# Agrupamento de Dados Visual Interactivo


Vitor Hugo Moreira dos Santos Barros

Departamento de Engenharia Informática

Instituto Superior de Engenharia do Porto

Tese submetida para a obtenção do grau

*Mestre*

Outubro de 2011

Júri:

Maria de Fátima Coutinho Rodrigues, Professora Coordenadora, ISEP (Presidente)

Paulo Alexandre Ribeiro Cortez, Professor Auxiliar, Universidade do Minho (Vogal)

Fernando Jorge Ferreira Duarte, Professor Adjunto, ISEP (Vogal e Orientador)

Dedico esta dissertação aos meus avós, por serem o exemplo de integridade, trabalho, dedicação, perseverança e amor que sigo na minha vida.

Estarão sempre comigo.

# Agradecimentos

Gostaria de agradecer a todas as pessoas que de alguma forma contribuíram para a realização deste trabalho. Agradeço especialmente:

Ao meu orientador, Doutor Fernando Jorge Ferreira Duarte, por toda a ajuda prestada, pela disponibilidade que teve sempre e sobretudo pelo enorme trabalho de revisão desta dissertação;

Ao investigador, professor e amigo, Mestre João Manuel Maia Duarte pela confiança em indicar-me para a realização deste trabalho, pelo conhecimento transmitido e pelo apoio constante;

Ao GECAD (Grupo de Investigação em Engenharia do Conhecimento e Apoio à Decisão) do Instituto Superior de Engenharia do Porto, pelos meios disponibilizados para a realização deste trabalho;

À Samantha e a toda a minha família por todo o seu apoio e compreensão ao longo do período no qual decorreu este trabalho;

A todos os meus amigos, em especial ao grupo XT, que nunca se esqueceram de mim, apesar de muitas vezes ter negado a minha presença em eventos importantes.

A todos, muito obrigado!

# Resumo

Com a crescente geração, armazenamento e disseminação da informação nos últimos anos, o anterior problema de falta de informação transformou-se num problema de extracção do conhecimento útil a partir da informação disponível.

As representações visuais da informação abstracta têm sido utilizadas para auxiliar a interpretação os dados e para revelar padrões de outra forma escondidos. A visualização de informação procura aumentar a cognição humana aproveitando as capacidades visuais humanas, de forma a tornar perceptível a informação abstracta, fornecendo os meios necessários para que um humano possa absorver quantidades crescentes de informação, com as suas capacidades de percepção.

O objectivo das técnicas de agrupamento de dados consiste na divisão de um conjunto de dados em vários grupos, em que dados semelhantes são colocados no mesmo grupo e dados dissemelhantes em grupos diferentes. Mais especificamente, o agrupamento de dados com restrições tem o intuito de incorporar conhecimento *a priori* no processo de agrupamento de dados, com o objectivo de aumentar a qualidade do agrupamento de dados e, simultaneamente, encontrar soluções apropriadas a tarefas e interesses específicos.

Nesta dissertação é estudado a abordagem de Agrupamento de Dados Visual Interactivo que permite ao utilizador, através da interacção com uma representação visual da informação, incorporar o seu conhecimento prévio acerca do domínio de dados, de forma a influenciar o agrupamento resultante para satisfazer os seus objectivos. Esta abordagem combina e estende técnicas de visualização interactiva de informação, desenho de grafos de forças direccionadas e agrupamento de dados com restrições. Com o propósito de avaliar o desempenho de diferentes estratégias de interacção com o utilizador, são efectuados estudos comparativos utilizando conjuntos de dados sintéticos e reais.

**Palavras chave:** *Agrupamento de Dados Visual Interactivo, Visualização de Informação, Agrupamento de Dados, Agrupamento de Dados com Restrições*.

# Abstract

With the rising generation, storage and dissemination of information in recent years, the previous problem of lack of information has become a problem of extracting useful knowledge from the information available.

The visual representations of abstract information have been used to assist in interpreting the data and reveal otherwise hidden patterns. Information visualization seeks to enhance human cognition by leveraging human visual capabilities to make sense of abstract information, providing means by which humans with constant perceptual abilities can absorb increasing amounts of information.

Data clustering techniques purpose is to partition a data set into several clusters, in which similar data is placed in the same cluster and dissimilar data in different clusters. More specifically, constrained clustering methods are intended to incorporate *a priori* knowledge in the clustering process, in order to improve data clustering quality and, simultaneously, find appropriate solutions to specific tasks or interests .

This thesis studied the interactive visual clustering approach that allows the user, through interaction with a visual representation of information, to incorporate prior knowledge about the data domain in order to influence the resulting grouping to meet its objectives. This approach combines and extends interactive information visualization, force directed graph layout and constrained clustering techniques. With the purpose of evaluating the performance of different user interaction strategies, comparative studies using sets of synthetic and real data are performed.

**Keywords:** *Interactive Visual Clustering, Information Visualization, Data Clustering, Constrained Clustering*.

# Conteúdo

# Lista de Figuras

# Lista de Tabelas

# Notação

**Conjunto de Números**

$\mathbb{N}$ – Conjunto dos números naturais, $\mathbb{N}=\{1,2,...\}$

$\mathbb{R}$ – Conjunto dos números reais

$x \in [a, b]$ – Intervalo $a \leq x \leq b$

$x \in ]a, b]$ – Intervalo $a < x \leq b$

$x \in ]a, b[$ – Intervalo $a < x < b$

**Dados**

$X$ – Conjunto de dados

$d$ – Dimensionalidade de $X$

$n$ – Número de objectos de dados de $X$

$K$ – Número de grupos do conjunto de dados

$x_i$ – Objecto de dados $x_i \in X$

$l_i$ – Rótulo atribuído a $x_i$

$P$ – Agrupamento/partição do conjunto de dados

$C_k$ – $k$-ésimo grupo de um agrupamento de dados $P$

$\{\overline{x_1}, \ldots, \overline{x_K}\}$ – Centros dos $K$ grupos que formam o agrupamento de dados $P$

$Rest_{=}$ – Conjunto de restrições de ligação obrigatória

$Rest_{\neq}$ – Conjunto de restrições de ligação proibida

$w_{=_{ij}}$ – Ponderação da restrição da ligação obrigatória entre $x_i$ e $x_j$

$w_{\neq_{ij}}$ – Ponderação da restrição da ligação proibida entre $x_i$ e $x_j$

$W_{=}$ – Conjunto de ponderações das restrições de ligação obrigatória

$W_{\neq}$ – Conjunto de ponderações das restrições de ligação proibida

$\epsilon$ – Máximo de pixeis para geração de restrições de ligação obrigatória

$\delta$ – Mínimo de pixeis para geração de restrições de ligação proibida

**Vectores, Matrizes e Normas**

$A^T$ – Matriz transposta da matriz $A$

$det(A)$ – Determinante da matriz $A$

**Funções**

$d(x_i, x_j)$ – Distância entre os objectos $x_i$ e $x_j$

$g_i$ – índice do grupo mais próximo de $x_i$

$h_i$ – índice do grupo mais próximo do centro de grupo $\overline{x_{l_i}}$, a que pertence $x_i$

$I(\ldots)$ – Função que devolve 1 caso a expressão seja verdadeira, devolvendo 0 no caso contrário

# 1 Introdução

## *1.1 Enquadramento*

A revolução da informação está a mudar a forma como muitas pessoas pensam e vivem. Vastas quantidades e diversos tipos de informação são gerados, guardados e disseminados, levantando sérias questões acerca de como tornar essa informação útil. É cada vez mais evidente a necessidade de extrair e compreender o conhecimento a partir da informação armazenada. Podem-se referir a título de exemplo:

- Os estudantes têm a possibilidade de aceder a conteúdos digitais educacionais infindáveis.
- Os compradores *online* têm constantemente de decidir entre dezenas de produtos, marcas, modelos e preços distintos.
- Novas disciplinas como a bioinformática estão a liderar a revolução na utilização de informação intensiva aplicada à ciência, utilizando tecnologias de alta capacidade de recolha de dados, como os *microarrays* e os repositórios de dados *online*.

A *Visualização de Informação* evoluiu como uma abordagem que pretende tornar inteligíveis elevados volumes de informação complexa. Uma visualização de informação constitui uma interface visual da informação, com o objectivo de permitir ao utilizador *percepcionar* a informação [21]. Para isto, são geradas representações visuais interactivas da informação que exploram as capacidades de percepção do sistema visual humano e as capacidades do cérebro humano de resolução interactiva de problemas.

Ainda que representada de forma a facilitar a percepção pelos sensores humanos, existe informação de difícil interpretação, normalmente a associada a dados dispersos e de elevada dimensionalidade. As técnicas de *Agrupamento de Dados* são bastante úteis para descobrir distribuições com significado ou classes em dados. Desta forma, o agrupamento de dados contribui para a consolidação da informação ao agrupar de forma automática instâncias do domínio de dados

semelhantes entre si e ao separar objectos distintos entre si.

Diferentes tipos de utilizadores pretendem obter percepções distintas, muitas vezes no mesmo domínio de dados. Por exemplo, numa base de dados de artigos de um retalhista, um director comercial estaria interessado em perceber a evolução da margem de lucro do artigo, enquanto que um gestor de stocks procuraria situações de rotura ou excesso de stocks.

O *Agrupamento de Dados com Restrições* procura endereçar esta problemática permitindo que seja fornecida informação *a priori* acerca da estrutura dos dados de forma a influenciar a descoberta de grupos nos dados. Geralmente, o uso de restrições permite que a estrutura de dados encontrada vá de encontro a interesses específicos, contrariamente às soluções encontradas pelos algoritmos de agrupamento de dados tradicionais, em que são optimizados critérios gerais em todos os problemas.

O *Agrupamento de Dados Visual Interactivo* combina técnicas de visualização interactiva com o agrupamento de dados, mais especificamente com o agrupamento de dados com restrições, com o intuito de permitir ao utilizador explorar conjuntos de dados relacionais de uma forma interactiva e com o propósito de produzir um agrupamento que satisfaça os seus objectivos. Esta abordagem é desenvolvida e explorada nesta dissertação, efectuando-se estudos comparativos com outras abordagens de interacção com o utilizador.

## *1.2 Objectivos e Principais Contribuições*

Esta dissertação tem como tema principal a abordagem de Agrupamento de Dados Visual Interactivo. Neste contexto, foram definidos para este trabalho os seguintes principais objectivos:

**Revisão do estado da arte em Visualização de Informação.** Estudo dos diferentes tipos de estruturas de visualização de informação, assim como dos principais algoritmos de cada uma delas. São apresentados os conceitos fundamentais da visualização como o pipeline da visualização e as

diferentes estratégias de representação e interacção com o utilizador.

**Revisão do estado da arte em Agrupamento de Dados.** Estudo das fases e tipos de agrupamentos de dados, incluindo alguns dos principais algoritmos. São apresentados os conceitos fundamentais da aprendizagem supervisionada, não supervisionada e semi-supervisionada para contextualizar o agrupamento de dados com restrições. Os diferentes tipos e estratégias de aquisição de restrições são também abordados.

**Descrição da abordagem Agrupamento de Dados Visual Interactivo.** É descrita a abordagem de Agrupamento de Dados Visual Interactivo como uma combinação de técnicas de agrupamento de dados com restrições com abordagens de visualização de informação e desenho de grafos. O funcionamento e as diferentes fases da abordagem são apresentados.

**Avaliação de várias abordagens de interacção com o utilizador.** É efectuado um estudo comparativo entre o Agrupamento de Dados Visual Interactivo e outras abordagens de interacção com o utilizador. São utilizados diferentes conjuntos de dados (sintéticos e reais), com o objectivo de validar o desempenho das diferentes abordagens, evidenciando as vantagens da utilização do Agrupamento de Dados Visual Interactivo.

**Desenvolvimento de plataforma para a aplicação de abordagens de interacção com o utilizador.** De forma a realizar o estudo comparativo entre as diferentes abordagens de interacção com o utilizador é desenvolvida uma plataforma que implementa cada um dos métodos e heurísticas aplicados. Esta plataforma possibilita também a extracção de estatísticas representativas da eficácia de cada abordagem, sendo aplicada aos diferentes conjuntos de dados referidos.

Além da revisão do estado da arte nas diferentes áreas associadas ao Agrupamento de Dados Visual Interactivo, as principais contribuições deste trabalho são:

**Estudo de desempenho entre diferentes abordagens de interacção com o utilizador.** São apresentados os resultados e considerações de estudos comparativos entre diferentes abordagens de interacção com o utilizador. Estas avaliações medem a eficácia da aplicação da combinação de técnicas de visualização interactiva de informação, com técnicas que permitam a incorporação de conhecimento *a priori* no processo de agrupamento de dados, na obtenção de um agrupamento final que vá de encontro a interesses particulares. São também referidas limitações e novas direcções de investigação.

**Plataforma de análise de desempenho entre diferentes abordagens de interacção com o utilizador.** É disponibilizada a plataforma Preka que aplica implementa a combinação de técnicas de visualização, com métodos de agrupamento de dados com restrições e com a interacção com o utilizador, de forma a representar a abordagem de Agrupamento de Dados Visual Interactivo, assim como das restantes abordagens comparadas.

## *1.3 Guia de Leitura*

Nesta secção é apresentado o guia de leitura do presente documento. Esta dissertação é composta por 6 capítulos:

- Neste primeiro capítulo, *Introdução*, é efectuado o enquadramento do tema central desta dissertação, apresentado a abordagem do Agrupamento de Dados Visual Interactivo, sendo também referidos os principais objectivos deste trabalho, bem como as principais contribuições.
- No segundo capítulo, *Visualização de Informação*, exploraram-se os conceitos fundamentais da Visualização de Informação. São apresentados: a Pipeline de Visualização, sendo descritas as principais fases do processo de visualização; as Estruturas de Informação, contendo os diferentes tipos de estruturas a mapear numa visualização; as Estratégias de Representação Visual, focando nas técnicas de redução de objectos de dados; as Estratégias de Navegação, com as diferentes abordagens de navegação em visualizações; e as Estratégias de Interacção, representando as formas de interacção com o utilizador.

- No terceiro capítulo, *Agrupamento de Dados*, é apresentada uma síntese da evolução da aprendizagem automática, com foco na aprendizagem não supervisionada e semi-supervisionada, mais concretamente no agrupamento de dados com restrições. Além das fases e dos tipos de agrupamento de dados, são também descritos os vários tipos de restrições que podem ser incorporadas no agrupamento de dados, assim como alguns algoritmos de agrupamento de dados com restrições, representativos das diferentes ideias e intuições em que se baseiam.
- No quarto capítulo, *Agrupamento de Dados Visual Interactivo*, é introduzido o tema do Agrupamento de Dados Visual Interactivo como a combinação das técnicas de visualização de informação e de agrupamento de dados com restrições. São descritos os diferentes passos aplicados pela técnica, assim como uma exemplificação do seu funcionamento.
- No quinto capítulo, *Avaliação de Abordagens de Interacção com o Utilizador* , é realizado um estudo comparativo entre diferentes abordagens de interacção com o utilizador. São descritos os vários conjuntos de dados reais e sintéticos utilizados, sendo apurados e discutidos os resultados finais.
- As conclusões deste trabalho são apresentadas no sexto e último capítulo, *Conclusões*. Neste capítulo são descritas as principais limitações deste trabalho, sendo apresentadas direcções para trabalho futuro relacionado com o Agrupamento de Dados Visual Interactivo.

# 1 INTRODUÇÃO

# 2 Visualização de Informação

## *2.1 Introdução*

A área de pesquisa denominada como visualização da informação (também conhecida como *InfoVis*, proveniente do inglês *Information Visualization*) tem evoluído como uma abordagem que pretende tornar inteligíveis grandes quantidades de informação abstracta através da produção de representações visuais interactivas da mesma [1]. Pretende-se tornar visualmente perceptível a estrutura interna dos dados, assim como relações existentes entre eles, possibilitando a extracção do conhecimento útil associado à informação. A visualização de informação tem várias aplicações, tais como, a organização de conteúdos académicos para estudantes e investigadores, a comparação de gamas de produtos por retalhistas, o estudo de padrões em bioinformática, o apoio a serviços de inteligência governamental, entre outros. Neste capítulo descrevem-se as fases principais do processo de desenho da visualização de informação, assim como as diferentes técnicas a aplicar em cada fase.

## *2.2 Percepção*

A visão humana contém milhões de foto-receptores e é capaz de efectuar um rápido processamento paralelo de imagens, bem como a descoberta de padrões [2]. A intensa aptidão da visão como forma de comunicação torna possível uma transferência eficiente da informação em formato digital para a mente humana. Além deste factor, e mais importante ainda, é a capacidade humana de racionalizar sobre as suas visualizações, extraindo o conhecimento de alto nível, isto é, percepcionando as visualizações, ao invés de apenas transferir a informação para a sua mente [3]. Isto permite a inferência de modelos mentais de fenómenos reais representados pela informação por parte dos utilizadores de sistemas de visualização.

A título de exemplo, a tabela 1 representa um extracto de uma base de dados de uma rede social. A

informação encontra-se em formato tabular e, a partir da sua visualização directa, apenas é possível perceber que se trata de um grupo de pessoas que se encontram relacionadas:

| Pessoas | |
|---|---|
| **Nome** | **Género** |
| Ben | Masculino |
| Alan | Masculino |
| Otomi | Feminino |
| ... | ... |
| **Ligações** | |
| **Nome 1** | **Nome 2** |
| Ben | Otomi |
| Alan | Otomi |
| Alan | Chris |
| ... | ... |

*Tabela 1: Extracto de base de dados de uma rede social*

Já a partir da representação visual destes dados (figura 1) é facilmente perceptível que o Ben é quem tem mais amigos, sendo que estes, na sua maioria, apenas são amigos dele. É também possível verificar que a Otomi é a única amiga que o Alan e o Ben têm em comum, ou que o Chris possui apenas duas amigas do sexo feminino.

Este tipo de informação não se encontra explicitamente descrita na base de dados, sendo antes inferido através do reconhecimento visual de padrões. Estas observações não são tão facilmente identificáveis a partir da sua representação textual. Assim, torna-se clara a importância do desenho da representação visual dos dados. Uma visualização representada de forma incorrecta pode impedir a percepção de informação, ou até sugerir percepções incorrectas da mesma.

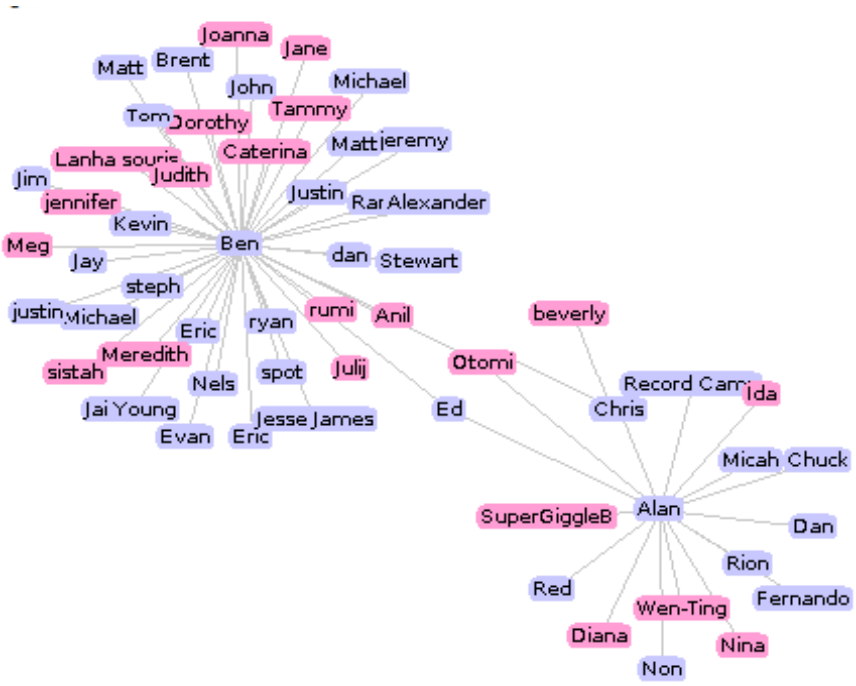



*Figura 1: Rede social – Representação visual*

A visualização permite diferentes percepções de acordo com a informação visualizada. É possível a identificação de vários tipos de percepções visuais, divididas em dois grandes grupos: **Percepções Simples** e **Percepções Complexas**.

**Percepções Simples**

- Resumos: mínimos, máximos, médias, percentagens
- Pesquisa: identificação de rótulos conhecidos

**Percepções Complexas**

- Padrões: distribuições, tendências, frequências, estruturas
- Excepções: dados destacados
- Relações: correlações, interacções múltiplas
- Desequilíbrios: balanceamento, mínimos/máximos combinados
- Comparações: selecções (1:1), contexto (1:M), conjuntos (M:N)

- Agrupamento de Dados: grupos, semelhanças
- Caminhos: distância, conexões múltiplas, decomposições
- Anomalias: erros de dados

O primeiro grupo corresponde a percepções simplistas e podem ser prontamente suportadas por interfaces textuais ou de procura, tais como folhas de cálculo e formulários de pesquisa. Isto porque estas interfaces são precisas e apresentam a informação através da representação de uma única entidade. Contudo, as restantes percepções são mais complexas sendo bem suportadas pela visualização. Estas envolvem questões abertas com respostas complexas que requerem que seja observado o todo. Um dos pontos fortes da visualização é a capacidade para descobrir novos e inesperados factores que permitem a aquisição de conhecimento potencialmente imprevisto inicialmente. O leitor poderá obter mais informação acerca deste tema em [4], [5], [6] e [7].

## *2.3 Desenho*

Tal como qualquer *user interface*, sistemas de visualização de informação eficazes são difíceis de desenhar. Fundamentalmente, este tipo de sistemas deve transformar a informação abstracta em informação perceptível. A informação abstracta não possui uma forma perceptual inerente, como no caso das bases de dados ou directórios computacionais. Dessa forma, não existem restrições naturais aos tipos de representações visuais que a criatividade pode produzir para a informação abstracta, sendo as possibilidades ilimitadas. Como resultado, têm surgido múltiplos esforços tanto na criação de novas representações visuais, como na identificação das representações mais eficazes de entre as infindáveis hipóteses. Em contraste, a visualização científica [8] enfatiza tipicamente a visualização de dados que representam fenómenos físicos tridimensionais (arquitectónicos, meteorológicos, médicos, biológicos, entre outros), o que limita a sua acção a restrições naturais e foca os desafios desta área no realismo.

Duas das mais desafiantes características que tornam difícil o desenho de sistemas de visualização de informação são:

**Complexidade.** O suporte de informação abstracta diversa que pode ter múltiplos tipos de dados e estruturas relacionados entre si.

**Escalabilidade.** O suporte de elevadas quantidades de informação.

Dadas estas características, as representações visuais não são suficientes sendo necessário o desenvolvimento de técnicas de interacção. Enquanto os princípios de grafos e ilustrações estáticas são fundamentais no desenho da visualização (ler em [9], [10], [11] e [12]), tornam-se também relevantes novos problemas inerentes à interacção homem-máquina relacionados com estas características.

O processo de desenho da visualização engloba requisitos iterativos de análise, desenho e validação [13]. Na fase de análise de requisitos, é importante identificar os dois *inputs* primários necessários para a fase de desenho: as características da informação a ser visualizada e os tipos de percepções que a visualização deve possibilitar. As características da informação incluem o esquema de dados, as suas estruturas e o seu volume. Uma vez que o número de atributos e das percepções pretendidas pode ser elevado, será importante efectuar a priorização destes, de forma a que seja possível o balanceamento de desequilíbrios no desenho. Incluem-se também como elementos da análise de requisitos as acções e o conhecimento base dos utilizadores, a semântica dos dados e os requisitos de hardware.

Na fase de desenho são definidas as principais directivas da visualização incluindo o mapeamento visual da informação, a representação de estruturas de informação, estratégias de contextualização visual, estratégias de navegação e as técnicas de interacção, descritos em detalhe ao longo deste capítulo.

A fase de validação deve ser continuamente considerada durante o processo de desenho [14]. Uma análise de falhas identifica os impactos positivos e negativos das características do desenho da visualização na sua capacidade de transparecer conhecimento, procurando ultrapassar ou equilibrar representações menos correctas através do desenho iterativo [13]. Deve iniciar-se com validações analíticas que determinem se o desenho responde a requisitos como a escalabilidade ao volume de

dados e a assertividade na produção das percepções pretendidas. Em iterações posteriores, devem ser efectuadas validações empíricas envolvendo utilizadores reais como a técnica do feiticeiro de Oz, testes de usabilidade ou outras experiências controladas (ler em [15] e [16]). As percepções identificadas como desejáveis na análise de requisitos devem ser medidas através da capacidade de realização de tarefas por parte dos utilizadores nas validações empíricas. Alternativamente, uma vez que tais medidas tendem a restringir o teste à capacidade de retenção de percepções simples, o que penaliza a descoberta global de conhecimento da visualização, a metodologia baseada em percepções proposta por *Saraiya et al.* [17] pretende medir as percepções geradas pela visualização utilizando um protocolo experimental que não recorre a listas de tarefas dos utilizadores.

## *2.4 Pipeline de Visualização*

A *pipeline* de visualização consiste no processo computacional de conversão da informação numa forma visual passível de interacção com os utilizadores [3] (figura 2).

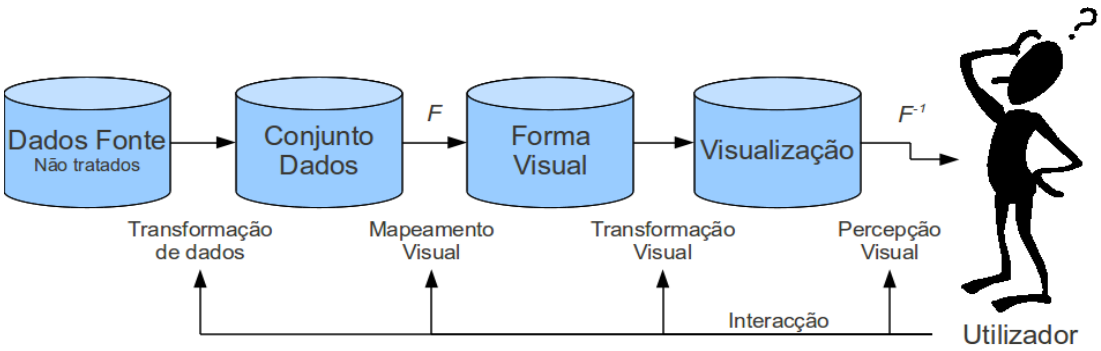



*Figura 2: Pipeline de Visualização*

A primeira fase reside na transformação da informação disponibilizada num formato bem organizado e adaptado às fases seguintes. O formato resultante consiste tipicamente num conjunto de dados contendo várias entidades com atributos idênticos, cada qual com os seus próprios valores. Podem ser efectuados tantos passos de processamento de dados quantos se considerar necessário. Informação derivada, através de resultados de *data mining* como o agrupamento de dados, podem ser muito úteis no apoio à geração de percepções e conhecimento [18]. O segundo passo, o coração do processo de visualização, consiste no mapeamento do conjunto de dados resultante na sua *forma*

*visual*. Esta forma visual engloba símbolos visuais que representam as entidades do conjunto de dados. A terceira etapa insere esta forma visual numa *visualização* que tem por objectivo mostrar ao utilizador a informação no formato visual, tornando possíveis, ao mesmo tempo, variadas transformações a essa visualização, como por exemplo a navegação. Após este processamento, a visualização é então apresentada e assimilada pelo utilizador. Através da visão humana, os utilizadores interpretam a visualização reconstruindo (parcialmente) a informação inerente. Finalmente, os utilizadores podem interagir com qualquer um dos passos da *pipeline* de forma a alterar a visualização e fazer novas interpretações. Este conjunto de passos constitui um sistema de visualização de informação.

### 2.4.1 Mapeamento Visual

O mapeamento visual no segundo passo do *pipeline* é, como já foi referido, o coração da visualização e deve ser alvo de um desenho cuidado. O meio de comunicação é a representação visual da informação. O conjunto de dados é mapeado computacionalmente numa forma visual usando uma determinada função F, que recebe como entrada o conjunto de dados gerando a sua representação visual como saída. Depois, quando a representação visual é comunicada aos utilizadores, estes devem reverter cognitivamente o mapeamento visual através da inversão dessa função F, de forma a descodificar a informação existente na representação visual. Ainda não é totalmente claro como, quando e até que ponto a função $F^{-1}$ é cognitivamente aplicada no processo perceptual, uma vez que existem vários modelos explicativos [2]. Enquanto alguma racionalização cognitiva opera na própria representação visual, o significado efectivo tem de ser descodificado. De qualquer forma, este processo de comunicação visual implica quatro características importantes da função de mapeamento visual F:

**Computável.** A função F deve ser uma função matemática que possa ser computada por um algoritmo. Apesar de existir espaço para a criatividade no desenho destas funções, a execução das mesmas deve ser algorítmica.

**Invertível.** Deve ser possível o uso de $F^{-1}$, o inverso da função de mapeamento F, de forma a ser possível a reconstrução da informação a partir da sua representação visual para um desejado grau de

certeza. Se isto não se verificar, a visualização será ambígua, não interpretável e poderá conduzir a interpretações erradas.

**Comunicável.** A função F (ou preferencialmente a função $F^{-1}$) deve ser conhecida pelo utilizador de forma a descodificar a representação visual. Este aspecto deve ser comunicado pela própria visualização, ou já conhecido pelo utilizador devido a experiências anteriores. Em termos de usabilidade, esta é uma questão de aprendizagem.

**Conhecível.** $F^{-1}$ deve minimizar a carga de conhecimento necessária para descodificar a representação visual. Esta é uma questão relacionada com a percepção humana e com o desempenho.

O passo de mapeamento visual é alcançado através de dois sub-passos [3] (figura 3).

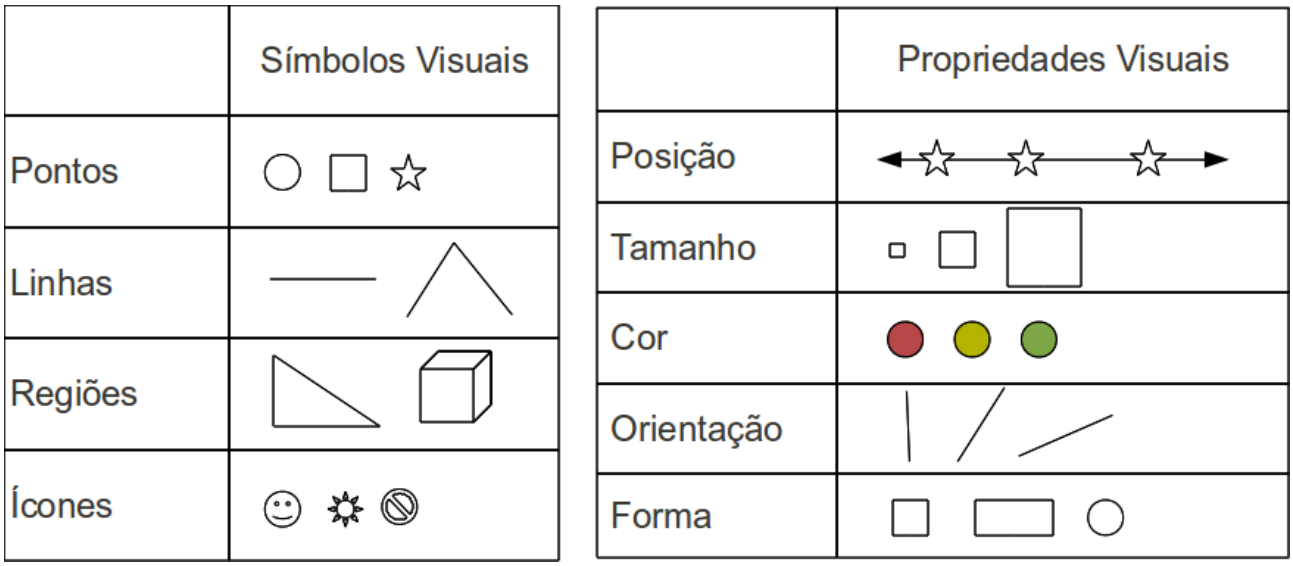



*Figura 3: Símbolos e propriedades visuais (adaptado de [3])*

Inicialmente, cada entidade é mapeada num determinado símbolo visual. O vocabulário de símbolos possíveis inclui pontos (ou formas simples), linhas (segmentos, curvas, caminhos), regiões (polígonos, áreas, volumes) e ícones (símbolos, imagens). Seguidamente, os atributos de cada entidade são mapeados para propriedades visuais do símbolo representativo da entidade. Propriedades visuais comuns de símbolos incluem a posição espacial (*n* dimensional), o tamanho (altura, área, volume), a cor (escalas do cor, saturação, intensidade), a orientação (ângulo, declive,

vector) e forma. Existem também outras propriedades visuais como a textura, o movimento, a intermitência, a densidade e a transparência.

### 2.4.2 Propriedades Visuais

Geralmente, os atributos devem ser priorizados de acordo com os requisitos do problema e com as percepções que se pretendem obter. A priorização pode então ser aplicada de forma a mapear os atributos de maior prioridade às propriedades visuais mais eficazes. As propriedades relacionadas com o posicionamento espacial são as mais eficazes, devendo ser reservadas para a representação visual dos atributos mais importantes.

As restantes propriedades visuais, denominadas propriedades retinais [19], poderão ser usadas de seguida. A efectividade destas propriedades é determinada por muitos factores interdependentes que incluem o processamento pré-focagem [20], a independência (separabilidade) perceptual [2], o tipo de dados (quantitativo, ordinal, categórico) [3], a polaridade (maior que, menor que) [2], a tarefa [21, 6] e a atenção [22]. A ordenação da efectividade destas propriedades aceite pela comunidade científica é baseada em evidências empíricas [23, 24], como também na experiência [19, 25]. A ordem representada na figura 3 é válida para dados quantitativos.

Para dados categóricos, a cor e a forma tornam-se mais predominantes. Existem alguns sistemas de desenho de visualizações que tentam usar este tipo de regras de forma a gerar mapeamentos visuais eficazes como o Apt [25] ou o Sage [26]. Finalmente, para os restantes tipos de atributos, poderão ser aplicadas técnicas de interacção.

Geralmente, o mapeamento visual directo da informação é o mais eficaz para uma rápida percepção, ao contrário das técnicas de interacção que requerem acções físicas mais lentas por parte do utilizador para que seja revelado conhecimento. Sempre que a representação visual de atributos adicionais implique a diminuição da compreensão de atributos mais importantes, devem ser aplicadas técnicas de interacção. Este tipo de técnicas permite que os utilizadores possam efectuar a

alteração da função de mapeamento visual ou de outros níveis da *pipeline* de visualização, baseados nos atributos adicionais. Através da observação das alterações resultantes na representação visual, os utilizadores podem inferir informação adicional relativamente a esses atributos. Por exemplo, a técnica de *queries* dinâmicas que permite pesquisas interactivas de outros atributos, pode ser utilizada para filtrar dinamicamente os símbolos que representam as entidades [27].

## *2.5 Estruturas da Informação*

O processo de mapeamento visual é um ponto de partida para o desenho da visualização, contudo, à medida que a complexidade aumenta, vão sendo necessários métodos mais avançados. A identificação de estruturas subjacentes à informação ajuda o direccionamento do processo de desenho. Estas estruturas descrevem uma organização a alto nível do conjunto de dados, o que normalmente contribui para uma representação apropriada das visualizações. Uma vez que estas estruturas são muito importantes na formação dos modelos mentais dos utilizadores, são tipicamente mapeadas a atributos de posicionamento espacial formando o *layout* primário da visualização. Podem ser considerados quatro tipos comuns de estruturas de informação [3, 28], descritos nas subsecções seguintes. De notar que estas não são classificações estritas ou mutuamente exclusivas, constituindo antes linhas orientadoras úteis.

### 2.5.1 Estrutura Tabular

As tabelas são constituídas por linhas (entidades) e colunas (atributos). Os dados em tabelas são muitas vezes referidos como sendo multi-dimensionais devido ao facto de cada atributo definir uma dimensão no espaço de dados que, para cada entidade, identifica um único ponto. Bases de dados e folhas de cálculo (como a apresentada em tabela 1) são alguns exemplos. Visualizações de tabelas que contêm um pequeno número de atributos podem ser desenhadas de uma forma relativamente simples através do processo de mapeamento visual. Contudo, este tipo de visualizações carece de escalabilidade a um número elevado de atributos devido ao número limitado de propriedades visuais disponíveis que não colidem entre si. Para endereçar este problema, foram desenvolvidos

vários e criativos métodos para tabelas com muitos atributos. Estes métodos assentam primariamente no uso de símbolos e *layouts* espaciais mais complexos.

O método TableLens [29] (figura 4) preserva a representação visual tabular de uma folha de cálculo, convertendo as células para barras horizontais e associando o tamanho da barra aos seu valores. Isto explora a propriedade visual tamanho, que é muito eficaz na representação da dados quantitativos. Além disto, e uma vez que as barras são muito finas, é possível a alocação de muitos valores na visualização, fornecendo uma vista geral dos dados. Neste método, cada entidade (linha) contém tantos símbolos visuais (barras) quantos os atributos (colunas). O utilizador pode seleccionar interactivamente uma ou várias linhas, de forma a revelar a informação textual a si associada. É possível efectuar também a ordenação por qualquer atributo. Desta forma, tornam-se perceptíveis distribuições e relacionamentos do atributo ordenado com os restantes atributos. O princípio de compatibilidade por proximidade [6] refere, contudo, que representações que utilizem apenas um símbolo visual por entidade são melhores do que métodos como o TableLens para o reconhecimento de relações entre atributos. Isto apesar de tais representações, como já referido anteriormente, serem mais limitadas a nível de escalabilidade. Assim, o TableLens constitui um bom método para conjuntos de dados para os quais seja importante revelar uma vista geral dos mesmos, oferecendo uma capacidade razoável para verificação de relacionamentos entre atributos distintos.

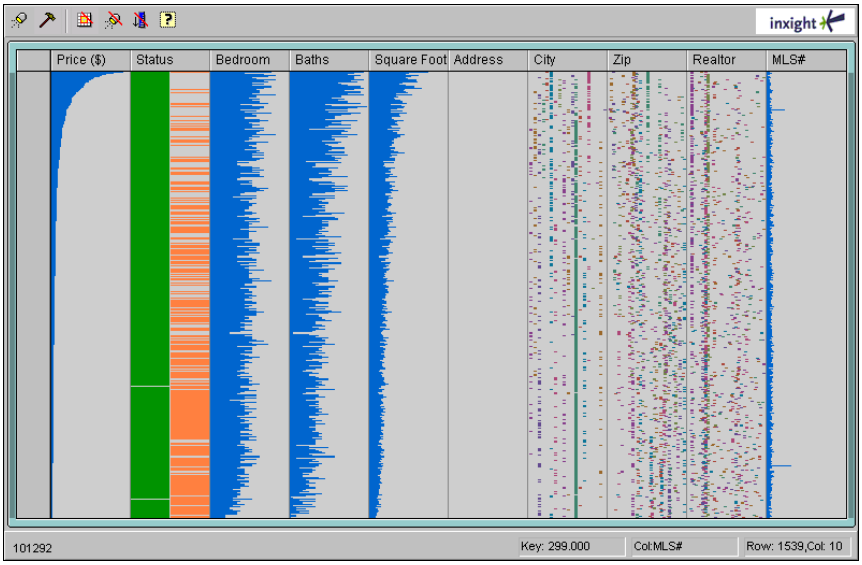



*Figura 4: Estrutura tabular – TableLens*

O sistema de coordenadas cartesiano utiliza os eixos ortogonais para mapear visualmente 2 ou 3 atributos de um conjunto de dados tabular para um plano espacial. São então estes eixos o factor limitativo da escalabilidade de atributos. Como alternativa, *Inselberg* propôs o método Parallel Coordinates [30] (figura 5). Este método mostra os eixos dos atributos através de linhas verticais paralelas. Cada entidade é mapeada para uma linha que liga os valores dos atributos em cada eixo. Desta forma, os valores dos atributos são mapeados consoante a posição vertical dos respectivos vértices na linha da entidade. Os utilizadores conseguem desta forma percepcionar grupos de dados semelhantes, bem como relações entre atributos adjacentes. Padrões de linhas cruzadas entre eixos adjacentes indicam uma relação inversa entre esses atributos, enquanto que linhas não cruzadas indicam uma relação proporcional. Contudo, o tipo de visualizações produzidas por este método são usualmente confusas. Para combater esta questão, o utilizador pode interactivamente seleccionar entidades resultando na coloração distinta da linha da entidade ao longo de todos os eixos de atributos.

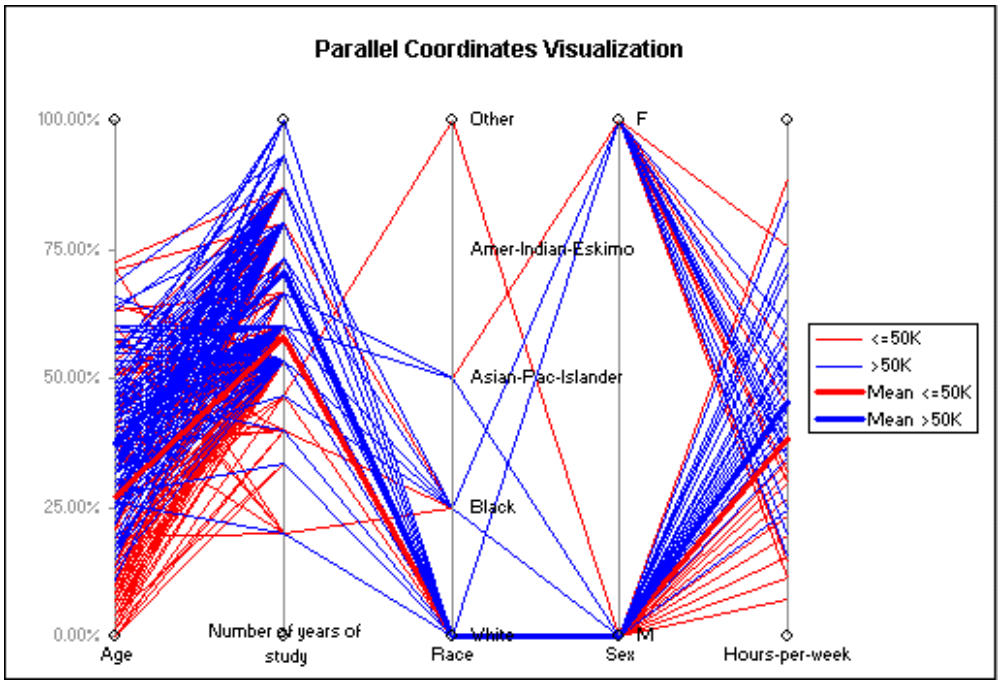



*Figura 5: Estrutura tabular – Parallel Coordinates*

Uma análise de escalabilidade ao método Parallel Coordinates resulta num valor semelhante ao método TableLens. Outras disposições de eixos incluem a disposição radial [31] ou a circunferencial [32].

### 2.5.2 Estrutura Espacial e Temporal

Esta estrutura possui uma forte componente multidimensional (uni, bi ou tridimensional) na qual normalmente são necessárias técnicas de navegação. Como exemplos unidimensionais podem ser referidos linhas temporais, música, listas, *streams* de vídeo, documentos lineares ou *slide shows*. Podem considerar-se exemplos de duas dimensões os mapas de estradas, imagens de satélite ou fotografias. Exemplos 3D incluem os exames médicos de ressonância magnética e tomografia computorizada, planos arquitectónicos CAD/CAM e ambientes virtuais. Funções contínuas, incluindo aquelas com domínios de dimensões superiores a 3, caem também nesta categoria [33]. Nestes casos, são geradas múltiplas visualizações, com um máximo de 3 dimensões cada. Estas estruturas espaciais e temporais constituem o mapeamento mais natural para ecrãs espaciais.

Por exemplo, o *Music Animation Machine* [34] (figura 6) apresenta uma representação em linha temporal de música que vai sendo percorrida à medida que a música toca. As notas são representadas por barras horizontais, com a posição vertical a representar a tonalidade, a posição horizontal a representar o tempo, o tamanho a duração e a cor a indicar outros atributos como o instrumento, o timbre ou a mão (no caso do piano). De forma semelhante, o LifeLines [35] representa eventos do historial médico de uma pessoa mas de uma forma mais compacta, recorrendo ao *zooming* para a navegação. Para linhas temporais que contenham ciclos temporais, como calendários, podem ser usadas espirais visuais para estimar os ciclos, garantindo ao mesmo tempo a manutenção de uma linha contínua [36].

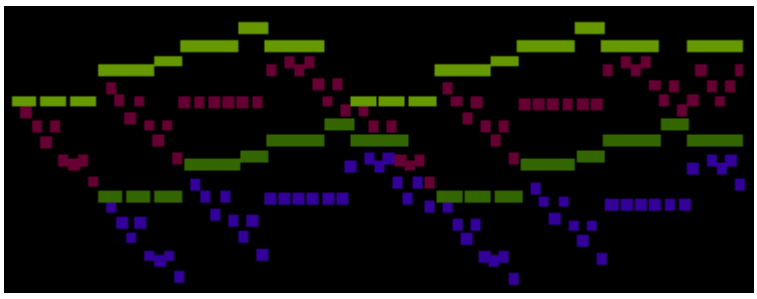


*Figura 6: Estrutura espacial e temporal - Music Animation Machine*

Em representações 3D, o maior desafio consiste em conseguir visualizar o interior da estrutura tridimensional quando existe oclusão. As aplicações relacionadas com a arquitectura (associadas a dados poligonais) usam tipicamente uma projecção em perspectiva de primeira pessoa, possibilitando uma navegação e rotatividade livres o suficiente para representar uma experiência semelhante à da vida real [37]. Para imagens médicas (associadas a dados volumétricos) são normalmente aplicadas técnicas como o corte e a transparência. Como exemplo pode ser referido o *Visible Human Explorer* [38] (figura 7) que apresenta cortes 2D que podem ser animados através dum corpo 3D. Na renderização de volumes 3D, pode ser disponibilizada aos utilizadores uma visão "raio-X" no espaço através do ajuste da opacidade dos diferentes materiais representados. Isto pode ser efectuado pelo controlo interactivo dos símbolos visuais através de uma função de transferência visual [39].

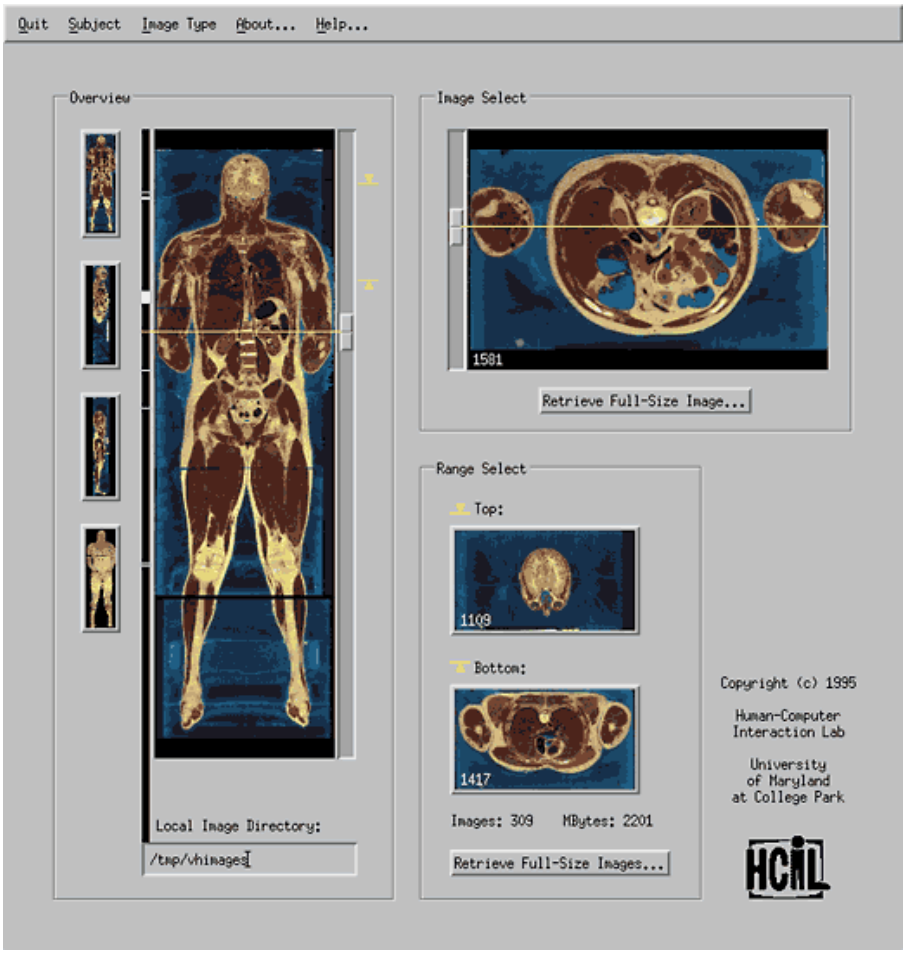



*Figura 7: Estrutura espacial e temporal - Visible Human Explorer*

Os espaços contínuos hiperdimensionais devem ser de alguma forma reduzidos a 3 ou menos dimensões para serem representados. Alguns métodos para conseguir isto incluem os eixos hierárquicos [40] e o corte [41].

### 2.5.3 Estrutura em Grafo

Neste tipo de estruturas é possível observar ligações específicas entre entidades. Em termos da teoria de grafos, uma rede consiste num conjunto de nós (entidades) ligados por um conjunto de arcos (ligações) que podem ser, ou não, direccionados. Tal como as entidades, as ligações também podem conter atributos. Alguns exemplos incluem as redes de comunicações, citações literárias ou ligações entre páginas web. Estruturas em árvore constituem um subconjunto especial de redes, sendo distintas o suficiente para garantir um tratamento separado. As árvores têm uma estrutura hierárquica que liga as entidades através de ligações pai-filho. Para ser uma árvore, cada entidade filho deve possuir apenas uma entidade pai. Exemplos deste tipo de estrutura são os directórios de ficheiros, menus, esquemas organizacionais e vários tipos de taxonomias. Existem também outras interessantes variantes da estrutura em árvore, como as multi-árvores [42] e as poli-hierarquias [43]. De forma a ser possível obter novas percepções, deve ser percebida a estrutura das ligações, como a largura ou a profundidade da árvore. O primeiro desafio para este tipo de visualizações reside na organização do *layout* espacial de forma a revelar a estrutura das ligações. O desafio secundário é o de possibilitar a visualização dos atributos das entidades e das ligações.

#### *2.5.3.1 Redes*

Na visualização de redes, a abordagem nó-ligação é dominante. Nesta abordagem as entidades são mapeadas para nós visuais e as ligações são representadas por ligações visuais entre os nós. Têm sido desenvolvidos muitos algoritmos para representação de diagramas em rede [52], sendo adaptados para tipos específicos de redes. O desenho destes algoritmos deve considerar factores como o número de nós e ligações, a direcção das ligações, os graus dos nós, padrões comuns na estrutura da rede e os atributos dos nós e das ligações que devem estar visíveis. As ligações devem

ser desenhadas como linhas rectas, arcos ou linhas ortogonais. Os algoritmos podem aplicar variadas restrições como a minimização dos cruzamentos das linhas, a minimização do tamanho da linhas ou a maximização das simetrias [52]. Geralmente, o objectivo é o de representar a rede de forma a revelar padrões escondidos, evitando ao mesmo tempo o fenómeno "prato de esparguete". Algumas técnicas aplicadas por algoritmos de representação de redes incluem desenhos circulares, concêntricos, modelados por forças físicas e agrupamentos de dados. O SeeNet [53] (figura 8) organiza os nós de acordo com a posição geográfica, evidenciando as ligações através de arcos tridimensionais. A cor e a espessura dos arcos são utilizadas para representar o tipo de comunicação e a largura de banda. O algoritmo H3 [54] acomoda os nós numa esfera 3D que pode ser rodada e explorada de uma forma semelhante às árvores hiperbólicas (Hyperbolic Tree). De forma a reduzir a complexidade da rede, podem ser usadas técnicas de navegação que permitam a representação da rede a partir do foco num único nó [55].

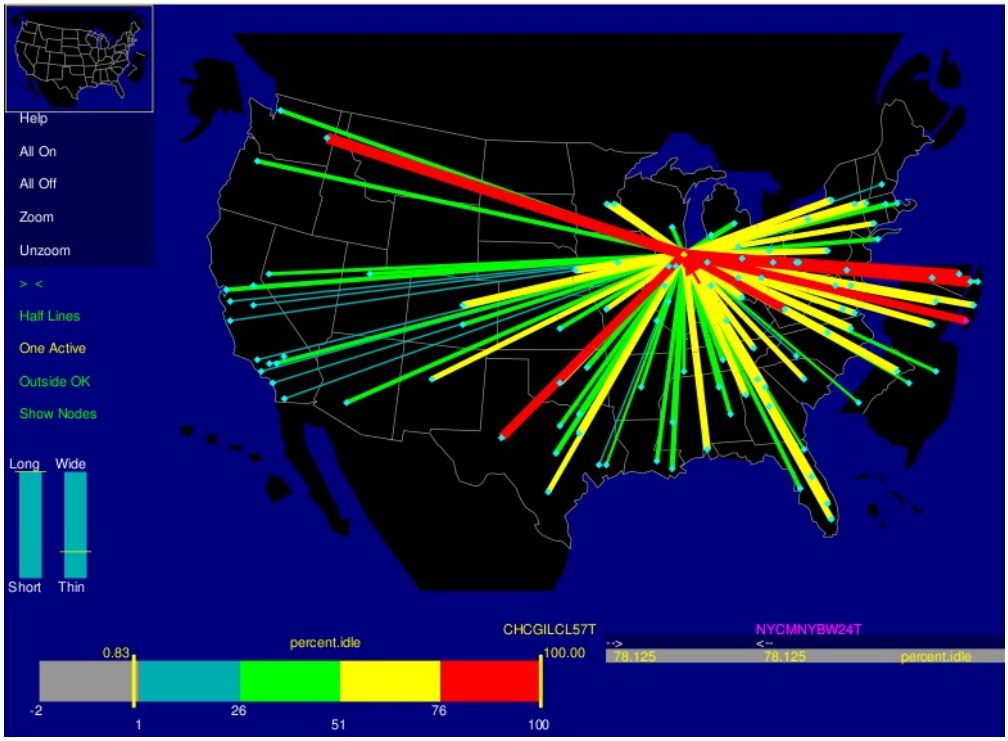
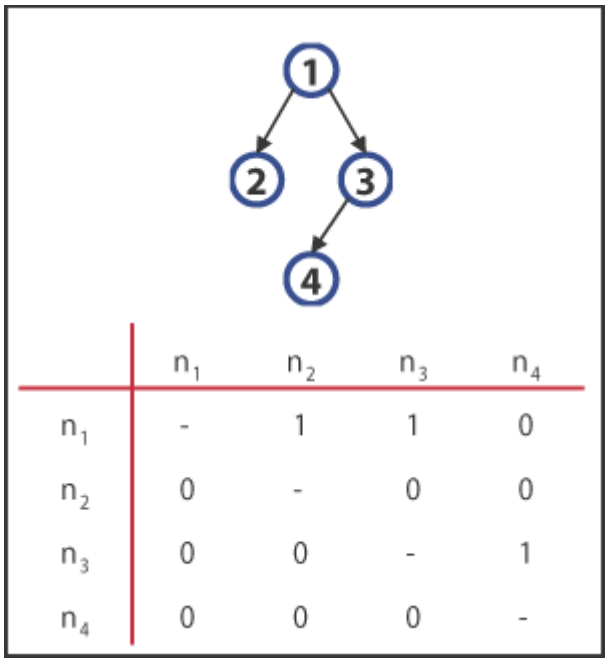



*Figura 8: Estrutura em grafo – Rede SeeNet*

Uma abordagem diferente consiste na visualização da rede como uma matriz de adjacências [56] (figura 9). Neste tipo de abordagem é dado ênfase às ligações em vez de aos nós, mapeando-se cada potencial ligação a uma célula na matriz.




| | $n_1$ | $n_2$ | $n_3$ | $n_4$ |
|---|---|---|---|---|
| $n_1$ | - | 1 | 1 | 0 |
| $n_2$ | 0 | - | 0 | 0 |
| $n_3$ | 0 | 0 | - | 1 |
| $n_4$ | 0 | 0 | 0 | - |

*Figura 9: Estrutura em grafo - Matriz de adjacências*

#### *2.5.3.2 Árvores*

Para estruturas em árvore, existem duas abordagens primárias para representar conexões pai-filho: ligação e contenção. A abordagem ligação utiliza diagramas nó-ligação. Em alternativa a este tipo de diagramas existem os diagramas indentados (como o GNOME Nautilus, o Mac finder ou o Windows Explorer), *top-down* ou *left-right* (SpaceTree [44]), radial (Hyperbolic Tree [45]) e combinações de *left-right* com radial (3D ConeTrees [46]) (figura 10). Este tipo de sistemas enfatiza a visualização de um único atributo como um rótulo textual em cada nó. Os diagramas nó-ligação tendem a consumir muito espaço devido à quantidade de espaço em branco necessário no interior de cada um destes *layouts* espaciais, tornado difícil a representação de mais de cerca de 100 nós. Dado não ser possível disponibilizar completamente estruturas em árvore muito densas, cada um destes *layouts* necessita ser capaz de efectuar navegação interactiva. A técnica Focagem+Contexto (*focus+context)* (secção 2.7.3) é a ideal para a navegação em árvores ao permitir que os utilizadores possam efectuar *drill-down* de um ramo específico, mantendo ao mesmo tempo o contexto do caminho percorrido. As árvores Hyperbolic Tree encolhem o tamanho dos nós da periferia de forma

a poderem mostrar mais nós. Abordagens 3D, como as ConeTrees, aproveitam a terceira dimensão para ganhar espaço na representação da árvore, contudo, devido à oclusão, não é claro se tal espaço extra é realmente útil. O factor mais importante em representações 3D é a navegação interactiva [47]. A rotação livre à volta de uma representação tridimensional não é eficaz nestas estruturas. As ConeTrees utilizam contudo uma técnica de interacção mais eficiente: a rotação de cones 3D de forma a mostrar os sub-nós desejados para a frente.

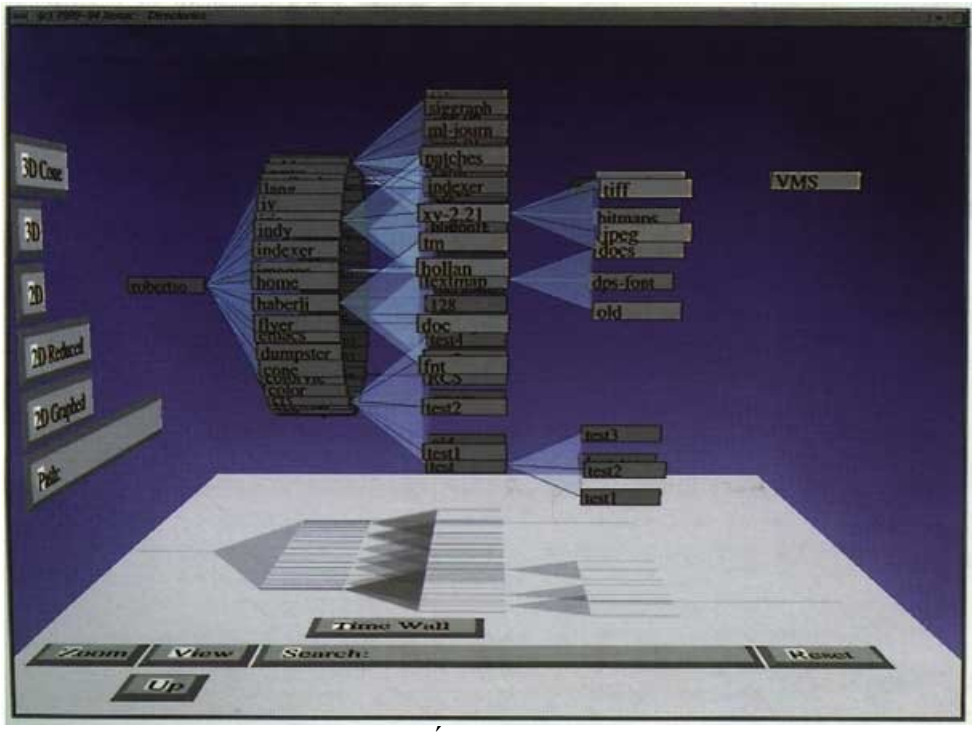



*Figura 10: Estrutura em grafo - Árvore ConeTree*

A abordagem de contenção para estruturas em árvore é exemplificada pelo Treemaps [48] (figura 11). Sub-nós, ou nós filho, representados por rectângulos, são contidos visualmente dentro dos seus nós pai como em diagramas Venn. Os Treemaps são orientados à maximização do aproveitamento do espaço disponível, sendo facilmente escalável a cerca de 10.000 entidades. Os atributos são mapeados para propriedades visuais retinais (facilmente identificadas pela retina ocular humana) dos nós rectangulares, como a cor ou o tamanho. Daí este tipo de representações enfatizar a visualização de atributos não textuais. Em Treemaps densas, não existe espaço extra para nós com rótulos textuais. Os nós podem ser organizados dentro do seu nó pai através de uma multiplicidade de algoritmos. O Treemap original utiliza um algoritmo *slice-and-dice*. Apesar de simples, este

algoritmo tende a gerar rectângulos com aspectos muito diferentes, o que torna a comparação visual difícil. Algoritmos mais recentes passaram a gerar Treemaps quadradas [49]. O algoritmo SunBurst [50] oferece uma versão radial da abordagem de contenção baseada em gráficos circulares empilhados. Em comparação com o Treemaps, o SunBurst reduz a curva de aprendizagem em utilizadores menos experientes, contudo consegue-o à custa da escalabilidade, uma vez que o número de sub-nós é limitado por um espaço circunferencial de 1 dimensão, ao contrário da área bidimensional disponível nas Treemaps.

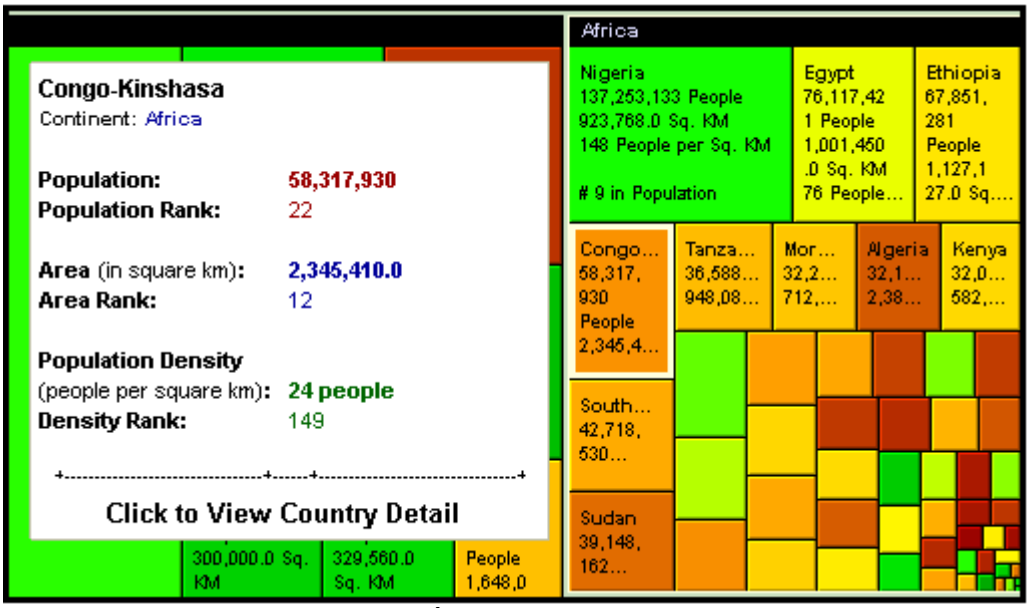



*Figura 11: Estrutura em grafo - Árvore Treemap*

### 2.5.4 Estrutura de colecções de texto e documentos

Esta estrutura consiste numa colecção arbitrária de documentos, normalmente textuais. Podem ser referidos como exemplos, as bibliotecas digitais, arquivos de notícias, repositórios de imagens digitais e código de *software*. Dos quatro tipos de estruturas de informação analisadas, este é o menos estruturado, o que torna mais difícil o desenho das visualizações. O mapeamento do texto para uma forma visual é particularmente desafiante uma vez que não é óbvio como o texto poderá ser o *input* de uma função de mapeamento visual, conforme descrito em 2.4.1. As funções de mapeamento devem aproveitar características do texto, de forma a gerar informação útil para a

computação de representações visuais. Estruturas de texto ou colecções de documentos externos como índices (estrutura em árvore), metadados (estrutura tabular) ou citações (estrutura em rede) são categorizados como outros tipos de estruturas de informação e encontram-se descritos em secções anteriores. A estrutura descrita nesta secção foca-se antes na informação implícita no texto ou nos próprios documentos como um todo. As diferentes soluções para esta problemática variam da macro-escala (*overview* de grandes colecções) à micro-escala (um único documento).

Uma grande parte das visualizações de texto focam-se na disponibilização de mapas semânticos de grandes colecções de documentos baseadas em tópicos. Geralmente, o objectivo é o de agrupar espacialmente os documentos de forma a que os semelhantes (documentos com conteúdo semelhante) fiquem perto uns dos outros, separando simultaneamente documentos distintos. Finalmente, é criado um mapa de documentos baseado na metáfora de distribuição dos livros numa biblioteca, em que os livros se encontram cuidadosamente organizados por tópicos. A similaridade entre documentos pode ser obtida de diversas formas, sendo geralmente domínio da área de recuperação de informação. Um método comum é a comparação da frequência da ocorrência de frases ou palavras de dicionário entre dois documentos. Podem ser analisadas as densidades de grupos de documentos de forma a possibilitar a extracção de tópicos que serão representados nos mapas. Os mapas auto organizáveis Kohonen [57] mapeiam documentos individuais para pequenos pontos que são agrupados pelo conteúdo textual. A selecção de um ponto no mapa revela um resumo do documento ou abre-o na sua totalidade. O ThemeView [58] (figura 12) enfatiza os tópicos dos documentos através da criação de um terreno tridimensional representativo dos diferentes temas da colecção. As montanhas representam os temas no mapeamento final, sendo que quanto maior for a montanha, maior é a relevância do tema. Montanhas adjacentes ou juntas indicam a presença de documentos que partilham ambos os temas.

A abordagem de pesquisa por palavras-chave permite um mapeamento mais focado, baseado em palavras especificadas pelo utilizador. O VIBE [59] torna visível a forma como os documentos se relacionam com os tópicos. Inicialmente, os tópicos são espalhados pela periferia da visualização de forma aleatória. Os documentos, representados por pontos, são mapeados através de pontos à volta dos tópicos de acordo com a sua relevância, utilizando um modelo baseado na força de molas. Existem métodos que invertem o mapeamento, como o TileBars [60] (figura 13) que mostra a

aproximação dos tópicos aos documentos, onde o critério de pesquisa é o tópico, ao invés de apresentar quais os tópicos associados a um determinado documento ou conjunto de documentos, onde o critério de pesquisa é o documento. Neste método os documentos são mostrados textualmente, como num motor de pesquisa, contudo possuem uma barra que é tanto maior quanto maior for a sua relação com o tópico pesquisado.

Por fim, os documentos podem também ser organizados pelos utilizadores da forma pretendida. Nesta abordagem os documentos devem ser representados por miniaturas, de forma a promover a pesquisa por conteúdo. O Web Book e Forager [61] agrupa as páginas web favoritas num livro tridimensional que os utilizadores podem "folhear" obtendo uma representação visual da página. Os livros podem ser organizados numa estante virtual. Com o DataMountain [62] (figura 14) os utilizadores conseguem organizar imagens das suas páginas web favoritas ou fotos num plano inclinado, tirando partido da memória espacial humana.

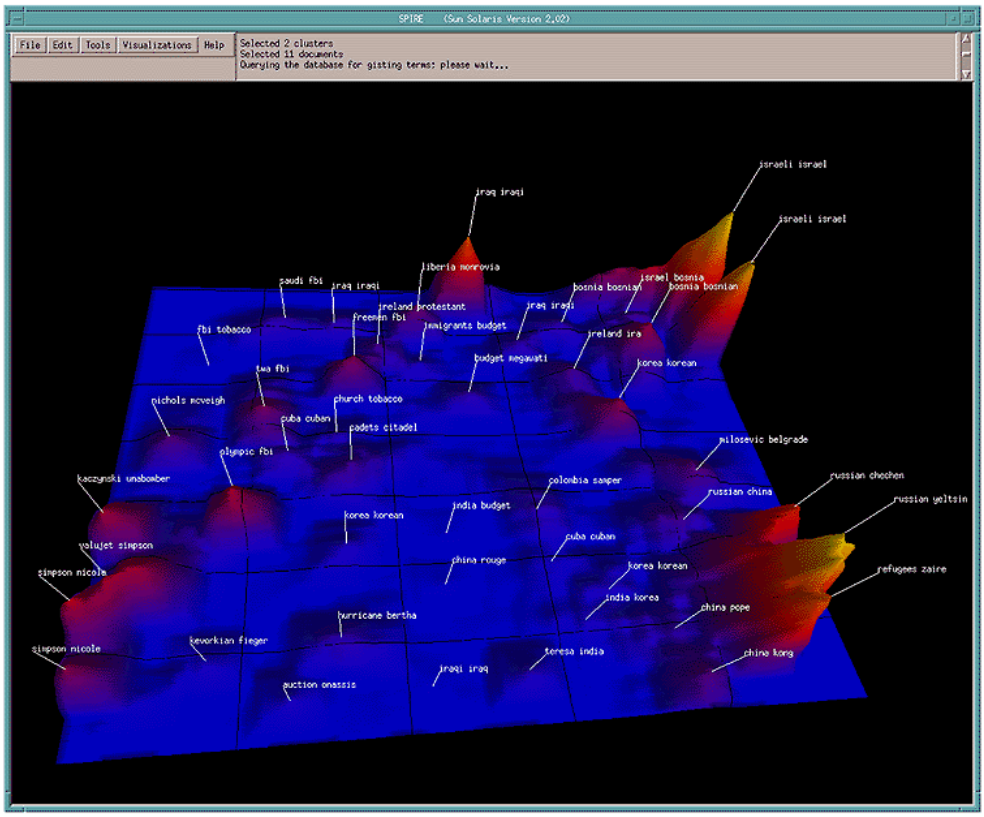



*Figura 12: Estrutura de colecções de texto e documentos – ThemeView*

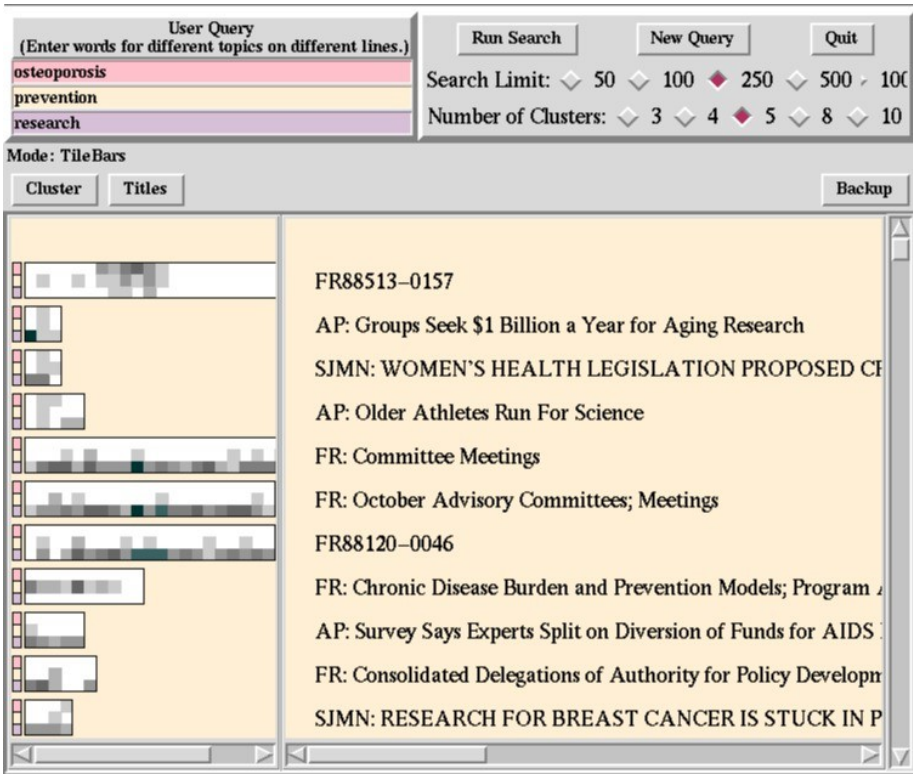



*Figura 13: Estrutura de colecções de texto e documentos - TileBars*


*Figura 14: Estrutura de colecções de texto e documentos - DataMountain*

## *2.6 Estratégias de Representação Visual*

À medida que a quantidade de informação disponível aumenta, é cada vez mais difícil apresentar toda a informação numa representação visual restringida pelo espaço disponível num ecrã. O desenho de métodos para a representação visual de grandes quantidades de informação é, desta forma, um dos problemas fundamentais na área de pesquisa da visualização. Ainda que fosse possível a inclusão de toda a informação no ecrã, a representação de todos os detalhes de uma colecção de dados num ecrã pode não ser visualmente apelativa. Uma abordagem ingénua na representação visual de dados poderia passar por consumir todo o ecrã com apenas um subconjunto dos mesmos, ainda que apresentando todos os detalhes destes. Este é o chamado *keyhole problem*, uma vez que é uma analogia ao acto de espreitar para uma fechadura para dentro de uma grande sala.

De forma a permitir a visualização de conjuntos vastos de informação, Shneiderman sugeriu o mote "obter uma visão geral primeiro, filtrar e fazer zoom e no final detalhar a pedido" [63]. Assim, a solução para o *keyhole problem* passa por apresentar inicialmente uma visão geral do conjunto de dados na sua totalidade sacrificando os detalhes da informação. De seguida devem ser fornecidos mecanismos de interacção que permitam que o utilizador possa fazer zoom na informação pretendida e que possa filtrar o que não é relevante. Por fim, deve ser facilitada a rápida obtenção de informação detalhada acerca de entidades seleccionadas pelo utilizador.

Existem muitas vantagens na apresentação de uma vista geral inicial da informação:

- Suporta a formação de modelos mentais do conjunto de dados
- Revela qual a informação que está e que não está presente
- Revela relações entre partes da informação, o que permite obter percepções mais amplas
- Possibilita acesso e navegação directos a partes específicas da informação apenas a partir da sua selecção na vista geral
- Encoraja a exploração dos dados

Existem alguns estudos [64] que confirmam que a utilização de vistas gerais resulta num aumento do desempenho do utilizador em várias tarefas da pesquisa da informação. Geralmente, ao desenhar uma visualização, deve ser encontrado o equilíbrio de colocar o máximo de informação possível na vista geral da forma mais clara possível. Uma das mais importantes decisões no desenho de uma visualização é determinar que detalhes da informação devem ser apresentados na vista geral e que informação deve ser apresentada apenas através da interacção do utilizador. A montra de uma loja que revela apenas alguns dos produtos da sua gama é uma boa analogia para demonstrar a importância desta decisão.

Para criar vistas gerais com o objectivo de englobar um vasto conjunto de dados num ecrã relativamente pequeno, existem duas abordagens possíveis no processo de mapeamento visual:

1. Redução da quantidade de dados antes do mapeamento, ou
2. Redução do tamanho físico dos símbolos visuais criados no mapeamento

### 2.6.1 Redução da quantidade dos dados

Um método para reduzir a quantidade dos dados, mantendo uma representação razoável dos dados originais, é a agregação. A agregação agrupa as entidades do conjunto de dados de forma a criar um novo conjunto de dados contendo um total de entidades menor. Cada agregado passa a ser uma entidade, substituindo temporariamente todas as entidades do agregado em questão. Por exemplo, um histograma aplica a agregação para representar a distribuição de dados assente num atributo.

Ao utilizar a agregação, é necessário decidir em primeiro lugar quais as entidades que serão agregadas. As entidades podem ser agregadas por atributos com valores comuns [65] ou por métodos mais avançados como algoritmos de agrupamento de dados [66] ou vizinho mais próximo. A decisão seguinte é a de determinar os novos valores dos atributos das entidades agregadas. Idealmente, os valores dos atributos da entidade agregada devem representar os das suas entidades constituintes. A agregação pode ser aplicada iterativamente para gerar estruturas em árvore de grupos e subgrupos [67]. A última decisão é acerca da representação visual das entidades agregadas que devem, idealmente, revelar o máximo de informação das suas entidades constituintes. A técnica

Aggregate Towers [68] (figura 15) agrega espacialmente as entidades se estas se sobrepuserem num mapa. As entidades agregadas são apresentadas como torres cuja altura representa o número de entidades do agregado. Ao ser efectuado o *zoom out* do mapa são ainda mais agregadas as entidades consoante o nível de detalhe pretendido. Como o *zoom in* o efeito é o contrário, à medida que a aproximação é efectuada, o mapa segrega as torres até ao ponto de não serem necessárias torres para representar os dados a visualizar.

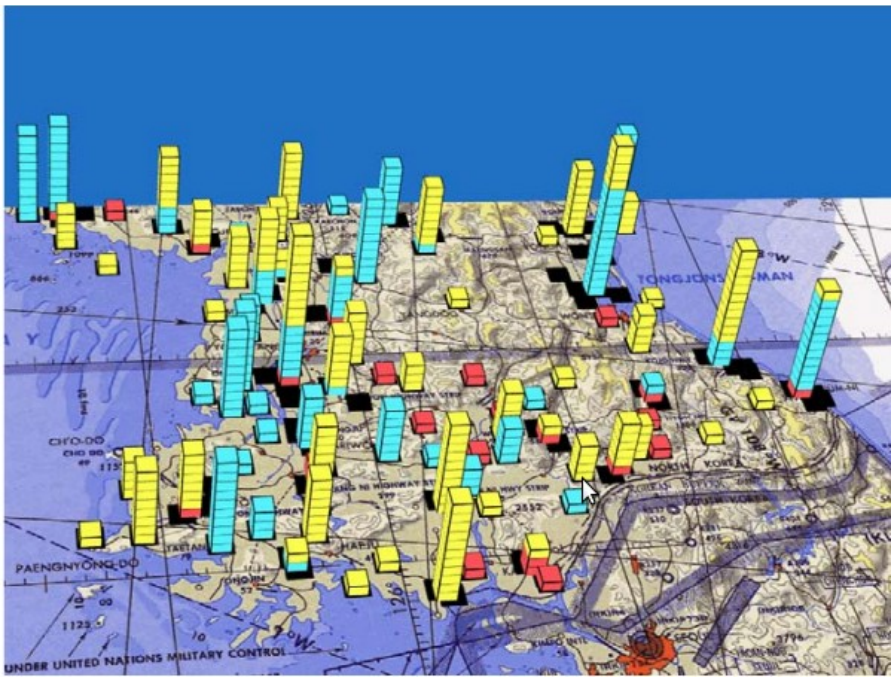



*Figura 15: Redução da quantidade de dados - Aggregate Towers*

A agregação pode também ser aplicada aos atributos das entidades. Os métodos de redução dimensional diminuem o número de atributos em grandes conjuntos de dados multidimensionais de forma a torná-los mais facilmente visualizáveis [69]. O conjunto reduzido de atributos deve capturar aproximadamente as mesmas tendências do conjunto de atributos original. Por exemplo, o método de Análise de Componentes Principais [70] projecta as entidades num novo espaço que melhor preserva a sua variância. O Escalamento Multidimensional [71] utiliza medidas de similaridade entre entidades, baseado nos valores dos seus atributos, para realizar um mapeamento uni, bi ou tridimensional que agregue espacialmente entidades similares.

Existem algumas técnicas para a redução da quantidade dos dados baseada em filtros. A VIDA [72] selecciona um subconjunto representativo de entidades, baseando-se na densidade dos dados e na importância das entidades. O Spotfire [27] relega os atributos menos importantes para métodos interactivos, como as consultas dinâmicas, eliminando-os dos parâmetros da função de mapeamento visual. A quantidade de informação estruturada em árvore é facilmente reduzida através do filtro dos níveis mais profundos da árvore. Isto permite a visualização dos níveis mais próximos da raiz que passam a constituir a vista geral.

### 2.6.2 Redução do tamanho dos símbolos visuais

Alternativamente, o ênfase pode ser colocado na miniaturização dos símbolos visuais gerados pelo processo de mapeamento visual. Tufte defende o aumento da densidade dos dados nas visualizações através da maximização dos dados por unidade de área do espaço disponível no ecrã e pela maximização do rácio dados/tinta [11]. Um índice dados/tinta mais alto é conseguido através da minimização da quantidade de "tinta" necessária para cada símbolo visual e da eliminação de atributos visuais desnecessários. O SeeSoft [74] (figura 16) permite obter uma visão geral de código textual de software usando miniaturização. De forma semelhante ao TableLens (figura 4), cada linha de código é reduzida a um segmento linear de pixeis coloridos cujo tamanho é proporcional ao número de caracteres na linha de código. Desta forma, grandes projectos de software de mais de 50000 linhas de código podem ser visualizados num único ecrã. A utilização de cores pode revelar outros atributos, como qual o programador que escreveu a linha ou se foi ou não testada.

O Pixel Bar Charts [75] reduz o tamanho dos símbolos visuais de linhas tabulares para pixeis únicos, coloridos por um atributo e ordenados no ecrã por outro. O Information Mural [76] leva a miniaturização ao nível sub-pixel. Quando muitos símbolos visuais se sobrepõem e se escondem entre si, o Information Mural visualiza a densidade dos símbolos visuais como uma imagem de raios-X.

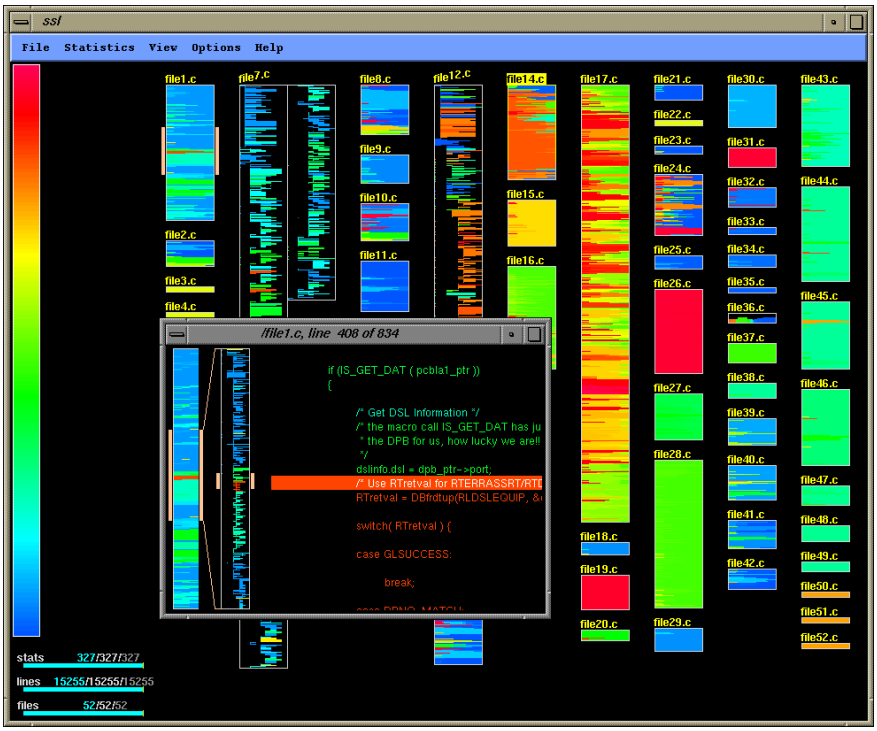



*Figura 16: Redução do tamanho dos símbolos visuais – SeeSoft*

## *2.7 Estratégias de Navegação*

Encontrando-se definida a estratégia de representação visual de grandes conjuntos de informação, surge o próximo desafio no desenho da visualização: a navegação. São necessários métodos interactivos que permitam a navegação desde a vista geral, até aos detalhes de toda a informação. De forma a responder a esta necessidade, assistiu-se à evolução de três estratégias primárias de navegação:

- *Zoom+Pan* (Aproximação+Deslocamento)
- *Overview+Detail* (Visão Geral+Detalhe)
- *Focus+Context* (Focagem+Contexto)

Estas estratégias residem na terceira etapa da *pipeline* de visualização, a transformação na visualização final (figura 2).

### 2.7.1 Aproximação+Deslocamento

As visualizações que possibilitam o *zoom* dos dados iniciam-se com uma visão geral dos mesmos, permitindo que o utilizador efectue aproximações de forma a detalhar a informação do seu interesse. A acção de *zoom out* converge para a vista geral dos dados, enquanto que o *zoom in* vai aumentando o nível de detalhe de um determinado subconjunto de dados. Os utilizadores poderão também, num determinado nível de *zoom*, efectuar o deslocamento (*pan*) no espaço de dados de forma a visualizar no ecrã outros subconjuntos de dados com o mesmo nível de detalhe. A aproximação pode ser uma navegação simples e contínua através do espaço, como no Pad++ [77], ou pode ser usada para fazer *drill down* de dados, como no Treemaps [48] (figura 11).

Apesar da estratégia de aproximação+deslocamento prever uma visão geral dos dados, o principal problema que permanece é a desorientação durante o zoom in. Os utilizadores perdem-se facilmente no espaço de dados num dado nível de zoom enquanto se deslocam, uma vez que não têm presente a visão geral dos dados.

### 2.7.2 Visão Geral+Detalhe

Esta estratégia utiliza várias visualizações para exibir simultaneamente uma visão global e uma visão de detalhe. Um campo indicador na visão geral indica a localização da vista de detalhes dentro do espaço de informação. As visualizações estão ligadas entre si de tal forma que a navegação na vista geral faz com que a vista de detalhes se mova em conformidade. Da mesma forma, quando os utilizadores navegam directamente na vista de detalhes, a visão global é actualizada. É muito comum encontrar implementações desta estratégia em *software* de mapas e gestão de imagens [78]. No SeeSoft [74], a vista geral miniaturizada serve de *scrollbar* para a vista de detalhe do texto propriamente dito (figura 16). O limite de ampliação que garante a usabilidade de imagens 2D é de 30:1 [78]. Podem contudo ser criadas vistas intermédias de forma a aumentar o factor global de ampliação possível. Esta estratégia pode também ser aplicada em ambientes 3D, como no caso do Worlds in Miniature [37]. Esta técnica permite a navegação em mundos

tridimensionais, mostrando uma vista geral do mapa (figura 17), de forma a ajudar os utilizadores a orientarem-se nesse mundo.

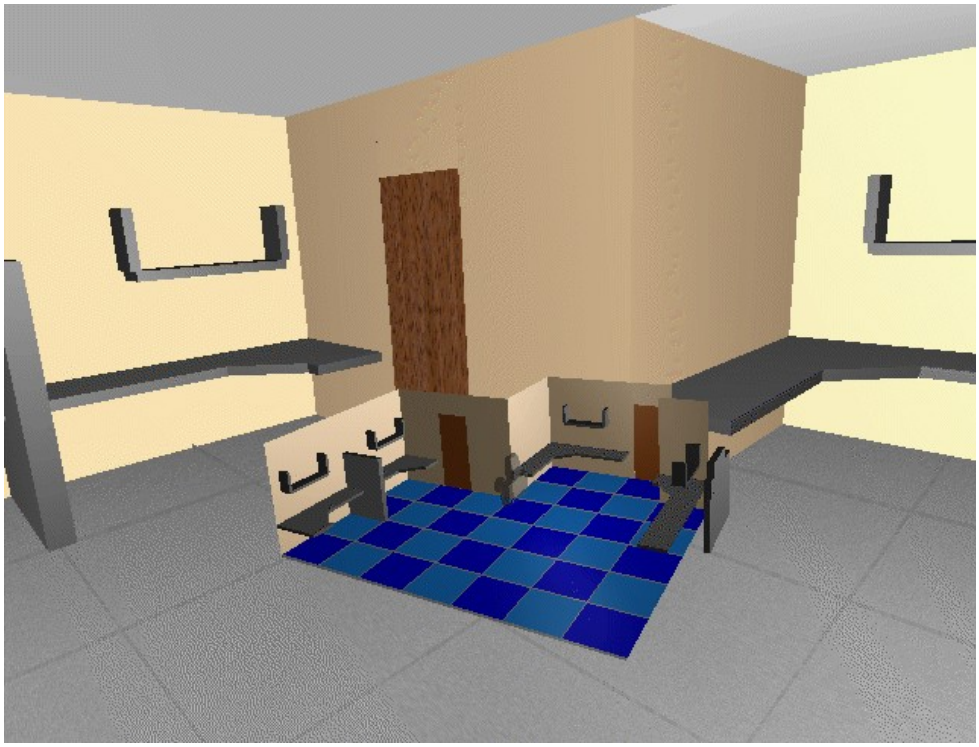


*Figura 17: Visão Geral+Detalhe – Worlds In Miniature*

A visão geral+detalhe preserva uma vista geral de forma a evitar a desorientação na vista de detalhe, contudo sofre de uma descontinuação entre a vista geral e a vista de detalhe.

### 2.7.3 Focagem+Contexto

Esta estratégia expande directamente uma região de focagem dentro da vista geral. O foco é ampliado de forma a fornecer informação detalhada acerca daquela porção do espaço de informação. Os utilizadores podem mover a zona de focagem dentro da vista geral de forma a visualizar os detalhes de outras zonas do conjunto de dados. De forma a arranjar espaço para a região de focagem, a zona circundante da vista geral tem de ser retraída ou pelo menos distorcida. Por esta razão, esta estratégia é também conhecida como "olho de peixe" (*fisheye*) [79] ou

visualização orientada à distorção [80]. Sem distorção, a região ampliada poderia ocluir o contexto adjacente, como acontece ao utilizar-se uma lupa. Uma vez que o contexto adjacente constitui a parte mais importante do contexto, o "efeito lupa" não é desejável, sendo necessário aplicar distorção de forma a preservar a vista geral. Assim, o ponto de focagem é o mais ampliado, reduzindo-se a ampliação com o aumento da distância ao ponto de focagem.

Foram desenvolvidas várias variantes da estratégia focagem+contexto para navegações em espaços de informação uni e bidimensionais, tais como: a Bifocal [28] que usa dois níveis distintos de ampliação, como no TableLens [29] (figura 4); a Perspectiva que apresenta a informação em superfícies com ângulos tridimensionais, como no Perspective Wall [46] (figura 18); as lentes de Ângulo Amplo que criam um efeito "olho de peixe" clássico, como nas Hyperbolic Trees [45]; a Não Linear que usa funções de ampliação mais complexas de forma a criar uma "bolha" ampliada .

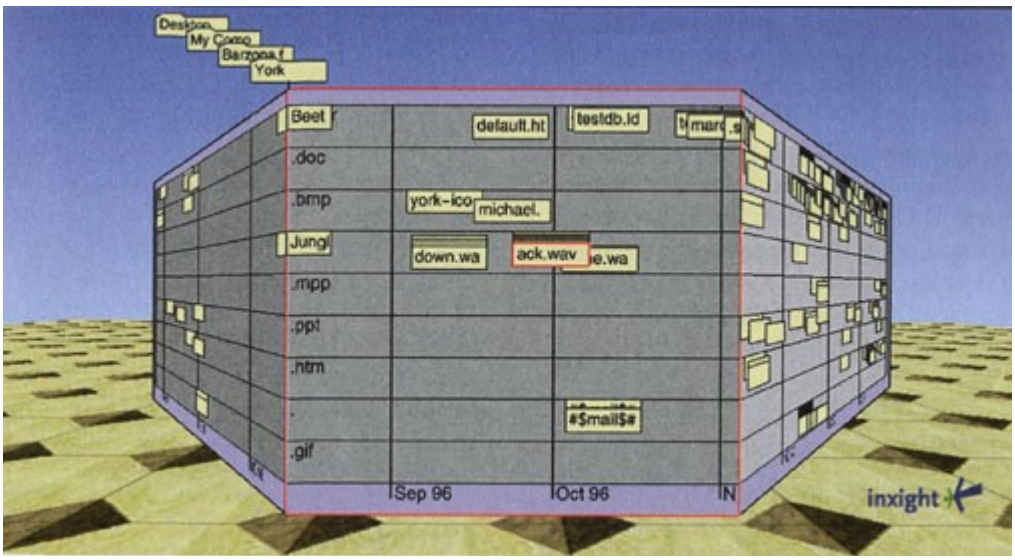



*Figura 18: Focagem+Contexto - Perspective Wall*

Foram também desenvolvidas técnicas "olho de peixe" para navegar em espaços 3D [82]. A estratégia focagem+contexto fornece uma visualização contínua entre a vista geral e o detalhe, contudo causa desorientação devido ao recurso à distorção.

### 2.7.4 Vantagens e desvantagens

Apesar da existência de múltiplos estudos (ex.: [64], [96] e [97]) que mostram as vantagens destas três estratégias em relação à visualização directa do detalhe da informação, as comparações entre elas são inconclusivas, estando normalmente muito ligadas às especificidades de cada desenho, domínios de dados e tarefas do utilizador. É apresentado de seguida um resumo analítico das vantagens e desvantagens de cada uma das estratégias:

| ***Zoom+Pan* (Aproximação+Deslocamento)** | |
|---|---|
| **Vantagens** | **Desvantagens** |
| Eficiente ocupação do espaço disponível no ecrã;<br>Escalabilidade infinita. | Perda da visão geral quando é efectuado a aproximação;<br>Navegação mais lenta. |

*Tabela 2: Vantagens e desvantagens da estratégia de navegação Aproximação+Deslocamento*

| ***Overview+Detail* (Visão Geral+Detalhe)** | |
|---|---|
| **Vantagens** | **Desvantagens** |
| Vista geral estável;<br>Escalável;<br>Visualizações encadeadas;<br>Várias vistas. | Desconexão entre vistas;<br>As vistas competem entre si pelo espaço do ecrã;<br>Vista geral reduzida. |

*Tabela 3: Vantagens e desvantagens da estratégia de navegação Visão Geral+Detalhe*

| ***Focus+Context* (Focagem+Contexto)** | |
|---|---|
| **Vantagens** | **Desvantagens** |
| Vista de detalhe conectada ao contexto envolvente. | Escalabilidade limitada;<br>Distorção. |

*Tabela 4: Vantagens e desvantagens da estratégia de navegação Focagem+Contexto*

## *2.8 Estratégias de Interacção*

Para possibilitar a visualização de informação mais complexa e escalável, poderão ser aplicadas estratégias de interacção. Idealmente, é preferível mapear visualmente toda a informação (incluindo detalhes) num ecrã, de forma a que sejam reveladas no imediato as percepções que de si advêm. Contudo, isto é impossível para dados com algum nível de complexidade. As estratégias de interacção ultrapassam esta limitação ao permitir que os utilizadores explorem mapeamentos e percepções diferentes interactivamente ao longo do tempo. Existem várias técnicas de interacção. No desenho de cada visualização devem ser consideradas algumas das categorias apresentadas de seguida.

### 2.8.1 Selecção

A selecção de entidades ou de subconjuntos de um conjunto de dados é uma das necessidades fundamentais na visualização. Os utilizadores seleccionam entidades para identificar a informação mais relevante. Este procedimento é útil em várias situações, tais como: visualizar informação detalhada acerca das entidades, destacar entidades "escondidas" entre várias outras existentes num dado ecrã, agrupar um conjunto de entidades ou extrair entidades para uso futuro.

Podem considerar-se dois critérios que os utilizadores seguem para a selecção de entidades:

- Selecção directa de entidades
- Selecção indirecta baseada em critérios de pesquisa

A selecção directa de entidades em visualizações é possibilitada pela utilização de várias técnicas [82], como a selecção de símbolos visuais, que representam as entidades pretendidas do conjunto de dados, ou a marcação de uma área onde todos os símbolos visuais (e respectivas entidades) são seleccionados. Os utilizadores podem também seleccionar indirectamente entidades através da aplicação de critérios de selecção em estruturas de informação (secção 2.5). Por exemplo, o

RadialGraphView (da *framework* Prefuse), baseado no GnuTellaVision [83] permite que os utilizadores efectuem selecções de entidades através da pesquisa do valor de um atributo (na figura 19 todas as pessoas com nomes começado por 'Pa' são destacadas).

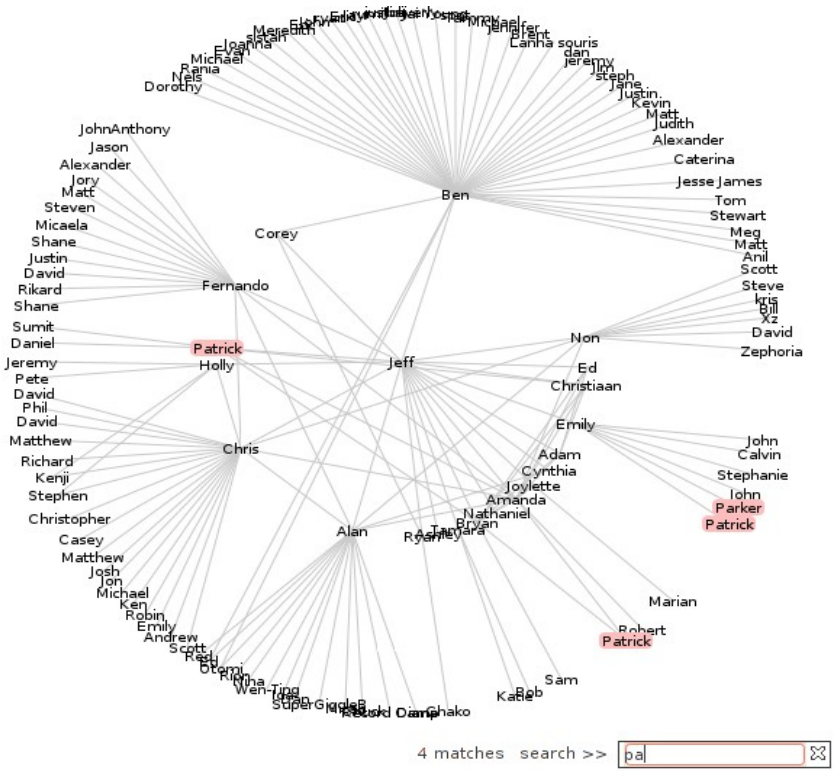



*Figura 19: Selecção de entidades por pesquisa de nome – RadialGraphView*

Existem outras técnicas de selecção associadas à especificidade de cada estrutura de informação, como a selecção de um ramo inteiro numa estrutura em árvore ou a selecção de um caminho numa estrutura em rede. As técnicas de selecção devem ser desenhadas de forma a permitir que os utilizadores possam facilmente seleccionar entidades, expandir a selecção actual a mais entidades e excluir entidades da selecção.

### 2.8.2 Ligação

A ligação é útil para relacionar informação entre visualizações de forma interactiva [84]. A informação pode ser mapeada de formas diferentes em vistas separadas de forma a revelar

diferentes perspectivas de diferentes porções da informação. A forma mais comum de ligação é o *brushing and linking* [85]. As selecções interactivas de entidades são propagadas a outras vistas destacando as entidades correspondentes de forma a tornar reconhecíveis relacionamentos entre elas. Esta estratégia permite que os utilizadores beneficiem simultaneamente das vantagens das várias representações visuais disponíveis. É particularmente útil no relacionamento entre diferentes estruturas de informação, ao utilizar a representação numa estrutura para pesquisar noutra.

Os utilizadores podem seleccionar entidades seguindo um critério numa estrutura, sendo apresentada a distribuição das mesmas noutra estrutura. A ligação também pode ser utilizada num contexto de múltiplas vistas, com várias representações assentes em diferentes estruturas. A combinação da ligação com estratégias de navegação possibilita a utilização de outras técnicas úteis como o *scrolling* sincronizado ou focagem+detalhe especializado para possibilitar o *drill down* em grandes conjuntos de dados [67].

### 2.8.3 Filtragem

A filtragem interactiva permite que os utilizadores reduzam dinamicamente a informação apresentada no ecrã, focando-se na informação do seu interesse. As pesquisas dinâmicas aplicam princípios de manipulação directa à pesquisa de valores de atributos [27]. São utilizados *widgets* visuais para facilitar um ajuste rápido dos parâmetros de pesquisa, assim como o seu resultado em tempo real (figura 20).

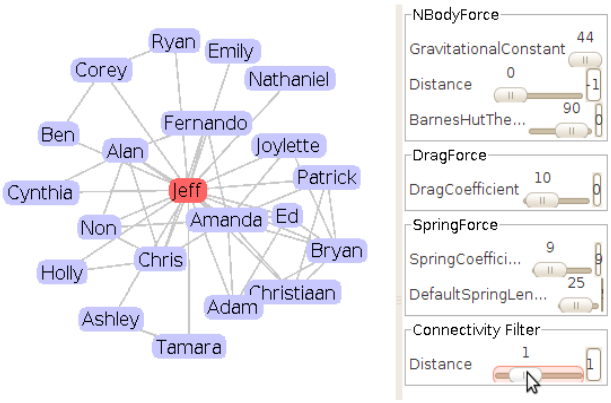



*Figura 20: Filtragem do nível de ligações numa rede social*

Devido ao rápido feedback, as pesquisas dinâmicas permitem não só a redução da quantidade de informação apresentada, como a exploração de relações entre os atributos mapeados e os atributos filtrados. Uma vez que os utilizadores podem rapidamente ajustar os parâmetros da pesquisa, o risco da ocorrência de filtragens excessivas ou escassas é reduzido, resultando numa quase inexistência de resultados com zero entidades ou com demasiadas entidades. O Magic Lenses [73] oferece um filtro específico que permite observar uma determinada localização espacial numa visualização. Para pesquisas mais avançadas envolvendo combinações de operações binárias complexas, são utilizadas metáforas como o Filter Flow [87] que permitem a construção de camadas virtuais de filtros pelo utilizador o que permite orientar o resultado da visualização à informação mais relevante.

### 2.8.4 Reorganização e Remapeamento

Uma vez que um mapeamento de informação para uma forma visual pode não ser o adequado, é essencial permitir que os utilizadores possam alterar esse mapeamento ou que possa seleccionar um de vários mapeamentos disponíveis. A forma mais poderosa de geração de diferentes percepções consiste na reorganização espacial da informação numa determinada representação visual. Por exemplo, o TableLens [29] (figura 4) consegue reordenar espacialmente a sua vista através de diferentes ordenações dos atributos. O Parallel Coordinates [30] (figura 5) pode inverter a ordem dos seus eixos. Isto possibilita que os utilizadores explorem relacionamentos entre diferentes

atributos.

Geralmente, qualquer parte do mapeamento ao longo da pipeline de visualização pode ser controlada pelo utilizador. No caso do Spotfire [27] é possível a edição do gráfico de dispersão através do mapeamento de atributos com várias propriedades visuais como a cor e o tamanho. Também fornece uma panóplia de diferentes representações visuais, incluindo mapas de calor, coordenadas paralelas, histogramas e gráficos de barras. O Visage [26] enfatiza uma técnica denominada *"data-centric interaction"* na qual os utilizadores podem seleccionar directamente entidades, arrastando-as para novas vistas de forma a visualizá-las de formas completamente diferentes. No extremo, existem sistemas como o Sage e SageBrush [26] que permitem que se desenhem novos mapeamentos visuais para o conjunto de dados, utilizando um conjunto de primitivas básicas, conforme descrito anteriormente no ponto 2.4. O Sage consegue inclusive gerar automaticamente certos mapeamentos visuais para determinados conjuntos de dados, através da utilização de um sistema pericial baseado em regras.

## *2.9 Sumário*

Neste capítulo introduziram-se os principais conceitos da visualização de informação como a percepção, o desenho e a pipeline de visualização. Foram também apresentadas as diferentes estruturas de informação que possibilitam o mapeamento visual dos dados, sendo referidos métodos de visualização representativos de cada uma delas. Finalmente, são descritas as diferentes estratégias de representação visual, interacção e navegação que têm como objectivo a optimização da representação visual da informação recorrendo à interacção com o utilizador.

# 3 Agrupamento de Dados

## *3.1 Introdução*

As técnicas de agrupamento de dados têm como objectivo organizar os objectos de um conjunto de dados em vários grupos homogéneos, onde os dados semelhantes pertencem ao mesmo grupo, enquanto os dissimilares pertencem a grupos distintos [91][111]. A divisão da informação em $K$ grupos possibilita a descoberta de relações e padrões na informação que seriam impossíveis através da observação das entidades do conjunto de dados independentemente. Esta área de pesquisa tem sido alvo de grande interesse nos últimos anos, o que tem resultado na criação de vários algoritmos de agrupamento de dados com características distintas. Nenhum algoritmo é adequado a todas as situações pois diferentes algoritmos têm desempenhos diferentes de acordo com o conjunto de dados ao qual se aplicam. O agrupamento de dados encontra-se em grande expansão em várias áreas como a tomada de decisão, estruturação de documentos e segmentação de imagem, *marketing*, genética, entre outros.

## *3.2 Aprendizagem automática*

O crescimento drástico de aplicações em áreas como a pesquisa na Internet, imagem e vídeo digital resultou na criação de vastos conjuntos de dados de elevado volume e alta dimensionalidade. É esperado que o universo digital consuma aproximadamente 2810 exabytes em 2011 [87] (1 exabyte equivale a 1.000.000 terabytes). A grande maioria dos dados é guardada digitalmente em suporte electrónico. Este factor revela um enorme potencial para o desenvolvimento de técnicas de análise automática de dados. Além do crescimento na quantidade de informação disponível, também se verificou uma maior diversificação de dados (texto, imagem e vídeo).

A aprendizagem automática é um campo da Inteligência Artificial que tem o objectivo de dotar o computador com a capacidade de aprendizagem, estudando para isso algoritmos e técnicas que

permitam ao computador aperfeiçoar-se no desempenho de uma determinada tarefa, sem que seja necessária intervenção humana.

Existe uma clara distinção entre problemas de aprendizagem **supervisionada** (muitas vezes referida apenas como classificação) e **não supervisionada** (como por exemplo, o agrupamento de dados). O primeiro tipo envolve apenas dados rotulados, isto é, o treino de padrões associados a categorias conhecidas previamente, enquanto que o o último opera apenas utilizando dados não rotulados [88]. Tem surgido um interesse crescente numa abordagem intermédia denominada aprendizagem **semi-supervisionada** [89]. Na classificação semi-supervisionada, apenas uma pequena porção do conjunto de dados de treino se encontra rotulada. Neste tipo de classificação, os dados não rotulados são também utilizados no processo de aprendizagem, ao invés de serem descartados.

A figura seguinte ilustra este espectro de diferentes tipos de problemas de aprendizagem de interesse para as áreas de reconhecimento de padrões e aprendizagem automática.

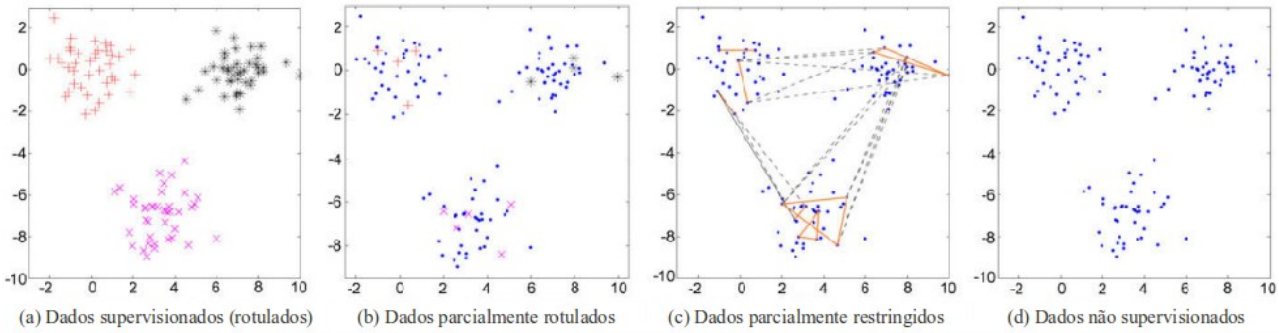



*Figura 21: Tipos de problemas de aprendizagem - adaptado de [90]*

### 3.2.1 Aprendizagem Supervisionada

A aprendizagem supervisionada, também denominada por classificação, é a área da aprendizagem automática que tem como objectivo inferir uma função ou regra de decisão (classificador) a partir de um conjunto de dados de treino, de forma a conseguir efectuar uma previsão sobre os novos dados. Cada um dos objectos desse conjunto de dados encontra-se associado a um atributo alvo, a

classe do objecto.

Um algoritmo de classificação atravessa as seguintes fases:

1. Análise do conjunto de dados de treino, previamente classificados
2. Inferência de um classificador, obtido como resultado da fase anterior
3. Aplicação do classificador gerado aos restantes dados de forma a obter uma previsão da sua classificação

O classificador inferido deve permitir uma previsão correcta de qualquer objecto válido no domínio de dados. Para isto, o algoritmo de aprendizagem deve utilizar um critério que permita a generalização a novos dados, baseando-se nas observações do conjunto de treino.

Um classificador é uma função $f$ que atribui a um objecto $x_i$ de um conjunto de dados $X\ (x_i \in X)$ caracterizado por $d$ atributos, uma classe $c_k \in C$ [91]:

$$f : \mathbb{R}^d \rightarrow C$$

em que $C = \{c_1, c_2, \ldots, c_k\}$ representa o conjunto de todas as $K$ classes a que um objecto de dados pode pertencer.

Existem vários algoritmos de classificação para treino de conjuntos de dados com o objectivo de aprender a função ou regra de decisão final. De entre as abordagens existentes, são destacados [91] [92] os métodos estatísticos (incluem o algoritmo $k$-vizinhos-mais-próximos – $k$-NN [141]), árvores de decisão (incluem os algoritmos ID3 [142] e CART [143]), redes neuronais artificiais (incluem o algoritmo MLP – *MultiLayer Perceptron* [144]) e as máquinas de suporte vectorial (ou SVM – *Support Vectorial Machines* [145]). Uma outra abordagem para a classificação de dados consiste na combinação de classificadores. O objectivo é o de aumentar o desempenho da classificação dos dados através da combinação de decisões individuais de cada classificador. Esta abordagem tem-se revelado eficiente, pelo que continua activa a investigação nesta área [94][95].

### 3.2.2 Aprendizagem Não Supervisionada

Na aprendizagem não supervisionada, mais especificamente no agrupamento de dados, é pretendido que seja efectuado o processo de aprendizagem com base num conjunto de dados não rotulado, onde não se conhece à partida qual o número de classes existentes no conjunto de dados nem a classe associada a cada objecto. O agrupamento de dados é um problema mais difícil e desafiante que a aprendizagem supervisionada [90].

O agrupamento de dados tem como objectivo associar dados a classes, descobrindo distribuições com significado. Isto é atingido através da divisão de um conjunto de dados em grupos, incluindo objectos de dados similares dentro do mesmo grupo e dissimilares em grupos distintos. O resultado final obtido por diferentes algoritmos de agrupamento de dados pode levar à descoberta de diferentes estruturas para um mesmo conjunto de dados (figura 22). Isto acontece pelo facto de não serem conhecidas quaisquer classes ou outro tipo de informação que indique a estrutura da informação a analisar. Assim, o resultado do agrupamento de dados está dependente de factores como os parâmetros de entrada, as inicializações do algoritmo de agrupamento de dados ou a própria estrutura inicial dos dados.

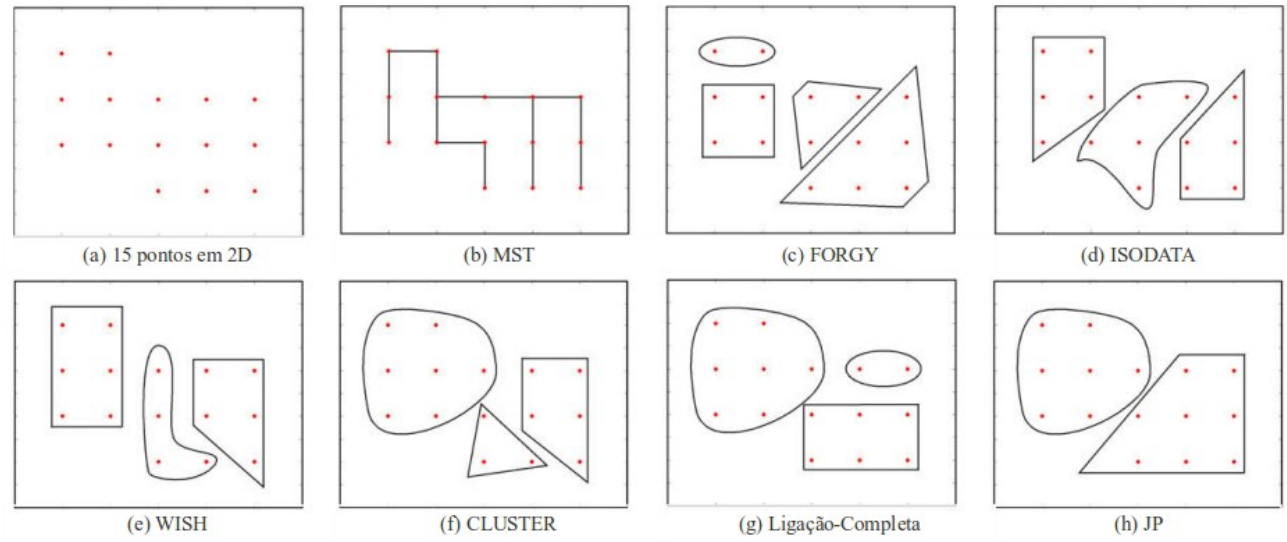



*Figura 22: Diferentes agrupamentos para o mesmo conjunto de dados – adaptado de [90]*

Algumas das principais áreas que aplicam técnicas de agrupamentos de dados incluem a compressão de dados, análise exploratória de dados, *webmining*, segmentação de imagem, genética, análise de informação espacial, *marketing* e estudos de mercado.

### 3.2.3 Aprendizagem Semi-Supervisionada

Como o nome sugere, a aprendizagem semi-supervisionada é uma abordagem híbrida das aprendizagens não supervisionada e supervisionada. Neste paradigma pressupõe-se que a aprendizagem é efectuada utilizando conjuntos de dados que incluem informação rotulada e não rotulada.

O objectivo da aprendizagem semi-supervisionada é combinar os dados pré-classificados com os não classificados de forma a afectar o processo de aprendizagem, influenciando positivamente a qualidade das previsões ou agrupamentos finais. A aprendizagem semi-supervisionada tem um elevado valor prático, uma vez que permite a utilização de conjuntos de dados onde a informação rotulada é escassa. A classificação dos dados pode ser difícil de obter uma vez que podem requerer intervenção humana, aparelhos especiais ou experiências caras e morosas. Por exemplo [93]:

- Na filtragem de *spam*, uma instância $x$ é um email, enquanto que o rótulo $l$ corresponde à categorização do utilizador (*spam* ou não). Neste caso, a dificuldade é a de conseguir que um utilizador comum categorize um elevado número de emails.
- Na vídeo-vigilância, uma instância $x$ corresponde a um *frame*, enquanto que o rótulo $l$ é a identidade do objecto no vídeo. Rotular manualmente objectos num número elevado de *frames* de vídeos de vigilância é uma tarefa morosa e repetitiva.

Enquanto os dados pré-classificados são difíceis de obter nestes domínios, a informação não rotulada é abundante e fácil de obter: os discursos verbais podem ser gravados a partir de emissões radiofónicas; as frases textuais podem ser extraídas da Internet; os *emails* encontram-se disponíveis no servidor de email; e as câmaras de vigilância gravam imagens 24h por dia.

## 3 AGRUPAMENTO DE DADOS

O crescente interesse na aprendizagem semi-supervisionada tem a ver com o facto deste paradigma utilizar tanto informação rotulada com não rotulada para atingir potencialmente melhores resultados que na aprendizagem supervisionada. De uma outra perspectiva, a aprendizagem semi-supervisionada pode atingir o mesmo nível de desempenho que a aprendizagem supervisionada, com menos instâncias rotuladas. A aprendizagem não supervisionada também beneficia da inclusão do conhecimento *a priori* (através das instâncias rotuladas) de forma a obter melhores agrupamentos de dados. De qualquer das formas, o resultado é um menor esforço na categorização de instâncias, o que resulta num menor custo e uma maior assertividade.

As estratégias de aprendizagem semi-supervisionada baseiam-se na extensão tanto da aprendizagem supervisionada, como da não supervisionada de forma a incluir informação adicional típica do outro paradigma. A primeira abordagem é também referida como **classificação semi-supervisionada**, enquanto que a segunda é conhecida como **agrupamento de dados com restrições**.

**Classificação semi-supervisionada.** Também conhecida como classificação com dados parcialmente rotulados, é uma extensão ao problema da classificação supervisionada, onde todos os objectos se encontram pré-classificados. O conjunto de treino inclui as instâncias rotuladas e as não rotuladas, sendo que normalmente é assumido que existem muitos mais dados não classificados que classificados. O objectivo da classificação semi-supervisionada é o de treinar um classificador $f$ tanto com os dados rotulados como os não rotulados até que o classificador obtido seja melhor que o gerado com recurso apenas aos dados pré-classificados.

**Agrupamento de dados com restrições.** É uma extensão à aprendizagem não supervisionada. No agrupamento de dados semi-supervisionado são geralmente definidas restrições entre pares de objectos de dados em vez de serem especificados rótulos de classes aos objectos, o que constitui uma forma mais frágil de representar o conhecimento prévio. Uma restrição de ligação obrigatória entre um determinado par de objectos, corresponde ao requisito que os dois objectos devem pertencer ao mesmo grupo. Já uma restrição de ligação proibida indica que os objectos afectados pela restrição não devem pertencer ao mesmo grupo. O objectivo do agrupamento de dados com restrições é o de obter um melhor agrupamento que o usado utilizando somente objectos não

rotulados e sem qualquer informação fornecida *a priori*.

Recentemente têm surgido abordagens baseadas na combinação da algoritmos de agrupamento de dados com restrições com o objectivo de aumentar o desempenho dos agrupamentos resultantes [156][157]. Este tema é explorado com maior pormenor na secção 3.5.

## 3.3 Fases do Agrupamento de Dados

Desde a selecção da informação a processar até ao agrupamento final, a actividade de agrupamento de dados atravessa diferentes fases. Jain et. al. [98] distinguem 5 fases típicas num agrupamento de dados:

1. Representação dos objectos de dados (opcionalmente inclui extracção de atributos e/ou selecção de atributos);
2. Definição de uma medida de proximidade (similaridade/dissimilaridade) entre os objectos de dados apropriada ao domínio dos dados;
3. Agrupamento de dados;
4. Abstracção de dados (se necessário);
5. Avaliação do agrupamento obtido (se necessário).

A figura 23 mostra a sequência típica dos primeiros três passos de um processo de agrupamento de dados, incluindo um ciclo de *feedback* em que o resultado do agrupamento pode afectar a selecção/extracção de atributos subsequentes e os cálculos de similaridade/dissimilaridade entre os dados.

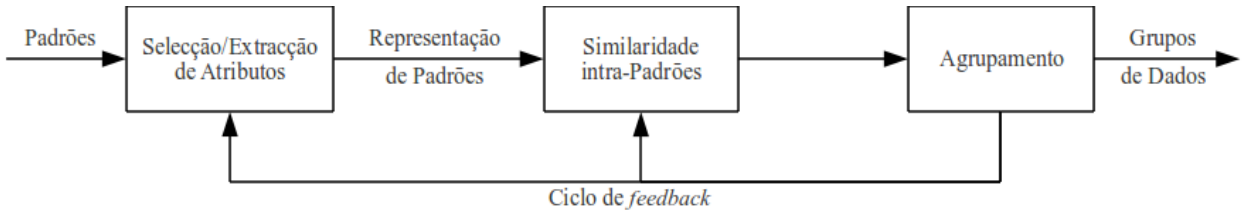



*Figura 23: Passos do Agrupamento de Dados – adaptado de [98]*

De seguida são resumidas cada uma das fases do agrupamento de dados.

### 3.3.1 Representação dos Objectos de Dados

O objectivo da fase de representação dos objectos de dados é o de seleccionar a informação efectivamente relevante para o problema em causa. As variáveis que devem ser consideradas nessa selecção incluem o número, tipo e escala dos atributos disponíveis para o algoritmo de agrupamento e, eventualmente , o número de grupos a obter. Desta forma, é possível dividir a representação dos objectos de dados em duas sub-fases: a **selecção e extracção de atributos** e a **selecção do algoritmo de dados**.

**Selecção e extracção de atributos**. A selecção de atributos consiste na identificação do subconjunto dos atributos originais que são mais descritivos e discriminatórios e que serão usados no subsequente processo de agrupamento, enquanto a extracção de atributos consiste no uso de uma ou mais transformações dos atributos originais, de forma, a obter novos atributos. Após esta análise inicial, pode chegar-se à conclusão que será necessária a mudança de escala dos atributos e/ou normalização. Este processo é denominado pré-processamento de dados. Em todos os casos, o objectivo consiste em melhorar a exactidão do agrupamento e a eficiência computacional. A redução do número de atributos é também benéfica pelo facto de ter a capacidade de produzir um resultado que pode ser inspeccionado visualmente por um ser humano.

**Selecção do algoritmo de dados**. De acordo com o conjunto de dados e tipos de atributos a processar, deve ser seleccionado um algoritmo de agrupamento de dados. O objectivo deste algoritmo será o de encontrar a estrutura do conjunto de dados. Um algoritmo de agrupamento de dados é caracterizado por uma medida de proximidade (secção 3.3.2) e por um critério de agrupamento de dados que induzem o resultado final (agrupamento de dados). Os diferentes tipos de agrupamentos de dados encontram-se descritos na secção 3.4.

### 3.3.2 Definição da Medida de Proximidade

A medida de proximidade quantifica a semelhança entre dois objectos de um conjunto de dados. Esta medida é normalmente definida por uma função de distância entre os pares de objectos. Normalmente, todos os atributos contribuem de igual forma para a definição da proximidade entre os objectos. Existem várias medidas de proximidade estudadas (principais medidas disponíveis em [99], [100], [101], [102], [103] e [104]), quer calculando a dissimilaridade entre dois objectos de dados utilizando uma medida de distância definida no espaço de atributos , como caracterizando a similaridade conceptual entre dois objectos.

A proximidade entre dados descritos unicamente por atributos contínuos é tipicamente calculada a partir da distância entre cada par de objectos, usando para isso uma medida de distância. A ideia de distância é formalizada e generalizada pela matemática através do conceito de métrica. Qualquer métrica obedece aos seguintes requisitos matemáticos:

1. $d(x_i, x_j) \geq 0$: a distância entre dois objectos de dados é um número não negativo;
2. $d(x_i, x_j) = 0$: a distância de um objecto de dados a ele próprio é 0;
3. $d(x_i, x_j) = d(x_j, x_i)$: a distância é uma função simétrica;
4. $d(x_i, x_j) \leq d(x_i, x_h) + d(x_h, x_j)$: a distância no espaço entre um objecto $x_i$ e o objecto $x_j$ não é superior à distância entre esses dois objectos com um desvio sobre qualquer outro dado $h$.

Todas as medidas de distância são formuladas de forma a permitir uma ponderação diferenciada dos atributos quantitativos. A ponderação $w$ do atributo $v$ deve ser definida tal que $w_v \in [0,1]$. Por omissão e se não houver nenhuma indicação em contrário, todos os atributos são ponderados igualmente ($w_v = 1$).

A medida de distância mais usada é a distância euclidiana. A distância euclidiana tem a particularidade apelativa de que $d(x_i, x_j)$ pode ser interpretada como a distância física entre dois objectos com $d$ dimensões ($x_i' = \{x_{i1}, \ldots, x_{id}\}$ e $x_j' = \{x_{j1}, \ldots, x_{jd}\}$) no espaço euclidiano. A distância

euclidiana é frequentemente usada para avaliar a proximidade entre objectos num espaço bidimensional ou tridimensional e funciona bem para conjuntos de dados com grupos compactos e isolados. A distância euclidiana é definida por:

$$d(x_i, x_j) = (\sum_{v=1}^{p} w_v{}^2 (x_{iv} - x_{jv})^2)^{\frac{1}{2}}$$

em que $x_{iv}$ e $x_{jv}$ são, respectivamente, os valores do atributo $v$ dos objectos $x_i$ e $x_j$, $w_v$ a ponderação do atributo $v$ e $p$ o número total de atributos.

### 3.3.3 Agrupamento de Dados

O agrupamento de dados pode ser realizado seguindo diferentes algoritmos. Cada algoritmo tem o seu próprio critério de agrupamento que poderá ser expresso por uma função de custo ou outro tipo de regra. Este critério determina a forma como o agrupamento do conjunto de dados é efectuado. Para a escolha do critério de agrupamento de dados deve ter-se em atenção, sempre que possível, a forma dos grupos.

Existem várias classificações propostas para os algoritmos de agrupamento. Na secção 3.4 são descritos os diferentes tipos de algoritmos de agrupamento de dados de acordo com uma delas.

### 3.3.4 Abstracção de Dados

A abstracção de dados é o processo de extracção de uma representação compacta e simples de um conjunto de dados. Por um lado, é pretendido simplificar a análise automática de forma a facilitar o processamento posterior dos dados. Por outro, para que a representação obtida seja intuitiva e de mais fácil compreensão, a simplificação deve ser orientada à análise humana. Uma abstracção típica de dados é uma descrição compacta de cada grupo, normalmente em termos de protótipos ou objectos representativos como o centróide [105].

### 3.3.5 Avaliação de Resultados

A avaliação de um agrupamento de dados consiste em validar a qualidade dos resultados obtidos por um algoritmo de agrupamento de dados. Esta acção é realizada aplicando medidas de validação de agrupamentos de dados baseadas em critérios externos [107] ou internos [106], denominados índices de validação de agrupamentos de dados. Um resumo das principais medidas de validação pode ser encontrado em [111].

Os critérios externos avaliam os resultados da aplicação de um algoritmo de agrupamento, comparando-os com uma estrutura de dados, definida anteriormente, que reflecte o conhecimento ou intuição de como os dados devem estar agrupados. Estes critérios são úteis para permitir uma avaliação e comparação objectiva entre diferentes algoritmos de agrupamento aplicados a repositórios de dados, para os quais as etiquetas das classes correspondem à estrutura real do agrupamento dos dados.

Um dos mais conhecidos índices de validação baseado em critérios externos é o Índice Rand [108]. Assumindo $l_i$ como o rótulo da instância $i$ do conjunto de dados resultante do agrupamento de dados e $l_i'$ como a classe da mesma instância no conjunto de dados pré-classificado, é possível indicar o número de instâncias agrupadas correctamente:

$$M_{=} = |\{i, j : ((C_i, C_j) \wedge (C_i', C_j'))\}|$$

e o número de instâncias agrupadas incorrectamente:

$$M_{\neq} = |\{i, j : ((C_i, C_j) \wedge (C_i', C_j'))\}|$$

Sendo $n$ correspondente ao número de objectos de $C$, o Índice Rand $RI$ é obtido por:

$$RI = \frac{M_{=} + M_{\neq}}{\binom{n}{2}}$$

assumindo $\binom{n}{2} = \frac{n(n-1)}{2}$.

Uma vez que o Índice Rand penaliza as partições com um maior número de grupos, o Índice Rand Ajustado (ou *ARI – Adjusted Rand Index*) [109] é normalmente mais utilizado. O ARI normaliza o Índice Rand de forma a ajustá-lo ao número de grupos através da comparação do número esperado de agrupamentos correctos com o número de agrupamentos correctos observados. Assume-se $U$ como o conjunto de classes *conhecidas a priori*, sendo $r$ o número de classes de $U$ e $V$ como o conjunto de grupos resultante do agrupamento de dados, sendo $c$ o número de grupos de $V$. É possível resumir as intersecções entre os dois conjuntos através da seguinte matriz de contingência:

| Classe / Grupo | $V_1$ | $V_2$ | ... | $V_c$ | Somatório |
|---|---|---|---|---|---|
| $U_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1c}$ | $a_1$ |
| $U_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2c}$ | $a_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $U_r$ | $n_{r1}$ | $n_{r2}$ | ... | $n_{rc}$ | $a_r$ |
| Somatório | $b_1$ | $b_2$ | ... | $b_c$ | $n$ |

onde $n_{ij}$ denota o número de objectos comuns nos dois conjuntos ($n_{ij}=|U_i \cap V_j|$), $a_i$ o total de objectos em comum para a classe $i$, $b_j$ o total de objectos em comum para o grupo $j$ e $n$ o número total de objectos. O ARI é então dado por:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

Tal como na maioria dos índices, o resultado do ARI é um número real entre 0 e 1 (inclusive). Um resultado $ARI = 1$ significa que todos os objectos foram correctamente agrupados, enquanto que o $ARI = 0$ significa que nenhum objecto integrou o grupo esperado. Mais informação acerca do ARI encontra-se disponível em [109] e [110].

## *3.4 Tipos de Agrupamento de Dados*

Com a existência de diversos algoritmos de dados com características diferentes e melhor adaptados a conjuntos específicos de dados, surgiram várias classificações com o intuito de facilitar a tarefa da selecção do algoritmo de agrupamento de dados mais adequado a determinado problema. Algumas das classificações existentes incluem [99] , [112] e [113] . Segundo a taxonomia definida por Duarte [111], as abordagens de agrupamento de dados podem ser divididas em cinco categorias: **abordagens de partição**, **hierárquicas**, **baseadas em densidade**, **baseadas em grelha** e **baseadas em modelos**.

### 3.4.1 Abordagens de Partição

Os algoritmos de agrupamento de dados de partição procuram estruturar um conjunto de dados $X$ num agrupamento de dados $P=\{C_{1,..},C_K\}$ com $K$ grupos, optimizando uma função-objectivo, $f:P\rightarrow\mathbb{R}$, que procura reunir objectos semelhantes no mesmo grupo e colocar objectos dissemelhantes em grupos diferentes. Inicialmente, os algoritmos de agrupamento desta categoria criam um primeiro agrupamento de dados e, em seguida, efectuam iterativamente a realocação de objectos de um grupo para outro, com o intuito de minimizar a função-objectivo $f$. O erro quadrático é a função-objectivo mais utilizada para o efeito

$$f=\sum_{k=1}^{K}\sum_{x_i\in C_k}\|x_i-\overline{x_k}\|^2$$

em que $\|x_i-\overline{x_k}\|$ consiste na distância euclidiana entre o objecto $x_i$ e o centro de grupo mais próximo $\overline{x_k}$.

O algoritmo de agrupamento de dados $K$-médias [114] é o algoritmo mais conhecido e usado das de todas as abordagens. Este algoritmo recebe como parâmetro de entrada o número de grupos pretendido, $K$, efectuando os seguintes passos:

1. Escolhe de forma aleatória $K$ objectos do conjunto de dados $X$ como os centros (*centróides*) iniciais de cada grupo, $\{\overline{x_1},...,\overline{x_K}\}$;

2. Atribui cada objecto $x_i$ ao grupo $C_k$ cujo centro $\overline{x_k}$ se encontra mais próximo;
3. Actualiza cada centro de grupo, $\overline{x_k}$, como sendo o vector médio dos $|C_k|$ objectos associados a esse grupo usando a equação $\overline{x_k}=\frac{\sum_{x_i \in C_k} x_i}{|C_k|}$.
4. Os passos 2 e 3 são repetidos até que não exista qualquer modificação nos grupos de uma iteração para a seguinte, garantindo que a função-objectivo converge para um óptimo (local).

Outros algoritmos inseridos nas abordagens de partição incluem derivados do $K$-Médias, como o ISODATA [115] e também o $K$-Medóides (e derivados, como o PAM (*Partitioning Around Medoids*) [116] e o CLARANS (*Clustering Large Application based upon RANdomized Search*) [117]) que estendem o algoritmo $K$-Médias utilizando o objecto mais centralmente localizado no grupo (*medóide*), em vez de considerar o valor médio dos objectos.

### 3.4.2 Abordagens Hierárquicas

Os algoritmos de agrupamento hierárquico têm por objectivo a construção de uma sequência hierárquica de agrupamentos encaixados. Esta sequência é criada através da aglomeração ou divisão gradual dos grupos. A estrutura hierárquica resultante é representada na forma de árvore, denominada dendograma.

Um algoritmo hierárquico pode ser classificado como **aglomerativo** ou **divisivo**, dependendo de como a decomposição hierárquica é realizada.

**Algoritmos aglomerativos**. A aproximação aglomerativa, também denominada *bottom-up*, inicia-se com cada um dos $n$ objectos $x_i \in X$ constituindo grupos separados $C_i=\{x_i\}$. Sucessivamente vão-se fundindo os grupos mais próximos ou mais similares (tendo em conta uma determinada função-objectivo). O algoritmo termina quando todos os objectos pertençam a um só grupo (o nível mais

elevado da hierarquia), ou até que se verifique uma determinada condição de término. O funcionamento dos algoritmos hierárquicos aglomerativos encontra-se ilustrado na figura 24, seguindo os passos da esquerda para a direita.

**Algoritmos divisivos**. Nos algoritmos divisivos (ou aproximação *top-down*) acontece o inverso. Inicialmente, todos os objectos encontram-se no mesmo grupo e em cada iteração, os grupos são sucessivamente divididos (bissecções sucessivas) em grupos mais pequenos. O algoritmo finaliza a sua execução assim que cada objecto esteja num só grupo ou até que se verifique uma determinada condição de término. A figura 24 ilustra também o funcionamento dos algoritmos de agrupamento hierárquicos divisivos, seguindo os passos da direita para a esquerda.

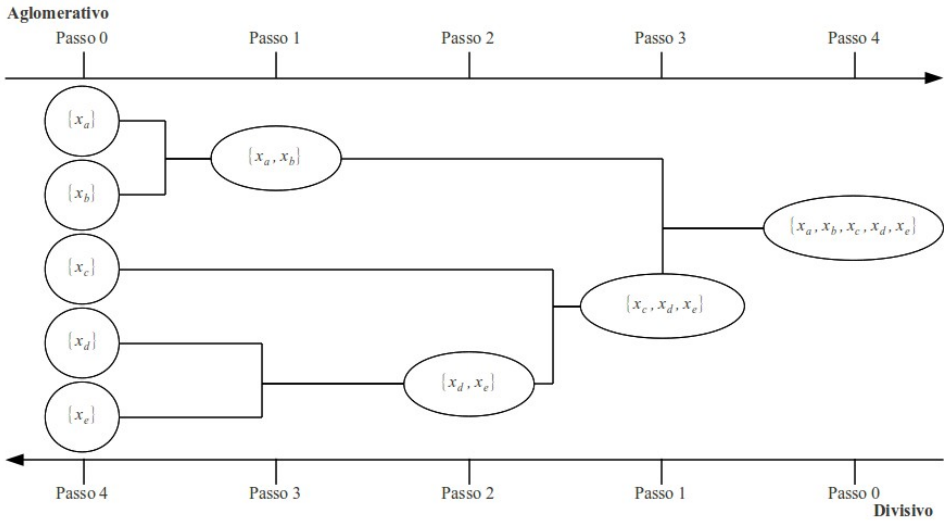



*Figura 24: Agrupamento de Dados Hierárquico – adaptado de [111]*

Os algoritmos aglomerativos são mais usados que os divisivos, já que são computacionalmente menos dispendiosos. Os algoritmos ligação simples (SL – *Single Link*) [118], ligação completa (CL – *Complete Link*) [119], média do grupo (AL – *Average Link*) [118], ligação centróide [118] e ligação Ward (WL – *Ward Link* ou mínima variância) [120] são exemplos clássicos da abordagem hierárquica aglomerativa.

Dois algoritmos hierárquicos aglomerativos mais recentes são o CURE (*Clustering Using*

*Representatives*) [121] e o Chameleon [122] que seguem uma estratégia diferente. O CURE adopta o meio-termo entre as abordagens baseadas em centróides e as abordagens baseadas em todos os objectos do conjunto de dados . Ao invés de utilizar um único centróide ou objecto para representar o grupo, é escolhido um número fixo $c$ de objectos representativos, de forma a identificarem a forma e o tamanho do grupo. Assim, no processo de agrupamento aglomerativo, a distância entre dois grupos é calculada considerando apenas os objectos representativos mais próximos entre cada grupo. O Chameleon explora a modelação dinâmica no agrupamento hierárquico. No processo de agrupamento, dois grupos são unidos se a inter-conectividade e proximidade entre esses grupos estiver altamente relacionada com a inter-conectividade e a proximidade dos objectos dentro dos grupos.

### 3.4.3 Abordagens Baseadas em Densidade

As abordagens anteriormente descritas agrupam os objectos de dados baseando-se em medidas de distância. Existem, contudo, outros algoritmos de agrupamento que se baseiam também na noção de densidade. Estes têm como intuito encontrar regiões densas no espaço dos dados e fazer a correspondência entre os objectos que se encontram em cada região densa e os grupos do agrupamento de dados. A ideia geral dos algoritmos de agrupamento baseados em densidade consiste em fazer crescer um determinado grupo enquanto houver objectos no grupo com densidade (um número de objectos) na sua vizinhança (num raio) superior a um determinado número limite.

Um exemplo de algoritmo baseado em densidade é o DBSCAN (*Density Based Spatial Clustering of Applications with Noise*) [123]. Este algoritmo recebe como parâmetros de entrada o número mínimo de objectos para uma região ser considerada densa $MinPts$ e o raio que define a vizinhança de cada objecto $Eps$, efectuando os seguintes passos:

1. É escolhido ao acaso um objecto $x_i$, que tenha $MinPts$ objectos a uma distância não superior a $Eps$;
2. É formado um grupo encontrando todos os objectos que se encontrem a uma distância não superior a $Eps$ de $x_i$;

3. O grupo é alargado, adicionando todos os objectos que se encontrem a uma distância igual ou inferior a *Eps* de qualquer objecto já incluído no grupo (desde que a densidade mínima seja satisfeita);
4. O passo 3 repete-se até que não seja possível adicionar mais objectos ao grupo. Nesta altura os objectos do grupo são rotulados;
5. Todos os passos anteriores são repetidos para um novo objecto não rotulado. O algoritmo termina quando não for possível formar mais nenhum grupo.

O DENCLUE (DENsity-based CLUstEring) [124] é um algoritmo de agrupamento baseado num conjunto de funções de distribuição de densidade. O algoritmo baseia-se nas seguintes ideias: a influência de cada objecto de dados pode ser formalmente modelada usando uma função matemática, chamada função de influência, que descreve o impacto de um objecto de dados na sua vizinhança; a densidade global do espaço de dados pode ser modelada analiticamente como a soma das funções de influência de todos os objectos de dados; e os grupos podem ser determinados matematicamente pela identificação de atractores de densidade, em que os atractores de densidade são os máximos locais da função de densidade global.

### 3.4.4 Abordagens Baseadas em Grelha

Os algoritmos de agrupamento baseados em grelha dividem o espaço dos dados em células, formando uma estrutura em forma de grelha. O objectivo é separar os valores possíveis de cada atributo num número de intervalos contíguos, criando um conjunto de células em grelha com múltiplas dimensões. Cada objecto é colocado numa célula cujo intervalo de valores para cada atributo contém os valores do objecto. No processo de agrupamento de dados apenas são consideradas as células da grelha e alguma informação estatística (por exemplo, o número de objectos e a densidade em cada célula). A definição da grelha tem um enorme impacto nos resultados de agrupamento. Ao contrário das abordagens anteriores, o custo de processamento dos algoritmos de agrupamento baseados em grelha é independente do número de objectos, dependendo antes do número de células que, por norma, é muito inferior ao número de objectos do conjunto de dados. Este factor representa a principal vantagem das abordagens em grelha, uma vez que garante

uma maior rapidez de processamento.

O STING (*STatistical INformation Grid*) [125] e o CLIQUE (*CLustering In QUEst*) [126] são exemplos de algoritmos de agrupamento baseados em grelha. O STING consegue descobrir alguns grupos com formas complexas em conjuntos de dados espaciais com ruído e valores isolados. Inicialmente é construída uma grelha através da divisão de cada atributo em intervalos de tamanho *TamAresta*, definido pelo utilizador. Em seguida, cada objecto é atribuído à célula que engloba os valores dos seus atributos. As células que não contêm objectos são descartadas. No final, os grupos de dados são formados pelas regiões densas da grelha. Este passo é bastante semelhante à forma com que o algoritmo DBSCAN descobre os seus grupos, contudo aplicado às grelhas. O CLIQUE por sua vez não pretende descobrir regiões densas que englobem todos os atributos de dados, mas antes correlações interessantes entre os objectos em sub-espaços de subconjuntos de atributos de dados.

### 3.4.5 Abordagens Baseadas em Modelos

Os algoritmos de agrupamento baseados em modelos têm como objectivo ajustar um modelo matemático ao conjunto de dados. Estes algoritmos subdividem-se em três abordagens: **agrupamento probabilístico**, **agrupamento conceptual** e **rede neuronal**.

**Agrupamento probabilístico**. No agrupamento probabilístico assume-se que os objectos de dados foram gerados a partir de um modelo de mistura, existindo uma distribuição de probabilidade associada a cada grupo de dados. Assim, após se assumir uma determinada distribuição para os grupos de dados, geralmente a distribuição gaussiana, o problema do agrupamento de dados resume-se à estimação dos parâmetros que definem a função de densidade de probabilidade para cada um dos grupos. O algoritmo EM (*Expectation-Maximization*) [127] é um algoritmo muito usado para o efeito.

**Agrupamento conceptual**. O agrupamento conceptual é uma forma de agrupamento que, dado um

conjunto de objectos não etiquetados, produz um esquema de classificação desses objectos não se baseando na distância entre eles. Ao contrário do agrupamento convencional, em que somente se identificam grupos de objectos idênticos, o agrupamento conceptual também encontra descrições para os atributos de cada grupo obtido. Assim, o agrupamento conceptual é um processo em duas fases: o agrupamento, em que são identificados grupos de objectos com base num ou mais critérios predefinidos e a caracterização, em que, é determinada uma descrição de cada um dos grupos obtidos na fase de agrupamento. O COBWEB [128] é um algoritmo simples e popular de agrupamento conceptual incremental. Os seus objectos de entrada são descritos por pares categóricos atributo-valor. O COBWEB cria uma estrutura hierárquica na forma de uma árvore de classificação com representação de conceitos (hierarquia de conceitos) e realiza uma procura heurística no espaço de possíveis árvores de classificação usando a "subida da colina".

**Rede Neuronal**. A abordagem de rede neuronal para o agrupamento tende a representar cada grupo como um exemplar (neurónio). Um exemplar actua como um protótipo do grupo e não tem necessariamente que corresponder a um exemplo de dados particular. Usando uma medida de distância, um novo objecto pode ser atribuído ao grupo cujo exemplar é mais semelhante. Um Mapa de Características Auto-Organizáveis (SOM – *Self Organizing Maps*) [129] consiste numa rede neuronal artificial, treinada de forma não supervisionada, que se organiza dinamicamente, respondendo aos estímulos de entrada, isto é, aos valores dos atributos dos objectos que são submetidos à rede. Esta acção é baseada nos mapas topológicos presentes no córtex cerebral, em que neurónios próximos no mapa devem responder por funções similares (específicas).

## *3.5 Agrupamento de Dados com Restrições*

O agrupamento de dados com restrições tem como objectivo utilizar o conhecimento sobre um determinado domínio na descoberta da estrutura do conjunto de dados. Esse conhecimento é representado na forma de restrições que expressam preferências, limitações e condições que o utilizador pretende impor. Com a inclusão de restrições é esperado que as soluções de agrupamento de dados se adequem da melhor forma a cada problema, uma vez que o conhecimento prévio representado através das restrições vai de encontro à resolução desse mesmo problema.

### 3.5.1 Tipos de Restrições

As restrições podem ser categorizadas pelo nível a que se encontram. Ao nível mais alto situam-se as **restrições globais** que se aplicam a todo o conjunto de dados. As **restrições ao nível dos grupos** e as **restrições ao nível dos atributos** encontram-se no nível intermédio de especificidade. Por fim, as **restrições ao nível dos objectos de dados** estão no nível mais baixo [91].

#### *3.5.1.1 Restrições Globais*

As restrições que se pretendem aplicar a um determinado conjunto de dados $X$ como um todo são designadas por restrições globais. Estas podem ter a forma de relações de vizinhança ou outro tipo de relações mais gerais entre os objectos de dados. Dos vários métodos para incorporar restrições globais são destacados o mapeamento de **obstáculos como restrições** e a **informação de vizinhança**.

**Obstáculos como Restrições**. Os algoritmos de agrupamento de dados com obstáculos efectuam o cálculo das distâncias entre os objectos do conjunto de dados, tendo em conta os obstáculos que estes têm de ultrapassar. Estes algoritmos são normalmente utilizados em conjuntos de dados onde uma distância euclidiana pequena entre os objectos pode não significar que estes pertençam ao mesmo grupo. Duas populações distintas (dois conjuntos de objectos próximos) separadas por um rio (obstáculo) são um exemplo de um conjunto de dados com este problema. O COE-CLARANS [130], baseado no algoritmo de agrupamento de dados CLARANS [131], constrói um grafo onde representa os objectos de dados e os obstáculos e utiliza técnicas de geometria computacional para calcular as distâncias mais curtas entre os objectos, tendo em conta que os obstáculos têm de ser contornados. Geralmente a quantidade de dados espaciais é bastante elevada, pelo que são realizados vários pré-processamentos e optimizações para reduzir o custo computacional do algoritmo.

**Informação de Vizinhança**. Os algoritmos de agrupamento de dados são por vezes aplicados a

conjuntos de dados em que os objectos de dados se encontram relacionados através de informação estrutural ou de vizinhança. Uma forma de incorporar este tipo de informação é o exemplo de uma imagem constituída por vários *pixels* organizados em posições bidimensionais [132]. A figura 25 (a) representa uma imagem com 25 *pixels* com três regiões distintas. As relações de vizinhança entre os *pixels* são representadas pelas ligações ilustradas na figura 25 (b). Finalmente, a figura 25 (c) exibe o agrupamento dos *pixels* em três grupos, considerando quer as relações de vizinhança, quer os atributos de cada *pixel* (neste caso, as posições e cor de cada *pixel*) .

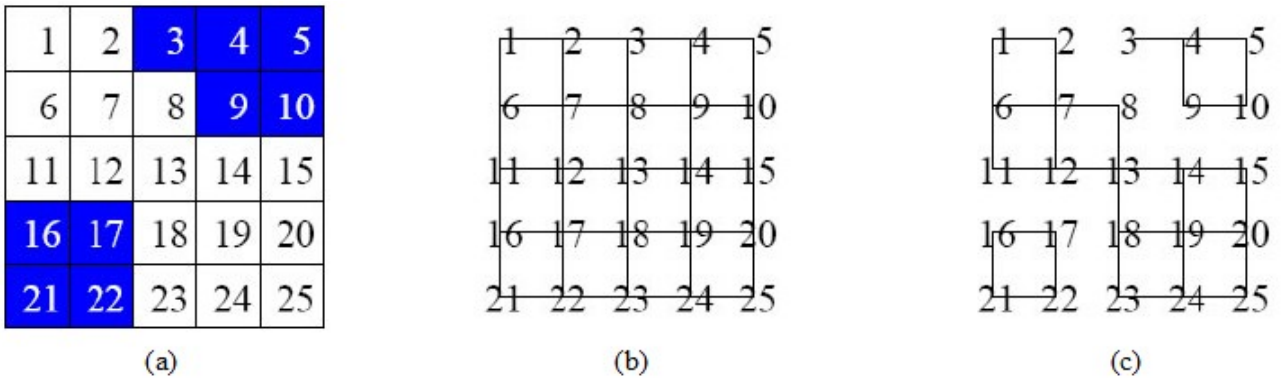



*Figura 25: Segmentação de imagem [91]*

### *3.5.1.2 Restrições ao Nível dos Grupos*

Por vezes, existe informação que se pretende aplicar aos grupos ($C_i$) de objectos de dados individualmente e não à totalidade do conjunto de dados $X$. Esta informação é representada usando restrições ao nível dos grupos e têm como objectivo restringir a forma, o tamanho, a variância ou outras características dos grupos. As restrições ao nível dos grupos mais frequentemente usadas são do tipo **capacidade mínima** e **capacidade máxima**.

**Restrições de Capacidade Mínima** . As restrições de capacidade mínima têm como objectivo limitar o número mínimo de objectos de dados que cada grupo pode conter. Este tipo de restrições é, por exemplo, utilizado na resolução de um problema do algoritmo de agrupamento de dados $K$-Médias, que consiste na obtenção de grupos vazios ou com um número reduzido de objectos. Para solucionar o problema, foi proposta [133] uma modificação ao algoritmo $K$-Médias que determina

que o número de objectos de dados em cada grupo não pode ser inferior a um valor especificado pelo utilizador (restrição de capacidade mínima).

**Restrições de Capacidade Máxima**. Pode também ser útil em alguns problemas definir a capacidade máxima, isto é, o número máximo de objectos em cada grupo de dados. As restrições de capacidade máxima são utilizadas frequentemente em problemas em que existe uma disponibilidade limitada de recursos, como na análise de localizações para infraestruturas [134].

#### *3.5.1.3 Restrições ao Nível dos Atributos*

Em determinados problemas, o utilizador pode pretender influenciar o agrupamento de dados mediante o valor de um atributo. Pode pretender-se, por exemplo, incluir no mesmo grupo todos os objectos de dados que possuam o mesmo valor para um determinado atributo. Uma das formas de o fazer consiste na conversão desta restrição ao nível dos atributos para várias restrições de relações entre pares de objectos de dados. As relações entre pares de objectos de dados são abordadas em seguida na secção 3.5.1.4.

#### *3.5.1.4 Restrições ao Nível dos Objectos*

As restrições ao nível dos objectos encontram-se ao nível mais específico do conhecimento de domínio. A informação relativa aos objectos de dados é determinante para definir se os objectos pertencem (ou não) a determinado grupo. Esta acção é efectuada através da imposição de relações entre pares de objectos. As restrições ao nível dos objectos de dados podem assumir várias formas: **rotulação parcial**, **restrições entre pares de objectos de dados** e **interactividade com o utilizador**.

**Rotulação Parcial**. Devido ao custo financeiro e humano associado à obtenção de objectos de dados rotulados, é frequente conseguir-se apenas subconjuntos deste tipo de objectos. Estes objectos

representam conhecimento prévio de um determinado domínio podendo, como tal, ser utilizados no agrupamento de dados. Duas abordagens de rotulação parcial são a votação maioritária com base nos objectos rotulados e grupos semeados com base em rótulos. Na primeira, os rótulos existentes em alguns dos objectos de dados restringem a colocação dos objectos de dados nos grupos com base numa medida de impureza [135]. Os grupos são rotulados de acordo com a maioria dos objectos rotulados que possuem. Nos grupos semeados com base em rótulos, os objectos rotulados podem também servir para determinar de forma inteligente os grupos iniciais. Basu et al. [136] propôs uma alteração ao algoritmo $K$-Médias no qual os $K$ grupos iniciais tivessem apenas objectos com o mesmo rótulo, efectuando uma selecção inteligente dos centros iniciais na inicialização do algoritmo, de forma a influenciar positivamente o agrupamento de dados.

**Relações entre Pares de Objectos de Dados**. A definição de um conjunto de relações entre pares de objectos de dados é a forma mais generalista e flexível para representar o conhecimento de domínio usando informação relacional [132]. Com este tipo de representação é possível codificar várias das outras formas de conhecimento de domínio apresentadas anteriormente. Tanto as restrições globais, que definem relações de vizinhança, como as restrições ao nível dos atributos, guiadas através de heurísticas definidas pelo utilizador, podem ser facilmente transformadas em relações entre pares de objectos. Apenas as restrições ao nível dos grupos podem ser dificilmente transformadas em relações entre pares de objectos, já que se focam principalmente nas capacidades mínimas e máximas de cada um dos grupos do agrupamento de dados a formar. Nas relações entre pares de objectos de dados, as restrições mais frequentemente utilizadas são do tipo ligação obrigatória ou ligação proibida. Uma ligação obrigatória, $(x_i, x_j) \in Rest_{=}$, indica que dois objectos de dados $x_i$ e $x_j$ devem ser agrupados no mesmo grupo, enquanto que uma ligação proibida, $(x_i, x_j) \in Rest_{\neq}$, indica que os dois objectos de dados $x_i$ e $x_j$ não devem ser agrupados conjuntamente.

**Interactividade com o Utilizador**. Este tipo de restrições é obtido através da interacção de um sistema de agrupamento de dados com o utilizador de forma iterativa. Em cada iteração, o sistema de agrupamento de dados produz uma solução de agrupamento e apresenta-a ao utilizador. Este avalia a solução apresentada e indica os erros que o sistema produziu, para que, nas iterações seguintes essa informação seja utilizada no sentido de evitar erros semelhantes. Este processo

encontra-se descrito com maior pormenor na secção 3.5.2.

### 3.5.2 Aquisição de Restrições

A aprendizagem activa para a aquisição de restrições tem como objectivo auxiliar o utilizador na definição de restrições sobre o conjunto de dados em análise. Desta forma, pretende-se que as restrições adquiridas sejam o mais informativas possíveis e que os algoritmos de agrupamento com restrições aproveitem essas restrições para obter partições de dados com qualidade superior.

**Explorar e Consolidar**. Esta abordagem consiste em dois passos para a aquisição de relações de ligação obrigatória e proibida: Explorar e Consolidar [137]:

1. Explorar. Neste passo o utilizador é consultado iterativamente para decidir se dois objectos devem (ou não) ser atribuídos ao mesmo grupo. O intuito é o de encontrar conjuntos de vizinhança entre os pares de objectos, em que cada conjunto de vizinhança pertence a um grupo *natural* do conjunto de dados;
2. Consolidar. Entre estas consultas, o conjunto de dados é pesquisado com o intuito de explorar a relação de vizinhança entre os pares de objectos para expandir o conjunto de restrições. A partir das consultas efectuadas ao utilizador pretende-se que os centros iniciais dos grupos do conjunto de dados sejam estimados o melhor possível.

**Interacção com o Utilizador**. O utilizador fornece iterativamente *feedback* ao algoritmo de agrupamento de dados que é incorporado sob a forma de restrições que o algoritmo de dados tenta satisfazer nas iterações seguintes [138]. O utilizador poderá assim, através da aplicação de restrições, guiar o agrupamento de dados de forma a torná-lo mais adequado ao seu objectivo. Consiste essencialmente em três etapas:

1. Agrupar o conjunto de dados com um algoritmo de agrupamento não supervisionado.
2. Analisar o resultado do agrupamento, em que o utilizador verifica em apenas alguns dos grupos se existem objectos mal agrupados. O utilizador dá informações ao sistema numa das

seguinte formas:

- o objecto não pertence ao grupo a que foi atribuído;
- o objecto deve ser movido para um determinado grupo;
- dois determinados objectos devem ser atribuídos ao mesmo grupo;
- e dois determinados objectos não podem ser atribuídos ao mesmo grupo.

3. Após a solução ter sido criticada, a medida de distância do algoritmo de agrupamento é modificada para que as restrições impostas pelo utilizador sejam satisfeitas.

Este processo repete-se enquanto o utilizador não estiver satisfeito com a solução apresentada.

### 3.5.3 Algoritmos

Com o evoluir dos algoritmos de agrupamento de dados, têm sido propostas bastantes abordagens com a capacidade de incorporar restrições. Na sua maioria, estas restrições são definidas ao nível dos objectos de dados, isto é, através de restrições entre pares de objectos. Este tipo de algoritmos divide-se em cinco categorias: **restrições invioláveis**, **restrições na forma de rótulos**, **penalização na violação de restrições**, **edição de distância** e **modificação do processo de geração** [91].

#### *3.5.3.1 Restrições Invioláveis*

O objectivo deste tipo de algoritmos de agrupamento de dados é garantir que todas as restrições especificadas pelo utilizador são satisfeitas. Dois algoritmos deste tipo são o COP-COBWEB, aplicado a conjuntos de dados com atributos categóricos, e o COP-$K$-Médias, que trata apenas conjuntos de dados com atributos numéricos.

**COP-COBWEB**. O algoritmo de agrupamento de dados com restrições COP-COBWEB [139] (*Constraint-Partitioning COBWEB*), baseado no COBWEB [128], emprega o conceito de *utilidade categórica* para produzir um agrupamento de dados que maximiza a dissimilaridade entre os grupos

e similaridade dos objectos pertencentes ao mesmo grupo. Existem quatro operações básicas que o COBWEB aplica na construção da árvore de classificação, nomeadamente *Adicionar*, *Novo*, *Fundir* e *Dividir*. A operação seleccionada é a cujo agrupamento resultante obtiver maior *utilidade categórica* [151]. O COP-COBWEB retorna como agrupamento de dados, o nó de topo da hierarquia produzida pelo COBWEB que satisfaz todas as restrições impostas pelo utilizador. Se existir alguma restrição de ligação obrigatória que indique que algum objecto $x_i$ tem de ser associado ao mesmo grupo que outro objecto $x_j$, já existente num grupo $C_k$ do agrupamento de dados corrente, o objecto $x_i$ é incluído no grupo $C_k$. Caso contrário, os operadores *Novo*, *Adicionar* e *Fundir* são aplicados para se determinar em que grupo $x_i$ vai ser incluído. As restrições do tipo ligação proibida são verificadas nos passos *Adicionar* e *Fundir*. Quando se considera incluir um objecto $x_i$ num grupo $C_k$, verifica-se se algum dos objectos $x_j$ pertencentes a $C_k$ tem uma relação de ligação proibida com $x_i$ e, se existir, $x_i$ não poderá ser incluído em $C_k$. O operador *Dividir* é sempre avaliado, quer tenha sido verificada a existência ou não alguma restrição. Este operador aplica recursivamente o COP-COBWEB a um subconjunto de dados que corresponde ao melhor grupo em que $x_i$ foi colocado, segundo a utilidade categórica. Finalmente, o agrupamento de dados resultante é aquele em que a utilidade categórica é maximizada.

**COP-*K*-Médias**. O algoritmo COP-$K$-Médias [140] (*COnstraint-Partitioning K-Means*) é uma modificação do $K$-Médias [114] para que este possa suportar restrições de ligação obrigatória e de ligação proibida entre pares de objectos de dados. A alteração ao algoritmo de agrupamento de dados $K$-médias incide na fase de atribuição de cada objecto de dados $x_i$ ao grupo $C_k$ mais próximo. Antes de se atribuir $x_i$ ao grupo mais próximo é realizado um passo de verificação de violações das restrições. Desta forma, o objecto $x_i$ vai ser incluído no grupo mais próximo em que todas as restrições sejam satisfeitas. Caso não exista nenhum grupo que satisfaça todas as restrições que envolvem $x_i$, o algoritmo aborta e retorna conjunto vazio.

#### *3.5.3.2 Restrições na Forma de Rótulos*

Nos algoritmos de agrupamento cujas restrições são expressas através de rótulos, são conhecidos os

grupos a que pertencem alguns dos objectos de dados e os algoritmos de agrupamento de dados com restrições usam essa informação para aumentar a qualidade de agrupamento de dados. Basu et al. [136] exploram esta abordagem propondo duas variantes do algoritmo de agrupamento de dados $K$-Médias: o $K$-Médias semeado (*Seeded $K$-Means*) e o $K$-Médias restringido (*Constrained $K$-Means*).

Considere-se $X=\{x_1;\ldots;x_n\}$ um conjunto de $n$ objectos de dados e $S\subseteq X$ o conjunto de sementes que é composto por todos os objectos rotulados. É também assumido que $S$ está dividido em $K$ partições, $\{S_1;\ldots;S_K\}$, tal que exista um subconjunto $S_j$, com pelo menos um objecto de dados, para cada um dos $K$ grupos naturais. No $K$-Médias semeado, o conjunto de sementes é apenas utilizado para definir os centros dos grupos iniciais do $K$-Médias, em vez de estes serem definidos aleatoriamente. Nesse sentido, o centro $\overline{x_j}$ do *j*-ésimo grupo inicial, $C_j$, é obtido pela média dos objectos que constituem o *j*-ésimo subconjunto de sementes, $S_j$. O resto do processo é igual ao $K$-Médias o que não garante que restrições impostas pelos objectos rotulados sejam totalmente satisfeitas.

No algoritmo $K$-Médias restringido, o conjunto de sementes $S$ é utilizado para inicializar os centros dos grupos, da mesma forma que no algoritmo $K$-Médias semeado. No entanto, os rótulos dos objectos sementes nunca são alterados nos passos seguintes do algoritmo, isto é, apenas serão alterados rótulos de objectos $x_i\notin S$. Como no algoritmo $K$-Médias restringido o rótulo de cada objecto semente é imutável, este deve ser apenas usado quando o conjunto de sementes não tem ruído, isto é, quando todos os rótulos atribuídos dos objectos sementes se encontram correctamente atribuídos. Caso contrário, o algoritmo $K$-Médias semeado é uma melhor opção, já que torna possível que um objecto semente troque de grupo no decorrer do processo de agrupamento.

#### *3.5.3.3 Penalização na Violação de Restrições*

A ideia principal deste tipo de algoritmos consiste na definição de uma função-objectivo que, para além de considerar as distâncias entre os objectos e respectivos centros de grupos, penaliza a violação de restrições. Dois algoritmos de agrupamento de dados representativos desta categoria são

o PC $K$-Médias e o CVQE.

**PC $K$-Médias**. O algoritmo de agrupamento de dados PC $K$-Médias (*Pairwise Constrained K-Means*) [137], considera apenas relações entre pares de objectos, mais precisamente, ligações obrigatórias e ligações proibidas, e respectivos custos de violação. Esta lógica é aplicada através de uma modificação do algoritmo $K$-Médias no passo da atribuição dos objectos de dados aos grupos. Como já referido na secção 3.5.3.2, uma inicialização apropriada do $K$-Médias, atendendo às restrições impostas pelo utilizador, aumenta o desempenho do agrupamento de dados. No passo de inicialização do PC $K$-Médias, o conjunto de restrições de ligação obrigatória, $Rest_{=}$, é expandido através de relações de transitividade entre os objectos de $Rest_{=}$ e os restantes objectos. O mesmo acontece para as restrições de ligação proibida $Rest_{\neq}$. Após este pré-processamento, são definidos os centros iniciais dos grupos tendo em conta os conjuntos de restrições gerados. O algoritmo segue então um processo iterativo em que, inicialmente, os objectos de dados são atribuídos a um grupo e, posteriormente, os centros de grupo são actualizados. No passo de atribuição dos objectos aos grupos, cada objecto de dados $x_i$ é atribuído ao grupo que minimiza a soma da distância de $x_i$ ao centro do grupo com um custo de violação de restrições imposta por essa atribuição. Este algoritmo não é determinístico, uma vez que os subconjuntos de objectos de $Rest_{=}$ e $Rest_{\neq}$ existentes em cada grupo podem variar com a ordem de atribuição dos objectos aos grupos. Desta forma, poderão ser produzidos agrupamentos distintos, mesmo para conjuntos de dados iniciais iguais.

**CQVE**. Outro algoritmo de agrupamento baseado na modificação da função-objectivo do algoritmo $K$-Médias é o CVQE (*Constrained Vector Quantization Error*) [146]. A função-objectivo do algoritmo $K$-Médias é equivalente ao *erro de quantificação vectorial*, ou VQE (*Vector Quantization Error*), pois tenta minimizar iterativamente o erro de quantificação vectorial, também denominado por distorção. O VQE é definido pelas seguintes equações:

$$VQE = \sum_{j=1}^{K} VQE_j$$

$$VQE_j = \frac{1}{2} \sum_{x_i \in C_j} d(\overline{x_j}, x_i)^2$$

O CVQE incorpora nesta função-objectivo as restrições do tipo ligação obrigatória e ligação

proibida. No primeiro passo deste algoritmo, os objectos não restringidos são associados ao grupo mais próximo, analogamente ao algoritmo $K$-Médias. Para os restantes objectos, para cada par de objectos envolvido numa restrição, são testadas todas as possibilidades de atribuição desses objectos a grupos e é escolhida a que menos aumenta o valor da função-objectivo. O segundo passo do algoritmo de agrupamento CVQE consiste na actualização dos centros dos grupos, com intuito de minimizar o erro de quantificação vectorial restringido.

#### *3.5.3.4 Edição de Distância*

Outra abordagem para a integração de restrições no agrupamento de dados consiste na edição da medida de distância. Os algoritmos de agrupamento desta categoria, para além de tentarem satisfazer as restrições impostas, tentam generalizar essas restrições ao nível do espaço dos atributos de dados. A figura 26 exemplifica a ideia na qual se baseia a edição de distância. A figura 26 a) representa um conjunto de dados que se pretende agrupar em dois grupos, existindo duas restrições de ligação obrigatórias. Uma possível solução é apresentada na figura 26 b). Como se pode verificar, as restrições impostas foram cumpridas na sua totalidade, apesar do agrupamento de dados apresentado não ser muito intuitivo. Por outro lado, a figura 26 c) apresenta também uma solução que satisfaz todas as restrições, mas que para além de apenas satisfazer as restrições impostas, generaliza essas restrições tendo em conta o seguinte raciocínio: se dois objectos de dados devem ser agrupados conjuntamente por estarem relacionados com uma ligação obrigatória, então os objectos de dados próximos destes também devem ser incluídos no mesmo grupo.

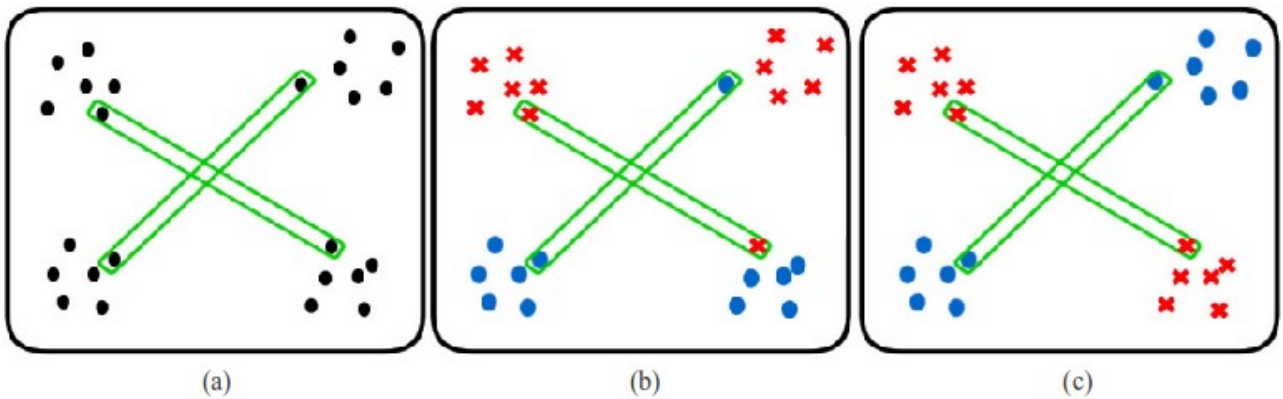



*Figura 26: Generalização ao nível do espaço de ligações obrigatórias [91]*

Dois algoritmos que exploram o conceito de edição de distância são o algoritmo de agrupamento hierárquico CCL e o algoritmo de agrupamento de partição de dados MPC $K$-Médias.

**CCL**. O algoritmo de Ligação Completa Restringido, ou CCL (*Constrained Complete-Link*) [147], expande o bem conhecido algoritmo hierárquico aglomerativo de Ligação Completa, ou CL (*Complete-Link*) [119], permitindo a especificação de ligações obrigatórias e proibidas entre pares de objectos de dados. A ideia deste algoritmo consiste na distorção do espaço de similaridade, aproximando os objectos de dados que se sabe que pertencem ao mesmo grupo e afastando objectos de dados que pertencem a grupos diferentes. Dada uma matriz de dissimilaridades, que represente as dissimilaridades entre os pares de objectos do conjunto de dados, e os conjuntos de relações obrigatórias $Rest_{=}$ e proibidas $Rest_{\neq}$, é criada uma nova matriz de dissimilaridades em que as distâncias entre objectos de dados são alteradas, reflectindo as restrições $Rest_{=}$ e $Rest_{\neq}$ e as suas implicações nos objectos do conjunto de dados. Inicialmente, são impostas as ligações obrigatórias entre pares de objectos na matriz de dissimilaridades $D \in \mathbb{R}^{n \times n}$, atribuindo o valor 0 a cada entrada na matriz entre pares de objectos com ligações obrigatórias, isto é, $\forall (x_i, x_j) \in Rest_{=}, D_{ij}, D_{ji} = 0$. A propagação das ligações obrigatórias ao resto do conjunto de dados é então efectuada através de uma versão modificada do algoritmo de Floyd-Warshall [152], que calcula os caminhos mais próximos entre todos os pares de objectos do conjunto de dados. Em seguida impõem-se as ligações proibidas atribuindo o valor $\infty$ às entradas na matriz $D$ correspondentes aos objectos de dados com ligações proibidas, ou seja, $\forall (x_i, x_j) \in Rest_{\neq}, D_{ij}, D_{ji} = \infty$. Note-se que não é necessário propagar as restrições de ligação proibida, já que, a propagação será realizada implicitamente na aplicação do algoritmo de Ligação Completa. Em seguida, é aplicado o algoritmo Ligação Completa para se obter o agrupamento do conjunto de dados.

**MPC $K$-Médias**. O algoritmo de agrupamento de dados MPC $K$-Médias [148] para além de penalizar violações de ligações obrigatórias e/ou proibidas entre pares de objectos, efectua também aprendizagem da medida de distância, estendendo assim o algoritmo PC $K$-Médias apresentado na secção 3.5.3.3.

A medida de distância euclidiana pode ser parametrizada através de uma matriz de ponderações

$A \in \mathbb{R}^{d \times d}$ simétrica e positiva da seguinte forma:

$$\left\| x_i - x_j \right\|_A = \sqrt{(x_i - \overline{x_{l_i}})^T A (x_j - \overline{x_{l_j}})}$$

em que $x_i$ é o vector de atributos de um objecto de dados e $\overline{x_{l_i}}$ o vector médio do grupo a que pertence. A inicialização dos centros dos grupos é realizada tal como definido no algoritmo PC $K$-Médias. O MPC $K$-Médias itera entre a atribuição dos objectos de dados a grupos, a estimação dos novos centros dos grupos e a aprendizagem da parametrização da medida de distância.

No passo de atribuição dos objectos a grupos, cada objecto $x_i$ é atribuído a um grupo de forma a minimizar a soma das distâncias de $x_i$ ao respectivo centro e o custo das violações de restrições. Geralmente, as ponderações $W_=$ e $W_\neq$ são uniformes, pelo que a violação das restrições são sempre tratadas de igual forma. No entanto, o custo de penalização de uma ligação obrigatória deve ser mais elevado se os dois objectos de dados se encontrarem próximos do que se os dois objectos se encontrarem afastados. O custo de penalização das ligações proibidas deve também ser elevado se a distância entre os objectos de dados for elevada e reduzido se essa distância for escassa. Para considerar esta intuição, as ponderações $W_=$ e $W_\neq$ são multiplicadas por funções de penalização, $f_=(x_i, x_j)$ e $f_\neq(x_i, x_j)$ respectivamente, definidas por:

$$f_=(x_i, x_j) = max(\alpha_{min}, \alpha_{max} - \left\| x_i - x_j \right\|_A^2)$$

$$f_\neq(x_i, x_j) = max(\alpha_{min} + \left\| x_i - x_j \right\|_A^2, \alpha_{max})$$

em que $\alpha_{min}$ e $\alpha_{max}$ são constantes não negativas que correspondem, respectivamente, aos valores mínimo e máximo que a penalização pode tomar.

No passo de estimação dos novos centros de grupos $\overline{x_k}$, as restrições $Rest_=$ e $Rest_\neq$ não são consideradas, sendo cada centro de grupo actualizado com o vector médio dos objectos que lhe foram atribuídos:

$$\overline{x_k} \leftarrow \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

Finalmente, no passo de aprendizagem da medida de distância, pelo facto de ser difícil realizar a

aprendizagem da matriz completa $A$, é apenas realizada a aprendizagem da diagonal de $A$, o que é equivalente a aprender a medida de distância através da ponderação dos atributos. Assim, o $m$-ésimo elemento da diagonal de $A$, $a_{mm}$, corresponde à ponderação do $m$-ésimo atributo de dados e é actualizado da seguinte forma:

$$\begin{aligned} a_{mm} = & \Bigg( \sum_{x_i \in X} (x_{im} - \overline{x_{l_i}})(x_{im} - \overline{x}_{l_i m})^T \\ & - \sum_{(x_{im}, x_{jm}) \in Res^*_{=}} w_{=_{ij}} (x_{im} - x_{jm})(x_{im} - x_{jm})^T I(l_i \neq l_j) \\ & + \sum_{(x_{im}, x_{jm}) \in Res^*_{\neq}} w_{\neq_{ij}} (x_{im} - x_{jm})(x_{im} - x_{jm})^T I(l_i = l_j) \Bigg)^{-1} \end{aligned}$$

em que $Res^*_{=}$ e $Res^*_{\neq}$ são subconjuntos de $Res_{=}$ e $Res_{\neq}$ que excluem os pares de objectos cujas funções de penalização $f_{=}$ e $f_{\neq}$ tomam os valores $\alpha_{min}$ e $\alpha_{max}$ respectivamente.

#### *3.5.3.5 Modificação do Processo de Geração*

Os algoritmos de agrupamento apresentados nesta secção assumem que os dados são gerados segundo um modelo probabilístico, sendo o objectivo dos algoritmos estimar os parâmetros desse modelo, considerando tanto os atributos de dados como as restrições existentes. Dois algoritmos que seguem este modelo são o PPC e o HMRF $K$-Médias.

**PPC**. O Agrupamento Probabilístico de Dados com Penalização, ou PPC (*Penalized Probabilistic Clustering*) [149], propõe um modelo de mistura Gaussiana para realizar o agrupamento de dados em que as preferências do utilizador são incorporadas na forma de restrições difusas entre pares de objectos. As preferências são representadas através da probabilidade Bayesiana de pares de objectos deverem, ou não, ser atribuídos ao mesmo grupo. Após ser treinada com o algoritmo EM [127], a informação expressa *a priori* no agrupamento inicial de dados é codificada com sucesso nas componentes da mistura Gaussiana. Assim, o modelo é generalizado de uma forma consistente com o conhecimento prévio.

**HMRF $K$-Médias**. Basu et al. propuseram [150] um algoritmo de agrupamento de dados em que

são consideradas, simultaneamente, relações de ligação obrigatória e proibida entre pares de objectos de dados e a aprendizagem de uma medida de distância. O algoritmo proposto, denominado HMRF $K$-Médias, tem como objectivo a minimização de uma função-objectivo derivada do modelo HMRF (*Hidden Markov Random Fields*). Esta abordagem aumenta o desempenho do agrupamento de dados não supervisionado em três aspectos, generalizando o algoritmo MPC $K$-Médias apresentado na secção 3.5.3.4:

- Inicialização melhorada – os centros iniciais dos grupos são obtidos com base em conjuntos de vizinhança de objectos induzidos pelas restrições;
- Atribuição dos objectos de dados aos grupos atendendo a restrições – os objectos são atribuídos aos grupos atendendo não só à minimização de uma medida de distorção, mas também minimizando o número de violações de restrições;
- Aprendizagem iterativa da medida de distância – a medida de distorção é actualizada durante o processo de agrupamento de dados, modificando o espaço para que as restrições sejam satisfeitas.

## *3.6 Sumário*

Neste capítulo introduziram-se alguns conceitos da aprendizagem automática, como os diferentes tipos de aprendizagem, fases e tipos de um agrupamento de dados, com o objectivo de enquadrar o agrupamento de dados e, mais especificamente, o agrupamento de dados com restrições. Foi apresentada uma visão geral dos vários tipos de restrições que podem ser incluídas no agrupamento de dados. O uso de restrições permite que o utilizador indique preferências, limitações e conhecimento de domínio no agrupamento de dados para que a solução obtida seja mais útil e vantajosa para os seus objectivos. São ainda apresentados algoritmos de agrupamento de dados representativos dos vários tipos de abordagens existentes e com as respectivas especificidades, tais como, a impossibilidade de violação de restrições, a utilização de um subconjunto de objectos rotulados para inicializar os centros de algoritmos de agrupamento de partição, a modificação da função-objectivo de forma a penalizar a violação de restrições, a aprendizagem da medida de distância e a modificação do processo de geração de dados em modelos probabilísticos.

# 4 Agrupamento de Dados Visual Interactivo

## *4.1 Introdução*

O Agrupamento de Dados Visual Interactivo (ou IVC – Interactive Visual Clustering) é uma abordagem inovadora que permite ao utilizador explorar conjuntos de dados relacionais de forma interactiva, com o intuito de produzir um agrupamento que satisfaça os seus objectivos. Para isso, é utilizada uma combinação de grafos aplicando forças físicas de "molas" e do agrupamento de dados com restrições com interacção com o utilizador. Resultados experimentais em conjuntos de dados reais e sintéticos demonstram que o agrupamento visual de dados interactivo tem melhor desempenho do que abordagens alternativas.

## *4.2 Motivação*

O objectivo deste estudo consiste no desenvolvimento de métodos interactivos de agrupamento de dados, que permitam a um utilizador particionar um conjunto de dados em grupos adequados às suas tarefas e interesses. No agrupamento de dados tradicional, pretendem-se dividir os dados em grupos que possuam uma elevada similaridade intra-grupo e uma reduzida similaridade inter-grupo. O resultado será um agrupamento de dados dependente da similaridade métrica usada no agrupamento de dados, da função-objectivo do algoritmo, do método de pesquisa e eventualmente dos parâmetros de pesquisa. Contudo, na prática, os “melhores” agrupamentos podem também depender dos interesses e objectivos do utilizador. Por exemplo, um conjunto de dados com informação de automóveis fabricados por uma empresa poderia servir para propósitos distintos. Por um lado, um responsável comercial poderia estar interessado em atributos como o custo de fabrico e o preço de venda dos automóveis, de forma a perceber quais os grupos de maior rentabilidade. Por outro lado, um responsável de *marketing* da mesma empresa poderia ter interesse em conhecer as características técnicas dos automóveis, de forma a dividi-los em segmentos para diferentes públicos-alvo. Neste exemplo, apesar do conjunto de dados ser o mesmo, os grupos finais esperados pelos dois utilizadores são diferentes. Um agrupamento de dados automático poderia encontrar um

destes agrupamentos, mas nunca ambos.

A investigação recente na área do agrupamento de dados com restrições endereça a problemática da obtenção de diferentes agrupamentos finais a partir do mesmo conjunto de dados, utilizando informação adicional fornecida pelo utilizador. O Agrupamento de Dados com Restrições (secção 3.5) baseia-se no princípio de que, apesar dos utilizadores poderem não ser capazes de definir explicitamente os critérios do agrupamento desejado, é normalmente possível que estes forneçam conhecimento parcial acerca da natureza dos agrupamentos pretendidos. Esta informação adicional é fornecida tipicamente sob a forma de restrições ao nível dos objectos de dados (secção 3.5.1.4) que é então utilizada para influenciar o agrupamento para a solução desejada.

Idealmente, o utilizador deveria fornecer algumas restrições para "semear" os grupos iniciais e depois ir adicionando restrições conforme necessário de forma a ajustar e melhorar os agrupamentos resultantes. A dificuldade desta acção deve-se ao facto de nem sempre ser fácil a interpretação dos grupos, particularmente em domínios de dados de elevada dimensionalidade, em que é difícil visualizar os grupos com clareza. Como resultado, o utilizador poderá não ser capaz de informar que restrições adicionais seriam mais úteis para melhorar o agrupamento de dados.

Em alguns domínios, poderá existir informação relacional além das restrições entre pares de objectos. Por exemplo, podem existir informações como estudantes pertencentes às mesmas turmas, informação de grupos em redes sociais ou artigos científicos que citem as mesmas fontes. Este tipo de informação relacional, que pode ser representada por arcos (ou ligações relacionais) num esquema em grafo, fornece informação adicional acerca da similaridade dos dados. Contudo, estas relações são normalmente mais fracas do que as restrições entre pares de objectos na medida em que não implicam rigorosamente que os objectos tenham (ou não) de pertencer ao mesmo grupo, embora indiquem uma maior correlação entre as instâncias interligadas. A maioria dos algoritmos de agrupamento de dados considera as informações dos atributos das instâncias ou a informação relacional entre as instâncias, mas raramente ambas.

A abordagem apresentada nesta dissertação tem como objectivo endereçar esta problemática, ao

permitir que um utilizador explore interactivamente conjuntos de dados relacionais de grandes dimensões, de forma a produzir um agrupamento que satisfaça os seus objectivos. Este objectivo é alcançado através da combinação das seguintes técnicas:

- Visualização de Informação
- Agrupamento de Dados com Restrições
- Interacção com o Utilizador

Na abordagem de Agrupamento de Dados Visual Interactivo (IVC), os dados relacionais são inicialmente apresentados num grafo sob o qual actuam forças direccionadas que têm como intuito favorecer as relações entre as instâncias. De seguida, o utilizador pode mover instâncias de forma a criar os grupos iniciais. É então aplicado um algoritmo de agrupamento de dados com restrições que gera grupos que combinam a informação dos atributos com as restrições inferidas pela acção do utilizador. Com base nos grupos resultantes são criadas *ligações de agrupamento* entre os objectos e os respectivos grupos. As ligações de agrupamento asseguram que objectos classificados no mesmo grupo são próximos entre si. Além destas, são também adicionadas as *ligações relacionais* entre as instâncias. Esta estrutura relacional deverá influenciar objectos menos bem classificados a aproximarem-se dos grupos correctos. Com base na nova visualização (distribuição dos objectos), o utilizador poderá identificar instâncias que se encontram "deslocadas" e mover essas instâncias para os grupos correctos.

No capítulo 5 é demonstrado experimentalmente, utilizando conjuntos de dados fictícios e reais, que o IVC converge para o agrupamento-alvo significativamente mais rápido que a execução isolada do grafo aplicando as forças direccionadas *spring embedded* ou do agrupamento de dados.

## *4.3 Trabalho Prévio*

Este trabalho combina e estende técnicas de visualização de informação com métodos de agrupamento de dados com restrições e a interacção com o utilizador. Na visualização, é aplicado o

paradigma de desenho de grafos de forças direccionadas [164]. No agrupamento de dados com restrições, é utilizado o algoritmo MPC$K$-Médias [148].

As abordagens de forças direccionadas encontram-se entre as técnicas de desenho de grafos mais populares [153]. O tema do desenho de grafos encontra-se referido com maior pormenor na secção 2.5.3. Neste trabalho é utilizado um tipo de desenho de forças direccionadas denominado *spring embedding* (molas embutidas) [154]. No *spring embedding*, os nós de um grafo interagem entre si com dois tipos de forças simuladas, modeladas em processos físicos. A primeira é uma força de repulsão entre os nós emitida por cada nó, simulando uma força gravitacional inversa. A segunda força corresponde à *força de mola* (*spring*) que atrai (ou repulsa, se os nós se tornarem muito próximos) os nós adjacentes. O desenho *spring embedded* é determinado iterativamente pelo cálculo da soma de todas as forças envolvidas em cada nó, movendo os nós incrementalmente na direcção da força resultante. Este processo é repetido até que seja encontrado um equilíbrio entre as forças.

O agrupamento de dados com restrições permite que seja fornecida informação adicional relativamente à natureza dos grupos, tipicamente na forma de restrições entre pares de objectos, indicando que dois pontos devem pertencer ao mesmo grupo (ligações obrigatórias) ou devem pertencer a grupos distintos (ligações proibidas). O algoritmo de agrupamento de dados mais popular é o $K$-Médias [114] que associa iterativamente objectos de dados ao centro de grupo mais próximo e recalcula os centróides até que se verifique um agrupamento estável. Na abordagem PC$K$-Médias [137] se um agrupamento resultante da aplicação do $K$-Médias violar as restrições de ligação obrigatória ou proibida, é aplicada uma penalização à solução. Esta penalização é incorporada na função objectivo do $K$-Médias, de forma a que o PC$K$-Médias procure de forma eficiente o agrupamento que maximize a coerência dos grupos e ao mesmo tempo minimizando as penalizações por violação de restrições. O MPC$K$-Médias [148] estende o PC$K$-Médias adicionando-lhe a capacidade de efectuar aprendizagem da medida de distância utilizando um conceito de custo de penalização de restrições. O peso de penalizações de ligações obrigatórias é elevado entre objectos próximos e reduzido entre objectos afastados. Nas ligações proibidas acontece o contrário, sendo o custo das penalizações elevado entre objectos distantes e reduzido entre objectos de dados próximos. Algumas técnicas de agrupamento de dados com restrições

encontram-se descritas em pormenor na secção 3.5.3.

Neste trabalho foi desenvolvido o método de Agrupamento de Dados Visual Interactivo com recurso a ferramentas existentes que implementam as técnicas de visualização de informação, agrupamento de dados e interacção com o utilizador estudadas e apresentadas nesta dissertação. Nas componentes de visualização de informação e interacção com o utilizador foi utilizada a *toolkit* de visualização de informação Prefuse [155]. Para o agrupamento de dados foi utilizada a *framework* WEKA [159], especificamente uma extensão para suportar aprendizagem semi-supervisionada desenvolvida pelos autores do algoritmo MPC$K$-Médias: WEKAUT [160]. Estas ferramentas são descritas sumariamente de seguida.

### 4.3.1 Prefuse

A plataforma Prefuse consiste num conjunto de ferramentas gráficas e de interacção que tem como objectivo possibilitar o desenvolvimento flexível de visualizações de informação, sofisticadas e altamente interactivas. É desenvolvida em tecnologia Java sob uma licença *open source* (licença BSD). A arquitectura do Prefuse foi definida com base no conceito da *pipeline* de visualização, uma recomendação de como implementar uma visualização de informação, descrita na secção 2.4. O *toolkit* foi desenvolvido segundo um modelo extensível que permite o desenvolvimento de componentes "à medida", consoante as necessidades da aplicação desenvolvida.

O Prefuse é utilizado tanto para o desenho de visualizações de informação inter-relacionada, gravada numa estrutura de grafo ou árvore, ou informação não relacionada, armazenada em tabelas de dados. O desenho de itens visuais (forma, cor, posição, ...) é efectuado por um componente específico, *renderer*, que possui acesso ao item propriamente dito, assim como ao contexto da biblioteca *Graphics2D* Java da visualização. Esta abordagem possibilita a utilização dos métodos de desenho Java directamente através do item visual do Prefuse. Entre outros, os itens visuais do Prefuse são responsáveis por gerir as propriedades visuais (secção 2.4.2) dos elementos da visualização.

A plataforma Prefuse fornece também um conjunto de algoritmos de desenho (*layout* – secção 2.3) que definem a forma como os elementos visuais se apresentam e evoluem na visualização. Entre os desenhos mais interessantes destacam-se a visualização e interacção de grafos e árvores, bem como a implementação do desenho de forças direccionadas *spring embedded*, e o *layout* em círculo.

Além de fornecer um conjunto alargado de elementos predefinidos para a visualização de dados, o Prefuse disponibiliza diversas ferramentas para a usabilidade da visualização. Estão disponíveis várias técnicas de interacção como *tooltips*, ou o arrastamento de elementos visuais. São também suportadas técnicas mais sofisticadas como o *zooming*, *panning* ou o *zooming* semântico (secções 2.7 e 2.8). Adicionalmente, o Prefuse distingue claramente as coordenadas absolutas das coordenadas de visualização. Esta separação permite o desenho de todos os elementos visuais de uma forma lógica sem a preocupação que técnicas de visualização aplicadas *à posteriori* alterem a visualização.

É apresentada de seguida a estrutura de pacotes do Prefuse, assim como a sua correspondência com a *pipeline* de visualização:

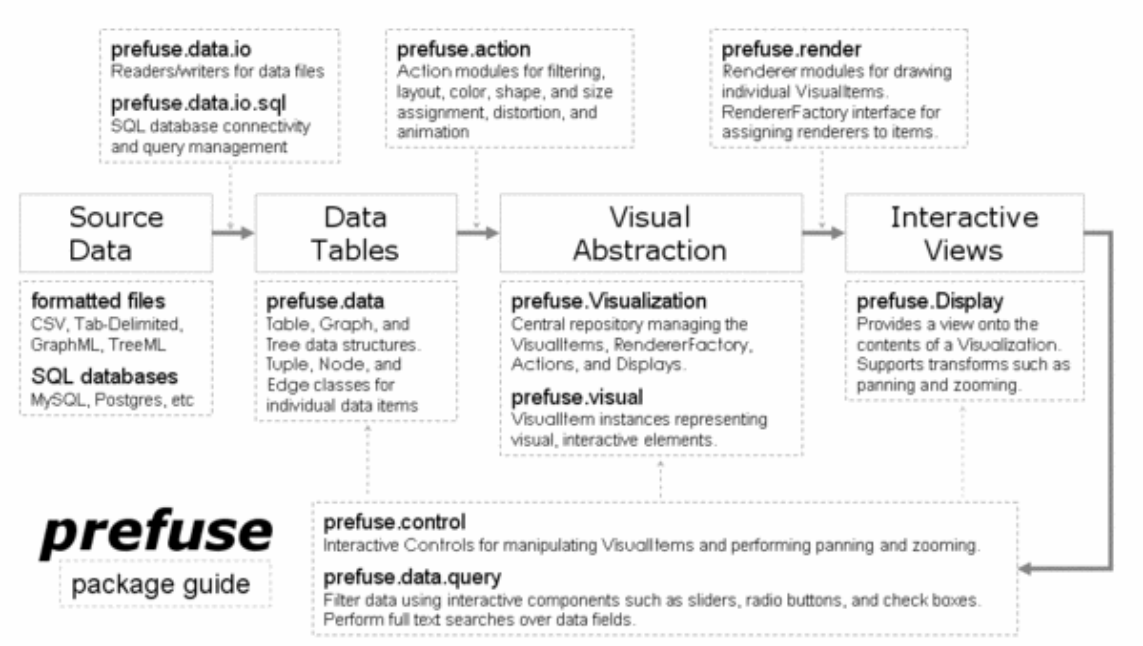



*Figura 27: Guia de pacotes da plataforma Prefuse [155]*

Numa aplicação baseada na plataforma Prefuse, o primeiro passo é a obtenção de um conjunto de dados fonte a visualizar. Este poderá ser um conjunto de dados referente a uma tabela, grafo ou árvore. O pacote *prefuse.data.io* disponibiliza métodos para carregar conjuntos de dados a partir de ficheiros de diferentes formatos. É também possível o carregamento de conjuntos de dados a partir de bases de dados com recurso ao pacote *prefuse.data.io.sql*.

Este conjunto de dados fonte é então usado na construção de tabelas de dados, que constituem a representação interna dos dados a visualizar. O processo desde o conjunto de dados fonte até às tabelas de dados pode envolver apenas a leitura a partir de uma ficheiro formatado ou uma base de dados, como pode também implicar transformações de dados. O pacote *prefuse.data.expression* implementa uma linguagem interpretada de expressões que permite a transformação de dados. As tabelas de dados finais ficam armazenados em estruturas do tipo tabela (*Table*), grafo (*Graph*) e árvore (*Tree*). Uma linha da tabela é representada por um tuplo (*Tuple*) que, numa estrutura de grafo ou árvore, poderá ser um nó (*Node*) ou arco (*Edge*).

As tabelas de dados resultantes (que, apesar do nome, podem também representar estruturas de dados interligados como grafos e árvores) são então sujeitas a mapeamentos visuais para a criação de uma abstracção visual (*Visualization*), um modelo de dados que inclui propriedades visuais como o desenho espacial, cor, tamanho e forma. A abstracção visual é responsável por conter toda a informação necessária para o desenho de uma representação visual dos dados. O pacote *prefuse.visual* inclui os elementos da abstracção visual como os itens visuais (*VisualItem*, *NodeItem* ou *EdgeItem*).

A renderização dos dados da abstracção visual propriamente dita, é efectuada através de um processo de transformações visuais, no qual os conteúdos da abstracção visual são desenhados em diferentes vistas interactivas. Estas vistas podem fornecer perspectivas variadas dos dados, como por exemplo, operações de *panning* e *zooming*. Podem também ser efectuados mapeamentos visuais avançados que possibilitem o desenho e a lógica de movimento dos itens visuais em *layouts*, a atribuição de cores e muitas outras operações sob os itens visuais (*VisualItem*) da abstracção visual (*Visualization*). O pacote *prefuse.action* inclui uma série de acções (*Action*) para a realização de

várias operações visuais predefinidas.

A interacção do utilizador com a visualização (normalmente através do rato ou teclado) pode ser integrada neste processo, causando actualizações em qualquer fase da *pipeline* de visualização. Como exemplo destacam-se o arrastamento de um item, o aumento (*zoom*) de uma vista ou a abertura de um ficheiro de dados diferente.

Neste trabalho foi utilizada toda a extensão da plataforma Prefuse. Especificamente no desenho da visualização, foi utilizado como base o *ForceDirectedLayout* correspondente ao modelo de forças direccionadas de "molas" *spring embedded* utilizado no método de Agrupamento de Dados Visual Interactivo.

### 4.3.2 WEKAUT

O Waikato Environment for Knowledge Analysis (WEKA) é uma plataforma uniformizada para a aprendizagem automática que suporta as várias actividades dos investigadores da área. É desenvolvida em tecnologia Java e está disponível numa licença *open source* (licença GPL). O WEKA inclui implementações de algoritmos de classificação, agrupamento de dados e regras de associação, assim como interfaces gráficas e ferramentas de visualização para exploração de dados e avaliação de algoritmos.

Entre as diversificadas características da plataforma WEKA destacam-se o **pré-processamento de dados**, a **classificação**, o **agrupamento de dados**, a **selecção de atributos**, a **visualização de dados** e a avaliação de resultados [166].

**Pré-processamento de dados.** Além de um formato de ficheiros nativo (.arff), o WEKA suporta vários outros formatos (como ficheiros .csv, Matlab e ASCII), assim como ligação à base de dados através da camada JDBC. Os dados podem ser filtrados através de vários métodos, desde a remoção

de atributos particulares, até operações avançadas como a análise de componentes principais.

**Classificação.** O WEKA contém mais de 100 métodos de classificação. Os classificadores estão divididos em métodos "Bayseanos" (Naïve Bayes, redes Bayseanas), métodos preguiçosos (vizinho mais próximo e variantes), métodos baseados em regras (tabelas de decisão, OneR, RIPPER), aprendizagem em árvore (C4.5, árvores Naïve Bayes, M5), aprendizagem baseada em funções (regressão linear, SVMs, processos Gaussianos), entre outros.

**Agrupamento de Dados.** A aprendizagem não supervisionada é suportada pelo WEKA através de diferentes algoritmos, incluindo os modelos de mistura EM, $K$-Médias e variados algoritmos de agrupamento de dados hierárquicos. Apesar de não existirem tantos métodos como os existentes para a classificação, a maior parte dos algoritmos clássicos do agrupamento de dados estão incluídos no WEKA.

**Selecção de Atributos.** O conjunto de atributos utilizado é essencial para o desempenho do agrupamento de dados ou da classificação. Estão disponíveis vários métodos de pesquisa e critérios de selecção de atributos.

**Visualização de Dados.** Os dados podem ser inspeccionados visualmente através da comparação dos valores dos atributos com a classe ou com os valores de outros atributos. O resultado do classificador pode ser comparado com os dados do conjunto de treino, de forma a detectar *outliers*, a observar características do classificador e barreiras de decisão. Para métodos específicos existem ferramentas de visualização especializadas, como um visualizador de árvores para os métodos que produzam árvores de classificação, um visualizador de redes Bayseanas com um *layout* automático e um visualizador de dendogramas para o agrupamento de dados hierárquicos.

O WEKA também inclui suporte para regras de associação, comparação de classificadores, geração de conjuntos de dados e conversão de dados.

De seguida é apresentada a estrutura de pacotes Java da *framework* WEKA:

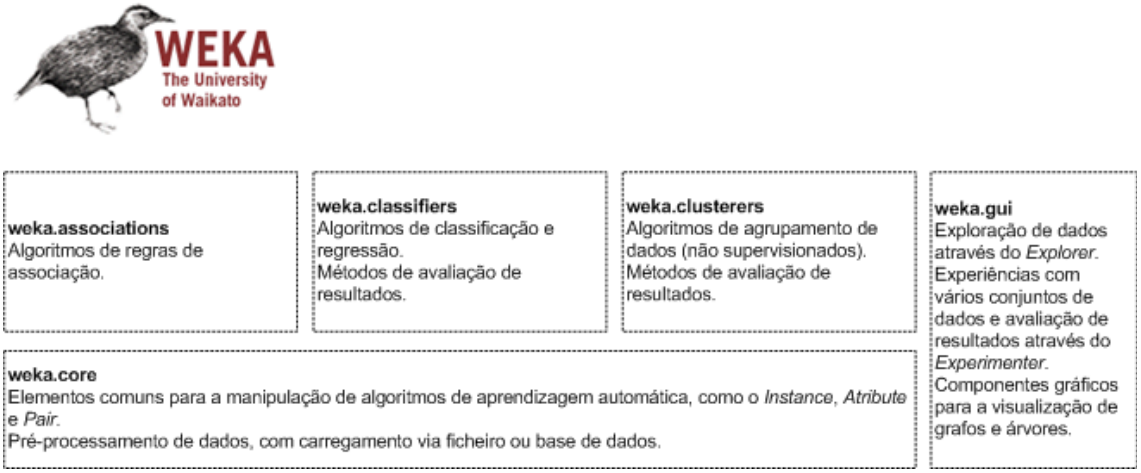



*Figura 28: Guia de pacotes da plataforma WEKA*

Neste estudo, é utilizada uma extensão da *framework* WEKA para suportar aprendizagem semi-supervisionada desenvolvida pelos autores do algoritmo MPC$K$-Médias: WEKAUT [160]. O WEKAUT, estende o pacote *weka.clusterers* para incluir ferramentas para o desenvolvimento de algoritmos de agrupamento de dados com restrições, incluindo métodos para determinar instâncias iniciais para os centros dos grupos, mecanismos de gestão de restrições e medidas de avaliação de agrupamentos de dados. Foram também desenvolvidos alguns algoritmos de agrupamento de dados, entre os quais o MPC$K$-Médias, utilizado neste trabalho.

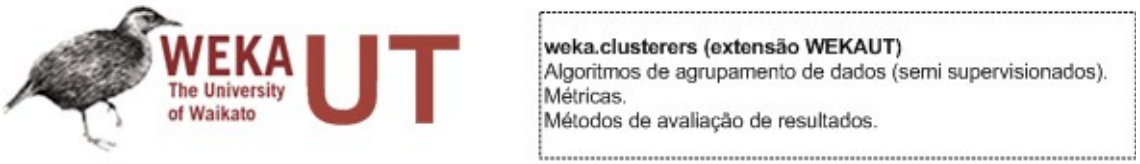



*Figura 29: Extensão WEKAUT*

## *4.4 Framework Preka*

A *framework* Preka [165], desenvolvida no âmbito deste trabalho, reúne um conjunto de

ferramentas base que implementam o método de Agrupamento Visual de Dados, assim como de outras abordagens de interacção (descritas na secção 5.2). O Preka é desenvolvido em Java, sob a licença *open source* GPL versão 2. Esta plataforma surge da necessidade de acrescentar funcionalidade a ambas as *frameworks* Prefuse e WEKA e de uma camada comum que permita a implementação de métodos que combinem técnicas de visualização de informação, interacção com o utilizador e aprendizagem automática, como o apresentado nesta dissertação. O nome da plataforma deriva da combinação dos nomes das plataformas base Prefuse + WEKA = Preka.


*Figura 30: Construção do nome da plataforma Preka*

As características desenvolvidas no Preka separam-se em três grupos distintos: componentes de visualização, aprendizagem automática e comuns. Esta divisão é clara no diagrama de pacotes Java da plataforma Preka:

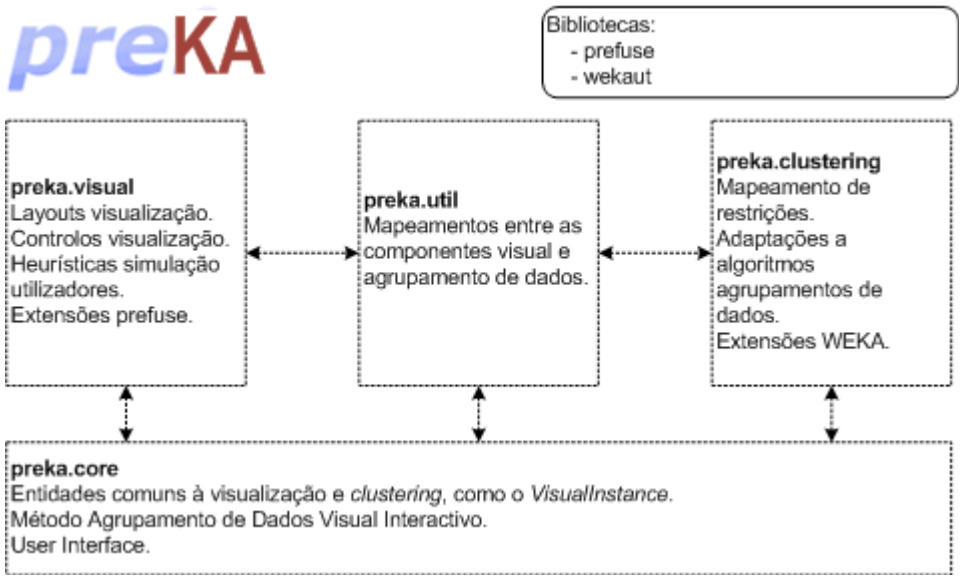



*Figura 31: Guia de pacotes da plataforma Preka*

Nos componentes de visualização podem ser identificadas as extensões ao Prefuse, incluindo controlos específicos para a usabilidade da informação, acções que implementam heurísticas para a simulação do comportamento dos utilizadores e a adaptação do desenho de forças direccionadas *spring embedded*. Estas componentes podem ser encontradas no pacote *preka.visual*.

Nas componentes de aprendizagem automática, foram desenvolvidas ferramentas de ligação entre o agrupamento de dados e a visualização. Estas incluem a detecção e geração de restrições e um método de representação visual para agrupamentos de dados. O algoritmo MPC$K$-Médias, utilizado neste estudo, foi adaptado de forma a tirar partido destes mecanismos. É também nesta componente que se encontra o método de definição de grupos para os métodos sem agrupamento de dados (*NoClustering*). Nestes casos, os "grupos" são definidos pelas instâncias com determinada proximidade com o centro. As medidas de avaliação foram também componentes consideradas, especificamente o índice ARI (descrito na secção 3.3.5). Este trabalho incidiu sobre a aprendizagem semi-supervisionada, mais concretamente no agrupamento de dados com restrições, pelo que as funcionalidades desenvolvidas foram incluídas no pacote *preka.clustering*.

Nas componentes comuns incluem-se: o carregamento de dados fonte da visualização; a partir de conjuntos de dados no formato .arff; a geração automática de ligações relacionais; uma entidade comum a objectos da visualização (*VisualItems* – Prefuse) e instâncias do agrupamento de dados (*Instances* – WEKA) denominada *VisualInstance*; o método de Agrupamento de Dados Visual Interactivo e as diferentes abordagens de visualização; e ferramentas para obtenção de resultados comparativos experimentais. Estas funcionalidades estão disponíveis nos pacotes *preka*, *preka.core* e *preka.util*.

Ao longo deste capítulo, será descrita em detalhe a abordagem de Agrupamento de Dados Visual Interactivo, em conjunto com a respectiva implementação em Preka.

## *4.5 Abordagem*

O paradigma do Agrupamento de Dados Visual Interactivo consiste nos seguintes passos essenciais:

1. **Inicialização da visualização**. É utilizado o algoritmo de desenho de grafos *spring embedded* para gerar a visualização inicial.
2. **Interpretação das acções do utilizador**. Movimentação das instâncias atendendo às preferências/indicações do utilizador. À medida que o utilizador move instâncias são geradas restrições entre pares de objectos que indicam a relação entre eles.
3. **Agrupamento de dados com restrições**. Após o movimento de uma instância são adicionadas novas restrições ao conjunto de restrições e é aplicado um algoritmo de agrupamento de dados com restrições para produzir o novo agrupamento de dados.
4. **Actualização da visualização**. Utilizando o agrupamento produzido, a visualização é actualizada de forma a que os novos grupos sejam visualmente percepcionados.

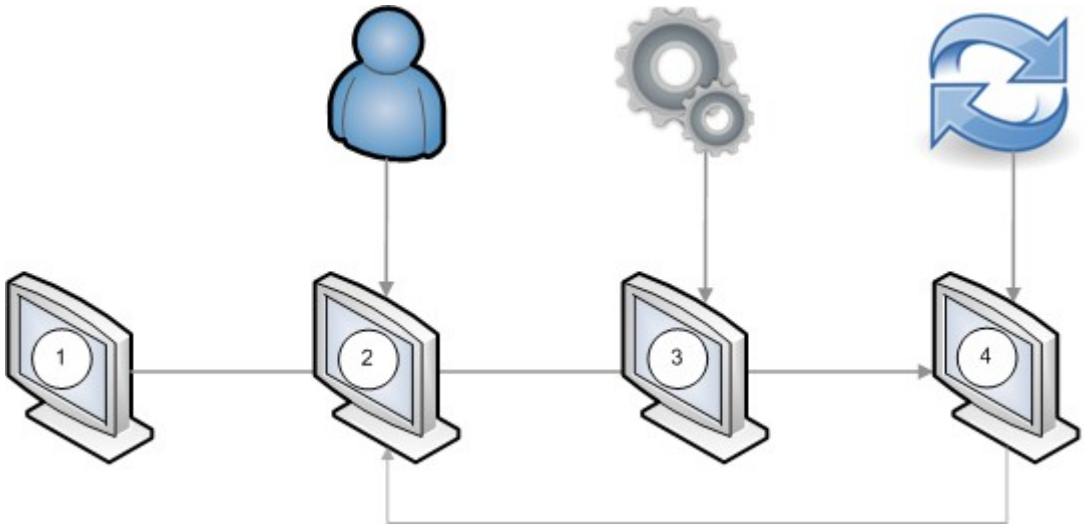



*Figura 32: Passos essenciais do Agrupamento de Dados Visual Interactivo*

Conforme esquematizado na figura 32, assim que é efectuado o carregamento da visualização (passo 1), o ciclo principal do Agrupamento de Dados Visual Interactivo inicia-se:

1. Assim que é detectado o movimento de uma instância pelo utilizador (passo 2) é suspenso o desenho da visualização;

2. De acordo com a posição das instâncias são geradas as restrições que alimentam o algoritmo de agrupamento de dados;
3. A visualização é actualizada de acordo com o novo agrupamento e é retomado o seu desenho.

Os passos repetem-se até ser obtido o agrupamento pretendido. De seguida são detalhadas estas diferentes fases do Agrupamento de Dados Visual Interactivo.

### 4.5.1 Inicialização da Visualização

O processo de Agrupamento de Dados Visual Interactivo inicia-se com o carregamento do grafo a partir de um conjunto de dados a estudar.

Para validar qual o efeito da informação relacional na convergência para os agrupamentos esperados, pode ser incluída informação relacional no conjunto de dados. Neste estudo é utilizada a heurística do vizinho mais próximo. A geração de arcos utilizando esta heurística cria uma ligação entre cada instância e o seu vizinho mais próximo, usando distância Euclidiana no espaço de atributos. Em conjuntos de dados cuja pertença nos grupos esteja fortemente relacionada com a distribuição Euclidiana das instâncias no espaço de atributos, esta acção resultará em ligações bem correlacionadas com os grupos reais. A geração de ligações relacionais com recurso à heurística vizinho mais próximo resultam num número de arcos igual ou inferior ao número de instâncias, uma vez que é criado um arco para cada instância, no entanto duas instâncias podem ser o vizinho mais próximo uma da outra, situação na qual apenas é criada uma ligação.

Após o carregamento do conjunto de dados a visualizar, são definidas as propriedades visuais dos itens visuais a apresentar. Entre outros, estas propriedades visuais podem incluir a cor, o tamanho, a forma e o rótulo (secção 2.4.2). Estas propriedades devem reflectir os atributos de cada instância de forma a permitir que o utilizador efectue o agrupamento que vá de encontro ao seus interesses. Por exemplo, num conjunto de dados de roupa de diferentes cores, uma t-shirt amarela do tamanho L

poderia ser apresentada por um círculo (tipo da peça) amarelo (côr da peça) com um rótulo (tamanho) “L”, enquanto que uma camisa castanha de tamanho M seria um quadrado castanho com o rótulo “M”.

De seguida, é atribuído um desenho (*layout*) à visualização. No método IVC o desenho utilizado é o de forças direccionadas *spring embedded* que actuará em todos os nós e arcos criados na visualização, de acordo com o descrito anteriormente na secção 4.3. Os nós são distribuídos aleatoriamente na visualização. No final, a visualização é exibida e o desenho começa a actuar sobre os itens visuais.

Na *framework* Preka, foi desenvolvida uma componente que possibilita a leitura de ficheiros de conjuntos de dados a partir de ficheiros com o formato .arff, da *framework* WEKA, utilizado extensivamente pela comunidade científica da área da aprendizagem automática. O conjunto de dados carregado é directamente instanciado em itens visuais na camada de visualização fornecida pela plataforma Prefuse. Nesta altura poderá também ser activada a geração de ligações relacionais com recurso à implementação da heurística vizinho mais próximo do Preka (*NNRelationalEdges*). As propriedades visuais são definidas com recurso ao gestor de renderização da visualização da *framework* Prefuse (*RendererFactory*). O desenho é carregado através de uma versão adaptada do *ForceDirectedLayout* do Prefuse e inicia-se o ciclo principal do Agrupamento de Dados Visual Interactivo. Neste estudo foi utilizada uma visualização com as dimensões de 1066x1066 pixeis.

### 4.5.2 Interpretação das Acções do Utilizador

Sempre que um utilizador move uma instância do conjunto de dados, esta é fixa no local, não sendo afectada pelo desenho *spring embedded*. Contudo, estas instâncias fixas exercem forças nas restantes instâncias do grafo.

O processo de agrupamento de dados com restrições é iniciado após o utilizador ter movido duas instâncias. As restrições são geradas a partir do cálculo da distância entre cada par de instâncias. Se

as instâncias se encontrarem a, pelo menos, δ pixeis (em que δ é um parâmetro definido pelo utilizador) são consideradas como pertencentes a grupos diferentes, sendo adicionada uma restrição de ligação proibida entre elas. Se as instâncias se encontrarem a menos do que ϵ pixeis (em que $\epsilon \leq \delta$ é também um parâmetro definido pelo utilizador), as instâncias são consideradas como pertencentes ao mesmo grupo e é adicionada entre elas uma restrição de ligação obrigatória. Sempre que é definido um destes tipos de ligações, os objectos mais próximos entre si que ϵ pixeis são considerados como pertencendo ao mesmo grupo, enquanto que objectos mais afastados entre si que δ pixeis são considerados como pertencendo a grupos distintos. Se a distância no ecrã entre as instâncias for superior a ϵ mas inferior a δ, então a situação é ambígua e não são geradas restrições.

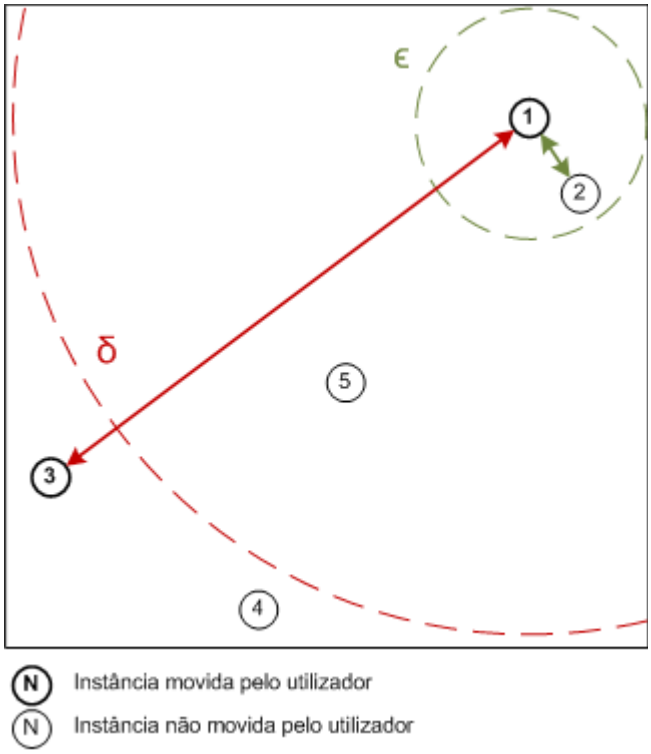



*Figura 33: Geração de restrições*

No exemplo ilustrado na figura acima, as instâncias 1 e 3 foram movidas pelo utilizador. No processamento da instância 1 é verificado que a instância 2 se encontra no raio de ϵ, pelo que é assumido que ambas pertencem ao mesmo grupo e é gerada uma restrição de ligação obrigatória entre as duas instâncias. Já a instância 3, movida pelo utilizador, encontra-se a uma distância superior a δ da instância 1, pelo que é detectado que as mesmas não pertencem ao mesmo grupo,

sendo gerada uma restrição de ligação proibida elas. A instância 4, apesar de estar nas mesmas condições, não foi movida pelo utilizador, pelo que não é criada uma restrição de ligação proibida entre as instâncias 1 e 4. A instância 5 não está abrangida por nenhuma destas regras, logo não são geradas quaisquer restrições envolvendo esta instância.

A acção de ciclo principal do Preka fornece um mecanismo de detecção das instâncias movidas pelo utilizador. Estas instâncias têm de ficar fixas no local após o movimento, razão pela qual foi criado um controlo específico no Preka, *DragFixControl*, que suporta este comportamento. A primeira interacção apenas ocorre após a movimentação de duas instâncias, sendo as restantes interacções despoletadas a cada movimentação de uma instância. Nesta altura, é invocada a componente de detecção de restrições do método *IVC*, responsável por definir ligações obrigatórias entre as instâncias movidas e todas as outras instâncias afastadas de si num máximo de $\epsilon$ pixeis. As ligações proibidas são também detectadas neste passo, mas apenas entre as instâncias movidas separadas por, pelo menos, $\delta$ pixeis. De notar, que a distância calculada é visual (em pixeis) e não a distância Euclidiana no espaço dos atributos. Estes parâmetros ($\epsilon$ e $\delta$) podem ser definidos na classe principal do método *IVC*. Neste estudo foram usados os parâmetros $\epsilon$ = 113 pixeis e $\delta$ = 533 pixeis.

### 4.5.3 Aplicação do Agrupamento de Dados com Restrições

Considerando as restrições inferidas no passo anterior, é então aplicado um algoritmo de agrupamento de dados com restrições que determinará o novo agrupamento a apresentar.

Neste estudo é utilizada a técnica de agrupamento de dados com restrições MPC$K$-Médias. Tendo em conta as restrições definidas no passo anterior, cada objecto de dados $x_i$ é atribuído a um grupo minimizando a soma das distâncias de $x_i$ ao respectivo centro $\overline{x_i}$ e o custo pesado das violações de restrições. O custo da violação das ligações obrigatórias $W_=$ e das ligações proibidas $W_\neq$ é calculado tendo em conta a distância Euclidiana entre os objectos. No final, é efectuada a aprendizagem da medida de distância através da ponderação dos atributos. O funcionamento deste algoritmo, incluindo o método de aprendizagem da medida de distância, encontra-se descrito com

maior pormenor na secção 3.5.3.4.

A plataforma Preka fornece a listagem de restrições ao algoritmo MPC$K$-Médias implementado pela *framework* WEKA e este devolve um novo agrupamento. Este resultado é incorporado nos itens visuais que o representam na visualização. Esta acção é conseguida através das instâncias visuais Preka (*VisualInstances*) que mapeiam instâncias WEKA (Instances) com itens visuais Prefuse (*VisualItems*).

### 4.5.4 Actualização da Visualização

Após a geração de novas restrições, no passo da interpretação das acções do utilizador, e o novo agrupamento produzido, no passo do agrupamento de dados com restrições, a visualização deve ser actualizada de forma a reflectir os grupos inerentes ao novo agrupamento. As ligações relacionais entre os dados é preservada. Se esta estrutura relacional estiver correlacionada com o novo agrupamento, então as ligações do grafo e as ligações do agrupamento reforçar-se-ão, levando a uma rápida convergência para o agrupamento final.

Para actualizar o grafo é utilizada uma abordagem descrita por Brockenauer e Cornelson [158] para representar visualmente agrupamentos de dados em grafos. Inicialmente é criado um nó fictício para representar o centro de cada grupo. A posição deste nó é calculada como a posição média dos nós movidos pertencentes a esse grupo. De seguida, é adicionada uma ligação entre cada centro e cada instância pertencente ao grupo. As ligações relacionais utilizam a *constante de mola* (*spring constant*) definida por omissão no Prefuse $(2.0\times10^{-5})$. As ligações do agrupamento de dados possuem uma constante de mola igual ao dobro do valor por omissão $(4.0\times10^{-5})$. Como resultado, as ligações das instâncias com os grupos têm um efeito bastante mais significativo que as ligações relacionais, não dominando no entanto completamente o desenho da visualização.

O desenho *spring embedded* é então aplicado no novo grafo combinado, que inclui as ligações relacionais do conjunto de dados original e as novas ligações do agrupamento de dados. O grafo

resultante é apresentado, mas apenas as ligações relacionais são apresentadas ao utilizador, não sendo desenhados os nós fictícios que representam o centro dos grupos. Seria possível a inclusão do desenho das ligações do agrupamento de dados, no entanto isto resultaria num grafo demasiado confuso, ofuscando as ligações relacionais.

Sempre que existam grupos atribuídos às instâncias visuais, o Preka invoca uma componente de desenho de grafos num contexto de agrupamento de dados. Neste trabalho foi desenvolvido o método *BrockenauerCornelsonGraph* que aplica a abordagem referida acima. Este método é alimentado pela informação de ligações de agrupamento de dados e ligações relacionais para efectuar a construção do grafo a visualizar. O desenho de forças direccionadas *spring embedded* do Prefuse não permite a aplicação de diferentes forças por tipo de ligação. Foi por esta razão necessário o desenvolvimento de uma especialização deste desenho na *framework* Preka que permitisse a aplicação de diferentes forças, a ligações de agrupamento de dados e a ligações relacionais. Após a aplicação do novo agrupamento do MPC$K$-Médias ao grafo, a visualização é actualizada com as novas ligações, a aplicação do desenho de forças direccionais é efectuada e o utilizador poderá mover uma nova instância. O processo repete-se até que o utilizador esteja satisfeito com o agrupamento obtido.

### 4.5.5 Simulação do Utilizador

A técnica IVC apresentada é direccionada a utilizadores humanos. Contudo, de forma a permitir uma quantidade superior de testes experimentais que levassem a aferir resultados mais assertivos, foram implementadas e utilizadas duas heurísticas de selecção de instâncias: *selecção aleatória* e *selecção da mais afastada*. O método de selecção aleatória de instâncias simplesmente identifica uma instância aleatoriamente para mover em cada passo. Já o método de selecção da mais afastada selecciona a instância mais afastada (no ecrã) do grupo correcto, movendo-a para perto desse grupo em cada interacção. A intuição inerente a esta heurística é a de que o utilizador tenderá a aperceber-se com maior facilidade das instâncias anómalas, isto é, das instâncias que surjam mais afastadas do local onde deveriam estar. Para ambas as heurísticas de movimentação de instâncias, são utilizadas localizações predefinidas (junto aos cantos do ecrã) para os centros dos grupos. Após a selecção da

instância a mover com base numa das heurísticas descritas, esta é reposicionada para perto do centro correcto. Esta movimentação possibilita a criação de restrições de ligação obrigatória com outros objectos pertencentes a esse grupo.

A plataforma Preka implementa as duas heurísticas através dos métodos *RandomMove* e *FarthestFirstMove* correspondentes a novas acções Prefuse aplicadas à visualização. Estes métodos detectam que uma instância não pertence ao grupo correcto através do resultado do agrupamento obtido na execução do algoritmo MPC$K$-Médias. Após a selecção da instância a mover, é desencadeada uma animação que coloca a instância numa posição aleatória num círculo de raio $\epsilon$ pixeis desde o centro do grupo real da instância. A cada interacção é calculado o ARI (descrito na secção 3.3.5). A simulação pára quando o ARI atingir o valor 1 (todos os elementos foram instanciados nos grupos correctos).

## *4.6 Funcionamento*

De forma a clarificar o processo de Agrupamento de Dados Visual Interactivo, é apresentada uma sequência de figuras que representam a acção de um utilizador movendo os nós para os grupos pretendidos. A figura 34 mostra a visualização inicial produzida pelo conjunto de dados *Iris*, que contém objectos de dados pertencentes a três grupos distintos, conforme descrito na secção 5.3.3. Para efeitos de ilustração do processo, as cores indicam a que grupos pertencem as instâncias na realidade. Os rótulos numéricos indicam o número da instância. Para simplificar a visualização destas representações, as ligações relacionais não são mostradas. Repare-se que os nós de todos os grupos se encontram dispersos no ecrã.

Os nós assinalados na figura 34 correspondem aos dois primeiros nós seleccionados pelo utilizador. A visualização resultante é apresentada na figura 35. Nesta representação, os grupos do canto superior esquerdo e do canto inferior direito (onde os primeiros dois nós foram colocados) começam a definir-se.

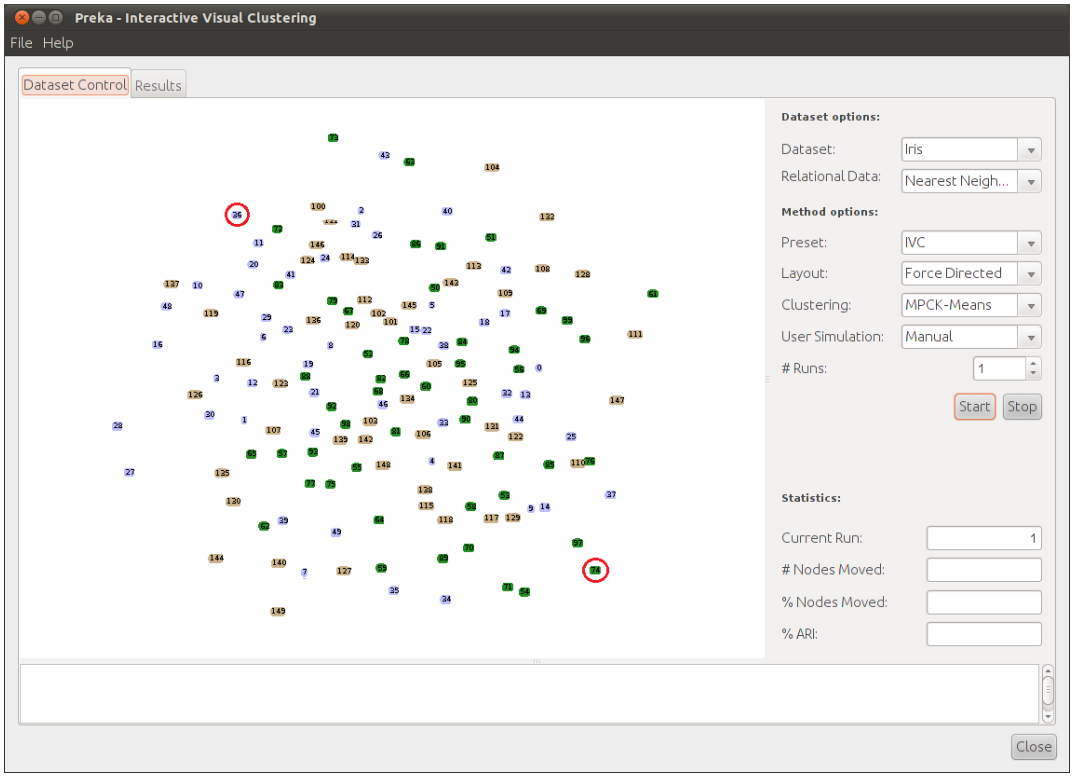



*Figura 34: Desenho inicial do conjunto de dados Iris*

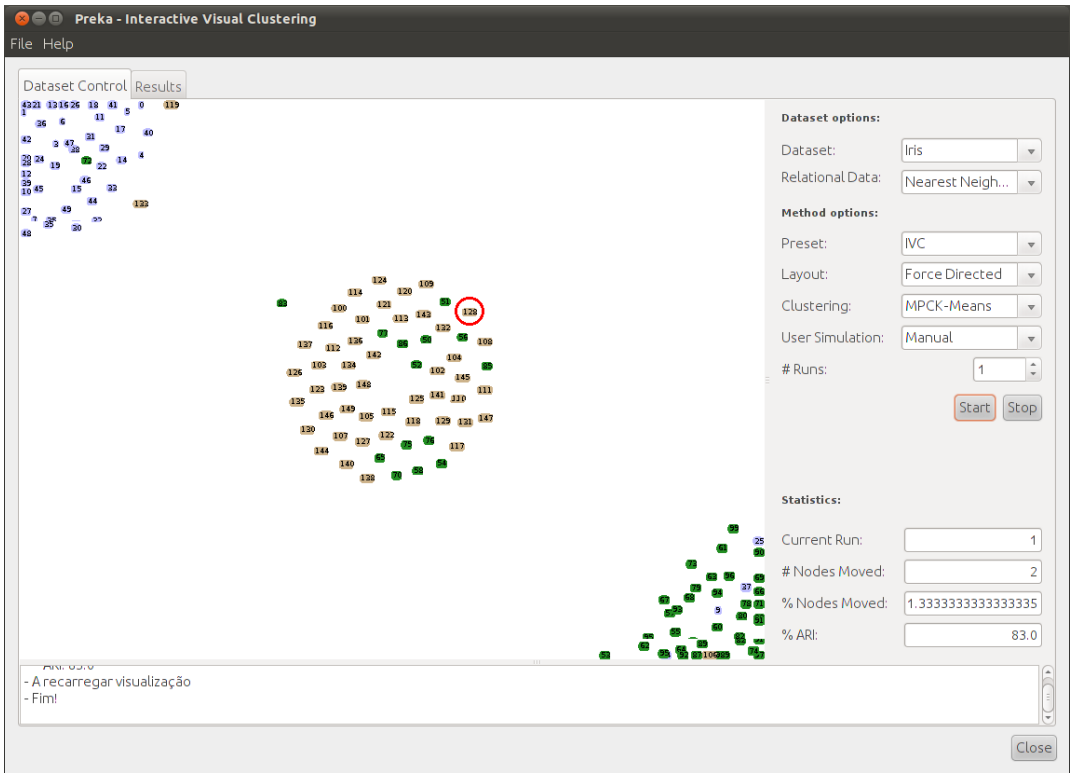



*Figura 35: Desenho do conjunto de dados Iris após duas instâncias movidas*

A figura 36 mostra a visualização após o movimento do terceiro nó. Por não se enquadrar com nenhum dos grupos cuja formação se iniciou com os dois primeiros nós, esta instância deu origem à criação de um novo grupo. Este foi o último grupo gerado, uma vez que o conjunto de dados representado atingiu o limite de grupos definido.

A figura 37 mostra a visualização após 14 movimentações de nós. Neste ponto, conforme é possível comprovar no gráfico da figura 46, a maioria das instâncias encontra-se agrupada correctamente nos seus grupos "reais". Os grupos são visualmente muito distintos, apenas com alguns nós espalhados entre os grupos.

## *4.7 Sumário*

Neste capítulo foi apresentado o paradigma do Agrupamento de Dados Visual Interactivo como uma combinação e extensão de técnicas de visualização de informação, interacção com o utilizador e de agrupamento de dados com restrições. Foi também apresentada a sequência de passos necessária à implementação de um processo de Agrupamento de Dados Visual Interactivo, desde a inicialização da visualização com recurso ao desenho de forças direccionadas *spring embedded*, passando pela interacção com o utilizador na aquisição de restrições, até à integração das restrições num algoritmo de agrupamento de dados com restrições. No caso deste estudo, utilizando o MPC$K$-Médias.

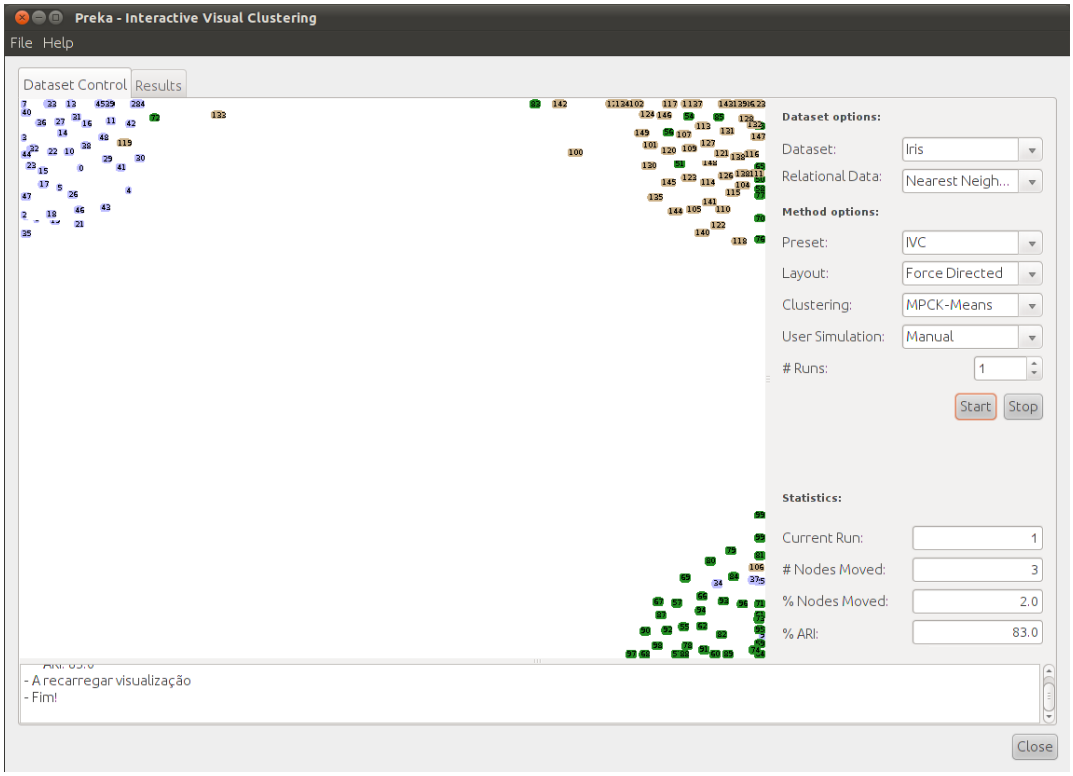



*Figura 36: Desenho do conjunto de dados Iris após três instâncias movidas*

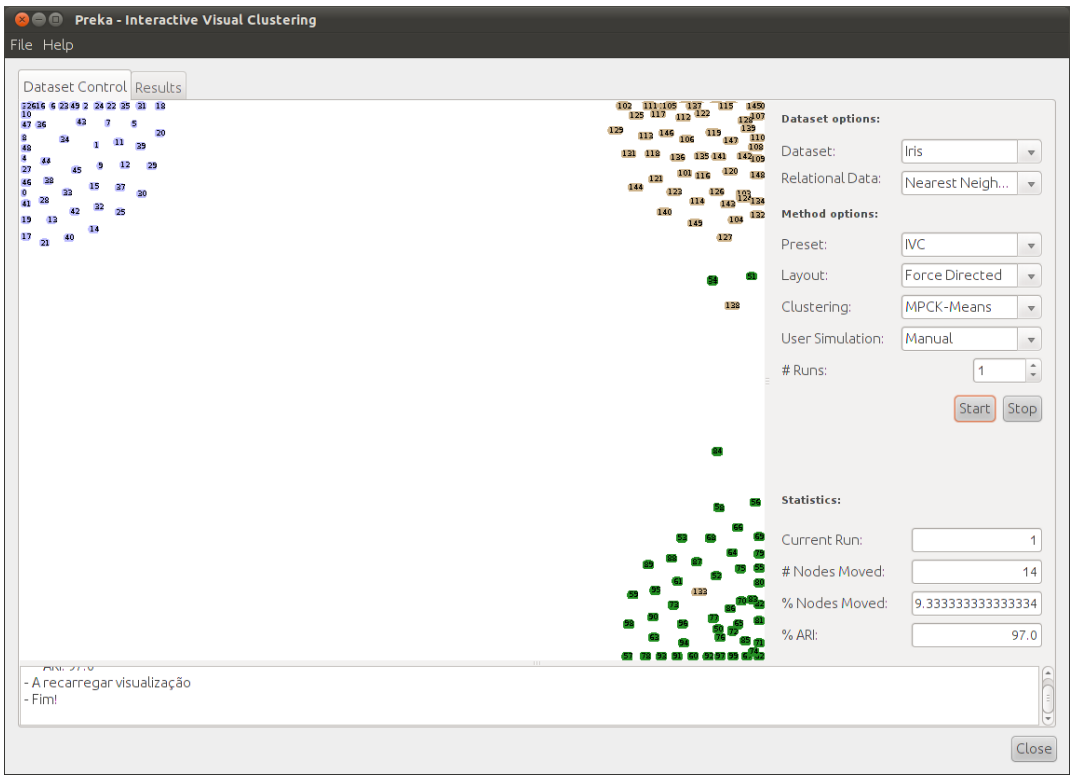



*Figura 37: Desenho do conjunto de dados Iris após catorze instâncias movidas*

# 5 Avaliação de Abordagens de Interacção com o Utilizador

## 5.1 Introdução

O Agrupamento de Dados Visual Interactivo é uma técnica centrada na interacção com o utilizador para acelerar o processo de convergência de conjuntos de dados eventualmente com informação relacional para um agrupamento que represente os interesses e objectivos pretendidos. A finalidade é permitir que se consiga atingir o agrupamento desejado com o mínimo de pedidos ao utilizador. De forma a comparar o desempenho do Agrupamento de Dados Visual Interactivo e de outras abordagens de interacção com o utilizador, é efectuado neste capítulo um estudo comparativo entre os diferentes paradigmas onde é medida a eficácia da convergência para o agrupamento óptimo em cinco conjuntos de dados.

## 5.2 Metodologia

Neste capítulo, é comparada a abordagem do Agrupamento de Dados Visual Interactivo com quatro abordagens alternativas. As cinco abordagens testadas são apresentadas na tabela 5:

| Abordagem | Layout? | Clustering? | Inf. Relacional? | Simulação Utilizador |
|---|---|---|---|---|
| Agrupamento de Dados Visual Interactivo (IVC) | **Sim** | **Sim** | **Sim** | Selecção da mais afastada |
| Agrupamento de Dados com Restrições (ADR) | **Sim** | **Sim** | **Não** | Selecção aleatória |
| Forças Direccionadas (FD) | **Sim** | **Não** | **Sim** | Selecção da mais afastada |
| Forças Direccionadas Aleatório (FDA) | **Sim** | **Não** | **Sim** | Selecção aleatória |
| Desenho Manual (Manual) | **Não** | **Não** | **Não** | Selecção aleatória |

*Tabela 5: Abordagens de interacção com o utilizador testadas*

A coluna "*Layout?*" indica se o desenho de forças direccionadas é, ou não, utilizado. Quando não é utilizado o desenho de forças direccionadas, a visualização apenas é movimentada quando o utilizador move as instâncias. Sendo utilizado o desenho (*layout*) de forças direccionadas, as instâncias movem-se constantemente, de acordo com o cálculo das forças das molas e das instâncias. A coluna "*Clustering?*" indica se é utilizado o agrupamento de dados com restrições. Sendo utilizado, de cada vez que uma instância é alterada é realizado um novo agrupamento de dados e as ligações dos objectos aos grupos são actualizadas. Se não for utilizado agrupamento de dados, não são consideradas as ligações de agrupamento de dados. Os "grupos" são determinados pela proximidade das instâncias ao centro. Se as instâncias estiverem no raio do parâmetro definido para as restrições de ligação obrigatória ($\epsilon$), então é considerado que se encontram no mesmo "grupo". A coluna "Inf. Relacional?" indica se o método utiliza a versão do conjunto de dados com informação relacional. Em caso afirmativo, no passo de carregamento do conjunto de dados são geradas ligações relacionais entre os objectos utilizando a heurística vizinho mais próximo (secção 4.5.1). Estas ligações geradas inicialmente permanecem sempre presentes ao longo de todas as interacções. Apesar da abordagem de Agrupamento de Dados Visual Interactivo ser direccionada a utilizadores humanos, por escassez de recursos, são utilizadas heurísticas de simulação de utilizadores na execução das diferentes abordagens. A coluna "Simulação Utilizador" indica a heurística de movimentação automática de instâncias utilizada: selecção aleatória ou selecção da mais afastada (secção 4.5.5).

A abordagem Agrupamento de Dados Visual Interactivo (IVC) é o tema central deste trabalho e encontra-se detalhada no capítulo 4. Esta abordagem combina o desenho de forças direccionadas com o agrupamento de dados, utilizando, sempre que esteja disponível, a informação relacional dos conjuntos de dados. A heurística de movimentação de instâncias utilizada é a "selecção da mais afastada", uma vez que simula mais proximamente o comportamento de um utilizador humano. A abordagem Agrupamento de Dados com Restrições (ADR) é equivalente a efectuar um agrupamento de dados com restrições normal com atribuição de restrições gradual à medida que as instâncias são movimentadas. Contudo, o *layout* exerce influência nos resultados deste método, na medida em que as restrições obrigatórias são definidas pelas instâncias mais próximas da instância movida, no espaço do ecrã (secção 4.5.2). Não é utilizada informação relacional, uma vez que o algoritmo MPC$K$-Médias, usado neste estudo, não considera este tipo de ligações. É utilizada a heurística "selecção aleatória", pois os algoritmos de agrupamento de dados são indiferentes à

posição das instâncias na visualização. As abordagens de *layout*, Forças Direccionadas (FD) e Forças Direccionadas Aleatório (FDA), são idênticas ao IVC, sem a componente do agrupamento de dados. A diferença entre as duas abordagens está na heurística de movimentação de instâncias: no primeiro é usada a "selecção da mais afastada"; e no último a "selecção aleatória". Estas configurações permitirão verificar, por um lado, qual a influência do agrupamento de dados no método IVC, por outro, o efeito da aplicação das diferentes heurísticas de simulação do utilizador. A abordagem Desenho Manual (Manual) reflecte a movimentação das instâncias isoladamente, numa visualização estanque, sem o apoio de técnicas auxiliares como o *layout* de forças direccionadas ou o agrupamento de dados. A informação relacional é irrelevante em cenários onde as instâncias não são afectadas pelo desenho da visualização, pelo que não é utilizada neste método. Esta abordagem permitirá quantificar a vantagem das diferentes abordagens face à organização manual das instâncias.

Uma vez que a incorporação de conhecimento *a priori* assente em restrições entre pares de objectos de dados contribui de forma significativa para a obtenção de um agrupamento que corresponda aos interesses do utilizador, é esperado que as abordagens que incluam o agrupamento de dados com restrições obtenham melhores resultados.

O desenho de forças direccionadas deverá ter um desempenho superior ao desenho manual devido à utilização das ligações relacionais do conjunto de dados. Estas ligações tendem a agrupar as instâncias visualmente, uma vez que aproximam instâncias que estão relacionadas. Desta forma, quando uma instância mal posicionada é movida para o grupo correcto, esta deve também aproximar instâncias semelhantes, resultando na possibilidade de várias instâncias serem movidas para o grupo correcto. O reposicionamento destas instâncias no novo grupo poderá resultar na geração de novas restrições de ligação obrigatória, fazendo com que no agrupamento gerado na iteração seguinte, estas pertençam ao mesmo grupo. Com isso, são geradas as respectivas ligações do agrupamento de dados que exercem uma influência ainda mais forte, uma vez que a sua *constante de mola* é superior, fazendo com que o desenho represente cada vez mais o agrupamento aprendido.

De notar que o paradigma IVC está assente no pressuposto de que as ligações relacionais estão correlacionadas com a associação das instâncias ao grupos. Se isto não se verificar, estas ligações não serão úteis e resultarão num decréscimo de desempenho.

Para medir o desempenho de cada uma das abordagens apresentadas neste trabalho, é utilizado o Índice Rand Ajustado (ou *Adjusted Rand Index* – ARI) [109]. O ARI é usado para avaliar a "proximidade" de um agrupamento de dados ao agrupamento real dos dados. O Índice Rand (ou *Rand Index*) [108] mede a proporção das correspondências do agrupamento. Uma correspondência consiste num par de instâncias que se encontram no mesmo grupo tanto no agrupamento aprendido, como no agrupamento-alvo ou se encontram em grupos diferentes nos dois agrupamentos. O Índice Rand penaliza as partições com maior número de grupos, pelo que o Índice Rand Ajustado é mais utilizado. O valor do ARI varia entre 0 e 1. Um ARI de 1 significa que todas as instâncias estão correctamente agrupadas. Neste trabalho é utilizada a implementação do ARI disponibilizada pelo sistema WEKA [159]. O ARI é descrito com maior pormenor na secção 3.3.5.

Nos resultados experimentais, o desempenho do agrupamento de dados é sempre apresentado em função do número de instâncias movidas. As posições iniciais das instâncias no desenho são definidas aleatoriamente, pelo que, para cada experiência, é apresentado o desempenho médio de 20 execuções.

Foram utilizados os seguintes parâmetros de sistema na *framework* Preka, neste estudo:

- Área visual – $1066 \times 1066$ pixeis
- Constante de Mola de Agrupamento de Dados – $4 \times 10^{-5}$
- Constante de Mola de Ligações Relacionais – $2 \times 10^{-5}$
- Distância para Restrições de Ligação Obrigatória $\epsilon$ – 113 pixeis
- Distância para Restrições de Ligação Proibida $\delta$ – 533 pixeis
- Tempo entre movimentos (com *layout*) – 3000 milissegundos
    - Período entre execuções do agrupamento de dados sobre a visualização, para permitir a

estabilização do desenho (*layout*) da visualização

- Tempo de animação de movimentos (com *layout*) – 1000 milissegundos
  - Período de animação do "arrastamento" automático de itens visuais, para permitir a influência do desenho (*layout*) da visualização
- Tempo entre movimentos (sem *layout*) – 300 milissegundos
  - Período entre execuções do agrupamento de dados sobre a visualização
- Tempo de animação de movimentos (sem *layout*) – 100 milissegundos
  - Período de animação do "arrastamento" automático de itens visuais

## *5.3 Conjuntos de Dados*

As diferentes abordagens foram testadas experimentalmente utilizando cinco conjuntos de dados: dois criados no âmbito deste estudo (*Circles* e *Overlapping Circles*); o *Iris* do repositório de aprendizagem automática da UC Irvine [163]; e dois conjuntos de dados utilizados noutros estudos de agrupamento de dados *Cigar* e *Half Rings* obtidos em [98][91]. Para possibilitar os testes com informação relacional, foi utilizada a heurística "vizinho mais próximo" [161] (secção 4.5.1) na geração de ligações relacionais que resultou em duas versões de cada conjunto de dados: uma sem ligações relacionais e outra com ligações relacionais geradas com base na heurística referida.

### *5.3.1 Circles*

O conjunto de dados *Circles* inclui 120 instâncias distribuídas por dois grupos não sobrepostos:

- Grupo 1
  - São gerados aleatoriamente 50 pontos no plano cartesiano com centro na coordenada [50, 50], num raio de 50

- Grupo 2
    - São gerados aleatoriamente 50 pontos no plano cartesiano com centro na coordenada [150, 150], num raio de 50

Pelo facto da distância entre os centros dos grupos ser superior à soma dos seus raios, os grupos não se intersectam. São ainda gerados aleatoriamente 20 pontos fora da área definida para a criação dos objectos nos grupos. Estes pontos são associados ao centro do grupo mais próximo. Os dois atributos de cada instância são os pontos cartesianos $(x, y)$.

O conjunto de dados *Circles* é criado com o objectivo de tornar possível a observação do comportamento das diferentes abordagens num conjunto de dados em que os grupos estão claramente separados.

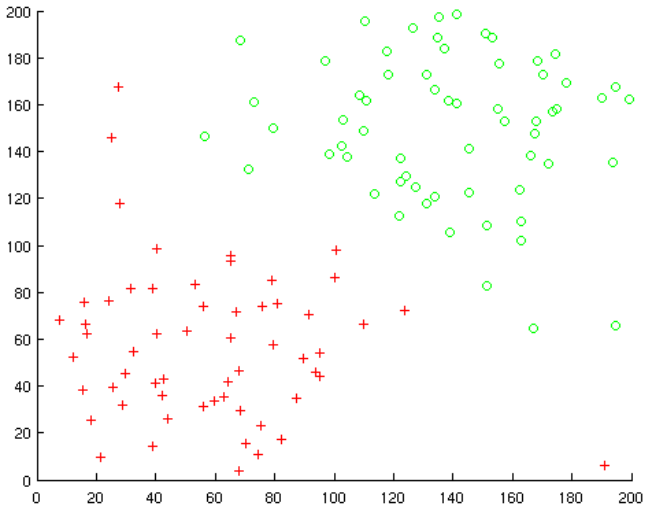



*Figura 38: Conjunto de dados Circles*

Na versão com ligações relacionais deste conjunto de dados, existe um número elevado de ligações (36), sendo que todas ligam instâncias do mesmo grupo.

### *5.3.2 Overlapping Circles*

O conjunto de dados *Overlapping Circles* inclui 100 instâncias distribuídas por quatro grupos sobrepostos. Os centros dos grupos são posicionados no centro de cada quadrante (formando um quadrado de tamanho 2) de um plano cartesiano $(x, y)$, colocados a distâncias iguais entre eles. São colocadas 25 instâncias em cada grupo, de forma aleatória, num raio igual à distância entre os centros. Assim, pelo facto da distância entre os centros dos grupos ser inferior à soma dos seus raios, estes sobrepõem-se, existindo instâncias de vários grupos na área de sobreposição.

O intuito da criação do *Overlapping Circles* é o de possibilitar a observação da evolução das diferentes abordagens num conjunto de dados em que os grupos não estão claramente separados entre si.

Uma vez mais, a versão deste conjunto de dados com ligações relacionais possui ligações de instâncias pertencentes ao mesmo grupo (31), mas também pertencentes a grupos distintos (4).

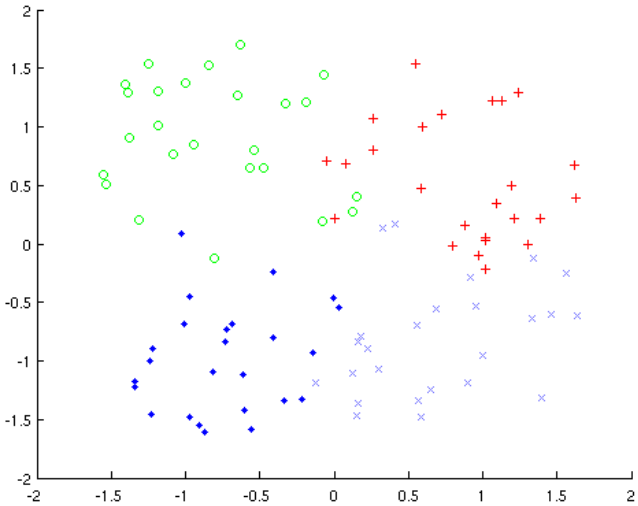



*Figura 39: Conjunto de dados Overlapping Circles*

### 5.3.3 *Iris*

O conjunto de dados *Iris*, proveniente do repositório de aprendizagem automática da UC Irvine [163], é extensamente utilizado na classificação e no agrupamento de dados. Este conjunto de dados contém 150 instâncias distribuídas por três grupos de 50 instâncias cada. Cada instância é descrita por quatro atributos numéricos: tamanho e comprimento das sépalas e tamanho e comprimento da pétalas. Os três grupos correspondem às três espécies diferentes de íris: *Iris Setosa*, *Iris Versicolour* e *Iris Virginica*. Este conjunto de dados é conhecido por ser um dos mais difíceis de tratar para a maioria dos algoritmos de agrupamento de dados, uma vez que duas das classes são linearmente separáveis uma da outra, ao contrário da terceira.

A versão com ligações relacionais deste conjunto de dados apresenta um total de 41 ligações entre instâncias do mesmo grupo e 1 ligação entre instâncias de grupos distintos.

### *5.3.4 Cigar*

O conjunto de dados Cigar é constituído por quatro grupos, possuindo dois dos grupos 100 objectos cada e os restantes dois grupos 25 objectos cada. Estes quatro grupos não se intersectam no plano cartesiano, como é possível verificar na figura 40. Por questões de optimização do tempo de processamento, foi utilizada uma versão deste conjunto de dados reduzida a 150 instâncias, com a seguinte distribuição: dois grupos com 60 instâncias e dois grupos com 15 instâncias. A selecção das instâncias a remover de cada grupo foi aleatória.

Este conjunto de dados, na sua versão com informação relacional possui um elevado número de ligações (42), sendo todas entre instâncias pertencentes ao mesmo grupo.

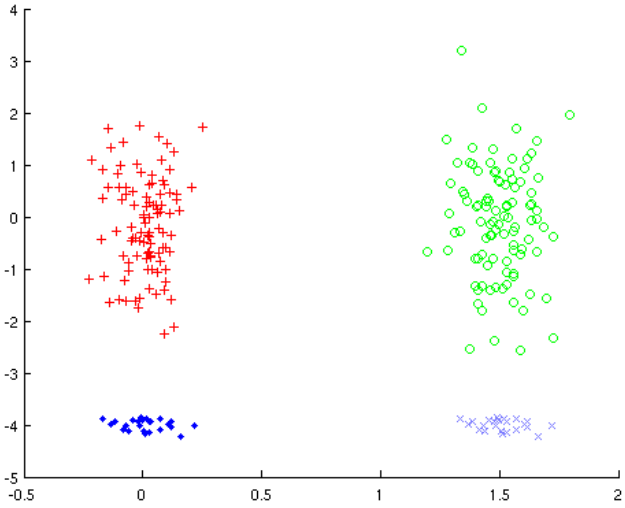



*Figura 40: Conjunto de dados Cigar*

### *5.3.5 Half Rings*

O conjunto de dados Half Rings é formado por três grupos, dois contendo 150 objectos cada e um terceiro com 200 objectos. Estes grupos, apesar de não se intersectarem, partilham áreas próximas, como é possível verificar na figura 41. A versão deste conjunto de dados utilizada neste estudo foi também reduzida a 150 instâncias, com a seguinte distribuição: dois grupos com 45 instâncias cada e um grupo com 60 instâncias. Uma vez mais, o critério de selecção das instâncias a excluir em cada grupo foi a aleatoriedade.

Mais uma vez, na versão com informação relacional deste conjunto de dados, todas as ligações (42) são referentes a instâncias pertencentes ao mesmo grupo.

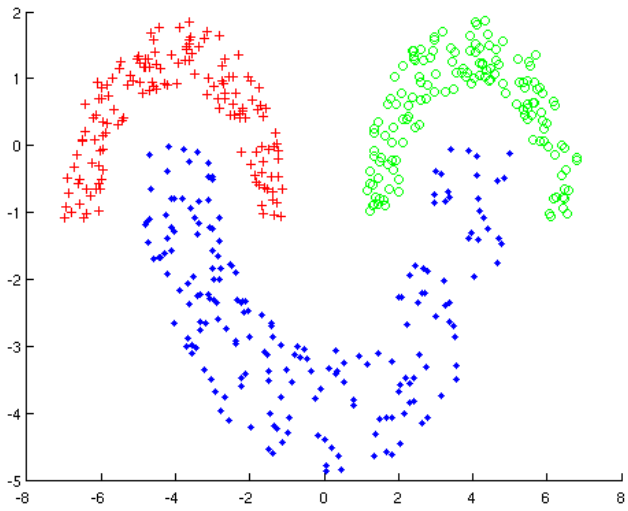



*Figura 41: Conjunto de Dados Half Rings*

## 5.4 Resultados e Discussão

São apresentados nesta secção os resultados experimentais das diferentes abordagens de interacção com o utilizador nos cinco conjuntos de dados apresentados. O estudo é dividido numa primeira comparação entre o desempenho dos cinco métodos, e numa avaliação das versões do conjunto de dados com e sem informação relacional nos três métodos que utilizam este tipo de ligações.

### 5.4.1 Circles

| **Abordagem \ Nós Movidos** | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 94.66% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| ADR | 90.68% | 97.21% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| FD | 49.64% | 59.02% | 71.38% | 94.43% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| FDA | 49.64% | 58.82% | 71.38% | 87.99% | 96.66% | 100.00% | 100.00% | 100.00% | 100.00% |
| Manual | 49.64% | 49.65% | 49.79% | 52.36% | 56.73% | 61.84% | 77.92% | 89.91% | 100.00% |

*Tabela 6: ARI nos 100 primeiros movimentos do conjunto de dados Circles*

Conforme é possível verificar na tabela 6 e no gráfico da figura 42, os resultados deste conjunto de dados são os previstos. O movimento manual de instâncias (Manual – linha castanha) é o que

mostra a pior evolução, tendo apenas atingido o agrupamento desejado após 90 nós movidos. O desenho de Forças Direccionais Aleatório (FDA – linha amarela) evidencia uma melhoria significativa sobre o método anterior. A inclusão da heurística "selecção da mais afastada" no método Forças Direccionadas (FD – linha laranja) resultou num desempenho ainda superior. O método de Agrupamento de Dados Visual Interactivo (IVC – linha azul) é o que tem o melhor desempenho, contudo, o Agrupamento de Dados com Restrições (ADR – linha verde) apresenta um resultado praticamente idêntico. Este resultado é esperado na medida em que as instâncias estão bem separadas, sendo este um problema simples para os algoritmos de agrupamento de dados.

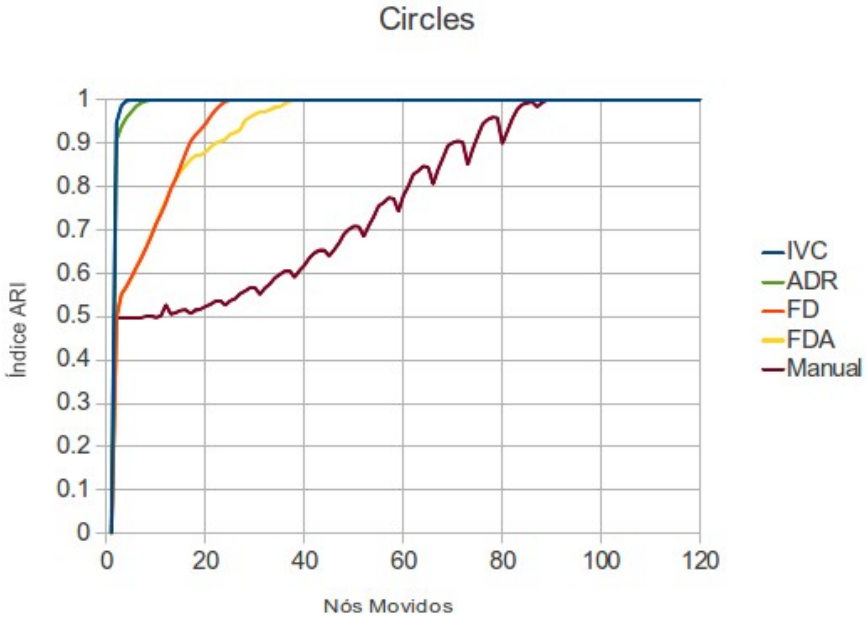



*Figura 42: Resultados experimentais no conjunto de dados Circles*

É também interessante avaliar qual a influência da informação relacional neste conjunto de dados.

| **Abordagem / Nós Movidos** | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 94.66% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| IVC s/Inf. Relacional | 91.20% | 97.46% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| FD | 49.64% | 59.02% | 71.38% | 94.43% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% |
| FD s/Inf. Relacional | 49.64% | 51.11% | 52.50% | 58.06% | 65.39% | 74.87% | 97.97% | 100.00% | 100.00% |
| FDA | 49.64% | 58.82% | 71.38% | 87.99% | 96.66% | 100.00% | 100.00% | 100.00% | 100.00% |
| FDA s/Inf. Relacional | 49.64% | 50.80% | 52.32% | 57.48% | 65.47% | 75.38% | 83.96% | 89.56% | 95.83% |

*Tabela 7: ARI nos 100 primeiros movimentos do conjunto de dados Circles com ligações relacionais e sem ligações relacionais*

Conforme esperado, os métodos utilizando o conjunto de dados com ligações relacionais geradas pelo método "vizinho mais próximo" (secção 4.5.1) obtêm melhores resultados. No gráfico da figura 43 é possível notar-se a diferença de desempenho entre os métodos que utilizam o conjunto de dados com informação relacional (linhas contínuas) e sem informação relacional (linhas tracejadas).

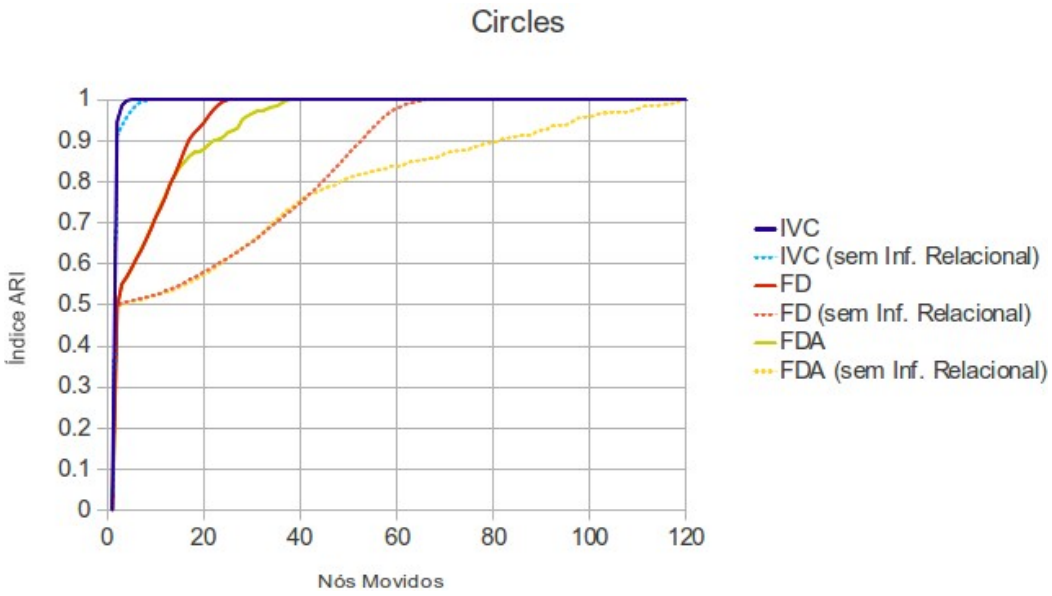



*Figura 43: Efeito das ligações relacionais no conjunto de dados Circles*

### *5.4.2 Overlapping Circles*

| **Abordagem \ Nós Movidos** | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 78.19% | 79.25% | 81.77% | 90.51% | 98.05% | 99.85% | 100.00% | 100.00% | 100.00% |
| ADR | 76.98% | 77.12% | 77.92% | 80.91% | 86.28% | 92.67% | 99.85% | 100.00% | 100.00% |
| FD | 24.29% | 65.82% | 67.85% | 75.72% | 87.55% | 99.56% | 100.00% | 100.00% | 100.00% |
| FDA | 24.29% | 66.62% | 69.58% | 79.64% | 90.13% | 98.84% | 100.00% | 100.00% | 100.00% |
| Manual | 24.24% | 58.79% | 62.72% | 64.69% | 66.26% | 71.62% | 81.31% | 91.81% | 100.00% |

*Tabela 8: ARI nos 100 movimentos do conjunto de dados Overlapping Circles*

A figura 44 mostra os resultados experimentais para o conjunto de dados *Overlapping Circles*. O método Agrupamento de Dados Visual Interactivo é novamente o que apresenta melhor desempenho, desta vez com um distanciamento bastante mais significativo do Agrupamento de Dados com Restrições. Este comportamento é justificado pelo facto dos grupos não estarem tão

bem definidos como no conjunto de dados *Circles*, tornando-o um problema mais complexo para o agrupamento de dados. Pela mesma razão e pelo facto de incluírem informação relacional, os métodos sem agrupamento de dados, Forças Direccionadas e Forças Direccionadas Aleatório, atingiram o agrupamento real com menos movimentações de instâncias que o Agrupamento de Dados com Restrições. Destaca-se também que o método Forças Direccionadas Aleatório, apesar de evoluir mais consistentemente que o método Forças Direccionadas, convergiu com uma ligeira desvantagem, o que demonstra uma influência positiva da heurística de selecção de instâncias "selecção da mais afastada" neste estudo. Sem surpresa, o método Manual teve o pior desempenho.

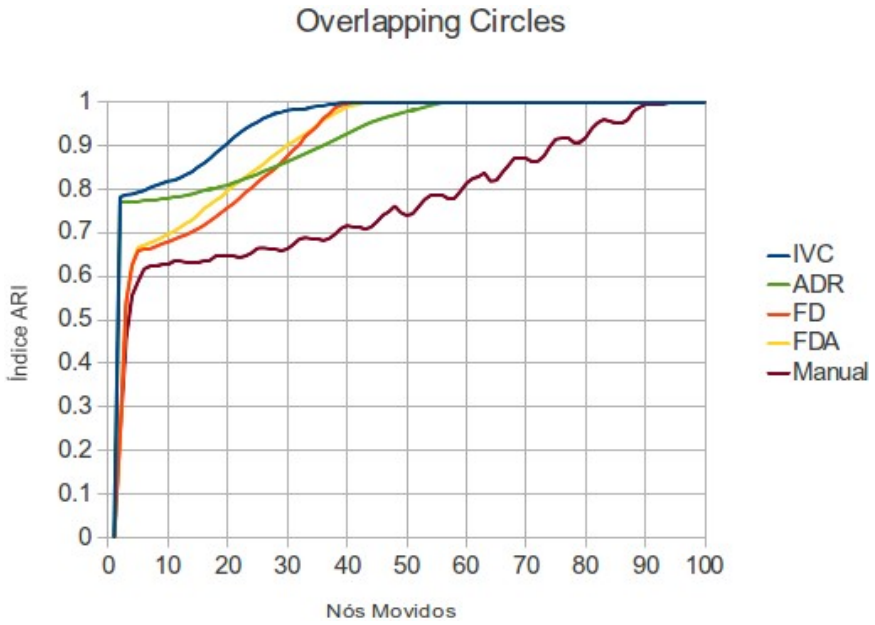



*Figura 44: Resultados experimentais no conjunto de dados Overlapping Circles*

É também analisada a influência das ligações relacionais no conjunto de dados *Overlapping Circles*.

| **Abordagem / Nós Movidos** | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 78.19% | 79.25% | 81.77% | 90.51% | 98.05% | 99.85% | 100.00% | 100.00% | 100.00% |
| IVC s/Inf. Relacional | 77.53% | 78.00% | 78.80% | 81.85% | 86.66% | 93.44% | 99.80% | 100.00% | 100.00% |
| FD | 24.29% | 65.82% | 67.85% | 75.72% | 87.55% | 99.56% | 100.00% | 100.00% | 100.00% |
| FD s/Inf. Relacional | 24.50% | 63.27% | 63.97% | 67.15% | 71.13% | 75.73% | 89.61% | 99.80% | 100.00% |
| FDA | 24.29% | 66.62% | 69.58% | 79.64% | 90.13% | 98.84% | 100.00% | 100.00% | 100.00% |
| FDA s/Inf. Relacional | 24.50% | 63.42% | 64.12% | 66.58% | 70.38% | 75.52% | 88.59% | 100.00% | 100.00% |

*Tabela 9: ARI nos 100 primeiros movimentos do conjunto de dados Overlapping Circles com ligações relacionais e sem ligações relacionais*

Como no conjunto de dados *Circles*, a utilização da versão do conjunto de dados sem ligações relacionais nas diferentes abordagens resulta num desempenho global dos métodos inferior. De notar que, os métodos Forças Direccionadas e Forças Direccionadas Aleatório apresentam um desempenho com evolução e resultados finais semelhantes. Esta situação verifica-se uma vez que, sem informação relacional, o critério de selecção das instâncias torna-se pouco relevante. O facto de ser movimentada a instância mais afastada ou ser movimentada outra instância qualquer para o seu grupo real sem que esta acção implique a movimentação de outras instâncias por "arrasto", não influencia a convergência para o agrupamento óptimo. A aproximação entre os métodos com e sem informação relacional, relativamente ao conjunto de dados anterior, indica também que as ligações relacionais criadas foram de pior qualidade. Esta é uma situação prevista, na medida em que os grupos se sobrepõem.

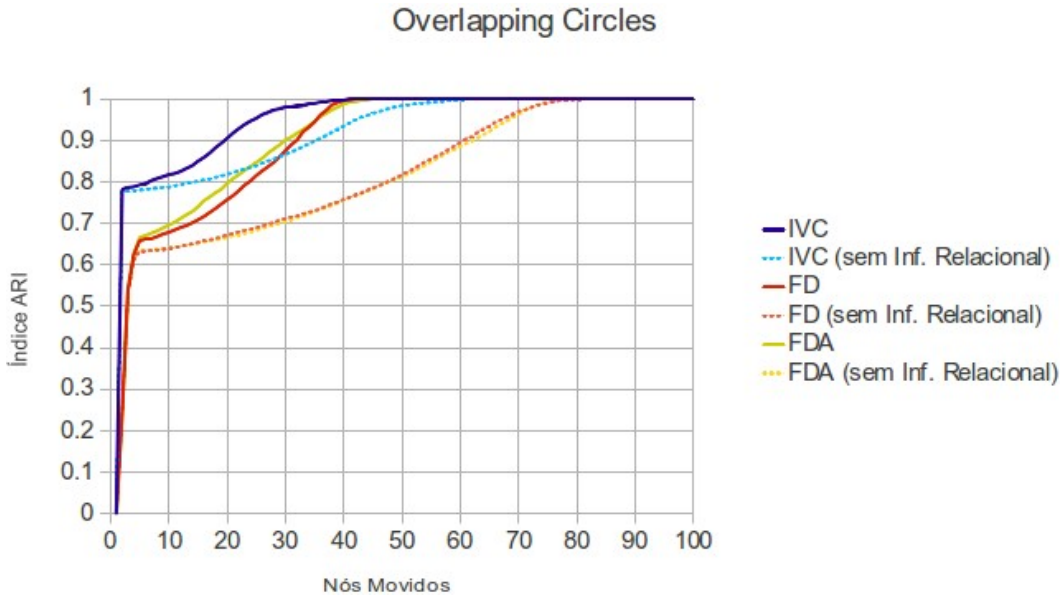



*Figura 45: Efeito das ligações relacionais no conjunto de dados Overlapping Circles*

### *5.4.3 Iris*

| Abordagem \ Nós Movidos | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 90.77% | 91.54% | **98.58%** | 99.03% | 99.46% | 100.00% | 100.00% | 100.00% | 100.00% |
| ADR | 88.71% | 89.99% | 94.88% | **98.99%** | 98.90% | 98.88% | 99.06% | 99.60% | 100.00% |
| FD | 32.91% | 56.22% | 56.63% | 66.38% | 80.13% | 91.35% | 100.00% | 100.00% | 100.00% |
| FDA | 32.89% | 55.95% | 57.04% | 66.13% | 80.05% | 93.65% | 100.00% | 100.00% | 100.00% |
| Manual | 34.03% | 52.15% | 55.90% | 56.15% | 57.74% | 62.07% | 67.09% | 77.37% | 90.44% |

*Tabela 10: ARI nos 100 primeiros movimentos do conjunto de dados Iris*

No gráfico da figura 46, a abordagem de Agrupamento de Dados Visual Interactivo evidencia também o melhor desempenho de todas as abordagens testadas no conjunto de dados *Iris*. Esta abordagem atinge um agrupamento quase perfeito muito rapidamente, com um ARI muito próximo de 1 ao fim de cerca de 10 movimentos. O Agrupamento de Dados com Restrições atingiu o mesmo patamar já com quase 20 instâncias movidas, evidenciando-se a influência positiva da heurística "selecção da mais afastada". As abordagens Forças Direccionadas e Forças Direccionadas Aleatório, sem agrupamento de dados, apresentam desempenhos semelhantes. A diferença na evolução entre as abordagens IVC e ADR e as abordagens FD e FDA indicam que, neste conjunto de dados, o agrupamento de dados tem uma influência significativamente mais positiva do que o *layout* na convergência para o agrupamento óptimo.

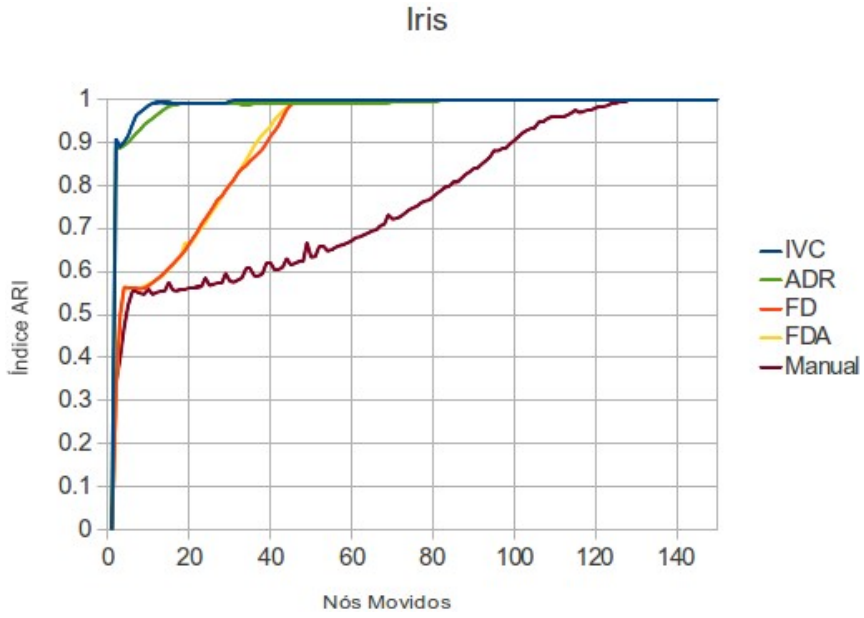



*Figura 46: Resultados experimentais no conjunto de dados Iris*

É apresentado de seguida o efeito das ligações relacionais no conjunto de dados *Iris*.

| Abordagem / Nós Movidos | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 90.77% | 91.54% | 98.58% | 99.03% | 99.46% | 100.00% | 100.00% | 100.00% | 100.00% |
| IVC s/Inf. Relacional | 87.28% | 86.96% | 88.72% | 88.92% | 87.70% | 87.19% | 91.28% | 96.77% | 100.00% |
| FD | 32.91% | 56.22% | 56.63% | 66.38% | 80.13% | 91.35% | 100.00% | 100.00% | 100.00% |
| FD s/Inf. Relacional | 32.93% | 55.48% | 55.61% | 56.29% | 57.58% | 60.20% | 67.36% | 77.20% | 83.21% |
| FDA | 32.89% | 55.95% | 57.04% | 66.13% | 80.05% | 93.65% | 100.00% | 100.00% | 100.00% |
| FDA s/Inf. Relacional | 32.89% | 55.50% | 55.55% | 56.49% | 58.20% | 60.83% | 68.56% | 78.44% | 90.64% |

*Tabela 11: ARI nos 100 primeiros movimentos do conjunto de dados Iris com ligações relacionais e sem ligações relacionais*

Também no conjunto de dados *Iris*, as ligações relacionais influenciam positivamente a evolução para o agrupamento real. Destaca-se contudo uma evolução significativamente mais rápida para a convergência no método Forças Direccionadas Aleatório sem Informação Relacional do que noutros conjuntos de dados. Este resultado não está de acordo com o esperado na medida em que as abordagens de *layout* apenas deverão diferir entre si em conjuntos de dados com ligações relacionais. Sem as ligações relacionais, o movimento de instâncias não implica o "arrastamento" de outras ligadas a si, tornado o critério de selecção de instâncias a mover pouco relevante. Esta questão específica será abordada em trabalho futuro.

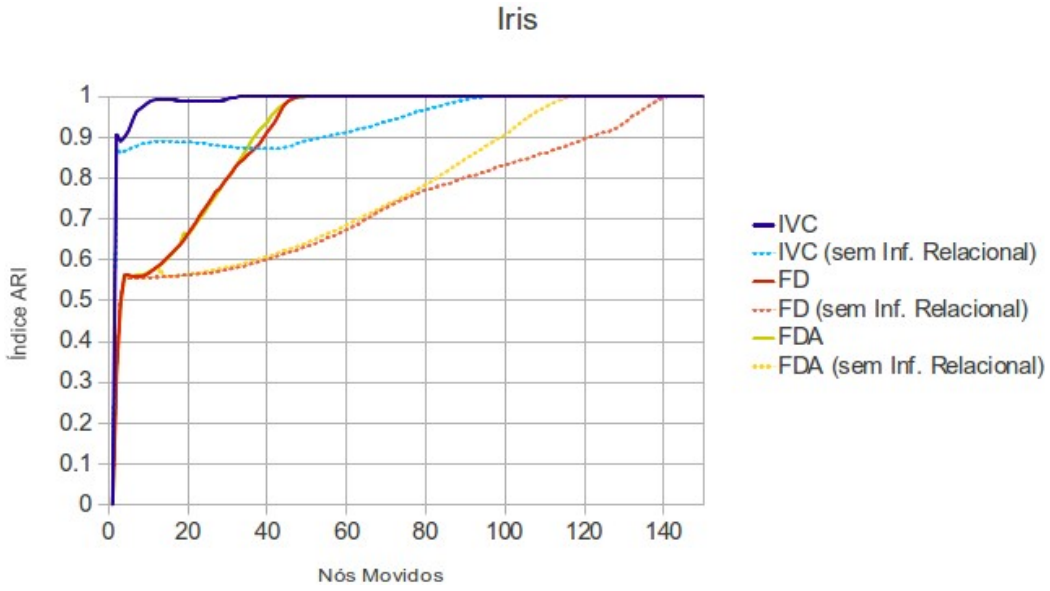



*Figura 47: Efeito das ligações relacionais no conjunto de dados Iris*

### *5.4.4 Cigar*

| Abordagem \ Nós Movidos | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 81.86% | 83.20% | 86.53% | 91.26% | 97.19% | 99.08% | **100.00%** | 100.00% | 100.00% |
| ADR | 83.32% | 87.49% | 86.31% | 84.17% | 82.89% | 82.49% | 85.39% | 91.99% | 96.91% |
| FD | 33.65% | 59.77% | 62.54% | 74.09% | 85.04% | 99.44% | **100.00%** | 100.00% | 100.00% |
| FDA | 33.67% | 59.56% | 61.72% | 71.48% | 83.81% | 97.99% | **100.00%** | 100.00% | 100.00% |
| Manual | 33.56% | 55.61% | 57.92% | 58.83% | 60.05% | 61.74% | 66.79% | 74.42% | 83.57% |

*Tabela 12: ARI nos 100 primeiros movimentos do conjunto de dados Cigar*

No conjunto de dados *Cigar* (figura 48), todas as abordagens requerem que muitas instâncias sejam movidas até que seja obtido um agrupamento correcto. A abordagem Agrupamento de Dados Visual Interactivo é a que evolui mais rapidamente, atingindo um ARI superior a 0,9 com menos de 20 instâncias movidas. Este desempenho nos primeiros movimentos é justificado pelo facto dos grupos estarem bem separados no espaço de atributos. Contudo, desde cerca das 5 instâncias movidas, até que atinge o agrupamento óptimo, o Agrupamento de Dados Visual Interactivo evolui de forma mais lenta. Este comportamento ocorre já que o número de instâncias é desequilibrado nos grupos, obrigando a um maior número de movimentos para atingir o agrupamento real dos dados. O Agrupamento de Dados com Restrições beneficia de uma evolução inicial semelhante. No entanto, a partir de cerca das 5 instâncias verifica-se uma regressão e uma lenta evolução para a convergência no agrupamento real. A ausência de ligações relacionais foi a responsável por este comportamento (em comparação com a abordagem anterior, este é o factor diferenciador). Os métodos de *layout*, Forças Direccionadas e Forças Direccionadas Aleatório tiveram, a seguir ao Agrupamento de Dados Visual Interactivo, o melhor desempenho neste conjunto de dados. Este facto indica uma influência positiva das ligações relacionais e da heurística "selecção da mais afastada". Contudo, o facto do Agrupamento de Dados Visual Interactivo não se apresentar com um desempenho muito superior às outras abordagens, está relacionado com os parâmetros de obtenção de restrições. Neste caso, uma distância para a captação de restrições de ligação obrigatória $\epsilon$ superior implicará que mais restrições serão criadas tirando partido das ligações relacionais (todas entre instâncias do mesmo grupo, conforme referido na secção 5.3.4), resultando num melhor desempenho desta abordagem.

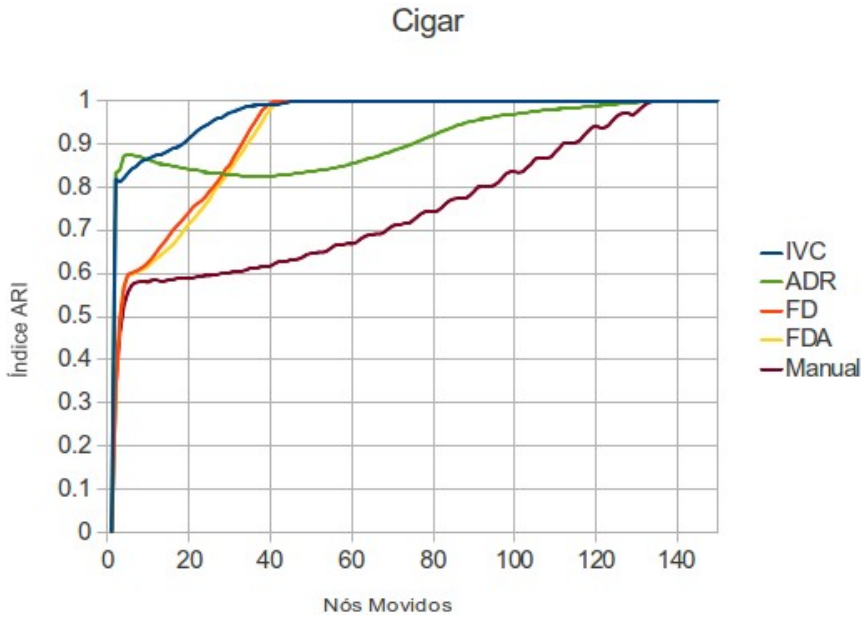



*Figura 48: Resultados experimentais no conjunto de dados Cigar*

É analisada de seguida a influência das ligações relacionais no conjunto de dados *Cigar*.

| **Abordagem / Nós Movidos** | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 81.86% | 83.20% | 86.53% | 91.26% | 97.19% | 99.08% | 100.00% | 100.00% | 100.00% |
| IVC s/Inf. Relacional | 81.70% | 86.52% | 88.24% | 88.61% | 87.67% | 86.95% | 89.77% | 95.80% | 99.58% |
| FD | 33.65% | 59.77% | 62.54% | 74.09% | 85.04% | 99.44% | 100.00% | 100.00% | 100.00% |
| FD s/Inf. Relacional | 33.65% | 58.15% | 58.31% | 59.60% | 62.43% | 64.85% | 70.57% | 77.78% | 84.37% |
| FDA | 33.67% | 59.56% | 61.72% | 71.48% | 83.81% | 97.99% | 100.00% | 100.00% | 100.00% |
| FDA s/Inf. Relacional | 33.63% | 58.11% | 58.29% | 59.06% | 60.64% | 62.85% | 68.51% | 75.49% | 83.07% |

*Tabela 13: ARI nos 100 primeiros movimentos do conjunto de dados Cigar com ligações relacionais e sem ligações relacionais*

Conforme esperado, também no conjunto de dados *Cigar* com informação relacional, as abordagens têm um melhor desempenho global. De notar, a especial influência das ligações relacionais no método Agrupamento de Dados Visual Interactivo onde o facto das instâncias seleccionadas "arrastarem" outras ligadas a si provocou a definição de restrições em maior número e com maior qualidade. Esta situação acontece essencialmente em conjuntos de dados com grupos desequilibrados do ponto de vista de número de instâncias.

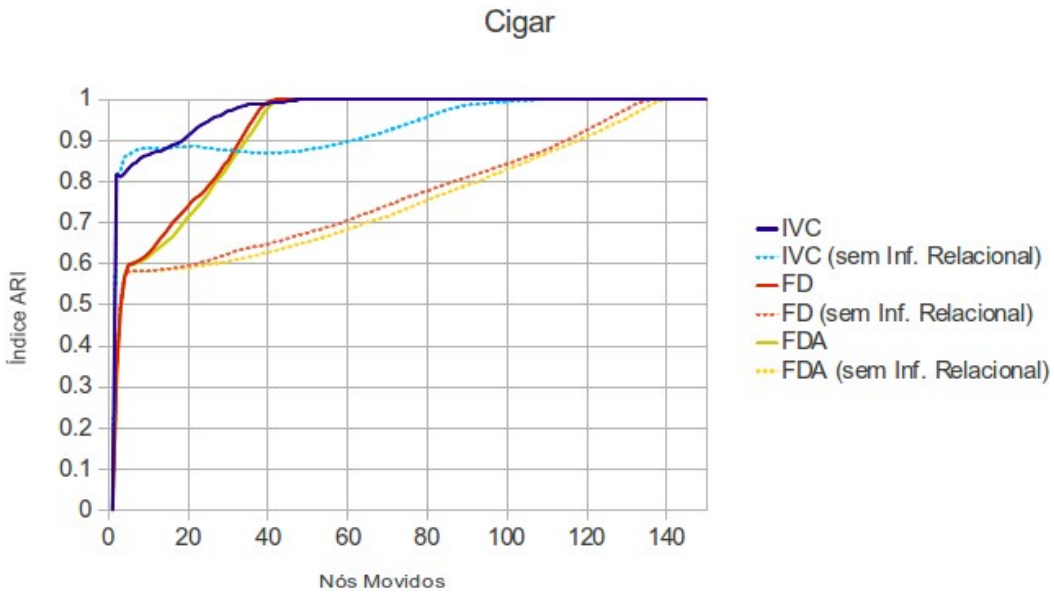



*Figura 49: Efeito das ligações relacionais no conjunto de dados Cigar*

### *5.4.5 Half Rings*

| Abordagem \ Nós Movidos | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | **83.11%** | **84.05%** | **84.84%** | **86.43%** | 90.54% | 96.29% | 100.00% | 100.00% | 100.00% |
| ADR | **82.97%** | **83.31%** | **84.09%** | **86.12%** | **86.42%** | **86.30%** | **88.54%** | 93.87% | 99.95% |
| FD | 33.57% | 56.76% | 59.17% | 76.36% | 86.76% | 98.81% | **100.00%** | 100.00% | 100.00% |
| FDA | 33.56% | 56.91% | 59.00% | 72.55% | 89.75% | 99.91% | **100.00%** | 100.00% | 100.00% |
| Manual | 33.58% | 51.78% | 54.84% | 55.38% | 57.02% | 58.15% | 63.73% | 72.06% | 83.62% |

*Tabela 14: ARI nos 100 primeiros movimentos do conjunto de dados Half Rings*

São apresentados na figura 50, os resultados experimentais no conjunto de dados *Half Rings*. Dado tratar-se de um problema de agrupamento de dados com alguma complexidade, os métodos Agrupamento de Dados Visual Interactivo e Agrupamento de Dados com Restrições tiveram uma evolução lenta entre os valores aproximados de ARI 0,8 e 0,9, apesar de serem os métodos com o menor número de instâncias movidas até então. A partir daí, as ligações relacionais aliadas à aplicação da heurística "selecção da mais afastada" permitiram uma rápida convergência do Agrupamento de Dados Visual Interactivo para o agrupamento esperado. O Agrupamento de Dados com Restrições voltou a demorar a convergir, mais uma vez, pela ausência de informação relacional. O método Forças Direccionadas Aleatório teve o melhor desempenho, muito próximo do método Forças Direccionadas, tendo obtido os grupos esperados com cerca de 40 instâncias

movidas. Ambos os métodos de *layout* evoluíram consistentemente, o que demonstra a forma como o *layout* é influenciado positivamente pela informação relacional neste conjunto de dados. Mais uma vez, o facto do Agrupamento de Dados Visual Interactivo não se apresentar com um desempenho superior estará relacionado com os parâmetros de dependência com o agrupamento de dados.

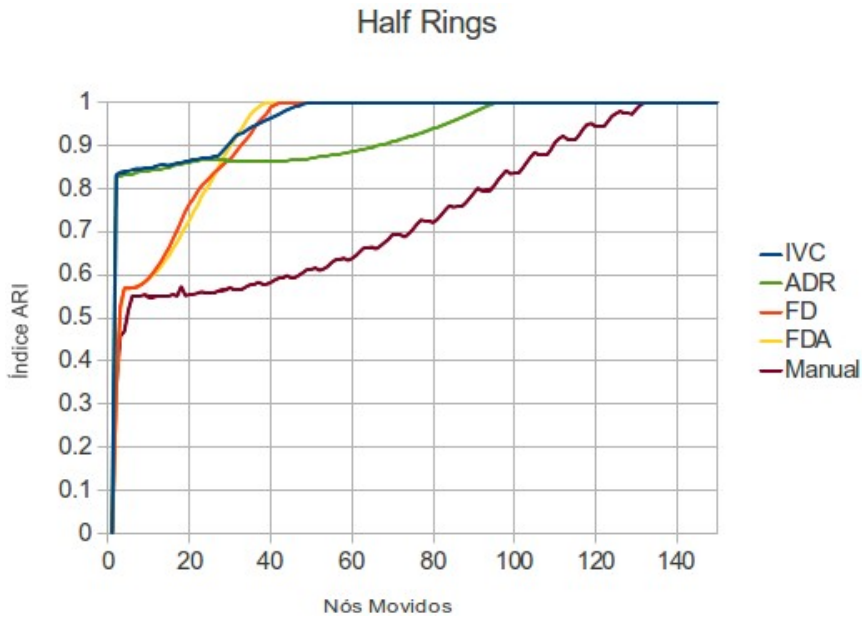



*Figura 50: Resultados experimentais no conjunto de dados Half Rings*

De seguida é analisado o efeito das ligações relacionais no conjunto de dados *Half Rings*.

| **Abordagem / Nós Movidos** | 2 | 5 | 10 | 20 | 30 | 40 | 60 | 80 | 100 |
|---|---|---|---|---|---|---|---|---|---|
| IVC | 83.11% | 84.05% | 84.84% | 86.43% | 90.54% | 96.29% | 100.00% | 100.00% | 100.00% |
| IVC s/Inf. Relacional | 82.44% | 82.79% | 82.07% | 80.86% | 80.10% | 79.51% | 84.10% | 93.18% | 98.49% |
| FD | 33.57% | 56.76% | 59.17% | 76.36% | 86.76% | 98.81% | 100.00% | 100.00% | 100.00% |
| FD s/Inf. Relacional | 33.64% | 55.45% | 55.79% | 56.61% | 58.16% | 60.76% | 68.58% | 76.25% | 82.50% |
| FDA | 33.56% | 56.91% | 59.00% | 72.55% | 89.75% | 99.91% | 100.00% | 100.00% | 100.00% |
| FDA s/Inf. Relacional | 33.65% | 55.60% | 55.93% | 56.97% | 58.81% | 61.70% | 68.40% | 77.35% | 88.33% |

*Tabela 15: ARI nos 100 primeiros movimentos do conjunto de dados Half Rings com ligações relacionais e sem ligações relacionais*

Uma vez mais, como é visível no gráfico da figura 51 as ligações relacionais têm uma influência positiva nas abordagens estudadas. Destaca-se contudo o fenómeno não esperado da diferenciação de resultados entre os métodos Forças Direccionadas e Forças Direccionadas Aleatório, com

vantagem do último. Este efeito é também verificado no conjunto de dados Iris (secção 5.4.3) e será alvo de análise futura. No conjunto de dados Half Rings nota-se também uma aproximação maior do desempenho entre os métodos com e sem informação relacional, do que nos conjuntos de dados *Cigar*, *Iris* e *Circles*. O facto dos grupos não se encontrarem claramente separados neste conjunto de dados levam à criação de ligações relacionais com menor qualidade, o que prejudica os algoritmos que utilizam restrições baseadas na posição das instâncias: Agrupamento de Dados Visual Interactivo e Agrupamento de Dados com Restrições.

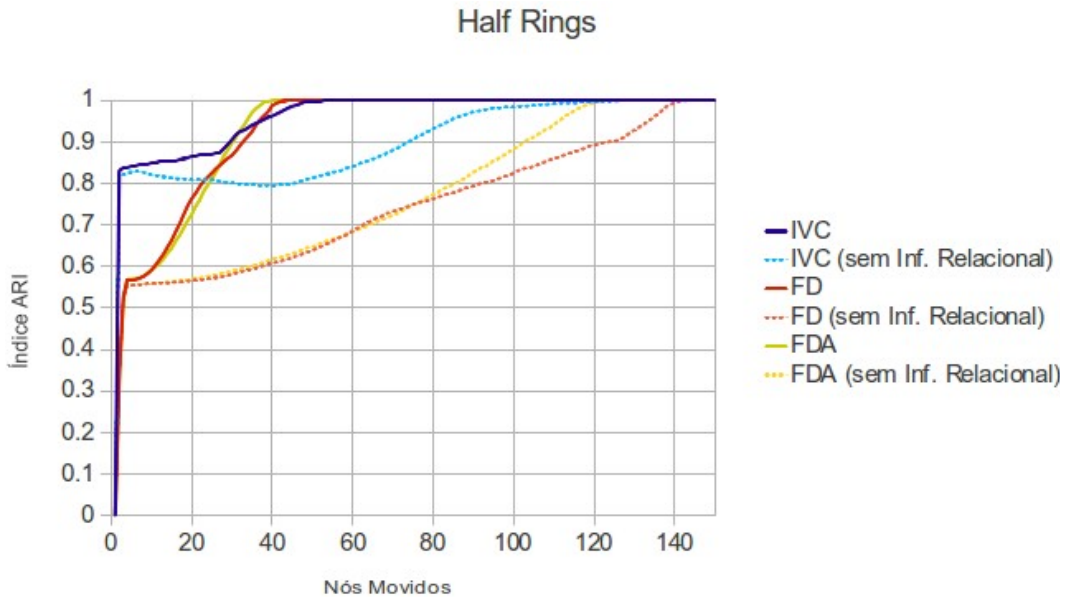



*Figura 51: Efeito das ligações relacionais no conjunto de dados Half Rings*

## *5.5 Sumário*

Neste capítulo foi apresentado um estudo comparativo que teve como objectivo avaliar o desempenho do Agrupamento de Dados Visual Interactivo, apresentado no capítulo 4, e de outras abordagens de interacção com o utilizador. Foram apresentados os cinco conjuntos de dados utilizados no estudo, assim como os resultados das experiências efectuadas e as respectivas considerações. Na generalidade, estes resultados confirmam a teoria de que a abordagem do Agrupamento de Dados Visual Interactivo aumenta o desempenho do agrupamento de dados, quando comparado com os outros paradigmas. Foram, contudo, identificadas algumas questões que serão alvo de trabalho futuro.

# 6 Conclusões

## *6.1 Resumo*

Nesta dissertação estudou-se o Agrupamento de Dados Visual Interactivo, uma abordagem inovadora que permite a exploração de conjuntos de dados relacionais de forma interactiva, produzindo um agrupamento de dados que satisfaça os objectivos pretendidos pelo utilizador. Esta abordagem baseia-se na combinação de técnicas de visualização de informação, interacção com o utilizador e de agrupamento de dados.

No contexto da *visualização de informação* foram apresentados conceitos base como: a *percepção*, que consiste na aptidão humana para absorver informação e extrair o conhecimento de alto nível de uma determinada visualização; o *desenho*, que representa a transformação da informação abstracta em informação perceptível; e a *pipeline de visualização*, pela qual se rege a maioria das abordagens de visualização e que consiste na representação dos passos do processo computacional de conversão da informação não estruturada numa forma visual passível de interacção com os utilizadores.

Foram referidos alguns métodos representativos das diferentes *estruturas de informação* aplicadas no *mapeamento visual* do processo de visualização de informação, como a *estrutura tabular*, que representa a informação em tabelas, a *estrutura espacial e temporal*, utilizada na representação de informação em *n* dimensões (até 3), a *estrutura em grafo*, que enfatiza as ligações entre as instâncias de dados e a *estrutura de colecção de texto e documentos*, que consiste na colecção arbitrária de documentos, normalmente textuais.

Associadas ao processo de visualização de informação estão ainda múltiplas estratégias que possibilitam a representação visual e organização de grandes conjuntos de dados, muitas vezes recorrendo à interacção com o utilizador. São descritas estratégias de *representação de dados*, de *navegação* e de *interacção*.

# 6 CONCLUSÕES

O *agrupamento de dados* foi introduzido como uma área da aprendizagem automática, enquadrando-se na aprendizagem não supervisionada. Foram apresentadas as diferentes *fases do agrupamento de dados*, bem como as diferentes abordagens de agrupamento de dados: *abordagens de partição*, *abordagens hierárquicas* e abordagens *baseadas em densidade*, *baseadas em grelha* e *baseadas em modelos* matemáticos.

Foi destacado o *agrupamento de dados com restrições*, um campo da aprendizagem semi-supervisionada em que é utilizada informação *a priori* na forma de restrições. Esta informação é incorporada no agrupamento de dados com o intuito de proporcionar soluções adaptadas a tarefas ou interesses específicos. Existem vários níveis de restrições, desde as *restrições globais* que se aplicam a todo o conjunto de dados, passando pelas *restrições ao nível dos grupos* e ao *nível dos atributos*, até restrições a níveis mais específicos, as *restrições ao nível dos objectos*. As restrições ao nível dos objectos são as mais abordadas dado que grande parte das restrições dos níveis superiores podem ser expressas neste tipo de restrições.

Foram apresentados os principais algoritmos de agrupamento de dados que usam restrições, organizando-os em cinco categorias distintas: de *restrições invioláveis*, que garantem que as soluções encontradas satisfazem completamente todas as restrições; de *restrições na forma de rótulos*, em que a um subconjunto dos objectos de dados se encontram associados rótulos, servindo normalmente esses rótulos para inicializar os centros dos grupos dos algoritmos de agrupamento de dados de partição; de *penalização de violações de restrições*, em que para além de considerarem as distâncias entre os objectos de dados e os respectivos centros de grupo, é geralmente adaptada uma função-objectivo para penalizar a violação de restrições, não sendo necessário que todas as restrições sejam satisfeitas; de *edição de distância*, que aprendem uma medida de distância com o intuito de generalizar as restrições entre objectos de dados ao nível do espaço dos atributos de dados, propagando as restrições entre pares de objectos a outros objectos próximos, que podem não ter sido incluídos nos conjuntos de restrições; e, finalmente, de *modificação do processo de geração*, que pressupõem que os objectos de dados foram gerados segundo um modelo probabilístico, sendo esse modelo modificado com o objectivo de se considerar relações entre objectos na estimação dos parâmetros do modelo.

Como tema central desta dissertação foi apresentado o *Agrupamento de Dados Visual Interactivo*, uma abordagem inovadora que permite ao utilizador explorar conjuntos de dados relacionais de forma interactiva, com o intuito de produzir um agrupamento que satisfaça os seus objectivos e interesses. Esta abordagem é conseguida combinando técnicas de visualização de informação e agrupamento de dados com restrições. As forças direccionadas *spring embedded* sofrem a influência das ligações relacionais e das ligações do agrupamento obtidas a partir do agrupamento de dados, convergindo para um equilíbrio das forças. São também aplicadas técnicas de interacção com o utilizador que permitem que este efectue a associação de objectos a grupos, de forma a alimentar o conjunto de restrições de um algoritmo de agrupamento de dados que será executado iterativamente até se atingir o agrupamento final.

São apresentados os principais passos do Agrupamento de Dados Visual Interactivo que consistem na: *inicialização da visualização*, em que o desenho *spring embedded* é inicializado na geração da visualização inicial; *interpretação das acções do utilizador*, onde a movimentação de instâncias pelo utilizador é traduzida em relações de pertença de objectos a grupos, isto é, restrições de ligação obrigatória e de ligação proibida; execução do *agrupamento de dados com restrições*, que utiliza como base as restrições definidas no passo anterior para representar o conhecimento *a priori* do utilizador; *actualização da visualização*, que representa visualmente o agrupamento obtido de forma a facilitar a percepção deste pelo utilizador.

Finalmente, foi realizado um estudo comparativo entre diferentes abordagens de interacção com o utilizador com o objectivo de avaliar os desempenhos destes na convergência para o agrupamento real dos dados, em função do número de instâncias movidas pelo utilizador. Os desempenhos das diferentes abordagens foram avaliados em cinco conjuntos de dados. Na generalidade, concluiu-se que o método de Agrupamento de Dados Visual Interactivo estudado nesta dissertação apresentou um desempenho superior às outras abordagens. Contudo, nos conjuntos de dados *Cigar* e *Half Rings*, identificou-se uma possível melhoria a aplicar a este método com o objectivo de aumentar o seu desempenho.

## *6.2 Objectivos Alcançados*

O primeiro objectivo desta dissertação consistiu na revisão do estado da arte da visualização de informação, tendo sido estudadas as diferentes fases de um processo de visualização e apresentados os conceitos base da área. Foram detalhadas as diferentes estruturas de informação passíveis de representação visual, assim como alguns métodos representativos de cada uma delas. As estratégias de interacção com o utilizador e representação de dados foram também estudadas com o intuito de perceber a sua utilidade na optimização da apresentação da informação ao utilizador, uma vez mais, em conjunto com abordagens de cada uma das estratégias apresentadas.

Foi também estudado o tema do agrupamento de dados, segundo objectivo desta dissertação, com foco especial no agrupamento de dados com restrições, tendo sido apresentados os vários tipos de restrições usadas no agrupamento de dados e os principais algoritmos de agrupamento de dados com restrições. Foram também apresentados os conceitos fundamentais das aprendizagens supervisionada, não supervisionada e semi-supervisionada com o intuito de enquadrar o agrupamento de dados com restrições na aprendizagem automática.

O estudo dos dois primeiros temas teve como propósito a realização do terceiro objectivo da dissertação, que consistiu no estudo e implementação da abordagem de Agrupamento de Dados Visual Interactivo, que combina técnicas desses dois temas apresentados. Esta proposta pretende resolver o problema da adequação do resultado de agrupamentos de dados aos interesses e tarefas específicos de um utilizador, independentemente do conjunto de dados. Com esta finalidade, o utilizador é consultado interactivamente para indicar relações de pertença (ou não pertença) de objectos a grupos, que representam o conhecimento do utilizador sobre o domínio de dados. Estas operações traduzem-se na geração de restrições que alimentam o agrupamento de dados com restrições aplicado a esses conjuntos de dados. O efeito pretendido é que o agrupamento de dados final seja atingido com o mínimo de informação *a priori* necessária, optimizando-se desta forma processos de rotulação de objectos, em muitas situações dispendiosos e morosos.

O quarto objectivo desta dissertação consistiu na avaliação do desempenho de cinco abordagens de

interacção com o utilizador. De forma a avaliar a rapidez da convergência para um agrupamento-alvo foi comparado o desempenho de diferentes abordagens com diferentes características (desenho – *layout*, agrupamento de dados – *clustering* e heurísticas de simulação de utilizadores). Esta comparação foi aplicada a duas versões de cinco conjuntos de dados: uma com ligações relacionais entre as instâncias e outra sem esta informação.

Finalmente, o último objectivo deste trabalho teve como propósito a implementação de uma plataforma de aplicação das abordagens de interacção com o utilizador estudadas nesta dissertação. Esta plataforma possibilita o uso de uma das diferentes abordagens em diferentes conjuntos de dados, permitindo também a extracção de informação estatística relativamente à eficácia das abordagens aplicadas. A plataforma tem a denominação de Preka e foi utilizada como base para os estudos apresentados nesta dissertação, encontrando-se disponível em [165].

## *6.3 Limitações e Trabalho Futuro*

Nesta dissertação foi demonstrado que o Agrupamento de Dados Visual Interactivo melhora o desempenho do agrupamento de dados, conseguido pela integração de técnicas de desenho de forças direccionadas *spring embedded* com o agrupamento de dados com restrições e com a interacção do utilizador.

Foram, no entanto, identificadas algumas oportunidades de melhoria nas diferentes fases do Agrupamento de Dados Visual Interactivo:

**Inicialização da Visualização**

- Foi utilizada neste estudo uma dimensão de ecrã fixa, independentemente do conjunto de dados utilizado. Será interessante o desenvolvimento de métodos de ajuste dinâmico da área de visualização, tendo em conta o número de instâncias a apresentar, as distâncias para a criação de restrições e o número de grupos (quando possível).

- A geração de informação relacional nos conjuntos de dados foi efectuada com recurso à heurística "vizinho mais próximo". Ainda que esta heurística faça sentido em conjuntos de dados sem informação relacional, não foram efectuados estudos em conjuntos de dados com informação relacional "real".

**Interpretação das Acções do Utilizador**

- A obtenção de restrições baseou-se em medidas fixas ($\epsilon$ e $\delta$) para todos os conjuntos de dados. Conforme descrito nos resultados dos conjuntos de dados *Cigar* e *Half Rings*, estas distâncias terão condicionado o desempenho do Agrupamento de Dados Visual Interactivo. Será interessante o estudo de abordagens para determinação dinâmica destes parâmetros que garantam o melhor desempenho possível em determinado conjunto de dados.

**Aplicação do Agrupamento de Dados com Restrições**

- Os resultados apresentados consideram a aplicação do algoritmo MPC $K$-Médias. Será interessante avaliar estas abordagens com outros algoritmos de agrupamento de dados com restrições, que utilizem diferentes métricas além da euclidiana (por exemplo, Hamming para atributos binários e co-seno para texto).

**Simulação do Utilizador**

- A técnica de Agrupamento de Dados Visual Interactivo é direccionada a utilizadores humanos. Nesta dissertação, por escassez de recursos (pessoas e tempo), foram implementadas heurísticas de simulação de utilizadores nos resultados apresentados. Será pertinente a avaliação de resultados com recurso a utilizadores humanos.
- Foi observado que a heurística de movimentação de instâncias "selecção da mais afastada" tende a trocar instâncias entre os grupos posicionados em cantos opostos da visualização (fazendo um movimento diagonal, por se encontrarem mais afastados). Para reduzir este efeito nos grupos dos cantos não opostos da visualização, foi utilizada uma área de visualização quadrada neste estudo. Será interessante melhorar esta heurística no sentido de tornar os centros dos grupos equidistantes.

Será também interessante, o alargamento do estudo desta abordagem a outros conjuntos de dados com maior dimensionalidade, com diferentes tipos de atributos e com informação relacional "real".

Apesar do sucesso dos resultados obtidos, a abordagem descrita constitui apenas um primeiro passo na direcção de uma abordagem de agrupamento de dados mais centrada nos interesses do utilizador. Uma das possíveis direcções futuras será o de desenhar uma abordagem de agrupamento de dados mais interactiva do que as abordagens já existentes.

Desta forma, para além das sugestões apresentadas anteriormente, no sentido de superar as limitações identificadas, será interessante no futuro conduzir este trabalho, em âmbito de doutoramento, nas seguintes direcções:

- Estudo de métodos para o mapeamento de atributos de objectos de conjuntos de dados em propriedades visuais de objectos em visualizações que permitam uma representação visual simples de conjuntos de dados multidimensionais
- Exploração de novas heurísticas de selecção e posicionamento de instâncias que acelerem a convergência para os agrupamentos reais dos dados
- Integração de novas abordagens de agrupamento de dados com restrições que agrupem os dados simultaneamente no espaço de atributos e no espaço relacional
- Inclusão de pesos nas ligações de agrupamento de dados que diferenciem ligações de maior ou menor intensidade na visualização
- Combinação de abordagens guiadas pelo utilizador (como o Agrupamento de Dados Visual Interactivo) com métodos de selecção activa de restrições guiados pelo sistema
- Garantir a evolução da plataforma Preka pela implementação de inovações neste suporte

# Bibliografia


1. Keim, D.A.; Mansmann, F., and Schneidewind, J., and Ziegler, H., *Challenges in Visual Data Analysis,* Proceedings of Information Visualization (IV 2006), IEEE, p. 9-16, 2006.
2. Ware, C. (2004). *Information Visualization: Perception for Design*, Morgan Kaufmann.
3. Card, S., Mackinlay, J., Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann.
4. Shneiderman, B., Plaisant, C. (2005). *Designing the User Interface: Strategies for effective human-computer interaction*, Fourth Edition, Addison-Wesley.
5. Wehrend, S., and Lewis, C. (1990). *A problem-oriented classification of visualization techniques*. Proceedings of IEEE Visualization (Vis'90), p. 139-143.
6. Wickens, C., and Hollands, J. (2000). *Engineering Psychology and Human Performance*, Prentice-Hall.
7. Zhou, M., and Feiner, S. (1998). *Visual task characterization for automated visual discourse synthesis.* Proceedings of the ACM Human Factors in Computing Systems Conference (CHI'98), p. 392-399.
8. Rosenblum L., Earnshaw R., Encarnação J., Hagen H., Kaufman A., Klimenko S., Nielson G., Post F., Thalmann D. (1994). *Scientific Visualization: Advances and Challenges*. Academic Press in association with IEEE Computer Society, USA.
9. Cleveland, W. (1993). *Visualizing Data*, Hobart Press.
10. Gillan, D., Wickens, C., Hollands, J., Carswell,C. (1998) *Guidelines for presenting quantitative data in HFES publications*. Human Factors 36, p. 419-440.
11. Tufte, E. (2001). *The Visual Display of Quantitative Information*, Graphics Press, 2nd edition.
12. Wilkinson, L. (1999). *The Grammar of Graphics*, New York: Springer-Verlag.
13. Rosson, M.B., Carroll J. (2001) *Usability Engineering: Scenario-based Development of Human Computer Interaction.* Morgan Kaufmann.
14. Plaisant, C. (2004). *The Challenge of Information Visualization Evaluation.* Proceedings of Advanced Visual Interfaces (AVI'04), p. 109 – 116.

15. Chen, C., and Yu, Y. (2000). *Empirical studies of information visualization: a meta-analysis.* International J. Human-Computer Studies, 53(5): 851-866.

16. Tory, M., Möller, T. (2004). *Human Factors in Visualization Research.* IEEE Transactions on Visualization and Computer Graphics, TVCG 10(1): 72-84.

17. Saraiya, P., North, C., Duca, K., (2004). *An Evaluation of Microarray Visualization Tools for Biological Insight.* Proceedings IEEE Symposium on Information Visualization 2004, p. 1-8.

18. Fayyad, U.M., Grinstein, G., Wierse, A., Fayyad, U. (2001). *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann.

19. Bertin, J.B. (1983). *Semiology of Graphics: Diagrams, Networks, Maps*, The University of Wisconsin Press.

20. Healey, C. G., Booth, K. S., and Enns, J. T. (1996). *High-Speed Visual Estimation Using Preattentive Processing.* ACM Transactions on Human Computer Interaction 3(2): 107-135.

21. Carswell, C. (1992). *Reading graphs: Interactions of processing requirements and stimulus structure.* In B.Burns (Ed.), Percepts, Concepts and Categories: The Representation and Processing of Information, p. 605-645. Amsterdam: Elsevier Science Publishers.

22. Chewar, C. M., McCrickard, D. S., Ndiwalana, A., North, C., Pryor, J., and Tessendorf, D. (2002). *Secondary task display attributes: optimizing visualizations for cognitive task suitability and interference avoidance.* Proceedings of the Symposium on Data Visualization 2002, p. 165-171, Barcelona, Spain.

23. Cleveland, W. S., and McGill, R. (1984). *Graphical Perception: Theory, Experimentation and Application to the Development of Graphical Methods.* Journal of American Statistical Association 79(387): 531-554.

24. Nowell, L., Schulman, R., Hix, D. (2002). *Graphical encoding for information visualization: an empirical study.* IEEE Symposium on Information Visualization, 2002, p. 43- 50.

25. Mackinlay, J. (1986). *Automating the design of graphical presentations of relational information*, ACM Transactions on Graphics, 5(2): 110-141.

26. Roth, S.F., Kolojejchick, J., Mattis, J., and Goldstein, J. (1994). *Interactive Graphic Design Using Automatic Presentation Knowledge.* Proceedings of the Conference on Human Factors in

Computing Systems (SIGCHI '94), Boston, MA, p. 112-117.

27. Ahlberg, C., Wistrand, E. (1995). *IVEE: An Information Visualization and Exploration Environment.* Symposium on Information Visualization 1995, Atlanta, GA, p. 66-73.
28. Spence, R. (2001). *Information Visualization*, Addison-Wesley.
29. Rao, R., Card, S. (1994). *The Table Lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information.* Proceedings of the SIGCHI conference on Human Factors in Computing Systems '94, p. 318-322.
30. Inselberg, A. (1997). *Multidimensional detective.* Proceedings IEEE Symposium on Information Visualization '97, p. 100-107.
31. Kandogan, E., (2000). *Star coordinates: a multi-dimensional visualization technique with uniform treatment of dimensions.* LBHT Proceedings IEEE Symposium on Information Visualization 2000, p. 9-12.
32. Miller, J. (2004). *Daisy Analytics*, http://www.daisy.co.uk/.
33. Tory, M., Möller, T. (2004). *Rethinking Visualization: A High-Level Taxonomy*. Proceedings of IEEE Symposium on Information Visualization (InfoVis) 2004, pp 151- 158.
34. Malinowski, S. (2004). *Music Animation Machine*, http://www.musanim.com/.
35. Plaisant, C., Milash, B., Rose, A., Widoff, S., Shneiderman, B. (1996). *LifeLines: Visualizing personal histories.* Proceedings of the SIGCHI conference on Human Factors in Computing Systems '96, p. 221-227.
36. Carlis, J., Konstan, J. (1998). *Interactive visualization of serial periodic data.* Proceedings ACM User Interface Software and Technology, UIST '98, p. 29-38.
37. Stoakley, R., Conway, M. J., and Pausch, R. (1995). *Virtual reality on a WIM: Interactive worlds in miniature.* Proceedings of the Conference on Human Factors in Computing Systems '95, p. 265-272.
38. North, C., Shneiderman, B., Plaisant, C. (1996). *User Controlled Overviews of an Image Library: A Case Study of the Visible Human.* Proceedings ACM Digital Libraries '96 Conference, p. 74-82.
39. Kniss, J., Kindlmann, G., and Hansen, C. (2001) *Interactive volume rendering using multi-dimensional transfer functions and direct manipulation widgets.* In

Proceedings of IEEE Visualization, p. 255-262, San Diego, CA.

40. Mihalisin, T., Timlin, J., Schwegler, J. (1991). *Visualizing multivariate functions, data and distributions.* IEEE Computer Graphics and Applications, 11(3): 28-35.

41. van Wijk, J., van Liere, R. (1993). *HyperSlice: visualization of scalar functions of many variables.* Proceedings IEEE Visualization '93, p. 119-125.

42. Furnas, G., and Zacks, J. (1994). *Multitrees: enriching and reusing hierarchical structure.* In Conference proceedings on Human Factors in Computing Systems (CHI `94), p. 330-336.

43. Robertson, G., Cameron, K., Czerwinski, M., and Robbins, R., (2002). *Polyarchy visualization: visualizing multiple intersecting hierarchies.* Proceedings of the SIGCHI conference on Human factors in computing systems, p. 423-430, Minneapolis, Minnesota.

44. Grosjean, J., Plaisant, C., Bederson, B. (2002). *SpaceTree: Supporting Exploration in Large Node Link Tree, Design Evolution and Empirical Evaluation.* Proceedings of IEEE Symposium on Information Visualization, p. 57-64, Boston.

45. Lamping, J., Rao, R., and Pirolli, P. (1995). *A focus+context technique based on hyperbolic geometry for visualizing large hierarchies.* Proceedings of the SIGCHI conference on Human Factors in Computing Systems, p. 401-408, Denver, Colorado.

46. Robertson, G., Card, S., Mackinlay, J. (1993). *Information visualization using 3D interactive animation*, Communications of the ACM, v.36 n.4, p.57-71.

47. Wiss, U., Carr, D.A. (1999). *An empirical study of task support in 3D information visualizations.* Proceedings IEEE International Conference on Information Visualization, p. 392-399.

48. Johnson, B., Shneiderman, B. (1991). *Treemaps: a space-filling approach to the visualization of hierarchical information structures.* Proceedings of the 2nd International IEEE Visualization Conference, p. 284-291, San Diego.

49. Bederson, B., Shneiderman, B., and Wattenberg, M. (October 2002). *Ordered and Quantum Treemaps: Making Effective Use of 2D Space to Display Hierarchies.* ACM Transactions on Graphics, 21(4):833-854.

50. Stasko, J., Catrambone, R., Guzdial, M., Mcdonald, K. (2000). *An evaluation of*

*space-filling information visualizations for depicting hierarchical structures.* Int. J. Human-Computer Studies, 53(5): 663-694.

51. Herman, I., Melancon, G., and Marshall, M. S. (2000). *Graph visualization and navigation in information visualization: A survey.* IEEE Transactions on Visualization and Computer Graphics, 6(1):24-43.
52. Ware, C., Purchase, H., Colpoys, L., and McGill, M. (2002). *Cognitive measurements of graph aesthetics.* Information Visualization, 1(2), 103-110.
53. Becker, R. A., Eick, S. G., and Wilks, A. R. (1995). *Visualizing network data.* IEEE Transactions on Visualization and Graphics, 1(1):16-28.
54. Munzner, T. (1998). Exploring Large Graphs in 3D Hyperbolic Space. *IEEE Computer Graphics and Applications*, 18(4): 18-23.
55. Andrews, K., Kappe, F., and Maurer, H. (1995). Hyper-G and harmony: towards the next generation of networked information technology. *Conference companion on Human factors in computing systems 1995*, 33-34, Denver, Colorado.
56. Hetzler, B., Harris, W.M., Havre, S., Whitney, P. (1998). Visualizing the Full Spectrum of Document Relationships. In *Structures and Relations in Knowledge Organization. Proc. 5th Int. ISKO Conf.* Wurzburg: ERGON Verlag, pp. 168-175.
57. Lin, X. (1992). *Visualization for the document space*. Proceedings of the 3rd conference on Visualization '92, p. 274-281, Boston, Massachusetts.
58. Wise, J., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., Crow, V. (1995). *Visualizing The Non-Visual: Spatial Analysis And Interaction With Information From Text Documents*. Proceedings of IEEE Symposium on Information Visualization, p. 51-58.
59. Korfhage, R. (1995). *VIBE: Visual Information Browsing Environment*. SIGIR 95, p. 363.
60. Hearst, M. (1995). *TileBars: Visualization of Term Distribution Information in Full Text Information Access.* Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, p. 59-66, Denver, CO.
61. Card, S., Robertson, G., York, W. (1996). *The WebBook and the Web Forager: an information workspace for the World-Wide Web*. Proceedings of the SIGCHI conference on Human factors in computing systems 1996, p. 111,

Vancouver, British Columbia.

62. Robertson, G., Czerwinski, M. Larson, K., Robbins, D., Thiel, D., van Dantzich, M. (1998). *Data mountain: using spatial memory for document management*. Proceedings of the 11th annual ACM symposium on User Interface Software and Technology, p. 153-162, San Francisco, California.
63. Shneiderman, B., Plaisant, C. (2005). Designing the User Interface: Strategies for effective human-computer interaction, Fourth Edition, Addison-Wesley.
64. Hornbæk, K., Bederson, B., and Plaisant, C. (2002). Navigation patterns and usability of zoomable user interfaces with and without an overview. ACM Trans. Comput.-Hum. Interact, 9(4): 362-389
65. Stolte, C., Tang, D., and Hanrahan, P. (2002). Polaris: A System for Query, Analysis and Visualization of Multi-dimensional Relational Databases. IEEE Transactions on Visualization and Computer Graphics, 8(1): 52-65
66. Yang, J., Ward, M., and Rundensteiner E. (2003). Interactive Hierarchical Displays: A General Framework for Visualization and Exploration of Large Multivariate Data Sets. Computers and Graphics Journal, 27(2): 265-283
67. Conklin, N., Prabhakar, S., North, C. (2002). Multiple Foci Drill-Down through Tuple and Attribute Aggregation Polyarchies in Tabular Data. Proc. IEEE Symposium on Information Visualization 2002, pg. 131-134
68. Rayson, R. (1999). Aggregate Towers: Scale Sensitive Visualization and Decluttering of Geospatial Data. Proceedings of the 1999 IEEE Symposium on Information Visualization, pp 92-99.
69. Rencher, A. (2002). Methods of Multivariate Analysis, Second Edition, Wiley.
70. J. Shlens (2009). A Tutorial on Principal Component Analysis. Center for Neural Science, New York University.
71. Van Deun, K., Delbeke, L. (2000). Multidimensional Scaling, http://www.mathpsyc.uni-bonn.de/doc/delbeke/delbeke.htm. University of Leuven, Belgium.
72. Woodruff, A., Landay, J., Stonebraker, M. (1998). Constant Information Density in Zoomable Interfaces. Proceedings of the Advanced visual interfaces-- AVI'98, pp 57 – 65, L'Aquila, Italy.
73. Fishkin, K., and Stone, M. (1995). Enhanced Dynamic Queries via Movable

Filters. ACM Conference on Human Factors in Computing Systems, Denver, Colorado, pp. 415-420.

74. Eick, S. G., Steffen, J. L., and Sumner, E. E. Jr. (1992). SeeSoft---A Tool for Visualizing Line Oriented Software Statistics. IEEE Transactions on Software Engineering, 18 (11): 957—968.

75. Keim, D., Hao, M., Dayal, U., Hsu, M. (2002). Pixel bar charts: a visualization technique for very large multi-attribute data sets. Information Visualization, 1(1): 20—34.

76. Jerding, D. F., and Stasko, J. T. (1998). The Information Mural: A Technique for Displaying and Navigating Large Information Spaces. IEEE Transactions on Visualization and Computer Graphics, 4(3): 257—271.

77. Bederson, B.B., Hollan, J., Perlin, K., Meyer, J., Bacon, D., and Furnas, G. (1996). Pad++: A Zoomable Graphical Sketchpad for Exploring Alternate Interface Physics. *Journal of Visual Languages and Computing*, 7(1):3-32.

78. Plaisant, C., Carr, D., Shneiderman, B. (1995). Image browsers: taxonomy, guidelines, and informal specifications. *IEEE Software*, 12(2), pp. 21-32, (March 1995).

79. Furnas, G. (1986). Generalized fisheye views. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 1986, pp. 16—23, Boston, Massachusetts.

80. Leung, Y. K., and Apperley, M. D. (1994). A review and taxonomy of distortion-oriented presentation techniques. ACM Transactions on Computer-Human Interaction, 1(2):126—160.

81. Keahey, T.A., Robertson, E. (1996). Techniques for non-linear magnification transformations. In Proceedings IEEE Symposium on Information Visualization '96, pp 38-45.

82. Wills, G. (1996). Selection: 524,288 ways to say `this is interesting'. In Proceedings of the IEEE Symposium on Information Visualization, pp 54—60.

83. Yee, K., Fisher, D., Dhamija R., and Hearst M. (2001). Animated Exploration of Dynamic Graphs with Radial Layout . In Proceedings of the IEEE Symposium on Information Visualization 2001.

84. Baldonado, M., Woodruff, A., Kuchinsky, A. (2000). Guidelines for using multiple views in information visualization. Proc. ACM Advanced Visual Interfaces `00, pp. 110-119.

85. Becker R., and Cleveland, W. (1987).

Brushing scatterplots. Technometrics, 29(2):127-142.

86. Young, D., and Shneiderman, B. (1993). A Graphical Filter/Flow Representation of Boolean Queries: A Prototype Implementation and Evaluation. Journal of the American Society of Information Science, 44(6), 327-339.
87. Gantz, J. F., (2008). The diverse and exploding digital universe, http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf
88. Duda, R., Hart, P., Stork, D., (2001). Pattern Classification, second ed. John Wiley and Sons, New York, pp. 16-17.
89. Chapelle, O., Schölkopf, B., Zien, A. (Eds.), (2006). Semi-Supervised Learning. MIT Press, Cambridge, MA.
90. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters 31, 651–666 .
91. Duarte J. (2008). Agrupamento de Dados com Restrições. Master thesis, ISEP, Porto. 6, 17-26, 27-55.
92. Caruana R., Alexandru N. M. (2006), An Empirical Comparison of Supervised Learning Algorithms. Proc. of International Conference on Machine Learning.
93. Zhu X., GoldGoumas S. K., Dimou I. N., Zervakis M. E. (2010). Combination of multiple classifiers for post-placement quality inspection of components: A comparative study. Information Fusion, Volume 11, Issue 2. pp. 149-162berg A.B., Brachman R., Diettrich T. (2009). Introduction to Semi-Supervised Learning (Synthesis Lectures on Artificial Intelligence and Machine Learning). pp. 9-11
94. Sun J., Li H. (2009). Financial distress prediction based on serial combination of multiple classifiers. Expert Systems with Applications: An International Journal, Volume 36 Issue 4, Tarrytown, NY, USA. pp. 8659-8666
95. Goumas S. K., Dimou I. N., Zervakis M. E. (2010). Combination of multiple classifiers for post-placement quality inspection of components: A comparative study. Information Fusion, Volume 11, Issue 2. pp. 149-162
96. Cockburn A., Karlson A., and Bederson B. B. (2008). A review of overview+detail, zooming, and focus+context interfaces. ACM Computing Surveys (CSUR), Volume 41, Issue 1, New York, NY, USA.
97. Hornbæk K., Bederson B. B., Plaisant C. (2002). Navigation patterns and usability

of zoomable user interfaces with and without an overview. ACM Transactions on Computer-Human Interaction (TOCHI), Volume 9, Issue 4, NY, USA, pp. 362 - 389

98. Jain A., Murty M., and Flynn. P. (1999). Data Clustering: A Review. ACM Computing Surveys, Vol. 31, No. 3, 264-323.

99. Jain A., and Dubes R. (1988). Algorithms for Clustering Data. Prentice Hall.

100. Anderberg M. (1973). Cluster Analysis for Applications. Academic Press, Inc., New York, NY.

101. Diday E., and Simon J. (1976). Clustering analysis. In Digital Pattern Recognition, K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ, pp. 47–94.

102. Gower J. (1985). Measures of similarity, dissimilarity and distance. Encyclopaedia of Statistical Sciences, Volume 5. Wiley, New York.

103. Gower J., and Legendre P. (1986). Metric and Euclidean properties of dissimilarity coefficients. Journal of classification, 5, pp. 5-48.

104. Michalski R., Stepp R., and Diday E. (1983). Automated construction of classifications: conceptual clustering versus numerical taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5, No. 4, pp. 396-410.

105. Diday E., and Simon J. (1976). Clustering analysis. In Digital Pattern Recognition, K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ, pp. 47–94.

106. Roth V., Braun M., Lange T. and Buhmann J. (2002). A resampling approach to cluster validation. Computational Statistics. COMPSTAT'02.

107. Dom B. (2002). An Information-Theoretic External Cluster-Validity Measure. IBM Research Report.

108. Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, pp. 846–850.

109. Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, pp. 193–218.

110. Yeung K. Y., Ruzzo W. L. (2001). Details of the Adjusted Rand index and Clustering algorithms. Supplement to the paper “An empirical study on Principal Component Analysis for clustering gene expression data”.

111. Duarte F. J. (2008). Optimização da combinação de agrupamentos baseado na acumulação de provas pesadas por índices

de validação e com uso de amostragem. Ph.D. dissertation, Universidade de Trás-os-Montes e Alto Douro. 17-24, 25-76.

112. Berkhin P. (2002). Survey of Clustering Data Mining Techniques. Accrue Software.

113. Han J., and Kamber M. (2006). Data Mining, Concepts and Techniques. Morgan Kaufmann Publishers, 2nd Edition.

114. Macqueen J. B. (1967). Some methods of classification and analysis of multivariate observations. in Proceedings of the Fifth Berkeley Symposium on Mathemtical Statistics and Probability, 1967, pp. 281–297.

115. Ball G., and Hall D. (1965). ISODATA, a novel method of data analysis and classification. Tech. Rep., Stanford University, Stanford, CA

116. Kaufmann L., and Rousseeuw P. J. (1987). Clustering by means of medoids. In Dodge, Y. (Ed.) Statistical Data Analysis based on the L1 Norm, Elsevier/North Holland, Amsterdam. pp. 405-416.

117. Ng R. T., and Han J. (1994). Efficient and effective clustering methods for spatial data mining. in VLDB ’94: Proceedings of the 20th International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 144–155.

118. Sneath P., and Sokal R. (1973). Numerical Taxonomy. Freeman, London, Uk.

119. King B. (1963). Step-wise clustering procedures. Journal of the American Statistical Association, no. 69, pp. 86–101.

120. Ward J. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, no. 58, pp. 236–244.

121. Guha S., Rastogi R., and Shim K. (1998). Cure: an efficient clustering algorithm for large databases. In SIGMOD ’98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 1998, pp. 73–84.

122. Karypis G., Han E.-H. S., and Kumar V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. Computer, vol. 32, no. 8, pp. 68–75.

123. Ester M., Kriegel H.-P., Sander J., and Xu X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. in Proc. of 2nd International Conference on Knowledge,

Discovery and Data Mining (KDD-96), pp. 226–231.

124.Hinneburg A., and Keim D. (1998). An Efficient Approach to Clustering in Large Multimedia Databases with Noise. Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining.

125.Wang W., Yang J., and Muntz R. (1997). STING: A Statistical Information Grid Approach to Spatial Data Mining. Internacional Conference on Very Large Databases. Los Angeles, California.

126.Agrawal R., Gehrke J., Gunopulos D., and Raghavan P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. Proceedings of ACM SIGMOD Conference on Management of Data. 94-105, Seattle, WA.

127.Dempster A. P., Laird N. M., and Rubin D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, no. 1, pp. 1–38.

128.Fisher D. H. (1987). Knowledge acquisition via incremental conceptual clustering. Mach. Learn., vol. 2, no. 2, pp. 139–172.

129.Kohonen T. (1990). The self-organizing map. in Proceedings of the IEEE, vol. 78, no. 9, pp. 1464–1480.

130.Hou J. F.-J. (1999). Clustering with obstacle entities. Available: citeseer.ist.psu.edu/hou99clustering.html

131.Ng R., and Han J. (2002). Clarans: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering, vol. 14, no. 5, pp. 1003-1016

132.Wagstaff K. L. (2002). Intelligent clustering with instance-level constraints. Ph.D. dissertation, Ithaca, NY, USA, chair-Claire Cardie.

133.Tung A. K. H., Han J., Lakshmanan V. S., and Ng R. T. (2001). Constraint-based clustering in large databases. Lecture Notes in Computer Science, vol. 1973. [Online]. Available: citeseer.ist.psu.edu/tung00constraintbased.html

134.Shmoys D. B., Tardos E., and Aardal K. (1997). Approximation algorithms for facility location problems. In Proceedings of the 29th Annual ACM Symposium on Theory of Computing, pp. 265–274.

135.Demiriz A., Bennett K. P., and Embrechts M. J. (1999). Semi-supervised clustering using genetic algorithms. In Artificial Neural Networks in Engineering (ANNIE-

99).ASME Press, pp. 809–814.

136.Basu S., Banerjee A., and Mooney R. (2002). Semi-supervised clustering by seeding. [Online]. Available: citeseer.ist.psu.edu/basu02semisupervised.html

137.Basu S., Banerjee A., and Mooney R. (2004). Active semi-supervision for pairwise constrained clustering. In Proceedings of the 2004 SIAM International Conference on Data Mining (SDM-04), pp. 333–344.

138.Cohn D., Caruana R., and McCallum A. (2003). Semi-supervised clustering with user feedback. [Online]. Available: citeseer.ist.psu.edu/cohn03semisupervised.html

139.Wagstaff K., and Cardie C. (2000). Clustering with instance-level constraints. In Proceedings of the Seventeenth International Conference on Machine Learning, pp. 1103-1110. [Online]. Available: citeseer.ist.psu.edu/wagstaff00clustering.html

140.Wagstaff K., Cardie C., Rogers S., and Schrödl S. (2001). Constrained k-means clustering with background knowledge. In ICML ’01: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 577–584.

141.Seidl T., and Kriegel H. P. (1998). Optimal multi-step k-nearest neighbor search. University of Munich, Germany. ACM SIGMOD Record, vol. 27, pp. 154-165.

142.Quinlan J. (1986). Introduction of decision trees. Machine Learning, vol. 1, pp. 81–106.

143.Breiman L., Friedman J., Olshen R., and Stone C. (1984). Classification and regression trees. Wadsworth, Belmont, CA.

144.Rumelhart D. E., Hinton G. E., and Williams R. J. (1986). Learning internal representations by error propagation. pp. 318–362.

145.Cristianini N., and Shawe-Taylor J. (2000). An Introduction to Support Vector Machines. Cambridge: Cambridge University Press.

146.Davidson I., and Ravi S. (2005). Clustering with constraints feasibility issues and the k-means algorithm. in 2005 SIAM International Conference on Data Mining (SDM’05), Newport Beach,CA, pp. 138–149.

147.Klein D., Kamvar S. D., and Manning C. D. (2002). From instance-level constraints

to space-level constraints: Making the most of prior knowledge in data clustering. in ICML ’02: Proceedings of the Nineteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 307–314.

148. Basu S., Bilenko M., and Mooney R. (2003). Comparing and unifying search based and similarity-based approaches to semi-supervised clustering. [Online]. Available: citeseer.ist.psu.edu/article/basu03comparing.html

149. Lu Z., and Leen T. K. (2007). Penalized probabilistic clustering. Neural Comput., vol. 19, no. 6, pp. 1528–1567.

150. Basu S., Bilenko M., and Mooney R. J. (2004). A probabilistic framework for semi-supervised clustering. in KDD ’04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, pp. 59–68.

151. Corter, J. E. and Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. Psychological Bulletin 111 (2): pp. 291–303.

152. Floyd R. W. (1962). Algorithm 97: Shortest path. Commun. ACM, vol. 5, no. 6, p. 345.

153. Kaufmann M. and Wagner D. (2001). Drawing Graphs: Methods and Models. Springer.

154. Eades P. (1984). A heuristic for graph drawing. Congressus Numerantium, 42: pp. 149–160.

155. prefuse.org (2006). Prefuse: Interactive information visualization toolkit.

156. Duarte F. J., Fred A., Rodrigues M. F. and Duarte J. (2006). Weighted Evidence Accumulation Clustering using Subsampling. Sixth International Workshop on Pattern Recognition in Information Systems (PRIS-2006).

157. Duarte J., Fred A., Duarte F. J. (2009). Combining data clusterings with instance level constraints. Proceedings of the 9th Intl. Workshop on Pattern Recognition in Information Systems, Milan, Italy.

158. Brockenauer R. and Cornelsen S. (2001). Drawing clusters and hierarchies. In Drawing Graphs - Methods and Models, LNCS 2025, pp. 193-227.

159. University of Waikato (2006). Weka 3: Data mining with open source machine learning software in Java.

160. Basu S. (2006). WekaUT: Extensions to Weka. Available: http://www.cs.utexas.edu/users/ml/risc/co

de/

161. Desjardins M., MacGlashan J. and Ferraioli J. (2007). Interactive Visual Clustering. Proceedings of the 12th international conference on Intelligent user interfaces, Honolulu, Hawaii, USA.

162. Desjardins M., MacGlashan J. and Ferraioli J. (2008). Interactive Visual Clustering. Constrained Clustering: Advances in Algorithms, Theory, and Applications . Chapman & Hall/CRC Ed. Minneapolis, Minnesota, U.S.A.

163. Newman D. J., Hettich S., Blake C. L. and Merz C. J. (1998). UCI repository of machine learning databases.

164. Fruchterman T., and Reingold E. (1991). Graph Drawing by Force-Directed Placement. Software-Practice & Experience, vol. 21, pp. 1129-1164.

165. Barros, V. H. (2011). Preka: Toolkit for implementation of combined visualization and machine learning techniques. Available: http://code.google.com/p/preka/