INDEXAÇÃO AUTOMÁTICA E VISUALIZAÇÃO DE INFORMAÇÕES: UM ESTUDO BASEADO EM LÓGICA PARACONSISTENTE

Carlos Alberto Correa,Nair Yumiko Kobashi

Resumo: Apresentação dos resultados de pesquisa de utilização da lógica paraconsistente em procedimentos de indexação automática. A utilização dessa lógica e de métodos dela derivados, por serem flexíveis, comportam estados lógicos que vão além das dicotomias sim e não. Essas características permitem 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 tratar incerteza e imprecisão, desenvolvido no âmbito da lógica paraconsistente, para modificar os valores dos pesos atribuídos aos termos de indexação. O experimento foi realizado em corpus disponível em sistema de indexação e visualização de informações com código fonte aberto. Os resultados foram avaliados por meio de critérios e índices embutidos no próprio sistema de visualização e demonstram que há ganhos mensuráveis de qualidade na construção das visualizações. Confirma-se, assim, a hipótese de que a lógica paraconsistente tem aplicação promissora em indexação automática por sua potencialidade para tratar situações que envolvem incerteza, imprecisão e vagueza.

Palavras-chave: Indexação automática. Visualização de informação. Lógica paraconsistente.

1 INTRODUÇÃO

O volume crescente dos estoques de registros externos à memória humana, aumenta consideravelmente as dificuldades para a manutenção, organização e manipulação dos dispositivos informacionais. Uma das alternativas para enfrentar esse desafio tem se ancorado na indexação automática, processo que envolve diversas disciplinas, tais como Linguística, Estatística, Terminologia e diversas teorias desenvolvidas no âmbito da Matemática e da Lógica, como as teorias de tratamento de incerteza e imprecisão. Cabe acrescentar que, tradicionalmente, os sistemas de recuperação de informação são o ambiente preferido para testes de indexação automática e de visualização de informação

Neste texto, são apresentados os resultados de uma pesquisa de Indexação automática e visualização de informações baseada em Lógica paraconsistente (CORRÊA, 2011). Antes da apresentação dos resultados serão discutidos, de forma breve, o modelo do espaço vetorial e propostas baseadas em lógicas não clássicas, como também as propostas de visualização gráfica.

2 INDEXAÇÃO AUTOMÁTICA E O MODELO DO ESPAÇO VETORIAL

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). As pesquisas para armazenar informação se voltaram, com o tempo, para a busca de soluções de indexação automática. Os primeiros experimentos tiveram início na década de 1950, com Luhn e Baxendale (LANCASTER, 2004), sendo o modelo do espaço vetorial (MEV) um dos mais citados na literatura da área. O MEV é também utilizado em diferentes tipos de sistemas de visualização de informações.

O MEV foi desenvolvido por Luhn e popularizado por Salton no âmbito do sistema SMART – *System for the Manipulation and Retrieval of Text* (RAGHAVAN, 1997). Ele utiliza a mesma representação tanto para os documentos de uma coleção quanto para as consultas ao sistema. Essa representação é baseada no conceito matemático de vetor. Considere-se, por exemplo, o documento D, com os termos de indexação t_1, t_2 e t_3. No MEV atribui-se um peso para cada termo de indexação, cuja representação do documento é um vetor que pode ser visualizado em espaço tridimensional. O MEV pode ser representado por meio de uma matriz, 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	$w_{2,1}$	$W_{2,3}$	$w_{2,3}$	...	$W_{2,n}$
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 2 – Matriz documentos X pesos
Fonte: Adaptada de Salton (1989)

A abordagem vetorial permite que conceitos desenvolvidos no âmbito da teoria matemática de vetores sejam utilizados. Um deles é o de distância entre dois vetores. Salton e McGill (1983) consideram que o cálculo da distância entre dois vetores, que representam documentos da coleção, indica seu grau de similaridade. O cálculo da distância entre dois vetores x e y é efetuado pela fórmula (FERNEDA, 2003; SALTON e McGILL,1983):

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

onde $w_{i,x}$ refere-se ao 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. Uma expressão de busca também pode ser representada vetorialmente, atribuindo-se pesos aos termos utilizados. Esse procedimento permite calcular o grau

de similaridade entre uma consulta e os documentos da coleção. Uma vez compreendidos os aspectos básicos do MEV, uma questão continua em aberto: como definir os valores dos pesos para os termos de indexação e para os termos de busca?

Várias critérios podem ser utilizados para definir o cálculo dos pesos (SALTON; McGIL, 1983; SALTON; BUCKLEY, 1988). O método mais utilizado é baseado na frequência estatística de ocorrência dos termos em duas instâncias diferentes: o documento, individualmente e a coleção. No primeiro caso, efetua-se a contagem pura e simples do número de vezes em que o termo aparece no documento. Esse valor é denominado tf (expressão: *term frequency*). No segundo, 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 dado pela fórmula $idf_t = N / n_t$, onde N é o numero de documentos da coleção e n_t é o número de documentos da coleção que contém ao menos uma ocorrência do termo t. Uma variante dessa fórmula utiliza o logaritmo de base 10, ou seja: $idf_t = Log (N / n_t)$. Uma vez definidos os valores das instâncias locais e globais do termo t, seu peso é calculado como: $W_{t,d} = tf_{t,d}$ x idf_t.

3 INDEXAÇÃO AUTOMÁTICA E LÓGICAS NÃO CLÁSSICAS

As soluções desenvolvidas em indexação automática utilizam, muitas vezes, teorias para lidar com situações imprecisas, vagas e ambíguas, tais como as lógicas não clássicas. Três problemas, ao menos, motivaram o desenvolvimento das lógicas não clássicas: (i) a identificação de certos paradoxos lógico-matemáticos que alimentavam dúvidas quanto à validade dos princípios gerais dessas disciplinas; (ii) o surgimento de matemáticas "não tradicionais" como, por exemplo, as geometrias não euclidianas; (iii) as situações do mundo real que não se ajustam à dicotomia – verdadeiro ou falso –da lógica clássica

A lógica clássica foi desenvolvida com base em três princípios fundamentais: a) identidade, que estabelece que todo objeto é idêntico a si mesmo; b) não-contradição – que estabelece que uma proposição não pode ser verdadeira e falsa ao mesmo tempo; c) o do terceiro excluído – que considera que uma proposição é verdadeira ou falsa, não havendo uma terceira possibilidade. Dois exemplos de lógicas não clássicas são a lógica difusa e a lógica paraconsistente. A lógica difusa foi desenvolvida em contraposição à lógica clássica, com base na teoria dos conjuntos difusos, proposta por Loft Zadeh (BOJADZIEV; BOJADZIEV, 1995). A lógica paraconsistente, por sua vez, teve seus fundamentos desenvolvidos, independentemente, pelo polonês Stanislaw Jaskowiski e pelo brasileiro Newton da Costa. Podem ser citadas duas de suas principais características: (i) derroga o princípio da não-contradição; (ii) possui variantes de sua formulação original que permitem desenvolver vários estados lógicos além dos estados dicotômicos Verdadeiro (V) e Falso (F).

3.1 Indexação automática e lógica difusa

A teoria clássica de conjuntos e a lógica clássica são utilizadas, implícita ou explicitamente, na

grande maioria dos sistemas de recuperação de informação (SRI) (FERNEDA, 2003) e, similarmente, nos procedimentos de indexação automática. Os defensores da lógica difusa, por outro lado, consideram que a teoria dos conjuntos difusos pode ser empregada com sucesso para tratar a imprecisão e a subjetividade inerentes aos processos de indexação, bem como para gerenciar a vagueza embutida nas consultas formuladas pelos usuários (HERRERA-VIEDMA; PASI, 2003).

Bordogna e Pasi (1995) aplicaram a lógica difusa em um SRI. As autoras partiram do princípio que os documentos são organizados em partes estruturais, tais como título, autor(es), palavras-chave, resumo e referências. O papel informativo de cada termo depende da parte ou seção em que ele ocorre. Além disso, as seções de um documento podem ter diferentes graus de importância para cada usuário e, assim, o cálculo do grau de significância passa a depender da intervenção do usuário (BORDOGNA; PASI, 1995).

Outra utilização da abordagem difusa 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. Similarmente ao trabalho de Bordogna e Pasi (1995), esse modelo atribui um grau de importância a cada seção.

3.2 Indexação automática e lógica paraconsistente

A utilização da lógica difusa na indexação automática estimulou esta pesquisa de uso da lógica paraconsistente. Adotou-se, neste estudo, uma de suas variantes - a Lógica paraconsistente anotada. A anotação é feita com dois valores que, de agora em diante, será referida como LPA2v, conforme utilizado por Da Costa et al. (1999). Nessa lógica, atribuem-se duas variáveis a uma proposição: os graus de crença e de descrença, com valores que variam no intervalo [0,1]. Essa abordagem permite estabelecer quatro estados lógicos principais (inconsistente, verdadeiro, falso, indeterminado) e uma quantidade variável de estados lógicos intermediários. Os estados lógicos são estabelecidos a partir da análise efetuada sobre os dois valores da anotação, descritos por um par (μ_1, μ_2), que representa os graus de crença e descrença atribuídos a uma proposição. Dessa forma, para uma dada proposição, se os valores dos graus de crença e descrença (μ_1, μ_2), forem estabelecidos ou calculados com os valores (1.0, 0.0), o significado será: crença total e ausência de descrença na proposição ou, seja, seu estado lógico será verdadeiro. De forma similar podem-se descrever outros três estados lógicos principais: (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).

Outros valores pertencentes ao intervalo [0,1] podem ser atribuídos aos graus de crença e descrença, de modo a estabelecer estados lógicos intermediários. Esses estados ls podem ser representados em um gráfico, em que os eixos horizontal e vertical indicam os respectivos graus. Esse procedimento permite estabelecer descrições de situações em que os valores extremos não se aplicam, conforme a figura abaixo, chamada por Da Costa et al. (1999) de quadrado unitário do plano

cartesiano – QUPC.

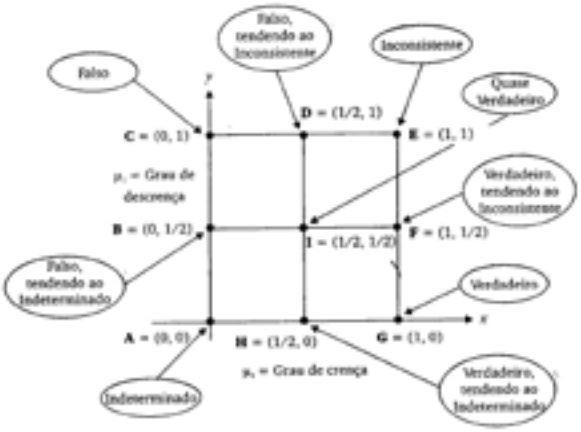


Figura 3 - Representação de alguns estados lógicos intermediários
Fonte: Da Costa et al. (1999), p 50

O gráfico acima pode ser expandido traçado-se alguns segmentos de reta de modo a definir, em vez de pontos intermediários, regiões de ocorrência dos estados lógicos extremos e intermediários. Ou seja, além do ponto que descreve o estado extremo, tem-se uma região que pode ser considerada a ocorrência de um estado extremo ou intermediário. 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.

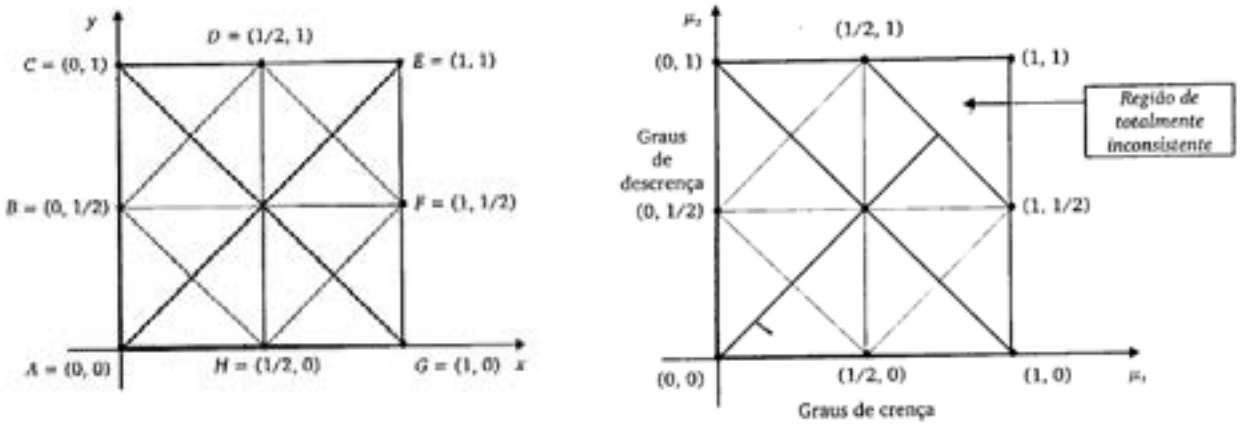


Figura 4 - QUPC com linhas de limitação de regiões caracterização da região do totalmente inconsistente
Fonte: Da Costa et al. (1999), p. 65 e 66

Alem das regiões associadas aos estados extremos, outras podem ser observadas. Elas

descrevem estados intermediários e tendências que se situam entre os estados extremos definidos. A figura 5, a seguir, descreve todas as regiões do QUPC. Essa configuração é chamada por Da Costa et al. (1999) de QUPC de resolução 12, pois podem ser identificadas, no gráfico, 12 regiões.

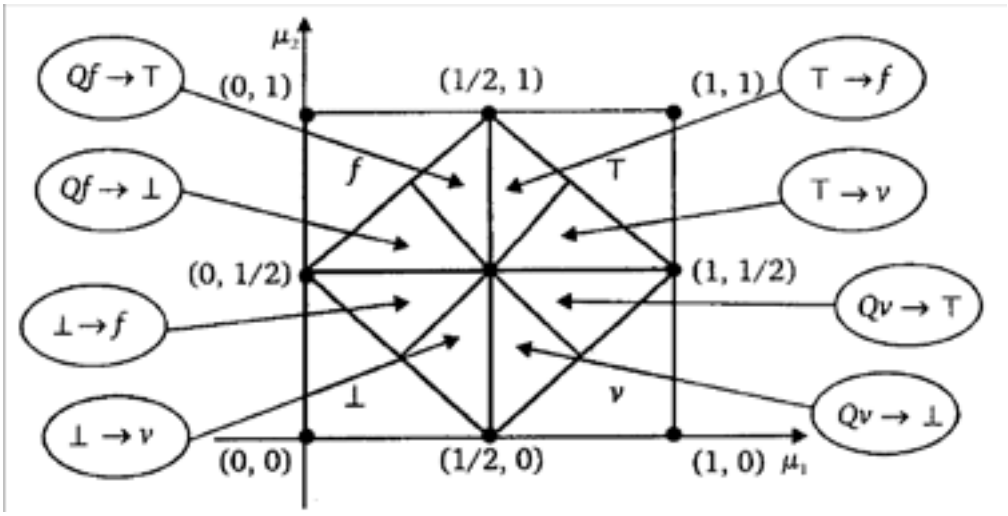


Figura 5 - QUPC de resolução 12 destacando todas as regiões possíveis – intermediárias e extremas
Fonte: Da Costa et al. (1999), p. 89

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; Qf □ T– Quase Falso, tendendo ao Inconsistente; Qf □ ⊥– Quase Falso, tendendo ao Indeterminado; Qv □ ⊥– Quase Verdadeiro, tendendo ao Indeterminado.

Da Costa et al. (1999) descrevem 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 de controle de contradição, 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 por serem artifícios matemáticos que permitem fazer ajustes no tamanho das regiões delimitadas no QUPC, de forma a restringir a sensibilidade do procedimento a certos valores lógicos. Na figura 6, a seguir, vemos variações da figura 5, em que os valores das variáveis V_{scc}, V_{icc}, V_{scct}, V_{icct} é modificado.

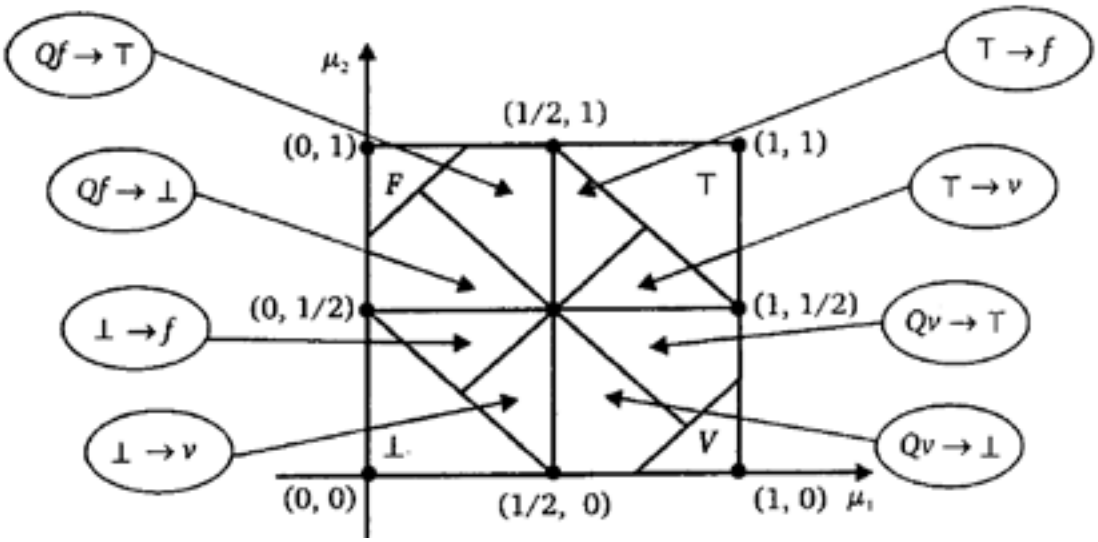


Figura 6- QUPC de resolução 12- destaque para regiões com variáveis delimitadoras modificadas
Fonte: Da Costa et al. (1999), p. 91 e 93

No segmento à direita, da figura 6 os valores das variáveis delimitadoras são: $V_{scc} = 3/4$, $V_{icc} = -3/4$, $V_{scct} = 3/4$, $V_{icct} = -3/4$. Pode-se observar a redução no tamanho das regiões extremas (verdadeiro, falso, inconsistente e indeterminado). Nesse caso, a análise efetuada a partir dos valores de crença e descrença (μ_1 e $\mu_{2)}$ será mais exigente para os valores extremos. Ou seja, o número de situações em que o par μ_1 - μ_2 será considerado verdadeiro (ou falso, indeterminado e inconsistente) será menor do que o número de situações analisadas na figura 5. No gráfico à esquerda da figura 6, os valores das variáveis delimitadoras são: $V_{scc} = 3/4$, $V_{icc} = -3/4$, $V_{scct} = 1/2$ e $V_{icct} = -1/2$. Pode-se ver no gráfico 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. Em ambas as situações descritas nas figuras 5 e 6, o tamanho das regiões intermediárias também é modificado, significando que pode haver diferentes tolerâncias para os valores intermediários.

Apesar de as figuras serem boas para visualizar o estado lógico de uma proposição, elas não são úteis para fazer avaliações de maneira contínua ou repetitiva. 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 é chamado de para-analisador, que permite calcular o estado lógico de um enunciado a partir dos valores de μ_1, μ_2, V_{scc},V_{icc},V_{scct}, e V_{icct}.

4 VISUALIZAÇÃO, SIMILARIDADE, DISTÂNCIA E PROXIMIDADE

A Computação Gráfica nasceu com o intuito de melhorar a interação humano-computador e a manipulação computacional de imagens de diversos tipos. Esses recursos são potencialmente úteis para minimizar as dificuldades com os processos de Recuperação de Informação. Essas soluções

permitem conhecer melhor o espaço de busca, de modo a aumentar a efetividade dos resultados porque procuram retratar a estrutura semântica global de uma coleção de documentos (BORNER; CHEN; BOYACK, 2003).

Os sistemas de visualização de apresentação de coleções de documentos, de acordo com a abordagem escolhida, adotam metáforas diversas. 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. Por outro lado, o VxInsight (BORNER; CHEN; BOYACK, 2003) utiliza uma metáfora baseada em mapas geográficos. A tela principal do sistema exibe um agrupamento de "montanhas", de altura variável para representar agrupamentos de objetos de informação. A altura da montanha é proporcional à quantidade de documentos similares que formam os agrupamentos.

Outro sistema com capacidade para criar visualizações de coleções de documentos é o Projection Explorer – PEx, desenvolvido no Instituto de Ciências Matemáticas e de Computação da USP/São Carlos. O PEx utiliza pequenos círculos coloridos para representar documentos individuais de uma coleção (PAULOVICH; OLIVEIRA; MINGHIM, 2007; PAULOVICH et al., 2008)[1].

Independentemente da metáfora utilizada, os sistemas de apresentação de coleções de documentos realizam procedimentos e cálculos voltados a dois objetivos específicos: (i) o agrupamento de documentos considerados similares; (ii) a separação dos grupos de documentos dissimilares entre si. Para isso, os sistemas devem ser capazes de efetuar cálculos que indiquem o grau de similaridade entre os documentos.

Um dos modelos mais utilizados para este fim é o MEV, descrito na seção 2. A figura 7, a seguir, é um exemplo de visualização construída com o sistema PEx. Os círculos de mesma cor indicam documentos de um mesmo assunto, todos pertencentes a uma mesma coleção de teste. Na figura, os retângulos coloridos destacam grupos de documentos que, apesar de serem de assuntos distintos, aparecem próximos na visualização. Este efeito é uma conseqüência dos critérios que o sistema usa para construir a visualização. Nesse caso, os 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.

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

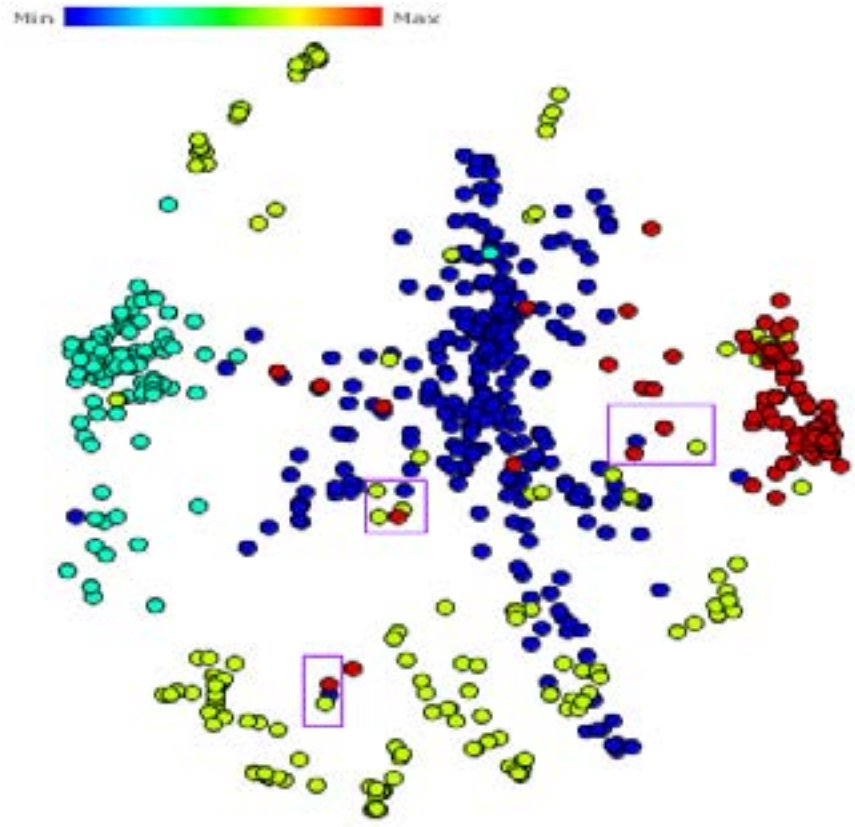


Figura 7 – Visualização de documentos no Projection Explorer – destaque p/ documentos de assuntos diferentes porem com algum grau de similaridade entre si
Fonte: Visualização criada, com o Projection Explorer

5 EXPERIMENTO COMPUTACIONAL DE INDEXAÇÃO COM CRITÉRIOS DA LÓGICA PARACONSISTENTE.

Os testes realizados para verificar os efeitos da utilização da Lógica paraconsistente nos procedimentos de indexação automática consistiram na inclusão do algoritmo para-analisador no código fonte do sistema PEx. Optou-se por utilizar, nos testes, as coleções de documentos disponibilizadas com o PEx, pois seu uso facilitaria a comparação entre as visualizações originais e as construídas com as alterações introduzidas no sistema.

Os testes efetuados podem ser sintetizados nos seguintes passos: (1) gerar a visualização de uma coleção de documentos com as facilidades de indexação/cálculos de similaridade embutidas originalmente no PEx; (2) utilizar o algoritmo para-analisador para avaliar e redefinir os índices e/ou os respectivos pesos estabelecidos para os documentos da coleção; (3) obter uma visualização da coleção a partir desses novos pesos;(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. As próximas seções detalharão esses aspectos bem como os resultados obtidos.

5.1 Coleções de teste utilizadas

A principal coleção de teste, constituída de 675 documentos é chamada pelos criadores do PEx de CBR-ILP-IR-SON, composto pelas siglas dos assuntos dos documentos que a compõem, ou seja: CBR – *Case-based reasoning* (raciocínio baseado em casos); ILP – *Inductive logic programing* (Programação lógica indutiva); IR – *Information Retrieval* (Recuperação de informação); e SON – *Sonification* (Sonificação). De acordo com Paulovich (2008), os documentos dos assuntos CBR e ILP foram retirados de periódicos dessas áreas. Os documentos dos assuntos IR e SON foram obtidos por meio de buscas na Internet, classificados de maneira aproximada, de acordo com a fonte onde foram obtidos (MINGHIM; LEVKOVITS, 2007).

Uma análise preliminar da coleção foi efetuada por Paulovich (2008) e uma das conclusões refere-se à não homogeneidade do conjunto de documentos de IR (*Information retrieval*). Segundo este autor, a não homogeneidade decorre do fato de os documentos classificados para este assunto terem sido obtidos por buscas genéricas na Internet, utilizando-se as palavras "information" and "retrieval". Nesse caso, elas não são suficientes para identificar precisamente a área. Em decorrência, espera-se que os objetos componentes desse 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.

Outras coleções foram construídas a partir da coleção original. Foram mantidas as abreviaturas dos assuntos da coleção original e, obtiveram-se as seguintes coleções: CBR-ILP-IR, com 574 documentos; CPR-ILP-SON, com 496 documentos; CBR-ILP, com 395 documentos. Todas as coleções são construídas a partir dos títulos, resumos e referências de cada documento.

5.2 Avaliação de visualizações

O PEx possui funcionalidades que permitem avaliar a qualidade das visualizações geradas, levando-se em conta diferentes aspectos.

Uma dessas técnicas é chamada de *neighborhood hit*. Ela objetiva analisar se, para uma pré-classificação efetuada na 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-existentes e a facilidade de encontrar limites bem definidos entre elas.

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 8, 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. Nesse gráfico, a situação representada pela curva vermelha é considerada a pior

de todas, pois apresenta os menores valores de precisão calculados (eixo vertical) para as quantidades de vizinhos especificadas (eixo horizontal). Analogamente, a curva azul é a situação com melhor avaliação de *neighborhood hit*.

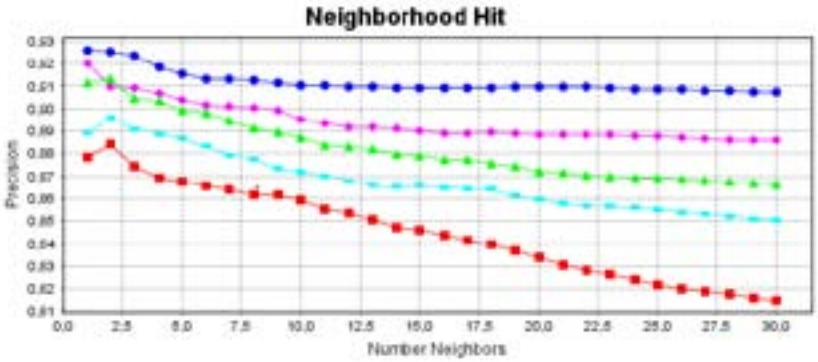


Figura 8 – Exemplo de curvas de Neighborhood Hit
Fonte – Produzida com a utilização do Projection Explorer

Outro método de avaliar as visualizações consiste em utilizar índices para medir a qualidade dos agrupamentos produzidos pelos diferentes algoritmos. Muitos desses métodos são baseados em análises estatísticas, utilizados na Análise de agrupamentos (*Cluster Analysis*). Um dos métodos, conhecido por coeficiente de silhueta (*silhouette coefficiente*) é um índice obtido numericamente. 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 (pior valor), e 1 (melhor valor).

O Projection Explorer utiliza outro método de visualização, ferramenta estatística chamada histograma de distâncias. Um histograma permite observar as distribuições de frequência de valores que ocorrem numa classe de variáveis observadas. O gráfico 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 é possível observar a distribuição das frequências de distâncias que ocorrem para a coleção. A figura 9 a seguir, é um exemplo de histograma de distância obtido no PEx.

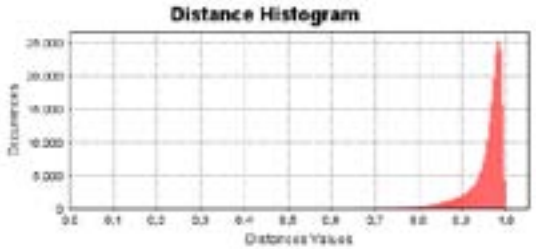


Figura 9 – Exemplo de histograma de distancias
Fonte: produzido no PEx

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

5.4 Uso do para-analisador no Projection Explorer

O MEV é o modelo básico utilizado para elaborar os cálculos de similaridade que irão definir a apresentação final (*layout*) da visualização. Nesse modelo, conforme exposto na seção 3.5, é montada uma matriz de 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. 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.

Para a 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 indicar. Para esse procedimento foi escolhido o valor de tf, do termo, para o grau de crença e valor df para o grau de descrença. 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.

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 Qv □ ⊥ (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, 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 anteriormente. 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.

5.5 Resultados obtidos

Os resultados do experimento foram compilados para cada coleção. São apresentados os

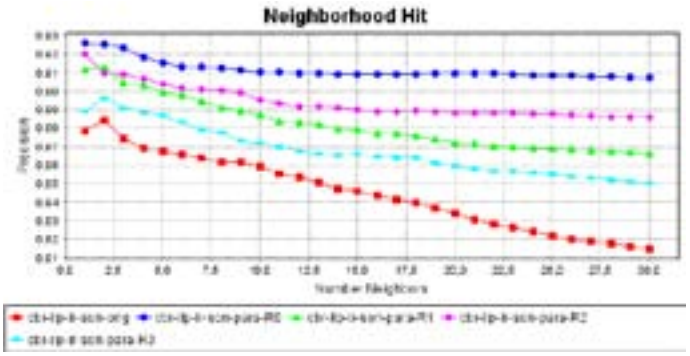
gráficos de *neighborhood hit*, os histogramas de distância e os valores dos coeficientes de silhueta obtidos nas visualizações construídas. As curvas são identificadas pelos nomes da coleção seguidas dos sufixos: *orig*, para visualizações construídas com o procedimento original do PEx; e *para-Rx*, 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. Também são mostradas as figuras geradas, para cada coleção. Foram destacados, em vermelho, os melhores valores obtidos para os coeficientes de silhueta. Devidos as restrições de espaço para esse trabalho, optamos por mostrar as regiões que apresentaram melhores valores de *neighborhood hit* e de coeficiente de silhueta.

Os histogramas de distâncias das coleções utilizadas indicam uma redistribuição das frequências. Podem-se observar 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 diminuição dos valores máximos de ocorrências de distância (ou similaridade) calculados para os documentos.

Coleção CBR-ILP-IR-SON

Legenda de cores para esta coleção:

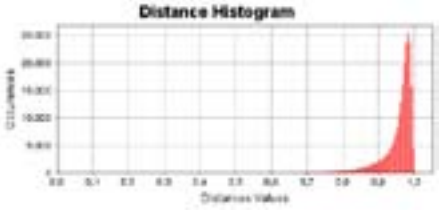
Azul – CBR (*Case based reasoning*)
Vermelho – SON (*Sonification*)
Lima – IR (*Information Retrieval*)
Verde – ILP (*Inductive Logic Programing*)



Visualização

Figura 10 – Visualização original
Fonte: produzida no PEx

Histograma de distancias



Coeficiente de Silhueta: - 0,058435094

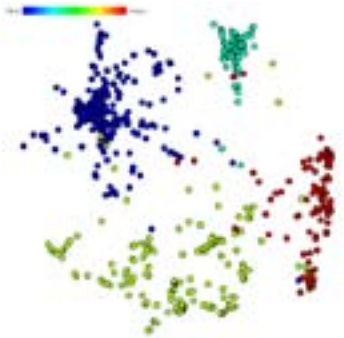


Figura 11 – Visualização da Região R0 do QUPC
Fonte: produzida no PEx

Coeficiente de Silhueta: 0,18285953

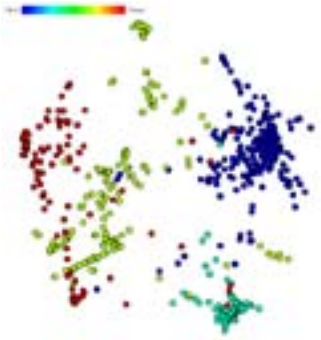
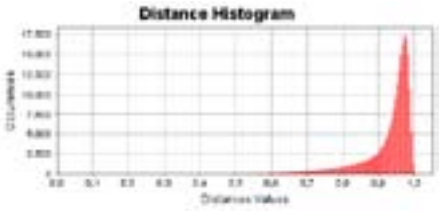


Figura 12 – Visualização da Região R1 do QUPC
Fonte: produzida no PEx

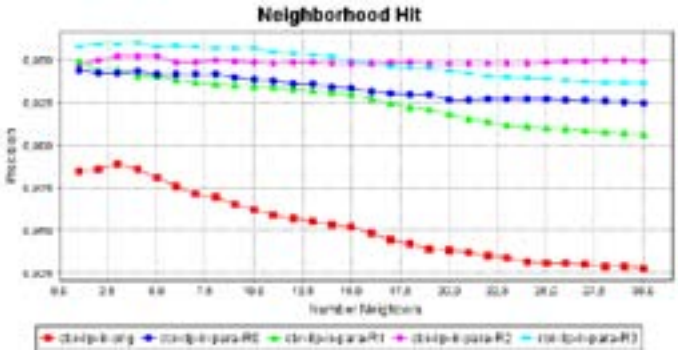


Coeficiente de Silhueta: **0,2880726**

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

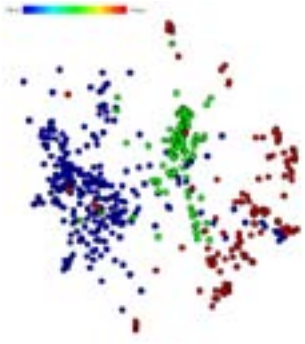
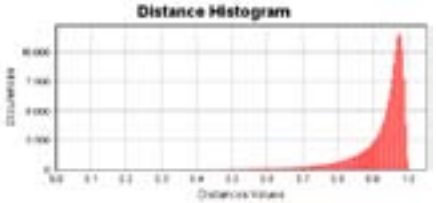
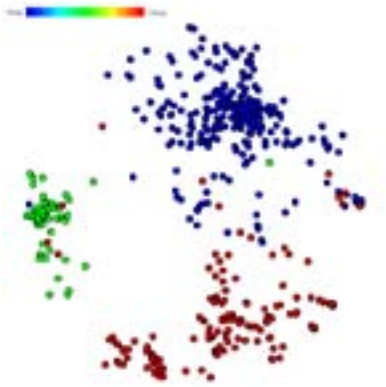


Figura 13 – Visualização original
Fonte: produzida no PEx

Histograma de distancias



Coeficiente de Silhueta: 0,19003317



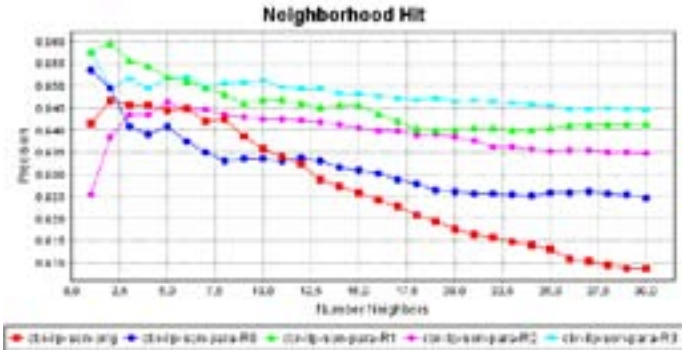
Coeficiente de Silhueta: **0,65129006**

Figura 14 – Visualização da Região R2 do QUPC
Fonte: produzida no PEx

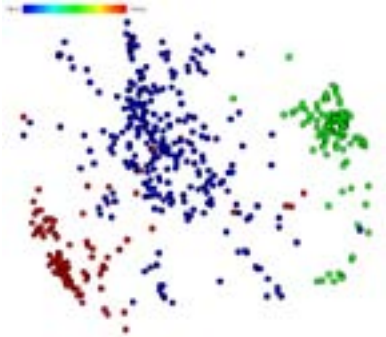
Coleção CBR-ILP-SON

Legenda de cores para esta coleção:

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



Visualização **Histograma de distancias**



Coeficiente de Silhueta: 0,41561985

Figura 15 – Visualização original
Fonte: produzida no PEx

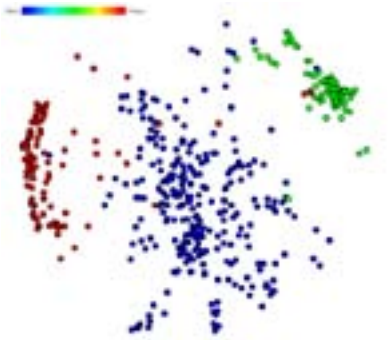
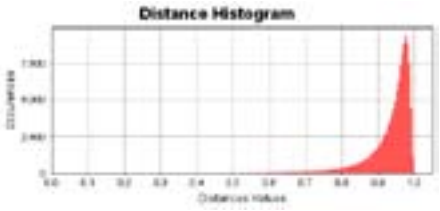


Figura 16 –Visualização da Região R3 do QUPC
Fonte: produzida no PEx

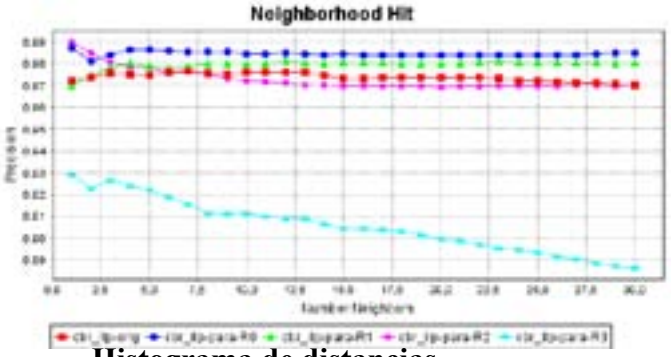


Coeficiente de Silhueta: **0,6606122**

Coleção CBR-ILP

Legenda de cores para esta coleção:

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



Visualização

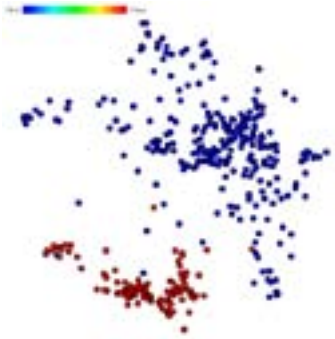


Figura 17- Visualização original
Fonte: produzida no PEx

Histograma de distancias

Coeficiente de Silhueta: 0,67798316

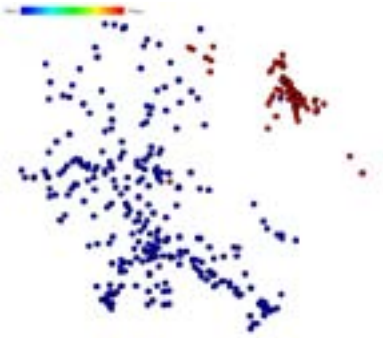


Figura 18 - Visualização da Região R2 do QUPC
Fonte: produzida no PEx

Coeficiente de Silhueta: **0,74633**

5.6 Análise dos resultados

Os documentos dos assuntos IR (*Information retrieval*) e SON (*Sonification*), conforme dito acima, não apresentam o mesmo grau de pureza quando comparados com os documentos de CBR (*Case-Based Reasoning*) e ILP (*Inductive Logic Programing*), pois foram obtidos a partir de consultas na Internet, ao passo que estes últimos foram selecionados de periódicos das respectivas áreas. É razoável supor que, para os documentos dos assuntos IR e SON, os termos extraídos dos documentos apresentam menor grau de representatividade do que aqueles extraídos para os assuntos CBR e ILP. Por outro lado, pode-se considerar que os documentos dos assuntos CBR e ILP, extraídos de periódicos especializados, utilizam com maior rigor a terminologia das respectivas áreas sendo portanto mais representativos. Pode-se considerar que as coleções CBR-ILP-IR-SON, CBR-ILP-IR, CBR-ILP-SON são coleções que apresentam em maior ou menor quantidade, certo grau de ruído. Por outro lado, a coleção CBR-ILP pode ser considerada como a que apresenta maior grau de pureza, pois foi criada com documentos de repositórios especializados.

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, em maior ou menor grau, algum ruído (documentos classificados sem rigor). Por outro lado, também se pode 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, 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, têm-se os seguintes valores:
Coleção CBR-ILP-IR-SON: total de documentos = 675; documentos ruído = 280
Coleção CBR-ILP-IR: total de documentos = 574; documentos ruído = 179
Coleção CBR-ILP-SON: total de documentos = 496; documentos ruído = 101.

Nas considerações anteriores, observa-se 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.

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 sugerem 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 de assuntos. 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, pode ser observado, nos testes realizados, uma melhora nos valores do coeficiente de silhueta dos agrupamentos produzidos pois o procedimento produz agrupamentos mais efetivos do que aqueles obtidos com os procedimentos originais do sistema.

6 CONSIDERAÇÕES FINAIS

O objetivo principal desta pesquisa - verificar o potencial e os efeitos da utilização da lógica paraconsistente em procedimentos de indexação automática -, procurou buscar interações entre diversas disciplinas ou campos de estudo, tais como: recuperação de informação, indexação automática, lógicas não-clássicas e visualização de informações. Tal objetivo foi desenvolvido por meio de dois procedimentos básicos. O primeiro consistiu em modificar o calculo tradicional de pesos, efetuado pelo modelo do espaço vetorial. Nesse contexto, utilizou-se o algoritmo para-analisador, desenvolvido no âmbito da lógica paraconsistente, para modificar os pesos atribuídos aos termos de indexação. 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.

Os resultados 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). Também se observaram modificações nas distribuições de frequência representadas nos histogramas de distâncias obtidos.

Os efeitos produzidos nos agrupamentos, demonstrados nos valores dos coeficientes de silhueta 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 consideram os efeitos da hipótese do *cluster*: de que documentos similares tendem a ser relevantes para uma questão formulada a um sistema de recuperação de informação (VAN RIJSBERGEN, 1979).

Os resultados 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.

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 devem se apoiar em ferramentas com capacidade para o tratamento desses aspectos, como o para-analisador.

Para Anderson e Perez-Caballo (2001), a atribuição de pesos aos termos de indexação é, muitas vezes, feita através de tentativa e erro, sem maiores justificativas teóricas. Os resultados desta pesquisa 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.

Pode-se concluir que a utilização de procedimentos para tratamento de incerteza, imprecisão e vagueza tem potencial para produzir ganhos mensuráveis nos processos de indexação automática. O experimento realizado mostrou o potencial positivo de utilização do para-analisador. 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.

De todo modo, é necessário aprofundar os aspectos teóricos do modelo do espaço vetorial, os critérios de atribuição de pesos aos termos de indexação e, finalmente, a concepção e realização de novos experimentos que capturem, com a devida precisão, os efeitos individuais do uso do para-analisador e da lógica paraconsistente em projetos de indexação automática.

Abstract: 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 use 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, was used 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 in the system. The results were evaluated by criteria and indexes built in the information visualization system itself, and demonstrate measurable gains in the construction, the quality of the displays, thus confirming the hypothesis of this research.

Keywords: Automatic indexing. Information visualization. Paraconsitent logic.

REFERÊNCIAS

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 documet similarities. *Information Visualization*, v. 1, n. 3/4, p. 166-181, 2002.

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, aplications*. Singapore: World Scientific Publishing Co Pte Ltd, 1995.

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

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

CORRÊA, C. A. *Indexação automática e visualização de informações: Um estudo baseado em lógica paraconsistente*. 2011. 152p. Tese (Doutorado em Ciência da Informação) – Escola de Comunicações e Artes, Programa de pós graduação em Ciência da Informação, Universidade de São Paulo, São Paulo 2011.

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.

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.

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.

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

LANCASTER F. W. *Indexação e Resumos: teoria e prática*, Brasília: Briquet de Lemos, 2004.

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

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

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

PAULOVICH, F.; OLIVEIRA, M. C. F.; MINGHIM, R. The Projection Explorer: A Flexible tool for projection-based multidimensional visualization, *Proc. of XX Braziliam Sumposium 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.

RAGHAVAN, P. Information retrieval algorithms: a survey. *Proceedings of the eight annual ACM-SIAM - Symposium on Discret algorithms*, Society for industrial and applied mathematics, Philadelphia, PA, USA, p. 11-18, 1997.

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

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

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

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.
TAN, P.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. Boston: Pearson Education Inc, 2006.
TOLEDO, G. L.; OVALLE, I. I. *Estatística Básica*. São Paulo: Editora Atlas S.A., 1986.
VAN RIJSBERGEN, C. J. *Information retrieval*. London: Butterworths, 1979.