**COPPE/UFRJ**

# QUALIDADE DE INFORMAÇÃO NA *WEB*: UM PROGNÓSTICO *FUZZY* BASEADO EM METADADOS

Ricardo Oliveira Barros

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Ciências em Engenharia de Sistemas e Computação.

Orientadores: Jano Moreira de Souza
Geraldo Bonorino Xexéo

Rio de Janeiro
Março de 2009

QUALIDADE DE INFORMAÇÃO NA *WEB*: UM PROGNÓSTICO *FUZZY* BASEADO EM METADADOS

Ricardo Oliveira Barros

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Aprovada por:

_____

Prof. Jano Moreira de Souza, Ph.D.

_____

Prof. Geraldo Bonorino Xexéo, D.Sc.

_____

Prof. Geraldo Zimbrão da Silva, D.Sc.

_____

Prof. Luís Alfredo Vidal de Carvalho, D.Sc.

_____

Profa. Ana Maria de Carvalho Moura, D.Ing.

_____

Prof. Ricardo de Almeida Falbo, D.Sc.

RIO DE JANEIRO, RJ - BRASIL

MARÇO DE 2009

Barros, Ricardo Oliveira

Qualidade de Informação na *Web*: Um Prognóstico *Fuzzy* Baseado em Metadados/ Ricardo Oliveira Barros. – Rio de Janeiro: UFRJ/COPPE, 2009.

XXI, 210 p.: il.; 29,7 cm.

Orientadores: Jano Moreira de Souza

Geraldo Bonorino Xexéo

Tese (doutorado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2009.

Referências Bibliográficas: p. 149-162.

1. Qualidade de Informações na *Web*. 2. Lógica *Fuzzy*. 3. Metadados de Qualidade. I. Souza, Jano Moreira, *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

*Aos meus filhos Aline, Ricardinho e*
*Maria Eduarda, pelo amor de cada dia.*

# Agradecimentos

Certamente, não tenho como agradecer a todas as pessoas que tiveram alguma participação na realização deste projeto. Porém, muitas delas tornaram menos difíceis essa longa jornada, e a elas, em especial, quero expressar meus verdadeiros agradecimentos.

Primeiramente faço um agradecimento institucional a UFRJ, em especial, a COPPE/PESC, pelo seu importante papel desempenhado no contexto do país, em vista de nossa atual conjuntura econômica e social. Particularmente, reconheço e agradeço a oportunidade de realização de um curso de alto nível científico, do qual muito me orgulho de ter participado.

Seguem, então, meus agradecimentos pessoais.

Ao Professor Xexéo, que ao longo desses anos, além de orientador tornou-se um amigo com quem aprendi que fazer ciência pode ser divertido e prazeroso. Obviamente a confiança, a compreensão e o seu espírito fraterno demonstrados durante toda jornada que, aliados a sua orientação cirúrgica e participativa, constituíram fatores decisivos à boa condução deste trabalho.

Ao Professor Jano pelo exemplo e pela conduta impecáveis mantidas, tanto como professor quanto como chefe da linha de Banco de Dados. A sua direção e soluções tempestivas aos mais variados tipos de problemas, indiscutivelmente, contribuíram para o crescimento pessoal, acadêmico e profissional de todos que participam dessa complexa rede social. Agradeço, também, por sua orientação, sugestões e correções enriquecedoras ao meu trabalho, e pela colaboração exercida por meio de indicações, sempre oportunas, de publicações, de pessoas e de lugares.

A professora Ana Maria, amiga de longos anos e, certamente, a grande incentivadora e responsável por eu ter desejado e continuado a buscar o caminho do estudo e da pesquisa. A ela agradeço, também, por ter, prontamente e de forma gentil, aceitado a participação nesta banca.

Aos professores Luís Alfredo, Geraldo Zimbrão e Ricardo Falbo, que ao comporem a banca examinadora, emprestam seus conhecimentos para o aperfeiçoamento do trabalho desenvolvido.

Ao amigo Wallace, grande companheiro de jornada, pelas longas horas de dedicação e por tudo que produzimos e publicamos juntos, vencendo o cansaço e os

*deadlines*. Com seu entusiasmo e disposição característicos, muito contribui nas especificações e soluções técnicas desta tese. Obrigado Wallace por estar ao meu lado até o último momento de trabalho.

Ao amigo Bruno que, de forma amiga e prestativa, e a partir de suas experiências e conhecimentos, não mediu esforços ao me prestar a ajuda necessária. Certamente, nossas discussões em busca da melhor solução para os problemas técnicos e conceituais, bem como a sua ajuda prática pouparam-me incontáveis horas de trabalho na fase de conclusão da pesquisa.

Ao Cadú pela amizade, pelo companheirismo e pela cooperação incondicional demonstrada durante todo tempo que trabalhamos no LABBD.

As amigas Jonice e Adriana pelas boas palavras de amizade e incentivo, e pelo ambiente onde a convivência cordata e agradável sempre esteve garantida.

Ao Fabrício e ao Heraldo, por suas fundamentais participações, principalmente, nas especificações técnicas e implementações, sem dúvida, decisivas para a finalização da pesquisa.

As amigas Patrícia e Ana Paula que tão gentilmente se multiplicam para nos atender com paciência e boa vontade nos muitos momentos de dificuldade, sempre dando um jeitinho de resolver tudo da melhor maneira possível.

A equipe da secretaria Solange, Cláudia, Mercedes, Sônia, Carol e Lúcia pelo apoio e orientação indispensáveis durante a nossa atribulada vida acadêmica.

Aos amigos da Marinha, pelo incentivo prestado e pelo apoio e confiança demonstrados durante longos anos de convivência.

Agora os agradecimentos a minha querida família.

Aos meus pais, pela formação moral e pela educação que têm balizado a minha vida.

Aos meus amados irmãos(ãs) Cléa, Zezinho (*in memmorian*), Fernando, Valéria, Toninho, Claudinha, cunhados(as) e sobrinhos(as), pela incansável torcida ao longo de todo esse tempo.

Ao Zé, meu irmão de coração que, do princípio ao fim, tem dividido comigo todas as etapas, boas e más, e com seu espírito fraterno, companheiro e muito sincero, tem me ajudado a tornar menos áridas todas às adversidades que tenho vivido.

A Monica que esteve presente em boa parte desse complicado percurso de vida e de trabalho. A ela meu reconhecimento e gratidão pelo carinho e pelo incentivo dedicados durante todos esses anos.

Aos meus filhos a quem mais amo na vida, Aline, Ricardinho e Duda pelo amor e compreensão a mim dedicados. Vocês, por serem as pessoas que são, me dão muito orgulho de ser seu pai. Por isso mesmo, fazem também de mim, uma pessoa melhor a cada dia.

Não poderia deixar de agradecer a Deus, porque acima de tudo, *nós somos o Seu campo e a Sua construção* (Coríntios 3:9).

A todos: MUITO OBRIGADO!

"*Vinum intra, subsidant mella, superstet oliva"*.
(Vinho do meio, mel de baixo, azeite de cima)
e ...informação de qualidade.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

# QUALIDADE DE INFORMAÇÃO NA *WEB*: UM PROGNÓSTICO *FUZZY* BASEADO EM METADADOS

Ricardo Oliveira Barros

Março/2009

Orientadores: Jano Moreira de Souza
Geraldo Bonorino Xexéo

Programa: Engenharia de Sistemas e Computação

Os usuários na *Web* têm que lidar com um grande volume de informações provenientes de fontes heterogêneas, particularmente, quando não têm conhecimento preciso das suas necessidades de busca. Partes dessas informações são, ou se tornarão em algum momento, inadequadas ao uso, pois muitas vezes são recuperadas informações desatualizadas, imprecisas, inválidas, intencionalmente erradas, falsas ou tendenciosas que, *a priori*, não há como serem avaliadas.

Esta tese propõe um modelo, uma metodologia e uma arquitetura para o prognóstico da qualidade de informações na *Web* com base nos seus metadados. Na metodologia proposta, a lógica *fuzzy* foi adotada como abordagem para a implementação do mecanismo de avaliação automática, em razão da sua habilidade para lidar com conceitos diferenciados e capturar o conhecimento impreciso dos seres humanos.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

WEB INFORMATION QUALITY: A METADATA BASED–*FUZZY* PREDICTION

Ricardo Oliveira Barros

March/2009

Advisors: Jano Moreira de Souza
Geraldo Bonorino Xexéo

Department: Computer Science and Engineering

Web users have to deal with a huge amount of information that is spread out at many sources, particularly, if they have no precise knowledge of their information needs. They constantly retrieve information that is, or may become, outdated, imprecise or invalid, also not forgetting the intentionally wrong, and false or biased information. Considering that users, *a priori*, have no means to assess the quality of the retrieved information, part of it is, or will become at some point, useless.

This thesis proposes a model, a methodology and an architecture for Web information quality prediction, based on its metadata. A *fuzzy* logic approach has been adopted, to implement the automated evaluation mechanism of the proposed methodology, due to its ability to deal with diverse concepts and capture humans' imprecise knowledge.

# Índice

Índice de Figuras ..... xv
Índice de Tabelas ..... xviii
Lista de Termos e Abreviações ..... xx
Capítulo 1 – Introdução ..... 1
1.1 – Motivação ..... 1
1.2 – Definição do Problema ..... 7
1.3 – Hipótese ..... 8
1.4 – Objetivos do Trabalho ..... 9
1.5 – Organização do Trabalho ..... 9
Capítulo 2 – Principais Tópicos de Pesquisa em Qualidade de Informação ..... 13
2.1 – Qualidade de Informação ..... 13
2.2 – Qualidade de Informação na *Web* ..... 17
2.3 – Perspectiva do Consumidor de Informação ..... 20
2.4 – Dimensões de Qualidade de Informação ..... 23
2.5 – Metadados de Qualidade de Informação ..... 33
2.6 – Contextos ..... 39
2.7 – Modelos, Metodologias e Categorias de Ferramentas de Qualidade de Informação ..... 41
2.7.1 – Modelos de Qualidade de Informação ..... 41
2.7.2 – Metodologias de Qualidade de Informação ..... 43
2.7.3 – Categorias de Ferramentas de Qualidade de Informação ..... 46
Capítulo 3 – Enfoques Sobre a Teoria *Fuzzy* ..... 53
3.1 – Breve Introdução sobre a Teoria *Fuzzy* ..... 53
3.2 – Conceitos Básicos dos Conjuntos *Fuzzy* ..... 54
3.2.1 – Conjuntos Nítidos e Conjuntos *Fuzzy* ..... 54
3.2.2 – Função *Fuzzy* de Pertinência ..... 57
3.2.3 – Representação de um Conjunto *Fuzzy* ..... 58
*3.2.3.1 – Conjunto fuzzy* ..... 58
*3.2.3.2 – Suporte* ..... 58
*3.2.3.3 – Supremo* ..... 59
*3.2.3.4 – Normalização* ..... 59

*3.2.3.5 – Conjuntos de corte-α* ..... 59
*3.2.3.6 – Cardinalidade* ..... 60
*3.2.3.7 – Fuzificação* ..... 61
*3.2.3.8 – Defuzificação* ..... 61
*3.2.3.9 – Funções fuzzy* ..... 62
*3.2.3.10 – Agregação de conjuntos fuzzy* ..... 63
3.3 – Números *Fuzzy* ..... 65
3.4 – Variáveis Lingüísticas ..... 66
3.5 – A Lógica *Fuzzy* ..... 68
3.6 – Sistemas Baseados em Conhecimento *Fuzzy* ..... 70
3.7 – Sistemas *Fuzzy* ..... 70
Capítulo 4 – Abordagem Proposta para o Prognóstico de Qualidade de Informações na *Web* ..... 73
4.1 – Modelo Proposto ..... 73
4.1.1 – Pacote Documento *Web* ..... 75
*4.1.1.1 – Classe Documento Web* ..... 76
*4.1.1.2 – Classe Conteúdo* ..... 76
*4.1.1.3 – Classe Metadado* ..... 76
*4.1.1.4 – Classe Contexto* ..... 78
4.1.2 – Pacote Dimensão de Qualidade ..... 79
*4.1.2.1 – Classes Dimensão de Qualidade* ..... 79
*4.1.2.2 – Classe Variável Lingüística e Classe Termo Lingüístico* ..... 80
4.1.3 – Pacote *Type* ..... 81
*4.1.3.1 – Classe Tipo Curva Pertinência* ..... 81
*4.1.3.2 – Classe Escala Grau Importância* ..... 81
4.1.4 – Pacote Avaliação ..... 82
*4.1.4.1 – Classe Pertinência* ..... 82
4.1.5 – Pacote Usuário ..... 83
*4.1.5.1 – Classes Usuário e Especialista* ..... 83
4.1.6 – Pacote Importância Dimensão de Qualidade ..... 84
*4.1.6.1 – Classes Importância Dimensão Qualidade Especialista e Importância Dimensão Qualidade Usuário* ..... 84
4.1.7 – Pacote Regra ..... 88

*4.1.7.1 – Classe Regra* ..... 88
4.2 – Metodologia Proposta ..... 90
4.2.1 – Instanciação do Modelo de QI na *Web* ..... 92
4.2.2 – Passos da Metodologia para o Prognóstico de QI na *Web* ..... 93
Capítulo 5 – Arquitetura da Aplicação para o Prognóstico de Qualidade de Informação na *Web* ..... 95
5.1 – Arquitetura e Detalhes Técnicos da Implementação ..... 95
Capítulo 6 – Aplicação e Validação da Abordagem Proposta ..... 99
6.1 – Provas de Conceito ..... 100
6.1.1 – Prova de Conceito Usando Funções *Fuzzy* de Transformação ..... 101
6.1.2 – Prova de Conceito Usando Regras *Fuzzy* de Inferência ..... 109
6.2 – Aplicação Colaborativa para Construção de uma Base de Testes Contendo Páginas *Web* Avaliadas ..... 115
6.2.1 – Visão Geral ..... 116
6.2.2 – Avaliação Colaborativa ..... 116
6.2.3 – FoxSet ..... 117
6.2.4 – Resultados Preliminares ..... 125
6.3 – Experimento para Avaliação da Abordagem Teórica Proposta ..... 128
6.3.1 – Definição ..... 128
6.3.2 – Planejamento ..... 129
*6.3.2.1 – Análise de Ameaças à Validade dos Resultados* ..... 129
6.3.3 – Avaliação Automática do Dataset ..... 132
6.3.4 – Avaliação Manual do Dataset ..... 132
6.3.5 – Método Analítico de Comparação dos Resultados ..... 134
*6.3.5.1 – Análise quantitativa dos resultados* ..... 134
*6.3.5.2 – Heurísticas Relacionadas à Execução do Experimento* ..... 136
*6.3.5.3 – Análise qualitativa dos resultados* ..... 138
6.4 – Análise Comparativa entre os Resultados de Ordenação do Google® e da Avaliação da Qualidade ..... 139
6.4.1 – Definição ..... 139
6.4.2 – Cálculo de Precisão e Cobertura ..... 140
Capítulo 7 – Conclusão e Trabalhos Futuros ..... 143
7.1 – Conclusão ..... 143

7.2 – Trabalhos Futuros ..... 146
Referências Bibliográficas ..... 149
Anexo I – Descrição das Dimensões de Qualidade ..... 163
Anexo II – Resultados das Avaliações Automática e Manual ..... 169
Anexo III – Questionário de Avaliação Manual ..... 173
Anexo IV – Cálculo Precisão e Cobertura pelos Valores de Ordenação do Google® ..... 175
Anexo V – Cálculo Precisão e Cobertura pelos Valores de Ordenação da Avaliação de Qualidade ..... 193

# Índice de Figuras

Figura 1-1: Esquema Produtor-Consumidor de Dados e os Filtros de Seleção ..... 3
Figura 1-2: Organização da Tese ..... 11
Figura 2-1: Tópicos de Pesquisa em Qualidade de Informação (Fonte: Batini & Scannapieco, 2006) ..... 13
Figura 2-2: Modelo de Qualidade de Dados (Fonte: Berti-´Equille, 2007) ..... 42
Figura 2-3: Principais Fases das Metodologias de Avaliação de QI (Adaptada de Amicis & Batini, 2004) ..... 44
Figura 3-1: Comparação de um número real e um intervalo nítido com um número *fuzzy* e um intervalo *fuzzy* respectivamente (Fonte: Klir & Yuan, 1995) ..... 66
Figura 3-2: Variável lingüística "Idade" (Fonte: Klir & Yuan, 1995) ..... 67
Figura 4-1: Modelo para o Prognóstico de QI na *Web* ..... 74
Figura 4-2: Modelo para o Prognóstico de QI na *Web* Organizado em Pacotes ..... 75
Figura 4-3: Pacote Documento *Web* ..... 75
Figura 4-4: Pacote Dimensão de Qualidade ..... 79
Figura 4-5: Pacote *Type* ..... 81
Figura 4-6: Pacote Avaliação ..... 82
Figura 4-7: Pacote Usuário ..... 83
Figura 4-8: Pacote Importância Dimensão de Qualidade ..... 84
Figura 4-9: Pacote Regra ..... 88
Figura 4-10: Metodologia para o Prognóstico de QI na *Web* ..... 91
Figura 4-11: Instanciação do Modelo de Qualidade de Informação na *Web* ..... 92
Figura 5-1: Arquitetura da Aplicação para o Prognóstico de Qualidade de Informação na *Web* ..... 96
Figura 5-2: Grafo de um Conjunto de Páginas *Web* no Formato Pajek ..... 97
Figura 6-1: Variável Lingüística "Atualidade" (Adaptada de Klir & Yuan, 1995) ..... 103
Figura 6-2: Modelo para o Mapeamento das Funções de Pertinência para os Subconjuntos *Fuzzy* ruim, regular e bom para as Variáveis Lingüísticas reputação e completeza ..... 104
Figura 6-3: Modelo para o Mapeamento das Funções de Pertinência para os Subconjuntos *Fuzzy* ruim, regular e bom para a Variável Lingüística atualidade 104

Figura 6-4: Resultados da Defuzificação para http://www.economist.com/surveys/... 107
Figura 6-5: Resultados da Defuzificação para http://www.economistconferences.com/ ..... 107
Figura 6-6: Resultados da Defuzificação para http://www.theworldin.com/ ..... 108
Figura 6-7: Resultados da Defuzificação para http://www.economist.com/opinion.... 108
Figura 6-8: Resultados da Defuzificação para http://www.scottrade.com/index.asp? supbid=68597 ..... 109
Figura 6-9: Resultados da Defuzificação LOM para http://www.economist.com/surveys/ ..... 112
Figura 6-10: Resultados da Defuzificação LOM para http://www.economistconferences.com/ ..... 112
Figura 6-11: Resultados da Defuzificação LOM para http://www.theworldin.com/ ... 113
Figura 6-12: Resultados da Defuzificação LOM para http://www.economist.com/opinion ..... 113
Figura 6-13: Resultados da Defuzificação LOM para http://www.scottrade.com/index.asp? supbid=68597..... 114
Figura 6-14: Contribuição de Completeza e Atualidade para PC ..... 114
Figura 6-15: Contribuição de Reputação e Atualidade para PC..... 115
Figura 6-16: Contribuição de Reputação e Completeza para PC ..... 115
Figura 6-17: Plugin Foxset para o Firefox ..... 117
Figura 6-18: Processo de Construção de Datasets no Foxset ..... 118
Figura 6-19: Perfis do Foxset ..... 119
Figura 6-21: Definição das Perguntas para um *Dataset* ..... 120
Figura 6-20: Parâmetros de Criação do *Dataset*..... 120
Figura 6-22: Parâmetros de Criação do *Dataset*..... 121
Figura 6-23: Definição Números Fuzzy Triangulares..... 122
Figura 6-24: Definição das Escalas de Avaliação para um *Dataset*..... 123
Figura 6-25: Avaliação Manual de um Dataset ..... 124
Figura 6-26: Seleção do Conjunto de Perguntas para Avaliação Manual de um *Dataset* ..... 124
Figura 6-27: Finalização do *Dataset* ..... 125
Figura 6-28: Subconjunto em XML Resultante do Dataset “Economia”..... 127
Figura 6-29: Distribuição de Freqüências das Avaliações Manuais..... 134

Figura 6-30: Distribuição de Freqüências das Avaliações Automáticas ..... 135
Figura 6-31: Percentual de Interseção das Freqüências das Avaliações Automáticas e Manuais ..... 135
Figura 6-32: Interseção das Freqüências das Avaliações Automática e Manual ..... 136
Figura 6-33: Cobertura versus Precisão..... 141
Figura 6-34: Cobertura versus Média Harmônica ..... 141

# Índice de Tabelas

Tabela 2-1: Iniciativas de Avaliação de Qualidade e suas Abordagens ..... 16
Tabela 2-2: Questões Específicas de Qualidade de Informações na *Web* (Adaptada de Caro & Calero, 2008) ..... 18
Tabela 2-3: Domínios e Estruturas das Abordagens de Avaliação de Qualidade de Informações e Dados (Adaptada e estendida de Burgess & Gray, 2004)..... 26
Tabela 2-4: Atributos de Qualidade Propostos para Diferentes Domínios no Contexto da *Web* (Fonte: Caro & Calero, 2008)..... 32
Tabela 2-5: Níveis e Categorias de Metadados (Propostas por Rothenberg, 1996)..... 34
Tabela 2-6: Fontes de Captura de Metadados na *Web*..... 37
Tabela 2-7: Exemplos de Ferramentas de QI na *Web* (Adaptada de Eppler & Muenzenmayer, 2002) ..... 50
Tabela 3-1: Exemplo de conjuntos *fuzzy* (Fonte: Klir & Folger, 1988) ..... 57
Tabela 3-2: Categorias genéricas de aplicações de sistemas *fuzzy* (Fonte: Munakata & Ani, 1994)..... 71
Tabela 4-1: Restrições de Integridade do Modelo..... 73
Tabela 4-2: Metadados originais e funções de derivação..... 78
Tabela 4-3: Escala de Graus de Importância das Dimensões de Qualidade por Contextos (Adaptada de Xexéo, 1996)..... 85
Tabela 6-1: Valores dos Metadados Originais e Metadados Derivados..... 101
Tabela 6-2: Resultados da Fuzificação com os Graus de Pertinência dos Subconjuntos Fuzzy ruim, regular e bom para Variáveis Lingüísticas atualidade, reputação e completeza..... 105
Tabela 6-3: Resultados da Defuzificação e da Ordenação dos Documentos Web..... 109
Tabela 6-4: Base de Regras *Fuzzy* para o Contexto Economia ..... 110
Tabela 6-5: Resultados da Defuzificação pelos métodos LOM, MOM, BOA, SOM e COG (PC) ..... 111
Tabela 6-6: Melhores Resultados Selecionados pelo Coordenador do *Dataset*..... 126
Tabela 6-7: Avaliação Colaborativa para a pergunta Q1 ..... 126
Tabela 6-8: Avaliação Automática com Ordenação Básica ..... 128
Tabela 6-9: Avaliação Automática com Ordenação P de N ($P = 2$ e $N = 3$)..... 128

Tabela 6-10: Amplitude das Avaliações por Intervalos de Classificação ..... 136

# Lista de Termos e Abreviações

| TERMO | DESCRIÇÃO |
|---|---|
| API | Interface de Programação de Aplicativos *(Application Programming Interface)* |
| BOA | Método do Bissetor de Área |
| BRI | Busca e Recuperação de Informações |
| CMS | Sistema de Gerenciamento de Conteúdo (*Content Management System*) |
| COG | Método do centro de gravidade, do centróide ou centróide da área (*Center of Gravity*) |
| DNS | Sistema de Nomes de Domínios (*Domain Name System*) |
| HITS | *Hyperlink Induced Topic Search* |
| HTML | *HyperText Markup Language* |
| IP | *Internet Protocol* |
| JUNG | *Java Universal Network/Graph Framework* |
| LOM | Método do Maior Valor Absoluto dos Máximos |
| MOM | Método do Valor Médio dos Máximos |
| SOM | Método do Menor Valor Absoluto dos Máximos |
| MySQL | É um sistema de gerenciamento de banco de dados (SGBD), que utiliza a linguagem SQL (*Structured Query Language* - Linguagem de Consulta Estruturada) como interface |
| OWA | Média de Pesos Ordenados (*Ordered Weighted Averaging*) |
| PC | Prognóstico Composto de Qualidade de Informação |
| PHP | *Hypertext Preprocessor* |
| PICS™ | Plataforma para Seleção de Conteúdo na Internet (*Platform for Internet Content Selection*) |
| PS | Prognóstico Singular de Qualidade de Informação |
| QD | Qualidade de Dados |

| | |
|---|---|
| QI | Qualidade de Informações ou Qualidade de Informações |
| RI*nn* | Restrição de Integridade (*nn* – número da RI) |
| ROI | Retorno sobre o Investimento (*Return on Investment*) |
| SBC | Sistemas Baseados em Conhecimento |
| TDWI | *The Data Warehouse Institute* |
| TIC | Tecnologia de Informação e Comunicações |
| UML | Linguagem de Modelagem Unificada (*Unified Modeling Language*) |
| URL | Localizador de Recursos Universal (*Uniform Resource Locator*) |
| VV&C | Verificação, Validação e Certificação |
| XML | *eXtensible Markup Language* |

# Capítulo 1 – Introdução

Uma busca no Google® com os termos "qualidade de informação" retorna algo em torno de 19 milhões de páginas. Se for em inglês – "*information quality*"–, serão 35 milhões de páginas. Esses são indicadores de que as questões relacionadas à qualidade de informação estão bastante difundidas.

A nossa pesquisa demonstra que ao longo dos últimos anos, as técnicas de gerenciamento e consultas de dados na *Web* têm amadurecido significativamente. No entanto, abordagens práticas e melhor fundamentadas para avaliar, ou mesmo garantir aos usuários, a qualidade das informações ainda são insuficientes.

Neste capítulo introdutório, são apresentadas as seguintes seções: a motivação, o problema a ser estudado, a hipótese da proposta da pesquisa, o objetivo deste trabalho e a estrutura e a organização do texto.

## 1.1 – Motivação

Já há algum tempo, o volume de informação disponível aos usuários da *Web* tem aumentado de forma considerável[1]. A maior disponibilidade dos meios físicos de armazenamento e de novas tecnologias, bem como as alternativas de inclusão digital francamente em expansão[2] trouxeram à maioria desses usuários uma incrível capacidade de acessar e acumular informação[3] (GERTZ, OZSU *et al.*, 2004).

Com a difusão das redes e da Internet, novos tipos de dados surgiram e foram investigados. A própria definição de "dados" foi modificada e estendida para abranger "qualquer tipo de informação que seja analisada sistematicamente" (BATINI, BARONE *et al.*, 2008).

Dentre esses novos tipos, os Dados *Web* são de maior interesse para nossa pesquisa. Esses tipos de dados são caracterizados por possuírem formatos não

[1] Google® anunciou em 25JUL2008 que havia indexado mais de um trilhão de URL únicas na Internet, marco que surpreendeu até mesmo os seus engenheiros, segundo *post* no blog oficial da companhia. (http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html).

[2] http://www.digitalinclusion.net/.

[3] Apesar da diferença entre os termos "dado" (reapresentação/notação) e "informação" (significado/denotação), nesta tese eles são usados indistintamente.

convencionais e pouco controle sobre eles. Apesar disso, freqüentemente constituem a fonte primária de informações para inúmeras atividades (DASU & JOHNSON, 2003).

Em particular, os dados *Web* não correspondem necessariamente aos dados extraídos dos bancos dados accessíveis na *Web*. Gertz & Ozsu (2004) os consideram como formas heterogêneas de dados coletados por um *Web crawler*. No contexto desta tese, são as páginas *Web*.

Nos ambientes de intranets corporativas e da Internet, a *Web* possibilita a busca de informações sobre um número ilimitado de categorias ou assuntos, numa vasta quantidade de domínios de aplicações, tais como: sistemas comerciais, governo eletrônico, ciências da vida e sistemas de bibliotecas eletrônicas dentre outros (BATINI & SCANNAPIECO, 2006). Esse fato provoca a sobrecarga de informações, por vezes de baixa qualidade, em que partes das informações criadas são, ou se tornarão em algum momento, inadequadas ao uso. Na maioria dos casos, a velocidade de criação dessas informações é relativamente maior que a capacidade dos usuários de utilizá-las (LYMAN & HAL, 2003) (O'NEIL, LAVOIE *et al.*, 2003) (ISC, 2007).

Além disso, as informações acumuladas apresentam diferentes níveis de qualidade de acordo com suas fontes de origem, que podem variar desde empresas multinacionais até pessoas com pouco ou nenhum conhecimento. A heterogeneidade dessas fontes também resulta numa diversidade de formatos, nos quais a informação é armazenada e visualizada (BURGESS, GRAY *et al.*, 2004).

Com o intuito de melhor caracterizar, ou mesmo de formalizar as dimensões e aspectos de qualidade de informação, é importante reconhecer que o tema não pode ser estudado isoladamente, tomando-se como exemplo o contexto de apenas uma aplicação. Ressalta-se, portanto, que existem processos e fluxos de trabalho tipicamente complexos no gerenciamento de dados e que, de forma necessária, a qualidade tem sido estudada considerando todos os ciclos desse gerenciamento.

Os componentes desses ciclos podem ser melhor representados pelo esquema mostrado na Figura 1-1. Esse esquema, cuja idéia é propor um Sistema Produtor-Consumidor de Dados, foi estendido e adaptado de Gertz & Ozsu (2004).

A infra-estrutura de sistemas complexos na *Web* compreende muitos Produtores, Repositórios e Consumidores. Em cada um deles, a análise e a caracterização das

questões da qualidade de dados, naturalmente, tornam-se mais difíceis por causa das complexidades subjacentes de cada sistema.

Obviamente, o esquema é bastante simples e apresenta somente os componentes mais importantes, comumente encontrados nesses ambientes, pois dependendo do domínio da aplicação, pode não ser necessária a representação de todos eles. Por exemplo, num contexto de sistemas de informação baseado na *Web*, o Produtor e o Repositório de Dados são, quase sempre, a mesma entidade.

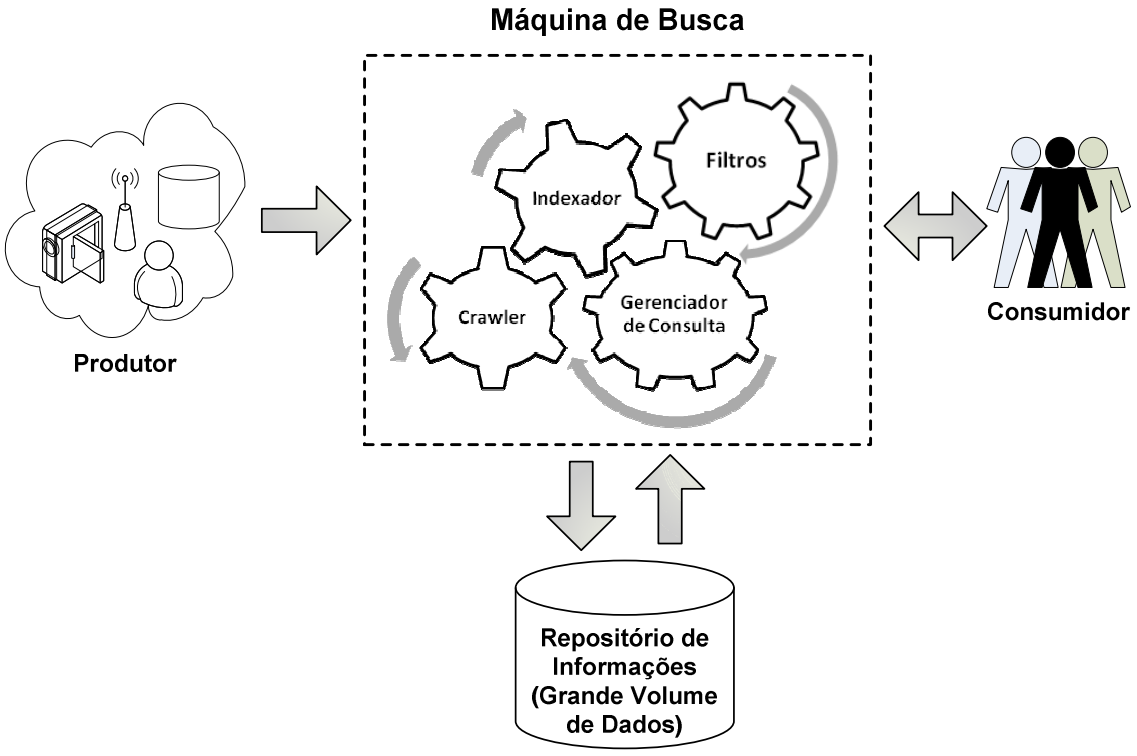



**Figura 1-1: Esquema Produtor-Consumidor de Dados e os Filtros de Seleção**

A seguir são detalhados todos os componentes da Figura 1-1.

**Produtor dos Dados** – Todas as fontes de dados são representadas por um único elemento, que produz os dados nos diversos assuntos, em formatos heterogêneos e com velocidade aleatória. Eles podem ser pessoas, sistemas, bancos de dados, satélites meteorológicos que enviam uma seqüência de imagens de uma determinada região, câmeras de segurança, etc. Além dos dados, existem os metadados relacionados como, por exemplo, a *data de publicação*, a *fonte*, os *autores*, o *tamanho* e o *assunto* de um documento publicado na *Web*. Esses metadados também devem ser recuperados, uma vez que serão úteis para organizar e filtrar informações de acordo com os requisitos propostos pelos usuários.

**Consumidor dos Dados** – É externo aos sistemas e representa todos os tipos de usuários que acessam dados. Eles podem ser pessoas, navegadores, sistemas específicos de organizações, bancos de dados, etc. A fim de auxiliar na avaliação da qualidade dos dados, é importante, em alguns casos, acompanhar e armazenar as preferências dos consumidores de dados como, por exemplo, sítios e domínios visitados.

**Repositório de Informações** – Esse componente representa o repositório que provê todas as informações aos Consumidores de Dados. Ele armazena todos os tipos de dados, estruturas de dados e metadados, por um período ilimitado de tempo, dependente dos requisitos dos usuários. Como o volume disponível de informações tem aumentado de forma considerável, a sua gerência tornou-se mais difícil, acarretando, conseqüentemente, muitos problemas que são discutidos mais adiante.

**Máquina de Busca de Informações** – Esse componente representa o mecanismo de busca das informações providas pelo Produtor de Dados. Em geral possui um *crawler* que percorre a *Web* e captura os documentos através de seus *links*, e um indexador que processa e indexa os documentos capturados. Além disso, dispõe de um gerenciador de consultas que trata as consultas dos consumidores de dados e retorna os documentos relevantes a partir da base de documentos indexados (BRIN & PAGE, 1998).

Resumidamente, a relevância é entendida como o grau de importância atribuído pelos mecanismos de busca da *Web*. A relevância de uma página (*site*) irá determinar a posição que ela ocupará no resultado de uma busca para um determinado termo (palavra-chave) pesquisado. Assim, quanto maior a relevância de uma página, melhor será a posição ocupada por ela na listagem resultante da busca (SHARMA, 2008).

Os mecanismos de busca Google®, Yahoo®, Live Search®, Alta Vista® e outros grandes portais funcionam de maneira parecida e seguem alguns critérios básicos, a seguir exemplificados, para determinar o grau de relevância de uma página para um determinado termo buscado (LEE, LEONARD *et al.*, 2008):

- Termos utilizados pelos usuários ao fazer a busca;
- O número de *links* (vindos de outros *sites*) que apontam para a página, bem como *links* de outros *sites* que são referenciados;

- A freqüência com que o termo buscado aparece na página, em relação a outras palavras e termos contidos na página;
- A localização do termo buscado na página. Por exemplo, se o termo aparece desde a parte inicial do código HTML da página, ela se torna automaticamente mais relevante para aquele termo específico em relação a páginas que contenham o termo na parte final do código HTML; e
- Códigos HTML 'limpos', sem imagens ou *links* quebrados e páginas sempre disponíveis, garantem uma relevância maior em relação àquelas páginas com códigos HTML mal escritos.

Essas máquinas de busca pressupõem um comportamento pró-ativo. Por exemplo, um *crawler* pode acessar páginas *Web* imediatamente após a sua geração, ou a atualização das páginas existentes, com a finalidade de minimizar as diferenças entre as datas de atualização e de busca. Além disso, elas provêem os filtros de pré ou pós-seleção dos resultados.

O filtro de pré-seleção realiza uma análise de conteúdo após a fase de busca, de acordo com um conjunto de especificações do usuário. Pode-se citar como exemplo, a Máquina de Busca, ao analisar o conteúdo das páginas para filtrar e recuperar somente as páginas relevantes aos usuários, de acordo com um argumento de pesquisa.

O filtro de pós-seleção seleciona os resultados fornecidos pelo Gerenciador de Consultas e já armazenados no Repositório de Informações. Ele realiza uma análise de conteúdo de acordo com os metadados, o contexto e as características definidas dos usuários. Por exemplo, a pós-seleção para filtrar o conteúdo das páginas realmente relevantes aos usuários, de acordo com as dimensões de qualidade: *completeza*, *atualidade* e *reputação* para o contexto de *economia*.

Com tanta informação disponível, a qualidade tornou-se um importante discriminador na decisão de quais informações podem ser usadas e quais devem ser descartadas (TARAPANOFF, 2002) (BURGESS, GRAY *et al.*, 2004). Os filtros, portanto, eliminam as informações de baixa qualidade, reduzindo o volume de dados armazenados, contribuindo, conseqüentemente, para a minimização dos problemas e dos seus efeitos relacionados (BELKIN & CROFT, 1992) (MAES, 1994).

Não é fato recente que as dificuldades para identificar, separar e avaliar a qualidade da informação têm causado prejuízos financeiros e comprometido os processos de tomada de decisão, tanto para as organizações, quanto para as pessoas de um modo geral (REDMAN, 1998).

Informações de baixa qualidade ocasionam falhas nos processos de negócio e custos adicionais relacionados às pessoas, aos materiais, ao tempo, ao dinheiro e até mesmo a perda definitiva de clientes. Apesar de algumas empresas entenderem a importância dessa qualidade, ainda assim, a maioria delas não está atenta aos verdadeiros impactos causados em decorrência da baixa qualidade dos dados (ECKERSON, 2002).

English (1999) ressalta que: "…os custos para negócio com a falta de qualidade de dados, incluindo custos irrecuperáveis, retrabalho com produtos e serviços e perdas e extravios de receitas podem ser maiores que 10% a 25% da receita ou do orçamento total da organização". O TDWI (*The Data Warehouse Institute*)[4] estima que problemas relacionados à qualidade de dados custem, para os Estados Unidos, mais de US$ 600 bilhões por ano. Apesar disso, decisões baseadas em informações resultantes de consultas analíticas são tomadas diariamente nas empresas, sem que haja um conhecimento prévio do grau de qualidade dos dados envolvidos, aumentando o risco da produção de resultados inesperados, em função dessa má qualidade (ECKERSON, 2002).

Outras questões mais específicas e não menos importantes também precisam ser direcionadas, como por exemplo: quais são as dimensões de qualidade de interesse? As técnicas atualmente adotadas pelo Google®, Yahoo®, Live Search® e Alta Vista® são satisfatórias, ou existem outras técnicas de avaliação melhores e mais precisas? (GERTZ, OZSU *et al.*, 2004).

A sobrecarga de informação e a qualidade de informação são, portanto, importantes aspectos que têm de ser considerados para a prevenção dos usuários na busca por informações relevantes.

Assegurar a qualidade da informação é igualmente importante e difícil. Assim, obter informação de alta qualidade é uma batalha que jamais será totalmente vencida.

[4] http://tdwi.org/.

Em parte porque o que constitui essa vitória não está claro e, também, porque as diversas partes envolvidas possuem visões diferentes a respeito da definição desse sucesso.

Apesar de extensa discussão na literatura, não há consenso sobre uma abordagem mais apropriada para melhorar a qualidade da informação obtida por meio dos mecanismos de busca na *Web*, tanto pela eficácia das propostas, quanto pelos benefícios esperados (RAMOS-LIMA, MAÇADA *et al.*, 2006). Contudo, todos os interessados concordam que o esforço para alcançar um padrão de informação de alta qualidade exige prioridade. Caso contrário, pode haver conseqüências indesejáveis para as pessoas e para a existência das organizações (BALLOU, MADNICK *et al.*, 2004).

## 1.2 – Definição do Problema

Os usuários da *Web* têm de lidar com um grande volume de informações provenientes de fontes heterogêneas, particularmente, quando não têm conhecimento preciso das suas necessidades de busca (BURGESS, GRAY *et al.*, 2004). Muitas vezes, são recuperadas informações desatualizadas, imprecisas, inválidas, intencionalmente erradas, falsas ou tendenciosas, que, *a priori*, não há como serem avaliadas.

Há ainda os problemas considerados legítimos, como no caso de informações que variam quanto a sua qualidade em razão do tempo, ou seja, mesmo valores que eram bons e corretos em uma base de dados, num dado momento, podem tornar-se ruins e incorretos num momento subseqüente[5] (TWIDALE & MARTY, 1999) (STVILIA, 2007).

Outro problema bastante difundido é o de manter os dados atualizados, tomando-se por premissa a dinâmica da atualização, que pode variar desde os casos extremos – nos quais os dados mudam constantemente –, até os que se modificam com baixíssima freqüência (BATINI & SCANNAPIECO, 2006).

É importante ressaltar que os problemas descritos também ocorrem, de forma semelhante, em ambientes coorporativos complexos que mantêm sua *Web* própria na forma de uma ou mais intranets. Por exemplo, o BNDES[6] possui dados em SGBD

[5] Entropia de Dados.

[6] Banco Nacional de Desenvolvimento Econômico e Social - BNDES, ex-autarquia federal criada pela Lei nº 1.628, de 20 de junho de 1952, foi enquadrado como uma empresa pública federal, com

distintos e diferentes portais de acesso para as mais variadas aplicações. Há, também, a transferência de informações de interesse, entre o Banco e inúmeros parceiros, que, por sua vez, também possuem seus portais, seus sítios *Web* e suas aplicações.

Existem outros problemas igualmente importantes, aliados aos fatos já mencionados, quanto à sobrecarga de informação e às diferentes abordagens das fontes para atribuição de qualidade. Entre eles podem ser resumidos e destacados (KIM, KISHORE *et al.*, 2005):

- A sobrecarga do processamento realizado, em vista da grande quantidade de dados armazenados;
- A maior dificuldade no processo cognitivo e a desorientação do usuário durante a busca de informações; e
- A variação nos critérios de avaliação em virtude dos perfis de usuários e dos contextos de utilização.

A conscientização do problema e de seus impactos é o primeiro passo crítico em direção à resolução desses problemas (REDMAN, 1998) e, nesse sentido, algumas políticas podem ser postas em prática para lidar com as questões identificadas. Porém, os problemas podem permanecer quando as mudanças são lentas e consideradas imperceptíveis, ou quando não justificam o custo do acompanhamento (TARAPANOFF, 2002).

A criação de estratégias e mecanismos computacionais de auxílio aos usuários, no tratamento dos problemas e melhoria de qualidade da informação obtida por meio dos mecanismos de busca na *Web*, possibilita a diminuição do custo e do tempo despendidos no curso dessas atividades. Além disso, a avaliação automática dos recursos da *Web*, valendo-se dos vários critérios e dimensões de qualidade, é um novo campo de pesquisa emergente de várias disciplinas (MANDL, 2006).

## 1.3 – Hipótese

Nos processos de BRI (Busca e Recuperação de Informações) na *Web* a qualidade das informações pode afetar, de forma significativa, os resultados da seleção

personalidade jurídica de direito privado e patrimônio próprio, pela Lei nº 5.662, de 21 de junho de 1971. O BNDES é um órgão vinculado ao Ministério do Desenvolvimento, Indústria e Comércio Exterior e tem como objetivo apoiar empreendimentos que contribuam para o desenvolvimento do país. http://www.bndes.gov.br/.

realizada sobre os conjuntos de páginas recuperadas. Se as páginas puderem ser adequadamente avaliadas, identificadas e separadas por subconjuntos, de acordo com a sua qualidade, é mais provável que aquelas com baixa qualidade possam ser descartadas, e que haja a diminuição das dificuldades depreendidas pelos usuários durante a seleção.

Se forem criados mecanismos automáticos que possibilitem o fornecimento de prognósticos de qualidade que sejam semelhantes às percepções do julgamento humano, se tornará viável apresentá-los junto com os resultados das buscas. Conseqüentemente, os usuários – organizações ou as pessoas de um modo geral –, contarão com um instrumento a mais para a orientação das suas seleções e para o atendimento dos seus requisitos de pesquisa, aumentando assim, a confiabilidade no processo de BRI como um todo.

## 1.4 – Objetivos do Trabalho

Concernente ao que foi apresentado, nossa pesquisa enfatizou a investigação das estratégias de auxílio aos usuários no tratamento dos problemas anteriormente destacados, aplicáveis ao contexto da qualidade de informações resultantes dos mecanismos de busca na *Web*.

Esta tese, portanto, tem por objetivo propor um modelo, uma metodologia e uma arquitetura para o prognóstico de qualidade de informações na *Web*, baseado em seus metadados. A lógica *fuzzy* foi adotada como abordagem para implementação do mecanismo de avaliação automática da metodologia proposta, em razão da sua habilidade para lidar com conceitos diferenciados e capturar o conhecimento impreciso dos seres humanos.

## 1.5 – Organização do Trabalho

Para alcançar os nossos objetivos, foram realizadas durante a pesquisa as atividades a seguir sumarizadas. Essas atividades, bem como os artefatos e produtos gerados, estão especificadas nos capítulos da tese:

- Revisão bibliográfica sobre as áreas de estudo acima mencionadas. Esses estudos serviram de direcionamento e subsídios para a definição da

abordagem e especificação dos mecanismos na construção da proposta para o prognóstico de qualidade de informação;

- Estudo, definição e implementação de um modelo e de uma metodologia para o prognóstico *fuzzy* de qualidade de informações de páginas *Web*, com base nos seus metadados;
- Estudo, definição e implementação dos componentes e da arquitetura tecnológica, como solução para suportar a abordagem teórica proposta;
- Implementação dos protótipos para realização das provas de conceito e dos estudos de caso, visando à verificação e ao refinamento das idéias, inicialmente propostas;
- Desenvolvimento e a implementação de uma aplicação colaborativa com o objetivo de tornar mais fácil e mais ágil o trabalho de construção das bases de testes. O propósito dessa aplicação é aliar ao processo de alimentação manual das páginas, a alimentação e o prognóstico de qualidade realizados automaticamente por meio da abordagem proposta;
- Avaliação da abordagem teórica proposta realizada com o auxílio de um método experimental de comparação dos resultados de avaliação sobre uma base de testes. Durante a análise foram comparados os resultados oriundos das avaliações realizadas automaticamente, em contrapartida aos resultados obtidos por meio de julgamento humano; e
- Avaliação comparativa entre os resultados de ordenação das páginas obtidos pelo prognóstico de qualidade e pelo Google®, com base nos cálculos de precisão e cobertura, e das suas médias harmônicas.

A Figura 1-2 ilustra a organização da tese. O conteúdo dos capítulos é detalhado a seguir.

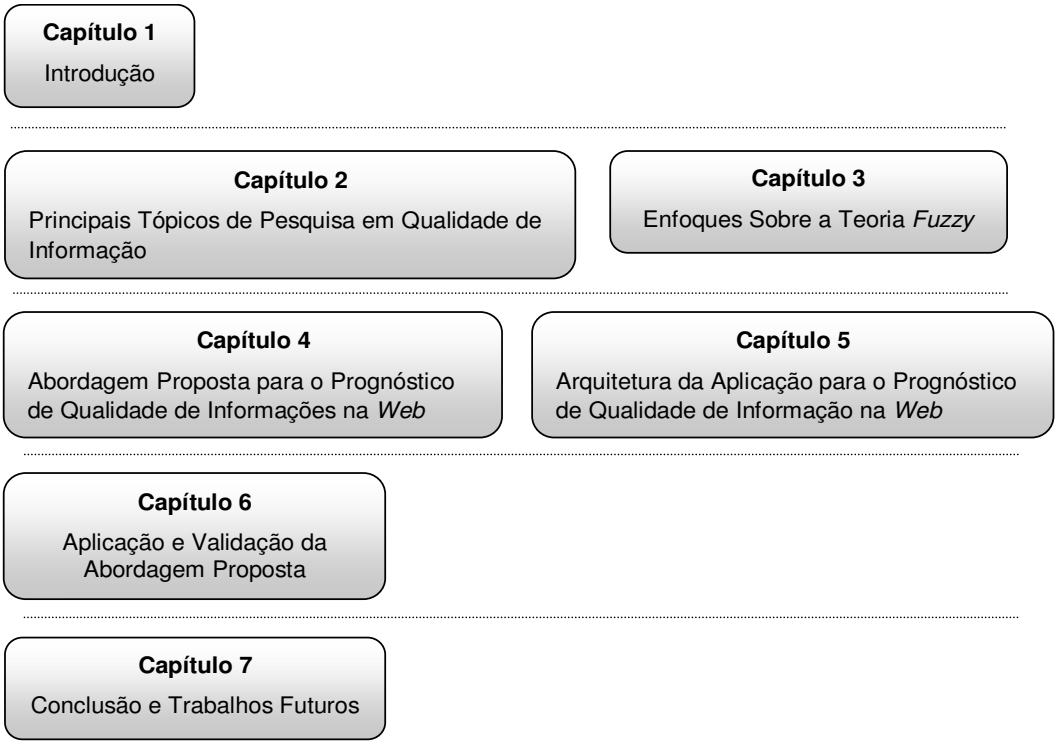



**Figura 1-2: Organização da Tese**

O **Capítulo 2** apresenta uma revisão dos aspectos mais importantes dos assuntos relacionados à qualidade de dados e informações, a metadados e contextos, dentro do escopo pesquisado.

O **Capítulo 3** sumariza os enfoques da Teoria *Fuzzy*, inerentes ao entendimento da solução.

O **Capítulo 4** refere-se à apresentação da nossa proposta de trabalho de tese como uma abordagem aos problemas destacados e descreve um exemplo de aplicação do prognóstico de qualidade de páginas *Web*, baseado em seus metadados.

O **Capítulo 5** explora a arquitetura e seus detalhes técnicos de implementação, desenvolvida para suportar o modelo e a metodologia proposta.

O **Capítulo 6** demonstra funcionalmente os resultados obtidos no emprego das estratégias propostas neste trabalho. Inicialmente, foram implementadas duas provas de conceito em duas abordagens *fuzzy*. Em seguida, é demonstrada a aplicação colaborativa desenvolvida como uma extensão do FoxSet[7], e a descrição dos estudos de caso que

[7] O FoxSet é o protótipo de uma ferramenta para a construção de *datasets* desenvolvido no PESC/COPPE, na disciplina de BRI, ministrada pelo Prof. Geraldo Xexéo. Ele consiste em um *plugin*

comparam os nossos resultados preliminares e os resultados obtidos pelo Google®, pelo Yahoo® e pelo Live Search®. O capítulo também descreve uma avaliação realizada com o auxílio de um método experimental para a comparação de resultados em maior escala. Além disso, há uma outra avaliação comparativa dos resultados de ordenação das páginas baseado nos cálculos de precisão e de cobertura das buscas realizadas.

Finalmente, o **Capítulo 7** sumariza o trabalho proposto e descreve algumas questões ainda em aberto. Além disso, faz uma objetiva identificação de propostas de trabalhos futuros e direções nas quais a pesquisa pode ser estendida, em consideração às tendências emergentes.

para o Firefox – *backend* em *PHP* – e usa o *MySQL* como repositório central. Tem como objetivo facilitar o gerenciamento do processo de construção de *datasets,* além de armazenar e disponibilizar conjuntos de páginas *Web* relevantes em relação a um determinado contexto.

# Capítulo 2 – Principais Tópicos de Pesquisa em Qualidade de Informação

Devido a sua importância, a sua natureza e a variedade de tipos de dados e de aplicações, a qualidade de informação tornou-se uma área de investigação complexa e multidisciplinar (BATINI & SCANNAPIECO, 2006). A Figura 2-1 mostra alguns aspectos referentes aos modelos, às técnicas e às ferramentas em suas dimensões e metodologias, bem como os domínios de aplicação envolvidos.

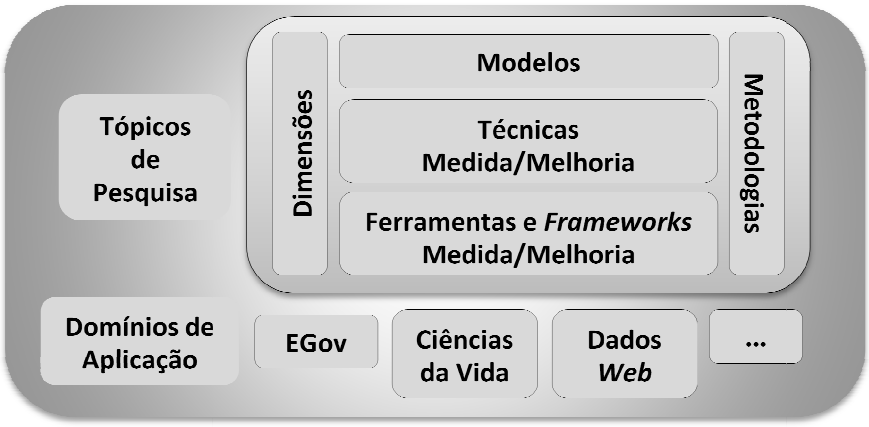



**Figura 2-1: Tópicos de Pesquisa em Qualidade de Informação (Fonte: Batini & Scannapieco, 2006)**

Neste capítulo são apresentados os principais tópicos de pesquisa em qualidade de informação, importantes aos desdobramentos da tese. As seções a seguir abordam: a qualidade de informação, a qualidade de informação na *Web*, os metadados, as dimensões de qualidade, o contexto, os modelos e as metodologias de qualidade.

## 2.1 – Qualidade de Informação

Apesar de as pessoas intuitivamente saberem o que significa o termo "qualidade", a solicitação de uma definição explícita implicará debate. Esse é um dos principais problemas encontrados quando se discute qualidade: todos sabem o que é, mas poucos conseguem definir. A partir daí, há uma grande dificuldade quando se torna essencial o tratamento da qualidade em um sistema computacional, mediante a representação quantitativa de termos (BURGESS, GRAY *et al.*, 2004).

A norma ISO 9000:2005 define qualidade como: "A totalidade de características de uma entidade que lhe confere a capacidade de satisfazer às necessidades explícitas e implícitas". Além disso, descreve as necessidades explícitas como aquelas expressas na definição dos requisitos propostos pelo produtor de dados. Elas são compostas pelas condições de utilização do produto, seus objetivos, funções e desempenho esperado. As necessidades implícitas são aquelas que, embora não expressas nos documentos do produtor de dados, ainda assim são importantes para os usuários (ISO 9000:2005, 2005).

O termo "Qualidade de Informação" descreve a qualidade da informação quanto ao seu conteúdo. Muitas vezes, é definido como – a condição de uso ou uma medida do valor da informação fornecida aos seus usuários – (STRONG, LEE *et al.*, 1997) (CAPPIELLO, FRANCALANCI *et al.*, 2004). Essa definição sugere a relatividade do conceito porque os dados considerados apropriados para um uso específico podem não ser para outro. Ela também sugere que a qualidade não pode ser avaliada independentemente das pessoas que usam os dados (STRONG & WANG, 1996). Essa perspectiva será tratada em seguida.

A maioria dos profissionais de sistemas de informação emprega o termo como sinônimo de qualidade de dados. No entanto, como muitos acadêmicos fazem a distinção entre dados e informações, alguns irão insistir em uma distinção entre a qualidade de dados e a qualidade de informações, o que não é o caso deste trabalho, como já foi observado.

A garantia de qualidade das informações é a confiança de que determinada informação cumpre alguns requisitos contextuais específicos de qualidade. Embora essas definições sejam utilizadas de forma mais corrente, alguns especialistas, muitas vezes, utilizam modelos mais complexos de qualidade de informação para tratá-las.

Numa abordagem mais tradicional, a qualidade de dados é definida como um conceito multidimensional e complexo. É multidimensional porque os usuários lidam com as percepções subjetivas dos indivíduos envolvidos com os dados, bem como com as medidas objetivas baseadas nos conjuntos de dados que estão sendo avaliados. É complexo porque possui significados diversos para diferentes pessoas (WAND & WANG, 1996).

A avaliação subjetiva (ou qualitativa) da qualidade reflete as necessidades e experiências dos usuários, sejam eles produtores ou consumidores dos dados. Nesse caso são utilizados indicadores subjetivos para julgar a qualidade dos dados e sua aptidão para o uso. A avaliação objetiva (ou quantitativa) da qualidade utiliza indicadores objetivos para medir a qualidade dos dados que podem ser dependentes ou independentes da tarefa. Os indicadores independentes da tarefa refletem o estado dos dados sem o conhecimento contextual da aplicação e podem ser aplicados a qualquer conjunto de dados. Já os indicadores dependentes da tarefa dependem de regras de negócio ou restrições existentes e são desenvolvidos para contextos de aplicações específicas (WANG, REDDY *et al.*, 1995).

A qualidade dos dados abrange os diferentes tipos e classificações de dados, incluindo dados estruturados e semi-estruturados, documentos textuais, multimídia e cadeias de dados, bem como os dados *Web* investigados por Dasu & Johnson (2003), que foram mencionados anteriormente.

Batini & Scannapieco (2006) classificam os dados quanto à freqüência de modificação, em três categorias: dados estáveis, dados modificados em longo prazo e dados freqüentemente modificados:

- Dados estáveis são aqueles cuja modificação seja improvável como, por exemplo, publicações científicas que apesar da ocorrência das novas publicações, não há mudanças nas antigas publicações;
- Dados modificados em longo prazo são aqueles atualizados com baixa freqüência como, por exemplo, endereços, moedas e listas de preços de hotéis. O conceito de baixa freqüência é dependente do domínio. Se o preço das ações na bolsa de valores for atualizado a cada hora, essa modificação é considerada de baixa freqüência, ao passo que, se o preço das mercadorias de uma loja for atualizado semanalmente, essa modificação é considerada de alta freqüência pelos clientes;
- Dados freqüentemente modificados são aqueles de intensa modificação como, por exemplo, o tráfego de informações em tempo real, sensores de medidas de temperaturas e quantidade de vendas. Nessa categoria, a freqüência de modificações pode ocorrer de forma definida ou aleatória.

Dentre outras formas adotadas para descrever a qualidade dos dados, há ainda os metadados, tratados mais adiante. Eles estão se tornando cada vez mais importantes, visto que podem proporcionar às aplicações e aos usuários informações sobre o valor e a confiabilidade dos dados (integrados) na *Web*.

Gerar, portanto, uma medida de qualidade aceitável para todos, não é trivial. Para isso é preciso definir as características de qualidade que interessam e, então, decidir como a avaliação da qualidade será feita por intermédio delas (KITCHENHAM, 1996).

Muitos pesquisadores têm despertado o seu interesse pela qualidade de informação em um grande número de disciplinas, abrangendo a ciência da computação, a biblioteconomia, a ciência da informação e os sistemas de gerência de informação. Naturalmente, existe um maior foco comercial nessa última área, com ênfase nos custos e no impacto para as organizações em conseqüência da baixa qualidade de dados (REDMAN, 1998). Esses impactos podem influenciar diretamente alguns paradigmas de diferencial competitivo na grande maioria dos empreendimentos (MORESI, 2000).

A Tabela 2-1 mostra algumas iniciativas que têm sido empreendidas na tentativa de organizar e avaliar a qualidade das informações disponíveis aos usuários.

**Tabela 2-1: Iniciativas de Avaliação de Qualidade e suas Abordagens**

| Iniciativas | Abordagens |
|---|---|
| (STRONG & WANG, 1996) | Apresentam um quadro que captura os aspectos da qualidade dos dados que são importantes para os Consumidores de Dados. |
| (WAND & WANG, 1996) | Sugerem uma definição rigorosa das dimensões de qualidade dos dados, fundamentando-as em bases ontológicas, e mostra como essas dimensões podem fornecer orientações aos desenvolvedores de sistemas nas questões de qualidade dos dados. |
| (REDMAN, 1998) | Categoriza todas as questões relacionadas com o impacto da baixa qualidade dos dados em três níveis organizacionais: operacional, tático e estratégico |
| (ENGLISH, 1999) | Propõe um método para reduzir os custos e aumentar os lucros, melhorando o *Data warehouse* e a qualidade das informações comerciais. |
| (TWIDALE & MARTY, 1999) | Delineiam uma nova abordagem colaborativa de gestão de qualidade dados |
| (LOSHIN, 2001) | Analisa a qualidade dos dados sob o ponto de vista da |

| Iniciativas | Abordagens |
|---|---|
| | gestão do conhecimento. Ele define a qualidade dos dados como "adequação ao uso" e destaca que a avaliação da qualidade é dependente do contexto do usuário. |
| (PIPINO, LEE *et al.*, 2002) | Descrevem as avaliações subjetivas e objetivas de qualidade dos dados e apresentam três formas funcionais para o desenvolvimento de métricas objetivas de qualidade dos dados. |
| (LEE, 2004) | Assume que profissionais experientes resolvem problemas de qualidade dos dados analisando e explicitando seus conhecimentos sobre os contextos dos dados que foram incorporados ou perdidos. |
| (BURGESS, GRAY *et al.*, 2004) | Propõem um modelo hierárquico genérico de qualidade que pode ser utilizado pelo consumidor de informações para ajudar na busca da informação, focando o resultado retornado nas preferências pessoais de qualidade definidas. |
| (KIM, KISHORE *et al.*, 2005) | Aplicam os conceitos de qualidade de dados no âmbito de sistemas de comércio eletrônico. |
| (AKOKA, BERTI-´EQUILLE *et al.*, 2007) | Abordam o problema no contexto dos sistemas de integração de dados, utilizando modelos de gráficos de custo que permitem a definição de métodos de avaliação e demonstração de proposições em termos de propriedades gráficas. |
| (BATISTA & SALGADO, 2007) | Apresentam uma abordagem para análise de qualidade da informação de esquemas em ambientes de integração de dados. |
| (BATINI, BARONE *et al.*, 2008) | Provêem uma tentativa inicial de unificação dos aspectos de qualidade de informação para os tipos heterogêneos de informação. Eles apresentam uma categorização geral das dimensões e das subdimensões de qualidade, que por sua vez são especializadas para os dados estruturados, textos semi e não estruturados e imagens. |

## 2.2 – Qualidade de Informação na *Web*

As pesquisas sobre a qualidade de dados tiveram início no contexto dos sistemas de informação (STRONG, LEE *et al.*, 1997) (LEE, STRONG *et al.*, 2002) e foram estendidas para os demais contextos como: sistemas cooperativos, data warehouses, comércio eletrônico e outros. Em virtude das características próprias das aplicações *Web* e as suas diferenças dos sistemas de informação tradicionais, a comunidade de pesquisa começou a investigar a qualidade de informação para esse contexto específico (GERTZ,

OZSU *et al.*, 2004) (KNIGHT & BURN, 2005). Atualmente, ela tornou-se um fator crítico de sucesso para as aplicações na *Web* (SCHAUPP, FAN *et al.*, 2006).

Na verdade, a natureza particular da Internet forçou a atenção a uma série de questões, que podem afetar ou influenciar a qualidade de informações nesse contexto. Uma delas refere-se ao fato de que os critérios de avaliação de qualidade propostos na literatura, também precisam ser adaptados e estendidos para representar as características das informações na *Web*, em específico, seus aspectos dinâmicos (TILLMAN, 2003). A Tabela 2-2 sumariza algumas dessas questões.

**Tabela 2-2: Questões Específicas de Qualidade de Informações na *Web* (Adaptada de Caro & Calero, 2008)**

| Questões | Descrição | Autores |
|---|---|---|
| **QD sob a perspectiva do usuário** | Implica que a QD (qualidade de dados) não pode ser avaliada de forma independente das pessoas que usam os dados. | (STRONG, LEE *et al.*, 1997) (CAPPIELLO, FRANCALANCI *et al.*, 2004) (GERTZ, OZSU *et al.*, 2004) (KNIGHT & BURN, 2005) |
| **Demanda por serviços em tempo real** | Aplicações *Web* interagem com diferentes fontes de dados externas cuja carga de trabalho não é conhecida. Essa situação pode influenciar drasticamente o tempo de resposta por vezes afetando aspectos de QD como a oportunidade ou a atualidade. | (AMIRIJOO, HANSSON *et al.*, 2003) |
| **Desenvolvimento de comércio eletrônico** | QD é essencial para alcançar o desenvolvimento do comércio eletrônico na *Web*, bem como para angariar a confiança da clientela. | (LIM & CHIN, 2000) (DAVYDOV, 2001) (HAIDER & KORONIOS, 2003) |
| **Dinamismo na *Web*** | O dinamismo com que ocorrem as mudanças dos dados, das aplicações e das fontes pode afetar a QD. | (PERNICI & SCANNAPIECO, 2002) (GERTZ, OZSU *et al.*, 2004) |
| **Integração de Dados estruturados e não-estruturados** | O uso de dados não-estruturados (e-mails, documentos de trabalho, manuais, etc.) e sua integração com dados estruturados é um importante desafio, porque ambos contêm conhecimentos de grande valor para as | (FINKELSTEIN & AIKEN, 1999) |

| Questões | Descrição | Autores |
|---|---|---|
| | organizações. | |
| **Integração de dados de diferentes fontes** | O acesso às diversas fontes de dados que provavelmente não têm o mesmo nível de QD pode prejudicar a QD do produto durante a integração que é realizada para os usuários. | (NAUMANN & ROLKER, 2000) (ZHU & BUCHMANN, 2002) (ANGELES & MACKINNON, 2004) (GERTZ, OZSU *et al.*, 2004) (WINKLER, 2004) |
| **Necessidade de compreender os dados e a sua qualidade** | Uma linguagem comum que facilite a comunicação entre pessoas, sistemas e programas é essencial e deve ser capaz de avaliar a QD. É necessário compreender os dados e os critérios utilizados para determinar a sua qualidade. | (FINKELSTEIN & AIKEN, 1999) (GERTZ, OZSU *et al.*, 2004) |
| **Problemas típicos de uma página *Web*** | Informação desatualizada, a publicação de informações inconsistentes, links obsoletos e outros. | (EPPLER & MUENZENMAYER, 2002) |
| **Fidelidade dos usuários** | Isso implica uma adequada gestão dos dados por usuário ou dos tipos de usuário; e uma permanente elaboração de dados de saída que mantenham o interesse e a fidelidade dos usuários. | (DAVYDOV, 2001) |

Como ficou anteriormente caracterizado, as informações na *Web* apresentam-se em grande volume para um extenso número de usuários e possuem qualidade heterogênea. Existem, nesse caso, algumas razões para essa variedade. Primeiramente, qualquer pessoa ou organização pode criar um sítio na *Web* e carregá-lo com todo tipo de informação, sem qualquer controle de qualidade, e algumas vezes com más intenções. A segunda razão refere-se a um conflito entre duas necessidades. Por um lado, os sistemas *Web* têm que publicar as informações no menor período de tempo possível, após a liberação pelas suas fontes. Por outro lado, as informações precisam ser verificadas quanto a sua acurácia, a sua atualidade e quanto à confiabilidade dessas mesmas fontes.

Esses dois requisitos são contraditórios em muitos aspectos. Pode-se tomar como exemplos o custo e o tempo necessários ao desenvolvimento de projetos mais precisos

referentes às estruturas de dados, aos melhores caminhos de navegação entre as páginas e à verificação dos dados para certificação da sua correção, em contrapartida, a premência de tempo para publicação das informações.

Existem ainda dois aspectos adicionais relacionados à qualidade de informações, diferentes dos que ocorrem tradicionalmente. O primeiro aspecto é que um sítio *Web* é uma fonte de informação abrangente e contínua e não está ligado a versões temporalmente fixadas das informações. Em segundo lugar, durante a atualização, algumas informações adicionais podem ser produzidas, possibilitando até as correções das que foram previamente publicadas. Isso cria de certa forma, a necessidade de verificações adicionais de qualidade.

Como argumento final, nos sistemas *Web* é praticamente impossível individualizar o "dono do dado", ou seja, o responsável ou responsáveis por determinadas categorias. Tipicamente, os dados são replicados dentro das organizações participantes, sendo difícil dizer qual delas é mais ou menos responsável por alguma informação específica. Em especial, quando observado o conceito de interatividade da *Web* 2.0, o consumidor também é um produtor de dados (ex: Wikipedia[8], comentários, *blogs*, etc.).

Todos esses aspectos tornam difíceis a certificação da qualidade das fontes e a avaliação da qualidade de informações pelos usuários na *Web* (BATINI & SCANNAPIECO, 2006).

Apesar da diversidade de propostas para avaliação de qualidade de dados, em sua maioria, elas enfatizam a importância de uma definição para qualidade. Entretanto, os esforços evidenciados nessas pesquisas, ainda não denotaram um padrão ou definição unificada (SMART, 2002).

## 2.3 – Perspectiva do Consumidor de Informação

Atualmente, a maior parte do trabalho de investigação realizado na área de qualidade refere-se à qualidade sob as perspectivas organizacionais ou dos produtores de informações. A Tabela 2-1 exemplifica algumas dessas iniciativas.

[8] http://en.wikipedia.org/wiki/Web_2.0.

Burgess & Gray (2004) mostram que a perspectiva do consumidor de informação distingue-se dessas outras por duas importantes razões:

*i.* O consumidor não tem qualquer controle sobre a qualidade das informações disponíveis;

*ii.* O objetivo do consumidor é encontrar a informação que corresponda às suas necessidades pessoais, em vez de prover informações que atendam às necessidades dos outros.

Como resultado dessas diferenças de foco, as definições de qualidade que foram previstas pelos produtores de informação não são adequadas aos seus consumidores. O consumidor de informação típico quer encontrar a melhor informação disponível que atenda às suas necessidades no momento da busca e dentro do seu domínio de interesse (LIU, GAO *et al.*, 2008). Esses resultados podem não ser necessariamente os melhores em vista das possíveis restrições do consumidor de dados, como aquelas relativas ao tempo disponível para a BRI. O consumidor pode precisar da informação imediatamente e não pode esperar várias horas, enquanto todas as possíveis fontes de informação sejam investigadas, para encontrar o melhor resultado entre elas.

Nesses casos, os consumidores estarão dispostos a aceitar os melhores resultados obtidos consideradas certas restrições, tais como: os dados atualmente disponíveis, os dados dentro de uma faixa de preços ou todos os dados que podem ser obtidos dentro de um determinado intervalo de tempo (BURGESS, GRAY *et al.*, 2004).

Como foi previamente mencionado, muitas definições para a qualidade de informação têm sido discutidas. No caso da *Web*, Mandl (2008) ressalta que alguns pesquisadores ainda sustentam que nenhuma noção objetiva de qualidade de informação foi encontrada. Uma das razões apontadas é que, por um lado, o conteúdo e a interface são componentes de difícil separação e conseqüentemente, nem sempre suas avaliações são facilmente distinguidas quando as páginas são avaliadas. Por outro lado, a qualidade e a relevância podem ser tratadas de forma separada. A relevância descreve o valor situacional de uma página em uma pesquisa específica, enquanto alguns aspectos das páginas descrevem a sua qualidade independentemente de qualquer necessidade corrente de informação. Conseqüentemente, o usuário deve ser capaz de julgar a qualidade sem considerar quaisquer informações concretas e seus parâmetros pragmáticos (AMENTO, TERVEEN *et al.*, 2000) (MANDL, 2008).

Amento & Terveen (2000) tratam a qualidade e a relevância como noções distintas, em vez de considerar a visão de qualidade como – apenas mais um aspecto dos julgamentos de relevância –. Para salientar essa distinção eles recorrem a um exemplo ilustrativo envolvendo um artigo escrito por um estudante e uma coleção de críticas literária sobre os sonetos Shakespeare. Ambos podem ser julgados igualmente relevantes, mas a coleção será julgada de muito maior qualidade.

Ainda não está bem entendido como os seres humanos avaliam a qualidade global das páginas *Web*. No entanto, as experiências mostram que o *layout* e o projeto das páginas são aspectos muito importantes para a avaliação humana da qualidade. Em diversos experimentos, as páginas bem concebidas foram preferidas pelos usuários em detrimento de outras páginas que possuíam conteúdos comparativamente bons (DHAMIJA, TYGAR *et al.*, 2006). A atribuição de qualidade das páginas *Web* não parece ser uma constante universal. A interface e o conteúdo – características consideradas importantes para os seres humanos nas suas decisões sobre qualidade – são culturalmente dependentes (MANDL, 2008).

Muitas listas de critérios de qualidade de páginas *Web* foram desenvolvidas sob a perspectiva da ciência da informação, a fim de apoiar o usuário nos processos de avaliação. No entanto, sob a perspectiva da avaliação automática de qualidade, essas listas são de pouco auxílio. Por vezes, seus critérios são vagos e não fica muito claro se uma regra indica alta ou baixa qualidade.

Mandl & de la Cruz (2007) demonstraram que algumas listas de critérios de qualidade de alguns países podem conter, de forma parcial, diferentes critérios e diferentes atribuições de importância a critérios idênticos, ou seja, as orientações para a avaliação dos *sites* diferem de cultura para cultura. Em sua pesquisa há um levantamento realizado por mais de 300 usuários da Internet, no Peru e na Alemanha. Nesse levantamento os critérios típicos constantes de várias listas foram classificados de formas substancialmente diferentes. Sobretudo, ficou evidenciado que não existe na Internet nenhuma cultura global e que culturas locais ainda dominam o comportamento dos usuários da *Web*.

Uma avaliação apropriada precisa investigar a qualidade média de uma lista de resultados obtidos mediante diferentes abordagens. Nesse sentido os usuários de teste precisam julgar as páginas resultantes de acordo com suas percepções subjetivas de

qualidade (MANDL, 2006). Em resumo, o usuário ou usuários são os árbitros finais da qualidade (REDMAN, 1996).

## 2.4 – Dimensões de Qualidade de Informação

A norma ISO 9126-4 (2004) fornece o significado de característica de qualidade, como "a referência básica à qualidade de um produto de software[9], utilizada em uma avaliação". Essa mesma norma fornece um modelo de propósito geral, que define seis amplas categorias de características de qualidade de software (ISO 9126-1, 2001) (ISO 9126-4, 2004).

A primeira etapa na avaliação da qualidade é a seleção das características aplicáveis, com base em um modelo de qualidade que as represente (BATINI & SCANNAPIECO, 2006). O conjunto de características mais adequado depende da aplicação do usuário, da seleção das métricas e da implementação de algoritmos de medida ou estimativa de avaliação de cada dimensão de qualidade (PERALTA, RUGGIA *et al.*, 2004). A escolha dessas dimensões é primariamente baseada no entendimento intuitivo, na experiência da indústria ou na revisão da literatura (WAND & WANG, 1996).

Tillman (2003) enfatiza que a atual condição da Internet deve ser considerada na adoção de critérios genéricos para a avaliação de qualidade da informação. Esse entendimento é bastante importante na determinação do melhor conjunto de dimensões da qualidade em razão do constante estado de mudanças da *Web* (TILLMAN, 2003).

As dimensões de qualidade são aplicadas de diferentes maneiras em modelos, técnicas, ferramentas e arquiteturas. Apesar de as medidas de qualidade em TIC, artefatos, processos e serviços não serem novos tópicos de pesquisa, por muitos anos algumas instituições de padronização têm trabalhado a fim de estabelecer a maturidade de conceitos relacionados às características de qualidade, indicadores e procedimentos de medida confiáveis.

A Tabela 2-3 sumariza um conjunto de pesquisas realizadas no mercado e na área acadêmica[10]. Essa Tabela foi agregada e adaptada de (BURGESS, GRAY *et al.*,

[9] Os requisitos de qualidade para programas e dados utilizam as mesmas definições das características de qualidade da norma ISO 9126.

[10] Evidentemente essa lista não é exaustiva.

2004) e (CARO, CALERO *et al.*, 2008). Ela mostra uma variedade de abordagens que usam diferentes terminologias como dimensões, critérios, métricas e fatores denotando que a qualidade é uma entidade de muitos atributos. Há, porém, um consenso que esses múltiplos atributos usados para definir a qualidade podem ser agrupados em categorias relacionadas, representando uma estrutura hierárquica.

Essas abordagens são capazes de representar as expectativas de qualidade dos usuários, considerando uma base de dados como o produto a ser avaliado (STRONG & WANG, 1996) (PIPINO, LEE *et al.*, 2002). Dentre elas estão alguns autores já citados neste capítulo (ISO 9126-1, 2001) (REDMAN, 1996) (GERTZ, OZSU *et al.*, 2004) (BOUZEGHOUB & PERALTA, 2004) e outros que também desenvolveram pesquisas sobre a definição de critérios de qualidade de informações na *Web* (ALADWANI & PALVIA, 2002), (CHEN, ZHU *et al.*, 1998), (OLSINA, LAFUENTE *et al.*, 2001) e (ZHU & GAUCH, 2000).

Foram acrescentadas ao sumário criado as abordagens propostas por (ROCHA, 1983), (COMERLATO, XEXEO *et al.*, 1994) (BELCHIOR, 1997), (CARVALHO, 1997), (ROTHENBERG, 1996), (WAND & WANG, 1996), (REDMAN, 1998), (MALETIC & MARCUS, 2000), (AMENTO, TERVEEN *et al.*, 2000), (PINHO, 2001), (ECKERSON, 2002), (PIPINO, LEE *et al.*, 2002), (BURGESS, GRAY *et al.*, 2003), (PIERCE, 2004), (KIM, KISHORE *et al.*, 2005), (MANDL, 2006), (MADNICK & ZHU, 2006), (STVILIA, GASSER *et al.*, 2006), (AKOKA, BERTI-´EQUILLE *et al.*, 2007), (CARO, CALERO *et al.*, 2008), (BATINI, BARONE *et al.*, 2008) e (BARROS, XEXÉO *et al.*, 2008a).

Em seguida a Tabela 2-4 apresenta uma revisão da literatura pertinente que tem por objetivo identificar os atributos de qualidade propostos para diferentes domínios no contexto da *Web* (CARO, CALERO *et al.*, 2008):

- *Sites Web*: (KATERATTANAKUL & SIAU, 2001), (EPPLER, 2001), (MOUSTAKIS, LITOS *et al.*, 2004);
- Integração de Dados: (NAUMANN & ROLKER, 2000), (BOUZEGHOUB & PERALTA, 2004);
- Comércio Eletrônico: (KATERATTANAKUL & SIAU, 2001);
- Portais de Informação *Web*: (YANG, CAI *et al.*, 2004);
- Serviço-Eletrônico Cooperativo: (FUGINI, MECELLA *et al.*, 2002);

- Tomada de Decisão: (GRAEFE, 2003);
- Redes Organizacionais: (MELKAS, 2004);
- Qualidade de Dados na *Web*: (GERTZ, OZSU *et al.*, 2004);
- Sistemas de Informação *Web* (evolução de dados): (PERNICI & SCANNAPIECO, 2002).

Por fim, o Anexo I descreve os atributos da Tabela 2-4 e estende esse conjunto com outras descrições de dimensões de qualidade mais comumente adotadas nas pesquisas referenciadas (NAUMANN, FREYTAG *et al.*, 2003).

A despeito da freqüência de uso de certos termos para indicar qualidade de dados, ainda não existe um conjunto de dimensões de qualidade de dados rigorosamente definido, nem um consenso quanto ao seu emprego. Observam-se ambigüidades de definições, mesmo para dimensões relativamente óbvias como "acurácia" (WAND & WANG, 1996).

**Tabela 2-3: Domínios e Estruturas das Abordagens de Avaliação de Qualidade de Informações e Dados (Adaptada e estendida de Burgess & Gray, 2004)**

| Fonte | Domínio | Estrutura da Abordagem |
|---|---|---|
| **Definições de Qualidade de Software** | | |
| (BARBACCI, KLEIN *et al.*, 1995) | Qualidade de Software | 4 modelos para cada um dos 4 atributos primários, com um total de 13 tópicos de interesse |
| (BOEHM, BROWN *et al.*, 1976) | Qualidade de Software | Estrutura de Árvore hierárquica composta por 10 categorias e 15 métricas |
| (DROMEY, 1995) | Qualidade de Software | 3 modelos, contendo 17 atributos e 42 subatributos únicos (repetida entre os modelos) |
| (HYATT & ROSENBERG, 1996) | Qualidade de Software | 4 metas e 13 atributos |
| (ISO 9126-1, 2001) | Qualidade de Software | 2 modelos: 1) qualidades internas e externas de software - 6 dimensões e 34 métricas 2) Qualidade em uso - 4 métricas |
| (LIU, ZHOU *et al.*, 2000) | Projeto de Software OO | 3 fatores e 8 critérios |
| (MCCALL, RICHARDS *et al.*, 1977) | Qualidade de Software | 3 classes, 11 fatores e 23 critérios |
| (ORTEGA, PÉREZ *et al.*, 2002) | Qualidade de Software | 6 métricas |
| (ROYCE, 1990) | Produtos de software | 4 métricas |
| (RUBEY & HARTWICK, 1968) | Qualidade de Software | Descrições de 7 atributos |

| Fonte | Domínio | Estrutura da Abordagem |
|---|---|---|
| (ROCHA, 1983) | Qualidade de Software | Compatível com a norma (ISO 9126-1, 2001) |
| (BELCHIOR, 1997) | Qualidade de Software | Particularização do modelo (ROCHA, 1983) para sistemas financeiros |
| (COMERLATO, XEXEO *et al.*, 1994) | Qualidade de Software | Particularização do modelo (ROCHA, 1983) para software científico |
| (CARVALHO, 1997) | Qualidade de Software | Particularização do modelo (ROCHA, 1983) para sistemas de informação hospitalar e para o prontuário médico computadorizado |

**Definições de Qualidade de Dados**

| | | |
|---|---|---|
| (ABATE, DIEGERT *et al.*, 1998) | Qualidade de Dados | 4 categorias e 15 dimensões |
| (CYKANA, PAUL *et al.*, 1996) | Qualidade de Dados | 6 características |
| (GARDYN, 1997) | *Data warehouse* | 5 dimensões |
| (LONG & SEKO, 2002) | Qualidade de Dados Médicos | 5 dimensões e 24 características |
| (NAUMANN, 2002) | Planejamento de Consultas | 4 dimensões e 22 métricas |
| (REDMAN, 1996) | Qualidade de Dados | 3 categorias e 27 dimensões |
| (STRONG & WANG, 1996) | Qualidade de Dados | 4 categorias e 15 dimensões |

| Fonte | Domínio | Estrutura da Abordagem |
|---|---|---|
| (WAND & WANG, 1996) | Qualidade de Dados | 4 categorias e 16 dimensões |
| (PINHO, 2001) | Qualidade de Dados | Particularização do modelo (ROCHA, 1983) para avaliação da qualidade de dados pela não conformidade |
| (REDMAN, 1998) | Qualidade de Dados | 1 dimensão sobre a qualidade dos dados em três níveis organizacionais: operacional, tático e estratégico |
| (PIPINO, LEE *et al.*, 2002) | Qualidade de Dados | 16 dimensões |
| (MALETIC & MARCUS, 2000) | Qualidade de Dados | 1 categoria e 4 dimensões |
| (MADNICK & ZHU, 2006) | Qualidade de Dados | 2 grupos semânticos relacionados a qualidade de dados |
| (ROTHENBERG, 1996) | Qualidade de Dados | 1 categoria e 2 dimensões |
| (ECKERSON, 2002) | Qualidade de Dados | 7 critérios |

**Definições de Qualidade de Informação**

| | | |
|---|---|---|
| (BOVEE, SRIVASTAVA *et al.*, 2001) | Qualidade de Informações | 4 critérios e 10 componentes |
| (DEDEKE, 2000) | Sistemas de Informação | 5 dimensões e 28 métricas |
| (EPPLER, 2001) | Qualidade de Informações | 4 níveis de qualidade e 16 critérios |
| (MATSUMURA & | Qualidade de Informações | 2 categorias e 4 atributos |

| Fonte | Domínio | Estrutura da Abordagem |
|---|---|---|
| SHOURABOURA, 1996) | | |
| (MILLER, 1996) | Qualidade de Informações | 10 dimensões |
| (BURGESS, GRAY *et al.*, 2003) | Qualidade de Informações | Estrutura genérica hierárquica composta por 3 categorias e 6 métricas |
| (PIERCE, 2004) | Qualidade de Informações | 4 dimensões |
| (BATINI, BARONE *et al.*, 2008) | Qualidade de Informações | 3 dimensões e 3 subdimensões |
| **Definições de qualidade na *Web*** | | |
| (ALADWANI & PALVIA, 2002) | Qualidade do *Site* | 4 dimensões e 25 itens |
| (CHEN, ZHU *et al.*, 1998) | Processamento de Consulta | 10 parâmetros de qualidade |
| (OLSINA, LAFUENTE *et al.*, 2001) | *Sites* Acadêmicos | Modelo hierárquico contando com +100 métricas |
| (ZHU & GAUCH, 2000) | Qualidade do *Site* | 15 métricas |
| (AMENTO, TERVEEN *et al.*, 2000) | Qualidade de Documentos *Web* | 4 características |
| (KIM, KISHORE *et al.*, 2005) | Qualidade de Dados no comércio eletrônico | 3 categorias e 9 dimensões |
| (MANDL, 2006) | Qualidade de Páginas *Web* | 7 categorias e 113 características |

| Fonte | Domínio | Estrutura da Abordagem |
|---|---|---|
| (CARO, CALERO *et al.*, 2008) | Qualidade de Portais *Web* | 34 atributos de qualidade |
| (BARROS, XEXÉO *et al.*, 2008a) | Qualidade de Informações na *Web* | Modelo formal com 13 componentes e 3 dimensões de qualidade |
| (KATERATTANAKUL & SIAU, 1999) | *Web sites* Pessoais | 4 categorias e 7 construtores |
| (NAUMANN & ROLKER, 2000) | Integração de Dados | 3 classes e 22 critérios de qualidade |
| (ABOELMEGED, 2000) | Comércio Eletrônico | 7 etapas de modelagem de problemas de QD |
| (KATERATTANAKUL & SIAU, 2001) | Comércio Eletrônico | 4 categorias associadas a outras 3 categorias de requisitos de dados de usuários |
| (PERNICI & SCANNAPIECO, 2002) | Sistemas de Informação *Web* (evolução de dados) | 4 categorias e 7 atividades de projeto e arquitetura de QD para gestão de QD |
| (FUGINI, MECELLA *et al.*, 2002) | Serviço-Eletrônico Cooperativo | 8 dimensões |
| (GRAEFE, 2003) | Tomada de Decisão | 8 dimensões e 12 aspectos relacionados com (produtores / consumidores) |
| (EPPLER, ALGESHEIMER *et al.*, 2003) | *Web sites* | 4 dimensões e 16 atributos |
| (GERTZ, OZSU *et al.*, 2004) | Qualidade de Informações na | 5 dimensões |

| Fonte | Domínio | Estrutura da Abordagem |
|---|---|---|
| | *Web* | |
| (MOUSTAKIS, LITOS *et al.*, 2004) | *Web sites* | 5 categorias e 10 subcategorias |
| (MELKAS, 2004) | Redes Organizacionais | 6 etapas para análise de QD com várias dimensões associadas a cada um delas |
| (BOUZEGHOUB & PERALTA, 2004) | Integração de Dados | 2 fatores e 4 métricas |
| (YANG, CAI *et al.*, 2004) | Portais *Web* de informação | 2 dimensões dentro do modelo global |
| (STVILIA, GASSER *et al.*, 2006) | Qualidade de Informações Dublin Core e Wikipedia | 22 dimensões |
| (AKOKA, BERTI-´EQUILLE *et al.*, 2007) | Integração de Dados | 2 fatores e 4 métricas |

**Tabela 2-4: Atributos de Qualidade Propostos para Diferentes Domínios no Contexto da *Web* (Fonte: Caro & Calero, 2008)**

| Dimensões / Autores | (NAUMANN & ROLKER, 2000) | (KATERATTANAKUL & SIAU, 2001) | (EPPLER, 2001) | (FUGINI, MECELLA *et al.*, 2002) | (PERNICI & SCANNAPIECO, 2002) | (GRAEFE, 2003) | (BOUZEGHOUB & PERALTA, 2004) | (GERTZ, OZSU *et al.*, 2004) | (MELKAS, 2004) | (MOUSTAKIS, LITOS *et al.*, 2004) | (YANG, CAI *et al.*, 2004) | Número de Referências |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Acessibilidade** | | x | x | | | x | | | x | | | 4 |
| **Acurácia** | x | x | x | x | x | | | | x | | x | 7 |
| **Volume de dados** | x | | | | | | | | x | | | 2 |
| **Aplicabilidade** | | | x | x | | | | | | x | | 3 |
| **Atratividade** | | x | | | | | | | | | | 1 |
| **Disponibilidade** | x | | | | | x | | | | | | 2 |
| **Credibilidade** | x | | | x | | x | | | x | x | x | 6 |
| **Completeza** | x | | x | x | x | | | x | x | | x | 7 |
| **Representação concisa** | x | | x | | | | | | x | | | 3 |
| **Representação Consistente** | x | | x | | | | | | x | | | 3 |
| **Custo Adequado** | | | | | | | | | x | | | 1 |
| **Suporte ao Usuário** | x | | | | | | | | | | | 1 |
| **Oportunidade** | | | x | | | | x | x | | | x | 4 |
| **Documentação** | x | | | | | | | | | | | 1 |
| **Duplicidade** | | | | | | | | x | | | | 1 |
| **Facilidade de Operação** | | | x | | | | | | x | | | 2 |
| **Expiração** | | | | | x | | | | | | | 1 |
| **Flexibilidade** | | | | | | | | | x | | | 1 |
| **Granuralidade** | | | | | | | | x | | | | 1 |
| **Interatividade** | | | x | | | | | | | | | 1 |
| **Consistência Interna** | | | x | x | | | | | | | | 2 |
| **Interpretabilidade** | x | | | | | x | | | x | | | 3 |
| **Latência** | x | | | | | | | | | | | 1 |
| **Manutenibilidade** | | | x | | | | | | | | | 1 |
| **Novidade** | | | | | | x | | | | | | 1 |
| **Objetividade** | x | | | | | | | | x | | | 2 |
| **Ontologia** | | | | | | | | x | | | | 1 |
| **Organização** | | x | | | | | | | | | | 1 |
| **Preço** | x | | | | | | | | | | | 1 |
| **Relevância** | x | | | x | | x | | | x | x | x | 6 |
| **Confiabilidade** | x | | | | x | | | | | | | 2 |

| Dimensões / Autores | (NAUMANN & ROLKER, 2000) | (KATERATTANAKUL & SIAU, 2001) | (EPPLER, 2001) | (FUGINI, MECELLA *et al.*, 2002) | (PERNICI & SCANNAPIECO, 2002) | (GRAEFE, 2003) | (BOUZEGHOUB & PERALTA, 2004) | (GERTZ, OZSU *et al.*, 2004) | (MELKAS, 2004) | (MOUSTAKIS, LITOS *et al.*, 2004) | (YANG, CAI *et al.*, 2004) | Número de Referências |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Reputação** | x | | | | | | | | x | | | 2 |
| **Tempo de Resposta** | x | | x | | | | | | x | | | 3 |
| **Segurança** | x | | x | x | | | | | x | | | 4 |
| **Especialização** | | | | | | | | | | x | | 1 |
| **Informações da Fonte** | | x | | | | | | | | | | 1 |
| **Atualidade** | x | | x | x | | | x | | x | | | 5 |
| **Rastreabilidade** | x | | x | | | | | | x | | | 3 |
| **Inteligibilidade** | x | x | x | | | | | | x | | | 4 |
| **Validade** | | | | | | x | | | | | | 1 |
| **Valor Agregado** | x | | | | | x | | | x | | | 3 |
| **Número de Atributos** | 22 | 6 | 16 | 8 | 4 | 8 | 2 | 5 | 20 | 4 | 5 | |

## 2.5 – Metadados de Qualidade de Informação

A definição mais comum de metadados (ou metainformação) vem de uma tradução literal: – metadados são dados sobre dados –. Metadados da *Web* – são as descrições inteligíveis de máquina de coisas (sobre) e da *Web*[11] –.

Existem algumas diferentes abordagens em que a avaliação da qualidade tem como base os metadados associados aos dados armazenados, requeridos para melhorar ou aperfeiçoar a qualidade de dados, de acordo com as dimensões mostradas na Tabela 2-4 da seção anterior (STRONG & WANG, 1996), (WAND & WANG, 1996), (PIPINO, LEE *et al.*, 2002), (ROTHENBERG, 1996), (TWIDALE & MARTY, 1999), (BOUZEGHOUB & PERALTA, 2004), (AKOKA, BERTI-´EQUILLE *et al.*, 2007), (MANDL & DE LA CRUZ, 2007) e (CARO, CALERO *et al.*, 2008).

[11] http://www.w3.org/Metadata/.

Rothenberg (1996) define um conjunto de metadados que é usado com esse propósito, voltado para o caso de dados estruturados. Nessa abordagem os metadados são organizados em categorias que são apresentadas em três diferentes níveis: banco de dados, elementos de dados e valores (instâncias) de dados. No seu entendimento, a qualidade de dados não é um atributo binário, mas está diretamente relacionada ao contexto e ao propósito de uso. Uma abordagem abrangente de qualidade de dados requer a evolução da qualidade dos valores de dados, realizando VV&C (Verificação, Validação e Certificação) e a evolução dos processos que geram e modificam o dado, visando o aumento da qualidade dos dados que eles produzem. A Tabela 2-5 apresenta as categorias de metadados propostas por ele. Essas categorias são consideradas como necessárias, ou para evolução, ou para registro da qualidade de dados (ROTHENBERG, 1996).

**Tabela 2-5: Níveis e Categorias de Metadados (Propostas por Rothenberg, 1996)**

| Nível dos Metadados | Categoria dos Metadados |
|---|---|
| **Banco de Dados** | **Geral:**<br>▪ Descrição e significado do banco de dados (BD);<br>▪ Uso ou variedade de propósitos e restrições do banco de dados;<br>▪ Requisitos para acesso e uso;<br>▪ Descrição e razão para estrutura ou projeto do banco de dados;<br>▪ Relacionamentos globais com outros bancos de dados;<br>▪ Informação do ciclo de atualização do BD. |
| | **Fonte de informação para o BD:**<br>▪ Fonte e credibilidade da fonte;<br>▪ Informação de classificação, acessibilidade, capacidade de reprodução;<br>▪ Autoridade de versão para o banco de dados. |
| | **Caracterização:**<br>▪ Granularidade (nível de detalhe) e qualidades (exatidão, completeza, etc.) pretendidas. |
| | **Informação sobre os elementos de dados:**<br>▪ Restrições e informação da dimensão de distribuição. |
| | **Qualidade da medida (total e para cada uso):**<br>▪ Exatidão, consistência, completeza, clareza, flexibilidade, robustez do |

| Nível dos Metadados | Categoria dos Metadados |
|---|---|
| | projeto de banco de dados, apropriação para o que se pretende usar. |
| | **Informação sobre o controle de processos** |
| **Elementos de Dados** | **Significado do elemento de dado, seus meta-valores e metadados:**<br>▪ Definição do que o elemento de dados representa; significados de nulos (valor desconhecido, aplicabilidade de atributo, valores especiais, etc.);<br>▪ Significados de exceções. |
| | **Origem e informação do ciclo de atualização para o elemento de dado:**<br>▪ Permissão de múltiplas origens com ciclos de atualização múltiplos, irregulares;<br>▪ "Modo de degradação" esperado;<br>▪ Classificação, acessibilidade, possibilidade de reprodução, autoridade de versão. |
| | **Informação de derivação / transformação:**<br>▪ Agregação ou outra informação de derivação;<br>▪ Informação do processo de transformação;<br>▪ Dados de controle do processo. |
| | **Restrições, relacionamentos com outros dados / BDs:**<br>▪ Completeza de entidade e atributo, etc. |
| | **Domínios, tipos de dado e unidades de medida:**<br>▪ Justificativas para portabilidade, flexibilidade, etc.;<br>▪ Uso específico de restrições para os domínios, incluindo as justificativas. |
| | **Resolução, precisão, exatidão pretendida e esperada:**<br>▪ Justificativas, representações de dependências e portabilidade. |
| | **Elemento de dado apropriado para o uso pretendido:**<br>▪ Significado, derivação, restrições, domínio, resolução, exatidão pretendida, etc. |
| | **Histórico de mudanças:**<br>▪ Auditoria da evolução das escolhas de domínio, tipos e unidades;<br>▪ Detalhes do momento e origens das modificações dos elementos de dados. |

| Nível dos Metadados | Categoria dos Metadados |
|---|---|
| | **Auditoria VV&C:**<br>▪ A respeito da adequação do elemento de dado, seu domínio, tipo, unidades, etc. |
| **Valores de dados** | **Qualidade (total e para cada uso):**<br>▪ Exatidão, consistência (resultados de validação), datas de expiração, "modos de degradação", adequação para o que se pretende usar, origens e qualidade do metadado. |
| | **Anotação:**<br>▪ Avisos, valores ou casos especiais, etc. |
| | **Informação da fonte:**<br>▪ Origem, derivação, tempo de geração e de entrada, etc. |
| | **Informação da próxima fonte:**<br>▪ Descrição de quando as alterações são esperadas e o que elas podem oferecer. |
| | **Informação de derivação e transformação:**<br>▪ Agregação ou outra informação de derivação;<br>▪ Informação do processo de transformação;<br>▪ Dados de controle do processo. |
| | **Auditoria de transformação:**<br>▪ Como esse valor tem sido transformado;<br>▪ Informação das transações de transformação em progresso;<br>▪ Detalhes do momento e origens das modificações dos elementos de dados. |
| | **Auditoria VV&C:**<br>▪ Finalidade da VV&C que tem sido feita nesse valor e "escopo" de validação e certificação. |

**Tabela 2-6: Fontes de Captura de Metadados na *Web***

| Google | PICS™ | Meta-Tags do Dublin Core |
|---|---|---|
| **Parâmetros de Consulta** | **Atributos de Serviço** | Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights, etc. |
| key, **q**: query terms, start, maxresults, filter, restricts, safesearch, **lr**: language restrict, **ie**: input encoding, **oe**: output encoding. | category, default, description, extension, icon, name, PICS-version, rating-service, and rating-system. | |
| **Operadores Especiais de Consulta** | **Atributos de Categorias** | |
| Special Query Capability, Include Query Term, Exclude Query Term, Phrase Search, Boolean OR Search, Site Restricted Search, Date Restricted Search Title Search (term), Title Search (all), URL Search (term), URL Search (all) Text Only Search (all), Links Only, Search (all), File Type Filtering, File Type Exclusion, *Web* Document Info, Back Links, Related Links, Cached Results Page. | description, extension, icon, integer, label, label-only, max, min, multivalue, name, transmit-as, and unordered. | |
| **Resultados Retornados** | | |
| <summary>, <URL>, <snippet>, <title> , <cachedSize>, <directoryTitle>, <hostName>, <relatedInformationPresent>, <directoryCategory>. | | |
| **Categorias de Diretórios** | | |
| <fullViewableName>, <specialEncoding> | | |

No caso da captura de metadados da Web existem algumas alternativas a serem consideradas, por exemplo, as APIs das máquinas de busca12, serviços de terceiros, o

[12] http://www.google.com/apis/.

emprego de protocolos como o PICS W3C13 e também os metadados fornecidos pelas fontes originais de dados que adotam padrões como o Dublin Core14. A Tabela 2-6 exemplifica os metadados capturados por cada uma dessas fontes.

Os *links* (ligações) existentes entre os hipertextos, ou páginas, são importantes metadados que permitem que a *Web* seja modelada por meio de grafos. A análise de *links* explora essas características estruturais para analisar a qualidade das páginas. A suposição básica é que o número de *links* que apontam para uma página é a medida para a popularidade e conseqüentemente para a qualidade dessa página. Esses *links* são denominados *in-links* ou *back-links* (MANDL, 2008).

Recentemente, os algoritmos de análise de *links* têm recebido bastante atenção, na maioria das vezes devido ao seu potencial para auxiliar nos problemas de avaliação de qualidade. A visão básica é a de que o *link* de um documento *A* para um documento *B* indica que o autor do documento *A* supõe que o documento B possui informações relevantes. Até então, existe uma pequena evidência empírica que os escores obtidos e as avaliações realizadas por esses algoritmos podem estar correlacionadas com o julgamento humano de qualidade (AMENTO, TERVEEN *et al.*, 2000).

O maior e mais utilizado motor de busca (*search engine*) atual – o Google® – é o precursor da utilização da análise de ligações entre páginas (nós em um grafo) da *Web* como um importante fator na definição da ordem na qual as páginas recuperadas são apresentadas para o usuário. Para tanto, o Google® utiliza o conceito denominado *PageRank* (PAGE, BRIN *et al.*, 1998) (BRIN & PAGE, 1998), implementado eficientemente (HAVELIWALA, 1999) através da análise de *links* entre páginas. O *PageRank* utiliza o conceito de *authority* (autoridade), ou seja, uma página que recebe *links* em um grafo, assim como o algoritmo denominado *HITS - Hyperlink Induced Topic Search*, que adicionalmente usa o conceito de *hub* (centralidade), que é uma página que emite *links*. O conceito de *authority*, por sua vez, é análogo ao conceito de *rank* ou *status* (prestígio) (KLEINBERG, 1998).

---

[13] A especificação PICS™ – *Platform for Internet Content Selection* permite que classificações (metadados) sejam associadas com o conteúdo da Internet. Ele foi originalmente concebido para ajudar os pais e professores a controlar o que as crianças acessam na Internet, e também facilitar outras utilizações, incluindo a definição de códigos e tratamentos de privacidade. O PICS é uma plataforma sobre a qual outros serviços de classificação e softwares de filtragem têm sido construídos. http://www.w3.org/PICS/# RDF).

[14] http://dublincore.org/.

Kleinberg (1998) observou que existe certo tipo natural de equilíbrio entre os *hubs* e os *authorities* no gráfico definido pela estrutura de rede de um ambiente de *links* entre páginas *Web*. Ele explorou essa característica para desenvolver um conjunto de algoritmos que identifica ambos os tipos de páginas simultaneamente. Ele também mostra em sua proposta que as melhores *authorities* serão as que apontam para os melhores *hubs*, e os melhores *hubs* serão os que apontam para as melhores *authorities*. Esse cálculo é repetido várias vezes. Em cada uma das repetições, o programa aumenta o peso do *authority* para os *sites* com *links* para os *sites* com mais peso de *hub* e aumenta o peso do *hub* para os *sites* com *links* para os *sites* com mais peso de *authority*. Por fim, ele sustenta que dez repetições são suficientes para convergir listas dos *authorities* e dos *hubs*. Na prática, apenas de 10 a 20 iterações são necessárias para que a convergência seja obtida.

O HITS opera nesses subgráficos da *Web* que são construídos a partir das saídas de um mecanismo de busca *Web* baseado em texto, como Google®, Yahoo®, Live Search® e Alta Vista®. A partir daí, o texto é ignorado e o aplicativo procura apenas a maneira como as páginas no conjunto expandido estão ligadas entre si.

## 2.6 – Contextos

Quanto mais pervasiva se torna a *Web*, cada vez mais ela representa todas as áreas da sociedade. A informação na *Web* tem a autoria de diferentes povos e é organizada por eles, cada um com distintos perfis, conhecimentos e expectativas. Ao contrário dos bancos de dados usados em sistemas tradicionais de recuperação de informação, a *Web* é muito mais diversificada em termos de conteúdo e de estrutura (LAWRENCE, 2000). Nesse sentido, o contexto é de grande utilidade e comumente usado para especificar o escopo ou os limites de uma área de estudo ou de uma discussão, ou mesmo um domínio de conhecimento.

Em Dey (2001), contexto é definido como “qualquer informação que pode ser usada para caracterizar a situação de uma entidade”. Uma entidade é uma pessoa, um lugar ou um objeto considerado relevante para a interação entre um usuário e uma aplicação, incluindo o próprio usuário e a aplicação. Essa definição faz com que seja mais fácil para um desenvolvedor enunciar um contexto para um determinado cenário de aplicação. Se parte da informação puder ser usada para caracterizar a situação de um participante em uma interação, então, essa informação é contexto (DEY, 2001).

Na prática, contextos estão implícitos na gestão da qualidade da informação e ainda têm sido considerados uma parte crítica na resolução dos problemas inerentes a essa gestão (DEY, 2001) (LEE, 2004) (PINHEIRO & MOURA, 2004), por exemplo, "História" e "Economia" como contextos de representação dessas áreas de conhecimento. No Google® e em outros *sites* de busca eles estão organizados em estruturas de diretórios por assuntos e em categorias como ilustrado a seguir:

- Artes e *Entretenimento*: Música, Televisão, Rádios, ...
- Ciência e Meio Ambiente: Engenharia, Física, Agropecuária, ...
- Estado e Governo: Embaixadas e Consulados, ...
- Estados: São Paulo, Rio Grande do Sul, ...
- Negócios e Economia:Informática, Compras, ...
- Notícias e Mídia:Revistas, Televisão, Rádio, ...
- Passatempos e Esportes: Futebol, Aquáticos, Artes Marciais, ...
- Regiões: Nordeste, Norte, Centro-Oeste, ...
- Saúde: Clínicas e Hospitais, ...
- Sociedade e Cultura: Religião e Espiritualidade, ...
- Transportes: Aéreas, Rodoviárias, ...
- Viagens e Turismo: Minas Gerais, Hospedagem, ...

Pipino, Lee *et al.* (2002) classificam as avaliações de qualidade em objetivas e subjetivas. Por sua vez, as avaliações objetivas são classificadas em métricas dependentes ou independentes da tarefa. As métricas independentes da tarefa refletem os estados dos dados sem o conhecimento contextual da aplicação e podem ser aplicadas a qualquer conjunto de dados, independentemente das tarefas em questão. Em contrapartida, as métricas dependentes da tarefa são desenvolvidas em contextos de aplicação específicos, que incluem as regras de negócio da organização, regulamentos governamentais e empresariais e as restrições fornecidas pelo administrador do banco de dados entre outros aspectos.

Esta pesquisa tem como foco as classificações referentes às métricas dependentes da tarefa, considerando que ela aborda uma contextualização e as perspectivas do usuário.

## 2.7 – Modelos, Metodologias e Categorias de Ferramentas de Qualidade de Informação

### 2.7.1 – Modelos de Qualidade de Informação

Os modelos são geralmente usados em bancos de dados para representar os dados e os esquemas de dados. Também são adotados pelos sistemas de informação para representar os processos de negócio das organizações. Eles devem ser estendidos com a finalidade de representar as dimensões e outros aspectos relacionados à qualidade de informação. Mais especificamente, os valores das dimensões de qualidade podem ser associados aos vários elementos dos modelos de dados, variando desde um valor de dados simples até a fonte de dados como um todo (BATINI, BARONE *et al.*, 2007).

Numa abordagem mais abstrata, Becker & McMullen *et al.* (2007) definem formalmente um metamodelo de qualidade de dados como uma ferramenta útil para garantir que as técnicas de gerenciamento de qualidade de dados a serem adotadas incorporem adequadamente os requisitos flexibilidade, generalidade e facilidade de uso, para cada situação de emprego em potencial. Além desses requisitos, o metamodelo proposto busca atender aos objetivos básicos de uma arquitetura de qualidade de dados, ou seja, primeiramente representar informações produzidas sobre os objetos de dados, medidas de QD, requisitos de QD, avaliações de QD e, finalmente, as ações de QD. Algumas extensões do metamodelo podem ser definidas para cobrir tópicos como agregação, métricas de credibilidade e julgamento dos usuários, rastreamento da linhagem dos dados, métricas parametrizadas e validação dos dados. Elas ainda atentam para o grande número de aplicações de interesse como suporte à decisão, gestão de dados e processos de melhoria contínua (BECKER, MCMULLEN *et al.*, 2007).

Outro modelo de qualidade genérico é proposto por Berti-Équille (2007). Esse modelo foi formalizado por meio de um diagrama de classes UML e é apresentado na Figura 2-2.

Tal modelo sintetiza os aspectos mais comuns que foram adotados pelas abordagens constantes da Tabela 2-3. Nele está representado que uma ou mais instâncias de qualidade dos dados podem estar associadas a um item de dados, isto é, valor do atributo, conjunto dos valores, registro, tabela, domínio, etc. A qualidade dos dados é composta por uma ou mais dimensões com visibilidade pública, para os

atributos que representam o tipo e a categoria da dimensão de qualidade. A dimensão de qualidade, por sua vez, é composta por uma ou mais medidas caracterizadas por seu tipo, métrica e descrição. Cada medida tem um ou mais valores de qualidade, com a data e a unidade de medida correspondentes (BERTI-´EQUILLE, 2007).

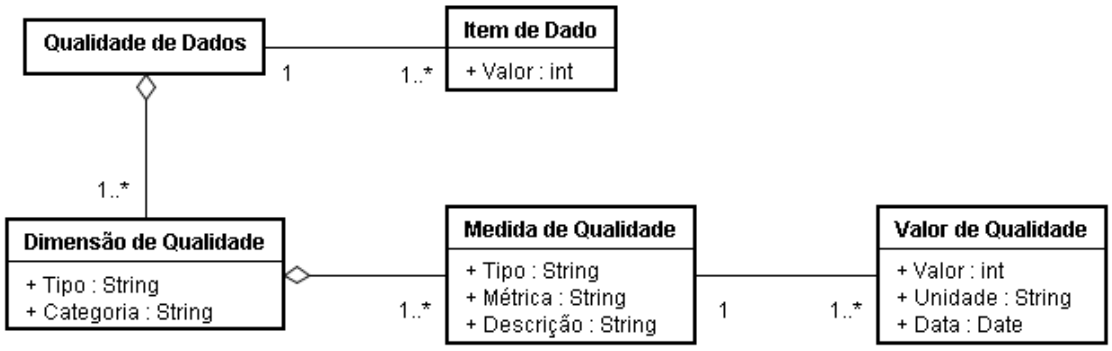



**Figura 2-2: Modelo de Qualidade de Dados (Fonte: Berti-´Equille, 2007)**

Adicionalmente aos modelos, existem as técnicas que correspondem aos algoritmos, às heurísticas, aos procedimentos baseados no conhecimento e aos processos de aprendizado que provêem solução aos problemas específicos relacionados às atividades de qualidade de informação (BATINI & SCANNAPIECO, 2006).

Além dos modelos e das técnicas, existem os métodos que devem ser definidos para a obtenção dos valores dos critérios de qualidade identificados, antes da utilização da qualidade como auxílio à busca de informação (BURGESS, GRAY *et al.*, 2004).

Naumann & Rolker (2000) definem um conjunto de classes para a avaliação de qualidade da informação e identificam os métodos que foram desenvolvidos para serem utilizados em cada classe. As fontes para avaliação dos critérios de qualidade são identificadas no seu artigo como: o usuário, a fonte de informação, bem como o processo de consulta adotado na obtenção da informação.

Essas três fontes são divididas nas seguintes classes de avaliação:

- Critérios subjetivos – quando os escores de qualidade de informação só podem ser obtidos a partir de usuários individuais, com base nas suas visões pessoais, experiências e práticas;
- Critérios Objetivos – quando os escores de qualidade de informação podem ser obtidos pela análise das informações;
- Critérios de processos – quando os escores de qualidade de informação são determinados pelo processo de consulta.

Para cada uma dessas classes de avaliação, Naumann & Rolker (2000) apresentam um conjunto de métodos de avaliação que pode ser usado para avaliar a qualidade de cada fonte de informação:

- Critérios subjetivos – experiência, previsões e avaliação contínua do usuário.
- Critérios Objetivos – contrato de qualidade, análise, previsão, entradas por especialista e avaliação contínua do conteúdo.
- Critérios de processos – limpeza de dados, avaliação contínua do processo e análise estrutural.

A Tabela 2-1 e a Tabela 2-3 mostram uma variedade de iniciativas e cada uma delas obedecendo ou propondo os modelos de qualidade que foram desenvolvidos para os seus domínios específicos.

### 2.7.2 – Metodologias de Qualidade de Informação

As metodologias de qualidade de informação são definidas em Batini & Scannapieco (2006) como um conjunto de orientações e técnicas que têm início nas informações de entrada relativas a uma determinada realidade de interesse. A partir dessas informações de entrada, elas definem um processo racional para medir e melhorar a qualidade dos dados de uma organização, obedecendo às fases e aos pontos de decisão determinados.

Essas metodologias têm por objetivo prover uma avaliação precisa ou um diagnóstico do estado dos sistemas de informação no que se refere às questões de qualidade de dados. Assim, as suas principais saídas são:

- Medidas de qualidade das bases de dados e dos fluxos de dados;
- Custos decorrentes da baixa qualidade dos dados; e
- Uma comparação dos níveis de qualidade dos dados considerados aceitáveis a partir do conhecimento existente ou mesmo de um *benchmarking* das boas práticas juntamente com as sugestões de melhoria.

Usualmente, os processos observados pelas metodologias de avaliação possuem três principais atividades:

*i.* As dimensões e métricas aplicáveis são inicialmente escolhidas, classificadas e medidas;

*ii.* Os julgamentos subjetivos dos especialistas são realizados; e

*iii.* As medições objetivas e os julgamentos subjetivos são comparados.

Dravis (2005) identificou uma série repetitiva de fases adotada pelas iniciativas de QI (Qualidade de Informação) no momento em que as pessoas ou as organizações necessitam obter soluções para os problemas de qualidade. Esse ciclo para solução dos problemas de QI possui três fases principais denominadas: percepção, quantificação e implementação, que as equipes de projeto de QD irão seguir, implícita ou explicitamente, no curso da implementação da solução. Até chegar a fase final de implementação da iniciativa de QI, ele identifica e explora os sucessivos estágios associados a cada uma das fases do ciclo, desde a identificação do problema e da tomada de decisão para o prosseguimento das ações, até a pesquisa da solução e a aprovação do projeto (DRAVIS, 2005).

A Figura 2-3 generaliza as principais fases das metodologias de avaliação da qualidade de informação em vista de suas principais atividades (DE AMICIS & BATINI, 2004).

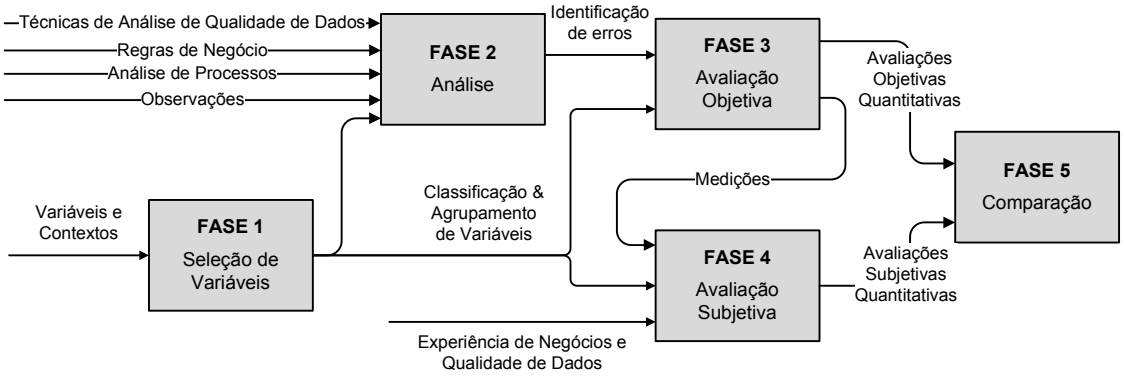



**Figura 2-3: Principais Fases das Metodologias de Avaliação de QI (Adaptada de Amicis & Batini, 2004)**

As metodologias de qualidade de dados podem ser classificadas de acordo com alguns critérios a seguir descritos (BATINI & SCANNAPIECO, 2006):

*i.* Orientada aos Dados *vs.* Orientada aos Processos – essa classificação está relacionada à estratégia geral escolhida para o processo de melhoria. As estratégias orientadas aos dados são baseadas exclusivamente no uso de fontes de dados para a melhoria de sua qualidade. Nas estratégias orientadas aos processos, o processo de produção de dados é analisado e possivelmente modificado para identificar e remover as causas originais dos problemas de qualidade;

*ii.* Medição *vs.* Melhoria – as metodologias são empregadas para a medição/avaliação da qualidade dos dados, ou para melhorar a sua qualidade. As atividades de medição e melhoria são intimamente relacionadas, considerando que, mesmo que somente as atividades de medição estejam disponíveis, é possível conceber técnicas a serem aplicadas e prioridades a serem estabelecidas. Conseqüentemente, o limite entre as duas propostas é vago em alguns casos;

*iii.* Propósito-Geral *vs.* Propósito- Especial – as metodologias de propósito-geral abrangem um amplo espectro de fases, dimensões e atividades, enquanto que as de propósito-especial estão focadas em atividades específicas (ex.: medição, identificação de objeto) em um domínio específico de dados (ex: um censo, uma base de endereços de pessoas) ou domínio de aplicações específico (ex.: biologia);

*iv.* Intra-Organizacional *vs.* Interorganizacional – as metodologias intra-organizacionais referem-se às atividades de medição e melhoria concernentes a uma organização específica, ou ao setor específico de uma organização, ou mesmo a um processo específico ou uma base de dados. As interorganizacionais referem-se a um grupo de organizações cooperando para um objetivo comum (ex: um grupo de agências públicas provendo melhores serviços aos cidadãos e às empresas).

Ainda no referencial literário, foi identificada outra classificação ortogonal à que foi mostrada. Ela descreve de forma básica, três tipos principais de abordagens para as propostas de metodologias de qualidade de dados: a intuitiva, a teórica e a empírica (STRONG & WANG, 1996) (BATINI & SCANNAPIECO, 2006) (GE & HELFERT, 2007).

A abordagem intuitiva incorpora um conjunto de atributos de qualidade de dados para alguma demanda específica, com base na experiência ou julgamento intuitivo dos pesquisadores. Nesses casos, os requisitos de qualidade são identificados de acordo com os contextos específicos das aplicações.

A abordagem teórica adota modelos formais para justificar as dimensões avaliadas. Ela enfatiza quão os dados podem se tornar deficientes durante seu processo de produção. Por exemplo, Wand & Wang (1996) definiram as dimensões de qualidade dos dados, usando conceitos ontológicos, baseados nos problemas que acontecem

durante o mapeamento dos dados do mundo real para os sistemas de informação. Esse estudo observa que o desenvolvimento e o uso da informação envolvem duas transformações: a transformação de representação e a transformação de interpretação. Dessa forma, assume-se que uma deficiência de dados pode acontecer durante a transformação de representação e/ou na transformação de interpretação, gerando, assim, uma falta de conformidade entre a visão do mundo real e aquela obtida do sistema de informação.

Na abordagem empírica são capturados os atributos de qualidade de dados que são importantes para os usuários. Os dados colecionados pelos usuários são analisados, para determinar as características que irão definir se os dados são adequados às suas tarefas ou não. Essa abordagem é usada quando a qualidade de dados está baseada na experiência ou no entendimento sobre quais são os atributos importantes do ponto de vista dos usuários (WANG, 1998), (ROCHA, 1983) e (WANG, KON *et al.*, 1993).

A Tabela 2-1 e a Tabela 2-3 mostram uma variedade de iniciativas e cada uma delas obedecendo ou propondo as metodologias de qualidade que foram desenvolvidas para os seus domínios específicos.

### 2.7.3 – Categorias de Ferramentas de Qualidade de Informação

A qualidade de dados tem demandado um contínuo crescimento na oferta e na busca de tecnologias inovadoras. Como resultado, o mercado de ferramentas de qualidade de dados tem crescido de forma sólida com a participação e a entrada de pequenos e grandes fornecedores (GOASDOUÉ, NUGIER *et al.*, 2007) (FRIEDMAN & BITTERER, 2007) .

No contexto da *Web*, novas ferramentas têm sido concebidas para auxílio aos usuários e aos desenvolvedores (*Webmasters*) na avaliação de qualidade da informação. Elas variam em termos de custos e níveis de auxílio, desde as ferramentas livres até as ferramentas customizadas para propósitos específicos.

Adicionalmente às ferramentas que coletam informações sobre métricas objetivas de qualidade, existem aquelas que podem recolher informações sobre métricas de qualidade obtidas por meio de pesquisas junto aos usuários. Essas informações são difíceis de serem medidas tecnicamente e, para reuni-las, há softwares de pesquisa que

apóiam as perguntas e as avaliações na *Web*, sobre aspectos como a usabilidade, conveniência, completeza, utilidade, relevância, etc.

As ferramentas de qualidade de informação no contexto da *Web* foram classificadas por Eppler & Muenzenmayer (2002) em cinco diferentes tipos:

*i.* Monitoramento de Desempenho;
*ii.* Analisadores de *Site*;
*iii.* Analisadores de Tráfego;
*iv.* Mineração na *Web*; e
*v.* Avaliação dos Usuários (para geração de pareceres com base nas opiniões dos utilizadores).

Eppler & Muenzenmayer (2002) também sustentam que o uso combinado dessas ferramentas pode medir as múltiplas dimensões de qualidade das informações no contexto da Internet ou da Intranet. Para tanto, é requerida uma metodologia clara baseada em uma sistemática de passos seqüenciais e em um *framework* de qualidade de informações, para o delineamento dos critérios de qualidade aplicáveis.

As cinco categorias são descritas a seguir:

*i.* Monitoramento de Desempenho – Teste de Monitoramento de Servidor e de Rede:

Nessa categoria encontram-se as ferramentas que monitoram a disponibilidade e o desempenho dos servidores e das redes (ex.: tempo de parada e tempo de resposta). Essa categoria está sendo sumariamente abordada, considerando que ela é bastante conhecida e subsidia apenas alguns poucos critérios de qualidade na *Web*, como a "velocidade" e a "confiabilidade".

*ii.* Analisadores de *Site*

Eles ajudam a analisar um *site* com base em diferentes critérios de qualidade. Vários aspectos da qualidade podem ser examinados e representados de forma automatizada e agregada em um relatório. Essas ferramentas foram desenvolvidas como *sites Web* e têm crescido em tamanho e complexidade, tornando cada vez mais difícil a sua manutenção e a sua gerência. Elas buscam identificar, por exemplo, *hiperlinks* obsoletos ou imagens perdidas e oferecem um conjunto de critérios e outras múltiplas funções para verificar a

qualidade nesses tipos de problemas. A seleção dessas ferramentas pode começar pelas que estão focadas em aspectos ou critérios de qualidade específicos, como metainformação ou o próprio código HTML, ou pelas que cuidam de um conjunto mais abrangente de critérios de qualidade e os representam em um relatório agregado com possibilidades de detalhamento. Os analisadores de s*ite* têm como foco prover funcionalidades padrão e especiais, identificar e analisar: *links* quebrados; arquivos órfãos; erros ortográficos; falta ou duplicação de palavras-chave ou títulos de página; inventário dos *sites* com *links* de estrutura; imagens usadas; tipos de documentos; mapas de imagem utilizados; páginas multimídia; proporção das páginas antigas *versus* novas páginas; controle de operações; capacidade de análise e planejamento; habilidade de busca com a capacidade de adicionar automaticamente *metatags*.

*iii.* Analisadores de Tráfego

A fim de obter avaliações sobre o tráfego e comportamento nos *sites*, essas ferramentas recolhem esses dados e os apresentam em relatórios, diagramas e tabelas representativas. Além disso, a integridade do *site* e a funcionalidade de trabalho na sua efetiva utilização são de grande interesse para as questões estratégicas e de qualidade.

Existem duas possibilidades diferentes para obtenção dos dados sobre o tráfego de usuários. A primeira baseia-se no *log* do servidor de arquivo e a outra num coletor de rede dedicado. O *log* é mais amplamente utilizado, pois a sua implementação é muito mais fácil e econômica. O coletor de rede permite uma avaliação mais detalhada e precisa, bem como uma medição de critérios de qualidade de informação não registrados no *log*. Outra vantagem dos coletores é a maior rapidez na avaliação. Porém, eles são mais caros, tanto em termos de hardwares quanto de softwares requeridos.

A funcionalidade padrão dos analisadores de tráfego inclui: acessos às páginas; visões; visitas; arquivos e páginas mais e menos solicitados; informações sobre os visitantes (ex.: segmentação geográfica; tipo de navegador; *plug-ins* instalados; endereços IP); a possibilidade de filtrar informações (ex.: por dia, semana, a região, navegador, etc.); pesquisa de DNS inversa para mostrar o domínio em vez do endereço IP e relatório

padrão (ex.: máquinas de busca utilizadas, palavras-chave e frases de pesquisa).

Alguns analisadores de tráfego também incluem os seguintes serviços: avaliação dos componentes especiais (ex.: *banners*, *links*); o tempo gasto nas páginas; relatórios sobre tendências dos visitantes; e o dinheiro gasto pelos clientes nas páginas (*ROI*[15]).

Os três tipos de ferramentas mostrados podem gerar uma enorme quantidade de informações. Essas informações necessitam ser integradas antes de serem analisadas nas questões subjacentes de qualidade.

As ferramentas de mineração na *Web*, descritas a seguir, viabilizam a integração dos dados relevantes das medições de qualidade de informação.

*iv*. Mineração na *Web*

Nessa categoria encontram-se as ferramentas que integram dados dos analisadores de *site*, dos analisadores de tráfego ou dos sistemas legados. Quanto mais ampla a base de dados mais precisa será a sua análise. Os dados, por exemplo, podem ser integrados a partir de um sistema de gerenciamento de conteúdo (CMS) com os dados dos analisadores de tráfego, a fim de obter uma melhor compreensão sobre os custos de manutenção de um *site* e do seu retorno para a melhoria do tráfego dos usuários. Outro exemplo, que conduz a uma visão valiosa sobre o comportamento da navegação dos usuários, é a integração dos dados dos analisadores de *site* (ex: a estrutura de um *site*) com os dados de tráfego do usuário.

*v*. Avaliação dos Usuários

Embora as ferramentas apresentadas até agora sejam bastante efetivas, ainda existem alguns critérios de qualidade (como a abrangência, clareza e precisão), que são difíceis ou mesmo impossíveis de serem tecnicamente medidos, ou a técnica de medição é demasiadamente cara na sua implementação. Nesses casos, a avaliação dos usuários é a forma adequada para avaliar os critérios de qualidade de informação. Existem diferentes possibilidades para receber as avaliações dos usuários, que vão desde uma ou duas questões simples de opinião, até formulários de avaliação completos

[15] Retorno sobre o Investimento.

com questões por ordem de seleção, entrevistas de seleção de nomes de parceiros e relatórios personalizados para diferentes papéis de usuários. Esses sistemas normalmente incluem as seguintes funcionalidades: representação gráfica dos resultados (ex.: gráficos de pizza, diagramas); métricas de avaliação dos resultados (ex.: valores médios); gráficos para a construção de questionários e modelos de *layout* e contexto.

Outros sistemas mais sofisticados oferecem funcionalidades adicionais, tais como: apoio a todo o processo de criação dos questionários; mala-direta; a avaliação das questões quantitativas para as avaliações dos usuários; exportação para diferentes formatos e sistemas; diferentes possibilidades para iniciar os questionários (ex.: por e-mail por meio de *links* nas páginas da *Web*) e a possibilidade de filtrar os resultados por hora, região, etc.

Existem softwares disponíveis para todas as cinco categorias de ferramentas, principalmente para as categorias *i, ii e iii*, que desempenham medições quantitativas, mediante avaliações de critérios de qualidade objetivos. Para as categorias *iv e v* existem ferramentas mais elaboradas, porém muito mais caras. Ao contrário das anteriores, elas desempenham medições qualitativas, mediante avaliações de critérios de qualidade subjetivos.

Todavia, apesar dessa diversidade de categorias e ferramentas, a percepção de qualidade, seja ela alta, média ou baixa, ainda depende do contexto das aplicações e dos requisitos definidos por seus usuários.

No Tabela 2-7 apresenta uma visão geral dos instrumentos que cobrem essas funcionalidades. Essa Tabela lista várias ferramentas em cada categoria e fornece os nomes dos produtos, seus fornecedores e o endereço do *site*.

**Tabela 2-7: Exemplos de Ferramentas de QI na *Web* (Adaptada de Eppler & Muenzenmayer, 2002)**

| Produto | Vendedor | URL |
|---|---|---|
| **Analisadores de *Site*** | | |
| Hypertrak Performance Monitor | Trio Networks | www.trionetworks.com |
| Watchfire Enterprise Solution | Watchfire | www.watchfire.com |

| Produto | Vendedor | URL |
|---|---|---|
| *Web*Analyzer 2.0 | InContext | www.incontext.com |
| *Web*master 5.0 | Coast | www.coast.com |
| **Analisadores de Tráfego** | | |
| Analog Logfile Analysis 5.03 | University of Cambridge Statistical Laboratory | www.analog.cx |
| Live Stats *Web* Analytics Server 6 | Deepmetrix | www.deepmetrix.com |
| Nedstat Basic | Nedstat | www.nedstat.com |
| PerfMan for *Web*servers | ISM | www.perfman.com |
| SiteStat | Nedstat | www.nedstat.com |
| Summary Plus 2.0 | Summary.net | http://summary.net |
| Surfreport 3.0 | netrics.com | www.surfreport.com |
| Urchin Multihome 3 | Quantified Systems | www.urchin.com |
| Website Analysis Suite | Hyperion | www.hyperion.com |
| *Web*Suxess 4.0 | Exody | www.exody.net |
| Wusage 7.1 | Boutell.com | www.boutell.com |
| Xcavate 1.9 | Expertise | www.exsoft.com |
| **Mineração na *Web*** | | |
| Accrue G2/ Hitlist | Accrue | www.accrue.com |
| C-Insight | Metaedge Corp. | www.metaedge.com |
| Clementine | SPSS | www.spss.com |
| EasyMiner 2 | Mine It | www.mineit.com |
| Synera ePack | Synera | www.synerasystems.com |
| Funnel *Web*Suite | Quest | www.quest.com |
| Esite | Informatica | www.informatica.com |
| Netgenesis 5 | Netgenesis | www.netgen.com |

| **Produto** | **Vendedor** | **URL** |
|---|---|---|
| NetTracker eBusiness Solution 5.5 | Sane Solutions | www.sane.com |
| *Web*Abacus | *Web*Abacus | www.webabacus.com |
| *Web*Feedback 3.0 | Liebhart Systems | www.cyberware-neotek.com/WFB |
| *Web*master Pro | Coast | www.coast.com |
| *Web*mining Genius | Novuweb | www.novuweb.com |
| *Web*trends Analysis Suite 7.0 | NetIQ | www.webtrends.com |
| **Avaliação dos Usuários** | | |
| Cont@xt Information | Factory | www.information-factory.com |
| Opinion Poll | Metrix Lab | www.opinionpoll.com |
| Infopoll Business Intelligence Suite I | InfoPoll | www.infopoll.com |
| *Web*Surveyor | *Web*Surveyor Corp. | www.websurveyor.com |

# Capítulo 3 – Enfoques Sobre a Teoria *Fuzzy*

A teoria dos conjuntos *fuzzy* (nebulosos) é usada para representar modelos de raciocínio impreciso, que possuem um papel essencial na notável habilidade humana, para tomar decisões racionais em ambientes de incertezas e imprecisões (ZADEH, 1988).

Este capítulo apresenta alguns enfoques sobre a teoria dos conjuntos *fuzzy*, em seus aspectos mais gerais, objetivando fornecer conceituações e propriedades básicas dessa teoria, juntamente com suas referências bibliográficas.

## 3.1 – Breve Introdução sobre a Teoria *Fuzzy*

A principal motivação da teoria dos conjuntos *fuzzy* é o desejo de construir uma estrutura formal quantitativa capaz de capturar as imprecisões do conhecimento humano, isto é, como esse conhecimento é formulado na linguagem natural. Essa teoria visa ser a ponte que une modelos matemáticos tradicionais, precisos, de sistemas físicos e a representação mental, geralmente imprecisa, desses sistemas (DUBOIS & PRADE, 1991).

A mente humana opera com conceitos subjetivos, tais como *alto*, *baixo*, *velho* e *novo*, que são incorporados em classes de objetos na teoria *fuzzy*, onde a pertinência ou não de um elemento a um conjunto ocorre de forma gradual e não abrupta (ZADEH, 1990). O advento dessa teoria viabilizou substanciais instrumentos para a representação de várias facetas cognitivas humanas (YAGER, 1991). Ela provê ferramentas robustas para a aplicação do conhecimento, da experiência e do pensamento humano em muitos sistemas industriais, de tráfego, ciência médica, entre outros (SUZUKI, 1993).

Giles (1988) levanta dois enfoques, igualmente importantes, para a formulação da teoria dos conjuntos *fuzzy*: o *axiomático* e o *semântico* (GILES, 1988).

No axiomático, uma função numérica é usada para modelar o conjunto *fuzzy*, fornecendo uma interpretação consistente para ele, embora não evidencie, diretamente, a estrutura sob esse conjunto (FRENCH, 1986; FRENCH, 1989). Para a adoção de certas definições de operações de conjuntos *fuzzy*, é condição necessária e suficiente que

essas definições sejam avaliadas por vários pesquisadores (BELLMANN & GIERTZ, 1973) e que se utilizem do enfoque axiomático.

No enfoque semântico, é analisado, empiricamente, o significado físico dos conceitos do sistema envolvido, modelando-os através de conjuntos *fuzzy*. Esses conceitos são precisamente formulados como axiomas quantitativos, produzindo não só uma teoria consistente, mas, também, uma interpretação específica deles. Esse enfoque pode, ainda, ser dividido em métodos *normativos* e *descritivos* (FRENCH, 1986). O método normativo conjectura como as pessoas poderiam organizar seus julgamentos em uma situação particular, através de conjuntos *fuzzy*. O descritivo (mais usado) investiga como de fato as pessoas realizam seus julgamentos (SDORRA, 1993).

Ambos os enfoques, *axiomático* e *semântico*, têm sido reportados na literatura atual em medidas de funções *fuzzy* ou simplesmente *funções de pertinência*. Enfoques *híbridos* têm sido mencionados, nos quais métodos experimentais foram usados em conjunção com os formais, testando e validando suas hipóteses e conseqüências (TURKSEN, 1991).

Esta tese investiga a utilização de um modelo *fuzzy* para atributos de qualidade de dados, visando interpretar, da melhor maneira possível, os indicadores de qualidade das informações acessadas na *Web* e apresentá-los aos seus usuários.

A abordagem também foi aplicada para identificar as expectativas dos usuários sobre a QI, por meio de pesos atribuídos às dimensões de qualidade de acordo com os contextos específicos e os requisitos por eles especificados.

A seguir, serão abordados alguns conceitos e propriedades da teoria dos conjuntos *fuzzy*, baseados no levantamento feito na literatura (BELCHIOR, XEXEO *et al.*, 1997; DUBOIS & PRADE, 1980; DUBOIS & PRADE, 1991; FUHRMANN, 1990; KLIR & FOLGER, 1988; KLIR & YUAN, 1995; ZIMMERMANN, 1991; ZADEH, 1965).

## 3.2 – Conceitos Básicos dos Conjuntos *Fuzzy*

### 3.2.1 – Conjuntos Nítidos e Conjuntos *Fuzzy*

Tradicionalmente, na Teoria dos Conjuntos, um elemento pertence ou não pertence a um conjunto. Os conjuntos *fuzzy*, uma generalização de um conjunto nítido

(ou ordinário), representado por $\tilde{A}$ em *X*, permitem a definição de um grau de dependência para cada elemento *x*, isto é, um número real no intervalo $[0,1]$, em decorrência de sua *função de pertinência* característica. Nesse caso, se o grau é zero, o elemento não pertence ao conjunto e, se é 1, o elemento pertence totalmente ao conjunto (TURKSEN, 1991; ZIMMERMANN, 1991).

Um elemento pode pertencer parcialmente a um conjunto *fuzzy*, por exemplo, o conjunto de pessoas "jovens". Quando uma pessoa não é mais jovem? A definição de um conjunto *fuzzy* pode-nos mostrar uma pessoa de 20 anos como "90 % jovem", enquanto que alguém de 60 anos seria apenas "30 % jovem".

Essas características permitem que a lógica *fuzzy* manipule os objetos do mundo real que possuem limites imprecisos. Utilizando *predicados fuzzy* (*velho, novo, alto, etc.*), *quantificadores fuzzy* (*muitos, poucos, quase todos, etc.*), valores verdade *fuzzy* (*completamente verdade, mais ou menos verdade*) e generalizando o significado dos conectores e operadores lógicos, a lógica *fuzzy* é vista como um meio de raciocínio aproximado (GRAUEL, 1999). Esses aspectos serão discutidos mais adiante.

Qualquer representação adequada de um conjunto *fuzzy* envolve o entendimento básico de cinco diferentes símbolos conceituais, relacionados entre si (TURKSEN, 1991):

*i.* *Conjunto de elementos* $\theta \in \Theta$: por exemplo, um "homem" em "homens" ou um "item" em "estoque";

*ii.* Variável *lingüística* ***V*****:** em um conjunto de variáveis lingüísticas, que é um rótulo para um atributo dos elementos $\theta \in \Theta$, como "*altura de homem"* ou o "*nível de estoque*" de uma empresa;

*iii.* *Termo* lingüístico ***T*****:** de uma *variável lingüística*, correspondendo a um adjetivo ou a um advérbio, em um conjunto de termos lingüísticos, como "*homem alto"* associado com a "*altura do homem*" ou "*estoque baixo"*, relacionado com possíveis "*níveis de estoque*" de uma empresa;

*iv.* *Intervalo numérico mensurável* $X \in [-\infty, \infty]$ conhecido como o *conjunto referencial* para um atributo particular ***V***, de um conjunto de elementos, como, por exemplo, "$[0,3]$ *metros*" para "*altura de homem"*, ou $[250,750]$ "*unidades*" para "*nível de estoque*";

*v.* *Atribuição numérica subjetiva* $\mu_{\tilde{A}}(\Theta)$, ou *valor de pertinência*, que é o grau com que um elemento pertence ao conjunto de elementos, rotulados por uma *variável lingüística* ***V*** e identificados pelo *termo lingüístico* ***T***. Por exemplo, o valor de pertinência dado a um "*homem*" em um grupo de homens por um observador, que usa o *termo lingüístico "alto"*, segundo sua visão de "*altura"* para homens, ou o valor de pertinência atribuído por um gerente para "*estoque"*, através do adjetivo "*baixo"*, englobando todos os níveis de estoque sob o seu gerenciamento.

A teoria *fuzzy* é usada basicamente para mapear modelos qualitativos de tomada de decisões e para métodos de representação imprecisa. Nesse contexto é que se pode utilizar a teoria *fuzzy* em dimensões (atributos) de qualidade de informações ou dados, uma vez que são, em sua maioria, conceitos subjetivos e de avaliação não trivial.

Uma avaliação de qualidade pode ter um escopo amplo e ser composta por outros atributos de qualidade. O modelo apresentado nesta tese propõe a utilização de funções *fuzzy* para interpretar medidas e definir processos de agregação de atributos.

Não basta, apenas, identificar as dimensões que determinam a qualidade dos dados, mas também que procedimentos adotar, para controlar seu processo de desenvolvimento (ciclo de vida dos dados), de forma a atingir o nível de qualidade desejado. Esse processo é realizado através da aplicação de métricas de qualidade, que são as medidas ou as avaliações das características de qualidade das informações.

Em resumo, a teoria dos conjuntos *fuzzy* é capaz de capturar as imprecisões do conhecimento humano, além de apresentar outras vantagens como: abordar gradativamente a solução, implementar um misto de matemática e objetividade com regras e subjetividade, empreender um misto de tratamento quantitativo com qualitativo, implementar o rigor matemático para temas normalmente tratados sem rigor (similaridade)[31].

[31] Curso Intensivo de Matemática Aplicada e Computação na Engenharia – COPPE/UFRJ. Conjuntos e Lógica *Fuzzy*. Prof. Geraldo Xexéo. http://ge.cos.ufrj.br/twiki/bin/view/Ufrj/WebHome.

### 3.2.2 – Função *Fuzzy* de Pertinência

A função de pertinência é o componente crucial de um conjunto *fuzzy* e muitas operações são definidas em conformidade com ela (ZADEH, 1965). Muitas vezes, a estimativa da função de pertinência $\mu_{\tilde{A}}(x)$, em situações do mundo real, vem de informações imprecisas e incompletas. Dado um conjunto *fuzzy* $\tilde{A}$, duas notações são comumente empregadas, para caracterizar essas funções:

$$\mu_{\tilde{A}}(x): X \rightarrow [0,1] \quad \text{ou} \quad \tilde{A}: X \rightarrow [0,1]$$

De acordo com a primeira notação, o identificador do conjunto *fuzzy* $\tilde{A}$ é distinguido do símbolo de sua função de pertinência $\mu_{\tilde{A}}(x)$. Na segunda, esta distinção não é feita. No entanto, não há ambigüidades nos resultados dessa dupla utilização, pois cada conjunto *fuzzy* é completo e unicamente definido por uma função de pertinência particular e, conseqüentemente, identificadores das funções de pertinência podem ser também usados como símbolos de seus conjuntos *fuzzy* associados.

*Exemplo 1*:

Seja *X* o conjunto universo de idades e os subconjuntos *fuzzy*, contidos em *X*, classificados como *infantil*, *adulto*, *jovem*, e *velho*, envolvendo todas as possibilidades de subconjuntos *fuzzy* de *X*, que é denotado por Ñ: $(X)$ , como mostra a Tabela 3-1.

$$X = \{5, 10, 20, 30, 40, 50, 60, 70, 80\}$$

**Tabela 3-1: Exemplo de conjuntos *fuzzy* (Fonte: Klir & Folger, 1988)**

| Idades | Infantil | Adulto | Jovem | Velho |
|---|---|---|---|---|
| **5** | 0 | 0 | 1 | 0 |
| **10** | 0 | 0 | 1 | 0 |
| **20** | 0 | 0,8 | 0,8 | 0,1 |
| **30** | 0 | 1 | 0,5 | 0,2 |
| **40** | 0 | 1 | 0,2 | 0,4 |
| **50** | 0 | 1 | 0,1 | 0,6 |
| **60** | 0 | 1 | 0 | 0,8 |
| **70** | 0 | 1 | 0 | 1 |
| **80** | 0 | 1 | 0 | 1 |

### 3.2.3 – Representação de um Conjunto *Fuzzy*

#### *3.2.3.1 – Conjunto fuzzy*

Os conjuntos *fuzzy* prestam-se às representações de conceitos vagos, expressados na linguagem natural, dependendo do contexto em que são usados. Usar-se-á, a seguir, a formalização de conjuntos *fuzzy* com suporte finito (uma das várias existentes na literatura).

Um conjunto *fuzzy* é denotado por um conjunto de pares ordenados, em que o primeiro elemento é $x \in X$, e o segundo, $\mu_{\tilde{A}}(x)$, é o grau de pertinência ou a função de pertinência de $x$ em $\tilde{A}$, que mapeia $X$ para o espaço de pertinência $M$. Quando $M$ contém somente os pontos 0 e 1, $\tilde{A}$ é *não-fuzzy* (ZIMMERMANN, 1991):

$$\tilde{A} = \left\{\left(x, \mu_{\tilde{A}}(x)\right) \middle| x \in X\right\}$$

*Exemplo 2*:

O conjunto *fuzzy* $\tilde{A}$, *jovens*, da Tabela 3-1, pode ser descrito como:

$$\tilde{A} = \{(5; 1), (10; 1), (20; 0{,}8), (30; 0{,}5), (40; 0{,}2), (50; 0{,}1)\}$$

#### *3.2.3.2 – Suporte*

O suporte de um conjunto *fuzzy* $\tilde{A}$, em um conjunto universo $X$, é o conjunto nítido que contém todos os elementos de $X$ com graus de pertinência diferentes de zero em $\tilde{A}$.

O suporte de um conjunto *fuzzy* $\tilde{A}$ em $X$, denotado por supp($\tilde{A}$) ou $S(\tilde{A})$, onde $\tilde{P}(X)$contém todos os subconjuntos *fuzzy* possíveis, é obtido pela função:

$$S: \tilde{P}(X) \rightarrow P(X)$$

Onde,

$$S\left(\tilde{A}\right) = \{(x \in X) | \mu_{\tilde{A}}(x) > 0\}$$

*Exemplo 3:*

O suporte do conjunto *fuzzy* $\tilde{A}$, jovem, da Tabela 3-1 é o conjunto nítido

$$S(\tilde{A}) = \{5, 10, 20, 30, 40, 50\}$$

O conjunto *fuzzy infantil* é um conjunto vazio, no conjunto universo escolhido.

Nesse caso, o suporte também é vazio, isto é, sua função de pertinência assinala 0 (zero) a todo elemento do conjunto universo.

#### *3.2.3.3 – Supremo*

O supremo, $\sup_{x \in X} \mu_{\tilde{A}}(x)$, de um conjunto *fuzzy* $\tilde{A}$ é o maior grau de pertinência obtido nesse conjunto por um desses elementos, isto é, sua altura, $h(\tilde{A})$. O contradomínio de uma função de pertinência é um subconjunto de números reais não negativos, cujo supremo é finito. Então,

$$h(\tilde{A}) = \sup_{x \in X} \mu_{\tilde{A}}(x)$$

#### *3.2.3.4 – Normalização*

Embora uma função de pertinência não esteja limitada ao contradomínio [0, 1], por conveniência, isto é, usualmente considerado como verdade, assume-se que um conjunto *fuzzy* $\tilde{A}$ é *normal* ou *normalizado*. Portanto, um conjunto *fuzzy* $\tilde{A}$ é chamado normal, quando $\sup_{x \in X} \mu_{\tilde{A}}(x) = 1$. Se $\sup_{x \in X} \mu_{\tilde{A}}(x) < 1$, ele passa a se denominar *subnormal*. A normalização de um conjunto *fuzzy* $\tilde{A}$, não vazio, é efetuada por:

$$\mu'_{\tilde{A}}(x) = \frac{\mu_{\tilde{A}}(x)}{\sup_{x \in X} \mu_{\tilde{A}}(x)}$$

#### *3.2.3.5 – Conjuntos de corte-* $\alpha$

Dado um conjunto *fuzzy* $\tilde{A}$, definido em X, a partir do grau de pertinência $\alpha \in [0,1]$, o *conjunto de corte*-α (α-*cut*) é o conjunto nítido *A*α, contendo todos os elementos de *X*, que possuem graus de pertinência em $\tilde{A}$ maiores ou iguais do que o valor especificado em α. Então,

$$A \propto = \{x \in X | \mu_{\tilde{A}}(x) \geq \alpha\}$$

O conjunto *de corte*-α *robusto* (*strong* α-*cut*), *A*'α, inclui apenas os elementos de graus de pertinência maiores que α. Então,

$$A' \propto = \{x \in X | \mu_{\tilde{A}}(x) > \alpha\}$$

Nesse caso, percebe-se que o suporte de $\tilde{A}$ corresponde, exatamente, ao conjunto de corte-α robusto de $\tilde{A}$ para α = 0.

*Exemplo 4*:

Ainda em referência à Tabela 3-1, os conjuntos de corte- α possíveis, para o conjunto *fuzzy* $\tilde{A}$, *jovem*, são:

$$A_{0,1} = \{5, 10, 20, 30, 40, 50\}$$
$$A_{0,2} = \{5, 10, 20, 30, 40\}$$
$$A_{0,5} = \{5, 10, 20, 30\}$$
$$A_{0,8} = \{5, 10, 20\}$$
$$A_{1,0} = \{5, 10\}$$

Nesse caso, o conjunto de corte- α robusto para α = 0,8 é *A'*α = {5, 10}.

O conjunto de todos os níveis $\alpha \in [0,1]$, que representa distintos conjuntos de corte-α de um dado conjunto *fuzzy* $\tilde{A}$, definido em *X*, é denominado *conjunto de nível de* $\tilde{A}, \Lambda_{\tilde{A}}$, dado por:

$$\Lambda_{\tilde{A}} = \{\propto | \mu_{\tilde{A}}(x) = \alpha \; para \; todo \; x \in X\}$$

A seguinte propriedade pode ser deduzida dos conjuntos de corte-α e corte-α robusto:

- Qualquer conjunto *fuzzy* $\tilde{A}$, com $\alpha_1, \alpha_2 \in [0,1]$ e $\alpha_1 \neq \alpha_2$, para $\alpha_2 < \alpha_2$, tem-se:

$$A_{\alpha_1} \supseteq A_{\alpha_1} \quad \text{e} \quad A'_{\alpha_1} \supseteq A'_{\alpha_1}$$

Em conseqüência dessa propriedade, todos os conjuntos de corte-α de um conjunto *fuzzy* $\tilde{A}$, em *X*, formam uma família de subconjuntos nítidos, aninhados em *X*.

### *3.2.3.6 – Cardinalidade*

A cardinalidade *escalar* ou, simplesmente, *cardinalidade*, $|\tilde{A}|$, de um conjunto *fuzzy* $\tilde{A}$, definido em *X*, é o somatório dos graus de pertinência de todos os elementos de *X* em $\tilde{A}$. Formalmente,

$$|\tilde{A}| = \sum_{x \in X} \mu_{\tilde{A}}(x)$$

Para um conjunto universo infinito *X*, a cardinalidade, que nem sempre existe (é necessário que $\mu_{\tilde{A}}(x)$ seja integrável), é dada por:

$$\left|\tilde{A}\right| = \int_x \mu_{\tilde{A}}(x)dx$$

A cardinalidade relativa, $\left\|\tilde{A}\right\|$, de um conjunto *fuzzy* $\tilde{A}$ depende da cardinalidade do conjunto universo considerado. Assim, deve-se escolher o mesmo conjunto universo *X*, caso se queira comparar conjuntos *fuzzy* através de sua cardinalidade relativa. Pode ser interpretada como a fração dos elementos de *X*, presentes em $\tilde{A}$, medidos por seus graus de pertinência:

$$\left\|\tilde{A}\right\| = \frac{\left|\tilde{A}\right|}{|X|}$$

*Exemplo 5*:

A cardinalidade escalar do conjunto *fuzzy* Ã, velho, da Tabela 3-1 é:

$$\left|\tilde{A}\right| = 0 + 0 + 0{,}1 + 0{,}2 + 0{,}4 + 0{,}6 + 0{,}8 + 1{+}1 = 4{,}1$$

A cardinalidade escalar do conjunto *fuzzy* Ã, infantil, é 0 (zero).

A cardinalidade relativa do conjunto *fuzzy* Ã, velho, é:

$$\left\|\tilde{A}\right\| = \frac{4{,}1}{9} = 0{,}456$$

#### *3.2.3.7 – Fuzificação*

A *fuzificação* ("*fuzzification*") acontece, quando um conjunto *fuzzy* $\tilde{A}$ é obtido pelo "*alargamento*" *fuzzy* de um conjunto nítido, isto é, um conjunto nítido é convertido em um conjunto *fuzzy* apropriado, para expressar medidas de incertezas.

*Exemplo 6*:

Seja o conjunto nítido $A = \{x|7 < x < 10\}$, então, pelo processo da *fuzificação*, teríamos o seguinte conjunto *fuzzyv* $\tilde{A} = \{x|7 \approx< x \approx< 10\}$, onde o símbolo $\approx$ é denominado um *fuzificador* e significa *aproximadamente*.

#### *3.2.3.8 – Defuzificação*

A *defuzificação* é a conversão de um conjunto *fuzzy* em um valor nítido (ou um vetor de valores) (ZIMMERMANN, 1991).

Teoricamente, qualquer função na forma $\tilde{A}: X \rightarrow [0,1]$ pode ser associada a um conjunto *fuzzy*, dependendo dos conceitos e das propriedades que precisam ser representadas no contexto em que o conjunto está inserido.

O processo de defuzificação pode ser definido como uma função que associa a cada conjunto *fuzzy* um elemento (do conjunto nítido subjacente) que o represente. O valor escolhido pode ser entendido como uma espécie de valor esperado, traçando uma analogia com as distribuições de probabilidade. Mas como fazer exatamente para obter o valor condensado a partir do conjunto *fuzzy*? Existem alguns métodos bastante utilizados (pelo menos 30) e o *COG – Center of Gravity* (método do centróide, centróide da área ou centro de gravidade) é o método mais comumente adotado. Ele fornece um valor correspondente à abscissa do baricentro do gráfico da função de pertinência.

A fórmula usada para o cálculo é a seguinte:

$$\mathbb{R}_{ij} = \sum_{j=1}^{k} w_j * r_{ij} \Big/ \sum_{j=1}^{k} w_j$$

Onde,

*wj* são os pesos *fuzzy* dos atributos;

*rij* é o grau de atendimento de cada atributo à característica avaliada; e

$\mathbb{R}_{ij}$ o grau de atendimento do ambiente a um padrão determinado.

Como exemplos de outros métodos de defuzificação podem ser citados o *BOA* (bissetor de área), *MOM* (valor médio dos máximos), o *SOM* (menor valor absoluto dos máximos) e o *LOM* (maior valor absoluto dos máximos) (COX, 1994; KANTROWITZ, HORSTKOTTE *et al.*, 1997; KLIR & YUAN, 1995; LEE, 1990).

#### *3.2.3.9 – Funções fuzzy*

Uma *função fuzzy* é uma extensão do conceito de uma função clássica *f*, podendo ser obtida através de diferentes "*graus*" de *fuzificação* (TURKSEN, 1991; ZIMMERMANN, 1991).

O *princípio da extensão* é utilizado para generalizar conceitos da matemática clássica para a teoria *fuzzy* (DUBOIS & PRADE, 1980; ZADEH, 1988). As funções *fuzzy* podem obedecer o seguinte mapeamento:

*i.* Mapeamento clássico de um conjunto *fuzzy*, que se realiza ao longo do processo de *fuzificação* do domínio da função, sendo que seu contradomínio seria nítido;

*ii.* Mapeamento *fuzzy* de um conjunto *fuzzy*, tornando seu contradomínio também *fuzzy*. Este processo é conhecido como "*fuzzifying* funções".

*iii.* Funções nítidas podem ter propriedades *fuzzy* ou estarem sujeitas a restrições *fuzzy*.

Dada uma função clássica $f$: $X \rightarrow Y$ e um domínio *fuzzy* $\tilde{A}$, em $X$, pelo princípio da extensão é gerada uma imagem *fuzzy* de $\tilde{B}$ com a função de pertinência (DUBOIS & PRADE, 1980).

$$\mu_{\tilde{B}}(y) = \sup_{x \in f^{-1}(y)} \mu_{\tilde{A}}(x)$$

### *3.2.3.10 – Agregação de conjuntos fuzzy*

A agregação é um processo utilizado em muitas tecnologias (YAGER, 1994), especialmente na tomada de decisão multicriterial (ZIMMERMANN, 1997). Têm sido propostas muitas alternativas para esse fim, como o processo de dedução e inferência lógica e outros tipos de conhecimento indutivo.

A idéia principal do processo de agregação é obter-se um grau de consenso entre as informações disponíveis, calculando-se um valor final. Se esses dados forem extraídos de especialistas, então se terá a taxa de aceitação ou rejeição entre eles, isto é, o grau pelo qual especialistas concordam em suas estimativas, tornando possível a elaboração de classificações das avaliações realizadas (KUNCHEVA & KRISHNAPURAM, 1996).

Os modelos de consenso são potencialmente férteis e podem ser usados em vários domínios de aplicação. Os métodos de consenso referem-se, principalmente, a esquemas de graduação, expandindo-se para relações de preferência, decisões de grupo e estimativas lingüisticamente definidas (DAY, 1989).

O grau de consenso tem sido estimado, usando-se alguns operadores de agregação *fuzzy*, através das preferências individuais de *n* especialistas, para cada atributo considerado, e gerando-se uma matriz de decisão, chamada em Kuncheva & Krishnapuram (1996) de 'perfil de decisão'.

Conceitualmente, operações de agregação *fuzzy* são combinações de vários conjuntos *fuzzy* (n ≥ 2), de uma forma desejável, para produzirem um único conjunto.

Em geral, uma operação de agregação é representada pela função:

$$h: [0,1]^n \rightarrow [0,1]$$

Quando a função *h* é aplicada para *n* conjuntos *fuzzy*, $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_n$, definidos em *X* e operando com os graus de pertinência de cada $x \in X$, é produzido um conjunto *fuzzy* agregado $\tilde{A}$. Assim, para cada $x \in X$:

$$\mu_{\tilde{A}}(x) = h\left(\mu_{\tilde{A}_1}(x),\ \mu_{\tilde{A}_2}(x), \dots,\ \mu_{\tilde{A}_n}(x)\right)$$

Yager (1988) desenvolveu a classe de *operações de média de pesos ordenados* ou *OWA* (*Ordered Weighted Averaging*) (YAGER R.R., 1988). Dado um vetor de pesos, $\boldsymbol{w}$, tal que $w_i \in [0,1]$ para todo $i \in \mathbb{N}_n$, então,

$$w = \langle w_1, w_2, \dots, w_n \rangle, \text{ e}$$

$$\sum_{i=1}^{n} w_i = 1$$

Em uma operação *OWA* associada ao vetor $\boldsymbol{w}$, onde $b_1$ é o maior *i*-ésimo elemento em $a_1, a_2, \dots, a_n$, para qualquer $i \in \mathbb{N}_n$, isto é, $\langle b_1, b_2, \dots, b_n \rangle$ é uma permutação do vetor $\langle a_1, a_2, \dots, a_n \rangle$, no qual os elementos estão ordenados: $b_i \geq b_j$, se $i < j$ para qualquer par $i, j \in \mathbb{N}_n$, então,

$$h_w(a_1, a_2, \dots, a_n) = w_1 b_1 + w_2 b_2 + \cdots + w_n b_n$$

*Exemplo 7*:

Dado $w = \langle 0{,}3;\ 0{,}1;\ 0{,}2;\ 0{,}4 \rangle$, tem-se que,

$$h_{\mathbf{w}}(0{,}6; 0{,}9; 0{,}2; 0{,}7) = 0{,}3 \times 0{,}9 + 0{,}1 \times 0{,}7 + 0{,}2 \times 0{,}6 + 0{,}4 \times 0{,}2 = 0{,}54$$

Em toda operação de agregação, os operadores *fuzzy* exercem uma grande influência no resultado final. Um número significativo de operadores foi proposto (DUBOIS, 1985; ZIMMERMANN, 1991), e a escolha apropriada desses operadores para uma determinada situação, é de grande relevância.

## 3.3 – Números *Fuzzy*

Em um processo de avaliação de resultados, os dados obtidos dos especialistas são geralmente imprecisos e contêm muitas ambigüidades, principalmente, em virtude de como foram capturados. A origem dessas ambigüidades pode ser (RÖMER & KANDEL, 1995):

- *A não acurácia dos dispositivos utilizados, envolvendo erros de medição de natureza fuzzy*;
- *A natureza lingüística dos dados observados*;
- *A natureza subjetiva dos dados obtidos*.

Muitas informações vagas podem ser convenientemente modeladas por números *fuzzy* (DUBOIS, 1985; KAUFMANN & GUPTA, 1991). O conceito de números *fuzzy* em incertezas *fuzzy* desempenha um papel semelhante ao de uma variável aleatória, nas relações de incertezas probabilísticas.

Um número *fuzzy* deve capturar a concepção intuitiva de números ou intervalos aproximados, tal como "*valores que estão próximos de um certo número real*", ou "*valores que estão em torno de um dado intervalo de números reais*". Tais conceitos são essenciais para a caracterização dos estados das variáveis *fuzzy* e, conseqüentemente, são importantes para aplicações tais como controle *fuzzy*, tomada de decisão, raciocínio aproximado e estatística. Casos especiais de números *fuzzy* incluem números e intervalos reais ordinários, como mostra a Figura 3-1:

*(a)* Um número real 3;
*(b)* um intervalo nítido [3, 4];
*(c)* um número *fuzzy* dado pela proposição "próximo a 3" (número *fuzzy* triangular);
*(d)* um número *fuzzy* com uma região plana (um intervalo *fuzzy* ou número *fuzzy* trapezoidal).

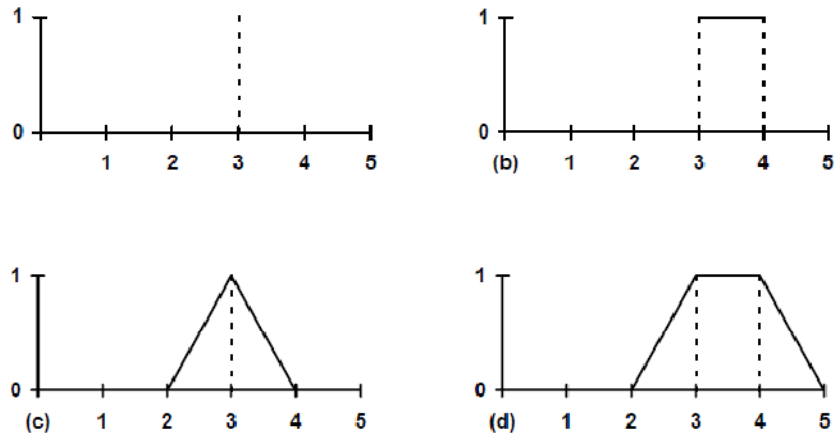



**Figura 3-1: Comparação de um número real e um intervalo nítido com um número *fuzzy* e um intervalo *fuzzy* respectivamente (Fonte: Klir & Yuan, 1995)**

Embora as funções de pertinência de números *fuzzy* tenham, usualmente, as formas triangular ou trapezoidal, existem outras formas para representá-las, que nem sempre são simétricas, dependendo do contexto da aplicação, como funções com "*forma de sino*", funções estritamente crescentes ou decrescentes.

Dados imprecisos podem ser modelados pelo significado de números *fuzzy L-R*, que é uma outra importante forma de representação de funções de pertinência desses números (DUBOIS & PRADE, 1980).

## 3.4 – Variáveis Lingüísticas

Uma variável lingüística é totalmente caracterizada por uma quíntupla

$(x, T(x), U, G, \widetilde{M})$.

Onde, o nome da variável é *x*;

O conjunto dos termos lingüísticos de *x* é *T(x)*, ou simplesmente *T*, que se refere a uma variável base u, cujos valores estão no conjunto universo *U*;

*G* é uma regra sintática, para a geração dos termos lingüísticos;

*M* é uma regra semântica, que associa a cada termo lingüístico $t \in T$ o seu significado, $\widetilde{M}(t)$, que é um conjunto *fuzzy* em *U* (ZIMMERMANN, 1991).

*Exemplo 8:*

Seja *X* uma variável lingüística identificada por "*Idade*" com $U = [0, 100]$ e seus termos lingüísticos, que também são conjuntos *fuzzy*, "*velho*", "*jovem*", "*muito*

*velho*", etc. A variável base *u* é a idade em anos de vida. $\widetilde{M}(t)$ é a regra que atribui um significado, isto é, um conjunto *fuzzy*, para esses termos.

$$\widetilde{M}(velho) = \{(u, \mu_{velho}(u)) | u \in [0,100]\}$$

Onde,

$$\mu_{velho}(u) = \begin{cases} 0 & \text{para } u \in [0,\ 50] \\ \left(1 + \left(\frac{u-50}{5}\right)^{-2}\right)^{-1} & \text{para } u \in [50,\ 100] \end{cases}$$

$T(x)$ define o conjunto de termos da variável *x*, que, nesse caso, é:

$T(idade) =$ {*velho*, *muito velho*, *não tão velho*, *mais ou menos jovem*, *inteiramente jovem*, *muito jovem*}

Onde, *G(x)* é uma regra que gera os rótulos dos termos no conjunto de termos, conforme a Figura 3-2 abaixo.

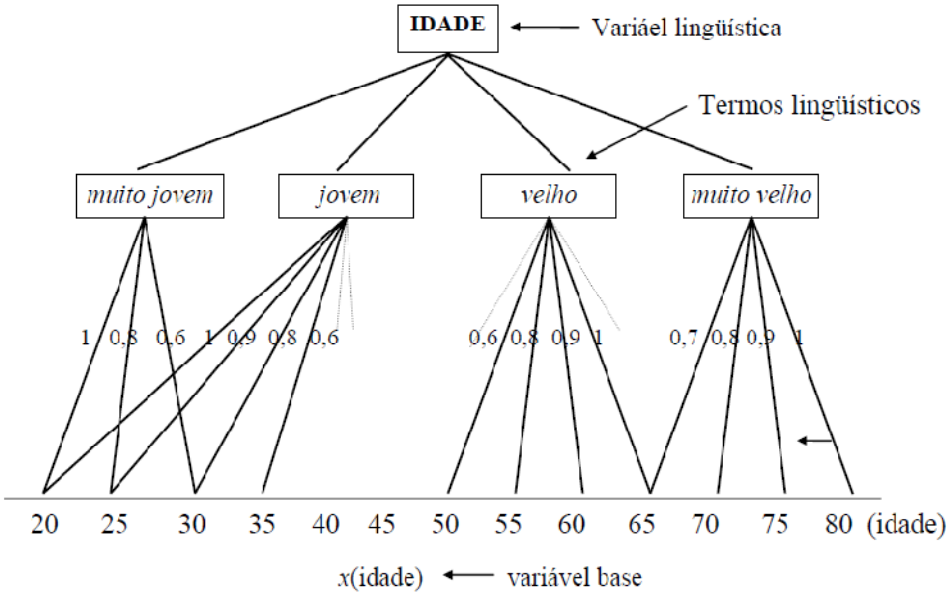



**Figura 3-2: Variável lingüística "Idade" (Fonte: Klir & Yuan, 1995)**

Portanto, cada variável lingüística, definida em termos de uma *variável base*, tem seu estado denotado por *termos lingüísticos*, que são interpretados como números *fuzzy* específicos. Os termos lingüísticos representam valores aproximados de uma variável lingüística, relacionados a uma aplicação particular. Uma variável base é uma variável no sentido clássico, exemplificada por uma variável física (temperatura,

pressão, velocidade) ou por uma variável numérica (desempenho, confiabilidade, probabilidade) (KLIR & YUAN, 1995).

Zadeh (1977) observou que os *termos lingüísticos* podem ser modelados através de funções, cujos valores são graus no domínio de uma *função de pertinência*, e que cada representação é fundamental para a modelagem do raciocínio aproximado (ZADEH, 1977).

*Limitadores lingüísticos* são termos lingüísticos especiais, que modificam outros termos lingüísticos como, por exemplo, *menos*, *muito*, *razoavelmente*, *fracamente* e *extremamente*. Qualquer limitador lingüístico pode ser interpretado como uma operação unária (ou modificadora), *h*, em um intervalo unitário [0, 1]. Por exemplo, o limitador *muito* é freqüentemente interpretado como uma operação unária $h(a) = a^2$, enquanto que *razoavelmente* é interpretado como $h(a) = a(a \in [0,1])$. Conhecendo-se o significado de um termo lingüístico e de sua operação modificadora, podem-se estabelecer *regras semânticas*, que traduzam o significado desse termo.

## 3.5 – A Lógica *Fuzzy*

A lógica *fuzzy* (difusa) é uma extensão da lógica (*booleana*) convencional, que lida com conceitos de verdade parcial - valores verdade entre '*completamente verdade*' e '*completamente falso*' - modelando as incertezas da linguagem natural (KANTROWITZ, HORSTKOTTE *et al.*, 1997). Nesse contexto, uma decisão, tida como correta, poderá ser alterada posteriormente, quando novas informações adicionais estiverem disponíveis.

A lógica *fuzzy* utiliza *predicados fuzzy* (*velho*, *raro*, *perigoso*, etc.), *quantificadores fuzzy* (muito, pouco, quase tudo, usualmente e semelhante), *valores-verdade fuzzy* (totalmente verdadeiro, mais ou menos verdadeiro, muito falso, etc.) (ZIMMERMANN, 1991).

Os quantificadores *fuzzy*, de particular interesse para os termos lingüísticos *fuzzy*, podem ser (BOSC, 1995):

*i.* Absolutos: definidos em $\mathbb{R}$ e denotados por um número, como, por exemplo, "no mínimo 7";

*ii.* Relativos: definidos no intervalo [0,1], referindo-se a uma proposição quantificada lingüisticamente como, por exemplo, "*muitos especialistas são convincentes*".

Uma proposição quantificada lingüisticamente pode ser apresentada como (KACPRZYK, 1992):

*Q y's são F*,

*Q* é um *quantificador lingüístico* (por exemplo, *muitos*), $Y = \{y\}$ é um *conjunto de objetos* de um conjunto universo (por exemplo, *especialistas*), e *F* é uma *propriedade*, isto é, um predicado *fuzzy* (por exemplo, *convincentes*).

Pode-se atribuir a *y* particulares (objetos) uma importância diferente (relevância, competência, ...) *B*, e adicioná-la à equação anterior, gerando:

*Qby's são F*.

No caso da importância (um predicado *fuzzy*) ser adicionada, *B* é definido como um conjunto *fuzzy* em *Y* e $\mu_{\tilde{B}}(y_i) \in [0,1]$ é um grau de pertinência de $y_i$. Assim sendo, o valor de pertinência $\mu_{\tilde{B}}(y_i)$ também é definido como a possibilidade de *B* assumir o valor $y_i$. Por esse motivo, a lógica *fuzzy* é chamada também de teoria da possibilidade.

A *teoria da possibilidade (ZADEH, 1978)* está, intrinsecamente, ligada à linguagem natural, onde a "*possibilidade*" é melhor interpretada do que a "*probabilidade*". No entanto, a possibilidade não substitui a probabilidade – ambas lidam com incertezas e se complementam entre si. Observa-se que um alto grau de possibilidade não implica um alto grau de probabilidade, todavia se um evento não é possível, também é improvável, isto é, a possibilidade é o limite superior da probabilidade (ZIMMERMANN, 1991).

Várias aplicações da teoria *fuzzy* envolvem o uso de uma base de regras *fuzzy* para modelos complexos, como os sistemas de controle de lógica *fuzzy*, os sistemas de controle *fuzzy* multivariável (GEGOV.A.E, 1995) e os sistemas adaptativos (GÜRMAN, 1995).

Muitos sistemas de controle *fuzzy* (KAUFMANN & GUPTA, 1991) são baseados em regras e se aplicam, quase que exclusivamente, a sistemas de produção e controle (DUBOIS & PRADE, 1989). Geralmente, essas regras não são extraídas de especialistas humanos através do sistema, mas produzidas, explicitamente, por seus projetistas.

A idéia básica é incorporar a "*experiência*" de um processo executado por um operador humano ao projeto de um controlador. De um conjunto de regras lingüísticas, que descrevem a estratégia de controle dos operadores, um algoritmo de controle é construído, e seus termos são definidos como conjuntos *fuzzy*.

Portanto, a lógica *fuzzy* simula o pensamento humano, incorporando sua inerente falta de precisão a um sistema físico. A incerteza é o resultado da imprecisão não aleatória de medições ou da impossibilidade de se obter uma descrição numérica exata para quantidades observadas (NICULESCU & VIERTL, 1992).

## 3.6 – Sistemas Baseados em Conhecimento *Fuzzy*

Várias propostas têm sido publicadas sobre como se usar a teoria *fuzzy* em SBC (Sistemas Baseados em Conhecimento). Essas propostas têm, em comum, a preocupação com que esses sistemas sejam providos com mecanismos que manuseiem incertezas, otimizando o seu desempenho (FICKAS & HELM, 1992; HULL, 1991).

A incerteza das informações na base do conhecimento, por exemplo, induz incertezas nas conclusões. Assim sendo, a máquina de inferência deve ser equipada com capacidade computacional, para analisar a transmissão de incertezas das premissas para as conclusões, associando-as a algumas medidas de incertezas, que sejam inteligíveis e interpretadas convenientemente pelo usuário (ZIMMERMANN, 1991).

Um dos principais benefícios de um SBC *fuzzy* é sua habilidade de usar e assimilar o conhecimento de múltiplos especialistas, outorgando-lhes uma expressividade não existente em sistemas convencionais de suporte à decisão (COX, 1994). Tanto os SBCs *fuzzy*, quanto os sistemas de controle *fuzzy*, objetivam modelar a experiência e o ambiente humano para tomada de decisões (CARCHIOLO & MALGERI, 1995; ZADEH, 1973).

## 3.7 – Sistemas *Fuzzy*

Recentemente, a área de pesquisa apresentou também "o cálculo *fuzzy*", "equações diferenciais *fuzzy*", "sistemas *fuzzy*", "lógica *fuzzy* com aplicações da engenharia", e assim por diante (COX, 1994; KANTROWITZ, HORSTKOTTE *et al.*, 1997; KLIR & YUAN, 1995).

**Tabela 3-2: Categorias genéricas de aplicações de sistemas *fuzzy* (Fonte: Munakata & Ani, 1994)**

| Categorias | Aplicações |
|---|---|
| Controle | utilizado em larga escala em aplicações industriais |
| Reconhecimento de Padrões | processamento de imagem, áudio e sinal |
| Análise Quantitativa | pesquisa operacional, estatística e gerenciamento |
| Inferência | sistemas especialistas para diagnóstico, planejamento e predição |
| | processamento de linguagem natural |
| | interfaces inteligentes |
| | robôs inteligentes |
| | engenharia de software |
| Recuperação de Informações | base de dados |

Com o emprego de sistemas *fuzzy* (COX, 1994; YAGER, 1994) de forma apropriada, julga-se ter produzido respostas mais rápidas e "suaves" que os sistemas convencionais. A maneabilidade, a robustez e, sobretudo, o baixo custo são fatores de qualidades característicos dos sistemas *fuzzy*, contribuindo para um melhor desempenho desses. São úteis para problemas ou aplicações complexas, que envolvam descrições humanas ou pensamento indutivo. A Tabela 3-2 mostra suas principais áreas de aplicação.

Os principais passos para o desenvolvimento de um sistema *fuzzy* são (MUNAKATA & ANI, 1994):

*i.* Constatação de que o conhecimento sobre o ambiente da aplicação em questão é descrito de forma aproximada ou com regras heurísticas;

*ii.* Identificação das entradas e saídas do sistema e seus respectivos intervalos de valores;

*iii.* Definição de uma função para cada parâmetro de entrada ou saída. A quantidade de funções requeridas depende do desenvolvedor e do ambiente do sistema;

*iv.* Construção de uma base de regras pelo projetista, que determine quantas regras serão necessárias e quando parar de adicionar novas regras; e

*v.* Verificação das saídas da base de regras e se seus intervalos de valores estão corretos e em conformidade com o conjunto de entradas usado, permitindo uma posterior validação.

Vários métodos têm sido utilizados para o desenvolvimento e a implementação de sistemas *fuzzy*. Os mais comuns são: o método de desenvolvimento baseado em especialistas humanos e o método de desenvolvimento baseado em tentativa e erro.

Os principais problemas e limitações de sistemas *fuzzy* são (MUNAKATA & ANI, 1994):

- *Estabilidade*: não há garantias teóricas de que um sistema *fuzzy* não venha a atingir um estado caótico e que seja estável. No entanto, a experiência tem mostrado que uma expressiva maioria dos sistemas *fuzzy* são estáveis.
- *Capacidade de aprendizagem*: falta-lhes a capacidade de aprendizagem, não possuindo um conhecimento previamente especificado. Atualmente, sistemas híbridos, particularmente os sistemas *neurofuzzy*, vêm suprindo esta limitação.
- *Determinação ou refinamento de funções e regras fuzzy* apropriadas: mesmo depois de muitos testes, pode ser ainda difícil o estabelecimento conveniente das funções *fuzzy* requeridas.
- *Conceituação*: há ainda um entendimento errôneo do termo *fuzzy*, como significando imprecisão ou imperfeição e como se não tivesse uma fundamentação matemática sedimentada.
- *Verificação e validação*: geralmente requerem testes extensivos. No entanto, os benefícios intrínsecos na utilização de sistemas *fuzzy* excedem as suas limitações, e seus problemas podem ser sumariados na redução do custo do desenvolvimento, da execução e da manutenção da aplicação (COX, 1994).

# Capítulo 4 – Abordagem Proposta para o Prognóstico de Qualidade de Informações na *Web*

Em razão do que foi ressaltado nos capítulos anteriores, a qualidade de informação na *Web* é, portanto, um importante aspecto a ser investigado para auxílio aos usuários em suas buscas por melhores informações. Este capítulo apresenta a abordagem proposta para o prognóstico de qualidade de informações da *Web*.

## 4.1 – Modelo Proposto

No contexto desta tese, o uso de um modelo agrega os benefícios citados na seção 2.7.1, uma vez que serão documentados os conceitos relevantes referentes ao domínio – qualidade de informações na *Web* baseada em metatados –. As restrições aos conceitos também serão explicitadas no modelo mostrado na Figura 4-1, adiante descrito e formalizado. Baseado nesse modelo, os documentos *Web* recuperados na busca são avaliados, considerando os seus metadados, o contexto semântico e as perspectivas dos usuários. Ele foi elaborado por meio de um diagrama UML no qual foram definidos os termos e os relacionamentos entre esses termos e, adicionalmente, um conjunto de funções *fuzzy* de transformação e de pertinência, responsável por manter algumas restrições semânticas identificadas. A Tabela 4-1 descreve as restrições de integridade que foram identificadas durante a modelagem.

**Tabela 4-1: Restrições de Integridade do Modelo**

| Restrições de Integridade | Descrição |
|---|---|
| **RI01** | A participação das instâncias da *Classe Metadado* nas *Classes* especializadas *Metadado Original* e *Metadado Derivado* é Completa e Disjunta. |
| **RI02** | Um *Especialista* só poderá atribuir *graus de importância* às *Dimensões de Qualidade* nos *Contextos* de sua *especialidade*. |

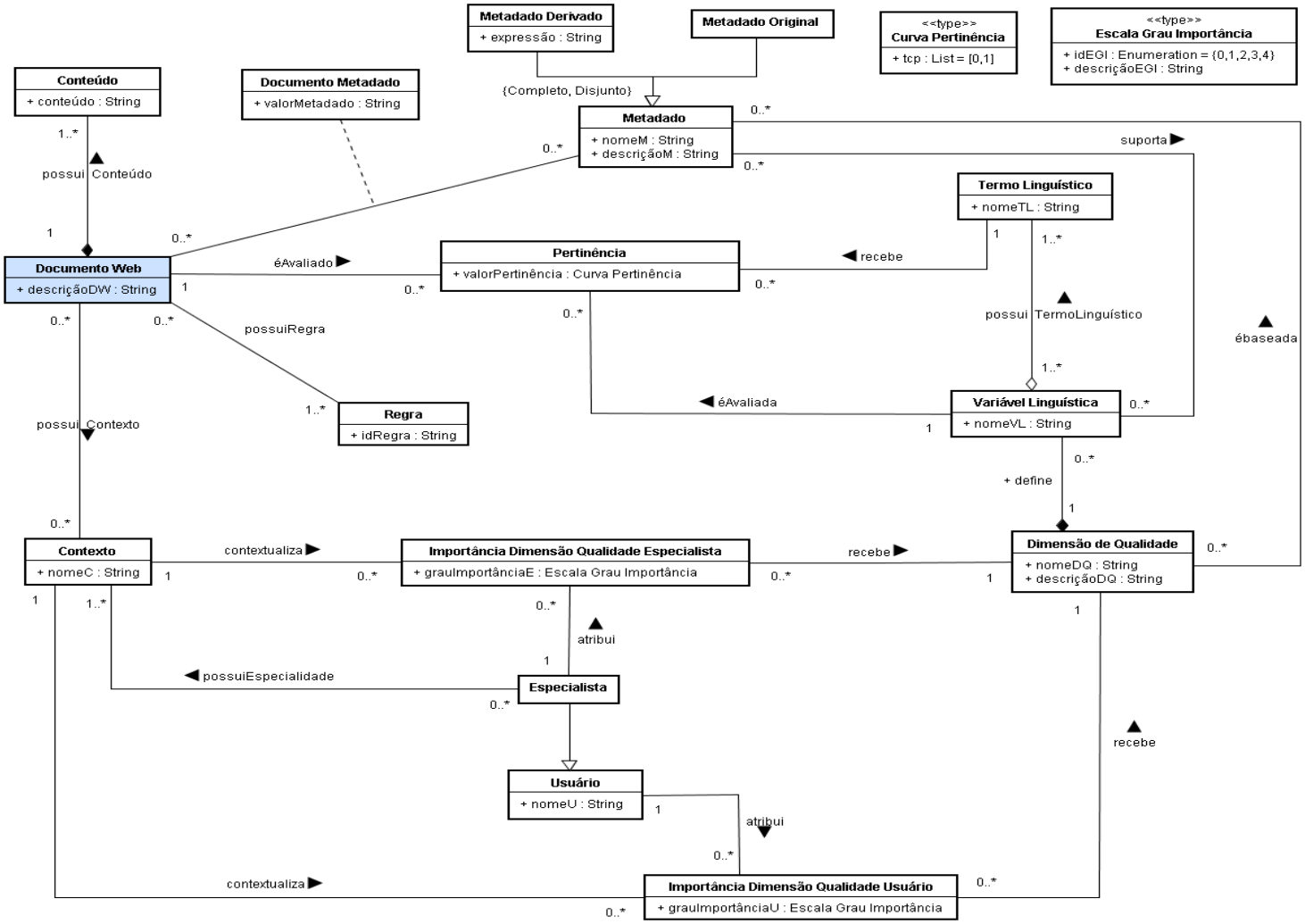



**Figura 4-1: Modelo para o Prognóstico de QI na *Web***

Com o intuito de facilitar a descrição e a leitura, a Figura 4-2 segmenta o modelo em pacotes de classes e suas dependências. Esses pacotes são detalhados a seguir.

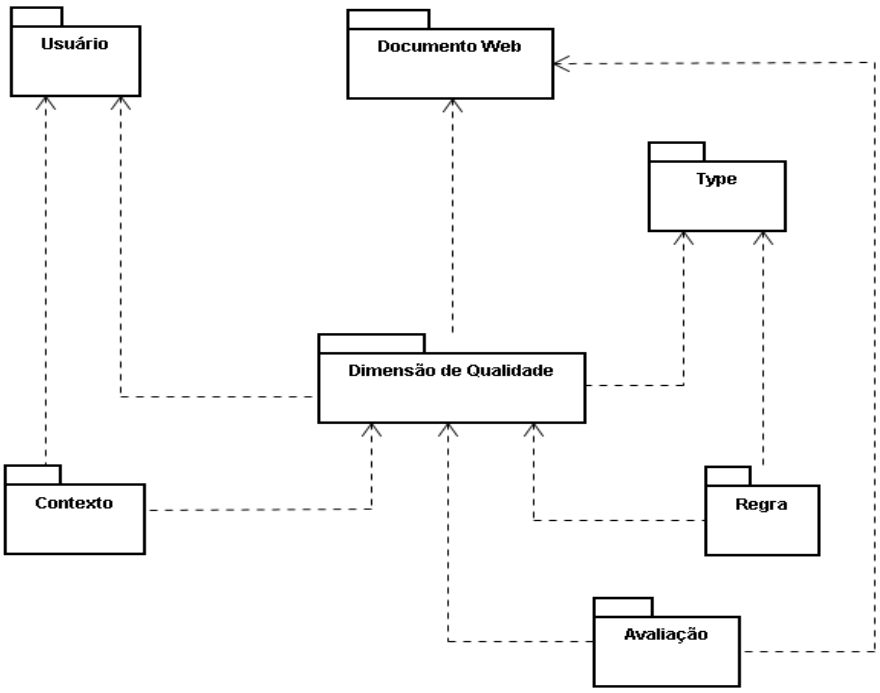



**Figura 4-2: Modelo para o Prognóstico de QI na *Web* Organizado em Pacotes**

### 4.1.1 – Pacote Documento *Web*

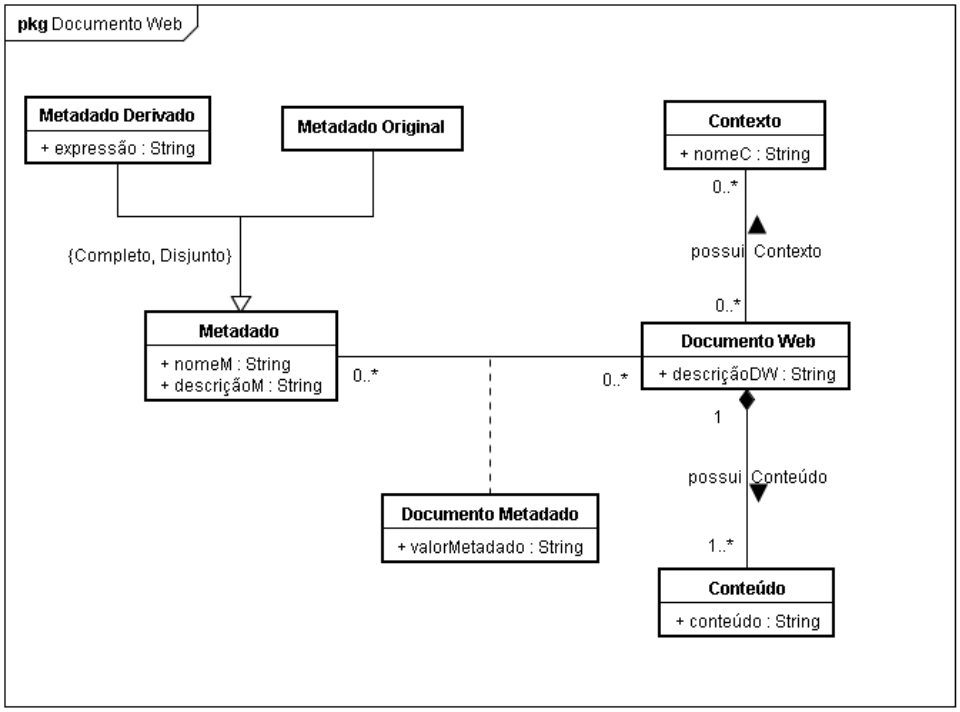



**Figura 4-3: Pacote Documento *Web***

#### 4.1.1.1 – Classe Documento Web

Essa classe define o conjunto de páginas *Web* recuperadas na busca, a ser avaliado. Ela é composta pelo seu conteúdo e possui os metadados e os contextos a ela associados.

**Definição 1**. A *Classe Documento Web* é representada pelo conjunto:

$$WebDoc = \{webdoc_1, webdoc_2, ...webdoc_n\}$$

Onde cada $webdoc_i, 1 \leq i \leq n$, é uma instância de um documento *Web*.

#### 4.1.1.2 – Classe Conteúdo

Essa classe define o conjunto dos conteúdos dos documentos *Web* (*Associação possuiConteúdo*) em seus aspectos intrínsecos de representação. Em geral os conteúdos, predominantemente desestruturados encontrado em documentos HTML, consistem de textos, imagens e outros componentes.

**Definição 2.** A *Classe Conteúdo* é representada pelo conjunto:

$$T = \{Cont_1, Cont_2, ...Cont_n\}$$

Onde cada $Content_i, 1 \leq i \leq n$, é uma instância de conteúdo.

#### 4.1.1.3 – Classe Metadado

Essa classe define o conjunto dos tipos de metadados *Web* usado como base para a avaliação da qualidade da informação. Os metadados originais são recuperados "*as-is*" juntamente com o documento *Web*, enquanto os metadados derivados são obtidos por funções da transformação, a partir de outros metadados. Cada um desses tipos de objetos criados é inserido na classe correspondente do modelo. Cada elemento possui um nome específico. Desse modo, ao ser criado um novo elemento, é verificado se ele já existe. Em caso afirmativo, seus dados podem ser redefinidos, mas o objeto não é recriado.

Uma *restrição de propriedade* dessa abstração define que a participação das instâncias da *Classe Metadado* nas *Classes* especializadas *Metadado Original* e *Metadado Derivado* é Total e Exclusiva (*RI01* na Tabela 4-1).

**Definição 3**. Os metadados de um documento (*Classe de Associação Documento Metadado*) são as informações sobre o documento e sobre os dados do documento, isto é, o seu conteúdo. O atributo *valorMetadado* expressa os seus valores.

A *Classe Metadado* é uma generalização das Classes *Metadados Originais* e *Metadados Derivados* e é representada pelo conjunto:

$$M = \{m_1, m_2, ...m_n\}$$

Onde cada $m_i, 1 \leq i \leq n$, é uma instância de metadado.

**Definição 4.** Os *metadados originais* são metadados que podem ser recuperados diretamente de um documento ou por meio de serviços disponíveis na *Web*, usando o documento como chave de busca.

A *Classe Metadado Original* é uma especialização da *Classe Metadado* e é representada pelo conjunto:

$$OM = \{ om_1, om_2, ...om_n\}$$

Onde cada $om_i, 1 \leq i \leq n$, é uma instância de um metadado original. Um exemplo simples de metadado é *data de atualização.*

**Definição 5**. Os *metadados derivados* são os metadados que não podem ser diretamente recuperados de um documento, ou por meio de serviços disponíveis na *Web*. Em vez disso, eles são derivados de outros metadados, por intermédio de algumas funções de transformação (*atributo expressão*). A notação adotada neste trabalho foi *m* para representar *metadado*, *om* para representar *metadados originais* e *dm* para representar os *metadados derivados*, indexados quando necessário. Para representar uma função que calcule o valor de um *metadado derivado* específico $dm_i$ foi usada a notação $fd_{(dm_i)}$. A Tabela 4-2, mostrada mais adiante, exemplifica algumas dessas funções.

A *Classe Metadado Derivado* é uma especialização da *Classe Metadado* e representada pelo conjunto:

$$DM = \{dm_1, dm_2, ...dm_n\}$$

Onde cada $dm_i, 1 \leq i \leq n$, é uma instância de um metadado derivado.

Também é definido: $M = OM \bigcup DM$ e $OM \bigcap DM = \phi$

A Tabela 4-2 fornece como exemplos os metadados originais e metadados derivados usados nesta tese. No primeiro exemplo (1), *Tempo de Atualização* é derivado dos metadados originais: *Data de Atualização* e *Data da Consulta*. O segundo (2) e terceiro (3) exemplos adotam os conceitos de *Autoridade* e de *Centralidade* (KLEINBERG, 1998) que são explorados mais adiante.

**Tabela 4-2: Metadados originais e funções de derivação**

| **Metadados originais** | **Funções para obtenção dos metadados derivados** $(\boldsymbol{fd_{(dm_i)}})$ |
|---|---|
| $\boldsymbol{ud_i}$ – data de atualização de um documento $Web_i$<br>$\boldsymbol{qt_i}$ – data da consulta de um documento $Web_i$<br>$\boldsymbol{BL_i}$ – número de *links* que apontam para o documento $Web_i$ em um contexto<br>$\boldsymbol{FL_i}$ – número de *links* externos do documento $Web_i$ em um contexto | **(1)** *Tempo desde a atualização (UT)* = $ut_i = qt_i - ud_i$<br>**(2)** *Autoridade (Authority)* = $a_i = \sum h_j, \quad onde: j \in BL_i$<br>**(3)** *Centralidade (Hub)* = $h_i = \sum a_j, \quad onde: j \in FL_i$<br>O cálculo de *authorities* e *hubs* considera um conjunto *S* de documentos em um contexto. É um processo iterativo onde todos os pesos são inicializados como 1. Posteriormente, os pesos de *hub* e *authority* são calculados e os resultados são normalizados. Esse processo é repetido até os valores *a* e *h* convergirem em todos os documentos. Adotamos o JUNG[33] para a obtenção dos valores de *hubs* e *authorities*. |

#### *4.1.1.4 – Classe Contexto*

Essa classe define o conjunto dos contextos que são recuperados juntamente com o documento *Web*, da mesma forma que os metadados (*Associação possuiContexto*). Eles especificam o escopo ou os limites de uma área de estudo ou de uma discussão, ou mesmo um domínio de conhecimento, de acordo com as definições da seção 2.4 do capítulo 2.

**Definição 6**. A *Classe Contexto* é representada pelo conjunto:

$$C = \{c_1, c_2, ...c_n\}$$

[33] http://jung.sourceforge.net/.

Onde cada $c_i, 1 \leq i \leq n$, é um nome associado a um contexto. Exemplos simples de contextos são *Economia* e *Matemática*.

### 4.1.2 – Pacote Dimensão de Qualidade

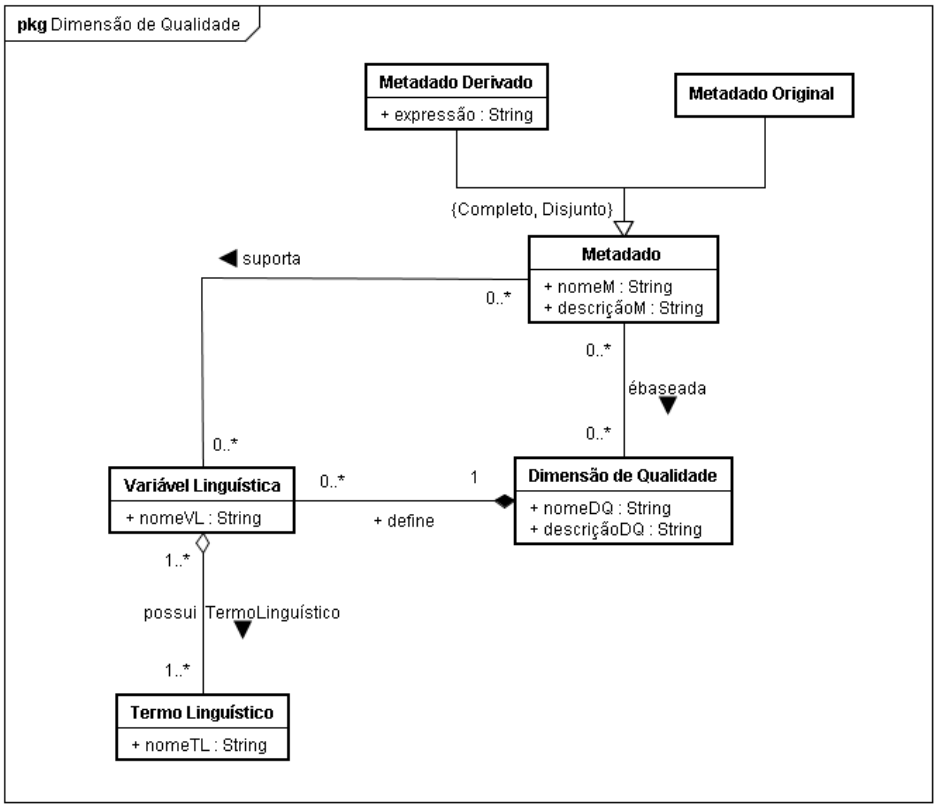



**Figura 4-4: Pacote Dimensão de Qualidade**

#### *4.1.2.1 – Classes Dimensão de Qualidade*

A *Classe Dimensão de Qualidade* define o conjunto das dimensões de qualidade. A dimensão de qualidade é percebida como uma perspectiva do usuário sobre a qualidade do documento. Elas são, de fato, os critérios e os fatores adotados na avaliação de qualidade da informação e servem de base para identificação das *variáveis lingüísticas* (*Associação define*).

**Definição 7**.. A *Classe Dimensão da Qualidade* é representada pelo conjunto:

$$DQ = \{dq_1, dq_2, ...dq_n\}$$

Onde cada $dq_i, 1 \leq i \leq n$ é uma instância de dimensão de qualidade. Um exemplo simples de dimensão de qualidade é a *Completeza*.

É essencial, neste momento, esclarecer um aspecto importante da modelagem. Enquanto os metadados descrevem os documentos *Web* e seus conteúdos, – que podem ser obtidos, diretamente ou indiretamente, pelos seus usuários – as dimensões de qualidade descrevem as perspectivas dos usuários sobre a qualidade esperada de um documento, independentemente da possibilidade de calculá-las.

Além disso, o modelo presume a viabilidade de associação direta entre uma dimensão da qualidade e um metadado específico (*Associação adota*).

#### *4.1.2.2 – Classe Variável Lingüística e Classe Termo Lingüístico*

A *Classe Variável Lingüística* define o conjunto das variáveis lingüísticas identificadas a partir da *Classe Dimensão de Qualidade* (*Associação define*).

**Definição 8**. A *Classe Variável Lingüística* é representada pelo conjunto:

$$VL = \{vl_1, vl_2, ...vl_n\}$$

Onde, cada $vl_i, 1 \leq i \leq n$ é uma instância de variável lingüística.

A *Classe Termo Lingüístico* define o conjunto dos adjetivos relacionados às variáveis lingüísticas (*Associação possuiTermoLingüístico*). Conseqüentemente, a associação entre as *Classes Variável Lingüística* e *Metadado* (*Associação suporta*) indica que uma variável lingüística pode ser representada, com algum grau de incerteza, a partir dos valores dos metadados. A operação de fuzificação para transformação dos valores dos metadados em instâncias de termos lingüísticos é uma função de pertinência dos conjuntos *fuzzy*.

**Definição 9**. A *Classe Termo Lingüístico* é representada por um conjunto:

$$TL = \{tl_1, tl_2, ...tl_n\}$$

Onde, cada $tl_i, 1 \leq i \leq n$ é uma instância de um termo lingüístico.

### 4.1.3 – Pacote *Type*

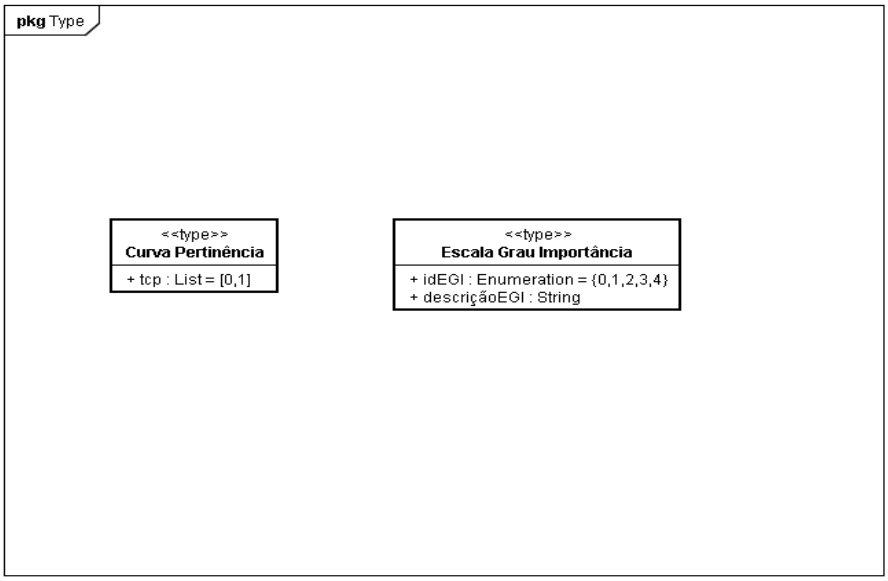



**Figura 4-5: Pacote *Type***

#### *4.1.3.1 – Classe Tipo Curva Pertinência*

A *Classe Tipo Curva Pertinência* estabelece uma curva de distribuição de pertinência ao longo de um intervalo dentro da faixa de [0,1]. Essa forma de representação permite que as pertinências dos termos lingüísticos possam ser usadas no domínio de qualquer variável lingüística.

#### *4.1.3.2 – Classe Escala Grau Importância*

A *Classe Escala Grau Importância* estabelece uma lista enumerada de graus de importância das dimensões de qualidade num contexto específico {0,1,2,3,4}. Essa forma de representação permite que os graus de importância sejam atribuídos às dimensões de qualidade pelos especialistas e usuários.

### 4.1.4 – Pacote Avaliação

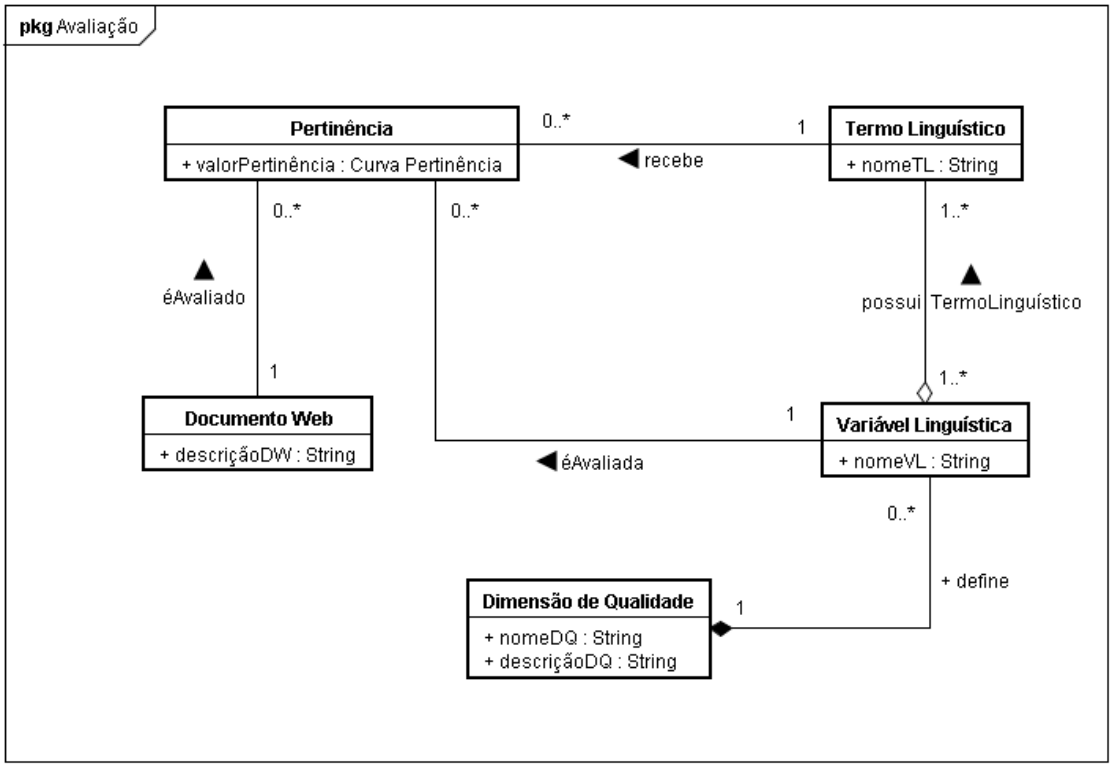



**Figura 4-6: Pacote Avaliação**

#### *4.1.4.1 – Classe Pertinência*

Ao avaliar um documento *Web* de acordo com uma dimensão da qualidade, o modelo possibilita uma interpretação lingüística dos valores dos metadados respectivos para esse documento (*Associação éAvaliado*). Para cada documento, o atributo *valorPertinência* da *Classe Pertinência*, expressa o valor *fuzzy* do termo lingüísticos de uma variável lingüística que foi calculada pelas funções *fuzzy* de pertinência. O atributo *valorPertinência* é um atributo do *Tipo Curva Pertinência*.

**Definição 10**. A dimensão de qualidade $dq_i$ pode ser representada com um conjunto de *variáveis lingüísticas* (*Classe Variável Lingüística*) que tem a possibilidade de assumir os valores *fuzzy* aplicáveis aos seus termos lingüísticos (*Classe Pertinência*): $tl_{ij}$ .

Onde *i* é uma variável lingüística específica (*Associação éAvaliada*) e *j* é um termo lingüístico específico (*Associação recebe*).

**Definição 11**. A *Classe Pertinência* é representada pelo conjunto:

$$P = \{(webdoc_k, vl_i, tl_{ij}, vp_{kij}), \ldots (webdoc_n, vl_i, tl_{ij}, vp_{nij})\}$$

Onde $vp_{kij}$ é o valor de pertinência de um termo lingüístico *j* específico de variável lingüística *i* específica de um $webdoc_k$, e $1 \leq k \leq n$.

### 4.1.5 – Pacote Usuário

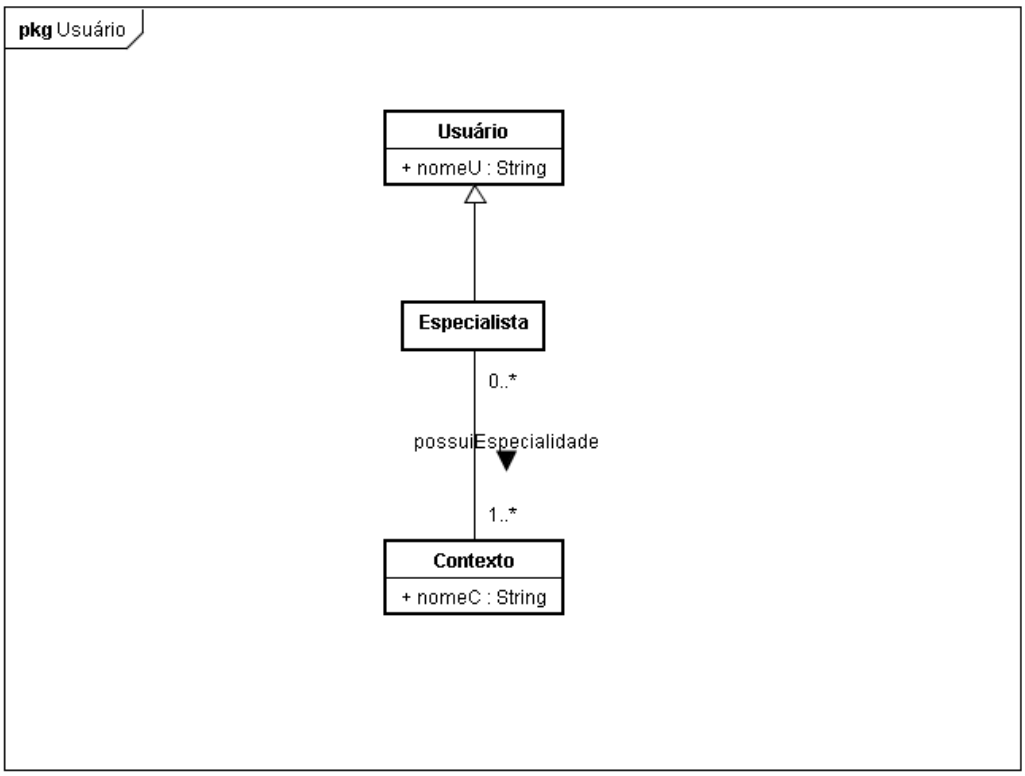



**Figura 4-7: Pacote Usuário**

#### *4.1.5.1 – Classes Usuário e Especialista*

A *Classe Usuário* define o conjunto de todos os usuários, dentre eles os especialistas.

A notação adotada neste trabalho foi *e* para representar os especialistas, e *u* para representar usuários, indexados quando necessário.

**Definição 12**. A *Classe Usuário* é representada pelo conjunto:

$$U = \{u_1, u_2, \ldots u_n\}$$

Onde cada $u_i, 1 \leq i \leq n$, é uma instância de um usuário.

**Definição 13.** A *Classe Especialista* é representada pelo conjunto:

$$E = \{e_1, e_2, \ldots e_n\}$$

Onde cada $e_i, 1 \leq i \leq n$, é uma instância de um especialista. A *Associação possuiEspecialidade* denota as suas especialidades. Exemplos simples de especialistas são *Economistas* e *Matemáticos*.

### 4.1.6 – Pacote Importância Dimensão de Qualidade

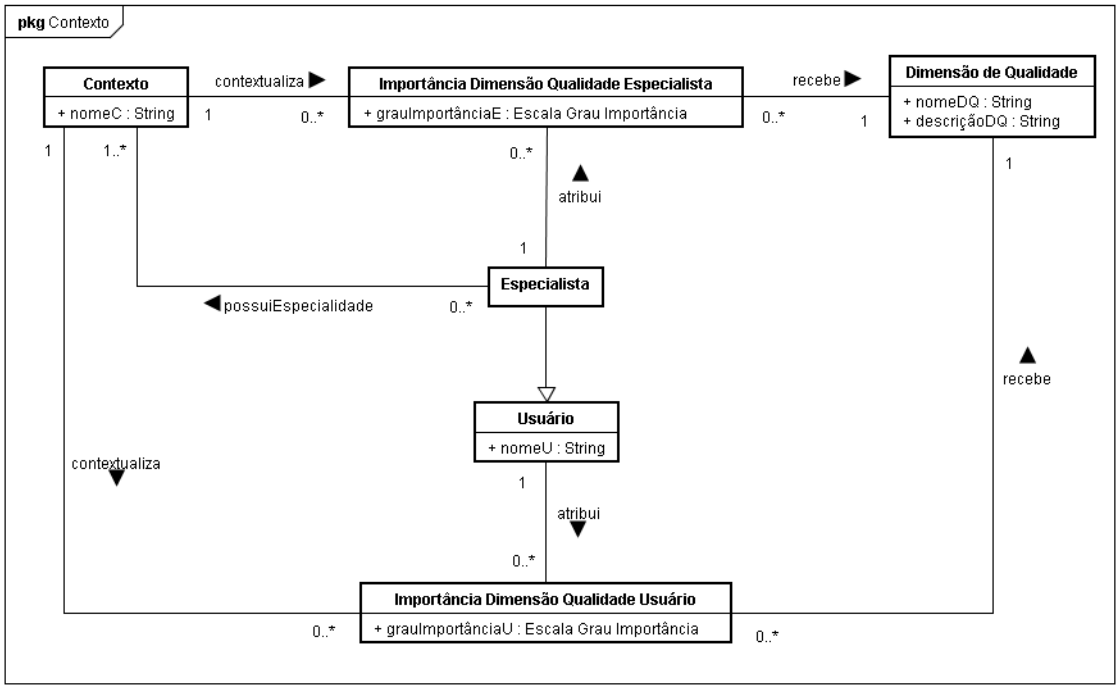



**Figura 4-8: Pacote Importância Dimensão de Qualidade**

#### *4.1.6.1 – Classes Importância Dimensão Qualidade Especialista e Importância Dimensão Qualidade Usuário*

Entendendo que as dimensões de qualidade possuem importâncias diferenciadas dependendo do contexto, os especialistas atribuem graus de importância para cada dimensão da qualidade, considerando um contexto específico. Essa atribuição é representada pelas associações *atribui*, *recebe* e *contextualiza* da *Classe Importância Dimensão Qualidade Especialista*. O atributo *grauImportânciaE* é um atributo do *Tipo Escala Grau Importância* e expressa o grau de importância atribuído pelo especialista.

Existe uma *restrição de integridade* dessa representação entre as *Classes Importância Dimensão Qualidade Especialista e a Associação atribui* (*RI02* na Tabela 4-1).

Como resultado dessa atribuição, é definido um vetor de importâncias para as dimensões de qualidade selecionadas. Esse vetor é usado na fase de defuzificação para ponderação e ajustes das avaliações das dimensões de qualidade.

A Tabela 4-3 mostra a escala dos graus de importância a serem usados pelos especialistas e usuários.

**Tabela 4-3: Escala de Graus de Importância das Dimensões de Qualidade por Contextos (Adaptada de Xexéo, 1996)**

| Graus | Descrição |
|---|---|
| 0 | Indica que a dimensão de qualidade avaliada não tem nenhuma importância. |
| 1 | Indica que a dimensão de qualidade avaliada tem pouca importância. |
| 2 | Indica que a dimensão de qualidade avaliada é importante em algumas circunstâncias, mas nem sempre. |
| 3 | Indica que a dimensão de qualidade avaliada é muito importante. |
| 4 | Indica que a dimensão de qualidade avaliada é essencial. |

**Definição 14**. O vetor de importâncias de dimensões de qualidade por contextos é representado por:

$$CDQ = \langle w(c, dq_1) \quad w(c, dq_2) \quad ... \quad w(c, dq_n) \rangle$$

Onde *c* é um contexto específico;

$dq_n$ é uma dimensão de qualidade;

$w(c, dq_n)$ é o grau de importância atribuído por especialistas a uma $dq_n$ em um *c* específico. Um exemplo simples é atribuir grau "3" para a dimensão de qualidade *atualidade* no contexto *Matemática.*

$CDQ$ representa o vetor de importâncias para cada $dq_n$ em *c*, atribuído pelos especialistas.

Além da contextualização considerada pelos especialistas, a perspectiva do usuário também é considerada. Eles atribuem o grau de importância de cada dimensão de qualidade, similarmente ao procedimento anterior, agora levando em conta seus requisitos específicos. Essa atribuição é representada pelas associações *atribui*, *recebe* e *contextualiza* da *Classe Importância Dimensão Qualidade Usuário*. O atributo

*grauImportânciaU* é um atributo do *Tipo Escala Grau Importância* e expressa o grau de importância atribuído pelo usuário.

Assim, é definido um vetor de importâncias para as dimensões de qualidade selecionadas pelos usuários. Esse vetor também será usado, mais tarde, na fase de defuzificação para ponderação e ajustes das avaliações das dimensões de qualidade.

**Definição 15**. O vetor de importâncias dos usuários é representado por:

$$UDQ = \langle w(c, dq_1) \quad w(c, dq_2) \quad ... \quad w(c, dq_n) \rangle$$

Onde $UDQ$ representa o vetor de importâncias para cada $dq_n$ em $c$, atribuído pelos usuários.

Considerando uma possível dificuldade em encontrar usuários especialistas para definir $CDQ$ para determinados contextos, a adoção de múltiplos $UDQ$ para compor $CDQ$ é percebida como uma solução alternativa.

As definições apresentadas até aqui tratam da obtenção dos resultados da avaliação, separadamente, por documento *Web* e por cada um dos termos lingüísticos que estejam relacionados a cada uma das variáveis lingüísticas consideradas. Esses resultados são denominados *PS – Prognóstico Singular de Qualidade de Informação*. Por exemplo, existem três *PS* ( $\mu_{\tilde{B}}(ut_i), \mu_{\tilde{R}}(ut_i)$ e $\mu_{\tilde{G}}(ut_i)$) para cada documento *Web*, referentes aos termos lingüísticos *ruim (bad), regular (regular)* e *bom (good)* relativos à variável lingüística *atualidade*.

Esses *PS* são os conjuntos *fuzzy* de entrada e a partir deles é calculado, para cada documento *Web*, o *PC – Prognóstico Composto de Qualidade de Informação*. A operação para obter o *PC* a partir dos *PS* é uma função de agregação *fuzzy* e o *PC* representa os conjuntos *fuzzy* da saída.

É importante ressaltar que a abordagem da lógica *fuzzy* permite tratar as composições das variáveis lingüísticas a partir de diferentes conceitos.

Em uma composição, todos os subconjuntos *fuzzy* relacionados a cada variável de saída são combinados para formar os subconjuntos *fuzzy* compostos para cada uma dessas variáveis. Usualmente, as inferências MAX (máximo) e SUM (soma) podem ser usadas.

Na composição MAX, o subconjunto *fuzzy* de saída combinado é construído tomando-se o grau máximo de pertinência sobre todos os subconjuntos *fuzzy* atribuídos à variável pela regra de inferência. Na composição SUM (soma), o subconjunto *fuzzy* de saída combinado é construído somando-se os graus de pertinência de todos os subconjuntos *fuzzy* atribuídos à variável de saída pela regra de inferência (KANTROWITZ, HORSTKOTTE *et al.*, 1997).

Na nossa proposta foi adotada a média aritmética ponderada pelos graus de importância atribuídos pelos usuários e especialistas para obter os subconjuntos *fuzzy* de saída. A definição 16 descreve esse conceito.

**Definição 16.** A composição de valores formada pelos subconjuntos *fuzzy* é representada por:

$$PC_j = \left[\frac{\sum_{i=1}^{n}\left(cqd_i * uqd_i * PS_{ij}\right)}{\sum_{i=1}^{n}(cqd_i * uqd_i)}\right]$$

Onde *j* é um termo lingüístico específico. No exemplo *ruim, regular e bom*;

*i* é uma variável lingüística específica, no exemplo *atualidade, reputação e completeza*;

$cdq$ representa o grau de importância de cada $dq_n$ num contexto específico *c*, estimado por especialistas;

$udq$ representa o grau de importância de cada $dq_n$ estimado pelos usuários;

$PS_{ij}$ representa os resultados da avaliação, separadamente, por documento *Web* e para um termo lingüístico *j* que esteja relacionado a uma variável lingüística *i*;

$PC_j$ representa o subconjunto *fuzzy* de saída para um termo lingüístico específico *j*.

Há ainda a defuzificação, que é usada quando é útil converter um conjunto *fuzzy* de saída em um valor nítido (ou num vetor de valores). Ela é executada após a inferência de agregação *fuzzy*.

No nosso caso, a defuzificação é muito importante, para a ordenação do conjunto de documentos *Web* de acordo com seus prognósticos de qualidade. Esses conjuntos de documentos *Web* ordenados podem ser empregados pelos consumidores

dos dados para filtrar ou personalizar a informação pesquisada, de acordo com seus níveis de qualidade preferidos. O capítulo anterior identificou alguns exemplos dos diversos métodos de defuzificação possíveis de serem adotados.

### 4.1.7 – Pacote Regra

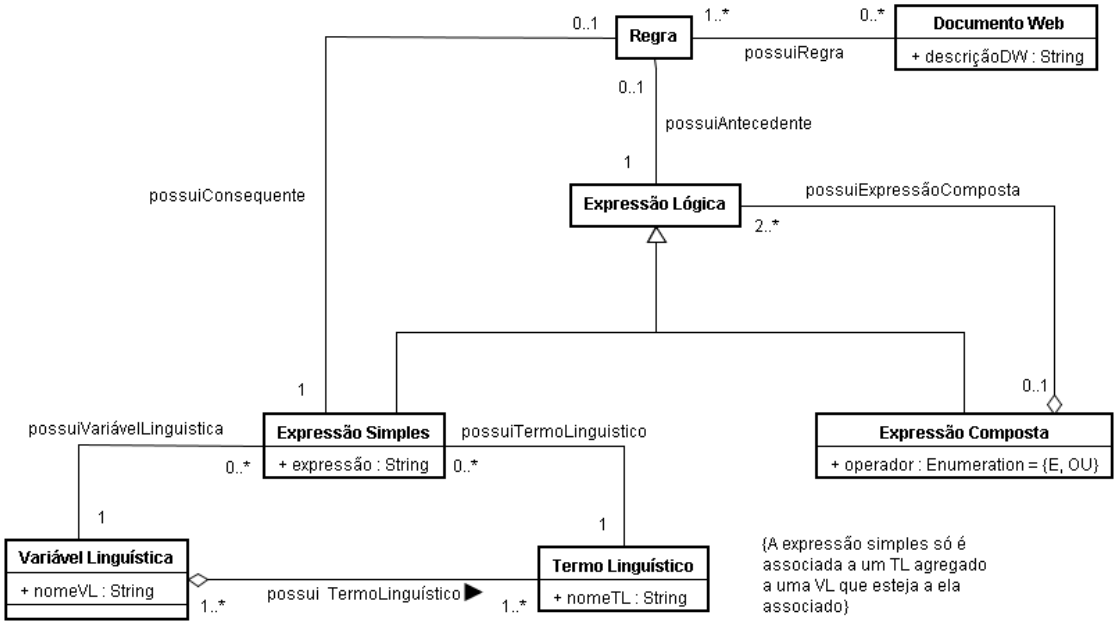



**Figura 4-9: Pacote Regra**

#### *4.1.7.1 – Classe Regra*

O conjunto de classes que forma o sistema de regras do modelo tem por finalidade construir as regras mediante a inferência de conhecimento. Essa inferência ocorre sobre uma variável lingüística, em função de uma associação lógica *fuzzy* entre outras variáveis lingüísticas.

**Definição 17**. Uma regra é representada por (ROSS, 2004):

REGRA *n*: SE *premissa* (*antecedente*) ENTÃO *conclusão* (*conseqüente*)

A *Classe Regra* representa essa estrutura pela *Associação possuiAntecedente* para a *Classe Expressão Lógica* e pela *Associação possuiConseqüente* para a *Classe Expressão Simples*. Uma expressão da *Classe Expressão Lógica* pode ser simples (*Classe Especializada Expressão Simples*) ou composta (*Classe Especializada Expressão Composta*).

As expressões simples associam as variáveis lingüísticas (*Associação possuiVariávelLinguistica* para a *Classe Variável Linguistica*) aos termos lingüísticos (*Associação possuiTermoLinguistico* para a *Classe Termo Linguistico*), por exemplo: <*reputação excelente*>.

As expressões compostas podem ser duas ou mais expressões simples conjugadas por meio dos operadores da *Classe Expressão Composta* (*Associação possuiExpressãoComposta).* O *atributo operador* é uma lista enumerada de operadores {E, OU} e tem por finalidade realizar as operações lógicas *fuzzy* entre as expressões simples. Esse processo recursivo flexibiliza a construção de qualquer tipo de expressão lógica, por exemplo: <*completeza = regular* E *reputação = boa* E *atualidade = boa* >[34].

A *Classe Regra* está associada à *Classe Documentos Web* pela associação *possuiRegra*.

Na seção anterior foi mostrado que o *PC – Prognóstico Composto de Qualidade de Informação* pode ser obtido pela função de composição *fuzzy* da **definição 16**. Entretanto, o *PC* também pode ser obtido pela inferência de regras *fuzzy*.

A contextualização considerada pelos especialistas na atribuição dos graus de importância de cada dimensão de qualidade é adotada de forma semelhante para definição das regras. A seguir um exemplo da definição de uma regra:

*REGRA 1*:

SE as variáveis de entrada < *completeza = regular* E *reputação = boa* E *atualidade = boa* >

ENTÃO a variável de saída *PC = boa*.

Durante o procedimento de inferência são executados os mapeamentos que estabelecem o comportamento das entradas e das saídas (COX, 1994). Esses mapeamentos podem estar baseados em:

- **Implicação** – Combinação dos conjuntos *fuzzy* de entrada que ativam uma regra específica. Dependendo dos resultados desejados e das aplicações envolvidas, um dos usos possíveis é o da implicação pelo mínimo. Ou seja, o

[34] As condições tratadas na nossa abordagem consideram somente a igualdade *fuzzy*, as inegualdades podem ser implementadas de forma similar.

conjunto de saída tem um grau de pertinência igual ao grau mínimo de pertinência, entre os graus de pertinência dos conjuntos de entrada. Por exemplo, se $\mu_{\tilde{B}}(h_i) = 0{,}3$ e $\mu_{\tilde{R}}(a_i) = 0{,}5$ e $\mu_{\tilde{G}}(ut_i) = 0{,}7$ então $\mu_{\tilde{B}}(pc_i) = 0{,}3$.

- **Agregação** – Combinação dos conjuntos *fuzzy* de saída gerados a partir das regras que foram ativadas. De forma semelhante, dependendo dos resultados desejados e das aplicações envolvidas, um dos usos possíveis é o da agregação pelo máximo. Ou seja, o grau máximo de pertinência é selecionado separadamente entre os graus de pertinência de cada conjunto de saída. Por exemplo, se $\mu_{\tilde{B}}(pc_i) = 0{,}2$ e $\mu_{\tilde{R}}(pc_i) = 0{,}4$ e $\mu_{\tilde{G}}(pc_i) = 0{,}3$ então $\mu_{\tilde{B}}(pc_i) = 0{,}4$.

Nesse caso, as funções de pertinência relacionadas às variáveis lingüísticas (variáveis da entrada) de cada página são analisadas pela inferência de implicação *fuzzy*. Em seguida, as funções de pertinência resultantes (variáveis de saída) de cada página são analisadas pela inferência de agregação *fuzzy*.

A defuzificação ocorre de forma semelhante ao que foi descrito na seção anterior. Ela é executada após a inferência de agregação *fuzzy*.

## 4.2 – Metodologia Proposta

A partir do conjunto de orientações e técnicas apresentadas, que se iniciam com a entrada de informações relativas ao modelo para o prognóstico de qualidade de informação proposto, foi definido um processo racional para avaliar a qualidade das informações na *Web*, obedecendo às fases e aos pontos de decisão determinados.

A Figura 4-10 mostra de forma organizada e seqüencial os passos da metodologia proposta.

Na seção 2.7.2 do capítulo 2 foi apresentada a generalização das principais fases das metodologias de avaliação de qualidade da informação em vista de suas principais atividades. A Figura 4-10 também destaca a correspondência de cada passo com cada uma daquelas fases.

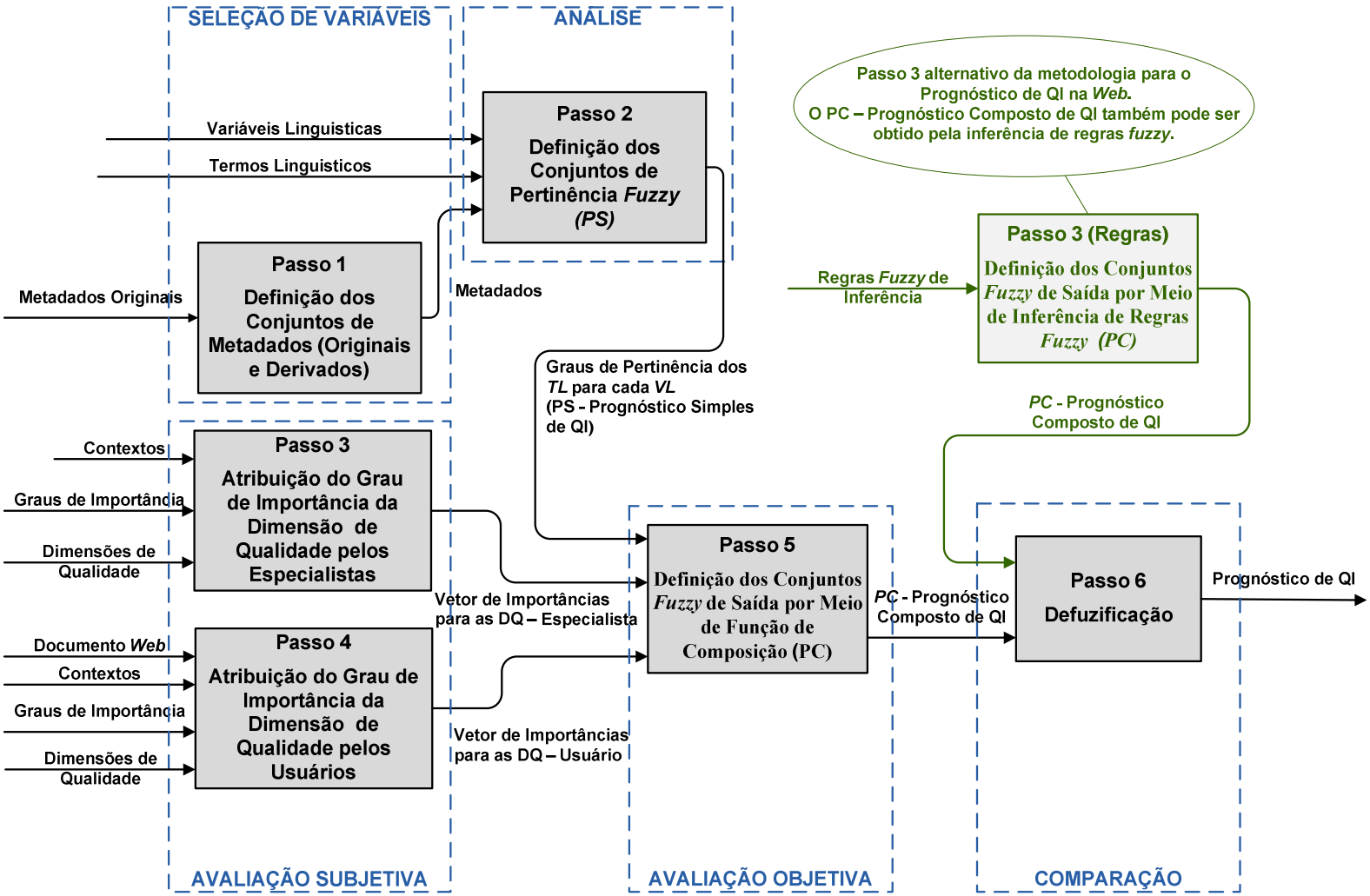



**Figura 4-10: Metodologia para o Prognóstico de QI na *Web***

### 4.2.1 – Instanciação do Modelo de QI na *Web*

O modelo de qualidade de informações na *Web* precisa ser instanciado, antes que se iniciem os passos da metodologia proposta. Essa fase corresponde ao ***Passo 0*** do processo e envolve as atividades de pré-implantação a seguir descritas e ilustradas pela Figura 4-11.

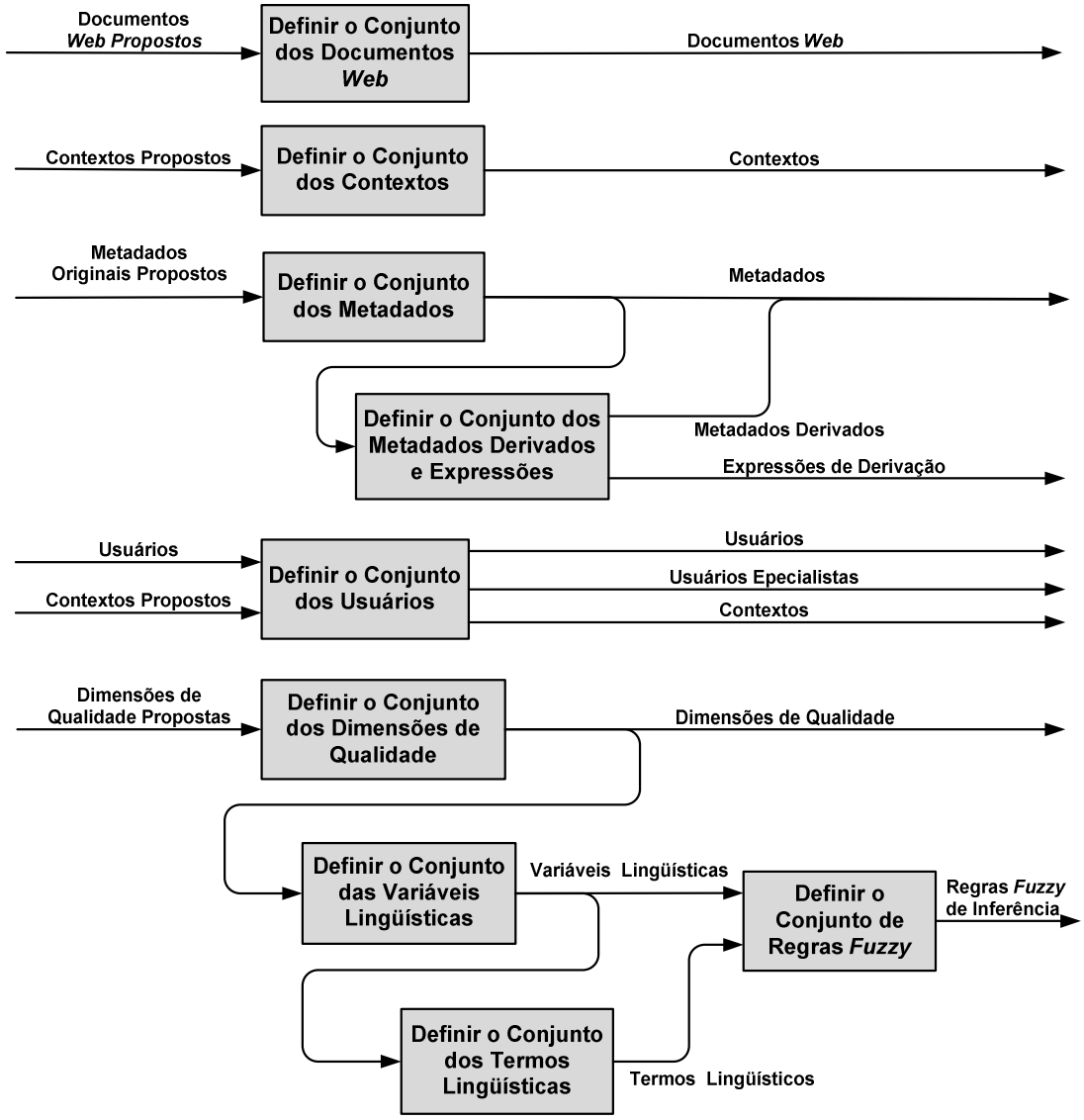



**Figura 4-11: Instanciação do Modelo de Qualidade de Informação na *Web***

Acompanhando as atividades da Figura 4-11, inicialmente é instanciado o conjunto dos Documento *Web* com os documentos oriundos da busca. São instanciados também os conjuntos Contexto e Metadado a partir de uma lista de contextos e de metadados originais propostos.

Com base no conjunto Metadados são instanciados os metadados derivados e definidas as expressões de derivação para essa classe de metadados.

Em seguida são definidos os usuários e entre eles os especialistas que são associados aos seus contextos de especialização.

Baseado numa lista de dimensões de qualidade propostas é instanciado o conjunto Dimensão de Qualidade, a partir do qual são definidos os conjuntos Variável Lingüística e Termo Lingüístico, respectivamente. Além disso, as dimensões de qualidade são associadas aos metadados cujos valores lhes servirão de base de avaliação.

Os conjuntos de variáveis lingüísticas e termos lingüísticos servem de base para definição e instanciação das regras *fuzzy* de inferência no conjunto Regra.

Ao término dessas atividades de pré-implantação pode ter início a avaliação de qualidade de informação, seguindo os passos da metodologia proposta, a seguir descrita.

### 4.2.2 – Passos da Metodologia para o Prognóstico de QI na *Web*

São descritos a seguir os passos da metodologia proposta e as suas correspondências com as principais fases das metodologias de avaliação de qualidade da informação: seleção de variáveis, análise, avaliação subjetiva, avaliação objetiva e comparação:

- **Seleção de Variáveis**

Essa fase corresponde ao ***Passo 1*** do processo e envolve a seleção do conjunto de metadados originais e a definição dos metadados derivados de acordo com as funções de transformação correspondentes.

- **Análise**

Essa fase corresponde ao ***Passo 2*** do processo e envolve a definição dos conjuntos *fuzzy* de pertinência *PS* para as variáveis lingüísticas especificadas, que representam os conjuntos *fuzzy* de entrada. O conjunto *PS* é calculado com base nos valores dos metadados adotados pelas variáveis lingüísticas.

- **Avaliação Subjetiva**

Essa fase corresponde aos ***Passos 3*** e ***4*** do processo. O ***Passo 3*** envolve a atribuição dos graus de importância para cada dimensão de qualidade, considerando um

contexto específico. Esses graus de importância são atribuídos pelos especialistas para as dimensões de qualidade. Como resultado dessa atribuição, é definido um vetor de importâncias para as dimensões de qualidade selecionadas. O ***Passo 3*** envolve a atribuição dos graus de importância para cada dimensão de qualidade, considerando os requisitos específicos do usuário. Esses graus de importância são atribuídos pelos usuários para as mesmas dimensões de qualidade e contexto do passo anterior, considerando agora as suas percepções e os seus requisitos específicos. Como resultado dessa atribuição, é definido um vetor de importâncias para as dimensões de qualidade selecionadas.

O ***Passo 3 – alternativo*** da metodologia destaca que o *PC – Prognóstico Composto de Qualidade de Informação* também pode ser obtido pela inferência de regras *fuzzy*. Na prova de conceito da seção 6.1.2 do capítulo 6 há um exemplo que emprega essas regras.

- **Avaliação Objetiva**

Essa fase corresponde ao ***Passo 5*** do processo e envolve a definição dos conjuntos *fuzzy* de saída *PC* que são obtidos por meio da função de composição *fuzzy*, a partir dos conjuntos *fuzzy* de pertinência *PS*.

- **Comparação**

Essa fase corresponde ao ***Passo 6*** do processo e envolve a defuzificação dos conjuntos *fuzzy* de saída *PC*. Esses resultados são empregados como prognósticos de qualidade de informação para ordenação do conjunto de documentos *Web*.

A fim de melhor explicar a abordagem proposta, o capítulo 6 descreve a prova de conceito por meio de dois estudos de caso reais que percorrem todos os passos do processo de obtenção do prognóstico de qualidade. Nesses exemplos são aplicadas duas possíveis alternativas de solução, uma usando as funções de transformação *fuzzy* e outra usando a inferência de regras *fuzzy*.

# Capítulo 5 – Arquitetura da Aplicação para o Prognóstico de Qualidade de Informação na *Web*

Não há nenhuma definição universalmente aceita sobre quais aspectos constituem uma arquitetura de sistemas, existindo, portanto, inúmeras delas oriundas das mais variadas fontes. Em seu contexto, é comum reconhecer três ou quatro tipos de arquiteturas, cada uma correspondendo a um domínio particular. Exemplos desses domínios são as arquiteturas de negócio, dos dados, de infra-estruturas ou técnicas e das aplicações. Essas últimas mostram a estrutura e o comportamento das aplicações, com foco na interação entre os seus componentes e com os usuários. Tais arquiteturas mostram, ainda, as atribuições de responsabilidades dos diferentes componentes e buscam garantir que as interações entre esses componentes satisfaçam os requisitos do sistema (CUNNINGHAM, MAYNARD *et al.*, 2002). Elas destacam também os dados consumidos e produzidos pelas aplicações, em vez de suas estruturas internas. Geralmente denotam as funções do negócio e as tecnologias da plataforma da aplicação e fornecem um plano a partir do qual esses componentes poderão ser adquiridos ou desenvolvidos, para trabalharem em conjunto na execução do sistema como um todo[35].

A arquitetura do domínio da aplicação, que está descrita neste capítulo, incorpora, para o nosso caso, todos os requisitos necessários à adaptação do sistema às características dos usuários e do ambiente.

## 5.1 – Arquitetura e Detalhes Técnicos da Implementação

Com base nas definições referenciadas, foi implementada uma arquitetura computacional, flexível e extensível, como uma coleção de componentes, interfaces e padrões destinados a atender ao conjunto de requisitos identificados no domínio da aplicação em estudo. A Figura 5-1 denota a arquitetura implementada, seguida da descrição de todos os seus componentes (BARROS, XEXÉO *et al.*, 2008a).

[35] http://en.wikipedia.org/wiki/TOGAF. O *Open Group Architecture Framework* (TOGAF) é um *framework* para a Arquitetura dos Empreendimentos (*EA*) que fornece uma abordagem detalhada ao projeto, ao planejamento, à implementação e a governança da arquitetura de informação da empresa.

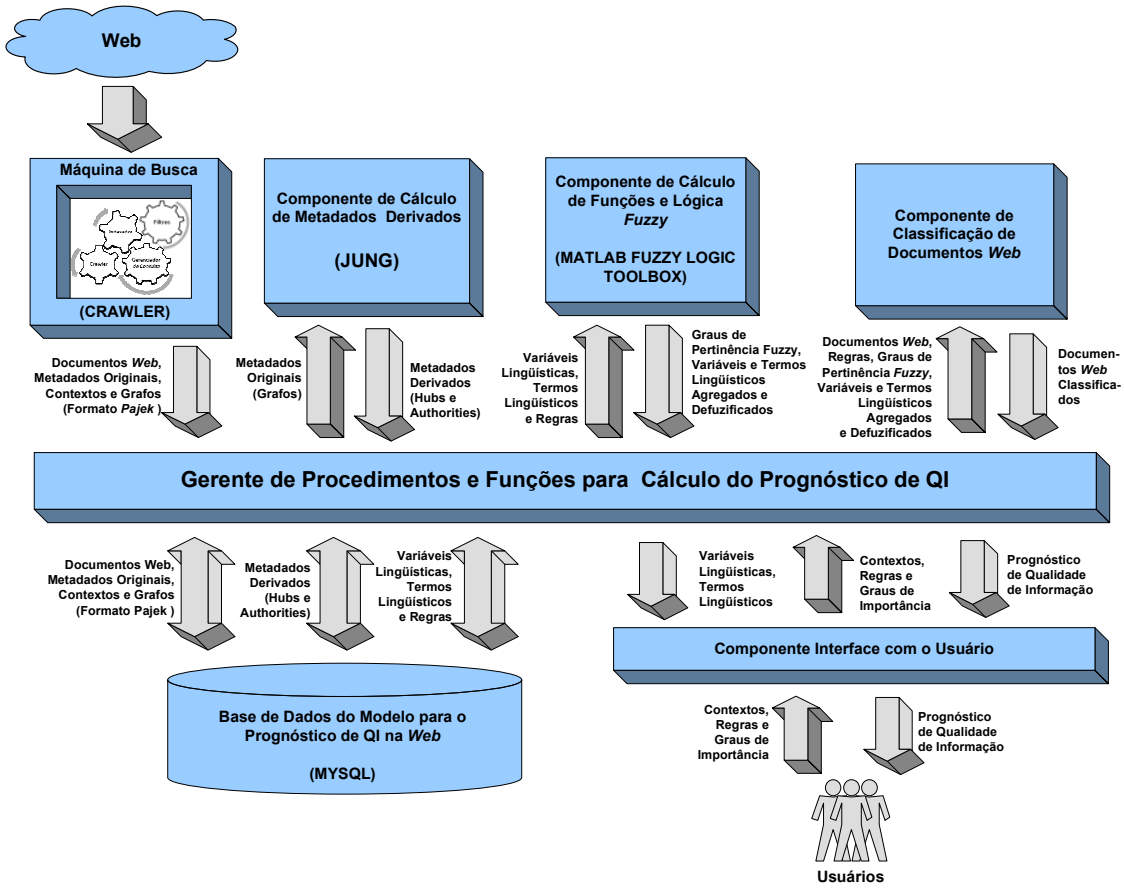



**Figura 5-1: Arquitetura da Aplicação para o Prognóstico de Qualidade de Informação na *Web***

- **Gerente de Procedimentos e Funções para Cálculo do Prognóstico de QI** – Esse componente é responsável pelo controle e mediação do processamento e do fluxo de informações entre todos componentes.

- **Máquina de Busca (*crawler*)** – Esse componente é responsável por percorrer a *Web* e capturar as páginas através de seus *links*, com base nos argumentos de pesquisa propostos pelos usuários. Os metadados e os contextos das páginas também são capturados juntamente com elas. Além disso, ele converte o conjunto de páginas *Web* capturadas em cada busca para o formato de grafos, nos quais as páginas são os vértices e os *links* são os arcos (formato Pajek[36]). Esses grafos são parâmetros de entrada para o Componente de Cálculo de Metadados Derivados. Inicialmente foi utilizado o *crawler* adaptado de (MAYWORM, 2007), adotando-se

[36] http://pajek.imfm.si/doku.php.

a política de busca em largura (*breadth-first strategy*) (BAEZA-YATES, CASTILLO *et al.*, 2005) e, posteriormente, foi desenvolvido um *crawler* específico para o projeto. A Figura 5-2 exemplifica o grafo que representa um conjunto de páginas *Web* em dois níveis de busca no formato Pajek, cuja a URL semente é HTTP://answer.com/topic/relational-database.

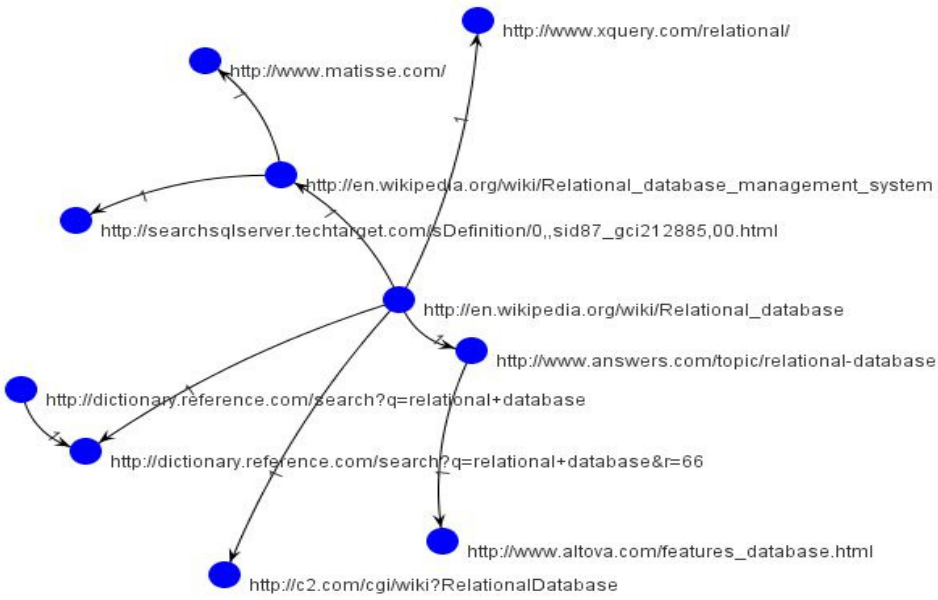



**Figura 5-2: Grafo de um Conjunto de Páginas *Web* no Formato Pajek**

- **Componente de Cálculo de Metadados Derivados** – Esse componente é responsável pela execução das funções de cálculo e transformação para obtenção dos metatados derivados a partir dos metadados originais, de acordo com o que foi previamente apresentado na Tabela 4-2. O JUNG[37] foi adotado para calcular os metadados derivados *hubs* e *authorities* (KLEINBERG, 1998) (AMENTO, TERVEEN *et al.*, 2000) usando os grafos gerados pelo *crawler*.

- **Componente de Cálculo de Funções e Lógica *Fuzzy* –** Esse componente é responsável pela execução de todas as funções *fuzzy*. A partir das variáveis

[37] O JUNG - *Java Universal Network/Graph Framework* é um software de biblioteca que fornece uma linguagem comum e extensível para modelagem, análise e visualização dos dados que podem ser representados como um grafo ou uma rede usando o formato de *Pajek*. Ele foi escrito em Java, permitindo que as aplicações baseadas no JUNG empreguem as potencialidades da *built-in* da API Java, bem como aquelas existentes em bibliotecas Java disponibilizadas por terceiros.
http://jung.sourceforge.net/.

lingüísticas, termos lingüísticos e regras definidas, ele calcula todos os graus de pertinência *fuzzy*, bem como os valores de agregação e a defuzificação desses valores. Ele foi implementado por meio de uma coleção de funções construídas no *MATLAB FUZZY LOGIC TOOLBOX*™[38]. Esse produto estende o ambiente computacional do MATLAB® com ferramentas e métodos para projetos de sistemas baseados na lógica *fuzzy*.

- **Componente de Classificação de Documentos *Web* –** Esse componente é responsável pela classificação dos documentos, a partir dos seus valores de agregação e de defuzificação para geração dos prognósticos da qualidade das informações.

- **Componente Interface com o Usuário –** Esse componente é responsável por retornar aos usuários os prognósticos gerados, bem como pelas demais interações entre os usuários e o sistema, no que se refere a atribuição dos graus de importância as dimensões de qualidade por contexto e a definição das regras *fuzzy* de inferência.

- **Base de Dados do Modelo para o Prognóstico de QI na *Web* –** Essa base armazena os conjuntos de dados e informações capturados, calculados e providos pelo modelo, durante todo o processo de geração do prognóstico de qualidade de informações. O MYSQL® foi adotado para implementação da base.

[38] http://www.mathworks.com/products/fuzzylogic/.

# Capítulo 6 – Aplicação e Validação da Abordagem Proposta

As avaliações explícitas são usadas em muitas aplicações comerciais de filtragem colaborativa como, por exemplo, a Amazon[47] e a Netflix[48], que as adotam para recomendar novos produtos aos usuários (LERMAN, 2007). Nesses casos, um usuário atribui uma avaliação, ou um voto positivo ou negativo a algum documento. Essa abordagem também tem sido bem sucedida nos agregadores sociais de notícia, como Digg[49], que se tornou popular por ser baseado nas opiniões distribuídas oriundas de muitos avaliadores independentes, para ajudar os usuários a encontrarem as notícias mais interessantes.

Uma avaliação de qualidade de informação apropriada precisa investigar a qualidade de uma lista de resultados obtidos mediante diferentes abordagens. Nesse sentido, os usuários de teste precisam julgar as páginas resultantes de acordo com suas percepções subjetivas de qualidade.

Neste capítulo, as seções a seguir descrevem os resultados obtidos por meio da aplicação e validação da abordagem proposta na tese.

Inicialmente, na seção 6.1, são demonstrados os resultados das provas de conceito implementadas por meio de dois estudos de casos práticos em duas abordagens *fuzzy*, uma adotando as funções *fuzzy* de transformação e outra as regras *fuzzy* de inferência.

Em seguida, a seção 6.2 descreve o desenvolvimento de uma aplicação colaborativa que foi desenvolvida como uma extensão do FoxSet[50]. Essa aplicação tem por objetivo a construção de bases de testes contendo páginas *Web* avaliadas

[47] http://www.amazon.com/.

[48] http://www.netflix.com/.

[49] http://digg.com/.

[50] O FoxSet é o protótipo de uma ferramenta para a construção de *datasets* desenvolvido no PESC/COPPE, na disciplina de BRI, ministrada pelo Prof. Geraldo Xexéo. Ele consiste em um *plugin* para o Firefox – *backend* em *PHP* – e usa o *MySQL* como repositório central. Tem como objetivo facilitar o gerenciamento do processo de construção de *datasets,* além de armazenar e disponibilizar conjuntos de páginas *Web* relevantes em relação a um determinado contexto.

manualmente e automaticamente por meio da nossa proposta. Os resultados preliminares foram obtidos por meio de dois estudos de casos que demonstram as diferentes formas de uso do Foxset para avaliação de um conjunto de páginas *Web*, em contextos específicos.

Há também, na seção 6.3, uma outra validação da abordagem proposta que foi realizada com o auxílio de um método experimental, para a avaliação de um maior número páginas. Nesse experimento, foram comparados os resultados obtidos automaticamente seguindo a nossa metodologia, em contrapartida aos resultados obtidos mediante a avaliação de usuários finais.

Finalmente, a seção 6.4 descreve uma última avaliação que foi realizada por meio da comparação entre os resultados de ordenação das páginas obtidos pelo prognóstico de qualidade e aqueles que foram obtidos pelo Google®. Tal comparação foi baseada nos cálculos de precisão e cobertura, e da média harmônica dos resultados obtidos.

## 6.1 – Provas de Conceito

Para provar a viabilidade da proposta e o conceito teórico estabelecido na nossa pesquisa, dois estudos de casos práticos são demonstrados a seguir, num cenário específico para o contexto *economia*.

Nesses exemplos foram adotadas a *atualidade,* a *reputação* e a *completeza* como variáveis lingüísticas, que foram definidas pelas dimensões de qualidade correspondentes. Essas variáveis lingüísticas são suportadas pelos metatados *atualidade, authority e hub* respectivamente.

O número de termos lingüísticos para uma avaliação subjetiva pode ser estabelecido de acordo com:

- a conveniência do projeto;
- as possíveis peculiaridades do domínio da aplicação; ou
- a determinação da equipe de gestão da qualidade.

Alguns trabalhos argumentam que para se obter uma boa classificação, devem ser definidos pelo menos cinco ou sete termos lingüísticos (COX, 1994; KANTROWITZ, HORSTKOTTE *et al.*, 1997; KLIR & YUAN, 1995; ROSS, 2004). Em razão da simplificação, nos estudos de caso das provas de conceito foram definidos

somente três termos lingüísticos para cálculo dos conjuntos *fuzzy* de pertinência que descrevem as variáveis lingüísticas. Os termos lingüísticos classificados são: *ruim (bad), regular (regular)* e *bom (good)*, envolvendo, no caso, todas as possibilidades dos subconjuntos *fuzzy* denotado por $\widetilde{N}(R), \widetilde{N}(C)\ e\ \widetilde{N}(T)$ para *atualidade, completeza e reputação*, respectivamente. Porém, para a aplicação colaborativa, bem como para o experimento de avaliação, o conjunto de termos lingüísticos foi estendido para cinco: *péssimo, ruim, regular, bom* e *excelente*.

Ambos os exemplos relativos as provas de conceito observaram os passos da metodologia descrita no capítulo 4. A Tabela 6-1 apresenta os cinco primeiros resultados obtidos pelo *crawler*, juntamente com os valores dos seus metadados.

**Tabela 6-1: Valores dos Metadados Originais e Metadados Derivados**

| **Contexto: *economia*** | **Metadados Originais** | | **Metadados Derivados** | | |
|---|---|---|---|---|---|
| **Sites *Web*** | **UD** | **QT** | **UT** | ***Authority*** | ***Hub*** |
| Economist.com Surveys[51] | 25 Nov | 28 Nov | 72 | 0.019793 | 0.026939 |
| Economist Conferences[52] | 23 Nov | 28 Nov | 120 | 0.018074 | 0.022909 |
| The World In 2007[53] | 21 Nov | 28 Nov | 168 | 0.017019 | 0.017008 |
| Economist.com Opinion[54] | 28 Nov | 28 Nov | 12 | 0.012019 | 0.053238 |
| Scottrade[55] | 22 Nov | 28 Nov | 144 | 0.007503 | 0.007331 |

### 6.1.1 – Prova de Conceito Usando Funções *Fuzzy* de Transformação

O exemplo a seguir demonstra, passo a passo, a prova de conceito que emprega as funções *fuzzy* de transformação (BARROS, XEXÉO *et al.*, 2008b):

**Passo 1 –** Seleção e cálculo do conjunto de metadados *OM* e *DM* calculados de acordo com a Tabela 4-2, cujos resultados estão expressos na Tabela 6-1:

$$OM_i = \{ud_i, qt_i, bl_i, fl_i\}$$

Onde, $1 \leq i \leq 5$ é uma instância de um documento *Web*;

[51] http://www.economist.com/surveys/

[52] http://www.economistconferences.com/

[53] http://www.theworldin.com/

[54] http://www.economist.com/opinion

[55] http://www.scottrade.com/index.asp?supbid=68597.

$ud_i$ – data de atualização de um documento $Web_i$;

$qt_i$ – data da consulta de um documento $Web_i$;

$bl_i$ – número de *links* que apontam para o documento $Web_i$;

$fl_i$ – número de *links* externos do documento.

$$DM_i = \{fd_{(ud_i,qt_i)}, fd_{(bl_i)}, fd_{(fl_i)}\} = \{ut_i, a_i, h_i\}$$

$ut_i$ – tempo desde a atualização de um documento $Web_i$;

$a_i$ – autoridade *(authority)* de um documento $Web_i$;

$h_i$ – centralidade *(hub)* de um documento $Web_i$.

**Passo 2 –** Definição dos valores *fuzzy* de pertinência *PS*[56] para as variáveis lingüísticas *reputação*, *completeza* e *atualidade* que representam os conjuntos *fuzzy* de entrada, lembrando que, a partir deles, os conjuntos *fuzzy* de saída *PC*[57] serão obtidos por meio de uma função de agregação *fuzzy*. Os metadados *autoridade* (*authority*), *centralidade* (*hub*) e o *tempo de atualização* da página foram assumidos como base de avaliação para as variáveis lingüísticas.

A descrição e o exemplo a seguir ilustram mais detalhadamente essa definição:

Seja *R* o conjunto referencial de todos os valores possíveis para a variável lingüística *reputação*. *Reputação* assume seu metadado correspondente a*uthority* como o dado base de avaliação. Para um dado documento *i*, baseado na a*uthority*, no contexto *economia*, o modelo deve prover os valores de pertinência para o conjunto dos termos lingüísticos definidos.

Seja *C* o conjunto referencial de todos os valores possíveis para a variável lingüística *completeza*. *Completeza* assume seu metadado correspondente *hub* como o dado base de avaliação. Para um dado documento *i*, baseado no *hub* no contexto *economia*, o modelo deve prover os valores de pertinência para o conjunto dos termos lingüísticos definidos.

Seja *T* o conjunto referencial de todos os valores possíveis para a variável lingüística *atualidade*. *Atualidade* assume seu metadado correspondente *tempo de*

[56] Prognóstico Singular de Qualidade de Informação.

[57] Prognóstico Composto de Qualidade de Informação.

*atualização* como o dado base de avaliação. Para um dado documento *i*, baseado no *tempo de atualização* no contexto *economia*, o modelo deve prover os valores de pertinência para o conjunto dos termos lingüísticos definidos.

A Figura 6-1 mostra a variável lingüística *atualidade* e seus valores possíveis, os termos lingüísticos: *ruim (bad), regular (regular)* e *bom (good)*, denotados como $\tilde{B}$, $\tilde{R}$ e $\tilde{G}$:

$$\tilde{B} = \left\{ \left( ut_i, \mu_{\tilde{B}}(ut_i) \right) \middle| \, ut_i \in UT \right\}$$

$$\tilde{R} = \left\{ \left( ut_i, \mu_{\tilde{R}}(ut_i) \right) \middle| \, ut_i \in UT \right\}$$

$$\tilde{G} = \left\{ \left( ut_i, \mu_{\tilde{G}}(ut_i) \right) \middle| \, ut_i \in UT \right\}$$

Onde, $\mu_{\tilde{B}}(ut_i) : UT \to [0,1]$, $\mu_{\tilde{R}}(ut_i) : UT \to [0,1]$, $\mu_{\tilde{G}}(ut_i) : UT \to [0,1]$ representam as funções *fuzzy* de pertinência que mapeiam o elemento $ut_i$ *(tempo de atualização* de um documento *Web)* em $\tilde{B}$, $\tilde{R}$ e $\tilde{G}$ respectivamente.

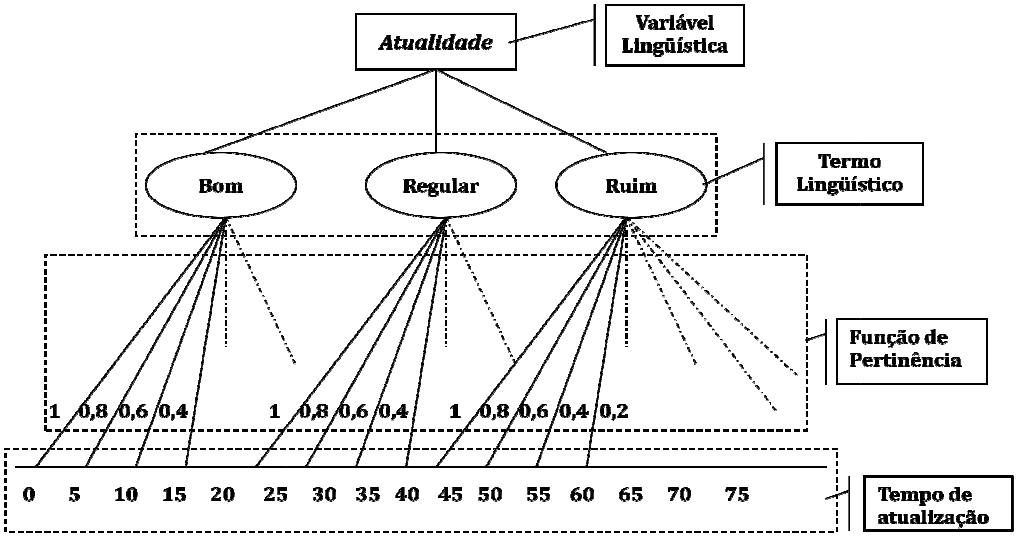



**Figura 6-1: Variável Lingüística "Atualidade" (Adaptada de Klir & Yuan, 1995)**

As funções de pertinência foram implementadas no MATLAB® usando os números *fuzzy* triangulares para distribuir os valores do domínio por cada termo lingüístico. A Figura 6-2 e a Figura 6-3 mostram graficamente o modelo do mapeamento das funções de pertinência para os subconjuntos *fuzzy*, ruim (*bad*), regular (*regular*) e bom (*good*). Os números triangulares, que compõem cada conjunto, têm a mesma largura de base. O número triangular relacionado ao conjunto *regular* está no meio do intervalo numérico dos metadados.

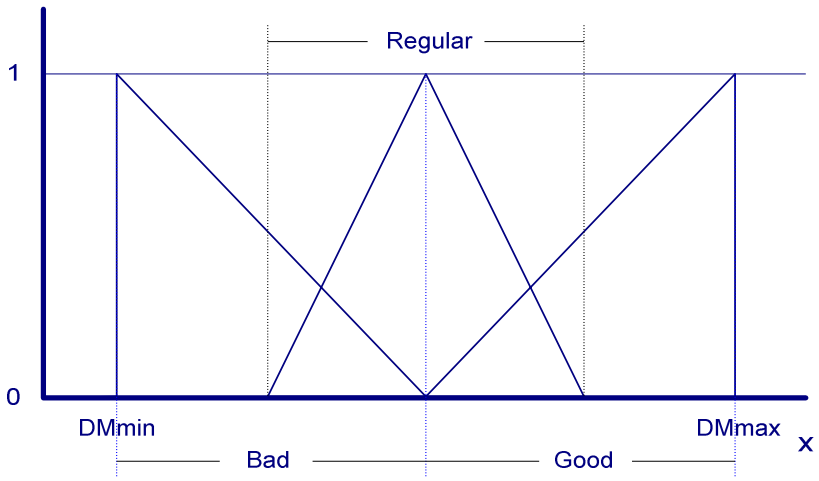



**Figura 6-2: Modelo para o Mapeamento das Funções de Pertinência para os Subconjuntos Fuzzy ruim, regular e bom para as Variáveis Lingüísticas reputação e completeza**

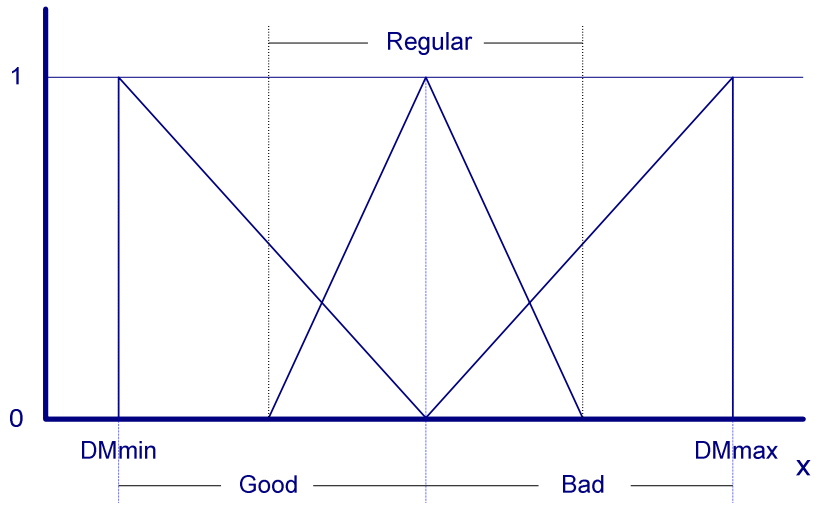



**Figura 6-3: Modelo para o Mapeamento das Funções de Pertinência para os Subconjuntos Fuzzy ruim, regular e bom para a Variável Lingüística atualidade**

O mapeamento *fuzzy* ilustrado nas Figura 6-2 e Figura 6-3 é realizado a partir dos metadados obtidos no **passo 1**.

Como resultados desse mapeamento são obtidos, para cada documento *Web*, os graus de pertinência dos subconjuntos *fuzzy ruim, regular e bom* para cada uma das variáveis lingüísticas. Esses resultados são mostrados na Tabela 6-2.

Observe que a Tabela 6-2 fornece *Dmin* e *Dmax* pelo menor e pelo maior valor de cada metadado derivado. Ela não adota os valores de *PC* para *Dmin* e *Dmax* e, nesse caso, foi pré-suposto *Dmin* = 0 e *Dmax* = 1.

**Tabela 6-2: Resultados da Fuzificação com os Graus de Pertinência dos Subconjuntos Fuzzy ruim, regular e bom para Variáveis Lingüísticas atualidade, reputação e completeza**

| **Contexto: *economia*** | **Resultados da Fuzificação (Graus de Pertinência)** | | |
|---|---|---|---|
| **Sites *web*** | **Atualidade** | **Reputação** | **Completeza** |
| **Economist.com Surveys** | $\mu_{\tilde{B}}(ut_i) = 0.25$<br>$\mu_{\tilde{R}}(ut_i) = 0.55$<br>$\mu_{\tilde{G}}(ut_i) = 0$ | $\mu_{\tilde{B}}(a_i) = 0$<br>$\mu_{\tilde{R}}(a_i) = 0$<br>$\mu_{\tilde{G}}(a_i) = 1$ | $\mu_{\tilde{B}}(h_i) = 0{,}167110$<br>$\mu_{\tilde{R}}(h_i) = 0{,}7157852$<br>$\mu_{\tilde{G}}(h_i) = 0$ |
| **Economist Conferences** | $\mu_{\tilde{B}}(ut_i) = 0$<br>$\mu_{\tilde{R}}(ut_i) = 0.25$<br>$\mu_{\tilde{G}}(ut_i) = 0.4$ | $\mu_{\tilde{B}}(a_i) = 0$<br>$\mu_{\tilde{R}}(a_i) = 0$<br>$\mu_{\tilde{G}}(a_i) = 0{,}727254$ | $\mu_{\tilde{B}}(h_i) = 0{,}33829$<br>$\mu_{\tilde{R}}(h_i) = 0{,}3734191$<br>$\mu_{\tilde{G}}(h_i) = 0$ |
| **The World In 2007** | $\mu_{\tilde{B}}(ut_i) = 0$<br>$\mu_{\tilde{R}}(ut_i) = 0$<br>$\mu_{\tilde{G}}(ut_i) = 1$ | $\mu_{\tilde{B}}(a_i) = 0$<br>$\mu_{\tilde{R}}(a_i) = 0$<br>$\mu_{\tilde{G}}(a_i) = 0{,}559862$ | $\mu_{\tilde{B}}(h_i) = 0{,}58895$<br>$\mu_{\tilde{R}}(h_i) = 0$<br>$\mu_{\tilde{G}}(h_i) = 0$ |
| **Economist.com Opinion** | $\mu_{\tilde{B}}(ut_i) = 1$<br>$\mu_{\tilde{R}}(ut_i) = 0$<br>$\mu_{\tilde{G}}(ut_i) = 0$ | $\mu_{\tilde{B}}(a_i) = 0{,}283470$<br>$\mu_{\tilde{R}}(a_i) = 0{,}4830675$<br>$\mu_{\tilde{G}}(a_i) = 0$ | $\mu_{\tilde{B}}(h_i) = 0$<br>$\mu_{\tilde{R}}(h_i) = 0$<br>$\mu_{\tilde{G}}(h_i) = 1$ |
| **Scottrade** | $\mu_{\tilde{B}}(ut_i) = 0$<br>$\mu_{\tilde{R}}(ut_i) = 0$<br>$\mu_{\tilde{G}}(ut_i) = 0.7$ | $\mu_{\tilde{B}}(a_i) = 1$<br>$\mu_{\tilde{R}}(a_i) = 0$<br>$\mu_{\tilde{G}}(a_i) = 0$ | $\mu_{\tilde{B}}(h_i) = 1$<br>$\mu_{\tilde{R}}(h_i) = 0$<br>$\mu_{\tilde{G}}(h_i) = 0$ |

**Passo 3 –** Atribuição dos graus de importância para cada dimensão de qualidade, considerando um contexto específico. Esses graus de importância são atribuídos pelo especialista para as dimensões de qualidade *atualidade (tim)*, *reputação (rep)* e *completeza (com)*, considerando o contexto *economia (eco)*. Como resultado dessa atribuição, é definido um vetor de importâncias para as dimensões de qualidade selecionadas:

$$CDQ = \langle w(eco,tim) \quad w(eco,com) \quad w(eco,rep) \rangle$$

O exemplo considera a importância como a média aritmética de todos os graus atribuídos pelos especialistas a uma $qd_n$. Dessa forma, são assumidos os seguintes graus de importância:

$$CDQ = \langle 2 \quad 3 \quad 4 \rangle$$

*atualidade* = 2 (é importante em algumas circunstâncias, mas nem sempre);

*reputação* = 3 (é muito importante);

*completeza* = 4 (é essencial).

**Passo 4 –** Atribuição dos graus de importância para cada dimensão de qualidade, considerando os requisitos específicos do usuário. Esses graus de importância são atribuídos pelos usuários para as mesmas dimensões de qualidade e contexto do passo anterior, considerando agora as suas percepções e os seus requisitos específicos. Como resultado dessa atribuição, é definido um vetor de importâncias para as dimensões de qualidade selecionadas:

$$UDQ = \langle w(eco,tim) \quad w(eco,com) \quad w(eco,rep) \rangle$$

O exemplo considera a importância como média aritmética de todos os graus atribuídos pelos usuários a uma $qd_n$. Dessa forma, são assumidos os seguintes graus de importância:

$$UDQ = \langle 3 \quad 2 \quad 4 \rangle$$

*atualidade* = 3 (é muito importante);

*reputação* = 2 (é importante em algumas circunstâncias, mas nem sempre);

*completeza* = 4 (é essencial).

**Passos 5 e 6 –** Definição dos conjuntos *fuzzy* de saída *PC* que são obtidos por meio da função de composição *fuzzy*, a partir dos conjuntos *fuzzy* de pertinência *PS* da Tabela 6-2 (definição 16 do capítulo 4). Os conjuntos *fuzzy* de saída *PC* são defuzificados pelo *método do Centróide* e empregados como prognósticos de qualidade de informação para ordenação do conjunto de documentos *Web*. As Figuras 6-4 a 6-8 mostram os conjuntos *fuzzy* de saída *PC* e os resultados da defuzificação para cada documento *Web*.

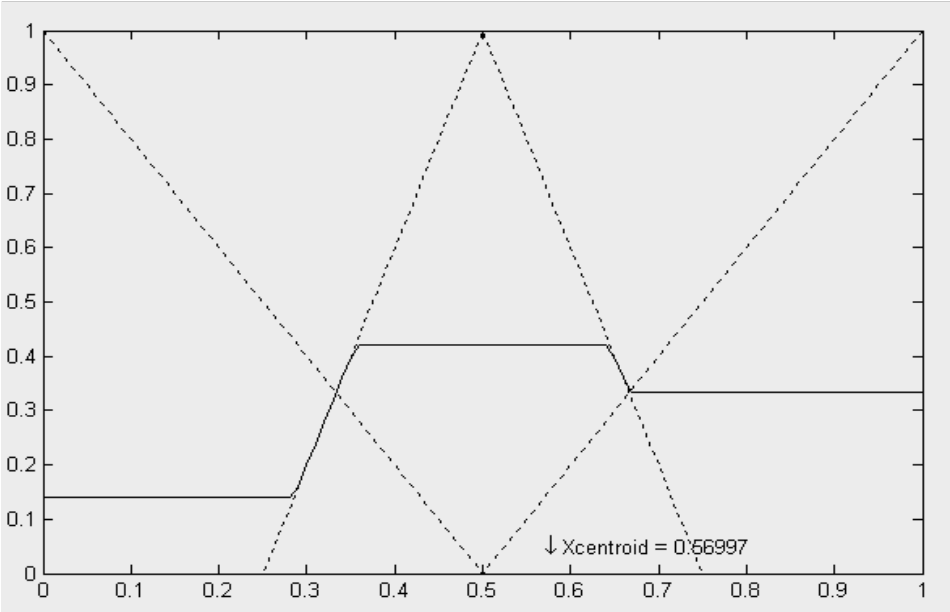



**Figura 6-4: Resultados da Defuzificação para http://www.economist.com/surveys/**

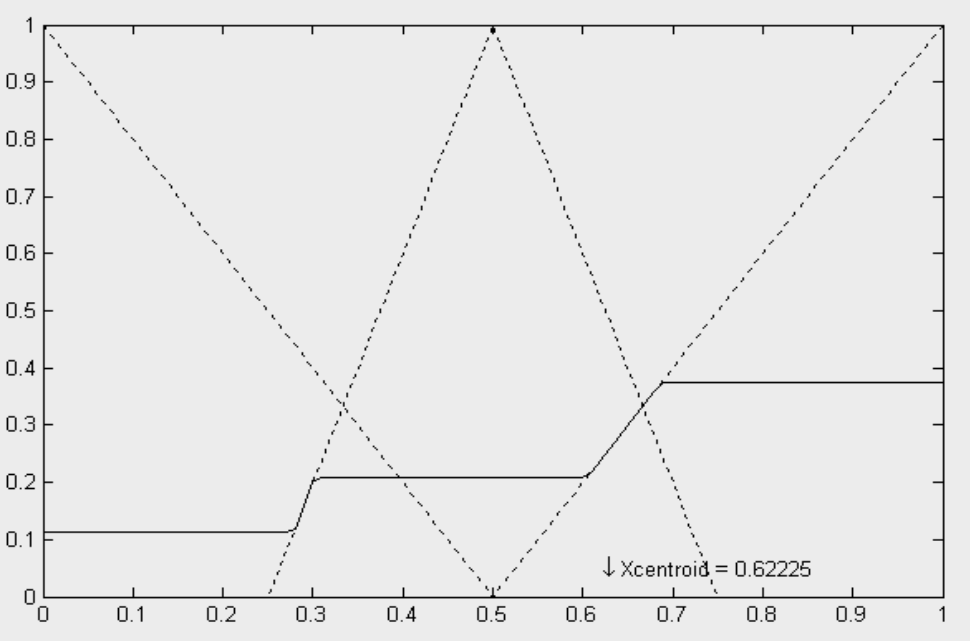



**Figura 6-5: Resultados da Defuzificação para http://www.economistconferences.com/**

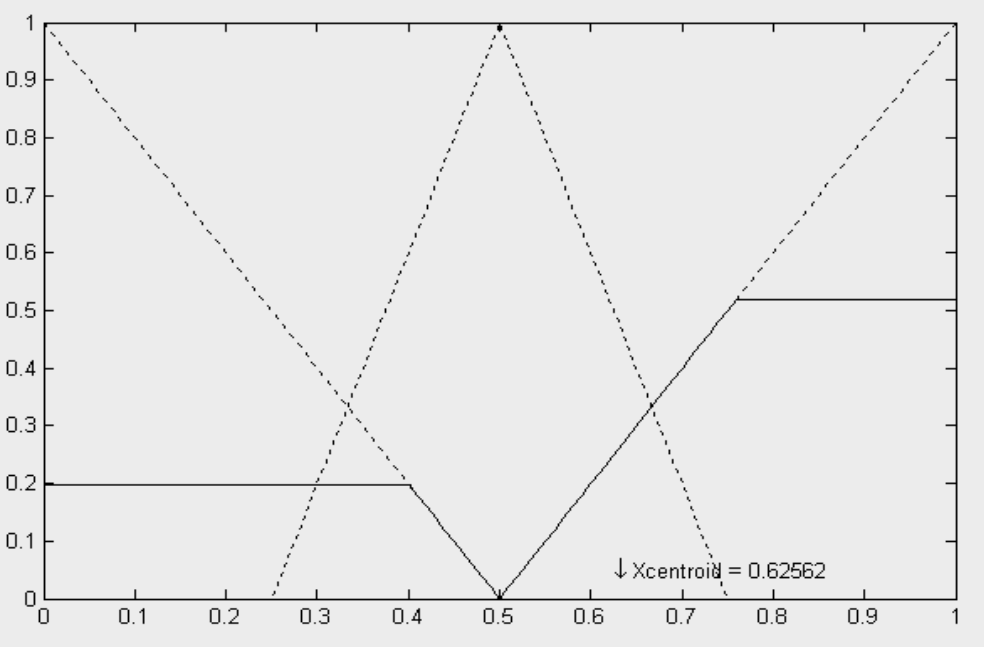



**Figura 6-6: Resultados da Defuzificação para http://www.theworldin.com/**

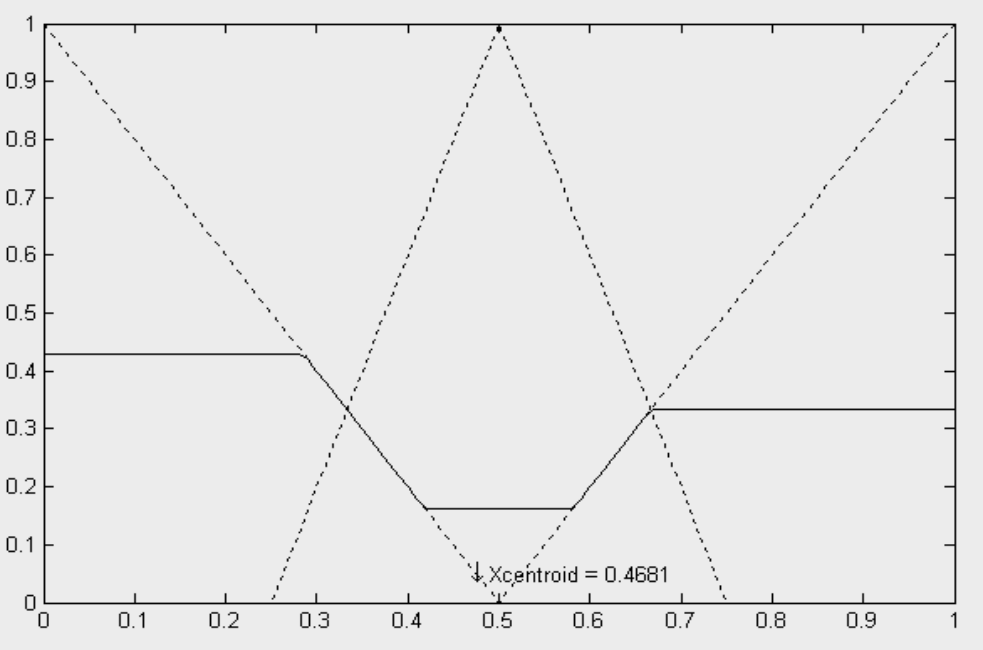



**Figura 6-7: Resultados da Defuzificação para http://www.economist.com/opinion**

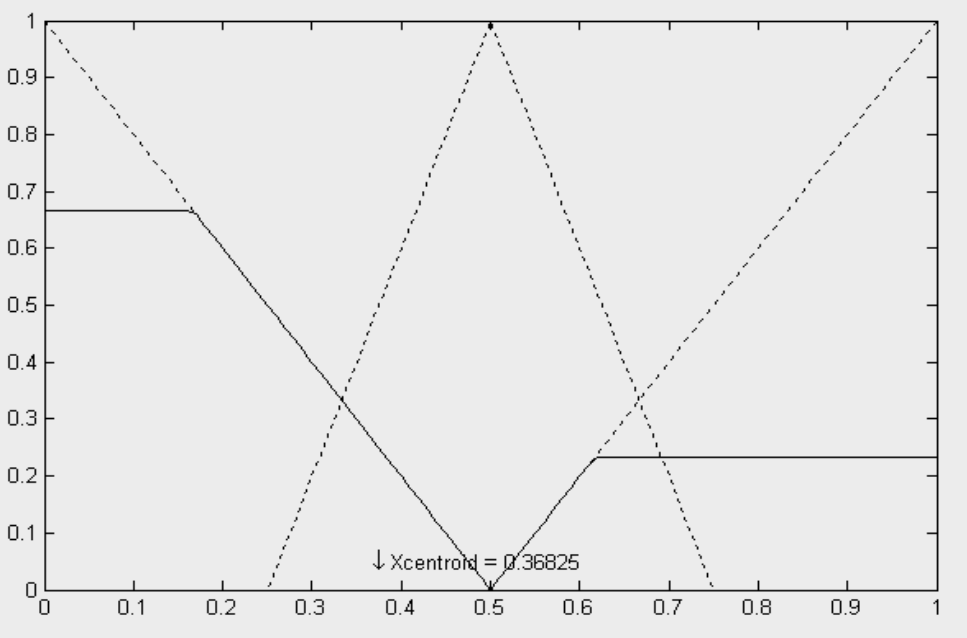



**Figura 6-8: Resultados da Defuzificação para http://www.scottrade.com/index.asp? supbid=68597**

A Tabela 6-3 mostra a ordenação das páginas obtidas a partir dos resultados da defuzificação. Observe que, nesses exemplos, os documentos *Web* com os maiores valores para o conjunto "bom" ocupam as primeiras posições. Entretanto, os graus de importância, quando são atribuídos, podem ou não afetar ou compensar os resultados finais, modificando a ordenação final.

**Tabela 6-3: Resultados da Defuzificação e da Ordenação dos Documentos Web**

| Contexto: *economia* | Resultados da Defuzificação | Ordenação |
|---|---|---|
| The World In 2007 | 0.62562 | 1º |
| Economist Conferences | 0.62225 | 2º |
| Economist.com Surveys | 0.56997 | 3º |
| Economist.com Opinion | 0.46810 | 4º |
| Scottrade | 0.36825 | 5º |

### 6.1.2 – Prova de Conceito Usando Regras *Fuzzy* de Inferência

Na seção 4.1.7.1 do capítulo 4 foi destacado que o *PC – Prognóstico Composto de Qualidade de Informação* também pode ser obtido pela inferência de regras *fuzzy*. O exemplo a seguir demonstra, passo a passo, a prova de conceito que emprega as regras *fuzzy* (BARROS, XEXÉO *et al.*, 2008a).

A contextualização considerada pelos especialistas na atribuição dos graus de importância de cada dimensão de qualidade é adotada de forma semelhante para definição das regras. A Tabela 6-4 fornece a base de regras *fuzzy* para o contexto *economia*. Ela contém 27 regras diferentes – as expressões lógicas *fuzzy* – que expressam os relacionamentos entre as variáveis lingüísticas e os conjuntos *fuzzy* compostos pelos termos lingüísticos. Nessa Tabela as variáveis de entrada (*PS*): a *completeza*, a *reputação* e a *atualidade* implicam a variável de saída *PC*.

Para relembrar, o exemplo a seguir ilustra a estrutura das regras *fuzzy*, destacado na Tabela 6-4:

*REGRA 1*:

SE as variáveis de entrada < *completeza* = *regular* E *reputação* = *boa* E *atualidade* = *boa* >

ENTÃO a variável de saída *PC* = *boa*.

**Passo 1 e Passo 2 –** Definições semelhantes às que foram realizadas na seção 6.1.1.

**Tabela 6-4: Base de Regras *Fuzzy* para o Contexto Economia**

| **Variáveis de Entrada (*PS*)** | | ***Atualidade*** | | |
|---|---|---|---|---|
| ***Completeza*** | ***Reputação*** | Ruim | Regular | **Boa** |
| Ruim | Ruim | ***Ruim*** | ***Ruim*** | ***Regular*** |
| | Regular | ***Ruim*** | ***Regular*** | ***Regular*** |
| | Boa | ***Regular*** | ***Regular*** | ***Regular*** |
| **Regular** | Ruim | ***Ruim*** | ***Regular*** | ***Regular*** |
| | Regular | ***Regular*** | ***Regular*** | ***Regular*** |
| | **Boa** | ***Regular*** | ***Regular*** | ***Boa*** |
| Boa | Ruim | ***Regular*** | ***Regular*** | ***Regular*** |
| | Regular | ***Regular*** | ***Regular*** | ***Boa*** |
| | Boa | ***Regular*** | ***Boa*** | ***Boa*** |
| | | **Variável de Saída (*PC*)** | | |

**Passo 3 –** Definição dos conjuntos *fuzzy* de saída *PC* que são obtidos por meio da inferência de regras *fuzzy*. Os conjuntos *fuzzy* de saída *PC* foram defuzificados pelos

*métodos LOM, MOM, BOA, SOM e COG* para obtenção dos valores finais mostrados na Tabela 6-5.

**Tabela 6-5: Resultados da Defuzificação pelos métodos LOM, MOM, BOA, SOM e COG (PC)**

| **Contexto: *economia*** | **Defuzificação (*PC*)** | | | | |
|---|---|---|---|---|---|
| ***Sites Web*** | ***LOM*** | ***MOM*** | ***BOA*** | ***SOM*** | ***COG*** |
| Economist.com Surveys | 0.61 | 0.5 | 0.5 | 0.39 | 0.5 |
| Economist Conferences | 1 | 0.5 | 0.5 | 0.34 | 0.5 |
| The World In 2007 | 0.61 | 0.5 | 0.5 | 0.39 | 0.5 |
| Economist.com Opinion | 0,63 | 0.87 | 0.81 | 0.74 | 0.805 |
| Scottrade | 0,57 | 0.075 | 0.16 | 0 | 0.176 |

Em vista do número reduzido de documentos da mostra, os resultados obtidos pelo método *LOM* foram escolhidos por apresentarem menos valores repetidos e serem, portanto, mais significativos para demonstração da fuzificação e da defuzificação pela inferência de regras *fuzzy*. Nesse contexto, foi possível observar, pelo conjunto de regras definidos, que a variável lingüística *atualidade* é considerada tão importante quanto a completeza e a reputação. Se os valores de *Authority* e dos *Hubs* forem considerados separadamente, isso resultará numa ordenação diferente dos documentos.

As Figuras 6-9 a 6-13 mostram o processo de fuzificação e de defuzificação dos documentos *Web* da Tabela 6-1. No topo das Figuras, podem ser identificados os valores dos metadados, bem como os *PC* que foram obtidos a partir da defuzificação. Além disso, são identificadas as 27 regras que foram previamente definidas, com destaque para aquelas que foram ativadas.

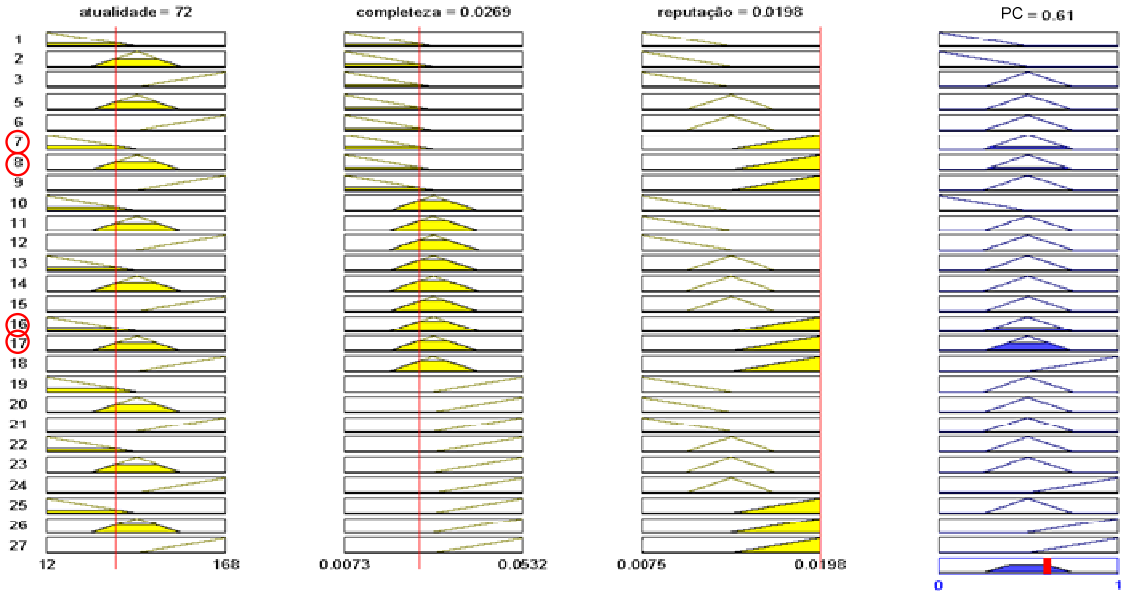



**Figura 6-9: Resultados da Defuzificação LOM para http://www.economist.com/surveys/**

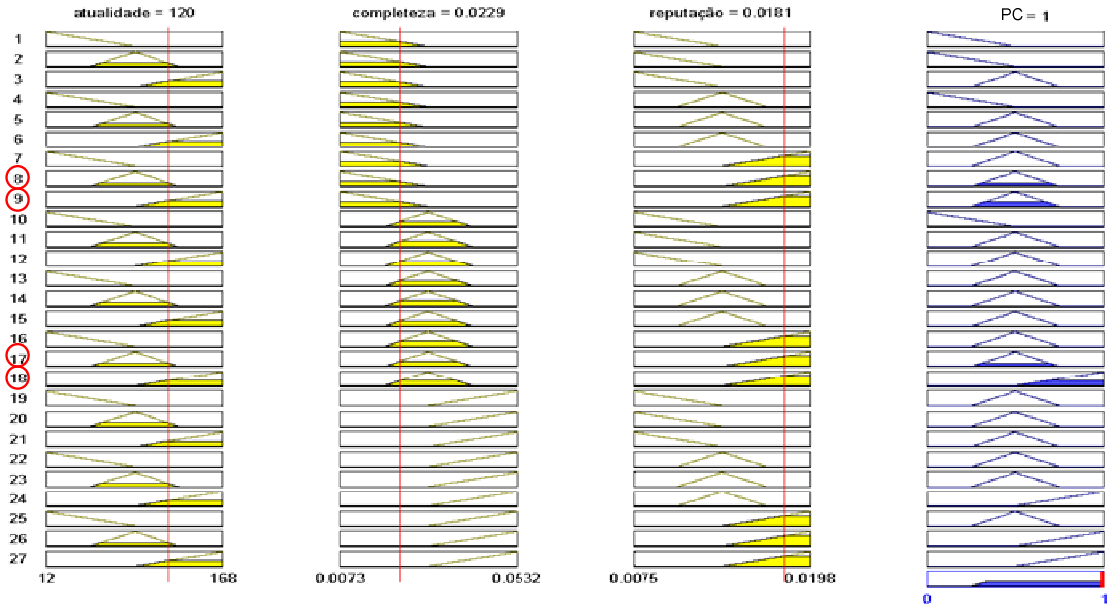



**Figura 6-10: Resultados da Defuzificação LOM para http://www.economistconferences.com/**

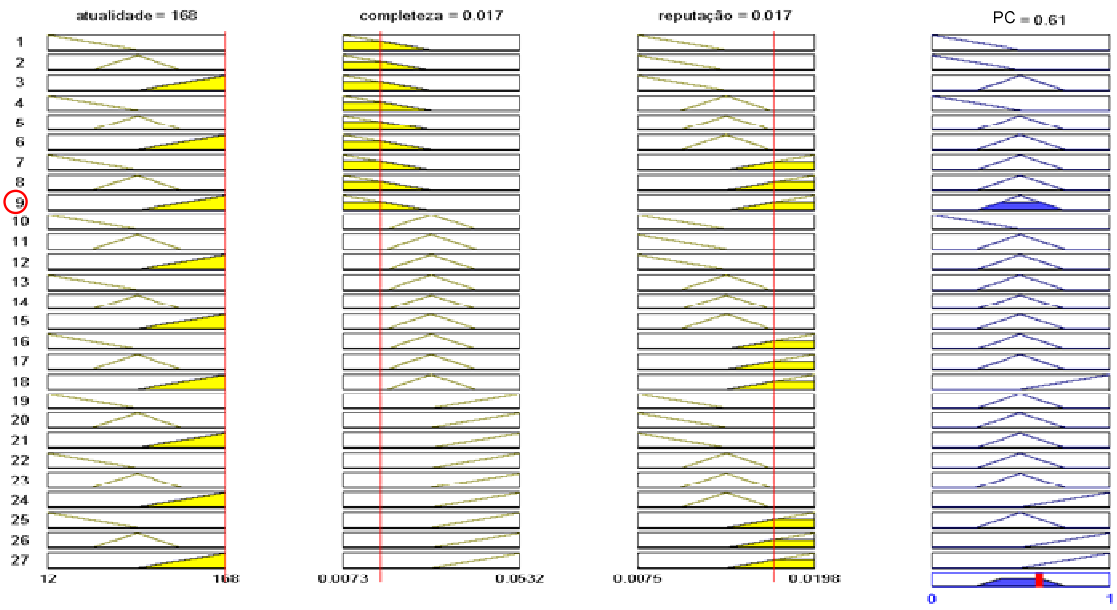



**Figura 6-11: Resultados da Defuzificação LOM para http://www.theworldin.com/**

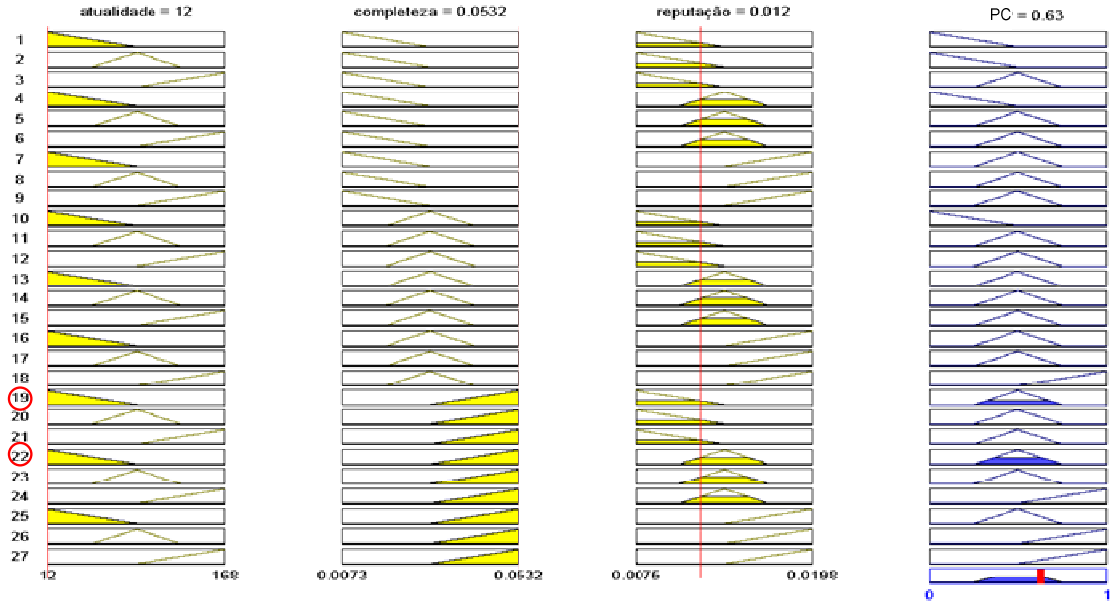



**Figura 6-12: Resultados da Defuzificação LOM para http://www.economist.com/opinion**

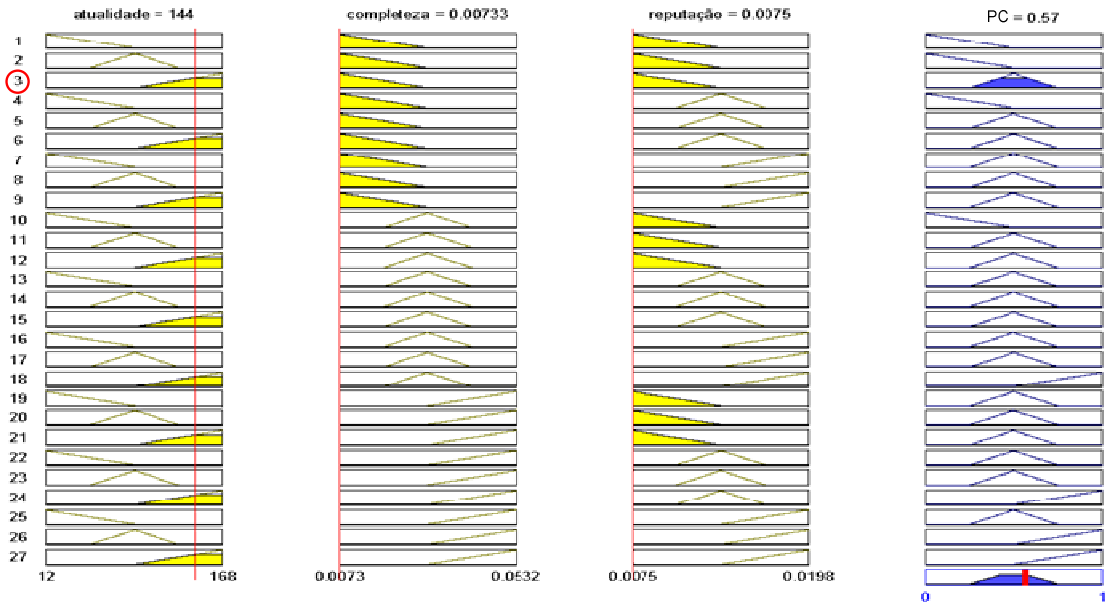



**Figura 6-13: Resultados da Defuzificação LOM para http://www.scottrade.com/index.asp?supbid=68597**

As Figuras 6-14 a 6-16 mostram as superfícies obtidas a partir das regras para a atualidade, reputação, completeza e *PC*. Cada figura mostra a associação entre duas variáveis lingüísticas. O resultado final corresponde ao *PC*. Baseado nessas figuras é possível analisar a contribuição de cada variável lingüística para o *PC* e, se necessário, as regras podem ser ajustadas considerando o contexto "economia".

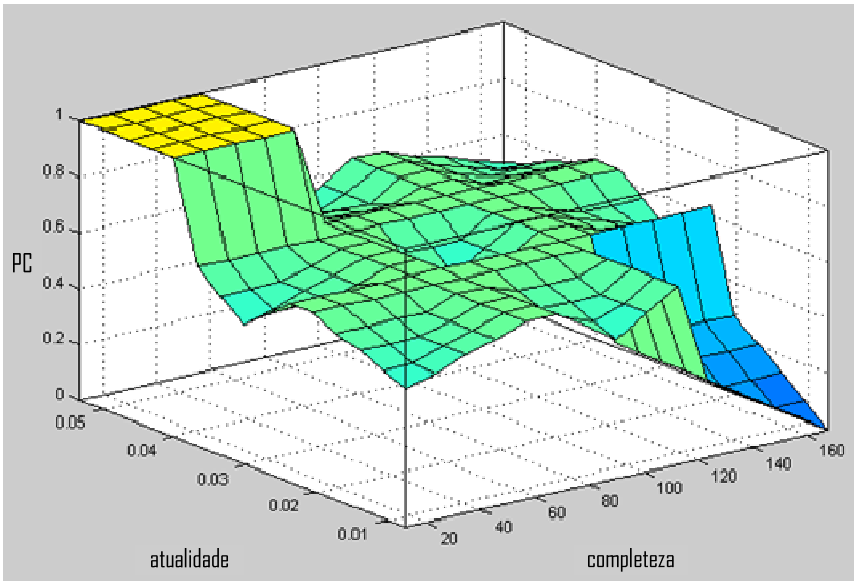



**Figura 6-14: Contribuição de Completeza e Atualidade para PC**

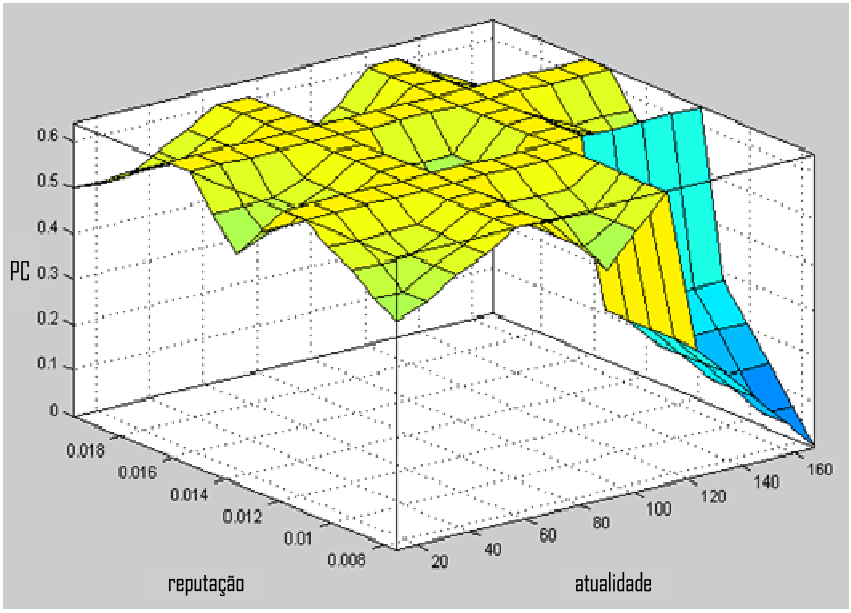



**Figura 6-15: Contribuição de Reputação e Atualidade para PC**

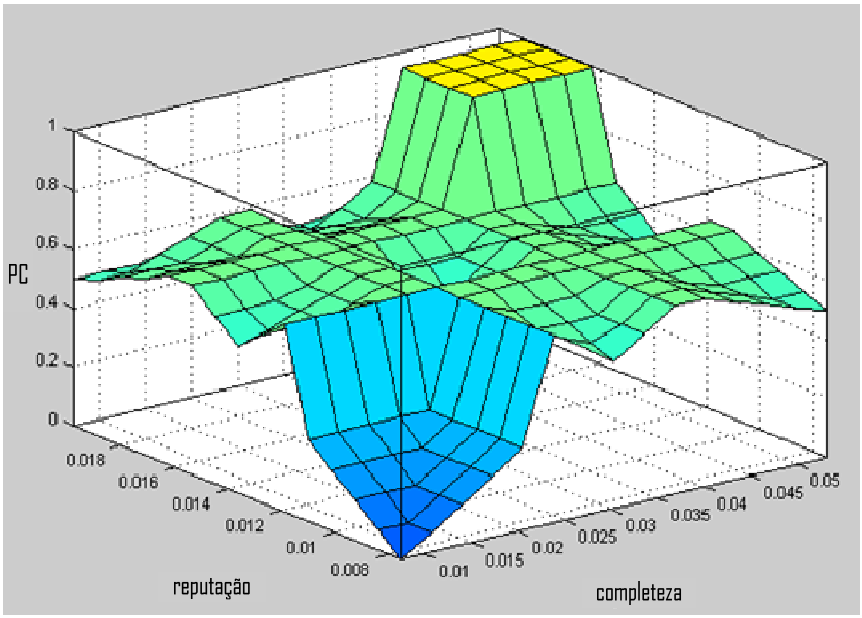



**Figura 6-16: Contribuição de Reputação e Completeza para PC**

## 6.2 – Aplicação Colaborativa para Construção de uma Base de Testes Contendo Páginas *Web* Avaliadas

Para que haja uma validação adequada da proposta para o prognóstico de qualidade de informações na *Web*, é fundamental dispor de um *dataset* (base de dados testes). Nesse sentido, percebe-se a necessidade de se ter em mãos uma coleção de páginas em que se conheçam exatamente as suas avaliações, dentro de um contexto

específico e pré-definido. Outro fator importante no qual a construção de *datasets* é útil é a comparação de diferentes mecanismos a fim de mostrar a eficácia de cada um deles. Assim, esses diferentes mecanismos podem ser testados e comparados de maneira independente, usando o mesmo conjunto de documentos.

Entretanto, a aquisição ou a construção de *datasets* demandam um custo elevado ou requerem muito trabalho. Isso se deve ao fato de que um bom *dataset* precisa conter uma grande quantidade de documentos (centenas ou milhares), além das informações que associam os documentos as suas avaliações. Normalmente, isso é feito através de um processo em que a alimentação e a avaliação de cada página são manuais.

Esta seção descreve o desenvolvimento de uma aplicação colaborativa para a construção de bases de testes contendo páginas *Web* que foram automaticamente avaliadas por meio da nossa proposta.

### 6.2.1 – Visão Geral

O desenvolvimento e a implementação de uma aplicação colaborativa tiveram como objetivo aliar ao processo de alimentação manual das páginas, a alimentação e o prognóstico de qualidade realizados automaticamente, com o propósito de tornar mais fácil e mais ágil o trabalho de construção das bases de testes propostas na especificação do FoxSet (BARROS, RODRIGUES-NT *et al.*, 2009).

Para um melhor entendimento do que foi desenvolvido, as seções a seguir descrevem os distintos papéis dos usuários, o processo de avaliação das páginas e como tal avaliação é integrada na construção dos *datasets*. Também são descritos os mecanismos de avaliação: um automático feito pelo sistema e outro manual realizado pelo avaliador do *dataset*.

### 6.2.2 – Avaliação Colaborativa

A colaboração na avaliação é viabilizada mediante a extensão do navegador Mozilla Firefox[58], por meio do qual é realizada a maioria das interações dos usuários.

Uma das tarefas que podem ser feitas de forma cooperativa é a coleta de documentos *Web*. Os usuários podem adicionar documentos a um *dataset* em particular

[58] http://www.mozilla.com/en-US/firefox/.

durante o processo de busca, por intermédio de uma opção no Foxset. Há também os diferentes perfis de usuários que irão definir as suas ações no processo de construção do *dataset*.

Alternativamente, a tarefa de coleta de documentos pode ser facilitada e realizada de forma semi-automática por um serviço do sistema. O serviço é capaz de percorrer a *Web* a partir de um documento inicial (semente) ou por meio de outros mecanismos de busca mais conhecidos, em um contexto definido. Nesse caso, obedecendo aos requisitos de qualidade pré-estabelecidos pelos usuários, os documentos são recuperados e automaticamente avaliados.

Depois dessa fase, os usuários podem colaborar na construção de diferentes subconjuntos de dados, a partir do *dataset* avaliado, e torná-los disponíveis para que outros usuários possam empregá-los em suas pesquisas.

### 6.2.3 – FoxSet

O FoxSet é uma ferramenta para a construção de *datasets* e consiste em um *plugin* para o Firefox, conforme mostrado na Figura 6-17. Ele tem como objetivo facilitar o gerenciamento do processo de construção de *datasets,* além de armazenar e disponibilizar os conjuntos de páginas *Web* avaliadas em relação a um determinado contexto.

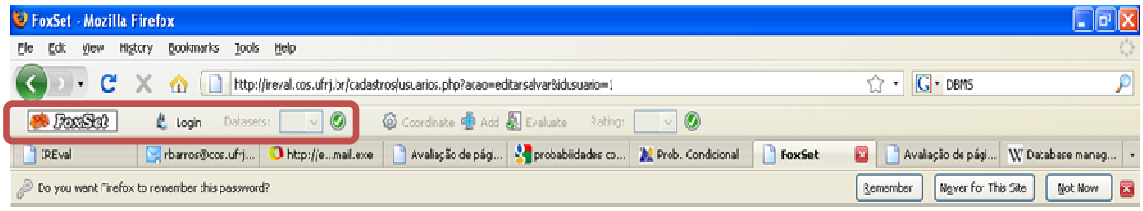



**Figura 6-17: Plugin Foxset para o Firefox**

A Figura 6-18 ilustra o processo de construção de *datasets* e suas etapas de criação, alimentação, avaliação, finalização e uso definidos no FoxSet.

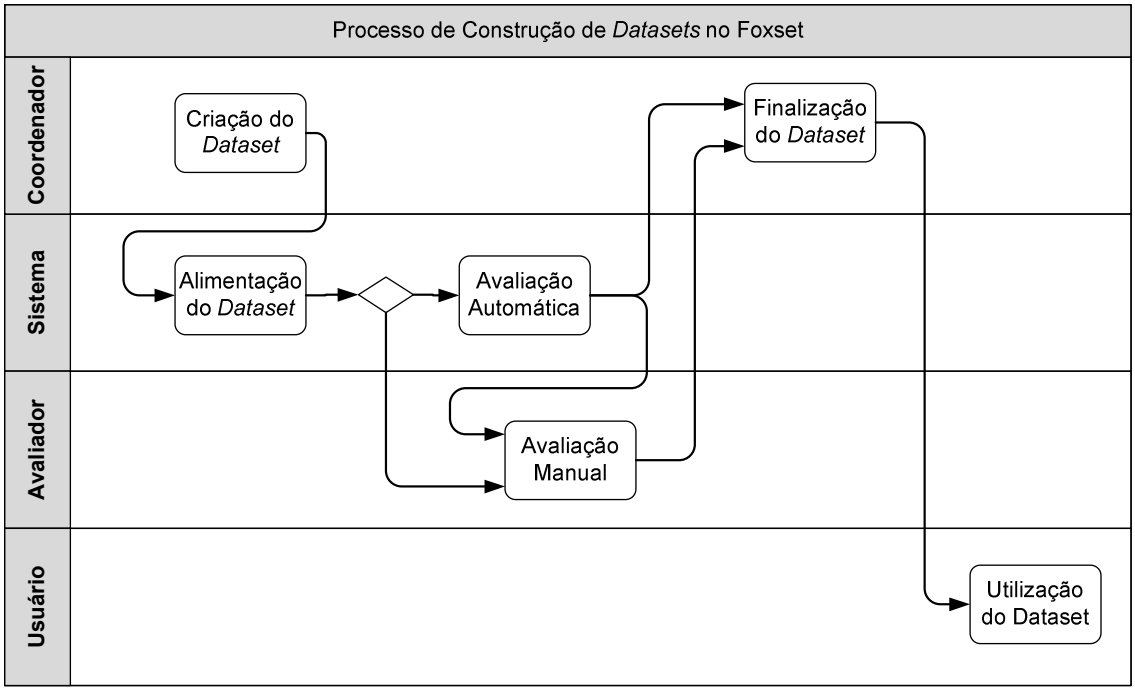



**Figura 6-18: Processo de Construção de Datasets no Foxset**

**Permissões e perfis –** Visando a um melhor entendimento, as pessoas que interagem diretamente com o sistema são denominadas colaboradores, visto que o termo "usuário" foi utilizado para definir um perfil de utilização do *dataset*. As permissões de acesso são concedidas para o sistema como um todo, mas os perfis são específicos para cada *dataset*. Existem somente duas permissões de acesso irrestrito ao sistema: a que concede direitos de administrador e outra para os direitos de coordenador. Os direitos de administrador dão a um colaborador a habilidade de criar, modificar e remover perfis de colaboradores. Os direitos do coordenador dão a um colaborador a habilidade de projetar um *dataset*, isto é, iniciar o seu processo de construção.

A Figura 6-19 mostra os três perfis existentes no FoxSet: coordenador, avaliador e usuário. Esses perfis são específicos para cada *dataset*, isto é, um colaborador pode coordenar um *dataset*, mas, ao mesmo tempo, ser usuário num outro *dataset*. Os coordenadores são responsáveis pelas atividades gerenciais no processo de construção dos *datasets*, tais como a criação, a atribuição de perfis e a finalização. Eles também são os únicos que podem executar as tarefas de escolher o contexto do dataset, definir como a avaliação automática será feita, selecionar as escalas de avaliação, definir as perguntas e remover da coleção os documentos *Web* inadequados. Quando um colaborador com direitos de coordenador cria um *dataset*, ele terá automaticamente o perfil de coordenador do *dataset* criado. Os avaliadores são os colaboradores que coletam

documentos *Web* cooperativamente (se a coleta e avaliação automática não estão sendo usadas) e, mais tarde, avaliam os documentos *Web* selecionados. Finalmente, os usuários são os colaboradores que estão interessados em usar um *dataset* que esteja finalizado. Eles estão habilitados a criar diferentes subconjuntos de um *dataset* ajustando alguns critérios e podem exportar estes subconjuntos em formatos diferentes, tais como XML, para um uso posterior.

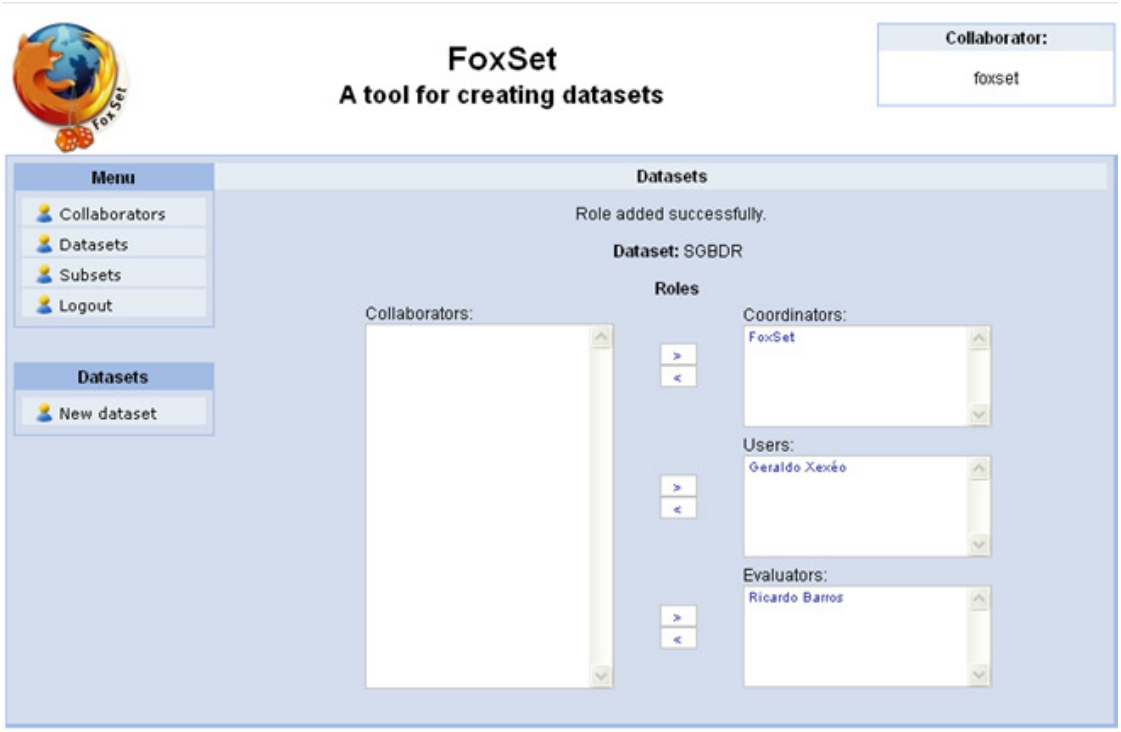



**Figura 6-19: Perfis do Foxset**

**Criação do *dataset* –** O coordenador é responsável por essa primeira etapa. Nela é definido a que contexto os documentos pertencem e se o *dataset* será populado manualmente ou automaticamente. Ele também escolhe o idioma, o número mínimo de páginas do *dataset* e as dimensões de qualidade que serão avaliadas com seus graus de importância, de acordo com o contexto definido e o nível de relevância desejado. Nessa primeira versão, foram implementadas as avaliações para as dimensões *completeness* (completeza), *reputation* (reputação) e *timeliness* (atualidade). Há também a definição das perguntas para o *dataset*, mas a implementação permite que elas sejam definidas em qualquer momento entre criação e a avaliação manual do *dataset*. As Figuras 6-20 a 6-22 mostram as definições dessa etapa.

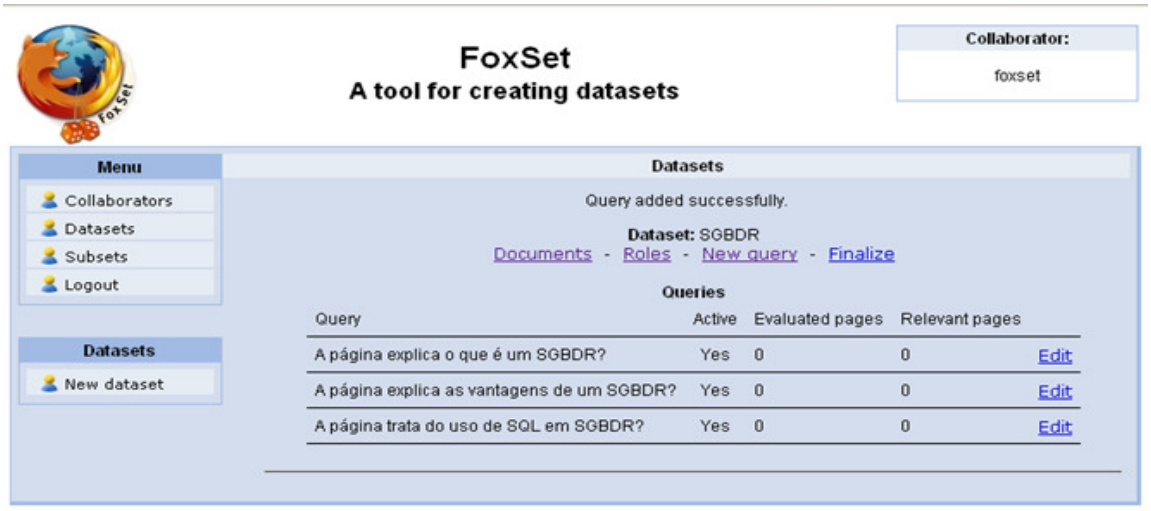



**Figura 6-20: Definição das Perguntas para um *Dataset***

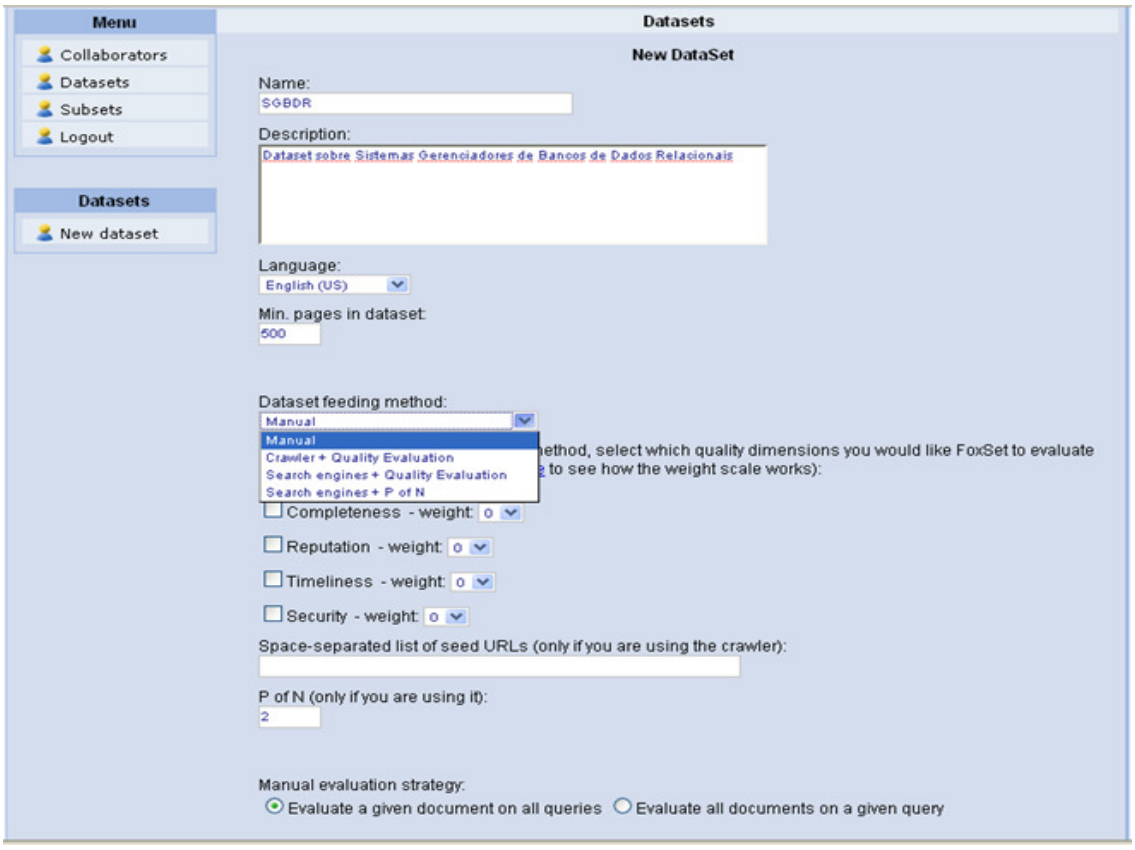



**Figura 6-21: Parâmetros de Criação do *Dataset***

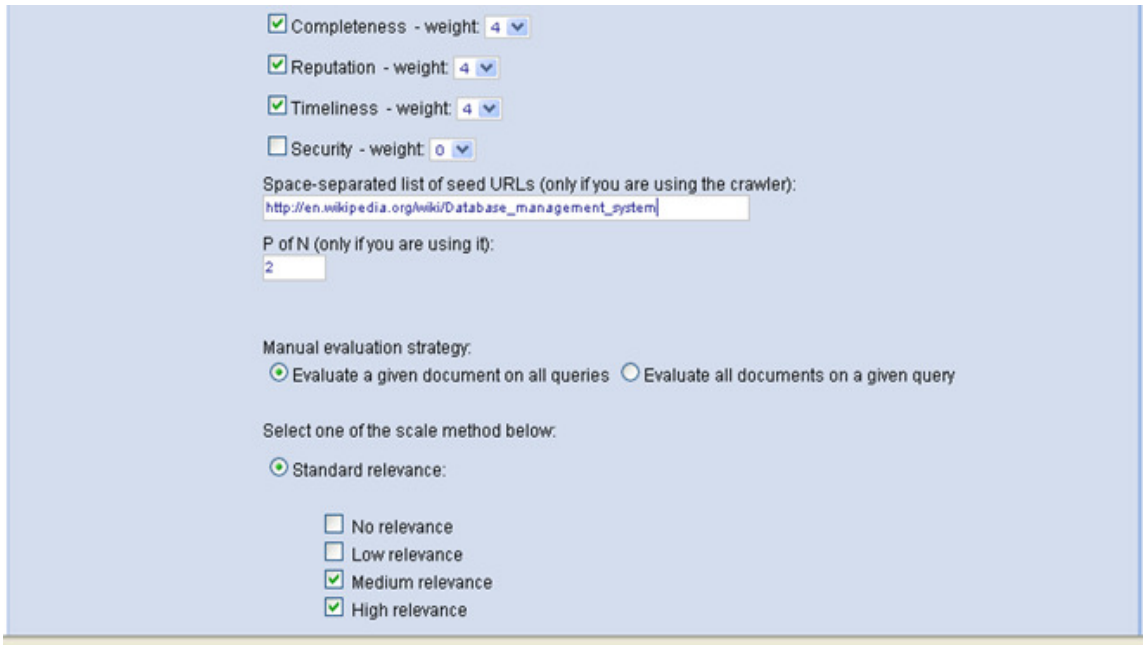



**Figura 6-22: Parâmetros de Criação do *Dataset***

**Alimentação do *dataset* –** Essa etapa pode ser feita manualmente pelos avaliadores ou automaticamente pelo *crawler*. A Figura 6-20 mostra essa definição (*dataset feeding method*).

**Avaliação automática –** O objetivo principal dessa etapa é fornecer ao coordenador uma avaliação automática de cada página baseada nos aspectos de qualidade considerados importantes para o contexto do *dataset*. Essa avaliação ajudará o coordenador na seleção de quais documentos devem pertencer ao *dataset*. Ela é feita usando as dimensões de qualidade descritas previamente e obedece aos passos definidos na seção 6.1.1. As funções de pertinência foram implementadas no MATLAB® usando os números *fuzzy* triangulares, para distribuir os valores do domínio por cinco termos lingüísticos, mostrados na Figura 6-23.

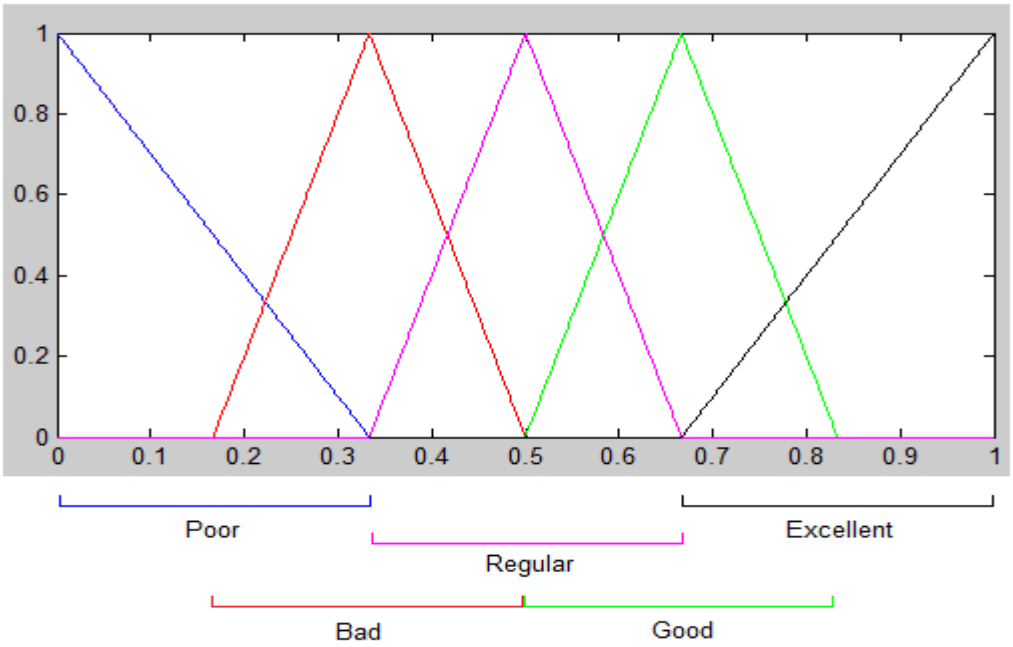



**Figura 6-23: Definição Números Fuzzy Triangulares**

**Avaliação manual –** Essa etapa é executada de forma cooperativa pelo coordenador e pelos diversos avaliadores. Ela começa com o coordenador que revê a coleção de documentos *Web*. Se a avaliação automática foi executada, o coordenador pode filtrar os documentos por meio das suas avaliações. Isso permite que uma boa quantidade de documentos irrelevantes sejam avaliados manualmente e, assim, tornem-se parte do *dataset* final. Depois disso, a avaliação colaborativa começa com cada avaliador recebendo um conjunto de documentos *Web* e um conjunto de perguntas, para avaliar cada documento em relação a cada pergunta. De acordo com o número de avaliadores e a quantidade de páginas presentes no *dataset* selecionado, o Foxset faz uma distribuição por igual da quantidade de páginas a serem avaliadas, uma seleção aleatória dessas páginas e então as fornece ao avaliador. O avaliador irá responder, para cada pergunta, se a página possui ou não a resposta para a pergunta formulada. Dependendo das necessidades de construção do *dataset*, as perguntas podem ser respondidas de duas formas arbitradas pelo coordenador: por meio de uma variável binária (“sim” ou “não”) ou por meio de uma variável nominal que pode ser composta por uma escala de valores, como por exemplo: *no info* – não responde; *mentions* – menciona; *partially explains* – responde parcialmente; e *fully explains* – responde completamente. Ambas as escalas indicam se o documento é ou não relevante em relação à questão proposta. As Figuras 6-24 a 6-26 ilustram essa tarefa.

Ainda que haja a liberdade para definir as escalas da avaliação, ressalta-se que qualquer escala definida pelo coordenador está associada aos níveis ordinais de medida, isto é, elas sempre podem ser ordenadas e permitem que sejam determinadas as medianas, indicadas as modas, mas não calculadas as médias das avaliações. Por exemplo, considerando que o coordenador pode definir que cada combinação de documento e pergunta seja avaliada por mais de um avaliador, uma mediana pode ser determinada para agregar essas diferentes avaliações. Isso ocorre, embora as questões referentes aos pesos e distinções de opiniões entre os usuários ainda não tenham sido tratadas nesta especificação do Foxset. Além disso, também não foram especificados e fornecidos mecanismos para identificar, nem para evitar comportamentos maliciosos dos usuários. Por hora, foi assumido que o FoxSet somente está sendo usado em ambientes controlados, tais como equipes de pesquisa. Na medida do amadurecimento e uso do FoxSet, as necessidades de especificação de tais mecanismos certamente irão surgir.

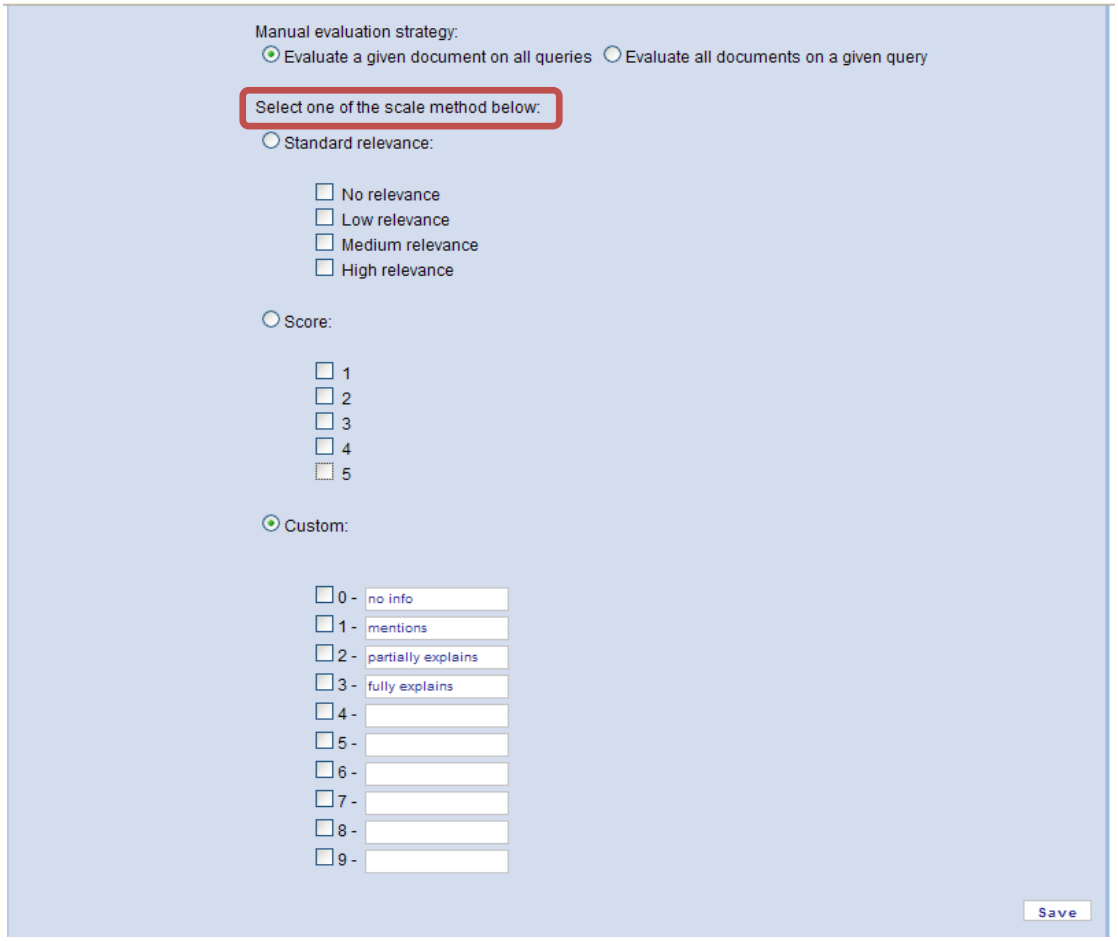



**Figura 6-24: Definição das Escalas de Avaliação para um *Dataset***

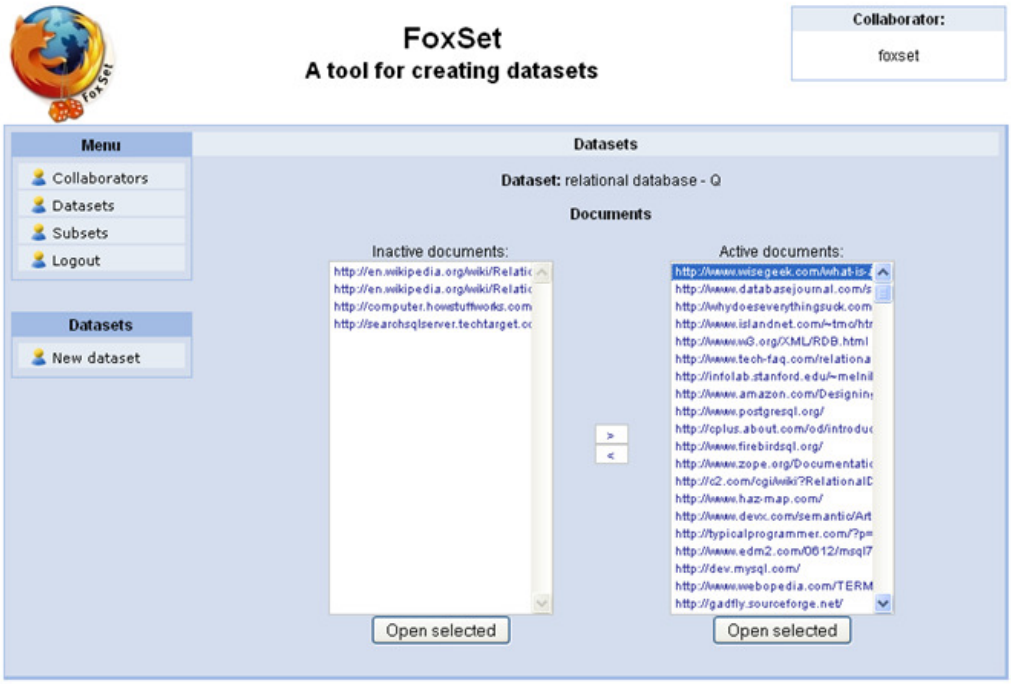



**Figura 6-25: Avaliação Manual de um Dataset**

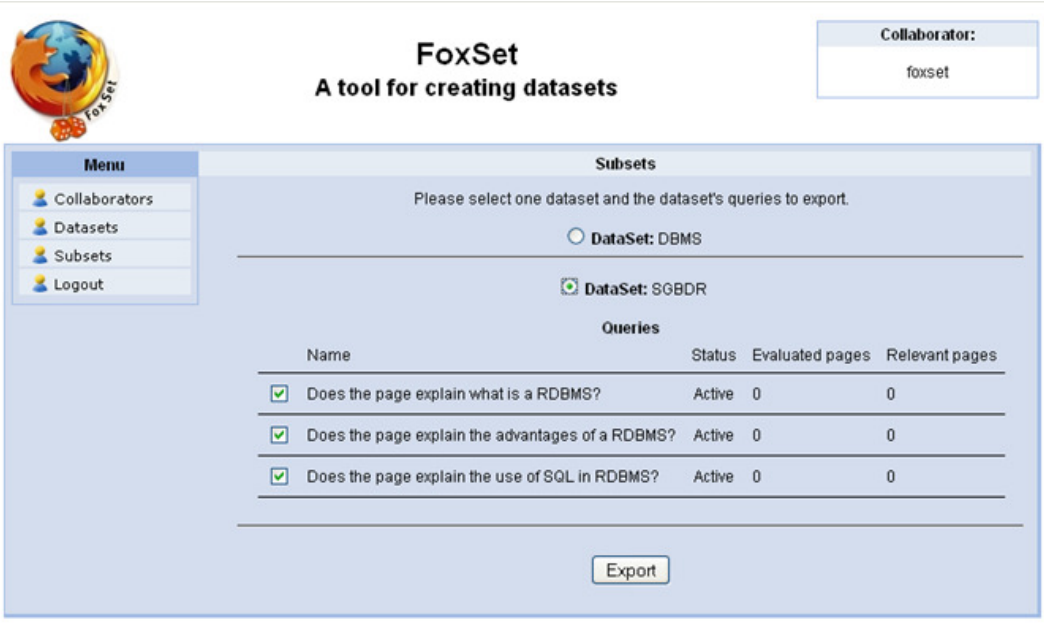



**Figura 6-26: Seleção do Conjunto de Perguntas para Avaliação Manual de um *Dataset***

**Finalização do Dataset –** A etapa de finalização do *dataset* é feita pelo coordenador quando todas as perguntas foram definidas e todos os documentos foram avaliados para cada pergunta. Uma vez finalizado, o *dataset* não pode ser modificado. Antes de finalizar, o coordenador pode rever o *dataset* e fazer algumas modificações

finais, removendo as páginas que não estejam de acordo com os critérios estabelecidos por ele, adicionando novas páginas ao *dataset* através do *crawler* e removendo algumas perguntas. A Figura 6-27 mostra a opção de finalização.

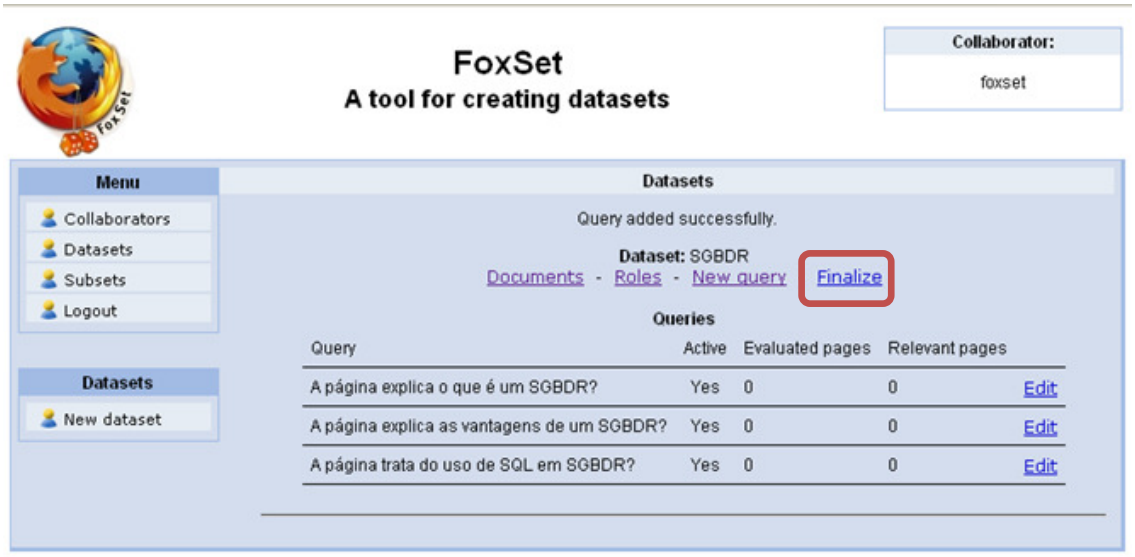



**Figura 6-27: Finalização do *Dataset***

**Uso do dataset –** Nessa etapa, o usuário pode refinar o *dataset* original removendo as perguntas ou filtrando documentos de acordo com alguns critérios. Uma vez criado o subconjunto, ele pode ser exportado para o uso externo, em formatos diferentes, tais como XML.

### 6.2.4 – Resultados Preliminares

A seção 6.3 a seguir descreve um experimento que foi conduzido para uma avaliação mais ampla da abordagem proposta. Paralelamente, foram obtidos alguns resultados preliminares por meio de um estudo de caso simples para demonstrar o uso do Foxset num contexto específico (BARROS, RODRIGUES-NT *et al.*, 2009). Primeiramente, o coordenador definiu o contexto "economia", selecionou três dimensões de qualidade disponíveis e atribuiu a cada uma delas um grau de importância – 3 para o *timeliness*, 4 para a *reputation* e 2 para a *completeness*. A decisão foi usar a alimentação automática do *dataset* e foi fornecida a *URL* de um documento como semente (*http://www.economywatch.com/*). Em seguida, foi definida a pergunta *Q1* "*Quais são os impactos da atual crise financeira?*". Com base nessas definições, um novo *dataset* foi criado, com o número de documentos estipulado pelo coordenador, no caso 500. Finalmente, foi realizada a avaliação automática com os resultados variando de 0.01244 até 0.43126. Antes de a avaliação manual ser iniciada, o coordenador reviu e

filtrou os documentos do *dataset* que possuíam resultados de 0.324 ou menos, para manter somente os cinco documentos melhores classificados. Esses documentos são mostrados na Tabela 6-6.

**Tabela 6-6: Melhores Resultados Selecionados pelo Coordenador do *Dataset***

| ID | URL do Documento | Resultado |
|---|---|---|
| D1 | http://www.economywatch.com/ | 0.43126 |
| D2 | http://www.brazzil.com/ | 0.32707 |
| D3 | http://www.scps.nyu.edu/areas-of-study/global-affairs/ | 0.32507 |
| D4 | http://www.hindustantimes.com/ | 0.32469 |
| D5 | http://www.csmonitor.com/world/ | 0.32466 |

Durante a avaliação colaborativa, dez avaliadores avaliaram esses documentos, para pergunta *Q1*. A

Tabela 6-7 mostra a quantidade de escalas de avaliação, atribuídas pelos avaliadores à pergunta *Q1* por documento. Ela mostra também as medianas determinadas por documento. Ressalta-se que os documentos com as medianas "*Fully explains*" e "*Partially explains*" – que foram selecionadas pelo coordenador como indicadores de relevância do documento – foram, como esperado, incluídos como respostas à pergunta *Q1*.

**Tabela 6-7: Avaliação Colaborativa para a pergunta Q1**

| Scale value | D1 | D2 | D3 | D4 | D5 |
|---|---|---|---|---|---|
| **No info** | 0 | 4 | 10 | 5 | 1 |
| **Mentions** | 0 | 6 | 0 | 4 | 1 |
| **Partially explains** | 3 | 0 | 0 | 1 | 6 |
| **Fully explains** | 7 | 0 | 0 | 0 | 2 |
| **Mediana** | **Fully explains** | **Mentions** | **No info** | **Mentions** | **Partially explains** |

Após a avaliação manual, o coordenador fez os ajustes finais e finalizou o *dataset*. A Figura 6-28 apresenta um subconjunto resultante do exemplo acima que foi exportado para XML.

```
<dataset id="1" name="Economy">
- <queries>
  - <query id="1" name="What are the impacts of the current financial crisis?">
      <document id="1" />
      <document id="5" />
    </query>
  </queries>
- <documents>
    <document id="1" url="http://www.economywatch.com/" />
    <document id="2" url="http://www.brazzil.com/" />
    <document id="3" url="http://www.scps.nyu.edu/areas-of-study/global-affairs/" />
    <document id="4" url="http://www.hindustantimes.com/" />
    <document id="5" url="http://www.csmonitor.com/world/" />
  </documents>
</dataset>
```

**Figura 6-28: Subconjunto em XML Resultante do Dataset "Economia"**

Além do estudo de caso demonstrado, foi conduzido outro estudo, igualmente simples, que consistiu na comparação dos resultados de ordenação de dois conjuntos de páginas para o contexto "bancos de dados relacionais".

A Tabela 6-8 mostra a ordenação do primeiro conjunto de páginas que foi obtida por meio do nosso mecanismo de avaliação automática baseada nas três dimensões de qualidade selecionadas *timeliness*, *reputation* e *completeness*. Foi atribuído o grau de importância 1 (a dimensão de qualidade avaliada tem pouca importância) para as três dimensões, visando a ordenação dos valores básicos.

A ordenação do segundo conjunto de páginas foi obtida por meio dos tradicionais mecanismos de buscas do Google®, do Yahoo® e do Live Search®. Para obtenção de cada um dos conjuntos, foram usadas expressões regulares.

Tomando-se os conjuntos de páginas resultantes dos três buscadores, o conjunto unificado resultante foi obtido por meio do critério "*P de N*", onde *P* é o número de conjuntos em que uma mesma página ocorre, e *N* é o número total de conjuntos. No nosso estudo, foram admitidas as páginas retornadas em pelo menos dois dos três mecanismos utilizados ($P = 2$ e $N = 3$). Na etapa seguinte, a avaliação automática e a ordenação do conjunto de páginas foram realizadas de forma semelhante ao tratamento dado ao primeiro conjunto.

A partir dos resultados obtidos da 1ª à 5ª posições de ordenação, observa-se que os documentos *D1*, *D2* e *D3* ocorreram nos dois conjuntos e foram mantidas as posições de ordenação para os documentos *D1* e *D2*. No segundo conjunto, dois novos

documentos *D6* e *D7* obtiveram a 3ª e 5ª posições, e os documentos *D4* e *D5* não aparecem, perdendo a 4ª e 5ª posições. O documento *D3* caiu da 3ª para 4ª posição.

**Tabela 6-8: Avaliação Automática com Ordenação Básica**

| Ordenação | ID | Resultado | URL do Documento |
|---|---|---|---|
| 1º | D1 | 0.30064 | 'http://en.wikipedia.org/wiki/Relational_database' |
| 2º | D2 | 0.28752 | 'http://en.wikipedia.org/wiki/Relational_database_ management_system' |
| 3º | D3 | 0.10886 | 'http://computer.howstuffworks.com/question599.htm' |
| 4º | D4 | 0.10785 | 'http://www.agiledata.org/essays/relationalDatabases.html' |
| 5º | D5 | 0.10782 | 'http://www.amazon.com/Server-Relational-Database-Design-Implementation/dp/143020866X' |

**Tabela 6-9: Avaliação Automática com Ordenação P de N (*P* = 2 e *N* = 3)**

| Ordenação | ID | Resultado | URL do Documento |
|---|---|---|---|
| 1º | D1 | 0.89218 | 'http://en.wikipedia.org/wiki/Relational_database' |
| 2º | D2 | 0.62313 | 'http://en.wikipedia.org/wiki/Relational_database_ management_system' |
| 3º | D6 | 0.44129 | 'http://www.answers.com/topic/relational-database' |
| 4º | D3 | 0.32015 | 'http://computer.howstuffworks.com/question599.htm' |
| 5º | D7 | 0.24542 | 'http://searchsqlserver.techtarget.com/sDefinition/0,,sid87_ gci212885,00.html' |

## 6.3 – Experimento para Avaliação da Abordagem Teórica Proposta

### 6.3.1 – Definição

A avaliação da abordagem teórica proposta nesta tese foi realizada com o auxílio de um método experimental de comparação dos resultados. O estudo foi desenvolvido sob a ótica de um pesquisador, avaliando a viabilidade de utilização da estratégia mencionada anteriormente. Durante a análise, foram comparados os resultados das avaliações automáticas de um *dataset* obtido seguindo a nossa metodologia, em contrapartida aos resultados derivados de um conjunto de avaliações obtidos por meio de julgamento humano para esse mesmo *dataset*.

A seguir são descritas as técnicas e estatísticas utilizadas no planejamento, execução e análise do experimento (TRAVASSOS, BARROS *et al.*, 2002).

### 6.3.2 – Planejamento

No planejamento inicial, a população do *dataset* utilizado em ambas as avaliações seria composta por 350 páginas, – aplicáveis a um contexto de "bancos de dados relacionais" –, a serem avaliadas por 105 pessoas. Um algoritmo foi desenvolvido para distribuir aleatoriamente dez páginas para cada avaliador, sendo que cada página receberia três avaliações de diferentes avaliadores. Os avaliadores são pessoas que possuem graduação, ou são alunos de graduação ou pós-graduação que conhecem ou atuam em atividades acadêmicas ou profissionais relacionadas ao contexto.

Em vista da dificuldade encontrada na obtenção das avaliações manuais para o conjunto completo de páginas, a amostra do *dataset* para o estudo foi composta pelas 84 páginas com pelo menos uma avaliação. Desse conjunto de 84 páginas, foram retiradas nove páginas inconsistentes (*links* quebrados e avaliações incompletas). O Anexo II apresenta essa amostra com os resultados das avaliações, automática e manual, utilizadas nas análises em estudo.

Essas análises foram realizadas com base em métodos estatísticos não-paramétricos, por meio dos quais foram adotadas as seguintes técnicas:

- Conversão e agregação dos valores ordinais da escala de Likert para as variáveis qualitativas (de ruim até excelente) em valores de escala intervalares para variáveis quantitativas;
- Gráfico de colunas para análise da distribuição das freqüências e para apresentação da freqüência relativa (ou percentual) de ocorrência dos dados, dividindo estes em um conjunto de classes distintas;
- Medidas de dependência - Coeficiente de correlação de Pearson para definição da correlação entre os conjuntos e os diagramas de dispersão;

#### *6.3.2.1 – Análise de Ameaças à Validade dos Resultados*

Em todos os estudos experimentais, existem ameaças que podem afetar a validade dos resultados. Por essa razão, devem ser analisadas as ameaças que um estudo sofre em decorrência do seu projeto. As ameaças relacionadas a esse estudo são apresentadas a seguir, classificadas em quatro categorias: validade interna, validade externa, validade de conclusão e validade de construção (WOHLIN, RUNESON *et al.*, 2000).

**Validade Interna**: Essas ameaças dizem respeito à identificação de outros fatores além dos tratamentos que possam ter provocado os resultados. Nessa categoria é questionado se a estrutura do estudo para verificar eventuais diferenças observadas podem ser realmente atribuídas aos tratamentos e não aos fatores fora do controle no estudo. Neste estudo, consideram-se três principais ameaças que representavam um risco de interpretação imprópria dos resultados: (1) Desgaste e casualidades do ambiente; (2) Moral e (3) Classificação de experiência e seleção dos participantes.

Em relação à primeira ameaça, alguns participantes podem reagir de forma negativa, pois a avaliação exige um esforço relativo. Como um mecanismo de proteção a essa ameaça, a aplicação desenvolvida para avaliação buscou a facilidade de acesso e de uso, reduzindo o número de respostas ao mínimo necessário para condução dos testes. Foram buscadas, também, a segurança e confiabilidade do ambiente, evitando-se ao máximo as falhas e interrupções.

Em relação à segunda ameaça, alguns participantes podem reagir de forma negativa quando o estudo não vai lhes proporcionar algum benefício. Isso foi observado, pois muitos participantes deixaram de responder ou completar as avaliações, ou, até mesmo avaliaram as páginas de forma inconsistente, apesar das insistentes solicitações formuladas durante as avaliações.

Em relação à terceira ameaça, os grupos de participantes do estudo devem estar balizados em termos de seu conhecimento e competências. Foi realizada, portanto, uma seleção dos avaliadores com perfis definidos no item 6.3.2, e uma auto-avaliação com base nas suas formações e experiências. Além disso, foi escolhido um tema de conhecimento básico – bancos de dados relacionais –, e as perguntas foram reduzidas a um nível mínimo de complexidade, para um experimento de curta duração.

Entretanto, ainda não foram especificados os mecanismos para o tratamento da gestão de competências para os avaliadores. Esse requisito foi identificado como uma perspectiva de trabalho futuro.

**Validade Externa**: Essas ameaças dizem respeito à capacidade de generalizar os resultados obtidos no estudo para uma população maior que a dos participantes. No que se refere a esse tipo de ameaça, duas questões precisam ser consideradas: (1) Normalmente ambientes acadêmicos não simulam totalmente as condições existentes em um ambiente do mundo real; (2) Restrições de tempo.

Apesar das limitações existentes no ambiente acadêmico no qual foi desenvolvido o estudo, o conjunto de páginas avaliado foi recuperado diretamente da *Web* e os avaliadores possuíam, no mínimo, conhecimento relativo em bancos de dados relacionais, com perfis tanto acadêmico, como profissional. Apesar disso, esse risco não pôde ser totalmente eliminado, em vista do baixo comprometimento de alguns avaliadores com o experimento.

No que se refere à segunda ameaça, existem riscos como impor ou suprimir restrições de tempo. Por essa razão, foi concedido aos usuários um tempo de três semanas, considerado razoável, para que pudessem completar as suas avaliações.

**Validade de Conclusão**: Essas ameaças dizem respeito à relação entre o tratamento e o resultado, em termos de significância estatística. Neste estudo, o maior problema é o tamanho da amostra, com um número reduzido de páginas no *dataset*, em relação à *Web* como um todo. Esse número não é o ideal do ponto de vista estatístico. Amostras reduzidas são um problema conhecido em estudos experimentais, principalmente nos casos em que eles envolvem julgamentos humanos. Em vista da dificuldade encontrada na obtenção das avaliações manuais para o conjunto completo de páginas, o *dataset* para o estudo foi composto pelas 84 páginas com pelo menos uma avaliação. Além disso, toda conclusão estatística possui uma margem de erro e a própria capacidade do método estatístico de chegar a uma conclusão incorreta. Uma taxa de erro alta pode implicar uma conclusão questionável. Devido a esses fatos, há a possibilidade de limitação nos resultados, sendo estes considerados não conclusivos e sim indícios.

Em razão dessas ameaças, foram escolhidos mecanismos de avaliação estatística adequados ao projeto do estudo e às escalas das métricas e variáveis. Os dados coletados e os resultados foram documentados e publicados para que as análises possam ser repetidas por outros pesquisadores.

**Validade de Construção**: Essas ameaças dizem respeito aos objetos e participantes do estudo, questionando se eles realmente refletem a questão que está sendo abordada. No que se refere a essa categoria de ameaça de validade, duas questões precisam ser consideradas: (1) Definição dos indicadores; (2) Expectativa do pesquisador.

A primeira delas questiona se a teoria, as variáveis e os tratamentos estão bem definidos, e se os instrumentos do estudo são adequados. Por essa razão, os indicadores

adotados neste estudo – os metatados e as dimensões de qualidade – são considerados plenamente adequados e utilizados em estudos de avaliação de qualidade de *sites* e páginas *Web* (MANDL, 2008) (AMENTO, TERVEEN *et al.*, 2000). As variáveis e termos lingüísticos da lógica *fuzzy* foram adotados como abordagem para a implementação do mecanismo de avaliação automatizada, em razão da sua habilidade para lidar com conceitos diferenciados e capturar o conhecimento impreciso dos seres humanos. Antes da realização do estudo, foram conduzidas as provas de conceito e os estudos de caso para verificação da eficácia dos instrumentos adotados.

Quanto à expectativa do pesquisador, de forma consciente ou não, ele pode influenciar os participantes, levando-os a resultados que "comprovem" sua percepção das hipóteses. Por esse motivo, as avaliações foram realizadas remotamente e os avaliadores receberam as mesmas instruções de preenchimento que foram fornecidas no formulário da aplicação.

### 6.3.3 – Avaliação Automática do Dataset

A avaliação automática do conjunto de páginas foi realizada de acordo com os passos do processo proposto e descrito em seções anteriores. Foram selecionadas três dimensões de qualidade *timeliness* (atualidade), *reputation* (reputação) e *completeness* (completeza), e foi atribuído, arbitrariamente, a cada uma delas um grau de importância 1 (a dimensão de qualidade avaliada tem pouca importância) com o objetivo de reduzir qualquer viés ou risco para a validade de construção.

Evidentemente que a atribuição de graus de importância diferentes para as dimensões de qualidade afetam a avaliação final das páginas. Isso permite que as avaliações realizadas de forma automática possam observar algumas perspectivas subjetivas referentes à contribuição de cada dimensão de qualidade para o resultado final. O Anexo II apresenta o *dataset* com os resultados da avaliação automática.

### 6.3.4 – Avaliação Manual do Dataset

Foi desenvolvida uma aplicação para coleta dos dados relativos às avaliações realizadas pelos usuários que foram utilizadas no experimento[59]. O Anexo III mostra o

[59] http://ireval.cos.ufrj.br/ireval/qualidade.php.

questionário de avaliação, por meio do qual cada avaliador respondeu às seguintes perguntas:

- **Sobre o avaliador**

Escolaridade: {ensino médio, graduação, especialização, mestrado e doutorado}

Atividade: {profissional, acadêmica, ambas}

- **Auto-avaliação**

Qual o seu conhecimento sobre o assunto (Bancos de dados relacionais)? {excelente, bom, regular, pouco e nenhum}

- **Avaliação das páginas**

Respostas a três perguntas {Sim ou Não**}**:

1. A página explica o que é um SGBDR?
2. A página explica as vantagens de um SGBDR?
3. A página trata do uso de SQL em SGBDR?

Avaliação de quatro atributos de qualidade da página {Péssima, Ruim, Regular, Boa ou Excelente}

Reputação: *avaliação da página considerando a sua fonte e o seu conteúdo.*

Completeza: *avaliação da página considerando a amplitude e profundidade do assunto tratado.*

Atualidade: *avaliação da página considerando se ela é suficientemente atualizada.*

Avaliação Global: a*valiação da página no geral.*

Por fim, você informará se a página é relevante para o assunto Bancos de dados relacionais {Sim ou Não**}**.

No caso de mais de uma avaliação por página, a fim de agregar as avaliações atribuídas individualmente a cada documento pelos avaliadores, os valores ordinais da escala de Likert para as variáveis qualitativas foram convertidos em valores de escala intervalares para variáveis quantitativas (péssimo = -2; ruim = -1; regular = 0; bom = 1 e excelente = 2). Essa escala intervalar é utilizada por viabilizar a classificação dos eventos com distâncias iguais, além de permitir um valor intermediário onde a dimensão

em questão não possui nem uma avaliação boa nem ruim (HAIR JR, BABIN *et al.*, 2005).

$$\sigma = \frac{(\sum_{i=1}^{n} \propto_i)}{n}$$

Onde,

$\forall \propto_i \in \{-2, -1, 0, 1, 2\}$;

$\sigma \Longrightarrow \sigma \in [-2, 2]$;

$\propto_i$ é cada uma das avaliações;

$\sigma$ é a agregação das três avaliações;

$n$ é o número de avaliações.

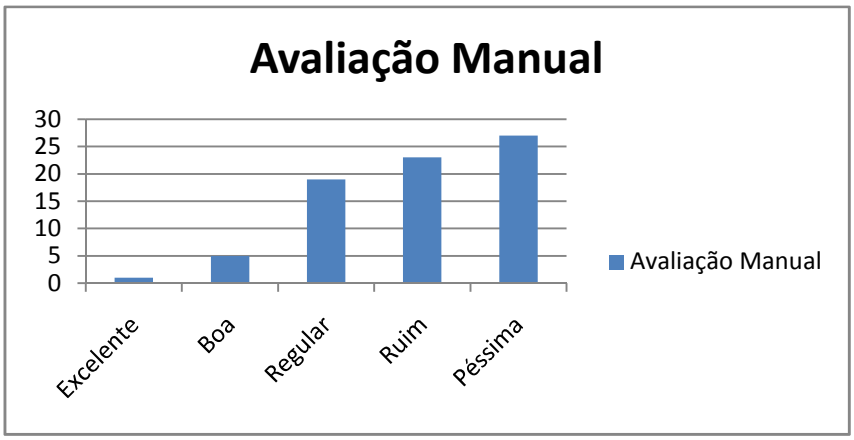
O Anexo II apresenta o *dataset* com os resultados da avaliação manual.

### 6.3.5 – Método Analítico de Comparação dos Resultados

#### *6.3.5.1 – Análise quantitativa dos resultados*

- **Análise Gráfica das Distribuições dos Dados**

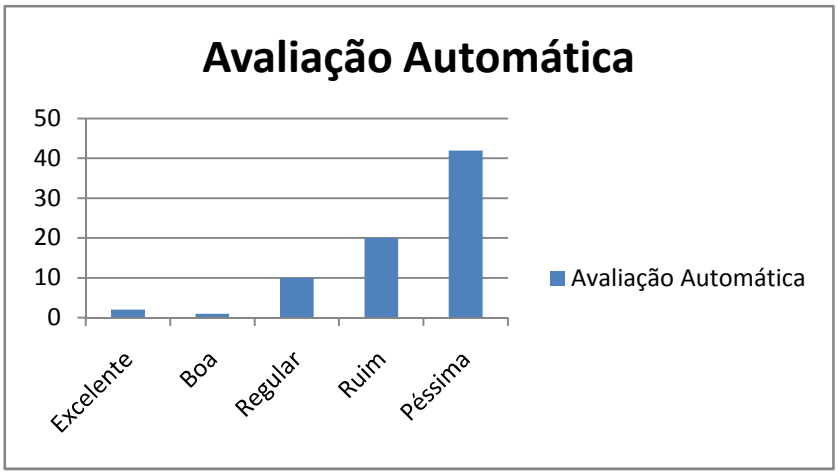
O gráfico de barras da Figura 6-29 mostra a distribuição de freqüências das avaliações manuais para as variáveis qualitativas em cada classe.




| Avaliação Manual | Freqüência |
|---|---|
| Excelente | 1 |
| Boa | 5 |
| Regular | 19 |
| Ruim | 23 |
| Péssima | 27 |
| Total | 75 |

**Figura 6-29: Distribuição de Freqüências das Avaliações Manuais**

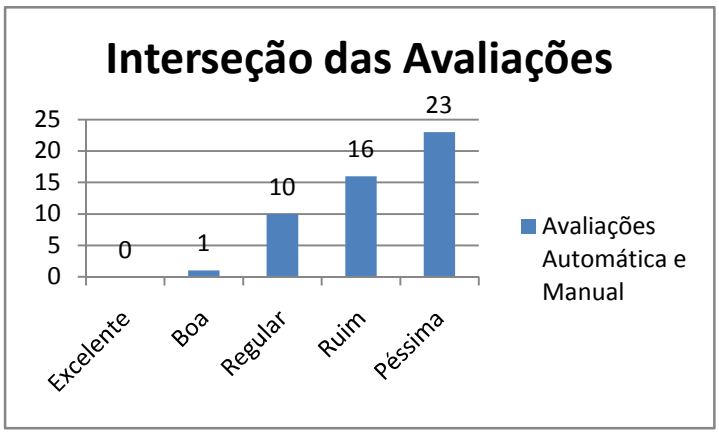
O gráfico de barras da Figura 6-30 mostra a distribuição de freqüências das avaliações automáticas para as variáveis qualitativas em cada classe.




| Avaliação Automática | Freqüência |
|---|---|
| Excelente | 2 |
| Boa | 1 |
| Regular | 10 |
| Ruim | 20 |
| Péssima | 42 |
| **Total** | **75** |

**Figura 6-30: Distribuição de Freqüências das Avaliações Automáticas**

O gráfico de barras da Figura 6-31 mostra a quantidade e o percentual da interseção das freqüências de acerto das avaliações automáticas e manuais para as variáveis qualitativas em cada classe.



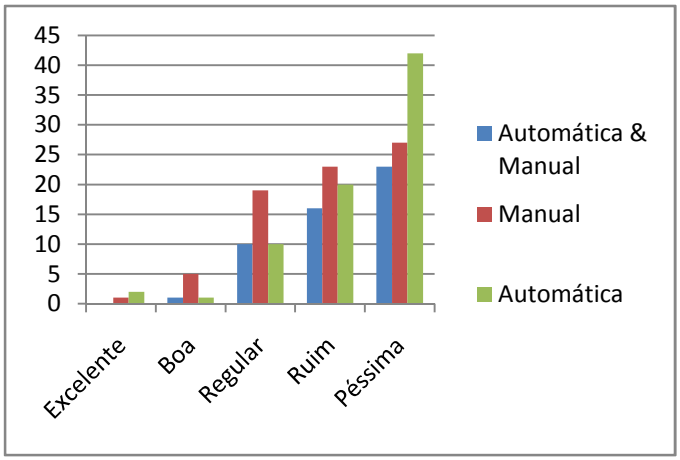

| Avaliação Automática | Freqüência | % Acerto |
|---|---|---|
| Excelente | 0 | 0% |
| Boa | 1 | 20% |
| Regular | 10 | 53% |
| Ruim | 16 | 70% |
| Péssima | 23 | 85% |
| **Total** | **50** | **67%** |

**Figura 6-31: Percentual de Interseção das Freqüências das Avaliações Automáticas e Manuais**

O gráfico de barras da Figura 6-32 mostra quantitativamente a interseção das freqüências das avaliações automáticas e manuais para as variáveis qualitativas.

Inicialmente, verificamos que a avaliação manual possui os maiores escores de classificação na qualidade da informação medida, porém a distribuição de freqüências da variável aleatória de tais avaliações tem moda *péssima*, isto é, a maior parte das páginas avaliadas recebeu essa avaliação, sendo 56% para as avaliações automáticas e 36% para as avaliações manuais. A variação do aumento da quantidade de páginas em relação à piora de sua qualidade é um comportamento esperado, comparativamente, ao estado da prática nas consultas na *Web*.

Temos ainda um percentual de acerto razoável em consideração aos seguintes percentuais por classes: 0% para *excelente*, 20% para *boa*, 53% para *regular*, 70% para *ruim*, 85% para *péssima* e 67% para o agrupamento de todas as classes.




| Avaliação | Automática & Manual | Manual | Automática |
|---|---|---|---|
| **Excelente** | 0 | 1 | 2 |
| **Boa** | 1 | 5 | 1 |
| **Regular** | 10 | 19 | 10 |
| **Ruim** | 16 | 23 | 20 |
| **Péssima** | 23 | 27 | 42 |
| **Total** | **50** | **75** | **75** |

**Figura 6-32: Interseção das Freqüências das Avaliações Automática e Manual**

A partir dos coeficientes de correlação entre as avaliações automáticas e manuais para as variáveis qualitativas, foi obtido o coeficiente de correlação de Pearson no valor de 0,250107, ou seja, as variáveis estão positivamente correlacionadas entre as duas avaliações.

A Tabela 6-10 mostra a amplitude das avaliações dentro de cada um dos intervalos de classificação. A classe que apresentou maior amplitude foi a de *valor ruim*, apesar de não ter sido aquela de maior freqüência quantitativa de avaliações, caracterizando ambas as medidas como variáveis independentes.

**Tabela 6-10: Amplitude das Avaliações por Intervalos de Classificação**

| Classificação | Escore Mínimo | Escore Máximo | Amplitude das Avaliações |
|---|---|---|---|
| **Péssima** | 0,27382844 | 0,281161635 | 0,007333 |
| **Ruim** | 0,474459466 | 0,574116285 | 0,099657 |
| **Regular** | 0,590521421 | 0,590803174 | 0,000282 |
| **Boa** | 0,720084678 | 0,720084678 | 0 |
| **Excelente** | 0,940707479 | 1 | 0,059293 |

### *6.3.5.2 – Heurísticas Relacionadas à Execução do Experimento*

A seguir são descritas algumas heurísticas simples, porém efetivas, que foram adotadas com o intuito de evitar certos problemas potenciais no tratamento de *links*

navegacionais (KLEINBERG, 1998). Também são descritas outras experiências aprendidas durante a realização do experimento.

Numa categoria de problemas técnicos, na execução do *crawler* ocorreu um problema relacionado ao tempo de espera para recuperação das páginas (ex.: *links* quebrados ou errados, dificuldades de acesso, *time out* da conexão, páginas inexistentes, exceções de acesso, etc.). Tal problema foi contornado por meio da criação de mecanismos para parametrização do "tempo máximo de espera", mediante várias execuções do *crawler* para população e persistência dos *bl* e *fl*.

Além disso, foram criados alguns filtros de limpeza para a construção do conjunto de páginas do *dataset*, antes da avaliação (ex.: eliminação das páginas /ad, blogs, .?, páginas duplicadas, conversão dos caracteres da URL para minúsculas, etc.).

Para 36 (trinta e seis) páginas com data de atualização anteriores a "01 de janeiro 1970", foi assumida a data "01 de janeiro de 2000, visando não distorcer os resultados da amostra. A data assumida foi a data válida mais antiga encontrada entre todas as datas da amostra.

Das três dimensões de qualidade, a que apresentou melhores resultados individuais de avaliação, considerando todas as avaliações feitas, foi a dimensão *timeliness*. Isso porque todas as páginas apresentaram as *datas de atualização*, que são os metadados necessários para que fosse realizada a avaliação dessa dimensão. As dimensões de qualidade *completeness* e *reputation* apresentaram resultados de avaliação individuais abaixo dos que foram obtidos para o *timeliness*. Isso ocorreu em razão da dificuldade observada na recuperação dos *forwardlinks* e dos *backlinks* externos para centenas de páginas.

Por questões de restrições de acesso às informações por meio das API do Google® e do Yahoo®, o *crawler* foi parametrizado para recuperação de, no máximo, 350 páginas, 100 *fl* e 100 *bl* por página, em 1 nível de navegação. Esses metadados foram recuperados por meio da API do Yahoo[60]. O grafo resultante somou 5423 vértices correspondentes a cada uma das páginas.

[60] http://developer.yahoo.com/.

As limitações referentes à otimização e à parametrização do *crawler* foram contempladas e sugeridas como uma perspectiva de trabalho futuro. Ela está sendo conduzida no escopo de um projeto de fim de curso[61].

Numa categoria de problemas conceituais, Mandl (2008) entende que o conteúdo e a interface com o usuário são inseparáveis na *Web* e, como conseqüência, as suas avaliações não podem ser facilmente separadas. Ele também acredita que ainda não está bem entendido como as pessoas decidem as avaliações globais das páginas, e que isso está relacionado a uma nítida dependência cultural (MANDL, 2008). Isso foi constatado, visto que ocorreram avaliações nas quais os usuários não consideraram os parâmetros pragmáticos estabelecidos.

Outro aspecto relevante refere-se ao fato das cinco escalas de classificação terem escores, subjetivamente próximos para "*péssimo e ruim*" e para "*bom e excelente*", o que num julgamento humano pode resultar em sobreposição, uma vez que no modelo de avaliação não são apresentados detalhes e explicações dos avaliadores de como concluíram seus julgamentos.

Há, também, algumas desvantagens e distorções relacionadas à avaliação da reputação e da atualidade das páginas, com base nos *links*, em relação ao julgamento humano. Foi observado que, em alguns casos, o julgamento humano atribui avaliação alta, simplesmente pela "fama" da página[62], independentemente da página estar ou não respondendo às questões, ou estar ou não atualizada.

### *6.3.5.3 – Análise qualitativa dos resultados*

De acordo com os resultados apresentados, podemos concluir, com certo grau de incerteza, que é possível utilizar a avaliação automática para identificar e separar as páginas por subconjuntos, conforme sua qualidade. Isso possibilita que as páginas com baixa qualidade sejam descartadas e que haja a diminuição das dificuldades depreendidas pelos usuários durante a seleção. Tais avaliações podem ser apresentadas junto com os resultados das buscas, fornecendo aos usuários um instrumento a mais

[61] COP*Crawler* – projeto de fim de curso orientado pelo Prof. Geraldo Xexéo.

[62] http://db.apache.org/derby.
http://www.vldb.org/conf/2004/ind5p2.pdf.

para a orientação das suas seleções e para o atendimento dos seus requisitos de pesquisa.

Outra conclusão obtida é que todos os resultados e aspectos encontrados neste trabalho são adequados somente para as condições apresentadas, limitadas principalmente pelos termos utilizados para as avaliações dos *datasets*. Essas definições não podem ser generalizadas, pois a dinâmica e a diversidade de possibilidades da *Web* poderia mudar o produto das avaliações geradas.

## 6.4 – Análise Comparativa entre os Resultados de Ordenação do Google® e da Avaliação da Qualidade

### 6.4.1 – Definição

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) (AMENTO, TERVEEN *et al.*, 2000). A avaliação da abordagem teórica proposta nesta tese também foi realizada por meio da comparação entre a ordenação das páginas pelos resultados do prognóstico de qualidade e pelo Google®.

A realização dessas avaliações foi baseada no cálculo da precisão e da cobertura das respostas obtidas. Apenas para relembrar, a precisão é definida pela razão entre a quantidade de documentos encontrados que são relevantes e a quantidade total de documentos encontrados, enquanto a cobertura é definida pela razão entre a quantidade de documentos encontrados que são relevantes e a quantidade de documentos relevantes de toda a coleção de documentos. Em síntese, essas duas definições são dadas pelas fórmulas abaixo (BAEZA-YATES, RIBEIRO-NETO, 1999):

$$PRECISÃO = \frac{R \cap E}{E}$$

$$COBERTURA = \frac{R \cap E}{R}$$

Onde $R$ é o conjunto de documentos relevantes e $E$ é o conjunto de documentos encontrados.

Dessa forma, percebe-se a necessidade de ter em mãos uma coleção de documentos na qual se conheçam exatamente aqueles que são relevantes dentro de um contexto específico ou para consultas pré-definidas.

### 6.4.2 – Cálculo de Precisão e Cobertura

As características de avaliação de qualidade para ordenação do *dataset* foram mantidas nos moldes descritos nas seções anteriores e a quantidade de páginas foi estabelecida pelo total de páginas avaliadas manualmente. Para realização do estudo comparativo, os avaliadores indicaram um conjunto de 232 documentos relevantes entre as 340 URLs avaliadas.

O Anexo IV lista o cálculo de precisão e cobertura pelos valores de ordenação do Google®, além de indicar as URLs relevantes, o número de acertos e as suas médias harmônicas, tratadas mais adiante.

O Anexo V lista o mesmo cálculo com os resultados obtidos por meio da avaliação de qualidade proposta.

A Figura 6-33 mostra o gráfico contendo as curvas resultantes correspondentes à ordenação Google® e à ordenação pela qualidade. A Figura 6-34 mostra o gráfico contendo as curvas correspondentes à média harmônica entre a precisão e a cobertura para cada uma das abordagens de ordenação. Incluímos, também, como parâmetro de comparação, os resultados de ordenação pela qualidade nos quais foi aplicada a ponderação com "grau de importância 4", para a dimensão de qualidade *timeliness* (atualidade).

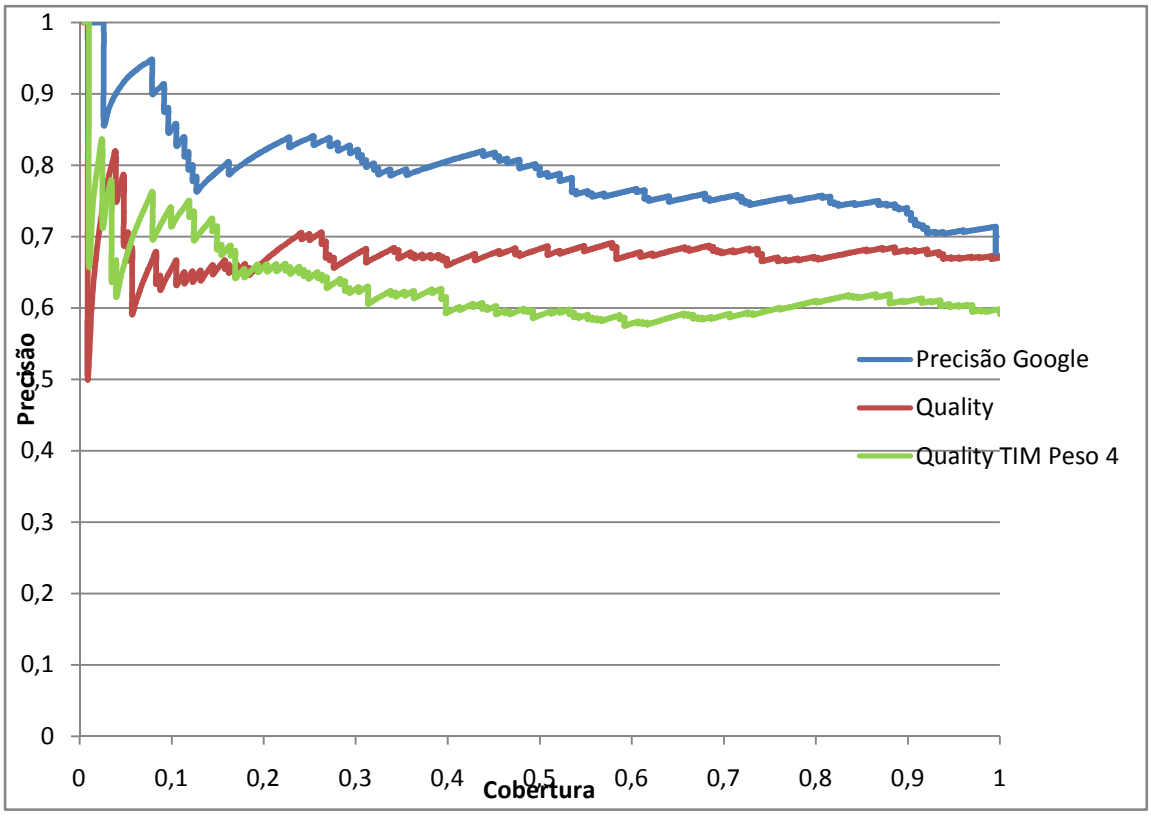



**Figura 6-33: Cobertura versus Precisão**

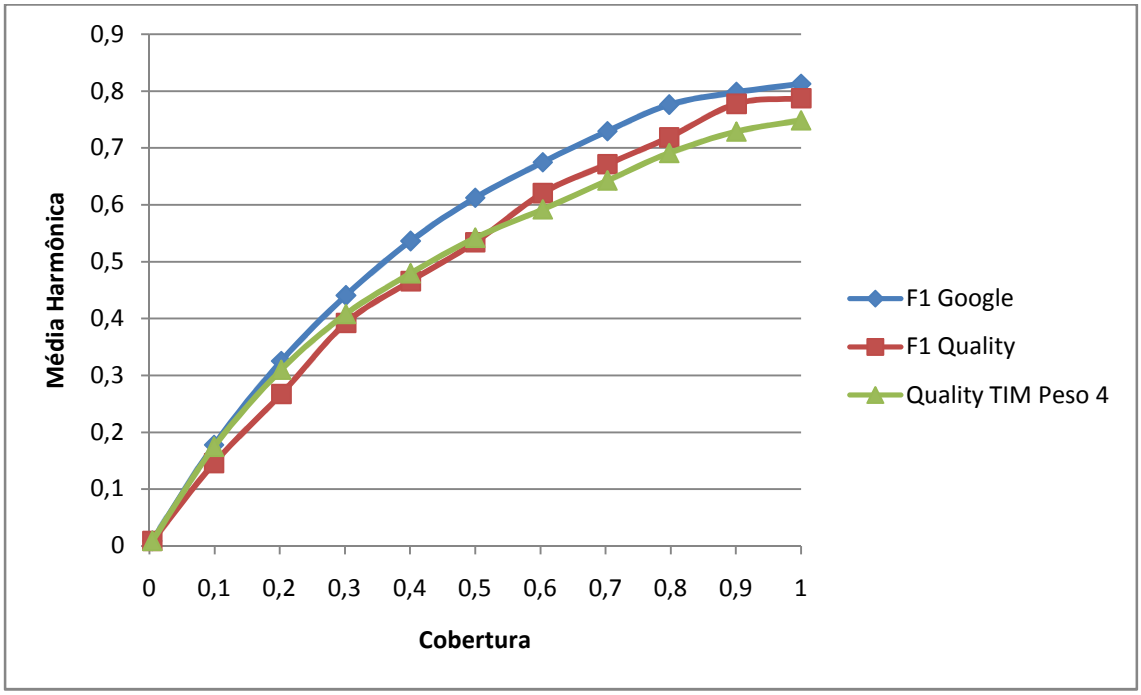



**Figura 6-34: Cobertura versus Média Harmônica**

A média harmônica é uma medida que combina precisão e cobertura e normalmente é expressa por *F*. O máximo valor de *F* é interpretado como uma tentativa

de achar a melhor relação entre a precisão e a cobertura (BAEZA-YATES, RIBEIRO-NETO, 1999). As médias harmônicas são expressas pela fórmula:

$$F = \frac{2PC}{P + C}$$

Onde: *F* é o valor da média harmônica, *P* é a precisão e *C* é a cobertura.

Observando-se na Figura 6-33 os valores de precisão obtidos pelo Google®, superaram os valores obtidos pela avaliação de qualidade e pelo *timeliness* ponderado. Entretanto, as curvas apresentadas no gráfico da Figura 6-34 mostram que, atribuindo-se peso a uma dimensão de qualidade, se consegue melhorar a média harmônica obtida na ordenação pela qualidade. Considerando-se a cobertura das 10 primeiras páginas, o *F* de qualidade pelo *timeliness* ponderado obteve melhor resultado que o do Google (*F* qualidade = 0,092592593 e *F* Google® = 0,091667). Tais resultados demonstram a sensibilidade do desempenho do algoritmo de qualidade nos casos em que as dimensões de qualidade são ponderadas. Isso é importante visto que, na maioria dos mecanismos de busca, as páginas inicialmente apresentadas aos usuários são as 10 primeiras.

Os valores base para ordenação foram obtidos por técnicas diferentes e, atualmente, as buscas realizadas pelo Google® apresentam os melhores resultados entre todos os buscadores disponíveis, sendo, de fato, muito difícil superá-los.

Vale ressaltar, portanto, que não é nosso objetivo superar os resultados de busca do Google®, mas somente mostrar que a nossa abordagem de avaliação de qualidade pode oferecer um mecanismo eficaz de auxílio aos usuários em suas buscas, haja vista que as médias harmônicas obtidas mostraram desempenho semelhante ao do Google®. Dessa forma, consideramos que os resultados obtidos foram satisfatórios.

Evidentemente, se forem estendidas as dimensões de qualidade avaliadas e refinados os atuais mecanismos de avaliação de qualidade propostos nesta tese, os resultados finais e a ordenação do *dataset* podem ser melhorados, com possíveis reflexos nos cálculos de precisão e cobertura. Esses refinamentos e melhorias foram contemplados como uma perspectiva de trabalho futuro.

# Capítulo 7 – Conclusão e Trabalhos Futuros

O fornecimento de informações sobre a qualidade de dados para os usuários representa um avanço em relação às consultas convencionais, pois possibilita que ele avalie até que ponto pode confiar nos dados apresentados. A fim de apoiar os usuários nessas atividades, muitos sistemas para o controle automático de qualidade ainda necessitam ser desenvolvidos. Modelos cada vez mais complexos de qualidade, baseados em diferentes características, irão surgir e a integração de suas diferentes definições de qualidade irá ocorrer gradativamente (MANDL, 2008).

Neste capítulo final, são apresentadas as conclusões e as principais contribuições da pesquisa, bem como as perspectivas de trabalhos futuros.

## 7.1 – Conclusão

Esta tese é relacionada a algumas áreas de estudo tais como qualidade de informações na *Web*, metadados e contextos; modelos e processos de avaliação de qualidade; conjuntos e lógica *fuzzy*. Ela investiga e aplica, de forma conjunta, as vantagens inerentes a cada uma dessas áreas, no intuito de obter soluções aos problemas inerentes à qualidade de informação, identificados no capítulo 1. Neste sentido, nossa expectativa é a de contribuir com uma abordagem inédita e alternativa, que traga resultados inovadores para a pesquisa em qualidade de informações na *Web*.

Em vista do grande volume de informações disponíveis, a possibilidade de avaliar e ordenar os documentos Web, com base em critérios de qualidade, representa novas oportunidades e estratégias mais eficientes, quando os consumidores de dados decidem personalizar a informação procurada, de acordo com suas perspectivas do nível de qualidade (LYMAN & HAL, 2003) (BURGESS, GRAY *et al.*, 2004).

Esta tese, portanto, é um avanço no estado da arte ao propor um modelo, uma metodologia e uma arquitetura para o prognóstico de qualidade de informações na Web. Conseqüentemente, os usuários – organizações ou pessoas de um modo geral – contarão com um instrumento a mais para a orientação das suas seleções e para o atendimento

dos seus requisitos de pesquisa de informações, sendo esta a principal contribuição deste trabalho.

Mais detalhadamente, outras contribuições da pesquisa incluem (BARROS, XEXÉO *et al.*, 2008a) (BARROS, XEXÉO *et al.*, 2008b) (BARROS, RODRIGUES-NT *et al.*, 2009):

*i.* A especificação de um modelo *UML*[55] para a representação, formalização, manutenção e compartilhamento dos conceitos e de suas instâncias relacionados a todas as fases da metodologia proposta para o prognóstico de qualidade de páginas *Web*;

*ii.* A proposta de uma metodologia para o prognóstico de qualidade de páginas *Web* baseado em seus metadados. A lógica *fuzzy* implementa o mecanismo de avaliação automática para valoração das dimensões de qualidade, com base nos valores dos metadados do conjunto de páginas *Web* recuperadas. Uma avaliação inicial é computada e posteriormente ajustada pela ponderação dos valores atribuídos ao contexto e das perspectivas dos usuários para obtenção das avaliações agregadas;

*iii.* Uma especificação teórica do modelo proposto com ênfase na avaliação de qualidade dos resultados fornecidos por máquinas de busca na *Web*. Para tanto, considerando que a qualidade de informação seja dependente do domínio no qual ela se insere, tanto os metadados, quanto os contextos em si servem como base de avaliação das dimensões de qualidade requeridas. Essa especificação pode ser adotada para implementação de mecanismos adicionais, como filtros de pré ou pós seleção, ou como um sistema de recomendação em colaboração com um navegador (SCHAFER, KONSTAN *et al.*, 2001) (TERVEEN & HILL, 2001);

*iv.* A construção de uma arquitetura a ser utilizada como ambiente para avaliação da qualidade de informações, com base nos metadados e no contexto de BRI na *Web*. Os diferentes componentes envolvidos no modelo e nesse processo de avaliação são representados na arquitetura;

*v.* A implementação das provas de conceito em duas abordagens *fuzzy* por meio de dois estudos de casos práticos, um adotando funções *fuzzy* de transformação e outro, regras *fuzzy* de inferência;

[55] www.uml.org.

*vi.* O desenvolvimento e a implementação de um protótipo funcional de uma aplicação colaborativa, com o objetivo de aliar ao processo de alimentação manual das páginas, proposto na especificação do FoxSet, o processo de alimentação e prognóstico de qualidade realizados automaticamente. Essa conjugação de processos tem como propósito tornar mais fácil e mais ágil o trabalho de construção das bases de testes. Há, também, dois estudos de caso que comparam alguns dos nossos resultados preliminares e os resultados obtidos pelo Google®, pelo Yahoo® e pelo Live Search®;

*vii.* A avaliação da abordagem teórica proposta nesta tese, realizada com o auxílio de um método experimental para a comparação de resultados em maior escala, incluindo a avaliação de usuários finais.

*viii.* A avaliação comparativa entre os resultados de ordenação das páginas obtidos pelo prognóstico de qualidade e pelo Google®, com base nos cálculos de precisão e cobertura, e das suas médias harmônicas. Todos os resultados obtidos estão disponíveis e podem ser utilizados como modelos de comparação em outras pesquisas semelhantes; e

*ix.* A apresentação de uma investigação das diversas abordagens existentes para a avaliação da qualidade de dados e informações, bem como a realização de um levantamento sobre os critérios e dimensões de qualidade adotados nessas abordagens. Ressaltando-se que ainda não há um consenso na comunidade científica sobre qual conjunto de dimensões melhor expressa o conceito de qualidade de dados (WAND & WANG, 1996) (RAMOS-LIMA, MAÇADA *et al.*, 2006).

Em síntese, a abordagem proposta formaliza os diferentes componentes envolvidos no processo de avaliação da qualidade de informações na *Web* e, certamente, outras pesquisas decorrentes são necessárias para explorar suas conseqüências computacionais e sociais, bem como os custos efetivos dessa melhoria de qualidade. Os custos imediatamente identificados estão relacionados à adaptação dos ambientes e dos sistemas durante o processo de adoção da estratégia apresentada.

## 7.2 – Trabalhos Futuros

Este trabalho, em sua especificação e na atual versão implementada, apresenta algumas limitações e deixa alguns problemas em aberto, que poderão ser investigados e resolvidos em trabalhos futuros. A seguir são apresentadas algumas sugestões:

- Expansão do modelo e do protótipo para incluir no processo de avaliação outras dimensões e metadados de qualidade, a fim de melhorar os resultados da avaliação de qualidade como um todo;
- O desenvolvimento ou a adoção de uma política para avaliação de competências dos especialistas. Essa questão aparece sempre que a avaliação é realizada por mais de um usuário. Algumas abordagens para distinguir a opinião dos usuários refletem suas colaborações anteriores, seu conhecimento avaliado por outros usuários ou as auto-avaliações (RODRIGUES NT, SOUZA *et al.*, 2006). Isto permitiria um melhor balanceamento entre níveis de habilidade diferentes. Os resultados da avaliação de competências seriam adotados como fatores de ponderação na definição do vetor de importâncias para as dimensões de qualidade selecionadas, em relação ao contexto. Além disso, essa política de avaliação seria de grande valia na especificação de mecanismos para identificar ou evitar comportamentos maliciosos dos usuários;
- Especificação de mecanismos para identificar ou evitar comportamentos maliciosos dos usuários. Se a avaliação estiver sendo feita em um ambiente público ou descontrolado – isto é, onde os usuários não são conhecidos de forma antecipada e puderem agir como quiserem – as questões relativas aos comportamentos maliciosos são de grande importância. Em tais ambientes, é comum existirem usuários tentando burlar o sistema intencionalmente, fornecendo informações falsas ou errôneas. Nesses casos, a análise desse comportamento pode ser bastante útil na identificação desses pontos fora da amostra. Foi assumido que o protótipo implementado somente está sendo usado em ambientes controlados, tais como equipes de pesquisa;

- Integração e/ou extensão do modelo de qualidade com a Ontologia da Informação proposta pelo TRASGO – Laboratório de Tratamento da Sobrecarga da Informação[56];
- Avaliação comparativa ou agregada dos documentos *Web* em múltiplos contextos, visto que, na atual especificação, o modelo trata um contexto de cada vez, em razão de seus níveis de abstração, hora muito genéricos, hora muito específicos;
- Tratamento da diversidade interdocumentos e intradocumentos. Esse tratamento permitiria determinar a consistência interna dos documentos e se os documentos pertencentes aos diferentes conjuntos estão consistentes entre si. Um exemplo são os documentos da Wikipédia[57] com partes bem definidas e partes mal definidas;
- Desenvolvimento de um novo *crawler* mais robusto, mais eficiente e parametrizável, visto que a atual implementação não contempla essas características, principalmente, quanto às questões de desempenho[58];
- Extensão da pesquisa para adoção de meta mecanismos de buscas ou emprego do maior número de mecanismos de busca possíveis, sejam eles genéricos ou verticais, visando a inclusão de conteúdos da *Web* profunda (*deep Web*), a fim de melhorar os prognósticos de qualidade para os conjuntos de páginas avaliados (BERGMAN, 2001) (WRIGHT, 2009);
- Desenvolvimento de um sistema de recomendação baseado nos resultados obtidos, integrando à atividade de recomendação como um passo a mais na metodologia proposta. Esses sistemas são serviços personalizados, desenvolvidos para ajudar as pessoas com a diversidade e a sobrecarga da informação, e possibilitam o compartilhamento de opiniões e de experiências (SCHAFER, KONSTAN *et al.*, 2001) (TERVEEN & HILL, 2001). A principal abordagem de tais soluções é encontrar informações interessantes

[56] O TRASGO – Laboratório de Tratamento da Sobrecarga da Informação é um projeto que está em fase de concepção, coordenado pelo Professor Geraldo Xexéo. Essa iniciativa busca caminhos de integração para as diversas pesquisas conduzidas na linha de BD, relacionadas à sobrecarga, à avaliação de qualidade e à BRI.

[57] http://www.wikipedia.org/.

[58] COP*Crawler* – projeto de fim de curso orientado pelo Prof. Geraldo Xexéo.

aos usuários, em vez de eliminar as que são irrelevantes. Nosso entendimento é de que essa é uma área fértil para a pesquisa, que ainda não foi explorada;

- Desenvolvimento de um sistema de filtragem de informação baseado na captura e armazenamento das informações do usuário (LAWRENCE, 2000) (MOURA, 2003) e nos resultados das avaliações. Nesse sentido, o uso dos filtros de informação pode ser útil na eliminação de informações irrelevantes (BELKIN & CROFT, 1992). Tais mecanismos geralmente usam os perfis dos usuários para representar as suas necessidades de informação e adotam agentes inteligentes nas tarefas de eliminação (MAES, 1994) (MIZZARO, 1997);
- Integração da proposta com aplicações de eliminação de dados. A eliminação de dados é uma das outras aplicações possíveis que podem ajudar na melhoria da qualidade dos dados. Numa das abordagens estudadas, esse tratamento inclui processos de agregação, organização e limpeza dos dados, modelados na forma de padrões que são voltados para a eliminação de dados não relevantes e desnecessários (PINHEIRO, BARROS *et al.*, 2008);

A questão da qualidade de informações de documentos *Web* tem se tornado cada vez mais importante, em razão do constante crescimento na quantidade de tais documentos. Alguns estudos mostram que 40% do material existente na *Web* desaparece em um ano, enquanto outros 40% são alterados e apenas 20% permanecem em sua forma original (BATINI & SCANNAPIECO, 2006). Outros estudos indicam que o tempo de vida médio de uma página *Web* é de 44 dias e que a *Web* muda completamente quatro vezes ao ano (LYMAN & HAL, 2003).

Considerando que a atual sociedade está centrada na informação, não há dúvida de que a qualidade dessas informações possui necessário e importante papel.

Devido ao rápido desenvolvimento das tecnologias da *Web*, que permite o compartilhamento e troca de dados de forma fácil e transparente, é mais provável que os problemas de qualidade de informações agravem-se, antes que se alcance uma solução definitiva. Tal condição continua a exigir, portanto, mais e mais pesquisas, propostas e abordagens voltadas para o auxílio aos usuários no gerenciamento e solução desses problemas (GERTZ, OZSU *et al.*, 2004).

# Referências Bibliográficas


ABATE, M. L., DIEGERT, K. V., ALLEN, H. W., 1998, "A Hierarchical Approach to Improving Data Quality", ***DataQuality***, v. 4, pp. 365-369.

ABOELMEGED, M., 2000, "A soft system perspective on information quality in electronic commerce". In: *Proceedings of the Fifth Conference on Information Quality*, pp. 318-319.

AKOKA, J., BERTI-ÉQUILLE, L., BOUCELMA, O., *et al.*, 2007, "A framework for quality evaluation in data integration systems". In: *Dans Proceedings of the 9th International Conference on Enterprise Information Systems, 2007 (ICEIS'07).*

ALADWANI, A. M.,PALVIA, P. C., 2002, "Developing and Validating an Instrument for Measuring User-Perceived Web Quality", ***Information and Management***, v. 39, n. 6, pp. 467-476.

AMENTO, B., TERVEEN, L., HILL, W., 2000, "Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents". In: *SIGIR 2000*, v. 7, pp. 296-303, Athens, Greece.

AMIRIJOO, M., HANSSON, J., SON, S., 2003, "Specification and management of QoS in imprecise real-time databases". In: *Proceedings of the Seventh International Database Engineering and Applications Symposium*, pp. 192-201.

ANGELES, P.,MACKINNON, L., 2004, "Detection and resolution of data inconsistencies, and data integration using data quality criteria". In: *QUAT - IC'2004*, pp. 87-93.

BAEZA-YATES, R.,RIBEIRO-NETO, B., 1999, ***Modern Information Retrieval***. 1 ed. USA, Addison-Wesley-Longman Publishing co.

BAEZA-YATES, R., CASTILLO, C., MARIN, M., *et al.*, 2005, "Crawling a country: Better strategies than breadth-first for Web page ordering". In: *Proceedings of the Industrial and Practical Experience Track of the 14th Conference on the World Wide Web*, pp. 864-872, Chiba, Japan.

BALLOU, D., MADNICK, S., WANG, R., 2004, "Assuring Information Quality", ***Journal of Management Information Systems / Winter 2003-4.© 2004 M.E.Sharpe, Inc.0742-1222 / 2004.***, v. 20, n. 3, pp. 9-11.

BARBACCI, M., KLEIN, M. H., LONGSTAFF, T. A., *et al.*, 1995, *Quality Attributes*

BARROS, R., XEXÉO, G., *et al.*, 2008a, "A Web Metadata Based-Model for Information Quality Prediction". In: Calero, C., Moraga M.A., and Piattini M., *Handbook of Research on Web Information Systems Quality*, chapter XIX, Information Science Reference, Hershey, PA.

BARROS, R., XEXÉO, G., *et al.*, 2008b, "User and Context-Aware Quality Filters Based on Web Metadata Retrieval". In: González, R. A., Chen, N., and Dahanayake A., *Personalized Information Retrieval and Access: Concepts, Methods and Practices*, chapter VIII, Information Science Reference, Hershey, PA.

BARROS, R., RODRIGUES-NT, J. A., CARNEIRO-FILHO, H. J. A., *et al.*, 2009, "A Collaborative Approach to Building Evaluated Web Pages Datasets". In: *Proceeding of the 13th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2009).*

BATINI, C.,SCANNAPIECO, M., 2006, ***Data Quality Concepts, Methodologies and Techniques***. 1 ed. New York, Springer.

BATINI, C., BARONE, D., MASTRELLA, M., *et al.*, 2007, "A Framework and a Methodology for Data Quality Assessment and Monitoring". In: *Proceedings of the 12th International Conference on Information Quality ICIQ '2007.*

BATINI, C., BARONE, D., CABITZA, F., *et al.*, 2008, "Toward a Unified Model for Information Quality". In: *VLDB '08*, Auckland, New Zeland.

BATISTA, M. C. M.,SALGADO, A. C., 2007, "Information Quality Measurement in Data Integration Schemas". In: *VLDB 2007 Proceedings*, Vienna, Austria.

BECKER, D., MCMULLEN, W., HETHERINGTON-YOUNG, K., 2009, "A Flexible and Generic Data Quality Metamodel". In: *Proceedings of the 12th International Conference on Information Quality ICIQ'2007.*

BELCHIOR, A., XEXEO, G. B., ROCHA, A. R. C., 1997, *Enfoques Sobre a Teoria dos Conjuntos Fuzzy*. In: COPPE, UFRJ, ES-430.

BELCHIOR, D. A., 1997, *Um Modelo Fuzzy para Avaliação da Qualidade de Software*, Tese de Doutorado, COPPE/UFRJ, Rio de Janeiro.

BELKIN, N. J.,CROFT, W. B., 1992, "Information filtering and information retrieval: Two sides of the same coin?". In: *Communications of the ACM*, v. 35, pp. 29-38.

BELLMANN, R. E.,GIERTZ, M., 1973, "On the analytic formalism of theory of fuzzy set", ***Information Science 5, in (SDORRA, 1993)***, v. 5, pp. 149-157.

BERGMAN, M. K., 2001, "The Deep Web: Surfacing Hidden Value", ***The Journal of Electronic Publishing***, v. 7, n. 1.

BERTI-´EQUILLE, L., 2007, "Measuring and Modelling Data Quality for Quality-Awareness in Data Mining", ***Studies in Computational Intelligence (SCI)***, v. 43, pp. 101-126.

BOEHM, B. W., BROWN, J. R., LIPOW, M., 1976, "Quantitative Evaluation of Software Quality". In: *Proceedings of the 2nd International Conference on Software Engineering*, pp. 592-605, San Francisco, CA, USA.

BOSC, P., 1995, "Quantified Statements in a Flexible Relational Query Language". In: *Proceedings of the 1995 ACM Symposium on Applied Computing*, Nashville, February.

BOUZEGHOUB, M.,PERALTA, V., 2004, "A framework for analysis of data freshness". In: *Proceedings of the International Workshop on Information Quality in Information Systems (IQIS2004)*, pp. 59-67, Paris, France.

BOVEE, M., SRIVASTAVA, R. P., MAK, B., 2001, "A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality". In: *Proceedings of the 6th International Conference on Information Quality (ICIQ-01)*, pp. 311-324, Cambridge, MA, USA.

BRIN, S.,PAGE, L., 1998, "The anatomy of a large-scale hypertextual Web search engine". In: *Proceedings of the 7th World Wide Web Conference*, Brisbane, December.

BURGESS, M. S. E., GRAY, W. A., FIDDIAN, N. J., 2003, "A Flexible Quality Framework for Use in Information Retrieval". In: *Proceedings of the 8th International Conference on Information Quality (ICIQ-03)*, pp. 297-313, Cambridge, MA, USA.

BURGESS, M. S. E., GRAY, W. A., FIDDIAN, N. J., 2004, "Quality measures and the information consumer". In: *Proceedings of the 9th International Conference on Information Quality (ICIQ-04)*, pp. 373-388, November.

CAPPIELLO, C., FRANCALANCI, C., PERNICI, B., 2004, "Data quality assessment from the user's perspective". In: *Proceedings of the International Workshop on Information Quality in Information Systems, (IQIS2004)*, pp. 68-73, Paris, France.

CARCHIOLO, V.,MALGERI, M., 1995, "A fuzzy approach to co-design system partitioning". In: *Proceedings of the 1995 ACM Symposium on Applied Computing*, Nashville.

CARO, A., CALERO, C., *et al.*, 2008, "A Data Quality Model for Web Portals". In: Calero, C., Moraga M.A., and Piattini M., *Handbook of Research on Web Information Systems Quality*, chapter VIII, New York, Information Science Reference, Hershey.

CARVALHO, D. O., 1997, *Qualidade de Sistemas de Informação Hospitalar*, Dissertação de Mestrado, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

CHEN, Y., ZHU, Q., WANG, N., 1998, "Query Processing with Quality Control in the World Wide Web", ***World Wide Web***, v. 1, n. 4, pp. 241-255.

COMERLATO, C., XEXEO, G. B., ROCHA, A. R. C., 1994, "Avaliação da Reutilizabilidade de Componentes de Software". In: *VIII Simpósio Brasileiro de Engenharia de Software*, pp. 23-26, Curitiba, Paraná.

COX, E., 1994, ***The fuzzy systems handbook: a practitioner's guide to building, using, and maintaining fuzzy systems***. 2 ed. London, Academic Press.

CUNNINGHAM, H., MAYNARD, D., BONTCHEVA, K., *et al.*, 2002, "GATE: an architecture for development of robust HLT applications". In: *Proceedings of the 40th Annual Meeting on Association For Computational Linguistics*, Philadelphia, Pennsylvania.

CYKANA, P., PAUL, A., STERN, M., 1996, "DOD Guidelines on Data Quality Management". In: *Proceedings of the Conference on Information Quality (IQ-1996)*, pp. 154-171, Cambridge, MA, USA.

DASU, T.,JOHNSON, T., 2003, "Exporatory Data Mining and Data Cleaning", ***J.Wiley Series in Probability and Statistics***, v. 11, n. 9.

DAVYDOV, M., 2001, ***Corporate portals and ebusiness integration***., McGraw-Hill Professional.

DAY, W. H. E., 1989, "Consensus methods as tools for data analysis", ***H.H.Bock***, v. 14, n. 1, pp. 53-69.

DE AMICIS, F.,BATINI, C., 2004, "A methodology for Data Quality Assessment on Financial Data", ***Studies in Commnunication***, v. 4, n. 2, pp. 115-136.

DEDEKE, A., 2000, "A Conceptual Framework for Developing Quality Measures for Information Systems". In: *Proceedings of the 2000 Conference on Information Quality (IQ-2000)*, pp. 126-128, Cambridge, MA, USA.

DEY, A. K., 2001, "Understanding and Using Context", ***Personal and Ubiquitous Computing Journal***, v. 5, pp. 4-7.

DHAMIJA, R., TYGAR, J. D., HEARST, M., 2006, "Why phishing works". In: *SIGCHI Conference on Human Factorsin Computing Systems CHI '06*, pp. 581-590, Montréal, Québec, Canada.

DRAVIS, F., 2005, "The IQ Solution Cycle". In: *Proceedings of the 12th International Conference on Information Quality ICIQ'2005*.

DROMEY, R. G., 1995, "A model for software product quality", ***IEEE Transactions on Software Engineering***, v. 21, n. 2, pp. 146-162.

DUBOIS, D.,PRADE, H., 1980, ***Fuzzy Sets and Systems: Theory and Applications***. New York, Academic Press.

DUBOIS, D., 1985, "A review of fuzzy set aggregation connectives", ***Information Science***, v. 36, pp. 85-121.

DUBOIS, D.,PRADE, H., 1991, "Fuzzy sets in approximate reasoning, Part 1: Inference with possibility distributions", ***Fuzzy Sets and Systems, IFSA, Special Memorial***, v. 25 years of fuzzy sets.

DUBOIS, D.,PRADE, H., 1989, "Fuzzy sets, probability and measurement", ***European J.Oper.Res.40 in (TURKEN, 1991).***, v. 40.

ECKERSON, W., 2002, "Data Quality and the Bottom Line: Achieving Business Success Through a Commitment to High Quality Data". In: *The Data Warehousing Institute Report Series*, v. 1, pp. 1-32.

ENGLISH, L. P., 1999, ***Improving Data Warehouse and Business Information Quality-Methods for Reducing Costs and Increasing Profits***. 1 ed. USA, Wiley.

EPPLER, M.,MUENZENMAYER, P., 2002, "Measuring information quality in the Web context: A survey of state-of-the-art instruments and an application methodology". In: *Proceedings of the Seventh International Conference on Information Quality*, pp. 187-196.

EPPLER, M., ALGESHEIMER, R., DIMPFEL, M., 2003, "Quality criteria of content-driven Web sites and their influence on customer satisfaction and loyalty: An empirical test of an information quality framework". In: *Proceedings of the Eighth International Conference on Information Quality*, pp. 108-120.

EPPLER, M. J., 2001, "A Generic Framework for Information Quality in Knowledge-Intensive Processes". In: *Proceedings of the 6th International Conference on Information Quality (IQ 2001)*, pp. 329-346, Cambridge, MA, USA.

FICKAS, S.,HELM, B. R., 1992, "Knowledge Representation and in the Design of Composite Systems", ***IEEE Transaction on Software Engineering***, v. 18, n. 6.

FINKELSTEIN, C.,AIKEN, P., 1999, "XML and corporate portals". In: *Building Corporate Portals Using XML*, New York.

FRENCH, S., 1986, ***Decision Theory: An Introduction to the Mathematics of Rationality***. New York, Halsted Press.

FRENCH, S., 1989, "Fuzzy sets: the unanswered questions", ***School of Computer***, n. 89.

FRIEDMAN, T.,BITTERER, A., 2007, *Magic Quadrant for Data Quality Tools*. In: Gartner Group, http://www.sas.com/news/analysts/gartner_df_dqt_2007.pdf.

FUGINI, M., MECELLA, M., PLEBANI, P., *et al.*, 2002, "Data quality in cooperative Web information systems". In: *CoopIS/DOA/ODBASE 2002*, pp. 486-502.

FUHRMANN, G., 1990, "Note on the Generality of Fuzzy Sets", ***Information Sciences***, v. 51, n. 2, pp. 143-152.

GARDYN, E., 1997, "A Data Quality Handbook for a Data Warehouse". In: *Proceedings of the Conference on Information Quality (IQ-1997)*, pp. 267-290, Cambridge, MA, USA.

GE, M.,HELFERT, M., 2007, "A Review Of Information Quality Research - Develop A Research Agenda". In: *Proceedings of the 12th International Conference on Information Quality ICIQ'2007*.

GEGOV.A.E, 1995, "Hierarchical fuzzy control of multivariable systems", ***Fuzzy Sets and Systems***, v. 72, n. 3, pp. 299-310.

GERTZ, M., OZSU, T. M., SAAKE, G., *et al.*, 2004, "Report on the Dagstuhl Seminar "Data Quality on the Web"", ***ACM SIGMOD Record On Line***, v. 33, n. 1, pp. 127-132.

GILES, R., 1988, "The concept of grade of membership". In: *Fuzzy Sets and Systems in (SDORRA, 93)*, v. 25, pp. 297-323.

GOASDOUÉ, V., NUGIER, S., DUQUENNOY, D., *et al.*, 2007, "An Evaluation Framework For Data Quality Tools". In: *Proceedings of the 12th International Conference on Information Quality ICIQ'2007*.

GRAEFE, G., 2003, "Incredible information on the Internet: Biased information provision and a lack of credibility as a cause of insufficient information quality". In: *Proceedings of the Eighth International Conference on Information Quality*, pp. 133-146.

GRAUEL, A., 1999, "Analytical and Structural Considerations in Fuzzy Modeling.", ***Fuzzy Sets and Systems***, v. 101, n. 2, pp. 205-206.

GÜRMAN, N. T., 1995, "Generation and Improvement of Fuzzy Classifier With Incremental Learning Using Fuzzy RuleNet". In: *Proceedings of the 1995 ACM Symposiumon Applied Computing*, Nashville, February.

HAIDER, A.,KORONIOS, A., 2003, "Authenticity of information in cyberspace: IQ in the Internet, Web, and e-business". In: *Proceedings of the Eighth International Conference on Information Quality*, pp. 121-132.

HAIR JR, J. F., BABIN, B., MONEY, A. H., *et al.*, 2005, ***Fundamentos de Métodos de pesquisa em Administração***. Porto Alegre, Bookman.

HAVELIWALA, T. H., 1999, "Efficient Computation of PageRank". In: <http://dbpubs.stanford.edu/pub/1999-31>, Accessed in 27/12/2008.

HULL, L. G., 1991, "Expert System Development Methodology and Management". In: *Proceedings of the IEEE/ACM International Conference on Development and Managing Expert System Programs*, Washington.

HYATT, L. E.,ROSENBERG, L. H., 1996, "A Software Quality Model for Identifying Project Risks and Assessing Software Quality". In: *Proceedings of the 8th Annual Software Technology Conference*, Utah, USA.

ISC, 2007, "Internet Systems Consortium". In: http://www.isc.org/index.pl?/ops/ds, Accessed in 12/11/2008.

ISO 9000:2005, 2005, *Quality management systems - Fundamentals and vocabulary*

ISO 9126-1, 2001, *ISO/IEC 9126-1:2001 - Software engineering -- Product quality -- Part 1: Quality model*

ISO 9126-4, 2004, *ISO/IEC TR 9126-4:2004 - Software engineering -- Product quality -- Part 4: Quality in use metrics*

KACPRZYK, J., 1992, "Group decision making and consensus under fuzzy preference and fuzzy majority", ***Fuzzy Sets and Systems***, v. 49, n. 1, pp. 21-31.

KANTROWITZ, M., HORSTKOTTE, E., JOSLYN, C., 1997, "Answers to Questions about Fuzzy Logic and Fuzzy Expert Systems". In: http://www-cgi.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/fuzzy/faq/fuzzy.faq., Accessed in 25/06/2006.

KATERATTANAKUL, P.,SIAU, K., 1999, "Measuring information quality of Web sites: Development of an instrument". In: *Proceedings of the 20th International Conference on Information System*, pp. 279-285.

KATERATTANAKUL, P.,SIAU, K., 2001, "Information quality in Internet commerce design". In: *Information and database quality Kluwer Academic Publishers*.

KAUFMANN, A.,GUPTA, M. M., 1991, "Introduction to Fuzzy Arithmetic: Theory and Applications", ***Van Nostrand Reinhold***.

KIM, Y. J., KISHORE, R., SANDERS, R. L., 2005, "DQ to EQ: understanding data quality in the context of e-business systems". In: *Communications of the ACM*, v. 48, pp. 75-81.

KITCHENHAM, B., 1996, "Software Quality: The Elusive Target". In: *IEEE Software*, v. 13, pp. 12-21.

KLEINBERG, J. M., 1998, "Authoritative sources in a hyperlinked environment". In: *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*, pp. 668-677.

KLIR, G. J.,FOLGER, T. A., 1988, ***Fuzzy Sets, Uncertainty and Information***. New Jersey, Prentice Hall.

KLIR, G. J.,YUAN, B., 1995, ***Fuzzy Sets and Fuzzy Logic: Theory and Applications***. New Jersey, Prentice Hall.

KNIGHT, S. A.,BURN, J. M., 2005, "Developing a framework for assessing information quality on the World Wide Web", ***Informing Science Journal***, pp. 159-172.

KUNCHEVA, L. I.,KRISHNAPURAM, R., 1996, "A fuzzy consensus aggregation operator", ***Fuzzy Sets and Systems***, v. 79, n. 3, pp. 347-356.

LAWRENCE, S., 2000, "Context in Web Search". In: *IEEE Data Engineering Bulletin*, v. 23, pp. 25-32.

LEE, C. C., 1990, "Fuzzy logic in control systems: fuzzy logic controller - parte I e II". In: *IEEE Transactions on Systems, Man, and Cybernetics*, v. 20, pp. 404-435, California.

LEE, H., LEONARD, D., WANG, X., *et al.*, 2008, "IRLbot: Scaling to 6 Billion Pages and Beyond". In: *WWW 2008*, Beijing, China.

LEE, Y. W., STRONG, D. M., KAHN, B., *et al.*, 2002, "AIMQ: a methodology for information quality assessment", ***Information & management***, v. 40, n. 2, pp. 133-146.

LEE, Y. W., 2004, "Crafting Rules: Context-Reflective Data Quality Problem Solving". In: *Journal of Management Information Systems*, v. 20, pp. 93-119.

LERMAN, K., 2007, "Dynamics of Collaborative Document Rating Systems". In: *Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, pp. 46-55.

LIM, O.,CHIN, K., 2000, "An investigation of factors associated with customer satisfaction in Australian Internet bookshop Web sites". In: *Proceedings of the 3rd Western Australian Workshop on Information Systems Research*, Autralia.

LIU, K., ZHOU, S., YANG, H., 2000, "Quality Metrics of Object Oriented Design for Software Development and Re-Development". In: *Proceedings of the 1st Asian Pacific Conference on Quality Software*, pp. 127-135, Hong Kong, China.

LIU, Y., GAO, B., LIU, T., *et al.*, 2008, "BrowseRank: Letting Web Users Vote for Page Importance". In: *SIGIR'08*, Singapore.

LONG, J. A.,SEKO, C. E., 2002, "A New Method for Database Data Quality Evaluation at the Canadian Institute for Health Information (CIHI)". In: *Proceedings of the 7th International Conference on Information Quality(ICIQ-02)*, pp. 238-250, Cambridge, MA, USA.

LOSHIN, D., 2001, ***Enterprise Knowledge Management - The Data Quality Approach***. 1 ed. USA, Morgan Kaufman.

LYMAN, P.,HAL, R. V., 2003, "How Much Information?". In: Retrieved from http://www.sims.berkeley.edu/how-much-info-2003, Accessed in 02/07/2006.

MADNICK, S.,ZHU, H., 2006, "Improving data quality through effective use of data semantics", ***Data Knowl.Eng.***, v. 59, n. 2, pp. 460-475.

MAES, P., 1994, "Social interface agents: acquiring competence by learning from users and other agents". In: *Working Notes of the AAAI Spring Syposium on Software Agents*, pp. 71-78, Stanford, California.

MALETIC, J. I.,MARCUS, A., 2000, "Data cleansing: beyond integrity checking". In: *Proceedings of The Conference on Information Quality (IQ2000)*, pp. 200-209, Boston, MA, USA.

MANDL, T., 2006, "Implementation and evaluation of a quality based search engine". In: *17th ACM Conference on Hypertext and Hypermedia (HT '06)*, pp. 73-84, Odense, Denmark.

MANDL, T.,DE LA CRUZ, T., 2007, "International differences in Web page evaluation guidelines", ***International Journal of Intercultural Information Management (IJIIM)***, v. 1, n. 2.

MANDL, T., 2008, "Automatic Quality Assessment for Internet Pages". In: Calero, C., Moraga M.A., and Piattini M., *Handbook of Research on Web Information Systems Quality*, chapter VI, Information Science Reference, Hershey, New York.

MATSUMURA, A.,SHOURABOURA, N., 1996, "Competing with Quality Information". In: *Proceedings of the 1996 Conference on Information Quality (IQ-1996)*, pp. 72-86, Cambridge, MA, USA.

MAYWORM, M. M., 2007, *Um Crawler Peer-To-Peer Baseado em Agentes*, Dissertação de Mestrado, COPPE/UFRJ, Rio de Janeiro, RJ, Brasil.

MCCALL, J. A., RICHARDS, P. K., WALTERS, G. F., 1977, "Factors in Software Quality", ***National Technical Information Service***, v. *I-III*.

MELKAS, H., 2004, "Analyzing information quality in virtual service networks with qualitative interview data". In: *Proceedings of the Ninth International Conference on Information Quality*, pp. 74-88.

MILLER, H., 1996, "The Multiple Dimensions of Information Quality", ***Information Systems Management***, v. 13, n. 2, pp. 79-82.

MIZZARO, S., 1997, "Relevance: the whole history", ***Journal of American Society for Information Science***, v. 48, n. 9, pp. 810-832.

MORESI, E. A. D., 2000, "Gestão da Informação e do Conhecimento". In: TARAPANOFF, K, *Inteligência Organizacional e Competitiva*, Brasília, Editora UnB.

MOURA, A., 2003, "The semantic Web: fundamentals, technologies, trends". In: *Anais do XVII Simpósio Brasileiro de Banco de Dados*, Gramado, Brasil.

MOUSTAKIS, V., LITOS, C., DALIVIGAS, A., *et al.*, 2004, "Web site quality assesment criteria". In: *Proceedings of the Ninth International Conference on Information Quality*, pp. 59-73.

MUNAKATA, J.,ANI, Y., 1994, "Fuzzy Systems: An Overview", ***Communications of the ACM***, v. 37, n. 3.

NAUMANN, F.,ROLKER, C., 2000, "Assesment methods for information quality criteria.". In: *Proceedings of the Fifth International Conference on Information Quality*, pp. 148-162.

NAUMANN, F., 2002, "Quality-Driven Query Answering for Integrated Information Systems", ***Lecture Notes in Computer Science***, v. 2261, n. 1.

NAUMANN, F., FREYTAG, J. C., BOETTCHER, J., *et al.*, 2003, "The HiQIQ - High Quality Information Query Project". In: http://www.hiqiq.de/quality.html, Accessed in 26/12/2008.

NICULESCU, S. P.,VIERTL, R., 1992, "Bernoulli's Law of Large Numbers for vague data", ***Fuzzy Sets and Systems***, v. 50, n. 2, pp. 167-173.

O'NEIL, E. T., LAVOIE, B. F., and et.al., 2003, "Trends in the Evolution of the Public Web:1998 - 2002", *D-Lib Magazine*.

OLSINA, L., LAFUENTE, G., ROSSI, G., 2001, "Specifying Quality Characteristics and Attributes for Web Sites", ***Web Engineering 2000***, pp. 266-278.

ORTEGA, M., PÉREZ, M. A., ROJAS, T., 2002, "A Systemic Quality Model for Evaluating Software Products". In: *Proceedings of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics (SCI 2002/ISAS2002).*

PAGE, L., BRIN, S., MOTWANI, R., *et al.*, 1998, "The PageRank Citation Ranking: Bringing Order to the Web". In: http://dbpubs.stanford.edu/pub/1999-66, Accessed in 27/12/2008.

PERALTA, V., RUGGIA, R., KEDAD, Z., *et al.*, 2004, "A Framework for Data Quality Evaluation in a Data Integration System". In: *SBBD - Simpósio Brasiliero de Banco de Dados*, pp. 134-147, Brasília - DF.

PERNICI, B.,SCANNAPIECO, M., 2002, "Data quality in Web information systems". In: *Proceedings of the 21st International Conference on Conceptual Modeling*, pp. 397-413.

PIERCE, E. M., 2004, "Assessing data quality with control matrices", ***Communications of the ACM***, v. 47, n. 2, pp. 82-86.

PINHEIRO, W. A.,MOURA, A. M. C., 2004, "An Ontology Based-Approach for Semantic Search in Portals". In: *Database and Expert Systems Applications (DEXA), IEEE Computer Society*, v. 15, pp. 127-131, Zaragoza.

PINHEIRO, W. A., BARROS, R., XEXEO, G. B., *et al.*, 2008, "Framework para Modelagem de Processos usando Redes de Petri de Alto-Nível e Regras Ativas: um Enfoque sobre a Eliminação de Dados". In: *SBSI 2008 - IV Simpósio Brasileiro de Sistemas de Informação*, pp. 199-210.

PINHO, S. F. C., 2001, *Avaliação da Qualidade de Dados pela Não Conformidade*, Dissertação de Mestrado, COPPE/UFRJ, Rio de Janeiro, Brasil.

PIPINO, L. L., LEE, Y. W., WANG, R. Y., 2002, "Data quality assessment". In: *Communications of the ACM*, v. 45, pp. 211-218.

RAMOS-LIMA, L. F. R., MAÇADA, A. C. G., VARGAS, L. M., 2006, "Research Into Information Quality: A Study of the State-Of-The Art in Iq and Its

Consolidation (Academic Paper)". In: *Proceedings of the 11th International Conference on Information Quality ICIQ'2006*, pp. 146-158.

REDMAN, T. C., 1997, ***Data Quality for the Information Age***. 1 ed. Boston, Artech House.

REDMAN, T. C., 1998, "The Impact of Poor Data Quality on the Typical Enterprise". In: *Communications of the ACM*, v. 41, pp. 79-81.

ROCHA, A. R. C., 1983, *Um Modelo para Avaliação da Qualidade de Especificações*, Tese de Doutorado, PUC-RJ, Rio de Janeiro, Brasil.

RODRIGUES NT, J. A., SOUZA, J. M., ZIMBRAO, G., *et al.*, 2006, "A P2P Approach for Business Process Modeling and Reuse. In: Business Process Management". In: *LNCS Business Process Management Workshops*, v. 4103, Berlin, Springer.

ROSS, T. J., 2004, ***Fuzzy Logic with Engineering Applications***. 2nd ed. ed. West Sussex, England.

ROTHENBERG, J., 1996, "Metadata to support data quality and longevity". In: *Proceedings of the 1st IEEE Metadata Conference*, pp. 16-18, Silver Spring, Md.

ROYCE, W., 1990, "Pragmatic Quality Metrics for Evolutionary Software Development Models". In: *Proceedings of the Conference on TRI-ADA '90*, pp. 551-565, Baltimore, Maryland, USA.

RÖMER, C.,KANDEL, A., 1995, "Statistical tests for fuzzy data", ***Fuzzy Sets Systems***, v. 72, n. 1, pp. 1-26.

RUBEY, R. J.,HARTWICK, R. D., 1968, "Quantitative Measurement of Program Quality". In: *Proceedings of the ACM National Conference 1968*, pp. 671-677.

SCHAFER, J. B., KONSTAN, J. A., RIEDL, J., 2001, "E-Commerce Recommendation Applications". In: *Proceedings of Data Mining and Knowledge Discovery*, pp. 115-153.

SCHAUPP, C., FAN, W., BELANGER, F., 2006, "Determining success for different Web site goals". In: *Proceedings of the 39th Hawaii International Conference on Systems Science*, v. 6, pp. 107.2-, Kauai, HI. USA.

SDORRA, P. B., 1993, "A measure-theoretic axiomatization of fuzzy sets", ***Fuzzy Sets and Systems 60***, v. 60, n. 1993, pp. 295-307.

SHARMA, S., 2008, "Information Retrieval in Domain Specific Search Engine with Machine Learning Approaches". In: *Proceeding of World Academy of Science, Engineering and Technology*, v. 32.

SMART, K. L., 2002, "Assessing Quality Documents", ***ACM Journal of Computer Documentation***, v. 26, n. 3, pp. 130-140.

STRONG, D.,WANG, R. Y., 1996, "Beyond Accuracy: What Data Quality Means to Data Consumers", ***Journal of Management Information Systems***, v. 12, n. 4, pp. 5-33.

STRONG, D. M., LEE, Y. W., WANG, R. Y., 1997, "Data Quality in Context", ***CACM***, v. 40, n. 5, pp. 103-110.

STVILIA, B., GASSER, L., TWIDALE, M. B., *et al.*, 2006, "A Framework for Information Quality Assessment", ***Journal of the American Society for Information Science and Technology***, v. 58, n. 12, pp. 1720-1733.

STVILIA, B., 2007, "A model for Information Quality Change". In: *Proceedings of the 12th International Conference on Information Quality ICIQ'2007*.

SUZUKI, H., 1993, "Fuzzy sets and membership functions". In: *Fuzzy Sets and Systems*, v. 58, pp. 123-132.

TARAPANOFF, K., 2002, ***Inteligência Organizacional e Competitiva***. Brasilia, UNB.

TERVEEN, L.,HILL, W., 2001, "Beyond Recommender System: Helping People to Find Each Other". In: *Proceedings of HCI in the New Millennium, Jack Caroll*.

TILLMAN, H. N., 2003, "Evaluating Quality on the Net". In: http://www.hopetillman.com/findqual.html, Accessed in 05/03/2006.

TRAVASSOS, G. H., BARROS, M. O., WERNER, C. M. L., 2002, "Um Estudo Experimental sobre a Utilização de Modelagem e Simulação no Apoio à Gerência de Projetos de Software". In: *XVI Simpósio Brasileiro de Engenharia de Software*, Gramado, RS.

TURKSEN, I. B., 1991, "Measurement of membership functions and their acquisition". In: *Fuzzy Sets and Systems IFSA*, v. Special Memorial Volume: 25 years of fuzzy sets, North-Holland, Amsterdam.

TWIDALE, M. B.,MARTY, P. F., 1999, *An Investigation of Data Quality and Collaboration*. In: University of Illinois at Urbana-Champaign, Technical Report UIUCLIS--1999/9+CSCW.

WAND, Y.,WANG, R., 1996, "Anchoring data quality dimensions in ontological foundations". In: *Communications of the ACM*, v. 39, pp. 86-95.

WANG, R. Y., REDDY, M. P., KON, H. B., 1995, "Toward Quality Data: An Attribute-Based Approach. Decision Support Systems". In: *Communications of the ACM*, v. 13, pp. 349-372.

WANG, R. Y., KON, H. B., MADNICK, S., 1993, "Data Quality Requirements Analysis and Modeling". In: *The Proceeding of the 9th International Conference on Data Engineering*, pp. 670-677.

WANG, R. Y., 1998, "A product perspective on total data quality management". In: *Communications of the ACM*, v. 41, pp. 58-65, New York, NY, USA.

WINKLER, W., 2004, "Methods for evaluating and creating data quality", ***Information Systems***, v. *29*, n. 7, pp. 531-550.

WOHLIN, C., RUNESON, P., HÖST, M., *et al.*, 2000, ***Experimentation in Software Engineering: an Introduction***. Norwell, MA, Kluver Academic Publishers.

WRIGHT, A., 2009, "Exploring a 'Deep Web' That Google Can't Grasp". In: http://www.nytimes.com/2009/02/23/technology/internet/23search.html?th&emc=th, Accessed in 23/02/2009.

YAGER R.R., 1988, "On ordered weighted averagingh aggregation operators in multicriteria decision making", ***IEEE Trans.on Systems, Man and Cybernectics,*** v. 18, n. 1, pp. 183-190.

YAGER, R. R., 1991, "Connectives and quantifiers in fuzzy sets". In: *Fuzzy Sets and Systems, IFSA, Special Memorial*, v. 25 years of fuzzy sets, North-Holland, Amsterdam.

YAGER, R. R., 1994, "Aggregation operators and fuzzy systems modeling", ***Fuzzy Sets and Systems***, v. 67, pp. 129-145.

YANG, Z., CAI, S., ZHOU, Z., *et al.*, 2004, "Development and validation of an instrument to measure user perceived service quality of information presenting Web portals", ***Information and Management***, v. *42*, pp. 575-589.

ZADEH, L. A., 1965, "Fuzzy sets", ***Information and Control***, n. 8, pp. 338-353.

ZADEH, L. A., 1973, "Outline of a new approach to the analysis of complex systems and decision process", ***IEEE Trans.on Systems, Man, Cybernectics, SMCL***, v. 1.

ZADEH, L. A., 1977, "A theory of approximate reasoning", ***Memorandoum no.UCB/ERLM 77/58, in (TURKSEN, 1991).***

ZADEH, L. A., 1978, "Fuzzy sets as a basis for the theory of possibiliy", ***Fuzzy Sets and Systems 1***, v. 100, n. supp., pp. 9-34.

ZADEH, L. A., 1988, "Fuzzy logic". In: *IEEE Transaction Comput*, v. 35.

ZADEH, L. A., 1990, "The Birth and Evolution of Fuzzy Logic. A Personal Perspective. Part 1 and 2.", ***Journal of Japan Society for Fuzzy Theory and Systems***, v. 11, n. 1, pp. 891-905.

ZHU, X.,GAUCH, S., 2000, "Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web". In: *Proceedings of the 23rd Annual International ACM SIGIR*, pp. 288-295, Athens, Greece.

ZHU, Y.,BUCHMANN, A., 2002, "Evaluating and selecting Web sources as external information resources of a data warehouse". In: *Proceedings of the 3rd International Conference on Web Information System Engineering*, pp. 149-160.

ZIMMERMANN, H. J., 1991, ***Fuzzy Set Theory and Its Applications***. 2 ed. Boston, Kluwer.

ZIMMERMANN, H. J., 1997, "Operators in Models of Decision Making". In: D.Dubois, H. Prade R. R. Yager, *Fuzzy Information Engineering: a guided tour of application*, chapter 29, New York, Jonh Wiley & Sons, Inc.

# Anexo I – Descrição das Dimensões de Qualidade

**ACESSIBILIDADE –** É o grau de disponibilidade dos dados, facilidade e rapidez com a qual eles podem ser recuperados.

**ACOMPANHAMENTO –** Determina se existe um responsável pelo dado.

**ACURÁCIA –** Determina se os dados estão corretos, confiáveis e certificados, ou seja, livres de erro.

**APLICABILIDADE –** Determina se os dados são específicos, úteis e facilmente aplicáveis pela comunidade a que se destinam.

**ARMAZENAMENTO EFICIENTE –** Determina se há uma gerência sobre o espaço ocupado em disco.

**ATRATIVIDADE –** Expressa o interesse despertado de um *site Web* para os seus visitantes.

**ATUALIDADE –** Determina se os dados são suficientemente atualizados ao propósito em questão.

**AUSÊNCIA DE AMBIGÜIDADES –** Expressa se cada item de informação tem um único significado.

**COMPLETEZA –** Expressa se os dados disponibilizados são suficientemente completos em largura, profundidade e escopo para a tarefa em questão.

**CONCISÃO –** Expressa se os dados são representados de forma compacta e não são prolixos.

**CONCORDÂNCIA SOBRE O USO –** É o grau no qual o dado atende às necessidades de grupos diferentes de usuários de forma que eles concordem utilizá-lo ao invés de criar elementos próprios.

**CONFIABILIDADE –** Expressa o grau de confiança que usuários têm nos dados e suas fontes.

**CONSISTÊNCIA INTERNA –** Determina se os dados pertencentes a diferentes conjuntos estão consistentes entre si.

**CONTROLE DE REDUNDÂNCIA –** Determina se as redundâncias existentes são apenas aquelas planejadas com propósitos específicos.

**CONTROLE DE VERSÕES –** Expressa o controle das diferentes versões dos dados.

**CREDIBILIDADE –** Determina se os dados e as suas fontes são aceitos como corretos.

**CUSTO ADEQUADO –** Determina se os custos de obtenção, gerência da qualidade, armazenamento e outros custos relacionados ao dado estão dentro do valor previsto.

**DISPONIBILIDADE –** É o grau de disponibilidade dos dados para o propósito em questão.

**DOCUMENTAÇÃO –** Determina a quantidade e utilidade dos documentos contendo metainformações.

**DUPLICIDADE –** Determina se os dados fornecidos contêm duplicatas.

**ESPECIALIZAÇÃO –** Expressa a especificidade dos dados disponibilizados em um *site Web*.

**EXISTÊNCIA DE METADADOS –** Expressa a existência de descrições sobre o significado dos dados.

**EXPIRAÇÃO –** Determina a data até a qual os dados permanecem reconhecidamente atualizados.

**FACILIDADE DE OBTENÇÃO –** Expressa a facilidade com que os dados são coletados e armazenados.

**FACILIDADE DE OPERAÇÃO –** Expressa se os dados são de fácil manuseio e manipulação (ou seja, atualizados, removidos, agregados, etc.) aplicados a diferentes propósitos.

**FLEXIBILIDADE –** Determina se os dados são expansíveis, adaptáveis e facilmente aplicados a outras necessidades.

**GRANURALIDADE –** Determina se a granularidade dos dados é adequada às necessidades.

**HOMOGENEIDADE –** Determina se os atributos possuem apenas um significado, ou seja, não existe mais de um tipo de informação sendo representado pelo mesmo atributo.

**IMPORTÂNCIA –** Expressa a importância dos dados para o seu destino. Está relacionada com a possibilidade de existirem processos no destino que dependam dos dados.

**INFORMAÇÕES DA FONTE –** Expressa se as informações sobre o autor/dono dos dados estão disponíveis aos consumidores dos dados.

**INTEGRIDADE –** Determina a manutenção consistente das estruturas de dados e do relacionamento entre entidades.

**INTELIGIBILIDADE –** Determina se os dados são claros, sem ambigüidades e facilmente compreensíveis.

**INTERATIVIDADE –** Expressa se a maneira pela qual os dados são acessados ou recuperados pode ser adaptada às preferências pessoais, por meio de elementos interativos.

**INTERPRETABILIDADE –** Expressa se os dados estão no idioma e unidades adequadas e as definições são claras de acordo com a capacidade dos seus consumidores.

**LATÊNCIA OU TEMPO DE RESPOSTA –** Determina a quantidade de tempo até a resposta completa chegar ao usuário.

**MANUTENIBILIDADE –** É a capacidade de modificação dos dados.

**NATURALIDADE –** Refere-se ao relacionamento entre o objeto modelado e o mundo real, significando que cada atributo representa um fato sobre um objeto do mundo real e os valores do domínio ao qual pertence espelham corretamente a realidade.

**NOVIDADE –** Expressa se os dados obtidos têm influência sobre o conhecimento e as novas decisões.

**OBJETIVIDADE –** Determina se os dados são não tendenciosos e imparciais sob o ponto de vista de uma única interpretação e independentemente do observador.

**ONTOLOGIA –** Expressa a descrição e o conhecimento de esquemas-fonte importantes na condução do processo de integração de dados.

**OPORTUNIDADE –** Expressa a disponibilidade dos dados no tempo esperado, de acordo com os requisitos de tempo especificados pelo destino.

**ORGANIZAÇÃO –** Determina a organização, configurações visuais ou características tipográficas (cor, texto, fonte, imagens, etc.) e as combinações consistentes desses diferentes componentes.

**ORIGEM IDENTIFICADA –** Explicita a origem dos dados.

**PORTABILIDADE –** Determina se a interface de apresentação dos dados suporta a migração da aplicação para plataformas diferentes.

**PRECISÃO –** É o grau no qual o valor do dado corresponde a um valor aproximado em relação ao valor real.

**PREÇO –** Determina normalmente qualquer encargo numa base de subscrição para acesso às informações ou numa base pagamento-por-consulta ou de pagamento-por-*byte* de fontes de dados comerciais. Muitas vezes existe uma compensação direita entre o preço e outros critérios de QI.

**PRIVACIDADE –** Determina a existência de mecanismos que garantam a privacidade de acesso aos dados (graus de sigilo, por exemplo).

**QUANTIDADE DADOS –** Determina se a quantidade ou o volume de dados fornecidos é adequado ao propósito em questão.

**RASTREABILIDADE –** Expressa se os dados são bem documentados, verificáveis e facilmente relacionados a uma fonte.

**REGISTROS DE APELIDOS –** Expressa se apelidos (alias) dos dados estão associados aos mesmos.

**RELEVÂNCIA –** Determina se os dados são aplicáveis e úteis para as necessidades dos usuários.

**REPRESENTAÇÃO CONCISA –** Determina se os dados são compactos sem elementos supérfluos ou não relacionados.

**REPRESENTAÇÃO CONSISTENTE –** Determina se os dados são apresentados de forma consistente com o seu domínio e com os demais dados similares.

**REPUTAÇÃO –** Expressa se os dados são confiáveis ou altamente recomendados em termos de sua origem ou conteúdo.

**ROBUSTEZ –** É a habilidade de refletir mudanças do mundo real no modelo de dados, alterando-o o mínimo possível.

**SEGURANÇA –** Determina a existência de mecanismos que garantam o controle de acesso aos dados.

**SIMPLICIDADE –** Expressa se o modelo de dados não é complexo, sendo de fácil compreensão.

**SUPORTE AO USUÁRIO –** Determina se o *site Web* oferece suporte on-line por meio de texto, de e-mail, de telefone, etc.

**TEMPO DE RESPOSTA OU LATÊNCIA –** Determina a quantidade de tempo até a resposta completa chegar ao usuário.

**TRATAMENTO DE VALORES NULOS –** Expressa a possibilidade de identificar o motivo pelo qual os dados não estão presentes: Se estão indisponíveis, se

não aplicáveis, se são desconhecidos ou se não existe valor no domínio que os represente corretamente.

**UBIQÜIDADE –** É o grau no qual diferentes consumidores de dados compartilham o dado.

**UNICIDADE –** Determina se cada objeto de dado é identificado de forma única.

**UTILIDADE –** Determina a utilidade dos dados para o propósito em questão.

**VALIDADE –** Expressa se os utilizadores podem avaliar e compreender os dados fornecidos.

**VALOR AGREGADO –** Expressa os benefícios proporcionados e vantagens de uso.

# Anexo II – Resultados das Avaliações Automática e Manual

| URL avaliada | Avaliação Manual | Escore Manual | Avaliação Automática | Escore Automático | Manual vs. Automática |
|---|---|---|---|---|---|
| http://2006.xmlconference.org/programme/presentations/188.html | Ruim | (-1) | Péssima | 0,281161635 | 0 |
| http://briefingsdirectblog.blogspot.com/2007/10/oracle-users-enjoyopen-source-benefits.html | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://cidoc.ics.forth.gr/docs/Implementing_the_CIDOC_CRM.rtf | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://codingforums.com/showthread.php?t=44774 | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://comjnl.oxfordjournals.org/cgi/content/abstract/39/2/124?ck=nck | Péssima | (-2) | Péssima | 0,275539265 | Péssima |
| http://computerprogramming.suite101.com/article.cfm/sql_server_training_first_steps | Ruim | (-1) | Ruim | 0,474459466 | Ruim |
| http://d3dnff.gat.com/D3DRDB/ | Péssima | (-2) | Ruim | 0,474459466 | 0 |
| http://db.uwaterloo.ca/OED/trdbms.html | Péssima | (-2) | Péssima | 0,278988836 | Péssima |
| http://duro.sourceforge.net/ | Ruim | (-1) | Ruim | 0,566737401 | Ruim |
| http://en.wikibooks.org/wiki/Category:Relational_Database_Design | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://encyclopedia.kids.net.au/page/re/Relational_database | Regular | (0) | Regular | 0,590521421 | Regular |
| http://foldoc.org/?relational+database | Ruim | (-1) | Ruim | 0,521276164 | Ruim |
| http://gamma.cs.unc.edu/DB/ | Regular | (0) | Péssima | 0,27382844 | 0 |
| http://gateway.nlm.nih.gov/MeetingAbstracts/102211909.html | Péssima | (-2) | Péssima | 0,274061965 | Péssima |
| http://gateway.nlm.nih.gov/MeetingAbstracts/102276723.html | Ruim | (-1) | Ruim | 0,566737401 | Ruim |
| http://gd.tuwien.ac.at:8050/H/1/ | Regular | (0) | Péssima | 0,27382844 | 0 |
| http://ilpubs.stanford.edu:8090/4/ | Ruim | (-1) | Ruim | 0,566737401 | Ruim |
| http://libra.msra.cn/papercited.aspx?id=352778 | Ruim | (-1) | Ruim | 0,566737401 | Ruim |
| http://lips.informatik.uni-leipzig.de/pub/1998-8 | Ruim | (-1) | Ruim | 0,540526751 | Ruim |

| URL avaliada | Avaliação Manual | Escore Manual | Avaliação Automática | Escore Automático | Manual vs. Automática |
|---|---|---|---|---|---|
| http://llc.oxfordjournals.org/cgi/content/abstract/2/2/89 | Ruim | (-1) | Ruim | 0,540526751 | Ruim |
| http://news.cnet.com/Rethinking-the-relational-database/2010-1015_3-5715457.html | Regular | (0) | Péssima | 0,27382844 | 0 |
| http://pd.acm.org/book_detail.cfm?isbn=0321305965 | Péssima | (-2) | Ruim | 0,540526751 | 0 |
| http://pooteeweet.org/files/phpconf05/relational_database_starter_da.pdf | Ruim | (-1) | Ruim | 0,528020469 | Ruim |
| http://publish.uwo.ca/~craven/558/558red.htm | Regular | (0) | Regular | 0,590521421 | Regular |
| http://relationaldatabasesoftware.surfpack.com/ | Regular | (0) | Péssima | 0,27382844 | 0 |
| http://research.microsoft.com/apps/pubs/default.aspx?id=64571 | Regular | (0) | Regular | 0,590521421 | Regular |
| http://scripts.iucr.org/cgi-bin/paper?S0108768102002458 | Péssima | (-2) | Péssima | 0,27728055 | Péssima |
| http://stats.oecd.org/glossary/detail.asp?ID=4520 | Regular | (0) | Regular | 0,590521421 | Regular |
| http://uttc.umn.edu/training/courses/description?designator=DB101 | Péssima | (-2) | Péssima | 0,274031506 | Péssima |
| http://www.100best-web-hosting.com/glossary674.html | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.15seconds.com/Issue/020522.htm | Regular | (0) | Regular | 0,590521421 | Regular |
| http://www.adobe.com/devnet/coldfusion/articles/display_dyn_data_02.html | Regular | (0) | Péssima | 0,27382844 | 0 |
| http://www.agiledata.org/essays/mappingObjects.html | Ruim | (-1) | Ruim | 0,574116285 | Ruim |
| http://www.altova.com/press/2003-01-13_hitsw.pdf | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.arcchip.cz/w05/w05_buitleir.pdf | Ruim | (-1) | Ruim | 0,574116285 | Ruim |
| http://www.athro.com/general/Phyloinformatics_7_85x11.pdf | Regular | (0) | Regular | 0,590803174 | Regular |
| http://www.auditmypc.com/acronym/ORDBMS.asp | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.blackwell-synergy.com/doi/abs/10.1111/j.1540-8159.1996.tb03421.x | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.bmrb.wisc.edu/search/rela_database.html | Ruim | (-1) | Ruim | 0,574116285 | Ruim |
| http://www.builderau.com.au/program/xml/soa/Transfer-and-store-data-from-an-XML-document-in-a-relational-database/0,339028469,339273299,00.htm | Regular | (0) | Péssima | 0,27382844 | 0 |
| http://www.cemml.colostate.edu/files/un5.pdf | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.codata.org/08conf/abstracts/ChangjunHu-A%20Visual%20Tool%20for%20Building%20MatML%20Data%20from%20Material%20Science%20Relational%20Database.htm | Ruim | (-1) | Péssima | 0,436151202 | 0 |

| URL avaliada | Avaliação Manual | Escore Manual | Avaliação Automática | Escore Automático | Manual vs. Automática |
|---|---|---|---|---|---|
| http://www.codeproject.com/KB/database/introtomatisse_part2.aspx | Ruim | (-1) | Péssima | 0,27382844 | 0 |
| http://www.codeproject.com/KB/showcase/object_relational_mapping.aspx | Regular | (0) | Péssima | 0,27382844 | 0 |
| http://www.defmacro.org/ramblings/relational.html | Regular | (0) | Regular | 0,590803174 | Regular |
| http://www.developers.net/enterprisedbshowcase/view/1323 | Ruim | (-1) | Ruim | 0,544753047 | Ruim |
| http://www.eecs.berkeley.edu/Pubs/TechRpts/1978/12384.html | Regular | (0) | Péssima | 0,27382844 | 0 |
| http://www.empress.com/ | Regular | (0) | Regular | 0,590803174 | Regular |
| http://www.funpecrp.com.br/gmr/year2007/vol4-6//pdf/xm0012.pdf | Ruim | (-1) | Ruim | 0,544753047 | Ruim |
| http://www.gridpp.ac.uk/papers/GGF3Rome2001.pdf | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.informationweek.com/news/management/showArticle.jhtml?articleID=159401656 | Ruim | (-1) | Excelente | 0,940707479 | 0 |
| http://www.laas.fr/~esorics/notices/Yazdanian90.html | Péssima | (-2) | Péssima | 0,274600088 | Péssima |
| http://www.lavoisier.fr/notice/gbBCO3AXSM3OP2RO.html | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.leavcom.com/db_08_00.htm | Regular | (0) | Péssima | 0,274072119 | 0 |
| http://www.midcarb.org/Documents/GSA-Nov-2000.shtml | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.mindmodel.com/ | Péssima | (-2) | Ruim | 0,544753047 | 0 |
| http://www.ohloh.net/tags/database | Péssima | (-2) | Péssima | 0,27382844 | Péssima |
| http://www.onesmartclick.com/engineering/relational-databases.html | Regular | (0) | Regular | 0,590803174 | Regular |
| http://www.oracle.com/database/docs/Berkeley-DB-v-Relational.pdf | Boa | (1) | Péssima | 0,275902244 | 0 |
| http://www.pinnaclekenya.com/ | Péssima | (-2) | Péssima | 0,274381793 | Péssima |
| http://www.postgresql.org/ | Boa | (1) | Boa | 0,720084678 | Boa |
| http://www.questia.com/PM.qst?a=o&se=gglsc&d=5001678243 | Péssima | (-2) | Péssima | 0,274092425 | Péssima |
| http://www.serc.iisc.ernet.in/ComputingFacilities/systems/cluster/vac-7.0/html/glossary/czgr.htm | Ruim | (-1) | Ruim | 0,544753047 | Ruim |
| http://www.springerlink.com/index/7442wt8574x44766.pdf | Ruim | (-1) | Péssima | 0,27382844 | 0 |
| http://www.sqlconsole.com/ | Péssima | (-2) | Péssima | 0,274021352 | Péssima |

| URL avaliada | Avaliação Manual | Escore Manual | Avaliação Automática | Escore Automático | Manual vs. Automática |
|---|---|---|---|---|---|
| http://www.stylusstudio.com/xml_database.html | Ruim | (-1) | Péssima | 0,27382844 | 0 |
| http://www.techbriefs.com/component/content/68?task=view | Péssima | (-2) | Excelente | 1 | 0 |
| http://www.thunderstone.com/site/texisman/relational_database_background.html | Ruim | (-1) | Péssima | 0,455217929 | 0 |
| http://www.vertica.com/relational-database-management-system | Ruim | (-1) | Ruim | 0,954422506 | Ruim |
| http://www.w3.org/XML/RDB.html | Boa | (1) | Péssima | 0,455217929 | 0 |
| http://www.warthman.com/projects-tymshare-b.htm | Péssima | (-2) | Péssima | 0,457121035 | Péssima |
| http://www.xml.com/pub/a/2007/07/12/xquery-and-data-abstraction.html | Boa | (1) | Ruim | 0,824454281 | 0 |
| http://www.yolinux.com/HOWTO/PostgreSQL-HOWTO.html | Excelente | (2) | Péssima | 0,455281225 | 0 |
| http://www.zope.org/Documentation/Books/ZopeBook/2_6Edition/Relational Databases.stx | Boa | (1) | Péssima | 0,455217929 | 0 |
| http://www3.open.ac.uk/courses/bin/p12.dll?C01M359 | Regular | (0) | Regular | 1 | Regular |

# Anexo III – Questionário de Avaliação Manual

# Avaliação de páginas - Bancos de dados relacionais

Bem-vindo, **rbarros@cos.ufrj.br**! Obrigado por aceitar participar da nossa pesquisa!

## Sobre você

Nome: Ricardo Barros

Escolaridade: Mestrado

Atividade: Ambos

## Auto-avaliação

Qual o seu conhecimento sobre o assunto (**Bancos de dados relacionais**)? Excelente

## Avaliação das páginas

Você será responsável por avaliar **10 páginas** no contexto de Bancos de dados relacionais. Para **cada página** apresentada, você deve:

1. Visitá-la clicando no link fornecido (você pode mantê-la aberta durante a avaliação e consultá-la sempre que preciso).

   Em seguida, você responderá às seguintes **3 perguntas**, escolhendo **Sim** ou **Não**:

2. A página explica o que é um SGBDR?
3. A página explica as vantagens de um SGBDR?
4. A página trata do uso de SQL em SGBDR?

   Depois, você avaliará **4 atributos** de qualidade da página, qualificando cada atributo como **Péssima**, **Ruim**, **Regular**, **Boa** ou **Excelente**:

5. Reputação
   *Avaliação da página considerando a sua fonte e o seu conteúdo.*
6. Completeza
   *Avaliação da página considerando a amplitude e profundidade do assunto tratado.*
7. Atualidade
   *Avaliação da página considerando se ela é suficientemente atualizada.*
8. Avaliação Global
   *Avaliação da página no geral.*

   Por fim, você informará se a página é relevante para o assunto (Bancos de dados relacionais).

   Se você tiver quaisquer dúvidas, por favor, envie um e-mail para **rbarros@cos.ufrj.br** com cópia para **bernardo.wpacheco@gmail.com**.

### Página 1

1. Visite a página clicando aqui.

   Perguntas:

2. A página explica o que é um SGBDR?
   Não
3. A página explica as vantagens de um SGBDR?
   Não
4. A página trata do uso de SQL em SGBDR?
   Não

   Atributos de qualidade:

5. Reputação
   Boa
6. Completeza
   Ruim
7. Atualidade
   Ruim
8. Avaliação Global
   Ruim

   A página é relevante para o assunto (Bancos de dados relacionais)? Não

**Página 2**

1. Visite a página clicando aqui.

   Perguntas:

2. A página explica o que é um SGBDR?
   Sim
3. A página explica as vantagens de um SGBDR?
   Sim
4. A página trata do uso de SQL em SGBDR?
   Sim

   Atributos de qualidade:

5. Reputação
   Boa
6. Completeza
   Excelente
7. Atualidade
   Excelente
8. Avaliação Global
   Excelente

   A página é relevante para o assunto (Bancos de dados relacionais)? Sim

# Anexo IV – Cálculo Precisão e Cobertura pelos Valores de Ordenação do Google®

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.dbmsmag.com/9804d13.html | 1 | 1 | 0,004310345 | 1 | 0,892178 | 0,008583691 |
| http://highscalability.com/search-source-data-how-simpledb-differs-rdbms | 1 | 2 | 0,00862069 | 1 | 0,892143 | 0,017094017 |
| http://www-10.lotus.com/ldd/lcdforum.nsf/DateAllThreadedweb/4655b9a7d2bdab49852572e300240dc9?OpenDocument | 1 | 3 | 0,012931034 | 1 | 0,892109 | 0,025531915 |
| http://www.turnkeylinux.org/appliances/postgresql | 1 | 4 | 0,017241379 | 1 | 0,89195 | 0,033898305 |
| http://connect.educause.edu/Library/Abstract/TheEffectofRelatioalData/30473 | 1 | 5 | 0,021551724 | 1 | 0,891883 | 0,042194093 |
| http://www.jumpbox.com/app/mysqld | 1 | 6 | 0,025862069 | 1 | 0,891817 | 0,050420168 |
| http://studyat.anu.edu.au/courses/COMP2400;details.html | 0 | 6 | 0,025862069 | 0,857142857 | 0,891514 | 0,050209205 |
| http://www.xml.com/pub/a/2003/03/05/tmrdb.html | 1 | 7 | 0,030172414 | 0,875 | 0,891225 | 0,058333333 |
| http://plone.org/documentation/faq/plone-relational-database | 1 | 8 | 0,034482759 | 0,888888889 | 0,891012 | 0,066390041 |
| http://www.ebookee.com/The-Relational-Database-Dictionary-Extended-Edition_181022.html | 1 | 9 | 0,038793103 | 0,9 | 0,890238 | 0,074380165 |
| http://gmod.org/wiki/Glossary | 1 | 10 | 0,043103448 | 0,909090909 | 0,889792 | 0,082304527 |
| http://en.wikipedia.org/wiki/Relational_model | 1 | 11 | 0,047413793 | 0,916666667 | 0,889642 | 0,090163934 |
| http://fyi.oreilly.com/2008/11/relational-database-technology.html | 1 | 12 | 0,051724138 | 0,923076923 | 0,889131 | 0,097959184 |
| http://www.sirdug.org/ | 1 | 13 | 0,056034483 | 0,928571429 | 0,888786 | 0,105691057 |
| http://www.openldap.org/faq/data/cache/378.html | 1 | 14 | 0,060344828 | 0,933333333 | 0,888411 | 0,113360324 |
| http://www.objenv.com/cetus/oo_db_systems_2.html | 1 | 15 | 0,064655172 | 0,9375 | 0,888218 | 0,120967742 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://en.wikipedia.org/wiki/Relational_database_management_system | 1 | 16 | 0,068965517 | 0,941176471 | 0,88783 | 0,128514056 |
| http://www.garret.ru/gigabase.html | 1 | 17 | 0,073275862 | 0,944444444 | 0,887633 | 0,136 |
| http://linuxfinances.info/info/rdbms.html | 1 | 18 | 0,077586207 | 0,947368421 | 0,886997 | 0,143426295 |
| http://criminaljustice.state.ny.us/crimnet/clf/oracle/oracle.htm | 0 | 18 | 0,077586207 | 0,9 | 0,886783 | 0,142857143 |
| http://ideas.repec.org/p/boc/nsug08/13.html | 1 | 19 | 0,081896552 | 0,904761905 | 0,886343 | 0,150197628 |
| http://simple.wikipedia.org/wiki/Relational_database | 1 | 20 | 0,086206897 | 0,909090909 | 0,885415 | 0,157480315 |
| http://www.geekgirls.com/database_dictionary.htm | 1 | 21 | 0,090517241 | 0,913043478 | 0,884931 | 0,164705882 |
| http://www.matisse.com/ | 0 | 21 | 0,090517241 | 0,875 | 0,884683 | 0,1640625 |
| http://www.hitsw.com/products_services/downloads.html | 1 | 22 | 0,094827586 | 0,88 | 0,884434 | 0,171206226 |
| http://www.techbookreport.com/tbr0273.html | 0 | 22 | 0,094827586 | 0,846153846 | 0,884182 | 0,170542636 |
| http://www.sapdb.org/ | 1 | 23 | 0,099137931 | 0,851851852 | 0,88393 | 0,177606178 |
| http://blog.terracottatech.com/2008/11/breaking_down_the_ relational_d.html | 1 | 24 | 0,103448276 | 0,857142857 | 0,883407 | 0,184615385 |
| http://www.constable.com/ | 0 | 24 | 0,103448276 | 0,827586207 | 0,883142 | 0,183908046 |
| http://its.psu.edu/training/handouts/gs_reldb_sp06.pdf | 1 | 25 | 0,107758621 | 0,833333333 | 0,882606 | 0,190839695 |
| http://db.apache.org/derby/ | 1 | 26 | 0,112068966 | 0,838709677 | 0,880639 | 0,197718631 |
| http://www.annauniv.edu/rcc/meseorSyllabus/SE072.pdf | 0 | 26 | 0,112068966 | 0,8125 | 0,879755 | 0,196969697 |
| http://www.altova.com/features_database.html | 1 | 27 | 0,11637931 | 0,818181818 | 0,878852 | 0,203773585 |
| http://download.microsoft.com/download/8/f/a/8fa3268a-d34f-4b3d-bb72-72e08701096f/Worldwide%20Relational%20 Database%20Management%20Systems%202007%20Vendor%20 Shares.pdf | 0 | 27 | 0,11637931 | 0,794117647 | 0,878545 | 0,203007519 |
| http://www.itjungle.com/tug/tug071008-story05.html | 1 | 28 | 0,120689655 | 0,8 | 0,877294 | 0,209737828 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://lists.xml.org/archives/xml-dev/200102/msg00148.html | 0 | 28 | 0,120689655 | 0,777777778 | 0,876657 | 0,208955224 |
| http://lists.xml.org/archives/xml-dev/200102/msg00009.html | 1 | 29 | 0,125 | 0,783783784 | 0,875353 | 0,215613383 |
| http://www.epa.gov/enviro/html/ef_home.html | 0 | 29 | 0,125 | 0,763157895 | 0,864889 | 0,214814815 |
| http://java.sun.com/docs/books/tutorial/jdbc/overview/database.html | 1 | 30 | 0,129310345 | 0,769230769 | 0,860507 | 0,221402214 |
| http://www.jegsworks.com/Lessons/databases/intro/database-relational.htm | 1 | 31 | 0,13362069 | 0,775 | 0,856173 | 0,227941176 |
| http://www.databasecolumn.com/2007/11/dbms-origins.html | 1 | 32 | 0,137931034 | 0,780487805 | 0,851764 | 0,234432234 |
| http://www4.wiwiss.fu-berlin.de/bizer/d2rq/ | 1 | 33 | 0,142241379 | 0,785714286 | 0,847564 | 0,240875912 |
| http://ora.ouls.ox.ac.uk/objects/uuid:839dd46f-57f4-49ef-9582-8c154764a962 | 1 | 34 | 0,146551724 | 0,790697674 | 0,843493 | 0,247272727 |
| http://www.idealliance.org/proceedings/xml04/papers/254/XQueryRDS.pdf | 1 | 35 | 0,150862069 | 0,795454545 | 0,839449 | 0,253623188 |
| http://www.jcc.com/DescRelationalDBDesign.htm | 1 | 36 | 0,155172414 | 0,8 | 0,835442 | 0,259927798 |
| http://www.o-xml.org/news/18-mar-2003.html | 1 | 37 | 0,159482759 | 0,804347826 | 0,820018 | 0,26618705 |
| http://www.objectarchitects.de/arcus/cookbook/relzs/index.htm | 0 | 37 | 0,159482759 | 0,787234043 | 0,816366 | 0,265232975 |
| http://www.swc.scipy.org/lec/db.html | 1 | 38 | 0,163793103 | 0,791666667 | 0,812715 | 0,271428571 |
| http://itc.ktu.lt/itc353/Vysnia353.pdf | 1 | 39 | 0,168103448 | 0,795918367 | 0,805643 | 0,277580071 |
| http://www.vldb2005.org/program/paper/thu/p1175-pal.pdf | 1 | 40 | 0,172413793 | 0,8 | 0,8023 | 0,283687943 |
| http://homepages.inf.ed.ac.uk/sviglas/pubs/OrderedXML.pdf | 1 | 41 | 0,176724138 | 0,803921569 | 0,798974 | 0,28975265 |
| http://www.ifi.uzh.ch/arvo/dbtg/vldbphd2007/Camera-Ready%20Papers/Paper%206/XQuery_Optimization.pdf | 1 | 42 | 0,181034483 | 0,807692308 | 0,792441 | 0,295774648 |
| http://csweb.bournemouth.ac.uk/aip/AIP%20Master%20Database.pdf | 1 | 43 | 0,185344828 | 0,811320755 | 0,783318 | 0,301754386 |
| http://www.db.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/relational_database.htm | 1 | 44 | 0,189655172 | 0,814814815 | 0,777507 | 0,307692308 |
| http://gadfly.sourceforge.net/gadfly.html | 1 | 45 | 0,193965517 | 0,818181818 | 0,77465 | 0,31358885 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.bkent.net/Doc/simple5.htm | 1 | 46 | 0,198275862 | 0,821428571 | 0,771977 | 0,319444444 |
| http://www.ornl.gov/sci/techresources/Human_Genome/publicat/99santa/99.html | 1 | 47 | 0,202586207 | 0,824561404 | 0,769361 | 0,325259516 |
| http://en.wikiversity.org/wiki/Topic:Object-relational_databases | 1 | 48 | 0,206896552 | 0,827586207 | 0,766795 | 0,331034483 |
| http://codex.cs.yale.edu/avi/db-book/slide-dir/ch7.pdf | 1 | 49 | 0,211206897 | 0,830508475 | 0,764252 | 0,336769759 |
| http://www.ambysoft.com/mappingObjects.html | 1 | 50 | 0,215517241 | 0,833333333 | 0,759343 | 0,342465753 |
| http://www.vldb.org/conf/2004/IND5P2.PDF | 1 | 51 | 0,219827586 | 0,836065574 | 0,757062 | 0,348122867 |
| http://www.mail-archive.com/accessvbacentral@yahoogroups.com/msg00555.html | 1 | 52 | 0,224137931 | 0,838709677 | 0,754794 | 0,353741497 |
| http://searchoracle.techtarget.com/tip/0,289483,sid41_gci1217363,00.html | 0 | 52 | 0,224137931 | 0,825396825 | 0,752561 | 0,352542373 |
| http://www.cs.dal.ca/news/def-1127.shtml | 1 | 53 | 0,228448276 | 0,828125 | 0,748163 | 0,358108108 |
| http://www.warriorforum.com/programming-talk/48802relational-database.html | 1 | 54 | 0,232758621 | 0,830769231 | 0,74611 | 0,363636364 |
| http://pubs.water.usgs.gov/ofr01359 | 1 | 55 | 0,237068966 | 0,833333333 | 0,744092 | 0,369127517 |
| http://searchoracle.bitpipe.com/plist/term/Relational-Database-Management-Software.html | 1 | 56 | 0,24137931 | 0,835820896 | 0,742073 | 0,37458194 |
| http://www.treesearch.fs.fed.us/pubs/4548 | 1 | 57 | 0,245689655 | 0,838235294 | 0,740084 | 0,38 |
| http://searchsystemschannel.techtarget.com/tip/0,289483,sid99_gci1255470,00.html | 1 | 58 | 0,25 | 0,84057971 | 0,734384 | 0,38538206 |
| http://biophysics.biol.uoa.gr/gpDB/ | 0 | 58 | 0,25 | 0,828571429 | 0,732571 | 0,38410596 |
| http://www.archive.org/search.php?query=mediatype%3Aeducation%20AND%20collection%3Aarsdigita%20AND%20subject%3A22Relational%20Database%20Management%20Systems%22 | 1 | 59 | 0,254310345 | 0,830985915 | 0,728968 | 0,389438944 |
| http://www.reviews.com/review/review_review.cfm?review_id=136349 | 1 | 60 | 0,25862069 | 0,833333333 | 0,722155 | 0,394736842 |
| http://searchsystemschannel.techtarget.com/topics/0,295493,sid_tax305101,00.html | 1 | 61 | 0,262931034 | 0,835616438 | 0,715687 | 0,4 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://searchoracle.techtarget.com/generic/0,295582,sid41_gci1091031,00.html | 1 | 62 | 0,267241379 | 0,837837838 | 0,70655 | 0,405228758 |
| http://www.bibsonomy.org/bibtex/2b871e617ebe9672da36fe27790d06e5f/dblp | 0 | 62 | 0,267241379 | 0,826666667 | 0,70209 | 0,403908795 |
| http://www.sigcse.org/cc2001/IM.html | 1 | 63 | 0,271551724 | 0,828947368 | 0,700685 | 0,409090909 |
| http://www-03.ibm.com/ibm/history/exhibits/vintage/vintage_4506VV3151.html | 1 | 64 | 0,275862069 | 0,831168831 | 0,695042 | 0,414239482 |
| http://www.chass.utoronto.ca/emls/iemls/mqlibrary/search.html | 0 | 64 | 0,275862069 | 0,820512821 | 0,689469 | 0,412903226 |
| http://ci.nii.ac.jp/naid/110003223310/ | 1 | 65 | 0,280172414 | 0,82278481 | 0,686659 | 0,418006431 |
| http://mac.softpedia.com/get/Font-Tools/FrontBase.shtml | 1 | 66 | 0,284482759 | 0,825 | 0,683958 | 0,423076923 |
| http://stinet.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix=html&identifier=ADA313447 | 1 | 67 | 0,288793103 | 0,827160494 | 0,66217 | 0,428115016 |
| http://agron.scijournals.org/cgi/content/full/93/4/923 | 0 | 67 | 0,288793103 | 0,817073171 | 0,659496 | 0,426751592 |
| http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=326282 | 1 | 68 | 0,293103448 | 0,819277108 | 0,656903 | 0,431746032 |
| http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347395 | 1 | 69 | 0,297413793 | 0,821428571 | 0,654533 | 0,436708861 |
| http://www.securityfocus.com/bid/23007 | 0 | 69 | 0,297413793 | 0,811764706 | 0,652194 | 0,43533123 |
| http://www.ouhk.edu.hk/WCM/?FUELAP_TEMPLATENAME=tcGenericPage&itemid=CC_COURSE_INFO_58222950&lang=eng | 1 | 70 | 0,301724138 | 0,813953488 | 0,647492 | 0,440251572 |
| http://www.codebeach.com/index.asp?authorName=Relational%20Database%20Consultants | 0 | 70 | 0,301724138 | 0,804597701 | 0,632739 | 0,438871473 |
| http://www.ingentaconnect.com/content/els/09666362/1995/00000003/00000004/art82904;jsessionid=213d6w9tu8vsp.alexandra | 1 | 71 | 0,306034483 | 0,806818182 | 0,625264 | 0,44375 |
| http://www.zdnet.com.au/whitepaper/0,2000063328,22462862p-16001235q,00.htm | 0 | 71 | 0,306034483 | 0,797752809 | 0,618236 | 0,442367601 |
| http://www.logisticsworld.com/logistics/glossary.asp?query=Relational+Database+Management+System&search=exactterm&form=show&acr=show&ref=show&rel=show&srl=show&llk= | 1 | 72 | 0,310344828 | 0,8 | 0,614774 | 0,447204969 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| show&wiz=show&num=&hst=show&mode= | | | | | | |
| http://bioinformatics.oxfordjournals.org/cgi/content/abstract/14/2/188 | 1 | 73 | 0,314655172 | 0,802197802 | 0,609911 | 0,452012384 |
| http://portal.acm.org/citation.cfm?id=77707 | 0 | 73 | 0,314655172 | 0,793478261 | 0,600471 | 0,450617284 |
| http://www.fujipress.jp/finder/xslt.php?mode=present&inputfile=JACII000600010005.xml | 1 | 74 | 0,318965517 | 0,795698925 | 0,598921 | 0,455384615 |
| http://xml.coverpages.org/blake94-ps.gz | 0 | 74 | 0,318965517 | 0,787234043 | 0,595827 | 0,45398773 |
| http://research.microsoft.com/apps/pubs/default.aspx?id=64535 | 1 | 75 | 0,323275862 | 0,789473684 | 0,594363 | 0,458715596 |
| http://www.kirupa.com/developer/php/relational_db_design2.htm | 1 | 76 | 0,327586207 | 0,791666667 | 0,588343 | 0,463414634 |
| http://technet.microsoft.com/en-us/library/ms189559(SQL.90).aspx | 1 | 77 | 0,331896552 | 0,793814433 | 0,58689 | 0,468085106 |
| http://unicode.org/iuc/iuc13/c12/slides.ppt | 0 | 77 | 0,331896552 | 0,785714286 | 0,585444 | 0,466666667 |
| http://portal.acm.org/citation.cfm?id=171128 | 1 | 78 | 0,336206897 | 0,787878788 | 0,58095 | 0,471299094 |
| http://www.javaworld.com/javaworld/jw-09-2007/jw-09-columndb.html | 1 | 79 | 0,340517241 | 0,79 | 0,579482 | 0,475903614 |
| http://shopping.msn.com/prices/relational-database-design-clearly-explained/itemid2439036/?itemtext=itemname:relational database-design-clearly-explained | 1 | 80 | 0,344827586 | 0,792079208 | 0,570532 | 0,48048048 |
| http://www.cs.vt.edu/node/4585 | 1 | 81 | 0,349137931 | 0,794117647 | 0,569044 | 0,48502994 |
| http://www.turnkeylinux.org/appliances/mysql | 0 | 81 | 0,349137931 | 0,786407767 | 0,567567 | 0,48358209 |
| http://casoilresource.lawr.ucdavis.edu/drupal/node/264 | 1 | 82 | 0,353448276 | 0,788461538 | 0,565992 | 0,488095238 |
| http://www.thestandard.com/news/2008/02/04/start-readies-easy-use-online-relational-database | 1 | 83 | 0,357758621 | 0,79047619 | 0,558182 | 0,492581602 |
| http://arnab.org/blog/web-20-and-relational-database | 1 | 84 | 0,362068966 | 0,79245283 | 0,551592 | 0,49704142 |
| http://developer.mimer.com/ | 1 | 85 | 0,36637931 | 0,794392523 | 0,527492 | 0,501474926 |
| http://www.zdnet.co.uk/tsearch/databases+relational+database.htm | 1 | 86 | 0,370689655 | 0,796296296 | 0,525349 | 0,505882353 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.onjava.com/pub/a/onjava/2006/04/12/object-to-relational-database-replciation-with-db40.html | 1 | 87 | 0,375 | 0,798165138 | 0,521028 | 0,51026393 |
| http://www.onlamp.com/pub/a/onlamp/2001/05/25/postgresql_mvcc.html | 1 | 88 | 0,379310345 | 0,8 | 0,518867 | 0,514619883 |
| http://computer.howstuffworks.com/question599.htm | 1 | 89 | 0,38362069 | 0,801801802 | 0,516548 | 0,518950437 |
| http://whydoeseverythingsuck.com/2008/02/death-of-relational-database.html | 1 | 90 | 0,387931034 | 0,803571429 | 0,501409 | 0,523255814 |
| http://www.miswebdesign.com/resources/articles/wrox-beginning-php-4-chapter-3-1.html | 1 | 91 | 0,392241379 | 0,805309735 | 0,498591 | 0,527536232 |
| http://troels.arvin.dk/db/rdbms/links/ | 1 | 92 | 0,396551724 | 0,807017544 | 0,483452 | 0,531791908 |
| http://cbbrowne.com/info/rdbms.html | 1 | 93 | 0,400862069 | 0,808695652 | 0,478972 | 0,536023055 |
| http://en.wikipedia.org/wiki/Relational_database | 1 | 94 | 0,405172414 | 0,810344828 | 0,464561 | 0,540229885 |
| http://hsqldb.org/ | 1 | 95 | 0,409482759 | 0,811965812 | 0,451804 | 0,544412607 |
| http://ec.europa.eu/ipg/standards/databases/standard_rdbms_en.htm | 1 | 96 | 0,413793103 | 0,813559322 | 0,443488 | 0,548571429 |
| http://jena.sourceforge.net/DB/index.html | 1 | 97 | 0,418103448 | 0,81512605 | 0,435606 | 0,552706553 |
| http://www.cetus-links.org/oo_db_systems_2.html | 1 | 98 | 0,422413793 | 0,816666667 | 0,430948 | 0,556818182 |
| http://www.databasedev.co.uk/database_normalization_process.html | 1 | 99 | 0,426724138 | 0,818181818 | 0,416003 | 0,560906516 |
| http://www.databasedev.co.uk/data_models.html | 1 | 100 | 0,431034483 | 0,819672131 | 0,408583 | 0,564971751 |
| http://www.relationalwizards.com/ | 0 | 100 | 0,431034483 | 0,81300813 | 0,407107 | 0,563380282 |
| http://www.geekgirls.com/databases_from_scratch_3.htm | 1 | 101 | 0,435344828 | 0,814516129 | 0,405637 | 0,56741573 |
| http://www.hackszine.com/blog/archive/2008/04/relational_database_using_jque.html | 1 | 102 | 0,439655172 | 0,816 | 0,402656 | 0,571428571 |
| http://ocw.mit.edu/NR/rdonlyres/Urban-Studies-and-Planning/11-208Introduction-to-Computers-in-Public-Management-IIJanuary--IAP-2002/64B3A7CB-FA1F-4749-869C-A5D96ABCBE50/0/lect52.pdf | 1 | 103 | 0,443965517 | 0,817460317 | 0,401071 | 0,575418994 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.contentdsi.com/content_management.htm | 0 | 103 | 0,443965517 | 0,811023622 | 0,397998 | 0,573816156 |
| http://www.snee.com/bobdc.blog/2008/07/devx-article-relational-databa.html | 1 | 104 | 0,448275862 | 0,8125 | 0,394965 | 0,577777778 |
| http://www.deskspace.com/ | 0 | 104 | 0,448275862 | 0,80620155 | 0,39333 | 0,576177285 |
| http://www.boinc-wiki.info/Relational_Data_Base_Management_System | 1 | 105 | 0,452586207 | 0,807692308 | 0,391687 | 0,580110497 |
| http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_74.pdf | 1 | 106 | 0,456896552 | 0,809160305 | 0,390089 | 0,584022039 |
| http://www.dhdursoassociates.com/ | 0 | 106 | 0,456896552 | 0,803030303 | 0,388484 | 0,582417582 |
| http://www.chass.utoronto.ca/epc/chwp/CHC2007/Liu_Smith/Liu_Smith.htm | 1 | 107 | 0,461206897 | 0,804511278 | 0,386897 | 0,58630137 |
| http://invasions.si.edu/nbic/search.html | 1 | 108 | 0,465517241 | 0,805970149 | 0,385217 | 0,590163934 |
| http://www.euclideanspace.com/software/information/relational/index.htm | 1 | 109 | 0,469827586 | 0,807407407 | 0,383478 | 0,59400545 |
| http://www.cosis.net/abstracts/9IKC/00196/9IKC-A-00196-1.pdf | 0 | 109 | 0,469827586 | 0,801470588 | 0,381754 | 0,592391304 |
| http://plc.inf.elte.hu/erlang/pub/refac_db_kent.ppt | 0 | 109 | 0,469827586 | 0,795620438 | 0,380034 | 0,590785908 |
| http://www.agiledata.org/essays/relationalDatabases.html | 1 | 110 | 0,474137931 | 0,797101449 | 0,378332 | 0,594594595 |
| http://www.chemcomp.com/journal/reldb.htm | 1 | 111 | 0,478448276 | 0,798561151 | 0,376601 | 0,598382749 |
| http://www.dspace.cam.ac.uk/handle/1810/14718 | 1 | 112 | 0,482758621 | 0,8 | 0,372854 | 0,602150538 |
| http://modperlbook.org/html/Chapter-20-Relational-Databases-and-mod_perl.html | 1 | 113 | 0,487068966 | 0,80141844 | 0,370975 | 0,605898123 |
| http://www.netl.doe.gov/publications/proceedings/01/carbon_seq/1a4.pdf | 0 | 113 | 0,487068966 | 0,795774648 | 0,369116 | 0,604278075 |
| http://lists.debian.org/debian-devel/2007/06/msg00106.html | 1 | 114 | 0,49137931 | 0,797202797 | 0,367251 | 0,608 |
| http://lists.w3.org/Archives/Public/public-rdf-dawg/2004JanMar/0208.html | 0 | 114 | 0,49137931 | 0,791666667 | 0,365252 | 0,606382979 |
| http://www.remcomp.fr/asmanet/asmapro/asmawork.htm | 0 | 114 | 0,49137931 | 0,786206897 | 0,363177 | 0,604774536 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.firstsql.com/ireldb.htm | 1 | 115 | 0,495689655 | 0,787671233 | 0,361112 | 0,608465608 |
| http://www.emis.de/journals/AUA/acta3/modelareadb.pdf | 1 | 116 | 0,5 | 0,789115646 | 0,359017 | 0,612137203 |
| http://www.almaden.ibm.com/cs/projects/iis/hdb/Publications/papers/sigmod98_dbi.pdf | 0 | 116 | 0,5 | 0,783783784 | 0,356963 | 0,610526316 |
| http://www.funpecrp.com.br/gmr/year2007/vol4-6/xm0012_abstract.html | 1 | 117 | 0,504310345 | 0,785234899 | 0,354805 | 0,614173228 |
| http://people.csail.mit.edu/jaffer/slib_6 | 1 | 118 | 0,50862069 | 0,786666667 | 0,352508 | 0,617801047 |
| http://www.haz-map.com/ | 1 | 119 | 0,512931034 | 0,78807947 | 0,350168 | 0,621409922 |
| http://www.ampl.com/NEW/tables.html | 0 | 119 | 0,512931034 | 0,782894737 | 0,347806 | 0,619791667 |
| http://cdlr.strath.ac.uk/pubs/dunsireg/alm03main.pps | 0 | 119 | 0,512931034 | 0,777777778 | 0,345455 | 0,618181818 |
| http://www.jot.fm/issues/issue_2003_09/article1.pdf | 1 | 120 | 0,517241379 | 0,779220779 | 0,343097 | 0,621761658 |
| http://www.hpss-collaboration.org/hpss/about/BoomerRDBMSHSM.pdf | 1 | 121 | 0,521551724 | 0,780645161 | 0,33783 | 0,625322997 |
| http://www.informatik.hu-berlin.de/Forschung_Lehre/wbi/publications/2005/dils05_ontologies.pdf | 1 | 122 | 0,525862069 | 0,782051282 | 0,335103 | 0,628865979 |
| http://www.agentjim.com/MVP/Excel/RelationalOffice.htm | 0 | 122 | 0,525862069 | 0,777070064 | 0,332787 | 0,627249357 |
| http://www.ebmt.org/4registry/Registry_docs/ProMISe%20Docs THE%20EBMT%20RELATIONAL%20DATABASE.pdf | 0 | 122 | 0,525862069 | 0,772151899 | 0,331364 | 0,625641026 |
| http://hms.liacs.nl/ilp.html | 0 | 122 | 0,525862069 | 0,767295597 | 0,329964 | 0,624040921 |
| http://www.fetac.ie/modules/C30147.PDF | 0 | 122 | 0,525862069 | 0,7625 | 0,327155 | 0,62244898 |
| http://gadfly.sourceforge.net/ | 1 | 123 | 0,530172414 | 0,763975155 | 0,325728 | 0,625954198 |
| http://sql.z3950.org/docs/zSQLgate.html | 0 | 123 | 0,530172414 | 0,759259259 | 0,324323 | 0,624365482 |
| http://www.mail-archive.com/accessvbacentral@yahoogroups.com/msg00567.html | 1 | 124 | 0,534482759 | 0,760736196 | 0,322922 | 0,627848101 |
| http://www.python.org/workshops/1997-10/proceedings/shprentz.html | 1 | 125 | 0,538793103 | 0,762195122 | 0,32016 | 0,631313131 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://leap.sourceforge.net/ | 1 | 126 | 0,543103448 | 0,763636364 | 0,318781 | 0,634760705 |
| http://www.cephbase.utmb.edu/ | 0 | 126 | 0,543103448 | 0,759036145 | 0,317403 | 0,633165829 |
| http://hal.archives-ouvertes.fr/docs/00/16/81/52/PDF/Hrivnac.pdf | 1 | 127 | 0,547413793 | 0,760479042 | 0,316042 | 0,636591479 |
| http://ferret.pmel.noaa.gov/HOMEPAGE/LAS/FAQ/relational_database_access.htm | 0 | 127 | 0,547413793 | 0,755952381 | 0,314683 | 0,635 |
| http://pnclink.org:8080/pnc2006/Abstract/Exemplary%20Atlas%20--%20Jiang%20Wu.pdf | 1 | 128 | 0,551724138 | 0,75739645 | 0,313341 | 0,63840399 |
| http://codex.cs.yale.edu/avi/db-book/selected-exer-dir/7-web.pdf | 1 | 129 | 0,556034483 | 0,758823529 | 0,311932 | 0,641791045 |
| http://www.cs.virginia.edu/papers/ismb02_sql.pdf | 1 | 130 | 0,560344828 | 0,760233918 | 0,310531 | 0,64516129 |
| http://www.hitsw.com/products_services/whitepapers/integrating_xml_rdb/ | 0 | 130 | 0,560344828 | 0,755813953 | 0,309128 | 0,643564356 |
| http://www.emis.de/journals/NSJOM/Papers/26_2/NSJOM_26_2_049_073.pdf | 1 | 131 | 0,564655172 | 0,757225434 | 0,306354 | 0,64691358 |
| http://www.idealliance.org/papers/extreme/proceedings/xslfo-pdf/2007/Ramalho01/EML2007Ramalho01.pdf | 1 | 132 | 0,568965517 | 0,75862069 | 0,303538 | 0,650246305 |
| http://www.research.ibm.com/journal/sj/424/telford.html | 1 | 133 | 0,573275862 | 0,76 | 0,302129 | 0,653562654 |
| http://www.duoconsulting.com/downloads/ContentManagement.pdf | 1 | 134 | 0,577586207 | 0,761363636 | 0,299315 | 0,656862745 |
| http://www.objectarchitects.de/ObjectArchitects/orpatterns/ | 1 | 135 | 0,581896552 | 0,762711864 | 0,297902 | 0,660146699 |
| http://eagle.cs.uiuc.edu/pubs/2005/ranksqldemo-vldb05-laci-jun05.pdf | 1 | 136 | 0,586206897 | 0,764044944 | 0,294931 | 0,663414634 |
| http://www.snee.com/xml/xml2006/owlrdbms.html | 1 | 137 | 0,590517241 | 0,765363128 | 0,29345 | 0,666666667 |
| http://www.jcc.com/DescImplementingDBUsingSQL.htm | 1 | 138 | 0,594827586 | 0,766666667 | 0,291964 | 0,669902913 |
| http://www.dds-lite.com/ | 0 | 138 | 0,594827586 | 0,762430939 | 0,290489 | 0,668280872 |
| http://articles.techrepublic.com.com/5100-22_11-5075453.html | 1 | 139 | 0,599137931 | 0,763736264 | 0,287418 | 0,671497585 |
| http://www.malacolog.org/ | 1 | 140 | 0,603448276 | 0,765027322 | 0,285863 | 0,674698795 |
| http://ajp.amjpathol.org/cgi/content/abstract/159/3/837 | 0 | 140 | 0,603448276 | 0,760869565 | 0,284313 | 0,673076923 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.ibm.com/developerworks/edu/dm-dw-dm-0611nelke-i.html | 0 | 140 | 0,603448276 | 0,756756757 | 0,282757 | 0,67146283 |
| http://jelle.druyts.net/2004/04/14/InheritanceModellingInARelational Database.aspx | 0 | 140 | 0,603448276 | 0,752688172 | 0,281178 | 0,669856459 |
| http://www.intersystems.com/cache/whitepapers/hybrid.html | 1 | 141 | 0,607758621 | 0,754010695 | 0,279511 | 0,673031026 |
| http://www.japmaonline.org/cgi/content/abstract/86/2/74 | 0 | 141 | 0,607758621 | 0,75 | 0,276162 | 0,671428571 |
| http://www.interpares.org/documents/interpares_cs_01_overview.pdf | 1 | 142 | 0,612068966 | 0,751322751 | 0,274488 | 0,674584323 |
| http://vsis-www.informatik.uni-hamburg.de/publications/view.php/164 | 1 | 143 | 0,61637931 | 0,752631579 | 0,272798 | 0,677725118 |
| http://macs.about.com/od/glossaryqt/g/relational.htm | 1 | 144 | 0,620689655 | 0,753926702 | 0,271032 | 0,680851064 |
| http://msdn.microsoft.com/en-us/library/bb245675.aspx | 1 | 145 | 0,625 | 0,755208333 | 0,269232 | 0,683962264 |
| http://www.omegahat.org/RSPostgres/Scenarios.pdf | 1 | 146 | 0,629310345 | 0,756476684 | 0,267429 | 0,687058824 |
| http://www.rpbourret.com/xml/DataTransfer.htm | 0 | 146 | 0,629310345 | 0,75257732 | 0,265606 | 0,685446009 |
| http://www.freepatentsonline.com/5905985.html | 0 | 146 | 0,629310345 | 0,748717949 | 0,263786 | 0,683840749 |
| http://www.mysql.com/ | 1 | 147 | 0,63362069 | 0,75 | 0,259916 | 0,686915888 |
| http://www.biodatamining.org/content/1/1/7 | 1 | 148 | 0,637931034 | 0,751269036 | 0,257916 | 0,68997669 |
| http://www.dbforums.com/ | 1 | 149 | 0,642241379 | 0,752525253 | 0,253879 | 0,693023256 |
| http://dictionary.reference.com/search?q=relational+database&r=66 | 1 | 150 | 0,646551724 | 0,753768844 | 0,251837 | 0,696055684 |
| http://www.zoominfo.com/Industries/software-mfg/software-development-design/relational-database-management-system.htm | 1 | 151 | 0,650862069 | 0,755 | 0,249645 | 0,699074074 |
| http://dictionary.reference.com/search?q=relational+database | 1 | 152 | 0,655172414 | 0,756218905 | 0,247439 | 0,702078522 |
| http://cll.emory.edu/it/classes.cfm?cla=-881825843&pt=3 | 1 | 153 | 0,659482759 | 0,757425743 | 0,245194 | 0,705069124 |
| http://www.mysql.fr/news-and-events/on-demand-webinars/display-od-216.html | 1 | 154 | 0,663793103 | 0,75862069 | 0,242938 | 0,708045977 |
| http://support.microsoft.com/kb/234208 | 1 | 155 | 0,668103448 | 0,759803922 | 0,240644 | 0,711009174 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://nar.oxfordjournals.org/cgi/content/full/26/1/335 | 0 | 155 | 0,668103448 | 0,756097561 | 0,23824 | 0,709382151 |
| http://www.membranetransport.org/ | 0 | 155 | 0,668103448 | 0,752427184 | 0,235734 | 0,707762557 |
| http://xml-and-relational-database.nuclearscripts.com/ | 1 | 156 | 0,672413793 | 0,753623188 | 0,233205 | 0,71070615 |
| http://adsabs.harvard.edu/abs/1993PASP..105.1482S | 0 | 156 | 0,672413793 | 0,75 | 0,228023 | 0,709090909 |
| http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243450 | 1 | 157 | 0,676724138 | 0,751196172 | 0,225334 | 0,712018141 |
| http://ieeexplore.ieee.org/iel2/3041/8637/00380352.pdf?arnumber=380352 | 1 | 158 | 0,681034483 | 0,752380952 | 0,222493 | 0,714932127 |
| http://peds.oxfordjournals.org/cgi/content/abstract/2/6/431 | 1 | 159 | 0,685344828 | 0,753554502 | 0,219597 | 0,717832957 |
| http://oreilly.com/pub/a/oreilly/frank/dbdesign_0701.html | 1 | 160 | 0,689655172 | 0,754716981 | 0,213709 | 0,720720721 |
| http://firebird.sourceforge.net/index.php | 1 | 161 | 0,693965517 | 0,755868545 | 0,210727 | 0,723595506 |
| http://www.boingboing.net/2009/01/21/keeping-up-with-lost.html | 1 | 162 | 0,698275862 | 0,757009346 | 0,207542 | 0,726457399 |
| http://www.actapress.com/PDFViewer.aspx?paperId=14993 | 1 | 163 | 0,702586207 | 0,758139535 | 0,204288 | 0,729306488 |
| http://www.aemj.org/cgi/content/abstract/7/5/472-a?ck=nck | 0 | 163 | 0,702586207 | 0,75462963 | 0,201008 | 0,727678571 |
| http://wdvl.com/Authoring/DB/Intro/relational_databases.html | 1 | 164 | 0,706896552 | 0,755760369 | 0,1977 | 0,730512249 |
| http://erx.sagepub.com/cgi/content/abstract/25/5/533 | 0 | 164 | 0,706896552 | 0,752293578 | 0,194339 | 0,728888889 |
| http://www.peopleware.net/0177/index.cfm?eventDisp=CDMAA | 0 | 164 | 0,706896552 | 0,748858447 | 0,190903 | 0,727272727 |
| http://www.oracle.com/database/berkeley-db/index.html | 1 | 165 | 0,711206897 | 0,75 | 0,187265 | 0,730088496 |
| http://www.textbooksrus.com/search/BookDetail/?isbn=0201752840&kbid=1067 | 0 | 165 | 0,711206897 | 0,746606335 | 0,183634 | 0,728476821 |
| http://intl.ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=20424&arnumber=943703&count=113&index=92 | 1 | 166 | 0,715517241 | 0,747747748 | 0,179962 | 0,731277533 |
| http://www.openoffice.org/servlets/ReadMsg?list=discuss&msgNo=39173 | 0 | 166 | 0,715517241 | 0,744394619 | 0,176252 | 0,72967033 |
| http://technet.microsoft.com/en-us/library/aa226072(SQL.80).aspx | 1 | 167 | 0,719827586 | 0,745535714 | 0,172548 | 0,73245614 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.itia.ntua.gr/en/projinfo/50/ | 1 | 168 | 0,724137931 | 0,746666667 | 0,168575 | 0,735229759 |
| http://relational-database-software.qarchive.org/ | 1 | 169 | 0,728448276 | 0,747787611 | 0,164537 | 0,737991266 |
| http://wiki.gxtechnical.com/commwiki/servlet/hwiki?Relational+Database+Theory | 1 | 170 | 0,732758621 | 0,748898678 | 0,160551 | 0,740740741 |
| http://education-portal.com/relational_database_fundamentals_online_course.html | 1 | 171 | 0,737068966 | 0,75 | 0,156484 | 0,743478261 |
| http://www.bibl.liu.se/liupubl/disp/disp96/tek452s.htm | 1 | 172 | 0,74137931 | 0,751091703 | 0,152436 | 0,746203905 |
| http://msdis.missouri.edu/presentations/gis_advanced/pdf/Relational.pdf | 1 | 173 | 0,745689655 | 0,752173913 | 0,148213 | 0,748917749 |
| http://havemacwillblog.com/2008/11/10/6-reasons-why-relational-database-will-be-superseded/ | 1 | 174 | 0,75 | 0,753246753 | 0,143827 | 0,75161987 |
| http://www.gsd.harvard.edu/gis/manual/relational/index.htm | 1 | 175 | 0,754310345 | 0,754310345 | 0,13947 | 0,754310345 |
| http://www.nationmultimedia.com/worldhotnews/30091764/Oracle-is-the-number-1-Relational-Database-in-Thailand | 1 | 176 | 0,75862069 | 0,755364807 | 0,135111 | 0,756989247 |
| http://www.basis-wien.at/index.php?id=72&L=2 | 0 | 176 | 0,75862069 | 0,752136752 | 0,130663 | 0,755364807 |
| http://hollywood.mit.edu/ | 0 | 176 | 0,75862069 | 0,74893617 | 0,127009 | 0,753747323 |
| http://www.utexas.edu/cc/database/datamodeling/ | 1 | 177 | 0,762931034 | 0,75 | 0,126663 | 0,756410256 |
| http://research.amnh.org/amcc/Freezerworks.html | 1 | 178 | 0,767241379 | 0,751054852 | 0,126324 | 0,759061834 |
| http://www.auditmypc.com/acronym/RDBMS.asp | 1 | 179 | 0,771551724 | 0,75210084 | 0,125989 | 0,761702128 |
| http://www.businessdictionary.com/definition/relational-database.html | 1 | 180 | 0,775862069 | 0,753138075 | 0,125652 | 0,76433121 |
| http://compbio.soe.ucsc.edu/rdb/index.html | 1 | 181 | 0,780172414 | 0,754166667 | 0,125311 | 0,766949153 |
| http://www.linux-mag.com/id/2093 | 1 | 182 | 0,784482759 | 0,755186722 | 0,124977 | 0,769556025 |
| http://www.surfermall.com/relational/lesson_1.htm | 1 | 183 | 0,788793103 | 0,756198347 | 0,124645 | 0,772151899 |
| http://awtrey.com/tutorials/dbeweb/database.php | 1 | 184 | 0,793103448 | 0,757201646 | 0,12432 | 0,774736842 |
| http://www.catalhoyuk.com/database/catal/Browse.asp | 0 | 184 | 0,793103448 | 0,754098361 | 0,123995 | 0,773109244 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.postgresql.org/ | 1 | 185 | 0,797413793 | 0,755102041 | 0,123667 | 0,775681342 |
| http://www.informationweek.com/news/management/showArticle.jhtml?articleID=159401656 | 1 | 186 | 0,801724138 | 0,756097561 | 0,123342 | 0,778242678 |
| http://www.vertica.com/relational-database-management-system | 0 | 186 | 0,801724138 | 0,753036437 | 0,123022 | 0,776617954 |
| http://dotnet.sys-con.com/node/175864 | 0 | 186 | 0,801724138 | 0,75 | 0,122704 | 0,775 |
| http://www.techbriefs.com/component/content/68?task=view | 0 | 186 | 0,801724138 | 0,746987952 | 0,122392 | 773388773 |
| http://www.xml.com/pub/a/2007/07/12/xquery-and-data-abstraction.html | 1 | 187 | 0,806034483 | 0,748 | 0,122076 | 0,77593361 |
| http://foldoc.org/?relational+database | 0 | 187 | 0,806034483 | 0,74501992 | 0,121764 | 0,774327122 |
| http://www.15seconds.com/Issue/020522.htm | 1 | 188 | 0,810344828 | 0,746031746 | 0,121453 | 0,776859504 |
| http://www.warthman.com/projects-tymshare-b.htm | 0 | 188 | 0,810344828 | 0,743083004 | 0,120844 | 0,775257732 |
| http://www.zope.org/Documentation/Books/ZopeBook/2_6Edition/RelationalDatabases.stx | 1 | 189 | 0,814655172 | 0,744094488 | 0,120544 | 0,777777778 |
| http://www.onesmartclick.com/engineering/relational-databases.html | 1 | 190 | 0,818965517 | 0,745098039 | 0,120243 | 0,780287474 |
| http://www.blackwell-synergy.com/doi/abs/10.1111/j.1540-8159.1996.tb03421.x | 1 | 191 | 0,823275862 | 0,74609375 | 0,119947 | 0,782786885 |
| http://www.w3.org/XML/RDB.html | 1 | 192 | 0,827586207 | 0,747081712 | 0,119651 | 0,785276074 |
| http://lips.informatik.uni-leipzig.de/pub/1998-8 | 0 | 192 | 0,827586207 | 0,744186047 | 0,119361 | 0,783673469 |
| http://www.leavcom.com/db_08_00.htm | 1 | 193 | 0,831896552 | 0,745173745 | 0,119071 | 0,786150713 |
| http://www.defmacro.org/ramblings/relational.html | 1 | 194 | 0,836206897 | 0,746153846 | 0,118786 | 0,788617886 |
| http://www.isprs.org/congresses/beijing2008/proceedings/2_pdf/1_WG-II-1/14.pdf | 1 | 195 | 0,840517241 | 0,747126437 | 0,118501 | 0,791075051 |
| http://www.agiledata.org/essays/mappingObjects.html | 1 | 196 | 0,844827586 | 0,748091603 | 0,118221 | 0,793522267 |
| http://duro.sourceforge.net/ | 1 | 197 | 0,849137931 | 0,74904943 | 0,117941 | 0,795959596 |
| http://www.funpecrp.com.br/gmr/year2007/vol4-6//pdf/xm0012.pdf | 1 | 198 | 0,853448276 | 0,75 | 0,117666 | 0,798387097 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.athro.com/general/Phyloinformatics_7_85x11.pdf | 0 | 198 | 0,853448276 | 0,747169811 | 0,117392 | 0,796780684 |
| http://www.codata.org/08conf/abstracts/ChangjunHu-A%20Visual%20Tool%20for%20Building%20MatML%20Data%20from%20Material%20Science%20Relational%20Database.htm | 0 | 198 | 0,853448276 | 0,744360902 | 0,117124 | 0,795180723 |
| http://www.eclipse.org/webtools/community/tutorials/RDB Tutorial/RDBTutorial.html | 1 | 199 | 0,857758621 | 0,745318352 | 0,116857 | 0,79759519 |
| http://pooteeweet.org/files/phpconf05/relational_database_ starter_day.pdf | 1 | 200 | 0,862068966 | 0,746268657 | 0,116593 | 0,8 |
| http://cidoc.ics.forth.gr/docs/Implementing_the_CIDOC_CRM.rtf | 0 | 200 | 0,862068966 | 0,743494424 | 0,116329 | 0,798403194 |
| http://gamma.cs.unc.edu/DB/ | 1 | 201 | 0,86637931 | 0,744444444 | 0,11607 | 0,800796813 |
| http://www.ietf.org/proceedings/94mar/mgt/rdbmsmib.html | 1 | 202 | 0,870689655 | 0,745387454 | 0,115816 | 0,803180915 |
| http://www.serc.iisc.ernet.in/ComputingFacilities/systems/cluster /vac-7.0/html/glossary/czgr.htm | 0 | 202 | 0,870689655 | 0,742647059 | 0,115566 | 0,801587302 |
| http://www.bmrb.wisc.edu/search/rela_database.html | 0 | 202 | 0,870689655 | 0,73992674 | 0,115316 | 0,8 |
| http://www.cemml.colostate.edu/files/un5.pdf | 1 | 203 | 0,875 | 0,740875912 | 0,115069 | 0,802371542 |
| http://2006.xmlconference.org/programme/presentations/188.html | 0 | 203 | 0,875 | 0,738181818 | 0,114822 | 0,800788955 |
| http://www.sqlconsole.com/ | 1 | 204 | 0,879310345 | 0,739130435 | 0,114118 | 0,803149606 |
| http://www.yolinux.com/HOWTO/PostgreSQL-HOWTO.html | 1 | 205 | 0,88362069 | 0,740072202 | 0,113657 | 0,805500982 |
| http://www.mindmodel.com/ | 0 | 205 | 0,88362069 | 0,737410072 | 0,113432 | 0,803921569 |
| http://www.gridpp.ac.uk/papers/GGF3Rome2001.pdf | 0 | 205 | 0,88362069 | 0,734767025 | 0,113217 | 0,802348337 |
| http://db.uwaterloo.ca/OED/trdbms.html | 0 | 205 | 0,88362069 | 0,732142857 | 0,113002 | 0,80078125 |
| http://www.oracle.com/database/docs/Berkeley-DB-v-Relational.pdf | 1 | 206 | 0,887931034 | 0,733096085 | 0,112789 | 0,803118908 |
| http://www.altova.com/press/2003-01-13_hitsw.pdf | 0 | 206 | 0,887931034 | 0,730496454 | 0,112577 | 0,80155642 |
| http://d3dnff.gat.com/D3DRDB/ | 0 | 206 | 0,887931034 | 0,727915194 | 0,112367 | 0,8 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://www.pinnaclekenya.com/ | 0 | 206 | 0,887931034 | 0,725352113 | 0,112169 | 0,798449612 |
| http://codingforums.com/showthread.php?t=44774 | 0 | 206 | 0,887931034 | 0,722807018 | 0,111975 | 0,796905222 |
| http://gd.tuwien.ac.at:8050/H/1/ | 1 | 207 | 0,892241379 | 0,723776224 | 0,111781 | 0,799227799 |
| http://www.stylusstudio.com/xml_database.html | 0 | 207 | 0,892241379 | 0,721254355 | 0,111589 | 0,797687861 |
| http://gateway.nlm.nih.gov/MeetingAbstracts/102211909.html | 0 | 207 | 0,892241379 | 0,71875 | 0,111396 | 0,796153846 |
| http://www3.open.ac.uk/courses/bin/p12.dll?C01M359 | 0 | 207 | 0,892241379 | 0,716262976 | 0,111214 | 0,79462572 |
| http://www.codeproject.com/KB/showcase/object_relational_ mapping.aspx | 1 | 208 | 0,896551724 | 0,717241379 | 0,111041 | 0,796934866 |
| http://www.ohloh.net/tags/database | 0 | 208 | 0,896551724 | 0,714776632 | 0,110869 | 0,79541109 |
| http://ilpubs.stanford.edu:8090/4/ | 1 | 209 | 0,900862069 | 0,715753425 | 0,110697 | 0,797709924 |
| http://www.auditmypc.com/acronym/ORDBMS.asp | 0 | 209 | 0,900862069 | 0,71331058 | 0,110526 | 0,796190476 |
| http://www.eecs.berkeley.edu/Pubs/TechRpts/1978/12384.html | 0 | 209 | 0,900862069 | 0,710884354 | 0,110357 | 0,794676806 |
| http://www.arcchip.cz/w05/w05_buitleir.pdf | 1 | 210 | 0,905172414 | 0,711864407 | 0,110208 | 0,796963947 |
| http://www.developers.net/enterprisedbshowcase/view/1323 | 0 | 210 | 0,905172414 | 0,709459459 | 0,110058 | 0,795454545 |
| http://www.questia.com/PM.qst?a=o&se=gglsc&d=5001678243 | 0 | 210 | 0,905172414 | 0,707070707 | 0,10991 | 0,793950851 |
| http://llc.oxfordjournals.org/cgi/content/abstract/2/2/89 | 0 | 210 | 0,905172414 | 0,704697987 | 0,109762 | 0,79245283 |
| http://www.codeproject.com/KB/database/introtomatisse_part2.aspx | 1 | 211 | 0,909482759 | 0,705685619 | 0,109614 | 0,79472693 |
| http://www.adobe.com/devnet/coldfusion/articles/display_dyn_ data_02.html | 1 | 212 | 0,913793103 | 0,706666667 | 0,109481 | 0,796992481 |
| http://uttc.umn.edu/training/courses/description?designator=DB101 | 0 | 212 | 0,913793103 | 0,704318937 | 0,109358 | 0,795497186 |
| http://stats.oecd.org/glossary/detail.asp?ID=4520 | 1 | 213 | 0,918103448 | 0,705298013 | 0,109234 | 0,797752809 |
| http://libra.msra.cn/papercited.aspx?id=352778 | 1 | 214 | 0,922413793 | 0,706270627 | 0,108988 | 0,8 |
| http://www.springerlink.com/index/7442wt8574x44766.pdf | 0 | 214 | 0,922413793 | 0,703947368 | 0,108871 | 0,798507463 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://research.microsoft.com/apps/pubs/default.aspx?id=64571 | 1 | 215 | 0,926724138 | 0,704918033 | 0,108774 | 0,800744879 |
| http://www.iop.org/EJ/abstract/1742-6596/119/4/042009 | 1 | 216 | 0,931034483 | 0,705882353 | 0,108678 | 0,802973978 |
| http://relationaldatabasesoftware.surfpack.com/ | 1 | 217 | 0,935344828 | 0,706840391 | 0,108581 | 0,805194805 |
| http://www.builderau.com.au/program/xml/soa/Transfer-and-store-data-from-an-XML-document-in-a-relational-database/0,339028469,339273299,00.htm | 1 | 218 | 0,939655172 | 0,707792208 | 0,108486 | 0,807407407 |
| http://nar.oxfordjournals.org/cgi/content/abstract/31/1/202 | 1 | 219 | 0,943965517 | 0,708737864 | 0,108389 | 0,80961183 |
| http://www.midcarb.org/Documents/GSA-Nov-2000.shtml | 0 | 219 | 0,943965517 | 0,706451613 | 0,108317 | 0,808118081 |
| http://www.100best-web-hosting.com/glossary674.html | 1 | 220 | 0,948275862 | 0,707395498 | 0,10825 | 0,810313076 |
| http://computerprogramming.suite101.com/article.cfm/sql_server_training_first_steps | 1 | 221 | 0,952586207 | 0,708333333 | 0,108183 | 0,8125 |
| http://en.wikibooks.org/wiki/Category:Relational_Database_Design | 1 | 222 | 0,956896552 | 0,709265176 | 0,108116 | 0,814678899 |
| http://pd.acm.org/book_detail.cfm?isbn=0321305965 | 1 | 223 | 0,961206897 | 0,710191083 | 0,10805 | 0,816849817 |
| http://comjnl.oxfordjournals.org/cgi/content/abstract/39/2/124?ck=nck | 1 | 224 | 0,965517241 | 0,711111111 | 0,107996 | 0,819012797 |
| http://gateway.nlm.nih.gov/MeetingAbstracts/102276723.html | 1 | 225 | 0,969827586 | 0,712025316 | 0,107961 | 0,821167883 |
| http://scripts.iucr.org/cgi-bin/paper?S0108768102002458 | 1 | 226 | 0,974137931 | 0,712933754 | 0,107926 | 0,823315118 |
| http://encyclopedia.kids.net.au/page/re/Relational_database | 1 | 227 | 0,978448276 | 0,713836478 | 0,107891 | 0,825454545 |
| http://www.lavoisier.fr/notice/gbBCO3AXSM3OP2RO.html | 0 | 227 | 0,978448276 | 0,711598746 | 0,107857 | 0,823956443 |
| http://www.empress.com/ | 0 | 227 | 0,978448276 | 0,709375 | 0,107822 | 0,822463768 |
| http://briefingsdirectblog.blogspot.com/2007/10/oracle-users-enjoy-open-source-benefits.html | 0 | 227 | 0,978448276 | 0,707165109 | 0 | 0,820976492 |
| http://infolab.stanford.edu/~ullman/fcdb/spr99/lec4.ps | 1 | 228 | 0,982758621 | 0,708074534 | 0 | 0,823104693 |
| http://news.cnet.com/Rethinking-the-relational-database/2010-1015_3-5715457.html | 0 | 228 | 0,982758621 | 0,705882353 | 0 | 0,821621622 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Google | FMGoogle |
|---|---|---|---|---|---|---|
| http://pages.cs.wisc.edu/~cs764-1/selinger.pdf | 1 | 229 | 0,987068966 | 0,706790123 | 0 | 0,823741007 |
| http://publish.uwo.ca/~craven/558/558red.htm | 1 | 230 | 0,99137931 | 0,707692308 | 0 | 0,825852783 |
| http://user.it.uu.se/~udbl/Theses/QinZhangMSc.pdf | 0 | 230 | 0,99137931 | 0,705521472 | 0 | 0,82437276 |
| http://www.cs.mcgill.ca/~kemme/papers/pakdd06.pdf | 0 | 230 | 0,99137931 | 0,703363914 | 0 | 0,822898032 |
| http://www.cs.ucla.edu/~zaniolo/papers/widm06.pdf | 0 | 230 | 0,99137931 | 0,701219512 | 0 | 0,821428571 |
| http://www.cs.uwaterloo.ca/~ashraf/pubs/cikm08recman.pdf | 0 | 230 | 0,99137931 | 0,699088146 | 0 | 0,819964349 |
| http://www.cse.lehigh.edu/~heflin/pubs/psss03-poster.pdf | 0 | 230 | 0,99137931 | 0,696969697 | 0 | 0,818505338 |
| http://www.dcs.bbk.ac.uk/~mARK/relational_book.html | 1 | 231 | 0,995689655 | 0,697885196 | 0 | 0,820603908 |
| http://www.eas.asu.edu/~winrdbi/author/ | 0 | 231 | 0,995689655 | 0,695783133 | 0 | 0,819148936 |
| http://www.informatik.uni-trier.de/~ley/db/conf/icde/Stonebraker86.html | 0 | 231 | 0,995689655 | 0,693693694 | 0 | 0,817699115 |
| http://www.informatik.uni-trier.de/~ley/db/conf/sigmod/SelingerACLP79.html | 0 | 231 | 0,995689655 | 0,691616766 | 0 | 0,816254417 |
| http://www.islandnet.com/~tmc/html/articles/orareln.htm | 0 | 231 | 0,995689655 | 0,689552239 | 0 | 0,814814815 |
| http://www.ispras.ru/~knizhnik/gigabase.html | 0 | 231 | 0,995689655 | 0,6875 | 0 | 0,813380282 |
| http://www.laas.fr/~esorics/notices/Yazdanian90.html | 0 | 231 | 0,995689655 | 0,685459941 | 0 | 0,811950791 |
| http://www.star.le.ac.uk/~cgp/adass2002.html | 0 | 231 | 0,995689655 | 0,683431953 | 0 | 0,810526316 |
| http://www.thunderstone.com/site/texisman/relational_database_background.html | 1 | 232 | 1 | 0,684365782 | 0 | 0,810526316 |
| http://www-db.in.tum.de/~teubnerj/publications/diss.pdf | 0 | 232 | 1 | 0,682352941 | 0 | 0,810526316 |

# Anexo V – Cálculo Precisão e Cobertura pelos Valores de Ordenação da Avaliação de Qualidade

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://support.microsoft.com/kb/234208 | 1 | 1 | 0,004545455 | 1 | 0,872933 | 0,009049774 |
| http://www.hackszine.com/blog/archive/2008/04/relational_ database_using_jque.html | 1 | 2 | 0,009090909 | 1 | 0,872933 | 0,018018018 |
| http://hollywood.mit.edu/ | 0 | 2 | 0,009090909 | 0,666666667 | 0,871006 | 0,01793722 |
| http://gateway.nlm.nih.gov/MeetingAbstracts/102211909.html | 0 | 2 | 0,009090909 | 0,5 | 0,871005 | 0,017857143 |
| http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/ Vol-401/iswc2008pd_submission_74.pdf | 1 | 3 | 0,013636364 | 0,6 | 0,871004 | 0,026666667 |
| http://hsqldb.org/ | 1 | 4 | 0,018181818 | 0,666666667 | 0,870998 | 0,03539823 |
| http://awtrey.com/tutorials/dbeweb/database.php | 1 | 5 | 0,022727273 | 0,714285714 | 0,870919 | 0,044052863 |
| http://codex.cs.yale.edu/avi/db-book/slide-dir/ch7.pdf | 1 | 6 | 0,027272727 | 0,75 | 0,870868 | 0,052631579 |
| http://eagle.cs.uiuc.edu/pubs/2005/ranksqldemo-vldb05-laci-jun05.pdf | 1 | 7 | 0,031818182 | 0,777777778 | 0,870788 | 0,061135371 |
| http://education-portal.com/relational_database_fundamentals_ online_course.html | 1 | 8 | 0,036363636 | 0,8 | 0,870544 | 0,069565217 |
| http://csweb.bournemouth.ac.uk/aip/AIP%20Master%20 Database.pdf | 1 | 9 | 0,040909091 | 0,818181818 | 0,869749 | 0,077922078 |
| http://agron.scijournals.org/cgi/content/full/93/4/923 | 0 | 9 | 0,040909091 | 0,75 | 0,869052 | 0,077586207 |
| http://firebird.sourceforge.net/index.php | 1 | 10 | 0,045454545 | 0,769230769 | 0,868933 | 0,08583691 |
| http://cbbrowne.com/info/rdbms.html | 1 | 11 | 0,05 | 0,785714286 | 0,868082 | 0,094017094 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://adsabs.harvard.edu/abs/1993PASP..105.1482S | 0 | 11 | 0,05 | 0,733333333 | 0,866564 | 0,093617021 |
| http://ajp.amjpathol.org/cgi/content/abstract/159/3/837 | 0 | 11 | 0,05 | 0,6875 | 0,86161 | 0,093220339 |
| http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=326282 | 1 | 12 | 0,054545455 | 0,705882353 | 0,85585 | 0,101265823 |
| http://d3dnff.gat.com/D3DRDB/ | 0 | 12 | 0,054545455 | 0,666666667 | 0,855188 | 0,100840336 |
| http://dictionary.reference.com/search?q=relational+database | 1 | 13 | 0,059090909 | 0,684210526 | 0,804616 | 0,108786611 |
| http://2006.xmlconference.org/programme/presentations/188.html | 0 | 13 | 0,059090909 | 0,65 | 0,671497 | 0,108333333 |
| http://hms.liacs.nl/ilp.html | 0 | 13 | 0,059090909 | 0,619047619 | 0,627544 | 0,107883817 |
| http://biophysics.biol.uoa.gr/gpDB/ | 0 | 13 | 0,059090909 | 0,590909091 | 0,571546 | 0,107438017 |
| http://comjnl.oxfordjournals.org/cgi/content/abstract/39/2/124?ck=nck | 1 | 14 | 0,063636364 | 0,608695652 | 0,443234 | 0,115226337 |
| http://computerprogramming.suite101.com/article.cfm/sql_server_training_first_steps | 1 | 15 | 0,068181818 | 0,625 | 0,441608 | 0,12295082 |
| http://en.wikiversity.org/wiki/Topic:Object-relational_databases | 1 | 16 | 0,072727273 | 0,64 | 0,441608 | 0,130612245 |
| http://gateway.nlm.nih.gov/MeetingAbstracts/102276723.html | 1 | 17 | 0,077272727 | 0,653846154 | 0,439272 | 0,138211382 |
| http://www.surfermall.com/relational/lesson_1.htm | 1 | 18 | 0,081818182 | 0,666666667 | 0,432231 | 0,145748988 |
| http://www.onesmartclick.com/engineering/relational databases.html | 1 | 19 | 0,086363636 | 0,678571429 | 0,432191 | 0,153225806 |
| http://www.objectarchitects.de/arcus/cookbook/relzs/index.htm | 0 | 19 | 0,086363636 | 0,655172414 | 0,432181 | 0,152610442 |
| http://portal.acm.org/citation.cfm?id=77707 | 0 | 19 | 0,086363636 | 0,633333333 | 0,432111 | 0,152 |
| http://technet.microsoft.com/en-us/library/aa226072(SQL.80).aspx | 1 | 20 | 0,090909091 | 0,64516129 | 0,431992 | 0,15936255 |
| http://sql.z3950.org/docs/zSQLgate.html | 0 | 20 | 0,090909091 | 0,625 | 0,431932 | 0,158730159 |
| http://research.microsoft.com/apps/pubs/default.aspx?id=64571 | 1 | 21 | 0,095454545 | 0,636363636 | 0,431922 | 0,166007905 |
| http://pooteeweet.org/files/phpconf05/relational_database_starter_day.pdf | 1 | 22 | 0,1 | 0,647058824 | 0,431892 | 0,173228346 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://lists.xml.org/archives/xml-dev/200102/msg00009.html | 1 | 23 | 0,104545455 | 0,657142857 | 0,431852 | 0,180392157 |
| http://wiki.gxtechnical.com/commwiki/servlet/hwiki?Relational+Database+Theory | 1 | 24 | 0,109090909 | 0,666666667 | 0,431823 | 0,1875 |
| http://nar.oxfordjournals.org/cgi/content/full/26/1/335 | 0 | 24 | 0,109090909 | 0,648648649 | 0,431603 | 0,186770428 |
| http://pages.cs.wisc.edu/~cs764-1/selinger.pdf | 0 | 24 | 0,109090909 | 0,631578947 | 0,431374 | 0,186046512 |
| http://www4.wiwiss.fu-berlin.de/bizer/d2rq/ | 1 | 25 | 0,113636364 | 0,641025641 | 0,431184 | 0,193050193 |
| http://www.chass.utoronto.ca/epc/chwp/CHC2007/Liu_Smith/Lu_Smith.htm | 1 | 26 | 0,118181818 | 0,65 | 0,430914 | 0,2 |
| http://www.netl.doe.gov/publications/proceedings/01/carbon_seq/1a4.pdf | 0 | 26 | 0,118181818 | 0,634146341 | 0,42931 | 0,199233716 |
| http://ora.ouls.ox.ac.uk/objects/uuid:839dd46f-57f4-49ef-9582-8c154764a962 | 1 | 27 | 0,122727273 | 0,642857143 | 0,428554 | 0,20610687 |
| http://oreilly.com/pub/a/oreilly/frank/dbdesign_0701.html | 1 | 28 | 0,127272727 | 0,651162791 | 0,426226 | 0,212927757 |
| http://unicode.org/iuc/iuc13/c12/slides.ppt | 0 | 28 | 0,127272727 | 0,636363636 | 0,42452 | 0,212121212 |
| http://whydoeseverythingsuck.com/2008/02/death-of-relational-database.html | 1 | 29 | 0,131818182 | 0,644444444 | 0,424445 | 0,218867925 |
| http://www.ornl.gov/sci/techresources/Human_Genome/publicat/99santa/99.html | 1 | 30 | 0,136363636 | 0,652173913 | 0,424423 | 0,22556391 |
| http://www-db.in.tum.de/~teubnerj/publications/diss.pdf | 0 | 30 | 0,136363636 | 0,638297872 | 0,423268 | 0,224719101 |
| http://msdn.microsoft.com/en-us/library/bb245675.aspx | 1 | 31 | 0,140909091 | 0,645833333 | 0,418047 | 0,231343284 |
| http://www.codeproject.com/KB/database/introtomatisse_part2.aspx | 1 | 32 | 0,145454545 | 0,653061224 | 0,404196 | 0,237918216 |
| http://duro.sourceforge.net/ | 1 | 33 | 0,15 | 0,66 | 0,399733 | 0,244444444 |
| http://www.ibm.com/developerworks/edu/dm-dw-dm-0611nelke-i.html | 0 | 33 | 0,15 | 0,647058824 | 0,397774 | 0,243542435 |
| http://www.mysql.com/ | 1 | 34 | 0,154545455 | 0,653846154 | 0,396686 | 0,25 |
| http://ci.nii.ac.jp/naid/110003223310/ | 1 | 35 | 0,159090909 | 0,660377358 | 0,396247 | 0,256410256 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://nar.oxfordjournals.org/cgi/content/abstract/31/1/202 | 1 | 36 | 0,163636364 | 0,666666667 | 0,383506 | 0,262773723 |
| http://www.developers.net/enterprisedbshowcase/view/1323 | 0 | 36 | 0,163636364 | 0,654545455 | 0,378674 | 0,261818182 |
| http://www.postgresql.org/ | 1 | 37 | 0,168181818 | 0,660714286 | 0,369972 | 0,268115942 |
| http://www.auditmypc.com/acronym/ORDBMS.asp | 0 | 37 | 0,168181818 | 0,649122807 | 0,36904 | 0,267148014 |
| http://www.mysql.fr/news-and-events/on-demand-webinars/ display-od-216.html | 1 | 38 | 0,172727273 | 0,655172414 | 0,35247 | 0,273381295 |
| http://people.csail.mit.edu/jaffer/slib_6 | 1 | 39 | 0,177272727 | 0,661016949 | 0,343766 | 0,279569892 |
| http://www.matisse.com/ | 0 | 39 | 0,177272727 | 0,65 | 0,342954 | 0,278571429 |
| http://www.jumpbox.com/app/mysqld | 1 | 40 | 0,181818182 | 0,655737705 | 0,341912 | 0,284697509 |
| http://pnclink.org:8080/pnc2006/Abstract/Exemplary%20 Atlas%20--%20Jiang%20Wu.pdf | 1 | 41 | 0,186363636 | 0,661290323 | 0,341882 | 0,290780142 |
| http://briefingsdirectblog.blogspot.com/2007/10/ oracle-users-enjoy-open-source-benefits.html | 0 | 41 | 0,186363636 | 0,650793651 | 0,33847 | 0,28975265 |
| http://www.leavcom.com/db_08_00.htm | 1 | 42 | 0,190909091 | 0,65625 | 0,33745 | 0,295774648 |
| http://dotnet.sys-con.com/node/175864 | 0 | 42 | 0,190909091 | 0,646153846 | 0,337287 | 0,294736842 |
| http://computer.howstuffworks.com/question599.htm | 1 | 43 | 0,195454545 | 0,651515152 | 0,329892 | 0,300699301 |
| http://modperlbook.org/html/Chapter-20-Relational-Databases-and-mod_perl.html | 1 | 44 | 0,2 | 0,656716418 | 0,324412 | 0,306620209 |
| http://simple.wikipedia.org/wiki/Relational_database | 1 | 45 | 0,204545455 | 0,661764706 | 0,322709 | 0,3125 |
| http://jena.sourceforge.net/DB/index.html | 1 | 46 | 0,209090909 | 0,666666667 | 0,308941 | 0,3183391 |
| http://developer.mimer.com/ | 1 | 47 | 0,213636364 | 0,671428571 | 0,303576 | 0,324137931 |
| http://peds.oxfordjournals.org/cgi/content/abstract/2/6/431 | 1 | 48 | 0,218181818 | 0,676056338 | 0,299273 | 0,329896907 |
| http://www.garret.ru/gigabase.html | 1 | 49 | 0,222727273 | 0,680555556 | 0,281134 | 0,335616438 |
| http://www.altova.com/features_database.html | 1 | 50 | 0,227272727 | 0,684931507 | 0,243401 | 0,341296928 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://www.bkent.net/Doc/simple5.htm | 1 | 51 | 0,231818182 | 0,689189189 | 0,236501 | 0,346938776 |
| http://msdis.missouri.edu/presentations/gis_advanced/pdf/Relational.pdf | 1 | 52 | 0,236363636 | 0,693333333 | 0,224192 | 0,352542373 |
| http://mac.softpedia.com/get/Font-Tools/FrontBase.shtml | 1 | 53 | 0,240909091 | 0,697368421 | 0,148303 | 0,358108108 |
| http://lists.debian.org/debian-devel/2007/06/msg00106.html | 1 | 54 | 0,245454545 | 0,701298701 | 0,148101 | 0,363636364 |
| http://wdvl.com/Authoring/DB/Intro/relational_databases.html | 1 | 55 | 0,25 | 0,705128205 | 0,143945 | 0,369127517 |
| http://www.springerlink.com/index/7442wt8574x44766.pdf | 0 | 55 | 0,25 | 0,696202532 | 0,143825 | 0,367892977 |
| http://www.interpares.org/documents/interpares_cs_01_overview.pdf | 1 | 56 | 0,254545455 | 0,7 | 0,143082 | 0,373333333 |
| http://its.psu.edu/training/handouts/gs_reldb_sp06.pdf | 1 | 57 | 0,259090909 | 0,703703704 | 0,142851 | 0,378737542 |
| http://lists.xml.org/archives/xml-dev/200102/msg00148.html | 0 | 57 | 0,259090909 | 0,695121951 | 0,14083 | 0,377483444 |
| http://www.cemml.colostate.edu/files/un5.pdf | 1 | 58 | 0,263636364 | 0,698795181 | 0,140612 | 0,382838284 |
| http://searchsystemschannel.techtarget.com/topics/0,295493,sid99_tax305101,00.html | 1 | 59 | 0,268181818 | 0,702380952 | 0,140038 | 0,388157895 |
| http://relationaldatabasesoftware.surfpack.com/ | 1 | 60 | 0,272727273 | 0,705882353 | 0,139786 | 0,393442623 |
| http://erx.sagepub.com/cgi/content/abstract/25/5/533 | 0 | 60 | 0,272727273 | 0,697674419 | 0,139314 | 0,392156863 |
| http://jelle.druyts.net/2004/04/14/InheritanceModellingInARelationalDatabase.aspx | 0 | 60 | 0,272727273 | 0,689655172 | 0,138326 | 0,390879479 |
| http://www.onjava.com/pub/a/onjava/2006/04/12/object-to-relational-database-replciation-with-db40.html | 1 | 61 | 0,277272727 | 0,693181818 | 0,138061 | 0,396103896 |
| http://llc.oxfordjournals.org/cgi/content/abstract/2/2/89 | 0 | 61 | 0,277272727 | 0,685393258 | 0,138033 | 0,394822006 |
| http://www.rpbourret.com/xml/DataTransfer.htm | 0 | 61 | 0,277272727 | 0,677777778 | 0,137684 | 0,393548387 |
| http://www.islandnet.com/~tmc/html/articles/orareln.htm | 0 | 61 | 0,277272727 | 0,67032967 | 0,137659 | 0,392282958 |
| http://linuxfinances.info/info/rdbms.html | 1 | 62 | 0,281818182 | 0,673913043 | 0,137507 | 0,397435897 |
| http://user.it.uu.se/~udbl/Theses/QinZhangMSc.pdf | 0 | 62 | 0,281818182 | 0,666666667 | 0,137507 | 0,396166134 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://pd.acm.org/book_detail.cfm?isbn=0321305965 | 1 | 63 | 0,286363636 | 0,670212766 | 0,137127 | 0,401273885 |
| http://www.informatik.uni-trier.de/~ley/db/conf/icde/Stonebraker86.html | 0 | 63 | 0,286363636 | 0,663157895 | 0,137013 | 0,4 |
| http://www.fetac.ie/modules/C30147.PDF | 0 | 63 | 0,286363636 | 0,65625 | 0,136857 | 0,398734177 |
| http://www.malacolog.org/ | 1 | 64 | 0,290909091 | 0,659793814 | 0,136031 | 0,403785489 |
| http://java.sun.com/docs/books/tutorial/jdbc/overview/database.html | 1 | 65 | 0,295454545 | 0,663265306 | 0,135562 | 0,408805031 |
| http://scripts.iucr.org/cgi-bin/paper?S0108768102002458 | 1 | 66 | 0,3 | 0,666666667 | 0,135406 | 0,413793103 |
| http://www.treesearch.fs.fed.us/pubs/4548 | 1 | 67 | 0,304545455 | 0,67 | 0,135348 | 0,41875 |
| http://www.swc.scipy.org/lec/db.html | 1 | 68 | 0,309090909 | 0,673267327 | 0,135052 | 0,423676012 |
| http://portal.acm.org/citation.cfm?id=171128 | 1 | 69 | 0,313636364 | 0,676470588 | 0,134824 | 0,428571429 |
| http://codex.cs.yale.edu/avi/db-book/selected-exer-dir/7-web.pdf | 1 | 70 | 0,318181818 | 0,67961165 | 0,134809 | 0,433436533 |
| http://www.ebookee.com/The-Relational-Database-Dictionary-Extended-Edition_181022.html | 1 | 71 | 0,322727273 | 0,682692308 | 0,134726 | 0,438271605 |
| http://www.dcs.bbk.ac.uk/~mARK/relational_book.html | 0 | 71 | 0,322727273 | 0,676190476 | 0,134715 | 0,436923077 |
| http://studyat.anu.edu.au/courses/COMP2400;details.html | 0 | 71 | 0,322727273 | 0,669811321 | 0,134581 | 0,435582822 |
| http://www.informatik.uni-trier.de/~ley/db/conf/sigmod/SelingerACLP79.html | 0 | 71 | 0,322727273 | 0,663551402 | 0,134255 | 0,434250765 |
| http://compbio.soe.ucsc.edu/rdb/index.html | 1 | 72 | 0,327272727 | 0,666666667 | 0,134055 | 0,43902439 |
| http://www.zdnet.co.uk/tsearch/databases+relational+database.htm | 1 | 73 | 0,331818182 | 0,669724771 | 0,133969 | 0,443768997 |
| http://leap.sourceforge.net/ | 1 | 74 | 0,336363636 | 0,672727273 | 0,133946 | 0,448484848 |
| http://cll.emory.edu/it/classes.cfm?cla=-881825843&pt=3 | 1 | 75 | 0,340909091 | 0,675675676 | 0,133945 | 0,453172205 |
| http://www.sigcse.org/cc2001/IM.html | 1 | 76 | 0,345454545 | 0,678571429 | 0,133787 | 0,457831325 |
| http://www.logisticsworld.com/logistics/glossary.asp?query=Relational+Database+Management+System&search=exactterm& | 1 | 77 | 0,35 | 0,681415929 | 0,133619 | 0,462462462 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| form=show&acr=show&ref=show&rel=show&srl=show&llk=show&wiz=show&num=&hst=show&mode= | | | | | | |
| http://www.jegsworks.com/Lessons/databases/intro/database-relational.htm | 1 | 78 | 0,354545455 | 0,684210526 | 0,132814 | 0,467065868 |
| http://www.catalhoyuk.com/database/catal/Browse.asp | 0 | 78 | 0,354545455 | 0,67826087 | 0,132752 | 0,465671642 |
| http://gadfly.sourceforge.net/gadfly.html | 1 | 79 | 0,359090909 | 0,681034483 | 0,132547 | 0,470238095 |
| http://www.bibsonomy.org/bibtex/2b871e617ebe9672da36fe27790d06e5f/dblp | 0 | 79 | 0,359090909 | 0,675213675 | 0,132547 | 0,46884273 |
| http://uttc.umn.edu/training/courses/description?designator=DB101 | 0 | 79 | 0,359090909 | 0,669491525 | 0,13237 | 0,467455621 |
| http://www.isprs.org/congresses/beijing2008/proceedings/2_pdf/1_WG-II-1/14.pdf | 1 | 80 | 0,363636364 | 0,672268908 | 0,13237 | 0,471976401 |
| http://shopping.msn.com/prices/relational-database-design-clearly-explained/itemid2439036/?itemtext=itemname:relational database-design-clearly-explained | 1 | 81 | 0,368181818 | 0,675 | 0,132344 | 0,476470588 |
| http://www.defmacro.org/ramblings/relational.html | 1 | 82 | 0,372727273 | 0,67768595 | 0,132024 | 0,480938416 |
| http://download.microsoft.com/download/8/f/a/8fa3268a-d34f-4b3d-bb72-72e08701096f/Worldwide%20Relational%20Database%20Management%20Systems%202007%20Vendor%20Shares.pdf | 0 | 82 | 0,372727273 | 0,672131148 | 0,131969 | 0,479532164 |
| http://relational-database-software.qarchive.org/ | 1 | 83 | 0,377272727 | 0,674796748 | 0,131884 | 0,483965015 |
| http://searchoracle.techtarget.com/tip/0,289483,sid41_gci1217363,00.html | 0 | 83 | 0,377272727 | 0,669354839 | 0,131825 | 0,48255814 |
| http://searchoracle.bitpipe.com/plist/term/Relational-Database-Management-Software.html | 1 | 84 | 0,381818182 | 0,672 | 0,131485 | 0,486956522 |
| http://gadfly.sourceforge.net/ | 1 | 85 | 0,386363636 | 0,674603175 | 0,131337 | 0,49132948 |
| http://www.empress.com/ | 0 | 85 | 0,386363636 | 0,669291339 | 0,131337 | 0,489913545 |
| http://en.wikipedia.org/wiki/Relational_database | 1 | 86 | 0,390909091 | 0,671875 | 0,131119 | 0,494252874 |
| http://www.o-xml.org/news/18-mar-2003.html | 1 | 87 | 0,395454545 | 0,674418605 | 0,130942 | 0,498567335 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://plc.inf.elte.hu/erlang/pub/refac_db_kent.ppt | 0 | 87 | 0,395454545 | 0,669230769 | 0,130835 | 0,497142857 |
| http://gmod.org/wiki/Glossary | 1 | 88 | 0,4 | 0,671755725 | 0,130827 | 0,501424501 |
| http://searchsystemschannel.techtarget.com/tip/0,289483, sid99_gci1255470,00.html | 1 | 89 | 0,404545455 | 0,674242424 | 0,130765 | 0,505681818 |
| http://cdlr.strath.ac.uk/pubs/dunsireg/alm03main.pps | 0 | 89 | 0,404545455 | 0,669172932 | 0,130625 | 0,504249292 |
| http://www.miswebdesign.com/resources/articles/ wrox-beginning-php-4-chapter-3-1.html | 1 | 90 | 0,409090909 | 0,671641791 | 0,130612 | 0,508474576 |
| http://lists.w3.org/Archives/Public/public-rdf-dawg/ 2004JanMar/0208.html | 0 | 90 | 0,409090909 | 0,666666667 | 0,130601 | 0,507042254 |
| http://dictionary.reference.com/search?q=relational+database &r=66 | 1 | 91 | 0,413636364 | 0,669117647 | 0,13051 | 0,511235955 |
| http://foldoc.org/?relational+database | 0 | 91 | 0,413636364 | 0,664233577 | 0,130381 | 0,509803922 |
| http://www.techbookreport.com/tbr0273.html | 0 | 91 | 0,413636364 | 0,65942029 | 0,130287 | 0,508379888 |
| http://www.ouhk.edu.hk/WCM/?FUELAP_TEMPLATENAME= tcGenericPage&itemid=CC_COURSE_INFO_58222950&lang=eng | 1 | 92 | 0,418181818 | 0,661870504 | 0,130217 | 0,512534819 |
| http://www.agiledata.org/essays/mappingObjects.html | 1 | 93 | 0,422727273 | 0,664285714 | 0,130214 | 0,516666667 |
| http://blog.terracottatech.com/2008/11/breaking_down_the_ relational_d.html | 1 | 94 | 0,427272727 | 0,666666667 | 0,130201 | 0,520775623 |
| http://www.xml.com/pub/a/2007/07/12/xquery-and-data-abstraction.html | 1 | 95 | 0,431818182 | 0,669014085 | 0,130201 | 0,524861878 |
| http://www.funpecrp.com.br/gmr/year2007/ vol4-6/xm0012_abstract.html | 1 | 96 | 0,436363636 | 0,671328671 | 0,130185 | 0,52892562 |
| http://research.amnh.org/amcc/Freezerworks.html | 1 | 97 | 0,440909091 | 0,673611111 | 0,130161 | 0,532967033 |
| http://www.cs.virginia.edu/papers/ismb02_sql.pdf | 1 | 98 | 0,445454545 | 0,675862069 | 0,13014 | 0,536986301 |
| http://www.stylusstudio.com/xml_database.html | 0 | 98 | 0,445454545 | 0,671232877 | 0,130105 | 0,535519126 |
| http://www.ispras.ru/~knizhnik/gigabase.html | 0 | 98 | 0,445454545 | 0,666666667 | 0,130075 | 0,534059946 |
| http://www.arcchip.cz/w05/w05_buitleir.pdf | 1 | 99 | 0,45 | 0,668918919 | 0,130007 | 0,538043478 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://fyi.oreilly.com/2008/11/relational-database-technology.html | 1 | 100 | 0,454545455 | 0,67114094 | 0,12999 | 0,54200542 |
| http://db.apache.org/derby/ | 1 | 101 | 0,459090909 | 0,673333333 | 0,129916 | 0,545945946 |
| http://www.builderau.com.au/program/xml/soa/Transfer-and-store-data-from-an-XML-document-in-a-relational database/0,339028469,339273299,00.htm | 1 | 102 | 0,463636364 | 0,675496689 | 0,129876 | 0,549865229 |
| http://www.databasedev.co.uk/database_normalization_ process.html | 1 | 103 | 0,468181818 | 0,677631579 | 0,129869 | 0,553763441 |
| http://searchoracle.techtarget.com/generic/0,295582,sid41_ gci1091031,00.html | 1 | 104 | 0,472727273 | 0,679738562 | 0,129866 | 0,557640751 |
| http://www.cephbase.utmb.edu/ | 0 | 104 | 0,472727273 | 0,675324675 | 0,129824 | 0,556149733 |
| http://research.microsoft.com/apps/pubs/default.aspx?id=64535 | 1 | 105 | 0,477272727 | 0,677419355 | 0,129821 | 0,56 |
| http://invasions.si.edu/nbic/search.html | 1 | 106 | 0,481818182 | 0,679487179 | 0,129671 | 0,563829787 |
| http://ocw.mit.edu/NR/rdonlyres/Urban-Studies-and-Planning/11-208Introduction-to-Computers-in-Public-Management-IIJanuary--IAP-2002/64B3A7CB-FA1F-4749-869C-A5D96ABCBE50/0/lect52.pdf | 1 | 107 | 0,486363636 | 0,681528662 | 0,129636 | 0,567639257 |
| http://www.sqlconsole.com/ | 1 | 108 | 0,490909091 | 0,683544304 | 0,129612 | 0,571428571 |
| http://www.lavoisier.fr/notice/gbBCO3AXSM3OP2RO.html | 0 | 108 | 0,490909091 | 0,679245283 | 0,129598 | 0,569920844 |
| http://publish.uwo.ca/~craven/558/558red.htm | 0 | 108 | 0,490909091 | 0,675 | 0,129535 | 0,568421053 |
| http://troels.arvin.dk/db/rdbms/links/ | 1 | 109 | 0,495454545 | 0,677018634 | 0,129528 | 0,572178478 |
| http://www.ohloh.net/tags/database | 0 | 109 | 0,495454545 | 0,672839506 | 0,129496 | 0,570680628 |
| http://itc.ktu.lt/itc353/Vysnia353.pdf | 1 | 110 | 0,5 | 0,674846626 | 0,129371 | 0,574412533 |
| http://pubs.water.usgs.gov/ofr01359 | 1 | 111 | 0,504545455 | 0,676829268 | 0,129363 | 0,578125 |
| http://stats.oecd.org/glossary/detail.asp?ID=4520 | 1 | 112 | 0,509090909 | 0,678787879 | 0,129361 | 0,581818182 |
| http://www.snee.com/bobdc.blog/2008/07/devx-article-relational-databa.html | 1 | 113 | 0,513636364 | 0,680722892 | 0,12933 | 0,585492228 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://www.warriorforum.com/programming-talk/48802-relational-database.html | 1 | 114 | 0,518181818 | 0,682634731 | 0,129314 | 0,589147287 |
| http://www.euclideanspace.com/software/information/relational/index.htm | 1 | 115 | 0,522727273 | 0,68452381 | 0,129232 | 0,592783505 |
| http://gamma.cs.unc.edu/DB/ | 1 | 116 | 0,527272727 | 0,686390533 | 0,129227 | 0,596401028 |
| http://db.uwaterloo.ca/OED/trdbms.html | 0 | 116 | 0,527272727 | 0,682352941 | 0,129221 | 0,594871795 |
| http://cidoc.ics.forth.gr/docs/Implementing_the_CIDOC_CRM.rtf | 0 | 116 | 0,527272727 | 0,678362573 | 0,129192 | 0,593350384 |
| http://news.cnet.com/Rethinking-the-relational-database/2010-1015_3-5715457.html | 0 | 116 | 0,527272727 | 0,674418605 | 0,129141 | 0,591836735 |
| http://www.sirdug.org/ | 1 | 117 | 0,531818182 | 0,676300578 | 0,129129 | 0,595419847 |
| http://technet.microsoft.com/en-us/library/ms189559(SQL.90).aspx | 1 | 118 | 0,536363636 | 0,67816092 | 0,129068 | 0,598984772 |
| http://en.wikipedia.org/wiki/Relational_model | 1 | 119 | 0,540909091 | 0,68 | 0,129065 | 0,602531646 |
| http://www.athro.com/general/Phyloinformatics_7_85x11.pdf | 0 | 119 | 0,540909091 | 0,676136364 | 0,129018 | 0,601010101 |
| http://www.jot.fm/issues/issue_2003_09/article1.pdf | 1 | 120 | 0,545454545 | 0,677966102 | 0,129018 | 0,604534005 |
| http://libra.msra.cn/papercited.aspx?id=352778 | 1 | 121 | 0,55 | 0,679775281 | 0,128954 | 0,608040201 |
| http://stinet.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix=html&identifier=ADA313447 | 1 | 122 | 0,554545455 | 0,681564246 | 0,128913 | 0,611528822 |
| http://macs.about.com/od/glossaryqt/g/relational.htm | 1 | 123 | 0,559090909 | 0,683333333 | 0,128905 | 0,615 |
| http://connect.educause.edu/Library/Abstract/TheEffectofRelationalData/30473 | 1 | 124 | 0,563636364 | 0,685082873 | 0,128793 | 0,618453865 |
| http://www.archive.org/search.php?query=mediatype%3Aeducation%20AND%20collection%3Aarsdigita%20AND%20subject%3A%22Relational%20Database%20Management%20Systems%22 | 1 | 125 | 0,568181818 | 0,686813187 | 0,128708 | 0,621890547 |
| http://www.dds-lite.com/ | 0 | 125 | 0,568181818 | 0,683060109 | 0,128696 | 0,620347395 |
| http://www.codata.org/08conf/abstracts/ChangjunHu-A%20Visual%20Tool%20for%20Building%20MatML%20Data%20 | 0 | 125 | 0,568181818 | 0,679347826 | 0,128655 | 0,618811881 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| from%20Material%20Science%20Relational%20Database.htm | | | | | | |
| http://plone.org/documentation/faq/plone-relational-database | 1 | 126 | 0,572727273 | 0,681081081 | 0,12859 | 0,622222222 |
| http://www.omegahat.org/RSPostgres/Scenarios.pdf | 1 | 127 | 0,577272727 | 0,682795699 | 0,128533 | 0,625615764 |
| http://ec.europa.eu/ipg/standards/databases/standard_rdbms_en.htm | 1 | 128 | 0,581818182 | 0,684491979 | 0,128513 | 0,628992629 |
| http://www.onlamp.com/pub/a/onlamp/2001/05/25/ postgresql_mvcc.html | 1 | 129 | 0,586363636 | 0,686170213 | 0,128494 | 0,632352941 |
| http://www.cs.vt.edu/node/4585 | 1 | 130 | 0,590909091 | 0,687830688 | 0,128475 | 0,635696822 |
| http://www.hitsw.com/products_services/downloads.html | 1 | 131 | 0,595454545 | 0,689473684 | 0,12837 | 0,63902439 |
| http://www.databasedev.co.uk/data_models.html | 1 | 132 | 0,6 | 0,691099476 | 0,128281 | 0,642335766 |
| http://www.codebeach.com/index.asp?authorName=Relational% 20Database%20Consultants | 0 | 132 | 0,6 | 0,6875 | 0,128252 | 0,640776699 |
| http://www.peopleware.net/0177/index.cfm?eventDisp=CDMAA | 0 | 132 | 0,6 | 0,683937824 | 0,128229 | 0,639225182 |
| http://www.gsd.harvard.edu/gis/manual/relational/index.htm | 1 | 133 | 0,604545455 | 0,68556701 | 0,128163 | 0,642512077 |
| http://www.ampl.com/NEW/tables.html | 0 | 133 | 0,604545455 | 0,682051282 | 0,128136 | 0,640963855 |
| http://www.freepatentsonline.com/5905985.html | 0 | 133 | 0,604545455 | 0,678571429 | 0,128127 | 0,639423077 |
| http://www.eecs.berkeley.edu/Pubs/TechRpts/1978/12384.html | 0 | 133 | 0,604545455 | 0,675126904 | 0,128125 | 0,637889688 |
| http://www.agentjim.com/MVP/Excel/RelationalOffice.htm | 0 | 133 | 0,604545455 | 0,671717172 | 0,128037 | 0,63663636 |
| http://www.eas.asu.edu/~winrdbi/author/ | 0 | 133 | 0,604545455 | 0,668341709 | 0,128023 | 0,634844869 |
| http://articles.techrepublic.com.com/5100-22_11-5075453.html | 1 | 134 | 0,609090909 | 0,67 | 0,127951 | 0,638095238 |
| http://www.firstsql.com/ireldb.htm | 1 | 135 | 0,613636364 | 0,671641791 | 0,12795 | 0,641330166 |
| http://www.zoominfo.com/Industries/software-mfg/software-development-design/relational-database-management-system.htm | 1 | 136 | 0,618181818 | 0,673267327 | 0,127939 | 0,644549763 |
| http://arnab.org/blog/web-20-and-relational-database | 1 | 137 | 0,622727273 | 0,674876847 | 0,127906 | 0,647754137 |
| http://bioinformatics.oxfordjournals.org/cgi/content/abstract/14/2/188 | 1 | 138 | 0,627272727 | 0,676470588 | 0,127906 | 0,650943396 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://casoilresource.lawr.ucdavis.edu/drupal/node/264 | 1 | 139 | 0,631818182 | 0,67804878 | 0,127906 | 0,654117647 |
| http://codingforums.com/showthread.php?t=44774 | 0 | 139 | 0,631818182 | 0,674757282 | 0,127906 | 0,65258216 |
| http://criminaljustice.state.ny.us/crimnet/clf/oracle/oracle.htm | 0 | 139 | 0,631818182 | 0,671497585 | 0,127906 | 0,651053864 |
| http://en.wikibooks.org/wiki/Category:Relational_Database_Design | 1 | 140 | 0,636363636 | 0,673076923 | 0,127906 | 0,654205607 |
| http://en.wikipedia.org/wiki/Relational_database_management_system | 1 | 141 | 0,640909091 | 0,674641148 | 0,127906 | 0,657342657 |
| http://encyclopedia.kids.net.au/page/re/Relational_database | 1 | 142 | 0,645454545 | 0,676190476 | 0,127906 | 0,660465116 |
| http://ferret.pmel.noaa.gov/HOMEPAGE/LAS/FAQ/ relational_database_access.htm | 0 | 142 | 0,645454545 | 0,672985782 | 0,127906 | 0,658932715 |
| http://gd.tuwien.ac.at:8050/H/1/ | 1 | 143 | 0,65 | 0,674528302 | 0,127906 | 0,662037037 |
| http://hal.archives-ouvertes.fr/docs/00/16/81/52/PDF/Hrivnac.pdf | 1 | 144 | 0,654545455 | 0,676056338 | 0,127906 | 0,665127021 |
| http://havemacwillblog.com/2008/11/10/6-reasons-why-relational-database-will-be-superseded/ | 1 | 145 | 0,659090909 | 0,677570093 | 0,127906 | 0,668202765 |
| http://highscalability.com/search-source-data-how-simpledb-differs-rdbms | 1 | 146 | 0,663636364 | 0,679069767 | 0,127906 | 0,671264368 |
| http://homepages.inf.ed.ac.uk/sviglas/pubs/OrderedXML.pdf | 1 | 147 | 0,668181818 | 0,680555556 | 0,127906 | 0,674311927 |
| http://ideas.repec.org/p/boc/nsug08/13.html | 1 | 148 | 0,672727273 | 0,68202765 | 0,127906 | 0,677345538 |
| http://ieeexplore.ieee.org/iel2/3041/8637/00380352.pdf?arnumber=380352 | 1 | 149 | 0,677272727 | 0,683486239 | 0,127906 | 0,680365297 |
| http://ilpubs.stanford.edu:8090/4/ | 1 | 150 | 0,681818182 | 0,684931507 | 0,127906 | 0,683371298 |
| http://infolab.stanford.edu/~ullman/fcdb/spr99/lec4.ps | 0 | 150 | 0,681818182 | 0,681818182 | 0,127906 | 0,681818182 |
| http://intl.ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=20424&arnumber=943703&count=113&index=92 | 1 | 151 | 0,686363636 | 0,683257919 | 0,127906 | 0,684807256 |
| http://lips.informatik.uni-leipzig.de/pub/1998-8 | 0 | 151 | 0,686363636 | 0,68018018 | 0,127906 | 0,683257919 |
| http://vsis-www.informatik.uni-hamburg.de/publications/view.php/164 | 1 | 152 | 0,690909091 | 0,68161435 | 0,127906 | 0,686230248 |
| http://www.100best-web-hosting.com/glossary674.html | 1 | 153 | 0,695454545 | 0,683035714 | 0,127906 | 0,689189189 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://www.15seconds.com/Issue/020522.htm | 1 | 154 | 0,7 | 0,684444444 | 0,127906 | 0,692134831 |
| http://www.actapress.com/PDFViewer.aspx?paperId=14993 | 1 | 155 | 0,704545455 | 0,685840708 | 0,127906 | 0,695067265 |
| http://www.adobe.com/devnet/coldfusion/articles/display_dyn_data_02.html | 1 | 156 | 0,709090909 | 0,68722467 | 0,127906 | 0,697986577 |
| http://www.aemj.org/cgi/content/abstract/7/5/472-a?ck=nck | 0 | 156 | 0,709090909 | 0,684210526 | 0,127906 | 0,696428571 |
| http://www.agiledata.org/essays/relationalDatabases.html | 1 | 157 | 0,713636364 | 0,68558952 | 0,127906 | 0,699331849 |
| http://www.almaden.ibm.com/cs/projects/iis/hdb/Publications/papers/sigmod98_dbi.pdf | 0 | 157 | 0,713636364 | 0,682608696 | 0,127906 | 0,697777778 |
| http://www.altova.com/press/2003-01-13_hitsw.pdf | 0 | 157 | 0,713636364 | 0,67965368 | 0,127906 | 0,696230599 |
| http://www.ambysoft.com/mappingObjects.html | 1 | 158 | 0,718181818 | 0,681034483 | 0,127906 | 0,699115044 |
| http://www.annauniv.edu/rcc/meseorSyllabus/SE072.pdf | 0 | 158 | 0,718181818 | 0,678111588 | 0,127906 | 0,697571744 |
| http://www.auditmypc.com/acronym/RDBMS.asp | 1 | 159 | 0,722727273 | 0,679487179 | 0,127906 | 0,700440529 |
| http://www.basis-wien.at/index.php?id=72&L=2 | 0 | 159 | 0,722727273 | 0,676595745 | 0,127906 | 0,698901099 |
| http://www.bibl.liu.se/liupubl/disp/disp96/tek452s.htm | 1 | 160 | 0,727272727 | 0,677966102 | 0,127906 | 0,701754386 |
| http://www.biodatamining.org/content/1/1/7 | 1 | 161 | 0,731818182 | 0,679324895 | 0,127906 | 0,704595186 |
| http://www.blackwell-synergy.com/doi/abs/10.1111/j.1540-8159.1996.tb03421.x | 1 | 162 | 0,736363636 | 0,680672269 | 0,127906 | 0,707423581 |
| http://www.bmrb.wisc.edu/search/rela_database.html | 0 | 162 | 0,736363636 | 0,677824268 | 0,127906 | 0,705882353 |
| http://www.boinc-wiki.info/Relational_Data_Base_Management_System | 1 | 163 | 0,740909091 | 0,679166667 | 0,127906 | 0,708695652 |
| http://www.boingboing.net/2009/01/21/keeping-up-with-lost.html | 1 | 164 | 0,745454545 | 0,680497925 | 0,127906 | 0,711496746 |
| http://www.businessdictionary.com/definition/relational-database.html | 1 | 165 | 0,75 | 0,681818182 | 0,127906 | 0,714285714 |
| http://www.cetus-links.org/oo_db_systems_2.html | 1 | 166 | 0,754545455 | 0,683127572 | 0,127906 | 0,717062635 |
| http://www.chass.utoronto.ca/emls/iemls/mqlibrary/search.html | 0 | 166 | 0,754545455 | 0,680327869 | 0,127906 | 0,715517241 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://www.chemcomp.com/journal/reldb.htm | 1 | 167 | 0,759090909 | 0,681632653 | 0,127906 | 0,71827957 |
| http://www.codeproject.com/KB/showcase/object_relational_mapping.aspx | 1 | 168 | 0,763636364 | 0,682926829 | 0,127906 | 0,721030043 |
| http://www.constable.com/ | 0 | 168 | 0,763636364 | 0,680161943 | 0,127906 | 0,719486081 |
| http://www.contentdsi.com/content_management.htm | 0 | 168 | 0,763636364 | 0,677419355 | 0,127906 | 0,717948718 |
| http://www.cosis.net/abstracts/9IKC/00196/9IKC-A-00196-1.pdf | 0 | 168 | 0,763636364 | 0,674698795 | 0,127906 | 0,71641791 |
| http://www.cs.dal.ca/news/def-1127.shtml | 1 | 169 | 0,768181818 | 0,676 | 0,127906 | 0,719148936 |
| http://www.cs.mcgill.ca/~kemme/papers/pakdd06.pdf | 0 | 169 | 0,768181818 | 0,673306773 | 0,127906 | 0,717622081 |
| http://www.cs.ucla.edu/~zaniolo/papers/widm06.pdf | 0 | 169 | 0,768181818 | 0,670634921 | 0,127906 | 0,716101695 |
| http://www.cs.uwaterloo.ca/~ashraf/pubs/cikm08recman.pdf | 0 | 169 | 0,768181818 | 0,66798419 | 0,127906 | 0,714587738 |
| http://www.cse.lehigh.edu/~heflin/pubs/psss03-poster.pdf | 0 | 169 | 0,768181818 | 0,665354331 | 0,127906 | 0,713080169 |
| http://www.databasecolumn.com/2007/11/dbms-origins.html | 1 | 170 | 0,772727273 | 0,666666667 | 0,127906 | 0,71578947 |
| http://www.db.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/relational_database.htm | 1 | 171 | 0,777272727 | 0,66796875 | 0,127906 | 0,718487395 |
| http://www.dbforums.com/ | 1 | 172 | 0,781818182 | 0,6692607 | 0,127906 | 0,721174004 |
| http://www.dbmsmag.com/9804d13.html | 1 | 173 | 0,786363636 | 0,670542636 | 0,127906 | 0,723849372 |
| http://www.deskspace.com/ | 0 | 173 | 0,786363636 | 0,667953668 | 0,127906 | 0,722338205 |
| http://www.dhdursoassociates.com/ | 0 | 173 | 0,786363636 | 0,665384615 | 0,127906 | 0,720833333 |
| http://www.dspace.cam.ac.uk/handle/1810/14718 | 1 | 174 | 0,790909091 | 0,666666667 | 0,127906 | 0,723492723 |
| http://www.duoconsulting.com/downloads/ContentManagement.pdf | 1 | 175 | 0,795454545 | 0,667938931 | 0,127906 | 0,726141079 |
| http://www.ebmt.org/4registry/Registry_docs/ProMISe%20Docs/THE%20EBMT%20RELATIONAL%20DATABASE.pdf | 0 | 175 | 0,795454545 | 0,66539924 | 0,127906 | 0,724637681 |
| http://www.eclipse.org/webtools/community/tutorials/RDBTutorial/RDBTutorial.html | 1 | 176 | 0,8 | 0,666666667 | 0,127906 | 0,727272727 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://www.emis.de/journals/AUA/acta3/modelareadb.pdf | 1 | 177 | 0,804545455 | 0,667924528 | 0,127906 | 0,729896907 |
| http://www.emis.de/journals/NSJOM/Papers/26_2/NSJOM_26_2_049_073.pdf | 1 | 178 | 0,809090909 | 0,669172932 | 0,127906 | 0,732510288 |
| http://www.epa.gov/enviro/html/ef_home.html | 0 | 178 | 0,809090909 | 0,666666667 | 0,127906 | 0,73100616 |
| http://www.fujipress.jp/finder/xslt.php?mode=present&inputfile=JACII000600010005.xml | 1 | 179 | 0,813636364 | 0,667910448 | 0,127906 | 0,733606557 |
| http://www.funpecrp.com.br/gmr/year2007/vol4-6//pdf/xm0012.pdf | 1 | 180 | 0,818181818 | 0,669144981 | 0,127906 | 0,736196319 |
| http://www.geekgirls.com/database_dictionary.htm | 1 | 181 | 0,822727273 | 0,67037037 | 0,127906 | 0,73877551 |
| http://www.geekgirls.com/databases_from_scratch_3.htm | 1 | 182 | 0,827272727 | 0,671586716 | 0,127906 | 0,741344196 |
| http://www.gridpp.ac.uk/papers/GGF3Rome2001.pdf | 0 | 182 | 0,827272727 | 0,669117647 | 0,127906 | 0,739837398 |
| http://www.haz-map.com/ | 1 | 183 | 0,831818182 | 0,67032967 | 0,127906 | 0,742393509 |
| http://www.hitsw.com/products_services/whitepapers/integrating_xml_rdb/ | 0 | 183 | 0,831818182 | 0,667883212 | 0,127906 | 0,740890688 |
| http://www.hpss-collaboration.org/hpss/about/BoomerRDBMSHSM.pdf | 1 | 184 | 0,836363636 | 0,669090909 | 0,127906 | 0,743434343 |
| http://www.idealliance.org/papers/extreme/proceedings/xslfo-pdf/2007/Ramalho01/EML2007Ramalho01.pdf | 1 | 185 | 0,840909091 | 0,670289855 | 0,127906 | 0,745967742 |
| http://www.idealliance.org/proceedings/xml04/papers/254/XQueryRDS.pdf | 1 | 186 | 0,845454545 | 0,671480144 | 0,127906 | 0,748490946 |
| http://www.ietf.org/proceedings/94mar/mgt/rdbmsmib.html | 1 | 187 | 0,85 | 0,672661871 | 0,127906 | 0,751004016 |
| http://www.ifi.uzh.ch/arvo/dbtg/vldbphd2007/Camera-Ready%20Papers/Paper%206/XQuery_Optimization.pdf | 1 | 188 | 0,854545455 | 0,673835125 | 0,127906 | 0,753507014 |
| http://www.informatik.hu-berlin.de/Forschung_Lehre/wbi/publications/2005/dils05_ontologies.pdf | 1 | 189 | 0,859090909 | 0,675 | 0,127906 | 0,756 |
| http://www.informationweek.com/news/management/showArticle.jhtml?articleID=159401656 | 1 | 190 | 0,863636364 | 0,676156584 | 0,127906 | 0,758483034 |
| http://www.ingentaconnect.com/content/els/09666362/1995/00000003/00000004/art82904;jsessionid=213d6w9tu8vsp.alexandra | 1 | 191 | 0,868181818 | 0,677304965 | 0,127906 | 0,760956175 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://www.intersystems.com/cache/whitepapers/hybrid.html | 1 | 192 | 0,872727273 | 0,67844523 | 0,127906 | 0,763419483 |
| http://www.iop.org/EJ/abstract/1742-6596/119/4/042009 | 1 | 193 | 0,877272727 | 0,679577465 | 0,127906 | 0,765873016 |
| http://www.itia.ntua.gr/en/projinfo/50/ | 1 | 194 | 0,881818182 | 0,680701754 | 0,127906 | 0,768316832 |
| http://www.itjungle.com/tug/tug071008-story05.html | 1 | 195 | 0,886363636 | 0,681818182 | 0,127906 | 0,770750988 |
| http://www.japmaonline.org/cgi/content/abstract/86/2/74 | 0 | 195 | 0,886363636 | 0,679442509 | 0,127906 | 0,769230769 |
| http://www.javaworld.com/javaworld/jw-09-2007/jw-09-columndb.html | 1 | 196 | 0,890909091 | 0,680555556 | 0,127906 | 0,771653543 |
| http://www.jcc.com/DescImplementingDBUsingSQL.htm | 1 | 197 | 0,895454545 | 0,6816609 | 0,127906 | 0,774066798 |
| http://www.jcc.com/DescRelationalDBDesign.htm | 1 | 198 | 0,9 | 0,682758621 | 0,127906 | 0,776470588 |
| http://www.kirupa.com/developer/php/relational_db_design2.htm | 1 | 199 | 0,904545455 | 0,683848797 | 0,127906 | 0,778864971 |
| http://www.laas.fr/~esorics/notices/Yazdanian90.html | 0 | 199 | 0,904545455 | 0,681506849 | 0,127906 | 0,77734375 |
| http://www.linux-mag.com/id/2093 | 1 | 200 | 0,909090909 | 0,682593857 | 0,127906 | 0,779727096 |
| http://www.mail-archive.com/accessvbacentral@yahoogroups.com/msg00555.html | 1 | 201 | 0,913636364 | 0,683673469 | 0,127906 | 0,782101167 |
| http://www.mail-archive.com/accessvbacentral@yahoo groups.com/msg00567.html | 1 | 202 | 0,918181818 | 0,684745763 | 0,127906 | 0,784466019 |
| http://www.membranetransport.org/ | 0 | 202 | 0,918181818 | 0,682432432 | 0,127906 | 0,782945736 |
| http://www.midcarb.org/Documents/GSA-Nov-2000.shtml | 0 | 202 | 0,918181818 | 0,68013468 | 0,127906 | 0,781431335 |
| http://www.mindmodel.com/ | 0 | 202 | 0,918181818 | 0,677852349 | 0,127906 | 0,77992278 |
| http://www.nationmultimedia.com/worldhotnews/30091764/ Oracle-is-the-number-1-Relational-Database-in-Thailand | 1 | 203 | 0,922727273 | 0,678929766 | 0,127906 | 0,782273603 |
| http://www.objectarchitects.de/ObjectArchitects/orpatterns/ | 1 | 204 | 0,927272727 | 0,68 | 0,127906 | 0,784615385 |
| http://www.objenv.com/cetus/oo_db_systems_2.html | 1 | 205 | 0,931818182 | 0,681063123 | 0,127906 | 0,786948177 |
| http://www.openoffice.org/servlets/ReadMsg?list= discuss&msgNo=39173 | 0 | 205 | 0,931818182 | 0,678807947 | 0,127906 | 0,785440613 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| http://www.oracle.com/database/berkeley-db/index.html | 1 | 206 | 0,936363636 | 0,679867987 | 0,127906 | 0,787762906 |
| http://www.oracle.com/database/docs/Berkeley-DB-v-Relational.pdf | 1 | 207 | 0,940909091 | 0,680921053 | 0,127906 | 0,790076336 |
| http://www.pinnaclekenya.com/ | 0 | 207 | 0,940909091 | 0,678688525 | 0,127906 | 0,788571429 |
| http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347395 | 1 | 208 | 0,945454545 | 0,679738562 | 0,127906 | 0,790874525 |
| http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243450 | 1 | 209 | 0,95 | 0,680781759 | 0,127906 | 0,79316888 |
| http://www.python.org/workshops/1997-10/proceedings/shprentz.html | 1 | 210 | 0,954545455 | 0,681818182 | 0,127906 | 0,795454545 |
| http://www.questia.com/PM.qst?a=o&se=gglsc&d=5001678243 | 0 | 210 | 0,954545455 | 0,67961165 | 0,127906 | 0,793950851 |
| http://www.relationalwizards.com/ | 0 | 210 | 0,954545455 | 0,677419355 | 0,127906 | 0,79245283 |
| http://www.remcomp.fr/asmanet/asmapro/asmawork.htm | 0 | 210 | 0,954545455 | 0,675241158 | 0,127906 | 0,790960452 |
| http://www.research.ibm.com/journal/sj/424/telford.html | 1 | 211 | 0,959090909 | 0,676282051 | 0,127906 | 0,793233083 |
| http://www.reviews.com/review/review_review.cfm?review_id=136349 | 1 | 212 | 0,963636364 | 0,677316294 | 0,127906 | 0,795497186 |
| http://www.sapdb.org/ | 1 | 213 | 0,968181818 | 0,678343949 | 0,127906 | 0,797752809 |
| http://www.securityfocus.com/bid/23007 | 0 | 213 | 0,968181818 | 0,676190476 | 0,127906 | 0,796261682 |
| http://www.serc.iisc.ernet.in/ComputingFacilities/systems/cluster/vac-7.0/html/glossary/czgr.htm | 0 | 213 | 0,968181818 | 0,674050633 | 0,127906 | 0,794776119 |
| http://www.snee.com/xml/xml2006/owlrdbms.html | 1 | 214 | 0,972727273 | 0,675078864 | 0,127906 | 0,797020484 |
| http://www.star.le.ac.uk/~cgp/adass2002.html | 0 | 214 | 0,972727273 | 0,672955975 | 0,127906 | 0,795539033 |
| http://www.techbriefs.com/component/content/68?task=view | 0 | 214 | 0,972727273 | 0,670846395 | 0,127906 | 0,79406308 |
| http://www.textbooksrus.com/search/BookDetail/?isbn=0201752840&kbid=1067 | 0 | 214 | 0,972727273 | 0,66875 | 0,127906 | 0,792592593 |
| http://www.thestandard.com/news/2008/02/04/start-readies-easy-use-online-relational-database | 1 | 215 | 0,977272727 | 0,669781931 | 0,127906 | 0,794824399 |
| http://www.thunderstone.com/site/texisman/ | 1 | 216 | 0,981818182 | 0,670807453 | 0,127906 | 0,79704797 |

| URL | Relevância | Acertos | Cobertura | Precisão | Ordenação Qualidade | FMQualidade |
|---|---|---|---|---|---|---|
| relational_database_background.html | | | | | | |
| http://www.turnkeylinux.org/appliances/mysql | 0 | 216 | 0,981818182 | 0,66873065 | 0,127906 | 0,79558011 |
| http://www.turnkeylinux.org/appliances/postgresql | 1 | 217 | 0,986363636 | 0,669753086 | 0,127906 | 0,797794118 |
| http://www.utexas.edu/cc/database/datamodeling/ | 1 | 218 | 0,990909091 | 0,670769231 | 0,127906 | 0,8 |
| http://www.vertica.com/relational-database-management-system | 0 | 218 | 0,990909091 | 0,668711656 | 0,127906 | 0,798534799 |
| http://www.vldb.org/conf/2004/IND5P2.PDF | 0 | 218 | 0,990909091 | 0,666666667 | 0,127906 | 0,797074954 |
| http://www.vldb2005.org/program/paper/thu/p1175-pal.pdf | 0 | 218 | 0,990909091 | 0,664634146 | 0,127906 | 0,795620438 |
| http://www.w3.org/XML/RDB.html | 0 | 218 | 0,990909091 | 0,662613982 | 0,127906 | 0,79417122 |
| http://www.warthman.com/projects-tymshare-b.htm | 0 | 218 | 0,990909091 | 0,660606061 | 0,127906 | 0,792727273 |
| http://www.xml.com/pub/a/2003/03/05/tmrdb.html | 1 | 219 | 0,995454545 | 0,66163142 | 0,127906 | 0,79491833 |
| http://www.yolinux.com/HOWTO/PostgreSQL-HOWTO.html | 0 | 219 | 0,995454545 | 0,659638554 | 0,127906 | 0,793478261 |
| http://www.zdnet.com.au/whitepaper/0,2000063328,22462862p-16001235q,00.htm | 0 | 219 | 0,995454545 | 0,657657658 | 0,127906 | 0,7920434 |
| http://www.zope.org/Documentation/Books/ZopeBook/2_6 Edition/RelationalDatabases.stx | 0 | 219 | 0,995454545 | 0,655688623 | 0,127906 | 0,790613718 |
| http://www-03.ibm.com/ibm/history/exhibits/vintage/ vintage_4506VV3151.html | 0 | 219 | 0,995454545 | 0,653731343 | 0,127906 | 0,789189189 |
| http://www-10.lotus.com/ldd/lcdforum.nsf/Date AllThreadedweb/4655b9a7d2bdab49852572e300240dc9?Open Document | 0 | 219 | 0,995454545 | 0,651785714 | 0,127906 | 0,787769784 |
| http://www3.open.ac.uk/courses/bin/p12.dll?C01M359 | 0 | 219 | 0,995454545 | 0,649851632 | 0,127906 | 0,786355476 |
| http://xml.coverpages.org/blake94-ps.gz | 0 | 219 | 0,995454545 | 0,647928994 | 0,127906 | 0,784946237 |
| http://xml-and-relational-database.nuclearscripts.com/ | 1 | 220 | 1 | 0,648967552 | 0,127906 | 0,787119857 |
| http://www.openldap.org/faq/data/cache/378.html | 0 | 220 | 1 | 0,647058824 | 0,127067 | 0,785714286 |