Universidade do Minho

Escola de Engenharia

Departamento de Informática

Mineração de Dados para suporte à decisão no processo da Recolha Selectiva – Optimização de Rotas

Nuno Ramos Matos

Dissertação de Mestrado

2008

Mineração de Dados para suporte à decisão no processo da Recolha Selectiva – Optimização de Rotas

Nuno Ramos Matos

Dissertação apresentada à Universidade do Minho para obtenção do grau de Mestre em Informática, elaborada sob orientação do Professor Doutor Orlando Manuel de Oliveira Belo.

2008

Para a Mónica

Agradecimentos

Agradeço ao Eng. Cândido Martins, director do Departamento de Engenharia e Inovação da Cachapuz pelo apoio prestado no desenvolvimento deste mestrado. Agradeço também à Cachapuz pelo financiamento do mestrado e pela disponibilização do SPAR como caso de estudo.

Agradeço à Resulima pela disponibilização da informação utilizada para criação desta dissertação.

Agradeço à ao Professor Doutor Orlando Belo pela sua orientação no decorrer desta dissertação.

Agradeço aos meus pais e irmãos pelo apoio que sempre me deram.

Finalmente agradeço à Mónica pela força que sempre me deu para chegar onde cheguei.

Obrigado!

Resumo

A crescente preocupação mundial com a gestão de resíduos impulsionou as empresas de recolha selectiva a adoptarem sistemas de gestão que os auxiliassem a gerir os seus processos. O principal objectivo nesta área consiste na recolha de todos os resíduos com a frequência suficiente, gastando o mínimo de recursos possível. As empresas passaram a registar elevados volumes de dados, originando dificuldades na extracção de informação importante para tomada de decisão [Sumathi e Sivanandam, 2006]. Perante este cenário, as ferramentas informáticas ditas tradicionais, embora respondam bem aos processos operacionais, não se adequam aos processos de tomada de decisão. Estas limitações levaram as empresas a procurarem no mercado novas áreas de investigação para a descoberta de conhecimento que auxilie os gestores na tomada de decisão na optimização dos seus processos. Uma dessas áreas é o *data mining*.

Nesta dissertação é apresentada a aplicação do *data mining* na optimização de rotas de recolha selectiva. Para isso, foram estudadas algumas das técnicas existentes e a sua aplicação na indústria, bem como a sua aplicação à recolha selectiva. Esta dissertação foi elaborada com base na metodologia *CRISP-DM* [Web CRISPDM], desde a fase de conhecimento da área da recolha selectiva até à modelação e análise dos resultados. Os modelos desenvolvidos permitiram reorganizar as rotas de recolha com base na informação histórica de um ano de recolhas.

Palavras-Chave: Optimização de Rotas, *Data Mining*, Recolha Selectiva, Gestão de Resíduos

Abstract

The crescent world concern regarding the waste management guided the companies operating in the selective collection to adopt management systems to help them manage their processes. The main objective in this area consists in the collection of all the waste with enough frequency and spending the minimum amount of resources possible. These companies began registering high volumes of information, increasing the difficulty to extract from that data important information that may be useful for decision making [Sumathi e Sivanandam, 2006]. Facing this scenario, the "traditional" informatics tools, although able to correctly answer to the operational processes, are not suitable for the decision making processes. These limitations leaded the companies to search the market for new investigation areas for the knowledge discovery that could help the managers in the decision making and in the optimization of their processes. One of such areas is data mining.

In this dissertation, an application of data mining in the optimization of selective collection routes will be presented. To accomplish that task, some of the existing techniques and their application in the industry were studied, as well as their possible adaptation to the selective collection. This dissertation was elaborated based on the CRISP-DM methodology [Web CRISPDM], from the comprehension of business area to the evaluation of the results. The models developed enabled the reorganization of collection routes based on the historical data of a year of collections.

Keywords: *Routing Optimization, Data Mining, Waste Collection, Waste Managment*

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1

Introdução

1.1 O conhecimento da área de negócio

A recolha selectiva é o processo através do qual são recolhidos os resíduos, previamente separados pelos cidadãos, para serem usados no processo de reciclagem [Web NetRes]. A recolha selectiva é efectuada, apenas por operadores de resíduos licenciados, através dos seguintes processos: a recolha de contentores uni-materiais espalhados pelo território nacional que são usados pelos cidadãos para deposição de resíduos previamente separados; e a recolha porta-a-porta que é feita através da visita ao domicílio ou a estabelecimentos comerciais para a recolha dos diferentes resíduos separados. Uma vez recolhidos os resíduos, estes são transportados para as estações de triagem onde serão triados e enfardados para serem posteriormente retomados para reciclagem pelas empresas produtoras de embalagens. Os resíduos reciclados dão origem a novos produtos que são introduzidos novamente no mercado.

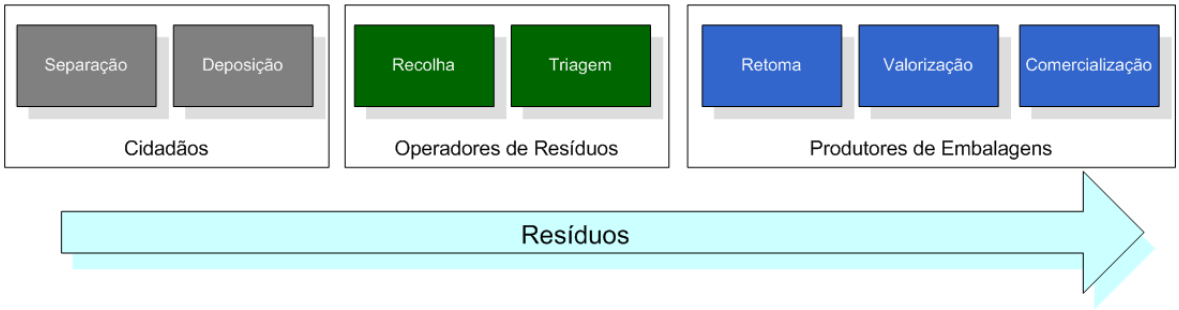

Figura 1 – Ciclo de vida dos resíduos sólidos urbanos separados

Os operadores de resíduos têm um importante papel na sociedade porque são eles os responsáveis pela recolha e manutenção dos contentores da sua zona de actuação, tornando ao mesmo tempo a sua actividade rentável. Os principais problemas enfrentados pelos operadores neste processo estão relacionados com os seguintes factores:

- **Recolha e Manutenção** – Todos os contentores devem ser recolhidos com uma frequência tal que permitam às pessoas, no momento de depositarem os seus resíduos, encontrem os contentores com espaço suficiente para o fazerem. Os contentores devem também ser mantidos em bom estado de conservação e higiene, sendo limpos e reparados sempre que necessário. Os cidadãos são os produtores da matéria-prima da reciclagem e por este facto devem ser satisfeitos com um serviço eficaz, evitando reclamações.
- **Gestão de Recursos** – Com o aumento dos preços dos combustíveis e dos preços da manutenção dos camiões de recolha, e como os lucros obtidos com esta actividade são muito baixos, torna-se essencial minimizar as viagens para recolher os resíduos, aumentando as quantidades recolhidas sempre que as rotas são executadas.

Para conseguirem alcançar estes objectivos, os operadores optimizaram alguns dos seus processos com a utilização de software de gestão que lhes permitem ter um melhor controlo do seu negócio. Estas ferramentas permitem actualmente a gestão de elevadas quantidades de dados que são gerados e que aumentam consideravelmente de dia para dia. Com este súbito aumento de informação disponível apareceu um novo objectivo nesta área: a adopção de ferramentas que

possibilitem a análise em tempo útil de elevados volumes de informação e ainda a transformação dessas análises em decisões que permitam melhorar continuamente os processos de gestão de resíduos.

1.2 Uma definição de mineração de dados

Para percebermos o significado do processo de mineração de dados é importante analisarmos literalmente as suas palavras: mineração (*mining*) e dados (*data*) [Giudici, 2003]. Mineração sugere extracção que é normalmente associada à extracção de recursos preciosos da Terra. A associação desta palavra com *Data* sugere uma exploração aprofundada de informação que era até então indetectável no elevado volume de dados existentes. Mas em termos gerais, podemos dizer que a mineração de dados é o processo de selecção, exploração e modelação de elevadas quantidades de dados para descobrir padrões ou relações que até no início eram desconhecidas com o objectivo de obter resultados claros e úteis para o proprietário das bases de dados.

Pela definição mais comum e tradicional, a mineração de dados é aplicada quando queremos extrair informação importante para um determinado negócio a partir de um elevado volume de dados [Bigus, 1996]. A palavra "elevado" pode ser relativa e variar consoante a área de negócio onde é usada. Por exemplo, uma empresa de pequenas dimensões que trabalhe sobre uma base de dados de um gigabyte, poderá afirmar que possui um elevado volume de informação mas se essa base de dados fosse comparada com uma base de dados de um terabyte de empresa de maior dimensão, o seu tamanho poderia ser considerado extremamente reduzido. Então quão grande deverá ser a nossa base de dados para que seja rentável a utilização de mineração de dados? Uma base de dados será suficientemente grande se ela contiver tanta informação que faça com que as relações entre os dados estejam escondidas. Desta forma, e após a aplicação do *data mining*, será possível extrair informações valiosas e não óbvias.

A *IBM* foi uma das primeiras empresas no mundo a utilizar os processos de *data mining* para detecção eficiente de padrões e relações em grandes volumes de informação e fê-lo através da invenção das regras de associação e padrões sequenciais [Web IBM], que lhe permitiram patentear alguns conceitos e processos nesta área [Web SourceWatch].

Segundo a metodologia *CRISP-DM – CRoss-Industry Standard Process for Data Mining* [Web CRISPDM], para aplicar processos de *data mining* é necessário ter em atenção as seguintes fases:

- ***Business Understanding*** – O conhecimento do negócio sobre o qual pretendemos desenvolver um projecto de *data mining* é essencial para todo o processo e para o sucesso desse projecto.
- ***Data Understanding*** – É necessário também conhecer e perceber toda a informação disponível para análise, para isso será ainda necessário desenvolver cada uma das seguintes sub-tarefas:
    - ***Collect Initial Data*** – Obtenção de toda a informação necessária.
    - ***Describe Data*** – Descrever e explicar a informação obtida.
    - ***Explore Data*** – Explorar a informação.
    - ***Verify Data Quality*** – Constatar se os dados têm os requisitos de qualidade necessários para a sua submissão aos modelos de *data mining*.
- ***Data Preparation*** – Os dados existentes devem ser preparados para sobre eles serem aplicados os modelos de *data mining*. Essa preparação é efectuada através das seguintes sub-tarefas:
    - ***Select Data*** – Seleccionar os dados pretendidos.
    - ***Clean Data*** – Limpar os dados, se necessário.
    - ***Construct Data*** – Construir e inferir novos atributos a partir dos dados existentes.
    - ***Integrate Data*** – Integrar informação com dados provenientes de várias tabelas ou registos.
    - ***Format Data*** – Formatar a informação ao nível da sintaxe.

- ***Modeling*** – Após terem sido analisados quer o negócio, quer os seus dados, devem ser criados e aplicados os modelos de *data mining* escolhidos através das seguintes fases:
    - ***Select Modeling Technique*** – Escolher os modelos de *data mining* a utilizar.
    - ***Generate Test Design*** – Definir a execução, teste e avaliação dos modelos de *data mining* implementados.
    - ***Build Model*** – Construir os modelos.
    - ***Assess Model*** – Interpretar os modelos criados.
- ***Evaluation*** – Uma vez executados os modelos de *data mining*, os resultados devem ser analisados de forma a verificar a sua utilidade para o negócio em causa e a sua adaptação aos donos desse mesmo negócio. É nesta fase que o processo é avaliado e que pode dar origem a ajustes nas suas fases anteriores.
- ***Deployment*** – Se os resultados obtidos com os modelos implementados forem validados, dá-se origem à disponibilização de relatórios que sejam facilmente entendidos pelos clientes finais e que possam auxiliar na tomada de decisões que melhorem os seus processos de negócio.

Os processos de mineração de dados podem ser aplicados a inúmeras áreas de negócio que lidam diariamente com elevados volumes de informação, como são os casos das áreas de Marketing de Vendas, Economia, Banca, Aplicações Económicas, Seguros, Ciência, Biomedicina, Gestão de Portais, Telecomunicações, Segurança, Investigação, entre outras [Sumathi e Sivanandam, 2006].

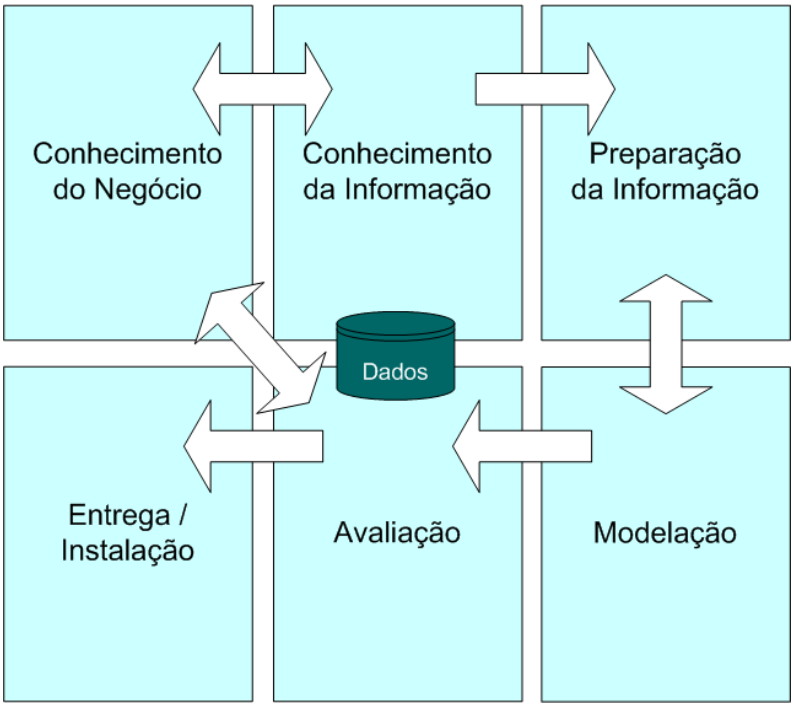

Figura 2 – Fases do *CRISP-DM*, baseadas da figura original [Web CRISPDM]

1.3 Motivação e objectivos

A crescente preocupação mundial com a gestão de resíduos levou a que todas as empresas a operar nesta área sentissem a necessidade de adoptarem sistemas de gestão que os auxiliassem a gerir e optimizar os seus processos. Neste grupo de empresas encontram-se os operadores de resíduos que são os responsáveis pela recolha selectiva dos resíduos previamente separados quer pelos cidadãos quer pelas empresas. Os resíduos são recolhidos diariamente por equipas de trabalho com o auxílio de camiões especiais para esta área de actividade e todo este trabalho deve ser planeado e executado com o objectivo de que todos os resíduos sejam recolhidos com a maior frequência possível, gastando o mínimo de recursos possível. Alguns dos recursos envolvidos na recolha selectiva são os gastos com as manutenções dos camiões de recolha que aumentam com a sua utilização, os consumos de combustível que têm aumentado com o aumento dos preços do

petróleo, horas de recursos humanos afectos a esta actividade que implicam gastos com os seus vencimentos. Se os gastos com estes recursos forem diminuídos os operadores de resíduos ficarão mais competitivos, serão mais eficazes nas suas tarefas, acabando por prestar um melhor serviço público aos cidadãos.

O objectivo desta dissertação assenta então sobre a optimização das rotas de recolha selectiva tendo em conta a reorganização dos contentores que as compõem. Ao reorganizar os contentores por rotas tendo em conta as suas características no que dizem respeito às taxas de enchimento, quantidades recolhidas, entre outras, poderá levar à redução dos recursos afectos a esta actividade, elevando as quantidades recolhidas na execução das rotas de recolha selectiva.

A possibilidade de aplicação do *data mining* nesta área possibilitou a abertura de uma porta que poderá, a médio e longo prazo, dotar os operadores de resíduos de informação vital, que até ao momento era difícil ou até mesmo impossível de obter com as ferramentas existentes actualmente. Para isso, será necessário enriquecer as bases de dados com mais informação e atributos complementares que permitam descobrir novos padrões e novas relações escondidas nos dados. Por exemplo, seria possível determinar, com base nas técnicas de *data mining* mais utilizadas e divulgadas, nomeadamente, a seguinte informação:

- Relações entre as quantidades recolhidas e informações meteorológicas.
- Previsão de enchimento dos contentores com base no dia do ano.
- Associação entre as quantidades recolhidas com as festas populares e feriados.

1.4 Organização do documento

Para além deste capítulo, esta dissertação é composta por mais quatro capítulos organizados da seguinte forma:

- **Capítulo 2 – Análise dos processos dos Operadores de Resíduos**

  Neste capítulo serão descritos os processos dos operadores de resíduos, serão descritas algumas das ferramentas que eles utilizam para gerir o seu negócio, qual a informação regista e tratada por estas empresas e ainda alguns problemas encontrados pelos operadores no momento da análise da informação existente para a correcta tomada de decisão.

- **Capítulo 3 – Mineração de dados na indústria**

  Aqui serão analisadas algumas das técnicas de *data mining* existentes e a sua aplicabilidade na indústria, nomeadamente na sua adequação à optimização de rotas de recolha selectiva.

- **Capítulo 4 – Reorganização de rotas através de técnicas de mineração de dados**

  Neste capítulo será explicada a aplicação das técnicas de *data mining* escolhidas no capítulo 3 na optimização de rotas de recolha selectiva através de um caso de estudo. Serão descritos todos os passos nessa implementação das técnicas de *data mining* e analisados os resultados obtidos, com base na metodologia *CRISP-DM*.

- **Capítulo 5 – Conclusões e Trabalho Futuro**

  No quinto e último capítulo desta dissertação será feita uma análise crítica a todo o trabalho, salientando os resultados obtidos. Serão ainda indicados objectivos para trabalhos futuros nesta área.

Capítulo 2

Análise dos processos dos Operadores de Resíduos

2.1 Processos de negócio

A gestão dos resíduos não é uma necessidade apenas dos tempos modernos, começou a sê-lo por volta do ano 10000 a.C., quando os povos deixaram de ser nómadas. As populações começaram a estabelecer-se no mesmo local e começaram a ter necessidades de gerir os resíduos que produziam. No ano de 200 d.C., os Romanos sentiram a necessidade de criar um conceito de recolha de resíduos nas suas cidades [Kimball, 1992]. Esta recolha era executada por equipas de dois homens que iam recolhendo o lixo das ruas, atirando-o para dentro de uma carroça. Segundo Deby Kimball [Kimball, 1992], em 1690 foi criada pela família Rittenhouse, a primeira unidade de reciclagem de papel, em Wissahickon Creek perto de Filadélfia. Apesar de ser não ser possível eleger uma pessoa em particular pela invenção da reciclagem, podemos concluir que foi um

processo colaborativo entre vários países do mundo com o objectivo de tornar o nosso planeta mais sustentável [Web Recycling].

Nesta área, os operadores de resíduos registam elevados volumes de informação diariamente que devem ser tratados convenientemente em prol dos seus objectivos enquanto prestadores de serviço público. Para conhecermos um pouco mais sobre o número de registos efectuados diariamente, serão apresentadas algumas estatísticas retirdadas do site do grupo Águas de Portugal (AdP), uma das entidades portuguesas que contribuem para a resolução de problemas relacionados com o tratamento e valorização de resíduos, entre outros assuntos de carácter ambiental. Segundo a AdP, em 2007 foram recolhidas 146000 toneladas de resíduos através de processos de recolha selectiva [Web ADP].

	2005	**2006**	**2007**
Resíduos sólidos urbanos tratados (milhões de toneladas)	2,4	2,5	2,6
Recolha selectiva (milhares de toneladas)	113	131	146

Tabela 1 – Quantidades de Resíduos recolhidas entre 2005 e 2007

A reciclagem é hoje em dia um processo importante na gestão dos resíduos produzidos já que estes podem agora ser recolhidos, tratados e reintroduzidos no mercado sob a forma de novos produtos. Até a chegada ao mercado destes produtos, os resíduos utilizados no seu fabrico, passam por vários processos de seguida descritos.

2.1.1 Produção de resíduos

As pessoas no seu quotidiano produzem enormes quantidades de resíduos que devem ser tratados convenientemente. Tomemos como exemplo Portugal. Segundo a Quercus[1] cada português produz em média 1,2Kg de lixo por dia [Web Quercus]. Já um Americano produz em média 2Kg de lixo por

dia [Web Garbage]. Este valor varia bastante de país para país estando relacionado com factores económicos, sociais, religiosos, entre outros. Com a criação do conceito de reciclagem, nasceu a necessidade dos resíduos serem separados em casa pelas pessoas o que permite que os resíduos passíveis de serem reciclados não sejam contaminados por outros resíduos que não o sejam. As pessoas separam os resíduos e depositam-nos em contentores próprios existentes no ecoponto, criado para auxiliar o processo de separação. Um ecoponto é um então um conjunto de contentores para deposição de resíduos previamente separados [Web EcoPonto]. No caso de Portugal, no ecoponto dito tradicional, existem quatro contentores, um vidrão (contentor verde), um embalão (contentor amarelo), um papelão (contentor azul) e um pilhão (contentor vermelho), para depositar vidro, embalagens de plástico e metal, papel e cartão e pilhas respectivamente. Os restantes resíduos indiferenciados são depositados no contentor de resíduos indiferenciados. A composição de um ecoponto varia consoante as necessidades da zona onde é colocado, podendo ter mais do que um determinado contentor, como acontece com o contentor de vidro nas zonas junto a restaurantes e cafés, ou até não ter um dos contentores, como acontece frequentemente com o pilhão.

2.1.2 Recolha Selectiva

Uma vez separados e depositados os resíduos nos contentores dos ecopontos, é necessário que empresas licenciadas e autorizadas façam a recolha dos contentores. Estas empresas têm a designação de Operadores de Resíduos [Web Operadores]. Cada operador tem zonas de recolha bem delimitadas onde opera e organiza os ecopontos dessa zona por rotas de recolha, ou seja, conjuntos de contentores a recolher de um determinado produto. Cada equipa de recolha composta por um motorista com ou sem ajudantes, efectua a recolha de uma ou mais rotas diariamente com o auxílio de camiões específicos para esta tarefa. Normalmente cada equipa apenas pode recolher um produto numa determinada rota de recolha, mas existem já operadores que dispõem de

[1] Quercus – Associação Nacional de Conservação da Natureza

camiões bi-compartimentados que são capazes de transportar dois produtos distintos sem os misturarem, tentando desta forma optimizarem os recursos afectos a esta área de actividade. Após as recolhas, os resíduos são levados para Aterros Municipais onde são descarregados e tratados em função do seu tipo, resíduos indiferenciados são descarregados e encaminhados para deposição no aterro sanitário enquanto os restantes produtos são reencaminhados para o centro de triagem [Web AMDE].

2.1.3 Triagem

Todos os materiais passíveis de serem valorizados [Web Valorização], com excepção do vidro chegam ao centro de triagem e passam por um processo de separação. Este processo pode ser mecânico ou manual e permite separar os resíduos tendo em conta o seu tipo: garrafas e frascos de PEAD, garrafas de PVC, garrafas de PET de óleo alimentar, filmes (sacos de PEBD/PEAD), poliestireno expandido (esferovite), embalagens Tetra Pak, embalagens ferrosas, embalagens de alumínio, plásticos mistos, etc. [Web AMARSUL]. Os produtos separados são enfardados e armazenados até à sua retoma.

2.1.4 Retoma, valorização e comercialização

Todos os resíduos separados e enfardados são retomados por empresas produtoras de embalagens que são responsáveis por introduzir estes resíduos nos seus processos produtivos, criando com eles novos produtos para comercializar. A este processo é dado o nome de valorização ou reciclagem. As empresas produtoras de embalagens encontraram uma nova área de negócio que lhes permite, por um lado adquirir matérias-primas provenientes dos resíduos a baixo preço, por outro lado escoar os resíduos produzidos pelas populações que deixam de ter como destino o aterro sanitário [Web Valorização].

A recolha selectiva é então o processo de recolha dos resíduos previamente separados pelos cidadãos. Existem dois grandes grupos de recolha selectiva:

- **Recolha de contentores** – Os cidadãos separam os resíduos que produzem e depositam-nos nos contentores do ecoponto. Estes contentores são recolhidos periodicamente pelos operadores e encaminhando-os para a central de triagem; Actualmente este tipo de recolha selectiva está mais banalizado em alguns produtos, como o papel e cartão, vidro, embalagens de plástico ou metal e pilhas.
- **Recolha porta-a-porta** – Os operadores de resíduos recolhem os resíduos previamente separados pelos cidadãos directamente ao domicílio ou nos estabelecimentos comerciais. Este tipo de recolha, apesar de ser mais recente, atinge já tanto os produtos normalmente depositados nos ecopontos, como outros produtos que estão agora a ter alguma importância para reciclagem, como o caso de óleos, madeiras, entre outros.

A recolha selectiva é efectuada tendo em conta as seguintes fases:

- **Análise** – Os responsáveis pela gestão da recolha selectiva dos operadores de resíduos analisam a informação histórica recolhida pelas equipas no terreno, quando existe, de forma a planearem o trabalho de recolha. Uma vez efectuada essa análise, é planeado o trabalho das equipas.
- **Planeamento** – O planeamento é efectuado com base todas as rotas de contentores existentes, atribuindo uma ou mais rotas a cada uma das equipas de trabalho.
- **Recolha** – As equipas de trabalho visitam e recolhem os contentores planeados, registando a informação necessária durante este processo, como o caso dos enchimentos dos contentores, o seu estado de higiene, necessidade ou não de reparação, entre outras informações relacionadas com o processo.

- **Descarga** – As equipas de trabalho regressam ao aterro para pesarem e descarregarem as viaturas na central de triagem, descarregando também a informação recolhida para nova análise e planeamento do trabalho seguinte.

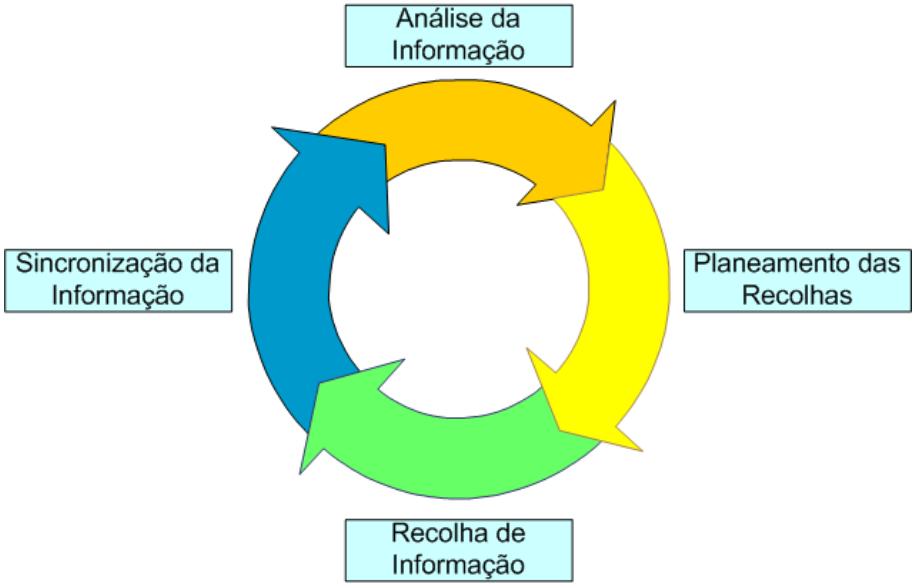

Figura 3 – Gestão da recolha selectiva

Os operadores de resíduos que são responsáveis pela recolha selectiva têm dois grandes objectivos a alcançar com os seus processos. O primeiro grande objectivo está relacionado com o serviço público que prestam na sua zona de actuação. Os operadores são responsáveis pela recolha de todos os contentores com a melhor frequência possível para que, quando os cidadãos tenham a necessidade de depositar os seus resíduos separados, tenham possibilidade de o fazer. São responsáveis também pela higiene de todos os contentores mantendo a zona limpa e são responsáveis ainda pela manutenção dos contentores, assegurando que estes funcionam correctamente, nomeadamente nos mecanismos de abertura e de descarga. O segundo grande objectivo dos operadores é o de optimizarem ao máximo os seus processos de forma a torná-lo o mais rentável possível.

Neste ponto foi coberta a percepção e conhecimento do negócio em estudo, a recolha selectiva, cobrindo a primeira das fases do modelo *CRISP-DM*.

2.2 A recolha de informação

Como descrito em capítulos anteriores, a utilização de ferramenta para gestão dos processos da recolha selectiva é uma necessidade. No âmbito desta dissertação e tendo em conta a metodologia *CRISP-DM,* foi analisada a ferramenta SPAR e a informação gerida pela mesma na fase de *Data Understanding.*

O SPAR – Sistema de Planeamento e Análise da Recolha [Web SPAR] é uma solução modular desenvolvida pela Cachapuz[2] que permite gerir e controlar todo o processo da recolha selectiva. Desde o planeamento inicial das recolhas a efectuar pelas equipas de trabalho, passando pelo registo de toda a informação inerente ao negócio, até à obtenção de indicadores, esta ferramenta cobre todos os requisitos desta actividade. O sistema é composto pelos módulos de BackOffice, Mobilidade, Cartografia, Business Viewer, Portal, Portal e SMS de seguida descritos.

2.2.1 Backoffice

O BackOffice é um módulo que corre em computadores com ambiente *Microsoft Windows* ligado a uma base de dados centralizada em *Microsoft SQL Server*. É o principal módulo de todo o sistema já que permite armazenar e gerir toda a informação relacionada com a recolha selectiva, tal como:

- Ecopontos / Pontos de Recolha e seus Contentores.
- Rotas de Recolha.
- Produtos.

[2] Cachapuz, empresa sedeada em Braga e que se dedica ao desenvolvimento de soluções de pesagem industrial

- Motoristas e Ajudantes.
- Viaturas de Recolha.

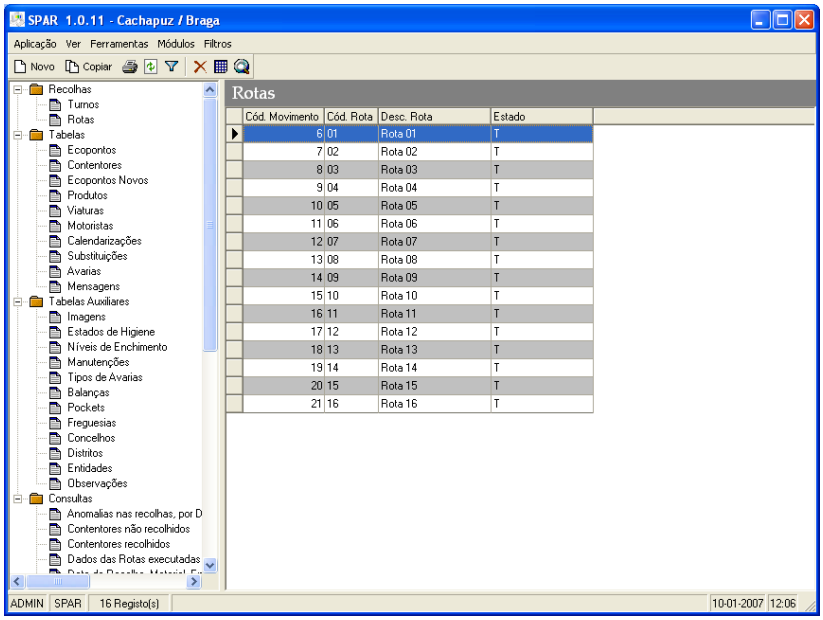

Figura 4 – *Screenshot* do módulo SPAR BackOffice

Este módulo possibilita aos encarregados pela recolha selectiva a completa análise histórica das recolhas efectuadas, consultar qual o último estado registado para todos os contentores existentes no sistema, permite consultar previsões de enchimento, a performance dos motoristas bem como analisar todos os restantes recursos afectos a esta actividade tais como:

- Quilómetros percorridos pelas viaturas de recolha.
- Quantidades recolhidas por produto.
- Horas de Serviço.
- Avarias registadas.

Toda a informação do BackOffice está organizada tendo em conta turnos de trabalho (dia de trabalho). Um turno de trabalho é o trabalho efectuado por uma determinada equipa de trabalho

(composta por um motorista com ou sem ajudantes) num dia, com uma viatura de recolha, e que indica ainda quais as rotas processadas, quais os contentores recolhidos, qual o enchimento e higiene de cada um dos contentores visitados, quais as quantidades recolhidas por produto e quantos quilómetros foram percorridos. Cada equipa de trabalho pode realizar várias descargas dos camiões durante a execução do seu turno sempre que este esteja cheio de produto. A cada uma das descargas efectuadas num determinado turno foi dada a designação de serviço. Logo, em cada turno, a equipa de trabalho poderá executar m rotas e descarregar o camião n vezes realizando x serviços.

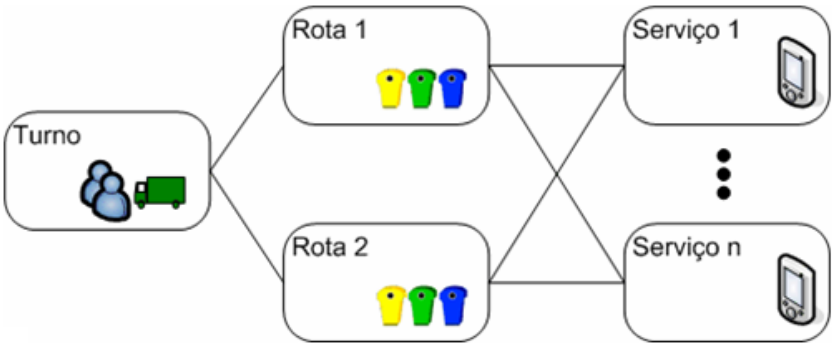

Figura 5 – Associação de rotas e serviços a um turno

Toda a informação operacional é recolhida no terreno pelas equipas de trabalho com o auxílio do módulo de Mobilidade.

2.2.2 Mobilidade

O módulo de Mobilidade é um software que executa em dispositivos portáteis como *PDA (Personal Digital Assistant)*, *Pocket PC* ou *Smartphone* (equipados com sistemas operativos *Windows Mobile*) e que permite registar toda a informação relacionada com o processo de recolha e manutenção de ecopontos/pontos de recolha por parte das equipas de trabalho no terreno.

Cada equipa de trabalho é munida de um *PDA* com o qual será capaz de receber do BackOffice (via *Wi-Fi*) toda a informação planeada para si pelo responsável ou encarregado pelo planeamento das recolhas.

Figura 6 – *Screenshot* do módulo SPAR Mobilidade

O *PDA* permite-lhe verificar quais as rotas planeadas (e respectivos ecopontos/pontos de recolha a visitar) e registar a informação inerente ao processo:

- **Estado de Enchimento** – Para cada um dos ecopontos visitados, a equipa regista o estado de enchimento de cada um dos contentores.
- **Estado de Higiene** – Para cada um dos contentores visitados, é registado pelas equipas, o seu estado de higiene, informação útil para futuros planeamentos de limpezas.
- **Avarias** – Sempre que são detectadas avarias, estas são registadas no módulo SPAR Mobilidade.

- **Observações** – As equipas de trabalho podem também registar outras informações importantes que ocorram durante a execução das recolhas, como o caso de viaturas mal estacionadas que impeçam a recolha de um determinado contentor, etc..
- **Quilómetros** – Sempre que uma equipa inicia e termina um turno de trabalho, regista no módulo de mobilidade os quilómetros que o camião de recolha tem naquele momento. Esta informação é útil para análise das performances das equipas de trabalho.
- **Quantidades** – Quando um camião regressa para efectuar uma descarga, cada equipa de trabalho termina um serviço de recolha registando qual o peso dos resíduos recolhidos.

Caso os *PDA* utilizados pelas equipas de trabalho estejam equipados com receptor *GPS* (*Global Positioning System*), o SPAR Mobilidade regista automaticamente e em intervalos predefinidos, a posição geográfica actual (coordenadas latitude e longitude). Desta forma, o responsável pela recolha pode conferir se as equipas de trabalho cumprem as rotas programadas.

No final do turno de trabalho todos os dados registados são enviados para o SPAR BackOffice, ficando imediatamente disponíveis para consulta. A arquitectura do sistema pode ser vista na Figura 7:

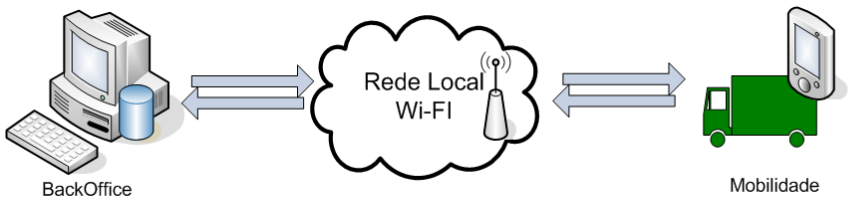

Figura 7 – Comunicação entre os módulos BackOffice e Mobilidade

2.2.3 Cartografia

O SPAR Cartografia é um módulo para visualização de informação geográfica sobre o processo de recolha de resíduos. Este módulo permite a visualização em vários motores SIG (*Google Earth*,

Microsoft MapPoint, entre outros) da localização dos ecopontos/pontos de recolha existentes no sistema. Permite ainda visualizar e conferir os percursos efectuados pelas equipas de trabalho com base nas coordenadas latitude e longitude registadas pelo módulo de Mobilidade.

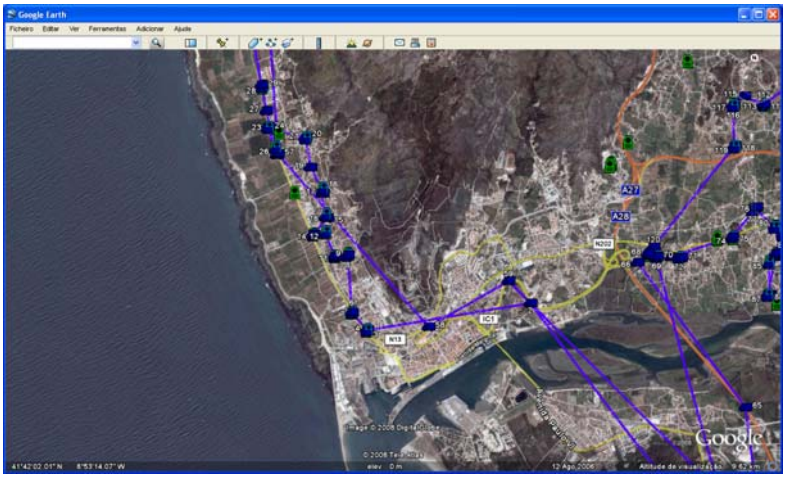

Figura 8 – *Screenshot* do módulo SPAR Cartografia

2.2.4 Business Viewer

O Business Viewer é um módulo com carácter pró-activo que executa, automaticamente e em períodos pré-configurados, um conjunto de *queries* que podem despoletar alertas sobre situações críticas que ocorram durante o processo da recolha selectiva. Este módulo envia automaticamente através de SMS (*Short Message Service*) e e-mail, relatórios e alertas a um conjunto pré-definido de destinatários, que podem desta forma agir mais rapidamente sobre o sistema.

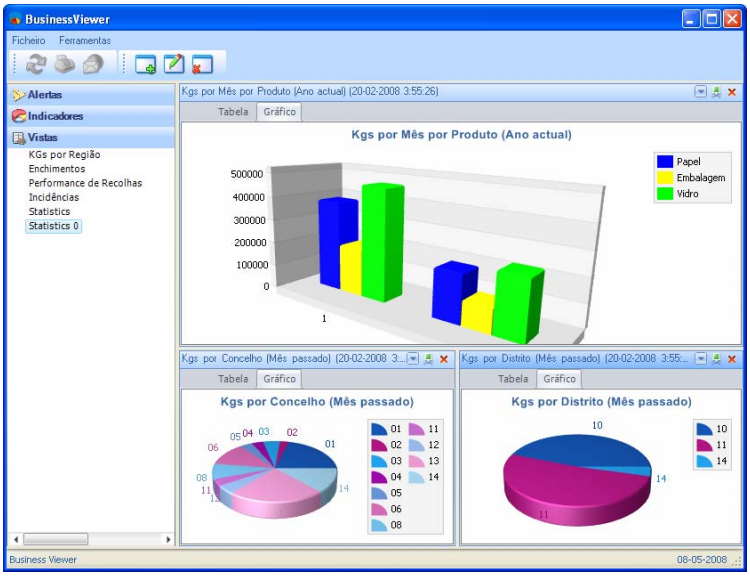

Figura 9 – *Screenshot* do módulo Business Viewer

Num sistema tão complexo como o SPAR e com o elevado volume de informação gerado diariamente, é necessário ter mecanismos que permitam facilmente, por um lado, confirmar que toda a informação é inserida correctamente pelas equipas de trabalho e, por outro lado, encaminhar os indicadores de negócio em tempo útil para destinatários certos. O Business Viewer permite ainda que a informação seja pré-processada com antecedência e em períodos configurados, garantindo assim que a informação é mostrada ou enviada de forma imediata, libertando tempo para tarefas realmente importantes.

2.2.5 Portal e SMS

Através da utilização do SPAR Portal e do SPAR SMS os cidadãos serão capazes de intervir activamente no processo enviando directamente para a entidade gestora informações vitais para o bom funcionamento das recolhas. Toda a informação flui automaticamente entre os módulos evitando a necessidade de intervenção de operadores. A entidade gestora ao receber e analisar a

informação submetida pelos cidadãos poderá prestar um melhor serviço, quer ao nível das recolhas, quer ao nível da manutenção de equipamentos.

O SPAR Portal é um site que integra com o SPAR BackOffice. A informação existente no SPAR BackOffice pode ser incluída no portal (de forma automática) e toda a informação introduzida no portal pelos cidadãos é recebida e tratada de forma integrada no SPAR BackOffice, diminuindo o tempo de resposta.

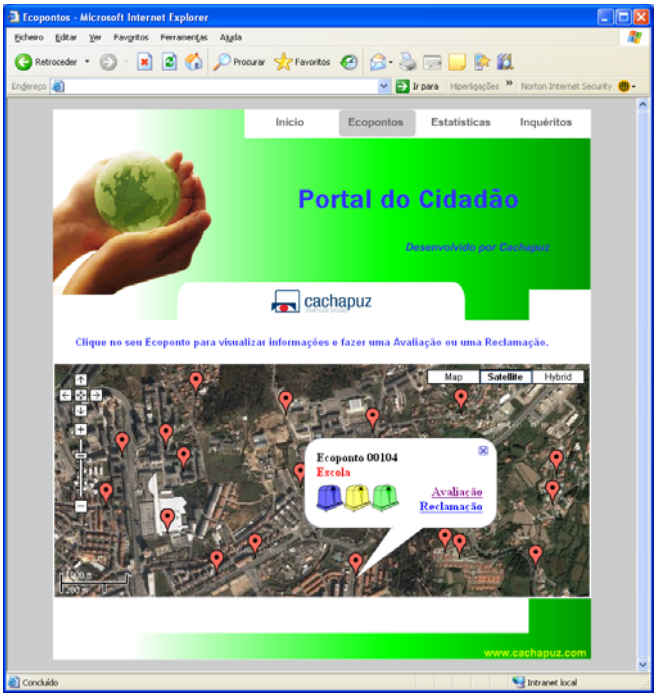

Figura 10 – *Screenshot* do módulo SPAR Portal

O SPAR SMS é um módulo que permite o envio de SMS pelos cidadãos e a sua recepção e tratamento no SPAR BackOffice. Essa interacção pode ser feita através de um SMS simples ou de uma aplicação JAVA[3] que pode ser descarregada para o telemóvel.

[3] Java é uma linguagem de programação orientada aos objectos

Figura 11 – *Screenshot* do módulo SPAR SMS

O SPAR contém na sua arquitectura, para além destes, o módulo de Sincronização, Integração, Business Intelligence e Internet que não serão detalhados uma vez que a sua funcionalidade não é essencial para o âmbito desta dissertação.

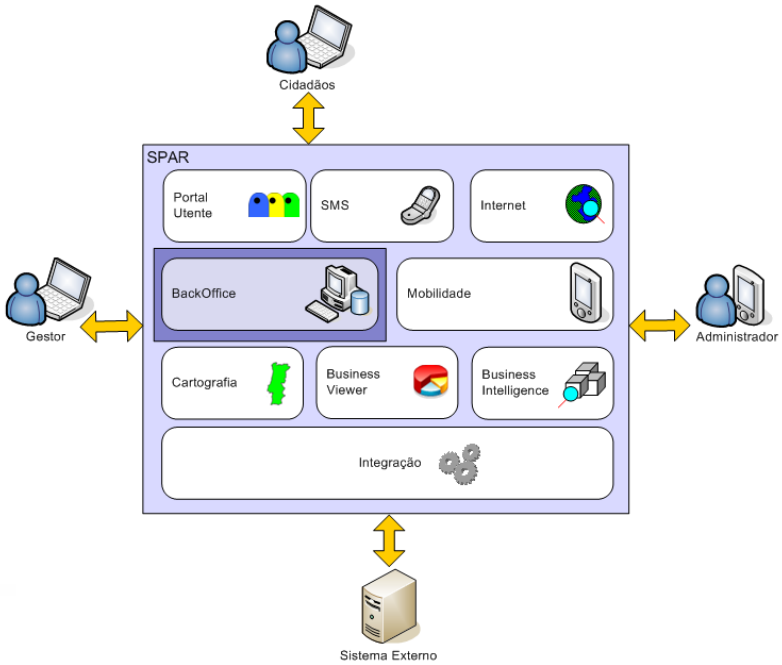

Figura 12 – Arquitectura do SPAR

2.3 Dados registados no processo de recolha selectiva

O SPAR é um sistema que armazena uma elevada quantidade de informação que é introduzida pelas seguintes entidades envolvidas no processo de recolha selectiva:

- **Encarregados/Responsáveis** – São os utilizadores que gerem todo o sistema (utilizando o módulo BackOffice) e que são os responsáveis pela introdução e gestão de toda a informação sobre os motoristas, ajudantes, viaturas de recolha, contentores, ecopontos, produtos e rotas. São estes os utilizadores responsáveis pela análise da informação existente no BackOffice e posterior planeamento de todas as recolhas que serão efectuadas pelas equipas.
- **Motoristas / Ajudantes** – Os motoristas registam para todos os contentores visitados, qual o estado de enchimento actual, se foi ou não recolhido, se necessita ou não de limpeza ou reparação em alguma dos seus constituintes, registam ainda os valores de quilometragem dos camiões da recolha no início e no fim de cada serviço (em alguns casos também registam os quilómetros parciais, ou seja, à chegada a cada ecoponto). Toda esta informação é agrupada em função do plano de trabalho, dando origem a um turno de trabalho. Cada turno, para além da informação para cada um dos contentores visitados, tem também informação das quantidades recolhidas nesse turno por cada um dos produtos (no caso dos camiões bi-compartimentados, as equipas podem recolher dois produtos em simultâneo).
- **Dispositivos** – Para evitar a introdução errada por parte dos utilizadores, o SPAR Mobilidade automatiza a obtenção de alguma informação essencial para o processo. Neste conjunto de dados introduzidos constam as datas e horas de determinados pontos de controlo, como o caso início e do final do turno, o início e o final de cada serviço, as horas de chegada a cada um dos ecopontos visitados. Outra informação que é também registada automaticamente está relacionada com os percursos efectuados por cada equipa de trabalho. Cada *PDA* (no caso de estar equipado com receptor *GPS*), regista

automaticamente e em intervalos de tempo pré-definidos, o pontos do percurso efectuados (coordenadas latitude e longitude).

- **Cidadãos** – Os cidadãos podem também interagir com o SPAR, introduzindo informação sobre o processo. Com o auxílio do módulo SPAR Portal ou SPAR SMS, os cidadãos podem enviar informação directamente para o Sistema, como o caso da necessidade de recolha de um determinado contentor de um ecoponto, a necessidade de reparar ou limpar um determinado contentor, efectuar reclamações ou sugestões e até preenchendo inquéritos, dando ao sistema mais informação essencial para análise.

Toda esta informação é enviada pelos módulos do SPAR para uma base de dados centralizada alojada em *Microsoft SQL Server 2000* ou *Microsoft SQL Server 2005*. No caso do módulo SPAR Mobilidade, os dados são armazenados nos *PDA* temporariamente, em base de dados *SQL Server Compact Edition*, até à sua sincronização com o módulo SPAR BackOffice. Tendo em conta a toda a complexidade do modelo de dados do SPAR e o contexto da sua utilização nesta dissertação, será analisada apenas uma pequena, mas importante parte do mesmo, tendo em conta os objectivos anteriormente descritos. De seguida serão descritas algumas das tabelas analisadas, explicando a sua função no sistema:

- **Turno –** A tabela *Turno* permite armazenar toda a informação referente a um dia de trabalho de recolha selectiva efectuado por uma equipa de trabalho.
- **Movimento –** A entidade *Movimento* permite armazenar todas as rotas de recolha do sistema.
- **Linhas Movimento –** A tabela *Linhas Movimento* permite armazenar toda a informação referente a um registo de uma determinada linha do movimento, por outras palavras, ao registo de informação relacionado com um contentor de um ecoponto associado à rota em causa.
- **Pesos Turno –** A tabela *Pesos Turno* permite armazenar a informação relacionada com os pesos líquidos registados na recolha de cada um dos produtos recolhidos.

- **Serviço –** A tabela *Serviço* armazena toda a informação referente a todos os serviços realizados num determinado turno (descargas).
- **Pesos Serviço –** Esta tabela armazena a informação relacionada com os pesos líquidos registados na recolha de cada um dos produtos recolhidos no serviço associado.
- **Ecoponto –** A tabela *Ecoponto* permite armazenar toda a informação associada a cada um dos ecopontos do sistema.
- **Contentor –** A tabela *Contentor* permite armazenar toda a informação associada a cada um dos contentores do sistema.
- **Produto -** A tabela permite armazenar toda a informação relacionada com os produtos existentes no sistema.

Operador de Resíduos	Contentores
Algar - Valorização e Tratamento de Resíduos Sólidos, S.A.	6369
Amarsul - Valorização e Tratamento de Resíduos Sólidos, S.A.	6519
Ersuc - Resíduos Sólidos do Centro, S.A.	8338
Rebat - Valorização e Tratamento de Resíduos Sólidos, S.A.	1262
Resat - Valorização e Tratamento de Resíduos Sólidos, S.A.	953
Residouro - Valorização e Tratamento de Resíduos Sólidos, S.A.	901
Resiestrela - Valorização e Tratamento de Resíduos Sólidos, S.A.	1279
Resioeste - Valorização e Tratamento de Resíduos Sólidos, S.A.	4866
Resulima - Valorização e Tratamento de Resíduos Sólidos, S.A.	2250
Suldouro - Valorização e Tratamento de Resíduos Sólidos Urbanos, S.A.	3878
Valnor - Valorização e Tratamentos de Resíduos Sólidos do Norte Alentejano, S.A.	2750
Valorlis - Valorização e Tratamento de Resíduos Sólidos, S.A.	2617
Valorminho - Valorização e Tratamento de Resíduos Sólidos, S.A.	1001
Valorsul - Valorização e Tratamentos de Resíduos Sólidos da Área Metropolitana de Lisboa (norte), S.A.	7014

Tabela 2 – Número de contentores por entidade do grupo AdP

Tendo em conta apenas os 14 operadores de resíduos pertencentes ao grupo AdP, existem em Portugal cerca de 50000 contentores para a recolha selectiva.

Se tivermos em conta que uma empresa deste grupo terá, em média, cerca de 3570 contentores para recolher, e se recolher cada um desses contentores duas vezes por semana, o volume de informação recolhida andará na ordem dos 90000 registos por ano, com todos os atributos associados.

À medida que o volume de dados nas empresas aumenta, aumenta também a dificuldade de extrair desses dados, informação importante que auxilie na tomada de decisão [Sumathi e Sivanandam, 2006]. Perante este cenário, podemos concluir que as ferramentas tradicionais de gestão da recolha selectiva, como o caso do SPAR, embora respondam bem aos processos operacionais, não se adequam aos processos de tomada de decisão. Estas limitações existentes no SPAR e nas outras ferramentas operacionais levaram as empresas a procurarem no mercado novas áreas de investigação e novos paradigmas para a descoberta de conhecimento na sua informação de negócio que auxilie os gestores na tomada de decisão. No Capítulo 3 desta dissertação serão abordadas algumas das técnicas de *data mining* existentes bem como a sua adaptabilidade na indústria, e em particular na recolha selectiva.

Capítulo 3

Mineração de dados na indústria

3.1 Estudo de técnicas de mineração de dados

Algumas das técnicas de mineração de dados [Olson e Delen, 2008] existentes são a Associação, Classificação, *Clustering* e Previsão e devido às suas características apenas podem ser aplicadas nos cenários correctos. De seguida serão explicadas brevemente estas técnicas, com um pequeno exemplo, realçando as vantagens da sua utilização.

3.1.1 A técnica de Associação

O modelo de associação, também denominado de regras de associação, foi desenvolvido no campo das ciências da computação e é intensivamente usado em áreas como a análise de cabazes de compras (*market basket analysis*) para medir as associações ou relacionamentos entre os produtos

adquiridos por um determinado cliente, e na análise de cliques em páginas Internet para medir associações ou relacionamentos entre páginas vistas sequencialmente pelos visitantes de um site [Giudici, 2003]. Em geral, o objectivo deste modelo é o de encontrar grupos de acontecimentos que normalmente ocorrem juntos num determinado conjunto de dados. A informação onde é aplicada a associação está armazenada em bases de dados de transacções. Um exemplo de cenário onde podemos aplicar a técnica de associação é na venda de produtos num estabelecimento comercial. Se numa determinada percentagem de compras do produto A, também tiver sido adquirido o produto B, e se essa percentagem for suficientemente representativa no número de compras existente, podemos admitir que os produtos A e B estão relacionados de alguma forma. Este tipo de conclusão indica que no futuro, é provável que este padrão seja mantido. No exemplo da Figura 13, podemos encontrar as associações existentes entre os vários produtos, por exemplo, o vinho está associado com a carne, o leite com o queijo, entre outras associações.

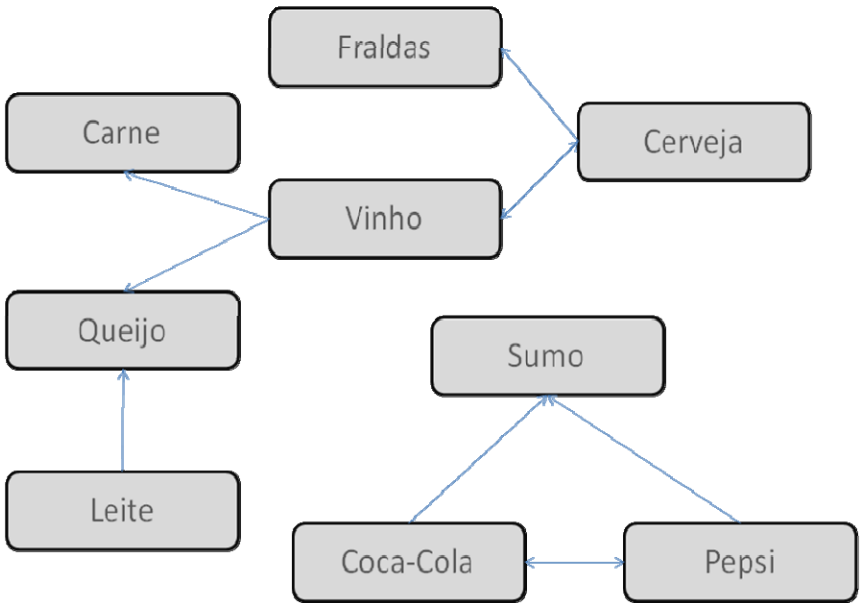

Figura 13 – Associação entre produtos

Com base nesta informação, os gerentes de superfícies comerciais poderão levar a cabo acções que permitam aumentar as vendas da seguinte forma:

- Reformular a disposição dos produtos nas suas lojas, colocando os produtos associados próximos uns dos outros.
- Utilizar a informação para optimizar as quantidades dos produtos em stock.
- Criar campanhas de publicidade para determinados produtos.
- Determinar as tendências dos consumidores.

3.1.2 A técnica de Classificação

A técnica de Classificação consiste em examinar as características de um novo objecto atribuindo-o a um conjunto predefinido de classes [Berry e Linoff, 2004]. Os objectos classificados são normalmente representados através de registos de uma tabela de uma base de dados ou de um ficheiro e o acto de classificação em adicionar uma nova coluna com o código de uma qualquer classe. A técnica de Classificação é então caracterizada pela existência de um conjunto finito de classes e um conjunto de registos previamente classificados que servirão de casos de treino. O modelo "aprende" com os dados de treino e será capaz de classificar a nova informação submetida, classificando-a com base nos seus atributos. Neste tipo de técnica são normalmente usadas redes neuronais. Na Figura 14 podemos ver uma tabela que indica, em função do peso, da altura e do sexo, a classificação de uma criança. As classes possíveis são Magro, Obeso, Ideal e foram calculadas em função do IMC (índice de massa corporal). Esta tabela poderia ser usada como dados de treino para um modelo de classificação. Após treinado e testado o modelo, este seria capaz de responder à questão "O Manuel, com 1,3 m de altura, com peso 50 kg, está obeso, está com o peso ideal ou está com peso abaixo do normal?".

Peso	Sexo	Altura (m)	Classe
50	F	1,2	Obeso
50	M	1,3	Obeso
35	M	1,2	Ideal
65	F	1,8	Ideal
35	F	1,8	Magro
40	M	1,5	Magro

Figura 14 – Dados de treino

3.1.3 A técnica de *Clustering*

O algoritmo de *Clustering* permite encontrar agrupamentos naturais na informação quando esses grupos não são óbvios [Tang e MacLennan, 2005]. Por outras palavras, podemos dizer que o algoritmo permite encontrar a variável escondida, que com maior precisão classifica os nossos dados. Tomemos como exemplo a observação de um conjunto elevado de viajantes a recolherem as suas malas nos tapetes de um aeroporto. Neste conjunto de pessoas existem dois grupos, as pessoas do primeiro grupo estão vestidas com *t-shirts* e calções e as pessoas do segundo grupo estão vestidas com casacos, calças e camisolas quentes. Neste exemplo conseguimos encontrar uma variável escondida: as pessoas do primeiro grupo regressaram de um clima tropical e as pessoas do segundo regressaram de um local molhado e de temperaturas baixas. Esta capacidade de agrupar a informação faz deste tipo de técnica de mineração, uma técnica muito usada, nomeadamente no marketing. Na Figura 15 podemos ver um exemplo em que, com base na análise da informação de uma base de dados, tendo em conta os atributos vencimento e idade,

pessoas foram agrupadas em três diferentes grupos, possivelmente desconhecidos até ao momento.

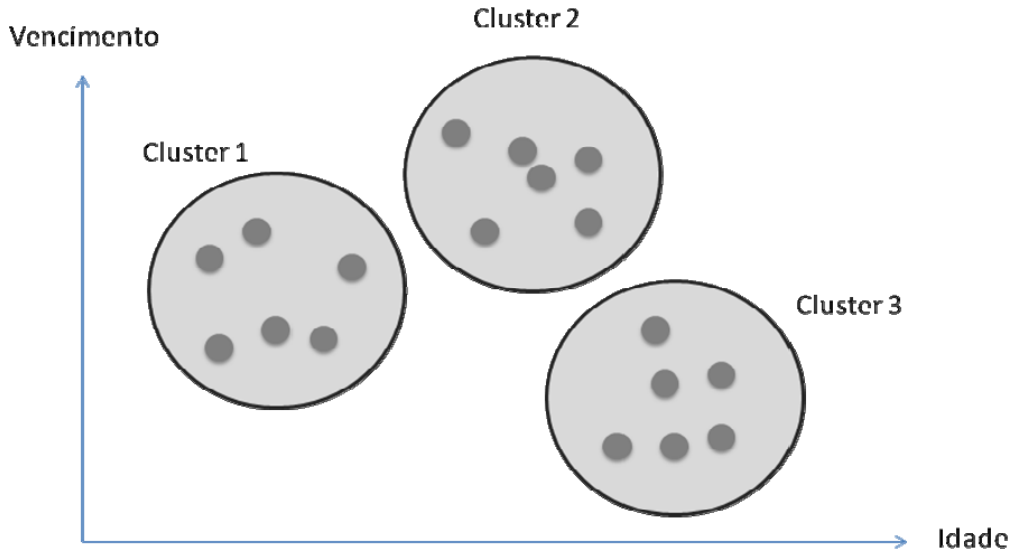

Figura 15 – Exemplo de *Clustering*

3.1.4 A técnica de Previsão

A técnica de Previsão é similar à técnica de Classificação, com a diferença que os registos são classificados de acordo com a previsão de comportamentos futuros ou com valores estimados futuros [Berry e Linoff, 2004]. Na técnica de previsão, a única forma de validar os resultados é a de esperar que os acontecimentos ou valores aconteçam. A principal razão pela qual a técnica de Previsão é mantida separada da técnica de Classificação é que o modelo de previsão tem características adicionais que estão relacionadas com a relação temporal entre os dados independentes e os dependentes. Como exemplos de aplicação desta técnica, é possível destacar a previsão dos lucros (variável dependente) com base nas vendas (variável independente), a previsão do comportamento das vendas no futuro tendo em conta o histórico de vendas, ou o comportamento futuro do valor das acções de uma determinada empresa.

3.2 Casos práticos de aplicação

As técnicas de *data mining*, devido às suas vantagens, têm sido aplicadas num crescente e diversificado número de indústrias. De seguida serão descritas algumas dessas aplicações de *data mining*, bem com as vantagens dessas aplicações nas indústrias em causa.

3.2.1 Publicidade e *Marketing*

Contrariamente ao que a maioria das pessoas pensam, o *marketing* não é composto apenas pela publicidade e pela comercialização de bens e serviços [Web MOTI]. Estas duas actividades fazem parte de todo um universo de actividades que fazem do marketing uma potente ferramenta de sucesso a que as empresas não podem estar indiferentes. No geral, o marketing envolve as seguintes actividades principais:

- Identificar as necessidades dos clientes no mercado alvo.
- Satisfazer essas necessidades melhor do que a concorrência.

Para que o marketing funcione, as empresas apostam em ferramentas que as auxiliem na investigação dos gostos, tendências e necessidades dos seus consumidores, analisando toda a informação disponível no seio da empresa, como o caso de encomendas, vendas, inquéritos, entre outros tipos de dados. É com base nos resultados dessa análise, que as empresas tomam decisões de negócio que englobam, por exemplo, a definição de preços, design, promoção e distribuição dos seus produtos. Segundo os autores [Boone e Kurtz, 1998], no seu livro *Contemporary Marketing Wired*, o "*Marketing é o processo de planear e executar a concepção, estabelecimento do preço, promoção e distribuição de ideias, bens, serviços, organizações e eventos para criar e manter relações que irão satisfazer os objectivos individuais e organizacionais.*".

Para atingir tais objectivos, as empresas necessitam de ferramentas que as auxiliem na tomada de decisão com base em toda a informação disponível. É neste contexto que, ultimamente, as empresas têm investido cada vez mais em soluções baseadas em *data mining*, provocando com que o *data mining* e a gestão de campanhas estejam cada vez mais interligados.

Algumas das vantagens da utilização de técnicas de *data mining* no marketing são as seguintes [Sumathi e Sivanandam, 2006]:

- Aumentar a rapidez do planeamento e execução as campanhas de marketing.
- Aumentar o grau de sucesso das campanhas criadas.
- Aumentar o retorno do investimento em marketing.
- Permitir a análise das vendas anteriores em função da idealização de novos produtos.
- Encontrar padrões nas compras efectuadas por parte dos consumidores para permitir executar campanhas orientadas às suas necessidades.

As técnicas de *data mining* mais utilizadas neste sector de actividade são:

- **Classificação** - Que permite classificar os consumidores em função dos seus atributos, como por exemplo, a idade, a região, o sexo, estado civil, emprego, entre outros.
- ***Clustering*** – Que permite organizar os consumidores em grupos com base nos seus padrões de compras.
- **Associação** – Que permite associar os produtos entre si com base em compras anteriores.
- **Previsão** – Que permite prever quando e em que quantidades os produtos serão consumidos.

3.2.2 Sistemas de Detecção de Intrusão

Um dos principais desafios na gestão de segurança das redes em larga escala e de elevada velocidade consiste na detecção de anomalias suspeitas nos padrões de tráfego de rede. Estas anomalias podem ser causadas por ataques DDoS[4] ou propagação de programas maliciosos (worms[5]) [Web IDS]. Uma rede é segura se garantir os seguintes pontos:

- **Confidencialidade dos dados** – A informação que circula na rede apenas poderá estar acessível a pessoas autorizadas.
- **Integridade dos dados** – A informação deverá manter-se intacta desde o seu envio até à sua recepção, isto é, não poderão existir corrupção ou perda de dados.
- **Disponibilidade dos dados** – A rede deverá ser resistente a ataques maliciosos.

Grandes empresas como a *Yahoo*, *eBay*, *Amazon*, entre outras, foram já vítimas, no passado, de ataques deste tipo [Web Attacks], e se as empresas não estão preparadas convenientemente para prevenir e impedir estes ataques, podem ser afectadas, dando origem a prejuízos avultados.

Devido ao enorme volume de pacotes de dados que circulam em redes deste tipo, a utilização do *data mining* é útil nos seguintes pontos:

- Eliminar a actividade normal da informação para permitir aos analistas focalizarem-se em ataques reais.
- Encontrar actividades anómalas que escondam ataques reais.
- Identificar prolongados padrões que aconteçam nas redes.

[4] *Distirbuted Denial of Service*
[5] *Worm* é um programa que se replica numa rede informática e que normalmente produz actividades maliciosas

As técnicas mais usadas neste sector de actividade são as seguintes:

- **Classificação** – Classificar a informação do tráfego existente em duas classes, tráfego normal e tráfego maliciosos.
- ***Clustering*** – Particionar a informação em subconjuntos, separando os dados potencialmente maliciosos dos dados normais.

3.2.3 Medicina

Segundo a definição encontrada no *American Cancer Society* [Web Cancer], o cancro é um grupo de doenças caracterizadas pelo aumento descontrolado de células anormais. Se este crescimento não for controlado a tempo, estas doenças provocam à morte dos pacientes. O cancro pode ser provocado pelo tabaco, químicos, radiação, organismos infecciosos, mutações internas, hormonas, problemas de imunidade, entre outras causas. A taxa de sobrevivência dos casos diagnosticados destas doenças têm aumentado nos últimos anos, já que entre 1975 e 1977 esta taxa rondava o valor de 51% e aumentou, no período de 1996 a 2002, para um valor de 66%. Este aumento da taxa de sobrevivência ao cancro é reflexo das melhorias nos processos de diagnóstico e nos tratamentos dos pacientes.

Investigadores desta área tentam, a todo o custo, encontrar mecanismos que os auxiliem na compreensão da doença para encontrarem a tão desejada cura. Recentemente, e com base em bases de dados com registos relativos a estas doenças, os investigadores utilizam as tecnologias emergentes para a obtenção de dados que lhes permitam combater o cancro. Uma destas tecnologias usadas é o *data mining*. O *data mining* pode ser utilizado para encontrar padrões escondidos na informação de casos de cancro passados tendo em conta os atributos dos pacientes, idade, tipo de cancro, sexo, tipo de sangue, entre outros e se estes sobreviveram ou não ao cancro.

Segundo o estudo "*Predicting Breast Cancer Survivability Using Data Mining Techniques*" levado a cabo por [Web Siam], é possível a aplicação do *data mining* em bases de dados médicas para prever a sobrevivência de pacientes. Neste estudo, os autores recorreram a técnicas de Classificação, nomeadamente utilizando os algoritmos Árvores de Decisão, Redes Neuronais e *Naïve Bayes* para comparar os resultados de previsão da taxa de sobrevivência de pacientes com cancro da mama.

3.3 Mineração de dados na recolha selectiva

A optimização dos processos de recolha de resíduos tem sido alvo de inúmeros estudos que tentam encontrar a melhor solução para este problema. Um desses estudos foi desenvolvido pelo [Simonetto e Borenstein, 2005]. Neste estudo, os autores, utilizaram a Investigação Operacional e um Sistema de Informação Geográfica (SIG) para tentar resolver o problema da recolha selectiva no Brasil, Rio Grande do Sul, tendo em conta os seguintes objectivos:

- Reduzir a quantidade de resíduos sólidos destinada à deposição em aterros sanitários.
- Garantir um a cadência de chegada de resíduos em cada unidade de triagem.
- Alocação de veículos de recolha.
- Definir os percursos de recolha óptimos.
- Estimar a capacidade de trabalho (produtividade) das unidades de triagem, em relação à chegada e ao processamento (separação) de resíduos.

Embora este estudo seja importante na optimização das entregas de resíduos tendo em conta a capacidade do local de entrega dos resíduos, a estação de triagem, o que acontece na realidade, por exemplo em Portugal, é que as empresas dispõem apenas de uma central de triagem com uma capacidade de escoamento bastante grande, ou dispõem de estações de transferência, que servem apenas como depósitos intermédios de resíduos. Logo, podemos admitir que têm uma capacidade

as centrais de triagem tem uma capacidade de produção (triagem) infinita. O estudo em causa foi desenvolvido sem ter em conta outro aspecto muito importante na gestão dos processos de recolha selectiva que é o facto de que os contentores terem de ser recolhidos no limiar da sua capacidade. Neste tipo de estudos, a taxa de enchimento dos contentores é de extrema importância e as recolhas tem que ser efectuadas de forma a garantir que os cidadãos sejam capazes de depositar os seus resíduos.

Outro estudo nesta área, mas desta feita levado a cabo pelos autores [Alves e Carvalho, 2004], também este no âmbito da área da Investigação operacional, os autores analisam o problema da optimização de rotas de recolha de desperdícios de madeira. É um problema similar à recolha selectiva, já que o objectivo passa por efectuar a recolha de contentores de resíduos, com necessidades semelhantes às dos contentores da recolha selectiva, isto é, necessitam de ser recolhidos tendo em conta o seu enchimento. Os autores derivaram o algoritmo *Vehicle Routing Problem* (VRP) [Balas, 1989], no algoritmo que denominaram de *Prize Collecting Vehicle Routing Problem with service restrictions* (PCVRPsr), para tentarem resolver o problema da optimização das rotas de recolha de desperdícios de madeira. Neste estudo foram assumidos números de camiões de recolha infinitos, contrariamente ao que acontece na realidade.

Em nenhum destes estudos foram utilizadas técnicas de *data mining*, tendo sido levados a cabo com base na Investigação Operacional. Não foi possível então encontrar na bibliografia aplicações do *data mining* aos processos de recolha selectiva, que permitissem efectuar comparações mais detalhadas com o trabalho que deu origem a esta dissertação.

Contudo e com base na análise dos mecanismos de *data mining* descritos neste capítulo, foi possível seleccionar as técnicas de *Clustering* e Associação para o desenvolvimento de dois casos de aplicação. Esta escolha foi apoiada tanto nas características destas técnicas, como no paralelismo das suas aplicações na indústria com o problema da reorganização dos contentores em rotas de recolha selectiva.

Capítulo 4

Optimização de rotas através de técnicas de mineração de dados

4.1 O caso de estudo

Neste capítulo será descrito todo o trabalho desenvolvido nesta dissertação, tendo por base a metodologia *CRISP-DM*, já explicada, desde a fase da Análise dos Dados *(Data Understanding)*, até à fase de Análise *(Evaluation)*, descrevendo cada uma das fases intermédias. Como caso de estudo, foram utilizados os dados da empresa Resulima[6], registados pelas equipas de trabalho com o auxílio da ferramenta SPAR. Estes dados são compostos por 595298 registos no total e foram registados entre os dias 02/01/2007 e 31/12/2007, ou seja um ano completo de informação sobre a recolha selectiva. A ferramenta usada neste projecto foi o *Microsoft SQL Server 2005*, para a análise e

tratamento da informação e o *Microsoft SQL Server 2005: Analysis Services* para modelação e visualização dos resultados.

4.2 Aplicação de Associação

4.2.1 Análise dos Dados

O primeiro passo na aplicação do *data mining* consistiu na Análise dos Dados, (*Data Understanding),* de toda a base de dados SPAR da Resulima para perceber toda a informação registada pelas equipas de trabalho no terreno, para isso, foram recolhidos, descritos, explorados e analisados todos os dados necessários à fase seguinte, a fase de Preparação dos Dados (*Data Preparation*).

Recolha Inicial de Dados *(Collect Initial Data)*

Nesta fase foi solicitada a base de dados do SPAR à Resulima que ocupava aproximadamente 312Mb e de seguida foi carregada para o *SQL Server 2005*. Os dados foram carregados correctamente e sobre eles foram efectuadas consultas preliminares, nomeadamente para saber a quantidade de registos existentes referentes às recolhas e em que datas foram criados. Esta consulta foi efectuada na tabela *Linhas Movimento*, onde é regista toda a informação dos contentores por parte dos motoristas e ajudantes, e revelou 595298 registos. Na Figura 16, podemos ver uma amostra dessa tabela, apenas com alguns dos atributos mais importantes para este trabalho. Existem ainda outras tabelas também usadas para recolher informação complementar, como o caso das tabelas *Ecoponto*, *Contentor*, *Produto*, *Turno*, *Serviço*, *Linhas Movimento*, *Movimento*, *Peso Turno*, *Pesos Serviço* e *Pesos Turno.*

[6] Resulima, empresa que efectua a recolha selectiva no Distrito de Viana do Castelo e parte do Distrito de Braga.

Data	Contentor	Ecoponto	Produto	Enchimento	Recolhido
2007-01-02 04:24:43.000	02659	00887	01	88	1
2007-01-02 04:24:43.000	02660	00887	02	88	1
2007-01-02 04:24:43.000	02661	00887	03	13	0
2007-01-02 04:30:48.000	02662	00888	01	13	0
2007-01-02 04:30:48.000	02663	00888	02	88	1
2007-01-02 04:30:48.000	02664	00888	03	88	1
2007-01-02 04:37:57.000	02665	00889	01	88	1
2007-01-02 04:37:57.000	02666	00889	02	63	1
2007-01-02 04:37:57.000	02667	00889	03	13	0
2007-01-02 04:45:40.000	02668	00890	01	63	1
2007-01-02 04:45:40.000	02669	00890	02	38	1
2007-01-02 04:45:40.000	02670	00890	03	13	0
2007-01-02 04:52:51.000	02671	00891	01	63	1
2007-01-02 04:52:51.000	02672	00891	02	38	1
2007-01-02 04:52:51.000	02673	00891	03	13	1
2007-01-02 04:59:10.000	02674	00892	01	63	1
2007-01-02 04:59:10.000	02675	00892	02	63	1
2007-01-02 04:59:10.000	02676	00892	03	13	0
2007-01-02 05:05:29.000	02677	00893	01	38	1

Figura 16 – Extracto da tabela *Linhas Movimento*

Descrição dos Dados *(Describe Data)*

A Tabela 3 representa o número de registos de cada uma das tabelas usadas neste trabalho:

A entidade *Linhas Movimento* contem os registos de todos os contentores visitados durante o processo de recolha selectiva. Na Tabela 4, é possível consultar os tipos de dados de cada um dos atributos da entidade *Linhas Movimento*. A tabela *Ecoponto* permitirá descobrir quais os Distritos e Concelhos de cada um dos contentores. Na tabela 5 é possível consultar os tipos de dados de cada um dos atributos desta entidade.

Tabela	Registos
Ecoponto	1119
Contentor	3249
Produtos	3
Movimento	16
Linhas Movimento	595289
Turno	2145
Serviço	4608
Pesos Turno	4280
Pesos Serviço	8735

Tabela 3 – Número de registos por tabela

Coluna	Tipo de Dados	Chave
Código do turno	Alfanumérico	X
Código do movimento	Alfanumérico	X
Código da linha	Numérico	X
Código do serviço	Numérico	
Ecoponto	Alfanumérico	
Contentor	Alfanumérico	
Produto	Alfanumérico	
Nível de enchimento	Numérico	
Estado de higiene	Alfanumérico	
Produto fora	Numérico	
Recolhido	Numérico	
Observações	Alfanumérico	
Data de início de operação	Data	
Data de início de operação	Data	
Quilómetros de operação	Numérico	

Tabela 4 – Tipos de dados da tabela *Linhas Movimento*

Coluna	Tipo de Dados	Chave
Código	Alfanumérico	X
Descrição	Alfanumérico	
Morada	Alfanumérico	
Localidade	Alfanumérico	
Freguesia	Alfanumérico	
Concelho	Alfanumérico	
Distrito	Alfanumérico	
Data de instalação	Data	
Latitude	Numérico	
Longitude	Numérico	

Tabela 5 – Tipos de dados da tabela *Ecoponto*

Exploração de Dados *(Explore Data)*

Durante a exploração dos dados disponíveis foi possível identificar que o Concelho do contentor como um atributo importante para a análise e que deveria ser incluído no modelo. Observando a Figura 17 é possível constatar que o número de registos por Concelho é bastante heterogéneo, indicando que as taxas de enchimento variam em função deste atributo.

	Distrito	Concelho	Registos
1	16	01	23601
2	03	06	46310
3	16	07	30898
4	03	02	85804
5	16	06	9255
6	16	09	102152

Figura 17 – Número de registos por Concelho

Foi identificado também que as quantidades recolhidas seriam importantes para a análise. Na Figura 18, é possível verificar as quantidades totais por produto.

	Produto	Quantidade
1	Embalagem	1011851
2	Papel	2785095
3	Vidro	4400819

Figura 18 – Quantidades totais por produto

Verificar a Qualidade dos Dados *(Verify Data Quality)*

Após análise dos dados, foi possível verificar que cerca de 50% dos registos existentes não poderiam ser usados no trabalho já que não continham qualquer informação. Este problema tem como origem o tamanho das rotas existentes, já que os motoristas não são capazes de visitar todos os contentores que lhes são atribuídos durante um turno de trabalho. Na Figura 19 é possível consultar alguns desses dados, cujo enchimento tem valor igual a "-1". Existiam inicialmente 294047 registos deste tipo que não foram carregados na fase de Selecção dos Dados, tendo sobrado apenas 298020 registos para análise.

	Contentor	Enchimento
1	02791	-1
2	02792	-1
3	02793	-1
4	02794	-1
5	02795	-1
6	02796	-1
7	02797	-1
8	02798	-1
9	02799	-1
10	02827	-1
11	02828	-1
12	02829	-1

Figura 19 – Registos sem o valor do Enchimento

4.2.2 Preparação dos Dados

A preparação dos dados para a fase de Modelação *(Modeling)*, teve inicio na criação de uma tabela auxiliar, com o nome *Registos*, e cuja estrutura pode ser consultada na Tabela 6, para sobre ela ser aplicado o modelo em estudo.

Coluna	Tipo de Dados
Data	Data
Contentor	Alfanumérico
Produto	Alfanumérico
Quantidade	Numérico
Enchimento	Numérico
Freguesia	Alfanumérico
Concelho	Alfanumérico
Distrito	Alfanumérico

Tabela 6 – Tipos de dados da tabela *Registos*

Selecção de Dados *(Select Data)*

Foram seleccionados para análise os registos existentes na tabela *Linhas Movimento* e *Ecoponto* cuja informação era válida, ou seja, cujo valor do atributo Enchimento tinha sido registado. Esta informação foi inserida na tabela *Registos*, passando apenas os registos válidos. O processo de Limpeza de Dados *(Clean Data)* não foi aplicado já que foram filtrados apenas os registos e atributos necessários.

Criação de Dados *(Construct Data)*

Nesta fase foi derivado o atributo *Quantidade* necessário para a tabela *Registos*, que não existia nas tabelas do SPAR:

- **Quantidade** – O valor do atributo Quantidade foi calculado através de uma estimativa dos em função dos pesos de todos os contentores recolhidos num determinado turno e dos valores dos enchimentos registados, já que não é possível saber quanto pesa cada contentor. Esta estimativa foi efectuada através da seguinte fórmula: Quantidade = (Peso Total * Enchimento)/ ∑ Enchimento.

Integridade de Dados *(Integrate Data)*

Toda a informação necessária foi inserida na tabela *Registos*, tendo por base as tabelas *Ecoponto*, *Contentor*, *Produto*, *Turno*, *Serviço*, *Linhas Movimento*, *Movimento*, *Peso Turno*, *Pesos Serviço* e *Pesos Turno*. Um extracto desta tabela pode ser consultado na Figura 20. Toda a informação foi formatada nesta fase, pelo que não houve necessidade da fase de *Format Data.*

DataPartida	Ecoponto	Contentor	Produto	Kgs	Enchimento	Distrito	Concelho	Freguesia
11-06-2007 4:02:37	00840	02518	01	36,46...	63	Braga	Esposende	Fão
11-06-2007 4:02:37	00841	02521	01	50,93...	88	Braga	Esposende	Fão
11-06-2007 4:45:26	00147	00440	02	15,55...	38	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00151	00452	02	25,78...	63	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00152	00455	02	15,55...	38	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00153	00458	02	15,55...	38	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00147	00439	01	34,72...	38	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00150	00448	01	34,72...	38	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00151	00451	01	57,56...	63	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00152	00454	01	57,56...	63	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00153	00457	01	80,40...	88	Viana do Castelo	Viana do Castelo	Afife
11-06-2007 4:45:26	00124	00371	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00125	00374	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00126	00377	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00128	00383	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00132	00395	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00133	00398	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00134	00401	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00135	00404	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00136	00407	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	01080	03252	02	15,55...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00125	00373	01	34,72...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00126	00376	01	34,72...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00128	00382	01	34,72...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00132	00394	01	34,72...	38	Viana do Castelo	Viana do Castelo	Areosa
11-06-2007 4:45:26	00133	00397	01	34,72...	38	Viana do Castelo	Viana do Castelo	Areosa

Figura 20 – Extracto da tabela *Registos*

4.2.3 Modelação

Toda a informação existente na base de dados do SPAR da Resulima foi analisada, compreendida, escolhida, filtrada, e completada durante as fases Análise e Preparação dos Dados. Essa informação foi inserida numa tabela auxiliar, a tabela *Registos* e foi sobre esta tabela que o modelo de Associação foi aplicado. Neste capítulo de modelação serão descritas brevemente as fases referentes à fase de Modelação *(Modeling)* do *CRISP-DM*, bem como todos os passos na criação, configuração e aplicação do modelo escolhido, sobre os dados.

Selecção da Técnica de Modelação *(Select Modeling Technique)*

O modelo de *data mining* escolhido foi o modelo *Associação*, baseado no algoritmo *Microsoft Association Rules*, existente no *Microsoft SQL Server 2005: Analysis Services*.

Criação e Interpretação do Modelo *(Build and Assess Model)*

Nesta fase serão descritos todos os passos na construção de um modelo de *Associação*, utilizando o algoritmo *Microsoft Association Rules*. Em primeiro lugar foi criada a origem de dados, a tabela *Registos*, através de uma vista de dados.

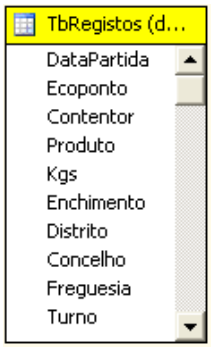

Figura 21 – Vista da origem de dados

De seguida foi criada uma estrutura de *mining* com os seguintes parâmetros:

- Selecção do *Microsoft Association Rules* como técnica de *data mining*.
- Selecção da tabela *Registos* como "*Case*".
- Selecção do atributo *DataPartida* como *Key*, e os atributos *Concelho*, *Contentor*, *Distrito*, *Enchimento*, *Freguesia* e *Produto* como *Input*.
- Selecção do atributo *Kgs* como *Predict*.
- Selecção do nome para o modelo, *Registos* e selecção da opção *Allow drill through* para permitir navegar pelas regras.
- Os parâmetros do algoritmo foram mantidos com valores por defeito.

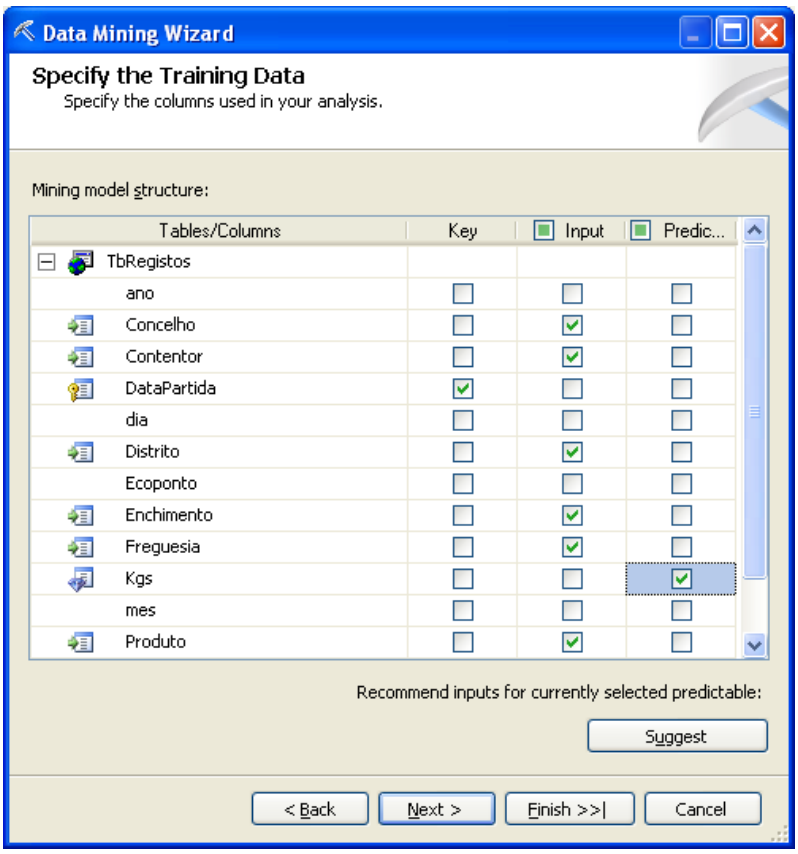

Figura 22 – Selecção dos atributos *Key*, *Input e Predict*

4.2.4 Avaliação

Após a execução deste modelo, foram descobertas regras que associam os atributos *Produto*, *Enchimento, Distrito* e *Concelho* com as quantidades previstas. Por exemplo, se o produto for vidro e o enchimento registado for superior a 75%, então o peso do conteúdo do contentor será superior a 567kg, com uma probabilidade superior a 0,5. Outra regra encontrada foi que a que relaciona o Distrito com o produto, que diz que se o produto foi igual a vidro e o Distrito foi igual a "Viana do Castelo", o enchimento terá valores entre 285kg e 437kg aproximadamente. Esta previsão tem uma probabilidade superior a 0,4.

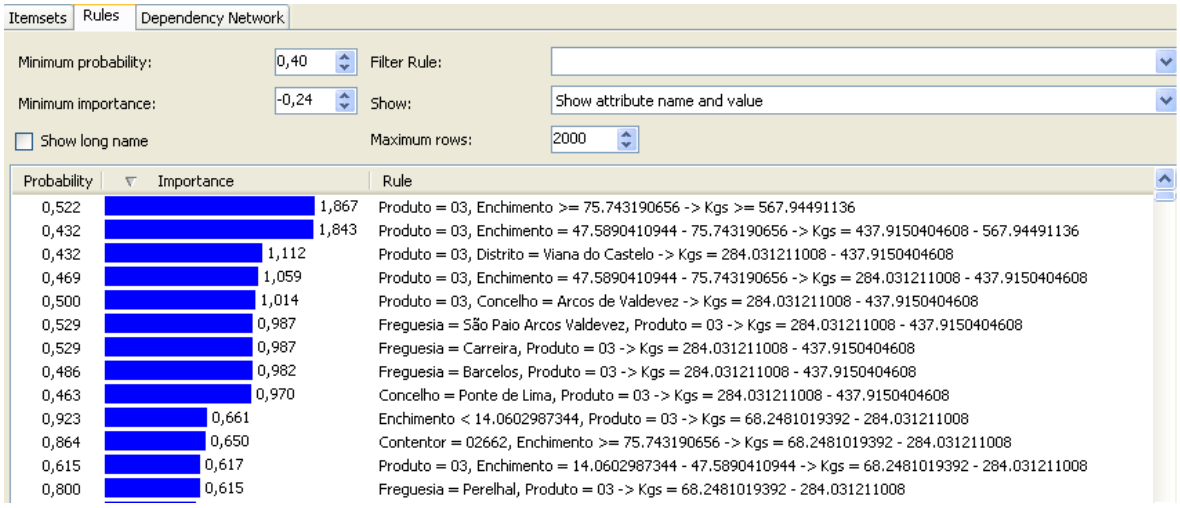

Figura 23 – Regras encontradas pelo modelo

4.3 Aplicação de *Clustering*

4.3.1 Análise dos Dados

A fase do *CRISP-DM* referente à Análise da dos Dados do SPAR não foi necessária para a aplicação do *Clustering* visto ter sido executada na aplicação da técnica anterior e descrita no Capítulo 4.2.1.

Por este motivo, apenas serão descritas as fases posteriores que tiveram início na fase de Preparação dos Dados.

4.3.2 Preparação dos Dados

A preparação da informação para análise teve inicio na criação de uma tabela auxiliar, com o nome *Recolhas*, e cuja estrutura pode ser consultada na Tabela 7, para sobre ela ser aplicado o modelo em estudo.

Coluna	Tipo de Dados
Ecoponto	Alfanumérico
Contentor	Alfanumérico
Produto	Alfanumérico
Taxa	Numérico
Quantidade	Numérico
Concelho	Alfanumérico
Distrito	Alfanumérico

Tabela 7 – Tipos de dados da tabela *Recolhas*

Selecção de Dados *(Select Data)*

Foram seleccionados para análise os registos existentes na tabela *Linhas Movimento* e *Ecoponto* cuja informação era válida, ou seja, cujo valor do atributo Enchimento tinha sido registado. Esta informação foi inserida na tabela *Recolhas*, passando apenas os registos válidos. O processo de *Clean Data* não foi aplicado já que foram filtrados apenas os registos e atributos necessários.

Criação de Dados *(Construct Data)*

Nesta fase foram derivados os atributos Taxa e Quantidade necessários para a tabela *Recolhas*, que não existiam nas tabelas do SPAR:

- **Taxa** – O valor da taxa foi calculado através da média dos enchimentos registados pelas equipas de trabalho, através da fórmula: Taxa = ∑ Enchimento/Nº Registos, da tabela *Linhas Movimento*:

Integração de Dados *(Integrate Data)*

Toda a informação necessária foi inserida na tabela Recolhas, tendo por base as tabelas *Ecoponto*, *Contentor*, *Produto*, *Turno*, *Serviço*, *Linhas Movimento*, *Movimento*, *Peso Turno*, *Pesos Serviço* e *Pesos Turno*. Um extracto desta tabela pode ser consultado na Figura 24. Toda a informação foi formatada nesta fase, pelo que não houve necessidade da fase de *Format Data.*

Ecoponto	Contentor	Produto	Taxa	Quantidade	Concelho	Distrito
00788	02363	02	38	15,5525238744...	Esposende	Braga
00672	02015	02	38	16,2442748091...	Barcelos	Braga
00019	00056	02	38	17,6744186046...	Ponte da Barca	Viana do Castelo
00588	01763	02	38	18,1952117863...	Esposende	Braga
00588	01762	01	38	18,8118811881...	Esposende	Braga
00061	00182	02	38	18,8288288288...	Arcos de Valdevez	Viana do Castelo
01062	03185	02	38	20,3208556149...	Viana do Castelo	Viana do Castelo
00215	00644	02	63	24,0983606557...	Viana do Castelo	Viana do Castelo
01065	03194	02	63	26,0901339829...	Viana do Castelo	Viana do Castelo
01071	03221	02	38	27,4140704656...	Esposende	Braga
00276	00827	02	63	28,0415430267...	Ponte de Lima	Viana do Castelo
00552	01655	02	88	28,8052373158...	Barcelos	Braga
00863	02588	02	38	30,0791556728...	Esposende	Braga
010897	03303	02	63	32,2700814901...	Viana do Castelo	Viana do Castelo
01062	03184	01	38	33,9908256880...	Viana do Castelo	Viana do Castelo
00966	02897	02	88	34,4032921810...	Viana do Castelo	Viana do Castelo
00145	00434	02	38	34,6468783672...	Viana do Castelo	Viana do Castelo
00278	00833	02	38	36,8334915683...	Ponte de Lima	Viana do Castelo
00003	00008	02	38	39,5017257641...	Arcos de Valdevez	Viana do Castelo
00407	01220	02	38	40,2268431001...	Ponte de Lima	Viana do Castelo
00016	00047	02	88	40,9302325581...	Ponte da Barca	Viana do Castelo
00020	00059	02	88	40,9302325581...	Ponte da Barca	Viana do Castelo
010897	03312	01	63	41,2438625204...	Viana do Castelo	Viana do Castelo
00367	01099	01	63	41,2776412776...	Barcelos	Braga
01106	03339	02	63	43,0176211453...	Ponte de Lima	Viana do Castelo
00033	00098	02	38	43,0729781933...	Ponte da Barca	Viana do Castelo
00189	00566	02	88	44,0918580375...	Viana do Castelo	Viana do Castelo

Figura 24 – Extracto da tabela *Recolhas*

4.3.3 Modelação

Toda a informação necessária foi inserida na tabela auxiliar *Recolhas* e foi sobre esta tabela que o *data mining* foi aplicado. Neste capítulo de modelação serão descritas brevemente as fases referentes ao *Modeling* do *CRISP-DM*, bem como todos os passos na criação, configuração e aplicação do modelo escolhido, sobre os dados.

Selecção da Técnica de Modelação *(Select Modeling Technique)*

O modelo de *data mining* escolhido para a realização deste trabalho foi o *Clustering*, baseado no algoritmo *Microsoft Clustering*, existente no *Microsoft SQL Server 2005: Analysis Services*.

Criação e Interpretação do Modelo *(Build and Assess Model)*

Nesta fase serão descritos todos os passos na construção de um modelo de *Clustering*, utilizando o algoritmo *Microsoft Clustering*. Em primeiro lugar foram criadas as origens de dados, a tabela *Recolhas*, bem como as vista sobre a tabela em causa.

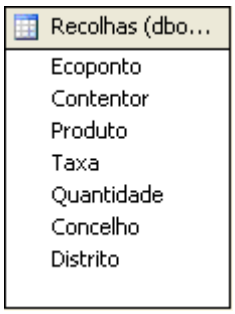

Figura 25 – Vista da origem de dados

De seguida foi criada uma estrutura de *mining* com os seguintes parâmetros:

- Selecção do *Microsoft Clustering* como técnica de *data mining*.
- Selecção da tabela *Recolhas* como "*Case*".
- Selecção do atributo Contentor como *Key*, e os atributos *Concelho*, *Distrito*, *Quantidade*, *Taxa* e *Produto* como *Input*.
- Selecção do nome para o modelo, *Recolhas* e selecção da opção *Allow drill through* para permitir navegar pelos membros dos *clusters*.
- Os parâmetros do algoritmo foram mantidos com os valores por defeito, com a excepção do *Cluster Count* (número de *clusters* a gerar) e do *Clustering Method* (foi usado o valor

referente ao método *K-Mean* [Tang e MacLennan, 2005], que obriga o modelo a não colocar o mesmo evento em diferentes *clusters*), atribuindo o valor 16 e 3, respectivamente.

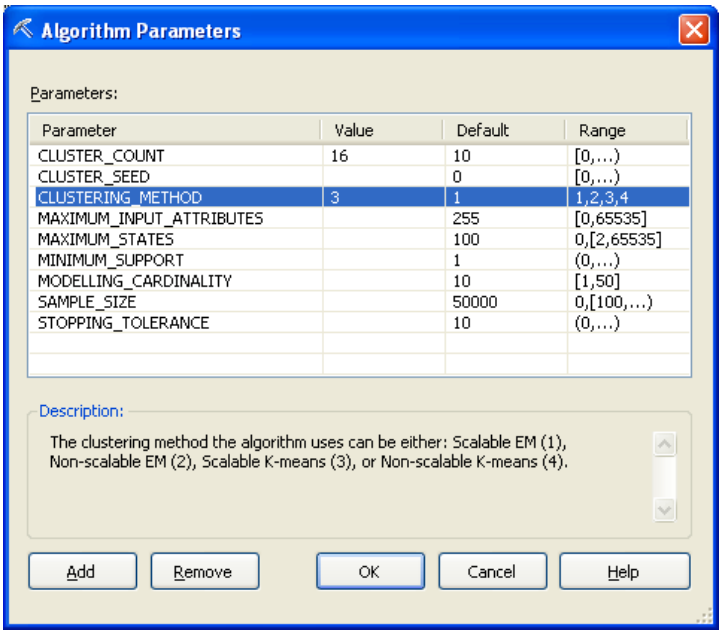

Figura 26 – Parâmetros escolhidos para o algoritmo

O modelo foi executado tendo dado origem ao número de *clusters* parametrizado, ou seja 16. Foi escolhido o valor 16 porque os contentores da Resulima estavam previamente divididos por 16 rotas.

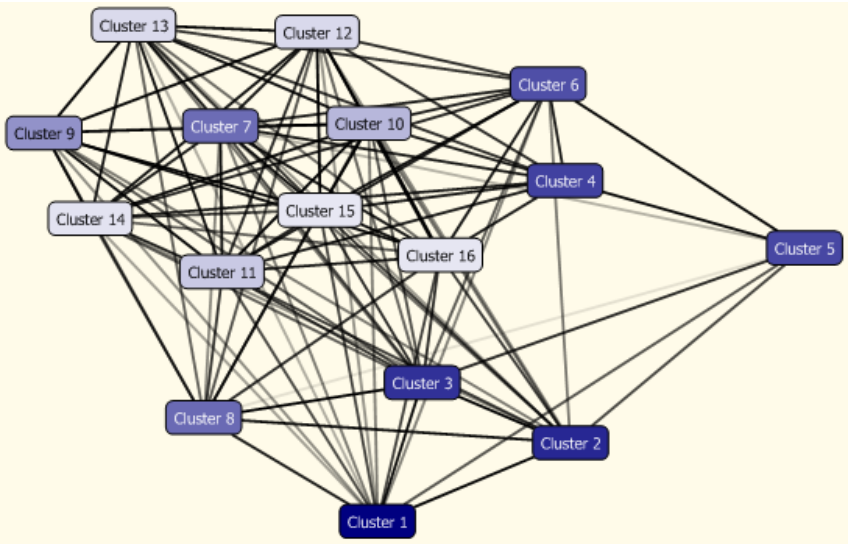

Figura 27 – *Clusters* gerados pelo modelo

Após a execução deste modelo, foi descoberto que os comportamentos dos contentores variam em função dos produtos. A Resulima actualmente tem as suas todas as suas rotas compostas por todos os contentores de todos os produtos, ou seja, a Rota 1 de papel tem os mesmos ecopontos que a Rota 1 de vidro e de que a Rota 1 de embalagem. A primeira execução do modelo, teve em conta todos os contentores de todos, pelo que foi necessário alterar criar três modelos, um referente a cada um dos três produtos.

Foram criadas três vistas na base de dados sobre a origem de dados Recolhas, cada uma filtrando apenas os contentores de um determinado produto. De seguida foram criados três novos modelos de *mining*, um para os contentores de papel, outro para os contentores de embalagem e outro para os contentores de vidro. Nestes modelos, foram usadas os mesmos parâmetros usados no primeiro modelo. Nas Figuras 28, 29 e 30, podem ser vistos os *clusters* gerados por cada um dos modelos criados.

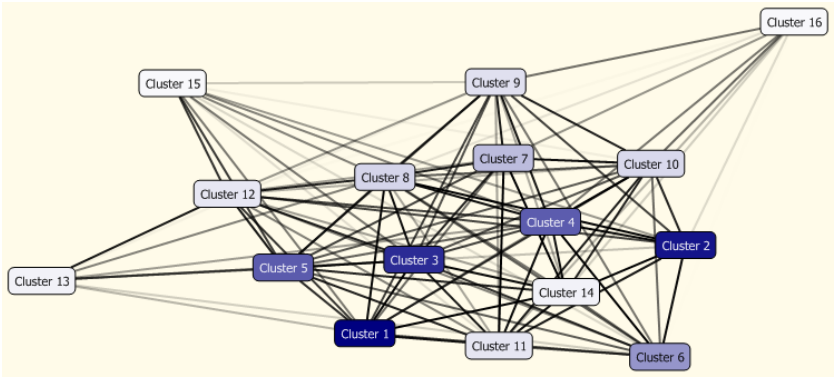

Figura 28 – *Clusters* de papel

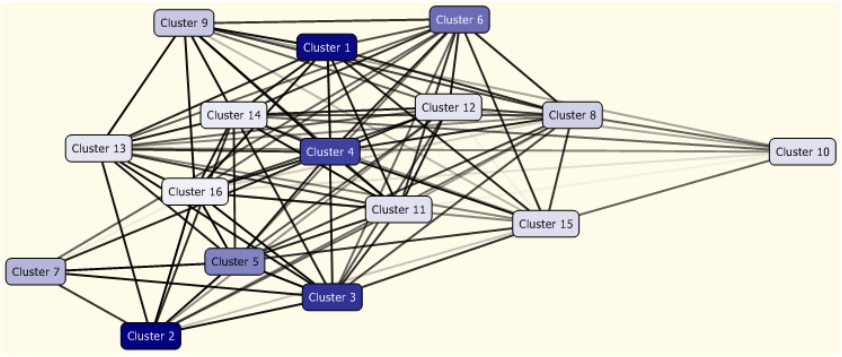

Figura 29 – Cluster de embalagens

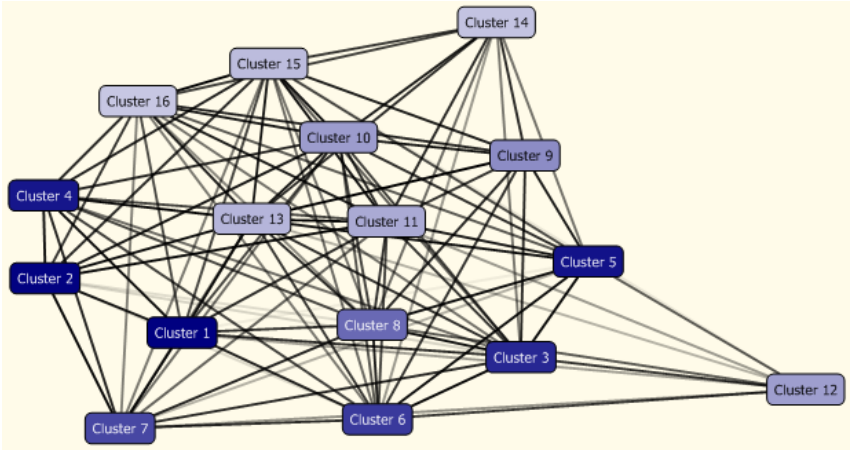

Figura 30 – *Cluster* de vidro

4.3.4 Avaliação

Pela análise efectuada aos resultados obtidos no primeiro modelo desenvolvido, permitiu concluir que contentores com produtos distintos não poderiam fazer parte do mesmo *cluster*. Esta conclusão comprova que os contentores têm comportamentos diferentes quer sejam papel, embalagem ou vidro. A Resulima, actualmente, tem organizados todos os seus ecopontos e contentores em rotas multi-produto. Os resultados da execução do primeiro modelo revelaram a necessidade de efectuar uma filtragem por produto, dando origem a três novos modelos de mineração.

Com base na análise dos resultados obtidos pelo processamento dos três modelos de *data mining* desenvolvidos (um por cada um dos produtos de recolha selectiva), foi possível constatar que os contentores foram agrupados tendo em conta as semelhanças nos seus atributos. O atributo mais forte na criação dos *clusters* foi o valor da quantidade, ou seja, os contentores que dão origem a maiores quantidades têm de ser agrupados nas mesmas rotas, e os que dão origem a menores quantidades têm de ser agrupados juntos.

Characteristics for Cluster 1

Variables	Values	Probability
Produto	03	
Quantidade	4190,1 - 6096,7	
Distrito	Braga	
Distrito	Viana do Castelo	
Taxa	40,3 - 46,9	
Concelho	Barcelos	
Taxa	46,9 - 53,5	
Concelho	Viana do Castelo	
Concelho	Esposende	
Taxa	17,6 - 40,3	
Taxa	53,5 - 76,3	
Concelho	Ponte de Lima	
Concelho	Arcos de Valdevez	
Quantidade	2283,6 - 4190,1	
Concelho	Ponte da Barca	

Figura 31 – Características do *Cluster* 1 de vidro

Tomemos como exemplo o *Cluster 1* de vidro. Neste *cluster* foram agrupados contentores que geram anualmente quantidades superiores a 4190,1 Kg e menores do que 6096,7 Kg (ver Figura 31). Este *cluster* é composto por 112 contentores o que faz com que na prática, devido à duração de um turno de trabalho e às capacidades dos camiões de recolham, não possa ser criada apenas uma rota a partir deste *cluster*. Em média, uma rota da Resulima tem aproximadamente 65 contentores. Este *cluster* poderá então dar origem a duas rotas e na hora da escolha dos contentores que poderão ou não fazer parte de uma rota, poderá ser tido em consideração outros atributos, por exemplo atributos baseados na sua localização geográfica, como o caso do Concelho e do Distrito.

4.4 Apreciação geral

Como apreciação final deste capítulo é possível afirmar que a aplicação da mineração de dados na área da recolha selectiva é viável e poderá tornar-se uma realidade no futuro. Os modelos foram desenvolvidos com base na metodologia *CRISP-DM*, que se revelou uma boa aposta na estruturação e orientação dos trabalhos e na obtenção de resultados.

Surgiram algumas dificuldades na aplicação dos modelos directamente sobre a base de dados do SPAR pelo que foi necessário criar e alimentar duas novas tabelas que serviram de vistas para a análise efectuada. Uma vez criadas estas vistas, foi possível a aplicação de duas técnicas de *data mining*, a Associação e o *Clustering*, aos dados de um ano de recolha selectiva da Resulima. Da aplicação destes modelos resultaram algumas regras importantes tendo em conta os atributos existentes e foi possível sugerir várias rotas que poderão ser uma mais-valia para a Resulima caso se venham a revelar adequadas à sua realidade.

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Conclusões

A recolha selectiva tem, nestes últimos anos, assumido um papel importante na gestão de resíduos, impedindo que uma grande parte dos resíduos produzidos pelas famílias e pelas empresas deixe de ser enviada para os aterros municipais. Em lugar disso, tanto os operadores de resíduos, que são os responsáveis pela recolha dos resíduos, como as empresas produtoras de embalagens, que fazem a retoma dos resíduos triados para valorização, encontraram nos resíduos uma oportunidade de negócio. Se juntarmos às vantagens económicas de cada uma destas empresas as vantagens ecológicas pela reutilização dos resíduos na criação de novos produtos estaremos perante uma solução adequada para a gestão dos resíduos e para a preservação do planeta.

Para dar resposta às necessidades de, por um lado efectuar um serviço público eficaz, recolhendo os resíduos com a frequência adequada, por outro efectuarem essas recolhas da forma mais optimizada possível como forma de contenção e redução de custos com viaturas, pessoal, entre outros, estas empresas foram obrigadas a inovar e utilizarem ferramentas informáticas para a gestão da sua actividade.

Com a adopção destas ferramentas informáticas para gestão dos processos de recolha selectiva, como o caso do SPAR analisado nesta dissertação, os operadores de resíduos foram capazes de, por um lado resolver os seus problemas operacionais, e por outro, foram capazes de gerar elevados volumes de dados relacionados com os seus processos. À medida que estes dados aumentam, aumenta a dificuldade de extracção de conhecimento e da descoberta de novos padrões e relacionamentos na informação.

Com esta dissertação foi possível demonstrar que a utilização do *data mining* sobre a informação da recolha selectiva poderá ser a resposta na descoberta de informação que auxilie estas empresas na obtenção de informação para a tomada de decisão sobre os seus negócios. Após a aplicação de dois mecanismos de mineração, a Associação e o *Clustering* nos dados da Resulima, foi possível encontrar relacionamentos entre os atributos dos contentores, como o caso da sua localização geográfica, produto, enchimentos e quantidades recolhidas que possibilitaram a sugestão da separação das rotas em função dos produtos dos contentores.

A aplicação da mineração de dados à recolha selectiva foi feita com apoio à metodologia *CRISP-DM*, tendo sido executadas as fases de Conhecimento do Negócio, Conhecimento dos Dados, Preparação dos Dados, Modelação e Avaliação dos modelos desenvolvidos.

A validação destes resultados como optimização ou não dos processos da Resulima, não fez parte do âmbito desta dissertação, já que apenas foi verificada a aplicabilidade da mineração de dados nesta área de negócio.

5.2 Trabalho Futuro

Após análise na bibliografia disponível não foi possível encontrar outros estudos que fossem baseados na aplicação do *data mining* à recolha selectiva. Esta constatação permite concluir que existe bastante trabalho para desenvolver no futuro na sua aplicação e adequação aos processos de recolha selectiva. Esta dissertação permite identificar algumas linhas orientadoras para esses trabalhos:

1. Aplicar as técnicas descritas nesta dissertação ou outras distintas a uma base de dados com dois ou mais anos de informação de recolha selectiva para permitir encontrar novos padrões, possivelmente sazonais.
2. Enriquecer a informação existente com novos atributos para encontrar novas associações e relacionamentos que influenciem as quantidades recolhidas, como sejam dados meteorológicos, informações sobre festas e feriados, dia da semana, eventos desportivos e informação demográfica.
3. Aplicar o *data mining* a base de dados de outros operadores de resíduos para verificar a sua adequação ao universo das empresas de recolha selectiva.
4. Aplicar o *data mining* a bases de dados provenientes de outras ferramentas de apoio à recolha selectiva.

Bibliografia

[Alves e Carvalho, 2004] Cláudio Manuel Martins Alves and José Manuel Valério de Carvalho, "Planeamento de Rotas num Sistema de Recolha de Desperdícios de Madeira", 2004

[Balas, 1989] Balas, E., The Prize Collecting Traveling Salesman Problem, Networks 19, 1989

[Berry e Linoff, 2004] Michael J. A. Berry and Gordon S. Linoff, "Data Mining Techniques", Second Edition, Wiley, 2004

[Bigus, 1996] Joseph P. Bigus, "Data Mining with Neural Networks, Solving Business Problems from Application Development to Decision Support", McGraw-Hill, 1996

[Boone e Kurtz, 1998] Boone and Kurtz, "Contemporary Marketing Wired", Dryden Press, 1998

[Giudici, 2003] Paolo Giudici, "Applied Data Mining - Statistical Methods For Business And Industry", John Wiley & Sons, 2003

[Kimball, 1992] Debi Kimball, "Recycling In America", ABC-CLIO Inc., 1992

[Olson e Delen, 2008] Olson D.L and Delen D., "Advanced Data Mining Techniques", Springer, 2008

[Simonetto e Borenstein, 2005] Eugênio de Oliveira Simonetto and Denis Borenstein, "Gestão Operacional da Coleta Seletiva de Resíduos Sólidos Urbanos – Abordagem utilizando um Sistema de Apoio à Decisão", 2005

[Sumathi e Sivanandam, 2006] S. Sumathi, S. Sivanandam, "Introduction to Data Mining and its Applications", Springer, 2006

[Tang e MacLennan, 2005] ZhaoHui Tang and Jamie MacLennan, "Data Mining with SLQ Server 2005", Wiley, 2005

Referências WWW

[Web NetRes] "Recolha Selectiva e Reciclagem", http://www.netresiduos.com/cir/rsurb/recrecicl.htm, Acedido em 07/10/2008

[Web Operadores] "Operadores de Resíduos Industriais Licenciados", http://www.netresiduos.com/operadores.php, Acedido em 07/10/2008

[Web IBM] "Online data continues to grow at an explosive pace - Knowledge Discovery & Data Mining", http://domino.watson.ibm.com/comm/research.nsf/pages/r.kdd.spotlight.html, Acedido em 07/10/2008

[Web SourceWatch] "Data mining Source Watch", http://www.sourcewatch.org/index.php?title=Data_mining#Another_Definition, Acedido em 07/10/2008

[Web CRISPDM] "CRoss-Industry Standard Process for Data Mining", http://www.crisp-dm.org/Process/index.htm, Acedido em 07/10/2008

[Web Recycling] "Who invented recycling?", http://www.professorshouse.com/your-home/environmentally-friendly/environmentally-friendly-article.aspx?id=3790, Acedido em 07/10/2008

[Web Quercus] "Netxplica – Ciências Naturais, Biologia e Geologia", http://forum.netxplica.com/viewtopic.php?t=2470&sid=8c57138d2d9d199b58bd226da25e6f3a, Acedido em 07/10/2008

[Web Garbage] "Smart Strategies & New Tech for Putting a Lid on Garbage", http://www.popularmechanics.com/home_journal/how_to/4234061.html, Acedido em 07/10/2008

[Web EcoPonto] "Ecoponto", http://pt.wikipedia.org/wiki/Ecoponto, Acedido em 07/10/2008

[Web AMDE] "Centro de Triagem", http://www.amde.pt/pagegen.asp?SYS_PAGE_ID=451797, Acedido em 07/10/2008

[Web Valorização] "Sistema de valorização de resíduos", http://pt.wikipedia.org/wiki/Sistema_de_valoriza%C3%A7%C3%A3o_de_residuos, Acedido em 07/10/2008

[Web AMARSUL] "Central de Triagem Amarsul, "http://www.amarsul.pt/listagem.aspx?sid=90b93ab4-e8a0-4032-8786-76567b53ca4e&cntx=EZXOFWZpyeNggHax73GMfmt%2FC3hkbP7JqCHHvoW3gaD6%2BgcRHd129U17IVOKN92G, Acedido em 07/10/2008

[Web SPAR] "Sistema de Planeamento e Análise da Recolha", http://www.cachapuz.com/cachapuzsolutions/PortalRender.aspx?PageID={cd9a065b-665b-4029-8498-e982bb34793b}, Acedido em 14/10/2008

[Web ADP] "Site do Grupo Águas de Portugal", http://www.adp.pt, Acedido em 14/10/2008

[Web MOTI] "Marketing on the Internet, What is Marketing?", http://iws.ohiolink.edu/moti/homedefinition.html, Acedido em 21/11/2008

[Web IDS] Theodoros Lappas and Konstantinos Pelechrinis, “Data Mining Techniques for (Network) Intrusion Detection System”, Department of Computer Science and Engineering UC Riverside, http://www.cs.ucr.edu/~kpele/dataIDS.pdf, Acedido em 21/11/2008

[Web Attacks] “A Short History of Computer Viruses and Attacks”, http://www.securityfocus.com/news/2445, Acedido em 21/11/2008

[Web Cancer] “Cancer Facts & Figures - 2007”, http://www.cancer.org/downloads/STT/CAFF2007PWSecured.pdf, Acedido em 22/11/2008

[Web Siam] Abdelghani Bellaachia and Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques”, Department of Computer Science, The George Washington University, http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf, Acedido em 22/11/2008