Múltiplas visões coordenadas para exploração de mapas de similaridade

Danilo Medeiros Eler

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Múltiplas visões coordenadas para exploração de mapas de similaridade

Danilo Medeiros Eler

Orientadora: ***Profa. Dra. Rosane Minghim***

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*.

USP – São Carlos
Maio de 2011

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

E37m

Eler, Danilo Medeiros
Múltiplas visões coordenadas para exploração de mapas de similaridade / Danilo Medeiros Eler; orientadora Rosane Minghim -- São Carlos, 2011.
138 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2011.

1. Visualização. 2. Múltiplas visões coordenadas. 3. Técnicas de coordenação. 4. Mapas de similaridade. I. Minghim, Rosane, orient. II. Título.

Agradecimentos

Agradeço a Deus por sua provisão em todas as etapas do doutorado. Pude ter tranquilidade mesmo nos momentos difíceis. Fiz a minha parte e Ele me ajudou.

Agradeço a minha esposa (Juliana) por sempre estar ao meu lado, cuidando de mim e me dando apoio para concluir o doutorado. Sua presença foi fundamental.

Agradeço a meus pais (Rubens e Glória) pela educação e amor que sempre me deram. Também agradeço a meus irmãos Daniela e Marcelo, que estiveram também me apoiando e torcendo por mim. Um agradecimento especial ao Marcelo que me ajudou com os diagramas do modelo de coordenação.

Agradeço a professora Rosane por ter me aceitado como aluno e ter confiando em mim. Aprendi muito durante desenvolvimento desta tese, observando sua seriedade com a pesquisa.

Agradeço os professores Cristina, Fernando, João e Gustavo por terem me ajudado em vários momentos do doutorado, nos trabalhos que fizemos em conjunto.

Agradeço a todos meus familiares e amigos que torceram por mim. E a toda a igreja daqui de São Carlos e lá de Martinópolis que sempre estiveram orando por mim.

Agradeço a todos amigos que fiz aqui em São Carlos. Em especial, agradeço os amigos de laboratório que também participaram do desenvolvimento deste trabalho.

Agradeço a todos os funcionários do ICMC; os professores que mostraram ser grandes conhecedores da área em que atuam; as secretárias da pós, sempre pacientes e atenciosas; os guardas da portaria, sempre mostrando companheirismo; o pessoal da biblioteca; e os funcionário da limpeza, sem eles seria difícil.

Agradeço a FAPESP pelo auxílio financeiro, que foi de grande importância para o desenvolvimento deste trabalho.

Não andeis ansiosos de coisa alguma; em tudo, porém, sejam conhecidas, diante de Deus, as vossas petições, pela oração e pela súplica, com ações de graças. E a paz de Deus, que excede todo o entendimento, guardará o vosso coração e a vossa mente em Cristo Jesus.
(Filipenses 4:6-7)

Resumo

Atualmente, diversas áreas de aplicação necessitam de mecanismos mais efetivos para analisar dados provenientes de naturezas distintas. Tipicamente, esses dados são abstratos, não estruturados e possuem uma natureza multidimensional (e.g., coleções de documentos). Dados que não possuem uma natureza multidimensional podem ser representados como tal por meio da aplicação de algoritmos extratores de características (e.g., coleções de imagens). Assim, técnicas de visualização de informação projetadas para interpretar dados multidimensionais podem ser aproveitadas para analisar dados não estruturados.

Esta tese empregou técnicas de visualização de informação para construir mapas de similaridade a partir de dados multidimensionais como uma forma de representação desses dados, uma vez que as técnicas para construi-los tem evoluído com a expansão dos campos de aplicação. Novas técnicas para coordenação de múltiplas visões foram desenvolvidas para permitir a exploração de conjuntos de dados, a partir de mapas de similaridade gerados por diferentes técnicas de construção de mapas, diferentes parâmetros ou ainda diferentes conjuntos de dados. As técnicas de coordenação desenvolvidas são baseadas em identificador, em distância, em tópicos, na identificação de tópicos em coleções que evoluem no tempo, e em uma técnica que combina o mapeamento de diferentes técnicas de coordenação.

Esta tese também apresenta aplicações das técnicas de coordenação desenvolvidas e das ferramentas construídas para análise de coleções de documentos, coleções de imagens e dados volumétricos, empregando coordenações de mapas de similaridade.

As técnicas de coordenação desenvolvidas são apoiadas por um modelo de coordenação que estende um modelo previamente proposto na literatura. O modelo estendido permite a configuração de técnicas de coordenação durante a exploração, admitindo diferentes tipos de mapeamentos. Uma característica importante do modelo é permitir o desenvolvimento de mapeamentos dinâmicos para técnicas de coordenação, isto é, mapeamentos que podem mudar o comportamento de acordo com a interação do usuário.

Como resultado desta tese, está disponível um arcabouço para visualização coordenada de múltiplos mapas de similaridade, composto por um modelo, um conjunto de técnicas e um conjunto de ferramentas que efetivamente permitem a análise visual de conjuntos de dados multidimensionais.

Abstract

Currently, various fields of application need effective mechanisms to analyse data differing in nature. Typically these data are abstract, unstructured and multidimensional (e.g. document collections). Data that do not present multidimensional description can be represented as such by means of feature extraction algorithms (e.g. image collections). Thus, information visualization techniques designed to interpret multidimensional data sets can be employed to analyse unstructured data.

This thesis employed information visualization techniques that build similarity maps from multidimensional data as a form of data representation, since the techniques to construct them have evolved lately with expanding fields of application. Novel techniques for coordination of multiple views were developed that allow exploration of data sets, from similarity maps generated using different techniques for building maps, different parameters or even different data sets. The developed coordination techniques are based on identity relationships, on distance relationships, on topic coverage (for text or other annotated data) and on evolution of topic coverage (also for text). An approach to combine different coordination techniques was also developed.

This thesis also reports on applications of the coordination techniques developed, and on tools built for analysis of image, text and volumetric data employing coordinated similarity maps.

The techniques developed in this work are supported by a coordination model that extends a model previously proposed in literature. The extended model allows the definition and configuration of coordination techniques during coordination tasks and performing various types of mappings. An important feature of the model is to support the development of dynamic mappings, which are mappings that may change behavior according to user interaction.

As a result of this thesis, a framework is available for coordinated visualization of multiple similarity maps, composed by a model, a set of techniques and a set of implemented tools that effectively support the visual analysis of multidimensional data sets.

Sumário

Resumo **i**

Abstract **ii**

1 Introdução **1**
1.1 Motivação 2
1.2 Objetivos e Contribuições 4
1.3 Nomenclatura 6
1.4 Organização 7

2 Mapas de Similaridade **9**
2.1 Considerações Iniciais 9
2.2 Métodos para Construção de Mapas de Similaridade 10
2.2.1 Mapas de Similaridade de Documentos 14
2.2.2 Mapas de Similaridade de Imagens 26
2.3 Considerações Finais 31

3 Múltiplas Visões Coordenadas **32**
3.1 Considerações Iniciais 32
3.2 Exemplos de Aplicações de Múltiplas Visões Coordenadas 33
3.3 Modelos de Coordenação 47
3.4 Considerações Finais 54

4 O Modelo de Coordenação **56**
4.1 Considerações Iniciais 56
4.2 Descrição do Modelo de Coordenação 56
4.2.1 Processo de Exploração 57
4.2.2 Descrição das Classes 58
4.2.3 Descrição do Processo de Coordenação 60
4.3 Instanciações do Modelo de Coordenação 63
4.3.1 Identity Coordination 64
4.3.2 Distance Coordination 65
4.3.3 Topic Coordination 65
4.3.4 Topic Time Coordination 66
4.3.5 Time Coordination 67

4.3.6 Joint Coordination 68
4.3.7 Nova Técnica de Coordenação 68
4.4 Considerações Finais 70

5 Técnicas de Coordenação entre Mapas de Similaridade 71
5.1 Considerações Iniciais 71
5.2 Projection Explorer 72
5.3 Exploração Coordenada entre Mapas de Similaridade 75
5.3.1 Coordenação baseada em Identificador 76
5.3.2 Coordenação baseada em Tópicos 79
5.3.3 Coordenação baseada em Tópicos no Tempo 82
5.3.4 Coordenação baseada em Atributos Temporais 85
5.3.5 Coordenação baseada em Distância 86
5.3.6 Combinação de Coordenações 88
5.4 Considerações Finais 90

6 Aplicações para Documentos e Imagens 92
6.1 Considerações Iniciais 92
6.2 Exploração de Coleções de Documentos 93
6.2.1 Coordenação Baseada em Identificador 94
6.2.2 Coordenação Baseada em Tópicos 101
6.2.3 Coordenação Baseada em Tópicos no Tempo 104
6.2.4 Coordenação baseada em Distância 106
6.2.5 Combinação de Coordenações 108
6.3 Exploração de Coleções de Imagens 110
6.3.1 Coordenação baseada em Identificador 113
6.3.2 Coordenação baseada em Distância 120
6.3.3 Coordenação entre Mapas de Imagens e Mapas de Textos 121
6.4 Exploração de Dados Volumétricos 125
6.4.1 Coordenação entre o mapa de similaridade e o espaço do objeto 126
6.5 Considerações Finais 130

7 Conclusões e Trabalhos Futuros 133
7.1 Conclusões 133
7.2 Trabalhos Futuros 136

A Lista de Publicações A1

Lista de Figuras

2.1 Mapa Auto-organizável de 2764 artigos. 10
2.2 Passos para criar um mapa de similaridade por meio de uma técnica de projeção. . 11
2.3 Mapas de Similaridade gerados por Projeção Multidmensional e Posicionamento de Pontos Baseado em Árvore. 13
2.4 Rede de co-citação com 380 autores com 9 ou mais citações. 14
2.5 Rede de co-citação apresentada na Figura 2.4. 15
2.6 Três tipos de nós especiais utilizados na rede de co-citação. 15
2.7 União de 624 nós de diferentes janelas de tempo. 16
2.8 Visualização de uma rede de co-citação de artigos sobre NSAID entre 1990 a 2004. 16
2.9 Visão das janelas de tempo da evolução de artigos sobre NSAID. 17
2.10 Rede de co-autoria. 17
2.11 Rede completa de co-autoria com componentes conexas com tamanho maior do que três. 18
2.12 Agrupamentos por família de patentes. 19
2.13 Exploração de família de patentes. 19
2.14 Visualização das 50 co-palavras com maiores freqüência e *burst* nas publicações dos PNAS. 20
2.15 Visualização de seis agrupamentos (tópicos) com o *3D Explorer*. 21
2.16 *Concept Density Maps* mostrando a correlação entre tópicos. 21
2.17 Visualição utilizando Galáxias. 22
2.18 Visualição utilizando *ThemeView*. 22
2.19 Visualização de um conjunto de documentos com o *VxInsight*. 23
2.20 Mapa de documentos gerado pela ferramenta *Projection Explorer*. 24
2.21 Mapa de 675 documentos de quatro áreas distintas gerado pela técnica *Hipp*. 25
2.22 Mapa de 300 documentos de quatro áreas distintas gerado pela técnica *incBoard*. . 25
2.23 Parte de um mapa de 675 documentos de quatro áreas distintas gerado pela técnica *hexBoard*. 26
2.24 Coleção de imagens organizadas pela similaridade do histograma de cor. 27
2.25 Recuperação de um conjunto de imagens baseada em uma ontologia. 28
2.26 Coleção de imagens de face com variação de brilho e orientação. 28
2.27 Projeção de amostras de uma coleção de imagens. 29
2.28 Construção de dois mapas da coleção de imagens ImageCLEF utilizando as técnicas MDS e ISOMAP. 29
2.29 Visualização de uma coleção de imagens da base MPEG-7. 30
2.30 Aplicação da técnica *incBoard* para exploração de imagens. 30

3.1 Visualizações coordenadas de um sistema de arquivos utilizando *Treemap*. 34
3.2 Visualização de um conjunto de dados por meio de Coordenadas Paralelas e Matriz de Permutação. 35
3.3 Interface do sistema PRISMA. 35
3.4 Técnicas de visualização e coordenação do sistema PRISMA. 36
3.5 *VisLink*: diferentes posicionamentos dos planos de visualização. 36
3.6 Visões para análise de dados estruturados e não estruturados. 37
3.7 *WireVis*: visão *Keyword Network*. 39
3.8 *WireVis*: visão *Heatmap*. 39
3.9 *WireVis*: visão *String and Bead*. 40
3.10 *WireVis*: ferramenta Busca por Exemplo. 40
3.11 *MetaCrystal*: visão *Category View*. 41
3.12 *MetaCrystal*: visão *Cluster Bulls-Eye*. 42
3.13 *MetaCrystal*: visão *RankSpiral*. 42
3.14 *Cluster Bulls-Eye* mostrando os 50 primeiros resultados de cinco diferentes motores de busca. 44
3.15 Painel *FreqCloud* mostrando as palavras mais freqüêntes. 44
3.16 Painel *Topic Panel* mostrando os tópicos dos resultados de busca. 45
3.17 Seleção de um tópico no painel *Topic Panel*. 45
3.18 Lista de ranking exibindo opções de *Ranking*, *Abstract* e *URL*. 45
3.19 Exploração de uma coleção de documentos por meio do arcabouço *VisTools*. 46
3.20 *Snap*: três visões coordenadas para exploração de um sistema de arquivos. 49
3.21 *Snap*: especificação da coordenação da visualização da Figura 3.20. 49
3.22 Modelo abstrato de coordenação de Boukhelifa e Rodgers. 51
3.23 Sistema *CViews* mostrando três visões coordenadas. 52
3.24 *Improvise*: coordenação direta. 53
3.25 *Improvise*: coordenação indireta. 54

4.1 Estados do processo exploratório. 58
4.2 Classes do modelo de coordenação. 59
4.3 Processo *Create View*. 61
4.4 Processo *Create Coordination*. 62
4.5 Processo *Coordinate*. 62
4.6 Descrição geral dos passos de execução do método *coordinate*. 63
4.7 Instanciações das técnicas de coordenação utilizadas. 64

5.1 Mapas de Similaridade gerados por Projeção Multidmensional e Posicionamento de Pontos Baseado em Árvore. 72
5.2 Técnicas de interação da *Projection Explorer*. 73
5.3 Exploração de um conjunto de notícias por meio de técnicas de detecção de tópicos. 74
5.4 Exploração coordenada de múltiplos mapas de similaridade utilizando três tipos de técnicas de coordenação e três conjuntos de dados. 76
5.5 Aplicação da *Identity Coordination* para coordenar dois mapas de similaridade. . . 77
5.6 Outra aplicação da *Identity Coordination* para coordenar dois mapas de similaridade. 77
5.7 Mapas construídos a partir de um conjunto de 138 sequências de proteína. 78
5.8 Partes do mapa da Figura 5.7(a), mostrando o padrão dos *contact maps*. 79
5.9 Exemplo de aplicação da técnica *Topic Coordination*. 82
5.10 *Topic Time Coordination* (Eler et al., 2009b) aplicada em três *TS Maps*. 84
5.11 Exemplo de aplicação da técnica *Time Coordination*. 85

5.12 Construção de novos espaços multidimensionais com atributos comuns. 87
5.13 Exemplo de aplicação da técnica *Distance Coordination*. 88
5.14 Exemplo da *Joint Coordination* definida com as técnicas *Topic Coordination* e *Time Coordination*. 90

6.1 Comparação entre mapas de similaridade gerados por diferentes configurações de medidas de similaridade. 95
6.2 Comparação da precisão de diferentes medidas de similaridade utilizadas para gerar os mapas da Figura 6.1. 96
6.3 Comparação de mapas gerados por diferentes espaços multidimensionais. 98
6.4 Gráfico dos valores de *Neighborhood Hit* calculado a partir dos mapas apresentados nas Figuras 6.3(a) e 6.3(b). 99
6.5 Comparação entre as técnicas LSP e NJ para gerar mapas de similaridade. 99
6.6 Exploração coordenada de notícias usando as técnicas LSP e NJ. 100
6.7 Exemplo de aplicação da técnica *Topic Coordination*. 102
6.8 Detalhamento da vizinhança dos documentos destacados nas Figuras 6.7(b) e 6.7(c). 103
6.9 *Topic Time Coordination* entre cinco *Time-Span Maps*. 105
6.10 Exemplo de aplicação da técnica *Distance Coordination*. 107
6.11 Aplicação da *Joint Coordination* configurada com as técnicas *Topic Coordination* e *Distance Coordination*. 109
6.12 Documentos que tiveram a maior cobertura do tópicos, mas não foram destacados. 110
6.13 *Projection Explorer for Images*: interações para exploração de coleções de imagens. 111
6.14 Mapas de similaridade de um conjunto de imagens *MRI-12* e gráfico de precisão. . 114
6.15 Mapas e comparação da precisão calculada por *Neighborhood Hit*. 115
6.16 Mapas de similaridade de um conjunto de imagens médicas construídos com diferentes métricas de distância. 116
6.17 Mapas construídos a partir da coleção *ImageCLEF-2006* utilizando características *Grupo Dois*, *Grupo Um* e wavelet. 118
6.18 Mapas de similaridade construídos a partir de um espaço de características completo e de um espaço reduzido. 119
6.19 Exploração da coleção *ImageCLEF-2006* por meio da técnica *Distance Coordination*. 120
6.20 Coordenação de mapas construídos a partir das imagens da coleção *FIGURES* e das descrições. 122
6.21 Coordenação entre mapa de texto e mapa de imagens. 123
6.22 *Topic Coordination* aplicada em mapa de texto e mapa de imagem. 124
6.23 Coordenação entre um mapa de similaridade e o espaço do objeto. 127
6.24 Projeções P-LSP de uma amostra de um volume de cabeça. 128
6.25 Projeção do *IEEE Visualization 2008 Contest*. 129
6.26 Exploração de um conjunto de dados de uma simulação. 132

CAPÍTULO

1

Introdução

Atualmente, muitas aplicações disponibilizam dados em grandes quantidades, os quais podem ser provenientes de simulações, sensores, coleções de documentos, censos diversos e coleções de imagens. Em contraste à capacidade de geração de dados, a tecnologia para análise e descoberta do conhecimento a partir das observações promovidas pela sua interpretação é ainda limitada. Pesquisas em algoritmos envolvendo estatística, mineração e visualização desses dados enfrentam o desafio de lidar com os números, isto é, tanto as grandes quantidades de elementos quanto a grande quantidade de variáveis (atributos, dimensões) observadas. A integração dessas áreas para exploração e análise de conjuntos de dados é conhecida como Análise Visual (*Visual Analytics*) (Wong e Thomas, 2004).

Em particular, abordagens que utilizam técnicas de visualização são largamente empregadas com o objetivo de dar sentido a esses dados, por meio de representações visuais e de formas de interação (Card et al., 1999). As representações visuais, criadas a partir das informações contidas em um conjunto de dados, procuram realçar o relacionamento entre as instâncias ou entre os atributos que as descrevem. Dessa maneira, o usuário pode interagir sobre o conjunto de dados e descobrir padrões ou confirmar e levantar hipóteses. Além disso, as representações visuais também buscam facilitar a descoberta de anomalias presentes em um conjunto de dados. Algumas dessas técnicas facilitam a análise das relações entre os atributos, por exemplo, Coordenadas Paralelas (Inselberg e Dimsdale, 1990; Inselberg, 2009); e um outro grupo de técnicas auxilia a análise das relações entre as instâncias de dados, por exemplo, técnicas de projeção multidimensional (Tejada et al., 2003). Um estudo sobre algumas das técnicas tradicionais de visualização pode ser encontrado em (de Oliveira e Levkowitz, 2003).

Técnicas de projeção multidimensional são comumente aplicadas para gerar representações gráficas de um conjunto de dados multidimensional, permitindo que o usuário compreenda como as instâncias estão relacionadas no espaço original. Essas técnicas trabalham projetando os dados de um espaço m-dimensional, também chamado de espaço original, para um espaço p-dimensional, também chamado de espaço projetado, onde $p << m$ ($p = 2$ ou $p = 3$). Para aplicações de visualização, as técnicas de projeção variam em sua abordagem, mas o objetivo comum é que a representação dos dados no espaço projetado deva refletir o espaço original, preservando as relações de distância definidas nesse espaço (Paulovich, 2008). Nos últimos anos, o grupo de Visualização, Imagens e Computação Gráfica do ICMC/USP [1] tem desenvolvido e utilizado técnicas de projeção multidimensional para gerar representações visuais de diferentes tipos de dados, por exemplo, coleções de documentos, séries temporais, coleções de imagens e outros. Este trabalho empregou as técnicas de projeção multidimensional e as técnicas de posicionamento de pontos desenvolvidas no grupo para construir mapas de similaridade de dados. Esses mapas são representações visuais que refletem as relações de similaridade entre as instâncias de um conjunto de dados.

Apesar das diferentes técnicas de visualização de informação colaborarem para a exploração e análise de diferentes conjuntos de dados, elas possuem pontos fortes e fracos, o que não permite a utilização de uma única técnica para produzir bons resultados para todo tipo de dados. Tomando como exemplo a bem conhecida técnica Coordenadas Paralelas, ela possui a vantagem de exibir muitos atributos simultaneamente e facilita a localização de agrupamentos de dados que tenham o mesmo comportamento. Um ponto fraco da técnica Coordenadas Paralelas é a dificuldade que o usuário tem em analisar um conjunto de dados com milhares de instâncias, não permitindo que seja encontrada com facilidade os agrupamentos. Na tentativa de suavizar os pontos fracos de uma técnica de visualização, diferentes técnicas podem ser utilizadas de forma complementar, possibilitando que o usuário tenha várias visões sobre um conjunto de dados. Uma visão, no contexto deste trabalho, significa uma representação visual de um conjunto de dados por meio de uma técnica de visualização. Utilizando várias técnicas de visualização sobre um conjunto de dados, têm-se várias visões, facilitando sua exploração. Adicionalmente, para auxiliar o usuário na troca entre as visões, pode-se utilizar um mecanismo de coordenação, o qual tem o objetivo de associar os elementos explorados nas diversas visões. A área que estuda como combinar múltiplas visões para obter resultados eficientes de associação entre informações é conhecida como Múltiplas Visões Coordenadas (*Coordinated and Multiple Views* – CMV), ou também *Linked Views*.

1.1 Motivação

Observando os trabalhos desenvolvidos na área de CMV, nota-se uma grande variedade de aplicações em diferentes áreas de pesquisa. Com o intuito de construir aplicações CMV com propósitos gerais, alguns modelos foram elaborados (North e Shneiderman, 2000a; Boukhelifa e

[1] http://infoserver.lcad.icmc.usp.br/

Rodgers, 2003; Weaver, 2006). Apesar da grande variedade de aplicações em CMV, poucas focaram na análise de mapas de dados bidimensionais e das que utilizam múltiplas visões, poucas são as que exploram novos mecanismos de coordenação para enriquecer o processo de exploração. Especificamente, poucas aplicações CMV são direcionadas para a exploração de coleções de documentos ou de mapas de similaridade. As técnicas de coordenação utilizadas na exploração utilizam somente a clássica técnica *brushing and linking*, que coordena as instâncias com base em seu identificador. Assim, a exploração baseada nesse tipo de coordenação é muito limitada, não permitindo, por exemplo, que diferentes conjuntos de dados sejam explorados.

Para que diferentes conjuntos de dados sejam simultaneamente explorados, em um mesmo processo exploratório de CMV, é necessário desenvolver novas técnicas de coordenação [2]. Alguns sistemas disponibilizam a coordenação baseada em tópico, com a qual é possível selecionar um tópico de uma ontologia ou um tópico extraído de um grupo de documentos. No entanto, na maioria dos sistemas, não é possível escolher quais documentos serão utilizados para gerar o tópico ou uma lista de tópicos. Além disso, é notável que, de um modo geral, os sistema utilizam coordenações fixas, isto é, eles não permitem a definição ou configuração das técnicas de coordenação durante o processo de exploração. Também, as técnicas já predefinidas no sistema, não permitem que parâmetros sejam configurados durante a interação. Em muitos casos, isso é uma limitação do modelo de coordenação empregado na construção do sistema de CMV.

Como foi anteriormente descrito, há alguns modelos de coordenação que apóiam o desenvolvimento de aplicações CMV. Esse modelos, no entanto, não estão totalmente preparados para desenvolver aplicações CMV que utilizem novas técnicas de coordenação voltadas para exploração de coleções de documentos, por exemplo. Esse tipo de exploração necessita de novas técnicas de coordenação que utilizem informações extraídas do conjunto de dados para auxiliar a exploração do mesmo conjunto ou de conjuntos distintos. Além disso, a criação e configuração das técnicas de coordenação durante o processo de exploração é fundamental para seu enriquecimento. Os modelos que permitem isso, não são de fácil extensão e não são eficientes para trabalhar com coleções de documentos, pois utilizam como base o modelo relacional (Weaver, 2004). Além de auxiliar a exploração de coleções de documentos, modelos que permitam o desenvolvimento de tais técnicas de coordenação também auxiliariam a exploração de outros tipos de conjuntos de dados.

Nos últimos anos, técnicas de visualização tem sido empregadas para exploração de coleções de imagens. No entanto, são raras as aplicações que utilizam múltiplas visões coordenadas para explorar esse tipo de dado. Utilizar CMV para explorar coleções de imagens é uma estratégia que aumenta a capacidade de entendimento do espaço de características, facilitando a identificação dos parâmetros ideais para construir um espaço de características que melhore a classificação do conjunto (Brandoli et al., 2010). Além disso, uma abordagem CMV poderia utilizar diferentes técnicas de coordenação para auxiliar a exploração simultânea de diferentes conjuntos de dados, sejam eles

[2] Ver final da Seção 1.3 para entender como os conceitos tipo de coordenação e técnica de coordenação são empregados no contexto deste trabalho

de mesma natureza ou não. Por exemplo, a exploração de diferentes coleções de imagens ou a exploração de coleções de imagens relacionadas a coleções textuais.

Um campo recente de pesquisa é utilizar a abordagem CMV para unir técnicas de visualização de informação com técnicas de visualização científica. Geralmente, nessas aplicações, as técnicas de visualização de informação são utilizadas para explorar o espaço do objeto indiretamente por meio de visualizações do espaço multidimensional que representa os dados (também chamado de espaço de características). Por meio da coordenação entre visualizações, as interações realizadas no espaço de características são refletidas no espaço do objeto, onde os dados são visualizados por meio de uma técnica de visualização científica (Blaas e Post, 2008; Linsen et al., 2008, 2009). As abordagens atualmente propostas utilizam técnicas que exploram a relação entre os atributos dos dados. Atualmente existem técnicas rápidas de projeção multidimensional que podem auxiliar nesse tipo de exploração, facilitando a identificação de grupos de dados similares. Por exemplo, consideranto a visualização de um volume, uma técnica de projeção construiria mapas com agrupamentos de voxels similares.

1.2 Objetivos e Contribuições

Visando propor soluções para os problemas levantados na seção anterior e também evoluir o número de ferramentas para auxiliar no processo de visualização exploratória, este trabalho integrou mapas de similaridade com uma abordagem CMV. O objetivo geral foi desenvolver técnicas de coordenação para auxiliar a exploração de coleções de dados baseada em mapas de similaridade e propor uma abordagem CMV para explorar concomitantemente mais de um mapa de similaridade. Com isso, o processo exploratório pode envolver mapas gerados por técnicas diferentes, mapas de conjuntos de dados diferentes ou mapas cuja natureza dos dados seja distinta. Além disso, as técnicas de coordenação são capazes de coordenar diferentes técnicas utilizadas para a construção do mapa de similaridade. Por exemplo, mapas baseados em projeção multidimensional e mapas baseados em árvore. Todas estas técnicas devem ser apoiadas por um modelo que permita a criação de configuração das técnicas de coordenação em tempo de execução. Com estes objetivos em foco, os resultados e contribuições alcançados neste trabalho foram:

- **Novas técnicas de coordenação entre mapas de similaridade, e entre esses e técnicas de visualização científica:** foram desenvolvidas novas técnicas baseadas em tópico, distância e elementos temporais para auxiliar na exploração de conjuntos de dados. Essas técnicas permitiram que diferentes coleções pudessem ser exploradas de maneira conjunta, utilizando o conhecimento de uma coleção para explorar outras ou focar em partes de interesse (Eler et al., 2008b, 2009b). Também foi desenvolvida uma ferramenta para exploração de dados volumétricos por meio de mapas de similaridade e a coordenação destes com o espaço do objeto de dados volumétricos (Paulovich et al., 2010);

- **Nova técnica para coordenação entre mapas de documentos baseadas em tópicos e na evolução temporal dos conjuntos de dados:** foi desenvolvida uma técnica que utiliza a vizinhança de documentos de interesse para identificar a evolução de tópicos no tempo. Os tópicos são extraídos da vizinhança, que é modificada conforme a coleção de documentos evolui no tempo, isto é, conforme novos documentos são incorporados à ela (Eler et al., 2009b);

- **Um sistema que especializa a visualização por meio de mapas de similaridade e CMV para o caso de coleções de imagens:** foi desenvolvido um sistema, denominado *Projection Explorer for Images* (PEx-Image), que permite explorar diferentes coleções de imagens e utilizar a coordenação no processo de exploração. Além da coordenação auxiliar a exploração, também foi possível comparar diferentes espaços multidimensionais, os quais foram construídos por diferentes algoritmos extratores de características; comparar diferentes métricas de distância para calcular a similaridade entre as instâncias; comparar resultados de seletores de características e de classificadores. Esse sistema também pode utilizar resultados de classificação para auxiliar a exploração de coleções de imagens não rotuladas e a exploração simultânea de mapas de textos e de imagens (Eler et al., 2008a, 2009a);

- **Técnicas de coordenação que podem ser configuradas durante o processo de exploração:** algumas das técnicas desenvolvidas neste trabalho podem ser configuradas pelo usuário, produzindo diferentes mapeamentos. Além disso, o conceito de mapeamento dinâmico é utilizado nas técnicas, o qual permite que o mapeamento criado entre as instâncias das visões envolvidas no processo de coordenação mude de acordo com a interação do usuário, que introduz elementos dinâmicos para as técnicas de coordenação (Eler et al., 2008b, 2009b);

- **Nova técnica de coordenação que permite a combinação de mapeamentos:** com o objetivo de unir a capacidade de exploração de técnicas de coordenação distintas, foi desenvolvida uma técnica que combina, em um único mapeamento, os mapeamentos de diferentes técnicas de coordenação escolhidas pelo usuário;

- **A adaptação de um modelo de coordenação:** foi adaptado um modelo de coordenação para que as técnicas propostas fossem desenvolvidas e aplicadas. Esse modelo possibilita que as técnicas de coordenação sejam definidas durante o processo de exploração e que diferentes tipos de mapeamentos possam ser criados e combinados, os quais podem ser estáticos ou dinâmicos.

A próxima seção descreve alguns termos que serão utilizados na tese.

1.3 Nomenclatura

O objetivo desta seção é explicar e relatar os significados de alguns termos que serão utilizados no restante desta tese. Esses termos são:

- **Instância, dado, ponto, elemento, amostra:** são os itens que formam um conjunto de dados. Por exemplo, um documento é uma instância de uma coleção de documentos;
- **Atributo, coordenada, dimensão, característica:** é a propriedade que descreve as instâncias de um conjunto de dados. Por exemplo, um termo (palavra) é um atributo de um espaço multidimensional construído a partir de uma coleção de documentos. O conjunto de atributos que descreve uma instância é um **vetor de atributos**;
- **Vetor de atributos, vetor multidimensional, vetor de características:** é a representação de uma instância por meio de atributos. Por exemplo, um documento é representado por um conjunto de atributos (termos), que forma um vetor de atributos;
- **Espaço multidimensional, espaço original, modelo de espaço vetorial, espaço de características:** é o espaço que descreve um conjunto de dados por meio de vetores de atributos;
- **Espaço projetado, espaço reduzido:** é o espaço resultante da aplicação de uma técnica de redução de dimensionalidade. Por exemplo, a aplicação de uma técnica de projeção sobre um espaço multidimensional resulta em um espaço de menor dimensão, que é o espaço projetado;
- **Espaço de visualização, espaço visual:** é o espaço utilizado por uma técnica de visualização para representar as informações das instâncias de um conjunto de dados. Por exemplo, um espaço projetado, quando a redução do espaço multidimensional é realizada para dimensões 2 ou 3;
- **Espaço do objeto, volume:** é o domínio de onde o conjunto de dados foi adquirido. No caso de uma simulação, a representação no domínio em que os dados foram criados;
- **Marcação visual:** é o elemento gráfico, ou forma (e.g., um círculo), utilizado para visualmente representar uma instância de um conjunto de dados no espaço de visualização;
- **Atributo visual:** é a propriedade gráfica (e.g., cor, tamanho) utilizada para representar algum atributo do conjunto de dados;
- **Atributos informados, atributos dados, metadados:** são informações específicas de uma instância de dados. Por exemplo, para um artigo científico, autor, referências e data de publicação. Em geral, eles fazem parte do conjunto de dados ou são complementares;

- **Atributos extraídos:** são atributos que são obtidos do conteúdo existente de uma instância por meio de algum pré-processamento. Por exemplo, para documentos, termos extraídos do conteúdo textual são atributos extraídos; para imagens, cor e intensidade são atributos válidos para pixel;

- **Dados estruturados:** são dados que estão organizados de alguma maneira que facilite a identificação de seus atributos e também a sua recuperação. Por exemplo, uma tabela com dados sobre o índice de desenvolvimento humano de cidades;

- **Dados não estruturados:** são dados que não possuem uma estrutura definida. Eles estão disponíveis, mas não há uma organização que facilite a sua identificação e recuperação sem que seja realizado um pré-processamento. Por exemplo, o conteúdo de documentos;

- **Dados de natureza distinta, dados multimodais:** são dados que descrevem uma mesma instância, porém são de tipo (natureza) diferente. Por exemplo, uma imagem pode ser descrita pelo seu próprio conteúdo (pixels) e também por uma descrição textual do que há na imagem. Neste caso, há dois tipos de dado: texto e imagem;

- **Técnica de coordenação:** é a técnica utilizada para gerar o mapeamento entre as instâncias das visões coordenadas. Dada a seleção de instâncias em uma visão, uma técnica de coordenação identifica quais são as instâncias que serão mapeadas nas outras visões. Este mapeamento é realizado por algum critério ou função, que é o que diferencia uma técnica de outra. O mapeamento gerado por uma técnica de coordenação é ecoado para as outras visões por meio de algum tipo de coordenação;

- **Tipo de coordenação:** é a técnica utilizada para ecoar as instâncias mapeadas por uma técnica de coordenação. Por exemplo, a clássica *brushing and linking* destaca em outra visão as instâncias mapeadas por uma técnica de coordenação. Um outro exemplo seria a aplicação de uma operação de *zoom* nas instâncias mapeadas.

1.4 Organização

O restante desta tese está organizado da seguinte maneira:

- No Capítulo 2 são apresentadas técnicas para a criação de mapas de similaridade. O foco é dado às técnicas que utilizam projeção multidimensional e posicionamento de pontos no plano. No capítulo são apresentadas aplicações para exploração de coleções de documentos e coleções de imagens;

- No Capítulo 3 é apresentada a abordagem de múltiplas visões coordenadas, destacando as aplicações que são voltadas para a exploração de dados textuais. Também são apresentados

os principais modelos de coordenação para a implementação de múltiplas visões coordenadas, dos quais um foi escolhido para ser adaptado aos objetivos deste trabalho;

- No Capítulo 4 é apresentado o modelo de coordenação adaptado para o contexto deste trabalho. Os componentes do modelo e os processos também terão seus nomes mantidos em inglês, pois aproximam-se mais do que está implementado na ferramenta e descrito em artigos;

- No Capítulo 5 são apresentadas as novas técnicas de coordenação desenvolvidas neste trabalho. O nome das técnicas serão mantidos em inglês, pois assim estarão mais próximos do que está implementado na ferramenta e descrito em artigos;

- No Capítulo 6 são apresentadas aplicações das técnicas de coordenação, voltadas para a exploração de coleções de documentos, de coleções de imagens e de dados volumétricos;

- No Capítulo 7 são apresentadas as conclusões do trabalho desenvolvido e também possíveis trabalhos futuros;

- O Apêndice A apresenta uma lista com os trabalhos publicados e com aqueles em processo de submissão.

CAPÍTULO

2

Mapas de Similaridade

2.1 Considerações Iniciais

Mapas de similaridade são representações visuais de conjuntos de dados que tem o objetivo de posicionar as instâncias de forma a preservar no espaço projetado noções de similaridade do espaço original. Para a construção desse tipo de representação, os atributos utilizados podem ser informados ou extraídos. Atributos informados são metadados que descrevem as instâncias. Atributos extraídos revelam o conteúdo das instâncias. Neste caso, geralmente é necessário construir um modelo de espaço vetorial onde as instâncias são representadas por esses atributos (Skupin, 2002; Andrews et al., 2002; Wise, 1999; Paulovich et al., 2007). A partir desses atributos define-se uma similaridade, que é a fonte principal de informação para uma grande parte das técnicas de mineração e de visualização de dados. No caso de documentos, esse espaço vetorial é construído a partir do conteúdo textual dos documentos, isto é, seus termos, como proposto por (Salton et al., 1975; Salton e Buckley, 1987). No caso de imagens, esse espaço vetorial é construído a partir de técnicas de extração de características [1] (da Costa e Junior, 2000). Alternativamente, tópicos são extraídos para estabelecer a similaridade entre os documentos ou entre os próprios tópicos (Perez e de Antonio, 2004; Spangler et al., 2002; Havre et al., 2002). No caso em que os atributos informados são utilizados, é possível construir uma rede de relacionamento entre as instâncias do conjunto de dados (Chen, 2006; Mane e Börner, 2004; Chen e Hsieh, 2007). Por exemplo, no caso de documentos, é comum a construção de redes de co-citação e co-autoria para representar a relação entre as instâncias (documentos).

[1] O espaço vetorial em aplicações que envolvem imagens é também conhecido como espaço de características.

A seguir são apresentadas abordagens para a construção de mapas de similaridade, com foco em técnicas de projeção multidimensional e em técnicas de posicionamento de pontos, pois são as abordagens utilizadas neste trabalho. Em seguida são apresentados exemplos de aplicação de mapas de similaridade para a exploração de coleções de documentos e coleções de imagens. Por fim, é apresentada uma conclusão sobre o conteúdo deste capítulo.

2.2 Métodos para Construção de Mapas de Similaridade

Como descrito anteriormente, a construção de mapas de similaridade pode ser realizada a partir de atributos informados ou de atributos extraídos. No primeiro caso, uma abordagem comum é a construção de redes de relacionamento entre as instâncias. Exemplos de aplicações dessa abordagem são apresentados no início da próxima seção. No segundo caso, em que o conteúdo das instâncias é a fonte primária de dados, é comum a construção de um modelo de espaço vetorial. Um exemplo de técnica que utiliza o modelo de espaço vetorial é de Mapas Auto-organizáveis (*Self-Organizing Maps – SOMs*) (Kohonen, 1990), ilustrado na Figura 2.1. SOMs utilizam algoritmos de redes neurais para gerar mapas visuais auto-organizáveis em uma malha bidimensional, os quais tendem a posicionar instâncias similares em uma mesma vizinhança. Nesse exemplo, o resultado visual é interpretado observando os grupos como se fossem fronteiras entre países desenhadas em mapas cartográficos. Também deve ser observada a superfície do mapa e relacionar as cores como se fossem alturas de um relevo. No caso, a informação mapeada para cor é o grau de dominância (*Dominance of Highest Components*) dos três atributos que mais influenciaram a rede neural para o posicionamento das instâncias.

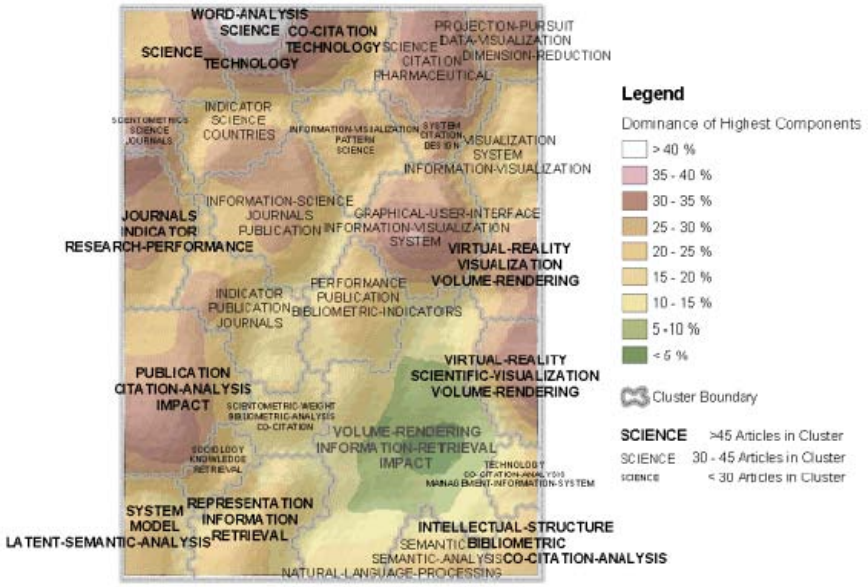


Figura 2.1: Parte de um Mapa Auto-organizável construído a partir de 2764 artigos. Extraído de (Börner et al., 2003)

Uma outra maneira recorrente de criar a representação visual no plano de visualização a partir de um modelo de espaço vetorial é por meio de Técnicas de Projeção e Técnicas de Posicionamento de Pontos. Os passos para a execução desse grupo de técnica são ilustrados na Figura 2.2. Primeiramente, o modelo de espaço vetorial é construído a partir de uma etapa de pré-processamento que extrai atributos do conteúdo das instâncias do conjunto de dados. Em seguida, são computadas as distâncias entre elas para gerar uma matriz de distância, sobre a qual é aplicada uma técnica de projeção para a construção do mapa de similaridade. Esse grupo de técnicas é utilizado neste trabalho. Alternativamente, alguns tipos de dados admitem cálculo de similaridade diretamente dos dados, sem necessidade de extrair atributos.

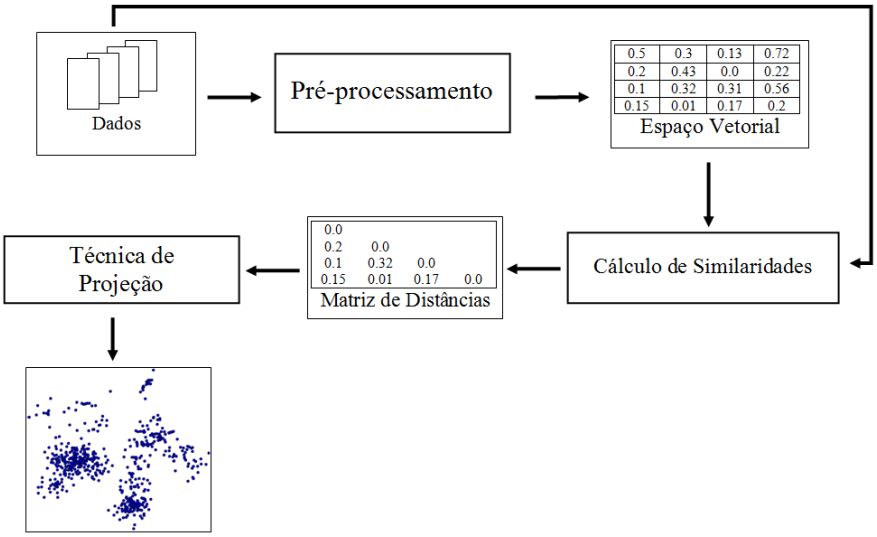


Figura 2.2: Passos para criar um mapa de similaridade por meio de uma técnica de projeção.

Técnicas de projeção podem ser divididas em dois grandes grupos, de acordo com a função de projeção empregada: técnicas de projeção lineares e técnicas de projeção não-lineares. As técnicas de projeção lineares criam combinações lineares dos atributos dos dados, definindo-os em uma nova base ortogonal com uma dimensão menor do que a original. Uma técnica muito conhecida desse grupo é a Análise de Componentes Principais (*Principal Component Analysis – PCA*) (Jolliffe, 1986), a qual cria uma matriz de covariância dos dados e a decompõe em m autovetores e m autovalores (m é a dimensão do espaço original). Os p primeiros autovetores associados aos maiores autovalores são selecionados para transformar o espaço m-dimensional em um espaço p-dimensional (p é a dimensão do espaço projetado). Outras técnicas lineares mais comuns são: Decomposição de Valores Singulares (*Singular Value Decomposition – SVD*) (Demmel, 1997) e *Local Linear Embedding* (LLE) (Roweis e Saul, 2000).

As técnicas de projeção não-lineares, em geral, visam minimizar uma função de perda de informação. Geralmente essa função se baseia na similaridade entre os objetos m-dimensionais e nas distâncias entre os pontos p-dimensionais. Em muitos casos, a aplicação de técnicas não-lineares não necessita que os dados originais tenham uma representação vetorial, sendo suficiente a definição de similaridades entre as instâncias e as distâncias entre os pontos no espaço p-dimensional.

Duas outras características importantes são a sua iteratividade, a qual permite que o usuário pare o processo no meio de sua execução, e a facilidade em adicionar novos elementos no espaço original e posicioná-los no espaço projetado. Neste grupo, encontram-se técnicas categorizadas como análise de agrupamentos (*clustering*), posicionamento baseado em força e projeções multidimensionais. As técnicas de projeção utilizadas neste trabalho pertencem a esse grupo de técnicas, por isso um maior destaque é dado às técnicas não-lineares.

No grupo das projeções não-lineares destacam-se as técnicas de Escalonamento Multidimensional (*Multidimensional Scaling – MDS*). Técnicas MDS exploram a similaridade ou dissimilaridade entre as instâncias dos dados para associar uma posição a cada instância em um espaço m dimensional. Para os casos em que o valor de m é pequeno, as posições resultantes podem ser exibidas no plano de visualização (Cox e Cox, 2000). Técnicas comuns de MDS são aquelas que utilizam posicionamento de pontos baseado em forças (*Force-Directed Placement* (FDP)) ou uma função que minimize a perda de informação que ocorre no posicionamento das instâncias no espaço m dimensional. Outras técnicas de projeção não-lineares que se destacam são a *Nearest-Neighbor Projection* (NNP) (Tejada et al., 2003) e a *Fastmap* (Faloutsos e Lin, 1995).

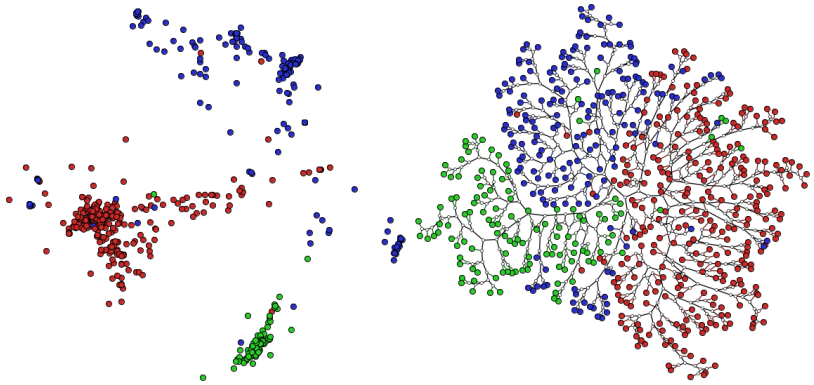
Uma técnica MDS muito conhecida é a *Sammon Mapping* (Sammon, 1964). O objetivo dessa técnica é minimizar o erro na tentativa de posicionar os pontos em uma curva ou em uma reta. Uma função indica a quantidade de informação perdida devido à projeção e então aplica-se uma otimização não-linear baseada no gradiente da função para encontrar um mínimo local e minimizar a função de perda (Pekalska et al., 1999).

As técnicas MDS mais simples são as que utilizam um modelo baseado em molas, frequentemente chamadas de posicionamento baseado em forças (Fruchterman e Reingold, 1991). Esse modelo foi originalmente proposto por Eades (Eades, 1984) como uma heurística para o desenho de grafos com o objetivo de estabilizar a força exercida pelas molas nos objetos conectados a elas. Dentre as abordagens que utilizam posicionamento baseado em forças (Fruchterman e Reingold, 1991; Chalmers, 1996; Frick et al., 1995), uma técnica denominada *Force Scheme* foi proposta por Tejada e outros (Tejada et al., 2003). Essa abordagem, também baseada no conceito de forças de atração e repulsão entre objetos, não trabalha empregando equações diferenciais, mas simplifica esse processo utilizando deslocamentos em determinadas direções (Tejada et al., 2003 apud Paulovich, 2008, p.24). A direção de deslocamento relaciona a distância dos pontos no espaço projetado com a sua similaridade no espaço original, criando uma perturbação em todos os pontos.

Geralmente, as técnicas baseadas em força são lentas, pois $n - 1$ operações são realizadas a cada iteração para recalcular o posicionamento dos pontos (n é o número de pontos). Com o objetivo de diminuir essa complexidade, Paulovich e Minghim propuseram uma técnica denominada *Projection by Clustering* (ProjClus) (Paulovich e Minghim, 2006). Nessa técnica de projeção, os dados são primeiramente agrupados por um algoritmo rápido de agrupamento e os centróides dos grupos são projetados (utilizando *NNP* ou *Fastmap*). As instâncias do conjunto são projetadas ao redor de seus respectivos centróides e por fim um algoritmo de força é aplicado localmente em cada grupo; com isso a complexidade cai de O(n^2) para O($n^{\frac{3}{2}}$).

Outra técnica de projeção rápida também foi proposta por Paulovich e outros (Paulovich et al., 2006, 2008), a qual foi denominada *Least-Square Projection* (LSP). A LSP foi generalizada de uma abordagem para reconstrução de malhas para mapeamento de dados no plano a partir de alguns pontos já conhecidos. Na LSP, um subconjunto de pontos no espaço original é projetado no plano (utilizando *NNP* ou *Fastmap*), os quais são conhecidos como pontos de controle. Na sequência, fazendo uso da relação de vizinhança dos pontos de controle no espaço original e das coordenadas dos pontos de controle no espaço projetado, é criado um sistema linear cujas soluções são as coordenadas dos pontos que serão projetados. Um exemplo de mapa de similaridade gerado por meio da técnica LSP é apresentado na Figura 2.3(a). No exemplo foi utilizado um conjunto de artigos científicos de três diferentes áreas, as quais são representadas pela informação de cor.

Uma outra abordagem para posicionamento de pontos que difere das abordagens convencionais é conhecida como *Neighbor-Joining* (NJ) (Cuadros et al., 2007). Essa técnica efetua o posicionamento de pontos baseado em árvore por meio de um algoritmo para construção de árvores filogenéticas. O algoritmo agrupa em um nó ancestral pares de pontos com características mais similares. Esse processo é realizado em todo o conjunto de dados, criando ramificações para formar a árvore. A técnica produz um bom resultado visual, mostrando com clareza os grupos formados pelas ramificações da árvore. No entanto é uma técnica de alto custo computacional. Um exemplo de mapa de similaridade construído por meio da técnica NJ é apresentado na Figura 2.3(b). Esse mapa foi construído com o mesmo conjunto utilizado para gerar o mapa da Figura 2.3(a).



(a) Mapa de Similaridade gerado por Projeção Multidimensional.

(b) Mapa de Similaridade gerado por Posicionamento de Pontos baseado em Árvores.

Figura 2.3: Mapas de similaridade gerados por uma técnica de projeção multidimensional (a), no caso, a técnica *Least Squares Projection*(Paulovich et al., 2008); e por uma técnica de posicionamento de pontos baseado em árvore (b), no caso, a técnica *Neighbour-Joining*(Cuadros et al., 2007).

Para maiores detalhes sobre técnicas de projeção e técnicas de posicionamento de pontos, é recomendada a leitura da tese “Mapeamento de dados multidimensionais – integrando mineração e visualização” (Paulovich, 2008).

Nas próximas seções são apresentados exemplos da utilização de técnicas para a construção de mapas de similaridade a partir de coleções de documentos e de coleções de imagens. Além de abordagens que utilizam técnicas de projeção e técnicas de posicionamento de pontos, serão apresentadas outras técnicas para a exploração da similaridade entre elementos. A descrição das abordagens é focada nas técnicas de visualização empregadas na representação do mapa de similaridade, a forma como foi definida a similaridade entre os elementos e as técnicas utilizadas para posicionar os pontos no plano de visualização.

2.2.1 Mapas de Similaridade de Documentos

A visualização auxilia o usuário a organizar mentalmente e acessar eletronicamente uma grande quantidade de informações. A seguir são apresentadas algumas técnicas para visualização e exploração de coleções de documentos baseada em atributos informados e atributos extraídos.

Um sistema que utiliza um ambiente virtual multi-usuário 3D para exploração de conjuntos de documentos é o *Starwalker* (Chen et al., 1999). Esse sistema permite a colaboração em um espaço semântico, o qual é construído utilizando ferramentas do arcabouço *Generalized Similarity Analysis* (GSA). O GSA utiliza informações do acerca dos documentos de uma coleção e constrói redes associativas a partir delas. Para a geração das redes associativas é aplicada a técnica *Pathfinder Network Scaling* (PFnets) (Schvaneveldt, 1990), que exibe nós conectados em uma rede de acordo com a proximidade e similaridade entre pares. A Figura 2.4 mostra uma rede de co-citação de autores construída pelo *Starwalker*. Esse mesmo mapa é apresentado na Figura 2.5, mas uma barra é utilizada como atributo visual para representar o número de citações de cada autor. A altura da barra indica o número de citações e a cor indica a data da citação.

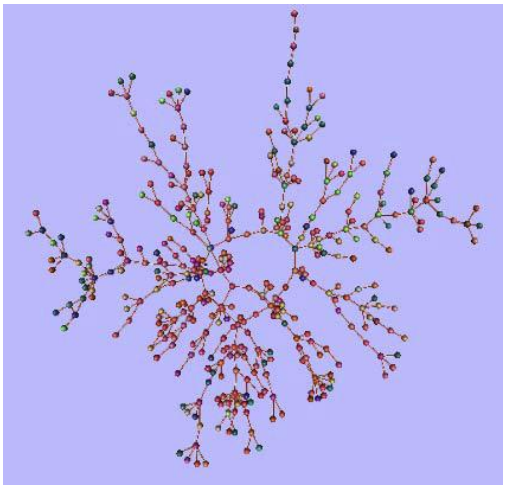


Figura 2.4: Rede de co-citação com 380 autores com 9 ou mais citações. Extraído de (Börner et al., 2003)

Chen também criou um método chamado *Progressive Knowledge Domain Visualization* (*Progressive KDViz*) (Chen, 2004) com o intuito de simplificar a identificação de pontos de variação

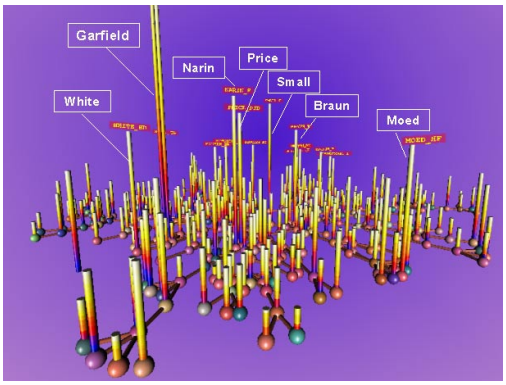


Figura 2.5: Rede de co-citação apresentada na Figura 2.4. Uma barra é utilizada para representar o número de citações de cada autor. A altura da barra indica o número de citações e a cor indica a data da citação. Extraído de (Börner et al., 2003)

(*turning points*) no domínio do conhecimento em redes de co-citação sobre o tempo. Esse método foi implementado no sistema chamado *CiteSpace II* (Chen, 2006), o qual divide todo o conjunto de dados em janelas de tempo que podem ser unitárias ou em intervalos de tempo. Cada janela contém uma rede de co-citação que se unirá às redes de co-citação das janelas vizinhas. Após a união das redes de co-citação, três tipos especiais de nós são utilizados para facilitar a identificação de artigos relevantes na literatura, salientando a visualização de algumas características dos nós. Os tipos de nós são: o nó de referência (*landmark*), que representa um artigo altamente citado; o nó *hub*, que representa um artigo com alta co-citação; e o nó pivô (*pivot*), que representa a junção entre duas redes de co-citação (ver Figura 2.6).

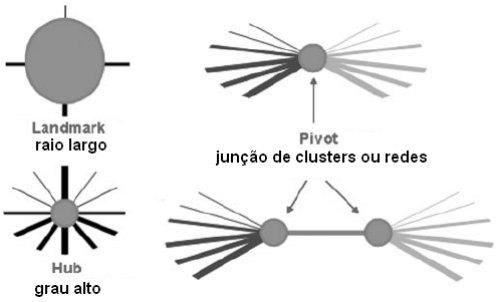


Figura 2.6: Três tipos de nós especiais utilizados na rede de co-citação (Extraído de (Chen, 2004)).

Bons candidatos para um ponto de mudança (*turning point*) são os nós que apresentam um rápido crescimento nas citações e também nós que se conectam a muitos *links* com diferentes cores (nó pivô). Provavelmente um nó pivô é um artigo único na literatura, pois ele conecta artigos de diferentes grupos. A Figura 2.7 mostra a visualização apresentada por esta técnica partir de um conjunto de artigos da física teórica.

Um outro domínio de conhecimento foi analisado por Chen e outros (Chen et al., 2005) com objetivo de compreender a evolução do campo farmacológico por meio da aplicação das técnicas do sistema *CiteSpace II*. Os dados utilizados são de artigos publicados em um campo de pesquisa

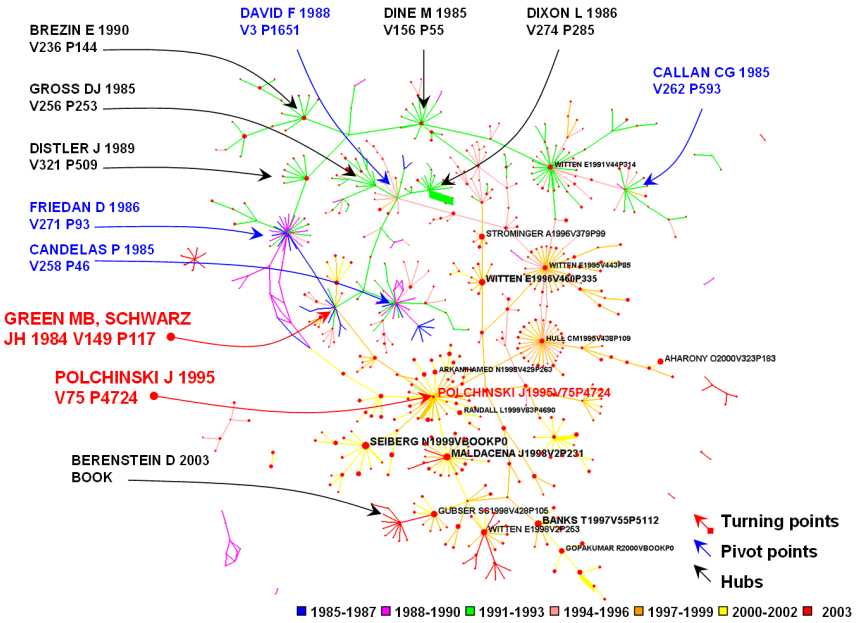


Figura 2.7: União de 624 nós de diferentes janelas de tempo (Extraído de (Chen, 2004)).

farmacológico sobre drogas anti-inflamatórias não-esteroidais (*Non-steroidal Anti-inflammatory Drug - NSAID*).

A Figura 2.8 mostra a visualização das redes de co-citação dos artigos sobre NSAID. As cores indicam as datas dos agrupamentos e as cores dos anéis indicam a idade das citações, isto é, o ano em que ocorreu a citação do artigo. A palavra com maior predominância é exibida após a seleção de um grupo, revelando a sua natureza. Uma outra visualização mostra as redes de co-citação organizadas em janelas de tempo, a qual também exibe as palavras que representam um *burst* (súbito aumento no seu uso) (ver Figura 2.9).

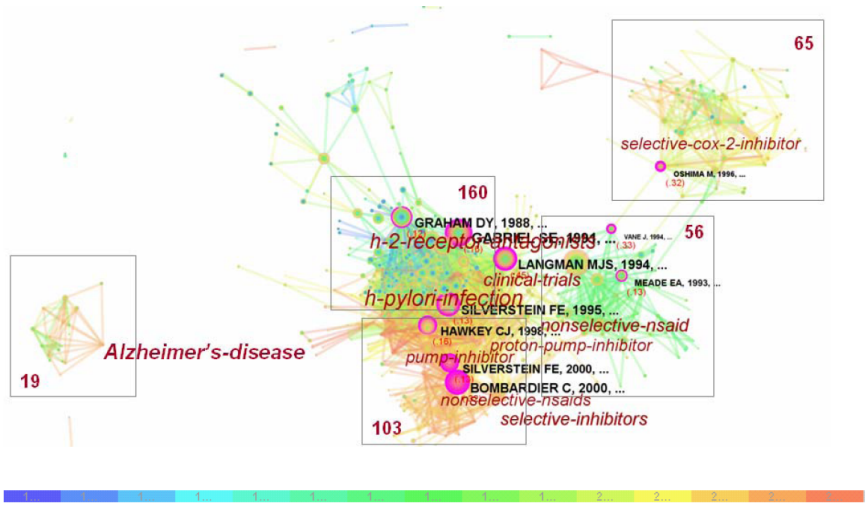


Figura 2.8: Visualização de uma rede de co-citação de artigos sobre NSAID entre 1990 a 2004 (Extraído de (Chen et al., 2005)).

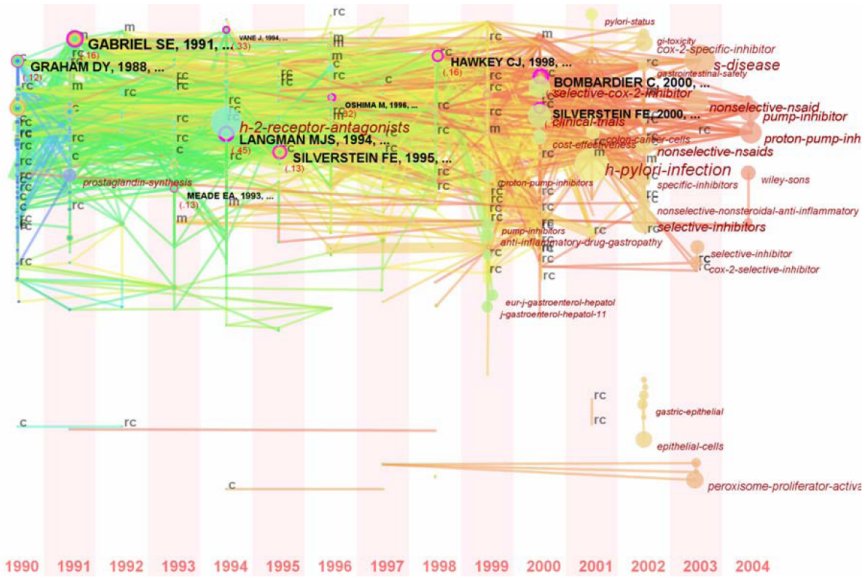


Figura 2.9: Visão das janelas de tempo da evolução de artigos sobre NSAID (Extraído de (Chen et al., 2005)).

Murray e outros (Murray et al., 2006) apresentaram uma abordagem para definir, analisar e mapear áreas científicas baseadas em redes de co-autoria. Segundo os autores, essa abordagem também auxilia a comparação e descoberta das habilidades de cada autor. A base de dados analisada contém as referências das publicações cedidas por vários autores (*bibtex* ou *EndNote*). A Figura 2.10 mostra uma rede de co-autoria construída pelo sistema *Geomi* (Ahmed et al., 2005), que é uma ferramenta para visualização e análise de redes grandes e complexas.

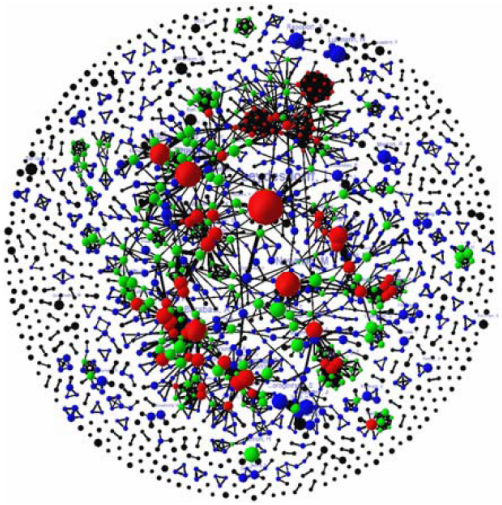


Figura 2.10: Rede de co-autoria (Extraído de (Murray et al., 2006)).

O mapa mostrado na Figura 2.10 representa todos os autores que escreveram pelo menos dois artigos, e suas respectivas co-autorias. O tamanho do nó representa o número de artigos escritos pelo autor e a cor descreve o grau do nó, isto é, vermelho representa grau maior ou igual a 10, verde representa grau entre 5 e 9, azul representa grau entre 2 e 4 e preto representa grau 1 ou 0. A

espessura da aresta é baseada no número de co-autorias. Os nós de autores com mais de 10 artigos são rotulados com o nome do autor.

A Figura 2.11 mostra a rede completa de co-autoria cuja componente conexa do grafo possui mais do que três nós, construída pelo sistema *Pajek* (de Nooy et al., 2005), que é um sistema para análise e visualização de grafos. Nessa visualização, cada componente conexa é representada por uma cor diferente e, como é possível notar, a maior está representada pela cor roxa.

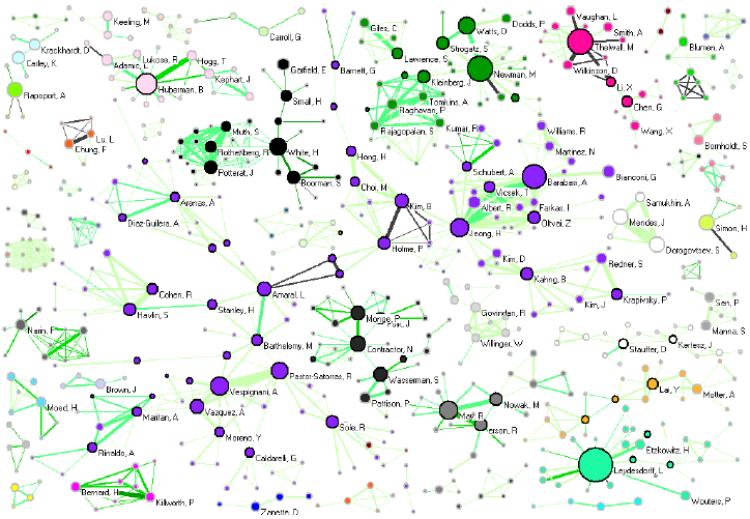


Figura 2.11: Rede completa de co-autoria com componentes conexas com tamanho maior do que três (Extraído de (Murray et al., 2006)).

Além de bases de artigos, esse tipo de representação visual pode ser aplicada a outros conjuntos de dados que possuam informações em comum. Giereth e outros (Giereth et al., 2007) utilizaram esse tipo de representação para visualizar conjuntos de patentes. No caso das patentes, várias informações podem ser compartilhadas, tais como: aplicação, inventor, classificação internacional (*International Patent Classification* (IPC)), *priority number*, dentre outras. Na Figura 2.12 é apresentado o relacionamento das patentes com seu *priority number*, possibilitando identificar famílias de patentes. Na Figura 2.13 as famílias de patentes são exploradas isoladamente, mostrando a data de sua criação (eixo X) e também o país de origem (eixo Y).

Além da análise de conjuntos de documentos com base em elementos comuns, a literatura descreve diferentes abordagens para exploração do conteúdo dos documentos. Uma delas é baseada nos tópicos e palavras freqüentes do conjunto de documentos. Mane e Börner (Mane e Börner, 2004) apresentam uma maneira de gerar mapas de associação de co-palavras (*co-word*) dos principais tópicos de um conjunto de textos. A geração é baseada na alta freqüência de palavras e palavras que tiveram um súbito aumento no seu uso (*burst*). A análise de co-palavras é uma técnica de análise de conteúdo utilizada para identificar uma forte associação entre palavras que co-ocorrem em um mesmo documento. Embora co-palavras sejam tipicamente geradas somente com base na mais alta freqüência das palavras, o trabalho desenvolvido também levou em conta a quantidade de *bursts* das palavras.

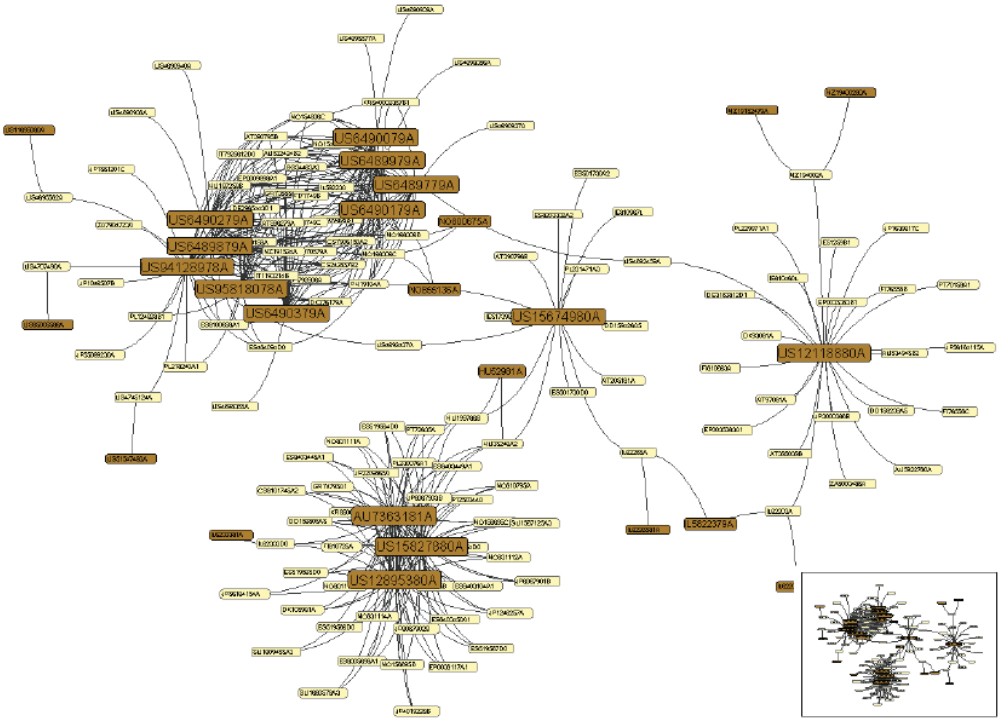


Figura 2.12: Agrupamentos por família de patentes (Extraído de (Giereth et al., 2007)).

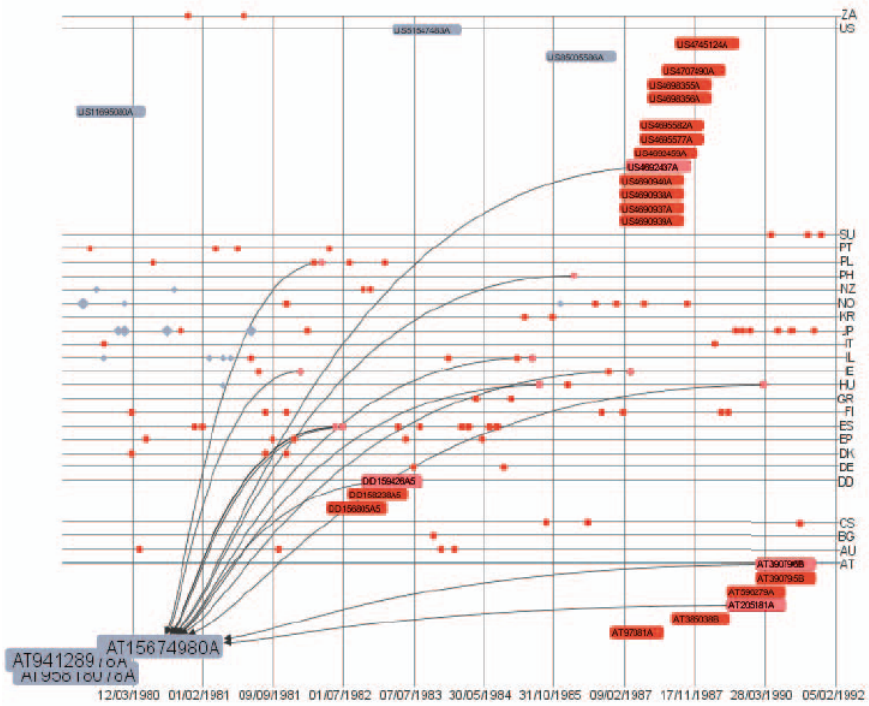


Figura 2.13: Exploração de família de patentes, exibindo data (eixo X) e país (eixo Y) (Extraído de (Giereth et al., 2007)).

A Figura 2.14 mostra as 50 primeiras palavras da intersecção dos conjuntos de mais alta freqüências e do mais alto *bursts*. Nesse exemplo foram tomadas as 742 palavras mais freqüentes e os 874 maiores *bursts* das palavras para obter a intersecção, da qual foram retiradas as primeiras

50 palavras. O tamanho do círculo corresponde ao nível máximo de *busrt* alcançado pela palavra. A cor é usada para denotar os anos em que a palavra foi usada com mais freqüência e também o ano do *burst* máximo. O ano que a palavra foi usada com freqüência máxima é representado pelo círculo da borda e o ano inicial do primeiro *burst* da palavra é representado pela cor do círculo interno. A espessura é proporcional à ocorrência da palavra. A base utilizada era formada por 20 anos de materiais dos *Procedings of the National Academy of Sciences* (PNAS).

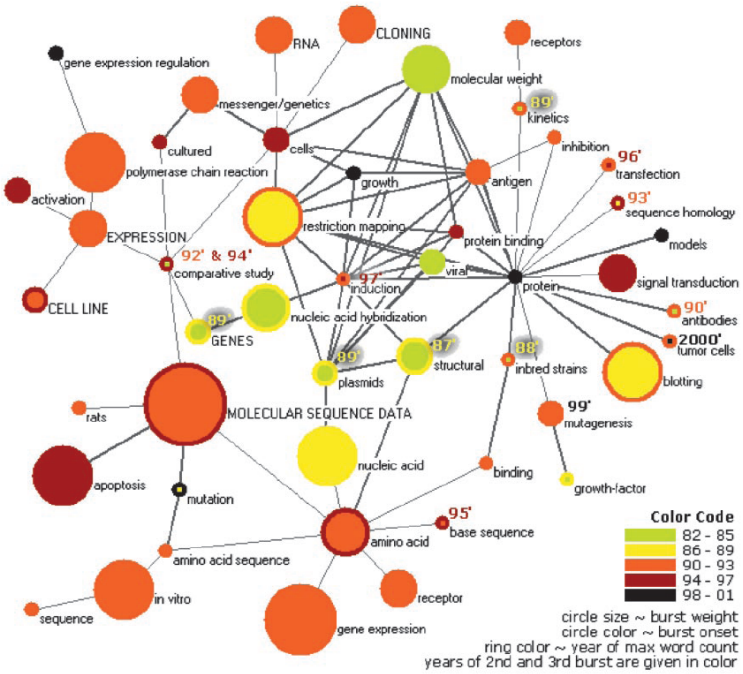


Figura 2.14: Visualização das 50 co-palavras com maiores freqüência e *burst* nas publicações dos PNAS entre 1982 e 2001 (Extraído de (Mane e Börner, 2004)).

Perez e de Antonio (Perez e de Antonio, 2004) propuseram um método para representar tópicos como eixos no espaço tridimensional, o qual foi denominado *3D Explorer*. Nesse método cada eixo agrupa um conjunto de documentos, os quais são representados por um tópico extraído do agrupamento. Inicialmente, é construído um modelo de espaço vetorial e um algoritmo de agrupamento (*clustering*) é aplicado para separar os documentos em um número determinado de agrupamentos, sendo que esse número é dependente do número de eixos escolhidos pelo usuário. Na seqüência, é utilizada uma técnica de redução de dimensionalidade (SVD (Demmel, 1997)) para fazer a projeção dos documentos. A Figura 2.15 mostra a visualização gerada pelo *3D Explorer* utilizando seis eixos.

Utilizando a exploração de um conjunto de documentos com base em seus tópicos, Eck e outros (van Eck et al., 2006) propuseram uma técnica de visualização denominada *Concept Density Maps*. Nessa técnica são extraídos tópicos do conjunto de documentos e é formada uma matriz $n \times n$, em que n é o número de tópicos. Os elementos dessa matriz são os valores de co-ocorrência dos tópicos no conjunto de documentos. Assim, um espaço n-dimensional é formado pela matriz de co-ocorrência de tópicos, a qual é projetada no espaço bidimensional por um algoritmo MDS

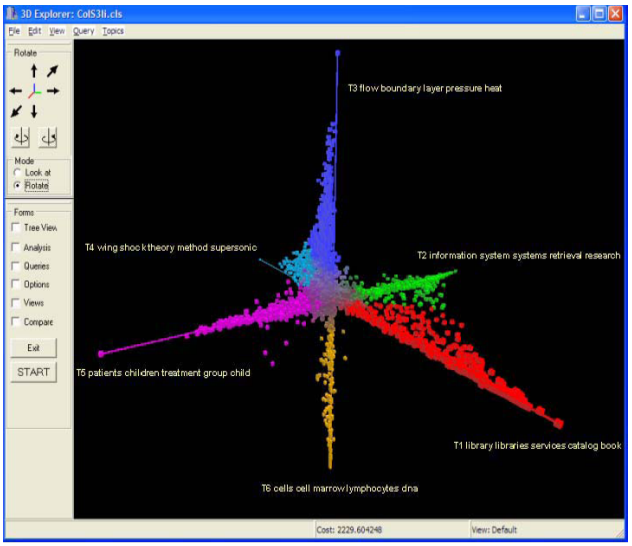


Figura 2.15: Visualização de seis agrupamentos (tópicos) com o *3D Explorer* (Extraído de (Perez e de Antonio, 2004)).

também proposto por Eck e outros (van Eck et al., 2005). No caso, o algoritmo de projeção procura agrupar os tópicos que possuem maior correlação. A visualização desse espaço bidimensional é mostrada na Figura 2.16, utilizando cores para indicar a correlação entre os tópicos, variando do vermelho (alta correlação) até o azul (baixa correlação).

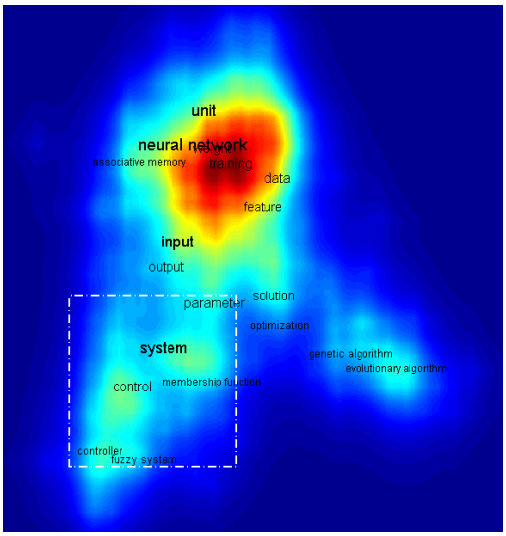


Figura 2.16: *Concept Density Maps* mostrando a correlação entre tópicos (Extraído de (van Eck et al., 2006)).

Outro sistema que utiliza o conteúdo dos documentos para explorar um conjunto de dados é o *IN-SPIRE* (ver `http://in-spire.pnl.gov/`). A diferença dos sistemas citados anteriormente é que o *IN-SPIRE* baseia-se em todo o conteúdo dos documentos para calcular o posicionamento espacial dos documentos no plano. Para isso, ele utiliza abordagens de redução de dimensionalidade, tais como SVD ou *Latent Semantic Indexing* (Drineas et al., 2004; Papadimitriou et al., 1998). A representação visual é dada por duas técnicas de visualização; a primeira

é baseada em uma metáfora de galáxias (*Galaxies*) (Wise et al., 1995; Wise, 1999), representando os documentos como se fossem estrelas (Figura 2.17). Nessa representação, documentos que possuem relação de conteúdo são agrupados próximos, enquanto os que possuem baixo grau de relação são posicionados a distâncias maiores. A outra forma de representar os documentos é a visualização *ThemeView*, a qual utiliza a representação das galáxias como montanhas suavizadas para indicar os temas dominantes no conjunto de documentos (Figura 2.18). Na *ThemeView* tópicos similares aparecem próximos, enquanto que os dissimilares são separados. A altura dos picos indica a ocorrência do tópico em um conjunto de documentos.

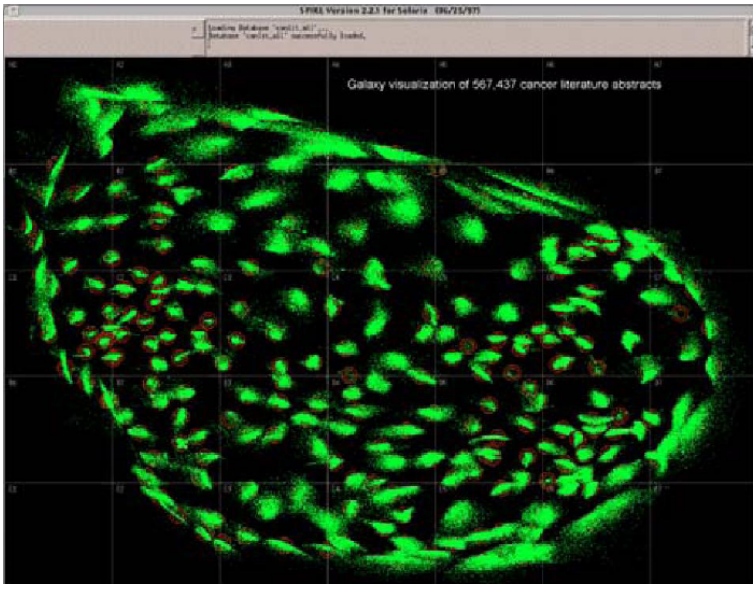


Figura 2.17: Visualição utilizando Galáxias (Extraído de (Eick e Cousins, 2006)).

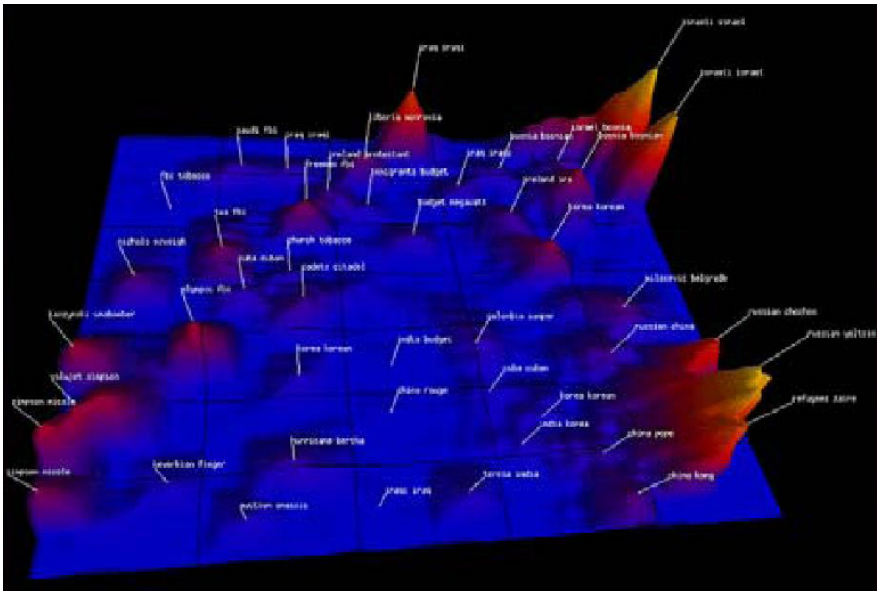


Figura 2.18: Visualição utilizando *ThemeView* (Extraído de (Eick e Cousins, 2006)).

Boyack e outros também propuseram um sistema, denominado VxInsight (Boyack et al., 2002), que se baseia no conteúdo dos documentos para construir mapas. Um algoritmo de posicionamento baseado em força (*VxOrd* (Davidson et al., 2001)) é utilizado para construir a representação bidimensional do conjunto de documentos com base na similaridade do seu conteúdo. Em seguida, é construída uma visualização com a aparência de montanhas, na qual a altura das montanhas está relacionada com a quantidade de elementos abaixo dela, como ilustrado na Figura 2.19, de forma muito similar ao *ThemeView*. Para auxiliar a exploração, rótulos são gerados no topo das montanhas mais significativas, revelando o conteúdo dos documentos que pertencem à montanha rotulada. Os rótulos são baseados na freqüência das palavras mais comuns.

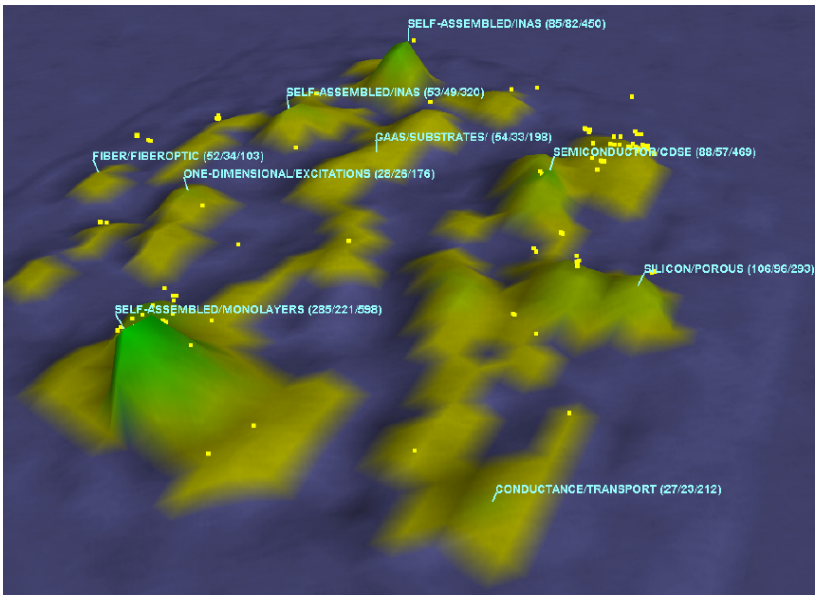


Figura 2.19: Visualização de um conjunto de documentos com o *VxInsight* (Extraído de (Boyack et al., 2002)).

Outro sistema que também organiza o conjunto de documentos com base no seu conteúdo é o *Projection Explorer* (PEX) PEx (Paulovich et al., 2007) (ver `http://infoserver.lcad.icmc.usp.br/infovis2/PEx`), o qual foi utilizado durante este trabalho de doutorado. O PEx constrói um modelo de espaço vetorial com os termos extraídos do conjunto de documentos, formando um espaço de dimensão elevada. Técnicas de projeção multidimensional, posicionamento baseado em força e posicionamento baseado em árvore são empregadas para fazer o mapeamento do espaço m-dimensional para o espaço bidimensional. O posicionamento dos pontos no plano de projeção é baseado na similaridade do conteúdo dos documentos. A representação visual é baseada em grafos, na qual os nós representam os documentos e as arestas representam algum relacionamento entre os nós (por exemplo, vizinhos mais próximos no espaço original ou no espaço 2D). Também são utilizados atributos visuais como tamanho e cor para indicar algum valor escalar associado ao documento (por exemplo, classe ou freqüência de um tópico). A Figura 2.20 mostra um mapa de documentos gerado pelo PEx, no qual as cores indicam as classes e as palavras são os termos dos tópicos identificados pela seleção de alguns documentos do mapa.

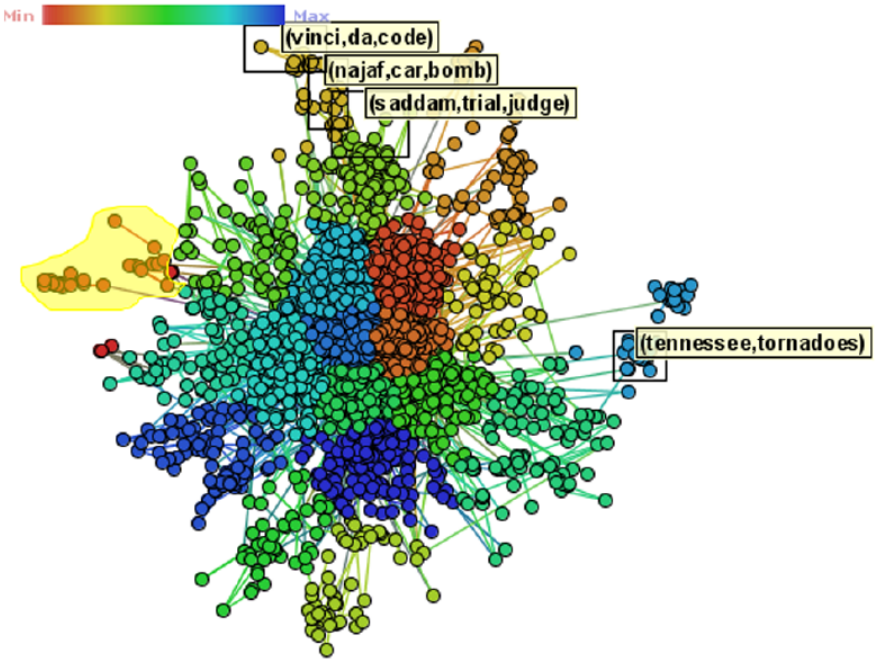


Figura 2.20: Mapa de documentos gerado pela ferramenta *Projection Explorer* (Extraído de (Paulovich et al., 2007)).

Paulovich e Minghim também propuseram uma técnica hierárquica de posicionamento de pontos denominada *Hipp* (*Hierarchical Point Placement*)(Paulovich e Minghim, 2008). A técnica proposta cria uma hierarquia do conjunto de dados e projeta essa hierarquia no plano de visualização. Primeiramente os nós de mais alto nível são projetados e os nós internos, ou seja, de mais baixo nível, são projetados conforme a interação do usuário. A projecão é realizada com a técnica *Least Squares Projection* (LSP) (Paulovich et al., 2008) e em seguida é realizado um espalhamento, proposto para evitar sobreposição dos pontos. Durante a exploração o usuário pode interagir com os nós de alto nível e verificar como os dados estão organizados hierarquicamente. Além disso, o usuário também pode agrupar nós conforme sua escolha. A Figura 2.21 mostra a visualização de uma coleção de 675 documentos científicos de quatro áreas distintas. A cor indica a área de cada documento, o tamanho do círculo indica o número de documentos dentro daquela nível da hierarquia e o texto revela o tópico dos documentos contidos naquele nó.

Pinho e outros propuseram uma abordagem incremental para posicionamento de pontos no plano, denominada Incremental Board (*IncBoard*) (Pinho et al., 2009). Nessa abordagem as instâncias são posicionadas em uma grade regular com base na similaridade entre elas. Para isso, cada instância do conjunto de dados que ainda não foi posicionada na grade regular é colocada na posição da instância mais similar que já está na grade. Em seguida, é verificado se a nova instância ocupará o lugar ou se ela será deslocada com base em um cálculo de erro acumulado. Em seguida, a instância que não ocupará a posição concorrida será deslocada e o processo se repete até que uma instância deslocada ocupe uma posição vazia na grade regular. A Figura 2.22 mostra um exemplo de visualização gerada pela técnica *incBoard* a partir de 300 documentos de quatro áreas distintas. A cor está indicando a área de cada documento. Recentemente essa técnica foi adaptada dando

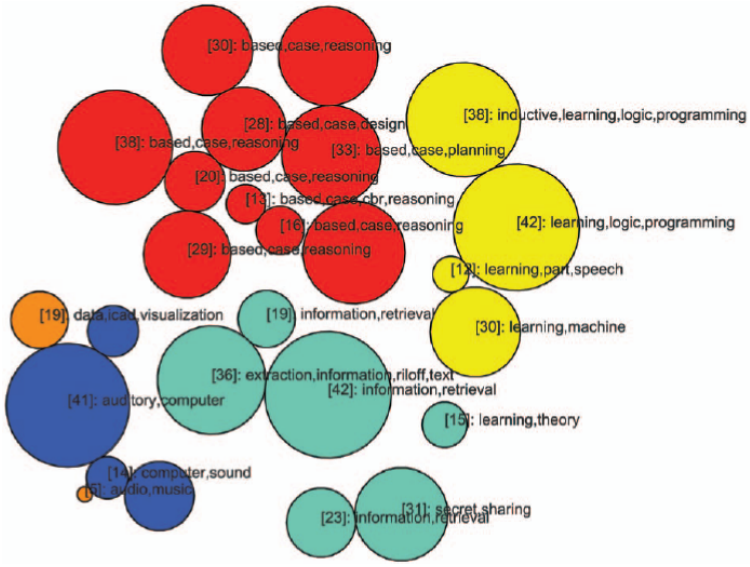


Figura 2.21: Mapa de 675 documentos de quatro áreas distintas gerado pela técnica *Hipp* (Extraído de (Paulovich e Minghim, 2008)).

origem à técnica denominada *HexBoard*, com a qual é possível representar uma melhor relação de vizinhança entre as instâncias de dados por meio de uma grade representada por elementos visuais hexagonais. A Figura 2.23 mostra parte de de uma mapa de 675 artigos científicos de quatro áreas distintas. Nessa visualização a cor também indica a área de cada documento.

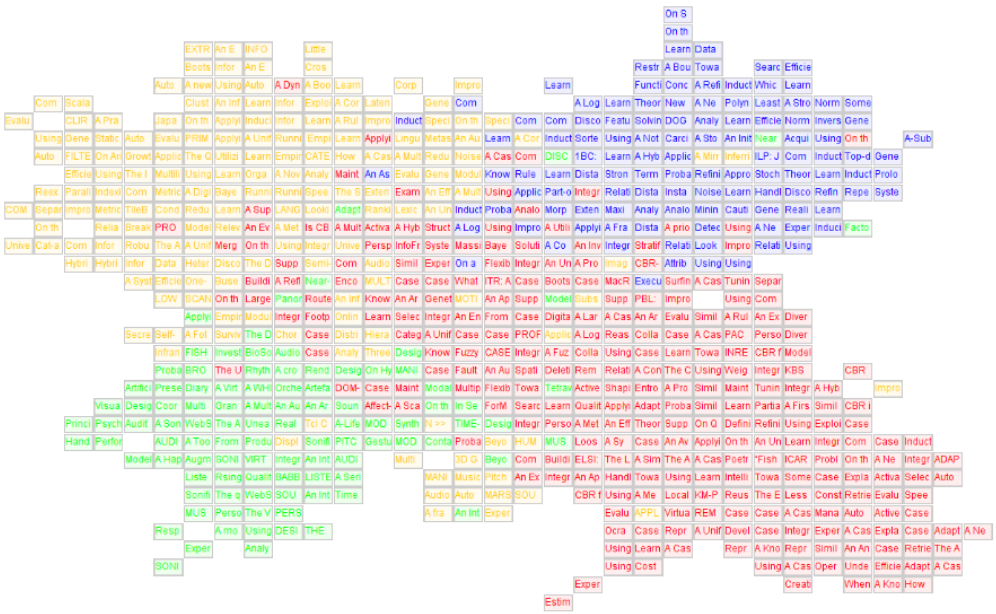


Figura 2.22: Mapa de 300 documentos de quatro áreas distintas gerado pela técnica *incBoard* (Extraído de (Pinho, 2009)).

Dos sistemas apresentados, a PEx se destaca na construção de mapas de similaridade, pois ela disponibiliza a maioria das técnicas de projeção e de posicionamento de pontos existentes na literatura. Além disso, a PEx possui um mecanismo de coordenação *brushing and linking* que

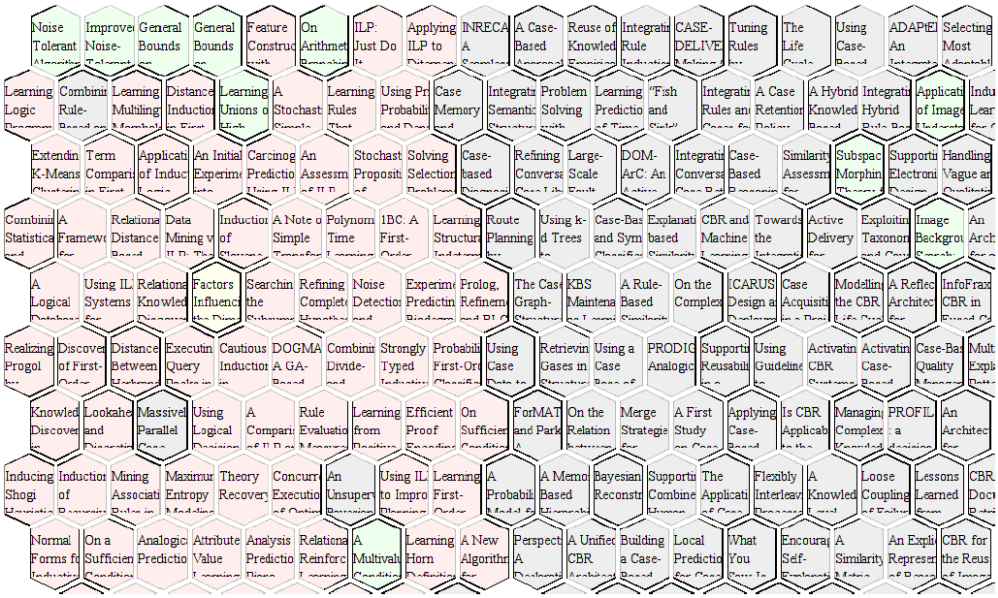


Figura 2.23: Parte de um mapa de 675 documentos de quatro áreas distintas gerado pela técnica *hexBoard* (Extraído de (Pinho e Oliveira, 2009)).

auxilia a exploração das coleções de documentos, diferenciado-se dos outros sistemas que utilizam diferentes visões (representações), mas não utilizam a coordenação entre as visões. No entanto, todos eles estão limitados à exploração de um único conjunto de dados. A exploração de diferentes conjuntos poderia ser realizada por meio de novos técnicas de coordenação, como as propostas neste trabalho.

2.2.2 Mapas de Similaridade de Imagens

Organizar grandes coleções de imagens visando sua efetiva busca e recuperação é um grande desafio das pesquisas de Visão Computacional e de Recuperação de Informação. Novas tendências investigam como técnicas de visualização e de mineração podem auxiliar o usuário na exploração de grandes coleções de imagens. Mapas de similaridade de imagens podem ser baseados em características calculadas a partir do seu conteúdo, das anotações textuais associadas, de atributos informados (metadados) ou de anotações semânticas.

Chen e outros (Chen et al., 2000), por exemplo, empregaram técnicas de visualização para comparar imagens com base em três métodos de extração de características: cor, forma e textura. Nessa abordagem, as imagens são classificadas por um sistema que emprega redes associativas, as quais são geradas por uma técnica chamada *Pathfinder Network Scaling* (PFnets) (Schvaneveldt, 1990). *PFnets* são utilizadas para identificar forte relação de similaridade entre os dados. Os nós dessas redes são conectados de modo a refletir a proximidade e a similaridade entre pares de imagens, restando somente as conexões mais importantes. Um exemplo da visualização dessa abordagem é mostrada na Figura 2.24, gerado com uma *PFnet* a partir da similaridade baseada

no histograma de cor das imagens. No exemplo foi utilizada uma coleção de imagens construída pelos autores.

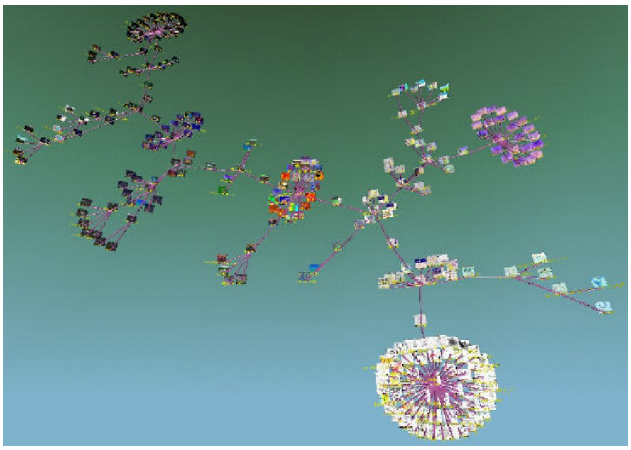
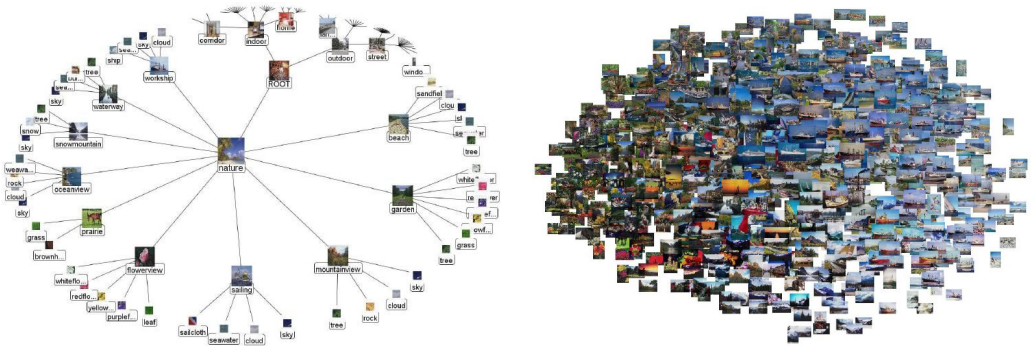


Figura 2.24: Coleção de imagens organizadas pela similaridade do histograma de cor (Chen et al., 2000).

Enquanto no trabalho realizado por Chen et al. são empregadas abordagens não supervisionadas para computar a similaridade das imagens, Fan et al. (Fan et al., 2007) classificaram imagens com base em anotações textuais. A partir das anotações, eles construíram uma ontologia com os objetos de interesse extraídos das imagens. Características são extraídas desses objetos para classificá-los e associá-los a um domínio específico determinado pela ontologia. Novas imagens adicionadas à coleção são classificadas de acordo com a ontologia, o que exige uma extração automática dos principais objetos das imagens e a extração de suas características. As características extraídas são comparadas com aquelas associadas a domínios da ontologia e classificadas por um algoritmo de *Support Vector Machines (SVM)* (Vapnik, 1995). A representação visual que agrupa as imagens com base na similaridade de conteúdo é construída por meio da técnica Análise de Componentes Principais (*Principal Component Analysis – PCA*) (Jolliffe, 1986). A interação sobre a ontologia é realizada por meio de uma árvore hiperbólica. Na Figura 2.25(a) é mostrada uma ontologia construída a partir de imagens de cenas naturais e na Figura 2.25(b) é mostrado o resultado de uma busca por imagens relacionadas com os termos “cenas naturais”.

Uma outra maneira de fazer a representação visual de coleções de imagens é abordada por Yang (Yang, 2004), que mostra um esquema de classificação de imagens projetando os dados em um espaço dimensional reduzido. Para isso, foi estendido o método de triangulação proposto por Lee e outros (Lee et al., 1977) para mapear um espaço de alta dimensão em um espaço de dimensão reduzida. Nessa abordagem não são extraídas características. Ao invés disso, todos os pixels são empregados para descrever o espaço de características da imagem. Yang descreve experimentos com coleções de imagens geradas de um único objeto, com variações na forma, brilho e orientação. Um exemplo utilizando uma coleção de imagens de face com variação de brilho e orientação é mostrado na Figura 2.26



(a) Ontologia sobre imagens de cenas naturais. (b) Resultado da busca por “cenas naturais”.

Figura 2.25: Recuperação de um conjunto de imagens (b) baseada em uma ontologia (a) (Fan et al., 2007).

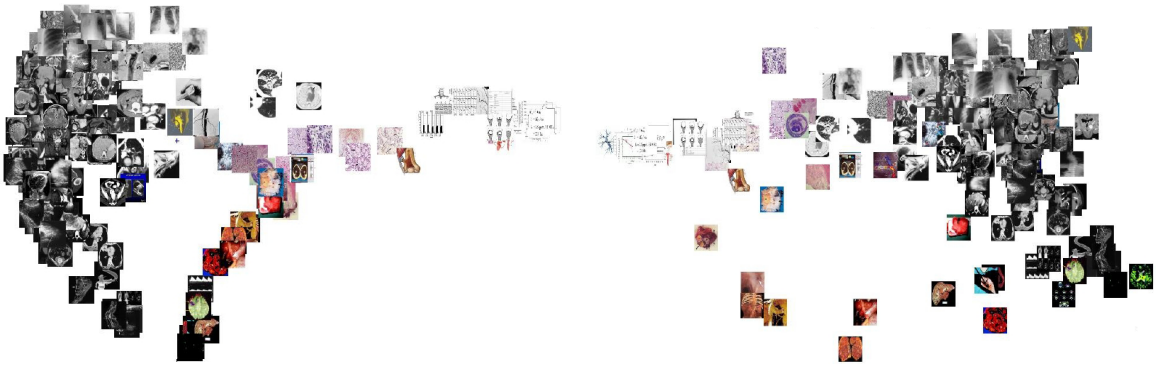
Figura 2.26: Coleção de imagens de face com variação de brilho e orientação (Yang, 2004).

Nguyen e Worring (Nguyen e Worring, 2008) propuseram uma abordagem para visualizar e anotar coleções de imagens. Nessa abordagem é proposta uma técnica de projeção que une as técnicas *Isometric Feature Mapping* (Isomap) (Tenenbaum et al., 2000) e *Stochastic Neighbor Embedding* (SNE) (Hinton e Roweis, 2002). Eles também propuseram uma maneira de visualizar imagens representativas ao invés de todo o conjunto, facilitando a visualização de grandes coleções de imagens. Para isso a coleção é particionada por um algoritmo de agrupamento (*k-means*) e somente as imagens mais próximas do centróide são exibidas. A Figura 2.27 mostra a visualização de parte da coleção de imagens (projeção da esquerda) e após a seleção de uma imagem (indicada por uma seta), as imagens do seu agrupamento são exibidas (projeção da direita).

Camargo e outros (Camargo et al., 2008) também propuseram uma abordagem para visualizar coleções de imagens baseada em projeção. Eles utilizaram a técnica *Classical Multidimensional Scaling* (Cox e Cox, 2000) e a técnica *Isometric Feature Mapping* (Isomap) (Tenenbaum et al., 2000) para projetar o espaço multidimensional no plano de visualização. Na Figura 2.28 são mostrados dois mapas construídos a partir da coleção de imagens *Cross-Language Retrieval in*

Figura 2.27: Projeção de amostras de uma coleção de imagens (esquerda) e projeção de um agrupamento da imagem selecionada (direita) (adaptado de (Nguyen e Worring, 2008)).

Image Collections 2007 (ImageCLEF 2007). O espaço de características representa informações de cor das imagens.



(a) Projeção MDS (b) Projeção ISOMAP.

Figura 2.28: Construção de dois mapas utilizando as técnicas MDS(a) e ISOMAP (b) a partir de características de cor extraídas da coleção de imagens ImageCLEF 2007 (Extraído de (Camargo et al., 2008)).

Schaefer (Schaefer, 2010) apresentou uma abordagem para visualizar e explorar coleções de imagens mapeando o plano de visualização em uma esfera. As imagens são projetadas com base na similaridade visual. Para isso, o espaço de características é composto pela média das informações de cor HSV das imagens, o qual é utilizado para posicionar as imagens na superfície da esfera. A matiz (*Hue*) indica o posicionamento na longitude e o valor (*Value*) indica o posicionamento na latitude da esfera. Dessa maneira, conforme ilustrado na Figura 2.29, a longitude representa diferentes cores e a latitude representa imagens mais claras ou mais escuras. As imagens são posicionadas em uma grade regular para evitar sobreposição de imagens. No caso de grande conjuntos de dados, uma abordagem hierárquica é utilizada para esconder ou mostrar as imagens conforme o usuário aproxima ou afasta do plano de visualização (esfera).

Figura 2.29: Coleção de imagens MPEG-7. Visão inicial (direita) e visão após rotacionar a esfera na longitude e latitude (esquerda) (Extraído de (Schaefer, 2010)).

Também é possível utilizar anotações manuais ou automáticas relacionadas com as imagens de uma coleção para construir mapas de similaridade. Nesse sentido, a técnica de posicionamento de pontos *incBoard* (Pinho et al., 2009) (descrita no fim da Seção 2.2.1) também foi aplicada para agrupar imagens similares com base nas anotações das imagens. Para calcular a similaridade entre as anotações foi utilizado o coeficiente de Jacquard (Romesburg, 1984). A Figura 2.30 ilustra parte de uma mapa de similaridade construído a partir das anotações de 1109 imagens.

Figura 2.30: Parte de um mapa de similaridade construído com a técnica *incBoard* a partir de anotações manuais de uma coleção de 1109 imagens (Extraído de (Pinho, 2009)).

Como apresentado nesta seção, os sistemas voltados para a exploração de coleções de imagens empregam técnicas de visualização para representar uma coleção de imagens no plano de visualização. Essa abordagem auxilia a compreensão do espaço de características utilizado para representar as imagens. Estes sistemas raramente utilizam múltiplas visões para enriquecer a exploração do conjunto de dados, portanto, também não utilizam mecanismos de coordenação. Nossa perspec-

tiva é que a exploração de coleções de imagens pode ser enriquecida com uma abordagem CMV. Por exemplo, seria possível analisar diferentes coleções de imagens em um mesmo processo de exploração, o que não é abordado pelos sistemas apresentados nesta seção.

2.3 Considerações Finais

A exploração de coleções de documentos e coleções de imagens são aplicações evidentes para a utilização de técnicas de visualização baseadas em mapas de similaridade. Como mostrado neste capítulo, existe uma grande quantidade de técnicas para explorar conjuntos de dados de formas variadas. A exploração de um conjunto pode ser realizada por meio de atributos informados (e.g. autor, citação, tópicos) ou de atributos extraídos (e.g. conteúdo da instância). Além disso, as representações visuais podem ser geradas por diferentes técnicas de visualização. Essas técnicas, herdadas da área de visualização de informação, criam representações visuais que auxiliam a análise e identificação de padrões existentes na coleção de dados, permitindo que o usuário entenda o comportamento do espaço de características.

Apesar da área de Múltiplas Visões Coordenadas (CMV) ter crescido nos últimos anos e ser aplicada em diversas áreas de conhecimento, das abordagens apresentadas neste capítulo, algumas utilizam múltiplas visões para explorar o conjunto de dados, mas poucas aplicam a coordenação entre as diferentes visões. É possível encontrar aplicações CMV que apóiam a exploração de coleções de documentos, mas é raro encontrar aplicações CMV para exploração de coleções de imagens. Este trabalho de doutorado propõe uma abordagem que utiliza múltiplas visões coordenadas para auxiliar a exploração de mapas de similaridade em geral, mas com especial atenção às coleções de documento e às coleções de imagens. Além disso, apesar de não ser discutido neste capítulo, este trabalho também utiliza essa abordagem CMV para explorar dados volumétricos, conforme é apresentado na Seção 6.4.

Este trabalho utilizou o arcabouço *Projection Explorer* (Paulovich et al., 2007) para gerar mapas de similaridade baseado em técnicas de projeção multidimensional e técnicas de posicionamento de pontos, pois ele disponibiliza a maioria das técnicas de projeção e de posicionamentos de pontos relatadas na literatura. Além disso, a PEx tem seu código livremente disponível e também foi desenvolvida no grupo de Visualização, Imagens e Computação Gráfica do ICMC/USP. Os mapas gerados pela PEx são utilizados para representar as relações entre coleções de documentos, imagens e dados volumétricos. Para auxiliar a exploração, técnicas de coordenação são propostas. Como resultado, um conjunto de ferramentas é produzido para adicionar novas possibilidades de análise às existentes, enriquecendo o arcabouço PEx. No próximo capítulo será discutida a área de Múltiplas Visões Coordenadas, mostrando exemplos de aplicações, com foco em dados textuais, e também modelos de coordenação para a elaboração de sistemas de múltiplas visões coordenadas.

CAPÍTULO

3

Múltiplas Visões Coordenadas

3.1 Considerações Iniciais

O processo de exploração e análise de dados pode ser fortalecido pela incorporação de técnicas de visualização como um instrumento para estimular a participação mais ativa do usuário, oferecendo-lhe suporte para monitoramento, avaliação e controle do processo, e elevando o grau de confiança nas observações (Ankerst, 2000; Keim, 2002). A flexibilidade nesse processo pode ser alcançada por meio de operações que permitam seleção de itens de dados individuais ou subconjuntos de interesse, destaque de determinados atributos, visão geral ou detalhada de trecho dos dados, navegação em estruturas hierárquicas e controle do mapeamento dos atributos (Grinstein e Ward, 2002; MacEachren et al., 1999; North, 1997).

As diferentes técnicas de visualização apresentam vantagens e desvantagens em sua utilização. Para amenizar as deficiências individuais de cada técnica, outras podem ser adicionadas ao processo de exploração. A adição de diferentes tipos de técnicas pode sobrecarregar o usuário na adaptação do contexto de cada uma. Então, para facilitar a associação que o usuário tem que fazer entre elas, técnicas de coordenação podem ser utilizadas para ajudar a conectar as diferentes visões [1]. A área que trata deste problema é conhecida como Múltiplas Visões Coordenadas (CMV - *Coordinated and Multiple Views*), ou *Linked Views*. Há muitos problemas em aberto na busca por sistemas de visualização baseados em CMV. Atualmente, esses sistemas têm como requisito a capacidade de manipular não só grandes quantidades de dados, mas também padrões complexos de um ou mais conjuntos de dados (Heijs, 2007). Segundo Heijs (Heijs, 2007) empresas dedicadas

[1] Visão é a representação visual de um conjunto de dados por meio de uma técnica de visualização

a aplicações comerciais não possuem o conhecimento necessário para desenvolver novas técnicas de visualização que utilizam múltiplas visões coordenadas. Universidades ou empresas que focam em pesquisas são mais adequadas em aceitar esse tipo de desafio em visualização. Heijs também diz que se tais empresas não focarem em CMV e *Data Mining* estarão perdendo oportunidades atuais de desenvolvimento.

A próxima seção apresenta exemplos de sistemas que utilizam CMV, com foco naqueles que utilizam como fonte dados textuais. Na Seção 3.3 são apresentados modelos para a construção de sistemas que utilizam múltiplas visões coordenadas. Por fim, na Seção 3.4 é apresentada uma discussão sobre a área de CMV e a sua relação com o presente trabalho de doutorado.

3.2 Exemplos de Aplicações de Múltiplas Visões Coordenadas

Múltiplas visões permitem observar um conjunto de dados sob várias perspectivas, bem como explorar os pontos fortes e minimizar o efeito dos pontos fracos das técnicas de visualização (Keim e Kriege, 1996). Geralmente, sistemas que utilizam CMV possuem mais de uma técnica de visualização para explorar um mesmo conjunto de dados, conjuntos de dados diferentes ou diferentes subconjuntos de um único conjunto de dados.

A Figura 3.1, gerada com o sistema *WinDirStat* [2], mostra um exemplo de múltiplas visões coordenadas. Nesse exemplo existem duas representações visuais sobre o mesmo conjunto de dados, isto é, o sistema de arquivos de um disco rígido. Em uma das visões, os dados são representados por uma *Folder Tree* e na outra visão são representados por uma *Treemap*. Um retângulo na *Treemap* indica um arquivo e um grupo indica um diretório. O tamanho do retângulo representa o tamanho do arquivo e a cor representa o tipo de arquivo. As duas visões estão coordenadas da seguinte forma, se um retângulo for selecionado, o arquivo representado por ele será destacado na visão da *Folder Tree*, ou vice versa. Essa seleção é conhecida na literatura como *Brushing* (Becker e Cleveland, 1987), ou *brushing-and-linking*, que é uma seleção e destaque do mesmo dado em duas visões distintas. Geralmente essa seleção se baseia em algum identificador da instância, no caso, o nome do arquivo.

Visualizar o mesmo conjunto de dados com diferentes visões aumenta a capacidade de aquisição de conhecimento. Fazer a conexão entre as diferentes visões, por meio da coordenação, reduz a carga cognitiva de migrar de uma visão para outra, permitindo que se façam com mais facilidade análises em duas visões separadas. Entretanto, fazer a conexão entre duas visões conceitualmente distintas pode não ser tão simples se os tipos de visões são complexos e muito diferentes. Siirtola (Siirtola, 2003) apresenta a comparação de duas técnicas de visualização conceitualmente distintas, Coordenadas Paralelas e Matriz de Permutação (ou matriz ordenável, do inglês, *Reorderable Matrix*).

[2] Disponível em http://windirstat.info/

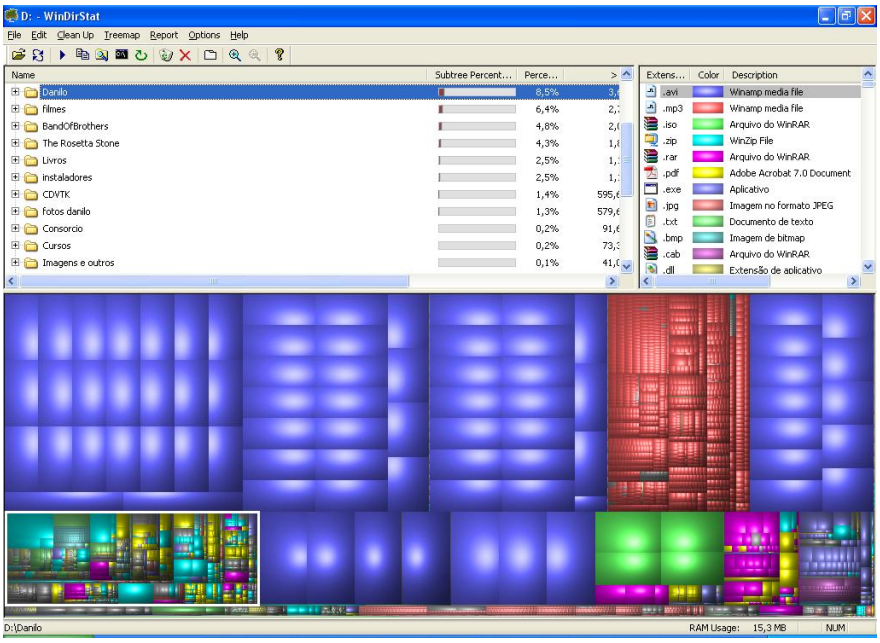


Figura 3.1: Visualizações coordenadas de um sistema de arquivos utilizando *Treemap*.

Na técnica de visualização Coordenadas Paralelas, ilustrada na Figura 3.2, cada instância do conjunto de dados é visualmente representado por uma linha. Cada eixo paralelo representa um atributo do conjunto de dados e a ligação entre os eixos (por meio das linhas) indica o valor da instância em cada um dos atributos (eixos). Na técnica de visualização Matriz de Permutação cada coluna representa uma instância do conjunto de dados e cada linha da matriz representa um atributo. O valor de cada atributo é representado pela altura de um símbolo gráfico que aparece nas células da matriz.

Segundo Siirtola, as coordenadas paralelas e a matriz de permutação se complementam em muitos aspectos. A matriz de permutação é boa para identificação de padrões, classificação, formação de hierarquia e facilidade de ordenar os dados, opondo-se à técnica coordenadas paralelas. Além de uma técnica suprir algumas desvantagens da outra, ambas possuem a capacidade de exibir toda a representação visual do conjunto de dados em uma única visão.

A Figura 3.2 mostra a visualização de um conjunto de dados utilizando as duas técnicas. A coordenação destaca as instâncias ou atributos no momento em que o usuário faz uma seleção, propagando-a de uma visão para a outra. Se uma linha na visão das coordenadas paralelas é selecionada, a coluna que representa a mesma instância é destacada na visão da matriz de permutação. Se uma célula ou mais são selecionadas na visão da matriz de permutação, então são feitas as ligações entre os eixos correspondentes na visão que utiliza as coordenadas paralelas.

Também utilizando diferentes técnicas de visualização para explorar um mesmo conjunto de dados, Godinho e outros (Godinho et al., 2007) desenvolveram um sistema para explorar conjuntos de dados, denominado PRISMA. O sistema utiliza três técnicas de visualização bem conhecidas, cada uma favorecendo diferentes aspectos da análise de dados. São elas: Gráficos de Dispersão, *Treemap* e Coordenadas Paralelas (Figura 3.3).

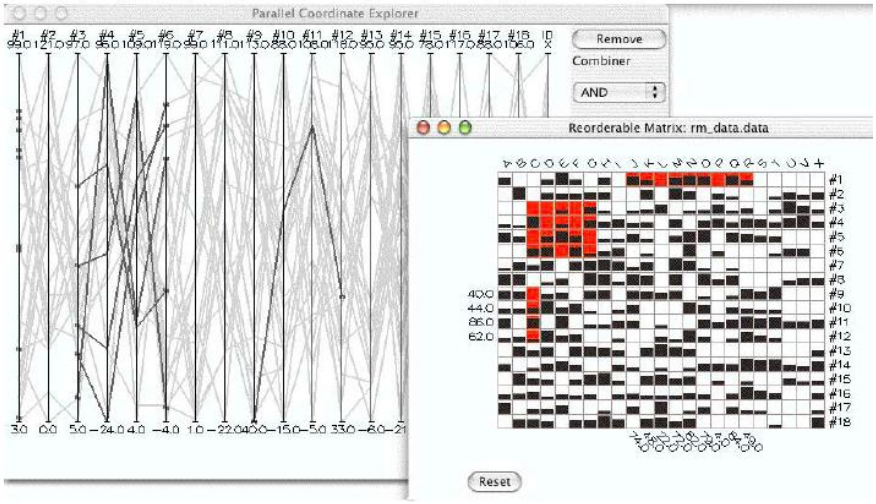


Figura 3.2: Visualização de um conjunto de dados por meio de Coordenadas Paralelas e Matriz de Permutação (Extraído de (Siirtola, 2003)).

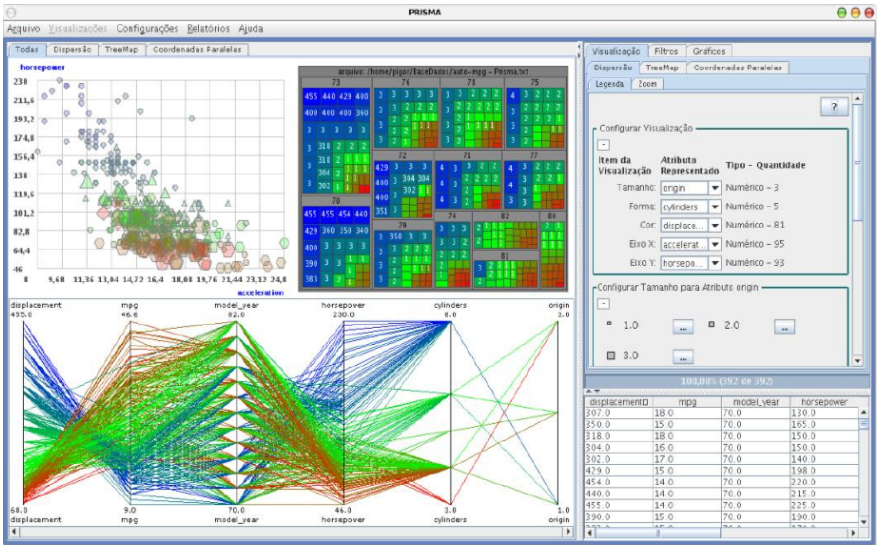


Figura 3.3: Interface do sistema PRISMA (Extraído de (Godinho et al., 2007)).

O sistema PRISMA utiliza um único conjunto de dados para todas as visões. O usuário pode analisar os dados individualmente em cada visão ou simultaneamente em todas elas. A coordenação entre as técnicas é estática, ou seja, a coordenação é predefinida para as visões, não podendo ser alterada. A coordenação de filtros, seleção, cor e detalhes sob demanda estão disponíveis para todas as visões, mas tamanho e forma estão disponíveis somente para algumas visões, conforme mostra a Figura 3.4.

Carpendale (Carpendale, 2007) propõe uma abordagem diferente das convencionais para mostrar o relacionamento entre as visões, a qual recebeu o nome de *VisLink*. Ela estende as abordagens existentes para visualizar múltiplos relacionamentos e mostrar o relacionamento entre as visualizações enquanto mantém suas posições espaciais. Essa abordagem utiliza múltiplas representações (*layouts*) 2D, desenhando cada uma em seu plano. Estes planos podem ser posicionados e reposici-

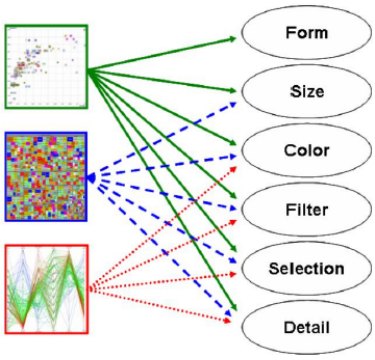


Figura 3.4: Técnicas de visualização do sistema PRISMA e suas respectivas coordenações disponíveis (Extraído de (Godinho et al., 2007)).

onados lado a lado, em paralelo, ou em lugares que favoreçam a visualização. Os relacionamentos entre as visões podem ser visualizados por meio de arestas, as quais conectam os elementos das visões, como ilustrado na Figura 3.5.

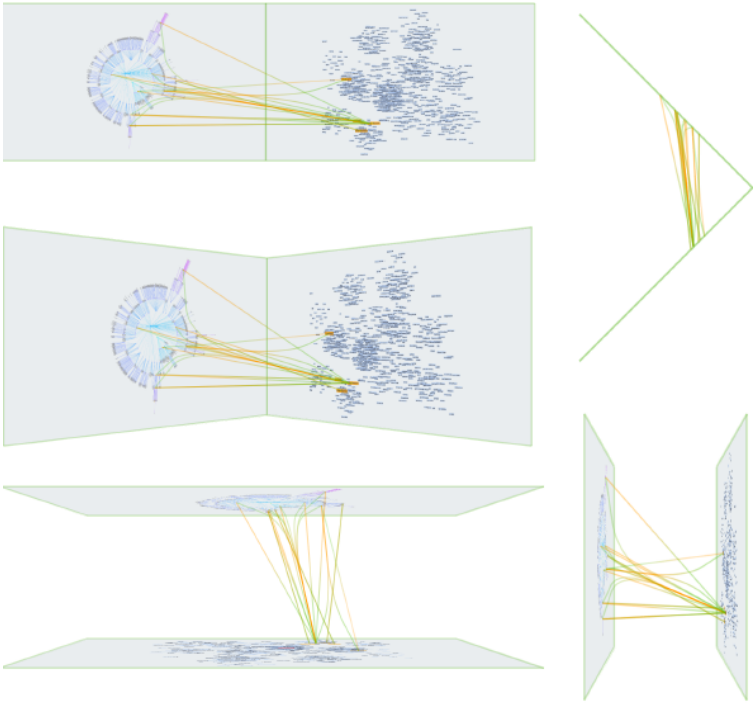


Figura 3.5: *VisLink*: diferentes posicionamentos dos planos de visualização (Extraído de (Carpendale, 2007)).

As arestas indicam a relação entre os elementos das visões, a direção do relacionamento é indica pela cor, variando de laranja (origem) para verde (destino). As relações estabelecidas podem ser diretas, nas quais os elementos que representam as mesmas instâncias são conectados, ou indiretas, nas quais os elementos entre os planos de visualização têm o relacionamento estabelecido por meio de regras.

Os sistemas anteriormente apresentados utilizam CMV para explorar conjuntos de dados estruturados, deixando de lado outros conjuntos que não possuem uma organização em atributos, como

as coleções de documentos que possuem dados não estruturados. Para utilizar esses sistemas seria necessário extrair atributos dos dados. No entanto, a grande quantidade de atributos é um problema para a maioria das técnicas utilizadas nesses sistemas.

Em algumas aplicações é necessário analisar a relação entre dados estruturados e dados não estruturados, por exemplo, áreas de atendimento ao cliente, em que uma pesquisa pode ser realizada via um formulário ou receber também a opinião por meio textual. Com a motivação de analisar a relação entre dados estruturados e não estruturados, Seeling e Becks (Seeling e Becks, 2004a) utilizaram múltiplas visões coordenadas para auxiliar o usuário nesse processo. No trabalho proposto o usuário pode explorar o conjunto de dados com três tipos de visões, ilustradas na Figura 3.6.

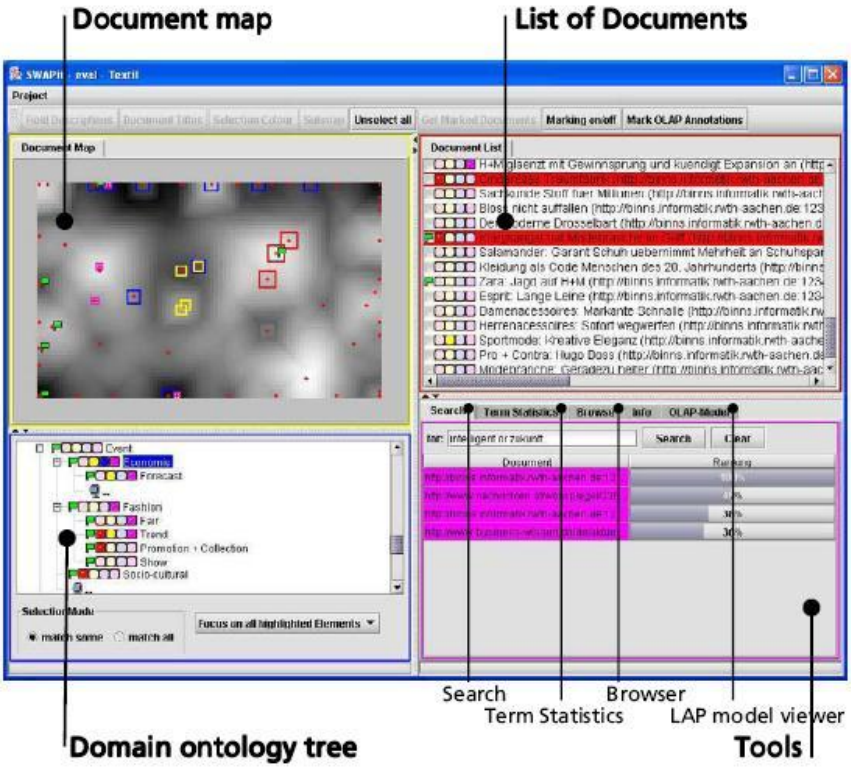


Figura 3.6: Visões para análise de dados estruturados e não estruturados (Extraído de (Seeling e Becks, 2004a)).

A visão *Document Map* é construída com base no conteúdo dos documentos (dados não estruturados). Os dados passam por um algoritmo de agrupamento para agrupar as instâncias por similaridade. Nessa visão, cada instância é visualmente representada por um ponto no plano de visualização. A visão *Domain Ontology Tree* trabalha no nível conceitual do conjunto de dados. A ontologia do domínio é exibida como uma árvore na qual cada item é um tópico e os filhos são sub-tópicos. A visão *List of Documents* exibe os dados estruturados. Também há uma janela disponibilizando ferramentas para busca no conjunto de dados, para classificação e para obtenção de estatísticas do conjunto.

A coordenação é utilizada para destacar o relacionamento entre os elementos das três visões. Quando um elemento é selecionado na visão *Document Map*, os elementos relacionados nas outras

visões são destacados com a cor amarela. Quando um elemento da visão *Domain Ontology Tree* é selecionado, os itens relacionados nas outras visões são destacados com a cor azul. Quando um elemento é selecionado na visão *List of Documents*, os elementos das outras visões são destacados com a cor vermelha. Utilizando a janela de ferramentas, quando uma busca é realizada os elementos cobertos pela pesquisa são destacados pela cor rosa.

Chang e outros (Chang et al., 2007) também utilizaram múltiplas visões coordenadas para analisar dados compostos de informações estruturadas e não estruturadas. Eles apresentam uma abordagem para auxiliar analistas na exploração de conjuntos de dados variando no tempo, contendo informações de transações financeiras (*wire transactions*). Essa abordagem foi chamada *WireVis*. O método é interativo e combina quatro visões: *keyword network view*, *heatmap*, Busca por Exemplo e *Strings and Beads*. Essas visões descrevem o relacionamento entre as contas das transações, o tempo e as palavras chaves das transações, permitindo agregar e organizar grupos para uma melhor investigação, análise e comparação de registros individuais. Das visões utilizadas, três estão coordenadas e são descritas a seguir:

- ***Keyword Network View*:** tem o objetivo de descrever relacionamentos entre palavras chave. Ela é importante para identificar transações questionáveis. Para mostrar o relacionamento entre palavras chaves, foi utilizado um grafo para conectá-las, como ilustrado na Figura 3.7. Uma palavra chave está relacionada a outra se ambas aparecem em uma mesma transação. O aparecimento de palavras chaves em uma mesma transação forma a base da matriz de relacionamento entre elas, na qual a distância entre as palavras é calculada com base no número de vezes que elas aparecem juntas na transação. Palavras chave próximas ao centro da *Keyword Network View* são as mais frequentes na transações, enquanto que as mais distantes do centro são as menos frequentes. Quando uma palavra chave é selecionada, linhas são desenhadas entre aquelas que estão relacionadas a ela, as quais também são destacadas.

- ***Heatmap*:** é utilizada para exibir medidas estatísticas relacionadas às palavras chave e às contas bancárias (ver Figura 3.8). Essa visão utiliza uma malha cujas colunas são as palavras chave de interesse e as linhas são os grupos de contas bancárias em análise. Na intersecção entre uma linha e coluna é mostrado o número de vezes que uma palavra chave aparece no grupo de transações em um determinado período de tempo. A saturação das cores é proporcional ao número de vezes que uma palavra chave aparece no grupo. Para evitar que o efeito de utilizar uma soma cumulativa no *Heatmap*, quando uma célula é selecionada um histograma mostra o número de palavras chave por transação. No histograma, cada linha horizontal na cor verde mostra o número de vezes que uma palavra chave apareceu para uma determinada transação. A seleção de uma palavra chave é refletida nas outras visões; os dias que possuem transações com a palavra selecionada ficam em destaque na visão *Strings and Beads*, permitindo verificar a ocorrência das palavras chave sobre a variação do tempo.

- ***Strings and Beads*:** mostra as atividades das transações no tempo, como ilustrado na Figura 3.9. Nessa visão as linhas do gráfico (*strings*) representam as transações ou grupos

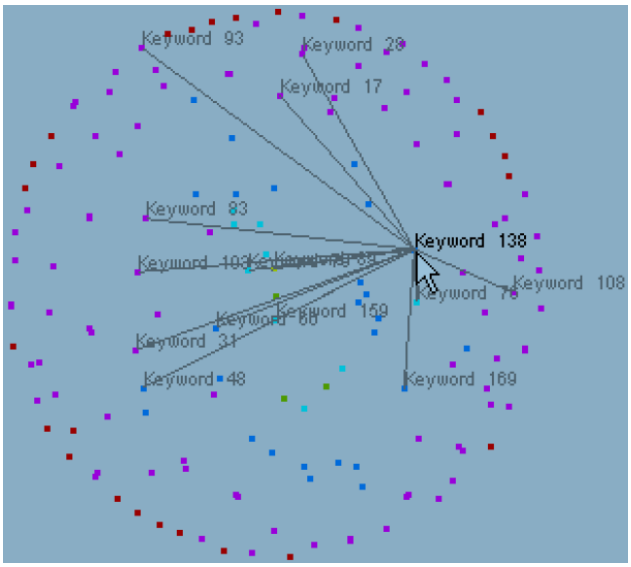


Figura 3.7: Visão *Keyword Network*: exibe o relacionamento entre as palavras chaves (Extraído de (Chang et al., 2007)).

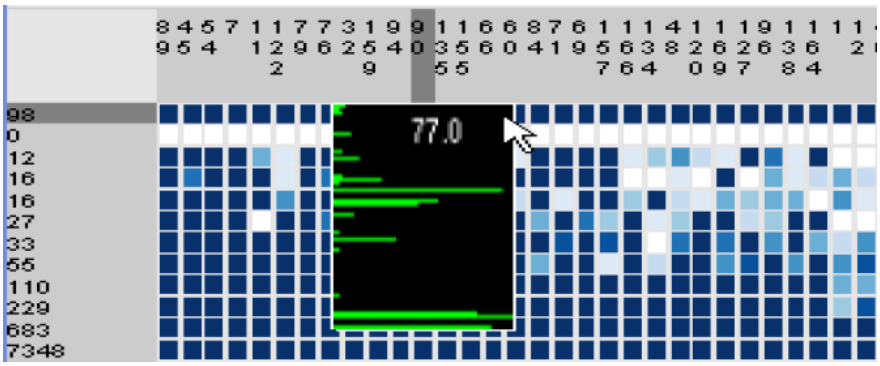


Figura 3.8: Visão *Heatmap*: exibe o relacionamento entre as palavras chave e os grupos de transações (Extraído de (Chang et al., 2007)).

de transações sobre o tempo e a menor porção da linha (*bead*) representa as transações em um dado dia. A visão mostra as tendências gerais das atividades, bem como de uma transação individual. O eixo X mostra a progressão do tempo e o eixo Y mostra o valor da transação, que pode ser a quantidade de transação, freqüência das atividades, número de vezes que uma determinada palavra chave aparece, entre outras opções. Como mostrado na Figura 3.9, quando um *bead* é selecionado, os detalhes específicos das transações daquele dia são mostrados em uma visão separada. Quando um *bead* é selecionado as transações agrupadas por ele são destacadas na visão *Heatmap*, permitindo saber quais palavras chaves estão relacionadas com esse *bead*.

A dificuldade em encontrar um padrão nas transações é grande. Depois de encontrado, pode-se utilizar a ferramenta Busca por Exemplo (Figura 3.10) para encontrar grupos de transações semelhantes. Em uma visão separada, é mostrado o grupo atualmente selecionado ou uma conta para utilizar como exemplo de busca. As barras representam o número de vezes que as palavras

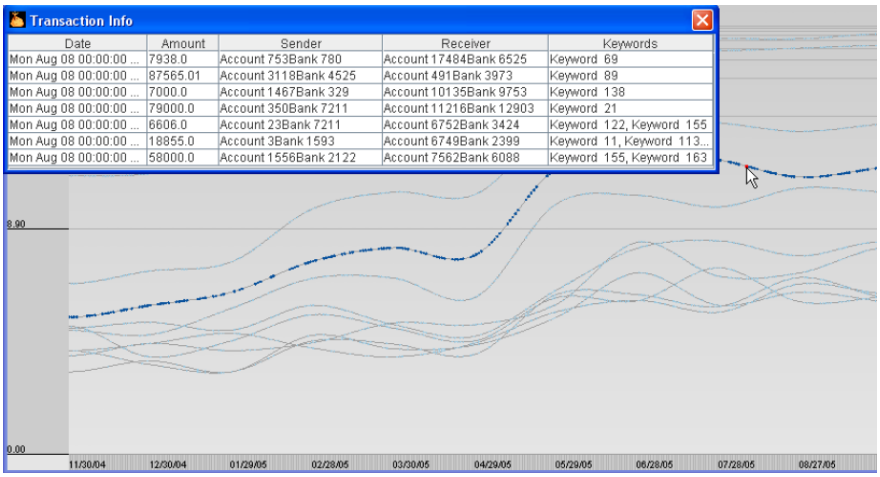


Figura 3.9: Visão *String and Bead*: exibe a visão geral e específica de agrupamentos de transações (Extraído de (Chang et al., 2007)).

chaves definidas aparecem. Um *slider* é usado para definir similaridade na identificação de um grupo ou transação. Conforme o usuário movimenta o *slider*, o número de transações similares aumenta ou diminui, dando a sensação do espaço a ser explorado. Em seguida, o usuário pode selecionar o grupo desejado para ser explorado em outra janela, utilizando as visões anteriormente descritas.

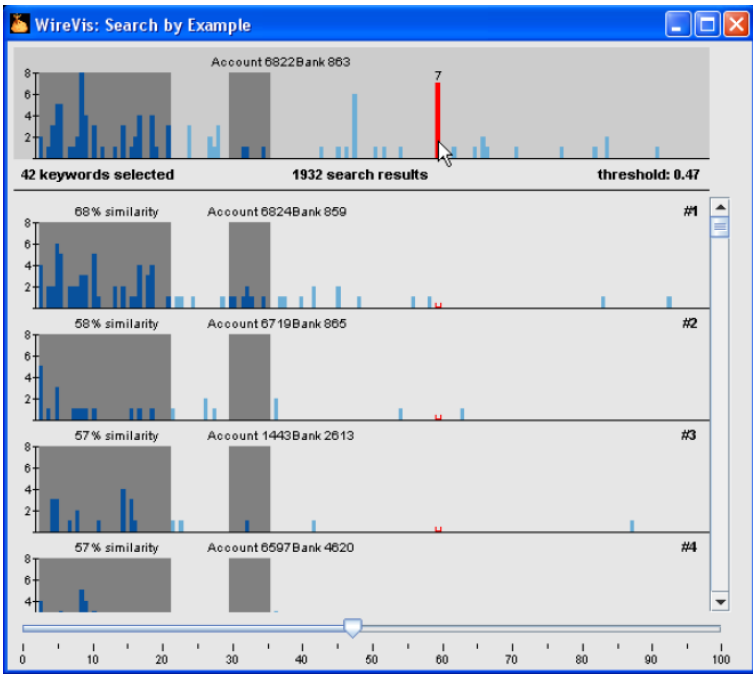


Figura 3.10: Ferramenta Busca por Exemplo: permite buscar transações semelhantes a um modelo selecionado pelo usuário (Extraído de (Chang et al., 2007)).

Também aplicando visualização para auxiliar o usuário a organizar dados não estruturados, Spoerri (Spoerri, 2004) propõe formas de visualizar dados provenientes de meta buscas, mas nesse caso não há relação de similaridade entre os documentos recuperados. A idéia é agrupar os do-

cumentos de forma que o usuário possa identificar por qual motor de busca ele foi recuperado, o seu *ranking* e por quantos motores distintos ele foi recuperado. Essa ferramenta recebeu o nome de *MetaCrystal*. A representação visual utiliza técnicas de interação que podem auxiliar o usuário a filtrar e combinar documentos retornados dos diferentes motores de busca. Também utiliza posição, tamanho, cor, orientação e forma como atributos visuais para os documentos recuperados. O trabalho desenvolvido por Spoerri utiliza três visualizações que organizam visualmente os documentos recuperados pelas meta buscas.

Uma das visões, denominada *Category View*, é mostrada na Figura 3.11. Essa visão mostra o número de documentos encontrados por diferentes combinações de motores de busca. Forma, cor, posição e orientação são utilizadas para visualizar as diferentes combinações. Na borda da figura são mostrados o nome dos motores de busca utilizados na meta busca e o número de documentos recuperados por cada um. Cada gráfico representa um conjunto de documentos e o número indica quantos documentos estão no conjunto. A forma do gráfico indica quantos motores de busca encontraram o mesmo conjunto de documentos. Por exemplo, o retângulo representa dois motores de busca, o triângulo representa três motores de busca e assim por diante. A cor indica quais são os motores de busca que encontraram o mesmo documento.

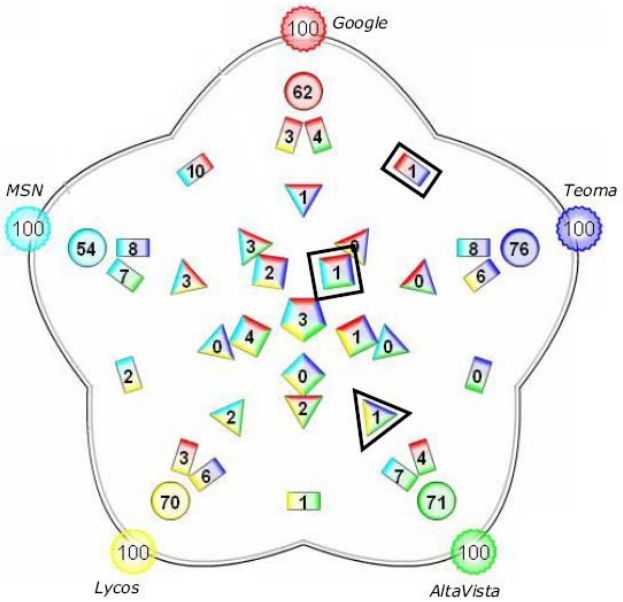


Figura 3.11: Visão *Category View* (Extraído de (Spoerri, 2004)).

Em uma outra visão, os gráficos da visão *Category View* são decompostos e representados individualmente, como exibido na Figura 3.12. Essa visão, denominada *Cluster Bulls-Eye*, mostra todos os documentos recuperados. Aqueles encontrados por um único motor de busca são posicionados na parte extrema do círculo. Enquanto que aqueles recuperados por vários motores de busca tendem em direção ao centro. A posição do documento dentro do círculo é baseada no motor de busca que possui melhor *ranking*. Em seguida, o documento é deslocado com base no *ranking* dos outros motores de busca para ter seu posicionamento final.

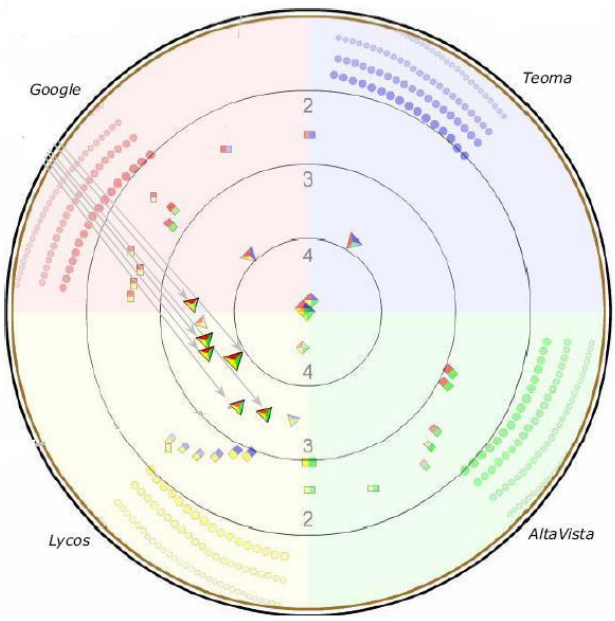


Figura 3.12: Visão *Cluster Bulls-Eye* (Extraído de (Spoerri, 2004)).

Uma outra visão também exibe todos os documentos recuperados, colocando-os sequencialmente em forma de espiral baseados em seu *ranking*. Essa visão, mostrada na Figura 3.13, é denominada *RankSpiral*. Ela mantém a idéia de que os documentos recuperados por um único motor de busca estão na periferia do círculo e os que foram recuperados por mais motores de busca rumam para o centro. Além disso, os documentos são organizados de acordo com o seu *ranking*, dando ao usuário uma noção dos documentos que possuem mais relevância com a busca efetuada.

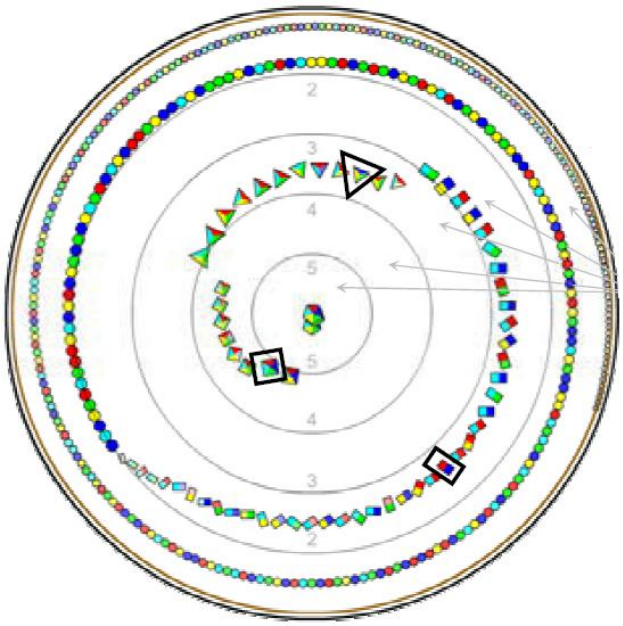


Figura 3.13: Visão *RankSpiral* (Extraído de (Spoerri, 2004)).

A coordenação foi utilizada para destacar os documentos nas diferentes visões. Se um grupo é selecionado na visão *Category View* os documentos pertencentes ao grupo são selecionados nas demais visões. Da mesma forma, se um documento é selecionado nas visões *RankSpiral* ou *Cluster Bulls-Eye* ele é destacado nas demais e o grupo ao qual ele pertence na visão *Category View* também é destacado. Quando um documento é selecionado na visão *RankSpiral* ou *Cluster Bulls-Eye* o seu ranking é mostrado em cada um dos motores de busca que o recuperou. Também é mostrado o conteúdo do documento. Ferramentas de filtragem também foram adicionadas ao processo, permitindo ao usuário destacar os documentos recuperados pelos motores de busca selecionados ou os documentos recuperados por um número determinado de motores de busca.

Recentemente Spoerri (Spoerri, 2007) retomou e aprimorou o seu trabalho sobre meta buscas, criando o chamado *searchCrystal*, o qual utiliza as informações textuais para auxiliar a exploração do conjunto de dados. O objetivo principal é utilizar texto e imagem para auxiliar o usuário em buscas exploratórias, como mostrado na Figura 3.14. Assim, o tamanho da figura e o tamanho da letra indicam a relevância do documento com base na média do ranking dos motores de busca. Linhas podem ser utilizadas para ligar as figuras, representando tópicos e a origem dos resultados de busca. Se os resultados estão ligados por uma linha azul, isso indica que possuem um mesmo tópico. Se estão ligados por uma linha laranja, indica que eles estão hospedados em um mesmo servidor.

Além da exibição de texto juntamente com figuras, a *searchCrystal* possui painéis que auxiliam a exploração dos resultados de busca. Um deles é conhecido como *FreqCloud* (Figura 3.15) o qual exibe as palavras que tiveram maior frequência nos títulos e resumos (*snippet*) dos resultados de busca. No *FreqCloud* são exibidas as 40 palavras mais frequentes, os termos utilizados na palavra de busca e que aparecem em todos os resultados são excluídos da lista de palavras freqüentes. As palavras são exibidas em ordem alfabética e o tamanho do texto indica a freqüência das palavras, quanto maior texto maior a freqüência da palavra. Quando uma palavra é selecionada, são mostradas aquelas que co-ocorrem nos resultados da busca. O tamanho da barra laranja que envolve a palavra indica se a co-ocorrência é forte ou fraca.

Um outro painel que auxilia a exploração é o *Topics Panel* (Figura 3.16). Ele mostra os tópicos que mais aparecem nos resultados da busca. Para identificar os tópicos são formados agrupamentos com o *Carrot2 Clustering Framework* [3]. Dos agrupamentos formados, são selecionados os que possuem mais elementos; se tiver mais de 10 agrupamentos, são selecionados os nove maiores e o restante é rotulado como 'outros'. Quando o *mouse* é colocado sobre um tópico são mostrados quais tópicos e quais palavras frequentes estão relacionados a ele. O número ao fim do nome do tópico indica quantos resultados estão relacionados com o tópico selecionado, os quais são destacados na visão principal, como mostrado na Figura 3.17.

A *searchCrystal* também possui um painel que exibe o *ranking* dos resultados de busca, como mostrado na Figura 3.18. Eles são ordenados pela quantidade de motores de busca que os encon-

[3]Disponível em http://project.carrot2.org

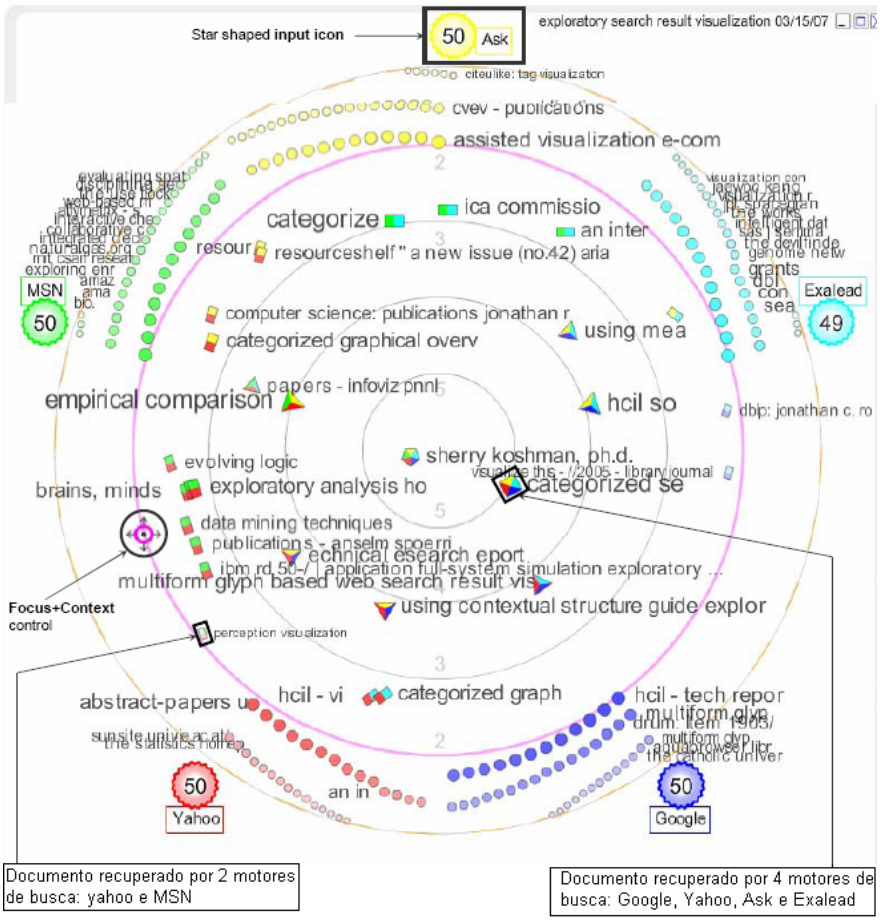


Figura 3.14: Técnica *Cluster Bulls-Eye* mostrando os 50 primeiros resultados de cinco diferentes motores de busca (Extraído de (Spoerri, 2007))

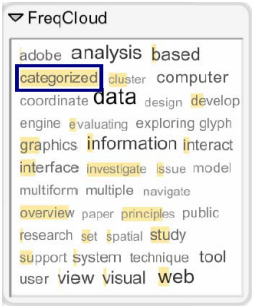


Figura 3.15: Painel *FreqCloud* mostrando as palavras mais freqüêntes (Extraído de (Spoerri, 2007))

traram. Os motores de busca são identificados pela cor da barra, que indica o *ranking*, e também pelo seu nome. Nessa visão também são colocados o título do resultado e o seu resumo.

Recentemente, Sabol e outros (Sabol et al., 2009) apresentaram a união de dois arcabouços para auxiliar a exploração e a visualização de coleções de documentos. O primeiro arcabouço é conhecido como *KnowMiner* (Klieber et al., 2009) e disponibiliza funcionalidades para automaticamente extrair conhecimento de grande coleções de documentos. O segundo, conhecido como *VisTools*, é um arcabouço baseado em múltiplas visões coordenadas e desenvolvido para atuar sobre as fun-

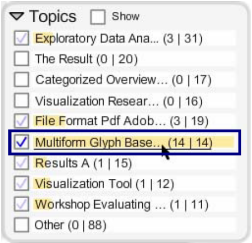


Figura 3.16: Painel *Topic Panel* mostrando os tópicos dos resultados de busca (Extraído de (Spoerri, 2007)).

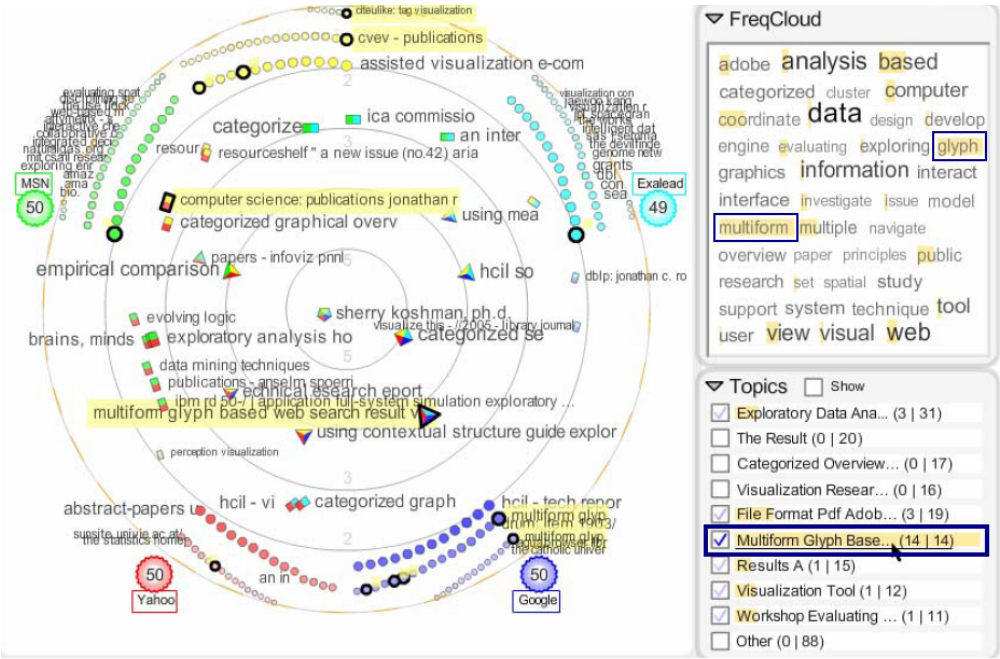


Figura 3.17: Seleção de um tópico no painel *Topic Panel*. Os resultados relacionados são destacados (Extraído de (Spoerri, 2007)).

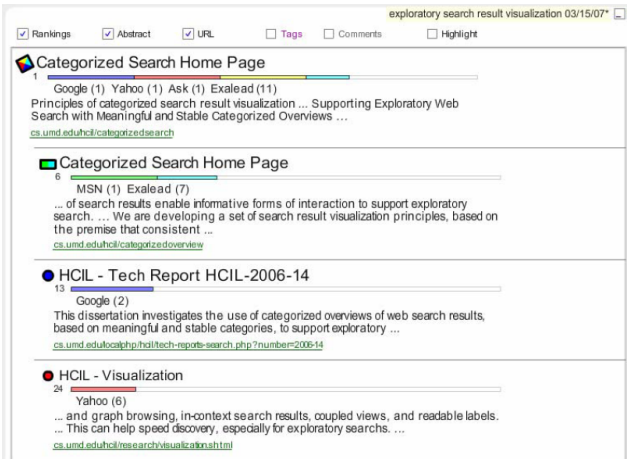


Figura 3.18: Lista de ranking, onde somente as opções *Ranking*, *Abstract* e *URL* foram escolhidas para aparecer na listagem (Extraído de (Spoerri, 2007))

cionalidades disponibilizadas pelo *KnowMiner*. O *VisTools* disponibiliza ferramentas para análise visual de coleções de documentos e exploração baseada em tópicos extraídos da coleção.

A Figura 3.19 mostra um exemplo de exploração de uma coleção de documentos por meio dos arcabouços *VisTools* e *KnowMiner*. No exemplo, quatro visões foram integradas por meio de múltiplas visões coordenadas. Uma das visões utiliza a representação visual *Landscape* para representar os relacionamentos entre os documentos da coleção, os quais foram construídos a partir de técnicas de projeção. Nessa visão são exibidos os tópicos dos principais grupos, que são os termos mais frequentes dos documentos agrupados. Uma outra visão, chamada de *StreamView*, é baseada na técnica *Theme River* (Havre et al., 2002). Ela realiza a visualização dos tópicos extraídos dos grupos e mostra a evolução temporal do tópico. A atividade de um tópico pode ser observada pela espessura da linha, a qual indica o número de documentos relacionados a ele. Também foram utilizadas duas outras visões baseadas em árvore (*TreeView*). Uma delas mostra os tópicos dos grupos e a outra mostra o nome das entidades agrupadas por classe (por exemplo, pessoas e localização). Apesar de não ser apresentado no exemplo, o sistema também disponibiliza duas visões de tabelas que exibem os sub-agrupamentos de um grupo selecionado em uma das visões. Enquanto que a outra exibe os documentos envolvidos na seleção. A coordenação pode ser iniciada em qualquer uma das visões. No exemplo foi selecionado um grupo identificado com os termos "russia eu nato". Em seguida, esse grupo é destacado na árvore que representa os tópicos e o tópico é identificado na visão *StreamView*.

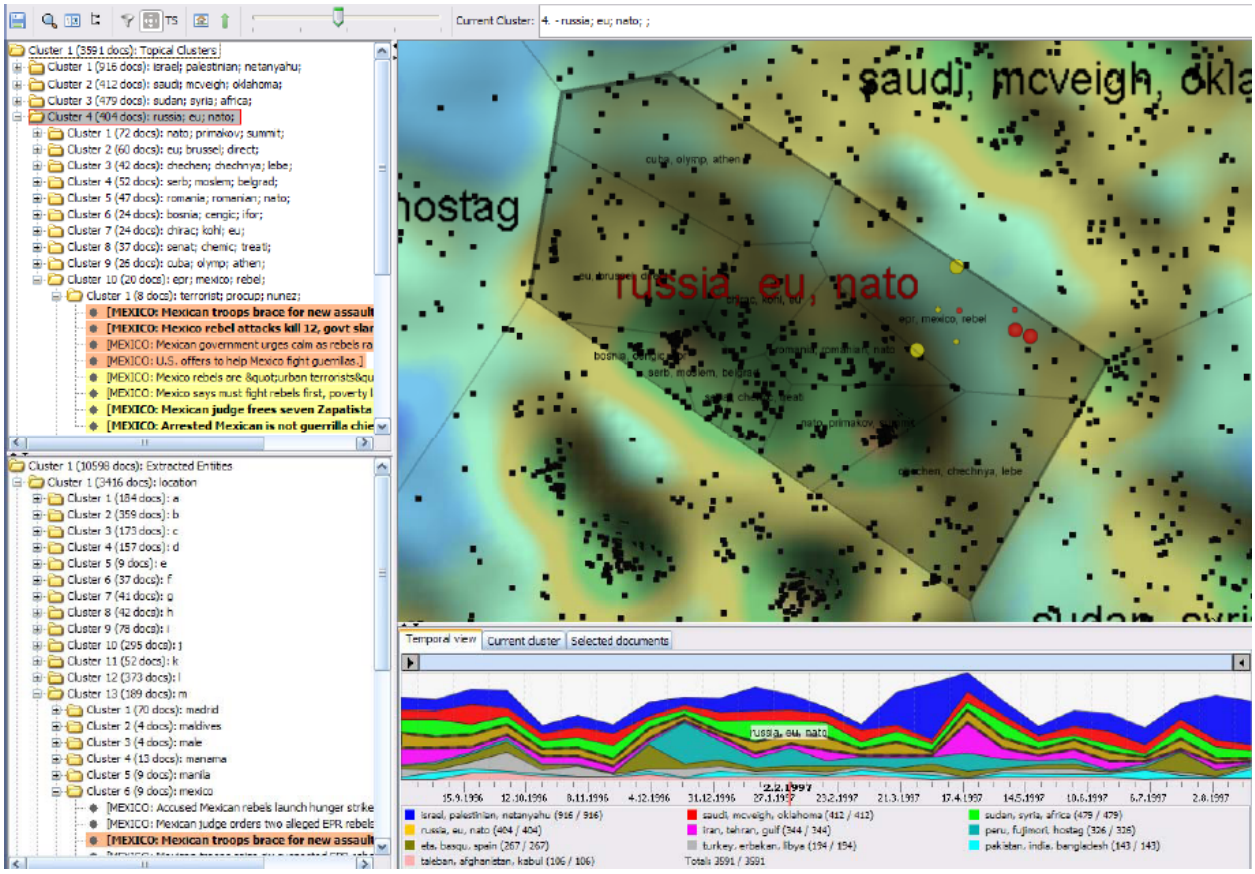


Figura 3.19: Exploração de uma coleção de documentos por meio do arcabouço *VisTools* (Extraído de (Sabol et al., 2009))

Os sistemas de CMV apresentados neste capítulo e que auxiliam a exploração de conjuntos não estruturados, em específico, coleções de documentos, se assemelham com a nossa proposta. Uma das diferenças é que nossa abordagem utiliza tópicos extraídos durante a interação do usuário para realizar a coordenação, enquanto que as abordagens apresentadas utilizam um ontologia do domínio de análise ou extraem tópicos de agrupamentos, não permitindo que o usuário indique os documentos que devam ser considerados para a extração dos tópicos. Além disso, os sistemas apresentam coordenações estáticas, ou seja, as coordenação não podem ser definidas ou configuradas pelo usuário. Neste trabalho, as técnicas de coordenação são criadas e configuradas pelo usuário durante o processo de exploração. Isso permite que diferentes técnicas de coordenação fossem propostas para auxiliar a exploração de coleções de documentos e também para a exploração de coleções de imagens.

A seguir são apresentados modelos para a construção de sistemas que utilizam a abordagem de múltiplas visões coordenadas.

3.3 Modelos de Coordenação

Na literatura, alguns modelos foram propostos para apoiar a construção de múltiplas visões (Schroeder et al., 1996; Fekete, 2004; Heer et al., 2005; Stolte e Hanrahan, 2000), mas não consideravam a coordenação entre elas. Um sistema de visualização com mecanismos de coordenação flexíveis requer um modelo apto em formalizar como as informações são trocadas entre as múltiplas visões, especificando como e o que pode ser coordenado. Nesse sentido, foram propostos modelos para permitir a coordenação entre diferentes técnicas de visualização (Jacobson et al., 1994; Takatsuka e Gahegan, 2002; Rundensteiner et al., 2002; Pillat, 2006; Moreno, 2006; Sanver e Yang, 2009). A seguir serão descritos os três principais modelos de coordenação encontrados na literatura. O primeiro e o terceiro se apóiam no modelo de dados relacional (North e Shneiderman, 2000a; Weaver, 2004) e o segundo no modelo *Model-View-Controler* (Boukhelifa e Rodgers, 2003), o qual foi adaptado para a construção das técnicas de coordenação propostas nesse trabalho de doutorado.

Modelo Proposto por North e Shneiderman

O primeiro sistema de visualização que adota um modelo que permite aos usuários a coordenação das ferramentas de visualização foi o *Snap Together Visualization* (North e Shneiderman, 2000a). O Snap engloba uma arquitetura e um sistema que possibilitam integrar diferentes ferramentas de visualização construídas por diferentes desenvolvedores. Os usuários podem construir suas próprias ferramentas e também a forma como elas serão coordenadas, facilitando a construção de ambientes de visualização extensíveis (North e Shneiderman, 1999).

O modelo conceitual do Snap é baseado no modelo de dados relacional e suas informações são armazenadas em um banco de dados relacional. Na visualização, uma instância é representada por

uma tupla, isto é, um registro no banco de dados relacional. Portanto, a visualização é composta por um conjunto de registros carregados de uma base de dados por meio de uma consulta (e.g., SQL). As visualizações apresentam uma tabela ou uma relação da base de dados. Os atributos das instâncias são mapeados por meio de atributos visuais. A coordenação dos registros de diferentes visões é baseada na junção relacional. Ações do usuário em uma visão iniciam ações nos registros associados por meio da junção relacional. A forma como uma junção relacional é mostrada em outra visão depende do tipo de coordenação escolhido.

Os conceitos de visualização e interface do Snap estão diretamente ligados com os conceitos de um modelo relacional (North e Shneiderman, 2000b). A seguir são listadas as relações entre a visualização e o modelo relacional.

- **Visualização:** relação de dados. Toda visualização representa uma tabela ou uma relação da base de dados.
- **Item da visualização:** registro da base de dados. Cada registro (tupla) é descrito como um item na visualização.
- **Identificador do Item:** chave primária. Os registros são identificados unicamente por meio das chaves primárias.
- **Atributos visuais:** propriedades visuais. Cada atributo visual (e.g. cor, tamanho, posição) pode mapear um atributo dos dados.
- **Coordenação:** junção relacional. Ações do usuário sobre os itens de visualização disparam ações nos registros associados pela junção relacional, a qual é exibida em outra visualização.

Um exemplo de visualização utilizando o Snap para explorar um sistema de arquivos é mostrado na Figura 3.20. No canto superior direito é utilizada uma visualização em árvore que representa os diretórios do sistema de arquivos. Ao clicar em um dos diretórios as informações dos seus arquivos são carregadas na tabela exibida na parte inferior da figura. Também é utilizada uma *Treemap* para exibir a estrutura de diretórios. Nessa representação o tamanho do retângulo mapeia o atributo de tamanho do arquivo e a cor mapeia o atributo referente à extensão do arquivo.

A relação entre as visualizações é construída por meio de uma interface gráfica baseada em grafos, na qual os nós representam as técnicas de visualização e as arestas representam o tipo de coordenação que será empregada, como ilustrado na Figura 3.21. Os quadrados identificados por letras representam os tipos de coordenação disponíveis para cada técnica de visualização; S (verde) indica seleção, S (amarelo) – *scroll*, L – carregar (*load*) e Z – *zoom*. O esquema descrito indica que quando houver a seleção de um item na visualização em árvore do sistema de arquivos, ocorrerá na visualização em tabela o carregamento das informações correspondentes ao item selecionado. Além disso, ocorrerá na visualização *Treemap* o destaque do mesmo registro.

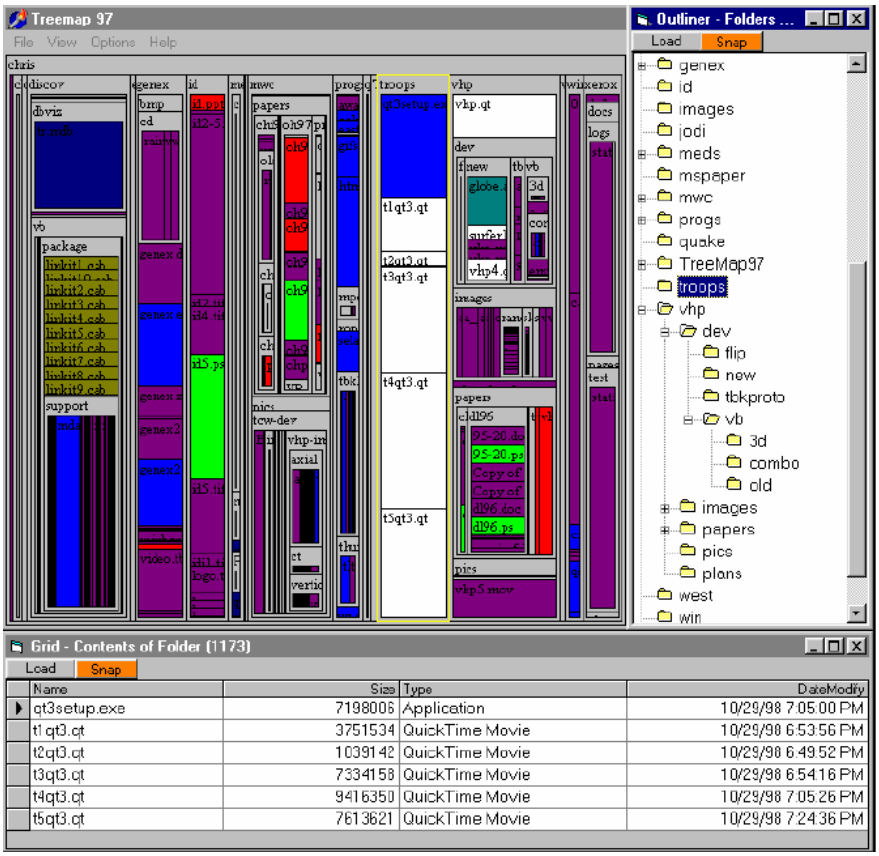


Figura 3.20: Três visões coordenadas para exploração de um sistema de arquivos (Extraído de (North e Shneiderman, 1999))

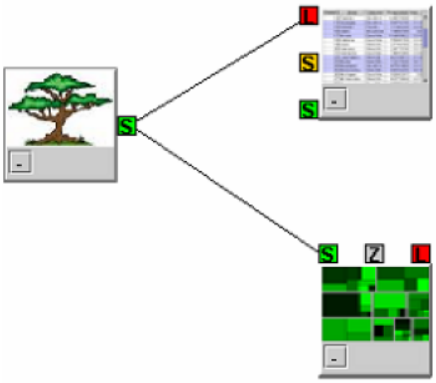


Figura 3.21: Especificação da coordenação da visualização da Figura 3.20 (Extraído de (Moreno, 2006))

Modelo Proposto por Boukhelifa e Rodgers

Outras tentativas em criar modelos gerais de visualizações coordenadas são relatados na literatura. O modelo proposto por Boukhelifa e Rodgers (Boukhelifa e Rodgers, 2003) manipula múltiplos aspectos em diferentes estágios do processo de visualização (e.g., pré-processamento e mapeamento dos dados). Objetos de coordenação implementam diferentes tipos de coordenação e funções de tradução são implementadas para as visões se comunicarem com os objetos de coordenação.

Naquele trabalho, é apresentado um estudo que procurou analisar como outras disciplinas tratavam questões de interdependência da informação. O estudo interdisciplinar de coordenação apresentou idéias que podem ser usadas para modelar múltiplas visões e entidades, tais como objetos, eventos, processos, funções, agentes e ontologias. Tudo pode ser coordenado (Boukhelifa et al., 2003). Alguns elementos que o modelo geral considera são listados a seguir:

- **Entidades de coordenação:** detalha os itens que estão atualmente sendo coordenados, tais como, aspectos da janela atual, visão, dados, registros, tempo, entre outros;
- **Tipo:** determina o método pelo qual as entidades de coordenação são coordenadas. O método de coordenação pode ser influenciado pelo tipo de parâmetro (inteiro, real ou estruturado) e também a direcionalidade da coordenação (unidirecional ou bidirecional);
- **Cronologia (tempo de vida e agendamento):** por quanto tempo as entidades são coordenadas é regido por seu tempo de vida. Elas podem ser coordenadas permanentemente, por uma dada ação, ou determinada por um escopo;
- **Escopo:** determina conexões globais e locais e o tempo de vida dos links. O escopo define grupos de visualizações coordenadas e sempre que uma nova visualização é adicionada ao grupo ela já estaria automaticamente coordenada;
- **Granularidade dos links:** muitas entidades podem ser conectadas por vários links. A granularidade determina o número de entidades em uma coordenação (2..n), o número de visões em uma coordenação (1..n), e o número de links que uma entidade contribui para a coordenação (0..n);
- **Inicialização:** determina como a coordenação é criada. Pode ser automática, especificada pelo usuário, ou especificada pelo grupo;
- **Atualização:** determina o tempo em que as visualizações serão atualizadas após uma ação de coordenação. Normalmente todas as visualizações são atualizadas em tempo real, mas pode-se admitir que algumas visualizações reflitam um estado anterior do processo;
- **Concepção (controle do usuário):** determina como o usuário especifica as coordenações e como esquema é apresentado (e.g. layout gráfico).

Os elementos descritos devem ser abordados pelo modelo, o qual não deve impor nenhuma restrição quanto ao conjunto de dados, navegação ou paradigma de comunicação, viabilizando um ambiente flexível, adaptável e extensível, que permita especificar formalmente as coordenações.

O modelo define **objetos de coordenação** (*coordination object*) que gerenciam combinações de entidades que controlam aspectos das visões coordenadas. Um único objeto de coordenação é associado com cada tipo de coordenação no sistema. Uma visualização é dita coordenada se ela

compartilha um objeto de coordenação comum. Todos os objetos de coordenação em um sistema de visualização são mantidos em um **espaço de visualização** (*coordination space*).

A Figura 3.22 ilustra duas visões coordenadas (V_1 e V_2) e dois objetos de coordenação, representando duas coordenações distintas. As visões que estão coordenadas necessitam definir uma função de tradução do objeto de coordenação para os parâmetros da visão coordenada. Por exemplo, considerando o objeto de coordenação 1, as funções de tradução são $f_{1,1}$ e $f_{2,1}$. As visões devem também ser informadas por um evento de notificação ($Notify_{x,y}$) quando um objeto de coordenação é modificado. Se um evento ocorre, ele altera o objeto de coordenação, o qual envia uma notificação para todas as janelas que estão ligadas a ele. Assim, as visões que foram notificadas de uma mudança, por meio da função de tradução, utilizam a informação fornecida pelo objeto de coordenação para atualizar a visão.

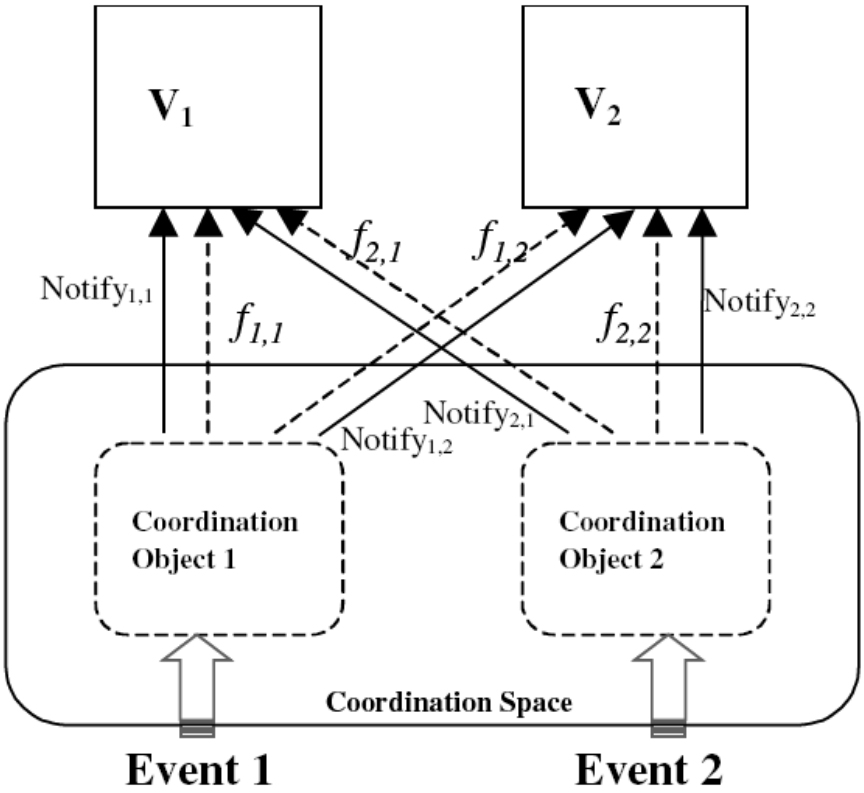


Figura 3.22: Modelo abstrato de coordenação. No caso, são mostrados dois tipos diferentes de coordenação e duas visões (Extraído de (Boukhelifa e Rodgers, 2003))

Esse modelo foi utilizado para implementar um sistema denominado *CViews*, no qual três tipos de visões e um tipo de coordenação foram implementados. As visões foram empregadas para representar dados textuais da base SUSANNE, a qual é uma base de taxonomia e anotações da gramática da língua inglesa. A visão *tabularView* exibe os dados no formato bruto; a visão *textView* exibe os dados originais das anotações; e a visão *concordanceView* exibe a ocorrência de uma palavra em diferentes contextos (anotações). A Figura 3.23 ilustra essas três visões e a coordenação entre elas. No exemplo foi selecionada a palavra “tax”. O tipo de coordenação

implementado foi o *brushing*, por meio do qual uma palavra selecionada em uma visão é destacada nas outras visões.

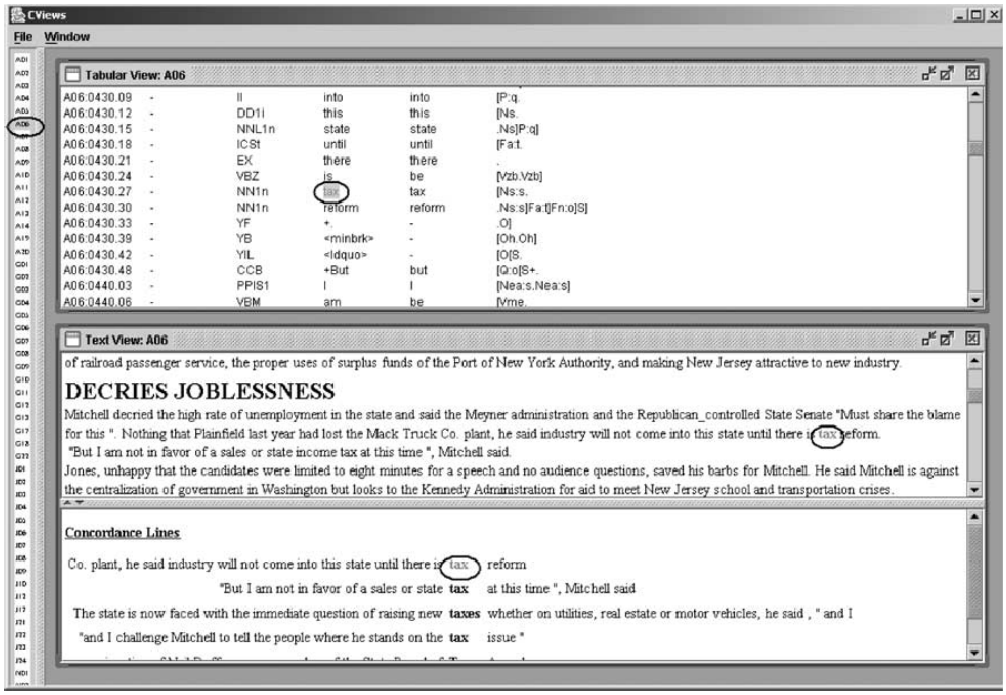


Figura 3.23: Sistema *CViews* mostrando três visões coordenadas. No caso, foi selecionada a palavra "tax" (Extraído de (Boukhelifa e Rodgers, 2003)).

Modelo Proposto por Weaver

Um modelo de coordenação mais recente foi proposto por Weaver (Weaver, 2006). Esse modelo foi desenvolvido para melhorar a flexibilidade do sistema *Devise* (Livny et al., 1997), criando um novo sistema chamado *Improvise* (Weaver, 2004). Esse modelo é similar ao *Snap-Together Visualization* e outros sistemas de visualização baseados no modelo de dados relacional. Ele combina um mecanismo de coordenação direta chamada de *Live Properties*, pela qual o usuário interage por meio de visões, barras de comando e outros controles da visualização; com o mecanismo de coordenação indireta chamado *Coordinated Queries*, que é uma linguagem de abstração visual. O resultado é um sistema no qual os usuários podem interativamente criar, destruir, coordenar e especificar os conteúdos das visões.

As *Live Properties* (LP) são a primeira característica importante do modelo de Weaver. Elas são definidas como um mecanismo simples de coordenação que utiliza um objeto de coordenação compartilhado. As LPs definem uma interface para coordenar controles da visualização por meio de objetos compartilhados chamados de variáveis. Cada controle define uma ou mais LPs, cada uma delas pode ser associada a uma variável de coordenação, as quais são utilizadas para compartilhar informações entre as diversas visões. As LP podem ser ativas (acessam e modificam variáveis) ou passivas (apenas acessam as variáveis). Quando ocorre a modificação em uma variável, estas alterações são propagadas aos controles pelas suas LPs, como ilustrado na Figura 3.24.

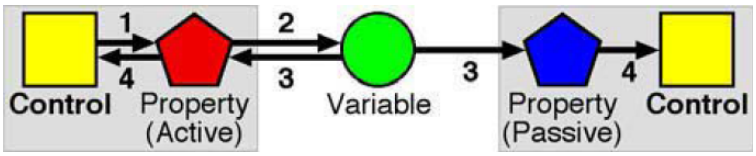


Figura 3.24: Coordenação direta (Extraído de (Weaver, 2004))

A Figura 3.24 ilustra o funcionamento da coordenação direta entre dois controles. Inicialmente, um controle modifica o valor de uma de suas LPs devido a alguma interação (1). A partir de sua alteração, a LP repassa seu novo valor a uma variável de coordenação (2), a qual envia um sinal a todas as LPs que estão relacionadas a ela para informar o seu novo valor (3). Por fim, as LPs avisam os seus controles para realizarem as mudanças ocorridas (4).

As LPs servem para dois propósitos. Primeiro, elas são compartimentos de valores que os controles usam para determinar sua aparência e comportamento. Por exemplo, um gráfico de dispersão tem duas LPs que especificam qual região do plano cartesiano exibir, e uma LP que especifica a cor usada para preencher o plano de fundo. Segundo, elas agem como portas pelas quais os controles comunicam entre si como resultado de uma interação.

A segunda característica importante do modelo de Weaver é as *Coordinated Queries*, as quais são uma linguagem de abstração visual baseada no modelo de dados relacional. Uma expressão é uma árvore de operadores que calculam o valor de um atributo de saída usando os atributos de um registro de entrada. As expressões fazem operações de consulta que as visões usam para codificar os registros de dados em atributos gráficos. Existem dois tipos de expressões:

- **Filtros:** usam uma única expressão para calcular um valor booleano para cada registro de entrada. Apenas os registros cuja expressão resulta em verdadeiro são processados pelas visões;
- **Projeções:** usam uma ou mais expressões para calcular atributos sucessivos de um registro de saída para cada registro de entrada. Visões desenham os registros usando informações gráficas (como cor e posicionamento) contidas nos atributos de um registro de saída.

Visualizações construídas com *coordinated queries* consistem de quatro tipos de componentes: controles, variáveis, *live properties* e *lexicals*. *Lexicals* são pacotes para conjuntos de dados externos, operações de busca (projeções, filtros e ordenações, e conjuntos de dados derivados). *Lexicals* podem ser atribuídos como valores de variáveis, como qualquer outro parâmetro interativo. Visões acessam e processam dados por meio de propriedades do tipo *lexical*.

Em cada visualização no Improvise, o *lexicon* é um repositório central para gerenciamento de conjunto de dados e operações de busca. O *lexicon* contém um *lexical* para cada dado, informação, projeção, filtro e ordenação na visualização. O *Lexical* pode ser criado, excluído e editado interativamente. Um *lexical* permite que cada conjunto de dados e operação de busca sejam reusados como um valor associado a múltiplas variáveis simultaneamente.

A Figura 3.25 mostra um esquema de coordenação indireta por meio das *Coordinated Queries*. Quando um controle depende de um valor de alguma variável que resulta de uma expressão, ele depende indiretamente das variáveis contidas dentro da expressão. Toda vez que um objeto altera o valor de uma das variáveis (1), ela notifica os valores léxicos das expressões que são calculados com base nos valores daquela variável (2). Cada expressão alterada notifica as variáveis que estão relacionadas com seu valor final (3). Por fim, a variável notifica os controles aos quais ela está associada para realizarem as novas alterações (4).

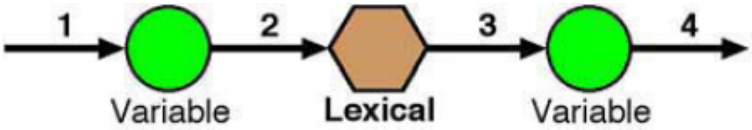


Figura 3.25: Coordenação indireta (Extraído de (Weaver, 2004))

3.4 Considerações Finais

Observando a área de Múltiplas Visões Coordenadas (CMV), pode-se notar um grande crescimento da sua aplicação em diversas áreas de pesquisa. Isto é constatado pela diversidade de aplicações apresentadas nas principais conferências em visualização. Boa parte dos trabalhos procura aplicar CMV em uma área de conhecimento específica (Bertini et al., 2004; Keller et al., 2005; Chastine et al., 2005; Craig et al., 2005; Matkovic et al., 2005; Lawrence et al., 2006), tentando melhorar a investigação em tal área. Também, muitos trabalhos procuram melhorar o processo de investigação unindo diferentes técnicas de visualização ou melhorando técnicas já existentes (Convertino et al., 2003; Piringer et al., 2004; Schafhitzel et al., 2005; Bertini et al., 2005; Ericson et al., 2005; Yang et al., 2006; Lamirel e Shehabi, 2006).

Na tentativa de construir sistemas CMV de propósito geral, modelos conceituais foram elaborados. Teoricamente, eles tentam permitir a elaboração de qualquer tipo de coordenação entre diferentes visões (North e Shneiderman, 2000a; Boukhelifa e Rodgers, 2003; Weaver, 2006), mas na prática eles são de difícil aplicação, pois aumentam a carga operacional da construção do sistema. Talvez esse seja o provável motivo das diversas aplicações apresentadas na literatura não criarem novas técnicas de coordenação, pois criam um modelo próprio para atender à necessidade básica de uma simples coordenação *brushing-and-linking* para destacar as instâncias que tenham o mesmo identificador.

Apesar das diversas áreas de aplicação, sistemas e modelos, poucos trabalhos utilizam múltiplas visões coordenadas para auxiliar a exploração de conjuntos de dados não estruturados, como coleções de documentos (Jern et al., 2003; Seeling e Becks, 2004b; Huang et al., 2005; Hoeber e Yang, 2006). As poucas abordagens que utilizam CMV para exploração desse tipo de conjunto de dados, geralmente manipulam somente um conjunto por vez. Além disso, não utilizam a coordenação das informações adquiridas em um conjunto para auxiliar a exploração de outros.

Um outro tipo de conjunto de dados que não é explorado pela abordagem de múltiplas visões coordenadas é a coleção de imagens. É difícil encontrar aplicações que auxiliem a exploração de coleções de imagens e a interpretação do espaço de características por meio de múltiplas visões coordenadas. Do ponto de vista da visualização existe uma série de requisitos que os métodos de coordenação existentes não tratam, ou uma sobrecarga de programação para que tratem, dificultando o desenvolvimento de técnicas para auxiliar a exploração desse tipo de conjunto de dados.

Este trabalho de doutorado continuou as pesquisas do grupo de visualização do ICMC-USP, expandindo as técnicas de múltiplas visões coordenadas já empregadas na exploração de conjuntos de documentos. Especificamente, foram propostas novas técnicas de coordenação, permitindo que vários conjuntos de dados sejam explorados simultaneamente. Por meio da coordenação, as descobertas realizadas em uma visão auxiliam a exploração entre as demais visões. A descrição detalhada das técnicas de coordenação desenvolvidas e exemplos de aplicação são apresentados Capítulo 5. Essa técnicas podem, em sua maioria, também ser utilizadas na exploração de mapas de similaridades construídos a partir de outros tipos de dados. Aplicações detalhadas das técnicas de coordenação são apresentadas no Capítulo 6.

Dos modelos de coordenação existentes na literatura, o mais geral e o que melhor se adequa às técnicas de coordenação propostas neste trabalho é aquele proposto por Boukhelifa e Rodgers (Boukhelifa e Rodgers, 2003), pois ele contempla todas as etapas e elementos do processo de exploração de múltiplas visões coordenadas. Neste trabalho, esse modelo foi adaptado para que as técnicas de coordenação propostas fossem desenvolvidas e aplicadas. A principal adaptação foi considerar técnicas de coordenação para criar o mapeamento entre as instâncias de diferentes visões, ao invés de tipos de coordenação. Essa adaptação possibilitou que as técnicas de coordenação fossem definidas durante o processo de exploração e que diferentes tipos de mapeamentos pudessem ser criados, os quais podem ser estáticos ou dinâmicos. Os modelos de coordenação baseados no modelo de dados relacional não foram considerados para serem utilizados ou estendidos, pois registros não seria a maneira ideal para representar os elementos de dados extraídos de coleções de documentos e o seus relacionamentos, conforme observado por Weaver (Weaver, 2006). O modelo adaptado é descrito no próximo capítulo.

CAPÍTULO

4

O Modelo de Coordenação

4.1 Considerações Iniciais

Sistemas de visualização que utilizam múltiplas visões coordenadas estão apoiados em algum modelo de coordenação responsável por gerenciar as técnicas de coordenação e por definir um padrão para futuras extensões. Neste trabalho, o sistema de visualização utiliza uma adaptação criada a partir do modelo proposto por Boukhelifa e Rodgers (Boukhelifa e Rodgers, 2003). Este capítulo visa detalhar a estrutura e funcionamento do modelo de coordenação desenvolvido. Em seguida são apresentadas as instanciações para cada uma das técnicas de coordenação propostas neste trabalho e também como um desenvolvedor deve proceder para adicionar uma nova técnica de coordenação ao modelo. Por fim, são apresentadas as conclusões sobre o modelo de coordenação utilizado.

4.2 Descrição do Modelo de Coordenação

Como apresentado na Seção 3.3, o modelo proposto por Boukhelifa e Rodgers (Boukhelifa e Rodgers, 2003) é um modelo geral de coordenação. Apesar disso, é considerado que cada tipo de coordenação (e.g., *brushing-and-linking*) seja representado por um único objeto de coordenação. Baseado nesse modelo os autores desenvolveram o sistema *CViews*, que apresentou uma inicialização estática das ligações entre os objetos de coordenação e as visões, não permitindo que novas coordenações sejam adicionadas ou configuradas durante o processo de exploração. Essa característica do sistema não permite a criação de diferentes tipos de mapeamentos. Esse modelo foi

utilizado como base para criar um novo modelo responsável por definir e gerenciar as técnicas de coordenação propostas neste trabalho. Neste contexto, técnicas de coordenação são responsáveis por criar os mapeamentos entre as visões, os quais são destacados por algum tipo de coordenação. Por exemplo, o *brushing-and-linking* realçaria as instâncias mapeadas pela técnica de coordenação. Um outro tipo de coordenação poderia ser empregado para, por exemplo, filtrar os dados ou efetuar *zoom* sobre a visualização. Para atender às necessidades dessas técnicas, a primeira adaptação foi não considerar tipos de coordenação, mas sim técnicas de coordenação para construção de mapeamentos. Assim, ao invés de um tipo de coordenação compartilhado entre todas as visões, é definida uma técnica de coordenação para gerar um mapeamento particular entre as instâncias das visões envolvidas. Conforme apresentado no Capítulo 5, cada técnica de coordenação possui uma forma particular de gerar o mapeamento entre as instâncias. Dessa maneira, novos mapeamentos podem ser criados com diferentes configurações durante a etapa de exploração. Mapeamento, neste contexto, é a relação criada entre as instâncias das visões envolvidas. Essa relação é criada por meio de uma técnica de coordenação, a qual mapeia, nas visões destino, uma ou mais instâncias relacionadas com cada instância selecionada na visão origem. Essa flexibilidade na criação dos mapeamentos exigiu a adição de um gerenciador de técnicas de coordenação (denominado *Coordinator*) para controlar a execução das técnicas que gerarão e aplicarão os mapeamentos entre as visões. A utilização de um gerenciador também seria um mecanismo usado por Boukhelifa e Rodgers para o caso de uma evolução do sistema *CViews* (Boukhelifa e Rodgers, 2003).

A próxima seção apresenta o processo exploratório do sistema de visualização, identificando os estados que envolvem as técnicas de coordenação.

4.2.1 Processo de Exploração

O processo de exploração empregado neste trabalho é apoiado por diferentes técnicas de coordenação, as quais geram diferentes tipos de mapeamento. Esse processo pode ser abstraido pelo diagrama de estados mostrado na Figura 4.1, o qual inclui três estados principais:

- **Criação da Visão (*View Creation*):** novas visões são criadas e adicionadas ao processo de exploração. Essa é a visão dos dados, ou seja, uma visualização em particular;
- **Criação da Coordenação (*Coordination Creation*):** novas técnicas de coordenação são definidas e adicionadas ao processo de exploração;
- **Exploração (*Exploration*):** ocorre a interação com as visões disponíveis, explorando os dados por meio das técnicas de visualização, técnicas de interação e técnicas de coordenação definidas.

O diagrama mostra um estado inicial (**Idle**) antes de qualquer interação. Uma vez que uma nova visão é criada, inicia-se o processo de exploração (**Exploration**), no qual ele pode navegar

entre as visões já criadas, apoiado por funcionalidades de interação. Durante o processo de exploração novas visões podem ser criadas (**View Creation**) para oferecer perspectivas alternativas sobre os dados. Além de criar novas visões, também é possível definir técnicas de coordenação para gerar um mapeamento entre as visões (**Coordination Creation**). Nessa etapa de definição de uma técnica de coordenação, são especificadas as visões envolvidas e, se necessário, os parâmetros de coordenação para gerar o mapeamento entre elas. Assim, o processo de exploração é enriquecido com a abordagem de múltiplas visões coordenadas.

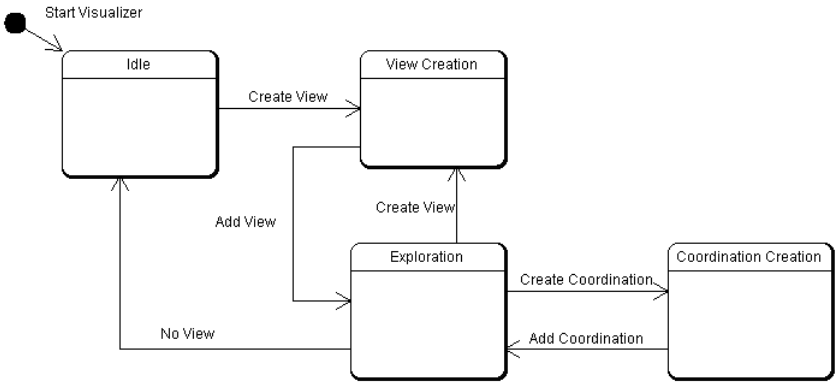


Figura 4.1: Estados do processo exploratório.

No Capítulo 5 foram apresentadas as diferentes técnicas de coordenação que podem ser utilizadas no processo de exploração ilustrado na Figura 4.1. Os exemplos deram uma idéia geral do funcionamento das técnicas de coordenações propostas, as quais são definidas entre duas visões, visão origem (**sourceView**) e visão destino (**targetView**), e o tipo de coordenação empregado é o *brushing-and-linking*. Vale lembrar que entre as etapas de *brushing-and-linking* cada técnica de coordenação tem sua forma particular de operar e gerar o mapeamento final. O mapeamento é realizado da visão origem (onde iniciou a seleção) para a visão destino (onde as instâncias mapeadas serão destacadas).

A seguir é descrito o modelo de coordenação que define e gerencia as técnicas propostas neste trabalho e outras futuras que atendam os mesmos quesitos. A descrição é feita considerando as classes envolvidas e o seu comportamento durante o processo de exploração. Primeiramente é descrita a estrutura de classes que define o modelo. Em seguida, são descritos os passos envolvidos na definição, criação e execução das técnicas de coordenação nos três principais estados de navegação do processo de exploração (Figura 4.1).

4.2.2 Descrição das Classes

A estrutura do modelo de coordenação utilizado neste trabalho pode ser abstraída pelo conjunto de classes ilustrado na Figura 4.2. A classe **Visualizer** representa o sistema de visualização, que compreende um conjunto de visões (**viewers**) e um gerenciador (**coordinator**), que é uma instância da classe **Coordinator**. A classe **Coordinator** é responsável por ecoar todos os eventos

de coordenação ativados pelas interações. Para isso, a classe **Coordinator** possui uma lista de todas as técnicas de coordenação definidas (**viewsCoordList**) entre as visões do sistema. A classe **View** representa uma visão do sistema. Ela inclui uma lista de coordenações (**coordination**) e um modelo de dados (**dataModel**), o qual armazena as informações do conjunto de dados utilizado. O método **highlightSelected** é responsável por destacar as instâncias indicadas, por meio do qual cada visão define sua maneira particular de destacar essas instâncias. A classe **Visualizer** e a classe **View** têm muitos outros atributos e métodos, mas aqui são destacados somente aqueles que interessam ao processo de coordenação. A classe **CoordinationList** é responsável por armazenar todas as técnicas de coordenação definidas para uma visão, isto é, ela representa o relacionamento entre uma visão e outra por meio de uma técnica de coordenação. A classe **Coordination** é a classe que representa uma técnica de coordenação, a qual é responsável por criar e aplicar um mapeamento entre as instâncias de duas visões (da **sourceView** para a **targetView**).

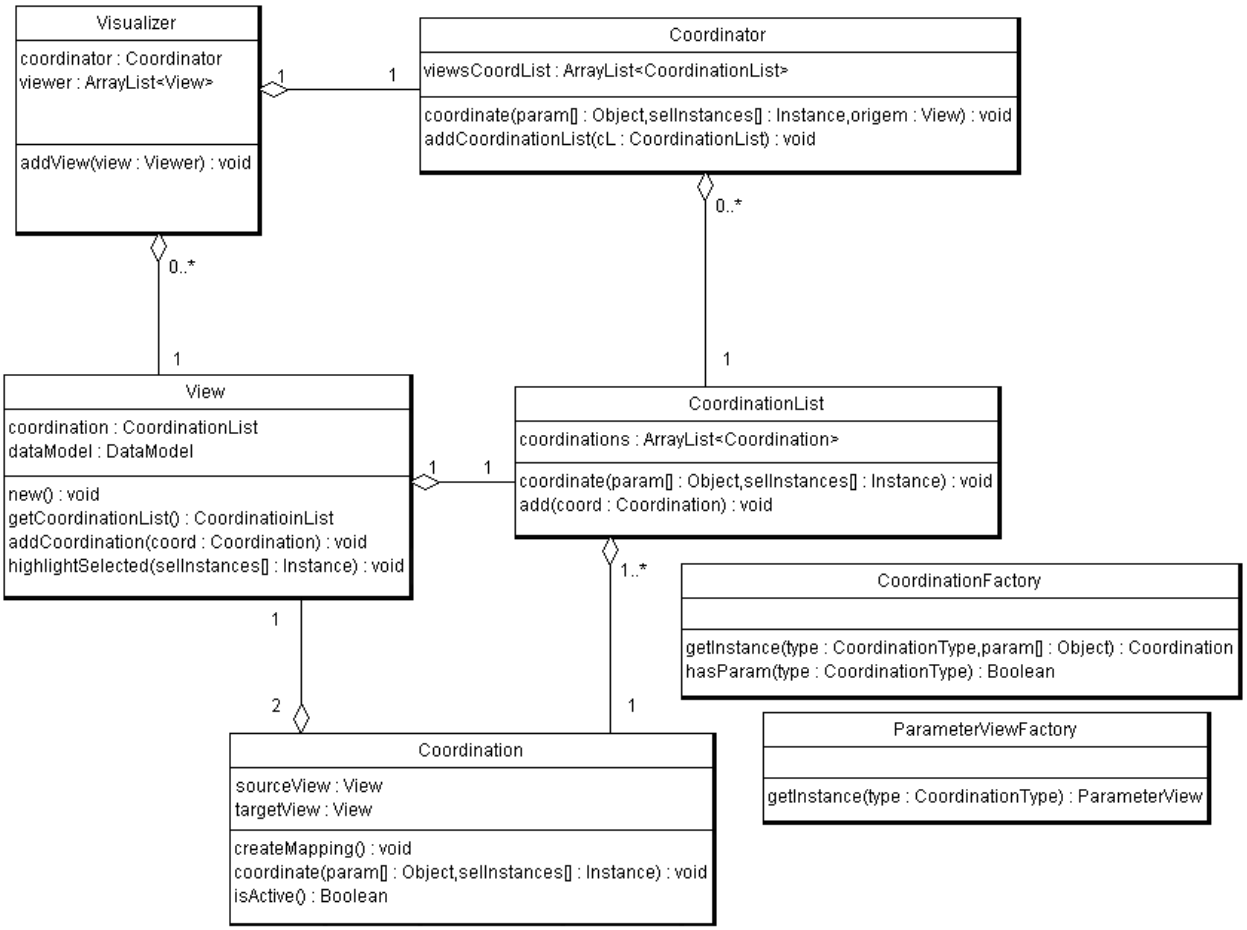


Figura 4.2: Classes do modelo de coordenação.

O mapeamento entre duas visões pode ser estático ou dinâmico. O primeiro ocorre quando o mapeamento é sempre o mesmo em todo o processo de exploração. Um exemplo é a coordenação baseada no identificador das instâncias de um conjunto de dados. O mapeamento dinâmico ocorre quando os elementos utilizados para sua geração podem mudar de acordo com a interação. Um exemplo é a coordenação baseada em tópicos gerados a partir da seleção de um grupo documentos.

O grupo de documentos pode mudar de acordo com a seleção efetuada, influenciando no tópico gerado e, consequentemente, no mapeamento.

Um mapeamento estático pode ser gerado na definição de uma técnica de coordenação, por meio do método **createMapping** da classe **Coordination**. Esse mapeamento é aplicado na invocação do método **coordinate** da classe **Coordination**. Para o caso de um mapeamento dinâmico, geralmente não há a criação de mapemanto na definição da técnica de coordenação, mas sim em sua execução, isto é, na invocação do método **coordinate** da classe **Coordination**. Nessa etapa, um pré-processamento é executado para criar o mapeamento entre as instâncias das visões de acordo com os elementos dinâmicos fornecidos (e.g., tópicos). De modo geral, a partir de uma seleção na visão origem (**sourceView**), o mapeamento de uma técnica de coordenação somente será executado se ela estiver ativa em alguma visão destino (**targetView**), seja ele estático ou dinâmico.

Algumas técnicas de coordenação necessitam de parâmetros em sua definição. Esses parâmetros são utilizados para criar o mapeamento entre as instâncias das visões envolvidas no processo de coordenação. Um exemplo é a técnica baseada em distância (*Distance Coordination*), a qual exige como parâmetros o número de vizinhos mais próximos que serão mapeados e também a métrica de distância que será utilizada para calcular os vizinhos mais próximos. Para isso o desenvolvedor deve criar uma interface para entrada de dados, a qual é executada antes da criação de uma técnica de coordenação.

Para facilitar e padronizar a instanciação das interfaces para entrada de parâmetros (classe **ParameterViewFactory**) e também a instanciação das técnicas de coordenação (classe **CoordinationFactory**), foi utilizado o padrão de projeto *Factory* (Gamma et al., 1995). Esse padrão define uma interface para criar um objeto da classe indicada no parâmetro **type**. No caso da **ParameterViewFactory** é necessário passar somente o tipo de técnica de coordenação para que a interface de entrada de parâmetros seja criada. No caso da **CoordinationFactory** é necessário passar o tipo de técnica de coordenação e os parâmetros fornecidos para que uma nova técnica de coordenação seja criada. A lista de parâmetros pode ser vazia, caso a técnica de coordenação não necessite de parâmetros para criar um mapeamento.

A seguir é descrito o comportamento do modelo de coordenação, apresentando o que ocorre na criação de uma visão, na criação de uma técnica de coordenação e na execução de um evento de coordenação gerado em alguma visão. Também serão descritos os passos de execução de um mapeamento estático e de um mapeamento dinâmico. As descrições focarão somente nos aspectos que envolvem o processo de coordenação.

4.2.3 Descrição do Processo de Coordenação

O comportamento do processo de coordenação é detalhado a seguir para os três estados principais do processo de exploração descrito na Figura 4.1. São eles: *Create View*, *Create Coordination* e *Exploration*. Como comentado na Seção 4.2.1, para iniciar o processo de exploração, é necessário criar uma visão que represente os dados por meio de uma técnica de visualização. Essa etapa

é apresentada no processo *Create View*, no qual também são executados passos para armazenar as técnicas de coordenação na *Coordination List*. Os passos desse processo são ilustrados na Figura 4.3. Inicialmente (**passo 1**) uma nova visão e uma lista de técnicas de coordenação (**passo 1.1**) são criadas. Em seguida, essa lista de técnicas (**CoordinationList**) é adicionada ao **Coordinator** (**passo 2**) para que ele possa gerenciar as técnicas de coordenação relacionadas com a visão criada.

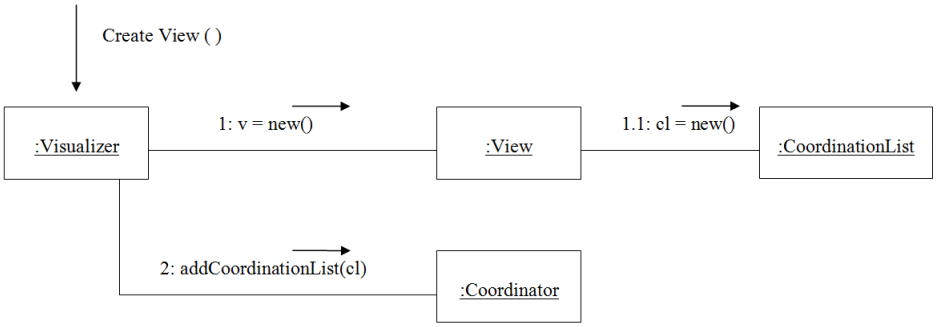


Figura 4.3: Processo *Create View*: passos executados quando uma nova visão é criada.

Em outro processo, denominado *Create Coordination*, são descritas a definição e a criação de uma técnica de coordenação. Seus passos são ilustrados na Figura 4.4. Para criar uma nova coordenação, três parâmetros são necessários: **Type**, que é o tipo de técnica de coordenação escolhida; **sourceView**, que é a visão onde as interações irão ocorrer para disparar o processo de coordenação; **targetView**, que é a visão onde o mapeamento será aplicado, ou seja, onde serão destacadas as instâncias mapeadas a partir das instâncias selecionadas na **sourceView**. Esse processo inicia verificando se a técnica de coordenação requer algum parâmetro (**passo 1**). Se não requer, uma nova técnica de coordenação é criada (**passo 2**) com nenhum parâmetro; caso contrário, uma janela para definição de parâmetros é apresentada (**passo 3**) para que sejam definidos os parâmetros relevantes (**passo 3.1**) e a técnica de coordenação seja criada (**passo 3.2**). Finalmente, a técnica de coordenação criada é adicionada à lista de técnicas de coordenação da **targetView** (**passo 4**). Note que a técnica de coordenação que criará o mapeamento entre duas visões está armazenada na lista de técnicas de coordenação da visão destino (**targetView**), pois ela é a visão que recebe o mapeamento de uma técnica de coordenação e é nela que se especifica qual técnica de coordenação está ativa para efetuar o mapeamento.

O processo *Coordinate* descreve a execução de uma técnica de coordenação entre as visões envolvidas no processo de exploração, conforme ilustrado na Figura 4.5. Esse processo ocorre quando um evento de coordenação é gerado em alguma visão. Para a execução desse processo, dois conjuntos de parâmetros são requeridos: **Param**, contendo os parâmetros da técnica de coordenação (pode ser vazio), e **selectedInstances**, contendo as instâncias selecionadas na visão origem (**sourceView**). Primeiramente, o **Coordinator** obtem a lista de técnicas coordenação de cada visão (**passo 1**), que está armazenada no objeto **viewsCoordList**, e invoca o método **coordinate** de cada uma delas (**passo 1.2**). Cada invocação do método **coordinate** faz com que seja realizada uma busca na lista de técnicas de coordenação da visão (**passo 1.2.1**). Durante essa busca

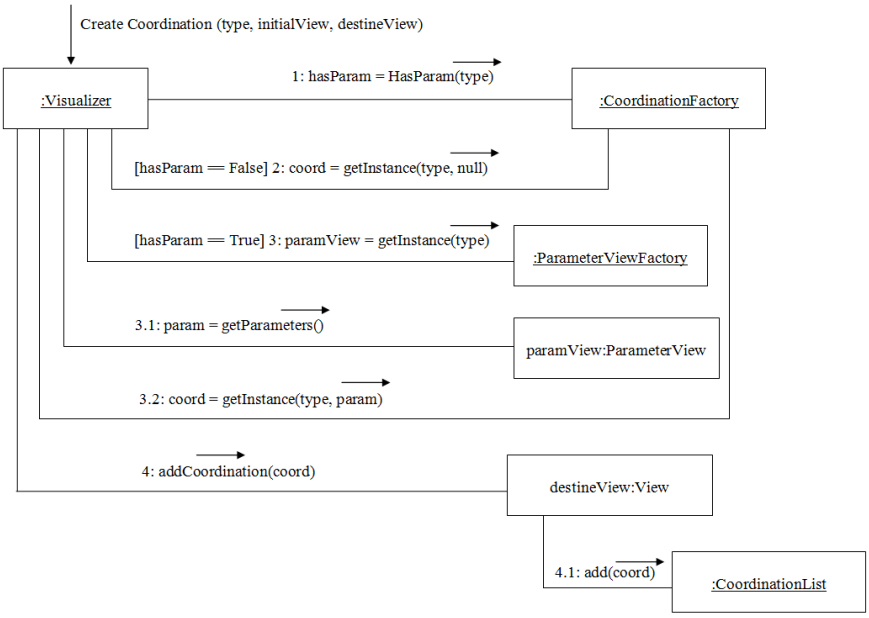


Figura 4.4: Processo *Create Coordination*: passos executados quando uma nova técnica de coordenação é criada.

é verificada qual é a técnica de coordenação que está ativa em cada visão destino (**targetView**), isto é, qual técnica de coordenação está selecionada para efetuar o mapeamento na visão. Assim, a técnica que estiver ativa invocará o método **coordinate** (**1.2.2**), fazendo com que o mapeamento seja aplicado na visão destino. É importante notar que uma lista de coordenação de uma visão pode armazenar muitas técnicas de coordenação, mas somente uma das técnicas definidas pode ser escolhida para aplicar o mapeamento. Este é o motivo da verificação (**passo 1.2.2**).

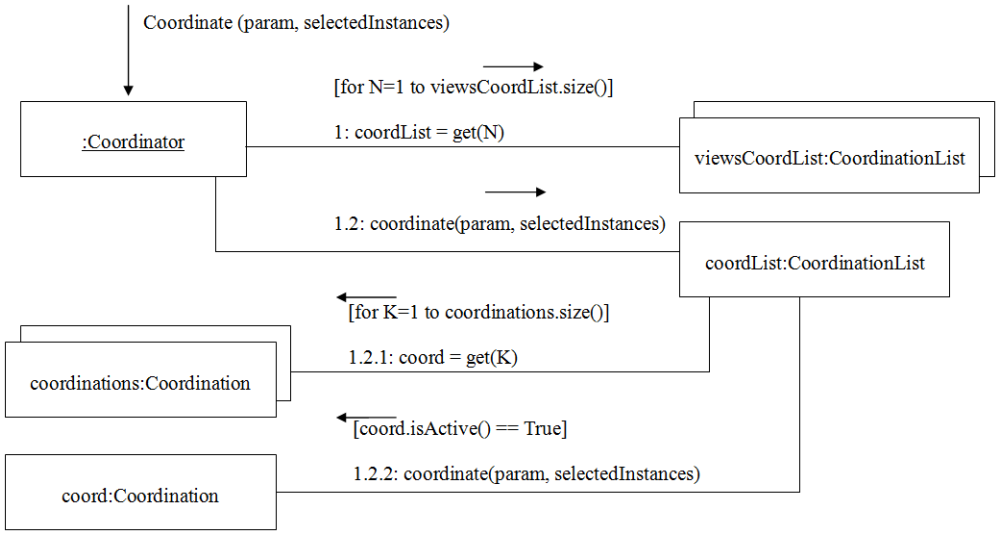


Figura 4.5: Processo *Coordinate*: passos executados durante a coordenação entre as visões.

Dependendo da técnica de coordenação (**Coordination**), o método **coordinate** pode ser executado de uma das duas formas ilustradas na Figura 4.6. Como mencionado na Seção 4.2.2, o mapeamento é criado logo que a nova coordenação é definida. No caso do mapeamento dinâ-

mico, o mapeamento entre as visões é somente definido durante o processo de coordenação, pois é dependente da interação. Assim, em geral, nenhum mapeamento prévio é criado quando a nova coordenação dinâmica é definida. No caso do mapeamento dinâmico, o método **coordinate** recebe as instâncias selecionadas, processa os dados obtidos ou gerados das instâncias selecionadas, cria e em seguida aplica o mapeamento criado com base nos dados processados. Por fim, as instâncias mapeadas são transmitidas para a visão destino realizar o destaque. No caso do mapeamento estático, o método **coordinate** recebe as instâncias de dados selecionadas na visão origem, aplica o mapeamento previamente criado e transmite as instâncias mapeadas para serem destacadas na visão destino.

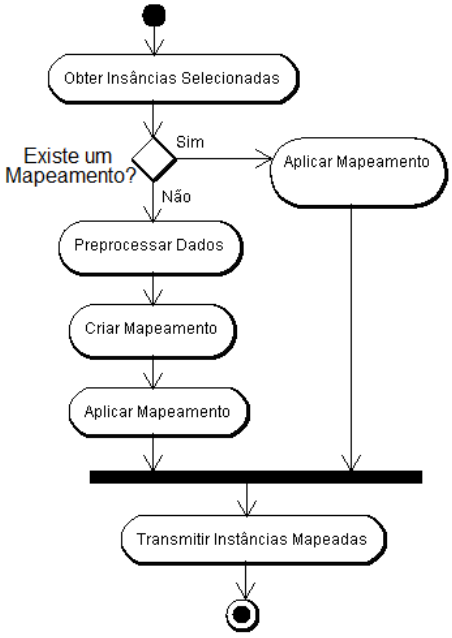


Figura 4.6: Descrição geral dos passos de execução do método *coordinate*.

A seguir são descritas as instanciações das técnicas de coordenação propostas neste trabalho.

4.3 Instanciações do Modelo de Coordenação

Com referência ao modelo descrito acima, nesta seção são apresentados detalhes de instanciação das técnicas de coordenação que são detalhadas no Capítulo 5. Conforme ilustrado na Figura 4.7, o modelo foi instanciado para descrever seis técnicas de coordenação: *Identity Coordination*, *Distance Coordination*, *Topic Coordination*, *Topic Time Coordination*, *Time Coordination* e *Joint Coordination*. O foco das descrições será os dois principais métodos utilizados para realizar o mapeamento entre as visões, pois são eles que diferenciam as seis técnicas de coordenação. Os dois principais métodos são o **createMapping**, o qual é responsável por criar um mapeamento estático; e o **coordinate**, o qual é responsável por aplicar um mapeamento ou, no caso de uma técnica com mapeamento dinâmico, por criar criar e em seguida aplicar o mapeamento (ver Figura 4.6 e explicação no fim da Seção 4.2.3).

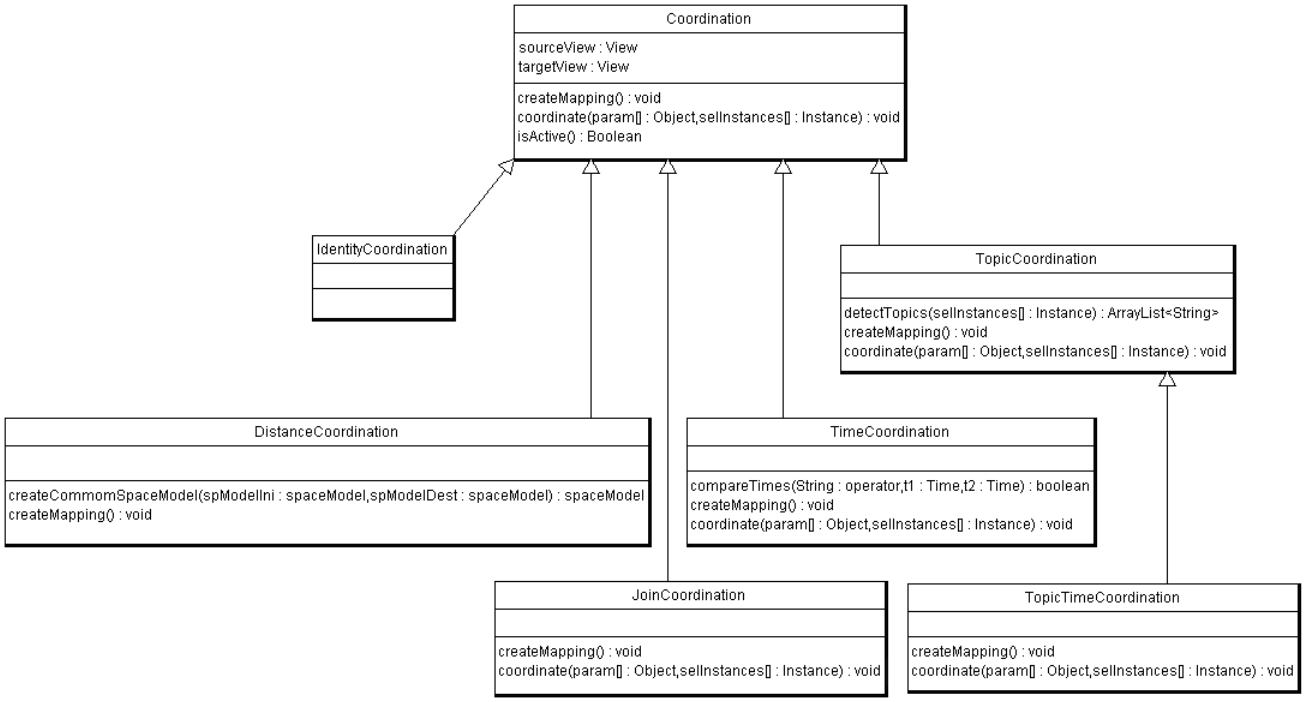


Figura 4.7: Instanciações das técnicas de coordenação utilizadas.

4.3.1 Identity Coordination

A *Identity Coordination* é uma técnica simples de coordenação que mapeia nas visões destino todas as instâncias que compartilham o mesmo identificador (ID) daquelas selecionadas na visão origem. O mapeamento é estático e pode ser gerado quando a coordenação é definida. Esta coordenação não requer nenhum parâmetro adicional além dos identificadores das instâncias. Os métodos **createMapping** e **coordinate** são descritos abaixo:

- **createMapping**
  - Para cada instância na visão inicial
    - Verificar se há uma ou múltiplas instâncias com o mesmo ID na visão destino
    - Se sim
      - Criar um mapeamento entre as instâncias que possuem o mesmo ID
- **coordinate**
  - Para cada instância selecionada na visão origem
    - Obter a lista de instâncias mapeadas na visão destino
  - Selecionar cada instância mapeada

4.3.2 Distance Coordination

O conceito de vizinhança é importante em mapas de similaridade, referindo aos objetos mais similares a um dado objeto. A *Distance Coordination* mapeia na visão destino as k instâncias mais próximas de cada instância selecionada na visão origem. De forma similar à *Identity Coordination*, esta coordenação por distância cria um mapeamento estático entre as visões origem e destino.

Os mapas de similaridade entre duas visões podem ser resultantes de diferentes conjuntos de dados ou de diferentes espaços multidimensionais. Portanto é necessário criar um novo espaço multidimensional que contenha somente os atributos comuns aos espaços utilizados nas visões envolvidas (ver Seção 5.3.5). Em seguida, para cada instância selecionada na visão origem, é criado um mapeamento para as k instâncias mais próximas na visão destino. Esta coordenação requer sejam definidos quantos vizinhos (k) serão mapeados entre as visões e também a métrica de distância utilizada para calcular a distância entre as instâncias. Os métodos **createMapping** e **coordinate** são descritos abaixo:

- **createMapping**
  - Obter o espaço multidimensional das visões origem e destino
    - Obter os atributos comuns entre os dois espaços multidimensionais
  - Construir um novo espaço multidimensional para cada visão, utilizando somente os atributos comuns
  - Para cada instância no novo espaço multidimensional da visão origem
    - Obter as k instâncias mais próximas no novo espaço multidimensional da visão destino
    - Criar um mapeamento entre estas k-instâncias
- **coordinate**
  - Para cada instância selecionada na visão origem
    - Obter a lista de instâncias mapeadas na visão destino
  - Selecionar cada instância mapeada

4.3.3 Topic Coordination

Ao contrário das duas técnicas de coordenação descritas acima, a *Topic Coordination* não pode criar um mapeamento estático entre duas visões, pois o relacionamento entre as instâncias é dado pelo tópico detectado quando uma seleção é efetuada sobre a visão origem. Portanto, o mapeamento é criado no método **coordinate**. Esta técnica de coordenação executa a detecção de um

tópico a partir de instâncias selecionadas na visão inicial. Em seguida, são mapeadas na visão destino as instâncias cobertas pelo tópico gerado. Na implementação atual é considerado o tópico cujos termos possuem o maior valor de covariância. Uma possível extensão é possibilitar que tópicos desejados sejam selecionados durante o processo de coordenação. Os métodos **createMapping** e **coordinate** são descritos abaixo:

- **createMapping**
  - Não há mapeamento para ser criado, pois os tópicos são gerados durante a interação. Portanto, não é possível criar um mapeamento prévio
- **coordinate**
  - Aplicar um algoritmo de detecção de tópicos sobre as instâncias selecionadas na visão origem e gerar uma lista de tópicos
  - Mapear na visão destino as instâncias que são cobertas pelo tópico principal
  - Selecionar cada instância mapeada

4.3.4 Topic Time Coordination

A *Topic Time Coordination* é uma técnica de coordenação que exibe a evolução de tópicos de um determinado grupo de instâncias de um conjunto de dados que evolui no tempo. Assim é possível identificar novos tópicos que são provenientes das novas relações adquiridas durante a evolução temporal. Para utilizar essa técnica de coordenação as visões devem conter diferentes fatias de tempo do conjunto de dados. Por exemplo, se considerarmos um conjunto de dados com documentos dos anos de 2000 a 2005, uma possível configuração das fatias de tempo seria: documentos dos anos de 2000 a 2001, de 2000 a 2003, de 2000 a 2005. Note que a fatia de tempo seguinte contém os documentos da fatia de tempo anterior.

Essa técnica de coordenação utiliza um mapeamento estático semelhante ao mapeamento da *Identity Coordination* para identificar na visão destino os elementos de mesmo identificador (ID) selecionados na visão origem. Essa técnica utiliza o método **coordinate** para aplicar esse mapeamento criado no método **createMapping**. Além disso, esse método também é utilizado para gerar uma lista de vizinhos mais próximos com base nas instâncias de mesmo ID mapeadas na visão destino. Em seguida, é aplicado um algoritmo para detectar tópicos a partir dessa lista de vizinhos mais próximos. O método utilizado para gerar o tópico depende da configuração definida para cada visão destino. Também depende do número de vizinhos mais próximos e o tipo de vizinhança que será considerada. Essa vizinhança pode ser baseada nos vizinhos do espaço projetado, vizinhos do espaço original ou outro tipo de vizinhança, por exemplo, vizinhança baseada nas ligações obtidas da triangulação de Delaunay. Os métodos **createMapping** e **coordinate** são descritos abaixo, considerando que as fatias de tempo já foram previamente organizadas em diferentes visões:

- **createMapping**
  - Para cada instância na visão inicial
    - Verificar se há uma ou múltiplas instâncias com o mesmo ID na visão destino
    - Se sim
      - Criar um mapeamento entre as instâncias que possuem o mesmo ID
- **coordinate**
  - Para cada instância selecionada na visão origem
    - Obter a lista de instâncias mapeadas na visão destino
      - Selecionar cada instância mapeada e obter os vizinhos mais próximos
      - Adicionar os vizinhos mais próximos em uma lista, sem repetição
  - Mapear as instâncias da lista de vizinhos
  - Selecionar as instâncias da lista de vizinhos
  - Aplicar um algoritmo de detecção de tópicos sobre a lista de vizinhos

4.3.5 Time Coordination

A *Time Coordination* é outra estratégia de coordenação cujo mapeamento criado entre duas visões é dependente da interação. Após a seleção de instâncias na visão origem, é requerido um operador relacional, o qual é utilizado para realizar comparações entre um atributo temporal (e.g., hora e data) das instâncias selecionadas e o atributo temporal das instâncias da visão destino. Portanto, o método **createMapping** não pode gerar um mapeamento estático. Os métodos **createMapping** e **coordinate** são descritos abaixo:

- **createMapping**
  - Não há um mapeamento para ser criado, pois o operador relacional para realizar as comparações entre as instâncias é fornecido durante a interação
- **coordinate**
  - Exibir uma interface para requerer um operador relacional
  - Para cada instância selecionada na visão origem
    - Verificar quais instâncias na visão destino satisfazem a relação baseada no atributo temporal de acordo com o operador relacional fornecido
      - Mapear as instâncias que são cobertas pela relação
  - Selecionar as instâncias mapeadas

4.3.6 Joint Coordination

A *Joint Coordination* é uma técnica de coordenação que permite combinar o mapamento de diferentes técnicas em um único mapeamento. Na implementação atual a combinação é realizada pela intersecção dos diferentes mapeamentos gerados. Para isso, deve-se definir e adicionar à *Joint Coordination* as técnicas de coordenação cujo mapeamento deseja combinar. Conforme apresentado na Seção 5.3.6 são executadas cada uma das técnicas de coordenação adicionadas e o mapeamento de cada uma delas é recuperado. Em seguida, as instâncias comuns aos mapeamentos das diferentes técnicas são as instâncias que compreendem o mapeamento da *Joint Coordination*, isto é, as instâncias que representam a intersecção dos diferentes mapeamentos. Essa técnica não pode criar um mapeamento prévio, pois ela é dependente do mapeamento resultante de todas as técnicas de coordenação envolvidas em sua definição. Os parâmetros desta técnica são as próprias técnicas de coordenação definidas e adicionadas para criar o mapeamento final. Os métodos **createMapping** e **coordinate** são descritos abaixo:

- **createMapping**
  - Não há um mapeamento para ser criado, pois cada técnica de coordenação definida invocará seu próprio método **createMapping**
- **coordinate**
  - Para cada técnica de coordenação definida e adicionada
    - Invocar o método **coordinate** e recuperar o mapemanto resultante
  - Gerar um mapeamento final com as instâncias comuns aos mapeamentos das técnicas de coordenação previamente executadas
  - Selecionar na visão destino as instâncias do mapeamento final

A seguir é descrito como um desenvolvedor deve proceder para criar uma nova técnica de coordenação.

4.3.7 Nova Técnica de Coordenação

Nesta seção é exemplificado como um desenvolvedor pode criar uma nova técnica de coordenação. Como exemplo será criada uma técnica de coordenação que mapeia em outras visões as instâncias que possuem atributos com valor que estejam dentro da mesma faixa de valores dos atributos das instâncias selecionadas na visão origem. Essa técnica de coordenação pode ser empregada para descobrir instâncias que possuam as mesmas características das instâncias selecionadas. Por exemplo, alguns conjuntos de dados que variam no tempo possuem estruturas que mantêm as mesmas características em toda a variação temporal. Este é o caso do olho de um furacão, o

qual costuma manter baixos valores de pressão e temperatura em toda a variação temporal (Blaas e Post, 2008).

Essa técnica de coordenação é dinâmica, pois a variação dos atributos dependerá das instâncias selecionadas durante a interação. Portanto, não será criado um mapeamento no método **createMapping**. Então o mapeamento será criado e aplicado no método **coordinate**. O mapeamento dessa técnica deve obter o menor e o maior valor de cada um dos atributos das instâncias selecionadas na visão origem, resultando em uma faixa de valores. Em seguida, deve verificar quais instâncias da visão destino possuem todos atributos com valor dentro da faixa criada com os valores dos atributos das instâncias selecionadas. A comparação é realizada entre os atributos correspondentes, por exemplo, o valor da pressão é comparado com o menor e o maior valor de pressão da faixa de valores criada. Por fim, serão mapeadas na visão destino as instâncias que possuem atributos com valor dentro da faixa de valores. Após criar a nova técnica de coordenação, o desenvolvedor deve adicionar um novo código de instanciação na classe **CoordinationFactory**. Esse código será responsável por criar um objeto da nova técnica de coordenação desenvolvida. Os métodos **createMapping** e **coordinate** dessa nova técnica são descritos abaixo:

- **createMapping**
  - Não há um mapeamento para ser criado, pois a variação dos atributos será obtida durante a interação
- **coordinate**
  - Verificar qual o menor e maior valor de cada atributo das instâncias selecionadas, gerando uma faixa de valores
  - Para cada instância da visão destino
    - Se o valor de cada atributo está dentro da faixa de valores dos atributos correspondentes
      - Criar um mapeamento para essa instância
  - Selecionar na visão destino as instâncias mapeadas

Considerando todos os atributos para gerar uma faixa de valores para filtrar as instâncias nas visões destino, não é necessário utilizar parâmetros para criar essa técnica de coordenação. No entanto, é interessante permitir a escolha de quais atributos serão utilizados para criar a faixa de valores. Se essa escolha for efetuada durante o processo de coordenação, basta adicionar uma interface para selecionar os atributos desejados antes de criar o mapeamento, isto é, no início do método **coordinate**. Além disso, é necessário alterar o código para que sejam considerados somente os parâmetros selecionados. Caso o desenvolvedor deseje que os atributos sejam sempre os mesmos em todo o processo de coordenação, pode-se criar uma interface de entrada de parâmetros na definição da técnica de coordenação. Essa interface seria exibida antes da criação da técnica de

coordenação. Além disso, no caso de permitir que o usuário final possa selecionar os atributos na definição da técnica de coordenação, o desenvolvedor da técnica de coordenação deve adicionar um novo código de instanciação na classe **ParameterViewFactory**, após criar a interface para entrada de parâmetros. Nesse caso, o código do método **coordinate** também deve ser alterado para considerar somente os atributos selecionados pelo usuário final.

4.4 Considerações Finais

Neste capítulo foi introduzido o modelo de coordenação que permitiu a criação de técnicas de coordenação para exploração de múltiplos conjuntos de dados. Esse modelo é uma adaptação daquele proposto por Boukhelifa e Rodgers (Boukhelifa e Rodgers, 2003). Também foram apresentadas neste capítulo instanciações das classes do modelo para as técnicas de coordenação definidas no Capítulo 5.

O diferencial do modelo desenvolvido é a possibilidade de criar e configurar as técnicas de coordenação durante o processo de exploração. Assim, uma técnica de coordenação pode ser configurada para gerar diferentes tipos de mapeamentos entre as visões. Além disso, o modelo permite que as técnicas de coordenação utilizem elementos que mudam de acordo com a interação (mapeamentos dinâmicos), o que também pode modificar o mapeamento entre as visões.

O modelo também permite que novas técnicas de coordenação sejam desenvolvidas e adicionadas com facilidade ao sistema de visualização. A dificuldade maior é codificar a criação dos mapeamentos entre as instâncias das visões. Um exemplo de adição foi apresentado na Seção 4.3.7. O modelo também possibilita a adição de outras técnicas de visualização para atender às necessidades de exploração do usuário final.

O modelo e o sistema de visualização desenvolvidos permitem que as técnicas de coordenação apresentadas sejam utilizadas no processo de exploração de conjunto de dados que não são textuais. Entretanto, no caso das técnicas de coordenação que lidam com dados textuais, o conjunto explorado deve possuir alguma coleção textual relacionada. As instanciações apresentadas neste capítulo foram empregadas nas aplicações realizadas com a ferramenta *Projection Explorer* (PEX) (Paulovich et al., 2007) para explorar coleções de documentos e na *Projection Explorer for Images* (PEx-Image) (Eler et al., 2009a) para explorar coleções de imagens. Também, baseado neste modelo, foi desenvolvida uma ferramenta para auxiliar a exploração de dados volumétricos. Essa ferramenta emprega duas técnicas de visualização: mapas de similaridade, para representar a similaridade entre as instâncias com base no espaço multidimensional (espaço de características) dos dados volumétricos; e técnicas de visualização volumétrica, para exibir o espaço do objeto. A coordenação entre as duas técnicas de visualização será realizada por meio de uma técnica de coordenação baseada em identificador, refletindo no espaço do objeto as seleções realizadas na visualização do espaço multidimensional. Essas aplicações são apresentadas no Capítulo 6.

CAPÍTULO

5

Técnicas de Coordenação entre Mapas de Similaridade

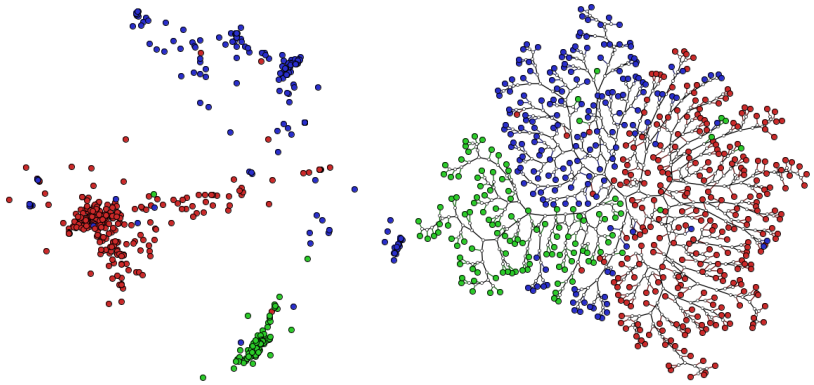
5.1 Considerações Iniciais

A coordenação entre mapas de similaridade aumenta a capacidade de exploração de conjuntos de dados, pois revela relações que são difíceis de serem representadas em uma única visualização. Os mapas podem ser construídos a partir de um único conjunto de dados, de diferentes conjuntos, ou a partir de diferentes representações de um conjunto de dados. Por exemplo, nesse último caso, os mapas podem ser construídos a partir de dados textuais relacionados a dados de imagens.

Este capítulo descreve as técnicas de coordenação desenvolvidas para apoiar a exploração de mapas de similaridade. Neste trabalho é utilizada a ferramenta *Projection Explorer* (PEx) (Paulovich et al., 2007) para construir os mapas de similaridade. Também são utilizadas as ferramentas de interação da PEx para auxiliar no processo de exploração e em algumas técnicas de coordenação. As coordenações desenvolvidas são implementadas como extensão da PEx, que é introduzida primeiro, destacando algumas técnicas de interação disponíveis para exploração de um conjunto de dados, especificamente, coleções de documentos. Em seguida, é apresentado como ocorre a exploração coordenada dos mapas por meio das técnicas desenvolvidas neste trabalho. Cada uma das técnicas desenvolvidas é detalhada e também são apresentados exemplos de aplicação.

5.2 Projection Explorer

Na abordagem proposta pela ferramenta *Projection Explorer* (PEx) (Paulovich et al., 2007) os mapas de similaridade são construídos a partir de projeções multidimensionais (Paulovich et al., 2008) e técnicas de posicionamento de pontos baseado em árvore (Cuadros et al., 2007). Em mapas gerados por projeções multidimensionais, pontos próximos no plano de visualização indicam maior similaridade (ver Figura 5.1(a)). No caso dos mapas gerados por técnicas de posicionamento de pontos baseado em árvore, os pontos de maior similaridade são aqueles posicionados em um mesmo ramo (ver Figura 5.1(b)). No exemplo da Figura 5.1 foi utilizada uma coleção de documentos de artigos científicos de três áreas: *Case-Based Reasoning* (azul), *Inductive Logic Programming* (Verde) e *Information Retrieval* (vermelho). Nesse exemplo a informação de classe é representada pela cor do círculo.



(a) Mapa de Similaridade gerado por Projeção Multidimensional.

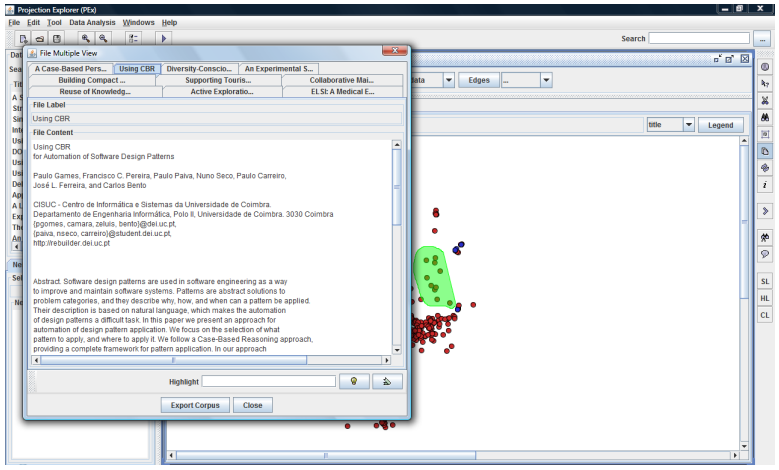
(b) Mapa de Similaridade gerado por Posicionamento de Pontos baseado em Árvores.

Figura 5.1: Mapa de similaridade gerado pela técnica LSP (a) e outro mapa pela técnica NJ (b).

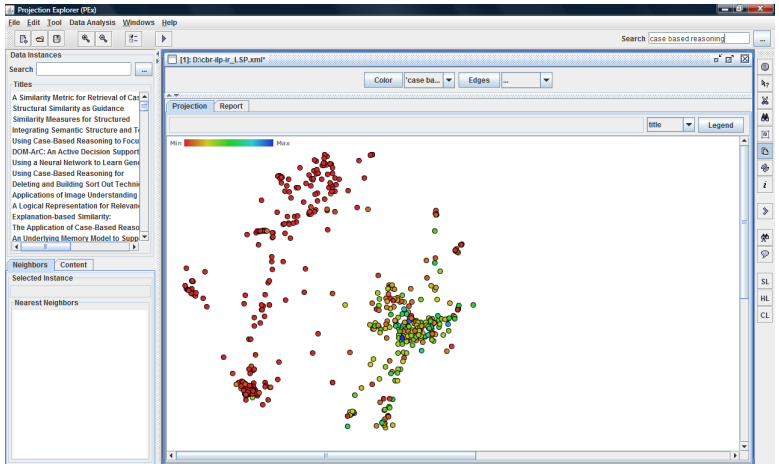
Após criar um mapa de similaridade o usuário pode explorar uma coleção de documentos por meio de técnicas de interação, conforme ilustrado na Figura 5.2. Com essas técnicas pode-se verificar quais são os nomes dos documentos que foram posicionados próximos. A Figura 5.2(a) ilustra isso para os documentos *"An Experimental Study of Increasing Diversity"*, *"Diversity-Conscious Retrieval"* e *"Using CBR"*. Outras interações são: visualizar o conteúdo de documentos selecionados (ver Figura 5.2(b)) e realizar buscas por palavras na coleção de documentos (ver Figura 5.2(c)). Nesse último tipo de interação a informação de cor é utilizada para representar o número de ocorrências de todos os termos da busca no conteúdo dos documentos. No exemplo, o vermelho indica nenhuma ocorrência dos termos e o azul indica a maior ocorrência dos termos.

Uma técnica essencial na exploração de conjuntos de documentos é a de detecção de tópicos, na qual os próprios documentos de um conjunto são os dados de entrada para a técnica de detecção. Esses documentos podem ser selecionados pelo usuário ou por algum processo automático (e.g., algoritmos de agrupamento). A Figura 5.3 mostra um exemplo de aplicação de duas téc-

(a) Verificação dos documentos que foram posicionados próximos.



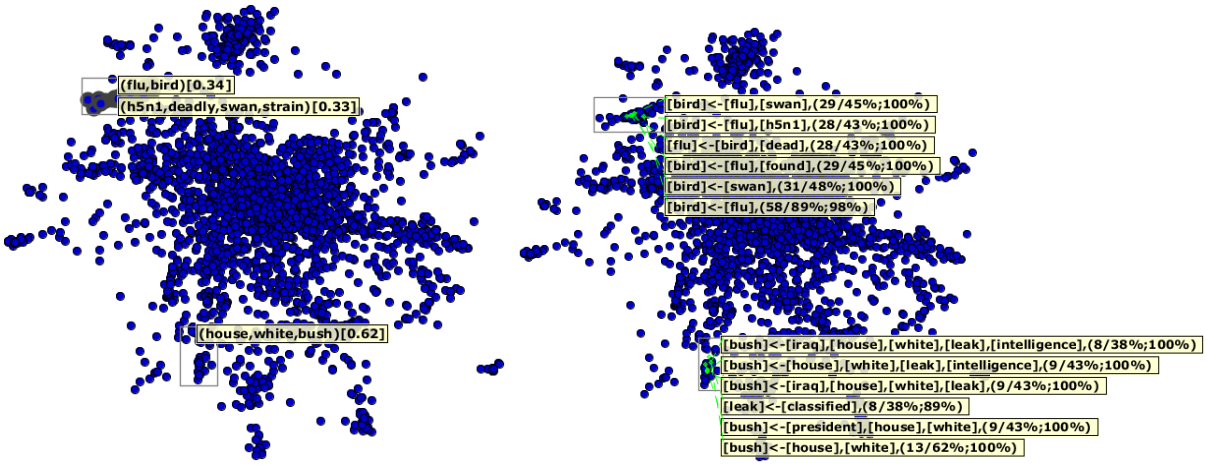
(b) Exibição do conteúdo de documentos selecionados.



(c) Busca no conteúdo dos documentos.

Figura 5.2: Técnicas de interação da *Projection Explorer*: (a) Exibição dos nomes dos documentos selecionados; (b) Exibição de conteúdo dos documentos selecionados; (c) Destaque dos documentos cobertos pelos termos da busca *"case based reasoning"*. A cor representa o número de ocorrências das palavras no conteúdo de cada documento.

nicas de detecção de tópicos, aplicadas em uma coleção de 2684 artigos de notícias mundiais [1]. A Figura 5.3(a) mostra o resultado da detecção de tópicos por meio de uma técnica baseada na covariância dos termos (Cuadros et al., 2007), revelando o conteúdo principal dos documentos selecionados. Essa técnica é apresentada com mais detalhes na Seção 5.3.2. A Figura 5.3(b) mostra o resultado gerado a partir da mesma seleção realizada no exemplo da Figura 5.3(a), mas é utilizada uma técnica de detecção baseada em regras de associação, conhecida como LWR (Lopes et al., 2007; Pinho, 2009). Essa técnica revela mais detalhes sobre o relacionamento entre os termos dos tópicos. A LWR aplica o algoritmo *Apriori* (Agrawal e Srikant, 1994) sobre os termos do espaço multidimensional para gerar regras de associação, as quais são exibidas como tópicos que descrevem o conteúdo de um grupo de documentos. Para evitar que um grande número de regras seja gerado e para garantir que sejam capturadas aquelas que melhor representam o conteúdo dos documentos, Lopes e outros (Lopes et al., 2007) apresentaram estratégias para restringir o espaço de busca do *Apriori*. Dada a sua complexidade, não serão apresentados detalhes da técnica LWR.



(a) Extração de tópicos baseada em covariância. (b) Extração de tópicos baseada em regras de associação.

Figura 5.3: Exploração de um conjunto de notícias por meio de técnicas de detecção de tópicos: (a) método da covariância (Cuadros et al., 2007) e (b) método das regras de associação (Lopes et al., 2007).

A técnica baseada em covariância revela tópicos que apresentam uma visão geral do conteúdo, isto é, os termos principais. Já a técnica baseada em regras de associação revela tópicos mais detalhados. Como apresentado na Figura 5.3(a), a seleção realizada na parte acima e à esquerda do mapa mostra tópicos sobre a gripe aviária, o subtipo do vírus (h5n1) e a morte de um cisne. As mesmas seleções são apresentadas na Figura 5.3(b), onde a técnica baseada em regras de associação revela detalhes sobre os tópicos, apresentando mais claramente como os termos estão relacionados no conteúdo dos documentos. Numa das seleções a gripe aviária é associada a um cisne encontrado morto e, na outra, o nome do Presidente Bush é associado a um vazamento de informações sobre o Iraque.

[1] Artigos de notícias publicados na internet pelas agências Reuters, BBC, CNN e Associated Press. Os artigos foram coletados durante dois dias do mês de Abril de 2006.

A ferramenta PEx foi desenvolvida em Java e tem seu código aberto. A ferramenta, manual e o código estão disponíveis para *download* em http://infoserver.lcad.icmc.usp.br/infovis2/PEx. A seguir são apresentadas novas técnicas de coordenação adicionadas à PEx para apoiar a exploração de diferentes mapas de similaridade. A maioria dos exemplos utilizam coleções de documentos, pois foram os dados que motivaram a criação dessas novas técnicas.

5.3 Exploração Coordenada entre Mapas de Similaridade

Durante o processo de exploração o usuário pode criar mapas de similaridade a partir de diferentes conjuntos de dados e definir técnicas de coordenação para apoiá-lo. As técnicas de coordenação criam uma relação entre as instâncias das visões envolvidas no processo de coordenação, isto é, geram um mapeamento entre as visões. A Figura 5.4 apresenta quatro mapas de similaridade construídos a partir de três diferentes coleções de documentos. Esse exemplo é somente uma ilustração de como as coordenações são executadas. O detalhamento das técnicas de coordenação e novos exemplos serão apresentados nas próximas seções. Os mapas indicados pelas letras **A** e **B** foram construídos a partir de uma coleção de artigos científicos da conferência *Information Visualization Conference* (IV) (artigos dos anos de 1997 a 2008); sendo o mapa indicado pela letra **A** gerado pela técnica *Least Squares Projection* (LSP) (Paulovich et al., 2008) e o mapa indicado pela letra **B** gerado pela técnica *Neighbour-Joining* (NJ) (Cuadros et al., 2007). O mapa indicado pela letra **C** foi gerado pela técnica LSP a partir de uma coleção de artigos da conferencia *Coordinated & Multiple Views* (CMV) (artigos dos anos 2003 a 2008). O mapa indicado pela letra **D** foi gerado pela técnica LSP a partir de uma coleção de artigos da conferencia *IEEE Symposium on Information Visualization* (IEEE InfoVis) (artigos dos anos de 1995 a 2008). Todos os mapas estão coordenados com o mapa indicado pela letra **A** por meio de diferentes técnicas de coordenação. A coordenação ativa, isto é, aquela que aplicará o mapeamento em cada mapa, é indicada ao lado do botão *Coordination*. O mapa indicado pela letra **A** não receberá o mapeamento de nenhuma técnica de coordenação, por isso está selecionada a opção *OFF*.

No exemplo apresentado na Figura 5.4, o usuário fez a seleção de alguns documentos no mapa indicado pela letra **A** (visão origem), destacando nos outros mapas (visões destino) as instâncias mapeadas, de acordo com a coordenação ativa em cada um. Cada uma dessas coordenações são descritas nas próximas seções. Elas permitem a exploração de um único conjunto de dados e também de diferentes conjuntos. As técnicas são baseadas em identificador, em tópicos, em tópicos identificados durante a evolução temporal de um conjunto, em atributos temporais das instâncias de um conjunto, em distâncias e na combinação dos mapeamentos gerados por diferentes técnicas de coordenação.

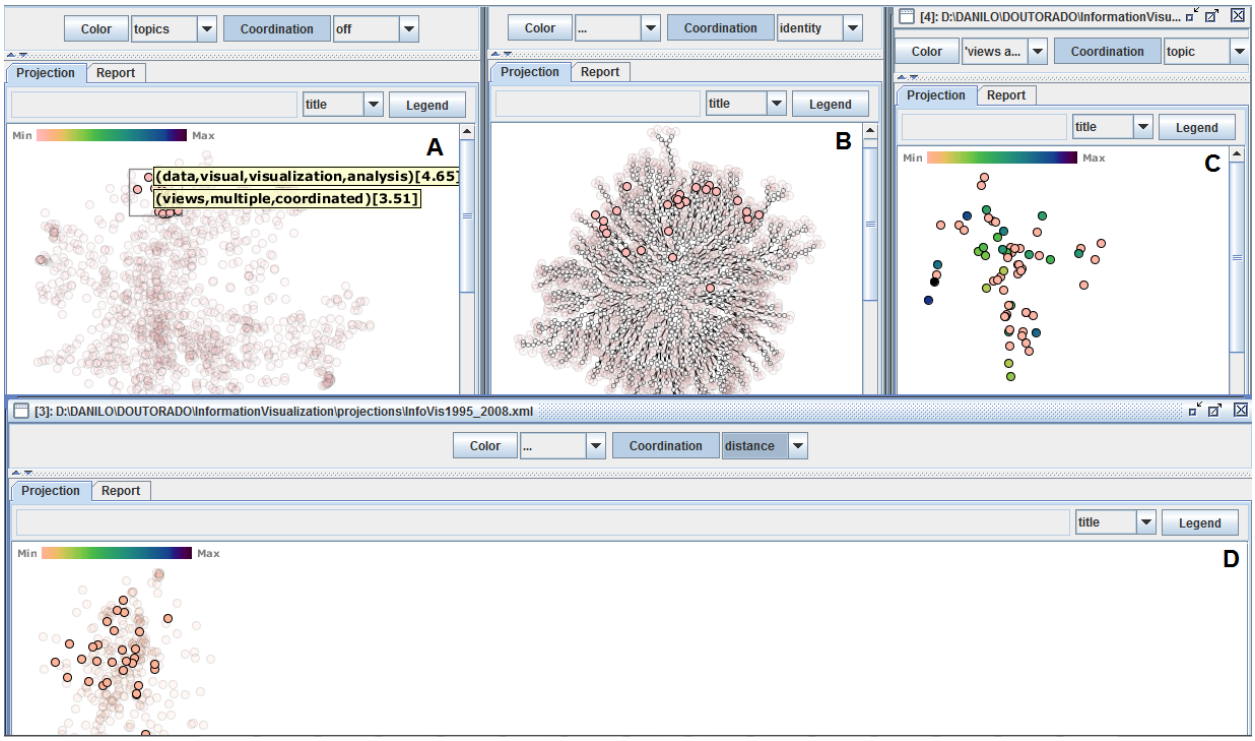
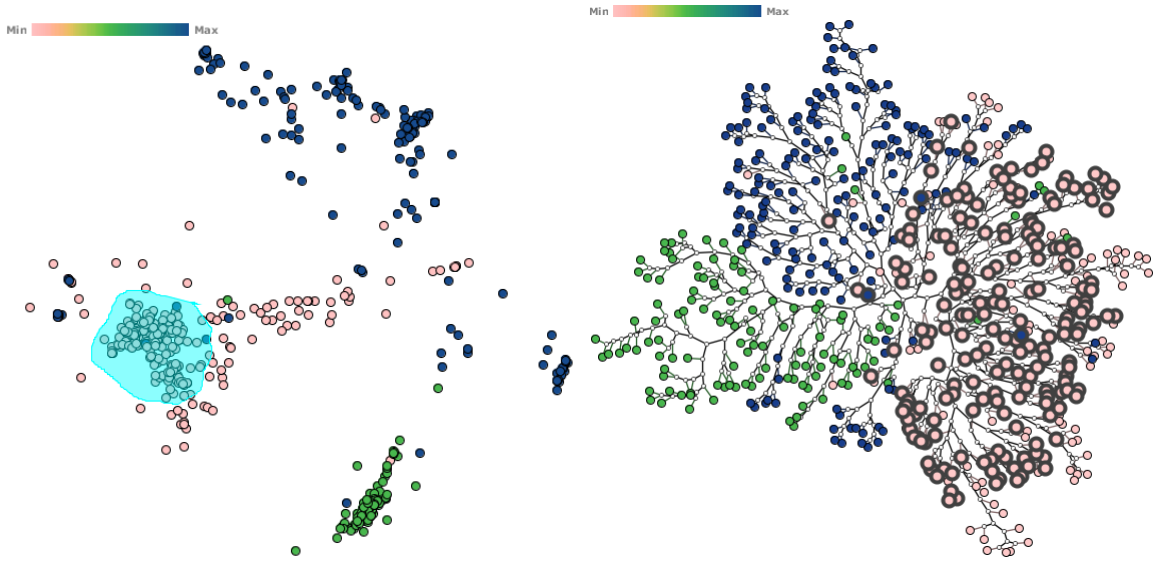


Figura 5.4: Exploração coordenada de múltiplos mapas de similaridade utilizando três tipos de técnicas de coordenação e três diferentes conjuntos de dados. O mapa indicado pela letra **A** é a seleção inicial do usuário. O mapa indicado pela letra **B** mostra o resultado do mapeamento da coordenação baseada em identificador. O mapa indicado pela letra **C** mostra o resultado do mapeamento da coordenação baseada em tópicos. O mapa indicado pela letra **D** mostra o resultado do mapeamento da coordenação baseada em distância.

5.3.1 Coordenação baseada em Identificador

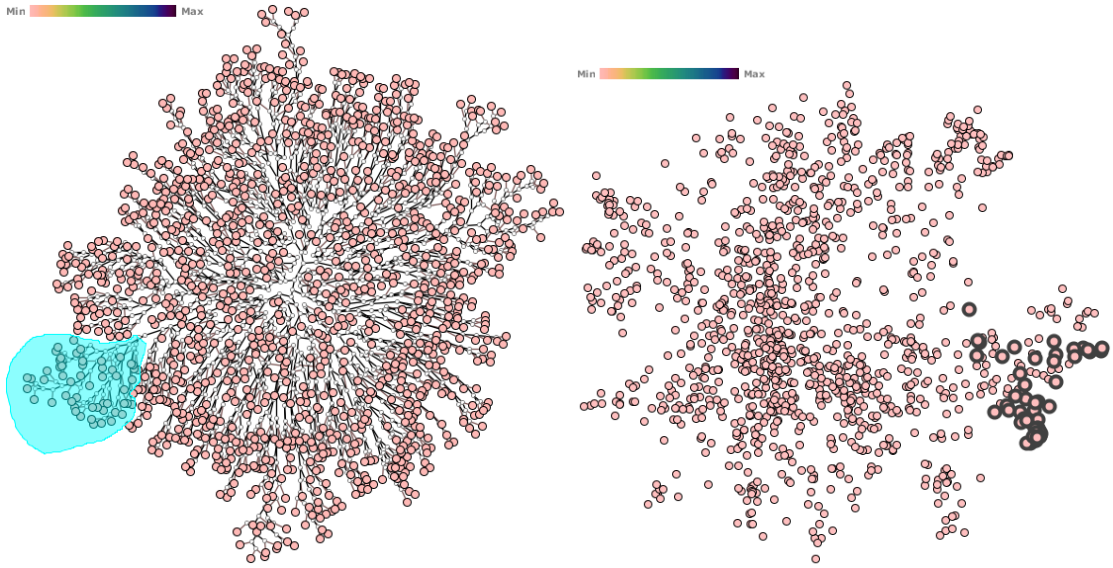
A coordenação baseada em identificador é a coordenação mais simples e mais implementada em sistemas que utilizam a abordagem de múltiplas visões coordenadas. Ela utiliza o identificador das instâncias selecionadas em uma visão (visão origem) para destacar em outras visões (visões destino) aquelas que possuem o mesmo identificador. Neste trabalho essa técnica foi denominada *Identity Coordination*. No exemplo da Figura 5.5 são utilizados os mesmos mapas mostrados na Figura 5.1 para ilustrar o funcionamento da coordenação por identidade. Primeiramente o usuário selecionou um grupo de documentos na visão origem (Figura 5.5(a)) e os documentos de mesmo identificador foram destacados na visão destino (Figura 5.5(b)). A cor dos círculos representa a informação de classe de cada documento.

A Figura 5.6 mostra um outro exemplo de coordenação por identificador, utilizando os mesmos mapas indicados pelas letras **B** e **A** na Figura 5.4. No exemplo, é selecionado um ramo inteiro da árvore NJ (Figura 5.6(a)), destacando no outro mapa (Figura 5.6(b)) os documentos de mesmo identificador. O destaque revela que as relações são mantidas tanto para a representação baseada em árvore, que agrupou as instâncias em um mesmo ramo, quanto para a representação criada por meio de projeção, que agrupou as instâncias em uma região próxima no espaço projetado.



(a) Visão Origem – Mapa gerado pela técnica LSP. (b) Visão Destino – Mapa gerado pela técnica NJ.

Figura 5.5: Aplicação da *Identity Coordination* para coordenar dois mapas de similaridade gerados pelas técnicas LSP (a) e NJ (b). O usuário selecionou documentos em (a) e aqueles de mesmo identificador foram destacados em (b).

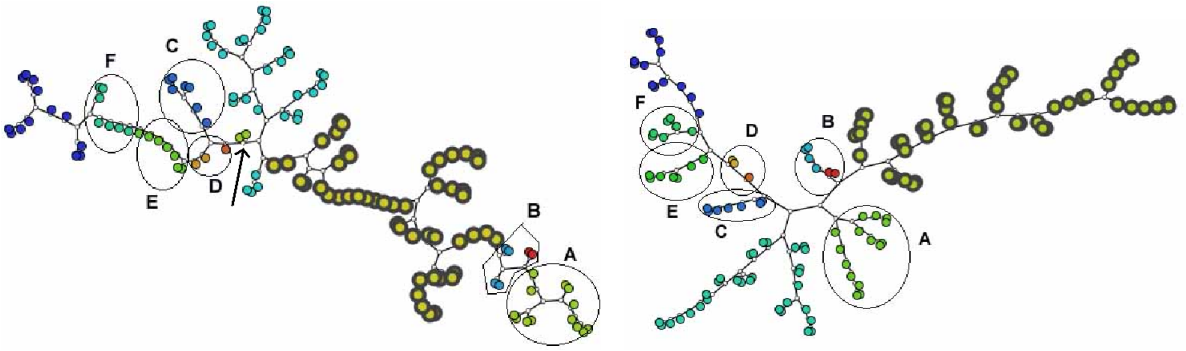


(a) Visão Destino – Mapa gerado pela técnica NJ. (b) Visão Origem – Mapa gerado pela técnica NJ.

Figura 5.6: Outra aplicação da *Identity Coordination* para coordenar dois mapas de similaridade gerados pelas técnicas NJ (a) e LSP (b). O usuário selecionou documentos em (a) e aqueles de mesmo identificador foram destacados em (b).

Neste trabalho os mapas de similaridade são visualmente representados por técnicas de posicionamento de pontos no plano. Ainda que sejam utilizadas diferentes técnicas para criar os mapas de similaridade, elas não mudariam significativamente a perspectiva do usuário sobre os relacionamentos das instâncias de um conjunto de dados, conforme apresenta o exemplo ilustrado na Figura 5.6. Nesse contexto, por meio da aplicação de uma simples técnica de coordenação por

identidade, o usuário pode destacar grupos e comparar a qualidade que as diferentes técnicas de projeção multidimensional têm em refletir a similaridade entre as instâncias. Além disso, essa técnica de coordenação pode auxiliar na comparação de diferentes espaços multidimensionais ou na aplicação de diferentes métricas de distância para a construção de mapas de similaridade. Outra utilidade desta técnica de coordenação é a análise a partir de representações diferentes de um mesmo conjunto de dados.



(a) (b)

Figura 5.7: Mapas construídos a partir de um conjunto de 138 sequências de proteína com 11 classes de enzimas. (a) Mapa construído a partir dos *contact maps* (imagens) e (b) Mapa construído a partir da sequência de proteínas (texto). O rótulo indica os mesmos grupos nos dois mapas (Extraída de (Eler et al., 2009a)).

A Figura 5.7 mostra um exemplo de exploração de parte de um conjunto de proteídas obtidas da base RSCB - *Protein Data Bank* (obtido em http://www.rcsb.org/pdb). Para construir a amostra foi aleatoriamente selecionada uma sequência de 11 diferente classes de enzimas. Em seguida foram selecionadas aquelas que apresentaram *e-value* [2] menor ou igual a e^{-100}, resultando em um total de 138 sequências de proteína com 11 classes de enzimas. As comparações foram realizadas por meio da ferramenta BLAST (Altschul et al., 1990), que é uma ferramenta largamente utilizada para alinhamento de sequências genéticas. Além das representações que são as próprias sequências genéticas (textos), uma outra representação foi utilizada para cada instância do conjunto. No caso, foram gerados *contact maps* [3] (imagens) a partir das sequências genéticas por meio da ferramenta PROTMAP2D (Michal J. Pietal e Bujnicki, 2007).

A Figura 5.7(a) apresenta um mapa construído por meio da técnica NJ aplicada sobre os *contact maps*. A Figura 5.7(b) apresenta um mapa construído a partir da técnica NJ aplicada sobre as sequências genéticas. Além da coordenação destacando um grupo selecionado, os demais grupos foram rotulados com a mesma letra em ambos os mapas. É possível notar que no mapa apresentado na Figura 5.7(a) duas instâncias, assinaladas por uma seta, foram posicionadas fora do grupo indicado pela letra **A**, que seria seu grupo ideal. Esse posicionamento fora do grupo revela que há

[2] *Expectation value* ou *e-value* é o resultado de cálculos estatísticos que indicam o grau de probabilidade de um alinhamento entre sequências genéticas ter ocorrido ao acaso.

[3] *Contact maps* representam a distância entre cada par de resíduos de uma proteína por meio de uma matriz bidimensional. Para dois resíduos i e j, é atribuido o valor 1.0 para o elemento $[i, j]$ da matriz se os resíduos estão mais próximos do que um determinado limiar; caso contrário, é atribuido o valor 0.0.

alguma diferença entre essas duas instâncias e as demais, considerando a representação por meio dos *contact maps*. A Figura 5.8 mostra partes do mapa da Figura 5.7(a), apresentando os padrões das *contact maps* gerados para as duas instâncias que foram posicionadas fora do grupo e para as demais instâncias do grupo indicado pela letra **A**. Essa exploração revela que há de fato diferenças entre os *contact maps* dessas sequências e das demais.

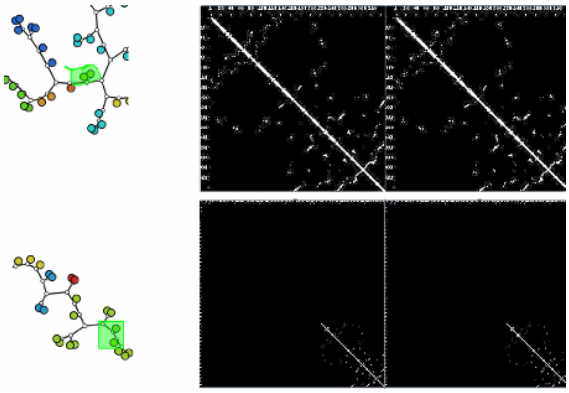


Figura 5.8: Partes do mapa da Figura 5.7(a), mostrando o padrão dos *contact maps* das instâncias posicionadas fora do grupo indicado pela letra **A** e o padrão das demais instâncias do grupo (Extraída de (Eler et al., 2009a)).

Essa estratégia de visualizar diferentes representações de um conjunto de dados também foi aplicada por Andery (Andery, 2010) para explorar conjunto de dados de redes de socias (e.g., orkut). Eles aplicaram a coordenação baseada em identificador para explorar representações visuais construídas a partir de dados de relacionamento entre membros de comunidades, gerados a partir de informações das comunidades que os membros participam. As representações visuais utilizaram diferentes informações para as mesmas instâncias e por meio da coordenação foi possível identificar novos relacionamentos entre elas.

A seguir são apresentadas técnicas capazes de coordenar mapas de similaridade construídos a partir de diferentes conjuntos de dados, os quais não estão explicitamente relacionados. A primeira delas é específica para a exploração de mapas construídos a partir de conjuntos que possuam alguma informação textual relacionada com as instâncias.

5.3.2 Coordenação baseada em Tópicos

Técnicas para detecção de tópicos são ferramentas importantes para auxiliar a exploração de coleções de documentos. Além disso, essas técnicas podem ser utilizadas no processo de exploração de diferentes coleções de documentos. Portanto, foi desenvolvida uma técnica de coordenação baseada em tópico denominada *Topic Coordination* (Eler et al., 2008b), a qual permite a coordenação e exploração de múltiplas coleções de documentos ou de outros conjuntos que possuam documentos relacionados com as instâncias. Por meio dessa técnica o usuário pode focar nos documentos de outras visões a partir de um tópico de interesse gerado durante a exploração de uma visão. Para isso, em uma visão origem o usuário faz a seleção de documentos de interesse, a partir

dos quais é gerado um tópico. Em seguida, o tópico gerado é propagado para as demais visões por meio da *Topic Coordination*, que destaca os documentos cobertos pelo tópico em todas elas. Um documento é considerado coberto por um tópico se ele possuir todos os termos do tópico em seu conteúdo.

O processo de extração de tópicos utilizado nessa técnica é a baseado na covariância dos termos. Nesse processo, uma matriz de "documentos x termos" (um espaço multidimensional) é construída para o conjunto de documentos selecionados. Nessa matriz cada linha representa um documento selecionado e cada coluna corresponde a um termo que ocorre no conteúdo dos documentos. Os valores dos elementos da matriz identificam a frequência de um termo específico no conteúdo de um determinado documento. Nessa abordagem para detecção de tópicos é executada uma análise de covariância dos termos na matriz de "documentos x termos". Primeiramente, os dois termos com maior covariância são escolhidos como um tópico inicial para descrever a seleção de documentos. A covariância mede o grau em que dois termos variam juntamente no grupo de documentos selecionados. A covariância é definida como

$$cov(t_i, t_j) = \frac{1}{n-1} \sum_{k=1}^{n} (t_{ki} - \bar{t_i})(t_{kj} - \bar{t_j}) \tag{5.1}$$

onde n é o número de documentos na seleção, $\bar{t_i}$ é a frequência média do i-ésimo termo t_i, e t_{ki} e t_{kj} são as frequências dos i-ésimo termo e j-ésimo termo no documento k.

Uma vez que os dois termos com maior covariância são encontrados, a covariância média entre eles e os termos restantes é computada. A covariância média é definida como

$$cov_m(t_i, t_j, t_s) = \frac{cov(t_i, t_s) + cov(t_j, t_s)}{2} \tag{5.2}$$

onde t_i e t_j são os termos com a maior covariância, e t_s é o termo para o qual queremos calcular a covariância média, dados t_i e t_j.

Em seguida, se a razão entre a covariância média de um termo t_s e o maior valor de covariância encontrado é maior ou igual a um limiar α, então o termo t_s é incluído na lista de termos do tópico. Devido à natureza do processo, a identificação de um tópico é fortemente influenciada pelos dois termos iniciais, podendo excluir termos que agregariam valor ao tópico gerado. Com o intuito de superar essa limitação, sempre que a razão entre a covariância de qualquer par de termos ainda não considerados e a maior covariância do par de termos inicial for acima de um certo limiar β, estes termos são considerados como um novo tópico e o processo de detecção é repetido com base nesses dois termos. O resultado final desse processo, descrito no Algoritmo 1, é uma lista de tópicos.

Uma vez que os tópicos foram detectados a partir de uma seleção realizada na visão origem, a *Topic Coordination* calcula sua ocorrência no conteúdo de cada documento da visão destino. Se todos os termos do tópico estão presentes no conteúdo de um documento, o valor de ocorrência do tópico é a média dos valores de ocorrência de cada termo; caso um dos termos não esteja presente, o valor de ocorrência do tópico é zero. Os valores de ocorrência são normalizados entre 0.0 e 1.0.

Algorithm 1: Detecção de Tópicos Baseada em Covariância (adaptado de (Eler et al., 2008b)).

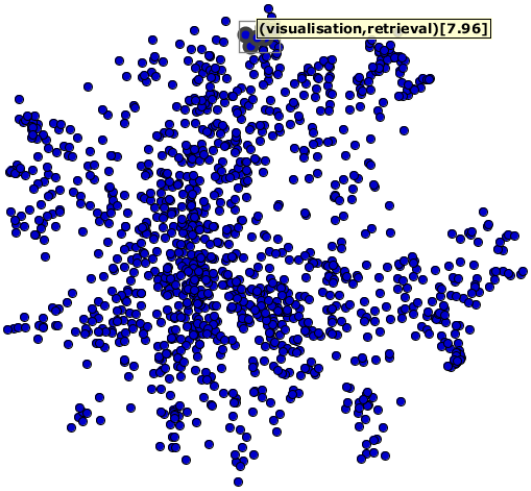
in: - $\mathbf{M}$: $n \times m$ matriz de "documentos x termos" calculada para os n documentos selecionados com m termos.
- α : limiar para adicionar novos termos a um tópico.
- β : limiar para adicionar novos tópicos à lista de tópicos.

out: - L: lista de tópicos identificados.

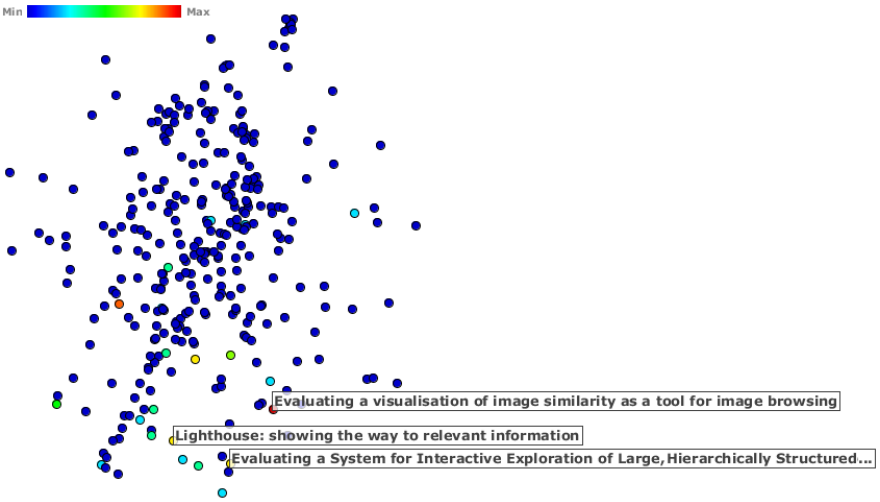
1: $t_{i_{max}}, t_{j_{max}}$ = $\max\{cov(t_i, t_j)\}$ $\forall t_i, t_j \in \mathbf{M}$
2: $cov = cov(t_{i_{max}}, t_{j_{max}})$
3: $cov_{max} = cov$
4: $L = \emptyset$
5: **while** $cov/cov_{max} \geq \beta$ **do**
6: cria um novo tópico $T = \emptyset$
7: adiciona os termos t_i e t_j a T.
8: **for all** $t_k \in M$ e $t_k \neq t_i$ e $t_k \neq t_j$ **do**
9: **if** $cov_m(t_i, t_j, t_k)/cov \geq \alpha$ **then**
10: adiciona o termo t_k a T.
11: **end if**
12: **end for**
13: adiciona o tópico T à lista de tópicos L.
14: t_i, t_j = $\max\{cov(t_i, t_j)\}$ $\forall t_i, t_j \in \mathbf{M}$ **e** ($t_i \notin l_b$ **ou** $t_j \notin l_b$, $\forall l_b \in L$)
15: $cov = cov(t_i, t_j)$
16: **end while**

Em seguida, esse valor é mapeado para cor, indicando a cobertura do tópico para cada documento da visão destino. A cor atribuída ao valor de ocorrência é dependente da tabela de cores escolhida pelo usuário. Na implementação atual da *Topic Coordination* é considerado somente o tópico principal, isto é, o tópico cujos termos apresentam os maiores valores de covariância.

Um exemplo de aplicação da *Topic Coordination* é apresentado na Figura 5.9. O mapa de similaridade apresentado na Figura 5.9(a) foi gerado pela técnica LSP, a partir de uma coleção de artigos científicos da conferência *Information Visualization Conference* (IV) (artigos dos anos de 1997 a 2008); e o mapa de similaridade da Figura 5.9(b) também foi gerado pela técnica LSP a partir de uma coleção de artigos científicos da conferência *IEEE Symposium on Information Visualization* (IEEE InfoVis) (artigos dos anos de 1995 a 2008). No exemplo, o usuário inicia a exploração com a seleção de alguns documentos no mapa apresentado na Figura 5.9(a) (visão origem), gerando o tópico *"visualization, retrieval"*. Em seguida, no mapa apresentado na Figura 5.9(b) são destacados os documentos cobertos pelo tópico detectado. No caso, a escala de cor escolhida é exibida no topo e à esquerda da Figura 5.9(b). Essa escala utiliza a cor azul para representar nenhuma ocorrência do tópico e a cor vermelho para representar a máxima ocorrência. Foram exibidos os nomes de alguns dos documentos que tiveram a maior cobertura do tópico, os quais tratam de navegação e recuperação de documentos e imagens. São eles: *"Evaluating a visualization of image similarity as a tool for image browsing"*, *"Lighthouse: showing the way to relevant information"* e *"Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories"*.



(a) Seleção em documentos da conferência IV.



(b) Destaque dos documentos da conferência InfoVis.

Figura 5.9: Exemplo de aplicação da técnica *Topic Coordination*(Eler et al., 2008b). No mapa de similaridade ilustrado em (a) foram selecionados alguns documentos e gerado o tópico "*visualization, retrieval*". Em (b) foram destacados os documentos cobertos pelo tópico. O destaque é realizado pela informação de cor, que indica a ocorrência dos termos do tópico.

Essa técnica pode ser aplicada a qualquer mapa de similaridade que tenha alguma informação textual relacionada com as instâncias. Além disso, outras formas de detecção de tópicos poderiam ser empregadas.

A seguir é apresentada uma técnica que também se baseia em tópicos para apoiar o processo de exploração.

5.3.3 Coordenação baseada em Tópicos no Tempo

Neste trabalho foi desenvolvida uma técnica de coordenação que utiliza as técnicas de detecção de tópicos para auxiliar na exploração de múltiplos mapas de similaridade. Essa técnica, chamada de *Topic Time Coordination* (Eler et al., 2009b), tem o objetivo de identificar tópicos a partir de

novas relações de vizinhança que documentos de interesse, selecionados pelo usuário, venham a ter durante a evolução temporal de uma coleção. Essa abordagem revela o assunto dos documentos que se relacionam aos documentos de interesse com o passar do tempo. Nesse contexto, evolução temporal significa adição de novos documentos à coleção conforme o tempo passa.

A *Topic Time Coordination* opera em diferentes mapas de similaridade gerados a partir de um determinado período de tempo de uma coleção de documentos. Esses mapas são denominados *Time-Span Maps* (*TS Maps*). Por exemplo, a partir de uma coleção de documentos correspondente aos anos de 1990 a 2005, pode-se criar três *TS Maps* incluindo, respectivamente, documentos dos anos de 1990 a 1994; de 1990 a 1999; e de 1990 a 2005. Note que o *TS Map* subsequente sempre incorpora os documentos de um *TS map* antecedente. Considerando um conjunto de documentos selecionados pelo usuário na visão origem, a *Topic Time Coordination* realiza a extração de tópicos com base na vizinhança desses documentos nas subsequentes visões destino.

Como as outras técnicas de coordenação, a *Topic Time Coordination* inicia a partir da seleção de algumas instâncias em um *TS Map* inicial, a partir das quais são gerados tópicos para o usuário conhecer o assunto dos documentos da seleção. Em seguida, as instâncias de mesmo identificador são destacadas nos mapas das visões destino, para que o usuário possa localizá-las em sua nova vizinhança, a qual é utilizada para a detecção de tópicos. As instâncias que pertencem à nova vizinhança são identificadas com base nos k vizinhos mais próximos daquelas de mesmo identificador, que são as instâncias selecionadas no *TS Map* inicial. O número de vizinhos (k) e a relação de vizinhança empregada são escolhidos pelo usuário, que pode selecionar uma vizinhança baseada no espaço original (espaço multidimensional), no espaço projetado, ou baseada em outro tipo de relação de vizinhança, como, por exemplo, as instâncias conectadas por meio de uma triângulação de Delaunay. Atualmente, esses são os três tipos de relação de vizinhança que podem ser geradas a partir da PEx. No entanto, a funcionalidade permite que outros tipos de vizinhança calculadas fora da ferramenta sejam carregados.

A partir dos k vizinhos mais próximos daqueles inicialmente selecionados são extraídos tópicos, os quais revelam o conteúdo das novas instâncias que estão relacionadas àquelas de interesse. A técnica de detecção empregada também é escolhida pelo usuário, que pode escolher técnicas diferentes para cada visão destino. Atualmente, as técnicas de detecção disponíveis na PEx são baseadas em covariância (Eler et al., 2009b) e em regras de associação (técnica LWR (Lopes et al., 2007)). Ao fim do processo, o usuário tem uma noção da evolução dos tópicos tratados por documentos relacionados ao grupo de documentos de interesse do usuário.

Um exemplo de aplicação da *Topic Time Coordination* é apresentado na Figura 5.10. Nesse exemplo são apresentados três *TS Maps* construídos a partir da coleção de artigos da conferência *Coordinated & Multiple Views* (CMV) (artigos dos anos 2003 a 2008). O mapa acima e à esquerda contém documentos dos anos 2003 e 2004; o mapa acima e à direita contem documentos dos anos 2003 a 2006; e o mapa da parte de baixo contem documentos dos anos de 2003 a 2008. Primeiramente o usuário selecionou dois documentos no *TS Map* inicial (visão origem), apresentado na parte acima e à esquerda da Figura 5.10. Os documentos selecionados são representados

em vermelho. Nesse mapa, é detectado e exibido o tópico dos documentos selecionados. Nos outros *TS Maps* são destacados os documentos que possuem o mesmo identificador. Em seguida, seus 2 vizinhos mais próximos também são destacados, mas com um menor grau de transparência, como apresentado nos mapas posicionados na parte de baixo e na parte acima e à direita da Figura 5.10. Os tópicos exibidos no *TS Map* inicial descrevem os documentos selecionados pelo usuário e os tópicos exibidos nos outros *TS Maps* descrevem os documentos que estão na vizinhança dos documentos da seleção inicial. Os tópicos dos mapas apresentados no topo da figura foram extraídos pelo algoritmo de detecção baseado na covariância dos termos. Aqueles tópicos mostrados no mapa apresentado na parte de baixo da figura foram obtidos por meio do algoritmo LWR (Lopes et al., 2007).

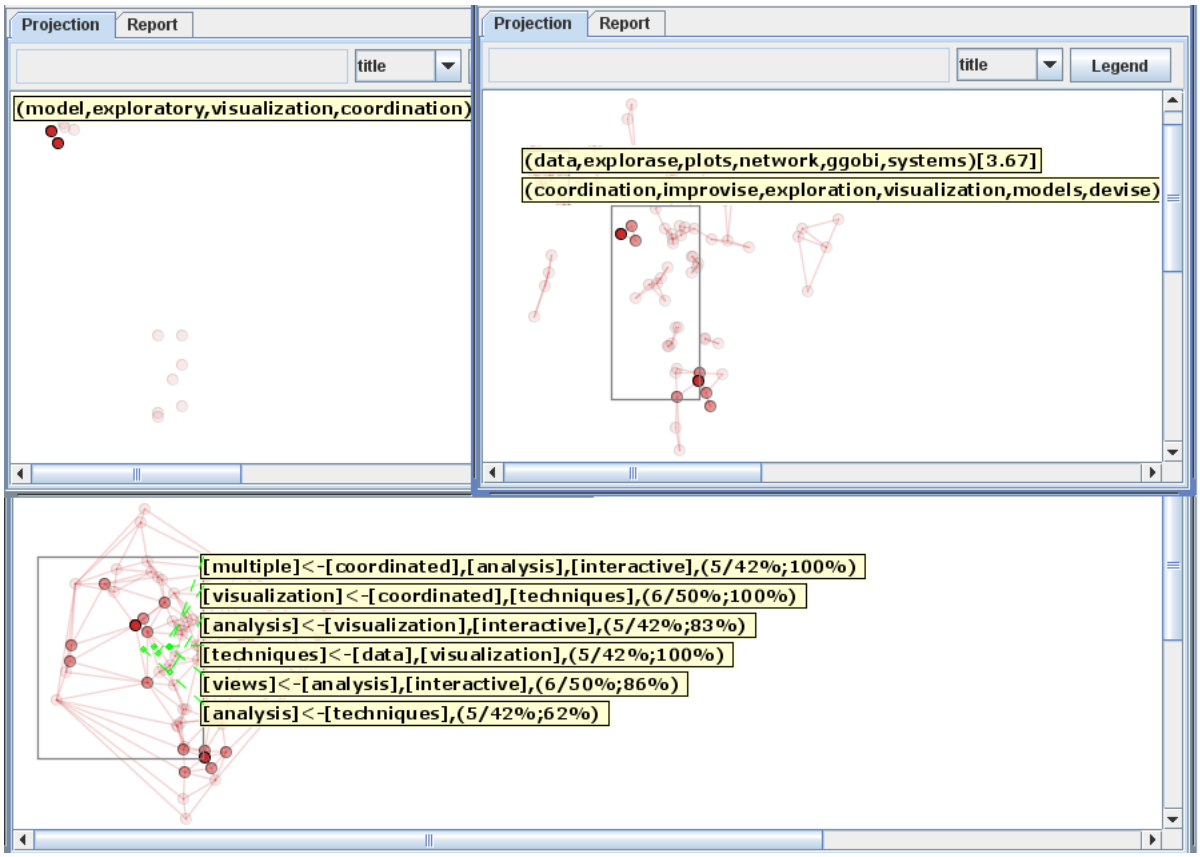


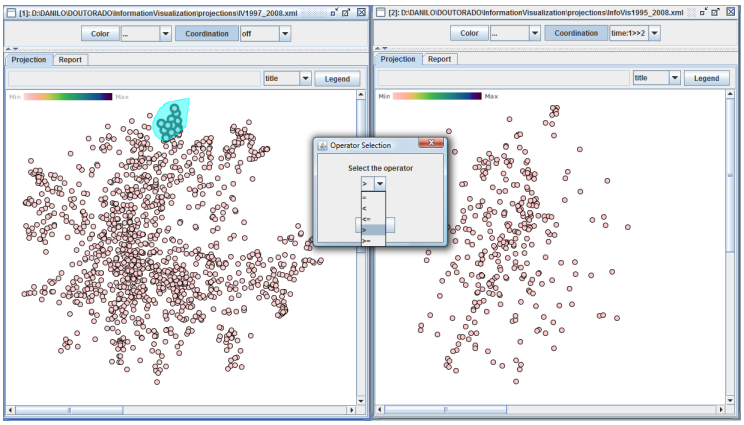
Figura 5.10: *Topic Time Coordination* aplicada em três *TS Maps* (Extraída de (Eler et al., 2009b)).

Além de alterar o algoritmo de detecção de tópicos, também é possível alterar a forma como a vizinhança é definida. Na Figura 5.10, o mapa acima e à direita utilizou uma vizinhança definida no espaço 2D e o mapa da parte de baixo utilizou uma vizinhança definida por meio de uma triangulação de Delaunay. No caso dessa triângulação, são consideradas todas as instâncias que estão conectadas às instâncias selecionadas.

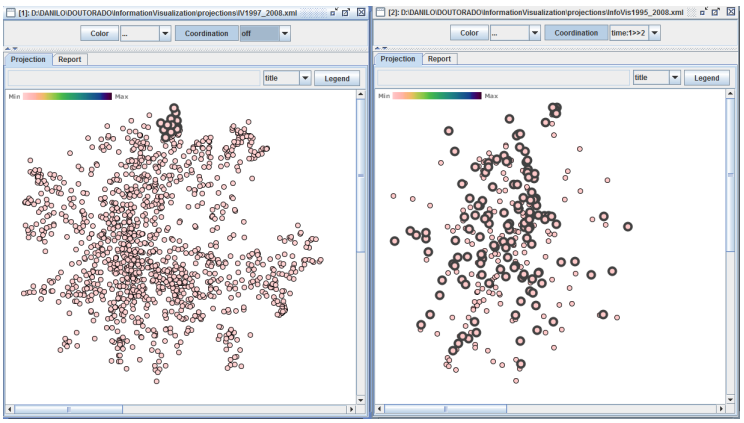
A seguir é apresentada uma técnica de coordenação que gera mapeamentos com base em atributos temporais das instâncias de um conjunto de dados.

5.3.4 Coordenação baseada em Atributos Temporais

Atualmente, muitos conjuntos de dados possuem informações temporais (e.g., data) relacionadas às instâncias. Essas informações podem ser utilizadas para identificar relacionamentos temporais entre as instâncias de diferentes mapas. Para isso foi criada uma técnica de coordenação denominada *Time Coordination*, que opera por meio da comparação entre os atributos temporais das instâncias selecionadas em uma mapa inicial com os atributos temporais das instâncias de outros mapas. A comparação é feita por meio de um operador relacional (e.g., maior, menor, menor ou igual), o qual é fornecido pelo usuário durante o processo de coordenação, isto é, após a seleção de algumas instâncias em um mapa inicial.



(a) Seleção de um artigo da conferência IV.



(b) Destaque de artigos da conferência InfoVis.

Figura 5.11: Exemplo de aplicação da técnica *Time Coordination*. No mapa apresentado à esquerda são destacados os documentos cujo atributo "ano" é maior do que o atributo "ano" daqueles selecionados em no mapa apresentado à direita.

Um exemplo de execução dessa técnica de coordenação baseada em atributos temporais é apresentado na Figura 5.11. O mapa da esquerda foi construído com a técnica LSP a partir de uma coleção de artigos da conferência *Information Visualization Conference* (IV) (artigos dos anos de 1997 a 2008). O mapa da direita foi construído com a técnica LSP a partir de uma coleção de artigos da conferencia *IEEE Symposium on Information Visualization* (IEEE InfoVis) (artigos dos

anos de 1995 a 2008). Nesse exemplo, as instâncias utilizadas possuem apenas o atributo temporal "ano". Durante a exploração o usuário faz a seleção de algumas instâncias no mapa apresentado à esquerda na Figura 5.11(a) (mapa inicial). Após a seleção, o usuário é questionado sobre qual operador relacional será empregado na comparação entre os atributos temporais das instâncias selecionadas e os atributos temporais das instâncias do outro mapa. No caso, foi selecionado o operador "maior". O mapa apresentado à direita na Figura 5.11(b) mostra o mapeamento criado pela *Time Coordination*, revelando quais são as instâncias que possuem o valor do atributo "ano" maior do que o valor do mesmo atributo nas instâncias selecionadas no mapa inicial.

O atributo utilizado nessa técnica de coordenação é de natureza temporal, no entanto outras técnicas podem ser criadas para explorar outros tipos de atributos dos dados. Um exemplo é o atributo relacionado à autoria de artigos científicos.

A seguir é apresentada uma técnica de coordenação baseada em distância que também permite a coordenação entre mapas criados a partir de diferentes conjuntos de dados.

5.3.5 Coordenação baseada em Distância

Esta seção apresenta uma técnica de coordenação baseada em distância denominada *Distance Coordination* (Eler et al., 2008b). Por meio dessa técnica, a partir da seleção de um grupo de instâncias em uma visão origem são destacadas nas visões destino as k instâncias mais próximas daquelas inicialmente selecionadas. O cálculo da distância para identificar os k vizinhos mais próximos é realizado no espaço original.

Os atributos de um espaço multidimensional construído para representar um conjunto de dados nem sempre são os mesmos utilizados no espaço multidimensional de um outro conjunto. Por exemplo, no caso de documentos, os termos (atributos) que formam um espaço multidimensional podem ser diferentes dos termos que formam o espaço multidimensional de uma outra coleção de documentos; no caso de imagens, os algoritmos e os parâmetros utilizados para extrair características (atributos) nem sempre são os mesmos para todas as coleções de imagens. Portanto, é necessário criar um mecanismo para calcular distâncias entre instâncias representadas em espaços que possuem diferentes atributos. Alternativamente, poderia ser construído um espaço multidimensional com atributos comuns, isto é, um espaço que seja representado pelos mesmos atributos. Atualmente, o segundo caso é a opção explorada pela *Distance Coordination*.

A Figura 5.12 apresenta o pré-processamento realizado para criar novos espaços multidimensionais com atributos comuns entre dois diferentes espaços. Primeiramente, é necessário identificar quais são os atributos comuns entre os diferentes espaços multidimensionais. Em seguida, novos espaços são criados com esses atributos comuns. Finalmente, é possível calcular as distâncias entre as instâncias representadas pelos novos espaços utilizando métricas tradicionais (e.g., Euclideana, cosseno, *City Block*).

A definição de uma *Distance Coordination* exige que o usuário forneça dois parâmetros, o valor de k e a métrica de distância empregada para calcular as distâncias para identificar os vizinhos mais

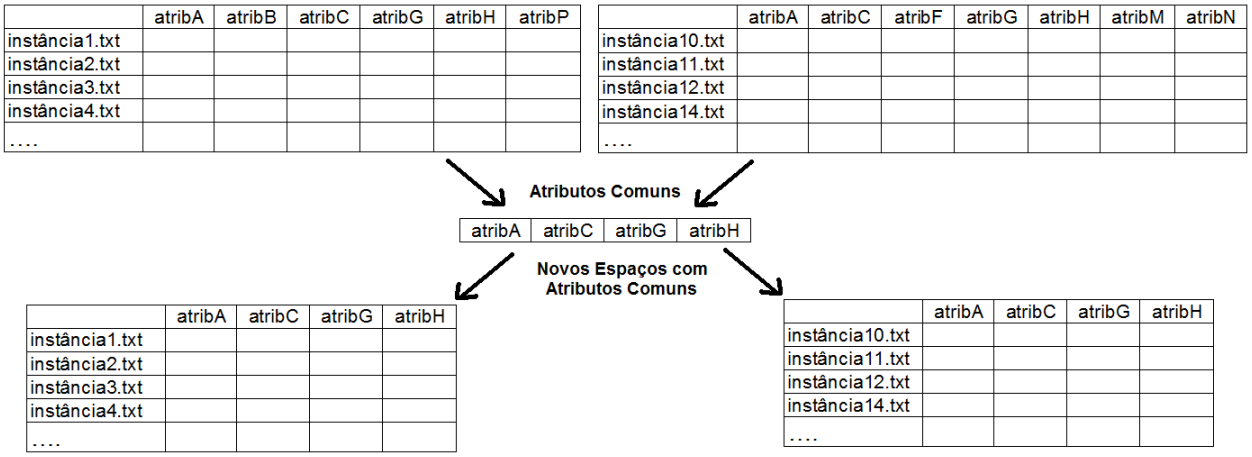
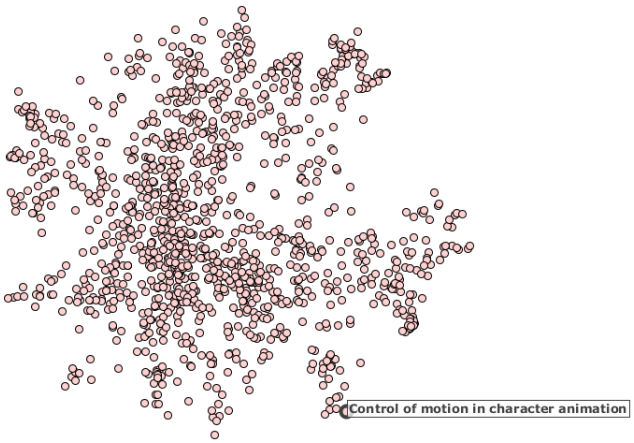


Figura 5.12: Construção de novos espaços multidimensionais com atributos comuns.

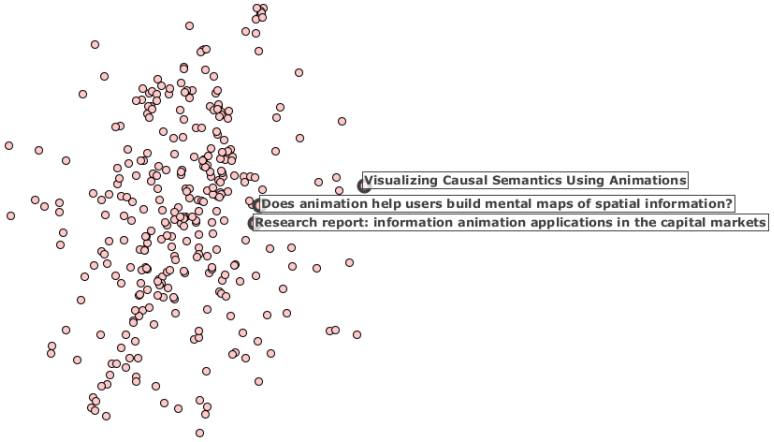
próximos. Após o usuário fornecer os parâmetros necessários, é executado o pré-processamento para criar os novos espaços multidimensionais. Em seguida, para cada instância da visão origem são identificados os k vizinhos (instâncias) mais próximos na visão destino. Ao fim desse processo, cada instância da visão origem tem um mapeamento criado para as k instâncias mais próximas da visão destino. Durante o processo de exploração, quando o usuário selecionar um grupo de instâncias na visão origem, o mapeamento criado para as instâncias selecionadas é aplicado, destacando na visão destino as k instâncias mais próximas daquelas selecionadas. É importante notar que as instâncias mapeadas na visão destino podem ser as mesmas para mais de uma instância selecionada na visão origem. Isso ocorrerá principalmente quando as instâncias selecionadas forem muito similares.

Um exemplo de aplicação da *Distance Coordination* é apresentado na Figura 5.13. Primeiramente, o usuário faz a seleção de um documento no mapa apresentado na Figura 5.13(a), o qual foi gerado da coleção de artigos da conferência *Information Visualization Conference* (IV) por meio da técnica LSP. No caso, o artigo selecionado está relacionado com animação de personagens. No mapa da Figura 5.13(b), gerado a partir de documentos da conferência *IEEE Symposium on Information Visualization* (InfoVis) por meio da técnica LSP, foram destacados os três vizinhos mais próximos do documento selecionado no mapa da Figura 5.13(a). O título dos artigos destacados foram exibidos, permitindo notar que todos eles têm alguma relação com animação em seu conteúdo. São eles, *"Visualizing Causal Semantics Using Animations"*, *"Does animation help users build mental maps of spatial information?"* e *"Research report: information animation applications in the capital markets"*.

A seguir é apresentada uma técnica de coordenação que permite combinar o mapeamento de diferentes técnicas, possibilitando a exploração de um ou vários conjuntos de dados.



(a) Seleção de um documento da conferência IV.



(b) Destaque dos três documentos mais próximos da conferência InfoVis.

Figura 5.13: Exemplo de aplicação da técnica *Distance Coordination*. Em (a) foi selecionado um documento e em (b) foram destacados os seus três vizinhos mais próximos.

5.3.6 Combinação de Coordenações

Algumas técnicas de coordenação, quando utilizadas de forma isolada, não produzem resultados que auxiliam significativamente na exploração de conjuntos de dados, como é o caso da *Time Coordination* apresentada na Seção 5.3.4. Para melhorar a exploração, foi desenvolvida uma técnica de coordenação denominada *Joint Coordination*, por meio da qual é possível combinar o mapeamento de diferentes técnicas adicionadas a ela, gerando um novo mapeamento final. Na implementação atual, esse mapeamento final contém somente as instâncias que foram mapeadas por todas as técnicas de coordenação, ou seja, a intersecção entre os diferentes mapeamentos. A seguinte expressão descreve a combinação dos diferentes mapeamentos:

$$Map_{Final} = Map_{Tecnica_1} \quad \otimes \quad Map_{Tecnica_2} \quad \otimes \quad \ldots \quad \otimes \quad Map_{Tecnica_n} \tag{5.3}$$

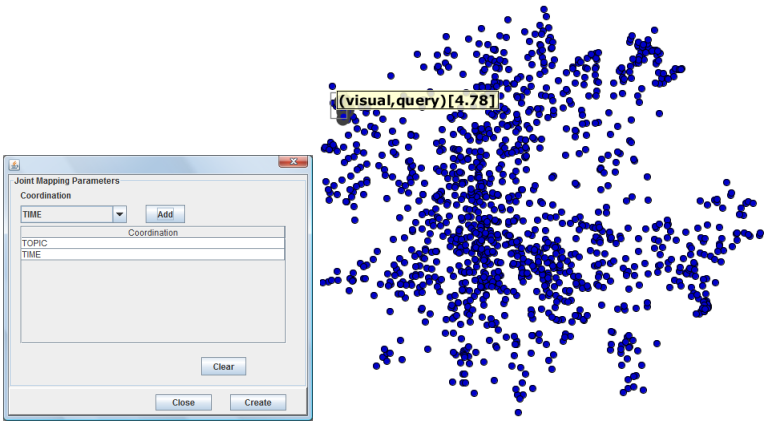
em que n é o número de técnicas de coordenação definidas e adicionadas à *Joint Coordination*, $Map_{Tecnica}$ é o mapeamento resultante de cada uma das n técnicas de coordenação adicionadas, e

Map_{Final} é o mapeamento da *Joint Coordination*, que é o mapeamento resultante da combinação ($\otimes$) dos mapeamentos de todas as técnicas de coordenação envolvidas. Apesar de atualmente ser utilizada a intersecção dos mapeamentos, é possível implementar outras formas de combinação, como a união.

Para criar uma *Joint Coordination* o usuário deve selecionar as técnicas de coordenação que deseja combinar. Para cada técnica de coordenação selecionada o usuário deverá fornecer os parâmetros de criação da técnica, caso ela necessite. Após essa configuração, a técnica de coordenação é criada e adicionada à *Joint Coordination*. A sua execução também é realizada durante o processo de exploração, a partir de um grupo de instâncias selecionadas em um mapa inicial, destacando no mapa destino as instâncias mapeadas. Nesse processo, cada técnica de coordenação criará o seu mapeamento específico entre as visões envolvidas, como se fosse isoladamente executada. Ao fim da execução de todas as técnicas de coordenação envolvidas, os diferentes mapeamentos criados por cada uma delas são combinados para gerar o mapeamento da *Joint Coordination*.

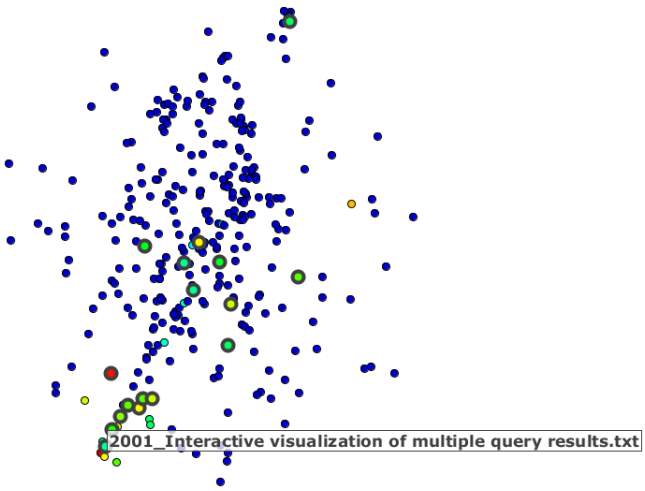
Um exemplo da *Joint Coordination* é ilustrado na Figura 5.14. A Figura 5.14(a) apresenta a definição de uma *Joint Coordination*, na qual foram definidas e adicionadas as técnicas *Topic Coordination* e *Time Coordination*. Primeiramente são selecionados alguns documentos no mapa de similaridade gerado pela técnica LSP dos documentos da conferência IV (Figura 5.14(b)). A seleção é transmitida à *Joint Coordination*, que executa as duas técnicas de coordenação adicionadas. A primeira técnica a ser executada é a *Topic Coordination*, que gerou um tópico que será propagado para o outro mapa. Em seguida é executada a *Time Coordination*, que questiona o usuário sobre qual operador relacional será empregado para criar o mapeamento entre as instâncias. Neste exemplo foi selecionado o operador maior ("$>$"). Ao fim da execução das técnicas adicionadas, o mapeamento de cada uma delas é combinado, gerando o mapeamento da *Joint Coordination*. No mapa da Figura 5.14(c), que é um mapa de similaridade gerado pela técnica LSP a partir de uma coleção de artigos científicos da conferência IEEE InfoVis, são destacados com borda em negrito os documentos cobertos pelo tópico e cuja data é maior do que a data dos documentos selecionados no mapa da Figura 5.14(b).

O mapa da Figura 5.14(c) também apresenta o mapeamento individual das técnicas de coordenação empregadas na *Joint Coordination*, isto é, o mapeamento da *Topic Coordination* (informação de cor) e o mapeamento da *Time Coordination* (bordas em negrito). No caso, a cor azul indica nenhuma ocorrência dos termos do tópico e a cor vermelha indica máxima ocorrência dos termos do tópico. No mapeamento final também é possível ver os documentos que são mapeados pelo tópico e cuja data não é maior do que a data dos documentos selecionados na visão origem. Esses são os documentos com cor diferente de azul (nenhuma ocorrência do tópico) e que não estão destacados pela borda em negrito. É importante notar que esse destaque por meio de cor é peculiar da *Topic Coordination*, o que permite que os mapeamentos das duas técnicas empregadas fossem individualmente identificados. No entanto, se fossem combinados os mapeamentos de duas técnicas de coordenação que utilizam somente a borda em negrito como destaque, seria possível ver somente o mapeamento final da *Joint Coordination*.



(a) Ajuste de parâmetros da *Joint Coordination*.

(b) Seleção na coleção de documentos da conferência IV.



(c) Destaque dos documentos da conferências InfoVis.

Figura 5.14: Exemplo da *Joint Coordination* definida em (a) com as técnicas *Topic Coordination* e *Time Coordination*. A partir da seleção em uma visão origem (b) e são destacados na visão destino (c) os documentos cobertos pelo tópico e cuja data é maior do que a data dos documentos selecionados em (b).

5.4 Considerações Finais

A ferramenta *Projection Explorer* (PEx) (Paulovich et al., 2007) provê um arcabouço de técnicas para gerar mapas de similaridade por meio de projeções multidimensionais e técnicas de posicionamento de pontos baseado em árvore. Além disso, algumas ferramentas de interação auxiliam o usuário no processo de exploração de coleções de documentos. Com o intuito de criar novas formas de exploração e associação entre mapas de similaridade, foram propostas técnicas de coordenação entre mapas de similaridade, as quais foram incorporadas à PEx. Essas técnicas permitem novas formas de explorar múltiplas coleções de documentos, o que até então não era possível na PEx e em outros sistemas. Por meio delas o usuário pode utilizar o conhecimento de uma coleção já explorada para descobrir novas relações em coleções ainda desconhecidas.

De um modo geral, as técnicas de coordenação existentes na literatura são diferentes das criadas neste trabalho para exploração de mapas de similaridade. As técnicas que mais se assemelham são aquelas aplicadas para apoiar a exploração de coleções de documentos, as quais utilizam representações visuais que relacionam palavras, tópicos ou ontologias extraídas de um conjunto de dados (Seeling e Becks, 2004a; Chang et al., 2007; Spoerri, 2007; Sabol et al., 2009). Essas representações são coordenadas com outras visões que representam uma coleção de documentos por meio de um mapa de similaridade, destacando os documentos com base em palavras, tópicos ou ontologias de um domínio. No entanto, as técnicas apresentadas neste capítulo permitem o controle do processo de coordenação, modificando quais são os documentos envolvidos na coordenação. Por exemplo, considerando a coordenação baseada em tópicos, diferentes seleções coordenadas resultam na identificação de diferentes tópicos e, consequentemente, de diferentes mapeamentos.

É notável que o arcabouço apresentado pela PEx é aplicável a outros tipos de conjuntos de dados, pois as técnicas empregadas para gerar os mapas de similaridade necessitam de um espaço multidimensional ou das relações de similaridade entre as instâncias. A exploração desses mapas também pode ser apoiada pelas técnicas apresentadas neste capítulo. Além disso, outras técnicas podem ser propostas para identificar relações específicas aos conjuntos explorados. Como apresentado neste capítulo, também é possível utilizar representações diferentes de um conjunto de dados para melhor entendê-lo, como os exemplos mostrados com imagens e texto; e também os apresentados por Andery (Andery, 2010) para exploração de redes sociais.

Conforme apresentado no Capítulo 3, existem poucos trabalhos que utilizam múltiplas visões coordenadas para auxiliar a exploração de coleções de imagens. Nesse sentido, a PEx foi adaptada no contexto deste trabalho, resultando na *Projection Explorer for Images* (PEx-Image) (Eler et al., 2008a, 2009a), que é uma ferramenta e abordagem que utiliza mapas de similaridades e múltiplas visões coordenadas para auxiliar na exploração de coleções de imagens. A PEx-Image e suas aplicações são descritas no próximo capítulo. Nessas aplicações é utilizada a maioria das técnicas de coordenação apresentadas neste capítulo.

Para que as técnicas de coordenação apresentadas nesse capítulo fossem implementadas e unificadas, foi necessário utilizar um modelo extensível de coordenação que descrevesse como as técnicas são definidas e como elas se comportam. O modelo desenvolvido, apresentado no capítulo anterior, é uma adaptação daquele proposto por Boukhelifa e Rodgers (Boukhelifa e Rodgers, 2003), o qual foi descrito na Seção 3.3. Este modelo permitir que as técnicas de coordenação desenvolvidas também pudessem ser aplicadas para outros tipos de dados, como as aplicação para imagens que serão apresentadas no capítulo seguinte.

Além das aplicações das técnicas de coordenação para explorar coleções de documentos e coleções de imagens, o próximo capítulo também apresenta uma ramificação importante gerada por este trabalho, a qual auxilia a exploração de dados volumétricos. Essa ramificação envolve mapas de similaridade e a coordenação dos mapas com alguma visualização do espaço do objeto, o qual representa os dados volumétricos.

CAPÍTULO

6

Aplicações para Documentos e Imagens

6.1 Considerações Iniciais

Este capítulo tem o objetivo de descrever aplicações das técnicas de coordenação descritas nos Capítulos 5 e 4, fornecendo evidência da sua aplicabilidade. Elas envolvem múltiplas visões coordenadas de mapas de similaridade gerados a partir de coleções de documentos, de coleções de imagens e de dados volumétricos. Os mapas foram gerados por meio de técnicas de projeção multidimensional e uma técnica de posicionamento de pontos baseada em árvore, para os dados textuais e dados de imagens; para os dados volumétricos foi utilizada uma nova técnica denominada *Piecewise Least Squares Projection* (P-LSP) (Paulovich et al., 2010). A ferramenta *Projection Explorer* (Paulovich et al., 2007), estendida para admitir o modelo de coordenação implementado, é utilizada para a exploração de coleções de documentos. Para a exploração de coleções de imagens foi construída uma nova ferramenta denominada *Projection Explorer for Images* (PEx-Image) (Eler et al., 2008a, 2009a), a qual é uma adaptação da *PEx*. A *PEx-Image* implementa técnicas de interação sobre os mapas de similaridade para auxiliar a exploração de coleções de imagens. Também foi desenvolvida uma ferramenta que auxilia a exploração de dados volumétricos a partir da coordenação de mapas de similaridade construídos de dados volumétricos. Nessa ferramenta, a seleção no mapa de similaridade reflete no espaço do objeto, que é observado por meio de uma técnica de visualização volumétrica.

6.2 Exploração de Coleções de Documentos

Esta seção apresenta aplicações que envolvem todas as técnicas de coordenação propostas neste trabalho. Essas técnicas de coordenação são aplicadas para explorar as seguintes coleções de documentos:

- **CBR-ILP-IR**: 574 artigos científicos de três áreas da inteligência artificial (três classes). São elas, *Case Based Reasoning* (CBR), *Inductive Logic Programming* (ILP) e *Information Retrieval* (IR). Os dados empregados são: o título, a autoria, a instituição, o resumo e as referências dos artigos;
- **NEWS**: 2684 extratos de notícias publicadas na internet pelas agências Reuters, BBC, CNN e *Associated Press*. Os arquivos foram coletados durante dois dias do mês de Abril de 2006. A fonte dos dados empregados são arquivos no formato RSS, recuperados dos *websites* das agências;
- **CMV**: 66 artigos da conferência *Coordinated & Multiple Views* (CMV) (publicações entre os anos 2003 a 2008). Os dados empregados são: o título, o resumo, as palavras chave e o ano de publicação do artigo;
- **IV**: 1263 artigos científicos da conferência *Information Visualization Conference* (IV) (publicações entre os anos de 1997 a 2008). Os dados empregados são: o título, o resumo, as palavras chave e o ano de publicação do artigo;
- **InfoVis**: 315 artigos da conferência *IEEE Symposium on Information Visualization* (IEEE InfoVis) (publicações entre os anos de 1995 a 2008). Os dados empregados são: o título, o resumo, as palavras chave e o ano de publicação do artigo.

Para a construção de um mapa de similaridade a partir de uma coleção de documentos é necessário realizar algumas etapas de pré-processamento (Paulovich e Minghim, 2006). Geralmente, elas são:

1. **Eliminação de *stopwords*:** esta etapa visa eliminar termos que não interessam ao usuário (e.g., artigos e preposições). Essa lista de termos é conhecida como Lista de *Stopwords*;
2. **Aplicação de uma técnica de *stemming*:** nesta etapa, por meio da aplicação de um algoritmo de *stemming* (Porter, 1997), é possível eliminar variações de palavras, utilizando somente o radical das palavras como atributo. Por exemplo, as palavras *projection* e *projecting* poderiam ser reduzidas para o radical *project*;
3. **Cálculo de frequência dos termos:** nesta etapa são realizadas as contagens de frequência (ocorrência) dos termos no conteúdo de cada documento;

4. **Corte baseado na frequência dos termos:** nesta etapa é possível utilizar a frequência dos termos (atributos) para realizar um filtro, podendo excluir termos que aparecem pouco ou muito na coleção de documentos. Esse corte é comumente conhecido como Corte de Luhn (Luhn, 1958);

O resultado é uma matriz documentos x termos com coeficientes que representam coordenadas. Este é o resultado do modelo de espaço vetorial para textos (Salton et al., 1975).

Neste capítulo, todos exemplos de aplicação envolvendo documentos utilizam as etapas de eliminação de *stopwords* e de aplicação de uma técnica de *stemming* antes do cálculo da frequência dos termos. Após a construção do espaço vetorial multidimensional alguma técnica de projeção ou de posicionamento de pontos é aplicada para gerar o mapa. Nessa etapa uma matriz de distância pode ser calculada com base em alguma medida de similaridade, a qual tem a finalidade de estabelecer o relacionamento entre as instâncias do conjunto de dados. Essa medida é escolhida pelo usuário e pode ser empregada sobre o espaço multidimensional ou sobre o conjunto de dados. As medidas de similaridade utilizadas nos próximos exemplos serão empregadas sobre o espaço multidimensional.

A seguir serão apresentadas explorações de coleções de documentos realizadas por meio de diferentes mapas de similaridade e da coordenação entre eles.

6.2.1 Coordenação Baseada em Identificador

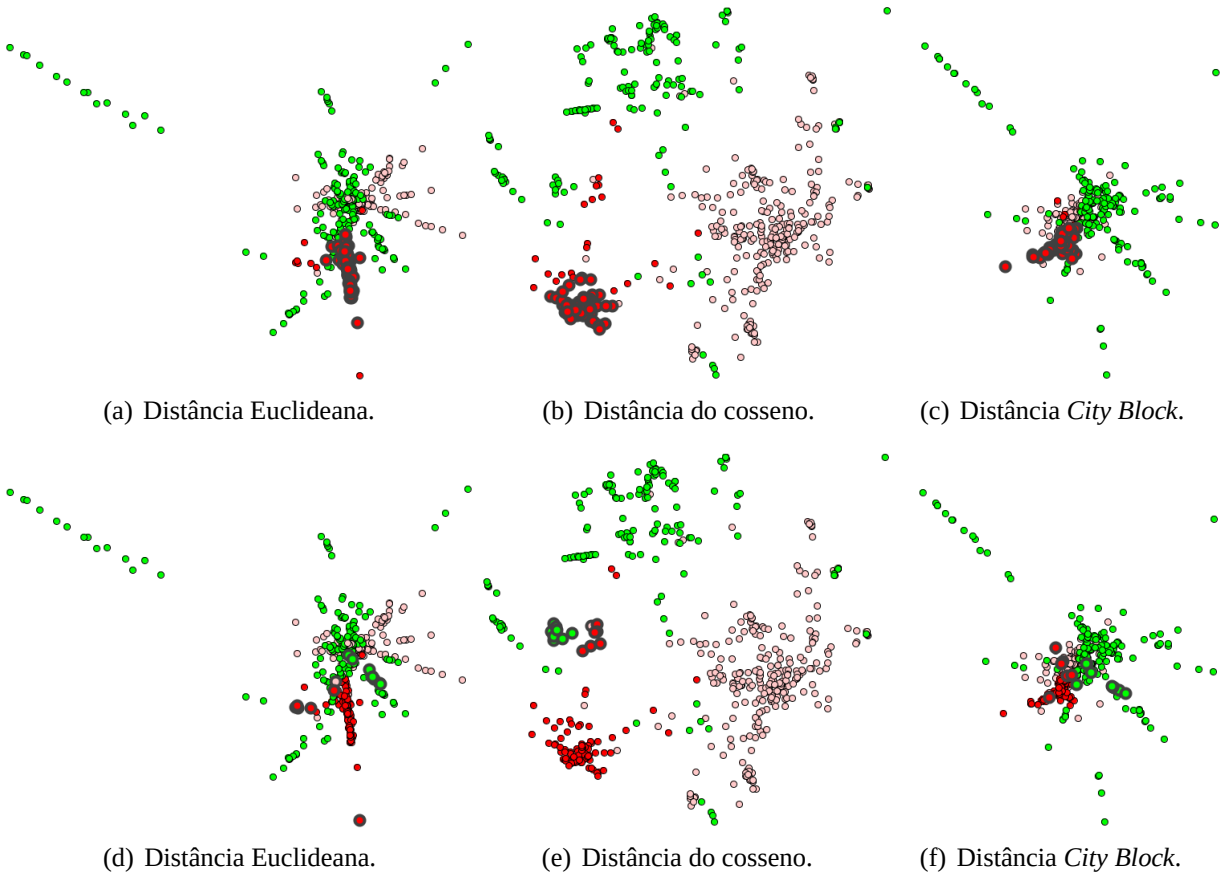
A técnica de coordenação baseada em identificador (*Identity Coordination*) é utilizada nas próximas aplicações para comparar diferentes configurações de parâmetros utilizados para gerar os mapas de similaridade. Esse tipo de análise permite confirmar qual configuração gera o mapa que melhor representa as similaridades entre as instâncias do conjunto de dados. Adicionalmente, por meio da coordenação, é possível focar em grupos específicos de um mapa e comparar o seu posicionamento em outros mapas. Também, a coordenação entre os mapas permite observar quais grupos estão contribuindo negativamente ou positivamente para a qualidade visual, conduzindo o usuário a rever o processo anterior à geração do mapa de modo a melhorar a qualidade dos grupos.

A qualidade de um mapa de similaridade pode ser subjetivamente avaliada por meio da inspeção visual, observando a formação e separação de grupos a partir de dados pré-classificados. Outra forma de avaliação é por meio de técnicas objetivas. Recentemente, Brandoli e outros (Brandoli et al., 2010) apresentaram uma maneira de utilizar o cálculo do coeficiente de silhueta para avaliar o resultado obtido de projeções multidimensionais. O coeficiente de silhueta é uma medida objetiva que revela o grau de separação e coesão de grupos. Na literatura também são apresentadas duas técnicas conhecidas como *Neighborhood Hit* (Paulovich et al., 2008) e *Neighborhood Preservation* (Paulovich e Minghim, 2008) utilizadas para avaliar a qualidade de mapas de similaridade baseados no posicionamento de pontos no plano. A técnica *Neighborhood Hit* avalia a vizinhança dos pontos de um mapa de dados pré-classificados, verificando qual a porcentagem de vizinhos

mais próximos que pertencem à mesma classe de um determinado ponto. O valor de precisão final dado pela *Neighborhood Hit* é a média das precisões de todos os pontos do mapa. A técnica *Neighborhood Preservation* avalia a preservação de vizinhança dos pontos no espaço projetado em relação à vizinhança no espaço original, ou seja, dado um ponto no mapa ela verifica qual a porcentagem de vizinhos mais próximos no espaço projetado que também são vizinhos mais próximos desse ponto no espaço original. O valor final de precisão indicado pela *Neighborhood Preservation* também é a média dos valores de precisão de todos os pontos do mapa. Essas duas técnicas são empregadas nos próximos exemplos para medir a qualidade dos mapas gerados e contrastá-la com a análise visual dos mesmos mapas.

Comparação entre diferentes de medidas de similaridade

No primeiro exemplo é utilizada a coleção *CBR-ILP-IR*. A Figura 6.1 apresenta três mapas de similaridade gerados pela técnica de projeção *Least Squares Projection* (LSP) (Paulovich et al., 2008), modificando somente a medida de similaridade utilizada para calcular a matriz de distância.



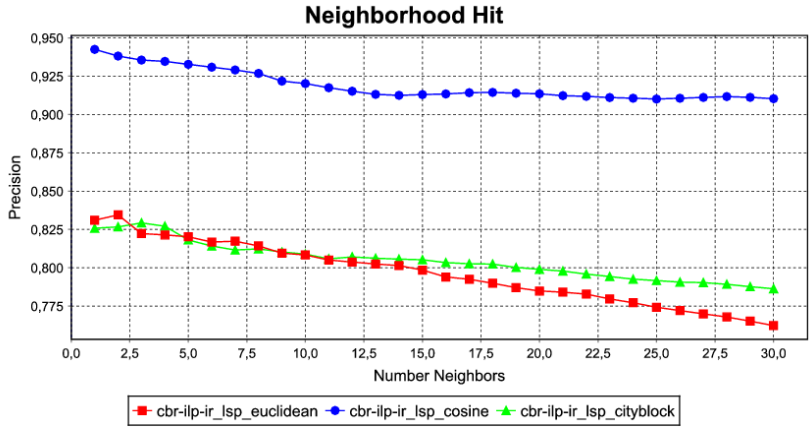
(a) Distância Euclideana. (b) Distância do cosseno. (c) Distância *City Block*.

(d) Distância Euclideana. (e) Distância do cosseno. (f) Distância *City Block*.

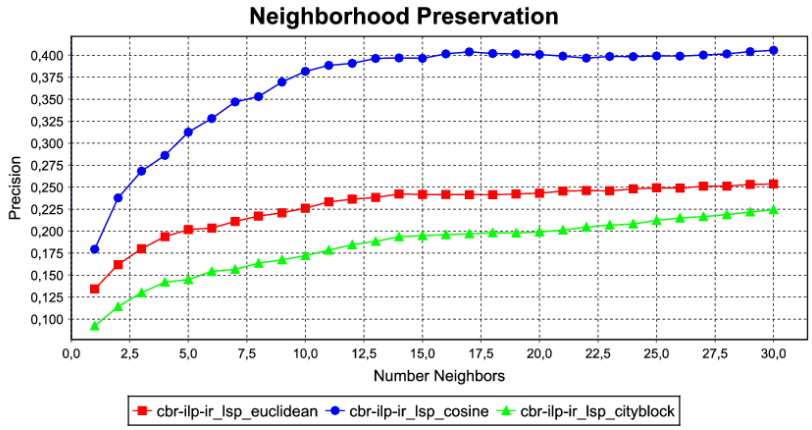
Figura 6.1: Comparação entre mapas de similaridade gerados por diferentes configurações de medidas de similaridade.

No mapa apresentado na Figura 6.1(b) foi realizada uma seleção em um grupo bem formado. Em seguida, as instâncias que possuem o mesmo identificador foram destacadas nos outros mapas por meio da coordenação baseada em identificador, conforme apresentado nas Figuras 6.1(a) e

6.1(c), comprovando que a distância do cosseno resulta em melhores grupos na geração do mapa. Nesse mesmo sentido a seleção de outros dois grupos também foi realizada no mapa apresentado na Figura 6.1(e). Após essa nova seleção, as instâncias de mesmo identificador foram destacadas nos mapas apresentados nas Figuras 6.1(d) e 6.1(f). Por meio da coordenação é possível notar que grupos mais definidos são formados no mapa que utilizou a distância do cosseno como medida de similaridade, enquanto que nos outros mapas as instâncias ficam mais espalhadas ou formam um grande grupo, não discriminando bem as instâncias do conjunto de dados. Esse exemplo ilustra a capacidade dos mapas, em conjunto com as coordenações, de apoiar a escolha de uma medida de similaridade apropriada.



(a) *Neighborhood Hit*.



(b) *Neighborhood Preservation*.

Figura 6.2: Comparação da precisão de diferentes medidas de similaridade utilizadas para gerar os mapas da Figura 6.1. As medidas utilizadas foram a distância Euclideana (vermelho), a distância do cosseno (azul) e a distância *City Block* (verde).

A Figura 6.2 apresenta dois gráficos que descrevem os valores de precisão calculados por meio das técnicas *Neighborhood Hit* e *Neighborhood Preservation* a partir dos mapas apresentados na Figura 6.1. O gráfico de precisão da técnica *Neighborhood Hit* indica que a distância do cosseno é a melhor opção para criar grupos discriminativos, isto é, grupos que contenham instâncias de uma mesma classe. Já o gráfico de precisão da técnica *Neighborhood Preservation* indica que a

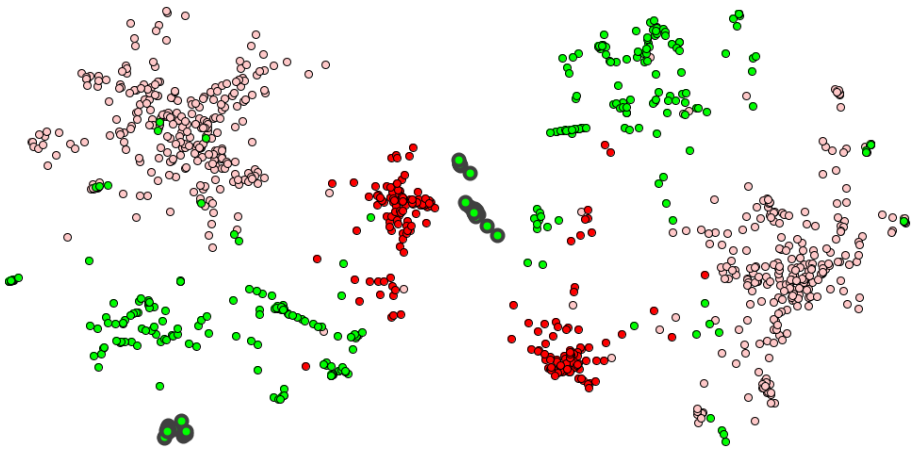
distância do cosseno é a melhor opção para preservar as relações de similaridade entre as instâncias do espaço original quando elas são posicionadas no espaço projetado. As técnicas de avaliação apresentam um valor que indica a qualidade geral da projeção, deixando de lado quais são os grupos que contribuem ou não para a qualidade do resultado visual. Por meio da coordenação por identidade é possível avaliar grupos específicos para comprovar qual é o melhor mapa em vários graus de vizinhança. Além disso é possível verificar quais são os melhores ou piores grupos, podendo conduzir o usuário a rever a configuração de parâmetros utilizada para gerar os mapas.

Comparação entre diferentes espaços multidimensionais

No exemplo anterior o espaço multidimensional utilizado para construir os mapas foi reduzido por meio da aplicação de um corte baseado nas freqüências dos termos. Este corte foi definido com limite inferior igual a 100 e limite superior igual a 2745, o valor da frequência máxima de termos para esse conjunto. Neste novo exemplo, apresentado na Figura 6.3, são comparados dois espaços multidimensionais construídos a partir da base *CBR-ILP-IR* com e sem cortes. No mapa apresentado na Figura 6.3(a) foram utilizados todos os termos para a construção do espaço multidimensional, aplicando somente as etapas de eliminação de *stopwords* e de redução dos termos por meio de um algoritmo de *stemming*. O mapa apresentado na Figura 6.3(b) é o mesmo mapa ilustrado na Figura 6.1(b), isto é, com os cortes mínimo e máximo em 100 e 2745.

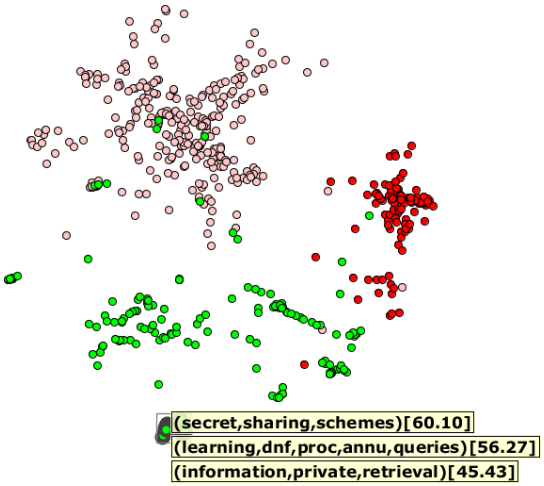
No mapa apresentado na Figura 6.3(a) foi selecionado um grupo compacto. Por meio da coordenação baseada em identificador, que destacou as mesmas instâncias no outro mapa (Figura 6.3(b)), é possível notar que as instâncias do grupo selecionado são posicionadas em dois grupos próximos no mapa apresentado na Figura 6.3(b). Por meio da exploração coordenada desses dois mapas observa-se que utilizando todos os termos do espaço multidimensional os grupos tendem a ser mais coesos, o que pode melhorar a qualidade visual; enquanto que o mapa construído com o espaço reduzido tende a discriminar melhor subgrupos de instâncias.

A Figura 6.4 apresenta a precisão calculada por meio da técnica *Neighborhood Hit* a partir dos mapas apresentados nas Figuras 6.3(a) e 6.3(b). É possível verificar que os dois mapas possuem grupos com alta precisão, sendo todos os valores acima de $0,91$. Apesar do mapa que utiliza o espaço multidimensional completo (Figura 6.3(a)) apresentar uma melhor precisão média, o mapa gerado com o espaço multidimensional reduzido (Figura 6.3(b)) é mais interessante do ponto de vista da exploração, pois ele tende a criar grupos mais discriminativos. Isso é comprovado analisando os grupos por meio da detecção de tópicos baseada em covariância (Eler et al., 2009b) (ver Seção 5.3.2). Neste exemplo, os tópicos do grupo selecionado no mapa ilustrado na Figura 6.3(c) apresentam os mesmos tópicos detectados em dois grupos distintos apresentados no mapa apresentado na Figura 6.3(d). Esse exemplo confirma que é preferível realizar uma redução do espaço dimensional por meio de cortes de frequência (e.g., Cortes de Luhn) a utilizar todos os termos para formar o espaço multidimensional, pois o espaço reduzido tende a criar grupos mais discriminativos. A técnica *Neighborhood Preservation* não foi utilizada nesse exemplo, pois diferentes espaços multidimensionais foram empregados para gerar os mapas de similaridade.

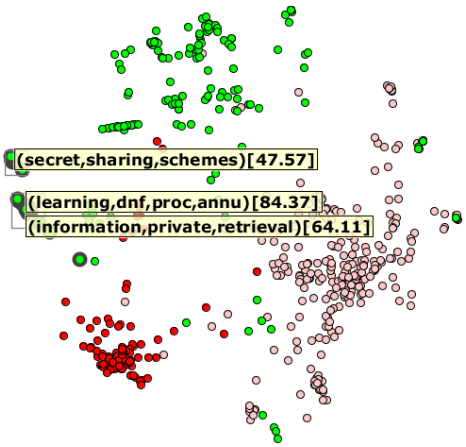


(a) Mapa construído de um espaço multidimenional com todos os termos.

(b) Mapa construído de um espaço multidimensional com número reduzido de termos.



(c) Tópicos detectados no grupo destacado em (a).



(d) Tópicos detectados nos grupos destacados em (b).

Figura 6.3: Comparação de mapas gerados por diferentes espaços multidimensionais.

Exploração simultânea de mapas gerados por diferentes técnicas

O próximo exemplo, apresentado na Figura 6.5, utiliza a técnica de coordenação baseada em identificador para simultaneamente empregar as técnicas *Least Squares Projection*(LSP) (Paulovich et al., 2008) e *Neighbor-Joining*(NJ) (Cuadros et al., 2007) na exploração de mapas de similaridade. Os mapas foram gerados com os mesmos parâmetros empregados na construção do espaço multidimensional reduzido que foi utilizado para gerar o mapa apresentado na Figura 6.3(b). A coordenação foi utilizada para destacar grupos e auxiliar a comparação das técnicas. Os grupos indicados pelas letras **A** e **D** foram selecionados no mapa apresentado na Figura 6.5(b). No caso, foram selecionados dois ramos completos, sendo possível identificar que eles correspondem aos grupos bem compactos no mapa apresentado na Figura 6.5(a). Já os grupos indicados pelas letras **B** e **C** foram selecionados no mapa apresentado na Figura 6.5(a) e destacados no mapa apresentado na Figura 6.5(b). É possível notar que as duas técnicas tiveram um comportamento parecido, posicionando próximas as instâncias mais similares. A diferença principal entre as duas técnicas

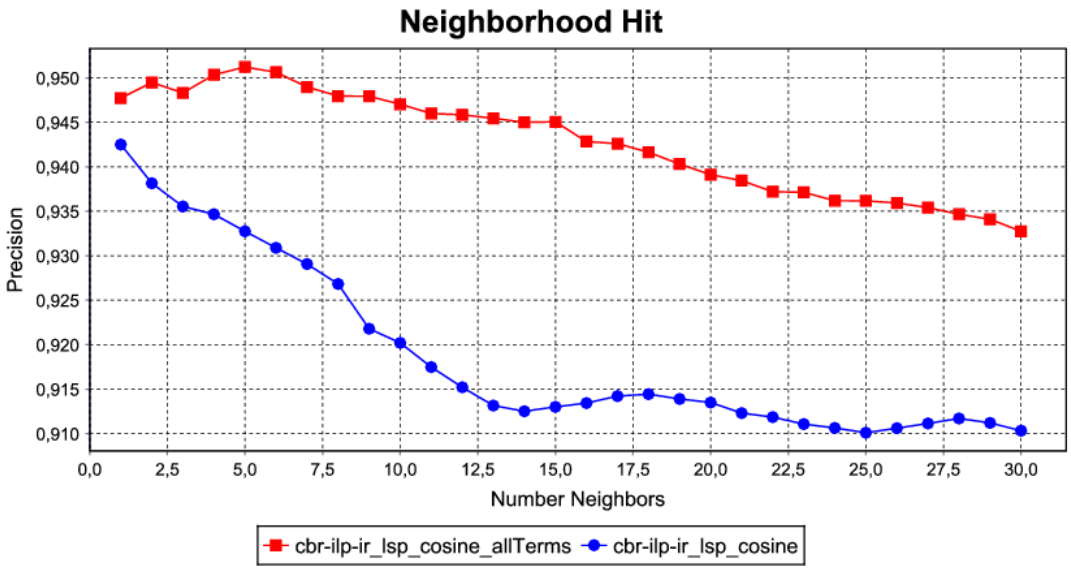


Figura 6.4: Gráfico dos valores de *Neighborhood Hit* calculado a partir dos mapas apresentados nas Figuras 6.3(a) e 6.3(b), coloridos respectivamente com as cores vermelho e azul.

é a forma particular que cada uma tem em posicionar os pontos no plano, isto é, no caso do mapa construído pela técnica de projeção, as instâncias foram posicionadas em um mesmo grupo; no caso do mapa construído por posicionamento de pontos baseado em árvore, as instâncias foram posicionadas em um mesmo ramo. Na abordagem baseada em árvore os ramos auxiliam a subdividir os grupos correspondentes em subgrupos, gerando uma hierarquia de similaridade e ajudando a focar em subgrupos de interesse.

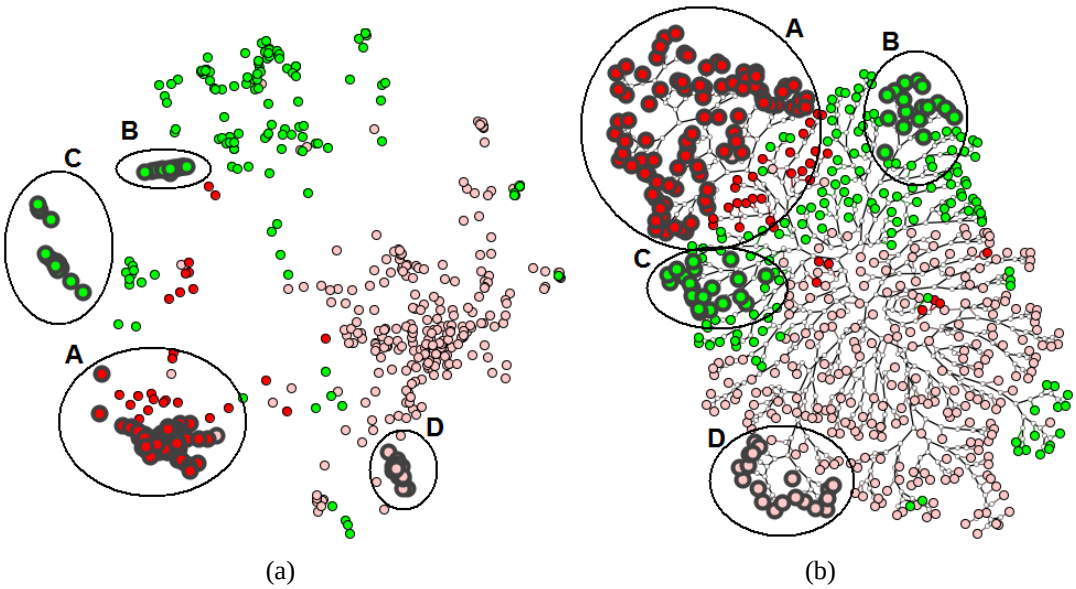


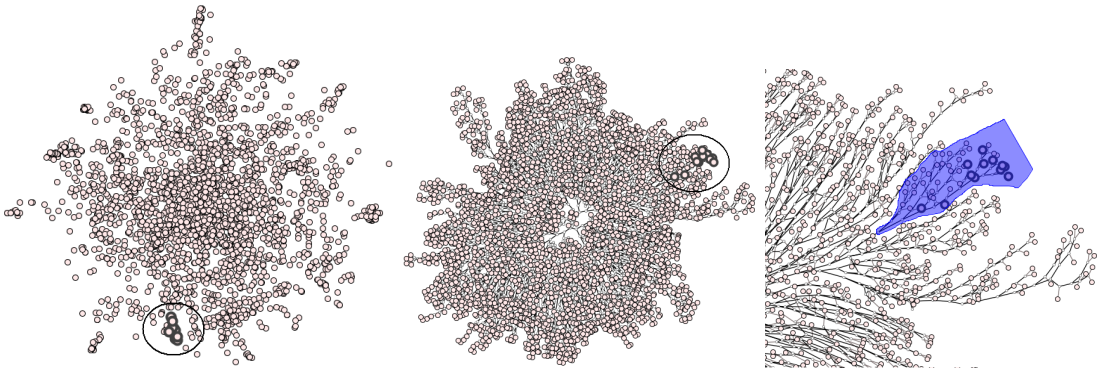
Figura 6.5: Comparação entre as técnicas LSP (a) e NJ (b) para gerar mapas de similaridade.

Com o uso da coordenação entre essas duas técnicas, é possível utilizar um mapa construído por meio de uma técnica de projeção, para localizar grupos compactos, e complementar a exploração

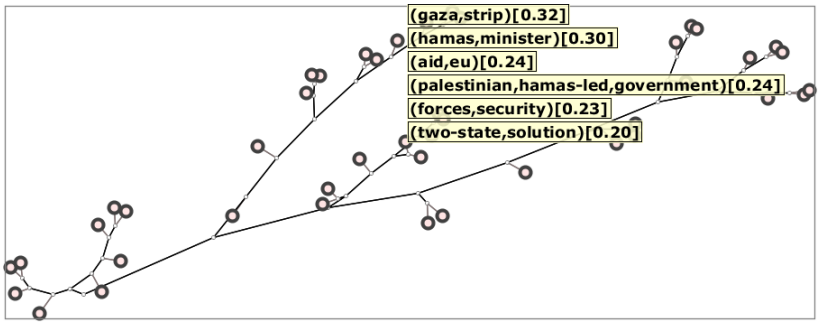
por meio de uma técnica baseada em árvore para estabelecer níveis de similaridade dentro dos grupos. Esse tipo de exploração é utilizado no próximo exemplo.

Mapa de notícias: exploração simultânea de diferentes mapas

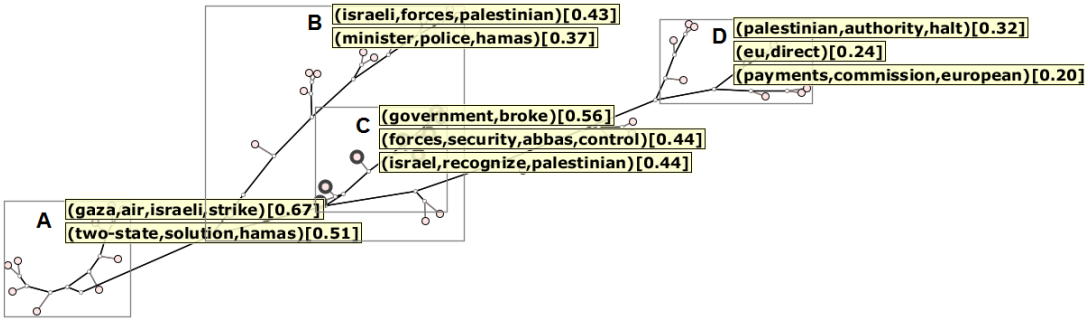
O próximo exemplo ilustra a exploração de diferentes mapas gerados a partir da coleção *NEWS*.



(a) (b) (c)



(d)



(e)

Figura 6.6: Exploração coordenada de notícias usando as técnicas LSP (a) e NJ (b) simultaneamente. Em (c) é realizada a localização de um grupo de interesse. Em (d) são exibidos os tópicos cobertos pelo ramo destacado em (c). Em (e) são apresentados sub-tópicos localizados no ramo destacado em (c).

Neste exemplo, é notável que as instâncias agrupadas no mapa construído por meio de uma técnica de projeção LSP continuam agrupados nos ramos da árvore. A Figura 6.6(c) apresenta um detalhamento com a seleção do ramo completo, no qual estão as instâncias destacadas por meio da técnica de coordenação. Esse ramo foi recortado para ser explorado por meio de uma técnica

de detecção de tópicos, a qual revelará a hierarquia de similaridade construída pela árvore. Esta hierarquia é baseada no conteúdo dos documentos. A Figura 6.6(d) mostra o tópico detectado sobre o conteúdo de todos os documentos do ramo. Já a Figura 6.6(e) apresenta os tópicos detectados na seleção individual dos ramos, revelando a hierarquia de similaridade construída pela árvore. É difícil notar essa hierarquia quando uma técnica de projeção multidimensional é exclusivamente utilizada. Na Figura 6.6(e) os documentos da seleção indicada pela letra **A** relatam um ataque aéreo na faixa de gaza e o acordo de paz do Hamas entre Israel e Palestina (*two-state solution*). Os documentos da seleção indicada pela letra **B** relatam o pronunciamento do primeiro ministro da Palestina sobre a crise e também relata a prisão de um ministro do Hamas pela polícia de Israel. Os documentos da seleção indicada pela letra **C** relatam que o primeiro ministro da Palestina tomará o controle das forças de segurança da Palestina, colocando o governo do Hamas contra o presidente da Palestina Mahmoud Abbas; eles também relatam que o Hamas diz que o governo da Palestina está falido e precisará de auxilio; e que o primeiro ministro do Hamas se recusa a reconhecer Israel. Finalmente, os documentos da seleção indicada pela letra **D** relatam que a União Européia (EU) e os Estados Unidos suspenderam o auxílio ao governo Palestino, colocando as autoridades Palestinas mais próximas da falência.

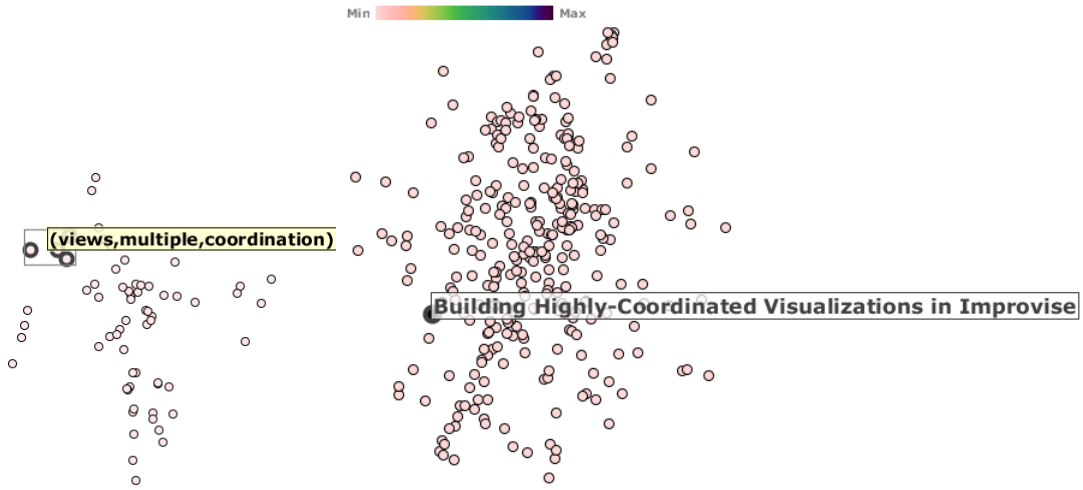
Por meio de uma simples técnica de coordenação por identificador é possível explorar e comparar diferentes mapas de similaridade, como nos exemplos anteriormente apresentados. No entanto, quando diferentes coleções de documentos estão envolvidas, outras técnicas são necessárias, as quais devem ser capazes de coordenar visões de diferentes coleções de documentos. A coordenação baseada em tópico, apresentada a seguir, é uma dessas técnicas.

6.2.2 Coordenação Baseada em Tópicos

Um exemplo de aplicação para explorar diferentes coleções de documentos é dado com a técnica de coordenação *Topic Coordination* (Eler et al., 2008b) (Figura 6.7), que foi introduzida na Seção 5.3.2. Todos os mapas apresentados na Figura 6.7 foram gerados por meio da técnica *Least Squares Projection* (LSP) (Paulovich et al., 2008). Essa técnica foi aplicada sobre as bases *CMV*, *InfoVis* e *IV* para gerar, respectivamente, os mapas das Figuras 6.7(a), 6.7(b) e 6.7(c).

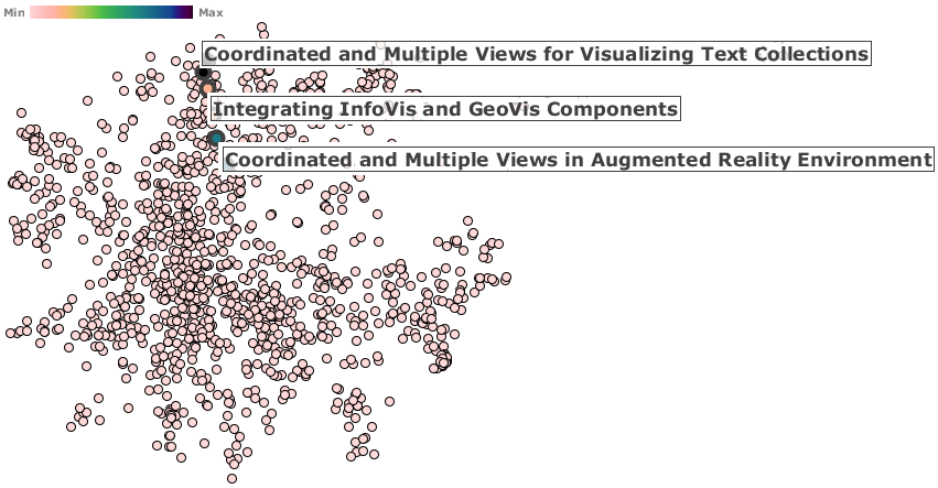
Durante a exploração o usuário selecionou alguns documentos do mapa apresentado na Figura 6.7(a). Após a seleção, o algoritmo de detecção de tópicos baseado em covariância (Cuadros et al., 2007) (ver Seção 5.3.2) detectou um tópico principal formado pelos termos *"views"*, *"multiple"* e *"coordination"*. Esse tópico é propagado para outras duas visões por meio da *Topic Coordination*, por meio da qual é possível descobrir quais são os documentos das outras bases que são cobertos pelo tópico detectado. Essa técnica de coordenação destaca os documentos por meio da informação de cor, a qual indica média dos valores de ocorrência de cada termo do tópico principal no conteúdo de cada documento. O mapeamento de cor pode variar de acordo com a escala de cor definida pelo usuário. No mapa apresentado na Figura 6.7(b) é destacado um único documento coberto pelo tópico. O documento é um artigo intitulado *"Building Highly-Coordinated Visualizations in Im-*

provise", que é um dos artigos sobre o sistema *Improvise*. Por outro lado, no mapa apresentado na Figura 6.7(c) são destacados três documentos cobertos pelo tópico detectado, todos eles relacionados com múltiplas visões coordenadas. Esses documentos são intitulados *"Coordinated and Multiples Views for Visualizing Text Collections"*, *"Integrating InfoVis and GeoVis Components"* e *"Coordinated and Multiple Views in Augmented Reality Environment"*. Isso mostra que na conferência *IV* há mais documentos relacionados ao tópico *"views"*, *"multiple"*, *"coordination"* do que na conferência *InfoVis*.



(a) Seleção em documentos da conferência CMV.

(b) Destaque dos documentos da conferência InfoVis.

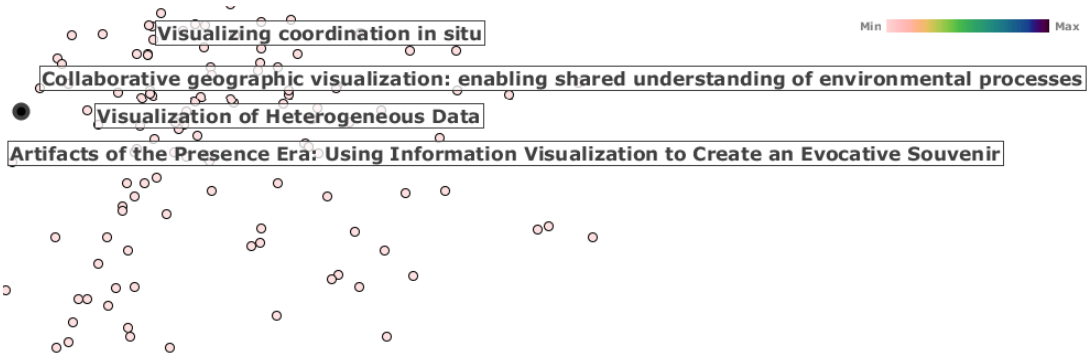


(c) Destaque dos documentos da conferência IV.

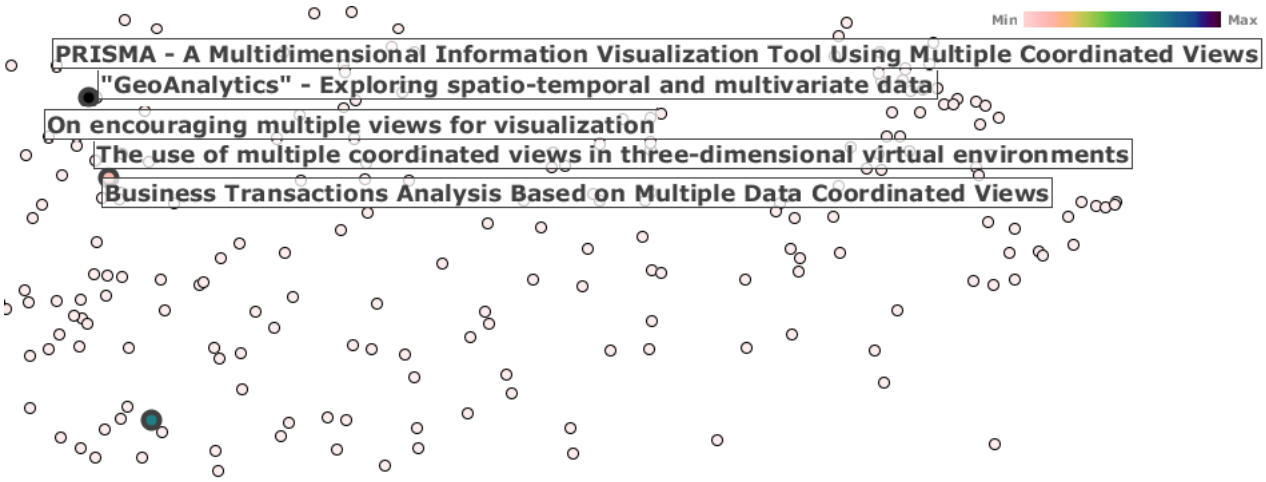
Figura 6.7: Exemplo de aplicação da técnica *Topic Coordination*. No mapa de similaridade ilustrado em (a) foram selecionados alguns documentos e gerado o tópico *"views"*, *"multiple"*, *"coordination"*. Nos mapas ilustrados em (b) e (c) foi destacada a cobertura do tópico gerado em (a). A informação de cor é utilizada para representar a frequência da cobertura do tópico.

A *Topic Coordination* conduz o usuário a focar no grupo de documentos próximos daqueles destacados. A Figura 6.8 mostra um destaque dos mapas apresentados nas Figuras 6.7(b) e 6.7(c). O título dos documentos é exibido para apresentar quais estão próximos daqueles desta-

cados pela *Topic Coordination*. Os títulos dos documentos do mapa apresentado na Figura 6.8(a) mostram que a vizinhança do documento destacado não está relacionada com o tópico *"views"*, *"multiple"*, *"coordination"*. Somente um único documento mais afastado, intitulado *"Visualizing coordination in situ"*, tem relação com o documento destacado. Por outro lado, no mapa apresentado na Figura 6.8(b), os títulos dos documentos vizinhos àqueles destacados pela coordenação estão bem relacionados com o tópico detectado na visão origem (Figura 6.7(a)).



(a) Zoom no mapa da Figura 6.7(b).



(b) Zoom no mapa da Figura 6.7(c).

Figura 6.8: Detalhamento da vizinhança dos documentos destacados nas Figuras 6.7(b) e 6.7(c). São exibidos os títulos de alguns dos documentos que estão na vizinhança daqueles destacados pela *Topic Coordination*.

Como apresentado nos exemplos anteriores, essa técnica de coordenação baseada em tópico é útil para explorar coleções de documentos ainda desconhecidas. Por meio dessa técnica é possível descobrir e focar em grupos de documentos presentes em outras bases, a partir de documentos de interesse de uma base já conhecida.

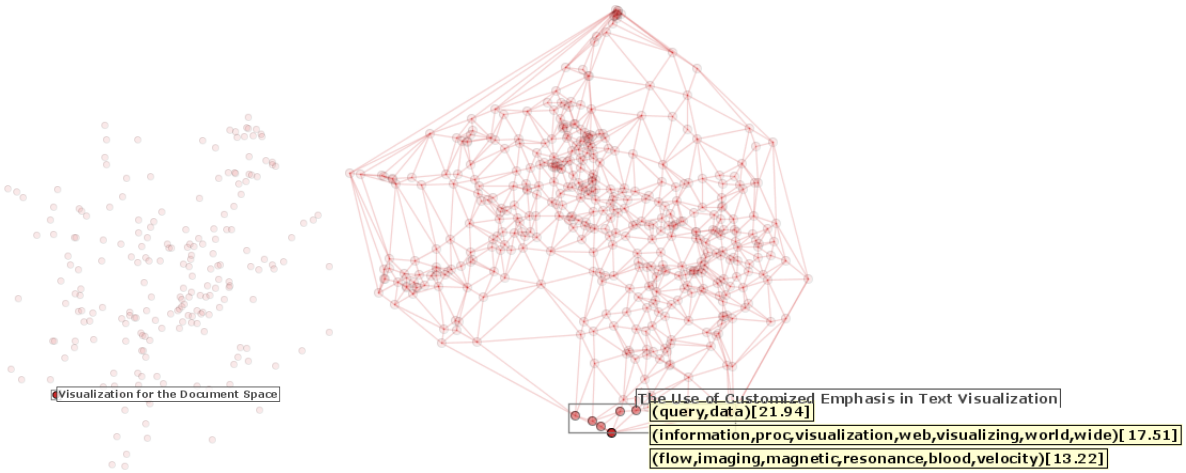
A seguir é apresentada a aplicação de uma técnica que utiliza a detecção de tópicos em mapas construídos a partir de coleções de documentos que variam no tempo.

6.2.3 Coordenação Baseada em Tópicos no Tempo

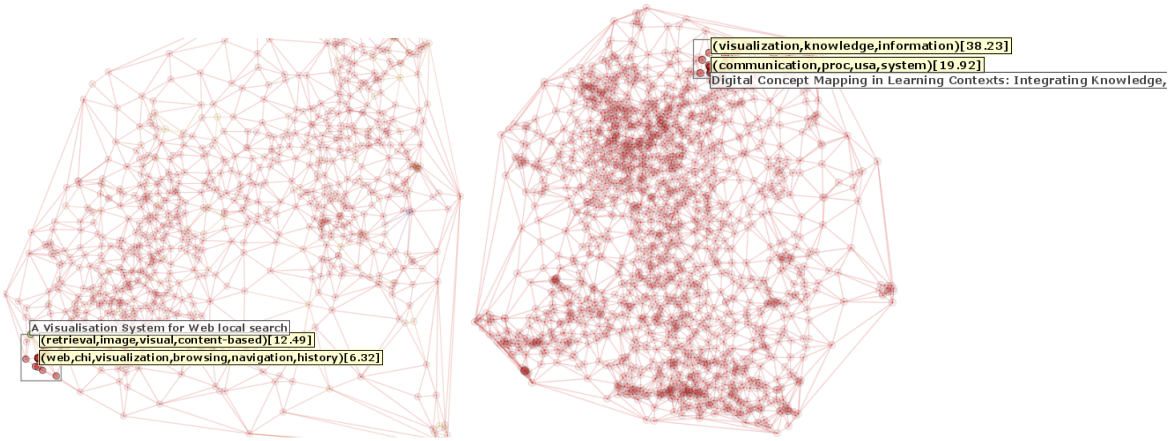
Muitas coleções de documentos crescem conforme o tempo passa. Uma informação de interesse é a evolução de temas relacionados a um artigo ou a um grupo de artigos ao longo do tempo. Na aplicação seguinte foram gerados mapas de alguns dos principais eventos em visualização, isto é, as conferências *Coordinated & Multiple Views* (CMV) (artigos de 2003 a 2008), a *Information Visualization Conference* (IV) (artigos de 1997 a 2007), o *IEEE Symposium on Information Visualization* (IEEE InfoVis) (artigos de 2000 a 2006) e a *IEEE Conference on Visualization* (IEEE Vis) (artigos de 1990 a 2007). Esta coleção de documentos contém 2408 artigos. Note que eles não fazem parte das bases descritas no início da Seção 6.2. A partir dessa coleção de documentos foram gerados mapas por meio da técnica *Least Squares Projection* (LSP) (Paulovich et al., 2008). Em todos os mapas foi utilizada a técnica de detecção de tópicos baseada na covariância dos termos (Eler et al., 2009b) (ver Seção 5.3.2). Nessa aplicação a vizinhança é dada com base em uma triangulação de Delaunay disponível na ferramenta *Projection Explorer*.

O intuito dessa aplicação é identificar os tópicos de documentos que se relacionam a um documento de interesse ao longo do tempo. Para isso foi utilizada a técnica *Topic Time Coordination* (Eler et al., 2009b), introduzida na Seção 5.3.3, a qual provê suporte para análise de tópicos de um conjunto de documentos que varia no tempo. A Figura 6.9 ilustra a exploração coordenada de cinco *Time-Span Maps* (*TS Maps*) distintos. Esses *TS Maps* foram numerados de 1 a 5, correspondendo aos respectivos períodos de tempo $1990 - 1993$, $1990 - 1997$, $1990 - 2001$, $1990 - 2005$ e $1990 - 2008$. Primeiramente, considere um artigo específico intitulado *"Visualization for the Document Space"*, o qual está relacionado com o tópico *"information retrieval systems"*. Este documento foi selecionado no *TS Map* 1, como apresentado na Figura 6.9(a). Após a seleção desse documento, a *Topic Time Coordination* detecta tópicos nos outros *TS Maps* com base na vizinhança do documento selecionado no *TS Map* 1. A vizinhança do documento selecionado muda ao longo do tempo e, consequentemente, também mudam os tópicos relacionados. No *TS Map* 2 (Figura 6.9(b)) os tópicos detectados referem a *"query data"*, *"information visualization"* e *"world wide web visualizations"*, os quais sugerem artigos relacionados com visualização de buscas na web. No *TS Map* 3 (Figura 6.9(c)) os tópicos detectados sugerem *"content-based image retrieval"*, que estão relacionados com sistemas de recuperação de informação, e particularmente com sistemas para recuperação de imagens. No *TS Map* 4 (Figura 6.9(d)) os tópicos referem a *"information visualization"* e *"knowledge visualization"*. Por fim, no *TS Map* 5 (Figura 6.9(e)) os tópicos novamente envolvem os termos *"information visualization"* e *"knowledge visualization"*. É interessante notar, por meio desses últimos tópicos detectados, que sistemas de visualização de informação e de visualização de domínios de conhecimento foram relacionados com o artigo inicialmente selecionado. Isso indica que os artigos de visualização relacionados descrevem sistemas que podem ser utilizados na recuperação de informação.

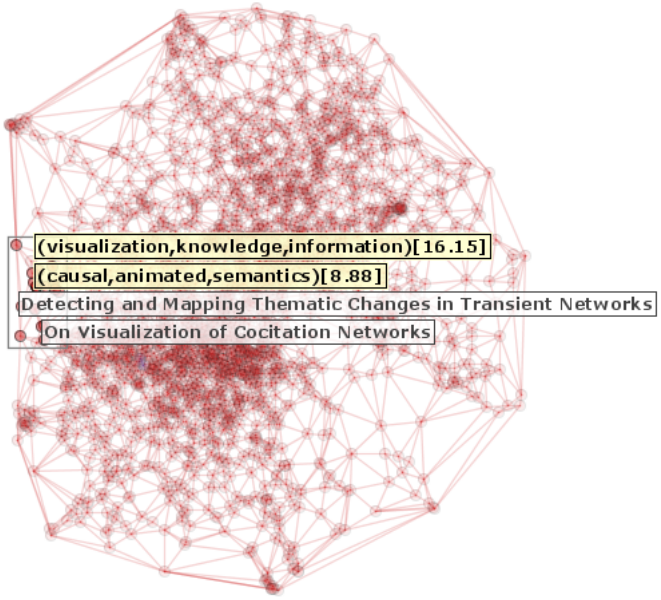
Como ilustrado na aplicação, a *Topic Time Coordination* sugere novos tópicos que aparecem ao longo do tempo e que estão relacionados aos documentos da seleção inicial do usuário. Assim,



(a) Time span 1 - 1990 a 1993. (b) Time span 2 - 1990 a 1997.



(c) Time span 3 - 1990 a 2001. (d) Time span 4 - 1990 a 2005.



(e) Time span 5 - 1990 a 2008.

Figura 6.9: *Topic Time Coordination* entre cinco *Time-Span Maps* contendo artigos publicados de 1990 a 2008 das conferências IEEE Vis, CMV, IV e IEEE InfoVis.

o usuário pode descobrir novos temas relacionados a um tema inicial, o que também pode levá-lo a focar nos grupos de interesse. Para alcançar bons resultados, esta técnica de coordenação deve operar em mapas baseado em similaridade que sejam bons, isto é, mapas capazes de preservar no espaço projetado a vizinhança definida no espaço original.

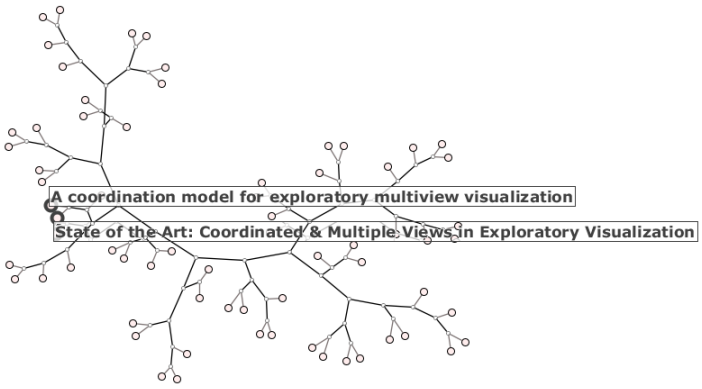
A seguir é apresentada outra aplicação que também utiliza uma técnica de coordenação para análise de diferentes coleções de documentos.

6.2.4 Coordenação baseada em Distância

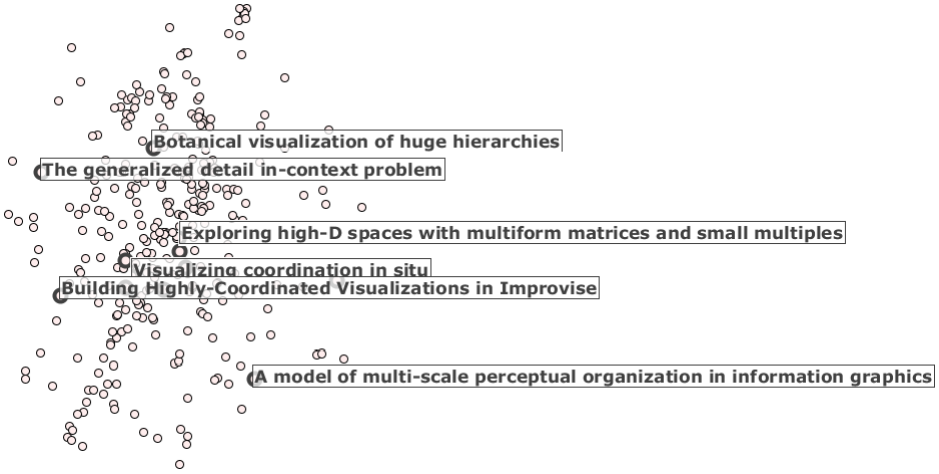
Com o mesmo propósito de auxiliar na exploração de coleções de documentos por meio da coordenação entre diferentes coleções, a aplicação ilustrada na Figura 6.10 utiliza a técnica *Distance Coordination* (Eler et al., 2008b) introduzida na Seção 5.3.5. Nessa aplicação a *Distance Coordination* foi definida para mapear em outra visão os 5 vizinhos mais próximos de cada instância selecionada na visão origem. A métrica de distância do cosseno foi utilizada para efetuar o cálculo de vizinhos mais próximos entre as instâncias dos mapas envolvidos.

O mapa de similaridade apresentado na Figura 6.10(a) foi gerado pela técnica *Neighbor-Joining* (Cuadros et al., 2007) a partir da base *CMV*. O mapa da Figura 6.10(b) foi gerado pela técnica *Least Squares Projection* (LSP) a partir da base *InfoVis*. O mapa da Figura 6.10(c) foi gerado pela técnica LSP a partir da base *IV*. No mapa apresentado na Figura 6.10(a) foram selecionados dois artigos relacionados a múltiplas visões coordenadas, intitulados *"A coordination model for exploratory multiview visualization"* e *"State of the Art: Coordinated & Multiple Views in Exploratory Visualization"*. Após a seleção foram destacados os 5 vizinhos (documentos) mais próximos nos mapas apresentados nas Figuras 6.10(b) e 6.10(c). Novamente, assim como na exploração baseada em tópicos apresentada na Seção 6.2.2, é possível notar que os documentos da base *InfoVis* não possuem muita relação com documentos que tratam de múltiplas visões coordenadas (Figura 6.10(b)). Vale lembrar que somente o resumo dos artigos é considerado. Dos documentos destacados, dois estão relacionados com o modelo e sistema *Impovise* proposto por Chris Weaver. Os títulos desses dois documentos são *"Visualizing Coordination In Situ"* e *"Building Highly-Coordinated Visualizations In Improvise"*. Por outro lado, é confirmado que o grupo de documentos da base *IV* estão em maior número relacionados a múltiplas visões coordenadas, conforme é possível observar pelo título dos documentos destacados (Figura 6.10(c)). Note também que os documentos destacados estão muito relacionados, pois foram todos posicionados em um grade grupo. Esse grupo conduz o usuário a focar também nos vizinhos dos documentos destacados, pois também estão relacionados com o tema "múltiplas visões coordenadas".

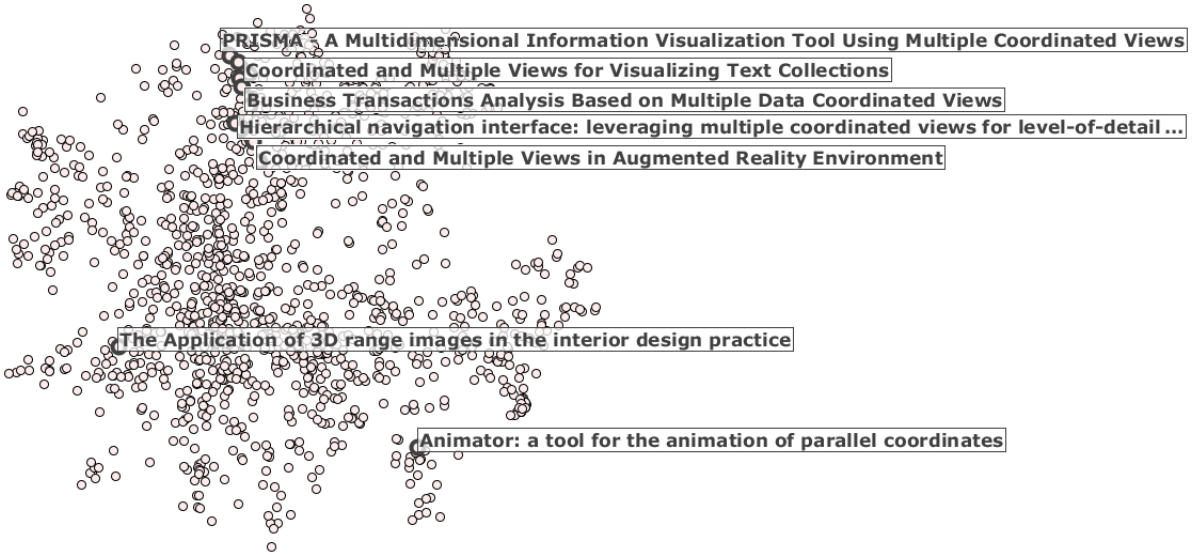
Diferente da técnica de coordenação baseada em tópicos, que mapeia na visão destino os documentos com base na ocorrência dos termos mais relevantes, a *Distance Coordination* é uma técnica que tem o objetivo de destacar os documentos que possuam relação em seu conteúdo. Essa relação é dada por meio das distâncias calculadas no espaço multidimensional, o qual é formado pelos termos extraídos dos documentos. Dessa maneira, como apresentado no exemplo anterior, é pos-



(a) Seleção de um artigo no mapa da coleção de documentos da conferência CMV.



(b) Destaque dos 5 artigos mais próximos no mapa da coleção de documentos da conferência InfoVis.



(c) Destaque dos 5 artigos mais próximos no mapa da coleção de documentos da conferência IV.

Figura 6.10: Exemplo de aplicação da técnica *Distance Coordination*. Em (a) foram selecionados dois documentos da conferência CMV; em (b) e em (C) foram destacados os 5 vizinhos (documentos) mais próximos de cada documento selecionado em (a).

sível coordenar mapas construídos de diferentes coleções de documentos e focar em documentos similares.

A seguir é apresentada uma aplicação que combina o mapeamento de diferentes técnicas de coordenação.

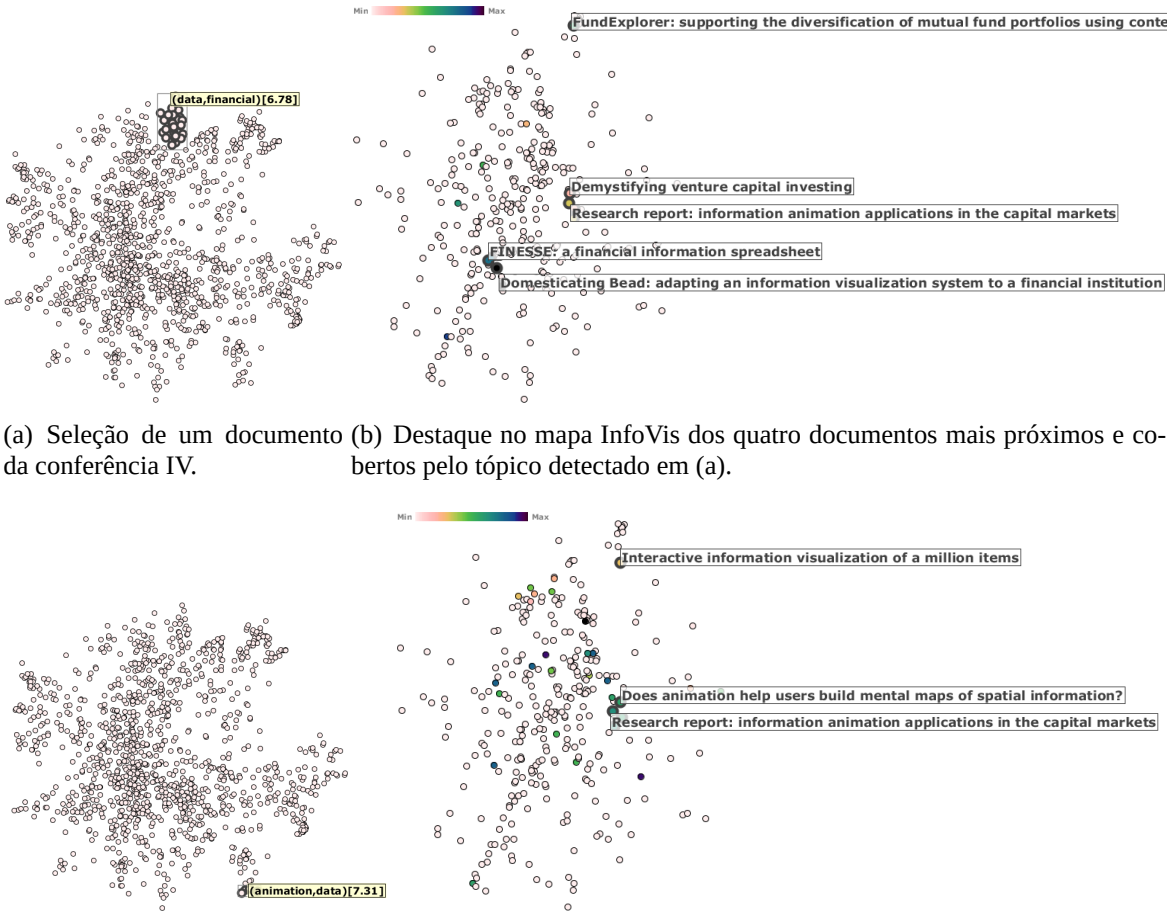
6.2.5 Combinação de Coordenações

A técnica de coordenação utilizada na próxima aplicação permite que o mapeamento de diferentes técnicas de coordenação sejam combinados. Essa técnica foi denominada *Joint Coordination*, introduzida na Seção 5.3.6, e também permite a exploração de uma ou mais coleções de documentos.

A Figura 6.11 apresenta um exemplo de aplicação de uma *Joint Coordination* configurada com as técnicas *Topic Coordination* e *Distance Coordination* para mapear os 4 vizinhos mais próximos. Através da combinação do mapeamento dessas duas técnicas é possível destacar os documentos que possuem uma relação baseada no tópico, por meio de sua cobertura, e uma relação baseada no conteúdo, mediante a distância calculada entre os documentos no espaço original. O mapa apresentado nas Figuras 6.11(a) e 6.11(c) foi construído utilizando a técnica *Least Squares Projection* (LSP) aplicada sobre a coleção de artigos da base *IV*; e o mapa apresentado nas Figuras 6.11(b) e 6.11(d) foi construído por meio da técnica LSP aplicada sobre a coleção de artigos da base *InfoVis*.

A seleção de alguns documentos é apresentada no mapa da Figura 6.11(a). A partir do conteúdo desses documentos foi gerado o tópico *"data, financial"*, o qual é utilizado pela *Topic Coordination*. Após a seleção, a *Joint Coordination* invoca a execução das duas técnicas adicionadas: *Topic Coordination* e *Distance Coordination*. Em seguida, o mapeamento das duas técnicas é combinado e as instâncias em comum são destacadas na visão destino, por meio da borda dos círculos em destaque, conforme apresentado no mapa da Figura 6.11(b). Vale lembrar que na implementação atual a combinação é a intersecção dos mapeamentos. Como a *Topic Coordination* utiliza informação de cor para destacar as instâncias mapeadas (documentos cobertos pelo tópico) é possível ver o seu mapeamento completo, representado pelos documentos destacados com cor diferente do rosa claro (nenhuma ocorrência dos termos do tópico). É possível notar que a *Joint Coordination* destacou três documentos que tiveram grande cobertura do tópico e que estão entre os vizinhos mais próximos das instâncias selecionadas. Por meio do título também é possível notar que todos eles estão relacionado com exploração de dados financeiros. Apesar da grande seleção realizada no mapa origem, poucas instâncias foram destacadas no mapa destino. Isso indica que poucas instâncias que são as vizinhas mais próximas foram cobertas pelo tópico detectado.

A seleção de outro grupo de documentos no mesmo mapa, apresentado na Figura 6.11(c), gerou o tópico *"animation, data"*. Após a seleção a *Joint Coordination* destacou as instâncias mapeadas no mapa apresentado na Figura 6.11(d). Nesse mapeamento, os vizinhos mais próximos não são aqueles que possuem a maior ocorrência do tópico. Apesar disso, os documentos mapeados estão



(a) Seleção de um documento da conferência IV.

(b) Destaque no mapa InfoVis dos quatro documentos mais próximos e cobertos pelo tópico detectado em (a).

(c) Seleção de um documento da conferência IV.

(d) Destaque no mapa InfoVis dos quatro documentos mais próximos e cobertos pelo tópico detectado em (c).

Figura 6.11: Aplicação da *Joint Coordination* configurada com as técnicas *Topic Coordination* e *Distance Coordination*. Os mapas ilustrados em (a) e (c) apresentam a seleção do usuário e o destaque é, respectivamente, apresentados nos mapas em (b) e (d). Os documentos cobertos pelo tópico e que estão entre os quatro vizinhos mais próximos dos documentos selecionados em (a) e em (c) foram, respectivamente, colocados em destaque em (b) e em (d). A cor indica a ocorrência dos termos do tópico principal no conteúdo dos documentos.

relacionados com animação, conforme observado pelo título dos documentos. Um dos documentos destacados nessa seleção também foi mapeado na seleção anterior, pois ele está relacionado com animação de dados financeiros. Este é o documento intitulado *"Research report: information animation application in the capital markets"*. O resultado dessa aplicação mostra que por meio da *Joint Coordination* é possível focar com mais precisão em grupos que estão relacionados com os documentos de interesse.

Devido à particularidade da *Topic Coordination* em destacar as instâncias mapeadas por meio de cor e não pela borda em destaque, é possível observar os documentos que foram mapeados por ela, mas não foram mapeados pela *Joint Coordination* por não pertencerem ao mapeamento da *Distance Coordination*. A Figura 6.12 mostra o mesmo mapa apresentado na Figura 6.11(d), porém exibindo os títulos dos documentos que tiveram a maior cobertura do tópico e não foram

destacados pela *Joint Coordination*. Todos eles estão relacionados com animação. Essa aplicação confirma que a *Joint Coordination* permite a criação de mapeamentos mais interessantes para auxiliar na exploração de diferentes coleções de documentos.

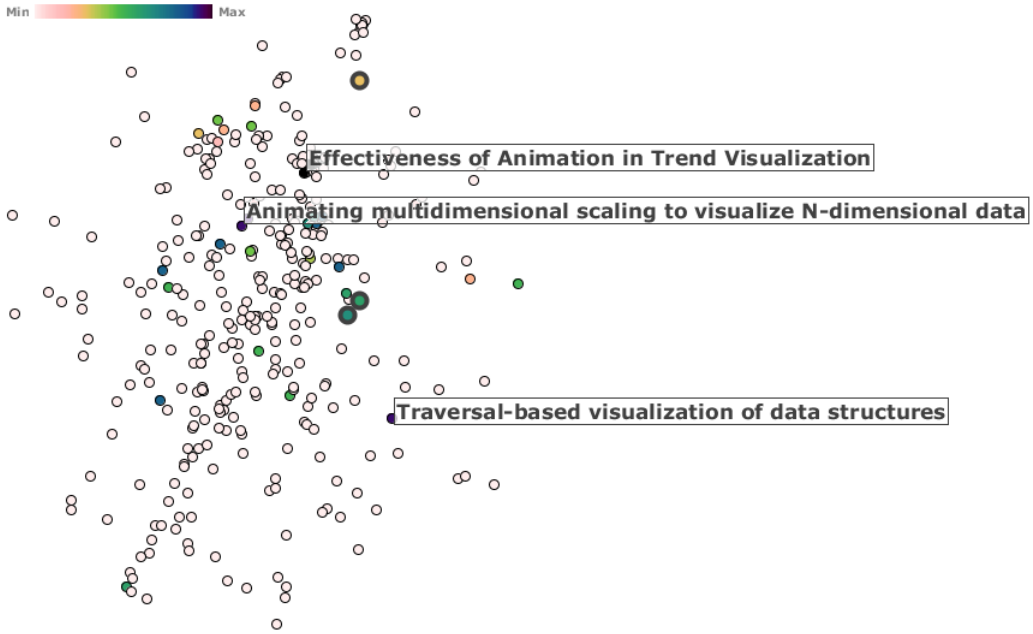
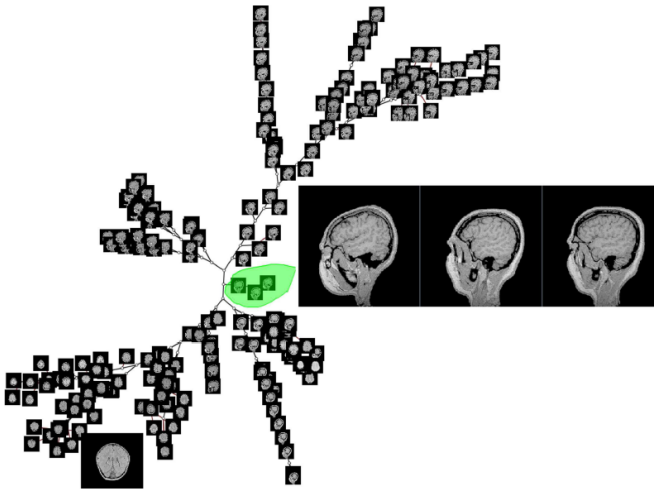


Figura 6.12: Documentos que tiveram a maior cobertura do tópicos, mas não foram destacados por não estar entre os 5 vizinhos mais próximos das instâncias.

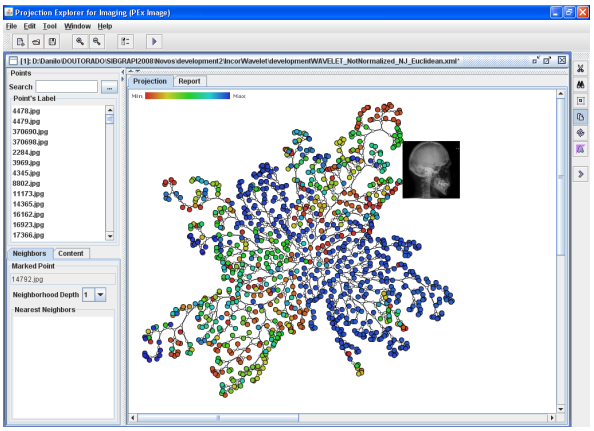
Na próxima seção será apresentada uma nova aplicação que implementa mapas de similaridade e múltiplas visões coordenadas para auxiliar a exploração de coleções de imagens.

6.3 Exploração de Coleções de Imagens

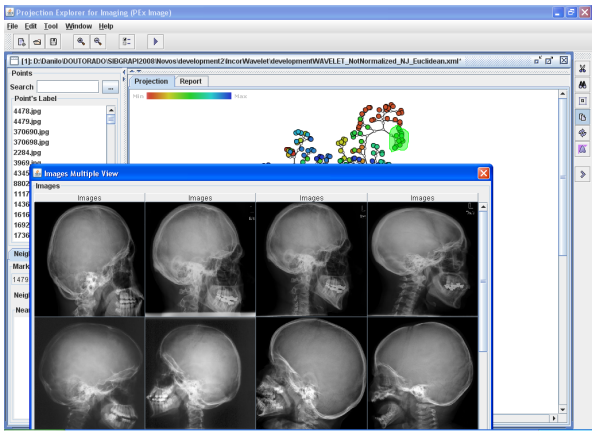
A ferramenta *Projection Explorer* (PEx) (Paulovich et al., 2007), que serviu de base para as implementações das técnicas de coordenação apresentadas, permite a construção de mapas de similaridade de qualquer conjunto de dados, para o qual exista um conjunto de atributos numéricos (espaço multidimensional) ou uma matriz de similaridade. Em aplicações que envolvem imagens, o espaço multidimensional, geralmente chamado de espaço de características, é construído por meio de algoritmos extratores de características. Nesse sentido, foi desenvolvida uma ferramenta para exploração de coleções de imagens com base em mapas de similaridade gerados a partir de espaços de características (ver Figura 6.13). Para isso, as técnicas de interação foram adaptadas e novas ferramentas específicas foram desenvolvidas. A nova ferramenta é denominada *Projection Explorer for Images* (PEx-Image) (Eler et al., 2008a, 2009a). Com a PEx-Image, é possível utilizar a própria imagem como marcação visual do mapa (Figura 6.13(a)). Adicionalmente, em qualquer uma das duas formas de marcação visual (imagens ou círculos) é possível visualizar o conteúdo das imagens sob o posicionamento do *mouse* (Figura 6.13(b)) e o conteúdo de um grupo de imagens através de uma janela dedicada (Figura 6.13(c)).



(a)



(b)



(c)

Figura 6.13: Interações do *Projection Explorer for Images* para exploração de coleções de imagens. (a) exibir o próprio conteúdo de cada amostra como marca visual; (b) exibir o conteúdo de uma amostra; (c) exibir o conteúdo de um grupo de amostras (extraída de (Eler et al., 2008a))

A PEx-Image também utiliza múltiplas visões coordenadas para auxiliar no processo de exploração. Assim, é possível comparar diferentes espaços de características e diferentes medidas de similaridade. Além disso, pode-se explorar diferentes conjuntos de imagens e também diferentes representações de um mesmo conjunto por meio de técnicas de coordenação. As aplicações que empregam as técnicas de coordenação propostas neste trabalho utilizam os seguintes conjuntos de dados:

- **MRI-12**: é uma coleção de 512 imagens médicas obtidas por ressonância magnética (MRI), as quais estão distribuídas em 12 classes distintas. A coleção contém imagens obtidas da cabeça nas posições sagital, axial e coronal; imagens da cabeça e pescoço na posição sagital; imagens da coluna na posição sagital; imagens do tórax na posição axial;

- **ImageCLEF-2006**: é uma coleção de imagens de raios X que está dividida em 116 classes desbalanceadas[1]. Esta coleção reúne radiografias que foram selecionados ao acaso dos casos de rotina do Departamento de Radiologia de Diagnóstico da Aachen University of Technology na Alemanha. As imagens representam diferentes idades, sexos, patologias e posições de visualização. Nesta tese, a coleção será dividida em quatro grupos. São eles:

  - **ImageCLEF-Train**: é a coleção de treino com 9000 imagens de raios X;
  - **ImageCLEF-Test**: é a coleção de teste com 1000 imagens de raios X;
  - **ImageCLEF-Train-Sub**: é uma amostra da coleção *ImageCLEF-Train*. Essa amostra compreende as classes que tenham entre 50 e 200 imagens, resultando em 3345 imagens de 35 classes diferentes;
  - **ImageCLEF-Test-Sub**: é uma amostra da coleção *ImageCLEF-Test*. Essa amostra é composta pelas imagens que pertencem às classes da coleção *ImageCLEF-Train-Sub*, resultando em um conjunto de 537 imagens.

- **FIGURES**: é uma coleção com 996 fotografias obtidas ao ar livre. Cada imagem também está relacionada a um arquivo texto, o qual contém uma breve descrição do que há na imagem [2].

Dessas coleções de imagens foram extraídas características para formar o espaço de características. Os extratores utilizados são:

- **Descritores de Fourier:** são extraídos 20 descritores do histograma da imagem e 6 descritores a partir da energia calculada da transformada de Fourier (da Costa e Junior, 2000; Huang e Aviyente, 2006) aplicada sobre a imagem 2D;

- **Filtros de Gabor:** são calculados 16 descritores por meio de filtros de Gabor (Manjunath e Ma, 1996; Daugman e Downing, 1998), utilizando 4 níveis de frequência (0.02 – 0.3) e 4 diferentes orientações. O algoritmo utilizado nesta tese foi implementado por Ahmad Poursaberi e está disponível em www.mathworks.com/matlabcentral/fileexchange/5237;

- **Matrizes de co-ocorrência:** são utilizadas 6 funções de Haralick (Haralick et al., 1973; Tuceryan e Jain, 1998) sobre 12 combinações de distâncias e direções, resultando em 72 características. O algoritmo utilizado nesta tese foi implementado e utilizado por Marco Aurélio Roncatti em sua dissertação de mestrado (Roncatti, 2008);

- **Média e desvio padrão:** são calculadas a média e desvio padrão das intensidades dos pixels da imagem;

[1] A base ImageCLEF está disponível em `http://ir.shef.ac.uk/imageclef/2006/`.

[2] A base FIGURES está disponível em `http://www.cs.washington.edu/research/imagedatabase/`.

- **Wavelet:** é utilizado um conjunto de características de wavelet, cedidas por Leandro Augusto da Silva e colaboradores do Instituto do Coração (INCOR) de São Paulo. Primeiramente, antes de extrair as características, todas as imagens foram redimensionadas para o tamanho de 256x256 pixels. Em seguida, foram extraídas características a partir do terceiro nível de decomposição, resultando em 1024 características (da Silva et al., 2007). Eles utilizaram a wavelet de Haar como função mãe e o pacote wavelet do Matlab da *MathWorks*;

- **Grupo Um:** é a combinação das características anteriores, com exceção das características de wavelets. Essa combinação será referida como *Grupo Um*;

- **Grupo Dois:** é a combinação das características de filtros de Gabor, descritores de Fourier, média e desvio padrão. Essa combinação será referida como *Grupo Dois*.

As aplicações apresentadas a seguir envolveram as técnicas de coordenação, de classificação e de seleção para auxiliar a exploração de coleção de imagens.

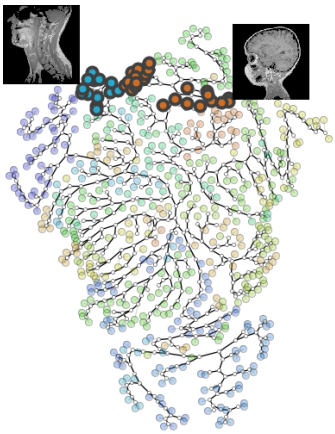
6.3.1 Coordenação baseada em Identificador

Os exemplos de aplicação apresentados aqui ilustram como a ferramenta PEx-Image pode ser usada para explorar coleções de imagens por meio de mapas de similaridade e de técnicas de coordenação. Os primeiros exemplos ilustram como a coordenação dos mapas pode contribuir na escolha de bons espaços de características e de uma medida de similaridade adequada. Para isso, foi utilizada a coordenação baseada em identificador (*Identity Coordination*). A técnica de precisão *Neighborhood Hit* (Paulovich et al., 2008), descrita na seção 6.2.1, será utilizada para identificar os melhores mapas. No caso das técnicas baseadas em árvore, o vizinho mais próximo de uma instância é aquele que apresenta o menor caminho na árvore até a instância.

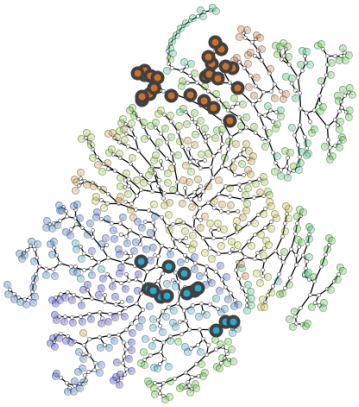
Mapa de imagens de ressonância magnética: comparação entre diferentes espaços de características

O primeiro exemplo de aplicação da PEx-Image mostra como a coordenação dos mapas de similaridade podem auxiliar a comparação de diferentes espaços de características. A Figura 6.14 ilustra quatro diferentes mapas gerados pela técnica *Neighbor Joining* (NJ) (Cuadros et al., 2007), utilizando a distância Euclideana como medida de similaridade. Nesse exemplo foi utilizada a coleção *MRI-12*. Cada imagem da coleção é representada por um círculo nos mapas de similaridade apresentados na Figura 6.14. Nesse caso, a cor indica a informação de classe (rótulo) da imagem. As representações visuais das Figuras (a), (b), (c) e (d) mostram os mapas gerados a partir de diferentes espaços de características.

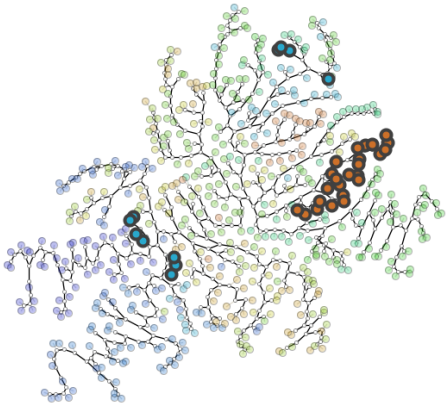
A Figura 6.14(e) apresenta o cálculo de precisão da técnica *Neighborhood Hit* aplicada em cada um dos mapas apresentados na Figura 6.14. A *Neighborhood Hit* indica que o espaço de características formado pelas matrizes de co-ocorrência, representado pela cor verde no gráfico,



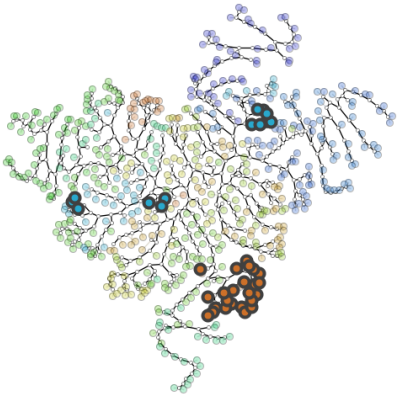
(a) Filtros de Gabor



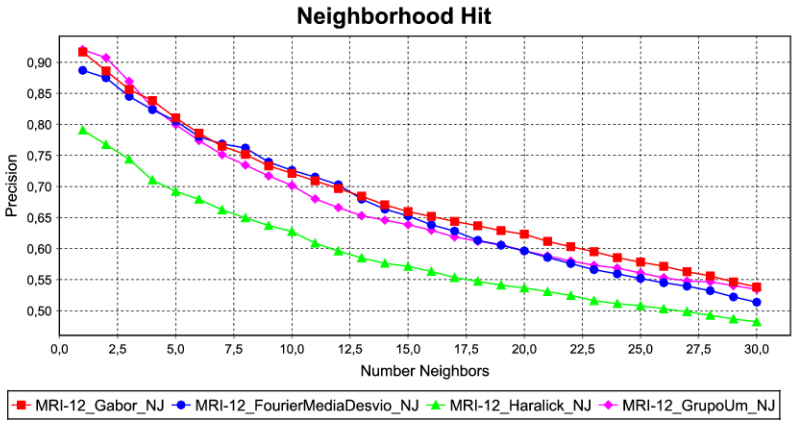
(b) Fourier, média e desvio padrão



(c) Matrizes de co-ocorrência



(d) *Grupo Um* (combinação de todas as características)

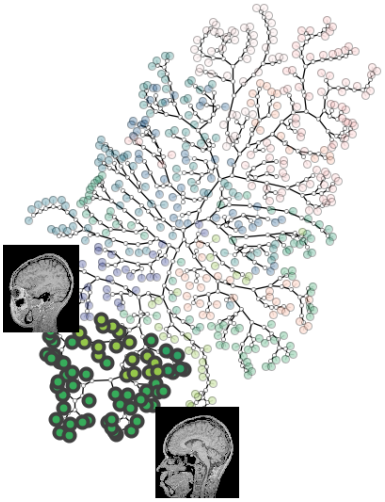


(e)

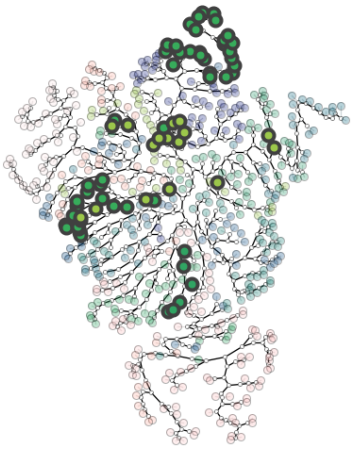
Figura 6.14: Mapas de similaridade de um conjunto de imagens *MRI-12*, utilizando quatro diferentes espaços de características e gráfico de precisão dos mapas (e).

é o pior espaço de características para diferenciar as classes dessa coleção de imagens. Já os outros três espaços possuem um comportamento similar, mas o espaço construído pelos filtros de Gabor (cor vermelha no gráfico) apresenta melhores valores de precisão a partir de 17 vizinhos. A inspeção visual realizada por meio de uma coordenação por identidade também confirma que o espaço construído pelos filtros de Gabor forma melhores grupos, conforme apresentado na Figura 6.14. Primeiramente, foram selecionadas algumas imagens de duas classes distintas no mapa apresentado na Figura 6.14(a). Uma imagem de cada classe foi exibida sobre a seleção. A seleção foi ecoada nas demais visões por meio da coordenação baseada em identificador. Esse exemplo mostra como as representações visuais podem auxiliar na investigação de qual conjunto de características possui melhor capacidade de discriminação, podendo revelar quais grupos estão contribuindo ou prejudicando a qualidade da representação visual.

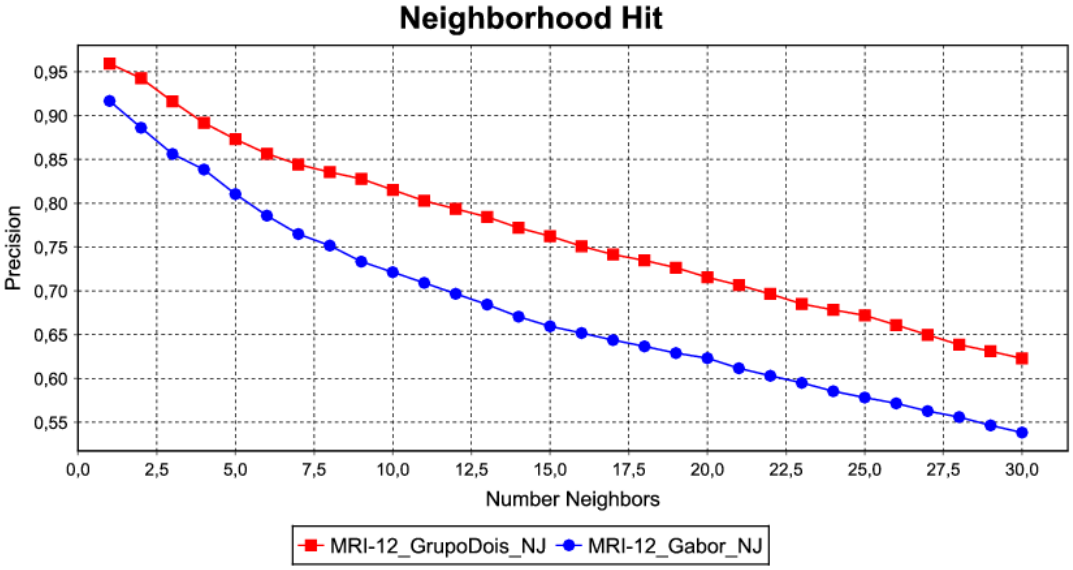
Como confirmado pelos valores de precisão da *Neighborhood Hit* apresentados no gráfico ilustrado na Figura 6.14(e), o espaço construído pelas matrizes de co-ocorrência gerou o mapa que apresentou a menor precisão para essa coleção de imagens. Isso sugere que essas características podem prejudicar a qualidade dos mapas gerados a partir do espaço *Grupo Um*. A Figura 6.15(a)



(a) *Grupo Dois*



(b) Filtros de Gabor



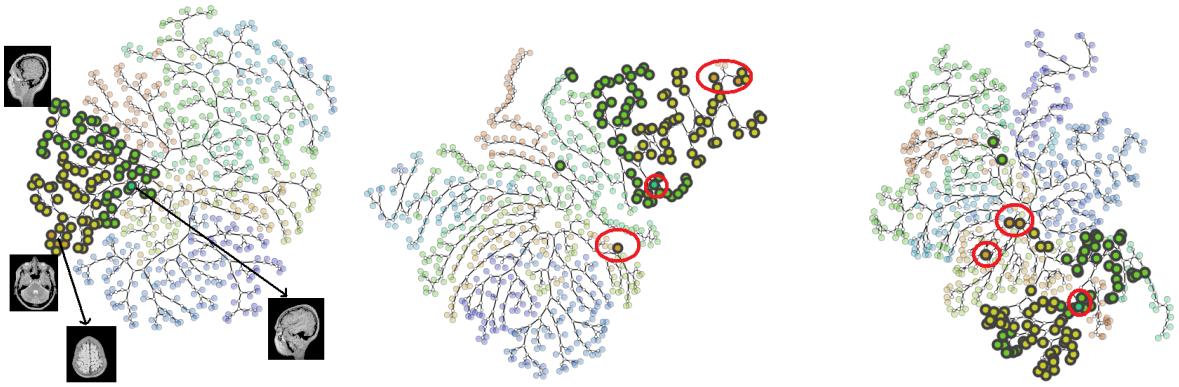
(c) *Neighborhood Hit* dos mapas (a) e (b).

Figura 6.15: Comparação da precisão calculada por *Neighborhood Hit* (c) e coordenação dos mapas construídos por meio dos espaços de características *Grupo Dois* (a) e filtros de Gabor (b).

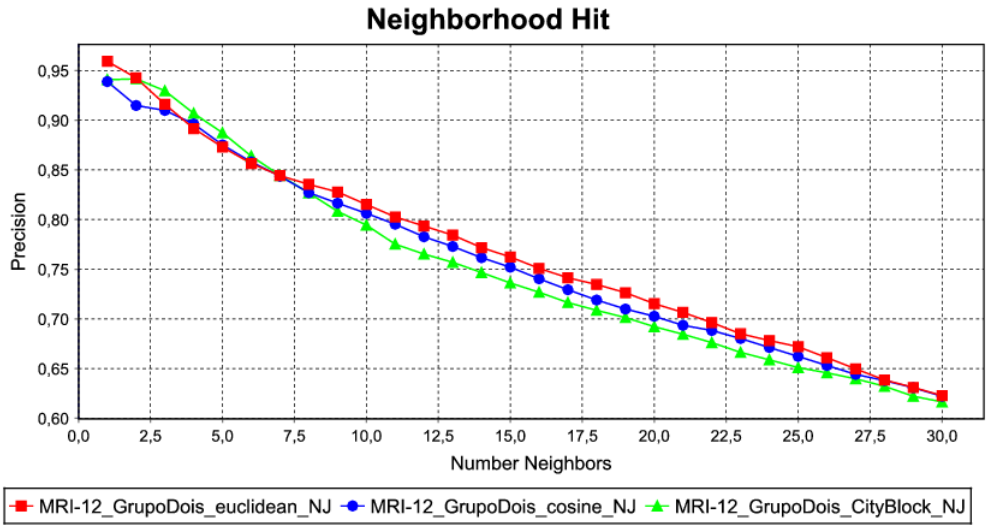
apresenta um mapa gerado pela combinação dos dois melhores espaços de características (*Grupo Dois*). Por meio da coordenação é possível comprovar que este mapa forma melhores grupos do que o mapa construído apenas com as características geradas pelos filtros de Gabor (ver Figura 6.15(b)). A Figura 6.15(c) apresenta um gráfico comparando a qualidade dos mapas gerados pelo espaço de características *Grupo Dois* (linha em vermelho) e pelo espaço de características construído pelos filtros de Gabor (linha em azul), confirmando que a combinação dos melhores espaços de características produziram um mapa com melhor precisão, ainda que variando entre 5% e 10% melhor.

Mapa de imagens de ressonância magnética: comparação entre diferentes medidas de similaridade

A coleção de imagens *MRI-12* também é utilizada no exemplo apresentado na Figura 6.16. Nesse exemplo são comparadas diferentes medidas de similaridade para a construção dos mapas. As Figuras 6.16(a), 6.16(b) e 6.16(c) apresentam mapas construídos utilizando o espaço de características *Grupo Dois* e, respectivamente, as distâncias Euclideana, cosseno e *City Block*.



(a) (b) (c)



(d)

Figura 6.16: Mapas de similaridade de um conjunto de imagens médicas construídos com o espaço de características *Grupo Dois* e diferentes métricas de distância: (a) Euclideana; (b) cosseno; e (c) *City Block*. Em (d) é apresentado o gráfico com a precisão de cada mapa.

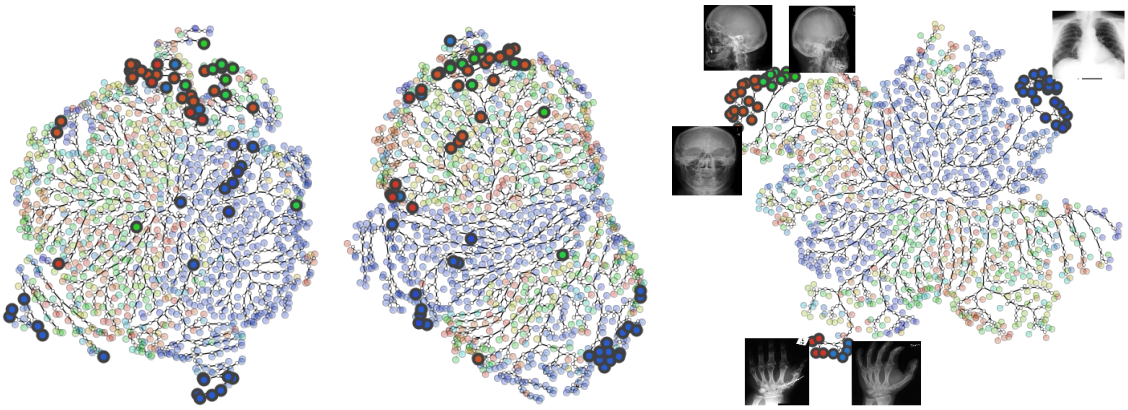
A Figura 6.16(d) apresenta o gráfico da precisão calculada pela técnica *Neighborhood Hit* a partir dos mapas apresentados na Figura 6.16. Por meio do gráfico é possível notar que de um modo geral todas as medidas de similaridade possuem um comportamento semelhante, mas que a distância Euclideana apresenta uma melhor precisão, considerando o número de vizinhos entre 8 e 22. Utilizando a coordenação é possível verificar que o mapa gerado com a distância Euclideana como medida de similaridade apresenta grupos com menos dispersão de elementos, conforme ilustrado pelo ramo selecionado na Figura 6.16(a) (mapa origem), composto por duas classes de

imagens da cabeça. Quatro instâncias de outras duas classes foram também posicionadas no ramo selecionado, sendo uma delas muito semelhante às imagens selecionadas. Essas duas classes são indicadas por uma seta, que aponta para uma imagem de sua respectiva classe. Nos outros mapas foram destacadas as instâncias que possuem o mesmo identificador daquelas selecionadas no mapa origem. É notável que o mapa que utilizou a distância do cosseno posicionou as instâncias representadas pela cor amarelo escuro em um mesmo ramo e que aquelas representadas pela cor verde são posicionadas em dois ramos (ver Figura 6.16(b)), seguindo o mesmo comportamento do mapa origem. No entanto, algumas instâncias estão espalhadas fora do grupo (ramo) correto. Observando o mapa que utilizou a distância *City Block* é possível notar o mesmo comportamento, mas as instâncias representadas pela cor verde ficam espalhadas em mais de dois grupos (ver Figura 6.16(c)). As quatro instâncias erroneamente posicionadas no ramo selecionada no mapa origem foram destacadas nos outros dois mapas por meio de círculos em vermelho. O erro no posicionamento dessas quatro instâncias foi maior no mapa que utilizou a distância do cosseno (ver Figura 6.16(b)).

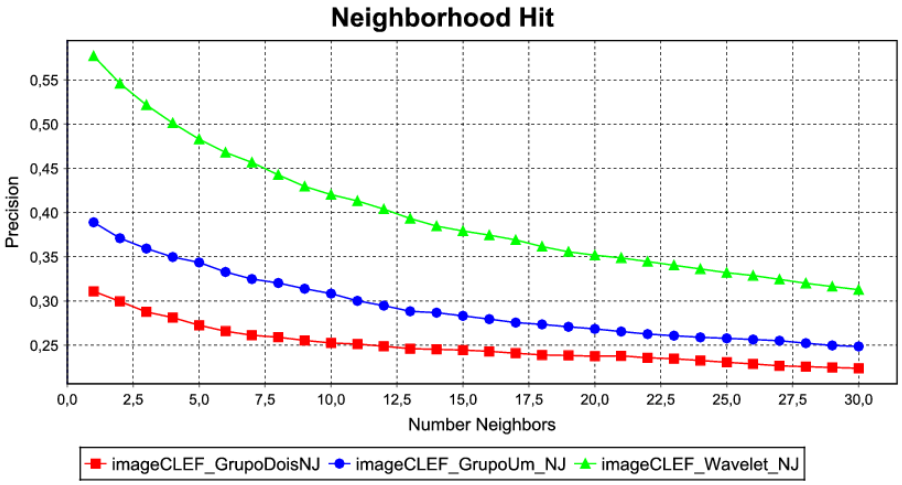
Mapa de imagens de raios X: comparação entre diferentes espaços de características

Os extratores de características utilizados na coleção de imagens de ressonância magnética *MRI-12* também foram utilizados na coleção de imagens conhecida como *ImageCLEF-2006*. A Figura 6.17 mostra três mapas construídos com a técnica NJ a partir da coleção *ImageCLEF-Test*. A Figura 6.17(a) apresenta o mapa construído com o espaço de características gerado pelos extratores *Grupo Dois*. A Figura 6.17(b) apresenta o mapa construído a partir de um espaço de características gerado pelos extratores *Grupo Um*. Por fim, a Figura 6.17(c) apresenta um mapa gerado a partir das características wavelet. Os mapas foram construídos utilizando a distância Euclideana como medida de similaridade. No mapa apresentado na Figura 6.17(c) (mapa origem) foram selecionados grupos de instâncias de diferentes classes. Por meio da coordenação foram destacadas nos outros dois mapas as instâncias que possuem o mesmo identificador daquelas selecionadas no mapa origem. É possível confirmar que o espaço de características de wavelet é a melhor opção para essa coleção de imagens, pois o mapa construído com esse espaço apresenta grupos mais coesos e discriminativos.

A inspeção visual pode ser confirmada por meio do cálculo de precisão dos mapas pela técnica *Neighborhood Hit*, conforme ilustra o gráfico apresentado na Figura 6.17(d). O cálculo da precisão mostra que o espaço construído por wavelets (linha verde) é melhor do que os outros dois espaços, *Grupo Dois* representado pela linha vermelha e *Grupo Um* representado pela linha azul. Também é interessante notar que para esse conjunto de dados o espaço *Grupo Um* é melhor do que o espaço *Grupo Dois*. Não foi possível realizar a comparação entre esses dois espaços por meio da inspeção visual, pois a qualidade dos dois é muito baixa, conforme também é observado na Figura 6.17. É importante notar que a coleção *ImageCLEF-2006* contém classes muito desbalanceadas e também um grau de dificuldade elevado para sua classificação. Isso é refletido nos valores de precisão



(a) *Grupo Dois* (b) *Grupo Um* (c) Wavelet



(d) Wavelet

Figura 6.17: Mapas construídos a partir da coleção *ImageCLEF-2006* utilizando diferentes espaços de características. Em (d) é apresentado o gráfico com a precisão de cada mapa.

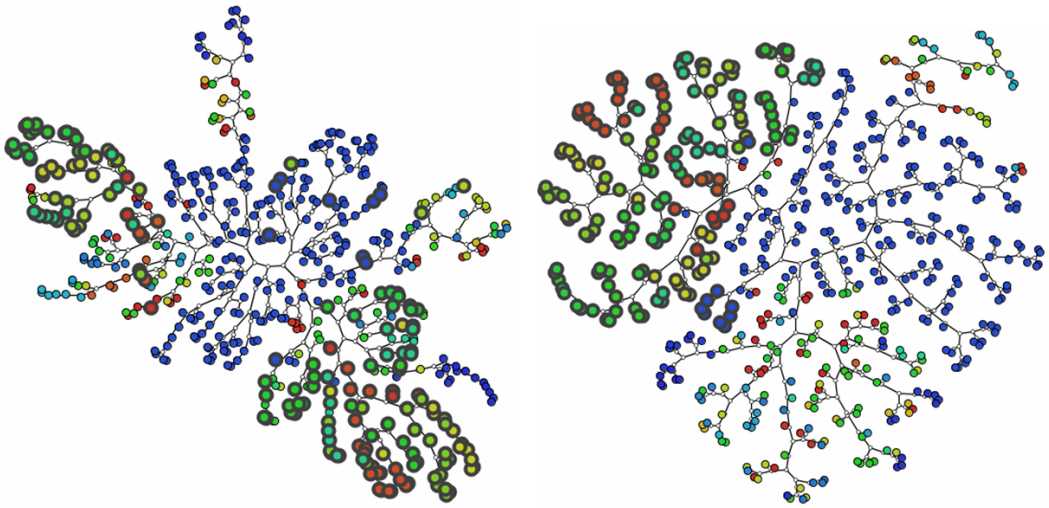
(todos abaixo de $0,6$) apresentado no gráfico ilustrado na Figura 6.17(d) e na qualidade visual dos mapas de similaridade. Apesar disso, o espaço de características construído com wavelets pôde gerar um mapa de similaridade que auxilia na exploração dessa coleção (ver Figura 6.17(c)).

Mapa de imagens de raios X: efeitos de um seletor de características para apoiar a exploração

Uma forma de melhorar o poder de discriminação do espaço de características é utilizar algoritmos seletores para escolher as características que melhor representam a coleção de imagens. O próximo exemplo, ilustrado na Figura 6.18, mostra o efeito de um seletor de características sobre a visualização. No caso, foi utilizado um seletor de características baseado na medida de saliência calculada a partir dos pesos de uma rede neural MLP (*Multi-Layer Perceptron*) (Santos e Batista, 2007). Na aplicação seguinte a MLP foi configurada com duas camadas ocultas. A saliência é uma medida proposta por Garson (Garson, 1991) que estima a importância de cada entrada da rede neural com base nos pesos dos diferentes caminhos de conexões possíveis até as saídas da rede

neural. A rede neural e o seletor de características foram implementados por Davi Santos em seu trabalho de mestrado (Santos, 2007).

Após o treino da rede neural com a coleção *ImageCLEF-Train-Sub* e o cálculo da Saliência para cada uma das características (entradas da rede neural), é possível efetuar a seleção com base no valor da Saliência. Quanto maior o seu valor, melhor é a contribuição da característica no processo de classificação. A Figura 6.18(a) mostra um mapa construído a partir da coleção *ImageCLEF-Test-Sub* com o espaço completo de características de wavelet (1024), enquanto que a Figura 6.18(b) mostra um mapa construído a partir da coleção *ImageCLEF-Test-Sub* com um espaço reduzido de características, ou seja, uma seleção de 37 características. Esse espaço reduzido inclui somente aquelas que possuem valor de Saliência maior do que 0.5. O valor de Saliência das características foi gerado do treino de um classificador baseado em rede neural com taxa de erro igual a 29.62%. Após uma exploração coordenada é possível notar que, de um modo geral, os mapas seguem um comportamento parecido para a formação de grupos, sendo diferenciados em alguns poucos grupos que são melhor distribuídos utilizando o conjunto reduzido de características, conforme ilustrado na Figura 6.18.



(a) (b)

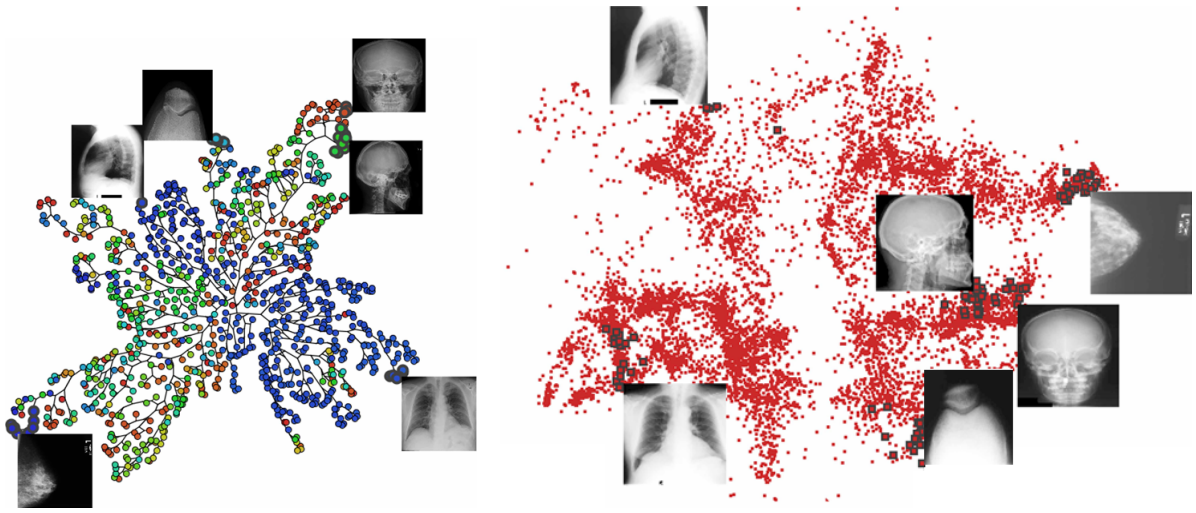
Figura 6.18: Mapas de similaridade construídos a partir de um espaço de características completo e de um espaço reduzido. A redução foi realizada por meio de uma seleção das características cujo valor de Saliência é maior do que 0.5. (a) 1024 características de wavelet (espaço completo); (b) 37 características de wavelet (espaço reduzido) (Extraída de (Eler et al., 2009a)).

A abordagem proposta mostra que a visualização de coleções de imagens por meio de mapas de similaridade é uma maneira de auxiliar a exploração e também a melhoria dos processos de classificação, pois as mudanças de parâmetros e algoritmos são refletidas nos grupos formados pelos mapas de similaridade. Com a visualização é possível localizar grupos e instâncias que possuem erros de classificação, auxiliando o especialista a modificar os parâmetros ou o algoritmo de extração de características (Brandoli et al., 2010).

A seguir é apresentado como é possível empregar a *Distance Coordination* (Eler et al., 2008b) para auxiliar a exploração diferentes coleções de imagens.

6.3.2 Coordenação baseada em Distância

Com a técnica de coordenação baseada em distância é possível utilizar uma coleção de imagens já conhecida para auxiliar a exploração de coleções desconhecidas. A Figura 6.19 mostra um exemplo de aplicação da *Distance Coordination* em duas coleções de imagens. A coleção utilizada para gerar o mapa apresentado na Figura 6.19(a) é a *ImageCLEF-Test*. A coleção utilizada para gerar o mapa apresentado na Figura 6.19(b) é a *ImageCLEF-Train*. O primeiro mapa foi gerado com a técnica *Neighbor Joining* (Cuadros et al., 2007) e o segundo com a técnica *Least Squares Projection* (Paulovich et al., 2008). Nos dois casos foram utilizadas 1024 características de wavelet. A *Distance Coordination* foi configurada para destacar os 5 vizinhos (imagens) mais próximos de cada imagem selecionada.



(a) 1000 imagens do conjunto de teste da coleção *ImageCLEF-2006*

(b) 9000 imagens do conjunto de treino da coleção *ImageCLEF-2006*

Figura 6.19: Exploração da coleção *ImageCLEF-2006* por meio da técnica *Distance Coordination* (Eler et al., 2008b). Após a seleção de algumas instâncias em (a) foram destacadas em (b), por meio da *Distance Coordination*, aquelas mais similares.

Durante a exploração foram selecionados seis grupos de imagens, conforme ilustrado no mapa apresentado na Figura 6.19(a) (visão origem). Uma imagem de cada grupo foi exibida para revelar quais classes de imagens foram selecionadas. Após a seleção, a *Distance Coordination* destacou no mapa apresentado na Figura 6.19(b) as imagens que são mais similares àquelas selecionadas na visão origem. Também foi exibida uma imagem de cada grupo de imagens destacadas. É possível observar que as imagens destacadas foram posicionadas próximas no mapa (ver Figura 6.19(b)), auxiliando o usuário a focar nos grupos de interesse. Os grupos de imagens de cabeça selecionados no mapa apresentado na Figura 6.19(a) estão posicionados em sub-ramos de um

mesmo ramo da árvore NJ. Esses dois grupos contêm instâncias de duas classes de imagens de cabeça. É interessante notar que essas instâncias também foram posicionadas próximas no mapa construído pela técnica de projeção LSP (Figura 6.19(b)). Isso também mostra que a projeção conseguiu representar bem as similaridades do espaço original (espaço de características) mesmo com um grande conjunto de imagens.

A seguir é apresentada como auxiliar a explorar de coleções de imagens com base em uma coleção textual relacionada.

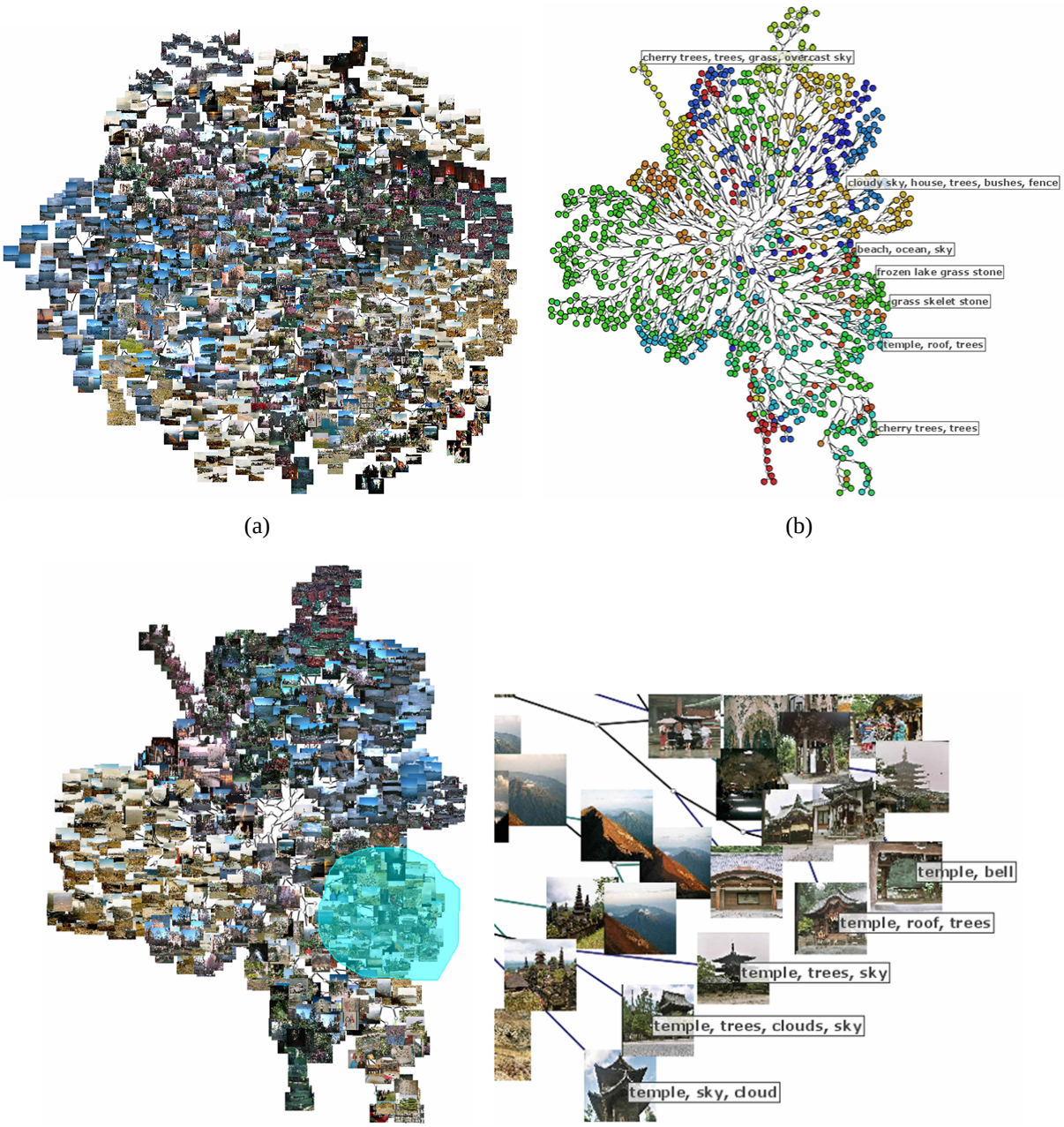
6.3.3 Coordenação entre Mapas de Imagens e Mapas de Textos

A ferramenta PEx-Image também permite que mapas de similaridade construídos a partir de coleções de imagens sejam coordenados com mapas construídos a partir de outro tipo de dado que esteja relacionado às imagens. Aqui são apresentados exemplos com duas representações do mesmo conjunto de dados, sendo uma por imagens e a outra por texto.

Exploração por meio da coordenação baseada em identificador

No exemplo apresentado na Figura 6.20 mapas de imagens são coordenados com mapas de texto construídos a partir da coleção *FIGURES*. Para isso é possível utilizar uma simples coordenação baseada em identificador (*Identity Coordination*). A Figura 6.20(a) apresenta um mapa construído pela técnica *Neighbor Joining* (NJ) a partir da coleção de imagens. Elas foram utilizadas como marcação visual ao invés dos pontos. Como características foram utilizados os descritores de Fourier, a média e o desvio padrão das intensidades da imagem. O resultado obtido pelo mapa mostra que imagens de um mesmo tipo são propriamente posicionadas em um mesmo ramo ou em ramos próximos, por exemplo, as imagens do céu azul que estão posicionadas à esquerda no mapa. Um outro mapa de similaridade foi construído utilizando a técnica NJ, mas a partir das descrições (dado textual) das imagens, conforme ilustrado na Figura 6.20(b). Nesse mapa, a cor indica a classe da imagem. As descrições de algumas instâncias foram exibidas no topo dos pontos, mostrando o conteúdo da descrição dessas imagens. Como a quantidade de termos das descrições é pequena foi utilizada uma métrica de distância conhecida como *Normalized Compressed Distance* (NCD_s) (Telles et al., 2007) para construir a matriz de distância das descrições das imagens. Esta métrica de distância foi aplicada sobre o conteúdo textual, isto é, as descrições das imagens. O mesmo mapa apresentado na Figura 6.20(b) é apresentado na Figura 6.20(c), mas dessa vez as fotografias correspondentes às descrições são utilizadas como marcação visual. Um detalhamento da área selecionada é apresentado na Figura 6.20(d) para mostrar o posicionamento das imagens no mapa. Como o mapa construído com as descrições apresenta um melhor resultado visual, essa é uma estratégia alternativa de melhorar a organização e a exploração da coleção de imagens.

Continuando a exploração da coleção *FIGURES*, a Figura 6.21 ilustra a coordenação entre os dois mapas apresentados nas Figuras 6.20(b) e 6.20(a). A Figura 6.21(a) mostra o mapa construído a partir das descrições das imagens e o mapa apresentado na Figura 6.21(b) corresponde ao mapa construído a partir da coleção de imagens. Por meio da coordenação baseada em identificador



(a) (b)

(c) (d)

Figura 6.20: Coordenação de mapas construídos a partir das imagens da coleção *FIGURES* e das descrições das imagens. (a) mapa construído a partir das imagens. (b) mapa construído a partir das descrições das imagens. A cor representa a informação de classe. (c) mapa construído das descrições das imagens, mas utilizando as imagens como marca visual. (d) detalhamento do galho selecionado em (c), exibindo algumas descrições no topo das imagens.

é possível confirmar que o mapa construído a partir das descrições apresenta melhores grupos. Primeiramente o usuário selecionou alguns pontos posicionados em um ramo do mapa construído a partir das descrições (ver Figura 6.21(a)), os quais são descrições de imagens de um jogo em um estádio de futebol americano. As instâncias correspondentes àquelas selecionadas são destacadas no mapa construído a partir das imagens (ver Figura 6.21(b)). A Figura 6.21(c) apresenta um

detalhamento da região das imagens destacadas no mapa da Figura 6.21(b), que correspondem a imagens de um jogo de futebol americano.

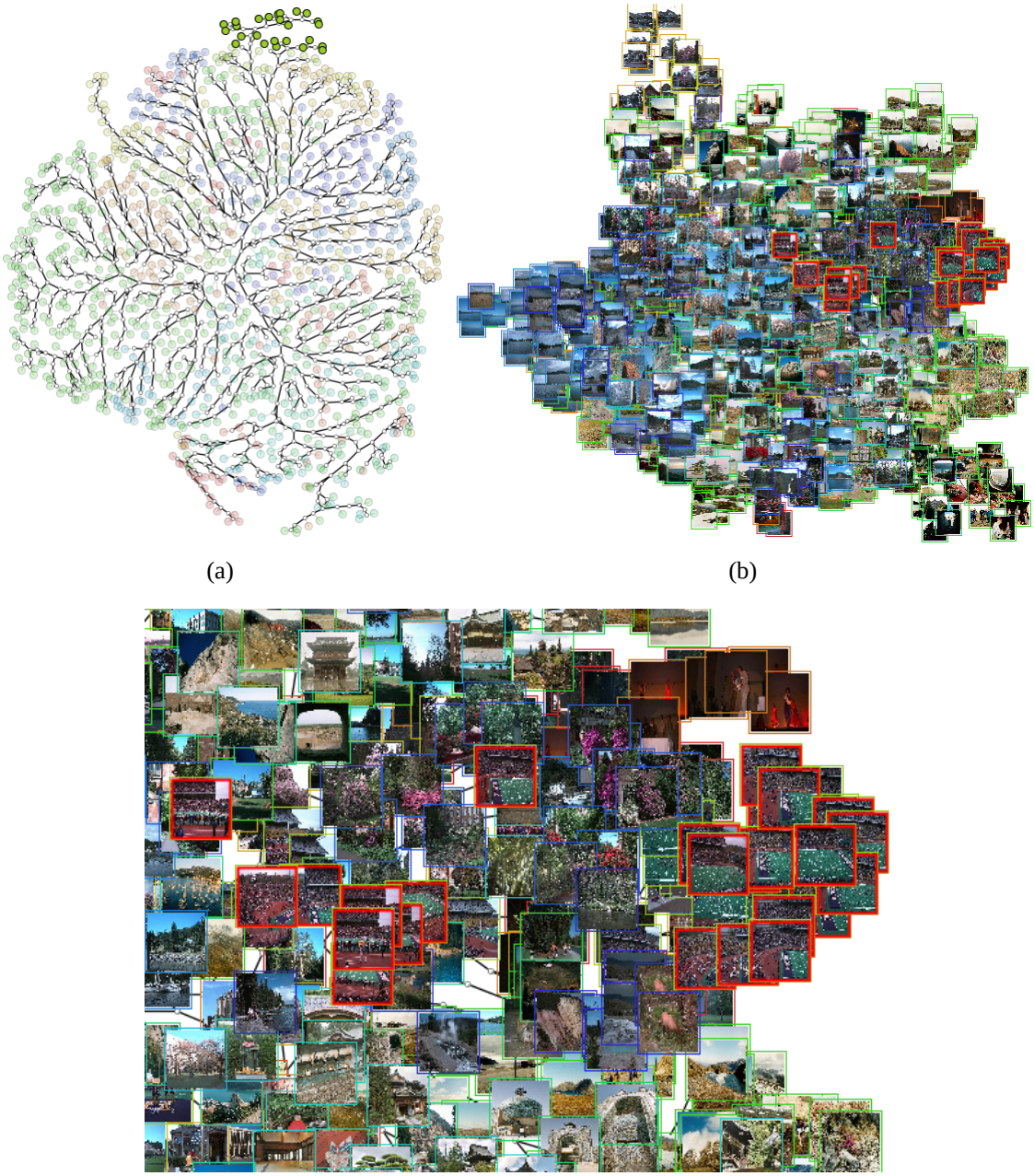
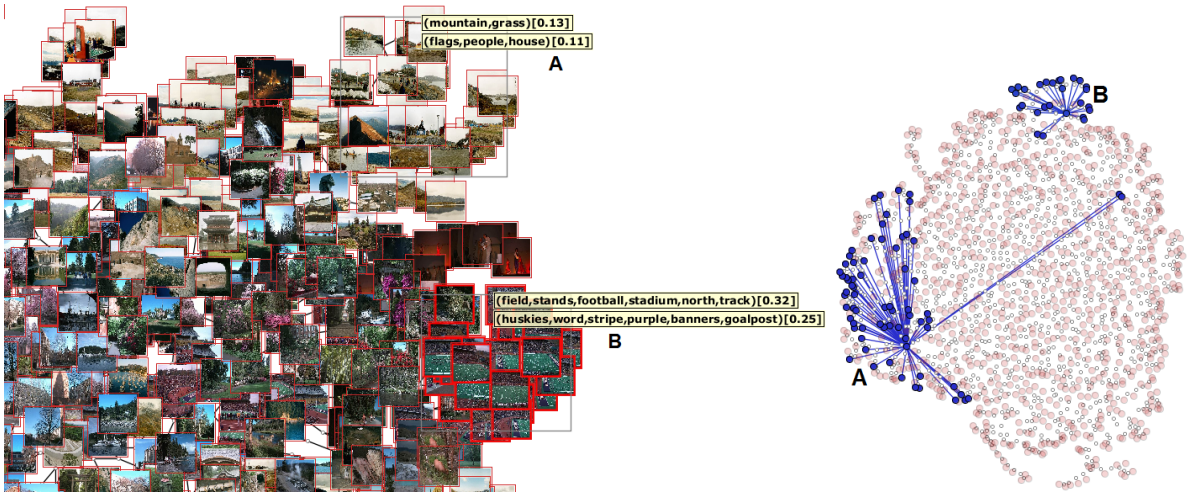


Figura 6.21: Ilustração da coordenação entre um mapa construído a partir de dados textuais (a) e outro construído a partir de dados de imagens (b). Em (c) é apresentado um detalhamento das figuras destacadas em (b).

Exploração por meio da coordenação baseada em tópicos e destaque por arestas

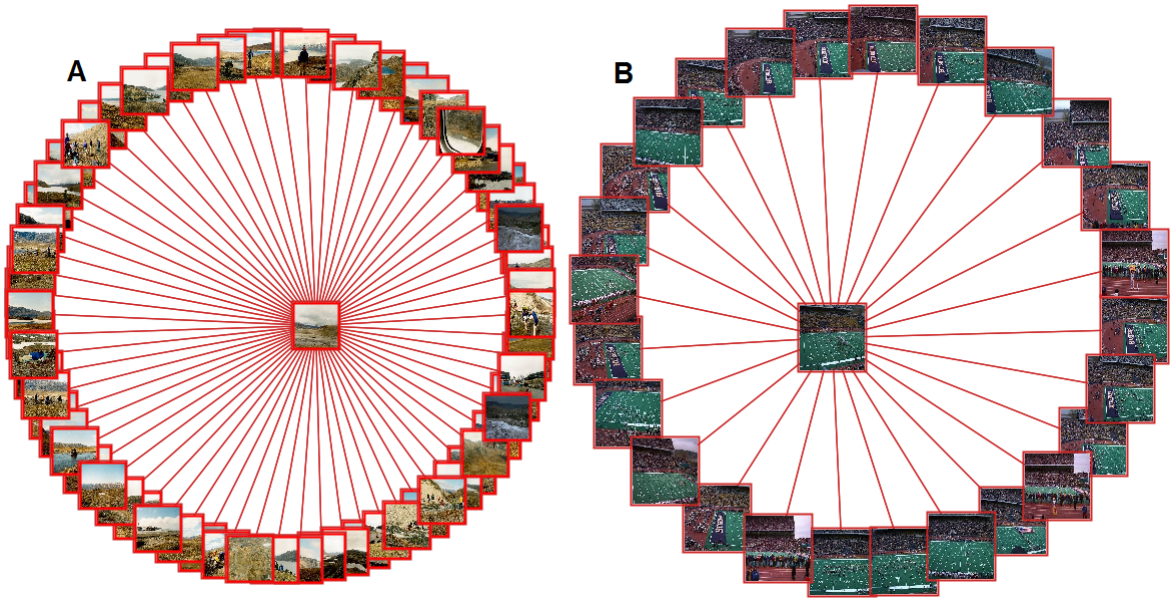
O mesmo conjunto de dados é utilizado no exemplo apresentado na Figura 6.22, para o qual foi adotado um modo alternativo de destacar as instâncias mapeadas pela técnica de coordenação.

O destaque das instâncias mapeadas é realizado por meio de arestas, unindo a primeira instância mapeada com as demais. A coordenação utilizada para criar o mapeamento é a *Topic Coordination* (Eler et al., 2008b) (ver Seção 5.3.2), a qual detecta tópicos a partir do texto relacionado às imagens e destaca em outra visão os documentos ou imagens cobertos pelo tópico detectado. Para isso é necessário carregar a informação textual de cada imagem.



(a) Seleção no mapa construído a partir da coleção de imagens.

(b) Destaque por meio de arestas no mapa construído a partir das descrições.



(c) Destaque por meio de arestas no mapa construído a partir das descrições.

Figura 6.22: Coordenação *Topic Coordination* entre mapas de similaridades construídos a partir de uma coleção de imagens e suas descrições. Em (a) foram selecionados dois grupos de imagens, sendo detectado o tópico correspondente. Em (b) foram destacadas por meio de arestas as anotações cobertas pelos tópicos detectados em (a). Em (c) as imagens destacadas são adicionadas em uma nova visão e posicionadas por meio de um algoritmo baseado em força.

Primeiramente, dois grupos de imagens foram selecionados no mapa gerado a partir da coleção de imagens (ver Figura 6.22(a)). Em seguida, os tópicos foram detectados a partir do conteúdo tex-

tual relacionado às imagens e as instâncias cobertas pelos tópicos são mapeadas e destacadas pela *Topic Coordination* no mapa construído a partir das informações textuais (ver Figura 6.22(b)). A seleção efetuada pelo usuário e o destaque realizado pela coordenação foram rotulados para facilitar a identificação. A seleção indicada pela letra **B** apresenta descrições relacionadas às imagens do estádio de futebol, gerando o tópico principal *"field, stands, footbal, stadium, north, track"*. As descrições correspondentes a esse tópico estão bem concentradas no mapa da Figura 6.22(b), indicando que poucas descrições e poucas imagens estão relacionadas a estádios de futebol. Por outro lado, a seleção indicada pela letra **A** gerou o tópico principal *"mountain, grass"* . O destaque no mapa apresentado na Figura 6.22(b) indica que as descrições relacionadas estão espalhadas em diferentes classes, pois estão em diferentes grupos (ramos). Portanto, imagens relacionadas a montanha e grama são mais comuns nessa coleção de imagens.

Uma outra maneira de analisar as imagens destacadas é adicioná-las em uma nova visão, conforme ilustrado na Figura 6.22(c). Para reposicionar as imagens foi aplicado um algoritmo baseado em força que está implementado na PEx-Image. Por meio dessa nova maneira de destaque, foi possível isolar as imagens de interesse para serem individualmente analisadas. Alternativamente, pode-se utilizar uma coordenação baseada em identificador para descobrir onde essas imagens estão posicionadas no mapa original, revelando o seu contexto no mapa.

Nossa perspectiva é que esses tipos de aplicações poderiam ser empregados na exploração de coleções de imagens médicas anotadas. Por exemplo, o usuário poderia analisar diferentes coleções de imagens com base nos tópicos extraídos das descrições ou anotações das imagens. Assim o usuário poderia selecionar um grupo de imagens de interesse e destacar aquelas cobertas pelo mesmo tópico. Essa estratégia foi também empregada para explorar sequências de proteínas (ver Seção 5.3.1). A exploração conjunta de duas ou mais representações de um mesmo conjunto de dados pode auxiliar sua compreensão.

A seguir são apresentadas aplicações para exploração de dados volumétricos apoiadas pela coordenação do mapa de similaridade e o espaço do objeto.

6.4 Exploração de Dados Volumétricos

Esta seção apresenta aplicações da técnica de coordenação baseada em identificador (*Identity Coordination*) para coordenar mapas de similaridade com a visualização do espaço do objeto de dados volumétricos. Os mapas de similaridade empregados foram construídos a partir de características extraídas dos dados volumétricos ou de um espaço de características já fornecido, que é o caso de dados de simulações. Como o número de instâncias é muito elevado, foi necessário utilizar uma extensão da *Least Squares Projection* (LSP) (Paulovich et al., 2008), denominada *Piecewise Least Squares Projection* (P-LSP) (Paulovich et al., 2010). A idéia geral da P-LSP é particionar o conjunto de dados em grupos representativos e aplicar a LSP em cada um dos grupos independentes. Por meio dessa estratégia é possível realizar a projeção de grandes conjuntos de dados,

isto é, conjuntos em torno de um milhão de instâncias. O relatório técnico da P-LSP (Paulovich et al., 2010) contém mais detalhes sobre a técnica, trabalhos relacionados com a abordagem apresentada aqui e também outras aplicações, as quais são as pesquisas iniciais do grupo de Visualização, Imagens e Computação Gráfica do ICMC/USP nessa frente de pesquisa.

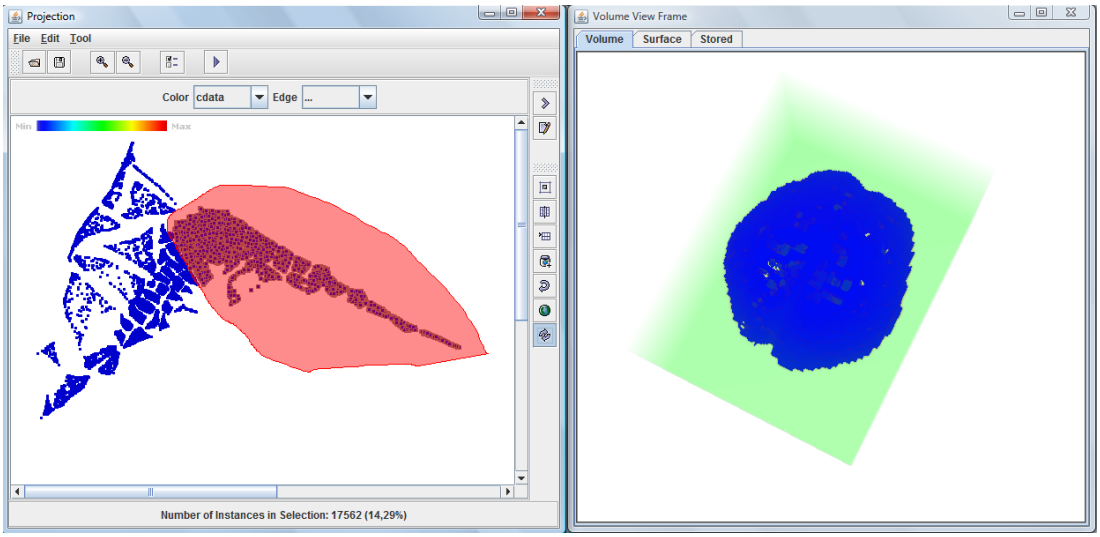
6.4.1 Coordenação entre o mapa de similaridade e o espaço do objeto

As aplicações seguintes se apóiam na coordenação entre os mapas de similaridade construídos a partir dos dados volumétricos e o espaço do objeto para auxiliar na análise dos dados volumétricos. Por meio da coordenação, quando são selecionadas instâncias no espaço projetado, no mapa de similaridade, estas mesmas instâncias são destacadas no espaço do objeto por meio de alguma técnica de visualização volumétrica, conforme ilustrado na Figura 6.23. Neste exemplo é apresentado um mapa construído a partir de um volume de cabeça que contém 122.880 voxels. Na Figura 6.23(a) é apresentada a seleção de um grande grupo de instâncias (direita) e no espaço do objeto (esquerda) são destacadas estas mesmas instâncias (voxels), as quais são visualizadas por meio de uma técnica de *rendering* volumétrico direto disponível no *Visualization Toolkit* (VTK) (Schroeder et al., 1996). A mesma seleção é apresentada na Figura 6.23(b), mas o espaço do objeto é visualizado por uma técnica de ajuste de superfície disponível no VTK. Este exemplo utiliza a amostragem de um volume cedido por colaboradores e não está disponível para *download*.

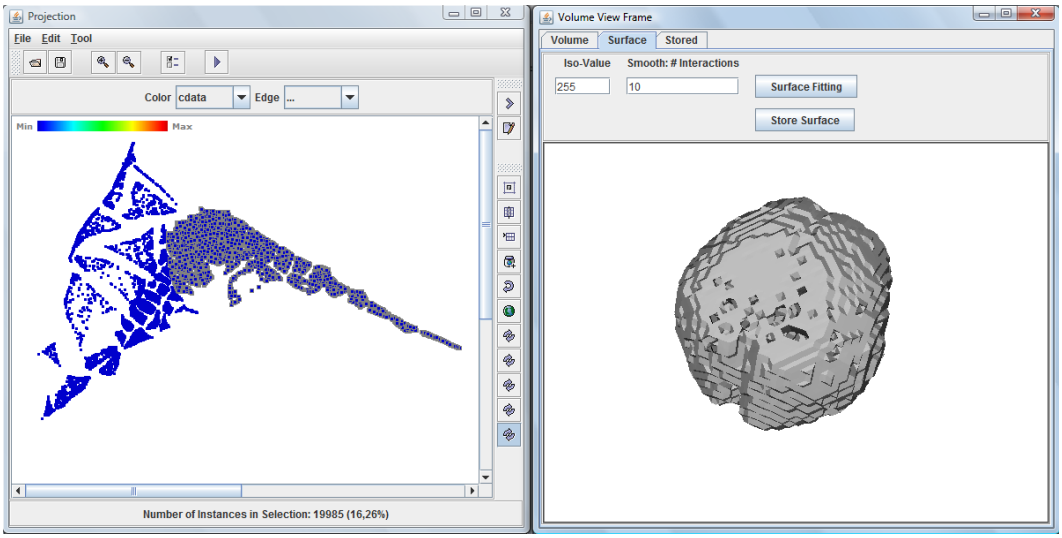
Para realizar o destaque no espaço do objeto, por meio da coordenação, as instâncias selecionadas recebem um valor escalar e as demais instâncias recebem outro valor. No caso deste trabalho, as instâncias selecionadas receberam o valor 255 e as instâncias não selecionadas receberam o valor 0. Assim, é possível configurar uma função de transferência que destaca somente as instâncias selecionadas, fazendo com que as estruturas de interesse sejam visualizadas, conforme ilustrado na Figura 6.23(a). Também é possível definir um iso-valor para ajustar uma superfície às estruturas destacadas, conforme ilustrado na Figura 6.23(b).

Exploração de um conjunto de dados volumétricos de uma cabeça

Contrário aos conjuntos de dados utilizados nas aplicações apresentadas no Capítulo 6, dados volumétricos tendem a ter um grande número de instâncias. A projeção de grandes conjuntos de dados é um problema para a visualização, pois a sobreposição de pontos é inevitável, uma vez que há mais pontos do que espaço disponível na tela. Esse problema impede a percepção de densidade em diferentes áreas do mapa de similaridade e da fronteira entre elas, o que dificulta a identificação de grupos, conforme pode ser percebido nos mapas apresentados nas Figuras 6.24(a), 6.24(b) e 6.24(c). Para reduzir esse problema, foi utilizada a transparência dos pontos, conforme ilustrado no mapa apresentado na Figura 6.24(d). Com a exploração do mapa de similaridade é possível verificar que as estruturas da parte interna e externa do cérebro são muito densas, indicadas com as cores bege e verde; ao contrário dos pontos azuis que variam por todo o mapa. Isso indica que



(a)



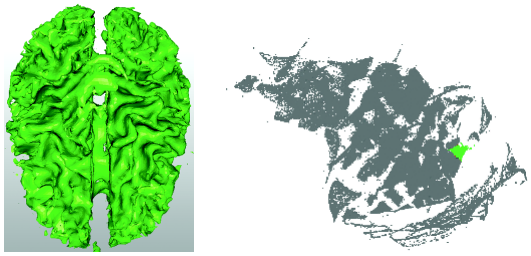
(b)

Figura 6.23: Coordenação entre um mapa de similaridade e o espaço do objeto (volume). A visualização do espaço do objeto foi realizada por meio de *rendering* volumétrico direto (a) e de ajuste de superfícies (b).

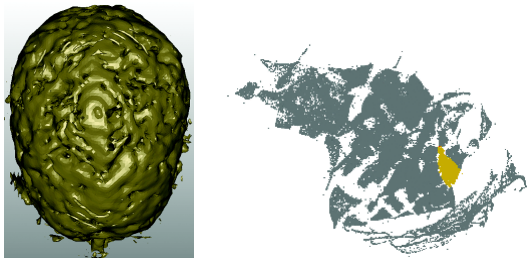
as características utilizadas e a medida de similaridade definem que os voxels que representam as partes mais internas do cérebro são mais similares do que os voxels que representam a pele, crânio e meninges (pontos em azul). Por meio dessa abordagem, utilizando a transparência, é possível localizar grupos e sub-grupos de voxels que são altamente relacionados, possibilitando o foco da exploração nessas áreas. Nesse exemplo, as projeções foram criadas a partir de um subconjunto de um volume de cabeça [3] obtido por meio de tomografica computadorizada. Foram considerados somente as fatias 30 a 97, pois o restante das fatias não continham informação relevante, somente parte de fundo. A amostra resultante contém 1.114.112 voxels. O espaço de características foi

[3] O volume original está disponível em `http://www9.informatik.uni-erlangen.de/External/vollib/`

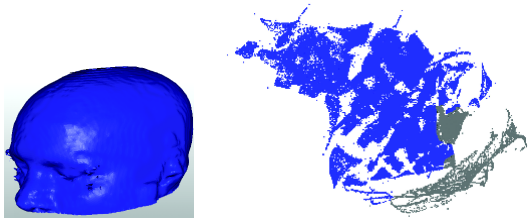
construído a partir da extração de características de intensidade, isto é, diferentes combinações da média, variância e desvio padrão das intensidades dos voxels; e o gradiente. Ao todo foram utilizadas 14 características.



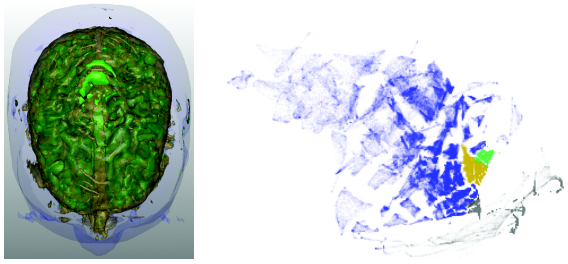
(a) Seleção de uma estrutura interna do cérebro.



(b) Seleção de uma estrutura externa do cérebro.



(c) Seleção que representa a pele, crânio e meninges.



(d) Combinação de todas as seleções anteriores.

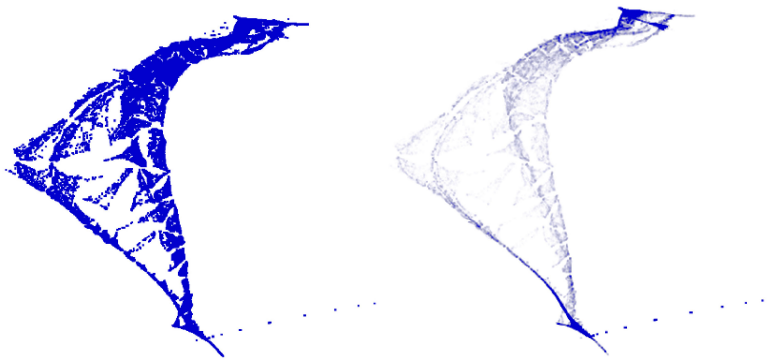
Figura 6.24: Projeções P-LSP de uma amostra de um volume de cabeça. As projeções mostram boa separação entre diferentes partes da cabeça (Paulovich et al., 2010).

As Figuras 6.24(a) e 6.24(b) mostram, respectivamente, a seleção do usuário representando a parte interna e a parte externa do cérebro. A seleção de pontos que representa a parte da pele, o crânio e as meninges pode ser vista na Figura 6.24(c). Por último, a Figura 6.24(d) mostra todas as partes juntamente selecionadas. Ainda que o conjunto de dados seja grande, a projeção P-LSP foi capaz de separar bem as diferentes estruturas presentes no volume. Além disso, as projeções indicam que o conjunto de características utilizado para representar os voxels consideram muito

similar a parte externa e a parte interna do cérebro, pois elas foram posicionadas próximas na projeção.

Exploração de um conjunto de dados de uma simulação

Um outro exemplo de exploração de dados volumétricos é apresentado na Figura 6.25, no qual foi utilizado um mapa de similaridade construído por meio da técnica P-LSP a partir de um conjunto de dados de uma simulação disponibilizado para a competição *IEEE Visualization 2008 Contest* (Whalen e Norman, 2008). Esse conjunto representa a simulação da propagação de uma frente de ionização durante a formação de uma galáxia. A simulação contém 200 fatias de tempo (*timesteps*) e cada uma delas é composta por um volume com $36.000.000$ instâncias, as quais são representadas por 10 diferentes atributos, ou seja, a densidade, temperatura e a massa de outros elementos químicos [4]. Como a grande quantidade de instâncias dificulta a interatividade sobre o mapa de similaridade, os dados foram regularmente amostrados, reduzindo o número de instâncias para $1.537.600$. Além disso, foi utilizada somente a fatia de tempo 99 da simulação. A Figura 6.25(a) mostra o resultado de uma projeção desse conjunto, na qual cada ponto do mapa representa uma instância da simulação. Novamente, para evitar o problema de sobreposição, foi adotada a transparência como estratégia para identificar grupos. A Figura 6.25(b) apresenta o mesmo mapa ilustrado na Figura 6.25(a), mas desenhado utilizando transparência, o que permite a identificação de duas regiões muito densas no mapa. Essas regiões são formadas por pontos muito similares entre si, mas muito dissimilares entre as duas regiões.



(a) Projeção original. (b) Projeção utilizando transparência.

Figura 6.25: Projeção do conjunto de dados do *IEEE Visualization 2008 Contest* (Whalen e Norman, 2008). A transparência pode auxiliar os usuários a localizar estruturas de interesse por meio dos grupos com pontos mais similares, os quais são os grupos com alta densidade.

Como mencionado, o conjunto de dados utilizado para gerar os mapas da Figura 6.25 é formado por 10 diferentes atributos. Para analisar o efeito desses atributos na similaridade dos pontos, foi escolhido o elemento hidrogênio (H) para colorir o mapa de acordo com sua concentração, conforme ilustrado na Figura 6.26. É interessante notar que o elemento químico hidrogênio é

[4] Os elementos químicos no conjunto de dados da simulação são: H, H+, He, He+, He++, H-, H_2, H_2+

um dos fatores que define a grande densidade da duas regiões de pontos similares, sendo um grupo com alta concentração e outro com baixa concentração de hidrogênio. As Figuras 6.26(a) e 6.26(b) apresentam, respectivamente, a seleção da região de pontos com alta concentração de hidrogênio e a superfície resultante dessa seleção. Por outro lado, a região oposta representa baixas concentrações de hidrogênio. As Figuras 6.26(c) e 6.26(d) apresentam, respectivamente, a seleção dos pontos com baixa concentração de hidrogênio e a superfície resultante da seleção. A Figura 6.26(e) apresenta a visualização das duas superfícies resultantes, utilizando transparência na superfície em amarelo. Por meio dessa abordagem, podemos dizer que instâncias com baixa concentração de hidrogênio compõem uma estrutura cercada por uma outra estrutura formada por instâncias com alta concentração de hidrogênio, no caso dessa simulação.

A seguir são apresentadas as considerações finais sobre as aplicações e as ferramentas apresentadas neste capítulo.

6.5 Considerações Finais

Este capítulo apresentou exemplos detalhados de aplicações envolvendo as técnicas de coordenação desenvolvidas neste trabalho para auxiliar a exploração de coleções de documentos e de coleções de imagens. O foco inicial foi as coleções de documentos, as quais deram origem às técnicas de coordenação desenvolvidas neste trabalho. As aplicações mostraram que elas auxiliam no processo de visualização exploratória em várias circunstâncias. Por meio da coordenação, o usuário pode focar em áreas de um mapa a partir de outro mapa, gerado por um outro conjunto de dados. Dessa maneira, é possível identificar novas relações entre diferentes conjuntos de dados. Além disso, a partir da área destacada é possível focar em sua vizinhança, pois ela pode apresentar alguma similaridade com as instâncias destacadas.

A PEx foi adaptada para que coleções de imagens pudessem se exploradas com mais facilidade, ou seja, novas técnica de interação foram adicionadas à ferramenta. Essa adaptação resultou em uma abordagem e ferramenta denominada *Projection Explorer for Images* (PEx-Image). A PEx-Image oferece uma nova forma de explorar coleções de imagens por meio de mapas de similaridade e múltiplas visões coordenadas. Por meio dela, pode-se observar que mapas de similaridade são uma abordagem interessante para representar o relacionamento entre imagens, principalmente os mapas construídos pela técnica *Neighbor-Joining*(NJ) (Cuadros et al., 2007). As técnicas de projeção, como a *Least Squares Projection* (LSP) (Paulovich et al., 2008), conseguem formar grupos compactos, o que causa sobreposição dos pontos. Por outro lado, as árvores NJ possuem uma forma diferenciada de representar os dados, evitando a sobreposição e mostrando detalhes da relação de similaridade dentro dos grupos. Por meio da coordenação entre essas técnicas é possível combinar suas vantagens no processo exploratório.

Além das aplicações apresentadas neste capítulo, as quais estão relacionadas com coordenação, a PEx-Image também permite que o usuário melhore o processo de exploração por meio de

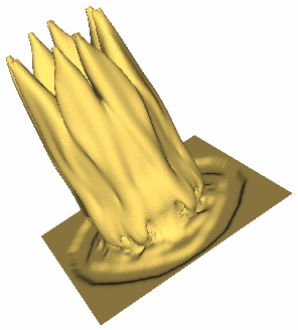
dados obtidos de classificadores. Em aplicações reais os rótulos não estão disponíveis com os conjuntos de dados. Assim, o mapa gerado a partir de um conjunto não rotulado resultaria em uma representação visual com uma única informação de cor. Para explorar o conjunto de dados o usuário utilizaria somente as relações de similaridade para explorar o mapa. No entanto, tendo um conjunto de treino, pode-se utilizar um classificador para rotular o conjunto desconhecido. Em seguida, os rótulos podem ser incorporados à representação visual, facilitando a identificação de grupos no mapa de similaridade. Essa e outras aplicações que não estão relacionadas com a coordenação podem ser encontradas no artigo "Visual analysis of image collections" (Eler et al., 2009a).

Essa nova abordagem, proposta com a PEx-Image, mostra que as técnicas de coordenação são importantes para a análise visual dos resultados das projeções, comparação de diferentes configurações de parâmetros para gerar as projeções, comparação de diferentes espaços multidimensionais e de diferentes medidas de similaridade. Também mostra que técnicas de coordenação podem ser utilizadas para auxiliar na interpretação de conjuntos de dados relacionados de qualquer natureza, como, por exemplo, imagens e textos descrevendo as mesmas entidades.

A exploração coordenada de mapas de similaridade gerou novas perspectivas de exploração. Neste trabalho foi desenvolvida uma ferramenta que utiliza os mapas de similaridade para explorar dados volumétricos, por meio da coordenação entre os mapas e o espaço do objeto de dados volumétrico. A união de técnicas de visualização de informação com técnica de visualização científica é uma tendência no processo de exploração de conjuntos de dados. É nossa perspectiva que os mapas de similaridade unidos com outras técnicas de visualização possam auxiliar o usuário na identificação de padrões em conjunto de dados volumétricos.

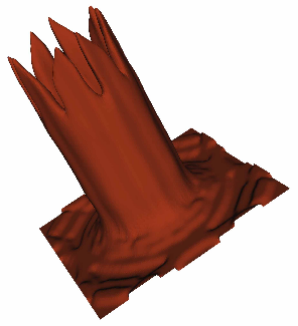
No próximo capítulo é apresentada uma conclusão sobre o trabalho desenvolvido e também sobre possíveis trabalhos futuros.

(a) Seleção de pontos com alta concentração de hidrogênio.

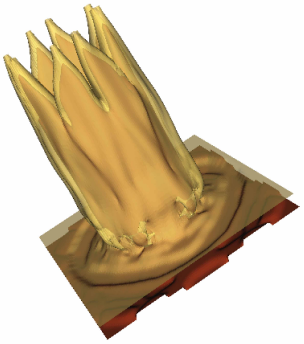


(b) Superfície que representa os pontos com alta concentração de hidrogênio selecionados.

(c) Seleção de pontos com baixa concentração de hidrogênio.



(d) Superfície que representa os pontos com baixa concentração de hidrogênio selecionados.



(e) Visualização das duas superfícies.

Figura 6.26: Exploração de um conjunto de dados de uma simulação. O mapa de similaridade gerado pela técnica P-LSP foi colorido com base no atributo que representa o elemento hidrogênio. As superfícies foram extraídas com base na seleção de duas regiões de alta densidade no mapa, as quais são regiões de alta (a) e baixa concentração (b) de hidrogênio.

CAPÍTULO

7

Conclusões e Trabalhos Futuros

7.1 Conclusões

Este trabalho apresentou algumas soluções para auxiliar no processo de exploração de conjuntos de dados. Essas soluções utilizam mapas de similaridade para visualmente representar o relacionamento entre as instâncias dos conjuntos de dados. Adicionalmente, elas também utilizam a abordagem de múltiplas visões coordenadas (CMV) para que o usuário tenha diferentes perspectivas de um ou vários conjuntos de dados. Nesse sentido, técnicas de coordenação foram desenvolvidas para auxiliar o usuário a descobrir novas relações entre os conjuntos representados pelas diferentes visões. Essas técnicas permitem a análise simultânea de um único conjunto de dados, de diferentes representações de um conjunto e de conjuntos distintos.

Com o desenvolvimento desse trabalho foi possível notar que mapas de similaridade são técnicas de visualização que podem produzir um melhor resultado exploratório quando são coordenados. Por meio de uma simples técnica de coordenação por identidade (*Identity Coordination*) foi possível confirmar quais os melhores parâmetros para construir o mapa de similaridade que melhor representasse as similaridades entre as instâncias de um conjunto de dados. Também foi possível verificar quais eram os melhores grupos formados e também quais grupos poderiam pejudicar a qualidade visual do mapa. Além disso, a coordenação de mapas de similaridade é uma forma interessante de auxiliar a exploração de conjuntos de dados que possuem diferentes representações para cada instância. Um exemplo apresentado foi a exploração de um conjunto de dados que possui representações de imagens e texto para cada instância.

Novas técnicas de coordenação foram desenvolvidas com o objetivo de auxiliar a exploração de conjuntos de dados distintos. Uma técnica que produziu resultados interessantes é a *Topic*

Coordination (Eler et al., 2008b), a qual é baseada em tópicos. Por meio dela é possível descobrir grupos de instâncias (e.g., documentos ou imagens) de interesse em mapas ainda não explorados a partir de um mapa já conhecido. Assim, o usuário pode focar nas instâncias destacadas pela coordenação e também em sua vizinhança, pois mapas de similaridade tendem a agrupar instâncias que são similares. Além de ser aplicada em coleções de documentos, essa técnica também foi aplicada em mapas construídos a partir de imagens e suas anotações textuais, auxiliando a exploração da coleção de imagens por meio do conteúdo textual relacionado às imagens.

Também baseada em tópicos, a *Topic Time Coordination* (Eler et al., 2009b) auxilia a descoberta da evolução de temas em coleções de documentos. A partir de uma base de dados com documentos de diferentes anos, foi possível descobrir o tópico de documentos de interesse em um intervalo de tempo e também a evolução desse tópico em outros intervalos de tempo, por meio dos documentos que vieram a se relacionar com os documentos da seleção inicial.

Uma técnica que também auxilia a explorar conjuntos de dados distintos é a coordenação baseada em distância *Distance Coordination* (Eler et al., 2008b). Por meio dela o usuário pode utilizar um conjunto de dados conhecido para explorar conjuntos ainda desconhecidos. Essa técnica revela em outros mapas as instâncias mais similares àquelas de interesse que foram selecionadas em um mapa. As instâncias reveladas são um passo inicial para a exploração por similaridade dos demais mapas disponíveis. Essa é uma grande vantagem de utilizar os mapas de similaridade para a visualização, pois tem-se o contexto específico no mapa pela proximidade entre as instâncias. Esse tipo de abordagem justifica a utilização da visualização ao invés de simplesmente utilizar um sistema de recuperação de informação tradicional.

Uma outra técnica desenvolvida foi a *Time Coordination*, que utiliza atributos temporais para realizar a coordenação entre os mapas. Essa técnica não produz bons resultados quando isoladamente utilizada, mas sua aplicação pode tornar-se interessante quando combinada com outras técnicas de coordenação (e.g., *Topic Coordination*). Nesse intuito de unir a capacidade de exploração de técnicas distintas, foi proposta uma técnica de coordenação que combina o mapeamento das técnicas de coordenação, isto é, ela gera um mapeamento final com as instâncias comuns entre os diferentes mapeamentos. Esta técnica foi denominada *Joint Coordination*. Por meio dela é possível criar mapeamentos mais precisos para explorar conjuntos de dados. Um exemplo de aplicação foi a união da *Topic Coordination* com a *Distance Coordination*, facilitando o foco do usuário em grupos de interesse que combinam tópico e similaridade de conteúdo no mapa destino. Outros tipos de combinação de mapeamento podem ser futuramente explorados pela *Joint Coordination*.

Na literatura, há inúmeras aplicações que utilizam CMV. No entanto, poucas são as que focam a exploração coordenada de coleções de documentos. Este trabalho utilizou as diferentes técnicas de coordenação propostas para auxiliar no processo de exploração desse tipo de conjuntos de dados, permitindo que diferentes coleções de documentos pudessem ser coordenadas. Assim, uma coleção pode contribuir para a exploração de outra coleção. Essas novas técnicas foram adicionadas à ferramenta *Projection Explorer* (PEx) (Paulovich et al., 2007), apoiando as ferramentas já existentes para exploração de coleções de documentos.

Alguns sistemas utilizam representações visuais para exploração de coleções de imagens, mas raramente empregam técnicas de coordenação para auxiliar no processo de exploração. Neste trabalho, foi criada uma abordagem denominada *Projection Explorer for Images* (PEx-Image) (Eler et al., 2008a, 2009a), a qual utiliza mapas de similaridade para representar o relacionamento entre as instâncias de uma coleções de imagens. Essa abordagem também utiliza CMV para auxiliar a exploração de diferente coleções e auxiliar a identificação dos melhores parâmetros para gerar bons mapa de similaridade. Além disso, a PEx-Image permite que o usuário utilize dados textuais relacionados às imagens para enriquecer o processo exploratório, possibilitando a coordenação entre mapas construídos das imagens e mapas construídos das informações textuais. Além de uma exploração coordenada com base no identificador das instâncias, é possível utilizar a coordenação baseada em tópicos para destacar imagens relacionadas. Nossa perspectiva é que esse tipo de exploração possa ser aplicado em dados da área médica, coordenando mapas construídos a partir de imagens com mapas construídos a partir de laudos ou anotações das imagens. A PEx-Image é uma contribuição significativa deste trabalho de doutorado e teve várias ramificações em trabalhos desenvolvidos no grupo de Visualização, Imagens e Computação Gráfica do ICMC/USP. A PEx-Image foi um passo inicial para muitas aplicações envolvendo exploração baseada em mapas de similaridade e também CMV.

Uma recente frente de pesquisa em CMV é unir técnicas de visualização de informação para explorar conjuntos de dados que eram explorados somente por técnicas de visualização científica. Pesquisas relacionadas com essa frente foram iniciadas no grupo de Visualização, Imagens e Computação Gráfica do ICMC/USP por meio deste trabalho. As pesquisas envolveram a criação de uma nova técnica de projeção para criar mapas de similaridade a partir de grandes conjuntos de dados, isto é, conjuntos com mais de um milhão de instâncias. Essa nova técnica de projeção foi denominada *Piecewise Least Squares Projection* (P-LSP) (Paulovich et al., 2010). Por meio da P-LSP é possível criar mapas que representassem o relacionamento entre as instâncias (voxels) de dados volumétricos. Para visualizar a seleção realizada nos grupos de estruturas formados pelos mapas de similaridade são empregadas técnicas de visualização, as quais são coordenadas com os mapas e refletiram no espaço do objeto (e.g., volume) as seleções realizadas neles.

A literatura apresenta diversas aplicações CMV, mas poucas relatam qual o modelo de coordenação utilizado. O modelo de coordenação é a base para a construção e definição de múltiplas visões coordenadas. Neste trabalho foi selecionado um modelo bem conhecido proposto por Boukhelifa e Rodgers (Boukhelifa e Rodgers, 2003), o qual foi adaptado para que as técnicas de coordenação propostas pudessem ser desenvolvidas e aplicadas. Esse modelo permite que técnicas possam ser criadas e configuradas durante o processo de exploração. Além disso, durante o processo de exploração, algumas das técnicas podem gerar diferentes tipos de mapeamentos, ou seja, o relacionamento entre as instâncias das visões é criado por meio de mapeamentos dinâmicos que variam de acordo com a interação do usuário. Esse modelo também possibilita que diferentes técnicas de visualização sejam adicionadas ao sistema, como a união das técnicas de visualização

de mapas de similaridade com as técnicas de visualização volumétrica para exploração de dados volumétricos.

A seguir são apresentados alguns possíveis trabalhos futuros identificados durante esse trabalho de doutorado.

7.2 Trabalhos Futuros

Após o desenvolvimento das técnicas de coordenação, a adaptação de um modelo de coordenação e a aplicação de todos eles, foi possível notar algumas carências que poderão ser futuramente abordadas. Algumas delas são:

- **Range Coordination:** técnica de coordenação que mapeia em outras visões as instâncias que possuem atributos com valores que estejam dentro da mesma faixa de valores dos atributos das instâncias selecionadas na visão origem. Essa técnica de coordenação pode ser empregada para descobrir instâncias que possuam características semelhantes às instâncias selecionadas. Essa técnica foi descrita na Seção 4.3.7 como exemplo de como adicionar uma nova técnica de coordenação;

- **Visual Attribute Coordination:** essa técnica de coordenação poderia transmitir atributos visuais (e.g., cor, forma, tamanho) para os elementos destacados na outras visões. Isso também facilitaria a identificação dos diferentes grupos ou estruturas destacados. Por exemplo, a exploração de um conjunto de dados volumétrico, as superfícies poderiam receber informação de cor e opacidade; ou então os voxels envolvidos em uma seleção em um mapa de similaridade, produzindo um efeito similar a uma função transferência;

- **Scalar Coordination:** essa técnica teria a finalidade de transmitir dados de campos escalares para outras instâncias. Por exemplo, alguns conjuntos de dados estão classificados (rotulados) e essa informação de classe pode ser utilizada para colorir os elementos ou estruturas visualizados. No entanto, alguns conjuntos não possuem essa informação, dificultando a discriminação de diferentes grupos. Isso é o que ocorre em conjuntos de dados reais e desconhecidos. Essa técnica poderia ser combinada com uma técnica de coordenação baseada em distância que auxiliaria a transmitir os escalares para instância similares. Nesse caso, essa coordenação atuaria como um pseudo-classificador manual;

- **"Disseminar o conhecimento" adquirido durante a exploração:** a exploração de alguns conjuntos de dados é um processo trabalhoso e que pode envolver vários especialistas. As descobertas durante a exploração poderiam ser armazenadas, catalogadas e disseminadas para outras pessoas envolvidas. Neste trabalho a exploração é dada por meio de coordenação, seria interessante que os mapeamentos pudessem ser disseminados para outras pessoas seguirem o caminho percorrido por quem os criou. Atualmente existem pesquisas sobre

o armazenamento dos passos realizados no processo exploratório e também pesquisas que utilizam a coordenação como um meio de colaboração;

- **Recuperação de imagens:** os mapas de similaridade revelam em uma única visualização os relacionamentos de similaridade entre as instâncias de um conjunto de dados. O resultado visual de um mapa pode ser utilizado como uma rápida recuperação de imagens ou grupos de imagens de interesse. Adicionalmente, técnicas de recuperação de imagens (*Content-based image retrieval* – CBIR) poderiam ser adicionadas ao processo de exploração desses mapas, revelando novos tipos de relacionamento em um mesmo mapa ou em diferentes mapas por meio da coordenação;

- **Arcabouço de visualização e de coordenação:** a construção de novas técnicas de coordenação a partir do modelo proposto neste trabalho é algo relativamente simples, sendo necessário o desenvolvedor se adaptar a algumas partes da estrutura da ferramenta de visualização e do modelo de coordenação. No entanto, adicionar novas técnicas de visualização não é uma tarefa que possui o mesmo grau de simplicidade. Portanto, é essencial a construção de um arcabouço que permita com mais facilidade a adição de novas técnicas de visualização e de coordenação;

- **Escalabilidade visual:** As coleções de documentos e imagens utilizadas neste trabalho podem ser consideradas pequenas, quando comparadas a algumas aplicações reais. O maior conjunto de dados manipulado continha 9000 elementos, o que já causa uma certa dificuldade no processo de exploração por meio dos mapas de similaridade. Esse problema deve ser confrontado o quanto antes, pois por meio de sua solução será possível criar ferramentas e técnicas mais próximas da realidade;

- **Auxilio ao processo de classificação:** as aplicações apresentadas neste trabalho indicam que os mapas de similaridade refletem o espaço multidimensional. Nesse sentido, foram executadas aplicações para comparar diferentes espaços multidimensionas e também parâmetros para os gerar. É nossa perspectiva que os mapas possam ser úteis para identificar problemas no processo de classificação, isto é, nas etapas anteriores à configuração de um classificador. Assim, técnicas rápidas de construção de mapas de similaridade podem contribuir para a escolha dos melhores parâmetros utilizados nas técnicas envolvidas no processo de classificação. Essa análise não pode ser baseada unicamente em coordenação, pois resultaria em uma análise subjetiva. Portanto, medidas devem ser calculadas para verificar quais são os melhores mapas. Com isso, o usuário teria a inspeção visual para identificar os problemas, podendo inferir sobre qual parâmetro modificar, e uma medida objetiva para o apoiar as comparações;

- **Avaliação e comparação de projeções e mapas:** foram apresentadas aplicações que comparavam diferente mapas de similaridade. No entanto, a comparação por meio da coordenação

não é a maneira mais confiável de fazer isso em muitas aplicações, pois ela adiciona a subjetividade do usuário em dizer o que é bom e o que não é. Portanto, é necessário criar formas de comparar diferentes mapas de similaridade por meio de medidas objetivas;

- **Avaliação com usuários:** as técnicas desenvolvidas neste trabalho foram claramente aplicadas para explorar coleções de documentos, coleções de imagens e dados volumétricos. Entretanto, é necessário realizar aplicações com diferentes usuários para identificar problemas, novas funcionalidades e o grau de eficácia das técnicas desenvolvidas;

- **Evolução temporal de dados volumétricos:** a aplicação desenvolvida para análise de dados volumétricos é o passo inicial para outras formas de exploração utilizando mapas de similaridade, outras técnicas de visualização e coordenação. Atualmente, parte dos conjuntos de dados volumétricos está disponível na forma temporal. Um campo de pesquisa atual é a identificação de padrões temporais por meio de técnicas de visualização de informação. Para isso, além de técnicas capazes de fazer esse tipo de identificação, é necessário a construção de técnicas rápidas de visualização, a melhoria da escalabilidade visual e da interação sobre os dados. Esse tipo de aplicação é um grande desafio para a área de visualização e de múltiplas visões coordenadas.

Referências Bibliográficas

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, p. 487–499.

AHMED, A.; DWYER, T.; FORSTER, M.; FU, X.; HO, J.; HONG, S.-H.; KOSCHÜTZKI, D.; MURRAY, C.; NIKOLOV, N. S.; TAIB, R.; TARASSOV, A.; XU, K. GEOMI: Geometry for maximum insight. In: *Proceeding of 13th International Symposium on Graph Drawing (GD '05)*, Limerick, Ireland, 2005, p. 468–479.

ALTSCHUL, S. F.; W. GISH, W. M.; MYERS, E. W.; LIPMAN, D. J. A basic local alignment search tool. *Journal of Molecular Biology*, v. 215, p. 403–410, 1990.

ANDERY, G. F. Integrando projeções multidimensionais à analise visual de redes sociais. Dissertação de mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 2010.

ANDREWS, K.; KIENREICH, W.; SABOL, V.; BECKER, J.; DROSCHL, 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.

ANKERST, M. *Visual data mining*. Tese de Doutoramento, Faculty of Mathematics and Computer Science, University of Munich, 2000.

BECKER, R. A.; CLEVELAND, W. S. Brushing scatterplots. *Technometrics*, v. 29, n. 2, p. 127–142, 1987.

BERTINI, E.; AQUILA, L. D.; SANTUCCI, G. SpringView: Cooperation of radviz and parallel coordinates for view optimization and clutter reduction. In: *Proceedings of the Third International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 22–29.

BERTINI, E.; CATARCI, T.; KIMANI, S.; SANTUCCI, G. Exploiting multiple views to support visual exploration and mining. In: *Proceedings of the Second International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'04)*, Washington, DC, USA: IEEE Computer Society, 2004, p. 15–23.

BLAAS, J.; POST, C. P. B. F. Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 6, p. 1436–1451, 2008.

BOUKHELIFA, N.; ROBERTS, J. C.; ROBERTS, P. J.; RODGERS, P. J. A coordination model for exploratory multi-view visualization. In: *Proceedings of the International Conference on Coordinated and Multiple Views In Exploratory Visualization (CMV '03)*, Washington, DC, USA: IEEE Computer Society, 2003, p. 76–85.

BOUKHELIFA, N.; RODGERS, P. A model and software system for coordinated and multiple views in exploratory visualization. *Information Visualization*, v. 2, n. 4, p. 258–269, 2003. Disponível em: `http://www.cs.kent.ac.uk/pubs/2003/1777`

BOYACK, K. W.; WYLIE, B. N.; DAVIDSON, G. S. Domain visualization using VxInsight for science and technology management. *Journal of the American Society for Information Science and Technology*, v. 53, n. 9, p. 764–774, 2002.

BRANDOLI, B.; ELER, D. M.; PAULOVICH, F. V.; MINGHIM, R.; BATISTA, J. Visual data exploration to feature space definition. In: *Proceedings of the XXIII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2010)*, Gramado, Brazil: IEEE Computer Society, 2010, p. 32–39.

BÖRNER, K.; CHEN, C.; BOYACK, K. W. Visualizing knowledge domains. *Annual Review of Information Science & Technology*, v. 37, p. 179–255, 2003.

CAMARGO, J.; GONZÁLEZ, F.; TORRES, R. Visualization of large collection of medical images. In: *Lecture Notes in Computer Science*, Universidad Autónoma de Bucaramanga, 2008, p. 615–622.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. *Readings in information visualization: using vision to think* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., p. 463–464, 1999.

CARPENDALE, S. Vislink: Revealing relationships amongst visualizations. *IEEE Transactions on Visualization and Computer Graphics*, v. 13, n. 6, p. 1192–1199, 2007.

CHALMERS, M. A linear iteration time layout algorithm for visualising high-dimensional data. In: *Proceedings of the 7th conference on Visualization (VIS '96)*, Los Alamitos, CA, USA: IEEE Computer Society Press, 1996, p. 127–132.

CHANG, R.; GHONIEM, M.; KOSARA, R.; RIBARSKY, W.; YANG, J.; SUMA, E.; ZIEMKIEWICZ, C.; KERN, D.; SUDJIANTO, A. WireVis: Visualization of categorical, time-varying data from financial transactions. In: *IEEE Symposium on Visual Analytics Science and Technology*, Sacramento, CA, USA, 2007, p. 155 – 162.

CHASTINE, J. W.; ZHU, Y.; BROOKS, J. C.; OWEN, G. S.; HARRISON, R. W.; WEBER, I. T. A collaborative multi-view virtual environment for molecular visualization and modeling. In: *Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 77–84.

CHEN, C. Searching for intellectual turning points: progressive knowledge domain visualization. *Proceedings of the National Academy of Sciences of the USA*, v. 101, Suppl 1, p. 5303–5310, 2004.

CHEN, C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, v. 57, n. 3, p. 359–377, 2006.

CHEN, C.; CHEN, Y.; MAULITZ, R. C. Understanding the evolution of NSAID: A knowledge domain visualization approach to evidence-based medicine. In: *Proceedings of the Ninth International Conference on Information Visualisation (IV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 945–952.

CHEN, C.; GAGAUDAKIS, G.; ROSIN, P. Similarity-based image browsing. In: *XVI IFIP World Computer Congress, International Conference on Intelligent Information Processing*, Beijing, China, 2000, p. 206–213.

CHEN, C.; THOMAS, L.; COLE, J.; CHENNAWASIN, C. Representing the semantics of virtual spaces. *IEEE MultiMedia*, v. 6, n. 2, p. 54–63, 1999.

CHEN, T. T.; HSIEH, L. C. On visualization of cocitation networks. In: *Proceedings of the 11th International Conference on Information Visualization (IV '07)*, Washington, DC, USA: IEEE Computer Society, 2007, p. 470–475.

CONVERTINO, G.; CHEN, J.; YOST, B.; RYU, Y.-S.; NORTH, C. Exploring context switching and cognition in dual-view coordinated visualizations. In: *Proceedings of the International Conference on Coordinated and Multiple Views In Exploratory Visualization (CMV '03)*, Washington, DC, USA: IEEE Computer Society, 2003, p. 55.

COSTA, L. F.; JUNIOR, R. M. C. *Shape analysis and classification: Theory and practice*. Boca Raton, FL, USA: CRC Press, Inc., 2000.

COX, T. F.; COX, M. A. A. *Multidimensional scaling*. Second ed. Chapman & Hall/CRC, 2000.

CRAIG, P.; KENNEDY, J.; CUMMING, A. Coordinated parallel views for the exploratory analysis of microarray time-course data. In: *Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 3–14.

CUADROS, A. M.; PAULOVICH, F. V.; MINGHIM, R.; TELLES, G. P. Point placement by phylogenetic trees and its application for visual analysis of document collections. In: *IEEE Symposium on Visual Analytics Science and Technology 2007*, Sacramento, CA, USA, 2007, p. 99–106.

DAUGMAN, J.; DOWNING, C. The handbook of brain theory and neural networks. cáp. Gabor wavelets for statistical pattern recognition, Cambridge, MA, USA: MIT Press, p. 414–420, 1998.

DAVIDSON, G. S.; WYLIE, B. N.; BOYACK, K. W. Cluster stability and the use of noise in interpretation of clustering. In: *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS '01)*, Washington, DC, USA: IEEE Computer Society, 2001, p. 23.

DEMMEL, J. W. *Applied numerical linear algebra*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1997.

DRINEAS, P.; FRIEZE, A.; KANNAN, R.; VEMPALA, S.; VINAY, V. Clustering large graphs via the singular value decomposition. *Machine Learning*, v. 56, p. 9–33, 2004.

EADES, P. A. A heuristic for graph drawing. In: *Congressus Numerantium*, 1984, p. 149–160.

ECK, N. J.; FRASINCAR, F.; BERG, J. Visualizing concept associations using concept density maps. In: *Proceedings of the International Conference on Information Visualization (IV '06)*, Washington, DC, USA: IEEE Computer Society, 2006, p. 270–275.

ECK, N. J.; WALTMAN, L.; BERG, J. A novel algorithm for visualizing concept associations. In: *Proceedings of the 16th International Workshop on Database and Expert System Applications*, IEEE Computer Society, 2005, p. 405–409.

EICK, S. G.; COUSINS A survey of techniques to visualize streaming textual datasets. In: *IEEE Aerospace Conference*, 2006, p. 8pp.

ELER, D.; NAKAZAKI, M.; PAULOVICH, F.; SANTOS, D.; ANDERY, G.; OLIVEIRA, M.; BATISTA, J. E. S.; MINGHIM, R. Visual analysis of image collections. *The Visual Computer*, v. 25, n. 10, p. 923–937, 2009a.

ELER, D. M.; NAKAZAKI, M. Y.; PAULOVICH, F. V.; SANTOS, D. P.; OLIVEIRA, M. C. F.; BATISTA, J. E. S.; MINGHIM, R. Multidimensional visualization to support analysis of image collections. In: *Proceedings of the XXI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI '08)*, Washington, DC, USA: IEEE Computer Society, 2008a, p. 289–296.

ELER, D. M.; PAULOVICH, F. V.; OLIVEIRA, M. C. F. D.; MINGHIM, R. Coordinated and multiple views for visualizing text collections. In: *Proceedings of the 12th International Conference Information Visualisation (IV '08)*, Washington, DC, USA: IEEE Computer Society, 2008b, p. 246–251.

ELER, D. M.; PAULOVICH, F. V.; OLIVEIRA, M. C. F. D.; MINGHIM, R. Topic-based coordination for visual analysis of evolving document collections. In: *Proceedings of the 13th International Conference Information Visualisation (IV '09)*, Washington, DC, USA: IEEE Computer Society, 2009b, p. 149–155.

ERICSON, D.; JOHANSSON, J.; COOPER, M. Visual data analysis using tracked statistical measures within parallel coordinate representations. In: *Proceedings of the Third International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 42–53.

FALOUTSOS, C.; LIN, K. Fastmap: A fast algorithm for indexing, data mining and visualization of traditional and multimedia databases. In: *Proceedings of the International Conference on Management of Data (SIGMOD '95)*, San Jose-CA, USA: ACM Press: New York, 1995, p. 163–174.

FAN, J.; GAO, Y.; H.LUO Hierarchical classification for automatic image annotation. In: *Proceedings XXX ACM International Conference on Research and Development in Information Retrieval*, New York, USA: ACM Press, 2007, p. 111–118.

FEKETE, J.-D. The infovis toolkit. In: *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS '04)*, Washington, DC, USA: IEEE Computer Society, 2004, p. 167–174.

FRICK, A.; LUDWIG, A.; MEHLDAU, H. A fast adaptive layout algorithm for undirected graphs. In: *Proceedings of the DIMACS International Workshop on Graph Drawing (GD '94)*, London, UK: Springer-Verlag, 1995, p. 388–403.

FRUCHTERMAN, T. M. J.; REINGOLD, E. M. Graph drawing by force-directed placement. *Software - Practice & Experience*, v. 21, n. 11, p. 1129–1164, 1991.

GAMMA, E.; HELM, R.; JOHNSON, R. E.; VLISSIDES, J. *Design patterns: Elements of reusable object-oriented software*. Reading, MA: Addison-Wesley, 1995.

GARSON, G. D. Interpreting neural-network connection weights. *AI Expert*, v. 6, p. 46–51, 1991.

GIERETH, M.; KOCH, S.; ROTARD, M.; ERTL, T. Web based visual exploration of patent information. In: *Proceedings of the 11th International Conference on Information Visualization (IV '07)*, Washington, DC, USA: IEEE Computer Society, 2007, p. 150–155.

GODINHO, P. I. A.; MEIGUINS, B. S.; MEIGUINS, A. S. G.; CARMO, R. M. C.; BRITO GARCIA, M.; ALMEIDA, L. H.; LOURENCO, R. PRISMA - a multidimensional information visualization tool using multiple coordinated views. In: *Proceedings of the 11th International Conference on Information Visualization (IV '07)*, Washington, DC, USA: IEEE Computer Society, 2007, p. 23–32.

GRINSTEIN, G. G.; WARD, M. O. *Information visualization in data mining and knowledge discovery. pp. 21–45*. Morgan Kaufmann Publishers Inc., 2002.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural features for image classification. *IEEE Transactions on System, Man and Cybernetics*, v. 3, n. 6, p. 610–621, 1973.

HAVRE, S.; HETZLER, E.; WHITNEY, P.; NOWELL, L. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, v. 8, n. 1, p. 9–20, 2002.

HEER, J.; CARD, S. K.; LANDAY, J. A. Prefuse: a toolkit for interactive information visualization. In: *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI '05)*, New York, NY, USA: ACM, 2005, p. 421–430.

HEIJS, A. Requirements for coordinated multiple view visualization systems for industrial applications. In: *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '07)*, Washington, DC, USA: IEEE Computer Society, 2007, p. 76–79.

HINTON, G.; ROWEIS, S. Stochastic neighbor embedding. In: *Advances in Neural Information Processing Systems 15*, MIT Press, 2002, p. 833–840.

HOEBER, O.; YANG, X. D. Exploring web search results using coordinated views. In: *Proceedings of the Fourth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '06)*, Washington, DC, USA: IEEE Computer Society, 2006, p. 3–13.

HUANG, K.; AVIYENTE, S. Rotation invariant texture classification with ridgelet transform and fourier transform. In: *Proceedings of the International Conference on Image Processing*, IEEE Computer Society, 2006, p. 2141–2144.
Disponível em: `http://dblp.uni-trier.de/db/conf/icip/icip2006.html#HuangA06a`

HUANG, S.; WARD, M. O.; RUNDENSTEINER, E. A. Exploration of dimensionality reduction for text visualization. In: *Proceedings of the Third International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 63–74.

INSELBERG, A. *Parallel coordinates: Visual multidimensional geometry and its applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2009.

INSELBERG, A.; DIMSDALE, B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: *Proceedings of the 1st conference on Visualization (VIS '90)*, Los Alamitos, CA, USA: IEEE Computer Society Press, 1990, p. 361–378.

JACOBSON, A. S.; BERKIN, A. L.; ORTON, M. N. Linkwinds: interactive scientific data analysis and visualization. *Communications of the ACM*, v. 37, n. 4, p. 42–52, 1994.

JERN, M.; RANLÖF, M.; PALMBERG, S.; NILSSON, A. Coordinated views in dynamic interactive documents. In: *Proceedings of the International Conference on Coordinated and Multiple Views In Exploratory Visualization (CMV '03)*, Washington, DC, USA: IEEE Computer Society, 2003, p. 95.

JOLLIFFE, I. T. *Principal component analysis* New York, USA: Springer-Verlag, 1986.

KEIM, D. A. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, v. 8, p. 1–8, 2002.

KEIM, D. A.; KRIEGE, H.-P. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering*, v. 8, n. 6, p. 923–938, 1996.

KELLER, R.; ECKERT, C. M.; CLARKSON, P. J. Multiple views to support engineering change management for complex products. In: *Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 33–41.

KLIEBER, W.; SABOL, V.; MUHR, M.; KERN, R.; ÖTTL, G.; GRANITZER, M. Knowledge discovery using the knowminer framework. *International Conference Information Systems (IADIS '09)*, v. Single, p. 307–314, 2009.

KOHONEN, T. The self-organizing map. *Proceedings of the IEEE*, v. 78, n. 9, p. 1464–1480, 1990.

LAMIREL, J.-C.; SHEHABI, S. A. MultiSOM: A multiview neural model for accurately analyzing and mining complex data. In: *Proceedings of the Fourth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '06)*, Washington, DC, USA: IEEE Computer Society, 2006, p. 42–54.

LAWRENCE, M.; LEE, E.-K.; COOK, D.; HOFMANN, H.; WURTELE, E. exploRase: Exploratory data analysis of systems biology data. In: *Proceedings of the Fourth International Conference on Coordinated an Multiple Views in Exploratory Visualization (CMV '06)*, Washington, DC, USA: IEEE Computer Society, 2006, p. 14–20.

LEE, R. C. T.; SLAGLE, J. R.; BLUM, H. A triangulation method for the sequential mapping of points from n-space to two-space. *IEEE Transactions on Computers*, v. 26, n. 3, p. 288–292, 1977.

LINSEN, L.; LONG, T. V.; ROSENTHAL, P.; ROSSWOG, S. Surface extraction from multi-field particle volume data using multi-dimensional cluster visualization. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 6, p. 1483–1490, 2008.

LINSEN, L.; VAN LONG, T.; ROSENTHAL, P. Linking multidimensional feature space cluster visualization to multifield surface extraction. *IEEE Computer Graphics and Applications*, v. 29, n. 3, p. 85–89, 2009.

LIVNY, M.; RAMAKRISHNAN, R.; BEYER, K.; CHEN, G.; DONJERKOVIC, D.; LAWANDE, S.; MYLLYMAKI, J.; WENGER, K. DEVise: integrated querying and visual exploration of large datasets. In: *Proceedings of the International Conference on Management of Data (SIGMOD '97)*, 1997, p. 301–312.

LOPES, A.; PINHO, R.; PAULOVICH, F.; MINGHIM, R. Visual text mining using association rules. *Computers & Graphics, Special Issue on Visual Analytics*, v. 31, n. 3, p. 316–326, 2007.

LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, v. 2, p. 159–165, 1958.

MACEACHREN, A.; WACHOWICZ, M.; EDSALL, R.; HAUG, D.; MASTERS, R. Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, v. 13, n. 4, p. 311–334, 1999.

MANE, K.; BÖRNER, K. Mapping topics and topic bursts in PNAS. In: *Proceeding of the National Academy of Sciences of the United States*, n. 101, 2004, p. 5287–5290.
Disponível em: `http://www.pnas.org/cgi/reprint/101/suppl_1/5287`

MANJUNATH, B. S.; MA, W.-Y. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 18, n. 8, p. 837–842, 1996.
Disponível em: `http://www.computer.org/tpami/tp1996/i0837abs.htm`

MATKOVIC, K.; JURIC, J.; KONYHA, Z.; KRASSER, J.; HAUSER, H. Interactive visual analysis of multi-parameter families of function graphs. In: *Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 54–62.

MICHAL J. PIETAL, I. T.; BUJNICKI, J. M. Protmap2d: visualization, comparison and analysis of 2d maps of protein structure. *Bioinformatics*, v. 23, n. 11, p. 1429–1430, 2007.

MORENO, V. Concepção e implementação de um modelo de coordenação para uma plataforma de visualização exploratória. Dissertação de mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 2006.

MURRAY, C.; KE, W.; BÖRNER, K. Mapping scientific disciplines and author expertise based on personal bibliography files. In: *Proceedings of the International Conference on Information Visualization (IV '06)*, Washington, DC, USA: IEEE Computer Society, 2006, p. 258–263.

NGUYEN, G. P.; WORRING, M. Interactive access to large image collections using similarity-based visualization. *Journal of Visual Languages and Computing*, v. 19, n. 2, p. 203–224, 2008.

NOOY, W.; MRVAR, A.; BATAGELJ, V. *Exploratory social network analysis with Pajek*. Cambridge: Cambridge University Press, 2005.

NORTH, C. *A taxonomy of multiple-window coordination*. Relatório Técnico #CS-TR-3854, University of Maryland Computer Science Department, 1997.

NORTH, C.; SHNEIDERMAN, B. *Snap-Together Visualization: Coordenating multiple views to explore information*. Relatório Técnico #CS-TR-4020, University of Maryland Computer Science Department, 1999.

NORTH, C.; SHNEIDERMAN, B. Snap-Together Visualization: A user interface for coodinating visualizations via relational schemata. In: *Advanced Visual Interfaces*, 2000a, p. 128–135.

NORTH, C.; SHNEIDERMAN, B. Snap-together visualization: can users construct and operate coordinated visualizations? *International Journal of Human-Computer Studies - Empirical evaluation of information visualizations*, v. 53, n. 5, p. 715–739, 2000b.

OLIVEIRA, M. C. F.; LEVKOWITZ, H. From Visual Data Exploration to Visual Data Mining: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, v. 9, n. 3, p. 378–394, 2003.

PAPADIMITRIOU, C. H.; RAGHAVAN, P.; TARNAKI, H.; VEMPALA, S. Latent semantic indexing: A probabilistic analysis. In: *Proceedings of the 17th ACM Symposium on the Principles of Database Systems*, 1998, p. 159–168.

PAULOVICH, F. V. *Mapeamento de dados multidimensionais – integrando mineração e visualização*. Tese de Doutoramento, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2008.

PAULOVICH, F. V.; ELER, D. M.; POCO, J.; NONATO, L. G.; BOTHA, C. P.; MINGHIM, R. *A fast projection technique and its applications to visualization of large data sets*. Relatório Técnico 349, Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, 2010.

PAULOVICH, F. V.; MINGHIM, R. Text Map Explorer: a tool to create and explore document maps. In: *Proceedings of the International Conference on Information Visualization (IV '06)*, Washington, DC, USA: IEEE Computer Society, 2006, p. 245–251.

PAULOVICH, F. V.; MINGHIM, R. HiPP: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of Information Visualization 2008 (INFOVIS '08))*, v. 14, n. 6, p. 1229–1236, 2008.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOVITZ, H. Visual mapping of text collections through a fast high precision projection technique. In: *Proceedings of the 10th International Conference on Information Visualisation, London, UK*, Washington, DC, USA: IEEE Computer Society, 2006, p. 282–290.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least Square Projection: a fast high precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 3, p. 564–575, 2008.

PAULOVICH, F. V.; OLIVEIRA, M. C. F.; MINGHIM, R. The Projection Explorer: A flexible tool for projection-based multidimensional visualization. In: *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI '07)*, Washington, DC, USA: IEEE Computer Society, 2007, p. 27–36.

PEKALSKA, E.; RIDDER R. P. W. DUIN, D.; KRAAIJVELD, M. A. A new method of generalizing Sammon mapping with application to algorithm speed-up. In: BOASSON, M.; KAANDORP, J. A.; TONINO, J. F. M.; VOSSELMAN, M. G., eds. *Proceedings of Fifth Annual Conference of the Advanced School for Computing and Imaging (ASCI '99)*, Delft, Netherlands, 1999, p. 221–228.

PEREZ, C.; ANTONIO, A. 3D visualization of text collections: An experimental study to assess the usefulness of 3d. In: *Proceedings of the Eighth International Conference on Information Visualization (IV '04)*, Washington, DC, USA: IEEE Computer Society, 2004, p. 317–323.

PILLAT, R. M. Coordenação dinâmica de visualizações de dados multidimensionais. Dissertação de mestrado, Instituto de Informática da Universidade Federal do Rio Grande do Sul. 2006.

PINHO, R. *Espaço incremental para a mineração visual de conjuntos dinâmicos de documentos*. Tese de Doutoramento, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2009.

PINHO, R.; LOPES, A.; OLIVEIRA, M. C. F. Incremental board: A grid-based space for visualizing dynamic data sets. In: *Proceedings of the 2009 ACM Symposium on Applied Computing*, 2009, p. 1757–1764.

PINHO, R.; OLIVEIRA, M. C. F. Hexboard: Conveying pairwise similarity in an incremental visualization space. In: *IV '09: Proceedings of the 2009 13th International Conference Information Visualisation*, Washington, DC, USA: IEEE Computer Society, 2009, p. 32–37.

PIRINGER, H.; KOSARA, R.; HAUSER, H. Interactive focus+context visualization with linked 2D/3D scatterplots. In: *Proceedings of the Second International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV'04)*, Washington, DC, USA: IEEE Computer Society, 2004, p. 49–60.

PORTER, M. F. *An algorithm for suffix stripping* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., p. 313–316, 1997.

ROMESBURG, H. C. *Cluster analysis for researchers*. Belmont, California: Lifetime Learning Publications, 1984.

RONCATTI, M. A. Avaliação de métodos ótimos e subótimos de seleção de características de texturas em imagens. Dissertação de mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 2008.

ROWEIS, S. T.; SAUL, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, v. 290, n. 5500, p. 2323–2326, 2000.

RUNDENSTEINER, E. A.; WARD, M. O.; YANG, J.; DOSHI, P. R. Xmdvtool: visual interactive data exploration and trend discovery of high-dimensional data sets. In: *Proceedings of the International Conference on Management of Data (SIGMOD '02)*, New York, NY, USA: ACM, 2002, p. 631–631.

SABOL, V.; KIENREICH, W.; MUHR, M.; KLIEBER, W.; GRANITZER, M. Visual knowledge discovery in dynamic enterprise text repositories. In: *IV '09: Proceedings of the 13th International Conference Information Visualisation*, Washington, DC, USA: IEEE Computer Society, 2009, p. 361–368.

SALTON, G.; BUCKLEY, C. *Term weighting approaches in automatic text retrieval*. Relatório Técnico TR87-881, Ithaca, NY, USA, 1987.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. *Communications of the ACM*, v. 18, p. 613–620, 1975.

SAMMON, J. W. A nonlinear mapping for data structure analysis. In: *IEEE Transactions on Computers*, 1964, p. 401–409.

SANTOS, D. P. Seleção de características: abordagem via redes neurais aplicada à segmentação de imagens. Dissertação de mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. 2007.

SANTOS, D. P.; BATISTA, J. E. S. Feature selection with equalized salience measures and its application to segmentation. In: *Proceedings of the XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI '07)*, IEEE Computer Society, 2007, p. 253–262.

SANVER, M.; YANG, L. A linking mechanism to integrate components of a visualization framework. In: *Proceedings of the 2009 13th International Conference Information Visualisation (IV '09)*, Washington, DC, USA: IEEE Computer Society, 2009, p. 92–97.

SCHAEFER, G. A next generation browsing environment for large image repositories. *Multimedia Tools Applications*, v. 47, n. 1, p. 105–120, 2010.

SCHAFHITZEL, T.; WEISKOPF, D.; ERTL, T. Interactive exploration of unsteady 3D flow with linked 2D/3D texture advection. In: *Proceedings of the Third International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '05)*, Washington, DC, USA: IEEE Computer Society, 2005, p. 96–105.

SCHROEDER, W. J.; MARTIN, K. M.; LORENSEN, W. E. The design and implementation of an object-oriented toolkit for 3D graphics and visualization. In: *Proceedings of the 7th conference on Visualization (VIS '96)*, Los Alamitos, CA, USA: IEEE Computer Society Press, 1996, p. 93–ff.

SCHVANEVELDT, R. W., ed. *Pathfinder associative networks: studies in knowledge organization*. Norwood, NJ, USA: Ablex Publishing Corp., 1990.

SEELING, C.; BECKS, A. Analysing associations of textual and relational data with a multiple views system. In: *Proceedings of the Second International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '04)*, IEEE Computer Society, 2004a, p. 61–70.

SEELING, C.; BECKS, A. Analysing associations of textual and relational data with a multiple views system. In: *Proceedings of the Second International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '04)*, IEEE Computer Society, 2004b, p. 61–70.

SIIRTOLA, H. Combining parallel coordinates with the reorderable matrix. In: *Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '03)*, IEEE Computer Society, 2003, p. 63–74.

SILVA, L. A.; MORENO, R. A.; FURUIE, S. S.; HERNANDEZ, E. D. M. Medical image categorization based on wavelet transform and self-organizing map. In: *Proceedings of VII International Conference on Intelligent Systems Design and Applications*, Washington, DC, USA: IEEE Computer Society, 2007, p. 353–356.

SKUPIN, A. A cartographic approach to visualizing conference abstracts. *IEEE Computer Graphics and Applications*, v. 22, n. 1, p. 50–58, 2002.

SPANGLER, W.; KREULEN, J.; LESSLER, J. Mindmap: Utilizing multiple taxonomies and visualization to understand a document collection. In: *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS '02)*, Washington, DC, USA: IEEE Computer Society, 2002, p. 102.

SPOERRI, A. Coordinated views and tight coupling to support meta searching. In: *Proceedings of the Second International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '04)*, IEEE Computer Society, 2004, p. 39–48.

SPOERRI, A. Coordinating linear and 2d displays to support exploratory search. In: *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '07)*, Washington, DC, USA: IEEE Computer Society, 2007, p. 16–26.

STOLTE, C.; HANRAHAN, P. Polaris: A system for query, analysis and visualization of multidimensional relational databases. In: *Proceedings of the IEEE Symposium on Information Vizualization 2000 (INFOVIS '00)*, Washington, DC, USA: IEEE Computer Society, 2000, p. 5.

TAKATSUKA, M.; GAHEGAN, M. Geovista studio: a codeless visual programming environment for geoscientific data analysis and visualization. *Computers & Geosciences*, v. 28, n. 10, p. 1131–1144, 2002.

TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, v. 2, n. 4, p. 218–231, 2003.

TELLES, G.; MINGHIM, R.; PAULOVICH, F. Visual analytics: Normalized compression distance for visual analysis of document collections. *Computers & Graphics, Special Issue on Visual Analytics*, v. 31, n. 3, p. 327–337, 2007.

TENENBAUM, J. B.; SILVA, V.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, v. 290, n. 5500, p. 2319–2323, 2000.

TUCERYAN, M.; JAIN, A. K. Texture analysis. In: *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, 1998, p. 235–276.

VAPNIK, V. N. *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag, 1995.

WEAVER, C. Building highly-coordinated visualizations in improvise. In: *Proceedings of the IEEE Symposium on Information Visualization (INFOVIS '04)*, Washington, DC, USA: IEEE Computer Society, 2004, p. 159–166.

WEAVER, C. E. *IMPROVISE: a user interface for interactive construction of highly-coordinated visualizations*. Tese de Doutoramento, University of Wisconsin, Madison, 2006.

WHALEN, D.; NORMAN, M. L. Competition data set and description. In: *2008 IEEE Visualization Design Contest*, http://vis.computer.org/VisWeek2008/vis/contests.html, 2008.

WISE, J. A. The ecological approach to text visualization. *Journal of the American Society for Information Science*, v. 50, n. 13, p. 1224–1233, 1999.

WISE, J. A.; THOMAS, J. J.; PENNOCK, K.; LANTRIP, D.; POTTIER, M.; SCHUR, A.; CROW, V. Visualizing the non-visual: spatial analysis and interaction with information for text documents. In: *Readings in information visualization: using vision to think*, San Francisco, CA - USA: Morgan Kaufmann Publishers Inc., 1995, p. 442–450.

WONG, P. C.; THOMAS, J. Visual analytics. *IEEE Computer Graphics and Applications*, v. 24, p. 20–21, 2004.

YANG, F.; GOODELL, H.; PICKETT, R.; BOBROW, R.; BAUMANN, A.; GEE, A.; GRINSTEIN, G. Data exploration combining kinetic and static visualization displays. In: *Proceedings of the Fourth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV '06)*, Washington, DC, USA: IEEE Computer Society, 2006, p. 21–30.

YANG, L. Distance-preserving projection of high-dimensional data for nonlinear dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 26, n. 9, p. 1243–1246, 2004.

APÊNDICE

A

Lista de Publicações

Artigos publicados que estão relacionados com o tema desta tese:

- PAULOVICH, Fernando Vieira; ELER, Danilo Medeiros; POCO, Jorge; BOTHA, Charl; NONATO, Luiz Gustavo; MINGHIM, Rosane. Piecewise Laplacian-based Projection for Interactive Large Data Exploration and Organization – journal Computer Graphics Forum – Eurographics / IEEE Symposium on Visualization (EUROVIS 2011);

- BRANDOLI, Bruno; ELER, Danilo Medeiros; PAULOVICH, Fernando Vieira; MINGHIM, Rosane; BATISTA NETO, João do E. S. Visual Data Exploration to Feature Space Definition. In: XXIII Brazilian Conference on Graphics, Patterns and Images (SIBGRAPI 2010), 2010, Gramado. Los Alamitos, CA : IEEE Computer Society, 2010. v. 1. p. 32 – 39

- ELER, Danilo Medeiros; Nakazaki, Marcel Y.; Paulovich, Fernando V.; Santos, Davi P., Andery, Gabriel F., Oliveira, Maria Cristina F., Batista Neto, João, MINGHIM, Rosane. Visual analysis of image collections. The Visual Computer. v.25, p.923 – 937, 2009

- ELER, Danilo Medeiros; PAULOVICH, Fernando Vieira; OLIVEIRA, M. C. F.; MINGHIM, Rosane. Topic-based Coordination for Visual Analysis of Evolving Document Collections In: XIII International Conference Information Visualization (IV09), 2009, Barcelona. Proceedings of XIII International Conference Information Visualization (IV09). Los Alamitos, CA: IEEE Computer Society, 2009. v.1. p.149 – 155

- ELER, Danilo Medeiros; PAULOVICH, Fernando Vieira; OLIVEIRA, M. C. F.; MINGHIM, Rosane. Coordinated and Multiple Views for Visualizing Text Collections In: XII Internati-

onal Conference Information Visualization (IV08), 2008, Londres. Proceedings of XII International Conference Information Visualization (IV08). Los Alamitos, CA: IEEE Computer Society, 2008. v.1. p.246 – 251

- ELER, Danilo Medeiros; NAKAZAKI, M. Y.; PAULOVICH, Fernando Vieira; SANTOS, D. P.; OLIVEIRA, M. C. F.; BATISTA NETO, J. E. S.; MINGHIM, Rosane. Multidimensional Visualization to Support Analysis of Image Collections In: XXI Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2008), 2008, Campo Grande. XXI Brazilian Symposium on Computer Graphics and Image Processing. Los Alamitos, CA: IEEE Computer Society, 2008. p. 289 – 296

Artigos publicados que não estão relacionados com o tema desta tese:

- BACKES, Andre Ricardo; ELER, Danilo Medeiros; MINGHIM, Rosane; BRUNO, Odemir Martinez. Characterizing 3D Shapes Using Fractal Dimension. XV Iberoamerican Congress on Pattern Recognition (CIARP 2010), São Paulo, Brazil, Lecture Notes in Computer Science, 2010. v. 6419. p. 14–21;

- SILVA, Sérgio Francisco; BRANDOLI, Bruno; ELER, Danilo Medeiros; BATISTA NETO, João do E. S.; TRAINA, Agma Juci. Silhouette-based feature selection for classification of medical images. XXIII IEEE International Symposium on Computer-Based Medical System (CBMS 2010), Perth, Australia, 2010. v. 1. p. 315–320.

Artigos em processo de submissão que estão relacionados com o tema desta tese:

- ELER, Danilo Medeiros; PAULOVICH, Fernando Vieira; OLIVEIRA, M. C. F.; MINGHIM, Rosane. A framework to coordinate multiple data similarity maps – (em definição);

- POCO, Jorge; ELER, Danilo Medeiros; PAULOVICH, Fernando Vieira; MINGHIM, Rosane. Real-time visual exploration of large DTI fiber tracking data – (IEEE Computer Graphics and Applications).