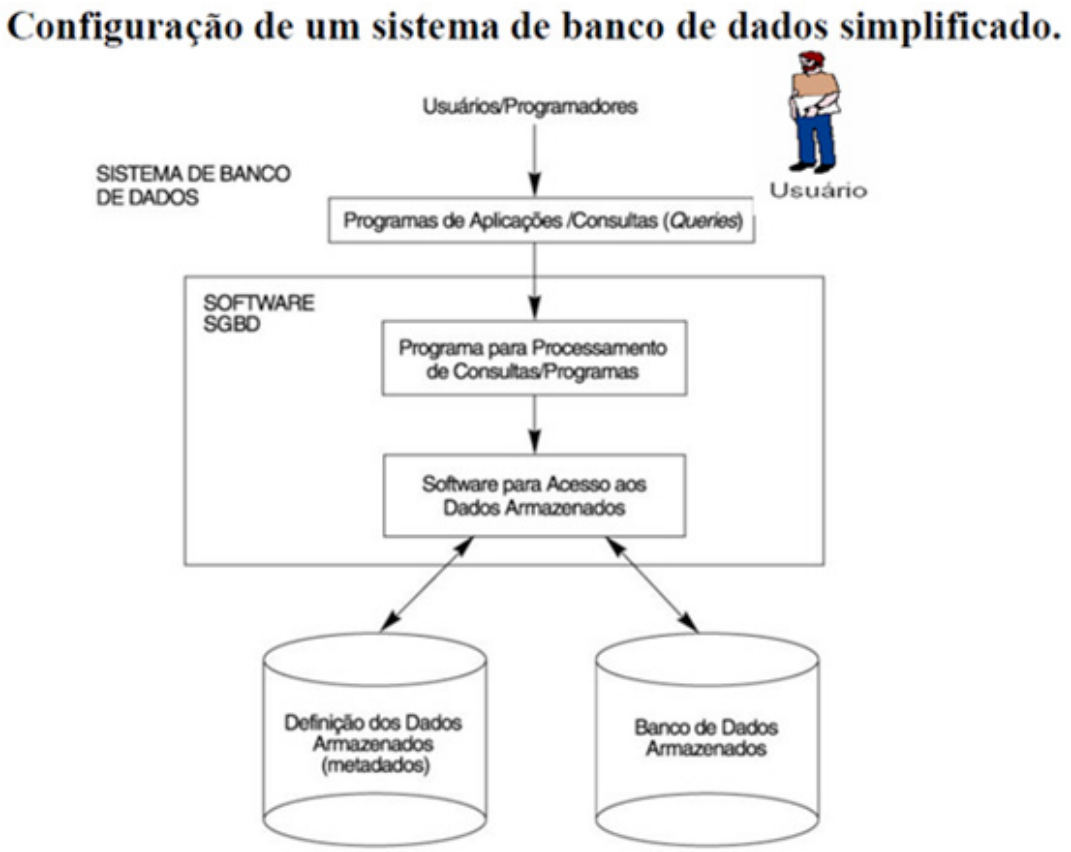


Figura 14:Componetes de Sistemas de Base de Dados

Os Objetivos de um sistema de banco de dados é isolar os utilizadores dos detalhes mais internos do banco de dados (abstração de dados) e também prover independência de dados ás aplicações (estrutura física de armazenamento e à estratégia de acesso).

As vantagens de utilizar um sistema de banco de dados é a rapidez na manipulação e no acesso á informação, redução do esforço humano (desenvolvimento e utilização), disponibilização da informação no tempo necessário, controle integrado de informações distribuídas fisicamente, redução de redundância e de inconsistência de informações, compartilhamento de dados, aplicação automática de restrições de segurança, redução de problemas de integridade.

Abstração de dados

O sistema de banco de dados (SBD) deve prover uma visão abstracta de dados para os utilizadores, isolando, desta forma, detalhes mais internos do BD. A abstracção se dá em três níveis como nos mostra a figura seguinte:

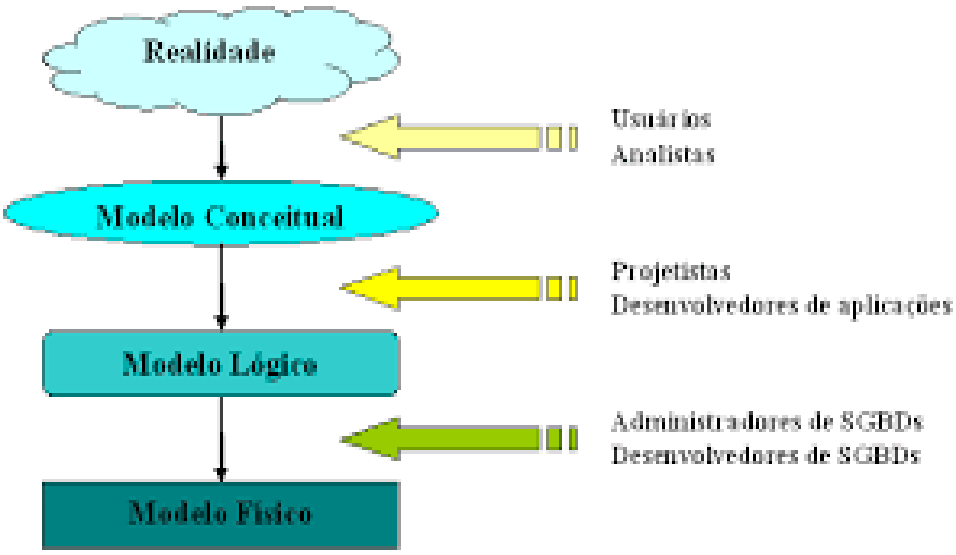


Figura 15: Modelo de Abstração de dados

Modelo conceitual de dados – Alto nível de conceitos, uma representação da realidade (Micro mundo), próximo de como o utilizador percebe a relação entre os objectos do seu micromundo, definindo assim quais os dados a serem armazenado e o grau de relacionamento entre os objectos.Por exemplo, modelo de Entidade-Relacionamento, modelo orientado a objecto,modelo semi-estruturado.

Modelo de dados de implementação (Lógico) – Conceitos que podem ser compreendidos pelo utilizador final, mas que não estão muito longe da organização de dados. Eles escondem alguns detalhes de armazenamento de dados, mas podem ser implementados em um sistema de computador de forma direta. Por exemplo, modelo hierárquico, modelo de rede, modelo relacional.

Modelo físico de dados – Conceitos de baixo nível que descreve detalhes de armazenamento dos dados,ou melhor efectivamente de que maneira os dados vão serem armazenados.

A maneira mais prática de classificar bancos de dados é de acordo com a forma que seus dados vão serem estruturados e armazenados. Diversos modelos foram e vem sendo utilizados ao longo da história, com vantagens para um ou para outro por determinados períodos.

Atualmente, a classificação mais comum citaria 5 modelos básicos:

- Modelo Hierárquico
- Modelo em Redes
- Modelo Relacional
- Modelo Orientado a Objectos
- Modelo Semi-Estruturado

Historicamente, o modelo de bancos de dados em rede foi implementado primeiro; porém o primeiro produto comercial usava o modelo de bancos de dados hierárquico, que nada mais é que uma versão simplificada do primeiro. Ambos os modelos foram resultado da pesquisa de usar mais efectivamente os novos dispositivos de memória secundária de acesso directo, que substituiam os cartões perfurados e as fitas magnéticas. Isso aconteceu na década de 1960.

Em 1970 E.F.Codd propôs o modelo de bancos de dados relacional que surgiu e ganhou destaque teórico imediato. Porém, a implementação do modelo exigia pesquisas e só na década de 1980 eles iam começar a ganhar o mercado, se estabilizando totalmente como líder do mercado a partir da década de 1990.

Podemos identificar o aparecimento do que pode ser chamado modelo plano (tabular) para fins mais diretos e simples.

No modelo relacional os dados estão estruturados em forma de tabela,que consiste em colunas,linhas e células.

Modelos Relacionais

Como o próprio nome exprime,trata se de um modelo que estrutura os dados em forma da tabela, de acordo com a relação existente entre os objectos do mundo real que fazem parte do micro mundo.

De acordo com a arquitetura ANSI / SPARC em três níveis, os bancos de dados relacionais consistem de três componentes:

1. Uma coleção de estruturas de dados, formalmente chamadas de relações, ou informalmente tabelas, compondo o nível conceitual;
2. Uma coleção dos operadores, a álgebra e o cálculo relacionais, que constituem a base da linguagem SQL;
3. Uma coleção de restrições da integridade, definindo o conjunto consistente de estados de base de dados e de alterações de estados. As restrições de integridade podem ser de quatro tipos:

    - Domínio (ou tipo de dados),
    - Atributo,
    - Restrições de base de dados.

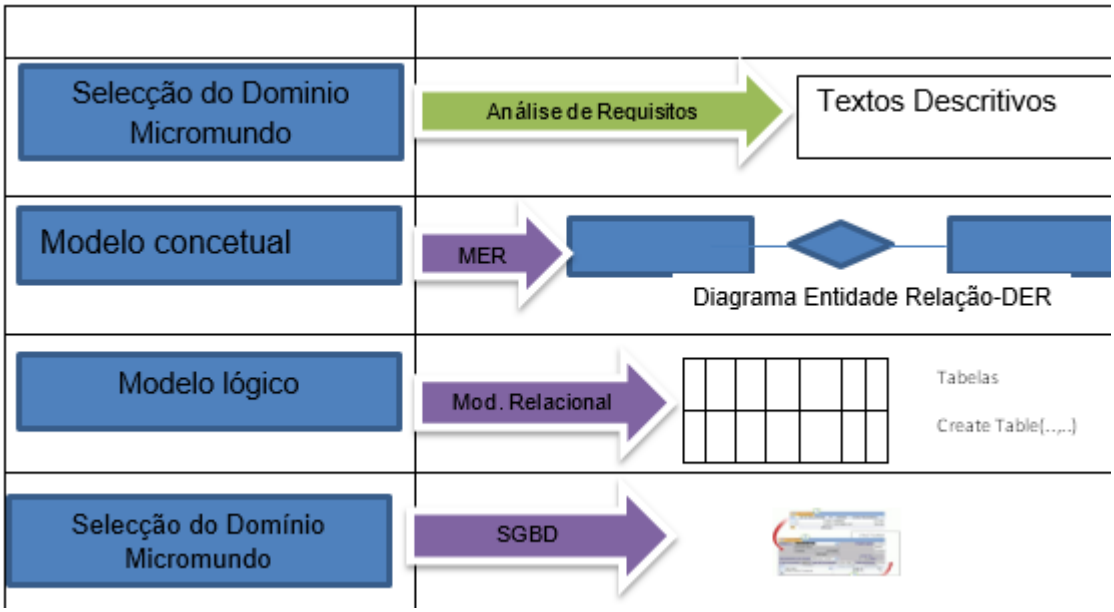


Figura 16: Relação entre Etapas de Modelação e Modelo Relacional

Modelo relacional

Relação ou entidade	Tabela
Atributo	Coluna
N Tupel-registo	Linhas

Diferentemente dos bancos de dados em rede, nos bancos de dados relacionais os relacionamentos entre as tabelas não são codificados explicitamente na sua definição. Em vez disso, se fazem implicitamente pela presença de atributos chave. As bases de dados relacionais permitem aos utilizadores (incluindo programadores) escreverem consultas (queries), reorganizando e utilizando os dados de forma flexível e não necessariamente antecipada pelos projetistas originais. Esta flexibilidade é especialmente importante em bases de dados que podem ser utilizadas durante décadas, tornando as bases de dados relacionais muito populares no meio comercial.

Um dos pontos fortes do modelo relacional de banco de dados é a possibilidade de definição de um conjunto de restrições de integridade.

Bancos de Dados Semi-Estruturados

Mais recentemente ainda, apareceram os bancos de dados semi-estruturados, onde os dados são armazenados independente do formato e que permite integração de dados de fontes heterogéneos. A estrutura dos dados depende da descrição dos objectos do mundo real e permitindo assim a representação de todos os tipos de dados.

Dados Semi-estruturados

Características principais dos dados semi-estruturados:

1. Sem imposição de tipos
2. Auto-descritivos:

- A descrição da estrutura está implícita nos dados, ou a descrição da estrutura acompanha os dados esquema conceitual:

3. Opcional descritivo, não prescritivo:

- Dados podem divergir da descrição do esquema

Caracteristicas principais do dados semi-estruturados

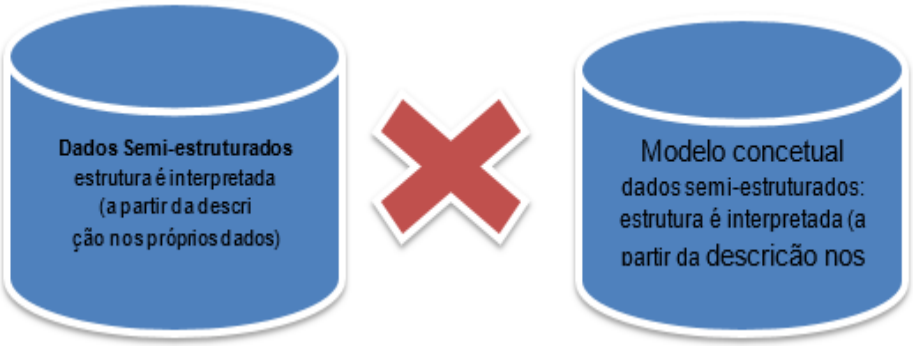


Figura 17:: Dados Semi-Estruturados versus dados estruturados

Indices

Todos os tipos de bancos de dados podem ter seu desempenho melhorado pelo uso de índices. O tipo mais comum de índice é uma lista ordenada dos valores de uma coluna de uma tabela, contendo ponteiros para as linhas associadas a cada valor. Um índice permite que o conjunto das linhas de uma tabela que satisfazem determinado critério seja localizado rapidamente. Há vários métodos de indexação utilizados comumente, como árvores B, hashes e listas encadeadas.

Modelo de base de dados (Bando de dados)

O modelo em rede permite que várias tabelas sejam usadas simultaneamente através do uso de apontadores (ou referências). Algumas colunas contêm apontadores para outras tabelas ao invés de dados. Assim, as tabelas são ligadas por referências, o que pode ser visto como uma rede. Uma variação particular deste modelo em rede, o modelo hierárquico, limita as relações a uma estrutura semelhante a uma árvore (hierarquia - tronco, galhos), ao invés do modelo mais geral direcionado por grafos.

Bases de dados relacionais consistem, mas de três componentes: uma coleção de estruturas de dados, nomeadamente relações, ou informalmente tabelas; uma coleção dos operadores, a álgebra e o cálculo relacionais; e uma coleção de restrições da integridade, definindo o conjunto consistente de estados de base de dados e de alterações de estados. As restrições de integridade podem ser de quatro tipos: domínio (também conhecidas como type), atributo, relvar e restrições de base de dados.

Diferentemente dos modelos hierárquico e de rede, não existem quaisquer apontadores, de acordo com o Princípio de Informação: toda informação tem de ser representada como dados; qualquer tipo de atributo representa relações entre conjuntos de dados. As bases de dados relacionais permitem aos utilizadores (incluindo programadores) escreverem consultas (queries) que não foram antecipadas por quem projetou a base de dados. Como resultado, bases de dados relacionais podem ser utilizadas por várias aplicações em formas que os projetistas originais não previram, o que é especialmente importante em bases de dados que podem ser utilizadas durante décadas. Isto tem tornado as bases de dados relacionais muito populares no meio empresarial.

Diferença entre modelos da base de dados

- O modelo relacional difere dos modelos hierárquico e em rede por não utilizar nem ponteiros nem links.
- Relaciona os registos por valores próprios a eles.
- Como não é necessário o uso de ponteiros, houve a possibilidade do desenvolvimento de fundamentos matemáticos para sua definição.

Os Bancos de Dados (BDs) são projetados para gerir grandes volumes de informações, e o gerenciamento dessas informações implica a definição de estruturas de armazenamento ede mecanismos de consultas adequados para recuperação de informações de um Banco deDados (BD) de forma eficiente.

Muitas vezes esta eficiência está relacionada com a complexidade interna dasestruturas de representação dos dados ou com as consultas submetidas para processamento.Não é esperado que o utilizador sempre escreva suas consultas de uma maneira eficiente, logo, éresponsabilidade do Sistema Gerenciador de Banco de Dados (SGBD) construir um plano deexecução que minimize o custo e maximize o desempenho das consultas.

Arquitetura da base de dados

Banco de dados é uma coleção de dados inter-relacionados, representando informações sobre um domínio específico.

Exemplos: Lista telefônica, controle do acervo de uma biblioteca, sistema de controle dos recursos humanos de uma empresa.

O modelo de sistema de computador nos quais os bancos de dados são executados écomposto por quatro categorias ou plataformas: as arquiteturas de banco de dadosCentralizados, Cliente/Servidor, Distribuído e Paralelo. Os quatro se distinguem,principalmente, no local onde realmente sucede o processamento dos dados.

Sistema De Gerenciamento De Dados DBMS

SGBD (Sistema de Gerenciamento de Banco de dados) é um software com recursos específicos que permite aos utilizadores a manipulação das informações, dos dados e o desenvolvimento de programas aplicativos. Exemplos : MS SQL Server , Oracle Database, IBM DB2, MySQL , PostgreSQL.

O processamento de consultas em um sistema de banco de dados é realizado quando o utilizador submete uma consulta a um SGBD, interativamente ou através de uma aplicação,utilizando uma linguagem de consulta de alto nível (SILBERSCHATZ, 2006). Este componente não trata somente da extração de informações de um banco de dados, mas também atividades com o objetivo de otimizar o processamento das consultas submetidas ao SGBD. Considerando a importância do processamento de consulta na recuperação das informações, vai ser os passos para descrever a estrutura de processamento e os passos de optimização de consultas realizados por alguns SGBDs.

Atualmente, o Banco de Dados é a tecnologia mais utilizada para armazenar dados e gerir informações para posterior recuperação ou atualização dessas informações por um utilizador, através de uma aplicação ou acessando um sistema de banco de dados.

Assim como os dados não constituem informação, a informação, isoladamente, não é significativa. Se os dados exigem processamento (classificação, armazenamento e relacionamento) para que possam realmente informar, a informação também exige processamento para que possa adquirir significado.

Características Gerais de um SGBD

1. Controle de Redundâncias
2. Compartilhamento dos Dados
3. Controle de Acesso
4. Interfaceamento
5. Esquematização
6. Controle de Integridade
7. Controle de Redundâncias:

A redundância consiste no armazenamento de uma mesma informação em locais diferentes, provocando inconsistências. Em um Banco de dados as informações só se encontram armazenadas em um único local, não existindo duplicação descontrolada dos dados. Quando existem replicações dos dados, estas são decorrentes do processo de armazenagem típica do ambiente Cliente-Servidor, totalmente sob controle do Banco de dados.

Interfaceamento:

Um Banco de dados deverá disponibilizar formas de acesso gráfico, em linguagem natural, em SQL ou ainda via menus de acesso, não sendo uma "caixa-preta" somente sendo passível de ser acessada por aplicações.

Esquematização:

Um Banco de dados deverá fornecer mecanismos que possibilitem a compreensão do relacionamento existente entre as tabelas e de sua eventual manutenção.

Controle de Integridade:

Um Banco de dados deverá impedir que aplicações ou acessos pelas interfaces pudessem comprometer a integridade dos dados.

Deste modo, os bancos de dados foram criados para operar em grandes quantidades de informação, propiciando um ambiente conveniente e eficiente para introdução, armazenamento, recuperação e gerenciamento das informações.

Esse gerenciamento envolve tanto a definição das estruturas de armazenamento da informação quanto o fornecimento de mecanismos para construí-los e manipulá-los. Além disso, um DBMS é responsável por manter a integridade e segurança dos dados armazenados e também para recuperação de informação se o sistema falhar. Um DBMS deve ser concebido em um sistema de multi-camadas, segundo relatório da ANSI/SPARC, como mostrado na figura 18.

A manipulação dos dados acontece por via de comandos (sintaxe SQL) bem definidas utilizando uma interface. Esses comandos são formulados utilizando a linguagem estruturada SQL. O utilizador utiliza uma interface para comunicar ao sistema de gerenciamento de dados SGDB que permite ao utilizador extrair as informações desejado da base de dados.

Mediante esta linguagem é possível dar ordens ao sistema para inserir,eleiminar, actualizar, seleccionar os dados da base de dados,o que por conseguinte permite recuperar informações. Por vezes,são desenvolvidas interfaces como o intuito de automatizar sistemas.

ARQUITECTURA DE TRÊS ESQUEMAS E A INDEPENDÊNCIA DOS DADOS

O Sistema de Banco de dados deve prover uma visão abstrata dos dados para os utilizadores. Essa Abstração se dá em três níveis, o primeiro nível é o externo, o segundo nível é o conceitual e o terceiro nível é o interno.

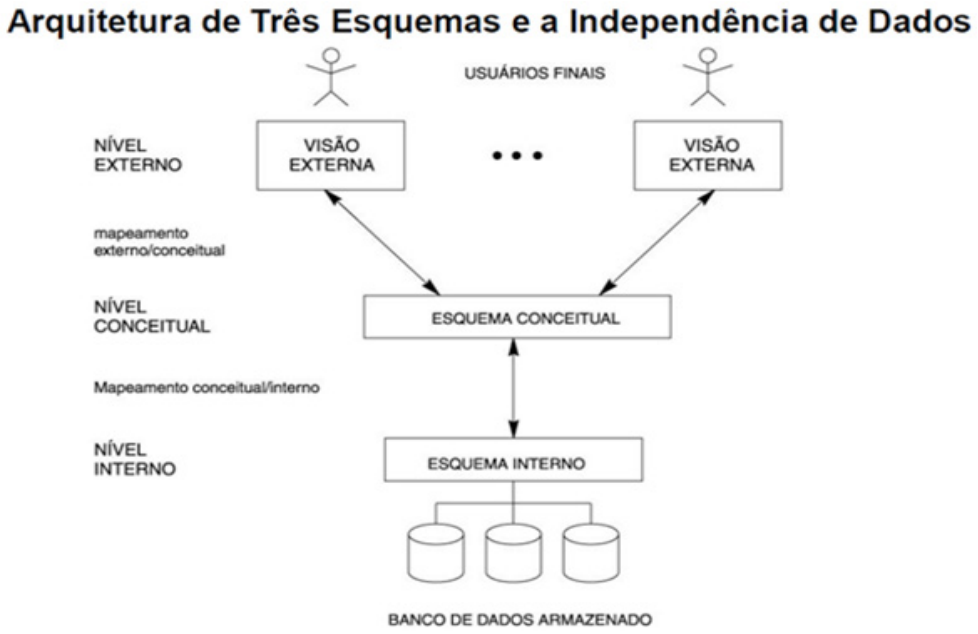


Figura 18:Arquitectura de Três Camadas

Nível Interno (Físico):

Nível mais baixo de abstracção. Descreve como os dados estão realmente armazenados, englobando estruturas complexas de baixo nível e descreve os detalhes completos do armazenamento de dados e caminho de acesso ao banco de dados.

Nível Conceitual:

Descreve quais dados estão armazenados e seus relacionamentos. Neste nível, o Banco de dados é descrito através de estruturas relativamente simples, que podem envolver estruturas complexas no nível físico. Concentra-se na descrição de entidades, tipos de dados, conexões, operações de utilizadores e restrições.

Nível Externo (visões do utilizador):

Descreve partes do banco de dados, de acordo com as necessidades de cada utilizador, individualmente ocultando o restante do banco de dados.

Instancias

As instâncias são formadas quando um dados são guardados no banco por um determinado tempo onde que se forma essas instâncias de banco de dados, sendo alterada toda vez que uma alteração na base de dados é realizada. O SGBD garante que todas instâncias satisfaçam ao esquema do banco de dados, respeitando sua estrutura e suas restrições.

Módulos Componentes Do SGBD

·Módulo de programa que fornece a interface entre os dados de baixo nível de armazenados num banco de dados e os programas aplicativos ou as solicitações submetidas ao sistema.

· Software que manipula todos os acessos ao banco de dados, proporcionando a interface de utilizador ao sistema de banco de dados.

Ilustrando o papel do sistema de gerência de banco de dados, de forma conceitual:

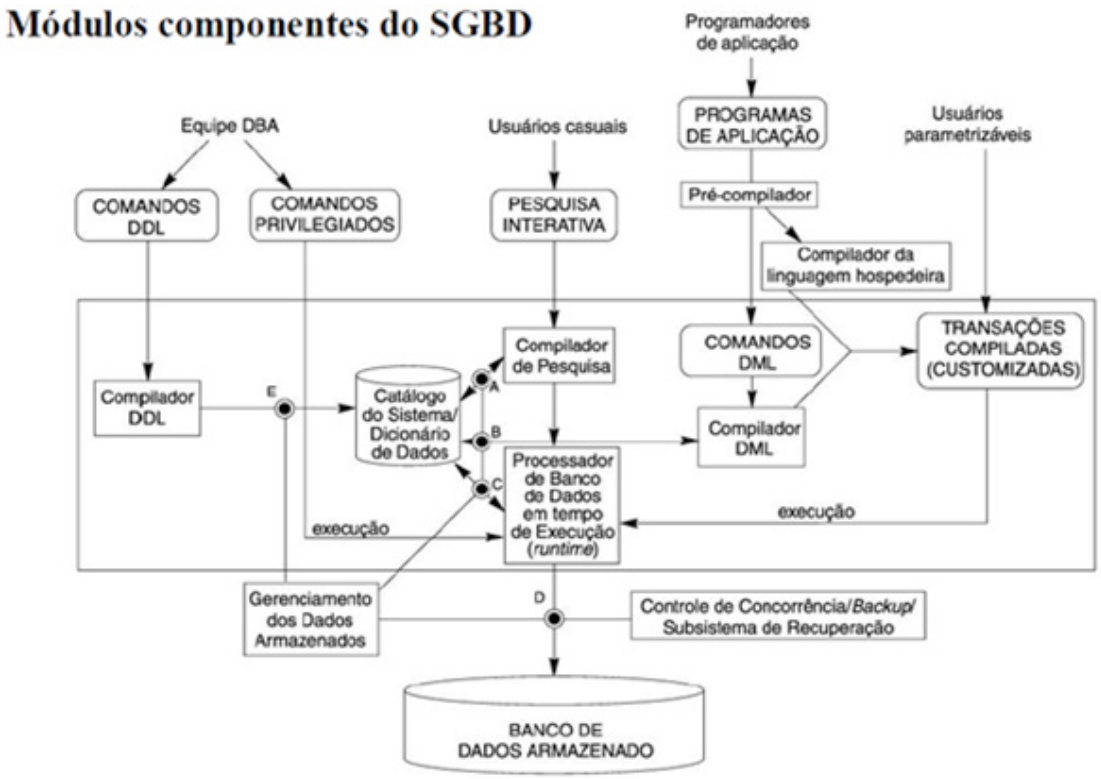


Figura 19: Módulos Componentes SGDB

- Utilizador emite uma solicitação de acesso.
- SGBD intercepta a solicitação e a analisa.
- SGBD inspeciona os esquemas externos (ou subesquemas) relacionados aquele utilizador, os mapeamentos entre os três níveis , e a definição da estrutura de armazenamento.
- SGBD realiza as operações solicitadas no banco de dados armazenado.

Classificação Dos SGBDs

Utilizadores: monoutilizadores, são usados em estações de trabalho, mini computadores e máquinas de grande porte.

Localização: possuem 2 estados localizado e distribuído. Quando é localizado todos os dados encontram-se em um único disco, se for distribuído os dados estarão em várias máquinas.

Ambiente: possui 2 tipos, os homogêneo que é o ambiente formado por um único SGBD e o heterogêneo que é o ambiente composto por diferentes SGBDs. Um exemplo é ter um sistema rodando 2 tipos de banco de dados.

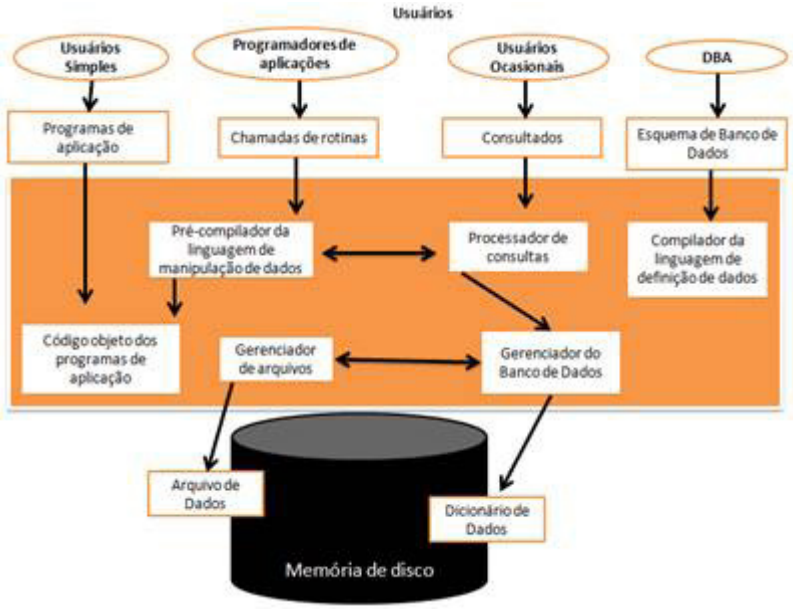


Figura 20:Ambiente mono utilizador

Abaixo, segue imagens com arquitecturas utilizadas em SGBD´s:

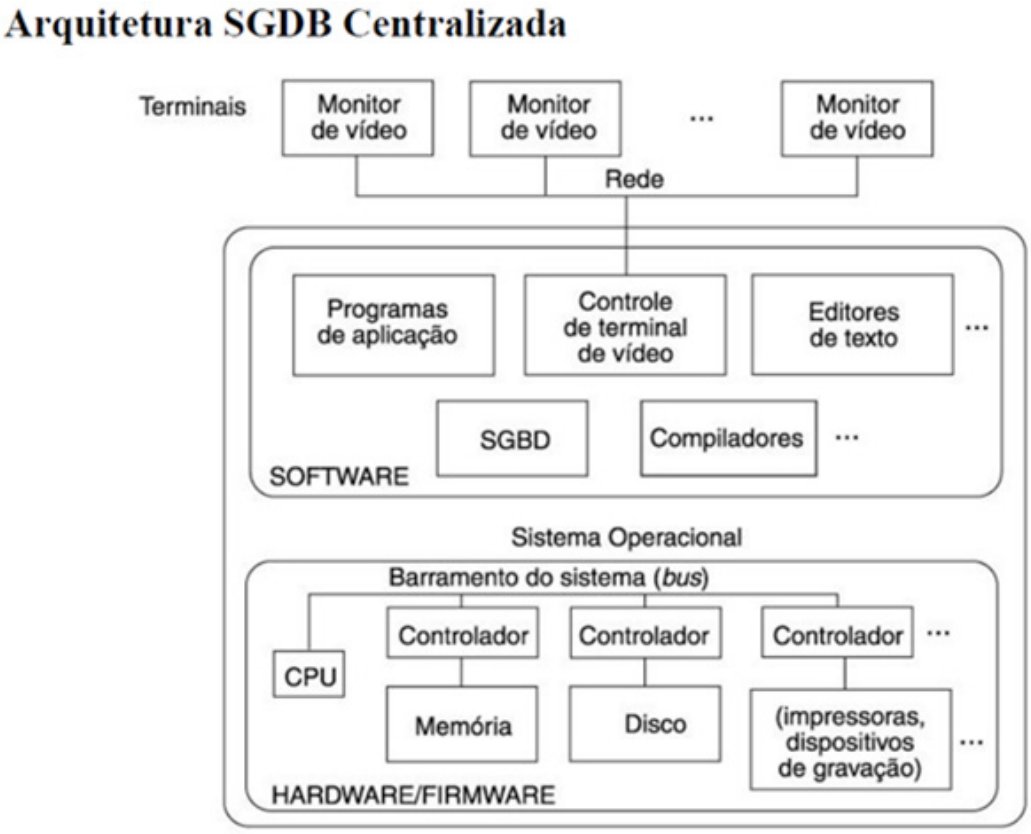


Figura 21:arquitectura SGBD Centralizada

Recuperação de informação na Base de dados

Uma das características da SGDB de extrema importância, destaca se a capacidade de realizar muitas consultas por segundo, grande capacidade de armazenamento de dados e a capacidade de indexação automática do conteúdos por meio de motor de armazenamento.

Analisando todo o processo de recuperação de informação, a última característica identificada nos permite falar de recuperação de informação e recuperação de dados, já que efectua por si só todos os processos prévios pertencentes a preparação dos dados, de modo a facilitar a recuperação da informação.

- Eliminação de palavras menos relevantes
- Indexação dos conteúdos
- Criação de arquivo inverso
- Análises de frequencias de cada termo
- Recuperação booleana
- Recuperação de texto completo
- Recuperação com linguagem natural
- Recuperação utilizando expressão da consulta

O processamento de consultas em um sistema de banco de dados é realizado quando o utilizador submete uma consulta a um SGBD, interativamente ou através de uma aplicação,utilizando uma linguagem de consulta de alto nível (SILBERSCHATZ, 2006). Este componente não trata somente da extracção de informações de um banco de dados, mas também actividades com o objectivo de optimizar o processamento das consultas submetidas ao SGBD. Considerando a importância do processamento de consulta na recuperação das informações, vai ser os passos para descrever a estrutura de processamento e os passos de optimização de consultas realizados por alguns SGBDs.

Actualmente, o Banco de Dados é a tecnologia mais utilizada para armazenar dados e gerir informações para posterior recuperação ou actualização dessas informações por um utilizador, através de uma aplicação ou acessando um sistema de banco de dados.

Assim como os dados não constituem informação, a informação, isoladamente, não é significativa. Se os dados exigem processamento (classificação, armazenamento e relacionamento) para que possam realmente informar, a informação também exige processamento para que possa adquirir significado.

Linguagens De SGBD

- SGBD deve oferecer linguagens e interfaces apropriadas e utilizadores de niveis diferentes.
- Uso da linguagem DDL (Data Definition Language - Linguagem de Definição de Dados) é definido pelo nível conceitual e interno. Esta linguagem é utilizada para permitir especificar o esquema do banco de dados através de um conjunto de definições de dados. Quando há uma separação do nível interno e conceitual que não é absorvido uma visão clara do utilizador, o SGBD tem a acção de compilar o DDL, tendo como acção, a permissão de executar as declarações identificadas pelas suas descrições dos esquemas/níveis onde irá armazená-las no catálogo do SGBD.

a. DDL (comandos que criam, alteram ou removem objetos) - CREATE, ALTER, DROP.

b. DCL (comandos que ajudam na segurança do banco de dados) - GRANT, REVOKE.

c. DML (comandos responsáveis pela manipulação dos dados) - SELECT, DELETE, UPDATE, INSERT.

A linguagem SQL permite a criação de bancos e estruturas de tabelas para executar tarefas básicas de gerenciamento de dados, além de possibilitar consultas complexas, transformando dados brutos em informações úteis, abstraindo assim a forma como os dados são armazenados no BD (ROB; CORONEL, 2011, p. 240 apud SANDRO OSCAR BUGMANN).

Uma vez que se está a tratar da questão recuperação de informação versus DBMS,deve se basear na definição da recuperação de informação,que se baseia no armazenamento e recuperação então é de extrema importância os capítulos de Data definition Lunguage e Data Manipulation Lunguage e DCL.

Pela seguinte razão:

Só é possível recuperar algo que tinha sido guardado,então para que se possa guardar dados estruturados, utilizando DBMS, obrigatoriamente tem se recorrer a linguagem DDL data Definition Lunguagem,caso não se queira utilizar o ambiente gráfico que DBMS oferece por vezes.

Linguagem de Manipulação de dados (DML)

Está Linguagem permite ao utilizador acessar ou manipular os dados, vendo-os da forma como são definidos no nível de abstração mais alto do modelo de dados utilizado.

Uma Consulta (“query”) é um comando que requisita uma recuperação de informação. A parte de uma DML que envolve recuperação de informação é chamada linguagem de consulta.

Manipulação de dados:

- Recuperação da informação armazenada,
- Inserção de novas informações,
- Exclusão de informações,
- Modificação de dados armazenados
- Linguagem de Manipulação de Dados (DML)

Permite ao utilizador acessar ou manipular os dados, vendo-os da forma como são definidos no nível de abstracção mais alto do modelo de dados utilizado.

- Uma consulta (" query") é um comando que requisita uma recuperação de informação.
- A parte de uma DML que envolve recuperação de informação é chamada linguagem de consulta.

Isso só nos permite armazenar e recuperar dados estruturados, ou melhor padronizados. Um problema relacionado com essa forma de recuperação tem sobretudo a ver com a sua forma de apresentação (Formatação) o que não é possível.

A SQL permite escrever consultas com operações expressadas em alto nível, assim o programador não precisa se preocupar com a forma como os dados estão armazenados e organizados e nem como a estrutura de armazenamento deve ser utilizada para executar a busca dos dados eficientemente (GARCIA-MOLINA; ULLMANN; WIDOM, 2002, p. 351 apud). Garcia-Molina, Ullmann e Widom (2002, p. 284) afirmam que seis cláusulas podem aparecer na consulta select-from-where, sendo elas: SELECT, FROM, WHERE, GROUP BY, HAVING e ORDER BY. Somente as duas primeiras cláusulas são obrigatórias e não é possível utilizar HAVING sem utilizar a cláusula GROUP BY. Qualquer outra cláusula adicional pode ser utilizada, porém deve seguir a ordem acima apresentada. A Quadro x seguinte mostra a sintaxe do comando SELECT.

Para recuperar informações na base de dados tem de se recorrer a sintaxe SELECT:

SELECT<NOMES de campos/atributosl*>

FROM<TABELA OU TABELAS>

WHERE<condição>

GROUP BY<lista de atributos>

HAVING<condição>

ORDER BY<lista de atributos>

a. Todas as linhas pertencentes a tabela chamada pela FROM são selecionadas; qualquer que seja a consulta a ser realizada tem de ter a estrutura da sintaxe que está dentro do rectângulo em cima.em caso de algumas restrições especiais pode se alargar sintaxe em função da informação a ser recuperada.

b. As linhas que não satisfazem a cláusula WHERE são eliminadas;

c. As linhas restantes são agrupadas de acordo com a cláusula GROUP BY;

d. Os grupos que não satisfazem a cláusula HAVING são rejeitados;

e. Os cálculos referentes a lista de atributos são executados;

f. Finalizando com o comando ORDER BY responsável por ordenar as linhas de acordocom os atributos listados.

Além das consultas simples, com restrições e agrupamentos, também é possível utilizar subconsultas dentro das cláusulas WHERE, FROM e HAVING. Segundo Elmasri e Navathe (2005, p. 165 apud BUGMANN,2012), "Algumas consultas necessitam da busca de valores presentes no BD para, então, usá-los na condição de comparação. Essas consultas podem ser formuladas de modo conveniente por meio de consultas aninhadas (nested queries)".

No BD a estrutura de dados é fixa, definida em um modelo de dados,como o relacional, assim é possível realizar consulta de dados através da linguagem SQL,obtendo uma resposta exata do estado atual dos dados. Em um sistema de RI, o modelo dedados é heterogêneo, os dados são os documentos indexados, onde as consultas tendem a serum conjunto de palavras- chave, ou uma frase na linguagem natural. O resultado da pesquisa RI é uma lista de identificadores de documentos ou algumas partes do texto (ELMASRI;NAVATHE, 2005, p. 672).

O resultado em uma consulta de banco de dados é uma resposta exata; se não foremencontrados registros na relação, o resultado é vazio (nulo). A resposta para umasolicitação do utilizador em uma consulta RI representa a melhor tentativa do sistemaRI de recuperar a informação mais relevante para sua consulta. (ELMASRI;NAVATHE, 2005, p. 672).

Os BDs possuem uma grande quantidade de informações, que possibilita a optimização das consultas, enquanto que nas operações com RI a consulta é feita com o valor do dado e afrequência com que este ocorre no documento. Muitas vezes uma análise estatística érealizada para determinar a relevância do documento com a solicitação do utilizado r(ELMASRI; NAVATHE, 2005, p. 672).A forma mais simples de procurar ocorrências em determinado texto é varrer osdocumentos disponíveis para a pesquisa sequencialmente. No entanto, este tipo de procura só é apropriado para uma colecção de documentos de pesquisa pequena. A grande maioria dos sistemas RI processa a colecção de documentos de entrada para gerar índices e operar sobre oíndice invertido (ELMASRI; NAVATHE, 2005, p. 682).

Avaliação da Unidade

Verifique a sua compreensão!

1. Cite algumas vantagens de se utilizar banco de dados.
2. Descreva situações de seu dia-a-dia em que você utiliza banco de dados.
3. Para que servem os diferentes níveis de abstração em bancos de dados?
4. Qual a vantagem de um SGBD utilizar uma linguagem de consulta declarativa (como o SQL) em vez de apenas uma biblioteca de funções em C ou C++ para realizar manipulação de dados?
5. Explica como como SGDB ajuda os profissionais da empresa na recuperação de informação?
6. Explica as diferenças entre SBD e SGBD?
7. Quais são os principais componentes SGDB?
8. O que SGBD em particular SGBD relacional?
9. Como os dados em um SGDB é manipulado e recuperado?
10. Qual é a diferença existente entre as arquiteturas de banco de dados Centralizados, Cliente/Servidor, Distribuído e Paralelo?
11. Os quatro se distinguem, principalmente, no local onde realmente sucede o processamento dos dados.
12. Em quantos níveis se procede a abstracção de dados?A abstração se dá em três níveis
13. Descreve cada uma das níveis?
14. Classifica os Sistemas de gestão da base de dados?
15. Responda Com V ou F as afirmações que se segue e justiça a sua escolha.

O modelo relacional de dados é o mais aconselhado para recuperar informações

A integridade e a inconsistência é garantida pela DB?

Num registo num base de dados pode ter sempre um campo que identifique os restantes campos .

Leituras e Outros Recursos

As leituras e outros recursos desta unidade encontram-se na lista de "Leituras e Outros Recursos do curso".

Unidade III - Recuperação de Informação na Web /Pesquisa Web

1. Sistemas de pesquisa Web;
2. Metadados e a recuperação de informações na web;
3. Motores de Busca;
4. Diretórios;
5. Metamotores de Busca;
6. Sistemas híbridos;
7. Arquitectura de um Motor de Busca.

Termos-Chaves

BUSCA: Processo que consiste na localização das informações necessárias a

Cada utilizador.

Browser: Programa que permite aceder e manipular toda a informação disponível na Internet (texto, som, vídeo, imagem, etc). Os browsers mais comuns são o Internet Explorer (da Microsoft) e o Netscape.

Hiperligação: (hyperlink): Ligação entre um elemento de um documento electrónico (palavra, imagem, etc) com outro elemento do mesmo ou de outro documento electrónico. Usam-se também os termos hipertexto (hypertext) ou hipermédia (hypermedia) quando os elementos em causa são texto, no primeiro caso, ou imagem, som, etc, no segundo caso.

Hipertexto: Documento constituído por texto, imagens e ligações que, quando seleccionados, possibilitam a visualização de outras partes do documento ou até de documentos diferentes, podendo o utilizador escolher o caminho a percorrer no hipertexto.

HTML (HiperText Markup Language): A extensão mais comum dos ficheiros legíveis na Internet, usando um vulgar browser.

HTTP (Hyper Text Transfer Protocol): Protocolo de comunicação entre computadores ligados pela Internet. Em particular, este protocolo permite a transferência universal de material digital em hipertexto.

Internet: Rede mundial de computadores ligados entre si, que partilham informação.

Link: Ligação. Num documento HTML ou num hipertexto identifica-se como uma palavra sublinhada ou em destaque que indica um ponto de ligação a outra informação.

URL (Uniform Resource Locator): Endereço de um siteda Internet. Por exemplo, o URL da Universidade do Porto é www.up.pt.

Servidor web / web sever: Servidor que, posto à disposição dos computadores descentralizados (clientes), processa os serviços e envia o docum ento solicitado.

Metadados: Elementos de descrição/definição/avaliação de recursos informacionais armazenados em sistemas computadorizados, organizado por padrões específicos, de forma estruturada

Multimídia: Tecnologia que permite ao computador trabalhar simultaneamente e de forma interativa com diversos tipos de registo informacional, como texto, som, imagens estáticas, animação e vídeo.

FORMULÁRIO:Documento estruturado que permite ao utilizador inserir informação específica com um objetivo determinado.

INTERFACE GRÁFICA: Forma de interação entre o utilizador e o computador, baseada no uso de imagens, ícones, janelas, botões e demais recursos gráficos.

INTERNET : Rede mundial de computadores, que se comunica por meio dos protocolos TCP/IP, permitindo o acesso à World Wide Web (WWW).

NAVEGADOR: Software que permite ao utilizador acessar, ver e navegar entre arquivos na web por meio de uma interface amigável, por exemplo, Internet Explorer e Mozilla Firefox. Sinônimo: BROWSER.

NAVEGAÇÃO: Ação de utilizar a Internet à procura de informação, por meio de um programa de navegação (browser), deslocando-se entre páginas do mesmo web site ou de sites diferentes, recorrendo a hiperligações ou hiperlinks.

PORTAL :Site Web que reúne produtos e serviços de informação de determinada área de interesse, em um único ponto de acesso. Normalmente oferecem, por exemplo, serviços gratuitos de correio eletrônico, conversa, notícias, informações sobre o tempo, cotação de ações, assim como facilidades para procurar outros sites.

SERVIDOR : Computador central, em uma rede, responsável pela administração e fornecimento de programas e informações aos demais computadores a ele conectados.

SITE: Um endereço dentro da Internet que permite acessar arquivos e documentos mantidos no computador de uma determinada instituição. Veja também a definição de PORTAL.

WWW: Teia de alcance mundial, ou seja, conjunto interligado de documentos escritos em linguagem HTML armazenados em servidores HTTP ao redor do mundo. Veja também as definições de INTERNET, HTML e HTTP.

URL: Acrónimo de Uniform Resource Locator, termo em inglês usado para designar o endereço de um recurso Internet, incluindo as páginas Web.

Métodos de recuperação de Informação na Web

Os motores de busca como paradigma da recuperação da Informação Internet.

Os motores de busca são semelhantes aos directórios nas possibilidades permitidas,mas divergem na forma como as informações são armazenadas.

Um motor de Busca periodicamente recolhe páginas da web e constrói um novo indíce que permite efectuar pesquisas rápidas sobre esta informação.

De todos os SRI que foram desenvolvidos pela web,os motores de busca são os que mais se incidem,devido a sua natureza dinâmica,uma vez que as tecnologias web estão em constante crescimento e os motores de busca se evoluem paralelo aos crescimentos das tecnologias e do número de utilizadores.

Funcionamiento de um Motor de Busca

Independentemente de seu método de rastreio e algoritmos utilisados no alinhamento dos documentos, todos os motores partem de uma situação inicial parecida: uma lista de endereço que serve de ponto de partida para o robot (o robots) do motor.

Os motores de busca utilizam software conhecido como ‘‘aranhas’’ ou “robots” que percorrem “toda” a Internet em busca da informação (documentos ou endereços de páginas web) que se pretende. Os dados são recolhidos para o index dos motores de busca, que cria uma base de dados com essa informação. A forma como a informação é indexada depende de cada motor de busca, podendo ser feita por palavras, títulos, URL’s ou por directorias. Assim, sempre que se introduz uma palavra ou um conjunto de palavras que se pretende pesquisar, as bases de dados são percorridas em busca de documentos ou sites que lhe correspondam. O resultado da busca é dado em hyperlinks, podendo clicar-se em cada uma das entradas para aceder à informação.

Arquitectura de um Motor de Busca.

A maioria dos motores utilizam uma arquitectura de tipo robot- indexador centralizado.

Os robôs não se movem pelas redes,nem serão executas nas máquinas remotas cujas páginas serão consultadas.

Orobôs funciona sobre o sistema de busca local do motor de busca e envia uma série de petições aos servidores remotos que alojam as apginas a analisar.

Um motor de pesquisa é composto por 5 componentes principais: o crawler o repositório, o indexador ,a interfacee o ordenador.

O crawler descobre e recolhe automáticamente conteúdos da web, seguindo os links contidos nas páginas. Apenas os conteúdos que o crawler seja capaz de encontrar e recolher poderão vir a constar em resultados de pesquisas no motor de pesquisa. Logo, é crucial escrever páginas amigas dos crawlers.

O repositório armazena as páginas recolhidas de modo a que possam ser indexadas e mostradas em cache.

O indexador extrai as palavras dos conteúdos web e cria um Indice Invertido. Caso não seja possível extrair correctamente as palavras de uma página, esta dificilmente será retornada como resultado de pesquisas.

O ordenador ordena as páginas que contenham os termos pesquisados por um utilizador de modo a que as mais relevantes sejam apresentadas nos primeiros lugares. As páginas que não tenham sido escritas considerando os requisitos dos motores de pesquisa são relegadas para posições mais baixas em relação apaginas optimizadas para motores de pesquisa.

O apresentador gere a interface de utilização do motor de pesquisa.

A prioridade do motor de pesquisa é satisfazer a necessidade de informação do seu utilizador. Os utilizadores de um motor de pesquisa podem sugerir sites para serem recolhidos. No entanto, um motor de pesquisa tem controlo absoluto sobre quais as páginas que irá recolher e em que posiçãoestas serão retornadas como resposta as pesquisas efectuadas por parte dos seus utilizadores.

Como exemplos de motores de pesquisa temos:

- Google,Exicite,Lycos,Infoseek,Hotbot
- AltaVista, o AllTheWeb ou o MSN Search.
- Que são metamotores de busca?

Os metamotores de busca (metasearch engines) são motores que pesquisam a informação de vários motores de buscas (Google, Bing entre outros) simultâneamente, isto é, a interrogação formulado pelo utilizador é enviado para os vários motores de buscas simultaneamente (daí o prefixo meta-, que é muito utilizado em informática) e apresentam o resultado dessa metapesquisa, de todo conjunto de motores de busca, numa lista de páginas que satisfazem os nossos critérios de interrogação, numa mesma página de respostas.

Contrariamente aos motores de busca tradicionais como o Google, osmetamotores de busca não possuem a sua própria base de dados, apenas recolhem os resultados de outros motores de busca e fundem-nos. A sua grande vantagem é a supressão de links repetidos, por se encontrarem armazenados em mais que um motor de busca, no entanto existem outros que ordenam os resultados de forma a apresentarem o os que melhor se adequam aos termos de pesquisa usados, ou organizam-nos por categorias.

Exemplos de metamotores de busca

Metacrawler –http://www.metacrawler.com

ixquick
Findloo
Dogpile
Kartoo

800go-http://www.800g0.com

Vantagens e limitações

Uma função útil de alguns metamotores é a supressão de (links) repetidas (dado estarem armazenadas em mais do que um motor de pesquisa).

Embora possam parecer muito interessantes para obter respostas ainda mais completas, a verdade é que, porque os motores de pesquisa têm sintaxes de interrogações específicas e discrepantes, os metamotores tendem a escolher o máximo denominador comum em termos de sintaxe, o que lhes faz perder muita informação e especificidade.

Mesmo assim, a sintaxe de interrogação dos metamotores é frequentemente algo mais complexa do que a dos motores simples. Além disso, podem não dispor de funções sintácticas de interrogação de que apenas dispõem alguns motores de pesquisa, pelo que não é possível interrogá-los tão especificamente.

Aumenta a probabilidade do utilizador recuperar a informação procurada

Mas também pode aumentar a probabilidades de mais ocorrências de informações sem relevância.

Directórios

Os diretórios foram a primeira solução proposta para organizar e localizar os recursos da Web, tendo precedido os motores de busca por palavras-chave. Foram introduzidos quando o conteúdo da Web ainda era pequeno o suficiente para permitir que fossecoletado de forma nãoautomática. Organizam os sites que compõem sua base de dados em categorias, as quais podem conter subcategorias, ou seja, os sites recebem uma organização hierárquica deassunto e permitem aos utilizadores localizar informações, navegando, progressivamente, para as subcategorias. Como são ferramentas genéricas, destinadas a um público variado, procuram incluir, em suas árvores hierárquicas de assunto, tópicos que são de interesse amplo. É comumque incluam, por exemplo, itens relacionados com educação, Desporte, entretenimento, viagens, compras ou informática. Cabeçalhos de assunto são atribuídos de forma consistente, de modo que os utilizadores podem contar com a ajuda de um vocabulário controlado.

Os sites coletados passam pela seleção, na maioria das vezes, por seres humanos, os editores, que tomam conhecimento de novos recursos por meio de sugestões de utilizadores, depesquisas na Internet (em listas de anúncios de novas páginas e atualizações, por exemplo), ou ainda, pelo uso de robôs para coletar novos URLs.

Os grandes diretórios podem conter dezenas de milhares de categorias e subcategorias e mais de um milhão de sites.

O primeiro diretório da Web foi o The World Wide Web Virtual Library (http://www.vlib.org/), lançado em novembro de 1992 e sediado no CERN, que também foi o local de nascimento da Web. Atualmente, o exemplo mais conhecido é o Yahoo!, que iniciou em 1994, a partir de um

hobby de estudantes de doutorado na Stanford University, e hoje é uma bem-sucedida empresa comercial. Outros exemplos de diretórios são Snap (http://www.snap.com),

LookSmart (http://www.looksmart.com), Open Directory (http://dmoz.org/), Yahoo Brazil (http://www.br.yahoo.com), Cadê (http://www.cade.com.br), Surf (http://www.surf.com.br) e Vai & Vem (http:// www.vaievem.com.br), sendo estes três últimos brasileiros

Embora os directórios forneçam serviços de pesquisa, estes procuram apenas entre as descrições dos sites, ao contrário dos motores de pesquisa que pesquisam sobre todos os textos de todas as páginas de um site. Pelo que, é muito importante fornecer boas descrições dos sites, focando os aspectos em que o site se distingue dos restantes namesma categoria, e mantendo-as actualizadas de modo a reflectirem os conteúdos actuais do site.

Os directórios permitem encontrar rapidamente listas exaustivas de links para sites sobre um determinado tema. Além disso, a qualidade dos sites registados é verificada pelos responsáveis pelo serviço de directório, pelo que os links contidos emdirectórios atestam a qualidade de um site. Assim sendo, os motores de pesquisa frequentemente iniciam as recolhas da web partindo dos links contidos nos directórios.

Ao longo do tempo muitas páginas referenciadas pelos directórios desaparecem e é necessário efectuar operações de manutenção para eliminar estes links inválidos, o que muitas vezes não é feito com a frequência necessária para garantir a qualidade dos resultados das pesquisas. Por sua vez, os motores de pesquisa eliminam URLs inválidos em cada nova recolha da web porque as páginas referenciadas não podem ser recolhidas e como tal não serão incluidas no novo índice. Algumas páginas desaparecem entre actualizações e são retornadas como resultados de pesquisas. Este problema é atenuado pela funcionalidade de cache, que permite visualizar as páginas arquivadas pelo motor de pesquisa.

Exemplos de directórios:Yahoo,sapo,IOL,DMOZ

	Descobrir Recursos	Representação Do Conteúdo	Representação da consulta	Apresentação dos resultados
Directórios	São realizadas pelos utilizadores	Classificação manual	Implícita navegação por categorias	Páginas criadas antes da consulta pouco exaustivo e muito preciso
Motores de Busca	Principalmente de forma automática por mei de robots	Indexação automática	Explicita (palavras chaves e operadores)	Páginas criadas dinamicamente em cada consulta muito exaustivo e muito preciso

Tabela traduzido Características de directorios e motores de Busca.Fonte: Delgado Domínguez apud FranciscoJavierMartínezMéndez, A. Mecanismos de recuperación deInformaciónen la WWW[En línea].Palma de Mallorca, Universitat de lesIlles Balears,1998. <http://dmi.uib.es/people/adelaida/tice/modul6/memfin.pdf> [Consulta:18 de septiembre de 2001]

Metadados e a recuperação de informações na web;

Nos dias de hoje em que a Internet (WWW) funciona como um obra de consulta, o problema da recuperação de informação que antes era especificamente para os profissionais das bibliotecas, dos arquivos e do centros de documentação, passa a ser um problema comum.

Para resolver esses problemas os profissionais criaram catálogos e formas de registos que permitem identificar os documentos com uma certa facilidade.

Profissionais de informação vêem criando, há séculos, metodologias para registo,

Inventario e descrição de documentos, como forma de controlar acervos e prover meios de acessar seletivamente os itens de uma coleção. Um instrumento de pesquisa de um arquivo ou um catálogo de uma biblioteca nada mais são que descrições de documentos de uma coleção, organizadas com a finalidade de facilitar sua recuperação e acesso, os agora chamados metadados Milsted (1999apud Marcondes Carlos Henrique).

O termo "metadados" surge no sentido de ajudar a identificar e recuperar o grande volume de informação que se encontram disponiveis na web, sem saber como chegar aos mesmos.

Um dos maiores objectivos do uso de metadados no contexto da Web é permitir não só descrever documentos eletrônicos e informações em geral, possibilitando sua avaliação de relevância por utilizadores, mas também permitir agenciar computadores e programas especiais, robôs e agentes de "software", para que eles compreendam os metadados associados a documentos e possam então recupera-los, avaliar sua relevância e manipulá-los com mais eficiência.

A identificação e recuperação de recursos informacionais torna -se assim uma das questões mais importantes da atual economia da Web. Como já dizíamos Marcondes (Marcondes,2001), "... a informação relevante para um dado problema tem que estar disponível no tempo certo. De nada adianta a informação existir se quem dela necessita não sabe da sua existência ou se ela não puder ser encontrada."

Metadados são definidos como "dados sobre dados" (Weibel, 1995apud Marcondes). São dados,associados a um recursos Web, um documento eletrônico, por exemplo, que permitem recupera-lo, descreve-lo e avaliar sua relevância, manipula-lo (o tamanho de um documentos, ao se fazer "downloading" ou o seu formato, para sabermos se dispomos do programa adequado para manipula-lo), gerencia-lo, utiliza-lo enfim.

A primeira tentativa de dar conta da explosão informacional em que se transformou a Web foram os catálogos, como o Yahoo (o primeiro catálogo da Web), e os chamados mecanismos de busca, como AltaVista, Lycos, WebCrawler, etc, e mais recentemente, o Google. Enquanto em catálogos como o Yahoo, a descoberta, avaliação e descrição e inclusão dos recursos Web na base de dados é feita por profissionais de informação, os mecanismos de busca, para indexarem a Web, possuem programas que visitam página por página da Web, percorrem o texto de cada página, extraindo daí palavras-chaves e armazenam numa base de dados estas palavras-chaves, associadas ao URL da página. É sobre esta base de dados que os utilizadores fazem suas buscas nos "sites" dos mecanismos de busca. Naturalmente, por ser uma indexação automática com base em palavras isoladas, sem nenhum controle terminológico, efetuada em páginas sobre os mais variados assuntos, diferentes idiomas e totalmente desprovida de qualquer informação contextual, os resultados tem baixíssima precisão (Sneiderman,1997). Estudos como o citado reforçam a dimensão do problema localização/ identificação colocado pela Internet.

Com o objetivo de ajudar a obter maior precisão nas buscas por páginas Web ajudando os robôs dos mecanismos de busca a fazerem uma indexação de maior qualidade, num primeiro momento foram incorporados metadados no texto mesmo destas páginas. Isto foi feito com o uso de "tags" especiais da linguagem HTML (a linguagem em que são escritas as páginas da Web), as "tags" META, como mostrado a seguir:

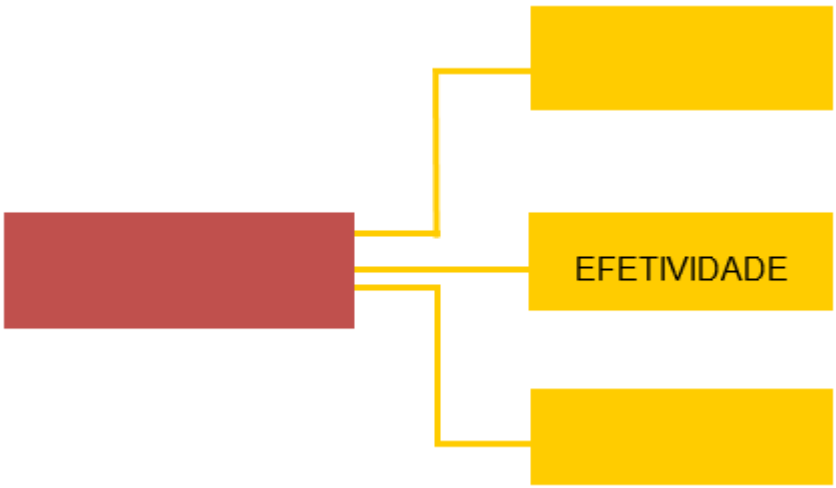
Interface

Interfaces gráficas podem auxiliar o utilizador final na Recuperação de Informação (RI),servindo de porta de entrada do utilizador para o sistema, para interagir com a base de dados,realizando interrogação com o objectivo de recuperar informações relevantes que atendam a necessidade informacional dos mesmos.

O estudo da interface deve ser abordado em duas perspectivas:

Interface que o sistema despõe para o utilizador expressar as suas necessidades de informação (Chang la denomina "interface de consulta (" [CHA 2001,apudMartínezMéndez,Francisco Javier)

E a interface de resposta que depõe um sistema de mostrar ao utilizador o resultado da sua operação de busca (BAE 1999, APUDMartínezMéndez, FranciscoJavier)



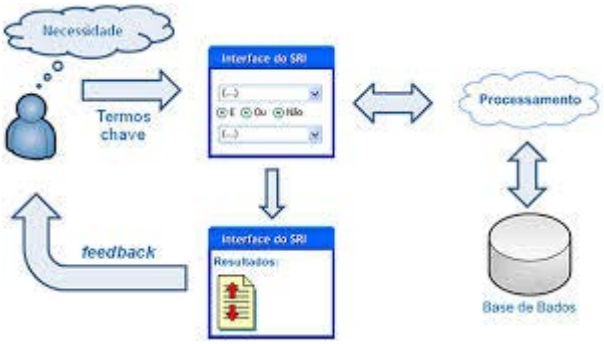


Figura 22:Interface do Utilizador

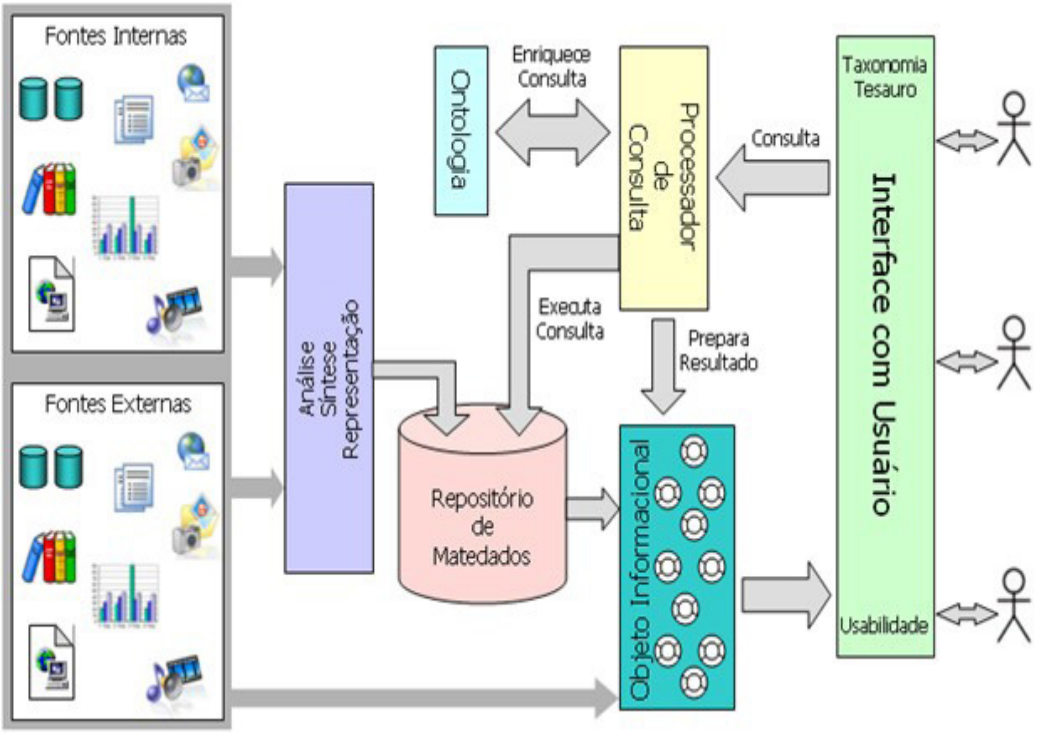


Figura 23:Interface do Utilizar Com as fontes informacionais

Nem todos os sistemas possuem a mesma capacidade de recuperação de informação.

Os motores de busca diferem também em relação às interfaces e recursos de busca que oferecem. Geralmente fornecem dois modos de busca, a busca simples ou a clássica para utilizadores leigos e a busca avançada para utilizadores mais experientes ou profissionais.

Na busca simples ou clássica, existe uma caixa de textos clássico, que permitemos utilizadores que digitem o termo ou a palavra-chave relacionada com a informação a ser procurada. Por vezes pode ser introduzido algumas restrições de busca como por exemplo o idioma da página a ser recuperado, o tipo de objecto que se deseja empregar na busca por frases literais ou busca exacta. Neste tipo de busca não é necessário conhecimento de lógica booleana.

A busca avançada fornece recursos mais poderosos, como expressões booleanas complexas. Muitas vezes, na busca simples, os conectivos booleanos são automaticamente colocados entre os termos de busca, e nem sempre os utilizadores sabem qual operador está sendo utilizado. Em alguns motores, por exemplo, um espaço entre os termos da consulta é interpretado como um conectivo booleano OR (Altavista e Excite, por exemplo), enquanto para outros tem o significado de AND (Google e Northernlight, por exemplo).

Podem oferecer recursos como truncamento, busca por frase, busca por proximidade de palavras, busca por campos e sensibilidade à caixa de caracteres (isto é, caixa-alta e caixa-baixa). É comum também haver opções para permitir a limitação por data, domínio, idioma ou tipo de arquivos (com base na extensão dos nomes dos arquivos).

Alguns motores fornecem opções mais sofisticadas, como a busca automática pela raiz das palavras, ou seja, se o utilizador entrar com a palavra "informática", ele encontrará também documentos com a palavra "informático". Em alguns casos, a pesquisa se estende também a outros termos sinônimos ou a termos com conteúdo semântico equivalente ao termo da consulta, como é o caso do Excite.

Esta busca estendida, quando existente, é geralmente automática, não sendo dada ao utilizador a possibilidade de desabilitá-la. São mais raros motores que permitem buscas em linguagem natural, na qual a consulta pode ser entrada na forma de uma sentença, em vez de termos isolados.

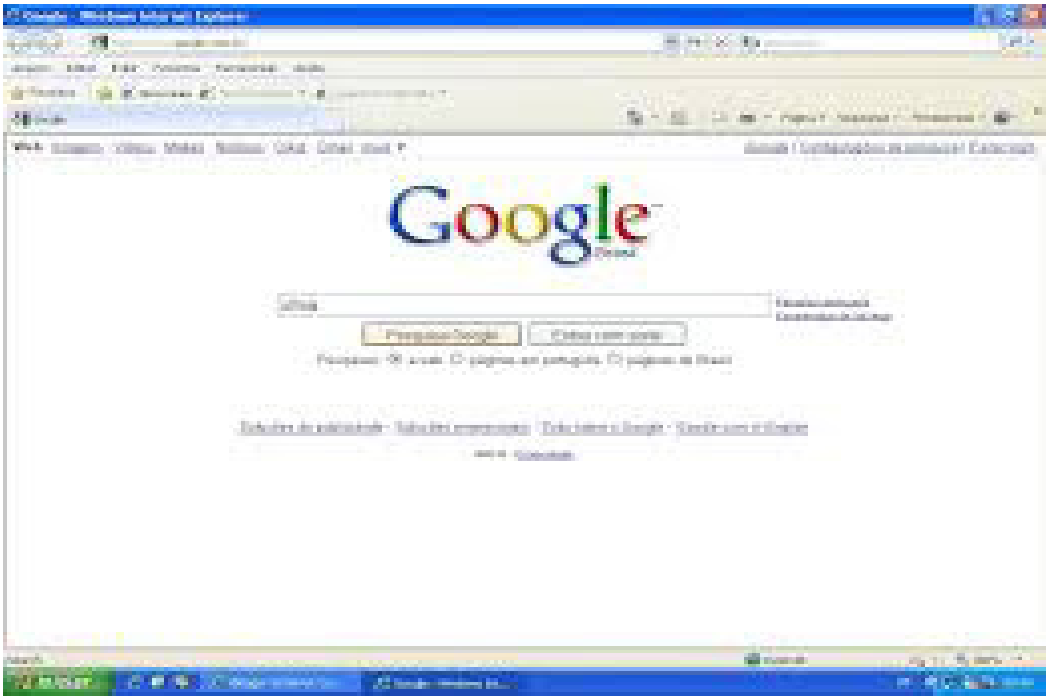


Figura 24:janela de Busca Simples

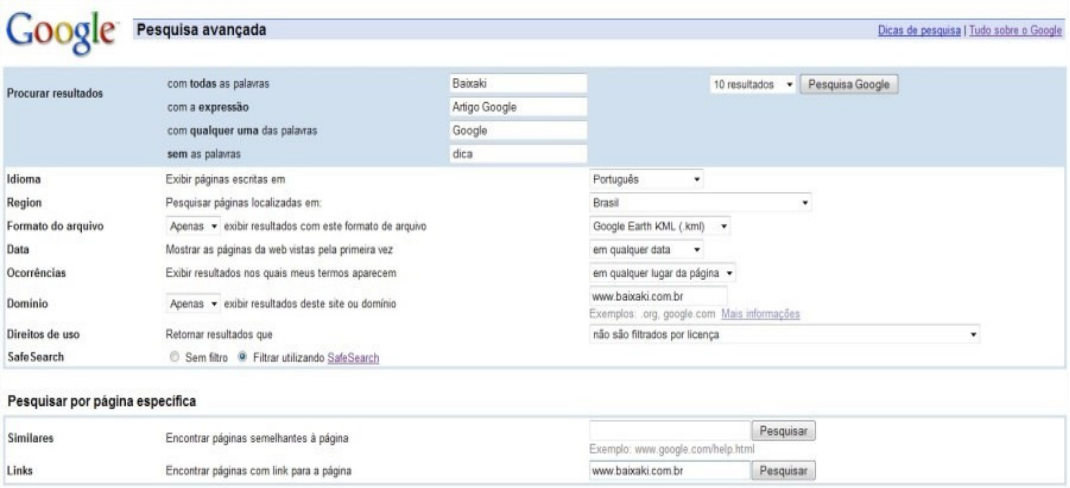


Figura 25:Interface de Busca Avançada

Sistema de Recuperação de Informação

Diante da grande quantidade de informações produzidas actualmente, os SRItem como objetivo armazenar e recuperar essas informações para uso posterior,localizando a informação relevante para satisfazer a necessidade de informação doutilizador. Como afirma Araujo Junior (2007) os SRI dizem respeito a um sistema de operações interligadas para identificar, dentre um grande conjunto de informações,aquelas que são de fato úteis, ou seja, que estão de acordo com a necessidade expressa pelo utilizador.

Os SRI tratam, basicamente, de indexar, armazenar e recuperar documentos potencialmente úteis com o objetivo de satisfazer as necessidades informacionais expressas pelo utilizador através da consulta.

A pesquisa ou consulta, consiste no momento em que o utilizador expressa sua necessidade ao sistema, a partir daí é realizada uma pesquisa por documentos que atendam a consulta do utilizador. Para que isso ocorra é necessário que haja uma interface, por meio da qual o utilizador irá traduzir sua necessidade de informação (SOUZA, 2006).

Os SRI fornecem dois modos através dos quais o utilizador pode realizar apesquisa: através da recuperação ou da navegação. Para cada um desses modos existem modelos específicos. Na recuperação o utilizador expressa sua necessidade ao sistema em forma de questões ou palavras-chave e na navegação o utilizador nãopropõe uma questão ao sistema, mas navega por categorias e entre os documentosem pesquisa de informações pertinentes.

Técnicas de Visualização da Informação

A Visualização da Informação pode ser definida como uma área da Ciência que tem por objetivo o estudo das principais formas de representações gráficas para apresentação de informações, a fim de contribuir para o melhor entendimento delas, bem como ajudar a percepção do consumidor a fim de deduzir novos conhecimentos baseados no que está sendo apresentado (FREITAS, 2001 apud DIAS; CARVALHO, 2007).

O processo de visualização de informação está relacionado com a transformação de dados abstratos em imagens que possam ser visualizadas. O objetivo das estruturas de visualização da informação é auxiliar no entendimento de determinado assunto, o qual, sem uma visualização, exigiria maior esforço para ser compreendido, o objetivo final é a inclusão informacional de seus utilizadores.

Na maioria dos casos, o oferecimento de imagens, figuras, estruturas gráficas equaisquer outros recursos gráficos, com a finalidade de apresentar uma informação, produz a compreensão da mensagem transmitida, pois esta se torna mais natural e exige menos esforço cognitivo, como afirma Nascimento e Ferreira (2005) umasimples visualização pode condensar uma grande quantidade de informações,facilitando a compreensão das mesmas, já que a visão é o sentido humano que possui maior capacidade de captação de informações por unidade de tempo.

No contexto da Recuperação de informação a visualização auxilia numa pesquisa efetiva, orientando o utilizador para a informação desejada, auxiliando no entendimento do contexto dos resultados retornados e facilitando a seleção de informações mais relevantes segundo sua necessidade.

A seguir são descritas algumas das diversas técnicas de visualização existentes, sendo abordadas técnicas atualmente mais utilizadas em interfaces de SRI. Serão abordados, também, alguns sites que utilizam interfaces visuais,demonstrando de forma prática que o uso de recursos de Visualização da Informação permite aos sistemas de recuperação da informação apresentarem, de forma amigável e prática, como estão organizadas e estruturadas suas informações, facilitando a localização e recuperação das mesmas.

Avaliação da Unidade

Verifique a sua compreensão!

Motores de Busca

1. Os motores de busca são semelhantes aos directórios. VER
2. Os motores de busca são diferentes dos directórios nas possibilidades permitidas, mas convergem na forma como as informações são armazenadas. FALS
3. Os motores de busca são semelhantes aos directórios mas diferem na forma como as informações as informações são armazenadas.VER
4. A construção de índice no motor de busca é realizada pelo homem. FALS
5. A capacidade de recuperação de informação dos motores de busca é superiora dos metamotores. FAL
6. O principio de funcionamento dos motores de busca é idêntico a dos metamotores VER
7. Os motores de busca permitem encontrar rapidamente lista exaustiva de link para site sobre determinado assuntos. FALS
8. Nos motores de busca é preciso efectuar manutenção de links de vez em quando,por motivo de desaparecimento de siste. FALS
9. Os motores de busca fazem recolhas de links contidos no directórios. VER
10. Os directorios eliminam URLs para cada nova recolha web,porque as paginas deixam de existir. FALS
11. Nos motores de busca as paginas que deixaram de existir não aparecem nos resultados da pesquisa. FAL
12. Nos directorios as paginas que deixaram de existir não aparecem no resultado da pesquisa. VER

Os directórios permitem encontrar rapidamente listas exaustivas de links para sites sobre um determinado tema. Além disso, a qualidade dos sites registados é verificada pelos responsáveis pelo serviço de directório, pelo que os links contidos em directórios atestam a qualidade de um site. Assim sendo, os motores de pesquisa frequentemente iniciam as recolhas da web partindo dos links contidos nos directórios.

Ao longo do tempo muitas páginas referenciadas pelos directórios desaparecem e é necessário efectuar operações de manutenção para eliminar estes links inválidos, o que muitas vezes não é feito com a frequência necessária para garantir a qualidade dos resultados das pesquisas. Por sua vez, os motores de pesquisa eliminam URLs inválidos em cada nova recolha da web porque as páginas referenciadas não podem ser recolhidas e como tal não serão incluidas no novo índice. Algumas páginas desaparecem entre actualizações e são retornadas como resultados de pesquisas. Este problema é atenuado pela funcionalidade de cache, que permite visualizar as páginas arquivadas pelo motor de pesquisa.

13. Refira os principais tipos de ferramentas de pesquisa de informação existente na Internet.

14. Qual é o grande objectivo de um directório? quais são as estrategias de buscas utilizadas para encontra um site num directório.

15. Diferença entre motores de busca e directórios?

16. Classifique os seguintes URL como Directórios,motores de Busca,Meta motores,site

URL	
Alta vista	
Yahoo	
Google	
Cadê	
Crawler	
Exicite	
Sapo	
Geocities	
unicv	

17. Os SRI fornecem dois modos através dos quais o utilizador pode realizar a pesquisa. Quais são?

E qual é a diferença existente entre elas?

Resposta:

Através da recuperação ou da navegação.

Para cada um desses modos existem modelos específicos. Na recuperação o utilizador expressa sua necessidade ao sistema em forma de questões ou palavras-chave e na navegação o utilizador não propõe uma questão ao sistema, mas navega por categorias e entre os documentos em pesquisa de informações pertinentes.

Resposta

7-ditectórios,motores de buscas,metamotores de busca

18. Explique como metadados pode contribuir para recuperação de informação na Web?

Resposta

Metadados são definidos como "dados sobre dados" . São dados, associados a um recursos Web, um documento eletrônico, por exemplo, que permitem recupera-lo, descreve-lo e avaliar sua relevância, manipula-lo (o tamanho de um documentos, ao se fazer "downloading" ou o seu formato, para sabermos se dispomos do programa adequado para manipula-lo), gerencia-lo, utiliza-lo enfim.

19. O que são metabuscadores? Como funcionam?
20. Como funcionam os diretórios Web?
21. Qual é a principal vantagem e desvantagem de um metamotor?
22. Compare os resultados devolvidos pelo Google e Altavista. Quais são as diferenças?
23. Quais são os estilos básicos de busca de um sistema de informação na Web?
24. Cite e explique pelo menos 3 desafios encontrados pelos sistemas de busca na Web?
25. Explique a arquitetura centralizada de um site de busca?
26. Explique a arquitetura distribuída de um site de busca?

Leituras e Outros Recursos

As leituras e outros recursos desta unidade encontram-se na lista de "Leituras e Outros Recursos do curso".

Bibliografias

- R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison-WesleyLongman,
- Boston, MA, 1999.
- Setzer Valdemar.Dado, Informação,Conhecimento e Competencia. DataGramaZero - Revista de Ciência da Informação - n. zero dez/99 .
- Goodoy Viera-Angel Freddy e Dos Santos Garrdo Isora: Folksonomia como uma estratégia para Recuperação Colaborativa da Informação. DataGramaZero - Revista de Ciência da Informação - v.12 n.2 abr11 ARTIGO 02
- Marcondes Carlos H. et al. Saberes e BIBLIOTECAS DIGITAIS SABERES E Práticas. Salvador/BrasíliaUFBA/IBICT 2005 http://livroaberto.ibict.br/bitstream/1/1013/1/Bibliotecas%20Digitais.pdf
- http://exame.abril.com.br/tecnologia/glossario/ - ACESSADO Novembro 2014
- http://marco.uminho.pt/~macedo/thesis/glossario.html ACESSADO NOV 2014
- CASSIOPEIA: UM MODELO DE AGRUPAMENTO DE TEXTOS BASEADO EM SUMARIZAÇÃO.http://nlx.di.fc.ul.pt/~guelpeli/Arquivos/Tese.pdf acessado Novembro 2014
- Implicações da categorização e indexação na recuperação da informação tecnológica contida em documentos de patentes. http://nlx.di.fc.ul.pt/~guelpeli/Arquivos/Tese.pdf acessado Novembro 2014
- Raghu Ramakrishnan,Johannes Gehrke: Sistemas de gerenciamento de banco de dados - 3.ed.
- De Medeiros Ericles Alves, TÉCNICA DE APRENDIZAGEM DE MÁQUINA PARA CATEGORIZAÇÃO DE TEXTOS.Trabalho de Conclusão de Curso- Engenharia da ComputaçãoRecife na Universidade politécnica Pernambuco, junho de 2004
- Hagar Espanha Gomes e Maria Luiza de Almeida Campos,Tesauro e normalização terminológica: o termo como base para intercâmbio de informações ,DataGramaZero - Revista de Ciência da Informação - v.5 n.6 dez/04
- Arquitectura de um SGDB http://www.devmedia.com.br/arquitetura-de-um-sgbd/25007 acessado Novembro 2014
- Base de dados, http://blog.gaudencio.net.br/2012/08/banco-de-dados-conceituando-banco-de_5437.html acessado Novembro 2014
- Ribeiro Rocha,Jaqueline et al.ESTUDO DE METODOLOGIAS PARA A CONSTRUÇÃO DE VOCABULÁRIOS CONTROLADOS NO ÂMBITO DA CIÊNCIA DA INFORMAÇÃO, Universidade Estadual de Londrina (UEL)
- Christopher D. Mannin et al.,Introduction to Information Retrieval ,Online edition (c),2009 Cambridge UP Online edition (c) Cambridge University Press Cambridge, England ,2009
- MARTÍNEZ MÉNDEZ, FranciscoJavier Recuperación de información:modelos,sistemas yevaluación / FranciscoJavier Martínez Méndez.– Murcia:KIOSKO JMC,2004. 106 p. ISBN: 1. Recuperación de Información.I. Título. 025.04.03

- Robson Luís Gomes dos Santos, Usabilidade de interfaces para sistemas de recuperação de informação na web Estudo de caso de bibliotecas on-line de universidades federais brasileiras
- AMORIM, Sergio R. Leusin e CHERIAF, Malik, SISTEMA DE INDEXAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO EM CONSTRUÇÃO BASEADO EM ONTOLOGIA, UFF - Universidade Federal Fluminense,
- Estratégia de busca na recuperação da informação: revisão da literatura, Ci. Inf., Brasília, v. 31, n. 2, p. 60-71, maio/ago. 2002
- ERICK BONFIM MARCELLO ,RECUPERAÇÃO DE DOCUMENTOS TEXTO USANDO UM MODELO PROBABILÍSTICO ESTENDIDO, tese de mestrado , SP 2006
- OSCAR BUGMANN SANDRO, SISTEMA TUTOR PARA ENSINO DE SQL, Trabalho de Conclusão de Curso II do curso de Ciência da Computação — Bacharelado. BLUMENAU,2012
- L e i v a Isi d o r o Gil e L o p e s F u j i t a Ma r i â n g e l a S p o t t i , Política de Indexação Cultura Academica Editora, Marília, 2012

Documentação

1. Sistemas de Recuperação da Informação (Slides Adaptados de Cristopher D. Manning)

2. Recuperação de Informação em Bases de Texto, http://www.di.uevora.pt/~pq/ribt_aula6.pdf acessado em Novembro 2014

3. Prof. Ulrich,Tópicos Especiais em Ciência da Computação:Sistemas de Recuperação da Informação, http://www.dsc.ufcg.edu.br/~ulrich/disciplinas/SRI.html Acessado Novembro de 2014

4. Prof. Ranato Corea, Slide Recuperação de Informação https://drive.google.com/viewerng/

s&srcid=ZGVmYXVsdGRvbWFpbnxyZW5hdG9jb3JyZWF8Z3g6MTI0ODYwMDc3MDIwYjQwOQ acessado Novembro 2014

5. Docent Joaquim Macedo, Bibliotecas Digitais , 2003 http://marco.uminho.pt/disciplinas/UCAN/BD/ acessado Novembro 2014

6. Cendón Beatriz Valadres,Ferramentas de Busca na Web,Ci. Inf., Brasília, v. 30, n. 1, p. 39-49, jan./abr. 2001

7. OLINDA NOGUEIRA PAES CARDOSO, Recuperação de Informação

8. Zuchini Márcio Henrique Modelos Computacionais Adaptativos para Recuperação de Informação Adaptative Computing Models for Information Retrieval , USF, ESAMC

9. Prof. Dr. Leandro Balby Marinho Vocabulário de Termos e Listas de Postings http//www.dsc.ufcg.edu.br/~lbmarinho acessado Novembro 2014

10. Professor: Ivon Rodrigues Canedo , BANCO DE DADOS I

11. Madalena Banhos, Vângela Tatiana,Usuabilidade na recuperação de Informação dissertação no program de pós graduação na UNESP,Marilia 2008

12. Silvana Drumond Monteiro et al AS CATEGORIAS DOS MECANISMOS DE BUSCA: OBJETO EM CONSTRUÇÃO E EM PERMANENTE MODIFICAÇÃO,

13. Kira Tarapanoff (ORG.),Intelegencia,Informação e Conhecimento Brasilia 2006

14. LEONARDO DAITX DE BITENCOURT,UM SISTEMA VOLTADO À INDEXAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO INTEGRADO À ONTOLOGIA Grau de Bacharel em Tecnologias da Informação Araranguá 2013

15. D. Antonio Gabriel López Herrera, Modelos de Sistemas de Recuperacion de Informaci´on Documental Basados en Informaci´on Ling¨ u´ıstica Difusa Memoria de Tesis presentada para optar al grado de Doctor en Informática Granada 2006

16. De Araújo Júnior Rogério Henriquee Kira Tarapanoff Precisão no processo de busca e recuperação da informação: uso da mineração de textos,Ci. Inf., Brasília, v. 35, n. 3, p. 236-247, set./dez. 2006

17. Lêda de Oliveira Monteiro, Igor Ruiz Gome, Thiago Oliveira

18. Etapas do Processo de Mineração de Textos – uma abordagem aplicada a textos em Português do Brasil, 2006

19. Fabricio Breve,Banco de Dados IIntrodução

Mineração de Textos

- 1E. A. M. Morais A. P. L. Ambrósio Technical Report - INF_005/07 - Relatório Técnico December - 2007 - Dezembro
- JEROCIR BOTELHO MARQUES DE JESUSTESAURO: UM INSTRUMENTO DE REPRESENTAÇÃODO CONHECIMENTO EM SISTEMAS DE RECUPERAÇÃO
- DA INFORMAÇÃO ,2002
- Cássio de Albuquerque Melo, Cássio Centro de Informática Mecanismos de Ranqueamento na Web aplicados na Busca e Recuperação de Componentes de Software, Universidade Federal de Pernambuco,2006
- JOHANNA WILHELMINA SMIT e NAIR YUMIKO KOBASHI,COMO ELABORAR VOCABULÁRIO CONTROLADO PARA APLICAÇÃO EM ARQUIVOS COMO FAZER VOL. 10 ,ARQUIVO DO ESTADO/IMPRENSA OFICIAL DO ESTADO ,São Paulo,2003
- Helder Joaquim Carvalheira Gomes ,Text Mining: Análise de Sentimentos na classificação de notícias dissertação de mestrado,Trabalho para obtenção do grau de Mestre em Estatística e Gestão de Informação,Universidade Nova de Lisboa 2012
- LIQUIO TAKAO EDUARDO ,UMA ANÁLISE DE MODELOS E SISTEMAS

- PROBABILÍSTICOS EM RECUPERAÇÃO DE INFORMAÇÃO EM BASES TEXTUAIS,Dissertação submetida à Universidade Federal de Santa Catarina para a obtenção do grau de Mestre em Ciências da Computação,Florianópolis, outubro de 2001
- C. J. van RIJSBERGEN,INFORMATION RETRIEVAL,Information Retrieval Group, University of Glasgow http://www.dcs.gla.ac.uk/Keith/Preface.htm acessado Novembro 2014
- MARCUS VINICIUS CARVALHO GUELPELI, CASSIOPEIA: UM MODELO DE AGRUPAMENTO DE TEXTOS BASEADO EM SUMARIZAÇÃO, Tese de Doutorado apresentada para obtenção do Grau deDoutor,Niterói 2012
- LEANDRO KRUG WIVESTECNOLOGIAS DE DESCOBERTA DE CONHECIMENTO EM TEXTOSAPLICADAS À INTELIGÊNCIA COMPETITIVA,PORTO ALEGRE, JANEIRO DE 2002.
- http://www.leandro.wives.nom.br/pt-br/publicacoes/eq.pdfAcessado Novembro 2014

Sede da Universidade Virtual africana

The African Virtual University
Headquarters

Cape Office Park

Ring Road Kilimani

PO Box 25405-00603

Nairobi, Kenya

Tel: +254 20 25283333

contact@avu.org

oer@avu.org

Escritório Regional da Universidade Virtual Africana em Dakar

Université Virtuelle Africaine

Bureau Régional de l'Afrique de l'Ouest

Sicap Liberté VI Extension

Villa No.8 VDN

B.P. 50609 Dakar, Sénégal

Tel: +221 338670324

bureauregional@avu.org