UNIVERSIDADE DE SÃO PAULO
ESCOLA DE COMUNICAÇÕES E ARTES

CARLOS ALBERTO CORRÊA

Indexação automática e visualização de informações: um estudo
baseado em Lógica paraconsistente

São Paulo

2011

CARLOS ALBERTO CORRÊA

Indexação automática e visualização de informações: um estudo
baseado em Lógica paraconsistente

Tese apresentada ao Programa de Pós-graduação
em Ciência da Informação da Escola de
Comunicações e Artes da Universidade de São
Paulo para obtenção do título de Doutor.

Área de Concentração: Cultura e Informação

Orientadora: Prof. Dra. Nair Yumiko Kobashi

São Paulo

2011

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo da Publicação
Serviço de Biblioteca e Documentação
Escola de Comunicações e Artes da Universidade de São Paulo

Correa, Carlos Alberto

Indexação automática e visualização de informações : um estudo baseado em lógica paraconsistente / Carlos Alberto Correa – São Paulo : C. A. Correa, 2011.

152 p. : il.

Tese (Doutorado) – Escola de Comunicações e Artes / Universidade de São Paulo.

Orientadora: Prof^a Dr^a Nair Yumiko Kobashi

1. Indexação automática 2. Visualização de informação 3. Lógica paraconsistente 4. Modelo do espaço vetorial 5. Atribuição de pesos I. Kobashi, Nair Yumiko II. Título.

CDD 21.ed. – 029.5

CORRÊA, C. A. Indexação automática e visualização de informações: um estudo baseado em Lógica paraconsistente. Tese apresentada ao Programa de Pós-graduação em Ciência da Informação da Escola de Comunicações e Artes da Universidade de São Paulo para obtenção do título de Doutor em Ciência da Informação.

Aprovada em:

Banca Examinadora

Prof. Dr. _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Prof. Dr. _____ Instituição: _____
Julgamento: _____ Assinatura: _____

Dedico esse trabalho,
aos meus pais, Norival e Geny,
ao meu irmão Ulisses,
aos meus filhos Victor e Marina.

Agradecimentos

A professora Nair Kobashi, pela oportunidade oferecida, pela confiança depositada, pela paciência exercida e pelos conhecimentos repassados.

A Sandra, pelo apoio constante, pela paciência com as inevitáveis e inúmeras horas ausentes, e com o stress que um trabalho desta envergadura causa, aos envolvidos direta ou diretamente com o mesmo.

Aos professores Rosane Minghim e Fernando Paulovich, do ICMC da USP, campus São Carlos, pelas inúmeras trocas de idéias, experiências e opiniões sem as quais este trabalho teria sido possível.

Aos colegas Marivalde, Mônica e Cibele, da USP, pelas dicas, conselhos e trocas de idéias indispensáveis ao desenvolvimento desse trabalho.

A José Guilherme Vasconcelos, Walmir Martinez Thomaz, Edenia Cipriano e Marly de Souza, respectivamente, Superintendente (8ª RF), Delegado, Delegada Adjunta e Chefe do Serviço de Tecnologia e Segurança da Informação (Delegacia da Receita Federal do Brasil), pelo apoio, incentivo e atos administrativos que possibilitaram a participação no curso de Doutorado concomitantemente ao desempenho de minhas atividades profissionais. Em particular, à Marly de Souza, pelo apoio recebido no final do curso.

A Gabriel de Faria Andery pela inestimável ajuda para compreensão e modificação do código fonte original do Projection Explorer.

Aos colegas, Kleber, Alex, Sergio, Cássio e Edgard, funcionários da Receita Federal do Brasil e SERPRO, pelo apoio, incentivo e ajuda na configuração do documento final.

A todas as outras pessoas que direta ou indiretamente contribuíram para que este trabalho alcançasse sua forma final.

À Causa Primária de Todas as Coisas. Independentemente do nome por que seja conhecida.

“Não entendo. Isso é tão vasto que ultrapassa qualquer entender. Entender é sempre limitado. Mas não entender pode não ter fronteiras. Sinto que sou muito mais completa quando não entendo. Não entender, do modo como falo, é um dom. Não entender, mas não como um simples de espírito. O bom é ser inteligente e não entender. É uma benção estranha, como ter loucura sem ser doida. É um desinteresse manso, é uma doçura de burrice. Só que de vez em quando vem a inquietação: quero entender um pouco. Não demais: mas pelo menos entender que não entendo”.

Clarice Lispector

”O homem cria instrumentos e depois molda-se à imagem deles. A realidade é fabricada pelo homem. O Universo é uma invenção, uma metáfora”.

John Brockman

Resumo

CORRÊA, C. A. Indexação automática e visualização de informações: um estudo baseado em Lógica paraconsistente

Pode-se afirmar que os processos de organização, recuperação e visualização de informações estão interligados por pelo menos um ponto em comum: a indexação de textos ou documentos. A indexação, um componente essencial do tratamento e organização de informações, é também importante tanto para a recuperação quanto para a visualização de informações. Nesse contexto, destacam-se as soluções que objetivam automatizar a indexação. As pesquisas que propõe soluções de indexação automática são desenvolvidas com base em diferentes pressupostos teóricos, entre os quais podem ser destacados: a Estatística, a Lingüística e o Controle de Vocabulários (Leiva, 1999). A grande maioria dessas soluções utiliza modelos híbridos que combinam dois ou três desses pressupostos. Outras abordagens para o problema da indexação automática utilizam, além dos pressupostos anteriormente citados, teorias que permitem o tratamento da incerteza, da imprecisão e da vagueza, tais como a teoria de Dempster-Shafer e a lógica difusa. O objetivo desta pesquisa é avaliar o potencial de utilização da lógica paraconsistente, uma lógica não clássica, com capacidade para tratar situações que envolvem incerteza, imprecisão e vagueza nos procedimentos de indexação automática. A hipótese de trabalho sustentada nesta pesquisa apóia-se teoricamente nos princípios da Lógica Paraconsistente. A utilização dessa lógica e de métodos derivados da mesma, por ser flexível e comportar estados lógicos que vão além das dicotomias sim e não, permite adiantar a hipótese de que os resultados da indexação poderão ser melhores do que os obtidos por métodos tradicionais. Do ponto de vista metodológico, optou-se pela utilização de um algoritmo para tratamento de incerteza e imprecisão, desenvolvido no âmbito da lógica paraconsistente, para modificar os valores dos pesos atribuídos aos termos de indexação. Os testes foram realizados em um sistema de visualização de informações, com código fonte disponível. As coleções utilizadas são disponibilizadas com o sistema. Os resultados obtidos foram avaliados por meio de critérios e índices embutidos no próprio sistema de visualização, e demonstram ganhos mensuráveis de qualidade na construção das visualizações, confirmando, assim a hipótese de pesquisa elencada.

Abstract

CORRÊA, C. A. Automatic indexing and information visualization: a research based on Paraconsistent logic

It can be argued that the processes of organization, retrieval and visualization of information are connected, presenting one point in common: the indexing of texts or documents. Indexing, an essential component of text analysis, is important to information retrieval and visualization. In this context, we highlight solutions to automate indexing. Researches that propose solutions for automatic indexing are developed based on different theoretical assumptions as, for example: Statistics, Linguistics and Controlled vocabularies (Leiva, 1999). Most of these solutions uses hybrid models combining two or three of these theoretical assumptions. Other approaches to the problem of automatic indexing uses, besides the aforementioned assumptions, theories that allow the treatment of uncertainty, imprecision and vagueness, as the Dempster-Shafer theory and the fuzzy logic. The aim of this research is to evaluate the use of paraconsistent logic, a nonclassical logic, capable of dealing with situations involving uncertainty, imprecision and vagueness, in the procedures of automatic indexing. The working hypothesis supported in this research is based theoretically on the principles of Paraconsistent Logic. The utilization of this logic, being flexible and containing logical states that go beyond the dichotomies yes and no, permit to advance the hypothesis that the results of indexing could be better than those obtained by traditional methods. From the methodological point of view, we chose to use an algorithm for treatment of uncertainty and imprecision, developed under the paraconsistent logic, to modify the values of the weights assigned to index terms. The tests were performed on a information visualization system, with source code available. The collections used are available with the system. The results were evaluated by criteria and indices built into the information visualization system itself, and demonstrate measurable gains in the construction quality of the displays, thus confirming the hypothesis listed.

Lista de Figuras

Figura 1 - Uma visão esquemática da Recuperação de informação.....	19
Figura 2 - Tipos de estudos de busca e recuperação de informação.....	28
Figura 3 - Importância da indexação: influências e interações.....	43
Figura 4 - Representação gráfica (ou vetorial) de um documento com três termos de indexação.....	51
Figura 5 - Representação gráfica (ou vetorial) de um documento com dois termos de indexação.....	52
Figura 6 - Matriz documentos X pesos.....	52
Figura 7 - Contexto pensamento – inferências X linguagem – argumentos.....	57
Figura 8 - Contexto realidade X descrição.....	58
Figura 9 - Representação difusa de documentos.....	66
Figura 10 - Representação dos estados lógicos extremos a partir dos graus de crença/descrença – QUPC.....	70
Figura 11 - Representação de alguns estados lógicos intermediários.....	70
Figura 12 - QUPC com linhas de limitação de regiões.....	71
Figura 13 - QUPC com linhas de limitação de regiões identificadas.....	71
Figura 14 - QUPC destacando a região de totalmente inconsistente.....	72
Figura 15 - QUPC destacando a região de totalmente indeterminado.....	72
Figura 16 - QUPC destacando a região de totalmente falso.....	72
Figura 17 - QUPC destacando a região de totalmente verdadeiro.....	73
Figura 18 - QUPC destacando a região de quase falso tendendo ao indeterminado.....	73
Figura 19 - QUPC destacando a região de quase verdadeiro tendendo ao inconsistente.....	74
Figura 20 - QUPC de resolução 12 destacando todas as regiões possíveis – intermediárias e extremas.....	74
Figura 21- QUPC de resolução 12- destaque para regiões com variáveis delimitadoras modificadas.....	76
Figura 22 - QUPC de resolução 12- destaque para regiões com variáveis delimitadoras modificadas.....	76
Figura 23 - Marcha de Napoleão na Rússia.....	82
Figura 24 - Modelo de referência de visualização.....	84
Figura 25 - Análise da informação e visualização.....	87

Figura 26 - Exemplo de visualização no sistema Infosky.....	88
Figura 27 - Visualização no sistema VxInsign.....	88
Figura 28 - Visualização de documentos no Projection Explorer – destaque p/ documentos de assuntos diferentes porem com algum grau de similaridade entre si.....	90
Figura 29 - Processo para construção da visualização de uma coleção de documentos.....	92
Figura 30 - Exemplo de visualização – PEx.....	93
Figura 31 - Exemplo de visualização no PEx com identificação de tópicos em agrupamentos de textos.....	94
Figura 32 - diferentes agrupamentos do mesmo conjunto original de dados.....	96
Figura 33 - Exemplo de agrupamentos produzidos por diferentes algoritmos.....	98
Figura 34 - Tela do PEx com destaque da opção da matriz termos x documentos utilizando o para-analisador.....	102
Figura 35 - Exemplo de curvas de Neighborhood Hit.....	104
Figura 36 - Comparação de <i>neighborhood hit</i> para a coleção CBR-ILP-IR-SON.....	109
Figura 37 - Comparação de <i>neighborhood hit</i> para a coleção CBR-ILP-IR.....	109
Figura 38 - Comparação de <i>neighborhood hit</i> para a coleção CBR-ILP-SON.....	109
Figura 39 - Comparação de <i>neighborhood hit</i> para a coleção IR-SON.....	109
Figura 40 - Comparação de <i>neighborhood hit</i> para a coleção CBR-ILP.....	109
Figura 41 - Exemplos de matrizes de similaridades para agrupamentos produzidos por diferentes algoritmos.....	116
Figura 42 - Matriz de similaridades original.....	117
Figura 43 - Matriz de similaridades - Região R0 do QUPC.....	117
Figura 44 - Matriz de similaridades - Região R1 do QUPC.....	117
Figura 45 - Matriz de similaridades - Região R2 do QUPC.....	117
Figura 46 - Matriz de similaridades- Região R3 do QUPC.....	117
Figura 47 - Matriz de similaridades original.....	118
Figura 48 - Matriz de similaridades - Região R0 do QUPC.....	118
Figura 49 - Matriz de similaridades - Região R1 do QUPC.....	118
Figura 50 - Matriz de similaridades - Região R2 do QUPC.....	118
Figura 51 - Matriz de similaridades - Região R3 do QUPC.....	118
Figura 52 - Matriz de similaridades original.....	119
Figura 53 - Matriz de similaridades - Região R0 do QUPC.....	119
Figura 54 - Matriz de similaridades - Região R1 do QUPC.....	119
Figura 55 - Matriz de similaridades - Região R2 do QUPC.....	119

Figura 56 - Matriz de similaridades - Região R3 do QUPC.....	119
Figura 57 - Matriz de similaridades original.....	120
Figura 58 - Matriz de similaridades - Região R0 do QUPC.....	120
Figura 59 - Matriz de similaridades - Região R1 do QUPC.....	120
Figura 60 - Matriz de similaridades - Região R2 do QUPC.....	120
Figura 61 - Matriz de similaridades - Região R3 do QUPC.....	120
Figura 62 - Matriz de similaridades original.....	121
Figura 63 - Matriz de similaridades - Região R0 do QUPC.....	121
Figura 64 - Matriz de similaridades - Região R1 do QUPC.....	121
Figura 65 - Matriz de similaridades - Região R2 do QUPC.....	121
Figura 66 - Matriz de similaridades - Região R3 do QUPC.....	121
Figura 67 - Exemplo de histograma de distâncias.....	122
Figura 68 - Histograma de distâncias.....	123
Figura 69 - Histograma de distâncias.....	123
Figura 70 - Histograma de distâncias.....	123
Figura 71 - Histograma de distâncias.....	123
Figura 72 - Histograma de distâncias.....	124
Figura 73 - Histograma de distâncias.....	124
Figura 74 - Histograma de distâncias.....	124
Figura 75 - Histograma de distâncias.....	124
Figura 76 - Histograma de distâncias.....	124
Figura 77 - Histograma de distâncias.....	124
Figura 78 - Visualização original.....	143
Figura 79 - Visualização da Região R0 do QUPC.....	143
Figura 80 - Visualização da Região R1 do QUPC.....	144
Figura 81 - Visualização da Região R2 do QUPC.....	144
Figura 82 - Visualização da Região R3 do QUPC.....	144
Figura 83 - Visualização original.....	145
Figura 84 - Visualização da Região R0 do QUPC.....	145
Figura 85 - Visualização da Região R1 do QUPC.....	146
Figura 86 - Visualização da Região R2 do QUPC.....	146
Figura 87 - Visualização da Região R3 do QUPC.....	146
Figura 88 - Visualização original.....	147
Figura 89 - Visualização da Região R0 do QUPC.....	147

Figura 90 -Visualização da Região R1 do QUPC.....	148
Figura 91 -Visualização da Região R2 do QUPC.....	148
Figura 92 -Visualização da Região R3 do QUPC.....	148
Figura 93- Visualização original.....	149
Figura 94 - Visualização da Região R0 do QUPC.....	149
Figura 95 - Visualização da Região R1 do QUPC.....	150
Figura 96 - Visualização da Região R2 do QUPC.....	150
Figura 97 - Visualização da Região R3 do QUPC.....	150
Figura 98- Visualização original 150.....	151
Figura 99 - Visualização da Região R0 do QUPC.....	151
Figura 100 - Visualização da Região R1 do QUPC.....	152
Figura 101 - Visualização da Região R2 do QUPC.....	152
Figura 102 - Visualização da Região R3 do QUPC.....	152

Lista de Tabelas

Tabela 1 - Interpretação dos valores do coeficiente de silhueta.....114

Tabela 2 - Coeficientes de silhueta das coleções, de acordo com a região do QUPC.....114

Lista de Siglas

HTML - *HyperText Markup Language*

LPA2v - Lógica paraconsistente anotada com anotação de dois valores

MEV - Modelo do espaço vetorial

NASA - *National Administration of Space and Aeronautic*

QUPC - Quadrado unitário do plano cartesiano

SMART - *System for the manipulation and retrieval of text*

Sumário

1 INTRODUÇÃO.....	18
2 ORGANIZAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO.....	23
2.1 Organização da informação.....	23
2.2 Recuperação da informação.....	25
2.3 Representação e linguagem.....	31
3 INDEXAÇÃO AUTOMÁTICA.....	40
3.1 Conceitos e princípios.....	40
3.2 Aspectos teóricos e características.....	43
3.3 Limites e dificuldades da indexação.....	44
3.4 Automação da indexação.....	47
3.5 O modelo do espaço vetorial.....	50
4 INDEXAÇÃO AUTOMÁTICA E LÓGICA PARACONSISTENTE.....	56
4.1 Introdução às lógicas clássica e não-clássica.....	56
4.2 Indexação e lógica difusa.....	61
4.3 Indexação e lógica paraconsistente.....	67
5 VISUALIZAÇÃO DE INFORMAÇÃO.....	81
5.1 Introdução à visualização.....	81
5.2 Similaridade, distancia e proximidade.....	89
5.3 Um sistema de visualização: Projection Explorer.....	91
6 INTRODUÇÃO A ANÁLISE DE AGRUPAMENTOS.....	95
7 EXPERIMENTO COMPUTACIONAL.....	100
7.1 Construção de visualizações no PEx.....	101
7.2 Coleções de teste utilizadas.....	102
7.3 Avaliação de visualizações.....	103
7.4 Uso do para-analisador no Projection Explorer - resultados obtidos.....	105

8 CONSIDERAÇÕES FINAIS.....	125
8.1 Sugestões para trabalhos futuros.....	127
REFERÊNCIAS.....	128
APENDICES.....	143

1 Introdução

Motivada principalmente por necessidades práticas, a humanidade desenvolveu um enorme conjunto de técnicas, “tecnologias” e saberes voltados a um objetivo central: armazenar, externamente à mente humana, os incontáveis registros mentais de conhecimento, criados e acumulados a partir de sua interação social e com o meio em que vive. Algumas técnicas ou “tecnologias” talvez não tenham sido criadas com esse objetivo em mente, mas puderam ser adaptadas e utilizadas para suprir as necessidades de armazenamento mnemônico do saber socialmente produzido.

Assim, não seria de todo impreciso considerar que a fala, a escrita, a tipografia, as bibliotecas e o computador eletrônico constituem exemplos de “dispositivos” e organizações criados ou adaptados para serem utilizados como mecanismos externos de armazenamento.

No curso dos acontecimentos que levaram ao desenvolvimento dos dispositivos informacionais, pode-se citar o surgimento de ciências e das práticas tecnológicas que modificaram a face do mundo e da estrutura das sociedades.

Por outro lado, o volume dos estoques de registros externos passou a constituir-se numa espiral ascendente, aumentando consideravelmente as dificuldades para sua manutenção, organização e manipulação. Nesse contexto, se inserem as discussões que envolvem os conceitos de dado, informação, conhecimento, mensagem e documento, que culminaram com a gênese da Biblioteconomia, da Documentação, e, recentemente, da Ciência da Informação, esta última, surgida em meados do século passado (CAPURRO, 2003).

Dado, informação e conhecimento são termos que apresentam um caráter e utilização essencialmente multi/inter disciplinares, sendo difícil, dessa forma, haver consenso sobre as definições e usos. Discussões específicas sobre as conexões entre informação conhecimento e documento podem ser encontradas, por exemplo, em Buckland (1991) e Le Coadic (1996). Aspectos sobre as origens da Documentação e as relações entre informação e documento também estão presentes em Rayward (1997). Questões sobre as conexões de informação e documento com outros conceitos estão presentes em trabalhos recentes de Zins (2007), ao discutir a configuração contemporânea da Ciência da Informação.

As interações entre o armazenamento e a recuperação de informações/documentos (a dualidade dos termos foi mantida com o objetivo de considerar a interação sob um ponto de vista abrangente) podem ser esquematizadas na figura a seguir.

Figura 1 – Uma visão esquemática da Recuperação de informação
Fonte: Belkin, 1993, p. 57

Nesse caso, a dualidade informação/documento foi substituída pelo termo texto. Por outro lado, os estudos e discussões relativos a esses elementos podem ser abordados sob vários pontos de vista, sendo que, na maioria das vezes, seus componentes podem ser reduzidos a três elementos básicos, que interagem, de alguma forma, entre si, como descrito a seguir.

Considerando a interação segundo o ponto de vista situacional, ou seja, **o contexto ou situação** de intermediação entre os estoques e a recuperação de informação, tem-se

Estoques de informação ← Busca por informação → Problema/Estado Anômalo do Conhecimento/Incerteza;

Considerando a interação segundo o ponto de vista teórico-conceitual, ou seja, **as áreas de estudos** que fornecem subsídios teóricos para pensar a intermediação entre os estoques e a recuperação de informação, tem-se:

Organização da informação ← Recuperação de informação → Necessidades de informação

Considerando a interação segundo o ponto de vista operacional, ou seja, a **materialização do contexto situacional**, com base nas áreas de estudo, em **artefatos manipuláveis ou representáveis** pelos seres humanos, tem-se:

Textos/Documentos (ou outros objetos informacionais) ← Sistema de recuperação de informação → Consultas.

A figura 1 e os esquemas acima dão apenas uma dimensão aproximada da problemática envolvida nos procedimentos de recuperação de informação. Na verdade, cada objeto ou processo, descrito na figura 1, dá origem a um ou mais campos de estudo que se ampliam à medida que os problemas e situações envolvidos são aprofundados.

Em relação à representação, presente na figura 1, um dos principais métodos, conhecido por indexação, consiste em atribuir termos, ou palavras chave, aos documentos

A indexação, dada a sua importância nos procedimentos de recuperação de informação, é motivo de muitas, diversificadas e multidisciplinares pesquisas e desenvolvimentos em todo o mundo. Contemporaneamente, um dos maiores aliados na busca por soluções para a indexação, são as tecnologias da informação e comunicação.

Além do uso dos computadores, a busca de soluções em indexação automática tem levado os pesquisadores a se envolverem com diversas disciplinas, tais como: Lingüística, Estatística, Terminologia e diversas teorias desenvolvidas no âmbito da Matemática e da Lógica, tais como teorias para o tratamento de incerteza e as lógicas não-clássicas.

A utilização da lógica difusa nos processos de recuperação de informação e de indexação propriamente ditos, suscita reflexões sobre a possibilidade de utilização de outras teorias, igualmente capazes de lidar com situações que envolvem incerteza e imprecisão. Nessa perspectiva, foi de particular importância a proposta de Sucolotti (2001), de utilização da Lógica paraconsistente (outro tipo de lógica não-clássica, assim como a lógica difusa), em procedimentos de recuperação de informação. Nesse trabalho, o autor propõe a utilização da lógica paraconsistente para expandir as consultas efetuadas a um sistema de recuperação de informação, com a utilização de tesauros. O autor não realizou testes de sua proposta e se limitou a um estudo teórico sobre a viabilidade de utilizar a lógica paraconsistente. O estudo citado deu origem à idéia de verificar o potencial de utilização da lógica paraconsistente em procedimentos de indexação automática, objeto da presente pesquisa.

Cabe acrescentar ainda que, tradicionalmente, os sistemas de recuperação de informação são o ambiente preferido para testes de soluções de indexação automática.

Contudo, outros sistemas também são fortemente dependentes desses procedimentos, tais como os sistemas de visualização de informação. Nesse caso, a indexação é um componente fundamental do sistema. A indexação automática é utilizada, para calcular os graus de similaridade entre os componentes da coleção a ser visualizada. Esse cálculo permite determinar, entre outras coisas, a distância entre os componentes da coleção que será visualizada. Nesses sistemas, são criadas metáforas visuais para representar os componentes da coleção como, por exemplo, pequenos círculos coloridos. Nessa situação, a distância se materializa através da representação, na tela do sistema de visualização, da proximidade entre os componentes visualizados.

Os argumentos anteriores estimulam a enunciação do seguinte problema de pesquisa, formulada sob a forma de uma questão: É possível utilizar a Lógica paraconsistente nos processos automáticos de indexação? Em caso positivo, de que maneira isso poderá ser feito?

Tendo em vista os trabalhos já realizados no campo da recuperação da informação, é bastante plausível a possibilidade de sucesso na utilização da Lógica paraconsistente em indexação automática. A utilização dessa lógica, por ser flexível e comportar estados lógicos que vão além das dicotomias sim e não, permite adiantar a hipótese de que os resultados da indexação poderão ser melhores do que os obtidos por métodos tradicionais. Tais métodos são baseados em cálculos estatísticos e manipulação de textos a partir do processamento da linguagem natural. Com base nessas reflexões e hipóteses de trabalho, podem ser enunciados os objetivos da presente pesquisa.

1.1 Objetivos

Objetivo geral

Verificar o potencial de utilização da Lógica paraconsistente, uma lógica não clássica, em procedimentos de indexação automática, utilizando um ambiente computacional para os testes de validação.

Objetivos específicos

- Verificar a possibilidade de utilização de heurísticas para tratamento da incerteza, imprecisão e vagueza, desenvolvidas no âmbito da Lógica paraconsistente, nos procedimentos de indexação automática.
- Implementar a aplicação de tais heurísticas em um sistema de visualização de informação.
- Comparar, através de critérios específicos, os resultados da utilização de um modelo de indexação tradicional e sua modificação utilizando-se os parâmetros e métodos da Lógica paraconsistente.

1.2 Organização do trabalho

O percurso da pesquisa para alcançar os objetivos elencados acima, está estruturado em 8 capítulos, resumidos a seguir.

Na Introdução são apresentados o contexto, as hipóteses e objetivos da pesquisa.

No Capítulo 2, são descritos os princípios e os pressupostos teóricos relativos à recuperação e organização de informação.

No Capítulo 3 são expostos os conceitos, correntes teóricas, teorias e algumas soluções de indexação automática utilizadas no âmbito de sistemas de recuperação de informação.

No Capítulo 4 é feita uma breve introdução aos fundamentos das lógicas clássica e não-clássica, especialmente a Lógica paraconsistente. Além disso, são descritas soluções que utilizam lógica difusa (outra lógica não clássica) e os aspectos metodológicos da utilização da Lógica paraconsistente nos processos de indexação automática.

No Capítulo 5 são apresentados os principais conceitos associados à visualização de informações bem como o sistema de visualização utilizado para a realização dos testes da proposta.

O Capítulo 6 introduz conceitos e métodos associados à análise de agrupamentos

O Capítulo 7 descreve a realização dos testes e os resultados obtidos com a aplicação da proposta efetuada no capítulo 4.

No Capítulo 8 são apresentadas as considerações finais e algumas perspectivas de continuidade das pesquisas sobre o tema desenvolvido na tese.

2. Organização e recuperação de informações

“... um homem não pode investigar sobre aquilo que já sabe nem sobre aquilo que ignora; pois, se sabe, não precisa investigar; e, se ignora, não pode, pois ele não sabe, sequer, sobre que assunto vai investigar.”

Platão

2.1 Organização da Informação

A estruturação dos estoques de informação, com o objetivo de proporcionar sua recuperação futura, desenvolveu-se no contexto da “explosão informacional” ocorrida a partir do século XIX. O advento dos computadores eletrônicos, principalmente, da Internet, só agravou esta situação.

As tecnologias eletrônicas foram vistas, num primeiro momento, como uma possível solução para os problemas advindos da “explosão informacional”. Contudo, essas tecnologias também podem ter contribuído para o agravamento de tal situação. Nesse sentido, Svenonius (2000), chama a atenção ao afirmar que essas tecnologias, isoladamente, não são suficientes, pois a efetividade de um sistema para acessar informação é uma função direta da inteligência utilizada para organizá-la.

No âmbito da Ciência da Informação existem diversas abordagens que envolvem as expressões *organização da informação*, *organização do conhecimento*, *representação da informação* e *representação do conhecimento*. Tais expressões denotam áreas de pesquisa criadas e desenvolvidas com o objetivo principal de propiciar acesso aos estoques de conhecimento produzidos. Antes de caracterizá-las ou defini-las, deve-se chamar a atenção para uma distinção básica e intrínseca a ser observada nessas expressões: os processos e os produtos resultantes. Nesse sentido, organização e representação da informação bem como organização e representação do conhecimento podem ser vistos como processos, com objetivos e métodos distintos, que dão origem a produtos igualmente distintos, geralmente utilizados de maneira inter-relacionada.

Cabe acrescentar, ainda, que não existe padronização quanto à utilização dessas expressões. Muitas vezes, os termos *organização da informação* e *organização do conhecimento* são utilizados sem definição precisa ou com os autores alternando sua utilização, em um mesmo texto, como se descrevessem os mesmos processos e métodos. O mesmo pode ser dito em relação às expressões *representação da informação* e *representação do conhecimento*. Foge aos objetivos deste trabalho um aprofundamento dessas definições ou a discussão de eventuais disputas teóricas. Contudo, é possível afirmar que, muitas das polêmicas são consequência das dificuldades encontradas para definir e caracterizar de forma precisa os conceitos de informação e de conhecimento.

No âmbito deste trabalho serão utilizadas as expressões *organização da informação* e *representação da informação* devidamente definidas. Esses termos serão distinguidos de *organização e representação do conhecimento*.

O processo de organização de informação tem um objetivo específico: possibilitar a recuperação de informação por alguém que dela necessite (Alvarenga, 2003; Taylor, 1999). Uma interessante caracterização e definição das áreas de organização da informação e do conhecimento foi efetuada por Bräscher e Café (2008). As autoras definem organização da informação como “*um processo que envolve a descrição física e de conteúdo de objetos informacionais*” (BRÄSCHER; CAFÉ, 2008. p. 5).

Sinteticamente, pode-se considerar que os estudos que envolvem a organização da informação estão associados a pelo menos três domínios específicos: a Classificação, a Análise Documentária e a Terminologia.

Pombo (1998) afirma que o processo de classificação deve ser visto como natural, óbvio e indiscutível, uma vez que classificar fatos, entes e acontecimentos é um mecanismo básico utilizado para por ordem no mundo cognoscível. Esta autora destaca quatro grandes orientações para o problema das classificações: (i) ontológica – classificação dos seres; (ii) gnosiológica – classificação das ciências; (iii) biblioteconômica – classificação dos livros; (iv) informacional – classificação de informações. Afirma, ainda, que a principal diferença entre as classificações das ciências e as classificações documentais e biblioteconômicas reside no caráter especulativo das primeiras em oposição aos objetivos funcionais da segunda.

A classificação não é utilizada apenas com o intuito de organizar os estoques de informação e conhecimento. Soergel (apud BRASCHER; CAFÉ, 2008) especifica que, entre outros objetivos, os sistemas de representação do conhecimento apresentam a propriedade de prover a classificação para a ação. Ou seja, a classificação tem uso prático em diferentes

atividades profissionais, tais como a classificação de doenças para diagnóstico médico e a classificação de mercadorias para o comércio.

Classificar os conteúdos dos diferentes registros do conhecimento (textos, documentos, imagens, entre outros) não é o bastante para torná-los acessíveis. Algum outro tipo de tratamento dever ser dado a eles. É nesse contexto que se identifica a atuação da Análise documentária.

De acordo com Kobashi (1994) a expressão Análise documentária foi conceituada pelo pesquisador francês Jean-Claude Gardin, como um conjunto de procedimentos realizados com o objetivo de expressar o conteúdo de documentos científicos, sob formas destinadas a facilitar a recuperação da informação. A Análise documentária é, desse modo, considerada uma área teórica e metodológica, que inclui as atividades de indexação, classificação e elaboração de resumos, tendo em vista as diferentes finalidades da recuperação de informação.

O principal produto da Análise documentária, a informação ou representação documentária, pode ser evidenciada, essencialmente, em dois subprodutos: os resumos e os índices. De acordo com Kobashi (1994), o resumo deve apresentar relação de contiguidade e semelhança com o texto original, sendo o mesmo obtido por meio de procedimentos de condensação. O índice é obtido pela utilização de um código comutador, tendo caráter generalizante, embora com a preocupação de evidenciar as semelhanças e diferenças entre os textos (CINTRA et al., 2002). Devido à importância da indexação no contexto deste trabalho, maiores detalhes sobre este processo, bem como sobre as tecnologias a ele associadas, serão desenvolvidos no capítulo 3.

2.2 Recuperação de Informação

A expressão recuperação de informação foi utilizada primeiramente em 1951, por Calvin Mooers, que a definiu como uma disciplina que *“trata dos aspectos intelectuais da descrição da informação e sua especificação para busca e também de qualquer sistema, técnicas ou máquinas que são empregadas para realizar a operação”* (MOOERS apud FERNEDA 2003, p.11).

A preocupação com os processos de recuperação de informação não é recente. Assim, pode-se observar que Shera (apud JACOB; SHAW, 1998), considerava que a recuperação é o ponto focal das teorias da Biblioteconomia e da Ciência da Informação. É o objetivo final para o qual todos os esforços deveriam ser direcionados. Além disso, enfatizava que o sucesso da recuperação dependia de desenvolvimentos teóricos que envolvessem: (1) uma compreensão da linguagem e dos processos comunicacionais e seus papéis na transmissão do conhecimento e; (2) a compreensão dos padrões de pensamento humano, expressos como recursos, segundo os quais derivam os registros de conhecimento. Essa abordagem também é conhecida como Epistemologia social.

Dessa forma, Shera (apud JACOB; SHAW, 1998) entende que a recuperação de informação envolve quatro perguntas fundamentais:

- Como a linguagem influencia a representação e organização cognitiva da informação nos indivíduos?
- Quais os efeitos que a necessidade de comunicação tem sobre a representação simbólica e/ou organização cognitiva de informação dos indivíduos?
- Quais padrões ou estruturas de organização cognitiva são compartilhados pelos indivíduos ou por grupos de indivíduos?
- Como os mecanismos de recuperação podem ser projetados para facilitar, aumentar ou imitar estes padrões naturais de organização cognitiva?

Outro autor preocupado com os aspectos linguísticos da documentação é Blair (1990), que enfatiza que o processo de representação de documentos para recuperação é, essencialmente, um processo linguístico. Dessa forma, o problema se concentra em analisar como a linguagem é usada ou, ainda, que a representação de documentos (ou informação) pressupõe o uso de uma teoria da linguagem e do significado.

Desviando um pouco do foco em linguística, pode-se citar Robins (2000), para quem a recuperação da informação é uma disciplina preocupada com o processo pelo qual as questões apresentadas a um sistema de informações são comparadas com um estoque de textos (ou imagens, sons, figuras, entre outros). O resultado final, nesse caso, é uma lista de textos constituída de um subconjunto do estoque total. Tal conjunto deverá ter sua relevância ou utilidade avaliadas pelo usuário que empreendeu o questionamento.

Este autor mostra, ainda, como a pesquisa em recuperação de informação foi se modificando, ao longo do tempo, tendo evoluído para uma dicotomia que se polarizava entre uma abordagem centrada no sistema [de recuperação] e uma abordagem centrada nos usuários dos respectivos sistemas. Assim, as propostas relacionadas à pesquisa em recuperação de

informação adquiriram abordagens essencialmente dinâmicas, descritas e representadas por modelos de recuperação interativa de informação. Basicamente, essa interação sugere que o processo de recuperação não é efetivado em uma única sessão, mas sim numa sequência de eventos concatenados no tempo, em que usuários e sistemas interagem dinamicamente. Nesse contexto, a própria necessidade de informação inicial do usuário vai se modificando à medida que documentos são recuperados, e assim, novos eventos de busca vão sendo efetuados (ROBINS, 2000).

Um bom exemplo de modelo de recuperação interativa de informação é descrito por Ingwersen (1996). Este modelo (ou teoria) descreve, de maneira bem detalhada, as relações e interações que envolvem os três principais componentes associados à recuperação de informação: os estoques de informação, os usuários (e suas necessidades de informação) e o sistema propriamente dito (a interface com o usuário, os mecanismos de comparação, as estruturas de armazenamento, entre outros).

Contemporaneamente, a recuperação da informação é objeto de estudo tanto da Ciência da Informação quanto da Ciência da Computação (BAEZA-YATES; RIBEIRO-NETO, 1999). Particularmente, sob o ponto de vista da Ciência da Informação, os procedimentos anteriormente citados, podem ser divididos em dois grandes grupos: a busca e a recuperação de informações. Uma distinção entre esses dois grupos é exposta por Vakkari (1999, p.820):

A recuperação de informação pode ser vista como parte de um amplo processo de busca por informação. Por **busca de informação** entende-se o processo de pesquisar, obter e usar informação para um propósito (por exemplo, produzir a solução para uma tarefa) quando a pessoa não tem suficiente conhecimento anterior. Por **recuperação da informação** entende-se o uso de um sistema de informação para obter informação relevante para um propósito (por exemplo, uma tarefa).

Vakkari (1999) descreve, em detalhes, os diversos tipos de estudos de recuperação da informação e busca por informação. Sua categorização envolve os estudos orientados e não orientados aos sistemas, orientados e não orientados às tarefas, orientados e não orientados aos processos e tipos de objetos observados, quais sejam documentos ou canais. A figura 2, a seguir, resume a categorização proposta.

		System oriented		Non system oriented	
		Documents*	Channels*	Documents*	Channels*
Non task oriented	Non process oriented	Traditional IR study	L&I service user study		Typical user study with multiple channels
	Process oriented	Interaction with texts within a search session (between sessions)	Library use study	Citation study	
Task oriented	Non process oriented			Use of channels and document types for a <i>job</i>	
	Process oriented	Interaction with texts within & between search sessions		Use of channels and document types for a <i>task</i> ----- Use of documents to understand a problem	

* = as objects of study

Figura 2 – Tipos de estudos de busca e recuperação de informação

Fonte: Vakkari, 1999, p. 821

É interessante observar que os tipos descritos anteriormente englobam uma grande variedade de situações sobre a recuperação de informações. A essas situações poderia ser acrescentado, contemporaneamente, a recuperação de informação no contexto da Internet.

Outra caracterização da recuperação de informação, conforme Turtle e Croft (1997), é considerá-la como o campo de estudos preocupados em identificar documentos, em uma dada coleção, que um usuário julgue ser útil ou relevante. Nesse sentido, um sistema de recuperação de informações apresenta três componentes principais: a identificação e representação dos conteúdos dos documentos, a aquisição e representação das necessidades de informação e a especificação da função de comparação que seleciona os documentos relevantes baseados nas representações.

A materialização dos componentes anteriormente citados envolve o desenvolvimento dos, assim chamados, modelos de recuperação. Com base nos trabalhos de Turtle e Croft (1997) e de Crestani e Lalmas (2001), os modelos podem ser classificados em quatro grandes grupos:

Modelos de ajuste exato – que formam a base da maioria dos sistemas comerciais;

Modelo de espaço vetorial – que vê os documentos e as consultas como vetores de um espaço multidimensional, que usa a distância como uma medida de similaridade;

Modelos probabilísticos – que vêem a recuperação como o cálculo da probabilidade da representação de um documento satisfazer uma consulta;

Modelos baseados em lógica – que vêem os documentos e as consultas como conjuntos de proposições e avaliam a similaridade baseada no grau em que uma consulta pode ser inferida de um documento.

Outros modelos de recuperação de informação podem ser encontrados na literatura, envolvendo a utilização de diferentes teorias computacionais ou matemáticas. Uma visão geral sobre tais soluções pode ser encontrada, por exemplo, em Fernald (2003).

Pode-se acrescentar, ainda, que, de acordo com Croft (1988), as pesquisas em recuperação de informação podem ser orientadas por três questões fundamentais:

- (i) Como se deve representar o conteúdo de documentos textuais?
- (ii) Como se deve representar o problema de informação dos usuários?
- (iii) Qual é o modo apropriado de comparar estas representações de modo a identificar os documentos que resolvem o problema de informação?

Os tópicos anteriores devem ser vistos como breves introduções aos fundamentos da recuperação de informações. Contudo, mesmo uma discussão introdutória que envolva recuperação de informação não pode deixar de abordar alguns aspectos relativos ao conceito de relevância. Esta noção, de acordo com Baeza-Yates e Ribeiro-Neto (1999) é central na recuperação de informações.

As principais teorias em que o tema relevância é tratado são: Lógica, Filosofia e Comunicação. Harter (1992), por exemplo, sumariza a teoria da relevância psicológica e estabelece conexão com a Ciência da Informação, com ênfase em possíveis aplicações na recuperação de informações.

Saracevic (2006) elaborou uma extensa revisão de literatura sobre o tema da relevância, focada principalmente em aspectos relativos à formalização do conceito. O referido texto leva em consideração aspectos que envolvem o significado, as teorias, modelos, manifestações, comportamento e efeitos da relevância.

Com relação aos significados atribuídos ao termo relevância, podem ser destacadas as seguintes caracterizações:

- . pode ser descrita como uma relação entre informação ou objetos de informação e estados afetivos e situações (necessidades de informação, intenções, tópicos, problemas, tarefas), baseada em alguma propriedade que reflete uma manifestação de relevância desejada. Ou, resumidamente, a relação entre um objeto de informação (documento) e um contexto (estados afetivos/cognitivos e situações);
- . pode ser vista como uma propriedade ou uma medida;
- . é inferida
- . pode ser descrita em graus

Ainda com relação a este item, Saracevic (2006) descreve o grande desafio a ser respondido pela Ciência da Informação: Como a relevância ocorre? Quem a produz, sob que circunstâncias e como?

O tópico seguinte do texto de Saracevic (2006) envolve considerações sobre o modo como a relevância se manifesta, das quais podem ser destacadas:

- . o ponto de vista dual, em que a relevância é observada ora do lado do usuário, ora do lado do sistema, respectivamente nomeada como subjetiva e objetiva;
- . o ponto de vista de superação da dualidade. Neste caso são citados os modelos de Mizzaro (1998) e Borlund (2003). O primeiro se ocupa em classificar a relevância num espaço de quatro dimensões: recursos de informação, representação do problema do usuário, tempo (interação entre as dimensões) e componentes (tópico, tarefa, contexto). O segundo desenvolve um modelo que envolve a relevância tópica, situacional e cognitiva.

Em resumo, no âmbito da Ciência da Informação, pode-se falar de dois tipos de relevância: a algorítmica (ou do sistema,) e a relevância cognitiva (ou pertinência),

Considerações que envolvem o comportamento de usuários em sua relação com a relevância são abordadas em seguida, pelo mesmo autor. Neste sentido, são efetuadas pesquisas com o intuito de responder às questões: O que faz com que uma informação ou objeto de informação seja relevante? O que as pessoas observam nas informações ou objetos de informação de modo a inferir relevância? Duas abordagens são utilizadas para responder a essas questões. A primeira, ou tópica, analisa o modo como se estabelece uma relação como pertencendo ou não a um tópico. A segunda, ou abordagem de pistas (vestígios, indícios), estuda os critérios que as pessoas utilizam para realizar avaliações de relevância. A primeira abordagem lida com a relevância tópica, enquanto a segunda inclui a relevância cognitiva, a relevância situacional e a relevância afetiva.

São efetuadas, ainda, considerações que envolvem: a dinâmica da relevância (as inferências e critérios se alteram ao longo do tempo, para o mesmo usuário e tarefa e, em caso afirmativo, como?) e a realimentação de relevância (que fatores afetam o processo de realimentação de relevância?). Segundo Saracevic (2006, as conclusões dos estudos efetuados nesta seção devem ser encaradas como hipóteses de trabalho em vez de afirmações sobre a realidade. Deve-se enfatizar que Saracevic (2006) chama a atenção para dois aspectos. O primeiro refere-se à importância dos títulos e resumos dos documentos na produção de maior quantidade de pistas para os usuários. O segundo, à importância da informação visual na produção de indícios para os usuários.

A título de conclusão deste tópico, pode-se citar Chen (2006), para quem, sob o ponto de vista do sistema, nenhum algoritmo de ordenação de relevância pode realizar o trabalho real da recuperação de informações. Para este autor, os usuários, por si mesmos, devem realizar o julgamento de relevância para uma lista de documentos recuperados.

2.3 Representação e linguagem

Nos estudos que envolvem organização e recuperação de informação, podem ser acrescentados, ainda, dois aspectos que não foram devidamente detalhados, embora anteriormente sugeridos. O primeiro diz respeito ao conceito ou processo de representação. O segundo consiste na percepção da importância da linguagem e suas respectivas teorias como aspectos fundamentais a serem utilizados nos processos de representação.

Deve-se observar que a expressão representação é utilizada tanto para o processo em si quanto para os objetos produzidos pela atividade de representar. Tal processo visa estabelecer correspondência sistemática entre um domínio objetivo e um domínio modelador e em capturar ou “re-presentar”, através dos meios do domínio modelador, o objeto, dado ou informação do domínio alvo (JACOB; SHAW, 1998).

Mais especificamente, pode-se falar em representação da informação, como um *“conjunto de elementos descritivos que representam os atributos de um objeto informacional específico”* (BRASCHER; CAFÉ, 2008, p. 5). Este conjunto, de acordo com estas autoras, é o produto do processo de organização da informação.

Também é possível falar em representação das necessidades de informação nos termos descritos, por exemplo, em Corrêa (2003), que descreve a possibilidade teórica de se utilizar mecanismos de representação de situações problemáticas (problemas) para representação de necessidades de informação. Nesse estudo, as características e os fundamentos teóricos de uma metodologia desenvolvida para representar situações problemáticas são confrontados com conceitos e processos associados à busca por informação e, particularmente, à recuperação de informação.

Os aspectos conceituais e teóricos relativos à linguagem são fortemente vinculados aos estudos que envolvem a representação. É utilizando-se da linguagem que são produzidos os diferentes tipos de textos. É a linguagem o suporte principal de estruturação dos objetos informacionais. É a linguagem a “matéria prima” utilizada para operacionalizar os processos organizacionais da informação e do conhecimento, bem como para materializar os produtos resultantes: a representação da informação e a representação do conhecimento (CINTRA et al. 2002).

O ato de organizar [e representar] informação [e conhecimento] pode ser visto, de acordo com Svenonius (2000) como um tipo particular de uso da linguagem. Segundo a autora, uma vantagem dessa perspectiva consiste em utilizar os constructos linguísticos para generalizar e avaliar diferentes métodos de organizar informação, além de permitir uma conceitualização que poderia unificar os até então distintos métodos (catalogação, classificação e indexação).

Sinteticamente, podem ser observadas vinculações entre as teorias lingüísticas e:

- Ciência da Informação de maneira geral (GRACIOSO, 2008; MENDONÇA, 2000);
- Aspectos organizacionais e/ou representacionais da informação e do conhecimento (ALVARENGA, 2003; BRASCHER; CAFÉ, 2008; GOMEZ, 1993; HJORLAND, 2007; JACOB; SHAW, 1998; KOBASHI, 2007; LARA, 2004; NOVELLINO, 1998; SVENONIUS, 2000);
- Sistemas de recuperação de informação (BLAIR, 1990, 1992, 2002; GOMEZ, 2004).

As próximas subseções discorrerão, de maneira resumida, sobre as vinculações acima elencadas.

2.3.1 Linguística e Ciência da Informação

Segundo Mendonça (2000), a Linguística tem como objeto de estudo a linguagem oral, escrita e representada. Assim, com a contribuição da Linguística e considerando-se a função social da Ciência da informação, através da comunicação da informação científica, surgem questões prioritárias para o tratamento documental. A mesma autora considera, ainda, que a ligação entre a Linguística e a Ciência da informação se materializa através da análise documentária.

Deve-se acrescentar que a concepção de linguagem tem se modificado com o passar do tempo. Assim, conforme Cintra et al. (2002), até o século XVIII predominou uma visão teológica e fortemente influenciada pelas regras universais de sua lógica. Por outro lado, uma visão historicista, que concebia a linguagem como um processo em evolução, através dos tempos, marcou o século XIX. Modernamente, predominam as visões da linguagem como um sistema em funcionamento.

Mais detalhadamente, mesmo a partir da concepção de linguagem como um sistema, é possível observar que há diferentes pontos de vista nos estudos sobre o objeto. Assim, segundo Pietroforte e Lopes (2003, p.113) pode-se formular a discussão sobre os estudos da linguagem nos termos de uma pergunta: *“Devemos tomar a segmentação do mundo em classes como qualquer coisa da ordem do ‘já pronto’ ou do ‘construído’?”* Ou, segundo estes autores e reformulando a pergunta, podemos questionar se as categorias que estruturam o mundo são algo previamente construído nas próprias coisas ou se elas dependem das diferentes maneiras pelas quais se olha para o próprio mundo?

A resposta escolhida direcionará o desenvolvimento de diferentes teorias. No primeiro caso (a ordem do “já pronto”), a teoria linguística terá por base um referente externo à linguagem. Ou seja, supõe-se que o homem tem acesso direto às coisas do mundo, independentemente de quaisquer filtros *“interpostos pela sua inserção sócio-histórica ou cultural”* (PIETROFOTE; LOPES, 2003, p. 113). A segunda hipótese (a ordem do “construído”) direciona a teoria de modo a privilegiar os diferentes modos de observar as coisas. Dessa forma, é concedida prioridade ao ponto de vista, em vez de privilegiar o objeto. Esta abordagem implica, por exemplo, que dois observadores diferentes não vêem exatamente o mesmo mundo (PIETROFOTE; LOPES, 2003).

Os modos de direcionar os estudos linguísticos revertem em diferentes estruturas dos campos a eles subordinados e, dessa forma, serão afetadas as abordagens eventualmente tomadas como marcos teóricos para os estudos de informação. Um desses campos, de vital importância para os estudos da Ciência da informação, é a semântica.

Essa importância foi enfatizada por Van Rijsbergen (1986), ao afirmar que as bases fundamentais de todo o trabalho anteriormente realizado, incluindo o dele mesmo, estavam erradas, pois tal trabalho se baseou na pressuposição de que uma noção formal de significado, na resolução dos problemas de recuperação [e muito provavelmente de representação] da informação, não era necessária

A semântica pode ser definida como o estudo sistemático da produção de sentido nas línguas naturais (PIETROFORTE; LOPES 2003). Outra maneira de caracterizar a semântica consiste em dizer que ela busca descrever o significado das palavras e das sentenças (OLIVEIRA, 2000). Essa autora acrescenta, ainda, que a problemática do significado ultrapassa as fronteiras da linguística, uma vez que ela está fortemente associada à questão do conhecimento.

Entre as diversas semânticas existentes Oliveira (2000) destaca: a semântica formal, a semântica da enunciação e a semântica cognitiva.

A semântica formal é uma legítima representante do paradigma linguístico do referente (da ordem do “já pronto”). Fortemente influenciada por e influenciando a lógica formal, para esta semântica a relação da linguagem com o mundo é fundamental. Desse modo, é dada ênfase aos aspectos objetivos do significado, ou seja, àqueles que estão disponíveis para serem inspecionados por qualquer um.

Utiliza-se a semântica formal para falar de indivíduos, fatos, eventos e propriedades, descritos como externos à própria língua. Assim, o significado é entendido como uma relação que se estabelece entre a linguagem e aquilo sobre o qual a linguagem fala, ou seja, um mundo que pode ser tomado como o mundo real, ficcional ou hipotético (MULLER; VIOTTI, 2003).

Para a semântica da enunciação, a linguagem constitui o mundo e, dessa forma, não é possível sair fora dela. Neste caso, a referência é considerada uma ilusão, criada pela própria linguagem. Esta, por sua vez, é vista como um jogo de argumentações, em que a fala é utilizada para construir um mundo e, a partir dele, tentar convencer um eventual interlocutor, sobre a nossa verdade (OLIVEIRA, 2000).

A semântica cognitiva considera fundamental, para os estudos linguísticos, a questão do significado. Contudo, para tal semântica, ele não é consequência da relação entre a linguagem e mundo. Nesse caso, o significado é consequência da cognição como um todo e não um fenômeno prioritariamente linguístico. Essa abordagem considera a metáfora como um de seus mecanismos fundamentais, sendo ela definida como um mapa, um conjunto de

correspondências matemáticas, entre um domínio da experiência e outro domínio (OLIVEIRA, 2000).

Foge aos objetivos deste trabalho um exame mais detalhado das teorias que envolvem o segundo modo de ver a linguagem (da ordem do “construído”). Maiores detalhes sobre estas teorias e sua influência na Ciência da informação podem ser vistos, por exemplo, em Gomez (1993), Novellino (1998) e em Gracioso (2008).

A título de conclusão deste tópico, podem ser citados Jacob e Shaw (1998) que, enfatizando o papel da pesquisa sociocognitiva na Ciência da Informação, afirmam que ela poderá apoiar o objetivo subjacente à recuperação eficiente e efetiva de informação. Este apoio facilitaria o desenvolvimento de linguagens de representação e estruturas de organização que reflitam as necessidades de jogos de linguagem específicos, ou de domínios de conhecimento específicos ou comunidades linguísticas. E, para o alcance deste objetivo, será necessário investigar os aspectos linguísticos, comunicacionais e organizacionais da representação.

2.3.2 Linguagem, representação, informação e conhecimento

Uma vez aceitas e compreendidas as relações entre informação e conhecimento, fica claro que as discussões que envolvem processos e soluções que se utilizam destes conceitos, a princípio isoladamente, tendem a produzir reflexos e efeitos sobre o outro. Assim, por exemplo, em Gomez (1993) há uma discussão sobre o *locus* do conhecimento na sociedade ocidental. Assim, com base nessa autora,

Pode-se dizer que o conhecimento, enquanto relação do pensamento com o real, muda três vezes de lugar no quadro da cultura ocidental.

Em primeiro momento, a excelência do conhecer não passa pelo representar. O solo do conhecimento é ontológico; a vizinhança do homem com o mundo, da ordem do ser, precede e legitima toda vinculação predicativa, na ordem do juízo.

Em um segundo momento, conhecer é representar, e o mundo só é enquanto é representado. O solo do conhecimento é a consciência.

Em um terceiro momento, o representado, produto do conhecimento em sua investidura semiótica, manifesta-se como autônomo, independente do sujeito e do objeto do conhecimento. Nesse momento, o solo do conhecimento é a linguagem, sistema de significados ou matéria sinalética. A questão do acesso aos objetos e à reflexão acerca das condições da

experiência são substituídos pela questão do *fatum* da linguagem e pela reflexão acerca de suas condições de interpretação. (Gomez, 1993, p. 219)

Ou seja, a linguagem é um ponto focal sobre o qual deverão se basear os estudos de organização e representação da informação e do conhecimento.

Para Gomez (2004), a relação entre a linguagem, a comunicação e a informação pode ser resumida em duas grandes linhas de questionamento: (1) a linguagem como dimensão dos dispositivos de tratamento da informação e (2) a linguagem como dimensão das práticas e ações de informação dos sujeitos e das organizações. A autora conclui que o estudo das diferentes funções da linguagem assume grande importância para uma melhor compreensão das ações de informação e para a concepção de componentes de busca e recuperação de informações.

Observando-se a primeira linha de questionamento anteriormente delineada pode-se exemplificar a utilização da linguagem, ou da ciência que lhe dá suporte – a Linguística – como dimensão em pelo menos dois contextos de aplicação, que serão detalhados em seguida: os sistemas de representação do conhecimento e as linguagens documentárias.

A utilização da Linguística no contexto das soluções que objetivam construir soluções computacionais para a representação do conhecimento é baseada, principalmente, nas teorias e conceitos da semântica. Nessa perspectiva, Hjørland (2007), em artigo de revisão, descreve as relações entre a semântica e a organização do conhecimento. Deve-se acrescentar que, embora não haja uma definição explícita de semântica no artigo citado, entende-se que o autor se refere à semântica formal, nos termos detalhados anteriormente, embora em alguns trechos de seu artigo, pareçam existir referências a outros tipos de semântica possíveis.

Em um primeiro momento, Hjørland (2007) afirma que um tesouro é um tipo de ferramenta semântica, assim como os sistemas de organização do conhecimento descritos por Hodge (apud HJORLAND, 2007).

O autor chama a atenção para o fato de que diferentes abordagens na organização do conhecimento refletem diferentes abordagens da semântica. A abordagem da análise de facetas, por exemplo, tende a basear-se em relações semânticas a priori, estruturadas na aplicação de princípios lógicos. Já a tradição da recuperação de informação vê as relações semânticas como relações estatísticas. A abordagem orientada ao usuário ou cognitiva tende a substituir a garantia literária por estudos de usuários e, dessa forma, baseia as relações semânticas nos usuários. Finalmente, a abordagem bibliométrica considera que os

documentos estão semanticamente relacionados se eles citam um ao outro, são co-citados ou estão bibliograficamente unidos.

Hjorland (2007) efetua alguns questionamentos associados à análise semântica, tais como: que tipo de informação é necessária para determinar as relações semânticas entre dois termos A e B? Que tipo de garantia literária (ou outro tipo de garantia) é necessária para identificar relações semânticas e classificar conceitos? E conclui sobre a necessidade de uma teoria semântica do significado das palavras sob a forma de práticas tipificadas, que poderiam, assim, ser utilizadas pelos buscadores de informação de forma a incluir ou excluir certos documentos. Ele enfatiza, ainda, a necessidade de identificar as relações semânticas na literatura do assunto em questão, em vez de identificá-las por meio de estudos psicológicos.

As colocações de Hjorland (2007), com respeito ao surgimento e estruturação de relações semânticas, podem ser resumidas em:

- diferentes interesses humanos conduzem à observação de diferentes propriedades nos objetos e, conseqüentemente, ao desenvolvimento de relações semânticas que refletem tais interesses;
- as estruturas semânticas dependem de nossas teorias e visões de mundo e, dessa forma, os conceitos não são inerentes à mente ou a linguagens específicas;
- modos conflitantes de definir os conceitos demandam a negociação entre alegações distintas em diferentes textos e direcionam à seleção daquele com maior grau de autoridade cognitiva;
- visões conflitantes de relações semânticas devem estar disponíveis e, dessa forma, os sistemas de organização do conhecimento devem ser desenvolvidos baseados na negociação entre as diferentes visões mais ou menos visíveis na literatura a ser representada;
- relações semânticas são relações entre conceitos, significados ou sentidos. Não devem ser confundidas com relações entre termos, palavras, expressões ou signos utilizados para expressar os conceitos;
- as relações tipicamente utilizadas nas ferramentas semânticas podem ser descritas como ontológicas, uma vez que derivam dos modelos ontológicos produzidos pelos pesquisadores de diferentes disciplinas. Por outro lado, uma disciplina pode ser vista como um conceito social, englobando as pessoas e diferentes organizações envolvidas, apresentando relações internas de citação muito fortes, quando comparadas com outras disciplinas. Nas duas situações são produzidos diferentes tipos de relações semânticas (ontológicas em sua natureza ou sociais, como resultado das práticas de citação) e,

cabe ao cientista da informação prover ferramentas semânticas baseadas em ambos os tipos de relações.

As linguagens documentárias compõem o segundo contexto em que a linguagem pode ser vista como uma dimensão dos dispositivos de tratamento da informação.

Conforme descrito anteriormente, os índices e os resumos são os produtos advindos dos processos de análise documentária. Os índices objetivam instalar uma nova ordem de sentido, neste caso, mediada por um instrumento comutador: a linguagem documentária (KOBASHI, 1994).

Uma linguagem documentária é construída visando-se, principalmente, a indexação, o armazenamento e a recuperação de informações e, dessa forma, acaba por se constituir em um sistema simbólico destinado a traduzir os conteúdos dos objetos informacionais (documentos) (CINTRA et al., 2002) de forma padronizada. Sendo objeto de construção, uma linguagem documentária se reveste das características de uma linguagem artificial.

A construção de tais linguagens implica o desenvolvimento de diferenciação ou categorização de objetos linguísticos. Assim, uma palavra consiste em uma entidade linguística utilizada corriqueiramente nos processos comunicacionais. Um termo pode ser descrito como uma designação correspondente a um conceito em uma linguagem de especialidade). Ou seja, sua utilização pressupõe um consenso entre os componentes da comunidade linguística considerada. Um descritor pode ser visto como uma palavra utilizada com o propósito específico de representar um conceito (objeto, processo) no contexto de uma linguagem documentária que, por sua vez, é construída para atuar em um contexto delimitado (LARA, 2004).

Pode-se acrescentar que, de acordo com Cintra et al. (2002), as linguagens documentárias são usadas para representar o conteúdo dos textos, e não os textos em si mesmos. Neste caso, a função de representação tem natureza eminentemente referencial. Ou seja, as unidades de linguagem documentária devem ser utilizadas como índices dos assuntos contidos nos textos, não tendo a intenção de substituí-los.

Os elementos básicos que compõem uma linguagem documentária são: (i) o léxico – trata-se de uma lista de elementos descritores filtrados e depurados; (ii) uma rede paradigmática – que objetiva traduzir as relações essenciais entre os descritores. Com caráter lógico-semântico, organiza os descritores sob a forma de uma classificação; e (iii) uma rede sintagmática – que objetiva expressar as relações contingentes entre os descritores. Relações estas que são válidas no contexto em que aparecem (GARDIN apud CINTRA et al., 2002).

De acordo com Cintra et al.(2002), as linguagens documentárias mais conhecidas são os tesouros e os sistemas de classificação bibliográfica. O grau de reprodução das relações presentes nas linguagens naturais e no universo de conhecimento que se pretende cobrir é que define as principais diferenças entre estes dois tipos. As autoras citadas consideram que as linguagens documentárias mais consistentes para a representação de informação devem ter um vocabulário que integre elementos da linguagem de especialidade e das terminologias bem como da linguagem natural, visto ser esta a linguagem dos usuários.

3. Indexação Automática

“Renda-se, como eu me rendi. Mergulhe no que você não conhece como eu mergulhei. Não se preocupe em entender, viver ultrapassa qualquer entendimento”.

Clarice Lispector

3.1 Conceitos e princípios

Antes de proceder à conceituação de indexação, deve-se chamar a atenção para o fato de existirem diferentes abordagens sobre o tema. Conforme Silva e Fujita (2004), podem-se observar duas grandes correntes intelectuais, relativas ao modo de conceber os processos e procedimentos de indexação. A primeira, de influência francesa, considera a indexação como um dos processos da Análise documentária, anteriormente definida. A segunda, de influência inglesa, considera a indexação com um processo composto de diferentes etapas. Nessa abordagem, a análise de assuntos é vista como a etapa inicial da indexação. Deve-se observar, ainda, que a abordagem da análise documentária pressupõe o estudo e desenvolvimento das linguagens documentárias, conforme descrito na seção 2.3.2.

Independentemente da concepção adotada, a indexação configura-se como “*o problema fundamental, bem como o obstáculo mais dispendioso da recuperação da informação*” (FAIRTHORNE apud LANCASTER (2004), p. 18).

Desde os primórdios do armazenamento de informações, surge a necessidade de desenvolver mecanismos que facilitem o acesso aos itens armazenados. Silva e Fujita (2004) descrevem o modo como se desenvolveram soluções para facilitar o acesso, ao longo da história – resumos, catálogos alfabéticos, cabeçalhos de capítulos, índice alfabético de assuntos, inventários de livros, comentários à margem dos livros (feito pelos copistas). As autoras afirmam que esse desenvolvimento levou ao surgimento dos primeiros índices, criados para indexar a Bíblia inglesa. Ainda, segundo as autoras, a difusão desses procedimentos levou ao surgimento, na Alemanha, da ideia de palavras-chave para representar itens de informação. No entanto, sob o ponto de vista atual, tais procedimentos e

respectivos produtos, não se encaixam na visão contemporânea do que devem ser os processos de indexação e os índices propriamente ditos.

Com relação à caracterização da indexação, devem ser diferenciados dois aspectos fundamentais: (i) o que é a indexação e (ii) qual o seu objetivo ou finalidade principal. Nos parágrafos seguintes serão apresentadas as definições de indexação propostas pelos principais autores da área.

Uma primeira caracterização consiste em considerar que o propósito principal do desenvolvimento de índices e de resumos é “*construir representações de documentos publicados numa forma que se preste a sua inclusão em algum tipo de base de dados*” (LANCASTER, 2004, p. 1). Este autor entende que os processos de indexação e de elaboração de resumos são fases que compõem a “entrada” de um sistema de recuperação de informações. Afirmar, ainda, que essas atividades estão intimamente relacionadas, pois ambas dizem respeito ao preparo de representações do conteúdo temático dos documentos. Acrescenta, ainda, que a diferenciação entre a indexação e a elaboração de resumos está se tornando cada dia mais difusa e que, se ambas fossem consideradas como atividades complementares, a natureza intrínseca da atividade de indexação deveria sofrer algum tipo de mudança.

Para Lancaster (2004), a indexação é efetuada em duas etapas principais: (i) análise conceitual e (ii) tradução. Na primeira, deve-se decidir do que trata um documento, o seu assunto. A segunda etapa consiste na conversão da análise conceitual para um conjunto de termos de indexação, utilizando-se de “vocabulário do sistema” ou de termos “livres” (extraídos do próprio documento).

Outra maneira de conceituar a indexação consiste em focá-la como um tipo de tratamento técnico de diversos documentos, de forma a permitir a sua utilização. Nesse caso, tal tratamento pode ser dividido em análise da forma e análise do conteúdo. A análise da forma também é conhecida como descrição bibliográfica ou catalogação e a análise de conteúdo envolve os processos de classificação, elaboração de resumos e indexação (LEIVA, 1999). Este autor considera necessário fazer distinção entre duas etapas: a indexação dos documentos e a indexação das perguntas efetuadas pelos usuários.

Assim, para Leiva (1999), na primeira etapa, voltada para a indexação dos documentos, deve ser realizada a leitura do documento para identificar e selecionar os conceitos presentes de maneira explícita no documento. Após a leitura, tanto os conceitos extraídos quanto os atribuídos, se necessário, devem ser armazenados por meio da linguagem natural ou normalizados e controlados com a ajuda de um vocabulário controlado.

A segunda etapa descrita por Leiva (1999), a da indexação da pergunta, refere-se à conversão da pergunta inicial formulada pelo usuário, que o autor nomeia de pergunta documental, para os termos utilizados no sistema de recuperação de informação. Essa conversão deve ser guiada pelos mesmos parâmetros da primeira etapa. Ou seja, a indexação da pergunta deve ser orientada para os termos que indexariam um documento.

Uma terceira visão da indexação, descrita por Mai (2005), considera que o objetivo da indexação é determinar o assunto dos documentos e expressá-los em termos de indexação tais como: descritores, cabeçalhos de assunto, números de chamada e códigos de classificação. Tais termos de indexação tornam possível a recuperação do assunto. Para esse autor, a partir de uma visão externa ao processo de indexação, com frequência assume-se, implicitamente, que uma simples análise permite estabelecer o assunto de um documento. Considera, ainda, que uma abordagem emergente para a indexação clama para que os indexadores analisem os documentos tendo as necessidades de informação dos usuários em mente. Sustenta, tal autor, que o assunto dos documentos é dependente do contexto e não algo facilmente identificável.

A figura a seguir resume a importância da indexação. As setas tracejadas mostram algumas teorias, modelos e campos de estudos que influenciam o desenvolvimento de técnicas de indexação. As setas contínuas mostram a influência da indexação, por meio de processos específicos (Análise da informação e Medidas de similaridade), nos estudos sobre visualização. As setas cheias indicam sua interação com campos de conhecimentos associados tanto à Ciência da Informação quanto à Ciência da Computação.

3.2 Aspectos teóricos e características da indexação

Alem da importância da indexação, anteriormente descrita, as conceituações e caracterizações dos processos de indexação deixam claro que ela não é uma tarefa simples ou trivial.

Embora existam diferentes abordagens teóricas, podem ser observadas várias convergências de idéias entre os diferentes autores. A título de exemplo, pode-se verificar um alinhamento entre a opinião de Lancaster (2004), de que ao se proceder à indexação, deve-se indexar as idéias do autor do texto, e não suas palavras; e a colocação de Leiva (1999) que prescreve uma “leitura vertical” do documento, em que os conceitos implícitos ao mesmo devem ser identificados pelo leitor-indexador.

Por outro lado, deve-se chamar a atenção para a dificuldade de estabelecer relação clara entre os conceitos humanos ou idéias e as palavras (WESSEL, 1975)

Diferentes aspectos podem ser utilizados para caracterizar a indexação. Assim, para Lancaster (2004), o modo como é realizada a segunda etapa da indexação - a conversão da

análise conceitual para um conjunto de termos autorizados pelo sistema - determina um tipo de caracterização. Nesse sentido, é possível distinguir entre indexação por extração e indexação por atribuição. Na primeira, utilizam-se como termos para representar seu conteúdo temático, as palavras ou expressões que ocorram no documento. Na segunda, a atribuição de termos ao documento é feita a partir de uma fonte que não é o próprio documento. Nesse caso, os termos podem ser escolhidos pelo indexador ou, como é mais frequentemente realizado, utiliza-se de alguma tipo de vocabulário controlado.

Um vocabulário controlado é, essencialmente, uma lista de termos autorizados que inclui algum tipo de estrutura ou ordem. Existem três tipos básicos de vocabulários controlados: esquemas de classificação bibliográfica, listas de cabeçalhos de assuntos e tesouros (LANCASTER, 2004). Os tesouros e classificações compõem o que os pesquisadores de linha francesa costumam nomear de linguagens documentárias, conforme descrito em Cintra et al. (2002).

Outra maneira de caracterizar a indexação é descrita por Leiva (1999), com a ajuda de quatro características básicas. A primeira delas é a exaustividade, que pode ser caracterizada como o ato de selecionar todos os conceitos essenciais implícitos e explícitos no documento e na pergunta documental.

A segunda característica refere-se a estimar a qualidade da seleção dos termos que equivalem aos conceitos presentes nos documentos e é chamada de especificidade.

A correção da indexação, terceira característica, corresponde à ausência de erros no processo de indexação. Tais erros podem ocorrer por omissão de um termo e/ou inclusão de termos desnecessários.

A última característica refere-se à consistência da indexação. Ela pode ser avaliada com relação a um único indexador ou a vários. O primeiro caso refere-se à situação em que um profissional indexa um mesmo documento em momentos temporais diferentes. E o segundo, quando vários profissionais indexam um mesmo documento de maneira diferente.

3.3 Limites e dificuldades da indexação

Uma vez reconhecida a importância da indexação, devem ser observadas as dificuldades associadas à sua operacionalização.

A descrição dos limites e dificuldades da indexação pode ser feita com base em diferentes pressupostos teóricos ou formais. Wessel (1975), por exemplo, enfatiza os limites da indexação fazendo analogia com os aspectos formais da construção de teorias e linguagens formais, ou seja, baseado na lógica formal. Afirma esse autor ser virtualmente impossível estabelecer todo o conjunto de regras linguísticas necessárias a uma análise textual precisa para o estabelecimento ou escolha dos termos a serem utilizados na indexação. Ou seja, Wessel (1975) considera que por mais preciso que seja o procedimento utilizado para estabelecer tais regras, o conjunto resultante sempre será incompleto. O autor afirma, ainda, que tal conclusão poderia ser antecipada a partir da observação do teorema de incompletude de Kurt Gödel³.

Independentemente de maiores formalismos ou analogias, uma coisa pode ser afirmada: indexar não é uma tarefa fácil. Não é uma tarefa trivial estabelecer a possível utilidade de um índice. A utilidade de um objeto informacional não é absoluta e o seu julgamento é repleto de subjetividade (LIPETZ, 1988).

A maioria dos autores pesquisados é categórica em afirmar as dificuldades inerentes aos procedimentos de indexação (LANCASTER, 2004; LEIVA, 1999; MAI, 2005). Tais procedimentos implicam uma tomada de decisão não apenas quanto ao assunto do documento, mas também quanto aos motivos pelos quais um documento poderia interessar a um determinado grupo de usuários. Assim, não é possível determinar um conjunto correto de termos de indexação para documento algum, pois diferentes usuários podem estar interessados no mesmo documento por razões diferentes (LANCASTER, 2004).

Outra dificuldade, descrita por Mai (2001), refere-se à impossibilidade de determinar todas as idéias e significados associados a um documento, pois sempre existirão idéias e significados potenciais que diferentes usuários, em momentos distintos, poderão descobrir no mesmo.

Alem disso, Moens et al. (apud LANCASTER, 2004), consideram que um texto não possui uma atinência intrínseca. Ao contrário, diferentes significados, de acordo com o uso particular que uma dada pessoa pode fazer da atinência em um dado momento. Desse modo, a inconsistência das representações documentárias não poderá ser solucionada pela descoberta das regras implícitas seguidas por indexadores. Ao contrário, autores como Frohmann (1990)

³ Em termos não formais, podemos afirmar que o teorema de incompletude de Kurt Gödel demonstra a impossibilidade de criar uma formalização que garanta a inexistência de inconsistências em uma teoria, independentemente da quantidade e qualidade de axiomas (ou regras fundamentais...) escolhidas para fundamentar a formalização e, conseqüentemente, a teoria. Wessel (1975) faz uma analogia entre a impossibilidade descrita por Gödel e a de estabelecer todas as regras linguísticas necessárias. Descrições menos formais do teorema de Gödel podem ser vistas em Teixeira (1998) e Leão et al. (1987).

e Cooper (1979), (apud KOBASHI, 2004) propuseram, enfaticamente, a necessidade de trocar os procedimentos empíricos por métodos e regras precisas que possam ser assimiladas e compartilhadas pelos indexadores dos diferentes sistemas de informação.

Alem das dificuldades anteriormente citadas, existem críticas quanto aos procedimentos e aos produtos resultantes da indexação. Uma dessas críticas foi feita por Weinberg (1988), que se empenha em demonstrar que a indexação limitada à representação da atinência do documento tem pouca utilidade para os pesquisadores. Para esta autora, os pesquisadores, ao buscarem por informação, objetivam encontrar idéias novas, pontos de vista inéditos ou aspectos ainda não observados. Ou seja, para eles interessam as eventuais novidades que um determinado documento contém, quando comparado com o corpo de documentos da área de pesquisa a que ele pertence. Mas, para a autora, tais aspectos não são contemplados pela indexação, pois ela é direcionada para a atinência ou sobre o que trata determinado documento. A autora sustenta que os novatos estão interessados em documentos sobre um tópico ou assunto, mas, os pesquisadores com maior experiência se interessam por documentos que possuem informações novas sobre o assunto que já lhes é familiar.

Os argumentos de Weinberg (1988) são consistentes, a despeito de seu artigo não apresentar evidências empíricas para sustentar suas afirmações. Considera-se pouco provável que pesquisadores experientes e novatos compartilhem o mesmo tipo de interesse sobre um determinado assunto. No entanto, não se pode ignorar que cada texto apresenta um todo de sentido que deve ser respeitado na indexação, como propõem as teorias contemporâneas sobre o tratamento da informação textual (KOBASHI, 1994).

Com o objetivo de melhorar a qualidade dos índices, foram desenvolvidas várias abordagens que modificam o procedimento de básico da indexação, ou seja, a atividade de definir e atribuir índices a um documento. Lancaster (2004) chama estas abordagens de dispositivos de precisão da linguagem de indexação, pois eles possibilitam que se aumente a precisão durante o processo de busca em uma base de dados. São eles: indexação ponderada, o estabelecimento de elos entre termos, os indicadores de função e o subcabeçalhos.

A indexação ponderada consiste em atribuir um valor numérico a um termo, de modo que este valor reflita, na opinião do indexador, a importância desse termo em um dado documento. Quanto mais dominante for o assunto ou maiores detalhes o documento contiver a respeito do assunto, maior deverá ser o peso.

Com o objetivo de evitar recuperações indesejáveis, causadas por falsas associações entre termos, pode-se estabelecer elos entre os termos de indexação. Neste caso, é criada uma segmentação do documento original em diversos subdocumentos, com cada um deles se

referindo a um assunto separado, embora os assuntos de cada um estejam intimamente relacionados. A partir disso, são estabelecidos termos de indexação para cada subdocumento, de modo que os mesmos acabem guardando uma relação direta entre si.

O desenvolvimento dos elos entre os termos pode não ser suficiente para garantir a eficácia da recuperação devido ao estabelecimento de possíveis relações incorretas entre os termos e a ambigüidade resultante. O método utilizado para eliminar tais situações consiste em empregar indicadores de função ou indicadores relacionais, que são códigos utilizados para tornar explícitas as relações entre os termos. Segundo Lancaster (2004), os elos e os indicadores de função foram introduzidos em sistemas de recuperação de maneira simultânea, no início da década de 60. Ainda, segundo este autor, estas técnicas não gozaram de estima por muito tempo, devido ao alto custo de sua aplicação e às dificuldades em aplicá-las com coerência.

3.4 Automação da indexação

A utilização maciça dos computadores como meio para o armazenamento dos estoques de informação e como ferramenta para recuperação de informação, acabou por influenciar os procedimentos executados pelos seres humanos. Assim, parece natural que os computadores começassem a ser utilizados para reproduzir ou executar as atividades antes exercidas pelos seres humanos neste contexto.

Dessa forma, a atividade de indexação passou a ser alvo dos mais diversos tipos de experimentos objetivando realizá-la de forma automatizada. Os primeiros experimentos, de acordo com Lancaster (2004), tiveram início na década de 1950, com os trabalhos pioneiros de Luhn e de Baxendale. Neste caso, os programas desenvolvidos eram utilizados para realizar contagens de frequência de ocorrência de palavras, eliminando da lista aquelas que não possuíam caráter informacional, tais como artigos, preposições e conjunções, entre outras. As palavras com maior grau de frequência eram as candidatas naturais a termos de indexação do documento.

Outra série de experimentos envolvendo indexação automática foi desenvolvida, na década de 1960, com as técnicas KWIC (*Keyword-in-context*) e KWOC (*Keyword-out-of-Context*).

De acordo com Borko e Bernier (1978), o índice KWIC - indexação pela palavra-chave no contexto - surgiu como um novo método de indexação automatizada, aplicada a documentos técnicos, baseando-se nas palavras significativas dos títulos dos documentos. Este índice se caracteriza pelo uso da linguagem natural e, dessa forma, não há um controle de termos significativos e nem identificação de sinônimos. Para Lancaster (2004), trata-se de um método barato de obter acesso temático ao conteúdo de uma coleção, desde que os títulos sejam bons indicadores do conteúdo. Ainda segundo este autor, o índice KWOC - palavra fora do contexto - é similar ao KWIC, exceto pelo fato de que as palavras-chave, no caso do KWOC, tornam-se pontos de acesso, mas não se repetem no título do documento.

A utilização de computadores na indexação pode, de acordo com Leiva (1999), se dar de três formas. Na primeira, nomeada de indexação assistida por computador, estes são utilizados para armazenar os termos de indexação extraídos por um indexador humano no processo de análise conceitual. Neste caso, o sistema funciona como um auxiliar do indexador humano, proporcionando acesso a notas explicativas sobre os termos e, eventualmente, a documentos anteriormente indexados para comprovação de quaisquer aspectos relacionados.

A segunda forma, chamada por Leiva (1999) de indexação semi-automática, refere-se ao uso de sistemas que analisam os documentos de forma automática, mas a decisão sobre os termos de indexação a serem utilizados é validada por um indexador.

Na terceira forma, indexação automática, os sistemas possuem total autonomia para definir e utilizar os termos de indexação de um determinado documento.

O uso da indexação automática é motivo de grandes debates entre os pesquisadores, divididos entre os que são favoráveis e os que são contra a sua utilização.

Alguns motivos contra a indexação automática são: (i) a incapacidade de a automação identificar os conceitos implícitos de um texto, ao contrário dos indexadores humanos; (ii) a limitação imposta aos diferentes tipos de técnicas de indexação (estatísticas ou semânticas), pelas lacunas existentes na compreensão do processo de indexação (LEIVA, 1999).

A favor da indexação automática pode-se citar: (i) a unificação dos procedimentos de indexação, impossível de ser realizada quando a indexação é realizada por diferentes indexadores com distintos e variáveis critérios; (ii) a possibilidade de obtenção de bons resultados tanto na exaustividade quanto na precisão da recuperação de informação, quando comparada com diferentes especialistas (BORDON y PAZIENZA apud LEIVA, 1999); (iii) a diminuição dos erros de indexação acaba repercutindo positivamente nas bases de dados no momento da recuperação de informação; (iv) maior objetividade, uma vez que se aplicam sempre os mesmos parâmetros. Isto acaba evitando a inconsistência produzida por um mesmo

indexador ou aquela provocada por diferentes profissionais ao analisar um mesmo documento (LEIVA, 1999).

A quantidade de experimentos computacionais envolvendo a indexação automática não parou de crescer ao longo do tempo e a diversificar-se quanto aos métodos e técnicas empregadas. Tais eventos levaram ao surgimento de diferentes abordagens de pesquisa e conduziram a indexação automática a interagir com diferentes disciplinas.

Os métodos empregados na indexação automática podem, de acordo com Leiva (1999) ser descritos como: estatísticos, linguísticos, uso de vocabulários controlados e sistemas híbridos.

Os métodos estatísticos evoluíram dos primeiros experimentos anteriormente descritos para a utilização de fórmulas mais complexas, na tentativa de melhorar a precisão dos procedimentos. Neste caso foram desenvolvidas várias soluções envolvendo a utilização de ponderações atribuídas aos termos selecionados. Também foram desenvolvidos métodos que avaliam a frequência de ocorrência de termos em uma dada coleção de documentos. Segundo Leiva (1999), três limitações básicas podem ser associadas aos métodos estatísticos: (i) a incapacidade de reconhecer relações semânticas, tais como sinonímia, anáfora e elipse; (ii) a incapacidade de reconhecer termos compostos e (iii) a incapacidade dos métodos estatísticos de normalizar os termos. Neste caso, por exemplo, os termos documentário e documental seriam computados separadamente.

O desenvolvimento de soluções, que utilizam métodos linguísticos e computacionais, busca automatizar o processamento das linguagem natural. Deve-se observar que as vantagens inerentes à linguagem natural, para a comunicação humana, se convertem em problemas no momento de seu tratamento por meios automáticos (VERDEJO MAILLO apud LEIVA, 1999). Esses sistemas empregam três tipos de análises: morfológica, sintática e semântica. A análise morfológica intenta segmentar a palavra com o objetivo apreender sua gramática e determinar sua estrutura e propriedades. A análise sintática procura detectar as relações sintáticas entre as palavras de uma frase e se incumbe de resolver os problemas identificados pelo analisador morfológico relativamente à ambiguidade gramatical das palavras. Finalmente, a análise semântica se ocupa da significação dos signos linguísticos e de suas combinações. Um analisador semântico objetiva obter o significado das frases de um texto e, como consequência, inferir a compreensão.

A maior parte dos sistemas de indexação automática faz uso de vocabulários controlados. Leiva (1999) descreve vários sistemas desenvolvidos por grandes organizações. Um dos sistemas citados pelo autor foi desenvolvido na NASA, com o objetivo de reduzir

custos e agilizar o processo de indexação. O sistema é composto por três módulos. O terceiro deles comporta o vocabulário controlado propriamente dito, com mais de 100.000 candidatos a termos de indexação. Outro sistema citado por Leiva (1999), chamado de *SAPHIRE*, foi desenvolvido para recuperação de documentos de biomedicina, e utiliza o tesauro desenvolvido pela Biblioteca Nacional de Medicina dos Estados Unidos.

Atualmente, a maioria dos sistemas de indexação automática utiliza pelos menos dois dos métodos anteriormente vistos. São os chamados sistemas híbridos. Algumas das soluções de indexação automática também utilizam outros pressupostos teóricos-formais, tais como: matemática, lógicas não-clássicas e teorias para tratamento de imprecisão e incerteza. Na próxima subseção será apresentada uma solução baseada em pressupostos matemáticos: o modelo do espaço vetorial. No capítulo 4 serão apresentadas soluções que envolvem a utilização de lógicas não-clássicas.

3.5 O modelo do espaço vetorial

Nessa subseção será apresentado um dos mais citados modelos de indexação automática: o modelo do espaço vetorial, ao qual nos referiremos, daqui em diante, por MEV. Tal modelo e o modelo probabilístico são considerados, por Anderson e Perez-Carballo (2001a), os principais modelos de indexação automática já desenvolvidos. Este é o modelo utilizado no sistema de visualização de informações que será usado para testar a hipótese e a proposta formuladas nesta tese.

O modelo do espaço vetorial foi desenvolvido no âmbito do sistema SMART – *System for the Manipulation and Retrieval of Text* (SALTON; MCGILL, 1983).

É interessante observar que tanto o MEV quanto o modelo probabilístico têm seus fundamentos baseados em conceitos matemáticos (vetores e probabilidades). Isto talvez não seja de todo surpreendente se nos lembrarmos do contexto onde tais soluções funcionam: os sistemas informatizados. O funcionamento desses sistemas exige uma descrição precisa de suas regras e procedimentos, exigência inerente à construção e ao funcionamento dos computadores eletrônicos, que são o ambiente final aonde os sistemas se materializarão. Desta forma, parece natural que as soluções e/ou modelos de indexação e recuperação se baseiem

em regras matemáticas ou lógicas, pois estas se encaixam no funcionamento dos computadores.

O MEV descreve a mesma representação básica tanto para os documentos de uma coleção quanto para as consultas efetuadas ao sistema (SMART). Tal representação é baseada no conceito matemático de vetor⁴.

Considere-se, por exemplo, o documento D, com os termos de indexação t_1 , t_2 e t_3 . No MEV atribui-se um número, denominado de peso, para cada termo de indexação. Considerem-se, como exemplos os pesos 0.5, 0.4 e 0.3, respectivamente, para os termos anteriormente descritos. Neste caso, a representação do documento seria:

$$D: (0.5, 0.4, 0.3)$$

Tem-se, aqui, um vetor que pode ser visualizado no espaço tridimensional, representado na figura abaixo (FERNEDA, 2003).

Figura 4 - Representação gráfica (ou vetorial) de um documento com três termos de indexação
Fonte – Ferneda, 2003, p. 28

Caso seja considerado um documento com apenas dois termos de indexação, ter-se-ia um vetor representado no plano cartesiano, como na figura a seguir.

⁴ As grandezas podem ser classificadas em dois tipos básicos: escalares e vetoriais. As grandezas escalares são aquelas que precisam de apenas um atributo para ser descritas. Tal atributo pode ser sua intensidade ou quantidade, descrita por um número. São exemplos de tais grandezas: peso, altura e idade. Outras grandezas exigem, para sua completa descrição, outros atributos além da quantidade, tais como: direção e sentido. São exemplos destas grandezas a velocidade, a força e a aceleração. Tais tipos são denominados de grandezas vetoriais, sendo representadas por um vetor, que pode ter uma representação geométrica. Tal representação pode se dar no plano (duas dimensões), no espaço (três dimensões), ou genericamente, num espaço com um número maior de dimensões (n dimensional). Além da utilização para descrever grandezas físicas, como as anteriormente descritas, o conceito matemático de vetor e sua representação algébrica podem ser utilizados e/ou adaptados para descrever outras situações, como a que foi feita por Salton.

Figura 5 – Representação gráfica (ou vetorial) de um documento com dois termos de indexação
Fonte: Ferneda, 2003, p. 28

Se um documento tiver n termos de indexação, $t_1, t_2, t_3, \dots, t_n$, e com os pesos $w_1, w_2, w_3, \dots, w_n$, respectivamente associados aos termos, não seria possível visualizar tal vetor, uma vez que o mesmo pertenceria a um espaço com n dimensões. Contudo, isto não impede a realização de manipulações matemáticas envolvendo todos os termos do documento. Assim, para esta situação, pode-se representar o documento como:

$$D: (w_1, w_2, w_3, \dots, w_n)$$

Considere-se, agora, uma coleção com k documentos, e um total de n termos de indexação, observando-se todos os termos utilizados na coleção. Neste caso, um termo que não seja encontrado em um dado documento da coleção terá seu peso igual a 0 (zero). Pode-se representar tal situação por meio de uma matriz documentos X termos, conforme descrito por Salton (1989):

	t_1	t_2	t_3	...	t_n
D_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	...	$w_{1,n}$
D_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	...	$w_{2,n}$
D_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$	...	$w_{3,n}$
...	...	...	...	...	...
D_k	$w_{k,1}$	$w_{k,2}$	$w_{k,3}$	...	$w_{k,n}$

Figura 6 – Matriz documentos X pesos
Fonte: Adaptada de Salton (1989)

A abordagem vetorial adotada no MEV permite que outros conceitos, desenvolvidos no âmbito da teoria matemática de vetores, sejam utilizados no modelo. Um desses conceitos é o de distância entre dois vetores. Neste caso, Salton e McGill (1983) consideram que o

cálculo da distância entre dois vetores, que representam dois documentos da coleção, indica o grau de similaridade entre eles. O cálculo da distância entre dois vetores x e y é efetuado pela fórmula (FERNEDA, 2003; SALTON e MCGILL, 1983):

$$\text{sim}(x, y) = \frac{\sum_{i=1}^t (w_{i,x} \times w_{i,y})}{\sqrt{\sum_{i=1}^t (w_{i,x})^2} \times \sqrt{\sum_{i=1}^t (w_{i,y})^2}}$$

Onde $w_{i,x}$ refere-se peso do i -ésimo termo (ou elemento) do vetor x e $w_{i,y}$ refere-se ao peso do i -ésimo termo (ou elemento) do vetor y .

Da mesma forma, o MEV considera que uma expressão de busca pode ser representada vetorialmente, atribuindo-se pesos aos termos utilizados. Seja, por exemplo, a expressão de busca Q , efetuada com as palavras $q_1, q_2, q_3, \dots, q_n$, as quais se atribuem pesos $e_1, e_2, e_3, \dots, e_n$. Neste caso, pode-se representá-la como segue:

$$Q: (e_1, e_2, e_3, \dots, e_n)$$

Assim, da mesma forma que se pode calcular a similaridade entre dois documentos (ou suas representações vetoriais), podem-se calcular o grau de similaridade entre uma consulta e um ou mais documentos da coleção. Este procedimento permite que para uma dada consulta, seja comparada com todos os documentos da coleção e crie, por exemplo, uma lista de documentos (crescente ou decrescente) de acordo com o grau de similaridade entre estes últimos e a expressão de consulta.

Uma vez compreendidos os aspectos básicos do MEV, uma questão continua em aberto: como definir os pesos para os termos de indexação e para os termos de busca?

Várias considerações podem ser efetuadas e várias justificativas utilizadas para definir o cálculo dos pesos. Ao longo do tempo, diferentes abordagens foram construídas, e uma gama de diferentes métodos para cálculo dos pesos podem ser encontrados na literatura (SALTON; MCGIL, 1983; SALTON; BUCKLEY, 1988).

O método mais utilizado para o cálculo dos pesos é baseado na frequência estatística de ocorrência dos termos em duas instâncias diferentes: o documento, individualmente, e a coleção como um todo.

No primeiro caso, efetua-se uma contagem da frequência de ocorrência de cada termo no documento em questão, ou seja, faz-se a contagem pura e simples do número de vezes em

que o termo aparece no documento. Tal valor é denominado *tf* (abreviatura da expressão: *term frequency*).

No segundo caso, verifica-se a quantidade de documentos da coleção em que o termo ocorre, comparando-o com a quantidade total de documentos da coleção. Esta medida é denominada de *idf* (da expressão: *inverse document frequency*). Nesse caso, o cálculo é descrito pela seguinte fórmula:

$$idf_t = N / n_t$$

onde, *N* é o número de documentos da coleção e *n_t* é o número de documentos da coleção que contém pelo menos uma ocorrência do termo *t*. Pode-se observar que, quanto menor o número de documentos em que ocorre o termo *t*, maior é o seu *idf*. Pode ser encontrada uma variante da fórmula anterior, em que o cálculo é efetuado utilizando-se o logaritmo de base 10, como segue:

$$idf_t = \text{Log} (N / n_t)$$

Uma vez definidos os valores das instâncias locais e globais para o termo *t*, seu peso é calculado como:

$$W_{t,d} = tf_{t,d} \times idf_t$$

Estas são as operações básicas efetuadas no sistema SMART para recuperação de documentos a partir de uma expressão de busca dada. Outros procedimentos são efetuados antes da contagem puramente estatística de ocorrência dos termos. De acordo com Ferneda (2003), tais procedimentos são:

1. Identificação e isolamento de cada palavra do texto do documento ou de sua representação (resumo, palavras-chave);
2. Eliminação de palavras com grande freqüência e pouco valor semântico, tais como preposições, artigos, conjunções, etc (*stop-words*);
3. Remoção de afixos (prefixos e sufixos) das palavras restantes, reduzindo-as ao seu radical (processo denominado na língua inglesa de *stemming*);

Após estes passos os radicais formarão os vetores que representam os documentos com os respectivos pesos calculados da forma anteriormente descrita.

Neste capítulo foi efetuada uma descrição sucinta do modelo do espaço vetorial. Apesar de eventuais críticas, o MEV é estudado, avaliado e utilizado até hoje. Esse modelo serve de base para o sistema de visualização utilizado, neste trabalho, para teste das hipóteses de pesquisa elencadas.

4. Indexação automática e lógica paraconsistente

“O contrário de uma verdade profunda não é um erro, mas outra verdade”.

Niels Bohr

“O contrário de uma verdade não é um erro, mas uma verdade contrária”.

Blaise Pascal

4.1 Introdução às lógicas clássica e não-clássicas

Este capítulo apresentará o conteúdo principal desta tese: uma proposta de utilização da Lógica paraconsistente nos processos de indexação automática e as justificativas dessa escolha. Para este fim, serão introduzidos, inicialmente, os conceitos de lógica clássica, lógicas não-clássicas, lógica difusa (*fuzzy*) e Lógica paraconsistente, de interesse para esta pesquisa. Também serão apresentadas aplicações dessas lógicas no contexto da indexação automática.

4.1.1 A Lógica clássica

4.1.1.1 Contexto e conceituação

Muitas são as dificuldades para uma definição precisa do que seja a Lógica. Os livros sobre o assunto apresentam diferenças de abordagem e conceituação. Apesar das

divergências, os autores são, em geral, unânimes em creditar a origem da Lógica aos trabalhos do filósofo grego Aristóteles (MORTARI, 2001; DA COSTA et al., 1999).

A história da lógica pode ser dividida em três períodos: (i) Aristotélico; (ii) Booleano e (iii) Contemporâneo. O período aristotélico compreende as primeiras sistematizações efetuadas pelo filósofo grego Aristóteles, alguns trabalhos desenvolvidos na Idade média e uma relativa ausência de novidades até o início de século XIX. O período Booleano apresenta um grande desenvolvimento da lógica, sua formalização e o desenvolvimento de uma conexão entre ela e a Álgebra. O período contemporâneo compreende o início do século XX até os dias atuais. Este período é responsável pelo crescimento vertiginoso tanto de aspectos teóricos quanto de aplicações da lógica. Surgem também os primeiros trabalhos que descrevem o surgimento de lógicas alternativas à lógica de origem Aristotélica (DA COSTA et al., 1999).

A compreensão efetiva sobre o que é a Lógica exige que sejam considerados dois contextos distintos: o contexto mental e o contexto linguístico.

O contexto mental diz respeito ao pensamento, mais precisamente, uma classe específica de pensamentos, que será definida em seguida: as inferências.

O contexto linguístico refere-se à enunciação ou representação linguística das inferências. Tal representação é denominada de argumento.

A figura a seguir descreve tal contexto e suas respectivas interações.

Figura 7 – Contexto pensamento – inferências X linguagem – argumentos

Fonte: autor deste trabalho

Além dessa contextualização outra se faz necessária. Uma que faça distinção e indique outro tipo de interação: entre a realidade observável e a descrição da realidade, conforme descrito a seguir. Nesse contexto surgem dois domínios: o domínio sensorial, relacionado à recepção de sinais da realidade através dos sentidos. O outro domínio refere-se à descrição ou enunciação da realidade percebida, conforme figura a seguir.

Figura 8 – Contexto realidade X descrição
Fonte: autor deste trabalho

As inferências podem ser descritas como um tipo de atividade mental. Pode ser entendida como um raciocínio, ou seja, uma sequência de passos mentais que permitem chegar a uma afirmação a partir de outras afirmações, no contexto do pensamento. No momento em que se enuncia ou se escreve essa sequência de passos, muda-se o contexto: passa-se do contexto do pensamento para o contexto lingüístico, quando a inferência se transforma num argumento. Nesse sentido, um argumento pode ser descrito como uma coleção de enunciados relacionados entre si (SALMON, 1978). É nesse segundo contexto – no lingüístico e no universo dos argumentos - que atua a Lógica. Os enunciados de um argumento podem ser divididos em dois tipos básicos: os enunciados propriamente ditos (ou afirmações iniciais, evidências) e a conclusão (afirmação final ou resultado da inferência).

A partir dessa contextualização, surgem as diversas conceituações de lógica. Para Salmon (1978, p. 15), por exemplo, A Lógica “*procura estudar o tipo de relação que pode existir entre a evidência e a conclusão*”.

A Lógica tem sido freqüentemente definida como a ciência das leis do pensamento. Mas esta definição não é exata. O pensamento é um dos processos estudados pelos psicólogos. Se o pensamento é qualquer processo mental que se produz na psique das pessoas, nem todo pensamento é objeto de estudo da Lógica. Contudo, pensar e raciocinar é, em grande parte, inferir. Ou seja, uma atividade psicológica que consiste em obter uma conclusão a partir de uma ou mais evidências. Consiste em chegar a certas crenças e opiniões à custa de outras crenças e opiniões (COPI, 1968).

Como os argumentos são representações linguísticas das inferências, alguns autores preferem definir a Lógica associando-a às inferências. Assim, a Lógica pode ser vista como *“a ciência que estuda princípios e métodos de inferência, tendo o objetivo principal de determinar em que condições certas coisas se seguem (são consequência), ou não, de outras”* (MORTARI, 2001).

Entende-se que a conceituação de lógica efetuada nos parágrafos anteriores é suficiente para a compreensão do restante dessa pesquisa. Contudo, outros aspectos fundamentais podem ser observados para uma maior compreensão desse campo de estudo. Esses aspectos podem ser encontrados no Apêndice A.

4.1.1.2 Princípios fundamentais da Lógica clássica

A lógica clássica teve seu desenvolvimento embasado por alguns princípios fundamentais. Alguns autores consideram a existência de quatro princípios (DA COSTA et al., 1999), enquanto outros consideram apenas três (D’OTTAVIANO, 1990). De qualquer forma, três princípios são comuns aos diferentes autores. De acordo com Da Costa et al. (1999, p.13) são eles:

- (1) Princípio da identidade – todo objeto é idêntico a si mesmo;
- (2) Princípio do terceiro excluído – de duas proposições contraditórias (de tal forma que uma é a negação da outra), uma delas é verdadeira;
- (3) Princípio da não-contradição – entre duas proposições contraditórias, uma é falsa.

Outra maneira, menos vinculada ao linguajar formal contemporâneo da lógica, de descrever os princípios (2) e (3), atribuída a Aristóteles, conforme D’Ottaviano e Feitosa (2003, p. 10) é:

- (2) Princípio do terceiro excluído – toda proposição é verdadeira ou falsa, não havendo uma terceira possibilidade
- (3) Princípio da não-contradição – uma proposição não pode ser verdadeira e falsa ao mesmo tempo

A capacidade de estabelecer conclusões a partir de enunciados primitivos fornece à Lógica o potencial para servir de base para o desenvolvimento de teorias diversas, pois as conclusões podem ser obtidas dos princípios básicos admitidos para a teoria. Pode-se observar, ainda, que nesse caso, a teoria será desenvolvida a partir de procedimentos exclusivamente dedutivos, como é o caso da matemática ou das diferentes disciplinas que a compõem. Este é o caso, por exemplo, da geometria euclidiana, anteriormente citada.

4.1.1.3 As Lógicas não-clássicas

Durante o seu desenvolvimento, tanto a Matemática quanto a Lógica se depararam com situações inexplicáveis e paradoxais. Em D’Ottaviano e Feitosa (2003), várias dessas situações são apresentadas. Resumidamente, podem ser consideradas três influências básicas que motivaram o desenvolvimento das lógicas ditas não clássicas: (i) a identificação de certos paradoxos lógico-matemáticos que alimentavam dúvidas quanto à validade dos princípios gerais que davam sustentação a essas disciplinas; (ii) o surgimento de matemáticas “não tradicionais”: um exemplo clássico é o surgimento das geometrias não euclidianas, que foram construídas modificando-se um ou mais dos axiomas fundamentais postulados por Euclides; (iii) a ocorrência de situações do mundo real, bem como de sua descrição linguística, que não se ajustam, perfeitamente, na dicotomia estabelecida pela lógica clássica – verdadeiro ou falso.

As lógicas não clássicas podem ser classificadas, de acordo com Haak (apud D’OTTAVIANO; FEITOSA, 2003) e Da Costa et al. (1999), em complementares à Lógica clássica e alternativas ou rivais à Lógica clássica.

As lógicas complementares não derogam os princípios básicos da lógica clássica e também não questionam sua validade. Nesse caso, são efetuados procedimentos que complementam a lógica clássica permitindo, em geral, um aumento de seu escopo de utilização. Por outro lado, as lógicas alternativas foram estruturadas para substituir a lógica clássica em certos domínios de conhecimento e, para tal, derogam um ou mais de seus princípios fundamentais da mesma. (D’OTTAVIANO; FEITOSA, 2003).

Nas seções seguintes serão apresentados os fundamentos de duas lógicas alternativas à lógica clássica, a lógica difusa e a lógica paraconsistente, tendo em vista suas aplicações no contexto da indexação automática.

4.2 Indexação automática e lógica difusa

4.2.1 Introdução à teoria dos conjuntos difusos e lógica difusa

Muitas situações do mundo real e suas respectivas representações linguísticas não podem ser classificadas ou organizadas com a rigidez imposta pela Lógica clássica.

Um dos exemplos mais citados para descrever tais situações consiste no conceito embutido na palavra alto. Afinal, o que é uma pessoa “alta”? Provavelmente, muitos concordarão que alguém com 1,90 m (metros) de altura é, indubitavelmente, uma pessoa alta. É provável também que outros concordarão que uma pessoa com 1,80 é uma pessoa alta. Contudo, e quanto a uma pessoa com 1,79 m de altura? Sob o ponto de vista da lógica clássica, uma pessoa é ou não é alta. Isto é consequência imediata do fato de que, nesta lógica, existem apenas dois estados lógicos possíveis para uma proposição ou enunciado: verdadeiro ou falso. Esse exemplo, clássico nos textos de lógica difusa, mostra como certas palavras, de uso intensivo nas situações reais apresentam certo grau de ambiguidade ou vagueza, sendo difícil defini-las de modo preciso. Ou seja, pode-se concordar com Bojadziev e Bojadziev (1995, p. xi), para quem a *“realidade é mais ou menos incerta, vaga e ambígua”*.

A situação anterior pode ser descrita de maneira mais formal por meio da teoria matemática clássica de conjuntos. Consideramos um conjunto de elementos descrito pela propriedade: pessoa alta. Nesse caso, uma pessoa com 1,80 m de altura pertence ao referido conjunto, assim como uma pessoa com 1,90 m. Por outro lado, uma pessoa com 1.79 m não pertence ao conjunto das pessoas altas.

Matematicamente, pode-se dizer que existe uma função que descreve se certo elemento pertence ou não a um conjunto dado. Esta função é chamada de função de pertinência dos elementos do conjunto. Seja, por exemplo, um conjunto qualquer A, composto de vários elementos. Considere-se ainda U um conjunto universo, e A um subconjunto de U (U é por vezes denominado de universo de discurso). Seja x um específico elemento de U. A função de pertinência pode ser descrita como:

$$\mu(x) = 1 \text{ se } x \text{ pertence ao conjunto } A \text{ e } 0 \text{ se } x \text{ não pertence ao conjunto } A,$$

Ou, usando linguagem matemática, esta sentença pode ser escrita como:

$$\mu(x) = 1, \text{ se } x \in A$$

$$0, \text{ se } x \notin A.$$

Para o exemplo das pessoas altas, citado, tem-se:

$$\mu(1,90) = \mu(1,80) = 1$$

$$\mu(1,79) = 0$$

Evidentemente, esta maneira de descrever as “pessoas altas” está sujeita às restrições anteriormente descritas. Ou seja, nessa abordagem uma pessoa com 1,79m não é, definitivamente, uma pessoa alta.

A teoria dos conjuntos difusos, originalmente desenvolvida por Lofti Zadeh, pode ser vista como uma generalização da teoria clássica dos conjuntos (BOJADZIEV;BOJADZIEV, 1995). Em um conjunto difuso, sua função de pertinência não se restringe aos valores 0 e 1. Nesse caso, a função pode assumir qualquer valor no intervalo [0,1]. Assim, os elementos de um conjunto difuso possuem valores que estão incluídos em todo o intervalo e não apenas os seus valores extremos, considerando, obviamente, os valores fracionários contidos no referido intervalo.

A título de exemplo, considere-se uma descrição simplificada do conjunto das pessoas altas, anteriormente citado. Seja A o conjunto contendo as alturas das pessoas consideradas altas. Matematicamente pode-se descrever A como:

$$A = \{ 1.80; 1.82; 1.83; 1.84; 1.85; 1.86; 1.87; 1.88; 1.89; 2.00 \}.$$

Agora, considere-se A o conjunto difuso das pessoas altas (a utilização de letras em itálico é muito comum para representar conjuntos difusos). Nesse caso, a descrição dos elementos de A é composta por um par de valores, a altura propriamente dita e um número entre 0 e 1 que indica o valor da função de pertinência de cada altura. Ou seja, por exemplo,

$$A = \{ (1.80; 0.1), (1.82; 0.2), (1.83; 0.3), (1.84; 0.4) (1.85; 0.5), (1.86; 0.6),$$

$$(1.87; 0.7), (1.88; 0.8), (1.89; 0.9), (2.00; 1) \}.$$

Esta representação pode ser informalmente descrita da seguinte forma: uma pessoa com 1,80m é alta com um grau de pertinência 0,1; uma pessoa com 1,85m é alta com grau 0,5

de pertinência; e uma pessoa com 2,00m é alta com total pertinência. É interessante observar que o questionamento efetuado no início da seção quanto ao fato de uma pessoa com 1,79m, ser ou não alta, pode ser resolvido acrescentando-se esta altura ao conjunto A com um valor para a função de pertinência como, por exemplo, 0,09. Assim o conjunto A , difuso, resultante será:

$$A = \{ (1.79; 0.09), (1.80; 0.1), (1.82; 0.2), (1.83; 0.3), (1.84; 0.4), (1.85; 0.5), \\ (1.86; 0.6), (1.87; 0.7), (1.88; 0.8), (1.89; 0.9), (2.00; 1) \}.$$

Nesse caso, uma pessoa com 1,79m seria considerada alta, com um grau de pertinência especificado. Muitas situações semelhantes a esta, descritas por palavras que ensejam interpretações ambíguas, podem ser descritas através de conjuntos difusos, tais como, por exemplo: quente, frio, caro, barato, novo e velho.

Usando linguagem estritamente matemática (BOJADZIEV; BOJADZIEV, 1995), pode-se definir formalmente um conjunto difuso. Seja A um conjunto clássico, subconjunto de um conjunto universo U . O conjunto difuso A pode ser definido como um conjunto de pares ordenados, compostos por x e $\mu_A(x)$, isto é, como uma relação binária, da seguinte forma:

$$A = \{ (x, \mu_A(x)) \mid x \in A, \mu_A(x) \in [0,1] \}$$

onde $\mu_A(x)$ é o grau pertinência, relativa ao elemento x . Tal função especifica, para um dado elemento x de A , o “quanto” ele pertence ao conjunto difuso A .

A teoria clássica de conjuntos apresenta uma grande quantidade de teoremas e propriedades com extensa gama de aplicações. A teoria dos conjuntos difusos possui propriedades semelhantes às da teoria clássica e de outras que esta última não possui. Além disso, pode ser utilizada nas mais variadas aplicações, tais com a descrição ou formalização das instancias linguísticas anteriormente citadas. Deve-se acrescentar que, conforme detalhado em Bojadziev e Bojadziev (1995), é possível estabelecer uma correspondência formal entre a teoria clássica de conjuntos e a lógica clássica. Tal correspondência é chamada, na Matemática, de isomorfismo. Da mesma forma, é possível estabelecer um isomorfismo entre a teoria dos conjuntos difusos e uma nova lógica, chamada de lógica difusa. Assim, surge uma nova teoria lógica com, resumidamente, as seguintes características:

- (i) é uma lógica multivalorada. Ou seja, admite inumeráveis valores entre os estados lógicos verdadeiro (V) e falso (F). Assim, tal lógica derroga o princípio do terceiro excluído;
- (ii) derroga o princípio da não-contradição;
- (iii) fornece um arcabouço teórico forte para o tratamento de situações ou fenômenos incertos, vagos, imprecisos e ambíguos.

Várias aplicações para a lógica difusa podem ser vistas em Bojadziev e Bojadzibev (1995) e em Pedrycz e Gomide (2007). Na próxima seção serão apresentados os seus usos no contexto da recuperação de informações.

4.2.2 Indexação, recuperação de informações e abordagem difusa

A teoria clássica de conjuntos bem como sua correspondente isomórfica, a lógica clássica, é utilizada, implícita ou explicitamente, na grande maioria dos sistemas de recuperação de informação (FERNEDA, 2003). Desta forma, parece natural que tentativas de utilização da teoria dos conjuntos difusos bem como da lógica difusa fossem efetuadas no âmbito dos sistemas de recuperação, como será visto a seguir.

Conforme visto na seção 2.2, entre as diferentes maneiras de descrever os sistemas de recuperação de informação, podem ser considerados dois aspectos básicos: a representação dos itens de informação e a especificação das expressões de consulta formuladas pelos usuários. Entre outros aspectos, os defensores da utilização da abordagem difusa consideram que a teoria dos conjuntos difusos pode ser empregada com sucesso para o tratamento da imprecisão e subjetividade inerentes aos processos de indexação, bem como para gerenciar a vagueza embutida nas consultas formuladas pelos usuários (HERRERA-VIEDMA; PASI, 2003)⁵.

Com efeito, vários experimentos utilizando a lógica difusa e envolvendo a utilização e construção de tesauros, ontologias e citações foram realizados com relativo grau de sucesso

⁵ A literatura que descreve a utilização da teoria dos conjuntos difusos e da lógica difusa no âmbito da recuperação de informação utiliza, em geral, as expressões “conjuntos difusos” e “lógica difusa” sem fazer distinção entre elas, apesar de serem objetos conceituais diferentes. Contudo, dada a característica de isomorfismo existente entre a teoria dos conjuntos (clássica e difusa) e a lógica (clássica e difusa), as expressões podem ser utilizadas, informalmente, como sinônimas. Nesta pesquisa será usada, a expressão “abordagem difusa”.

(OGAWA; MORIDA; BOBAYASHI, 1991; PEDRICZ; GOMIDE, 2007; PEREIRA, 2004; NOMOTO et al., 1990).

No contexto específico da representação de documentos, e particularmente na indexação, podem ser encontrados vários trabalhos que sintetizaremos nos parágrafos seguintes. Com a utilização da teoria clássica de conjuntos pode-se definir a representação dos documentos de uma coleção como um conjunto de termos e sua respectiva função de pertinência. Nessa situação, a função de pertinência assumirá o valor 1 quando o termo pertencer ao documento e 0 quando o termo não pertencer. Formalmente, com base em Bordogna e Pasi (1995), pode-se escrever:

$$R(d) = \{ t \mid t \in T, F(d,t) \neq 0 \}$$

Assim, R representa o conjunto composto por todos os termos t que aparecem na coleção (conjunto T) de documentos D . $F(d,t)$ é a função de pertinência do termo t para o documento d , que assume os valores anteriormente citados.

Outra maneira de definir tal representação é utilizar conjuntos difusos. Agora, a função de pertinência poderá assumir qualquer valor no intervalo $[0,1]$. Nessa situação, o valor 0 indica que o termo não possui significância para o documento e o valor 1 indica total significância para o documento. Formalmente,

$$R(d) = \{ \langle t, \mu_d(t) \rangle \mid t \in T \}$$

onde, nesse caso $F(d,t) = \mu_d(t)$.

Aqui surge a questão de como calcular o valor de $\mu_d(t)$. No modelo proposto por Bordogna e Pasi (1995) o valor atribuído é aquele calculado para um peso, conforme definido no modelo do espaço vetorial, ou seja,

$$F(d,t) = tf_t * idf_t$$

Nesse caso, as autoras calculam o valor normalizado de tf , ou seja, em vez de considerar o cálculo simples do número de ocorrências do termo, este é dividido pelo maior número de ocorrências do termo na coleção. Este procedimento garante que o valor de tf_t será, no máximo, igual a 1.

Outro ponto é considerado pelas autoras para o desenvolvimento da representação difusa dos documentos. Elas consideram que, em geral, os documentos são organizados em partes estruturais, tais como título, autor(es), palavras-chave, resumo e referências. Todas estas partes podem apresentar ocorrências de termos.

Acrescentam as autoras que o papel informativo de cada termo depende de seu significado na parte ou seção em que ocorre. Mais especificamente, afirmam que a simples ocorrência de um termo pode contribuir de modo diferente para o significado que o termo possui em todo o documento, dependendo da semântica da seção na qual o termo está localizado. Além desta dependência semântica, as seções de um documento podem ter diferentes graus de importância para cada usuário. Nesse caso, dois aspectos devem ser levados em consideração no cálculo da significância de um termo. O primeiro diz respeito à ocorrência do termo em diferentes seções do documento e, o segundo, ao fato de o interesse do usuário poder influenciar o cálculo da função $F(d,t)$, aumentando ou reduzindo a contribuição da ocorrência dos termos, dependendo da seção em que o termo ocorre. Dessa forma, o cálculo do grau de significância passa a depender da intervenção do usuário. Nesse caso, os usuários podem atribuir uma importância numérica que será utilizada para enfatizar a função $F_{Si}(d,t)$, que será associada a cada seção do documento.

Para calcular o grau de pertinência de um termo com relação a todo o documento, os graus de pertinência de cada seção são agregados através de uma nova função, configurada pelo usuário (BORDOGNA; PASI, 1995). A figura 9, a seguir, ilustra esse processo.

Figura 9 – Representação difusa de documentos
Fonte: Ferneda, 2003, p. 47

A abordagem desenvolvida foi testada em um sistema de recuperação de informações e avaliada com as medidas tradicionais para tal situação (precisão e cobertura).

Uma segunda utilização da abordagem difusa na indexação de documentos foi efetuada por Molinari e Pasi (1996). Nesse trabalho, as autoras propõem que a indexação de documentos HTML seja efetuada utilizando-se a estrutura sintática dessa linguagem. O documento também é dividido em seções, de acordo com as regras de construção da linguagem. Da mesma forma que na solução de Bordogna e Pasi (1995), esse modelo atribui a cada seção, um grau de importância. Só que, nesse caso, consideram-se as características utilizadas para a construção do documento HTML, ao invés de solicitar a intervenção do usuário. Assim, características como itálico, negrito, tamanho da fonte recebem valores diferentes conforme uma pré-definição das autoras. Novamente, uma função de agregação é utilizada para calcular a importância do termo no documento, a partir de sua importância em cada seção.

Outras utilizações da abordagem difusa podem ser encontradas, com diversos e sofisticados cálculos envolvendo o estabelecimento das funções de pertinência como, por exemplo, em Zadrozny e Nowacka (2009) e em Nowacka, Zadrozny e Kacprzyk. (2008), que serão brevemente descritos em seguida.

O trabalho de Zadrozny e Nowacka (2009) se propõe a ampliar trabalhos anteriores que utilizam a lógica difusa, preocupando-se principalmente com os procedimentos de indexação de documentos e das consultas a serem efetuadas. São realizados testes em várias coleções e comparações de resultados com outros modelos de recuperação de informação. Por outro lado, o trabalho de Nowacka, Zadrozny e Kacprzyk et al. (2008) é voltado para discussões teóricas sobre a utilização da lógica difusa na recuperação de informação, com ênfase no estudo da especificação das funções de pertinência a serem utilizadas no modelo.

A título de conclusão desta seção deve-se chamar a atenção para o fato de que a abordagem difusa associa um único valor ao grau de pertinência ou importância do termo, relativamente a um documento. Na próxima seção será visto outro tipo de abordagem que permite associar dois valores ao grau de importância de um termo para um documento.

4.3 Indexação automática e lógica paraconsistente

4.3.1 Lógica paraconsistente

Observou-se no item 4.1.1.3 que a modificação de um dos princípios básicos da lógica clássica dá origem a um novo tipo de lógica, com reflexos nos métodos que a compõem, entre os quais podem ser citados: o cálculo de predicados de primeira ordem, as tabelas verdade e os procedimentos de prova, sendo estes os métodos que formam o núcleo básico da Lógica clássica.

A Lógica paraconsistente, cujos fundamentos podem ser associados aos trabalhos iniciais do polonês Stanislaw Jaskowski e do brasileiro Newton da Costa, derroga o princípio da não contradição. Ou seja, ela tolera a existência de proposições contraditórias e permite o desenvolvimento de métodos semelhantes ao da lógica clássica, servindo de fundamento para o surgimento de cálculos e teorias em que o fenômeno da contradição ocorra. Além disso, a tolerância à existência das contradições possibilitou o desenvolvimento de métodos capazes de fornecer um arcabouço teórico e prático para o tratamento da incerteza, imprecisão e vagueza (DA COSTA et al., 1999).

O desenvolvimento de um corpo teórico sólido, ao longo do tempo, estimulou a percepção de várias aplicações para a lógica paraconsistente, entre as quais podem ser destacadas: na Matemática, com a axiomatização da teoria dos conjuntos, sem que surgissem os paradoxos anteriormente citados; na Física, para o tratamento de teorias que, num primeiro momento, parecem incompatíveis entre si, como é o caso da Mecânica Quântica e da teoria da Relatividade geral; em Computação, Inteligência Artificial e Robótica; em Psicanálise; em questões específicas da filosofia, como por exemplo, a Dialética (DA COSTA et al., 1999).

Uma das variações da Lógica paraconsistente é a Lógica paraconsistente anotada de anotação com dois valores que, de agora em diante, será referida como LPA2v, conforme utilizado em Da Costa et al. (1999). Nesta lógica, atribui-se, a uma proposição, duas variáveis: os graus de crença e de descrença, com valores que variam no intervalo $[0,1]$. Esta abordagem torna possível o estabelecimento de quatro estados lógicos (inconsistente, verdadeiro, falso, indeterminado), em vez dos dois estados previstos na lógica clássica (verdadeiro e falso). Os estados lógicos são determinados com base nos dois valores da anotação, descritos por um par (μ_1, μ_2) , que representam, respectivamente, o grau de crença e o grau de descrença atribuído a uma proposição.

A título de exemplo, seja considerada p uma proposição com graus de crença e descrença (μ_1, μ_2) . O par $(1.0, 0.0)$ significa crença total e ausência de descrença na proposição ou, dito de outra forma, ela é verdadeira. Analogamente, podem ser descritos os seguintes estados lógicos:

(1.0,1.0) – inconsistente (crença total e descrença total)

(0.0,1.0) – falso (ausência de crença e descrença total)

(0.0,0.0) – indeterminado (ausência de crença e ausência de descrença).

Pode-se observar que estes são os valores extremos passíveis de serem atribuídos aos graus de crença e descrença. Outros valores, pertencentes ao intervalo $[0,1]$ podem ser atribuídos, permitindo amplas possibilidades de estabelecer estados intermediários aos quatro estados lógicos anteriormente definidos. Uma análise pode ser efetuada de modo que, para uma dada proposição, pode-se avaliar sua tendência. Por exemplo, uma proposição com valores anotados de (0.8,0.2) pode ser entendida como sendo muito próxima de ser verdadeira.

Em Da Costa et al. (1999) podem ser encontrados teoremas e definições que completam a axiomatização da lógica paraconsistente anotada. Eles não serão transcritos aqui pois tal transcrição foge aos objetivos deste trabalho.

A título de conclusão deste tópico, pode-se afirmar que a teoria da LPA2v possibilitou o desenvolvimento de certos aspectos da Matemática com um tratamento diferente da abordagem tradicional. Foi possível, por exemplo, desenvolver uma teoria de conjuntos paraconsistente capaz de englobar a teoria dos conjuntos difusos (DA COSTA et al., 1999).

4.3.2 O algoritmo para-analisador

Na seção anterior foi descrito o modo como a LPA2v utiliza duas variáveis – os graus de crença e descrença – para definir o valor lógico de uma proposição. Conforme dito anteriormente, é possível estabelecer valores intermediários aos estados lógicos extremos. Assim, por exemplo, em vez de definir uma proposição, ou uma situação do mundo real representada sob a forma de uma proposição, como verdadeira ou falsa, podem-se estabelecer valores intermediários. Da mesma forma, situações que envolvem inconsistências ou contradições podem ser analisadas. No desenvolvimento desta seção será mostrado como esta análise pode ser efetuada de forma gráfica, como também um algoritmo que permite automatizar a análise.

Num primeiro momento, os estados lógicos da LPA2v, estabelecidos a partir dos valores dos graus de crença e descrença, podem ser representados em um gráfico simples, em

que os eixos horizontal e vertical indiquem os respectivos graus, conforme a figura abaixo. A figura representada, chamada de quadrado unitário do plano cartesiano será, de agora em diante, denominada QUPC (DA COSTA et al., 1999).

Figura 10 - Representação dos estados lógicos extremos a partir dos graus de crença/descrença - QUPC
Fonte: Da Costa et al. (1999), p. 46

Em um segundo momento, podem ser anexados outros pontos ao gráfico anterior, de modo a descrever estados intermediários aos quatro estados lógicos extremos. Este procedimento permite estabelecer descrições de situações em que os valores extremos não se aplicam, conforme a figura abaixo.

Figura 11 - Representação de alguns estados lógicos intermediários
Fonte: Da Costa et al. (1999), p 50

A partir do gráfico anterior, podem ser traçados alguns segmentos de reta de modo a definir, em vez de pontos intermediários, regiões de ocorrência dos estados extremos. Ou seja,

além do ponto que descreve o estado extremo, tem-se uma região que pode ser considerada a “ocorrência de um estado extremo”. A seguir estão listados os segmentos de reta traçados e as figuras que descrevem algumas das regiões que os segmentos delimitam.

Segmento BD - linha limite de falsidade

Segmento DF - linha limite de inconsistência

Segmento FH - linha limite de verdade

Segmento HB - linha limite de indeterminação

Figura 12 - QUPC com linhas de limitação de regiões
Fonte: Da Costa et al. (1999), p. 65

Figura 13 - QUPC com linhas de limitação de regiões identificadas
Fonte: Da Costa et al. (1999), p. 65

Figura 14 - QUPC destacando a região de totalmente inconsistente
 Fonte: Da Costa et al (1999), p. 66

Figura 15 - QUPC destacando a região de totalmente indeterminado
 Fonte: Da Costa et al. (1999), p. 67

Figura 16 - QUPC destacando a região de totalmente falso
 Fonte: Da Costa et al. (1999), p. 67

Figura 17 - QUPC destacando a região de totalmente verdadeiro
 Fonte: Da Costa et al. (1999), p. 68

Alem das regiões associadas aos estados extremos, outras podem ser observadas na figura. Essas regiões descrevem estados intermediários e tendências, que se situam entre os estados extremos definidos. As figuras 18, 19 e 20, a seguir, descrevem, respectivamente, duas regiões intermediárias isoladamente e uma compilação de todas as regiões do QUPC, intermediárias e extremas. Esta configuração é chamada por Da Costa et al. (1999) de QUPC de resolução 12, pois podem ser identificadas, no gráfico, 12 regiões.

Figura 18 - QUPC destacando a região de quase falso tendendo ao indeterminado
 Fonte: Da Costa et al. (1999), p. 70

Figura 19 - QUPC destacando a região de quase verdadeiro tendendo ao inconsistente
Fonte: Da Costa et al. (1999), p. 71

Figura 20 - QUPC de resolução 12 destacando todas as regiões possíveis – intermediárias e extremas
Fonte: Da Costa et al. (1999), p. 89

Para esta última figura, as regiões representadas são identificadas por diversas abreviaturas, cujos significados são:

T – Inconsistente

F – Falso

⊥ – Indeterminado

V – Verdadeiro

⊥ → f – Indeterminado, tendendo ao Falso

⊥ → v – Indeterminado, tendendo ao Verdadeiro

T → f – Inconsistente, tendendo ao Falso

T → v – Inconsistente, tendendo ao Verdadeiro

Qv → T – Quase Verdadeiro, tendendo ao Inconsistente

$Q_f \rightarrow T$ – Quase Falso, tendendo ao Inconsistente

$Q_f \rightarrow \perp$ – Quase Falso, tendendo ao Indeterminado

$Q_v \rightarrow \perp$ – Quase Verdadeiro, tendendo ao Indeterminado

As figuras anteriores, com suas regiões delimitadas, podem ser descritas matematicamente através de desigualdades que envolvem as variáveis mostradas nos gráficos. Da Costa et al. (1999) descrevem outras variáveis que podem ser utilizadas como delimitadores das regiões mostradas. São elas:

V_{scc} – valor superior de controle de certeza. Que limita o grau de certeza próximo ao verdadeiro;

V_{icc} – valor inferior de controle de certeza. Que limita o grau de certeza próximo ao falso;

V_{scct} – valor superior de controle de contradição. Que limita o grau de contradição próximo ao estado inconsistente;

V_{icct} – valor inferior que limita o grau de contradição próximo ao indeterminado.

Para todos os gráficos anteriores essas variáveis assumem os seguintes valores:

$$V_{scc} = 1/2$$

$$V_{icc} = -1/2$$

$$V_{scct} = 1/2$$

$$V_{icct} = -1/2$$

Essas variáveis não são visíveis diretamente nos gráficos mostrados, pois, na verdade, são artifícios matemáticos que serão utilizados mais adiante. Em Da Costa et al. (1999) existem outros gráficos em que essas variáveis são visíveis. Elas permitem que sejam realizados ajustes no tamanho das regiões delimitadas no QUPC. Tais ajustes permitem que o procedimento de análise descrito a seguir possa ser calibrado de modo a restringir a sensibilidade do procedimento a certos valores lógicos.

As figuras 21 e 22, a seguir, são variações da figura 20, em que os valores das variáveis V_{scc} , V_{icc} , V_{scct} , V_{icct} é modificado. É interessante observar que as alterações nesses valores extremos não modificam a quantidade de regiões no gráfico. Contudo, as novas configurações modificam a sensibilidade da análise a ser efetuada.

Figura 21- QUPC de resolução 12- destaque para regiões com variáveis delimitadoras modificadas
Fonte: Da Costa et al. (1999), p. 91

Na figura 21 os valores das variáveis delimitadoras são: $V_{scc} = 3/4$, $V_{icc} = -3/4$, $V_{sct} = 3/4$, $V_{ict} = -3/4$. Pode-se observar, no gráfico, a redução no tamanho das regiões extremas (Verdadeiro, Falso, Inconsistente e Indeterminado). Nesse caso, uma análise efetuada a partir dos valores de crença e descrença (μ_1 e μ_2) será mais exigente para os valores extremos. Ou seja, o número de situações em que o par $\mu_1 - \mu_2$ será considerado verdadeiro (ou falso, indeterminado e inconsistente) será menor do que o número de situações analisadas na figura 20.

Outras configurações podem ser efetuadas como, por exemplo:

Figura 22 - QUPC de resolução 12- destaque para regiões com variáveis delimitadoras modificadas
Fonte: Da Costa et al. (1999), p. 93

Nessa configuração, figura 22, os valores das variáveis delimitadoras são: $V_{scc} = 3/4$, $V_{icc} = -3/4$, $V_{sct} = 1/2$, $V_{ict} = -1/2$. Observando o gráfico, pode-se ver a redução no tamanho das regiões extremas para os valores lógicos verdadeiro e falso, quando comparadas com as regiões para o indeterminado e o inconsistente. Aqui, a exigência foi alterada para os valores

padrão da Lógica clássica e modificada para os valores lógicos que a LPA2v inclui. Ou seja, nessa configuração um número menor de pares crença/descrença irá satisfazer as condições de verdadeiro e falso. Por outro lado, um número maior de pares crença/descrença satisfará as condições indeterminado e inconsistente.

Em ambas as situações descritas pelas figuras 21 e 22, o tamanho das regiões intermediárias também é modificado, o que significa que podem haver diferentes tolerâncias para os valores intermediários. Nos próximos parágrafos será mostrado como esta análise pode ser efetuada com a utilização de um procedimento padronizado.

Um segundo conjunto de variáveis deve ser definido (DA COSTA et al., 1999), antes de explicitar o procedimento de análise. São elas:

G_{it} – Grau de inconsistência, onde:

$$G_{it} = \mu_1 + \mu_2 - 1, \text{ se e somente se } (\mu_1 + \mu_2) \geq 1$$

G_{id} – Grau de indeterminação, onde:

$$G_{id} = \mu_1 + \mu_2 - 1, \text{ se e somente se } (\mu_1 + \mu_2) \leq 1$$

G_v – Grau de verdade, onde:

$$G_v = \mu_1 - \mu_2, \text{ se e somente se } \mu_1 \geq \mu_2$$

G_f – Grau de falsidade, onde:

$$G_f = \mu_1 - \mu_2, \text{ se e somente se } \mu_1 \leq \mu_2$$

G_{ct} – Grau de contradição, onde:

$$G_{ct} = \mu_1 + \mu_2 - 1, \text{ se e somente se } 0 \leq \mu_1 \leq 1 \text{ e } 0 \leq \mu_2 \leq 1$$

G_c – Grau de certeza, onde:

$$G_c = \mu_1 - \mu_2, \text{ se e somente se } 0 \leq \mu_1 \leq 1 \text{ e } 0 \leq \mu_2 \leq 1$$

As figuras anteriores podem ser entendidas como ferramentas que auxiliam a avaliar o estado lógico de uma proposição ou enunciado, que possui valores de crença e descrença μ_1 e μ_2 , respectivamente. Apesar de as figuras serem boas para visualizar o estado lógico, elas não são úteis para fazer avaliações de maneira contínua ou repetitiva, pois seria necessário verificar de forma permanente em que região do QUPC um determinado par de μ_1 e μ_2 se situa. Nesse sentido, Da Costa et al. (1999) desenvolveram um algoritmo que permite, a partir dos valores de graus de crença e descrença, determinar o estado lógico da proposição. O algoritmo, chamado de para-analisador, será reproduzido a seguir.

Algoritmo para-analisador

/* Definições dos valores */

$V_{scc} = C_1$ /* Definição do valor superior de controle de certeza */

$V_{icc} = C_2$ /* Definição do valor inferior de controle de certeza */

$V_{sct} = C_3$ /* Definição do valor superior de controle de contradição */

$V_{ict} = C_4$ /* Definição do valor inferior de controle de contradição */

/* Variáveis de entrada */

μ_1

μ_2

/* Variável de saída */

S

/* Expressões matemáticas */

Sendo $0 \leq \mu_1 \leq 1$ e $0 \leq \mu_2 \leq 1$

$G_{ct} = \mu_1 + \mu_2 - 1$

$G_c = \mu_1 - \mu_2$

/* Determinação dos estados extremos */

Se $G_c \geq C_1$ então $S = V$ /* Verdadeiro */

Se $G_c \leq C_2$ então $S = F$ /* Falso */

Se $G_{ct} \geq C_3$ então $S = T$ /* Inconsistente */

Se $G_{ct} \leq C_4$ então $S = \perp$ /* Indeterminado */

/* Determinação dos estados não extremos */

Para $0 \leq G_c < C_1$ e $0 \leq G_{ct} < C_3$

Se $G_c \geq G_{ct}$ então $S = Qv \rightarrow T$

Se $G_c < G_{ct}$ então $S = T \rightarrow v$

Para $0 \leq G_c < C_1$ e $C_4 < G_{ct} \leq 0$

Se $G_c \geq |G_{ct}|$ então $S = Qv \rightarrow \perp$

Se $G_c < |G_{ct}|$ então $S = \perp \rightarrow v$

Para $C_2 < G_c \leq 0$ e $C_4 < G_{ct} \leq 0$

Se $|G_c| \geq |G_{ct}|$ então $S = Qf \rightarrow \perp$

Se $|G_c| < |G_{ct}|$ então $S = \perp \rightarrow f$

Para $C_2 \ 0 < G_c \leq 0$ e $0 \leq G_{ct} < C_3$

Se $|G_c| \geq G_{ct}$ então $S = Qf \rightarrow T$

Se $|G_c| < G_{ct}$ então $S = T \rightarrow f$

O algoritmo apresentado permite que uma análise, relativa ao valor lógico de uma proposição, possa ser efetuada em sistemas computacionais, desde que se tenha uma proposição conveniente para os fins propostos e sejam feitos os cálculos dos graus de crença e descrença.

4.3.3 Utilização do para-analisador na indexação

A utilização da lógica difusa e da teoria de Dempster-Shafer na indexação automática pressupõe o estabelecimento de algum tipo de função que estabeleça um grau de agregação entre um ou mais termos e um documento no qual eles ocorram. Este mecanismo é denominado função de pertinência na lógica difusa (BORDOGNA; PASI, 1995) e função de crença na Teoria de Dempster-Shafer (LALMAS; RUTHVEN, 1998). A solução que envolve a teoria de Dempster-Shafer avalia também o grau de plausibilidade para uma dada proposição. Contudo, ambas as abordagens desconsideram avaliações que indiquem a “desagregação” de um termo a um documento. Ou seja, o grau de “não pertinência” do termo ao documento em questão.

Conforme visto anteriormente, a LPA2v utiliza, formalmente, uma variável para o grau de crença (ou de agregação) e uma variável que indica o grau de descrença (ou “desagregação”) quanto a uma proposição. No contexto desta pesquisa, será considerada a proposição p_t – o termo t é um bom termo para indexar o documento D. Se forem atribuídos valores μ_1 e μ_2 a p_t , com ambos os graus estabelecidos por meio de algum cálculo ou heurística, pode-se estabelecer, com precisão, o quanto a proposição se aproxima de um dos quatro estados lógicos conhecidos. Ou seja, o quanto o termo t é um bom termo para indexar o documento.

Um termo t , para o qual os valores de crença e descrença forem, respectivamente, 1 e 0, poderia ser considerado como um bom termo de indexação, pois a proposição p_t seria considerada verdadeira. Já um termo com valores 0 e 1 não seria um bom termo de indexação, uma vez que a proposição seria considerada falsa. Pode-se destacar que esta sistemática possibilita estabelecer a adequação de um termo t levando-se em conta a utilização de duas variáveis, em vez de uma, conforme as soluções anteriormente descritas em Bordogna e Pasi (1995) e em Lalmas e Ruthven (1998).

Acredita-se que a utilização de duas variáveis - os fatores de crença e descrença - permitirá o desenvolvimento de uma sistemática de indexação que poderá ser comparada tanto com as soluções clássicas de indexação automática quanto com as soluções que se utilizam de outras teorias para o tratamento da incerteza, aqui descritas.

Uma proposta para utilização da Lógica paraconsistente no contexto da recuperação de informação foi efetuada por Sucolotti (2001) que sugeriu sua utilização para a expansão de consultas efetuadas em um sistema de recuperação de informações, com a utilização de tesouros. Esta proposta considerava a utilização dos índices atribuídos aos documentos e sua comparação com o conteúdo do tesouro.

Nesta pesquisa, testa-se a possibilidade de utilização da Lógica paraconsistente tanto na geração de índices quanto no seu refinamento, uma vez utilizada alguma metodologia já conhecida.

Assim, a idéia para testar a hipótese de trabalho subjacente aos objetivos elencados na seção 1.1 consiste em utilizar o para-analisador para avaliar a qualidade dos índices associados aos documentos de uma coleção. O teste propriamente dito deverá ser efetuado em um sistema computacional que utilize, direta ou indiretamente, a indexação. Neste caso, os resultados dos testes serão mostrados por meio de um sistema de visualização de informação. Assim, no próximo capítulo será feita uma introdução à teoria da visualização de informações e a apresentação do sistema que será utilizado nos testes.

5. Visualização de informações

5.1 Introdução à visualização de informação

A necessidade de avanços na interação humano-computador, o aumento de desempenho dos sistemas computacionais, e o surgimento de demandas que envolvem a manipulação computacional de imagens de diversos tipos são fatores que, entre outros, conduziram ao desenvolvimento da Computação Gráfica. Esta última pode ser descrita como a área da Ciência da Computação que se preocupa com a análise, síntese e processamento de imagens.

O desenvolvimento da Computação Gráfica modificou o modo como os usuários interagem com o computador. O surgimento das interfaces gráficas é um excelente exemplo dessa influência, que não se limitou apenas às interfaces, mas expandiu-se por diversas disciplinas da Ciência da Computação e Engenharia, entre outras.

As dificuldades com os processos de Recuperação de Informação, descritas no Capítulo 2, podem, em princípio, ser minimizadas à medida que os usuários passam a conhecer melhor o espaço de busca em que estão atuando, aumentando assim, a efetividade de sua procura. Muitos sistemas de visualização de informações retratam a estrutura semântica global de uma coleção de documentos. Assim, os usuários podem utilizar a visualização estrutural como base para suas buscas subsequentes (BORNER; CHEN; BOYACK, 2003).

A utilização de imagens, figuras ou gráficos, de diversos tipos e formatos é um recurso utilizado para transmitir ou sintetizar uma grande quantidade de informações num espaço reduzido.

A figura a seguir é um exemplo clássico de utilização de imagens para sintetizar uma grande quantidade de informações que, de outra forma, envolveria uma extensa descrição textual linear. Ela descreve a campanha de Napoleão, na Rússia, e foi elaborada em 1812.

Figura 23 - Marcha de Napoleão na Rússia
 Fonte: Chen (2006), p. 9

A visualização de informações estuda o desenvolvimento de técnicas de representação visual de conjuntos de dados que não apresentam uma geometria intrínseca ou natureza gráfica. Refere-se à representação gráfica de dados complexos, multidimensionais e volumosos, tais como hierarquias muito grandes, bases de dados ou dados estatísticos. A expressão também é associada à representação dos conjuntos de dados não estruturados ou abstratos (documentos, coleções, sites da WWW, entre outros).

Na sociedade contemporânea, a sobrecarga de informações tem sido um problema comum, consequência do crescimento exponencial da acessibilidade aos conteúdos informacionais e, dessa forma, faz-se necessário realizar filtragens eficientes de informações e o compartilhamento de facilidades.

A visualização de informações apresenta potencial para ajudar as pessoas a encontrarem a informação que precisam de forma mais efetiva e intuitiva (CHEN, 2006). Ela pode, de acordo com Chen (2006), ser vista como:

- uma maneira de diminuir o esforço necessário para explorar coleções de informações;
- uma ferramenta que auxilia a identificação de padrões e a obtenção de *insights* em situações que envolvem a manipulação e o acesso a grandes quantidades de informação;
- um modo de enxergar o “invisível” ou de “descobrir” o “desconhecido”.

Podem ser acrescentadas três questões subjacentes aos estudos de visualização de informações: O que os usuários querem visualizar? Que metáforas visuais devem ser utilizadas? Como avaliar e comparar os efeitos das diferentes técnicas ou soluções?

Nesse sentido, Chen (2006) questiona a utilização de visualizações baseadas em mapas geográficos, embora elas possam parecer intuitivas e naturais. Desse modo, formula as seguintes questões gerais sobre a visualização de informações e os ambientes virtuais:

- São os paradigmas de visualização geográfica extensíveis às informações que não são naturalmente geográficas ou astronômicas?
- Como se deve visualizar espaços de informação abstrata em geral?
- Quais são os critérios para uma visualização informativa e estimuladora de *insights* de um espaço de informações abstratas?
- Que fatores humanos devem ser levados em conta?

O autor afirma, ainda, que a navegação em um espaço de informação visual passa pela construção de um mapa cognitivo, tal como na navegação em um espaço real. Nesse caso, o desafio é construir visualizações que permitam ao usuário manter conexão com o seu mapa cognitivo interno sem que a complexidade do espaço de informação, representado e disponibilizado pelo computador, torne-se um obstáculo.

Resumidamente, a visualização de informação possui dois aspectos fundamentalmente relacionados: (i) a modelagem estrutural; e (ii) representação gráfica. O objetivo do primeiro é detectar, extrair e simplificar relacionamentos subjacentes. Tais relacionamentos formam a estrutura que caracteriza uma coleção de documentos ou de outros conjuntos de dados. Por outro lado, o objetivo da representação gráfica consiste em transformar a representação inicial da estrutura em uma forma gráfica, de maneira a tornar possível a interação e o exame da estrutura original manipulando sua representação gráfica (CHEN, 2006).

Outra maneira de descrever o processo para gerar uma visualização para um conjunto de dados abstratos está representado na Figura 24, composto por três etapas: Transformação dos Dados; Mapeamento Visual e Transformações Visuais.

Figura 24 – Modelo de referência de visualização
Fonte : Minghim, 2003, p. 3

Na primeira etapa, um conjunto de dados brutos (coletados ou gerados por algum processo) são processados e transformados em um conjunto de relações (tabelas de dados) mais apropriado para uso na visualização.

A etapa de Mapeamento Visual consiste em escolher estruturas visuais para mapear essas tabelas; assim, os itens de dados são associados a marcas visuais.

Na última etapa, Transformações de Visões, o usuário deve ser capaz de interagir com as visões para controlar a quantidade de informação exibida na tela e também para alterar as propriedades da visualização; assim, o usuário pode explorar diferentes cenários para um melhor entendimento dos dados visualizados.

Um grande número de sistemas de visualização pode ser visto em Chen (2006) e em Zhu e Chen (2005). Nesses textos também podem ser vistas descrições de técnicas e soluções diversas para resolver os problemas associados à visualização. Não existe uma padronização no que diz respeito à classificação de técnicas e soluções tecnológicas utilizadas nos diversos sistemas.

O desenvolvimento de sistemas de visualização de informações não estruturadas (documentos textuais ou de outro formato) é subsidiado, de acordo com Zhu e Chen (2005), por estudos realizados nas seguintes dimensões de pesquisa: representação de informação, interação usuário-interface e análise da informação. Nos próximos parágrafos serão detalhados os aspectos mais importantes que envolvem estas dimensões.

Existem diferentes maneiras de classificar os modos como a informação é representada em um sistema de visualização. Uma das abordagens mais citadas é a de Shneiderman (1996). Este autor propôs sete tipos básicos de métodos de representação: 1 dimensão, 2 dimensões, 3 dimensões, multidimensional, árvore (ou hierárquica), rede (ou grafo) e temporal.

Podem ser encontrados sistemas de visualização para todos os tipos básicos anteriormente citados. Em alguns exemplos, mais de um dos tipos descritos é utilizado pelo sistema. Essencialmente, os sete tipos se propõem a representar documentos textuais em objetos (ou metáforas visuais) que possam ser exibidos na tela de um computador. É importante ressaltar a necessidade de integrar os métodos de representação utilizados com a interface do usuário do respectivo sistema (ZHU; CHEN, 2005).

A interação entre a interface de um sistema de visualização e de seus usuários permite tanto a manipulação direta dos objetos visuais exibidos quanto à seleção do que será ou não exibido (CARD; MACKINLAY; SHNEIDERMAN, 1999).

Shneiderman (1996), por sua vez, resume seis tipos de funcionalidades para uma interface: *overview*, *zoom*, *filtering*, *details on demand*, *relate* e *history*. Diferentes técnicas foram desenvolvidas de modo a facilitar os vários tipos de interações. As mais citadas na literatura são: *overview + detail* e *focus + context* (CARD; MACKINLAY; SHNEIDERMAN, 1999).

A técnica de *overview + detail* disponibiliza múltiplas visões, a primeira com o objetivo de prover modelos globais. Posteriormente, detalhes sobre alguma parte de interesse do usuário podem ser exibidos. Quando uma visão detalhada é necessária, dois tipos de *zooming* podem ser utilizados: *spatial zooming* e *semantic zooming*. O primeiro diz respeito ao processo de ampliar os objetos visuais escolhidos para obter uma visão fechada de algum aspecto. O segundo provê informação adicional, sobre um objeto visual selecionado, alterando a sua aparência (ZHU; CHEN, 2005).

A técnica *focus + context* provê detalhes e resumo, dinamicamente, sobre uma mesma visão. Dois exemplos de utilização desta técnica são: Cone Tree e fish-eye (ZHU; CHEN, 2005). Um ponto a ser observado, nessa abordagem, é que os usuários geralmente podem mudar a região de foco dinamicamente.

A análise da informação é utilizada com o objetivo de reduzir a complexidade e extrair estruturas proeminentes de grandes quantidades de informação não estruturada. Para atingir esses objetivos, a análise é desenvolvida em dois estágios: indexação e análise propriamente dita (ZHU; CHEN, 2005).

A indexação, nesse contexto, consiste em extrair a semântica⁶ da informação para representar seu conteúdo. Assim, diferentes algoritmos são utilizados para diferentes tipos de

⁶ No artigo citado, não há uma definição formal de semântica, conforme visto na seção 2.3.1. Contudo, no parágrafo anterior desse texto, existe uma referência sobre o objetivo da análise da informação: reduzir complexidade e extrair estruturas proeminentes. Assim pode-se considerar que, informalmente, nesse contexto, a semântica está associada a esses dois objetivos.

mídias, tais como: textos, imagens, áudio e vídeo. Segundo Zhu e Chen (2005), as principais técnicas de indexação de documentos textuais são: indexação automática e processamento da linguagem natural. É possível identificar automaticamente informação útil em um documento por meio de operações de extração. Esta técnica consiste em identificar e extrair nomes de entidades de interesse, tais como: pessoas, locais, organizações, data, expressões numéricas, endereços de e-mail ou URLs da Internet.

No segundo estágio, de análise propriamente dita, são utilizadas técnicas de classificação e agrupamento (*clustering*). Na classificação, os objetos (de informação) são encaixados em grupos pré-definidos, enquanto o agrupamento aproxima objetos dinamicamente, com base em suas similaridades. Ambos os métodos geram grupos a partir da análise das características dos objetos, extraídas no estágio de indexação.

A análise da informação permite representar documentos textuais por meio de frases ou índices, e facilita a identificação de padrões interessantes, usando algoritmos de classificação e agrupamento. O desenvolvimento de sistemas de visualização baseados na utilização desses métodos permite lidar com coleções de documentos cada vez maiores e mais complexas. O esquema, a seguir, descreve sucintamente a influência da análise de informação no desenvolvimento de sistemas de visualização.

Figura 25 – Análise da informação e visualização
Fonte: autor desse trabalho

Os sistemas de visualização voltados para a apresentação de coleções de documentos adotam metáforas diversas para a construção da visualização final.

Assim, por exemplo, o sistema Infosky (ANDREWS et al. 2002) utiliza como metáfora a representação de uma visão noturna do céu. Nesse sistema, as ocorrências individuais dos documentos são apresentadas como “estrelas” no céu. Os “agrupamentos de estrelas” indicam agrupamentos de documentos similares, conforme pode ser visto na figura a seguir.

Figura 26 - Exemplo de visualização no sistema Infosky

Fonte: <http://know-center.tugraz.at/en/solutions/demos-screencasts-en#infosky>

Outro sistema de visualização, o VxInsight (BORNER; CHEN; BOYACK, 2003) utiliza uma metáfora visual baseada em mapas geográficos. A tela principal do sistema exibe um agrupamento de “montanhas”, de altura variável para representar agrupamentos de objetos de informação. Ou seja, as montanhas representam agrupamentos de documentos similares. A altura da montanha é proporcional à densidade ou quantidade de objetos que deram origem aos agrupamentos. O sistema também oferece facilidades para que o usuário realize consultas, com resultados realçados na imagem gerada. A figura a seguir é um exemplo de utilização do VxInsigh.

Figura 27 – Visualização no sistema VxInsigh

Fonte – Borner, Chen e Boyack, 2003, p. 46

Ambos os sistemas citados possuem facilidades para ampliação/aproximação das imagens (*zoom*) e visualização dos agrupamentos de documentos bem como para a realização de consultas.

5.2 Similaridade, distância e proximidade

Independentemente da metáfora visual utilizada, os sistemas desenvolvidos para apresentar coleções de documentos precisam realizar procedimentos e cálculos voltados a dois objetivos específicos: (i) o agrupamento de documentos considerados similares; e (ii) a separação dos grupos de documentos dissimilares entre si.

Além da separação dos diferentes grupos, os sistemas devem ser capazes de efetuar cálculos que indiquem o grau de similaridade entre os documentos componentes de cada grupo a ser visualizado. Esse cálculo determina a proximidade entre os itens individuais que compõem cada grupo a ser apresentado. A capacidade para realizar tais cálculos, de modo preciso, é fundamental em sistemas em que os itens da coleção são visualizados em grupos e individualmente, como é o caso do Infosky anteriormente mostrado e do Projection Explorer, que será descrito detalhadamente, na seção 5.3.

A similaridade entre documentos pode ser considerada subjetiva e dependente de características semânticas dos documentos, bem como do domínio subjacente às coleções (Minghim; Levkovitz, 2007). Contudo, para os fins práticos da criação de visualizações, torna-se necessário realizar cálculos de similaridade. Esse objetivo obriga os sistemas a utilizarem representações específicas para os documentos das coleções, de modo a facilitar a efetivação dos referidos cálculos. Um dos modelos mais utilizados para este fim, em diferentes sistemas de visualização, é o modelo do espaço vetorial (MEV), descrito na seção 3.5, apesar das restrições e críticas ao mesmo. Uma das principais críticas refere-se à pressuposição, assumida pelo modelo, de que os termos utilizados para representação vetorial dos documentos são independentes entre si. Segundo Chen (2006), tal pressuposição pode simplificar demasiadamente o inter-relacionamento entre termos particulares e seu contexto, e levar a resultados não intuitivos. Entretanto, para Baeza-Yates e Ribeiro Neto (1999) não existem evidências empíricas suficientes para sustentar essa crítica.

Diversos métodos foram estabelecidos ou adaptados para a efetivação do cálculo de similaridades entre documentos, tais como, por exemplo, a complexidade de Kolmogorov e a indexação semântica latente, que podem ser vistos com maiores detalhes em Minghim e Levkowitz (2007) e em Chen (2006), respectivamente.

A figura 28, a seguir, é um exemplo de visualização construída com o sistema Projection Explorer. Os círculos coloridos da mesma cor indicam documentos de um mesmo assunto, todos pertencentes a uma mesma coleção de teste. Na figura foram destacados, através de retângulos coloridos, alguns grupos de documentos que, apesar de serem de assuntos distintos aparecem próximos na visualização. Este efeito foi obtido com base nos critérios que o sistema utiliza para construir a visualização, ou seja, tais documentos possuem algum grau de similaridade. Nesse contexto, poder-se-ia perguntar se tal constatação seria facilmente observável em um sistema de recuperação de informações.

Figura 28 – Visualização de documentos no Projection Explorer – destaque p/ documentos de assuntos diferentes porem com algum grau de similaridade entre si

Fonte: Visualização criada, pelo autor, com o Projection Explorer

A título de conclusão desta subseção, podem ser citados Minghim e Levkovitz (2007), que consideram que a questão do que é importante para definir como similar, em uma particular configuração de uma aplicação ou tarefa, direciona o desenvolvimento e/ou aplicação de novas abordagens para o cálculo de similaridades entre pares de documentos. Consideram ainda, que estratégias próprias podem ser desenvolvidas para levar em conta a

homogeneidade e os níveis semânticos de uma coleção de textos tanto quanto o contexto no qual a coleção está sendo analisada.

5.3 Um sistema de visualização de informações: Projection Explorer

Ao longo das últimas décadas, muitos sistemas para visualização de informações e de domínios de conhecimento foram desenvolvidos. Os sistemas apresentam propósitos diferentes e se utilizam também de metáforas diferentes para representar, nas telas dos computadores, coleções de documentos ou alguns de seus atributos. Vários exemplos destes sistemas podem ser vistos em Borner, Chen e Boyack (2003), em Chen (2006) e em Zhu e Chen (2005).

No Brasil, foi desenvolvido no Instituto de Ciências Matemáticas e de Computação da USP/São Carlos um sistema de visualização que vem sendo utilizado por diferentes grupos de pesquisadores para testes, simulações e pesquisas. Esse sistema é conhecido por PEx – Projection Explorer (PAULOVICH; OLIVEIRA; MINGHIM, 2007; PAULOVICH et al., 2008)⁷.

O sistema PEx utiliza pequenos círculos coloridos como metáforas visuais para representar documentos individuais de uma coleção. Além disso, o aplicativo possibilita que os cálculos de similaridade entre os documentos da coleção sejam efetuados utilizando-se diferentes técnicas e métodos.

A figura a seguir descreve, resumidamente, as fases que compõem o processo de criação da visualização de uma coleção de documentos com o uso do Projection Explorer.

⁷ O sistema Projection Explorer pode ser obtido, gratuitamente, em <http://infoserver.lcad.icmc.usp.br/infovis2/PEx>.

Figura 29 – Processo para construção da visualização de uma coleção de documentos

Fonte: Paulovich, Oliveira e Minghim, 2007, p. 31

Na Figura 29 podem ser observadas as diferentes maneiras de alimentar o sistema com os dados a serem visualizados. Os dados podem ser: uma coleção de documentos (1); uma tabela estruturada de dados relativos a um conjunto de documentos (2); ou uma matriz de distâncias (3) (PAULOVICH; OLIVEIRA; MINGHIM, 2007). A projeção é realizada a partir do estabelecimento das distâncias entre os elementos individuais e, em seguida, apresentados sob a forma de objetos visuais.

Em termos não formais, a projeção pode ser descrita como um processo que consiste em estabelecer correspondência matemática entre a representação de um objeto de informação num espaço de m dimensões e a sua representação em um espaço de dimensão reduzida, em geral uma, duas ou três dimensões. Tal correspondência deve, tanto quanto possível, preservar as relações de distância e proximidade do espaço original no espaço projetado. Esse procedimento se faz necessário por, essencialmente, dois motivos: (i) a representação dos objetos de informação utilizando geralmente um espaço multidimensional, como consequência da extração do conjunto de índices utilizados para representar cada item da coleção; (ii) a visualização é efetivada nas telas dos computadores e, nesse caso, a representação gráfica final terá, necessariamente, uma, duas ou, no máximo, três dimensões.

Existe um grande conjunto de técnicas de projeção, cujo detalhamento foge aos objetivos deste trabalho, mas que podem ser consultados, por exemplo, em Minghim e Levkowitz (2007) e Paulovich (2008).

Pode-se observar na Figura 29 que, ao “alimentar” o sistema PEx com uma coleção de documentos, o sistema construirá uma tabela de dados relativos aos pesos que podem ser atribuídos aos termos de indexação de cada documento, de acordo com os cálculos estabelecidos no modelo de espaço vetorial, detalhado na seção 3.5, e efetua os cálculos de similaridade, também de acordo com o mesmo modelo.

Para a construção da representação dos documentos da coleção no MEV, o PEx efetua alguns procedimentos de modo a preparar a coleção. Os procedimentos são muito similares aos utilizados para indexação de documentos no sistema SMART, detalhados na seção 3.5. De acordo com Paulovich (2008), são eles:

1. Eliminação das palavras sem conteúdo representativo, tais como artigos e preposições (*stop words*);
2. Conversão das palavras restantes aos seus radicais (lematização). O PEx utiliza o algoritmo de Porter para textos na língua inglesa;
3. Contagem das palavras, dentro de cada documento, para estabelecer suas frequências de ocorrência;
4. Aplicação dos cortes de Luhn, para eliminação de palavras muito frequentes ou raras;
5. Cálculo do peso das palavras selecionadas através da fórmula $tf * idf$ definida na seção 3.5.

As Figuras 30 e 31, a seguir, mostram exemplos de visualizações de coleções de documentos criadas com o PEx.

Figura 30 – Exemplo de visualização – PEx
 Fonte: Paulovich, Oliveira e Minghim, 2007, p. 31

Figura 31 – Exemplo de visualização no PEx com identificação de tópicos em agrupamentos de textos
 Fonte: Paulovich, Oliveira e Minghim, 2007, p. 32

O PEx oferece várias facilidades para manipulação das visualizações, entre as quais podem ser destacadas:

- possibilidade de estabelecer rótulos para agrupamentos de objetos selecionados pelo usuário. Tais rótulos são obtidos a partir da extração de tópicos de assuntos dos objetos escolhidos;
- estabelecimento de cores diferentes para diferentes grupos de objetos;
- possibilidade de realização de consultas e indicação dos documentos com maior número de ocorrências com o termo consultado;
- comparação de mapas de documentos criados a partir de diferentes técnicas de projeção;
- coordenação entre visualizações: permite visualizar, simultaneamente, o mesmo objeto em duas projeções diferentes da mesma coleção.

As funcionalidades do sistema são portanto adequadas para o teste das hipóteses de trabalho desta pesquisa.⁸

⁸ Deve-se assinalar também a acolhida dos desenvolvedores do sistema PEx para a aplicação das hipóteses de trabalho desta pesquisa. Além disso, a equipe do PEx contribuiu generosamente para tornar possível a operacionalização das propostas do presente estudo.

6. Introdução à análise de grupos

Nessa seção serão introduzidos conceitos e métodos desenvolvidos para auxiliar a compreensão e avaliação de alguns aspectos observáveis nas visualizações produzidas. Conforme pode ser observado nas figuras 28, 30 e 31, as visualizações produzidas com o Projection Explorer apresentam grupos de objetos identificáveis por suas diferentes cores. Além disso, o Projection Explorer possui medidas que possibilitam avaliar a qualidade dos agrupamentos obtidos.

As técnicas para avaliação dos agrupamentos presentes nas visualizações estão descritas no corpo de conhecimentos conhecido como análise de grupos (*cluster analysis*). Este, por sua vez, é um dos tópicos de um assunto mais abrangente, conhecido como mineração de dados (*data mining*).

A análise de grupos⁹ divide conjuntos de dados em grupos que podem ser significativos, úteis ou as duas coisas ao mesmo tempo. Se o objetivo é obter grupos significativos, a análise deverá, de alguma forma, capturar a estrutura natural dos dados. Em outras situações, a análise de grupos é apenas de um ponto de partida para outros objetivos. Independentemente do objetivo a ser alcançado, a análise de grupos tem um papel importante em diferentes disciplinas, tais como: psicologia, ciências sociais, biologia, estatística, reconhecimento de padrões, recuperação de informação, aprendizagem de máquina e mineração de dados (TAN; STEINBACH; KUMAR, 2006).

Ainda, segundo Tan, Steibach e Kumar (2006), a análise de grupos aproxima objetos baseada nas informações encontradas nos dados que os descrevem e em seus relacionamentos. Nessa situação, o objetivo é obter grupos cujos objetos que o compõem sejam similares entre si, e diferentes ou não-relacionados aos objetos que compõem os outros grupos. Quanto maior a similaridade dentro do grupo e maior a diferença entre os diversos grupos, melhor o agrupamento. Para esses autores, a atividade de agrupamento pode ser referenciada como um tipo de classificação não supervisionada.

⁹ A literatura consultada nas áreas de mineração de dados e análise de grupos utiliza dois conceitos básicos: *cluster* e *clustering*. De acordo com Tan, Steinbach e Kumar (2006, pg. 491), no original em inglês, “An entire collection of clusters is commonly referred to as clustering...”. No âmbito dessa pesquisa, será adotada a tradução **grupo(s)** para a palavra inglesa *cluster(s)* e **agrupamento(s)** para a palavra inglesa *clustering(s)*. Ou seja, um agrupamento é composto por vários grupos individuais. Além disso, a palavra *cluster* também pode ser aplicada para o verbo **agrupar**.

Outra formalização de conceitos pode ser vista em Han e Kamber (2006), para os quais, o processo de agrupar um conjunto de objetos abstratos ou físicos em classes de objetos similares é chamado de agrupamento (*clustering*). Nesse caso, um grupo (*cluster*) é uma coleção de objetos que são similares uns aos outros, dentro do mesmo grupo, e dissimilares aos objetos de outros grupos.

Uma abordagem similar a descrita no parágrafo anterior pode ser vista em Kaufman e Rosseeuw (1990). Esses autores consideram que a análise de grupos é a arte de encontrar grupos em conjuntos de dados, com o objetivo de agrupar objetos similares entre si e diferenciá-los de outros grupos de objetos dissimilares. Afirmam ainda, que a análise de grupos pode ser usada tanto para identificar uma estrutura já presente no conjunto de dados, quanto para impor uma estrutura a um conjunto de dados mais ou menos homogêneo.

A figura 32, a seguir, exemplifica o objetivo principal da análise de grupos. Nela podemos ver como um conjunto original de dados pode ser dividido em subconjuntos diferentes, de acordo com as técnicas e métodos utilizados para gerar os agrupamentos finais. Nesse exemplo, dependendo da técnica utilizada, o conjunto de dados inicial foi dividido em dois, quatro ou seis grupos. Nessa figura os grupos são diferenciados pela utilização de diferentes símbolos para representá-los.

Figura 32 – diferentes agrupamentos do mesmo conjunto original de dados
Fonte - Tan, Steibach e Kumar (2006, pg 491)

Considerando que um agrupamento é composto de vários grupos de objetos, pode-se observar diferentes modos de categorizar os grupos ou os agrupamentos.

Para Tan, Steibach e Kumar (2006), os agrupamentos podem ser classificados observando-se diferentes características. Na primeira delas, os agrupamentos são classificados em hierárquicos ou particionais. Um agrupamento particional é obtido a partir da simples divisão do conjunto original de dados em subconjuntos não superpostos, de modo que cada objeto está alocado exatamente em um subconjunto. Por outro lado, se um grupo possuir

subgrupos nele aninhados, obter-se-á um agrupamento hierárquico. Nesse caso, os grupos são organizados como uma árvore em que os subgrupos aparecem como seus ramos ou galhos.

Uma segunda forma de classificar os agrupamentos, segundo Tan, Steibach e Kumar (2006), consiste em verificar se os agrupamentos são exclusivos, sobrepostos ou difusos. Agrupamentos exclusivos atribuem cada objeto a apenas um grupo. Contudo, existem situações em que um objeto pode ser colocado em mais de um grupo e, nesse caso, tais situações são mais bem representadas por grupos sobrepostos ou não exclusivos. Em um agrupamento difuso cada objeto pertence a cada grupo de acordo com um grau de pertinência, ou seja, o grupo é tratado como um conjunto difuso, nos termos vistos na seção 4.2.1.

A terceira e última forma de observar um agrupamento, de acordo com os autores citados, consiste em considerar os agrupamentos completos ou parciais. Em um agrupamento completo todo objeto é atribuído a um grupo. Por outro lado, um agrupamento parcial é aquele em que nem todos os objetos originais podem ser atribuídos a um grupo. Nesse caso, os objetos que ficaram afastados, ou excluídos dos grupos construídos podem ser vistos como ruídos (*outliers*).

Além dos agrupamentos, pode-se classificar os grupos individuais que os compõem. Assim, um grupo pode ser: bem separado, baseado em protótipo, baseado em grafo, baseado em densidade (ou conceitual, que possui uma propriedade compartilhada). Um grupo bem separado é aquele em que todos os objetos que o compõem estão mais próximos (ou são mais similares) de si mesmos do que dos outros objetos que compõem outros grupos. Um grupo baseado em protótipo é aquele em que cada um de seus componentes está mais próximo (ou são mais similares) a um protótipo, definido para representar o grupo em questão, do que dos protótipos definidos para os outros grupos. Um grupo baseado em grafo é aquele em que seus objetos são representados por um grafo com linhas representando as conexões entre os objetos e os nós representando os objetos propriamente ditos. Um grupo baseado em densidade é utilizado quando se observa uma distribuição irregular de seus objetos e quando ruído (ou objetos mais afastados) estão presentes. Uma maneira mais genérica de definir grupos consiste em considerá-los como conjuntos de objetos que compartilham uma mesma propriedade, ou seja, são conceitualmente similares. Essa definição engloba todas as anteriores, sem preocupação excessiva com a forma com que os grupos se apresentam (TAN; STEINBACH; KUMAR, 2006).

Existem diferentes métodos para estabelecer um agrupamento a partir de dados originais. Tais diferenças de metodologia são consequência tanto do tipo de agrupamento a ser obtido quanto do tipo de grupo sobre os quais os métodos irão atuar. Assim, existem

diferentes algoritmos desenvolvidos para obtenção de agrupamentos hierárquicos ou particionais. Foge aos objetivos dessa pesquisa um detalhamento desses métodos. Os detalhes podem ser vistos em Tan, Steibach e Kumar (2006), Han e Kamber (2006) e em Kaufman e Rosseeuw (1990). A figura 33, a seguir, exemplifica os resultados obtidos com a aplicação de diferentes métodos de agrupamento sobre um mesmo conjunto de dados original. Os algoritmos utilizados no exemplo são o *k-means*, o DBSCAN e o *complete link*. Nesse exemplo, todos os agrupamentos produzidos são compostos por três grupos de objetos. Contudo, pode-se observar, na figura, que a quantidade de objetos em cada grupo é modificada de acordo com o algoritmo utilizado. Novamente, os objetos que compõem os diferentes grupos são representados por símbolos diferentes.

Figura 33 – Exemplo de agrupamentos produzidos por diferentes algoritmos
 Fonte - Fonte - Tan, Steibach e Kumar (2006, pg 534)

Um dos algoritmos de agrupamento mais utilizados é o *k-means* (eventualmente traduzido para a expressão k-médias), Esse algoritmo foi desenvolvido para produção de agrupamentos particionais, sendo utilizado no sistema Projection Explorer.

Na próxima seção serão detalhados os testes efetuados com o uso o do algoritmo para-analisador nos processos de indexação efetuados pelo PEx, observando-se os efeitos desses procedimentos nas visualizações produzidas pelo sistema.

7. Experimento Computacional

“So, so you think you can tell, Heaven from Hell, blue skies from pain,
Can you tell a green field from a cold steel rail? A smile from a veil?
Do you think you can tell?”

Roger Waters

Esse capítulo descreve os testes realizados para verificar a validade da hipótese de trabalho da pesquisa, qual seja, o de verificar os efeitos da utilização da Lógica paraconsistente nos procedimentos de indexação automática. A abordagem escolhida consistiu na inclusão do algoritmo para-analisador no código fonte do sistema Projection Explorer.

Optou-se por utilizar, para os testes, as coleções de documentos disponibilizadas com o Projection Explorer. Tal escolha baseou-se em dois aspectos: (i) disponibilidade das coleções já utilizadas e testadas no sistema; e (ii) facilidade de comparação dos resultados. As coleções utilizadas são habitualmente utilizadas para testes do próprio sistema. Seu uso facilita a comparação entre as visualizações originais e as visualizações construídas com as alterações introduzidas no sistema.

Conforme visto no capítulo anterior, os sistemas de visualização de informações, quando alimentados com coleções de documentos, podem utilizar-se dos índices estabelecidos para cada documento da coleção ou realizar cálculos para o estabelecimento dos índices, da mesma forma como fazem os sistemas de recuperação de informações. Assim, os sistemas de visualização apresentam potencial para testes com diferentes abordagens de indexação, fornecendo um ambiente propício para a verificação dos efeitos dos diferentes modos de realizar a indexação.

A idéia básica dos testes efetuados neste trabalho pode ser sintetizada nos seguintes passos:

- (1) gerar a visualização de uma coleção de documentos utilizando o ferramental de indexação/cálculos de similaridade embutido originalmente no PEx;
- (2) utilizar o algoritmo para-analisador para avaliar, redefinir ou refinar os índices e os respectivos pesos estabelecidos para os documentos da coleção;

- (3) obter uma visualização da coleção a partir desses novos índices;
- (4) comparar a visualização obtida em (3) com a visualização obtida em (1).

Deve-se observar que existem diferentes métodos que possibilitam mensurar a qualidade de uma dada visualização, bem como alguns conjuntos de coleções disponibilizadas com o PEx. As próximas seções detalharão esses aspectos do experimento bem como os resultados obtidos.

7.1 Construção de visualizações no PEx

A construção de uma visualização no PEx demanda uma sequência de passos bem definidos, nos quais são efetuadas escolhas dos parâmetros que determinarão as características da visualização. Tais passos são efetuados utilizando uma série de telas do aplicativo, em que uma resposta do usuário é solicitada. A lista abaixo resume as principais etapas do processo:

- (1) indicação da coleção a ser visualizada;
- (2) definição da técnica de projeção e o tipo de cálculo de distância a serem utilizados;
- (3) definição da lista de *stopwords* bem como dos parâmetros que serão utilizados nos cálculos de similaridade;
- (4) Definição da técnica de redução de dimensionalidade e normalização de vetores;
- (5) Definição de parâmetros gerais;
- (6) Definição de parâmetros particulares da técnica de projeção escolhida.

A partir dessas definições a projeção é construída e mostrada na tela do sistema. Deve-se observar que para os fins propostos nesse trabalho, vários dos parâmetros necessários à construção da visualização foram utilizados aproveitando-se os valores padrão (*default*) oferecidos pelo sistema. A inclusão do para-analisador no código fonte do sistema deu origem a uma opção a mais a ser escolhida no passo 4, conforme tela mostrada a seguir.

Figura 34 – Tela do PEx com destaque da opção da matriz termos x documentos utilizando o para-analisador
Fonte: Reproduzida do sistema Projection Explorer

A opção TF-IDF, mostrada na figura 34, efetua os cálculos de similaridade, através da utilização da matriz termos X documentos, conforme a definição original do MEV, descrita na seção 3.5. Por outro lado, a opção Para TF-IDF, destacada em vermelho na figura 34, efetua a montagem da matriz conforme descrito detalhadamente na seção 7.4.

7.2 Coleções de teste utilizadas

A principal coleção de teste utilizada é chamada pelos criadores do PEx de CBR-ILP-IR-SON. Este nome é composto pelas siglas dos assuntos dos documentos que compõem a coleção, a saber: CBR – *Case-based reasoning* (raciocínio baseado em casos); ILP – *Inductive logic programming* (Programação lógica indutiva); IR – *Information Retrieval* (Recuperação de informação); e SON – *Sonification* (Sonificação). Essa coleção é composta por 675 documentos. De acordo com Paulovich (2008), os documentos dos assuntos CBR e ILP foram retirados de periódicos dessas áreas, enquanto que os assuntos IR e SON foram obtidos por meio de buscas na Internet. Os documentos foram pseudo-classificados, ou seja,

classificados de maneira aproximada, de acordo com a fonte de onde foram retirados (MINGHIM; LEVKOVITS, 2007).

Para essa coleção, a aplicação dos procedimentos para representação no MEV dá origem a um espaço com 1.423 dimensões (PAULOVICH, 2008), sendo esta, portanto, a quantidade total de termos extraídos dos documentos da coleção.

Uma análise preliminar da coleção foi efetuada por Paulovich (2008) e uma das conclusões obtidas refere-se à não homogeneidade do conjunto de documentos de IR (*Information retrieval* ou recuperação de informação). Segundo este autor, isto é consequência do fato de que os documentos classificados para este assunto foram obtidos por buscas genéricas na Internet, utilizando-se das palavras “information” and “retrieval” e que, nesse caso, elas não são suficientes para identificar precisamente a área. Uma consequência desta conclusão é que se espera que os objetos componentes deste conjunto de documentos se apresentem, na visualização, mais espalhados que os demais, não compondo um grupo separado e nem se misturando muito aos objetos dos outros assuntos.

Com o objetivo de diversificar os testes a serem efetuados, decidiu-se por construir outras coleções, a partir da coleção original. Esse procedimento foi efetuado realizando-se agrupamentos dos documentos em subgrupos da coleção original. Foram mantidas as abreviaturas que identificam os assuntos da coleção original e, dessa forma, foram obtidas as seguintes coleções:

- (i) CBR-ILP-IR, com 574 documentos
- (ii) CPR-ILP-SON, com 496 documentos
- (iii) CBR-ILP, com 395 documentos
- (iv) IR-SON, com 280 documentos.

Todas as coleções são construídas a partir dos títulos, resumos e referências contidas em cada documento. A coleção é armazenada em um arquivo compactado.

7.3 Avaliação de visualizações

O PEx possui algumas funcionalidades que permitem avaliar as projeções efetuadas pelo sistema. Pode-se afirmar que avalia-se, indiretamente, a qualidade da visualização gerada. Basicamente o sistema utiliza duas técnicas de avaliação: *neighborhood hit* e *neighborhood preservation*.

A técnica *neighborhood hit* objetiva analisar se, para uma pré-classificação efetuada em uma coleção, é possível identificar na projeção ou visualização construída a separação entre as diferentes classes do conjunto original de dados. Assim, quanto mais separados e agrupados os pontos estiverem na visualização gerada, de acordo com as classes originais, maior será a precisão. Este procedimento permite avaliar, numericamente, o quanto estão destacadas, na visualização final, as classes pré-existent e a facilidade de encontrar limites bem definidos entre elas. Por outro lado, esta técnica não permite avaliar a qualidade dos grupos gerados e nem o quanto as relações de distância foram preservadas na visualização final. Assim, foi desenvolvida uma segunda técnica, chamada de *neighborhood preservation*. Essa, por sua vez, avalia a preservação da vizinhança dos objetos multi-dimensionais na visualização final (PAULOVICH, 2008).

Para avaliação de coleções compostas de grupos de documentos de diferentes assuntos, Minghim (informação pessoal)¹⁰ sugere a utilização do índice *neighborhood hit*.

O índice *neighborhood hit* materializa-se por meio de um gráfico que avalia a precisão de acordo com o número de “vizinhos” analisados. A figura 35, a seguir, é um exemplo de como é efetuada uma comparação com a utilização do índice citado. As curvas de diferentes cores avaliam o índice para situações diferentes. No exemplo mostrado abaixo, a situação representada pela curva vermelha é considerada a pior de todas, pois possui os menores valores de precisão calculados (eixo vertical) para as quantidades de vizinhos especificadas (eixo horizontal). Analogamente, a curva azul é a que representa a situação com melhor avaliação de *neighborhood hit*.

Figura 35 – Exemplo de curvas de Neighborhood Hit
Fonte – Produzida com a utilização do Projection Explorer

¹⁰ Minghim, R. Comunicação efetuada em Dezembro de 2010.

7.4 Uso do para-analisador no Projection Explorer – resultados obtidos

7.4.1 Teste inicial

A idéia de utilizar o algoritmo para-analisador nos procedimentos de indexação, conforme descrito na seção 4.3.3, foi sendo reavaliada e modificada à medida que os testes efetuados no PEx foram sendo realizados.

Num primeiro momento, o algoritmo foi utilizado para testar a qualidade dos índices estabelecidos pelos procedimentos originais do sistema, levando-se em conta a proposição efetuada na seção 4.3.3 (o termo t é um bom termo para indexar o documento D). Assim, para este teste, considerou-se como fator de crença para avaliação da proposição o valor tf e para fator de descrença o valor df (ou seja, o inverso do idf). Esta pareceu ser a solução natural para utilização da para-análise, pois se um termo ocorre frequentemente em um documento, então ele deve ser um bom termo para indexação. Por outro lado, se o termo ocorre frequentemente na coleção, ele não deve ser um bom termo para indexar o documento. Nessa situação, a idéia foi eliminar, em cada documento analisado, os termos de indexação que não alcançassem o estado verdadeiro e suas tendências quando submetidos à para-análise, a partir dos fatores de crença e descrença estipulados. Nesse caso, a eliminação foi efetuada atribuindo o valor 0 (zero) como peso para os termos que não alcançassem os estados lógicos descritos.

Os resultados obtidos com esse procedimento foram sensivelmente piores do que aqueles obtidos com o procedimento original do PEx.

Nesse caso, o problema surgiu, segundo Paulovich (informação pessoal)¹¹ devido à eliminação dos índices. Para este autor, tal procedimento claramente prejudicaria os cálculos de similaridade/distância entre os documentos, deteriorando, assim, a visualização. O pesquisador sugeriu que fossem observados os histogramas de distâncias produzidos para cada visualização.

7.4.2 Segunda bateria de testes

¹¹ Paulovich, F. Comunicação pessoal efetuada em Novembro de 2010.

A partir dos resultados descritos na seção anterior, elaborou-se uma nova estratégia de utilização do para-analisador.

Conforme descrito na seção 5.2, o MEV é o modelo básico utilizado para elaboração dos cálculos de similaridade que irão definir a apresentação final (*layout*) da visualização. Nesse modelo, conforme exposto na seção 3.5, é montada uma matriz documentos - termos, que contém os pesos calculados para cada termo, de acordo com a fórmula $tf * idf$. Essa matriz é utilizada pelo PEx para efetuar os cálculos de similaridade entre os documentos.

De acordo com Salton e McGill (1983), os pesos calculados para um termo, em um dado documento, representam o grau de utilidade do termo para este documento. É interessante observar que este grau varia de acordo com o valor tf do termo em questão, uma vez que o idf representa a ocorrência do termo no conjunto da coleção e, dessa forma, terá o mesmo valor para todos os documentos em que o termo ocorre.

Para esta sequência de testes optou-se por atribuir uma penalização para o peso do termo (ou seu grau de utilidade), de acordo com a região do QUPC que a para-análise efetuada com os graus de crença (tf) e descrença (df) indicar. Os pesos originalmente calculados foram mantidos apenas para os termos cuja análise indicasse a região do totalmente verdadeiro. Os fatores de penalização foram arbitrados considerando-se uma diminuição de 0,15 em relação ao valor 1 (atribuído à região do totalmente verdadeiro), à medida que a região se afasta da região do totalmente verdadeiro.

Os fatores de penalização, arbitrados de acordo com a região do QUPC, são:

V – Verdadeiro == 1

T – Inconsistente == 0,25

F – Falso == 0,25

$\perp$ – Indeterminado == 0,25

$Q_f \rightarrow T$ – Quase Falso, tendendo ao Inconsistente == 0,4

$Q_f \rightarrow \perp$ – Quase Falso, tendendo ao Indeterminado == 0,4

$\perp \rightarrow f$ – Indeterminado, tendendo ao Falso == 0,55

$T \rightarrow f$ – Inconsistente, tendendo ao Falso == 0,55

$T \rightarrow v$ – Inconsistente, tendendo ao Verdadeiro == 0,70

$\perp \rightarrow v$ – Indeterminado, tendendo ao Verdadeiro == 0,70

$Q_v \rightarrow T$ – Quase Verdadeiro, tendendo ao Inconsistente == 0,85

$Q_v \rightarrow \perp$ – Quase Verdadeiro, tendendo ao Indeterminado == 0,85

Assim, supondo-se que o cálculo do peso de um termo, por meio da fórmula $tf * idf$, é igual a 100; por outro lado, se a para-análise indicar a região $Q_v \rightarrow \perp$ (quase verdadeiro, tendendo ao indeterminado), o peso a ser atribuído ao termo passa a ser $100 * 0,85$, ou seja, 85.

O objetivo é observar os efeitos que a alteração no peso dos termos provocará na construção das visualizações das coleções.

Além da penalização dos termos, os testes foram efetuados modificando-se os tamanhos das regiões do QUPC. Essa modificação é efetivada pela manipulação das variáveis V_{scc} (valor superior de controle de certeza), V_{icc} (valor inferior de controle de certeza), V_{scct} (valor superior de controle de contradição) e V_{icct} (valor inferior que limita o grau de contradição próximo ao indeterminado), conforme descrito na seção 4.3.2. As regiões utilizadas nos testes, e seus respectivos valores limite, são:

Região 0 (R0)

$$V_{scc} = 1/2 ; V_{icc} = -1/2 ; V_{scct} = 3/4; V_{icct} = -3/4$$

Região 1 (R1)

$$V_{scc} = 1/2; V_{icc} = -3/4; V_{scct} = 3/4; V_{icct} = -3/4$$

Região 2 (R2)

$$V_{scc} = 3/4; V_{icc} = -3/4; V_{scct} = 3/4; V_{icct} = -3/4$$

Região 3 (R3)

$$V_{scc} = 1/2; V_{icc} = -1/2; V_{scct} = 1/2; V_{icct} = -1/2$$

O procedimento anteriormente descrito, de alterar os pesos e modificar o tamanho das regiões do QUPC, produz visualizações diferentes daquelas produzidas com o cálculo tradicional de similaridades. A sistemática original, baseada no MEV, não foi modificada. Modificou-se apenas o cálculo dos pesos dos termos. A observação dos efeitos produzidos é efetuada calculando-se os valores de *neighborhood hit* obtidos nas visualizações construídas, comparando os resultados para uma mesma coleção.

Os procedimentos para a construção das visualizações foram padronizados de acordo com os passos exigidos no PEx, com os seguintes parâmetros:

Tipo de projeção – *Least Square Projection* (LSP)

Distancia – *Cosine-based* (baseada em cosseno – a distancia padrão do MEV)

Corte de Luhn – foram excluídos os termos que se encontram nos extremos das estatísticas. Ou seja, o(s) de menor ocorrência e o de maior ocorrência. Esta escolha esta baseada nas colocações de Salton e McGill (1988) que consideram que os melhores termos de indexação são os que apresentam valores intermediários de ocorrência.

Para os outros parâmetros passíveis de escolha foram adotados os valores padrão (*default*) oferecidos pelo sistema

Os testes foram realizados para cada uma das coleções descritas na seção 7.2. Construíram-se visualizações para cada coleção, utilizando-se as regiões R0, R1, R2 e R3. Depois disso, produziu-se um gráfico com os valores de *neighborhood hit* obtidos para cada região e para a visualização construída com o procedimento padrão do PEx. As visualizações obtidas podem ser vistas no apêndice B deste documento.

Os resultados obtidos são mostrados a seguir. As curvas são identificadas pelos nomes da coleção seguidas dos sufixos: *orig*, para o procedimento original do PEx; e *para-Rx*, para as visualizações obtidas com o uso do para-analisador na região Rx, onde x pode ser 0, 1, 2 ou 3, conforme descrito anteriormente.

Coleção CBR-ILP-IR-SON

Figura 36 – Comparação de *neighborhood hit*
Fonte: Visualização produzida com o Projection Explorer

Coleção CBR-ILP-IR

Figura 37 – Comparação de *neighborhood hit*
Fonte: Visualização produzida com o Projection Explorer

Coleção CBR-ILP-SON

Figura 38 – Comparação de *neighborhood hit*
Fonte: Visualização produzida com o Projection Explorer

Coleção IR-SON

Figura 39 – Comparação de *neighborhood hit*
Fonte: Visualização produzida com o Projection Explorer

Coleção CBR-ILP

Figura 40 – Comparação de *neighborhood hit*
Fonte: Visualização produzida com o Projection Explorer

Como pode ser observado nas figuras anteriores, o procedimento de penalizar os valores dos pesos através da para-análise permitiu melhorar a avaliação das visualizações originais em todas as coleções testadas. Contudo, não se encontrou um padrão quanto à região que obteve o melhor resultado. Apresenta-se, a seguir, a compilação dos resultados obtidos para cada coleção.

Coleção CBR-ILP-IR-SON

Melhor resultado: curva modificada e análise efetuada com a região R0

Pior resultado: curva original

Coleção CBR-ILP-IR

Melhor resultado: curva modificada e análise efetuada com as regiões R2 e R3

Pior resultado: curva original

Coleção CBR-ILP-SON

Melhor resultado: curva modificada e análise efetuada com a região R3

Pior resultado: curva original

Coleção IR-SON

Melhor resultado: curva modificada e análise efetuada com a região R3

Pior resultado: curva modificada e análise efetuada com a região R2

Coleção CBR-ILP

Melhor resultado: curva modificada e análise efetuada com a região R0

Pior resultado: curva modificada e análise efetuada com a região R3

7.4.3 Análise dos resultados obtidos – qualidade da visualização

Antes de proceder à análise dos resultados obtidos, algumas considerações são necessárias:

- (i) Conforme dito anteriormente, os documentos dos assuntos IR (*Information retrieval*) e SON (*Sonification*) não apresentam o mesmo grau de pureza quando comparados com os documentos de CBR (*Case-Based Reasoning*) e ILP (*Inductive Logic Programing*), pois aqueles foram obtidos a partir de consultas efetuadas na Internet, ao passo que estes foram selecionados de periódicos das respectivas áreas;
- (ii) Dessa forma, é razoável supor que, para os documentos dos assuntos IR e SON, os termos de indexação extraídos dos documentos apresentam menor grau de

representatividade do que aqueles extraídos para os assuntos CBR e ILP, visto que para dois primeiros conjuntos, o assunto foi determinado a partir do repositório em que os documentos foram encontrados. Por outro lado, pode-se considerar que os documentos dos assuntos CBR e ILP, por terem sido extraídos de periódicos das áreas, utilizam com maior rigor a terminologia de suas respectivas áreas e, dessa forma, os termos extraídos apresentam maior grau de representatividade dos assuntos.

- (iii) Pode-se considerar que as coleções CBR-ILP-IR-SON, CBR-ILP-IR, CBR-ILP-SON e IR-SON são coleções que apresentam em maior ou menor quantidade, certo grau de ruído. Sendo que, nesse caso, ruído deve ser entendido como documentos que não foram classificados da melhor maneira possível;
- (iv) Por outro lado, a coleção CBR-ILP pode ser considerada como a coleção com maior grau de pureza, pois foi criada com documentos extraídos de repositórios que garantem a aderência aos assuntos determinados para esses documentos.

Pode-se observar, nos gráficos de *neighborhood hit* obtidos, que o efeito de modificar os pesos dos termos é mais forte nas coleções que possuem, com maior ou menor grau, algum ruído (documentos classificados sem rigor). Por outro lado, também pode-se observar que a coleção CBR-ILP foi aquela em que os gráficos ou ficaram muito próximos da curva original (regiões R0, R1 e R2), ou muito aquém da mesma (região R3).

Os documentos da coleção CBR-ILP são os documentos bem classificados e, de acordo com seção 7.2, seu quantitativo é de 395. Analogamente, pode-se calcular o quantitativo de documentos ruído nas outras coleções, verificando-se a diferença entre os dois tipos de documentos. Nesse caso, tem-se os seguintes valores:

Coleção CBR-ILP-IR-SON

total de documentos = 675

documentos bem classificados = 395

documentos ruído = 280

Coleção CBR-ILP-IR

total de documentos = 574

documentos bem classificados = 395

documentos ruído = 179

Coleção CBR-ILP-SON

total de documentos = 496

documentos bem classificados = 395

documentos ruído = 101

Nas considerações anteriores, pode-se observar que os maiores efeitos sobre o índice de *neighborhood hit* foram obtidos nas coleções CBR-ILP-IR-SON e CBR-ILP-IR que são, respectivamente, as coleções com maior quantidade de documentos ruído. Para a coleção CBR-ILP-SON, a utilização da para-análise modificou os valores do *neighborhood hit*, mas as curvas resultantes se mantiveram, em alguns trechos, próximas da curva original. Efeito semelhante pode ser observado para o gráfico da coleção IR-SON composta, exclusivamente, por documentos ruído. Contudo, como a coleção IR-SON não é constituída de documentos bem classificados os gráficos obtidos para a mesma devem ser observados isoladamente das outras coleções.

Por outro lado, para a coleção CBR-ILP, com exceção da região R3 que se situa muito abaixo da curva original, as curvas resultantes se posicionaram extremamente próximas da curva original, ou seja, os efeitos da alteração de peso são mínimos quando comparados com a visualização original.

Essas considerações permitem concluir que o procedimento de modificar os pesos dos termos de indexação está, por meio da curva de *neighborhood hit*, avaliando a qualidade das coleções sob o ponto de vista da classificação dos assuntos dos documentos que a compõem. Ou seja, a utilização da para-análise e do índice de *neighborhood hit* fornece uma maneira indireta de avaliar a qualidade com que os documentos de uma coleção foram classificados.

Além disso, o procedimento está, aparentemente, avaliando a qualidade dos agrupamentos produzidos pela visualização das diferentes coleções.

Resumidamente, para os testes efetuados, quanto maior o efeito produzido na curva de *neighborhood hit*, maior a quantidade de documentos ruído na coleção.

Nas próximas seções serão verificadas algumas técnicas que permitem avaliar a qualidade dos agrupamentos obtidos nas visualizações das diferentes coleções. Tais técnicas estão disponíveis para utilização no próprio Projection Explorer.

7.4.4 Avaliação de agrupamentos

Vários métodos foram desenvolvidos para avaliar a qualidade dos agrupamentos produzidos pelos diferentes algoritmos. Muitos desses métodos são baseados em análises estatísticas.

Nessa seção serão apresentados dois métodos que estão disponíveis para utilização no sistema Projection Explorer. Um dos métodos é um índice obtido numericamente e o outro uma representação gráfica. São eles: o coeficiente de silhueta (*silhouette coefficient*) e a representação gráfica da matriz de similaridades.

7.4.4.1 Coeficiente de silhueta

O coeficiente de silhueta de um agrupamento é obtido a partir do cálculo do valor de silhueta de cada objeto pertencente aos diferentes grupos que compõem o agrupamento. Depois disso, uma média do valor é calculada e atribuída a cada grupo e, finalmente, as médias obtidas para cada grupo são utilizadas para calcular o coeficiente do agrupamento. O coeficiente de silhueta é descrito, com pequenas variações, em Kaufman e Rosseeuw (1990) e em, Tan, Steibach e Kumar (2006). O coeficiente de silhueta é um valor que se situa entre -1 (o pior valor), e 1 (o melhor valor).

O valor da silhueta de cada objeto é obtido da seguinte forma:

- 1) para cada objeto pertencente a um determinado grupo, calcular a média de suas distâncias (ou similaridades) a todos os outros objetos de seu próprio grupo. Esse valor é chamado de a_i .
- 2) para cada objeto pertencente a um determinado grupo, calcular a média de suas distâncias (ou similaridades) a todos os outros objetos pertencentes aos outros grupos que compõem o agrupamento. Achar o valor mínimo dessas medidas. Esse valor é chamado de b_i .
- 3) para cada objeto, seu valor de silhueta é dado pela fórmula:

$$s_i = (b_i - a_i) / \text{Max}(a_i, b_i)$$

Além de definirem o coeficiente de silhueta, Kaufman e Rosseeuw (1990) produziram uma tabela em que a qualidade de um agrupamento pode ser avaliada de acordo com valor do coeficiente obtido para o mesmo. Reproduz-se, abaixo, a tabela definida pelos autores.

Coefficiente de Silhueta	Interpretação
0,71 – 1,00	Uma estrutura forte foi encontrada
0,51 – 0,70	Uma estrutura razoável foi encontrada
0,26 – 0,50	A estrutura é fraca e pode ser artificial. Tente métodos adicionais para este conjunto de dados.
$\leq 0,25$	Nenhuma estrutura substancial foi encontrada

Tabela 1 – Interpretação dos valores do coeficiente de silhueta

Fonte: adaptada de Kaufman e Rosseeuw (1990)

Os resultados a seguir foram obtidos no Projection Explorer, verificando-se os valores dos coeficientes de silhueta para as visualizações construídas. Variou-se o cálculo para cada coleção e para as regiões do QUPC em que o para-analisador foi utilizado para modificar os pesos originalmente calculados. A expressão “original” refere-se às visualizações construídas com o procedimento original do Projection Explorer, ou seja, sem a utilização do para-analisador.

Coefficiente de silhueta

Coleção / Região	Original	R0	R1	R2	R3
CBR-ILP-IR-SON	0,05844	0,18286	0,28807	0,20883	-0,04546633
CBR-ILP-IR	0,19003	0,23321	0,32509	0,65129	0,22175422
CBR-ILP-SON	0,41562	0,41507	0,57681	0,51775	0,6606122
CBR-ILP	0,67798	0,66479	0,48078	0,74633	0,010869439
IR-SON	0,4099	0,40576	0,39995	0,54973	-0,040663127

Tabela 2 – Coeficientes de silhueta das coleções, de acordo com a região do QUPC

Fonte: obtido com a utilização do Projection Explorer

Utilizando as tabelas 1 e 2, anteriormente definidas, pode-se observar que a utilização do para-analisador possibilitou modificar a classificação dos agrupamentos obtidos em todas as coleções.

O gráfico a seguir permite observar, visualmente, os efeitos produzidos no coeficiente de silhueta.

Comparação dos coeficientes de silhueta obtidos

Gráfico 1 – Comparação dos coeficientes de silhueta obtidos

Fonte: Produzido com o sistema Excel

Como pode ser observado, o procedimento de alterar os pesos dos termos de indexação com a utilização do para-analisador permitiu melhorar, para certas regiões, o valor do coeficiente de silhueta. Como esse coeficiente é utilizado para avaliar a qualidade dos agrupamentos obtidos e como nenhuma outra alteração foi efetuada nos parâmetros que produzem as visualizações no Projection Explorer, pode-se concluir que, sob certas condições (a região do QUPC utilizada), a modificação dos pesos permite melhorar a qualidade dos agrupamentos obtidos.

7.4.4.2 Matriz de similaridades

Outra maneira de observar a qualidade dos agrupamentos produzidos consiste em visualizar a matriz de similaridades que pode ser construída para os conjuntos de objetos que compõem os grupos de um agrupamento.

Num primeiro momento, produz-se uma matriz numérica, em que as colunas e as linhas são compostas dos objetos que compõem todo o agrupamento. Nesse caso, algum tipo de identificador é associado a cada objeto (geralmente um número). Na matriz propriamente dita (ou seja, na junção linha-coluna), é colocado o valor numérico que representa a similaridade (ou distância) entre os objetos. Para que possa ser produzida uma visualização

dessa matriz, substituem-se os valores numéricos por, por exemplo, tons de cinza, de modo que uma gradação dos tons, do mais claro ao mais escuro, indica o quanto os objetos estão mais distantes ou próximos entre si. Nesse caso, tons claros indicam baixa similaridade (ou alta distância) e tons escuros indicam alta similaridade (ou baixa distância).

O procedimento anteriormente descrito permite construir uma figura em que os grupos de um agrupamento (nesse caso, o agrupamento se traduz pela coleção de documentos visualizada) são representados por blocos mais ou menos compactos, de acordo com as similaridades calculadas para seus componentes. A figura 41, a seguir, exemplifica esse procedimento, aplicado aos agrupamentos mostrados na figura 33.

Figura 41 – Exemplos de matrizes de similaridades para agrupamentos produzidos por diferentes algoritmos
Fonte - Fonte - Tan, Steibach e Kumar (2006, pg 545)

É interessante observar que as figuras reproduzem, proporcionalmente, os tamanhos dos agrupamentos produzidos pelos diferentes algoritmos utilizados. Ao lado de cada figura está colocada uma escala de tons que permite verificar, visualmente o comportamento das similaridades entre os objetos que compõem os diferentes grupos. Assim, grupos cujos elementos possuem maior grau de similaridade apresentam blocos com tonalidades mais escuras, enquanto que grupos cujos elementos apresentam menor grau de similaridade apresentam tonalidades que se aproximam mais dos tons de cinza. No exemplo da figura 41, as matrizes produzidas pelo algoritmo *k-means* são as que apresentam blocos mais coesos, sob o ponto de vista do “grau de escurecimento” observável.

A seguir, são reproduzidos os resultados obtidos no Projection Explorer para as diferentes coleções de teste utilizadas. Nas figuras apresentadas, as cores laterais indicam os

assuntos dos documentos que compõem as coleções. As cores são utilizadas para propiciar a identificação dos grupos obtidos.

Coleção CBR-ILP-IR-SON

Legenda de cores para esta coleção:

Azul – CBR (*Case based reasoning*)
Vermelho – SON (*Sonification*)
Lima – IR (*Information Retrieval*)
Verde – ILP (*Inductive Logic Programming*)

Figura 42 – Matriz de similaridades original
 Fonte: produzida no PEx

Figura 43 – Matriz de similaridades - Região R0 do QUPC
 Fonte: produzida no PEx

Figura 44 – Matriz de similaridades - Região R1 do QUPC
 Fonte: produzida no PEx

Figura 45 – Matriz de similaridades - Região R2 do QUPC
 Fonte: produzida no PEx

Figura 46 – Matriz de similaridades - Região R3 do QUPC
 Fonte: produzida no PEx

Para esta coleção podem ser observadas modificações nas representações dos grupos que compõem a coleção. Particularmente, para o agrupamento obtido na região R1 (aquela de maior coeficiente de silhueta) pode-se observar uma intensificação na tonalidade obtida para todos os grupos (assuntos) que compõem a coleção. Assim, pode-se considerar que os grupos obtidos na região R1 são mais coesos do que aqueles obtidos com a visualização original.

Neste caso, a visualização da matriz de similaridades é compatível com o valor do coeficiente de silhueta obtido para o agrupamento.

Coleção CBR-ILP-IR

Legenda de cores para esta coleção:

Azul – CBR (*Case based reasoning*)
Vermelho – IR (*Information Retrieval*)
Verde – ILP (*Inductive Logic Programming*)

Figura 47 – Matriz de similaridades original
 Fonte: produzida no PEX

Figura 48 – Matriz de similaridades - Região R0 do QUPC
 Fonte: produzida no PEX

Figura 49 – Matriz de similaridades - Região R1 do QUPC
 Fonte: produzida no PEX

Figura 50 – Matriz de similaridades - Região R2 do QUPC
 Fonte: produzida no PEX

Figura 51 – Matriz de similaridades - Região R3 do QUPC
 Fonte: produzida no PEX

Nesta coleção também pode ser observada a intensificação de tonalidades em várias regiões. Contudo, não é possível fazer o mesmo tipo de associação efetuado na coleção anterior (coeficiente de silhueta - matriz de similaridade). De qualquer forma, pode-se observar que a região para a qual foi obtido o maior valor de coeficiente de silhueta (R2) também apresenta maior intensificação nas tonalidades, quando comparada com a região

original. Interessante observar que essa região também apresenta uma melhor avaliação de *neighborhood hit*, quando comparada com a visualização original.

Coleção CBR-ILP-SON

Legenda de cores para esta coleção:

Azul – CBR (*Case based reasoning*)
 Verde – ILP (*Inductive Logic Programing*)
 Vermelho – SON (*Sonification*)

Figura 52 – Matriz de similaridades original
 Fonte: produzida no PEx

Figura 53 – Matriz de similaridades - Região R0 do QUPC
 Fonte: produzida no PEx

Figura 54 – Matriz de similaridades - Região R1 do QUPC
 Fonte: produzida no PEx

Figura 55 – Matriz de similaridades - Região R2 do QUPC
 Fonte: produzida no PEx

Figura 56 – Matriz de similaridades - Região R3 do QUPC
 Fonte: produzida no PEx

Também podem-se observar variações de tonalidade para as matrizes obtidas nessa coleção, com destaque para as matrizes das regiões R0 e R1. Contudo, assim como para a coleção anterior, não se observa correlação entre a visualização com o melhor valor de coeficiente de silhueta (região R3) e das matrizes visualizadas.

Coleção IR-SON

Legenda de cores para esta coleção:

Azul – IR (*Information Retrieval*)

Vermelho – SON (*Sonification*)

Figura 57 – Matriz de similaridades original
Fonte: produzida no PEx

Figura 58 – Matriz de similaridades - Região R0 do QUPC
Fonte: produzida no PEx

Figura 59 – Matriz de similaridades - Região R1 do QUPC
Fonte: produzida no PEx

Figura 60 – Matriz de similaridades - Região R2 do QUPC
Fonte: produzida no PEx

Figura 61 – Matriz de similaridades - Região R3 do QUPC
Fonte: produzida no PEx

Esta coleção também apresenta modificações nas intensidades das tonalidades que representam as matrizes de similaridades. Neste caso, podem-se destacar as regiões R0 e R1. Contudo, como esta coleção é composta pelos documentos ruído, não serão feitas maiores análises sobre os efeitos obtidos.

Coleção CBR-ILP

Legenda de cores para esta coleção:

Azul – CBR (*Case based reasoning*)

Vermelho – ILP (*Inductive Logic Programming*)

Figura 62 – Matriz de similaridades original
Fonte: produzida no PEx

Figura 63 – Matriz de similaridades - Região R0 do QUPC
Fonte: produzida no PEx

Figura 64 – Matriz de similaridades - Região R1 do QUPC
Fonte: produzida no PEx

Figura 65 – Matriz de similaridades - Região R2 do QUPC
Fonte: produzida no PEx

Figura 66 – Matriz de similaridades - Região R3 do QUPC
Fonte: produzida no PEx

Essa é a coleção composta exclusivamente pelos documentos bem classificados, conforme anteriormente detalhado. Também podem ser observadas variações nas intensidades das tonalidades das matrizes de similaridade. Nesse caso, os efeitos visuais são mais intensos para as regiões R0 e R1, com destaque para o grupo de assunto ILP (*inductive logic programming*).

7.4.5 Histograma de distâncias

Alem da matriz de similaridade, descrita na seção anterior, o Projection Explorer possui outro método para observação visual das características das visualizações criadas. É uma ferramenta estatística, chamada de histograma de distâncias.

Um histograma é um gráfico, que permite observar as distribuições de frequência de valores que ocorrem numa classe de variáveis observadas. Um histograma é construído a partir de dois eixos coordenados. No eixo horizontal são colocados os valores individuais da variável em estudo, no nosso caso, as distâncias. O eixo vertical apresenta a escala onde são lidos os valores relativos aos números de observações ou, mais comumente, as frequências de classe (TOLEDO; OVALLE, 1986).

Para as visualizações de coleções de documentos, construídas com o Projection Explorer, é possível observar a distribuição das frequências de distâncias que ocorrem para a coleção como um todo. A figura 67, a seguir, é um exemplo histograma de distância obtido no Projection Explorer.

Figura 67 – Exemplo de histograma de distancias
Fonte: produzido no PEx

No exemplo mostrado pode-se observar que a maioria das ocorrências de distâncias, entre os documentos que compõem a coleção, ocorrem no intervalo $[0,9;1,0]$. Também pode-se observar que existe um pico de 25000 ocorrências para determinado valor de distância.

Os histogramas de distâncias obtidos para as coleções utilizadas indicam uma redistribuição das frequências. Podem-se observar, neste caso, dois efeitos: (i) uma diminuição da concentração de distâncias, numa mesma classe, quando comparada com as visualizações criadas com o cálculo tradicional de tf-idf, ou seja, o histograma se apresenta mais espalhado e, (ii) uma diminuição no valor dos picos de frequências. Houve uma diminuição dos valores máximos de ocorrências de distância (ou similaridade) calculados para os documentos. A ausência de análises mais profundas sobre esses resultados impede a elaboração de maiores conclusões. Nas figuras, a seguir, são apresentados os histogramas de

distâncias para as visualizações originais e aquelas que apresentaram, bem visivelmente, os efeitos anteriormente descritos.

Os resultados das avaliações, realizadas com os métodos do Projection Explorer, bem como as respectivas visualizações das coleções, foram compilados, para cada coleção e cada região do QUPC e podem ser vistos no Apêndice B.

Coleção CBR-ILP-IR-SON

Visualização original

Figura 68 – Histograma de distâncias
Fonte: produzido com o PEx

Visualização usando região R0 do QUPC

Figura 69 – Histograma de distâncias
Fonte: produzido com o PEx

Coleção CBR-ILP-IR

Visualização original

Figura 70 – Histograma de distâncias
Fonte: produzido com o PEx

Visualização usando região R2 do QUPC

Figura 71 – Histograma de distâncias
Fonte: produzido com o PEx

Coleção CBR-ILP-SON

Visualização original

Figura 72 – Histograma de distâncias
Fonte: produzido com o PEx

Visualização usando região R3 do QUPC

Figura 73 – Histograma de distâncias
Fonte: produzido com o PEx

Coleção IR-SON

Visualização original

Figura 74 – Histograma de distâncias
Fonte: produzido com o PEx

Visualização usando região R2 do QUPC

Figura 75 – Histograma de distâncias
Fonte: produzido com o PEx

Coleção CBR-ILP

Visualização original

Figura 76 – Histograma de distâncias
Fonte: produzido com o PEx

Visualização usando região R2 do QUPC

Figura 77 – Histograma de distâncias
Fonte: produzido com o PEx

8. Considerações finais

O objetivo principal desta pesquisa - verificar o potencial e os efeitos da utilização da lógica paraconsistente em procedimentos de indexação automática -, direcionou os trabalhos de modo a buscar interações entre diversas disciplinas ou campos de estudo, tais como: recuperação de informação, indexação automática, lógicas não-clássicas e visualização de informações.

O teste da hipótese de trabalho que direcionou o desenvolvimento do estudo, a utilização de métodos para tratamento de incerteza e vagueza nos processos de indexação, se materializou por meio de dois procedimentos básicos. O primeiro deles consistiu em modificar o cálculo tradicional de pesos, efetuado pelo modelo do espaço vetorial. Esse modelo, originalmente desenvolvido para utilização em sistemas de recuperação de informação, mas também utilizado em sistemas de visualização de informação, engloba tanto conceitualmente quanto operacionalmente, os procedimentos básicos de indexação automática. O modelo constrói sua solução baseado em pressupostos linguísticos e estatísticos. Nesse contexto, utilizou-se o algoritmo para-analisador, desenvolvido no âmbito da lógica paraconsistente, para modificar os valores dos pesos atribuídos aos termos de indexação escolhidos. Dessa forma, aplicou-se um método desenvolvido para tratamento de situações em que a incerteza e a imprecisão são inerentes, a um processo de indexação automática.

O segundo procedimento consistiu em observar e mensurar os efeitos da modificação no cálculo dos pesos em um sistema de visualização de informação – o Projection Explorer, uma vez que esse sistema utiliza o modelo do espaço vetorial para determinar o grau de similaridade entre os documentos da coleção a ser visualizada.

A observação e mensuração dos efeitos foram efetivadas utilizando-se critérios embutidos no próprio Projection Explorer. Tais critérios são utilizados para: (i) avaliar as visualizações criadas com o aplicativo e (ii) avaliar a qualidade dos agrupamentos que se materializam na construção das visualizações.

Os resultados explicitados nas seções 7.4.3 e 7.4.4.1 demonstram que o uso do para-analisador permitiu ganhos quantificáveis tanto nas avaliações das visualizações como um todo (por meio do índice *neighborhood hit*), quanto na avaliação dos agrupamentos que representam as visualizações (por meio do coeficiente de silhueta). Além disso, observaram-

se os efeitos produzidos nas matrizes de similaridade. Contudo, nesse caso, os efeitos não são quantificáveis. Também observaram-se modificações nas distribuições de frequência representadas nos histogramas de distâncias obtidos.

O para-analisador é um algoritmo que permite o tratamento de dados (ou informações) e situações que não se ajustam, perfeitamente, à dicotomia rígida da lógica clássica (verdadeiro ou falso). Por outro lado, deve-se observar que a atribuição dos valores de pesos aos termos segue um procedimento estatístico, baseado em sua frequência nos documentos e na coleção como um todo. Tal procedimento, a princípio, não possui nenhum grau de incerteza, uma vez que se trata de uma contagem de ocorrências.

Os efeitos produzidos nos agrupamentos, demonstrados nos valores obtidos nos coeficientes de silhueta e observados nas matrizes de similaridade, indicam que o uso do para-analisador, sob as condições do experimento efetuado, tem a capacidade de gerar agrupamentos mais efetivos. Esse é um ponto que chama a atenção, uma vez que a formação de bons agrupamentos pode ser utilizada para aperfeiçoar procedimentos de recuperação de informação quando se levam em consideração os efeitos da hipótese do *cluster*: de que documentos similares tendem a ser relevantes para uma questão formulada por um usuário a um sistema de recuperação de informação (VAN RIJSBERGEN, 1979).

Conforme visto na seção 4.3.1, a Lógica paraconsistente derroga o princípio da não contradição e os métodos dela derivados, tal como o algoritmo para-analisador utilizado nesta pesquisa, fornecem subsídios teóricos e práticos para o tratamento da incerteza e vagueza.

Os resultados obtidos sugerem que a abordagem estatística deve ser relativizada ou, dito de outra forma, confrontada com informação adicional. Os resultados obtidos com a introdução do algoritmo para-analisador no sistema Projection Explorer permitiram relativizar os valores dos pesos conforme descrito na seção 7.4.2.

Assim, a pesquisa indica que a atribuição de pesos aos termos que representam um documento pode ser vista como um procedimento repleto de incertezas e vagueza e, dessa forma, os procedimentos para efetivação dessa atividade deve se apoiar em ferramentas com capacidade para o tratamento desses aspectos, como o para-analisador. O uso desse algoritmo possibilitou, sob as condições do experimento, a melhora no desempenho do sistema Projection Explorer, mantidas as variáveis padrão do sistema, para geração das visualizações.

Para Anderson e Perez-Caballo (2001a), a atribuição de pesos aos termos de indexação é, muitas vezes, feita através de procedimentos de tentativa e erro, sem maiores justificativas teóricas. Os resultados obtidos indicam que o uso do para-analisador, como uma ferramenta que avalia a qualidade dos termos de indexação escolhidos, pode ser visto como um modo de

justificar e relativizar o valor dos pesos dos termos. Ou seja, abre caminho para o desenvolvimento de uma teoria sobre a atribuição de pesos aos termos de indexação nos moldes definidos pelo modelo do espaço vetorial. Além disso, a relativização dos pesos encontra apoio nas colocações de Kaufman e Rousseeuw (1990), que entendem que algumas variáveis, escolhidas para representar os objetos a serem agrupados são, intrinsecamente, mais importantes que outras e, dessa forma, a atribuição de pesos deve se basear no conhecimento específico do assunto ao qual pertencem os objetos.

Pode-se concluir, com base nos argumentos anteriores, que a hipótese inicial de trabalho se confirmou. Ou seja, a utilização de procedimentos para tratamento de incerteza, imprecisão e vagueza tem potencial para produzir ganhos mensuráveis nos processos de indexação automática. Pode-se afirmar, portanto, que o uso de lógicas não-clássicas em sistemas de indexação automática, tal como a Lógica paraconsistente, indica novos caminhos para as pesquisas na área.

8.1 Sugestões para trabalhos futuros

O experimento computacional desenvolvido mostrou o potencial positivo de utilização do para-analisador na indexação automática. Contudo, não foi capaz de explicar o porquê e como ocorreram os efeitos observados. É possível afirmar, porém, que a adoção de parâmetros não-dicotômicos (não-excludentes) estabelece novas possibilidades de relacionar informação. A obtenção de tais conclusões implica um aprofundamento nos aspectos teóricos do modelo do espaço vetorial, da atribuição de pesos aos termos de indexação e na concepção e realização de outros experimentos que capturem, com a devida precisão, os efeitos individuais, nos termos de indexação, do uso do para-analisador.

Os experimentos conduzidos nesta pesquisa observaram os efeitos do uso do para-analisador no conjunto dos termos e não em instâncias isoladas.

Uma profunda compreensão dos efeitos sobre os termos individuais tem grande potencial de utilização nos aspectos teóricos da indexação automática e da recuperação de informação e, conseqüentemente, nos artefatos produzidos a partir deles, tais como os sistemas de recuperação e de visualização de informações.

Referências

ALVARENGA, L. Representação do conhecimento na perspectiva da Ciência da Informação em tempo e espaço digitais. **Encontros Bibli**: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, n. 15, p. 1-23, jan/jun 2003, Disponível em http://www.encontros-bibli.ufcs.br/Edicao_15.html Acesso em 28 de Mai. 2009.

ANDERSON, J. D.; PÉREZ-CARBALLO, J., The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. **Information Processing and Management**, 37, p. 231-254, 2001.

ANDERSON, J. D.; PÉREZ-CARBALLO, J., The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. **Information Processing and Management**, 37, p. 255-277, 2001a.

ANDREWS, K; KIENREICH, W.; SABOL, V.; BECKER, J.; DROSHCHL, G.; KAPPE, F.; GRANITZER, M.; AUER, P.; TOCHTERMANN, K. The InfoSky visual explorer: exploiting hierarchical structure and document similarities. **Information Visualization**, v. 1, n. 3/4, p. 166-181, 2002.

BAEZA-YATES, R.; RIBEIRO-NETO, B., **Modern Information Retrieval**. New York: Acm Press, 1999.

BELKIN, N. Interaction with Texts: Information Retrieval as Information-Seeking Behavior. **Proceedings of the Fifth International Symposium for Information Science**, 1993. Disponível em <http://citeseer.ist.psu.edu/belkin93interaction.html>.

BOJADZIEV, G.; BOJADZIEV, M. **Fuzzy sets, fuzzy logic, applications**. Singapore: World Scientific Publishing Co Pte Ltd, 1995.

BORDOGNA G.; PASI, G. Controlling information retrieval through a user adaptive representation of documents, **International journal of approximate reasoning**, 12, p. 317-339, 1995.

BORKO, H.; BERNIER, C. **Indexing Concepts and methods**. New York: Academic Press, 1978.

BORNER, K.; CHEN, C.; BOYACK, K. Visualizing Knowledge Domains. **Annual Review of Information Science and Technology (ARIST)**, v. 37, p. 179–255, 2003

BLAIR, D. C. **Language and representation for information retrieval**. Amsterdam: Elsevier Science Publishers, 1990.

BLAIR, D. C. Information retrieval and the philosophy of language. **The computer Journal**, v. 35, n. 3, p. 200-207, 1992.

BLAIR, D. C. Information retrieval and the philosophy of language. **Annual Review of Information Science and Technology (ARIST)**, v. 36, p. 3-50, 2002

BRÄSCHER, M.; CAFÉ, L. Organização da Informação ou Organização do Conhecimento? In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 9, 2008, São Paulo, Anais. São Paulo: ANCIB, 2008. Disponível em: < <http://www.enancib2008.com.br> >. Acesso em: 18 Mar. 2008.

BUCKLAND, M. K. Information as thing. **Journal of the American Society for Information Science**, v. 42, n.5, p.351-360, 1991.

CAPURRO, R. Epistemologia e Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 5, 2003, Belo Horizonte, Anais. Belo Horizonte: ANCIB, 2003. Disponível em: < <http://www.enancib2003.com.br> >. Acesso em: 21 Mai. 2008.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. **Readings in information visualization: Using vision to think**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999.

CHEN, C. *Information Visualization: Beyond the Horizon*. Springer, 2006.

CINTRA, A. M. M.; TÁLAMO, M. F. G. M.; LARA, M. L. G.; KOBASHI, N. Y. Para entender as linguagens documentárias. São Paulo: Editora Polis, 2002.

COHEN, D. M. **O consumidor da informação documentária: o usuário de sistemas documentários visto sob a lente da análise documentária**. São Paulo, 1995, 145p. Dissertação (Mestrado em Ciência da Comunicação) – Escola de Comunicações e Artes, Universidade de São Paulo.

COOPER, W.S. Utility theoretic indexing: a note on Wilson's note. **Journal of the American Society for Information Science**, n.30, p. 170-172.

COPI, I. M. **Introdução à lógica**. São Paulo: Editora Mestre Jou, 1968.

CORRÊA, C. A. **Estruturação de problemas e ações de informação: um estudo teórico-exploratório sobre o uso de ferramentas para definição das necessidades de informação**. Campinas, 2003, 106p. Dissertação (Mestrado em Ciência da Informação) – Programa de pós-graduação em Ciência da Informação, Pontifícia Universidade Católica de Campinas, Campinas, 2003.

CRESTANI, F.; LALMAS, M. Logic and uncertainty in information retrieval. Disponível em <http://citeseer.ist.psu.edu/crestani01logic.html>, 2001, acessado em 14/08/2006.

CRESTANI, F.; PASI, G. Handling vagueness, subjectivity, and imprecision in information access: an introduction to the special issue. **Information Processing and Management**, 39, p. 161-165, 2003.

CROFT, W.B.; THOMPSON, R. H. I³R: A new approach to the design of document retrieval systems. **Journal of the American Society for Information Science**, v. 38, n.6, p. 389-404, 1987.

CROFT, W.B Automatic Indexing. In:WEINBERG, B. H. (Ed.) **Indexing – the state of our knowledge and the state of our ignorance**. Proceedings of the 20th Annual Meeting of the American Society of Indexer, 1988

DA COSTA N. C.; ABE J. M.; DA SILVA FILHO , J. I; MUROLO A. C.; LEITE C. F. S. **Lógica paraconsistente aplicada**, São Paulo:Atlas,1999.

D'OTTAVIANO, I. M. L. Paradoxos auto-referenciais e as lógicas não-clássicas heterodoxas. **Ciência e Cultura**, 42(2), 1990.

D'OTTAVIANO, I. M. L.; FEITOSA, H. A. Sobre a história da lógica, a lógica clássica e o surgimento das lógicas não-clássicas. Disponível em <ftp://ftp.cle.unicamp.br/pub/arquivos/educacional/ArtGT.pdf>, 2003, Acessado em 17 de Agosto de 2010.

FERNEDA, E. **Recuperação de informação: Análise sobre a contribuição da Ciência da Computação para Ciência da Informação**. 2003. 147p. Tese (Doutorado em Ciência da Informação) – Curso de Ciências da Comunicação, Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo 2003.

FROHMANN, B. Rules of indexing: a critique of mentalism in information retrieval theory. **Journal of Documentation**, v. 46, n.2, p. 81-101.

GOMEZ, M. N. G. De. A representação do conhecimento e o conhecimento da representação: algumas questões epistemológicas, **Ciência da Informação**, v. 22, n. 3, p. 217-222, 1993.

GOMEZ, M. N. G. De. Novas fronteiras tecnológicas das ações de informação: questões e abordagens, **Ciência da Informação**, v. 33, n. 1, p. 55-67, 2004.

GRACIOSO, L. de S. **Filosofia da linguagem e Ciência da informação: Jogos de linguagem e ação comunicativa no contexto das ações de informação em tecnologias virtuais**. 2008. 175p. Tese (Doutorado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Instituto Brasileiro de Informação em Ciência e Tecnologia, Universidade Federal Fluminense, Rio de Janeiro, 2008.

HARTER, S. P. (1992). Psychological relevance and information Science. **Journal of the American Society for Information Science**, 53 (4), p. 229-243.

HJØRLAND, B. Semantics and Knowledge Organization. **Annual Review of Information Science and Techonolgy (ARIST)**, v. 41, p. 367-405, 2007.

HAN, J.; KAMBER, M. **Data mining – Concepts and Techniques**. San Francisco: Morgan Kaufmann Publishers, 2006.

HERRERA-VIEDMA, E.; PASI, G. Fuzzy approaches to access information on the web: recent developments and research trends. **Proceedings of the Third Conference of the European Society for Fuzzy Logic and Technologies**, 2003, p. 25-31.

JACOB, E. K.; SHAW, D. Sociocognitive perspectives on representation. **Annual Review of Information Science and Techonolgy (ARIST)**, v. 33, p. 131-185, 1998.

KAUFMAN, L.; ROUSSEEUW, P. J. Finding groups in data – An introduction to Cluster Analysis. New York: John Wiley & Sons, Inc., 1990.

KOBASHI, N. Y. **A elaboração de informações documentárias: Em busca de uma metodologia**. 1994. 260p. Tese (Doutorado em Ciências da Comunicação) – Curso de Ciências da Comunicação, Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo, 1994.

KOBASHI, N. Y. Fundamentos semânticos e pragmáticos da construção de instrumentos de representação de informação. **Datagramazero – Revista de Ciência da Informação**, v. 8, n. 6, 2007. Disponível em http://dgz.org.br/dez07/F_I_art.htm. Acesso em 10 de Mar 2009.

KURAMOTO, H. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, v. 25, n. 2, 1996.

LARA, M. L. G. de. Diferenças conceituais sobre termos e definições e implicações na organização da linguagem documentária. **Ciência da Informação**, v. 33, n. 2, p. 91-69, maio/ago, 2004.

LEÃO, E. C.; d'AMARAL, M. T.; SODRÉ, M.; DORIA, F. A. D. **A máquina e seu avesso**. Rio de Janeiro: Livraria Francisco Alves Editora S. A., 1987.

LE COADIC, Y. F. **A Ciência da Informação**. Brasília: Briquet de Lemos, 1996.

LEIVA, I. G. **La automación de la indización de documentos**. Gijón: Ediciones Trea, 1999.

LIPETZ, B. The usefulness of indexes. In: WEINBERG, B. H. (Ed.) **Indexing – the state of our knowledge and the state of our ignorance**. Proceedings of the 20th Annual Meeting of the American Society of Indexer, 1988

MAI, J. Semiotics and indexing: na analysis of the subject indexing process. **Journal of Documentation**, 57, p. 591-622, 2001.

MAI, J. Analysis in indexing: document and domain centered approaches. **Information Processing and Management**, 41, p. 599-611, 2005.

MENDONÇA, E. S. A lingüística e a ciência da informação: estudos de uma interseção. **Ciência da Informação**, v. 29, n. 3, p. 50-70, set/dez, 2000.

MINGHIM, R. Visualização de informação e mineração visual de dados. Texto disponibilizado para disciplina Visualização Computacional, no Insituto de Ciências Matemáticas e de Computação, p. 1-27, 2003, disponível em <http://www.lcad.icmc.usp.br/~rosane/TextoInfoVis.pdf>, acesso em 5 de Setembro de 2008.

MINGHIM, R.; LEVKOWITZ, H. Visual Mining of Text Collections. Tutorial Notes. In: **Proceedings of EUROGRAPHICS 2007** - Computer Graphics Forum, pp. 929-1021. 2007.

MOLINARI, A.; PASI, G. A fuzzy representation of HTML documents for information retrieval systems. **Proceedings of IEEE International Conference on Fuzzy Systems**, New Orleans, p. 8-12, 1996.

MORTARI, C. A. **Introdução à lógica**. São Paulo: Editora UNESP, 2001.

MÜLLER, A. L. P.; VIOTTI, E. C. Semântica formal. In: FIORI, J. L. (Ed.) **Introdução à lingüística II – Princípios de análise**. São Paulo: Editora Contexto, 2003.

NÉVÉOL A.; SHOOSHA, S.; HUMPHREY, S. M.; MORK, J. G.; ARONSON, R. A. A recent advance in the automatic indexing of the biomedical literature. **Journal of Biomedical Informatics** (2009), doi:10.1016/j.jbi.2008.12.007

NOMOTO, K.; WAKYAMA, S.; KIRIMOTO, T.; OHASHI, Y.; KONDO, M. A document retrieval system based on citations using fuzzy graphs. **Fuzzy Sets and Systems**, 38, p. 207-222, 1990.

NOVELLINO, M. S. F. A linguagem como meio de representação ou de comunicação da informação. **Perspectivas em Ciência da Informação**, v. 3, n.1, p.137-146, jul/dez. 1998.

NOWACKA, K.; ZADROZNY, S.; KACPRZYK, J. A new fuzzy logic based information retrieval model. **Proceedings of the 12th International Conference – Information processing and management of uncertainty in knowledge-base systems – IPMU**, Torremolinos (Málaga), 2008.

OGAWA, Y.; MORITA, T.; BOBAYASHI, K. A fuzzy document retrieval system system using connection matrix and a learning method. **Fuzzy Sets and Systems**, 39, p. 163-179, 1991.

OLIVEIRA, R. P. Semântica. In MUSSALIM, F.; BENTES, A. C. (Eds.) **Introdução à lingüística 2 – Domínios e Fronteiras**. São Paulo: Cortez Editora, 2000.

PAULOVICH, F.; OLIVEIRA, M. C. F.; MINGHIM, R. The Projection Explorer: A Flexible tool for projection-based multidimensional visualization, **Proc. of XX Brazilian Symposium on Computer Graphics and Image Processing**, p. 27-34, 2007.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least Square Projection: A Fast high-precision multidimensional projection technique and its application do documet mapping. **IEE Transactions on Visualization and Computer Graphics**, v. 14, n. 2, March/April, p. 1-12, 2008.

PAULOVICH, F.V. **Mapeamento de dados multidimensionais – integrando mineração e visualização**. 2008. 144p. Tese (Doutorado em Ciência da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2008.

PEDRYCZ, W.; GOMIDE, F. **Fuzzy systems engineering – Toward a human-centric computing**. New Jersey: John Wiley & Sons, Inc., 2007.

PEREIRA, R. C. **Modelo ontológico relacional fuzzy em sistemas de recuperação de informação textual**. Campinas, 2004, 99p. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, 2004.

PIETROFORTE, A. V. S. & LOPES, I. C. A semântica lexical. In: FIORIN, J. L. (Ed.) **Introdução à lingüística II – Princípios de análise**. São Paulo: Editora Contexto, 2003.

POMBO, O. Da classificação dos seres à classificação dos saberes. Leituras. Revista da Biblioteca Nacional de Lisboa, n. 2, Primavera, p. 19-33, 1998. Disponível em: <http://www.educ.fc.ul.pt/docentes/opombo/investigacao/opombo-classificacao.pdf>. Acesso em 10 de Março de 2009.

RAYWARD, W.B. The origins of information science and the International Institute of Bibliography/International Federation for Information and Documentation (FID). **Journal of the American Society for Information Science**, v.48, n.4, p.289-300, 1997.

ROBINS, David. Interactive information retrieval: context and basic notion. **Informing Science**, v. 3, no.2, 2000, p. 57-61.

SALMON, W. C. **Lógica**. Rio de Janeiro: Zahar Editores, 1978.

SALTON, G. **Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer**. New York: Addison-Wesley Publishing Company, Inc, 1989.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information**. New York: McGraw-Hill Publishing Company, 1983.

SALTON, G.; BUCKLEY, C. J. Term-Weighting approaches in automatic text retrieval. **Information Processing and Management**, v. 24, n. 5, p. 513-523, 1988.

SARACEVIC, T. Relevance: A review of the literature and framework for thinking on the notion in Information Science. Part II. **Advances in Libranrianship**, v. 30, 2006.

SHNEIDERMAN, B. The eyes have it: a task by data type taxonomy for information visualizations. **Proceedings of IEEE Symposium on Visual Languages**, Boulder, CO, September 3-6 pp. 336-343, 1996.

SILVA, M. R.; FUJITA, M. S. L. A prática da indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, 16(2), p.133-161, mai/ago 2004.

SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, n. esp., 1º sem 2006. Disponível em <http://www.encotros-bibli.ufcs.br/> Acesso em 11 de Mai. 2009

SUCOLOTTI A. A., **Recuperação de informação em bases textuais: uma abordagem baseada em lógica paraconsistente**. Dissertação (Mestrado) – Programa de pós-graduação em Ciência da Computação, Universidade Federal de Santa Catarina, Florianópolis, 2001.

SVENONIUS, E. **The intellectual foundation of information organization**. Cambridge: MIT Press, 2000.

TÁLAMO, M.F.G.M.; LENZI, L.A.F. Terminologia e documentação: e relação solidária das organizações do conhecimento e da informação no domínio da inovação tecnológica. **DataGramaZero – Revista de Ciência da Informação**, vol.7, n.4, agosto 2006. Disponível em [http://dgz.org.br/dez07/F I art.htm](http://dgz.org.br/dez07/F_I_art.htm). Acesso em 12 de Março 2009

TAN, P.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Boston: Pearson Education Inc, 2006.

TAYLOR, A. G. **The organization of information**. Englewood:Colorado, Libraries Unlimited Inc, 1999.

TEIXEIRA, J. F. **Mentes e máquinas – Uma introdução à ciência cognitiva**. Porto Alegre: Editora Artes Médicas Ltda, 1998.

TOLEDO, G. L.; OVALLE, I. I. **Estatística Básica**. São Paulo: Editora Atlas S.A., 1986.

TURTLE, H. R.; CROFT, W. B. Uncertainty in information retrieval system, In: MOTRO, A. & SMETS, P., (Eds.) **Uncertainty Management in Information Systems: from needs to solutions**. Boston/London/Dordrecht: Kluwer Academic Publisher, 1997.

VAKKARI, P. Task Complexity, problem structure and information actions. Integrating studies of information seeking and retrieval. **Information Processing and Management**, v. 35, p. 819-837, 1999.

VAN RIJSBERGEN, C. J. **Information retrieval**. London: Butterworths, 1979

VAN RIJSBERGEN, C. J. A new theoretical framework for information retrieval. **Proceedings of the Annual International ACM SIGIR – Conference on Research and Developments in Information Retrieval**, p. 194-200, 1986.

WEINBERG, B. H. Why indexing fails the researcher. **The Indexer**, v. 16, n.1, p. 3-6, 1988.

WESSEL, A. E. **Computer-Aided information retrieval**. Los Angeles: Mellville Publishing Company, 1975.

WILSON, T. D. Information behaviour: an interdisciplinary perspective. **Information Processing and Management**, v. 33, p. 551-572, 1997.

ZADROZNY, S.; NOWACKA, K. Fuzzy information retrieval model revisited. **Fuzzy sets and systems**, 160, p. 2173-2191, 2009.

ZINS, C. Conceptions of Information Science. **Journal of the American Society for Information Science and Technology**, v.58, n.3, p.335-350, 2007.

ZHU, B.; CHEN, H. Information Visualization, **Annual Review of Information Science and Technology**(ARIST), v. 39, p. 139–177, 2005.

Apêndices

Apêndice A – Fundamentos da lógica Clássica

Nesse apêndice veremos aspectos fundamentais da lógica clássica.

Um exemplo extraído de Salmon (1978) permite esclarecer a noção de argumento e conclusão. Esse autor utiliza a narrativa de uma das histórias do detetive inglês Sherlock Holmes. Nesta aventura, o herói encontra um velho chapéu de feltro. Embora não conheça o seu proprietário, Holmes conta a Watson, seu companheiro de aventuras, muita coisa a respeito do mesmo. Afirma, entre outras coisas, que o proprietário é um intelectual. A afirmação, como foi feita, não dispõe de qualquer apoio ou evidências observáveis. Holmes podia saber da existência das mesmas, mas não as deu. Watson, não percebe o que se passa e pede a Holmes que o esclareça. Holmes coloca o chapéu sobre a cabeça. O chapéu afunda, resvala pela sua testa, e vai apoiar-se no seu nariz. O detetive então afirma ser uma questão de volume, pois um homem com uma cabeça tão grande deve ter algo dentro dela.

O argumento de Holmes, conforme Salmon (1978, p.15), pode ser descrito da seguinte forma:

1. Há um chapéu grande.
2. Alguém é o proprietário deste chapéu.
3. Proprietários de chapéus grandes têm cabeças grandes.
4. Pessoas de grandes cabeças têm cérebros grandes.
5. Pessoas de cérebros grandes são intelectuais.
6. O proprietário deste chapéu é um intelectual.

Com esta descrição, podemos definir argumento como uma coleção de enunciados que estão relacionados uns com os outros segundo uma estrutura específica. Assim, um argumento consiste num enunciado que é a conclusão e num ou mais enunciados de evidência corroboradora. Os enunciados de evidência são chamados de premissas. Ou seja, a estrutura de um argumento pode ser assim explicitada da seguinte forma: argumento = premissas + conclusão.

No caso do exemplo citado teríamos: premissas - enunciados 1,2,3,4 e 5; conclusão - enunciado 6.

Nesse contexto, segundo Salmon (1978), algumas considerações podem ser feitas. Quando um argumento é enunciado para justificar sua conclusão, duas questões se colocam: (i) as premissas são verdadeiras? (ii) As premissas estão relacionadas de maneira adequada com a conclusão? Se uma das perguntas receber resposta negativa, a justificativa não pode ser satisfatória. Contudo, não devemos misturar as duas questões. Para a Lógica, só interessa a segunda. Assim, conclui o autor que “*a Lógica trata da relação entre premissas e conclusão, deixando de se importar-se com a verdade das premissas*” (SALMON, 1978, p. 16).

O principal objetivo da Lógica é elaborar métodos que permitam distinguir argumentos corretos e incorretos. A correção ou incorreção lógica de um argumento só depende da relação entre as premissas e a conclusão. Ou seja, a correção ou incorreção lógica é inteiramente independente da verdade das premissas. Num argumento logicamente correto, a relação que as premissas mantêm com a conclusão é: se as premissas forem verdadeiras, isso basta para admitir a veracidade da conclusão.

Normalmente, os argumentos são descritos usando-se a linguagem natural. A sua organização numa seqüência de passos é efetuada para facilitar a análise lógica.

Outras considerações são necessárias para um entendimento mínimo da Lógica, tais como a distinção entre descoberta e justificação e a entre verdade e validade. Os próximos itens tratarão dessas noções.

Descoberta e Justificação

Até aqui foram descritas as relações entre inferências e argumentos. Outras considerações outras podem ser efetuadas quando uma afirmação ou conclusão é enunciada. Segundo Salmon (1978), quando um enunciado é efetuado, duas questões principais podem ser imediatamente colocadas: (i) De que maneira tal enunciado chegou a ser concebido? (ii) Quais são as razões para aceitá-lo como verdadeiro?

Devemos observar que são duas questões distintas. A primeira relaciona-se com a descoberta, ou seja, “*as circunstâncias lembradas por ela formam o contexto da descoberta. A segunda relaciona-se com a justificação, ou seja, pertence ao contexto da justificação*” (SALMON, 1978, p. 24).

É importante estabelecer distinção entre os dois contextos quando se esta preocupado com o apoio que a(s) evidência(s) pode(m) oferecer a uma afirmação. A justificação de uma afirmação é feita por meio de uma entidade linguística, o argumento. Tal justificativa consiste na conclusão apontada e na sua evidência corroboradora, corretamente relacionadas. Por outro

lado, a descoberta de um enunciado, é um processo mental, psicológico, que conduz a sua defesa e eventual aceitação.

A atividade psicológica desenvolvida na inferência é um processo de descoberta. Quem infere deve pensar na conclusão. Por outro lado, o problema da descoberta, não se esgota neste ponto. O sujeito que infere precisa descobrir a evidência corroboradora e, mais ainda, descobrir a relação entre evidência e a conclusão (SALMON, 1978).

Para o autor citado, a Lógica não tem como objetivo descrever os modos de pensar e nem deve ensinar a inferir. Seu objetivo é indicar quais são as inferências que podem ser aceitas. Assim, pode-se concluir que a Lógica se interessa pela justificação, não pela descoberta. Ou seja, ela fornece métodos para analisar o discurso. Tal análise é fundamental para exprimir de maneira clara o pensamento e para possibilitar uma compreensão completa daquilo que se comunica e se aprende (SALMON, 1978).

Os argumentos não são expressos, necessariamente, em linguagem natural. Eles podem estar colocados sob a forma de uma linguagem artificialmente construída, como no caso dos enunciados da matemática. Nesse caso, o uso da Lógica não se resume em afirmar se, de fato, certa conclusão é consequência ou não de certas premissas. Aqui, a Lógica pode ser vista como um conjunto de técnicas que permitem produzir uma conclusão a partir da informação disponível (premissas) (MORTARI, 2001).

Verdade e Validade

Verdade e falsidade podem ser atributos dos enunciados que compõem um argumento, mas nunca dos argumentos em si mesmo. Da mesma forma, validade ou não-validade só podem ser associadas a argumentos, mas nunca a enunciados. Vejamos alguns exemplos de argumentos, retirados de Copi (1968, p.38):

i) premissas: Todas as baleias são mamíferos.

Todos os mamíferos têm pulmões.

conclusão: Todas as baleias têm pulmões.

(ii) premissas: Todas as aranhas têm seis pernas.

Todos os seres de seis pernas têm asas.

conclusão: Todas as aranhas têm asas.

Nesses exemplos, ambos os argumentos são válidos. Entretanto, não se pode garantir a verdade de todas as premissas e/ou conclusões. Os exemplos mostram que existem argumentos válidos com conclusões falsas, da mesma forma como podem existir argumentos não-válidos com conclusões verdadeiras. Assim, a verdade ou falsidade de uma conclusão não determina a validade ou não-validade de um argumento. Existem argumentos perfeitamente válidos que têm conclusões falsas. Mas, neste caso, devem ter pelo menos uma premissa falsa. Argumentos válidos com premissas verdadeiras devem, necessariamente, ter uma conclusão verdadeira. Determinar a verdade ou falsidade das premissas é uma tarefa que incumbe à Ciência, em geral, pois as premissas podem referir-se a qualquer tema. Nesse sentido, pode-se fazer referência à figura 4.2 anteriormente mostrada. A validação das premissas se encontra no contexto da realidade observável. Por outro lado, a validade da conclusão oriunda das premissas é feita analisando-se o argumento como um todo. Pode-se concluir, dessa forma, que a Lógica se preocupa em determinar a validade ou não dos argumentos (SALMON, 1978).

Dedução e Indução

Alem dos aspectos que envolvem a validade ou não-validade dos argumentos, outros podem ser observados, de modo a distingui-los. Nesse sentido, os argumentos podem ser separados em dois grandes grupos: os dedutivos e os indutivos. E, embora todos os argumentos tenham a pretensão de que suas premissas forneçam a prova da verdade de sua conclusão, somente nos argumentos dedutivos se encontra a pretensão de que suas premissas fornecem uma prova conclusiva. Um argumento dedutivo é válido quando suas premissas, se verdadeiras, fornecem provas convincentes para sua conclusão, isto é, quando as premissas e a conclusão estão de tal forma relacionadas, sendo absolutamente impossível que as premissas sejam verdadeiras se a conclusão não for verdadeira.

Por outro lado, um argumento indutivo, envolve a pretensão, não de que suas premissas proporcionem provas convincentes da verdade de sua conclusão, mas de que somente forneçam algumas provas disso. Os argumentos indutivos não são válidos ou não-válidos, no sentido em que estes termos se aplicam aos argumentos dedutivos. Os raciocínios indutivos podem, é claro, ser avaliados como melhores ou piores, segundo o grau de aproximação ou probabilidade que as premissas confirmam às respectivas conclusões.

A seguir, mais alguns exemplos retirados de Salmon (1978, p.29) (por simplicidade, e para utilizar uma notação comum entre os textos de lógica, de agora em diante, será usado o símbolo → para indicar a conclusão de um argumento).

Argumento dedutivo: Todo mamífero tem um coração.

Todos os cavalos são mamíferos.

→ Todos os cavalos têm um coração.

Argumento indutivo: Todos os cavalos que foram observados tinham um coração.

→ Todos os cavalos têm um coração.

Podem ser listadas algumas das características dos dois tipos de argumentos.

Argumentos dedutivos:

- se todas as premissas são verdadeiras, a conclusão deve ser verdadeira;
- toda a informação ou conteúdo fatural da conclusão já estava, pelo menos implicitamente, nas premissas.

Argumentos indutivos:

- se todas as premissas são verdadeiras, a conclusão é provavelmente verdadeira, mas não necessariamente verdadeira;
- a conclusão encerra informação que não estava, nem implicitamente, nas premissas.

Os argumentos dedutivos destinam-se a deixar explícito o conteúdo das premissas. Ou seja, tudo o que está embutido na conclusão já havia sido dito, mesmo que implicitamente, nas premissas. Por outro lado, os argumentos indutivos destinam-se a ampliar o alcance de nossos conhecimentos, ou seja, a conclusão diz mais do que o que estava afirmado nas premissas (MORTARI, 2001).

Deve-se observar que a relação que existe entre a generalização científica e a evidência sustentadora é do tipo indutivo. Assim, por exemplo, de acordo com a primeira lei de Kepler, a órbita de Marte é uma elipse. A evidência observacional para essa lei consiste em várias observações isoladas, destinadas a determinar a posição do planeta. A lei refere-se à posição do planeta, tenha sido este observado ou não. Em particular, tal lei afirma que o movimento futuro de Marte será elíptico, que o movimento passado, antes do planeta ser

observado, também era elíptico, e que o movimento é elíptico mesmo quando o planeta não pode ser visto. Está claro, nesse caso, que a lei, ou a conclusão do argumento, tem conteúdo muito mais amplo do que o dos enunciados que descrevem as posições observadas.

Por outro lado, pode-se afirmar que os argumentos matemáticos são dedutivos. Um exemplo muito citado desta situação é o da Geometria Euclidiana do plano. Nessa disciplina, os teoremas são demonstrados a partir dos axiomas e postulados enunciados. O método de demonstração consiste em deduzir os teoremas (conclusões) dos axiomas e postulados (premissas). Ou seja, o método de dedução garante que os teoremas devem ser verdadeiros se são verdadeiros os axiomas e postulados. Este exemplo serve para mostrar que os argumentos dedutivos nem sempre são triviais. Embora o conteúdo dos teoremas já esteja fixado nos axiomas e postulados, esse conteúdo está longe de ser óbvio ou facilmente dedutível (SALMON, 1978).

Os parágrafos anteriores introduzam aspectos fundamentais da lógica, sem detalhar as técnicas e métodos desenvolvidos para o tratamento ou manipulação lógica de diferentes argumentos. De acordo com Da Costa et al. (1999), existem dois métodos principais para tratamento de sistemas lógicos: o método lingüístico e o método algébrico. O texto anterior se restringiu a exemplos envolvendo o primeiro dos métodos citados. Maiores detalhes sobre o método algébrico podem ser encontrados, por exemplo, em Mortari (2001).

Apêndice B – Compilação de resultados

Comparação Visualização - Matriz de similaridades - Histogramas de distancias - Coeficiente de silhueta

Coleção CBR-ILP-IR-SON

Legenda de cores para esta coleção:

Azul – CBR (*Case based reasoning*)
 Vermelho – SON (*Sonification*)
 Lima – IR (*Information Retrieval*)
 Verde – ILP (*Inductive Logic Programing*)

Visualização

Matriz de similaridades

Histograma de distancias

Coeficiente de Silhueta:
 - 0,058435094

Figura 78 – Visualização original
 Fonte: produzida no PEx

Coeficiente de Silhueta:
 0,18285953

Figura 79 – Visualização da Região R0
 do QUPC
 Fonte: produzida no PEx

Coefficiente de Silhueta:
0,2880726

Figura 80 – Visualização da Região R1
do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta:
0,20882708

Figura 81 – Visualização da Região R2
do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta:
- 0,04546633

Figura 82 – Visualização da Região R3
do QUPC
Fonte: produzida no PEx

Coleção CBR-ILP-IR

Legenda de cores para esta coleção:

Azul – CBR (*Case based reasoning*)

Vermelho – IR (*Information Retrieval*)

Verde – ILP (*Inductive Logic Programing*)

Visualização

Matriz de similaridades

Histograma de distancias

Coefficiente de Silhueta:
0,19003317

Figura 83 – Visualização original
Fonte: produzida no PEx

Coefficiente de Silhueta:
0,23320624

Figura 84 – Visualização da Região R0
do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta:
0,3250948

Figura 85 – Visualização da Região R1
do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta:
0,65129006

Figura 86 – Visualização da Região R2
do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta:
0,22175422

Figura 87 – Visualização da Região R3
do QUPC
Fonte: produzida no PEx

Coleção CBR-ILP-SON

Legenda de cores para esta coleção:

Azul – CBR (*Case based reasoning*)

Verde – ILP (*Inductive Logic Programming*)

Vermelho – SON (*Sonification*)

Visualização

Matriz de similaridades

Histograma de distancias

Coefficiente de Silhueta:
0,41561985

Figura 88 – Visualização original
Fonte: produzida no PEx

Coefficiente de Silhueta:
0,4150745

Figura 89 – Visualização da Região R0 do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta:
0,5768113

Coefficiente de Silhueta:
0,5177454

Coefficiente de Silhueta:
0,6606122

Figura 92 –Visualização da Região R3 do QUPC
Fonte: produzida no PEx

Figura 90 –Visualização da Região R1 do QUPC
Fonte: produzida no PEx

Figura 91 –Visualização da Região R2 do QUPC
Fonte: produzida no PEx

Coleção IR-SON

Legenda de cores para esta coleção:

Azul – IR (*Information Retrieval*)

Vermelho – SON (*Sonification*)

Visualização

Figura 93- Visualização original
Fonte: produzida no PEx

Matriz de similaridades

Histograma de distancias

Coefficiente de Silhueta:
0,4098989

Figura 94 - Visualização da Região R0
do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta:
0,40576202

Min Max

Min Max

Coeficiente de Silhueta:
0,39995274

Min Max

Min Max

Coeficiente de Silhueta:
0,5497286

Min Max

Min Max

Coeficiente de Silhueta:
- 0,040663127

Figura 95 - Visualização da Região R1
do QUPC
Fonte: produzida no PEx

Figura 96 - Visualização da Região R2
do QUPC
Fonte: produzida no PEx

Figura 97 - Visualização da Região R3
do QUPC
Fonte: produzida no PEx

Coleção CBR-ILP

Legenda de cores para esta coleção:

Azul – CBR (*Case based reasoning*)
Vermelho – ILP (*Inductive Logic Programing*)

Visualização

Matriz de similaridades

Histograma de distancias

Coefficiente de Silhueta: 0,67798316

Figura 98- Visualização original
 Fonte: produzida no PEx

Coefficiente de Silhueta: 0,66479486

Figura 99 - Visualização da Região R0 do QUPC
 Fonte: produzida no PEx

Figura 100 - Visualização da Região R1 do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta: 0,48078427

Figura 101 - Visualização da Região R2 do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta: **0,74633**

Figura 102 - Visualização da Região R3 do QUPC
Fonte: produzida no PEx

Coefficiente de Silhueta: 0,010869439