O PADRÃO na descoberta de conhecimento em bases de dados demográficas

Maribel Santos
Universidade do Minho, Departamento de Sistemas de Informação e Centro de Investigação Algoritmi, Guimarães, Portugal
maribel@dsi.uminho.pt

Luís Amaral
Universidade do Minho, Departamento de Sistemas de Informação e Centro de Investigação Algoritmi, Guimarães, Portugal
amaral@dsi.uminho.pt

Resumo

A investigação na área da descoberta de conhecimento em bases de dados tem beneficiado inúmeros domínios de aplicação, caracterizados por possuírem bases de dados de grandes dimensões. Um destes domínios é a *Demografia Histórica*, que procura recolher, organizar e analisar os registos demográficos armazenados em diversas paróquias.

Os dados recolhidos são analisados pelos historiadores demógrafos, os quais verificam evoluções de natalidade, mortalidade, fecundidade, nupcialidade e movimentos da população, o que inclui também estudos de emigrações.

Este artigo aborda a problemática da exploração de bases de dados demográficas, utilizando os princípios subjacentes a descoberta de conhecimento em bases de dados. A abordagem proposta possui a particularidade de integrar na análise, a componente *espacial* associada aos dados demográficos explorados. Esta componente está implícita nos dados demográficos analisados, uma vez que os mesmos se encontram geograficamente referenciados através de identificadores qualitativos, tais como moradas.

A incorporação de mecanismos de *raciocínio espacial qualitativo*, no processo de descoberta de conhecimento, permitiu a identificação de padrões e outros relacionamentos implícitos, existentes entre os dados demográficos e os dados geo-espaciais analisados.

Palavras chave: descoberta de conhecimento, *data mining*, raciocínio espacial qualitativo.

1 Introdução

O acentuado desenvolvimento das capacidades informáticas ao nível do armazenamento de dados, tem sido acompanhado pela necessidade de manipular grandes quantidades de informação. A necessidade de recolher e armazenar dados de diversos tipos e proveniências superou a nossa capacidade de analisar, sintetizar e extrair conhecimento a partir desses dados. Enquanto que as bases de dados fornecem as ferramentas necessárias ao armazenamento e visualização de grandes quantidades de dados, a exploração dos mesmos requer a utilização de ferramentas apropriadas, que automatizem o processo de análise dos dados e descoberta de conhecimento [Fayyad e Uthurusamy 1996].

Os algoritmos utilizados para procurar padrões nos dados são denominados de *Data Mining*. O processo global de descoberta de conhecimento em bases de dados, que se desenrola em várias fases, inclui a gestão dos algoritmos de *Data Mining* e a interpretação dos padrões encontrados pelos mesmos, os quais são posteriormente utilizados no suporte à tomada de decisão [Fayyad, et al. 1996].

O PADRÃO [Santos 2000] é um *sistema de descoberta de conhecimento em bases de dados geo-referenciadas*, que permite a identificação de padrões ou outros relacionamentos implícitos, existentes entre dados geográficos e dados não geográficos. A sua arquitectura é baseada num sistema de raciocínio espacial qualitativo, que permite a incorporação da *componente espacial* dos dados no processo de descoberta de conhecimento. Esta componente está implícita nos identificadores geográficos utilizados na geo-referenciação da informação.

Neste artigo é descrita a análise de uma base de dados demográfica, salientando os padrões e relacionamentos implícitos existentes entre os dados demográficos e os dados geográficos analisados.

Este artigo é organizado da seguinte forma. A secção 2 destaca a importância da integração da componente espacial na exploração dos dados demográficos. A secção 3 apresenta a arquitectura do PADRÃO, destacando os seus principais componentes e ainda, o modo de funcionamento do mesmo. Na secção 4 é apresentado um exemplo de utilização do PADRÃO no domínio demográfico. A secção 5 culmina com a apresentação de algumas conclusões e propostas de trabalho futuro.

2 A componente geo-espacial associada aos dados demográficos

Os mapas constituem a forma mais natural de representação de dados geográficos. Agregam um conjunto de *pontos*, *linhas* e *polígonos*, cuja posição no espaço é definida recorrendo a determinado sistema de coordenadas. A legenda de um mapa permite a integração de dados não espaciais, como nomes de locais, símbolos, cores, etc., com dados espaciais, nomeadamente com a localização dos elementos geográficos representados no mapa [Aronoff 1989].

Os **dados geográficos** são caracterizados por possuírem pelo menos um **aspecto espacial**, que permite a definição geométrica e topológica dos mesmos [CEN/TC-287 1998b]. A componente geométrica descreve um objecto através de coordenadas ou funções matemáticas, caracterização quantitativa que permite estabelecer a *dimensão*, *posição*, *forma*, *tamanho* e *orientação* dos mesmos. A componente topológica permite descrever a conectividade existente entre diversas entidades geográficas, representando características que permanecem invariantes a transformações do espaço, como sejam mudanças de escala ou de sistema de coordenadas.

O termo geográfico está assim associado à localização, contextualizando determinado objecto ou acontecimento no espaço, enquanto que o termo espacial é utilizado para definir as características dessa localização, tais como a geometria e a topologia. Informação geográfica tem associada informação espacial, pelo que ao longo deste artigo, o termo *geo-espacial* é utilizado para explicitar essa associação.

Dados demográficos agrupam conjuntos de registos de acontecimentos, que caracterizam as várias fases da vida de um indivíduo. As fontes documentais básicas utilizadas na demografia histórica são os registos de nascimentos, casamentos e óbitos [Amorim 1992].

A quantidade de informação manipulada neste estudo, pelos historiadores demógrafos, envolve vários séculos (séc. XVII-XX), nos quais os dados são analisados por forma a verificar evoluções de natalidade, mortalidade, fecundidade, nupcialidade e movimentos da população, o que inclui também estudos de emigrações.

A Tabela 1 apresenta um excerto da tabela `indivíduos` existente na base de dados demográfica. Esta base de dados armazena os registos paroquiais datados entre 1690 e 1990 no

distrito de *Aveiro*. Pela análise da referida tabela é possível verificar que os dados se encontram geo-referenciados, factor que permite a integração da componente espacial no processo de descoberta de conhecimento.

Num	Name	S	Birth date	Birth place	Died	Died place	Occupation:	M	Ch
6224	JOAO ANTOI	M	18-03-1790	Arada	01-10-1847	Arada	Oleiro	1	12
6232	TERESA LOI	F	13-05-1790	Coimbrão	08-06-1830	Quinta do Picai	Oleira	1	8
6233	ANTONIO DA	M	24-05-1790	Quinta do Picai	16-09-1864	Quinta do Picai	Oleiro	2	10
6235	JOSE FRANC	M	28-05-1790	Quinta do Picai	05-10-1849	Verdemilho	Lavrador	1	10
6239	MANUEL FR	M	03-08-1790	Quinta do Picai	01-08-1830	Quinta do Picai	Lavrador	1	7
6241	ROSA DOS S	F	25-08-1790	Bom Sucesso	27-08-1830	Quinta do Picai	Lavradora	1	7
6249	MANUEL JOA	M	25-09-1790	Verdemilho	20-03-1841	Verdemilho	Lavrador	1	10
6250	ANTONIO SIN	M	21-09-1790	Arada	07-03-1874	Arada	Lavrador	1	9
6253	JOANA MARI	F	31-10-1790	Verdemilho	06-03-1863	Verdemilho	Lavradora	1	10
6257	FRANCISCO	M	10-11-1790	Bom Sucesso	28-05-1831	Bom Sucesso	Lavrador	1	4
6259	JOAQUINA M	F	17-12-1790	Verdemilho	24-03-1864	Verdemilho	Lavradora	1	9
6260	BERNARDO	M		Quinta da Gran	21-08-1843	Verdemilho	Lavrador	1	8
6261	JOAQUINA F	F	26-11-1767	Verdemilho	31-12-1823	Verdemilho	Lavradora	1	8
6267	PERPETUA M	F	06-03-1791	Verdemilho	10-12-1855	Verdemilho	Lavradora	1	10
6288	MARIA DE JE	F	06-06-1791	Arada	24-03-1877	Arada	Jornaleira	0	0
6299	JOSEFA DE	F		Quinta do Picai	30-03-1836	Quinta do Picai	Lavradora	1	9
6314	JOANA TERE	F	05-11-1791	Arada	17-04-1870	Arada	Mendicante	1	4
6331	MAURICIO FI	M	09-06-1792	Quinta do Picai	27-02-1858	Quinta do Picai	Lavrador	1	7
6335	MARIA ROSA	F	07-09-1792	Quinta do Picai	20-05-1870	Quinta do Picai	Lavradora	1	9
6337	MIGUEL FER	M	25-09-1792	Arada	24-12-1876	Arada	Lavrador	1	11
6338	MANUEL DA	M	27-09-1792	Bom Sucesso	28-09-1855	Bom Sucesso	Jornaleiro	1	6
6340	ANTONIA DO	F	28-10-1792	Bom Sucesso	25-05-1866	Arada	Lavradora	1	11

Tabela 1: Excerto da tabela `indivíduos`

3 A arquitectura e modo de funcionamento do PADRÃO

A arquitectura do PADRÃO agrega três componentes principais: i) Repositório de Dados e Conhecimento; ii) Análise de Dados e iii) Visualização de Resultados. A Figura 1 apresenta uma visão geral da arquitectura do sistema, a qual é de seguida brevemente descrita.

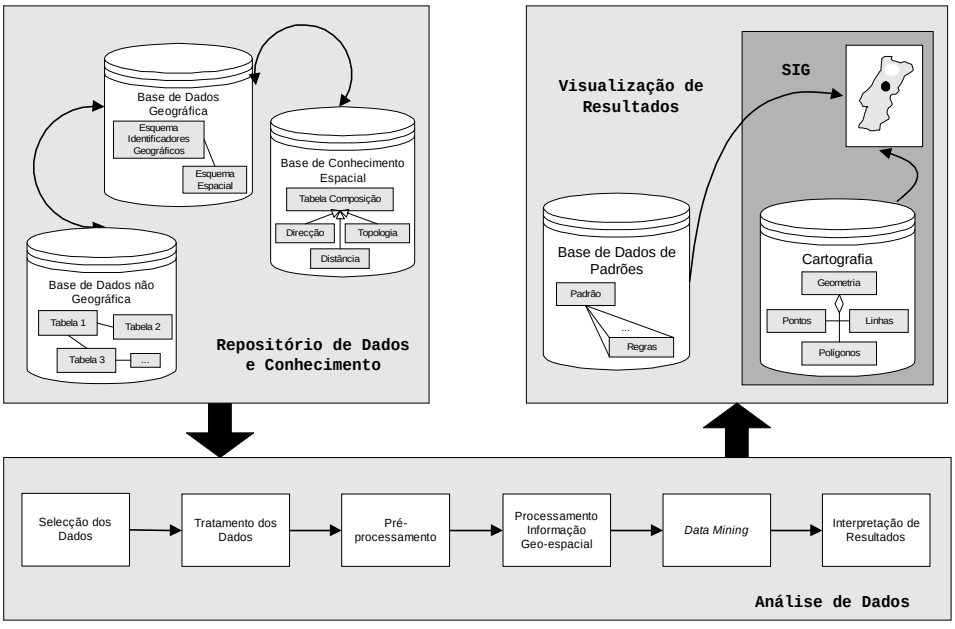

Figura 1: Arquitectura do sistema PADRÃO

O **Repositório de Dados e Conhecimento** integra três bases de dados:

1. Uma **Base de Dados Geográfica** (BDG), construída segundo os princípios estabelecidos pelo Comité Europeu de Normalização nas pré-normas[1] CEN TC 287 para Informação Geográfica. Seguindo as suas directivas, foi possível implementar uma BDG na qual o posicionamento espacial dos dados é conseguido recorrendo a um ***sistema de identificadores geográficos*** [CEN/TC-287 1998a]. Neste sistema, são caracterizadas subdivisões administrativas de Portugal Continental, ao nível dos Concelhos e Distritos. Entre outros atributos, esta base de dados armazena as relações espaciais do tipo direcção, distância e topologia [CEN/TC-287 1996] existentes entre subdivisões administrativas adjacentes, pertencentes ao nível hierárquico dos Concelhos [Santos e Amaral 1999].
2. Uma **Base de Conhecimento Espacial**[2] (BCE), que armazena os mecanismos de raciocínio qualitativo que permitem a inferência de relações espaciais desconhecidas. Dentre o conhecimento disponível nesta base, encontram-se as tabelas de composição que integram a direcção, distância e topologia no processo de raciocínio, os identificadores qualitativos utilizados, e ainda o intervalo de validade quantitativo associado a cada um dos mesmos.
3. Uma **Base de Dados não Geográfica**, que no caso particular da aplicação descrita neste artigo é representada por uma Base de Dados Demográfica (BDD). Esta base de dados é integrada no módulo de descoberta de conhecimento com as restantes bases de dados descritas, permitindo a identificação de padrões e outros relacionamentos implícitos, existentes entre os dados geo-espaciais e os dados demográficos analisados.

O componente de **Análise de Dados** é implementado no `Clementine` [ISL 1998], representando o módulo de descoberta de conhecimento. Este módulo é caracterizado por passar por 6 grandes etapas:

1. **Selecção dos dados**. Neste primeiro passo são seleccionados os dados demográficos e os dados geo-espaciais considerados relevantes para a tarefa de *Data Mining* a efectuar.
2. **Tratamento dos dados**. Esta fase preocupa-se com a limpeza dos dados, nomeadamente com o tratamento de dados omissos ou corrompidos.
3. **Pré-processamento dos dados**. Esta etapa visa essencialmente a redução do tamanho da amostra a analisar, através da generalização dos dados. Esta generalização é efectuada atendendo às hierarquias conceptuais definidas para o domínio de aplicação em causa.
4. **Processamento de informação geo-espacial**. Nesta fase verifica-se a informação geo-espacial disponível e necessária na etapa seguinte. Uma vez que a BDG apenas armazena as relações espaciais existentes entre entidades geográficas adjacentes, todas as restantes, e sempre que necessário, são inferidas recorrendo às regras e conhecimento espacial armazenado na BCE.
5. ***Data Mining***. Nesta etapa, algoritmos de *Data Mining* são utilizados para identificar padrões e relacionamentos implícitos existentes entre os dados demográficos e os dados geo-espaciais analisados.
6. **Interpretação de resultados**. Permite avaliar a utilidade/importância dos padrões encontrados. Sempre que esta avaliação catalogar uma descoberta como relevante, as respectivas regras podem ser armazenadas na Base de Dados de Padrões (BDP), permitindo a sua visualização num mapa.

O componente de **Visualização de Resultados** é responsável pela gestão da BDP e pela visualização do seu conteúdo em mapas das regiões analisadas. O PADRÃO recorre actualmente

[1] Uma vez que os documentos produzidos pelo referido comité se encontram em fase de votação final pelos seus respectivos membros, os mesmos são apelidados de pré-normas.

[2] Base de Conhecimento é utilizada neste contexto para referir uma base de dados que contém regras de inferência e informação referentes à experiência e perícia humana num dado domínio de aplicação [CT113 1999].

ao sistema de informação geográfica Geomedia Professional v3 [Intergraph 1999] como aplicação de suporte à visualização.

A arquitectura descrita, implementada através dos módulos descoberta de conhecimento e visualização de resultados, e suportada por três bases de dados centrais, pode ser caracterizada, em termos de funcionamento e interacção com o exterior, através do **diagrama caso de uso**[3] apresentado na Figura 2.

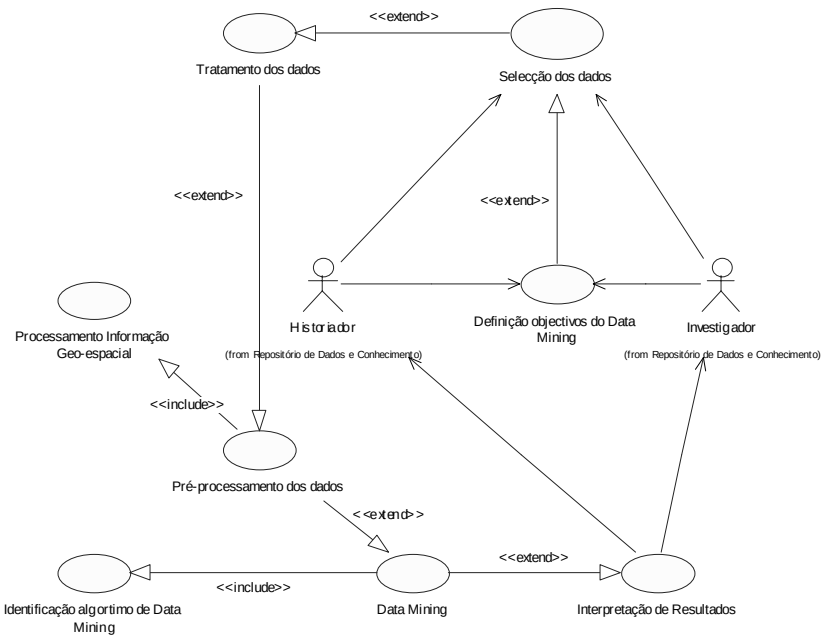

Figura 2: Funcionamento do PADRÃO no processo de descoberta de conhecimento

No diagrama caso de uso apresentado é possível identificar os actores que interagem com o sistema, no processo de descoberta de conhecimento. O *investigador*[4] permite a interacção do *historiador* com o PADRÃO, transformando os objectivos da análise (definidos pelo *historiador*) em objectivos de *Data Mining*.

O processo de descoberta de conhecimento é condicionado pela definição dos objectivos que o mesmo visa servir. A partir desta definição é possível proceder a selecção dos dados relevantes à análise, os quais são posteriormente tratados, pré-processados e integrados com a informação geo-espacial necessária à satisfação dos objectivos da tarefa. A identificação do algoritmo de *Data Mining* mais apropriado à tarefa em causa, permite ao *investigador* proceder com a análise e consequente identificação de padrões. Os mesmos são posteriormente disponibilizados ao *historiador* e ao próprio *investigador*, permitindo a este último avaliar o desempenho das decisões tomadas.

---

[3] Os **diagramas caso de uso** são um dos tipos de diagramas disponíveis em UML (*Unified Modeling Language*). Permitem agrupar um conjunto de entidades do tipo *caso de uso*, *actores* e seus respectivos *relacionamentos*. Estes diagramas são particularmente importantes na organização e modelação do funcionamento de um sistema, permitindo retratar a interacção existente entre este e os seus diversos *actores* (sejam eles utilizadores ou outros sistemas) [Booch, et al. 1999].

[4] A denominação *investigador* adoptada, para o actor responsável pela condução do processo de descoberta de conhecimento, decorre do facto de no caso da aplicação descrita neste artigo, este papel estar a ser desempenhado pelo investigador responsável pela implementação do sistema PADRÃO.

4 O PADRÃO na descoberta de conhecimento

A BDD a analisar armazena os registos paroquiais datados entre 1690 e 1990 no distrito de *Aveiro*. Estes registos são caracterizados através de atributos como `data de nascimento`, `local de nascimento`, `data de óbito`, `local de óbito`, `profissão`, `número de casamentos`, `número de filhos`, `idade ao casamento`, etc.

Tendo sido definido como objectivo do *Data Mining*, a caracterização da `idade ao casamento` dos indivíduos registados na base de dados, atendendo ao seu `sexo`, `número de casamentos`, e ao `século` e `município` no qual viveram, foi efectuada a selecção dos dados a analisar, e ainda o pré-processamento dos mesmos, atendendo às hierarquias conceptuais definidas para o domínio de aplicação em causa (as mesmas podem ser consultadas em Santos e Amaral [Santos e Amaral 2000a]).

A etapa de processamento da informação geo-espacial permitiu avaliar a informação geográfica disponível, e inferir relações espaciais desconhecidas. A inferência[5] é realizada consultando a tabela de composição que integra relações espaciais do tipo direcção, distância e topologia, segundo os princípios do raciocínio espacial qualitativo. Para concretizar a tarefa de *Data Mining* definida, foi necessário criar um modelo no `Clementine`, `dir_AVR`, que classifique a localização (em termos de direcção) de cada um dos municípios em relação ao distrito analisado. O modelo geográfico construído [Santos e Amaral 2000b] foi posteriormente integrado com os dados demográficos seleccionados, permitindo identificar um conjunto de regras que classificam os dados analisados. O algoritmo de *Data Mining* utilizado foi o `C5.0`, o qual permite construir árvores de decisão, nas quais é previsto o valor de um dado atributo de *saída* em função de diversos atributos de *entrada*.

A Figura 3 apresenta a *stream* construída no `Clementine` para a classificação do atributo `age_marr`, assim como o conjunto de regras obtido. Pela análise da figura é possível verificar que o modelo geográfico `dir_AVR` é integrado com os dados seleccionados da tabela `marriages` (acedidos via ligação ODBC[6]), que depois de devidamente balanceados são analisados pelo respectivo algoritmo de *Data Mining*. As regras encontradas são apresentadas à direita da figura, nas quais é possível verificar a classificação efectuada ao atributo `age_marr`. Pela análise das regras verifica-se que apenas no século XIX existe uma distribuição geográfica bem definida. Todos os indivíduos que viveram em concelhos localizados a `S` ou `SW` (*Sul* ou *Sudoeste*) de Aveiro apresentaram uma `idade ao casamento` representada pela classe `26-45`. Por outro lado, nos concelhos localizados a `N`, `NE` e `E` (*Norte*, *Nordeste* ou *Este*) do distrito, a idade ao casamento é representada pela classe `16-25`. Este padrão tem de ser analisado pelos historiadores demógrafos, mas uma possível justificação para este comportamento pode ser encontrada no facto das regiões localizadas a `S` e `SW` serem regiões maioritariamente piscatórias, nas quais os indivíduos do sexo masculino passam longas temporadas no mar, podendo este facto adiar a idade ao primeiro casamento.

O conjunto de regras obtido pode ser transferido, via ligação ODBC, para a BDP [Santos e Amaral 2000a]. Esta transferência permite a visualização das regras num mapa da região analisada, para além de permitir que as mesmas possam ser utilizadas em posteriores exercícios de *Data Mining*. Este procedimento permite a construção de meta-regras, que evidenciam a evolução das regras ao longo do tempo.

---

[5] O processo de inferência de relações espaciais é detalhadamente descrito em Santos e Amaral [Santos e Amaral 1999].

[6] *Open Database Connectivity*.

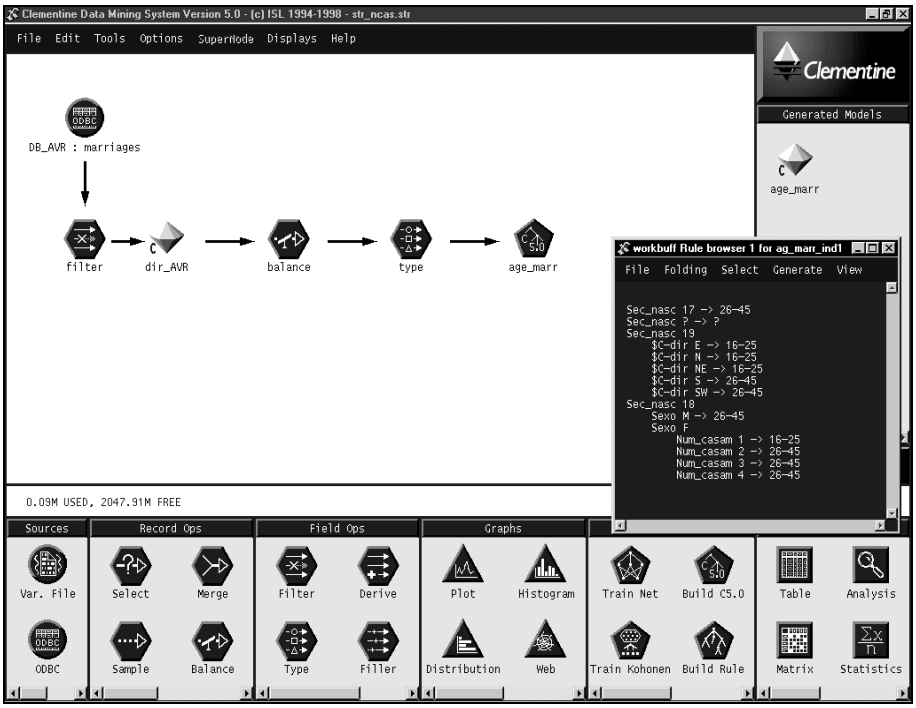

Figura 3: A descoberta de conhecimento no PADRÃO

5 Conclusões

Um caso particular da descoberta de conhecimento em bases de dados diz respeito à exploração de dados referenciados espacialmente, isto é, dados que representam objectos geográficos, localizações ou partes de uma divisão territorial. A análise destes dados impõe a verificação da componente espacial dos mesmos (posições relativas, adjacências, direcções, distâncias, etc.) e da influência que a mesma pode exercer nos restantes dados explorados.

Este artigo demonstrou a aplicabilidade do sistema PADRÃO na análise de dados demográficos. A abordagem proposta permitiu a inclusão da componente espacial, associada aos dados demográficos analisados, no processo de descoberta de conhecimento.

A integração da base de dados demográfica utilizada com a base de dados geográfica e a base de conhecimento espacial, permitiu ao PADRÃO a captura dos relacionamentos implícitos existentes entre os dados demográficos e os dados geo-espaciais analisados.

Em termos de trabalho futuro, e depois de conhecer o desempenho conseguido com a implementação do PADRÃO no `Clementine`, verificar-se-á se existem benefícios na sua implementação no `Kepler`, um sistema de *Data Mining* que permite *aprendizagem multi-relacional*.

6 Agradecimentos

Este trabalho tem sido parcialmente suportado por uma bolsa do PRODEP II (Acção 5.2, Concurso n.º 3/98 Doutoramentos).

Agradece-se também ao NEPS (Núcleo de Estudos da População e Sociedade) da Universidade do Minho, pela disponibilização dos dados necessários à realização deste estudo.

7 Referências

Amorim, N., *Evolução demográfica de três paróquias do sul do pico*, Universidade do Minho, Instituto de Ciências Sociais, 1992.

Aronoff, S., Geographic Information Systems: a management perspective, WDL Publications, Ottawa, 1989.

Booch, G., J. Rumbaugh, e I. Jacobson, *The Unified Modeling Language User Guide*, Addison Wesley Longman, Inc., 1999.

CEN/TC-287, Geographic Information: Data Description, Spatial Schema, Comité Europeu de Normalização, prENV 12160, 1996.

CEN/TC-287, Geographic Information: Referencing, Geographic Identifiers, Comité Europeu de Normalização, prENV 12661, 1998a.

CEN/TC-287, Geographic Information: Vocabulary, Comité Europeu de Normalização, CR 287003, 1998b.

CT113, Tecnologias da Informaçãon - Vocabulário. Parte 28: Inteligência Artificial - Conceitos básicos e sistemas periciais, Norma Portuguesa 3003, Instituto Português da Qualidade, prNP 3003-28, 1999.

Fayyad, U., e R. Uthurusamy, "Data Mining and Knowledge Discovery in Databases", Communications of the ACM, 39, 11 (1996), pp. 24-26.

Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, e R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, The MIT Press, Massachusetts, 1996.

Intergraph, Geomedia Professional v3, Reference Manual, Intergraph Corporation, 1999.

ISL, Clementine, User Guide, Version 5.0, Integral Solutions Limited, 1998.

Santos, M., PADRÃO: um sistema para a descoberta de conhecimento em bases de dados geo-referenciadas, Tese de Doutoramento (*em conclusão*), Universidade do Minho, 2000.

Santos, M., e L. Amaral, As Normas de Informação Geográfica e o Raciocínio Espacial Qualitativo na Inferência de Informação Geográfica Qualitativa, Proceedings of the V Geographic Information Systems Meeting, Lisboa, Portugal, 24-26 Novembro, 1999.

Santos, M., e L. Amaral, Knowledge Discovery in Spatial Databases through Qualitative Spatial Reasoning, PADD'00 Proceedings of the 4th International Conference and Exhibition on Practical Applications of Knowledge Discovery and Data Mining, Manchester, 11-13 Abril, 2000a, pp. 73-88.

Santos, M., e L. Amaral, A Qualitative Spatial Reasoning Approach in Knowledge Discovery in Spatial Databases, Proceedings of Data Mining 2000: Data Mining Methods and Databases for Engineering, Finance and Others Fields, Cambridge University, WIT Press, 5-7 Julho, 2000b, pp. 249-258.