Universidade Estadual de Campinas
Faculdade de Engenharia Elétrica e de Computação
Departamento de Comunicações

MAPAS AUTO-ORGANIZÁVEIS APLICADOS EM GOVERNO ELETRÔNICO.

Autor: Everton Luiz de Almeida Gago Junior

Orientador: Prof. Dr. Leonardo de Souza Mendes

Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica. Área de concentração: **Telecomunicações e Telemática.**

Banca Examinadora

Prof. Dr. Leonardo de Souza Mendes — DECOM/FEEC/UNICAMP
Prof. Dr. Rodolfo Miranda de Barros — DC/UEL
Prof. Dr. Bruno Bogaz Zarpelão — DECOM/FEEC/UNICAMP

Campinas – SP
Fevereiro/2012

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

G122m	Gago Junior, Everton Luiz de Almeida Mapas auto-organizáveis aplicados em governo eletrônico / Everton Luiz de Almeida Gago Junior. --Campinas, SP: [s.n.], 2012. Orientador: Leonardo de Souza Mendes. Dissertação de Mestrado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação. 1. Mineração de dados (Computação). 2. Internet na administração pública. I. Mendes, Leonardo de Souza. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Self-organizing maps applied to electronic government
Palavras-chave em Inglês: Data mining (Computing), Internet in public administration
Área de concentração: Telecomunicações e Telemática
Titulação: Mestre em Engenharia Elétrica
Banca examinadora: Bruno Bogaz Zarpelão, Rodolfo Miranda de Barros
Data da defesa: 16-02-2012
Programa de Pós Graduação: Engenharia Elétrica

COMISSÃO JULGADORA - TESE DE MESTRADO

Candidato: Everton Luiz de Almeida Gago Junior

Data da Defesa: 16 de fevereiro de 2012

Título da Tese: "Mapas auto-organizáveis aplicados em governo eletrônico"

Prof. Dr. Leonardo de Souza Mendes (Presidente): _____

Prof. Dr. Rodolfo Miranda de Barros: _____

Prof. Dr. Bruno Bogaz Zarpelão: _____

Resumo

Com as facilidades e resultados oferecidos pelos sistemas de gerenciamento automatizados, cada vez mais os municípios eliminam documentos físicos, e armazenam digitalmente suas informações. Uma das consequências diretas disso é a criação de um grande volume de dados. Em geral, estes dados são coletados por meio das Tecnologias da Informação e Comunicação (TIC) e armazenados em bases de dados transacionais. Nestes ambientes, os dados possuem relacionamentos complexos entre si, o que dificulta a identificação de padrões e comportamentos.

Diversas instituições utilizam técnicas de mineração de dados para identificar padrões e comportamentos ocultos em seus dados operacionais. Estes padrões podem auxiliar no planejamento de ações futuras e melhorar a gestão dos recursos financeiros, humanos e tecnológicos. A análise inteligente dos dados operacionais de uma instituição pode ser realizada através das Ferramentas de Apoio e Suporte a Tomada de Decisão (FASTD). Estas ferramentas permitem analisar um grande volume de dados a partir de regras previamente estabelecidas. Estas regras são apresentadas para as FASTD na fase de treinamento, quando a ferramenta aprende sobre os padrões que deve buscar.

Este trabalho apresenta um processo de suporte à tomada de decisão com base em mapas auto-organizáveis. Aplicado às ferramentas de governo eletrônico, este processo permite identificar padrões em um grande volume de dados de maneira autônoma, ou seja, dispensando o conjunto de regras para treinamento. Para realizar o estudo de caso, utilizaremos informações cedidas pela Prefeitura Municipal de Campinas, São Paulo, Brasil.

Palavras-chave: *Mineração de dados (Computação), Internet na administração pública.*

Abstract

Due to the characteristics offered by automated management systems, municipal administrations are now attempting to store digital information instead of keeping their physical documents. One consequence of such fact is the generation of large volume of data. Usually, these data are collected by ICT technologies and then stored in transactional databases. In this environment, collected data might have complex internal relationships. This may be an issue to identify patterns and behaviors.

Many institutions use data mining techniques for recognize hidden patterns and behaviors in their operational data. These patterns can assist to future activities planning and provide better management to financial resources. Intelligent analysis can be realized using the Support Tools and Support Decision Making (STSDM). These tools can analyze large volume of data through previously established rules. These rules are presented for STSDM in the training phase, and the tool learns about the patterns that should look.

This work proposes a process to support decision making based on self-organized maps. This process, applied to electronic government tools, can recognize patterns in large volume of data without the set of rules for training. To perform our case study, we use data provided by the city of Campinas, Sao Paulo.

Keywords: *Data mining (Computing), Internet in public administration*.

Aos meus pais,
Que me propiciaram uma vida digna onde eu pudesse crescer, acreditando que tudo é possível, desde que sejamos honestos e íntegros. Que sonhar e concretizar os sonhos só dependerá de nossa vontade.

A minha esposa,
Que sempre acreditou e apoiou os meus objetivos.

Agradecimentos

Ao Prof. Dr. Leonardo de Souza Mendes, pela orientação e constante estímulo transmitido durante o trabalho.

Aos amigos Gean e Bruno, que muito contribuíram com o meu aprendizado e a superação de todas as dificuldades.

A Prefeitura Municipal de Campinas, por ter cedido os dados utilizados em minha pesquisa.

Sumário

LISTA DE ABREVIAÇÕES ..... **XIII**

CAPÍTULO 1 ..... **15**

1.1 ORGANIZAÇÃO DO TRABALHO 18

CAPÍTULO 2 ..... **19**

2.1 GOVERNO ELETRÔNICO 19

2.1.1 Arquitetura de e-Gov 20

2.1.2 Classificações de e-gov 23

2.2 BUSINESS INTELLIGENCE 25

2.2.1 Extração, Transformação e Carga 26

2.2.2 Data Warehouse 27

2.2.3 Ferramenta Analítica de Processamento On-line 28

2.2.4 Mineração de Dados ou Análise Exploratória Automática 29

2.3 EXPLORAÇÃO NÃO SUPERVISIONADA OU CLUSTERIZAÇÃO 31

2.3.1 Redes Neurais Artificiais 32

2.3.2 Mapas Auto-organizáveis 34

2.3.3 Matriz de Distância Unificada 37

CAPÍTULO 3 ..... **41**

3.1 MODELO CONCEITUAL MULTIDIMENSIONAL 44

3.1.1 Camada ETL 45

3.1.2 Camada de Armazenamento e Disponibilização de Visões dos Dados 45

3.1.3 Camada de Aplicações para os Usuários Finais 47

3.2 MODELO GENÉRICO PARA REPRESENTAÇÃO DE AMOSTRAS E EXTRAÇÃO DE CONHECIMENTO 47

3.2.1 Preenchimento das Entidades de Representação de Amostras 52

3.2.2 Análise Exploratória Automática com Mapas Auto-organizáveis. 54

3.2.3 Rotina de Balanceamento e Carga 61

CAPÍTULO 4 64

4.1 ORIGEM DOS DADOS OPERACIONAIS 64

4.2 CARGA DOS DADOS NO MGRAEC 67

4.3 MINERAÇÃO DE DADOS PELO MAPA AUTO-ORGANIZÁVEL 69

4.4 ROTINA DE BALANCEAMENTO E CARGA 71

4.5 DISCUSSÃO TÉCNICA 73

CAPÍTULO 5 74

REFERÊNCIAS BIBLIOGRÁFICAS 76

Lista de Figuras

Figura 2.1 - Arquitetura das Aplicações de e-Gov [1]. 21
Figura 2.2 - Ambiente de BI. 26
Figura 2.3 - Neurônio Artificial [9]. 33
Figura 2.4 - Mapa Auto-Organizável [35]. 34
Figura 2.5 - Grau de adaptação aplicado aos neurônios vizinhos. 36
Figura 2.6 - Processo de adaptação do neurônio BMU e seus vizinhos [36]. 37
Figura 2.7 - Matriz de Distância Unificada (Matriz-U) [37] 37
Figura 2.8 - Visão dos dados no plano R_3 [36]. 38
Figura 2.9 - Visão dos dados no plano R_2 [36]. 39
Figura 2.10 - Redução de dimensionalidade do R_3 para o R_2 através da Matriz-U [36] 39
2.11 - Matriz-U [38]. 40
Figura 3.1 - Processo de Extração do Conhecimento (PEC). 41
Figura 3.2 – Preenchimento das Entidades do MGRAEC. 43
Figura 3.3 - Processo de Extração do Conhecimento 44
Figura 3.4 - Ambiente proposto por Marques [4]. 45
Figura 3.5 - Modelo Conceitual Multidimensional [4]. 46
Figura 3.6 - Modelo Genérico para Representação de Amostras e Extração de Conhecimento (MGRAEC). 48
Figura 3.7 - Exemplo de preenchimento das Entidades para Representação de Amostras (ERA). 50
Figura 3.8 – Extração dos dados do MCM e armazenamento no MGRAEC. 53
Figura 3.9 - Algoritmo para carga dos dados nas ERA. 54
Figura 3.10 - Algoritmo de mineração de dados, através de mapas auto-organizáveis. 56
Figura 3.11 - Diagrama de classes do mapa auto-organizável. 58
Figura 3.12 - Diagrama de sequência do algoritmo de mineração de dados. 60
Figura 3.13 - Exemplo de dados organizados em uma estrutura desbalanceada. 62
Figura 3.14 – Exemplo de dados organizados em uma estrutura do tipo árvore balanceada. 62
Figura 3.15 - Rotina de Balanceamento e Carga. 63
Figura 4.1 - Tecnologias utilizadas. 65

Figura 4.2 - Representação dos agrupamentos pela Matriz-U. 70
Figura 4.3 - Representação hierárquica do conhecimento. 72

Lista de Tabelas

Tabela 3.1 - Representação dos dados nas ERA. 49
Tabela 3.2 - Classificação Numérica. 51
Tabela 3.3 - Entradas para o Mapa Auto-Organizável. 52
Tabela 3.4 - Exemplo de redução para a dimensão Idade. 54
Tabela 4.1 - Redução de variáveis aplicada aos dados. 67
Tabela 4.2 - Características de convergência. 69
Tabela 4.3 - Épocas de Treinamento da Rede Neural Artificial. 70
Tabela 4.4 - Representação quantitativa dos agrupamentos. 72

Lista de Abreviações

BI	*Business Intelligence*
BMU	*Best Matching Unit*
CEF	Caixa Econômica Federal
DM	*Data Mart*
DW	*Data Warehouse*
e-gov	Governo Eletrônico
EJB	*Enterprese JavaBeans*
ERA	Entidades para Representação de Amostras
ETL	*Extract, Transform and Load*
FASTD	Ferramentas de Apoio e Suporte a Tomada de Decisão
IBGE	Instituto Brasileiro de Geografia e Estatística
IDF	Índice de Desenvolvimento Familiar
IPTU	Imposto Predial e Territorial Urbano
JDBC	*Java Database Connectivity*
Matriz-U	Matriz de Distância Unificada
MCM	Modelo Conceitual Multidimensional
MGRAEC	Modelo Genérico para Representação de Amostras e Extração de Conhecimento
MLP	*Multi-Layer Perceptron*
NIS	Número de Inscrição Social
OLAP	*On-line Analytical Processing*
PEC	Processo de Extração do Conhecimento
PNAD	Pesquisa Nacional por Amostra de Domicílios
RBF	*Radial Basis Function*
RMAA	Redes Metropolitanas de Acesso Aberto
RNA	Redes Neurais Artificiais
SF/BNDES	Secretaria para Assuntos Fiscais do Banco Nacional de Desenvolvimento Econômico e Social
SIGM	Sistema Integrado de Governança Municipal
SOM	*Self-Organizing Maps*
TIC	Tecnologias da Informação e Comunicação
XML	*Extensible Markup Language*

Trabalhos afins publicados pelo autor

1. GAGO JÚNIOR, Everton Luiz de Almeida; BREDA, Gean Davis; MARQUES, Eduardo Zanoni; MENDES, Leonardo de Souza. **SELF-ORGANIZING MAPS an Approach Applied to Electronic Government**. 8th International Conference on Web Information Systems and Technologies, Porto, Portugal, 2012.

2. MARQUES, Eduardo Zanoni; MIANI, Rodrigo Sanches; GAGO JÚNIOR Everton Luiz de Almeida; MENDES, Leonardo de Souza. **Development of a Business Intelligence Environment for e-Gov Using Open Source Technologies**. In: Data Warehousing and Knowledge Discovery, 2010, Bilbao. Lecture Notes in Computer Science. Berlim: Springer, 2010.

Capítulo 1

Introdução

AS facilidades e resultados oferecidos pelas Tecnologias da Informação e Comunicação (TIC) fazem com que as organizações públicas e privadas eliminem os documentos físicos, e passem a armazenar digitalmente suas informações [1]. O Governo Eletrônico (e-gov) emergiu como um termo popular na administração pública, para classificar o uso das TIC como ferramentas de gerenciamento de tarefas, atividades e acontecimentos [2]. As TIC vêm sendo utilizadas no setor público com o intuito de ajudar estas organizações a gerir seus recursos, viabilizando o monitoramento dos resultados da implantação de políticas públicas junto à sociedade [3].

Uma das consequências do uso das TIC é a criação de grandes volumes de dados, com relacionamentos complexos entre si. A grande quantidade de informações e a complexidade dos relacionamentos entre os dados dificultam a identificação de padrões e comportamentos, por parte dos administradores públicos e responsáveis pela tomada de decisão [4]. Outro agravante é que grande parte dos órgãos ou departamentos públicos utilizam bases de dados distintas, tanto em termos lógicos como físicos [1]. Isso dificulta não só a troca de dados entre os departamentos, como também a interpretação destas informações. Desta forma, as organizações públicas demandam por soluções de software que auxiliem na identificação de deficiências e oportunidades de negócio a partir da análise inteligente de seus dados operacionais [5][6]. A análise inteligente dos dados das instituições públicas pode ser realizada através de técnicas de mineração de dados.

As técnicas de mineração de dados se dividem em dois grupos: aprendizado supervisionado e aprendizado não supervisionado. O aprendizado supervisionado exige um

conhecimento prévio acerca dos dados a serem minerados. Este conhecimento é utilizado para treinar os algoritmos de mineração de dados. Os algoritmos de aprendizado não supervisionado dispensam qualquer informação prévia sobre os dados, já que estes algoritmos operam sobre as características e similaridades existentes nos registros. Em geral, as técnicas de aprendizado não supervisionado buscam por comportamentos e acontecimentos frequentes, ocultos nos dados operacionais das instituições [7][8][9].

A aplicação de técnicas de mineração de dados para os mais diversos fins, tem se intensificado nos países mais desenvolvidos. O governo norte-americano financia instituições privadas para que processe informações com o propósito de identificar indícios de infração, como o abuso na utilização de crédito do governo por servidores públicos e sonegação de impostos. A mineração de dados também é utilizada pelo setor militar norte-americano, que busca por indícios de atentados terroristas, e até mesmo na seleção de jovens para o serviço militar [7].

Braga [10], Oliveira [11], Mourady e Elragal [3] mostram que as plataformas de apoio ao planejamento público e a mineração de dados podem contribuir com o desenvolvimento econômico, fiscal e tributário das instituições públicas. Kum [12] mostra que a mineração de dados também pode ser utilizada para auto-avaliar o desempenho destas instituições, proporcionando melhor utilização dos recursos financeiros, humanos e tecnológicos. O trabalho de Kum [12] mostra que as técnicas de mineração de dados possibilitam otimizar os processos e agilizar a tramitação de documentos e protocolos administrativos, das instituições públicas.

Diversos trabalhos vêm sendo realizados com o objetivo de melhorar a gestão dos recursos públicos. Estes trabalhos utilizam técnicas de aprendizado supervisionado, que dependem de vocabulários controlados e dicionários de sinônimos. Outro problema comum nestes trabalhos é a capacidade de operar sobre um número pré-estabelecido de variáveis de análise, que limita o conjunto de treinamento e os padrões que podem ser obtidos com a mineração de dados.

A mineração de dados através das técnicas de aprendizado supervisionado eleva o custo operacional, pois necessitam de ontologias e analises de padrões, que serão utilizados durante o treinamento dos algoritmos de mineração de dados. Este tipo de solução dificulta a obtenção de

novas informações, pois os dados serão explorados de acordo com regras pré-estabelecidas. Nestes casos, os algoritmos de mineração de dados estão limitados a classificar os dados com rótulos já conhecidos, passando despercebido por novas informações.

Este trabalho propõe a construção de um Processo de Extração de Conhecimento (PEC) capaz de obter informação útil para a tomada de decisões. Este processo deve receber como entrada os dados operacionais das instituições públicas, que serão processados com o objetivo de identificar padrões ou comportamentos ocultos. A saída do PEC será tratada como o conhecimento obtido pela mineração de dados, que deve ser armazenado permitindo que outras aplicações do governo eletrônico utilizem tal informação para tomar decisões. Os resultados obtidos com a mineração de dados devem ser apresentados de forma organizada, de modo que facilite o entendimento por parte dos administradores públicos.

A técnica de mineração de dados utilizada para identificar padrões desconhecidos em grandes volumes de dados será Mapas Auto-Organizáveis (*Self-Organizing Maps* – SOM). Os mapas auto-organizáveis são redes neurais de aprendizado não supervisionado e competitivo. Neste tipo de rede, as unidades de processamento, denominadas neurônios, competem entre si pelo direito de representar um dado de entrada. Vence o neurônio cuja distância for menor em relação ao dado de entrada. O neurônio vencedor e seus vizinhos são adaptados em direção ao dado, porém os neurônios vizinhos são adaptados com menor intensidade [9][10]. Tendo em vista que o mapa auto-organizável é uma técnica de mineração de dados não supervisionada, é possível analisar um grande conjunto de dados, sem que se conheça qualquer característica sobre eles, possibilitando a descoberta de novos padrões.

O estudo de caso desenvolvido neste trabalho utiliza os dados de atendimentos sociais, cedidos pela Prefeitura Municipal de Campinas – SP. Estes dados são coletados pelo Módulo de Gestão Social do Sistema Integrado de Governança Municipal (SIGM).

1.1 Organização do Trabalho

O Capítulo 2 apresenta as seções técnicas do trabalho, que contextualiza o governo eletrônico e sua principal arquitetura de referência. Neste capítulo também será apresentado o Business Intelligence (BI) e os principais elementos que compõe este ambiente.

O Capítulo 3 apresenta o Processo para Extração de Conhecimento (PEC) e descreve com detalhes todas as fases realizadas por este processo.

O Capítulo 4 apresenta um estudo de caso real aplicado sobre uma ferramenta de governo eletrônico da Prefeitura Municipal de Campinas, SP – Brasil. Este capítulo apresenta também uma discussão técnica sobre os resultados obtidos.

O Capítulo 5 faz as considerações finais e apresenta as conclusões obtidas com este trabalho.

Capítulo 2

Seções Técnicas

ESTE capítulo apresenta as seções técnicas do trabalho, contextualizando o governo eletrônico e sua principal arquitetura de referência. Aqui também é apresentado o conceito de *Business Intelligence* (BI), juntamente com os elementos que o compõe, desde a extração dos dados operacionais, o tratamento e armazenamento dos dados pelo *Data Warehouse* (DW), até o reconhecimento de padrões através de análise exploratória não supervisionada. Aqui serão apresentados os principais tipos de padrões que podemos identificar em grandes volumes de dados, destacando a exploração não supervisionada através de mapas auto-organizáveis.

2.1 Governo Eletrônico

O sucesso da gestão pública é mensurado com base nos benefícios garantidos à sociedade. Organizações privadas, comunidades e cidadãos, exigem eficiência e responsabilidade na gestão dos recursos, além da garantia de entrega de melhores serviços e resultados [2].

Frente a este cenário, os países buscam revitalizar suas administrações públicas, inovando suas estruturas e procedimentos, além de qualificar os recursos humanos disponíveis. Neste contexto, a utilização das Tecnologias de Informação e Comunicação (TIC) constitui um papel fundamental para gestão e criação de um ambiente de crescimento econômico e social, proporcionando mais organização e eficiência na prestação de serviços aos cidadãos [2][13].

O termo Governo Eletrônico, em inglês *Electronic Government*, possui na língua inglesa uma expressão simplificada, *e-government*. A expressão também ganhou uma versão em português, e-governo. Uma abreviatura frequentemente encontrada, tanto em português como em inglês, é o termo e-gov. A Secretaria para Assuntos Fiscais do Banco Nacional de

Desenvolvimento Econômico e Social (SF/BNDES) define o governo eletrônico como o uso de TIC na prestação de serviços aos cidadãos, fornecedores e servidores [14].

As principais motivações para o uso das TIC na gestão e prestação de serviços públicos a comunidade, compreendem [13]:

- ✓ Troca rápida de informações entre os membros do governo;
- ✓ Facilidade de relacionamento entre o fisco e os contribuintes;
- ✓ Melhorar a qualidade dos serviços prestados aos cidadãos por meio do atendimento de demandas específicas;
- ✓ Prover maior transparência à gestão pública;

2.1.1 ARQUITETURA DE E-GOV

Para estabelecer e regulamentar os padrões de integração e troca de serviços entre governo, empresas e cidadãos, é importante definir uma arquitetura de e-gov. Esta arquitetura possibilita o entendimento do processo de implementação do governo eletrônico, e os requisitos acerca deste processo. Em geral, as arquiteturas de e-gov são dividas em camadas, e para cada camada devem ser adotadas soluções de TIC que contemple os requisitos existentes em cada ponto da arquitetura [1].

Embora existam trabalhos recentes que definem uma arquitetura genérica de e-gov, como o trabalho de Yan e Guo [6] e o trabalho de Carromeu [15], é possível notar que estes trabalhos se baseiam na proposta de Ebrahim e Irani [1], embora divirjam no número de camadas ou nas TIC empregadas em cada ponto da arquitetura. A arquitetura genérica de e-gov proposta por Ebrahim e Irani [1] é divida em quatro camadas: Camada de Acesso, Camada de e-gov, Camada de e-business e Camada de infraestrutura, como mostra a Figura 2.1:

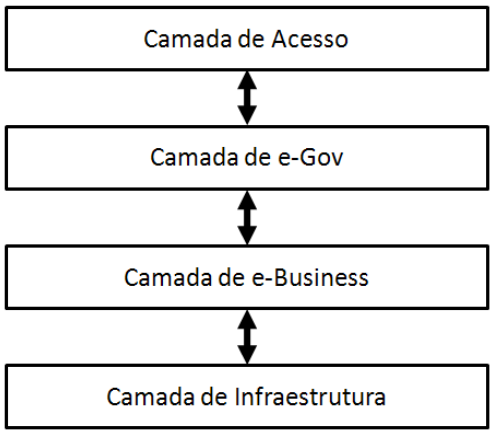


Figura 2.1 - Arquitetura das Aplicações de e-Gov [1].

A camada de acesso provê meios para distribuição de serviços, produtos e informações disponibilizados pelo e-gov. Estes meios consistem em canais de acesso on-line, como portais, que podem ser acessados através de computadores ou dispositivos móveis, e são capacitados para auxiliar na resolução de um problema [1].

Ebrahim e Irani [1] elencam alguns pontos que devem ser considerados para o sucesso no desenvolvimento desta camada:

- ✓ Criação de mecanismos para descoberta dos serviços e informações disponibilizadas, como catálogo de serviços e ferramentas de pesquisas;
- ✓ Estabelecer um padrão para navegação e apresentação das informações comuns, nos canais de atendimento *on-line*;
- ✓ Criar meios de atender usuários com diferentes conhecimentos, interesses e competências técnicas, garantindo que todos tenham suas necessidades atendidas.

A camada de e-gov pode ser vista como um repositório, onde estão alocados todos os serviços oferecidos pelo governo. O objetivo desta camada é estabelecer um ponto único de entrada para os usuários. Em geral, esta camada pode ser resolvida através de áreas restritas, onde os serviços são disponibilizados de acordo com o perfil do usuário, seja este uma empresa ou um cidadão. A definição do perfil do usuário permite a distribuição de serviços específicos, e facilita a navegação, por parte do usuário [1].

A camada de e-business é onde os sistemas dos diferentes órgãos e departamentos públicos podem ser integrados, no que se diz respeito a dados e serviços. Geralmente, os órgãos e departamentos públicos adotam soluções distintas de *software* e armazena seus dados em ambientes isolados, o que dificulta compartilhar informações e serviços entre estes departamentos. Nesta camada, os dados e serviços devem ser compartilhados através de uma interface distribuída, permitindo que os sistemas dos diferentes órgãos e departamentos públicos acessem as informações de um único lugar [6].

A camada de infraestrutura concentra as soluções de *hardware* que disponibilizam as informações e serviços pelos canais de acesso *on-line*. Podemos elencar como elementos da camada de infraestrutura os servidores de aplicação, roteadores e outros equipamentos que possibilitam distribuição dos serviços através da *Internet*, *Intranets* e *Extranets* [1]. Outro elemento muito importante desta camada é o meio de acesso aos serviços, disponibilizados pelo governo eletrônico. A Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2008, realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) mostra que apenas 34,8% dos entrevistados utilizaram a Internet nos últimos três meses, e destes, apenas 57,1% tem acesso a Internet em seu domicílio [16].

Alguns governos têm adotado alternativas para cobrir esta deficiência, criando Redes Metropolitanas de Acesso Aberto (RMAA). Estas redes são constituídas por diferentes meios de transmissão de dados, como fibra ótica, sem fio, etc. Esta solução possibilita a integração de órgãos governamentais, além de prover acesso a todas as pessoas do município, permitindo a distribuição dos serviços por canais de atendimento *on-line* [17].

Os serviços do governo eletrônico são acessados por meio da camada de acesso, que necessita da autenticação de seus usuários. Após autenticar o usuário, a camada de acesso fornece os dados da autenticação para a camada de e-gov, que verifica se o usuário é uma empresa, um cidadão ou uma agência do governo. Depois de verificar o perfil do usuário, a camada de e-gov obtém todos os serviços compartilhados pela camada de e-business. A camada de e-gov separa os serviços de acordo com o tipo de usuário, em seguida retorna estes serviços para a camada de acesso, onde serão apresentados.

A troca de informações entre as camadas do governo eletrônico é possível graças à camada de infraestrutura, que interliga os órgãos governamentais, os cidadãos e as empresas.

2.1.2 CLASSIFICAÇÕES DE E-GOV

Os sistemas e-gov são muito abrangentes e possuem usuários com necessidades distintas. No sentido de agrupar os serviços ofertados a cada grupo de usuários, surge a necessidade de classificar os sistemas de governo eletrônico. Esta classificação permite elencar os serviços e determinar os grupos de usuários aos quais estes serviços serão disponibilizados. Os sistemas de e-gov são classificados como: Governo para o Cidadão, Governo para Empresas, Governo para Governo, Aplicações de Eficiência e Efetividade Interna e Aplicações de Infraestrutura Global [1][4].

A classe de sistemas Governo para o Cidadão, concentra os serviços oferecidos aos cidadãos. Em geral, estes serviços são canais de comunicação que permitem ao cidadão solicitar aos órgãos públicos a execução de uma tarefa, como a limpeza e jardinagem de uma praça, ou simplesmente emitir a segunda via de impostos, como por exemplo, o Imposto Predial e Territorial Urbano (IPTU). Nesta classe de sistemas também se concentram serviços que possibilitam a emissão de documentos, como uma certidão negativa de débitos [1].

Como exemplo de um sistema de governo para o cidadão, podemos citar o trabalho de Ignatowicz [18], que propõe um modelo organizacional para distribuição e consumo de serviços. Neste modelo, as instituições governamentais podem disponibilizar serviços através de uma estrutura colaborativa, além de permitir o consumo destes serviços pelos cidadãos, através de um portal comunitário [18].

Outro exemplo de governo para o cidadão, é o trabalho de Silva [19]. Este trabalho propõe um espaço virtual que promove a interação social, para isso é utilizada a plataforma do *Second Life*. O objetivo é analisar a viabilidade da transposição de instâncias do governo eletrônico, quando aplicadas em ambientes virtuais. O espaço disponibiliza um serviço de votação, que permite aos cidadãos participar de projetos em estudo pelo governo. O trabalho mostra que a utilização de ambientes mais próximos da realidade, estimula o interesse dos usuários e tornam o uso dos serviços mais intuitivo.

A classe de sistemas Governo para Empresas concentra serviços que têm como objetivo facilitar a comunicação entre governo e empresas. Dentre os serviços ofertados para empresas, podemos citar os pregões eletrônicos, onde empresas fazem propostas pelo direito de executar serviços públicos ou vender equipamentos e materiais para o poder público [20].

A classe de sistemas Governo para Governo concentra os serviços que devem ser compartilhados entre órgãos e departamentos públicos. Em geral, os órgãos e departamentos públicos não adotam uma solução única de *software* e armazenamento de dados, e acabam mantendo as informações em ambientes separados e distintos. Frente a este cenário, o compartilhamento de informações e serviços é um desafio para as instituições públicas, que demandam por soluções de *software* e *hardware* que auxiliem na solução deste problema [6].

Como exemplo desta classe de sistemas, podemos citar o trabalho de Tilli [21], que apresenta uma arquitetura baseada na distribuição de serviços para desenvolvimento de aplicações de governo eletrônico. Este trabalho apresenta o conceito de cadastro unificado do cidadão, além de permitir compartilhar informações com outros sistemas do governo eletrônico, através de interfaces e objetos distribuídos.

A classe de sistemas de Efetividade e Eficiência Interna concentra as aplicações que visam melhorar a qualidade e eficiência dos processos internos aos órgãos e departamentos públicos. Como exemplo destas aplicações, podemos citar o trabalho de Kum [12], que propõe um sistema de descoberta de conhecimento para a auto-avaliação de resultados obtidos pelos órgãos e departamentos públicos, permitindo monitorar os resultados da implantação de políticas públicas em meio à sociedade.

A classe de sistemas de Infraestrutura Global compreende as questões de interoperabilidade entre as aplicações do e-gov, ou seja, garantem a integração dos dados e serviços oferecidos pelo governo eletrônico. As soluções empregadas nesta classe de sistemas abrangem recursos de *hardware* e *software*. Como exemplo de infraestrutura global, podemos citar o trabalho de Mendes [17], que estabelece meios de transmissão de dados possibilitando a integração de órgãos e departamentos governamentais, através da distribuição dos serviços por canais de atendimento *on-line*.

2.2 Business Intelligence

Business Intelligence (BI) compreende um conjunto de técnicas que permite identificar tendências de comportamento a partir de um conjunto de acontecimentos, auxiliando no processo de tomada de decisão de um negócio [22].

Com a evolução computacional e o aprimoramento dos mecanismos de armazenamento de dados, as organizações passaram a armazenar em meio digital todas as informações provenientes de suas atividades diárias, como tramitação de protocolos e documentos, registro de atividades realizadas por clientes, como compras e solicitações. As organizações começaram a enxergar estes dados como uma fonte de informações que poderia direcionar sua evolução e desenvolvimento. A crescente competição entre organizações, e a exigência de melhores serviços por parte dos clientes, impulsionaram o desenvolvimento de técnicas mais aprimoradas que permitem analisar grandes volumes de dados de forma inteligente. O BI é um conjunto de técnicas que permite interpretar as informações provenientes das atividades diárias de uma organização, por isso é classificado como sistema de suporte a decisão [22].

Os dados utilizados pelo sistema de suporte à decisão são coletados pelos sistemas de informação. Nos sistemas de informação, os dados são constantemente modificados e possuem relacionamentos complexos entre si, o que dificulta o entendimento dos dados e a extração de informação útil para a tomada de decisão. Surge então a necessidade de armazenar estes dados em ambientes simplificados, onde o grau de inter-relacionamento entre os dados é menor, proporcionando melhor desempenho em consultas e cruzamento de informações. Este cenário é atendido pelo conceito de *Data Warehouse* (DW), definido como bases de dados multidimensionais com um nível de normalização inferior às bases de dados transacionais, onde consultas podem ser realizadas mais rapidamente. No DW os dados não sofrem modificações constantes.

Os sistemas de suporte a decisão operam sobre os DW, e devem permitir aos usuários propor soluções, pesquisar histórico de decisões tomadas e simular situações diversificadas, utilizando apenas os dados operacionais coletados pelos sistemas de informação [22].

A Figura 2.2 apresenta o ambiente de BI e os elementos que compõe este ambiente:

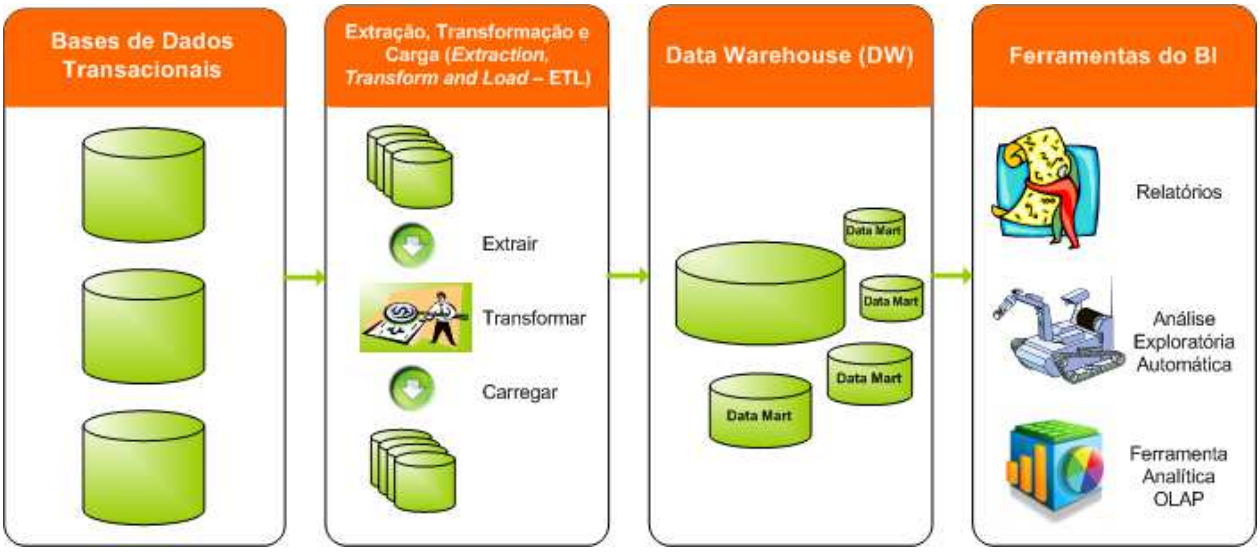


Figura 2.2 - Ambiente de BI.

Na Figura 2.2, o DW é alimentado com os dados trazidos das bases de dados transacionais. A inserção dos dados no DW é realizada por meio de ferramentas denominadas Extração, Transformação e Carga (*Extract, Transform Load* - ETL). As ferramentas ETL tratam os dados trazidos das bases de dados transacionais, antes de armazená-los no DW. Dentre os tratamentos realizados, podemos destacar a eliminação de registros duplicados, padronização de datas em um formato único, etc.

Portanto, os dados armazenados no DW estão mais consistentes que os dados presentes nas bases de dados transacionais, sem redundância e uniformizados, no que se diz respeito ao formato de números e datas. A unicidade dos registros e a uniformidade dos dados fazem com que as análises realizadas pelos sistemas de suporte a decisão, sejam mais eficientes.

As ferramentas de apoio e suporte a decisão se dividem em diferentes grupos, tais como: geradores de relatórios, ferramentas de exploração analítica e ferramentas de análise exploratória automática. As seções 2.2.1, 2.2.2 , 2.2.3 e 2.2.4 descrevem o processo de ETL, o DW, as ferramentas analíticas, e o processo de análise exploratória através da mineração de dados.

2.2.1 EXTRAÇÃO, TRANSFORMAÇÃO E CARGA

O processo de ETL é um elemento fundamental para a construção do DW, pois os dados a serem armazenados neste ambiente vêm de diferentes bases de dados transacionais, muitas

vezes não integrados e com modelagens distintas. O processo de ETL é divido em duas fases. Na primeira fase, os dados são extraídos das bases de dados transacionais e submetidos a uma série de transformações. Estas transformações integram os dados resolvendo as diferentes nomenclaturas de uma mesma variável. Um exemplo é a situação na qual um sistema define o gênero de uma pessoa como "Masculino" ou "Feminino" e o outro define o sexo como "M" e "F". No DW, todos os dados trazidos de sistemas diferentes devem ser uniformizados, e tal uniformização só é garantida no processo de transformação. Na segunda fase os dados tratados são armazenados no DW. O ETL é uma das tarefas de maior importância e complexidade no projeto de um sistema de BI. Uma carga de dados de má qualidade pode resultar em análises imprecisas [23][24].

2.2.2 DATA WAREHOUSE

Construir um ambiente de suporte a decisão utilizando dados armazenados em bases de dados transacionais é um processo complexo. As bases de dados transacionais fazem o controle operacional das organizações e em geral não estão interligadas. As bases de dados transacionais possuem modelagens distintas e estão em constante atualização, dificultando a realização de análises históricas sobre os dados. Neste contexto, surge a necessidade de uma base de dados orientada a assuntos, integrada, variante no tempo e não volátil, que facilite a tomada de decisão [24][25].

Inmon [25] elenca um conjunto de características do DW, que proporcionam melhor desempenho e qualidade nas análises realizadas pelos sistemas de suporte a decisão:

- ✓ Orientado a assuntos: Os dados não estão mais organizados de acordo com as regras de negócios dos sistemas, mas sim de acordo com as áreas de interesse da organização. Por exemplo: vendas de produtos a diferentes tipos de clientes, atendimentos e diagnósticos de pacientes, etc.
- ✓ Integrado: Diferentes nomenclaturas, formatos e estruturas das fontes de dados precisam ser agrupados em um único esquema, provendo uma visão unificada e consistente da informação. Por exemplo, um sistema pode tratar o gênero das

pessoas como Masculino e Feminino, outro como M e F. Ao serem passados para o DW, estes dados devem ser unificados para um único formato de apresentação.

- ✓ Não volátil: Os dados de um DW não são modificados como nas bases de dados transacionais, exceto para correções. Os dados são carregados e acessados para leituras.
- ✓ Variantes no Tempo: Os dados em um DW não sofrem modificações, o que permite a realização de análises históricas.

A principal função do DW é consolidar as informações a serem analisadas, unificando as diferentes nomenclaturas trazidas das bases de dados transacionais. No DW, as informações são organizadas a partir de áreas de interesse, apresentando os dados em um formato mais apropriado para a tomada de decisão [22][25].

Em alguns casos é possível obter DW específicos, denominados *Data Marts* (DM). Um DM é uma visão limitada dos dados armazenados no DW, normalmente empregado a uma determinada área ou assunto. O principal objetivo em ter um subconjunto das informações armazenadas no DW, é uma melhoria no desempenho das consultas realizadas neste ambiente. Para instituições públicas, por exemplo, poderíamos ter um DM para a área da saúde e outro para a educação. Em geral, as diferenças existentes entre um DM e um DW referem-se apenas ao tamanho do projeto. No entanto, um DM trata das questões departamentais, enquanto um DW trata as necessidades da instituição como um todo [22][24].

2.2.3 FERRAMENTA ANALÍTICA DE PROCESSAMENTO ON-LINE

As Ferramentas Analíticas de Processamento *On-line* (*On-line Analytical Processing* – OLAP) permitem aos usuários comparar e analisar um grande volume de dados com bastante flexibilidade e bom desempenho. A navegação intuitiva proporcionada por estas ferramentas permite aos usuários obter informações estratégicas, que podem facilitar a tomada de decisão. As ferramentas OLAP são conjuntos de soluções direcionadas à análise específica dos dados, com o intuito de transformar dados em informações gerenciais.

As ferramentas OLAP disponibilizam operações exclusivas que permitem consultar, manipular e analisar grandes volumes de dados armazenados nos DWs. A seguir, são apresentadas as principais operações realizadas pelas ferramentas OLAP [26][27]:

- ✓ *Slice and Dice*: Recurso que permite criar diferentes visões dos dados através da reorganização, permitindo analisar as informações por diferentes perspectivas.
- ✓ *Drill Down*: Esta operação permite que o usuário aumente o nível de detalhes da informação que está sendo consultada, aumentando assim o nível de granularidade dos dados.
- ✓ *Drill Up ou Roll Up*: É o contrário do *drill down*, neste caso o usuário diminui o nível de detalhe da informação, diminuindo assim o nível de granularidade dos dados.

2.2.4 MINERAÇÃO DE DADOS OU ANÁLISE EXPLORATÓRIA AUTOMÁTICA

Mineração de dados, do inglês, *Data Mining*, consiste na análise de grandes volumes de dados com objetivo de reconhecer novos padrões e tendências, a partir das informações de uma organização. Geralmente, estes dados são armazenados em bases de dados transacionais ou em DW. A mineração de dados utiliza técnicas de reconhecimento de padrões que buscam por similaridades existentes nos dados analisados.

Padrões são caracterizados por fatos reincidentes. Um exemplo de fato reincidente é quando várias pessoas apresentam a mesma enfermidade em uma época do ano. Se este evento voltar a ocorrer nos anos seguintes, este pode ser considerado um padrão. Se isso ocorrer, a secretaria da saúde pode se precaver comprando medicamentos, além de aumentar o efetivo de profissionais qualificados para aquele tipo de atendimento. A mineração de dados pode identificar este tipo de comportamento eliminando os fatos com menor incidência e destacando aqueles que ocorrem com mais frequência. O processo de descoberta realizado pela mineração de dados pode ser executado a partir de bases de dados transacionais, porém é mais eficiente realizá-lo a partir de um DW, onde os dados não apresentam erros ou duplicidade, são consistentes e possibilitam descobertas abrangentes e precisas [26][27][28].

Para reconhecer padrões, é preciso definir uma Tarefa de Mineração. Estas tarefas são tipos de padrões definidos com base em experiências práticas de exploração de dados. As duas principais tarefas são a preditiva e a descritiva. A primeira trata do uso de variáveis no banco de dados para predizer os valores das outras variáveis, já a segunda identifica padrões em dados históricos [26][28].

Os principais tipos de padrões que podem ser reconhecidos através da mineração de dados são: classificação de valores a partir de rótulos previamente estabelecidos, também conhecido como classificação supervisionada; identificação de padrões associativos; identificação de padrões seqüenciais; identificação de agrupamentos de dados, também conhecido como classificação não supervisionada ou clusterização.

A classificação de valores a partir de rótulos previamente estabelecidos pode ser definida como o processo no qual uma ou mais amostras de identidade conhecidas são utilizadas para classificar dados com características desconhecidas. Ela se baseia na disponibilidade de um conjunto de padrões anteriormente classificados, denominado conjunto de treinamento. O propósito desta tarefa de mineração é classificar objetos ainda não rotulados [9][28].

A identificação de padrões associativos utiliza as relações existentes nos dados. Estas relações podem ser da forma X → Y, onde X e Y pertencem ao conjunto de valores (artigos comprados por um cliente, sintomas apresentados por um paciente). Geralmente, este tipo de tarefa é aplicado em organizações varejistas, e permitem identificar conhecimentos não triviais que auxiliem nas estratégias de marketing e na reorganização das lojas [28].

Um padrão sequencial é uma expressão da forma <I1, I2, ..., In>, onde cada Ii é um conjunto de itens. A ordem de alinhamento destes conjuntos reflete a ordem cronológica em que os fatos ocorreram. A maior parte das pesquisas já realizadas sobre a mineração de padrões seqüenciais concentra-se na descoberta de séries temporais, normalmente aplicadas nos mercados financeiro, varejo, medicina e previsão do tempo [28].

A identificação de agrupamentos de dados é uma tarefa que permite agrupar dados a partir da similaridade entre eles. Normalmente, utilizamos esta tarefa de mineração quando não conhecemos nada sobre os dados a serem minerados. Diferente da classificação supervisionada,

esta tarefa de mineração dispensa um conjunto de treinamento, ou seja, amostras de identidade já conhecidas [9][28].

A seção 2.3 apresenta com detalhes a identificação de agrupamentos de dados através de redes neurais artificiais.

2.3 Exploração não Supervisionada ou Clusterização

A mineração de dados possibilita a descoberta de padrões desconhecidos, gerando novas informações a partir dos dados operacionais de uma instituição. Padrões ocultos nos dados podem auxiliar a tomada de decisões estratégicas, direcionando o negócio com base em ciclos de acontecimentos anteriores. As técnicas de exploração não supervisionada dispensam qualquer conhecimento prévio acerca dos dados, ou seja, operam sobre conjuntos de dados não classificados, de tipos, classes e grupos desconhecidos. Em geral, as técnicas de exploração não supervisionada são utilizadas como ponto de partida para futuras investigações, pois permitem compreender a forma em que os dados são divididos e organizados [29].

Existem diversas técnicas de exploração não supervisionada, as quais se dividem em métodos hierárquicos e não hierárquicos. Os métodos hierárquicos organizam os dados em uma estrutura tipo árvore, enquanto os métodos não hierárquicos localizam grupos de dados através de centros previamente estabelecidos. Um exemplo de método não hierárquico é o algoritmo *k-means*. Neste algoritmo, o número de agrupamentos é previamente estabelecido, e para cada agrupamento é atribuído um centro. O algoritmo consiste em adaptar os centros para que se aproximem das massas de dados, formando os grupos [30].

Os Mapas Auto-Organizáveis, do inglês *Self Organizing Mapas* (SOM) é o método de maior destaque para identificar agrupamentos de dados e classificar informações desconhecidas. Trata-se de um tipo de rede neural artificial de aprendizado não supervisionado e competitivo, que apresenta algumas vantagens com relação aos demais métodos. As principais vantagens são [29][31]:

- ✓ Normalmente, um grande volume de dados possui dimensionalidades mais altas, dificultando o entendimento de padrões e comportamentos ocultos. Os mapas

auto-organizáveis proporcionam uma redução de dimensionalidade dos dados a serem analisados, apresentando as classes ou grupos encontrados em grades bidimensionais.

- ✓ Os mapas auto-organizáveis exigem menor esforço computacional, se comparado aos métodos hierárquicos. A cada nova informação, os métodos hierárquicos necessitam percorrer toda a estrutura de dados já analisada e localizar o nó mais adequado para associar a nova informação. Os mapas auto-organizáveis operam sobre métodos iterativos, onde cada nova informação proporciona um grau de adaptação, ou aprendizado da rede neural. Desta forma, o mapa se organiza automaticamente sem a necessidade de investigar as informações fornecidas anteriormente.
- ✓ Os mapas auto-organizáveis não requerem uma estimativa prévia do número de agrupamentos existentes nos dados, como é necessário para o método *K-means*. O método *K-means* utiliza esta estimativa para organizar os dados em torno dos centros previamente estabelecidos, formando os grupos. O problema deste método é que o posicionamento inadequado dos centros pode levar a obtenção de agrupamentos errados, unindo elementos distintos ou até mesmo separando elementos que deveriam pertencer ao mesmo agrupamento;

2.3.1 REDES NEURAIS ARTIFICIAIS

As Redes Neurais Artificiais (RNA) compõem uma subárea da inteligência artificial que se baseia na estrutura do cérebro humano. Elas utilizam funções matemáticas não lineares e possuem a capacidade de adquirir, armazenar e utilizar o conhecimento. Uma RNA é constituída por um conjunto de células computacionais interligadas, denominada neurônios. Neurônios são unidades de processamento conectadas por canais de comunicação, que normalmente estão associados a respectivos pesos. A Figura 2.3 representa um modelo de neurônio artificial [9][32].

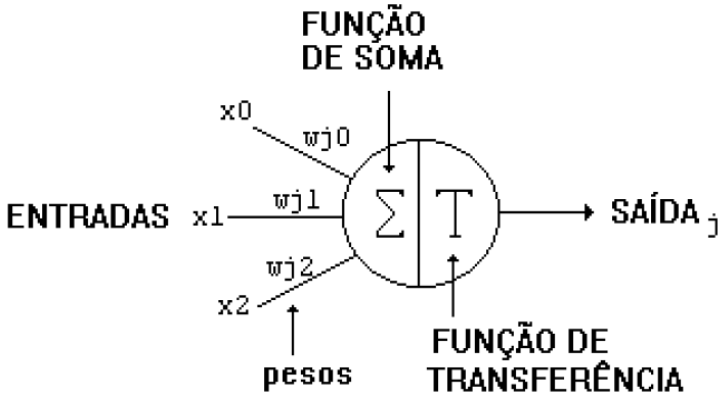


Figura 2.3 - Neurônio Artificial [9].

O neurônio artificial apresenta sinais de entrada, que são multiplicados pelos pesos sinápticos w. Somando-se ponderadamente os sinais, provenientes das diversas entradas e aplicando-se a função de transferência, é produzido um nível de ativação. Caso o nível exceda seu limite, inicia-se então a produção de sinais de saída [33].

A utilização de uma RNA na solução de uma tarefa passa inicialmente pela fase de aprendizagem, onde informações relevantes sobre os padrões são extraídas e apresentadas para a rede neural. A fase de aprendizagem consiste em um processo iterativo de ajuste de parâmetros da rede, armazenando o conhecimento nos pesos sinápticos das conexões entre as unidades de processamento. Kohonen [32] classifica as redes neurais em três categorias: redes de transferência de sinal, redes de transferência de estados e redes competitivas.

Nas redes de transferência de sinal, a saída da rede depende única e exclusivamente do valor de entrada. São exemplos deste tipo de rede os Perceptrons de Múltiplas Camadas – *Multi-Layer Perceptron* (MLP) e as redes de função de base radial – *Radial Basis Function* (RBF). Estas redes são usadas como classificadores supervisionados de padrões.

As redes de transferência de estado têm como característica a retroalimentação, garantindo que o estado de atividade convirja para um valor estável. Os valores de entrada acionam o estado inicial de atividade, após processar as saídas, estas serão apresentadas novamente como entradas para a rede neural, até chegar a seu estado final. As redes de Hopfield são exemplo de redes de transferência.

As redes de aprendizagem competitiva baseiam-se no processo competitivo entre seus neurônios. A aprendizagem competitiva é um processo adaptativo, onde os neurônios da rede se tornam sensíveis a diferentes categorias de entrada. É exemplo deste tipo de rede o Mapa Auto-Organizável – *Self Organizing Map* (SOM).

2.3.2 MAPAS AUTO-ORGANIZÁVEIS

O mapa auto-organizável foi desenvolvido com base na região do córtex cerebral humano. O córtex cerebral possui regiões especializadas em atividades específicas, por exemplo, os movimentos motores. Quando uma região do córtex cerebral é ativada, as demais regiões também sofrem um estímulo, porém com menor intensidade. A intensidade deste estímulo diminui à medida que a área se distancia da região ativada. [34].

Em 1982, Tuevo Kohonen desenvolveu um algoritmo competitivo denominado mapa auto-organizável, que possui a capacidade de organizar um conjunto de dados, separando-os em grupos de acordo com critérios de similaridade. Os mapas auto-organizáveis são redes neurais competitivas, organizadas em duas camadas: a camada de entrada e a camada de saída. Cada neurônio da camada de entrada está conectado a todos os neurônios da camada de saída por meio de vetores de pesos, como mostra a Figura 2.4:

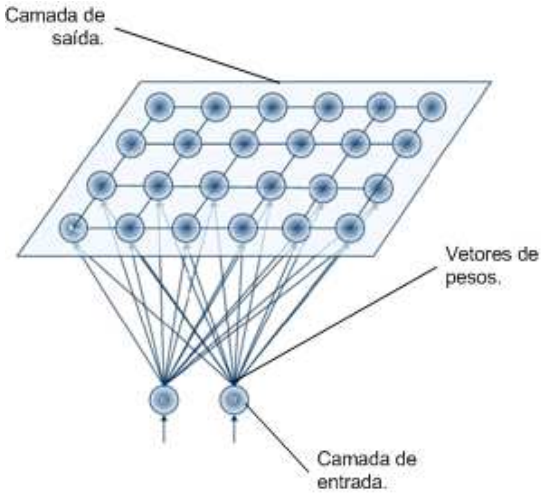


Figura 2.4 - Mapa Auto-Organizável [35].

O funcionamento do mapa auto-organizável inicia com a inserção do conjunto de dados a ser analisado. Este conjunto é apresentado a uma rede neural, e cada neurônio da rede possui

um vetor de pesos associado a todos os dados do conjunto. Em seguida, há uma competição entre todos os neurônios, que disputam o direito de representar os dados de entrada. Vence a competição o neurônio cujo vetor de pesos possuir menor distância, em relação ao dado de entrada. Este neurônio recebe o nome de *Best Matching Unit* (BMU) [36]. A distância do vetor de pesos com relação a cada neurônio é calculada de acordo com uma métrica, que pode variar de acordo com os dados de entrada. Em geral, utiliza-se a distância Euclidiana para a escolha do neurônio BMU [36]. A métrica Euclidiana é apresentada em (1), onde v_n é um dado de entrada e m_i é o peso da conexão entre cada neurônio i e a entrada n [36][37].

$$d(m_i, v_n) = \| m_i - v_n \| = \sqrt{\sum_{j=1}^{D} | m_{ij} - v_{nj} |^2} \quad \textbf{(1)}$$

Seja o conjunto de entrada $V = \{v_1 ... v_n\}, V \subseteq R^D$, de vetores $v_n = [v_{n1} ... v_{nD}] \in R^D$, onde cada vetor v_n representa um ponto no espaço dimensional D, através de seus D atributos. O mapa auto-organizável é definido por um conjunto de neurônios dispostos em uma matriz.

O neurônio BMU sofre um ajuste em seu vetor de pesos, para se aproximar ainda mais do dado analisado. Este ajuste aumenta a probabilidade de que este mesmo neurônio volte a vencer na próxima apresentação do mesmo dado. Os neurônios próximos ao BMU também terão seu vetor de pesos ajustado na direção do dado, embora com menor intensidade. Estes neurônios são considerados vizinhos do neurônio BMU. Ao longo do aprendizado, o ajuste dos pesos sinápticos do neurônio BMU e de sua vizinhança promovem a organização geral do mapa. O ajuste dos pesos sinápticos dos neurônios é definido a seguir por (2), onde t representa o instante de tempo e $\alpha(t)$ que define a taxa de aprendizado no instante de tempo t. A função de vizinhança no tempo t é representada por $h_{ic}(t)$:

$$m_i(t+1) = m_i(t) + \alpha(t).h_{ci}(t).[m_i(t) - v_n(t)] \quad \textbf{(2)}$$

A relação de vizinhança entre os neurônios é estabelecida segundo a função Gaussiana. O principal objetivo desta função é controlar o nível de atuação dos neurônios em torno do BMU. Segundo o modelo biológico, o nível de atuação dos neurônios vizinhos decai à medida

que eles se distanciam do neurônio BMU [9][36][37]. O grau de adaptação do neurônio BMU e de seus vizinhos depende da função de vizinhança e da taxa de aprendizado. Para ocorrer uma convergência do mapa, a função deve reduzir o grau de vizinhança relativo ao neurônio BMU ao longo do aprendizado [32].

A função gaussiana é definida em (3), onde r_c e r_i representam as posições dos neurônios c e i dentro do mapa, quando $\| r_c - r_i \|^2$ aumenta h_{ic} sofre uma redução exponencial. A largura da vizinhança, também conhecida como raio, é definida por $\sigma(t)$. Normalmente $\sigma(t)$ → 0 quando t → ∞. [36][37]:

$$h_{ci} = \exp\left(-\frac{\| r_c - r_i \|^2}{2\sigma^2(t)} \right) \quad \textbf{(3)}$$

A Figura 2.5 ilustra o cálculo do grau de adaptação dos neurônios vizinhos ao BMU. Quanto mais distante do BMU, menor será a adaptação aplicada ao neurônio vizinho.

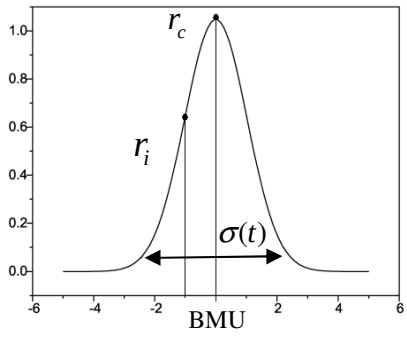


Figura 2.5 - Grau de adaptação aplicado aos neurônios vizinhos.

A Figura 2.6 ilustra o processo de adaptação do neurônio BMU e seus vizinhos em direção ao dado de entrada v_k:

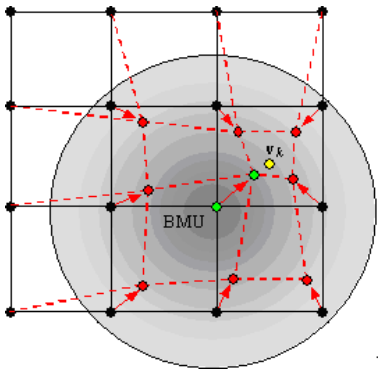


Figura 2.6 - Processo de adaptação do neurônio BMU e seus vizinhos [36].

Para cada época do aprendizado, os padrões de entrada devem ser apresentados de forma aleatória, garantindo a uniformidade da auto-organização. Uma época é definida pela apresentação de todo o conjunto de dados para o mapa auto-organizável [9]. O mapa auto-organizável geralmente precisa passar por várias épocas até atingir um ponto de convergência [37].

2.3.3 MATRIZ DE DISTÂNCIA UNIFICADA

A matriz de distância unificada, também conhecida como Matriz-U permite a representação gráfica dos agrupamentos encontrados pelo mapa auto-organizável. Esta matriz utiliza a métrica Euclidiana, descrita em (1), para calcular a distância entre os neurônios adjacentes. O resultado obtido com a aplicação da Matriz-U é uma imagem onde o nível de densidade de cada pixel corresponde à distância calculada. A partir de um mapa bidimensional, com topologia hexagonal, podemos encontrar a Matriz-U calculando as distâncias dx, dy e dz, para cada neurônio. O valor du da Matriz-U é a média ou mediana, calculado em função dos valores dos elementos circunvizinhos do neurônio du [38], como mostra a Figura 2.7:

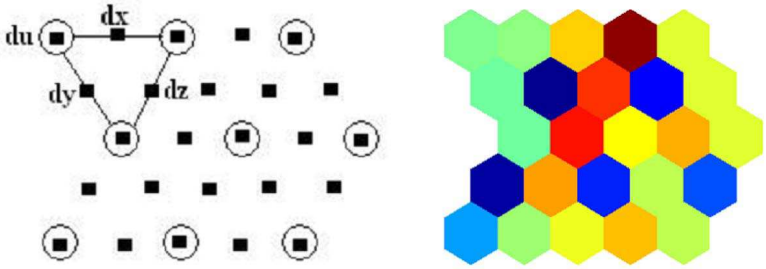


Figura 2.7 - Matriz de Distância Unificada (Matriz-U) [37].

A Matriz-U pode ser interpretada através da coloração dos pixels, de acordo com a intensidade de cada elemento da matriz. Os elementos de coloração mais escura correspondem a elementos dissimilares, já os elementos de coloração mais clara, ou seja, de menor intensidade corresponde aos elementos similares. Os elementos de maior intensidade são as fronteiras entre os agrupamentos [38].

A Matriz-U permite a visualização dos dados em uma grade bidimensional sem perdas de informação, mesmo que os dados estejam em dimensionalidades maiores. Por exemplo, podemos ter as seguintes informações em um conjunto de dados: gênero, idade e renda. Estas informações estão no plano R_3, como mostra a Figura 2.8:

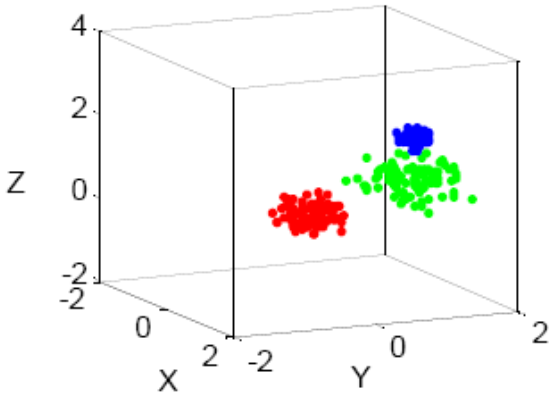


Figura 2.8 - Visão dos dados no plano R_3 [36].

O cluster em vermelho representa as pessoas do mesmo gênero. O cluster em verde representa as pessoas de mesma idade, enquanto o cluster em azul representa as pessoas com a mesma renda. Neste exemplo considera-se que não existe interpolação dos agrupamentos.

Ao tentar visualizar os dados apresentados pela Figura 2.8 em um plano bidimensional nos eixos X e Y, temos a impressão de que dois agrupamentos de dados estão sobrepostos, pois ao reduzir a dimensionalidade do R_3 para o R_2 perdemos parte da informação, neste caso a profundidade. A Figura 2.9 ilustra este processo:

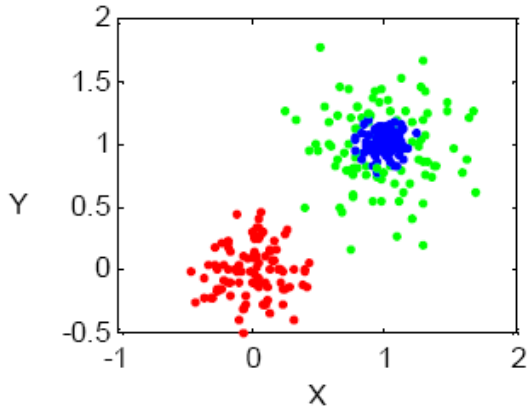


Figura 2.9 - Visão dos dados no plano R_2 [36].

A Matriz-U permite fazer uma redução de dimensionalidade, apresentando os dados sem que haja perda de informações. Esta redução facilita a identificação dos agrupamentos existentes nos dados. A Figura 2.10 apresenta a Matriz-U dos dados no R_2:

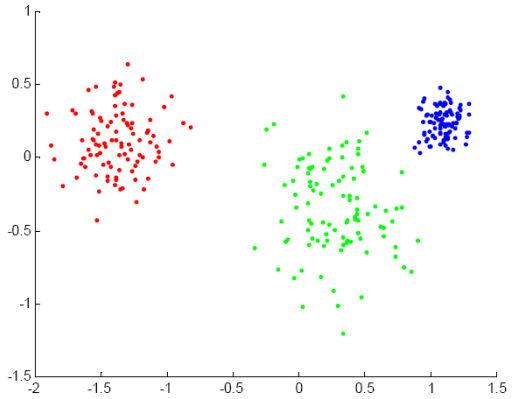
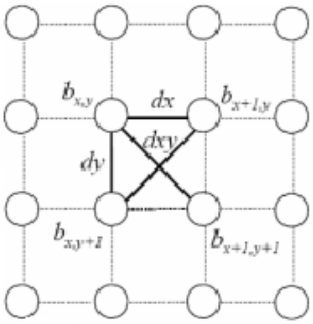


Figura 2.10 - Redução de dimensionalidade do R_3 para o R_2 através da Matriz-U [36].

A Matriz-U é representada por uma grade retangular N x M. Seja $b_{x,y}$ a matriz de neurônios do mapa auto-organizável e $w_{jx,y}$ a matriz dos vetores de pesos. Para cada neurônio b existem três distâncias d_x, d_y e d_{xy} na Matriz-U [38].



2.11 - Matriz-U [38].

As distâncias d_x, d_y e d_{xy} são obtidas pela métrica Euclidiana. Estas distâncias são calculadas no espaço de pesos, e apresentadas em uma Matriz-U de tamanho $(N-1)$ x$(M-1)$ [38]. Em geral, a utilização da Matriz-U é restrita a visualização dos agrupamentos, auxiliando a separação dos dados, com base na semelhança dos registros.

Capítulo 3

Processo de Extração de Conhecimento

ESTA seção apresenta o Processo de Extração de Conhecimento (PEC) proposto por este trabalho, capaz de obter informações que auxiliem na tomada de decisão. Este processo tem como entrada os dados operacionais de uma instituição, e após um processamento, apresenta os padrões e comportamentos ocultos nos dados operacionais. Estes padrões são apresentados em uma estrutura organizada, e permitem a compreensão dos resultados de maneira rápida e clara. A Figura 3.1 apresenta o PEC proposto:

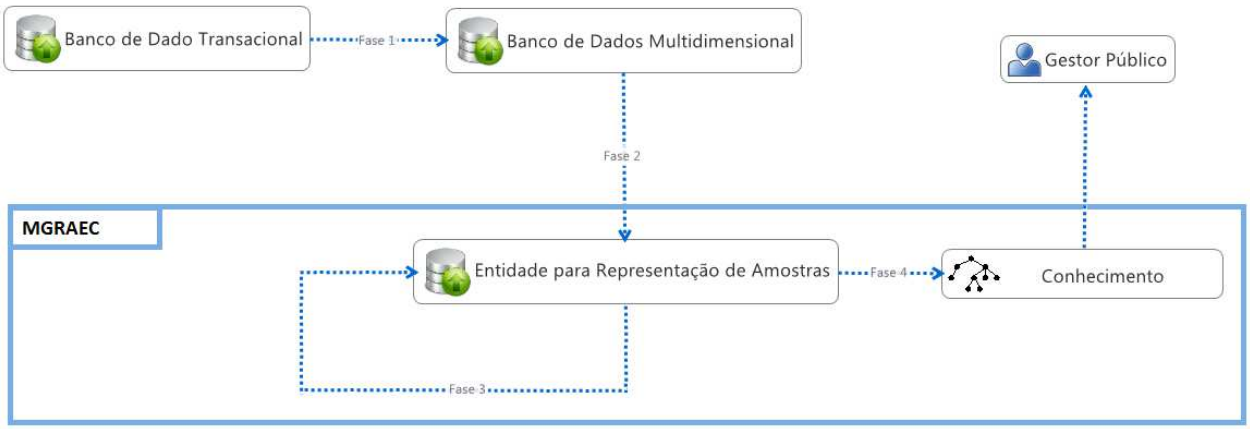


Figura 3.1 - Processo de Extração do Conhecimento (PEC).

Na Fase 1 do PEC, uma rotina ETL extrai os dados das bases de dados transacionais e armazena estes dados em um DW, representado pelo Modelo Conceitual Multidimensional (MCM). Esta rotina ETL elimina os registros duplicados e padroniza o formato de datas e valores numéricos.

Uma vez que os dados estão armazenados no MCM, daremos início a Fase 2. Nesta fase os dados passam por um segundo processamento ETL, para que sejam armazenados nas Entidades para Representação de Amostras (ERA). As ERA pertencem ao Modelo Genérico para

Representação de Amostras e Extração de Conhecimento (MGRAEC). Este modelo permite armazenar os dados que serão submetidos ao mapa auto-organizável e o conhecimento obtido pela mineração de dados.

O processamento realizado na Fase 2 associa valores numéricos aos dados, além de reduzir o número de variáveis de algumas dimensões. Os valores numéricos associados aos dados serão utilizados pelo mapa auto-organizável durante a Fase 3 do PEC.

Na Fase 3, o mapa auto-organizável recebe os dados armazenados nas ERA e da início a mineração de dados. Ao término desta fase, o mapa auto-organizável deve ter separado e agrupado os dados a partir da similaridade existente entre os registros. A divisão dos dados permitirá a associar cada registro armazenado nas ERA, com os agrupamentos encontrados pelo mapa auto-organizável.

A Fase 4 é responsável por resumir e organizar as características dos agrupamentos em uma estrutura do tipo árvore. As características dos agrupamentos serão tratadas como o conhecimento obtido pela mineração de dados. Este conhecimento será armazenado no MGRAEC e ficará disponível para a consulta, por parte dos responsáveis pela tomada de decisão.

As Fases 2, 3 e 4 do PEC, armazenam informações no MGRAEC. Na Fase 2 os dados extraídos do MCM são armazenados no MGRAEC, preenchendo as entidades: DIMENSAO, VARIAVEL_DIMENSAO, VARIAVEL e REGISTRO. Estas entidades são chamadas de Entidades para Representação de Amostras (ERA).

Após o término da Fase 3 do PEC, os agrupamentos encontrados pelo mapa auto-organizável serão armazenados na entidade GRUPO. Nesta fase os registros armazenados nas ERA serão associados aos agrupamentos encontrados pelo mapa auto-organizável.

Na Fase 4, as características dos agrupamentos serão resumidas e armazenadas na entidade CONHECIMENTO. A entidade CONHECIMENTO possui um auto-relacionamento, que caracteriza uma estrutura do tipo árvore.

A Figura 3.2 apresenta as entidades do MGRAEC e as fases do PEC em que elas são preenchidas.

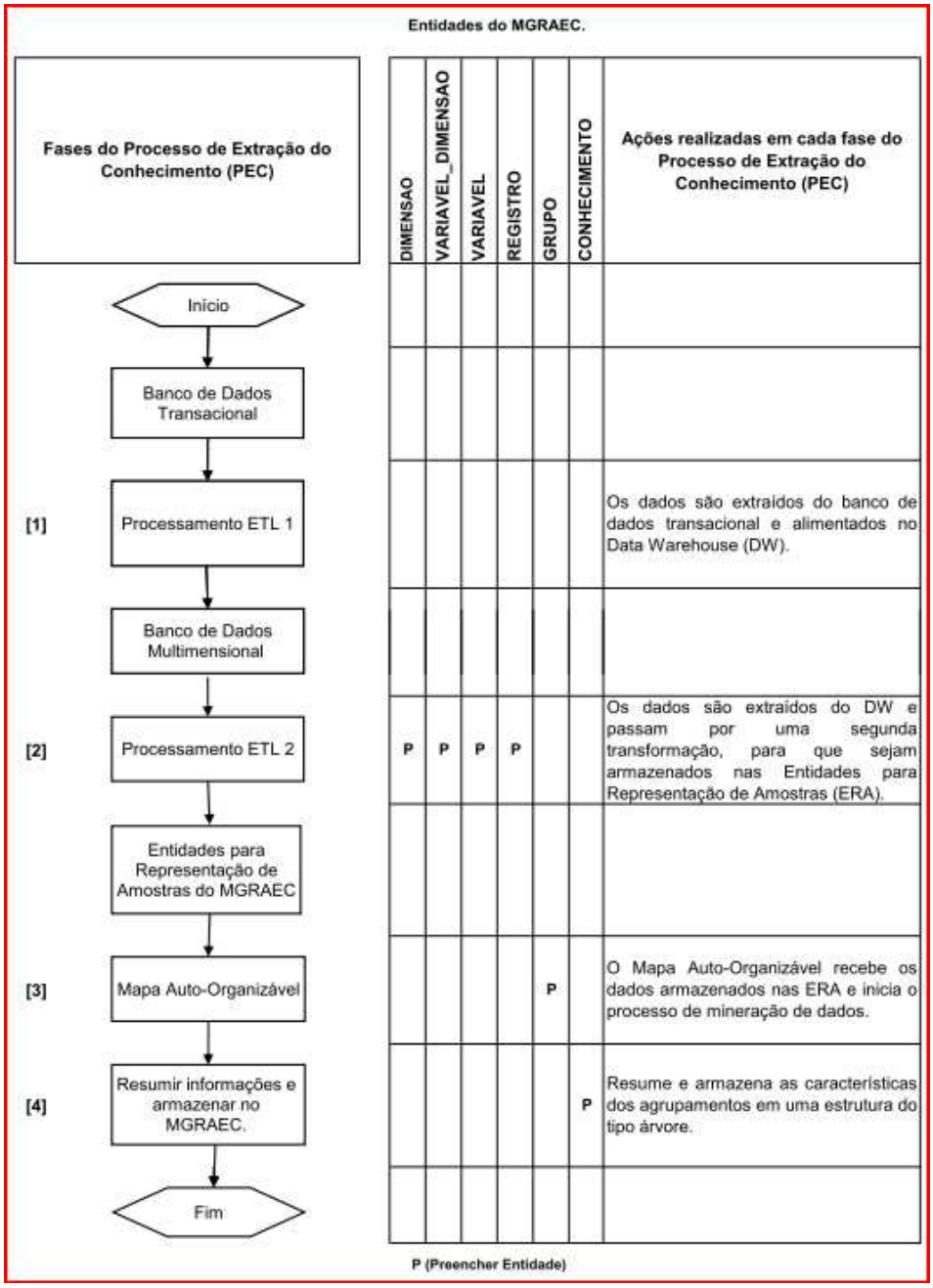


Figura 3.2 – Preenchimento das Entidades do MGRAEC.

A Figura 3.3 ilustra o PEC apresentado anteriormente. Nela é possível notar que os dados são extraídos das bases de dados transacionais e armazenados no MCM, através de um processo de ETL. Este processo de ETL corresponde a Fase 1 do PEC.

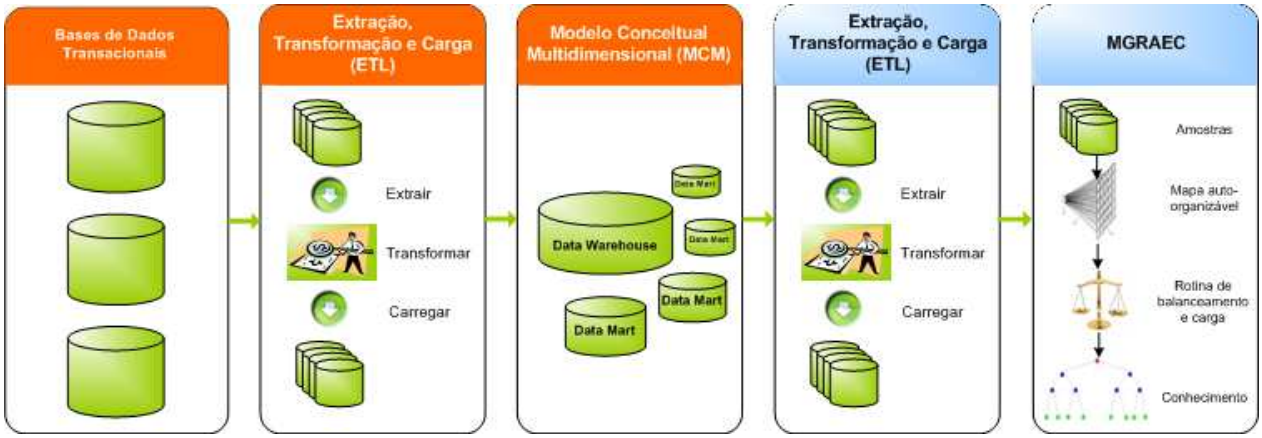


Figura 3.3 - Processo de Extração do Conhecimento.

3.1 Modelo Conceitual Multidimensional

O MCM utilizado no processo de extração do conhecimento é o modelo proposto por Marques [4], cujo objetivo é aplicar a inteligência de negócio em sistemas de governo eletrônico.

O ambiente proposto por Marques [4] é composto pela integração de diferentes ferramentas e tecnologias de código aberto, que contemplam desde a obtenção e transformação dos dados, até a disponibilização de ferramentas que permitem aos usuários finais analisar e manipular as informações armazenadas no MCM, seguindo uma navegação intuitiva. Marques [4] utiliza uma arquitetura divida em três camadas: A Camada ETL, a Camada de Armazenamento e Disponibilização de Visões de Dados e a Camada Aplicações para Usuários Finais.

As Figura 3.4 a e b apresentam a arquitetura utilizada e as ferramentas propostas no ambiente de BI de Marques [4].

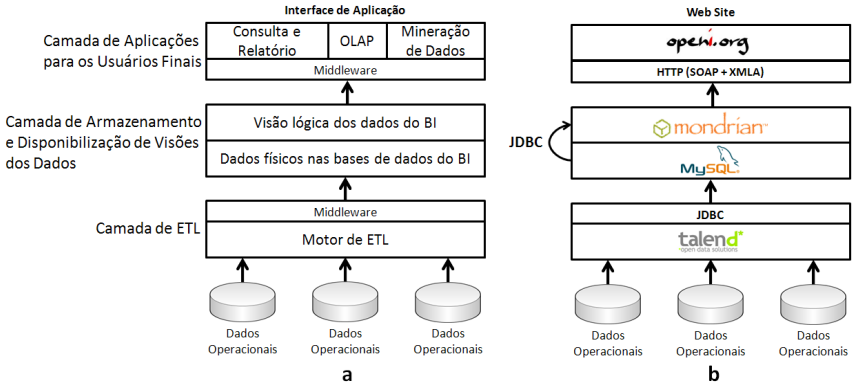


Figura 3.4 - Ambiente proposto por Marques [4].

A seguir serão descritos os objetivos de cada camada da arquitetura, e as soluções desenvolvidas por Marques [4]:

3.1.1 CAMADA ETL

A camada ETL é responsável pelo processo de extração, transformação e carga dos dados nos repositórios de dados operacionais, para as bases de dados do MCM. A camada ETL é divida em subcamadas: Motor ETL e Middleware. Na implementação da subcamada Motor ETL, foi adotada a ferramenta Talend Open Studio, que é especializada na integração e migração de dados. Para a escolha desta ferramenta, Marques [4] considerou a documentação disponível e a facilidade de disponibilização das rotinas de exportação em arquivos jar.

As transformações aplicadas aos dados incluem a remoção de registros duplicados e a convergência de nomenclaturas e valores como: valores monetários, datas e dados de domínio como sexo, tipos de deficiência, raça ou cor. Além das transformações citadas, os dados advindos das bases de dados transacionais passam por uma adequação estrutural, acomodando as informações em uma estrutura orientada a assuntos.

3.1.2 CAMADA DE ARMAZENAMENTO E DISPONIBILIZAÇÃO DE VISÕES DOS DADOS

Esta camada é responsável pela gestão dos dados armazenados no MCM, resultantes do processo de ETL. Esta camada é dividida nas seguintes subcamadas: Dados físicos nas bases de dados do BI, cuja função é prover mecanismos para o armazenamento de dados; Visão lógica dos dados do BI, responsável por gerar representações dos dados para camadas superiores. Para

o armazenamento de dados na subcamada Dados físicos nas bases de dados do BI, Marques utiliza o Sistema de Gerenciamento de Banco de Dados Relacional MySQL, devido ao suporte a diversos tipos de índices e a rapidez na carga dos dados.

A Figura 3.5 apresenta a implementação do MCM proposto por Marques [4]. Este modelo foi implementado para armazenar os dados sociais dos cidadãos.

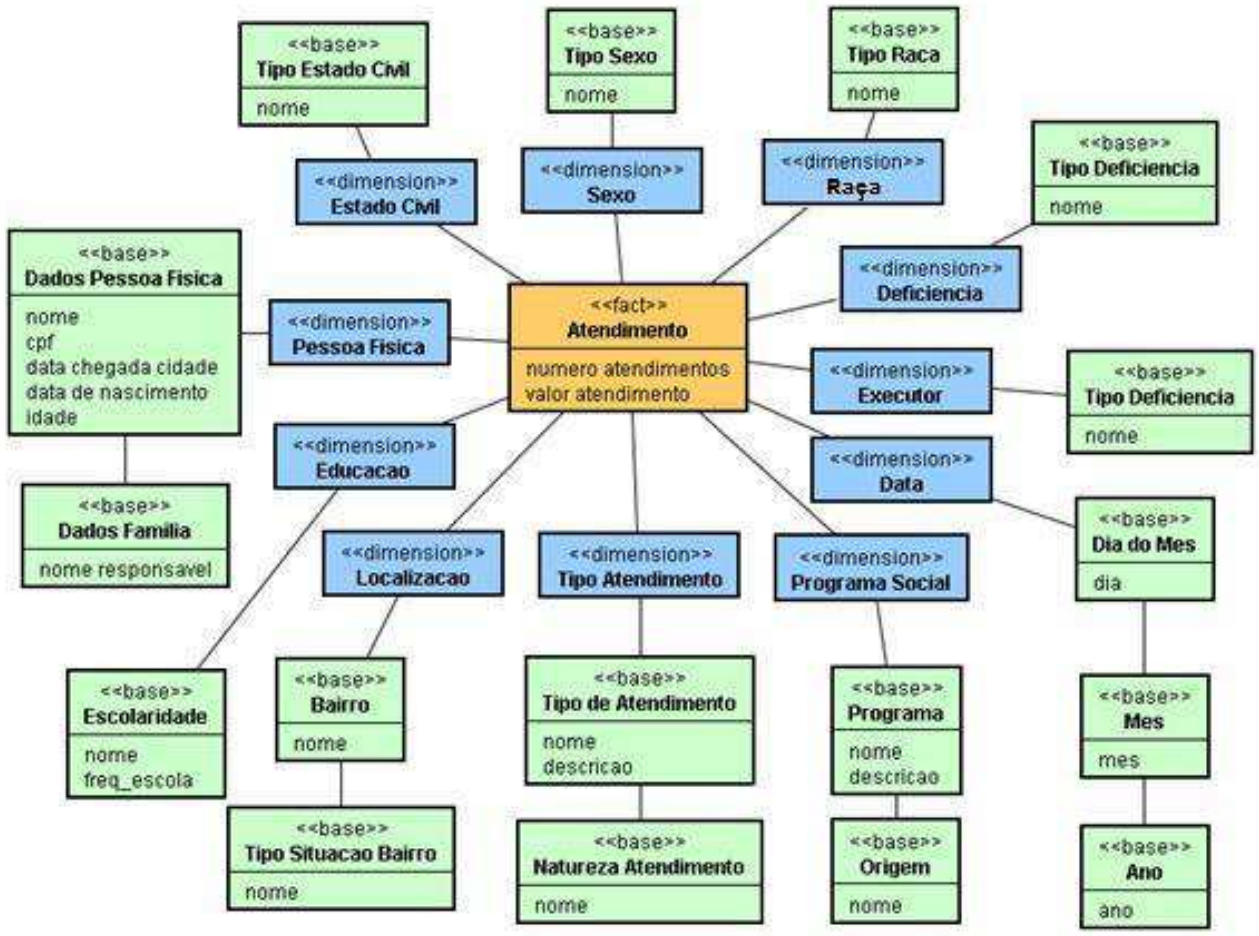


Figura 3.5 - Modelo Conceitual Multidimensional [4].

Já para a subcamada Visão lógica dos dados do BI, foi adotado o servidor OLAP Mondrian, que permite a execução de consultas multidimensionais em uma base de dados relacional. Juntamente com o servidor, a ferramenta Mondrian Schema Workbench é disponibilizada para auxiliar o mapeamento multidimensional dos dados relacionais, facilitando a confecção dos arquivos de mapeamento no formato XML [4].

3.1.3 CAMADA DE APLICAÇÕES PARA OS USUÁRIOS FINAIS

Esta camada tem como objetivo disponibilizar soluções que permitam aos usuários analisar intuitivamente os dados disponíveis no ambiente de BI através de visões pré-definidas, a ferramenta OpenI foi selecionada para este propósito. A ferramenta OpenI permite aos usuários consultar os dados do BI através de uma aplicação web, onde os resultados são apresentados no formato de tabelas multidimensionais ou gráficos. As consultas podem ser salvas e posteriormente acessadas e editadas [4]. As transformações e a carga dos dados no MCM, corresponde a Fase 1 do PEC.

3.2 Modelo Genérico para Representação de Amostras e Extração de Conhecimento

O Modelo Genérico para Representação de Amostras e Extração de Conhecimento (MGRAEC) oferece um ambiente centralizado capaz de armazenar um grande volume de dados, compostos por um número indeterminado de dimensões e valores. O MGRAEC é composto por um conjunto de entidades que se dividem em armazenar os dados que serão minerados, e resumir o conhecimento obtido pela mineração em uma estrutura do tipo árvore.

A Figura 3.6 apresenta o Modelo Genérico para Representação de Amostras e Extração de Conhecimento:

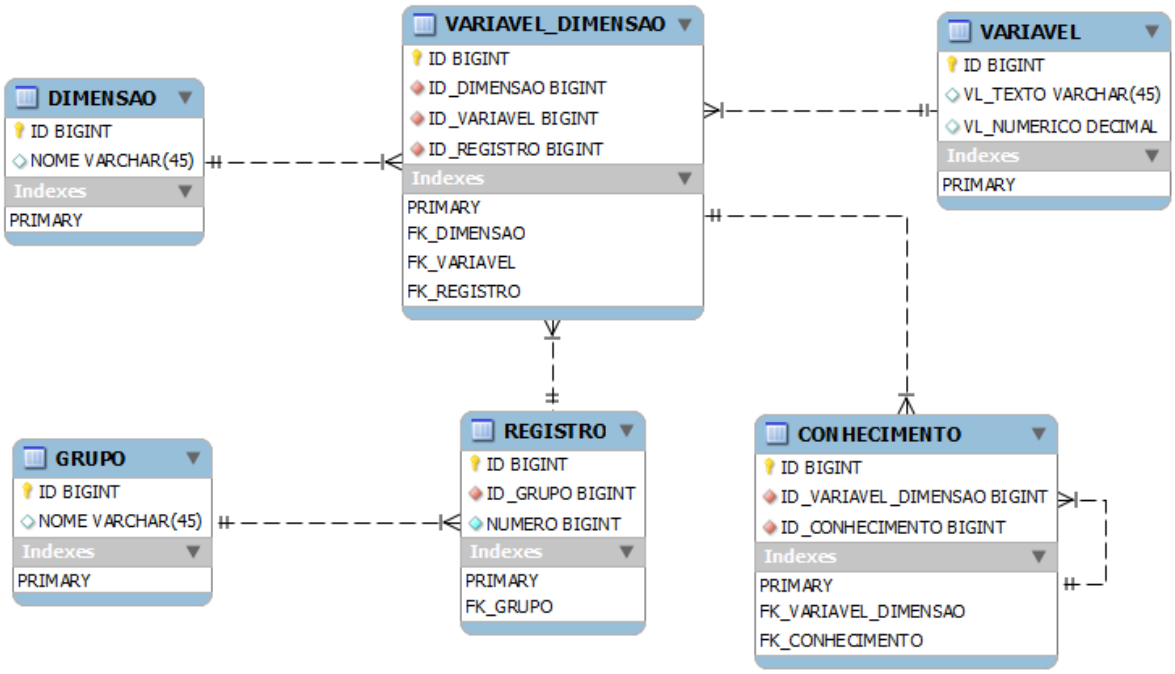


Figura 3.6 - Modelo Genérico para Representação de Amostras e Extração de Conhecimento (MGRAEC).

Na Figura 3.6 é possível notar seis entidades: DIMENSAO, VARIAVEL_DIMENSAO, VARIAVEL, GRUPO, REGISTRO e CONHECIMENTO. A entidade DIMENSAO armazena as colunas de uma amostra, identificadas de maneira única, enquanto a entidade VARIAVEL armazena os diferentes valores que cada dimensão pode assumir. Nesta última entidade, juntamente com o valor descritivo é armazenado uma constante numérica, que será utilizada pelo mapa auto-organizável durante a mineração de dados. Esta constante numérica será utilizada para calcular a similaridade entre regiões topológicas do mapa auto-organizável e os dados de entrada. A entidade VARIAVEL_DIMENSAO relaciona uma dimensão com um valor da entidade VARIAVEL, onde a dimensão e a variável pertencem à mesma entrada. Cada entrada possui um registro na entidade REGISTRO, que também se relaciona com a entidade VARIAVEL_DIMENSAO.

As entidades DIMENSAO, VARIAVEL, REGISTRO e VARIAVEL_DIMENSAO são as Entidades para Representação de Amostras (ERA), responsáveis por armazenar todos os dados a serem submetidos ao processo de mineração de dados através do mapa auto-organizável. A

Tabela 3.1 apresenta como um conjunto de dados de entrada pode ser representado através das ERA:

Tabela 3.1 - Representação dos dados nas ERA.

Númerador do Registro	*Sexo*	*Faixa Etária*
1	Masculino	Adulto
2	Masculino	Criança
3	Feminino	Jovem

A Tabela 3.1 apresenta três registros identificados de forma única através do Numerador do Registro. Note que no exemplo apresentado pela Tabela 3.1 existem duas dimensões: Sexo e Faixa Etária. Cada dimensão pode assumir valores distintos, a dimensão Sexo assume os valores: Masculino e Feminino, já a dimensão Faixa Etária assume os valores: Adulto, Criança e Jovem.

A Figura 3.7 ilustra a forma de preenchimento das ERA e como os dados de entrada se relacionam através de suas dimensões e variáveis, através da entidade VARIAVEL_DIMENSAO:

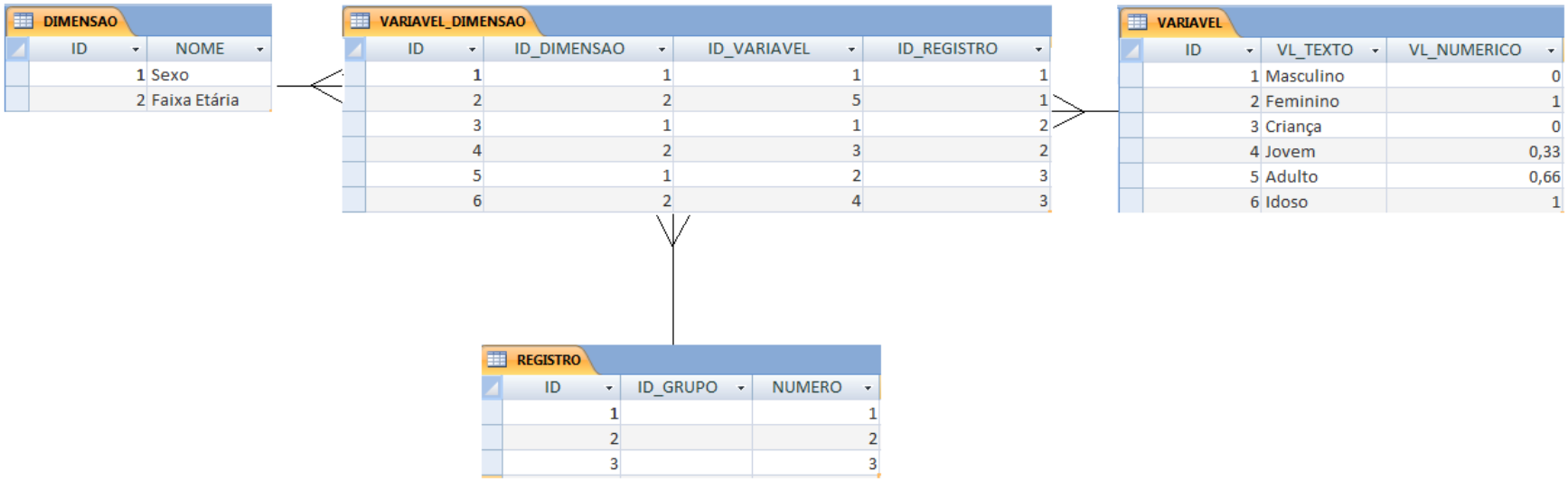


Figura 3.7 - Exemplo de preenchimento das Entidades para Representação de Amostras (ERA).

No exemplo apresentado pela Figura 3.7, a entidade DIMENSAO armazena as colunas da amostra, Sexo e Faixa Etária, enquanto a entidade VARIAVEL armazena os possíveis valores para cada coluna, associando a uma constante numérica. Para a coluna Sexo temos os valores: Masculino e Feminino, já a coluna Faixa Etária pode ter os valores: Criança, Jovem, Adulto e Idoso. A entidade VARIAVEL_DIMENSAO relaciona as dimensões e as variáveis, além de indicar o REGISTRO ao qual estas dimensões e variáveis pertencem. A coluna VL_NUMERICO da entidade VARIAVEL deve receber uma constante numérica que represente o dado. Esta informação corresponde ao dado de entrada, que será utilizado para calcular a similaridade com o vetor de pesos dos neurônios do mapa auto-organizável. As ERA devem ser preenchidas com os dados trazidos do MCM. Este preenchimento é realizado na Fase 2 do PEC.

Note que a coluna ID_GRUPO, da entidade REGISTRO está vazia. Os registros apenas serão associados aos seus devidos agrupamentos após o termino da Fase 3, depois de ter executado a mineração de dados. Na Fase 2 não é possível saber a qual grupo cada registro pertence.

A Tabela 3.2 apresenta a classificação numérica atribuída para os dados do exemplo ilustrado pela Figura 3.7:

Tabela 3.2 - Classificação Numérica.

Dimensão	*Dado*	*Valor Numérico Associado*
Sexo	Masculino	0
	Feminino	1
Faixa Etária	Criança	0
	Jovem	0,33
	Adulto	0,66
	Idoso	1

Na Tabela 3.2, os dados de cada dimensão possuem valores numéricos associados. Podemos observar que os valores são normalizados no intervalo aberto entre 0 e 1. Esta

normalização é necessária, pois a função de ativação utilizada pelo mapa auto-organizável tem melhor convergência quando os valores estão neste intervalo [9].

A Tabela 3.3 apresenta os dados que devem ser fornecidos ao mapa auto-organizável, nela os rótulos descritivos das dimensões Sexo e Faixa Etária foram substituídos pelos valores numéricos correspondentes. Desta forma, serão fornecidos ao mapa auto-organizável os valores numéricos associados a cada registro, que correspondem aos dados de entrada da rede neural. Estes dados serão utilizados para calcular a similaridade com os vetores de pesos de cada neurônio do mapa auto-organizável.

Tabela 3.3 - Entradas para o Mapa Auto-Organizável.

Registro	*Sexo*	*Faixa Etária*
1	0	0,66
2	0	0
3	1	0,33

3.2.1 PREENCHIMENTO DAS ENTIDADES DE REPRESENTAÇÃO DE AMOSTRAS

As ERA ilustradas pela Figura 3.7, serão preenchidas com os dados armazenados no MCM. Os dados serão extraídos do modelo conceitual através de uma rotina de conversão, e gravados nas ERA. Esta tarefa é realizada na Fase 2 do PEC.

A Figura 3.8 apresenta a extração dos dados do MCM e o preenchimento das Entidades para Representação de Amostras (ERA):

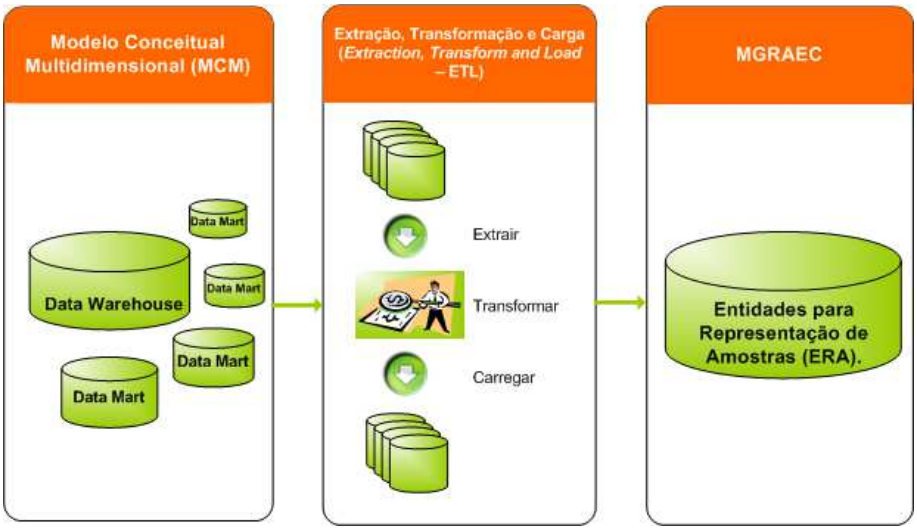


Figura 3.8 – Extração dos dados do MCM e armazenamento no MGRAEC.

O processamento ETL que lê os dados do MCM e armazena estes dados nas ERA, deve reduzir o número de variáveis de algumas dimensões. Esta redução aumentará a similaridade entre os registros e facilitará a convergência do mapa auto-organizável.

Nem todas as dimensões precisam ter suas variáveis reduzidas. Esta redução deve ser aplicada apenas nas dimensões que podem dificultar a separação dos dados em grandes agrupamentos. A dimensão Sexo, por exemplo, pode variar entre Masculino e Feminino. Como são poucas variações, esta dimensão não precisa ter suas variáveis reduzidas. Já a dimensão Idade pode ter um número muito grande de variáveis, dificultando a identificação de agrupamentos. Neste caso, os resultados serão mais satisfatórios se reduzirmos todas estas variáveis para valores mais abrangentes, por exemplo: Criança, Jovem, Adulto e Idoso.

Esta redução permitirá ao mapa auto-organizável formar agrupamentos maiores, permitindo a identificação de padrões. Posteriormente, se houver a necessidade de averiguar com detalhes a idade de cada pessoa, relatórios específicos podem ser extraídos a partir das bases de dados transacionais, utilizando os padrões encontrados pelo mapa auto-organizável como filtros de consulta. A Figura 3.9 apresenta o algoritmo para carga dos dados nas ERA:

```
1. Ler os dados de todas as dimensões do MCM;
2. Para cada dimensão
      a. Verificar se é necessário efetuar redução
           i. Efetuar redução para valores especificados;
      b. Armazenar dados nas Entidades para Representação de Amostras
         (ERA);
```

Figura 3.9 - Algoritmo para carga dos dados nas ERA.

A Figura 3.9 mostra o algoritmo ETL executado na Fase 2 do PEC, que lê os dados das dimensões do MCM, efetua as reduções e armazena estes dados nas ERA. Estas novas variáveis devem incorporar faixas de valores existentes na dimensão a ser reduzida, por exemplo, a dimensão Idade pode assumir a seguinte configuração:

Tabela 3.4 - Exemplo de redução para a dimensão Idade.

Faixa de valor	*Nova variável*
0 a 12 anos	Criança
13 a 20 anos	Jovem
21 a 60	Adulto
60 anos ou mais	Idoso

Após reduzir as variáveis da dimensão, os dados devem ser gravados nas Entidades para Representação de Amostras, como ilustra a Figura 3.7.

3.2.2 ANÁLISE EXPLORATÓRIA AUTOMÁTICA COM MAPAS AUTO-ORGANIZÁVEIS.

A mineração dos dados será realizada através de um mapa auto-organizável, devido a sua capacidade de classificação não supervisionada. A escolha desta técnica se deve ao fato de não conhecermos previamente os dados a serem minerados, de forma que não é possível especificar um conjunto de treinamento que compreenda todas as possíveis classes de objetos existentes nos dados. Os dados armazenados nas ERA devem ser submetidos ao mapa auto-organizável, dando início a Fase 3 do PEC.

A análise exploratória não supervisionada de dados consiste na busca de padrões nos dados amostrais, a partir de técnicas de análise de agrupamentos. Quanto maior o volume de dados amostrais melhor será a representação realizada pela análise exploratória. A análise exploratória através de mapas auto-organizáveis é divida em dois estágios, que compreendem:

- ✓ Parametrização da rede neural e escolha da topologia do mapa auto-organizável;
- ✓ Interpretação dos resultados obtidos pelo mapa auto-organizável;

Os estágios mencionados são relevantes para a geração de resultados confiáveis, destacando que nesta fase, o conjunto de dados a ser analisado deve estar armazenado nas ERA do MGRAEC.

A parametrização da rede neural e a escolha da topologia do mapa auto-organizável compreendem a definição de algumas variáveis como: raio inicial da função de vizinhança; número de épocas para o processo de aprendizagem; valor inicial para o passo adaptativo (taxa de aprendizagem) e o número de neurônios existentes no mapa auto-organizável. A escolha destes parâmetros é um processo empírico, onde o objetivo é chegar a um ponto de convergência com o menor número de neurônios possível, minimizando o tempo de processamento.

O ponto de convergência é atingido quando a configuração do mapa auto-organizável não sofre mudanças significativas de uma época para outra. Isso ocorre porque os vetores de pesos sinápticos atingiram os mínimos locais da função a ser representada [34]. A escolha do número de neurônios também é um processo empírico, ou seja, poucos neurônios podem não representar todos os agrupamentos existentes nos dados, enquanto um número excessivo de neurônios pode ser custoso computacionalmente. O número apropriado de neurônios é aquele que represente todos os agrupamentos existentes nos dados com o menor número de unidades no mapa auto-organizável.

Os resultados obtidos pelo processo de mineração de dados podem ser representados pela Matriz-U, porém esta representação indica apenas o número de agrupamentos existentes nos dados. Para compreender as características dos registros correspondentes a cada agrupamento, é necessário avaliar os resultados analíticos produzidos pelo mapa auto-

organizável. Após o término da mineração de dados, estes resultados poderão ser visualizados, através da relação estabelecida entre a entidade GRUPO e a entidade REGISTRO. Esta relação permite identificar o agrupamento ao qual cada registro pertence.

A Figura 3.10 apresenta o algoritmo de mineração de dados, através de mapas auto-organizáveis:

```
1. Inicializar pesos dos neurônios;
2. Definir o tamanho da rede;
3. Para cada valor de entrada
      a. Verificar se a rede não atingiu um ponto de convergência ou se
         não atingiu o limite de iterações
            i.  Identificar o neurônio BMU;
           ii.  Calcular o raio de vizinhança;
          iii.  Adaptar o neurônio BMU e seus vizinhos em direção ao dado
                de entrada;
           iv.  Reduzir o raio de vizinhança e o passo adaptativo;
4. Para cada agrupamento encontrado
      a. Inserir um registro na entidade GRUPO, e vincular os dados da
         entidade REGISTRO aos seus devidos agrupamentos;
```

Figura 3.10 - Algoritmo de mineração de dados, através de mapas auto-organizáveis.

Os pesos sinápticos presentes entre a camada de entrada e os neurônios da camada de saída, são inicializados aleatoriamente. O tamanho da rede é determinado pelo número de neurônios existentes na camada de saída. Quanto menor for este número, melhor será o desempenho do algoritmo.

Enquanto a rede neural não atingir um ponto de convergência ou um número limite de épocas, os dados devem ser apresentados a rede. Para cada dado apresentado devem ser realizados quatro procedimentos, são eles:

- ✓ Identificar o neurônio BMU. Este neurônio é aquele que possui menor distância, considerando o valor de seus pesos sinápticos e o dado de entrada. Uma vez encontrado o neurônio BMU, é preciso calcular o raio de vizinhança.

- ✓ Deve ser calculado o raio de vizinhança do neurônio BMU. A função Gaussiana atende aos requisitos para ajuste do neurônio BMU e seus vizinhos.
- ✓ Uma vez encontrado o neurônio BMU e calculado o raio de vizinhança, é preciso aplicar as adaptações ao neurônio BMU e seus vizinhos. Quanto mais distante o neurônio estiver, em relação ao BMU, menor será sua adaptação em direção ao dado de entrada.
- ✓ A cada iteração, é preciso reduzir o raio de vizinhança e o passo adaptativo. Estes valores devem ser reduzidos até que atinjam um valor mínimo.

Após o mapa auto-organizável atingir um ponto de convergência ou o número limite de épocas, será criado um registro na entidade *GRUPO*, do MGRAEC para cada agrupamento encontrado pela mineração de dados. Os registros da entidade *GRUPO* estarão vinculados aos dados da entidade *REGISTRO*, permitindo identificar a qual agrupamento cada dado de entrada pertence. Isto encerra a Fase 3 do PEC.

A Figura 3.11 apresenta o diagrama de classes que compreende os métodos e atributos pertencentes ao algoritmo de mineração de dados, ilustrado na Figura 3.10.

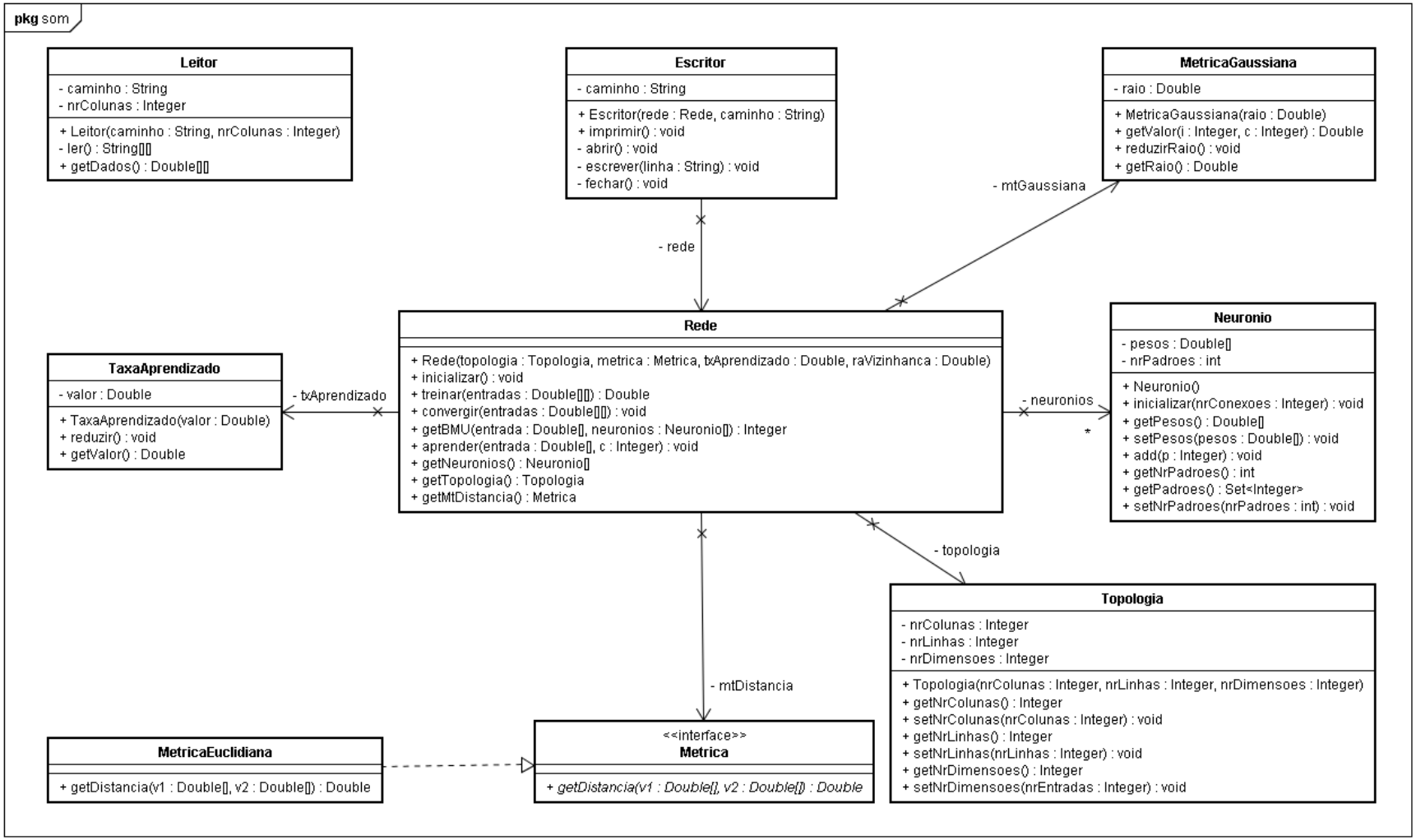


Figura 3.11 - Diagrama de classes do mapa auto-organizável.

A classe Leitor obtém as informações armazenadas nas ERA e fornece os dados ao mapa auto-organizável. A classe Escritor preenche a entidade GRUPO com os agrupamentos encontrados nos dados de entrada. A classe Escritor também relaciona todos os registros das entidades GRUPO com a entidade REGISTRO, permitindo identificar o agrupamento ao qual cada entrada corresponde.

A classe Rede instancia os elementos de processamento do mapa auto-organizável, tais como: Taxa Aprendizado; Métrica Gaussiana; Conjunto de Neurônios; Métrica de Distância Euclidiana e a Topologia da Rede. Cada elemento controla suas próprias características, como é o caso da redução do raio de vizinhança, na Métrica Gaussiana e a redução da Taxa de Aprendizado. O grau de adaptação dos pesos sinápticos é realizado pelo método aprender(), da classe Rede. A inicialização dos pesos sinápticos dos neurônios é realizada através da invocação do método inicializar(), da classe Neuronio. Já o neurônio que mais se aproxima de uma entrada fornecida à rede, é obtido através do método getBMU(), da classe Rede. O algoritmo representado pela Figura 3.10, que integra todos os métodos a fim de realizar a mineração de dados, é implementado pelo método convergir(), da classe Rede.

A Figura 3.12 representa um diagrama de sequência que indica a ordem de execução das tarefas, desde a leitura dos dados nas ERA, passando pelo processo de convergência dos dados, até a representação dos agrupamentos encontrados na entidade GRUPO.

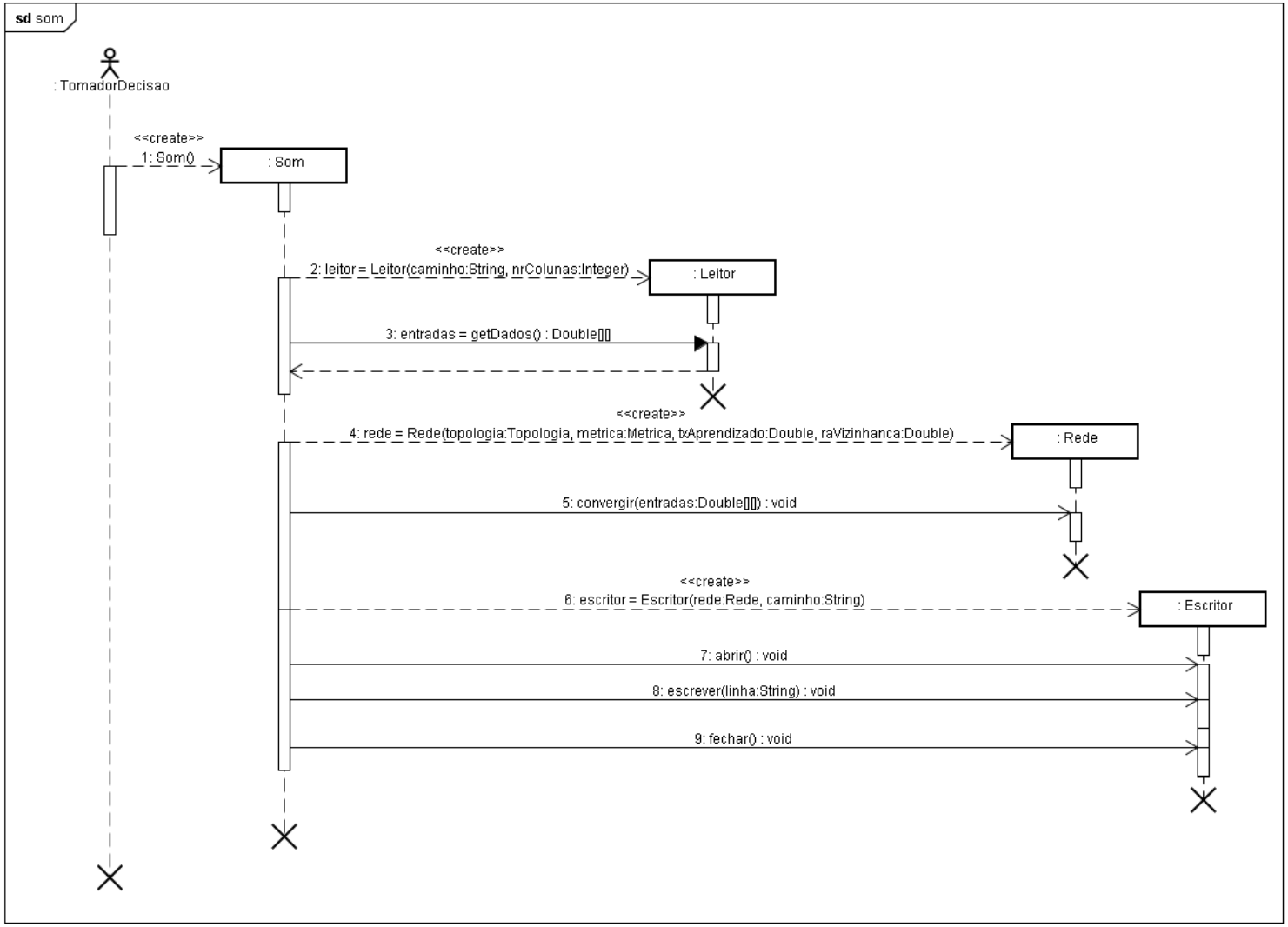


Figura 3.12 - Diagrama de sequência do algoritmo de mineração de dados.

Na Figura 3.12, a classe Leitor obtém no formato de matriz, os dados armazenados na ERA, por meio do método getDados(). Após obter os dados que servirão como entrada para o mapa auto-organizável é necessário instanciar a classe Rede, na qual indicamos: a Topologia, a Métrica de Distância a ser utilizada, a Taxa de Aprendizado e o Raio de Vizinhança inicial.

3.2.3 ROTINA DE BALANCEAMENTO E CARGA

Após concluir a mineração de dados, armazenar os agrupamentos na entidade GRUPO e relacionar cada agrupamento aos registros correspondentes, todo o conhecimento obtido pela mineração de dados está armazenado no MGRAEC. Embora o conhecimento esteja armazenado, consultar estas informações pode ser uma tarefa custosa, do ponto de vista computacional, tendo em vista que o conjunto de dados armazenados no MGRAEC pode ser muito grande. Surge então a necessidade de resumir as características dos agrupamentos e armazenar este resumo em uma estrutura organizada e de fácil acesso. Desta forma, outras aplicações do governo eletrônico podem utilizar o conhecimento obtido pela mineração de dados para tomar decisões.

Para resumir as características dos agrupamentos, é importante manter uma relação de dependência entre os dados. Desta forma é possível representar os agrupamentos sem perder o significado das informações. As estruturas do tipo árvore são conhecidas por sua capacidade representativa e facilidade de acesso, porém o desempenho durante o acesso a estas estruturas está diretamente relacionado ao balanceamento dos dados representados por ela. Os dados devem estar distribuídos de modo que a árvore de informações não cresça indiscriminadamente de um único lado, pois se isso ocorrer o desempenho de acesso será equivalente ao de uma lista, e não mais de uma árvore. O resumo dos resultados produzidos pela mineração de dados, e a organização destes resultados em uma estrutura do tipo árvore, são realizados na Fase 4 do PEC.

A seguir, um exemplo dos dados organizados em uma estrutura desbalanceada.

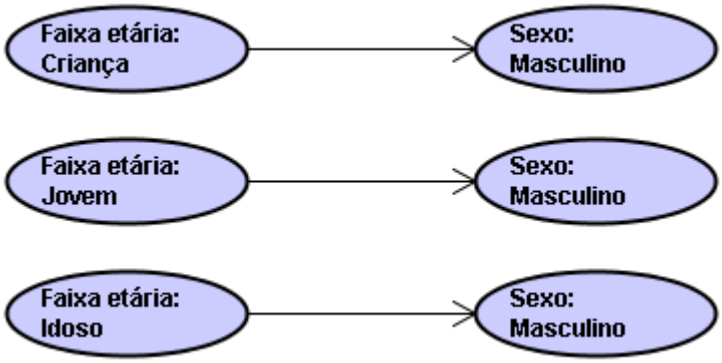


Figura 3.13 - Exemplo de dados organizados em uma estrutura desbalanceada.

Note que os dados estão dispostos em estruturas do tipo lista, e a dimensão Sexo é repetida para cada Faixa etária, ou seja, apresenta uma redundância de informações. Para garantir o balanceamento correto dos dados em uma estrutura do tipo árvore, é necessário definir como nó inicial aquele que possui o menor número de variações, como mostra a Figura 3.14:

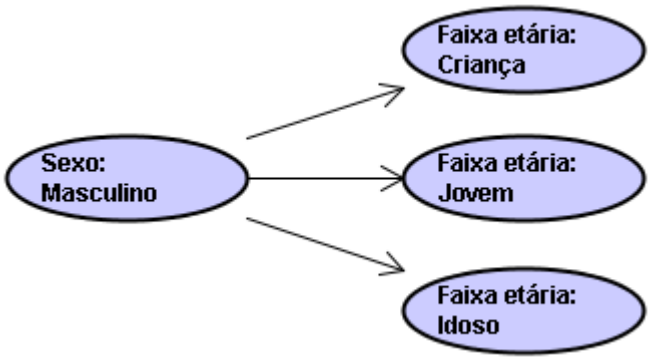


Figura 3.14 – Exemplo de dados organizados em uma estrutura do tipo árvore balanceada.

O exemplo ilustrado pelo Figura 3.14 apresenta três agrupamentos. É possível notar que os dados estão dispostos em uma estrutura do tipo árvore, e organizados de forma balanceada. O balanceamento dos dados é uma forma de organizar as informações e evitar redundâncias. Note que a dimensão Sexo possui apenas uma variação, enquanto a dimensão Faixa etária possui três.

A entidade CONHECIMENTO possui um auto-relacionamento que caracteriza uma estrutura do tipo árvore. Após o término da mineração de dados, daremos início a Fase 4 do PEC. Nesta fase, uma rotina de balanceamento e carga deve resumir as características dos agrupamentos e armazenar este resumo na entidade CONHECIMENTO. A Figura 3.15 ilustra a rotina de balanceamento e carga dos dados.

```
1. Obter todos os agrupamentos encontrados pelo mapa auto-organizável;
2. Ordenar os dados pelas dimensões que possuem menor número de
   variáveis
3. Para cada agrupamento
      a. Obter todos os registros associados ao grupo;
      b. Para cada registro
            i. Obter todas as dimensões e suas variáveis;
           ii. Para cada dimensão
                  1. Se for o primeiro
                        a. Armazenar na entidade conhecimento sem
                           vínculo;
                  2. Caso contrário
                        a. Armazenar na entidade conhecimento com vínculo
                           ao registro anterior;
```

Figura 3.15 - Rotina de Balanceamento e Carga.

Na Figura 3.15, para carregar os dados na entidade CONHECIMENTO, as demais entidades do MGRAEC devem estar preenchidas. O processo de carga da entidade CONHECIMENTO inicia com a obtenção de todos os registros da entidade GRUPO. Estes registros devem ser ordenados pelas dimensões que possuem menor número de variáveis. A dimensão que menos variar deve ser a raiz da árvore.

Observamos também, que o primeiro registro não possui vínculo com outros registros. Os demais registros são relacionados com o registro anterior, caracterizando uma estrutura do tipo árvore.

Capítulo 4

Estudo de Caso

ESTE capítulo apresenta os resultados obtidos em um estudo de caso realizado sobre os dados de atendimentos aos beneficiados por programas sociais, da Prefeitura Municipal de Campinas, SP – Brasil. Os dados utilizados durante os experimentos vêm do módulo de Gestão Social do Sistema Integrado de Governança Municipal (SIGM).

4.1 Origem dos Dados Operacionais

O Governo Federal Brasileiro mantêm o controle de atendimento aos beneficiados pelos programas sociais através de um instrumento de coleta de dados, com o propósito de caracterizar a situação das famílias a partir do Índice de Desenvolvimento Familiar (IDF) [39]. O instrumento utilizado para coletar os dados dos munícipes é o Cadastro Único, regulamentado pelo decreto nº 6.135, de 26 de Junho de 2007 e pela Portaria 376, de 16 de Outubro de 2008. Periodicamente os dados coletados pelo Cadastro Único são encaminhados para a Caixa Econômica Federal (CEF), em arquivos com formato previamente estabelecido. A CEF processa os registros e atribuir a cada pessoa o Número de Inscrição Social (NIS), de caráter único, pessoal e intransferível [40].

O Cadastro único não traz funcionalidades voltadas à gestão dos dados coletados. Logo as instituições públicas buscam soluções complementares que proporcionam maior eficiência na gestão dos dados operacionais, permitindo a visualização de relatórios e gráficos gerenciais referente aos atendimentos sociais. O Sistema Integrado de Governança Municipal (SIGM) é um exemplo destas soluções.

O SIGM é um aplicativo voltado para as necessidades de gestão das prefeituras, proporcionando mecanismos para o gerenciamento de todos os serviços, registros de cidadãos,

gerência de processos e dados relevantes para a administração do município. Este sistema é desenvolvido sobre uma arquitetura de múltiplas camadas, utilizando a tecnologia Enterprise Javabeans (EJB) [21] para a distribuição dos objetos de negócio, e o sistema de gerenciamento de banco de dados relacional ORACLE para o armazenamento dos dados [21]. O SIGM possui integração com os sistemas do Governo Federal Brasileiro, permitindo registrar os atendimentos dos beneficiados pelos programas sociais municipais, estaduais e federais.

O módulo do SIGM responsável por gerir os dados relativos a programas sociais é o módulo de Gestão Social. O estudo de caso utilizará como base o ambiente de BI construído no trabalho de Marques [4] para o módulo de Gestão Social do SIGM implantado, na Prefeitura Municipal de Campinas, SP.

A Figura 4.1 apresenta as fases do PEC e as tecnologias utilizadas na implementação de cada fase do processo:

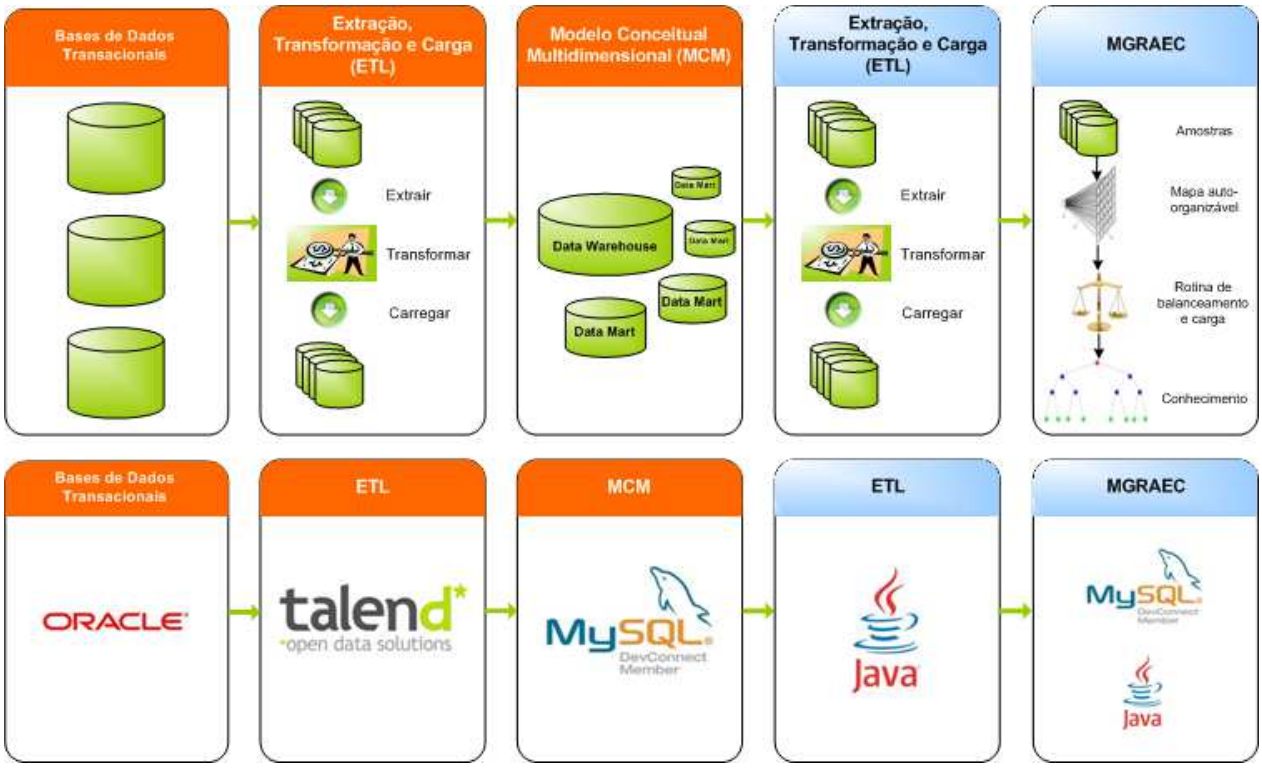


Figura 4.1 - Tecnologias utilizadas.

A Figura 4.1 apresenta as tecnologias utilizadas em cada fase do PEC. Para demonstrar a capacidade de integração entre soluções distintas, no estudo de caso foram utilizadas diferentes tecnologias em cada fase do processo.

As bases de dados transacionais do SIGM são desenvolvidas a partir do SGBD ORACLE. A tarefa ETL que lê os dados das bases de dados transacionais e armazena no MCM, é realizada a partir da ferramenta Talend. Já o MCM foi desenvolvido a partir do SGBD MySQL.

A segunda tarefa ETL foi realizada através de um software desenvolvido neste trabalho, utilizando a linguagem de programação Java. O mapa auto-organizável e a rotina de balanceamento e carga também foram desenvolvidos em Java. O MGRAEC foi desenvolvido a partir do SGBD My SQL.

O primeiro passo do estudo de caso é carregar os dados presentes nas bases de dados transacionais do SIGM para MCM por meio de rotinas ETL. Durante o processo de ETL os dados passaram por uma mudança de formato, de modo a serem armazenados no modelo multidimensional. Além da adequação de formato, os dados passaram por uma eliminação de duplicidades. Não houve necessidade de integrar valores, no que se diz respeito a nomenclaturas, visto que a única origem dos dados é a base transacional do SIGM. Esta tarefa foi realizada na Fase 1 do PEC, através da ferramenta Talend.

Cerca de 21.000 registros de atendimentos sociais foram carregados no MCM. Inicia-se então o segundo passo do processo, que é a aplicação de rotinas ETL para transportar os dados do MCM para as Entidades de Representação de Amostras (ERA) do MGRAEC. Dos 21.000 registros presentes no MCM, 1.621 foram carregados para o MGRAEC. Foram selecionados os registros que possuem as seguintes informações preenchidas:

- ✓ Sexo;
- ✓ Raça;
- ✓ Deficiente;
- ✓ Grau de instrução;
- ✓ Freqüenta escola;
- ✓ Tipo de benefício social;

- ✓ Região metropolitana;

As dimensões mencionadas foram selecionadas, pois são capazes de caracterizar um indivíduo sem comprometer sua privacidade. O preenchimento das ERA foi realizado na Fase 2 do PEC.

4.2 Carga dos Dados no MGRAEC

A Fase 2 do PEC foi realizada através da rotina de carga ilustrada pela Figura 3.9. Durante o processo de ETL, algumas dimensões tiveram seus valores agrupados em classes mais abrangentes, com o intuito de prover uma maior similaridade entre os dados. Portanto, foram desconsiderados os diferentes tipos de deficiência tornando esta uma dimensão booleana, ou seja, apenas indicando se a pessoa é deficiente ou não. Já o grau de instrução teve seus vários níveis agrupados em quatro classes: Nenhum, Baixo, Médio e Alto.

O tipo de benefício social também passou por uma redução de variáveis, acomodando os diferentes benefícios em cinco tipos. São eles: Transferência de renda, Benefício habitacional, Benefício sócio-educativo, Auxilio a criança e adolescente, Programas voltados ao público jovem. A redução para estas variáveis foi necessária para o processo de convergência, visto que o conjunto de dados selecionado dividia-se em 15 programas sociais.

A Tabela 4.1 apresenta os dados utilizados para o estudo de caso, e a redução de variáveis aplicada sobre estes:

Tabela 4.1 - Redução de variáveis aplicada aos dados.

Dimensão	**Variáveis**	**Variáveis reduzidas**
Sexo	Masculino	Masculino
	Feminino	Feminino
Raça	Amarela	Amarela
	Branca	Branca
	Parda	Parda
	Negra	Negra
Deficiente	Deficiência visual	Sim Não
	Deficiência auditiva	
	Mobilidade reduzida	
	Paralisia	

	Deficiência mental	
Grau de instrução	Analfabeto	Nenhum
	Ensino fundamental incompleto	Baixo
	Ensino fundamental completo	
	Ensino médio incompleto	
	Ensino médio completo	Médio
	Ensino superior incompleto	
	Ensino superior completo	Alto
	Pós graduação	
	Mestrado	
	Doutorado	
Freqüenta escola	Sim	Sim
	Não	Não
Tipo de benefício social	Bolsa família	Transferência de renda
	Renda cidadã	
	Programa de garantia de renda familiar mínima	
	Minha casa minha vida	Benefício habitacional
	Cadastro imobiliário	
	Serviço sócio-educativo de 0 a 6 anos	Benefício sócio-educativo
	Serviço sócio-educativo de 7 a 14 anos	
	Serviço sócio-educativo de 7 a 21 anos	
	Serviço para crianças de 0 a 5 anos	Auxílio a criança e adolescente
	Programa de acolhimento institucional e acolhimento familiar para crianças e adolescentes	
	Programa de erradicação do trabalho infantil	
	Programa de atenção e apoio à adolescente grávida	
	Ação jovem	Programas voltados ao público jovem
	Agente jovem	
	Protagonismo juvenil	
Região metropolitana	Central	Central
	Sudoeste	Sudoeste
	Noroeste	Noroeste
	Norte	Norte
	Sul	Sul
	Leste	Leste

Algumas dimensões passaram por uma redução do número de variáveis: *Deficiente, Grau de instrução* e *Tipo de benefício social*. Esta redução foi necessária para aumentar a similaridade entre os dados, proporcionando a mesma interpretação para registros parecidos.

4.3 Mineração de Dados pelo Mapa Auto-organizável

Após carregar os dados nas ERA, o próximo passo é fazer a análise exploratória utilizando o mapa auto-organizável. Esta tarefa é compreendida pela Fase 3 do PEC.

Durante esta etapa, foi possível notar que os parâmetros livres da rede neural e a escolha do número de neurônios influenciam diretamente no tempo de convergência dos resultados, algumas vezes, até mesmo nos resultados. Inicialmente o mapa auto-organizável foi definido com 841 unidades de processamento, ou seja, uma grade de 29 x 29 neurônios. Este número de neurônios é capaz de representar um número elevado de agrupamentos, levando em consideração a quantidade de dados disponíveis.

A análise exploratória levou 15 horas para atingir um ponto de convergência. No final, foi possível identificar 5 agrupamentos de dados. Devido ao tempo de convergência, o número de neurônios foi reduzido para 64 unidades, ou seja, 8 x 8, no intuito de obter o mesmo resultado com um número menor de neurônios. Como esperado, o tempo de convergência foi reduzido para aproximadamente 90 minutos, porém quatro agrupamentos ficaram evidentes e o quinto grupo foi incorporado aos demais. Em uma terceira tentativa de obter cinco agrupamentos com um número reduzido de neurônios, foi definida uma grade de 100 unidades de processamento, ou seja, 10 x 10 neurônios. Nesta última configuração foi possível obter os mesmos cinco agrupamentos, resultantes da primeira execução, com menos unidades de processamento, reduzindo o tempo de convergência para 3 horas.

A Tabela 4.2 apresenta as diferentes configurações testadas e as características de cada uma delas.

Tabela 4.2 - Características de convergência.

Tamanho do Mapa	**Tempo de Convergência**	**Nr. de Agrupamentos**
29 x 29 Neurônios	15 horas	5
08 x 08 Neurônios	1,5 horas	4
10 x 10 Neurônios	3 horas	5

Para que houvesse convergência dos resultados, a rede neural teve que passar por várias épocas de aprendizado, ou seja, o conjunto de dados foi apresentado à rede diversas vezes. Foi possível notar que ao modificar o passo de adaptação da rede neural, ou seja, a taxa de aprendizado da rede, o número de épocas necessárias para a convergência era diferente.

A Tabela 4.3 apresenta o número aproximado de épocas de treinamento, de acordo com a taxa de aprendizado da rede neural.

Tabela 4.3 - Épocas de Treinamento da Rede Neural Artificial.

Tamanho do Mapa	Taxa de Aprendizagem	Nr. de Épocas
10 x 10	0.9	2300
10 x 10	0.3	1000
10 x 10	0.6	700

Para a realização dos testes foi estabelecido um raio de vizinhança inicial tão grande quanto o tamanho do mapa, de modo de a ser reduzido gradativamente durante o processamento de cada época de aprendizado da rede, até atingir um ponto de estabilidade.

A Figura 4.2 apresenta a Matriz-U que ilustra os 100 neurônios do mapa auto-organizável e os agrupamentos encontrados:

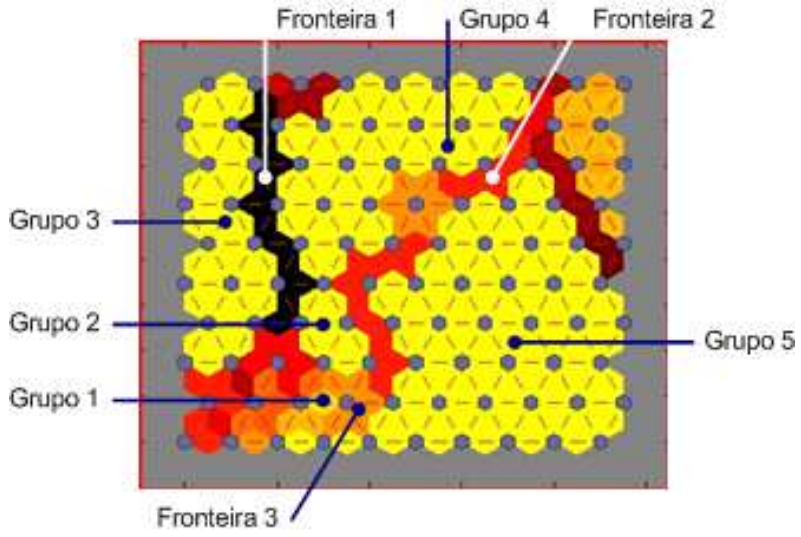


Figura 4.2 - Representação dos agrupamentos pela Matriz-U.

A Figura 4.2 é uma representação gráfica dos agrupamentos encontrados pelo processo de mineração de dados. Nesta figura, os dados são divididos em cinco grupos diferentes e entre um agrupamento e outro existe uma delimitação por fronteiras, de maior ou menor intensidade. As fronteiras mais escuras *Fronteira1* e *Fronteira2* são aquelas que representam menor similaridade entre os grupos. Sendo assim a fronteira mais clara *Fronteira3* representa maior similaridade entre os agrupamentos. É possível observar a existência de um pequeno agrupamento *Grupo1* composto de apenas um hexágono, cercado por fronteiras mais claras, na parte inferior esquerda da imagem. Este agrupamento pouco difere do agrupamento *Grupo5*, e do agrupamento *Grupo2*. Importante destacar que existe uma menor similaridade entre os agrupamentos *Grupo5* e *Grupo2*, delimitados pela fronteira mais escura *Fronteira2* entre eles. Outros dois agrupamentos, *Grupo3* e *Grupo4* podem ser observados na parte superior da Figura 4.2. A *Fronteira1* entre eles indica que existe pouca similaridade entre os dois agrupamentos.

4.4 Rotina de Balanceamento e Carga

Após o término da mineração de dados os agrupamento estavam identificados na entidade *GRUPO*. Embora fosse possível estabelecer uma relação entre estes agrupamentos e os dados, o volume de informações dificultou a compreensão destas relações. As relações entre os agrupamentos e os dados foram mais bem compreendidas após resumir estes relacionamentos em uma estrutura do tipo árvore. A estrutura do tipo árvore é provida pela entidade *CONHECIMENTO*, e o resumo pôde ser obtido através da rotina de balanceamento e carga. A rotina de balanceamento e carga é compreendida pela Fase 4 do PEC. A Figura 3.15 ilustra esta rotina.

Os dados sumarizados pela rotina de balanceamento e carga mostram que as pessoas que demandam por programas voltados ao público jovem geralmente são mulheres. Estas mulheres se dividem em dois grupos, deficientes e não deficientes. O primeiro grupo possui um elevado grau de instrução e se localiza na região leste da cidade. Já o segundo grupo, possui um grau de instrução médio e se localiza na região central da cidade. Como mostra a Figura 4.3:

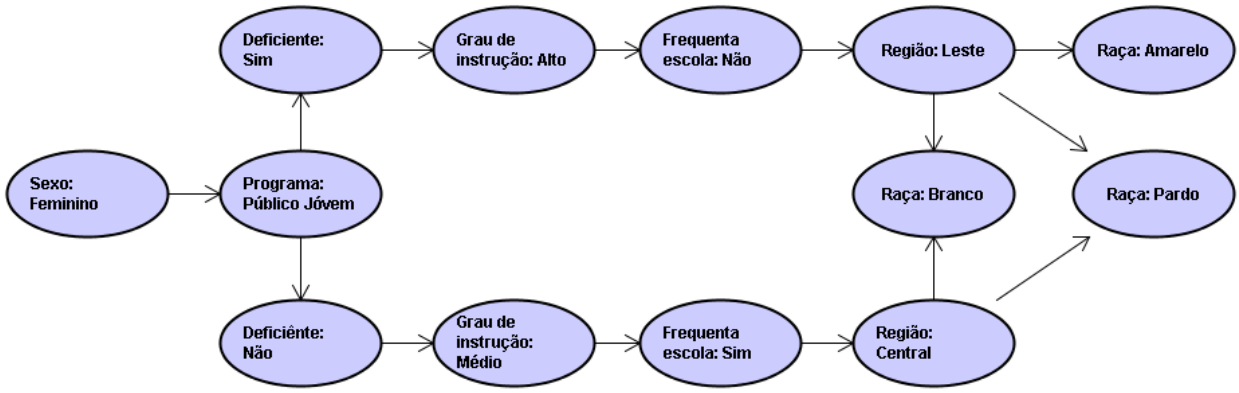


Figura 4.3 - Representação hierárquica do conhecimento.

A Tabela 4.4 apresenta o percentual de pessoas correspondente a cada agrupamento.

Tabela 4.4 - Representação quantitativa dos agrupamentos.

Representação quantitativa dos agrupamentos		
Grupo	**Descrição**	**Percentual**
1	Com deficiência, Grau de instrução alto, Região leste, Raça branca.	43,04%
2	Com deficiência, Grau de instrução alto, Região leste, Raça amarela.	1,27%
3	Com deficiência, Grau de instrução alto, Região leste, Raça parda.	6,33%
4	Sem deficiência, Grau de instrução médio, Região central, Raça branca.	22,78%
5	Sem deficiência, Grau de instrução médio, Região central, Raça parda.	26,58%

Na Tabela 4.4, os grupos 4 e 5 tratam de pessoas sem deficiência e são aproximadamente do mesmo tamanho. Já os grupos 1, 2 e 3 tratam de pessoas com deficiência.

Esta informação pode ser útil para as instituições municipais, pois permite que ações sejam tomadas no sentido de melhor atender estes beneficiados. O estudo de caso mostra que os pontos de atendimento social da região leste devem prover acessibilidade a deficientes, além de encaminhar os beneficiados mais qualificados para o mercado de trabalho, garantindo que outras pessoas sejam atendidas nesta região.

4.5 Discussão Técnica

A utilização de mapas auto-organizáveis como técnica de mineração mostrou que a escolha de algumas variáveis não tem solução definida, forçando o usuário a realizar baterias de testes até encontrar uma configuração de variáveis que proporcione um resultado satisfatório, em termos de resultado e tempo de processamento.

Durante os testes, notamos que o passo adaptativo quando iniciado em um valor alto, precisa de um número maior de iterações para convergir ao resultado esperado. Um valor muito baixo para o passo adaptativo resulta no mesmo cenário.

Outro fator que implica no tempo de convergência é o número de neurônios disponíveis no mapa auto-organizável. Um número muito grande de neurônios pode demorar a convergir, visto que as iterações são custosas, do ponto de vista computacional. No entanto, poucos neurônios podem não representar todas as características existentes nos dados.

No estudo de caso, inicialmente atribuímos muitos neurônios ao mapa auto-organizável, que demorou a convergir. Ao diminuir significativamente o número de neurônios, observamos uma convergência mais rápida, porém o número de agrupamentos encontrados foi menor, se comparado a mineração com mais neurônios. Frente a este cenário, estabelecemos um valor intermediário, que foi capaz de obter os resultados esperados em um tempo menor, se comparado ao primeiro teste.

Capítulo 5

Conclusão

NO decorrer deste trabalho, foi possível notar que a informação implícita nos dados operacionais das instituições é um elemento valioso para a tomada de decisão na gestão pública. A gestão do conhecimento e a mineração de dados permitem obter informações a partir dos dados operacionais, beneficiando as instituições e a sociedade.

Enquanto na esfera privada a gestão do conhecimento pode ser um diferencial competitivo, na esfera pública ela pode mudar a forma de interação do governo com a sociedade e com a economia, atendendo as verdadeiras necessidades dos cidadãos. Para tanto, o governo eletrônico deve possibilitar a interoperabilidade entre seus sistemas e realizar analises inteligentes sobre seus dados operacionais.

O mapa auto-organizável mostrou que é possível identificar padrões ocultos nos dados operacionais, mesmo sem um conjunto de treinamento. Estes padrões podem ser utilizados posteriormente em futuras investigações, sobre os mesmos dados operacionais. Por exemplo, podemos selecionar um dos agrupamentos encontrados pelo mapa auto-organizável, e avaliar o comportamento de variáveis que não foram consideradas durante a mineração de dados, como por exemplo, a renda per capta familiar.

Se houver uma semelhança na renda per capta familiar de todos os elementos daquele grupo, esta variável pode influenciar de alguma forma a situação na qual aquelas pessoas se encontram. De repente, uma mudança nesta variável pode contribuir para que as pessoas saiam daquela classificação.

Embora o mapa auto-organizável tenha agrupado os registros a partir das similaridades existentes nos dados, a visualização destes agrupamentos através da Matriz-U não é muito clara. A Matriz-U apresenta apenas a distribuição topológica dos grupos, mas não é capaz de mostrar

as características destes agrupamentos. Estas características são mais evidentes com o uso do MGRAEC, proposto neste trabalho, que permite associar cada registro ao seu devido grupo.

Mesmo que o MGRAEC associe os registros aos grupos, compreender as características destes agrupamentos pode ser difícil, quando há um grande volume de dados. Estas características podem ser compreendidas mais facilmente, após o resumo obtido pela rotina de balanceamento e carga. Esta rotina, também proposta neste trabalho, organiza os dados em uma estrutura do tipo árvore, que facilita a compreensão e permite uma boa percepção dos resultados, obtidos pela mineração de dados.

O estudo de caso mostrou que o PEC pode ser composto por diferentes soluções tecnológicas, uma vez que utilizamos ferramentas distintas em cada fase deste processo. No entanto, para que houvesse convergência dos dados foi necessário garantir a unicidade dos registros e a redução de algumas variáveis. Estas reduções proporcionaram uma mesma interpretação para registros parecidos.

Para que houvesse convergência dos dados, o mapa auto-organizável teve que se submeter a várias épocas de aprendizado, ou seja, o conjunto de dados foi apresentado diversas vezes à rede neural. Foi possível notar que passo adaptativo da rede pode influenciar no número de épocas necessárias para a convergência. Quando este passo é muito grande, o algoritmo pode ultrapassar o ponto de convergência da função e ter que voltar. No entanto, se este passo for pequeno demais o algoritmo pode demorar em atingir o ponto de convergência.

No estudo de caso realizado, os resultados mais satisfatórios foram obtidos quando iniciamos o passo adaptativo em um valor intermediário, e reduzimos gradativamente este passo após cada época da aprendizagem do mapa.

Novos trabalhos podem ser desenvolvidos, no sentido de identificar as variáveis correlatas a cada agrupamento obtido pelo PEC. Estas variáveis podem ter seus valores alterados, e através de projeções simular o comportamento dos dados operacionais. Isto permitiria prever os resultados de uma ação antes mesmo de executá-la.

Referências Bibliográficas

[1] Ebrahim, Z.; Irani, Z. E-government adoption: architecture and barriers. Business Process Management Journal, v. 11 n. 5, 2005.

[2] United Nations E-Government Survey From E-Government to Connected Governance, 2008.

[3] Mourady, A.; Elragal, A. Business Intelligence in Support of eGov Healthcare Decisions. European, Mediterraneam & Middle Eastern Conference on Information System, Athens, Greece, 2011.

[4] Marques, E. Z. Uma proposta de utilização das tecnologias de business intelligence para suporte à tomada de decisão no contexto de governo eletrônico. Dissertação de mestrado, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, 2011.

[5] Xu, L.; Zeng, L.; Shi, Z.; He, Q.; Wang, M. Research on Business Intelligence in Enterprise Computing Environment. ISIC. IEEE International Conference on Systems, Man and Cybernetics, 2007.

[6] Yan, P.; Guo, J. Researching and Designing the Architecture of E-government Based on SOA. Proceedings of the 2010 International Conference on E-Business and E-Government, 2010.

[7] Mohammed, A.; Goo, S. K. Government Increasingly Turning to Data Mining. The Washington Post, 2006. http://www.washingtonpost.com/wp-dyn/content/article/2006/06/14/AR2006061402063.html. Acessado em 09 de Janeiro de 2012.

[8] Campos, T. E. Técnicas de Seleção de Características com Aplicação em Reconhecimento de Faces. Dissertação de Mestrado. Universidade de São Paulo, 2001.

[9] Haykin, S. Redes Neurais: Princípios e Prática. 2. Ed. Porto Alegre: Bookman, 2001.

[10] Braga, C. V. Rede Neural e Regressão Linear: Comparativo entre Técnicas Aplicadas a um Caso Prático na Receita Federal. Dissertação de Mestrado. Faculdade de Economia e Finanças IBMEC, 2010.

[11] Oliveira, T. P. S. Sistemas Baseados em Conhecimento e Ferramentas Colaborativas para Gestão Pública: Uma proposta ao Planejamento Público Local. Dissertação de Mestrado. Universidade Federal de Santa Catarina, 2009.

[12] Kum, H.; Duncan, D. F.; Stewart, C. J. Supporting self-evaluation in local government via Knowledge Discovery and Data Mining. Government Information Quarterly, v. 26, issue 2, 2009.

[13] Aulich, C. From Citizen Participation to Participatory Governance in Australian Local Government. Commonwealth Journal of Local Governance, issue 2, 2009.

[14] Vilella, R. M. Conteúdo, Usabilidade e Funcionalidade: Três dimensões para avaliação de portais estaduais de Governo Eletrônico na WEB. Dissertação de Mestrado, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, 2003.

[15] Carromeu, C.; Paiva, D. M. B.; Cagnin, M. I.; Rubinsztejn, H. K. S.; Turine, M. A. S.; Breitman, K. Component-Based Architecture for e-Gov Web Systems Development. 17th IEEE International Conference and Workshops on the Engineering of Computer Based Systems, 2010.

[16] Pesquisa Nacional por Amostra de Domicilios. Disponível em: http://www.ibge.gov.br/home/estatistica/populacao/acessoainternet2008/default.shtm, Acesso em 29 de Outubro de 2011.

[17] Mendes, L. S.; Bottoli, M. L.; Breda, G. D. Digital cities and open MANs: A new communications paradigm, LATINCOM'09. IEEE Latin-American Conference on Communications, 2009.

[18] Ignatowicz, E. Criação de Modelos Organizacionais para Cidades Digitais Baseadas em uma Arquitetura Peer-to-Peer. Dissertação de Mestrado. Departamento de

Comunicação, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, 2009.

[19] da Silva, C. R. C.; Tavares, T. C.; Bicharra, A. C. Governo eletrônico em ambientes colaborativos virtuais. Instituto de Computação, Universidade Federal Fluminense, 2009.

[20] Isac, A.; Muntean, M.; Danaiata, D.; Soava, M. The Current Stage of the Development of G2B and B2G Eletronic Services in Romenia. Annals of University of Petrosani, 2010.

[21] Tilli, M.; Panhan, A. M.; Lima, O.; Mendes, L. S. A Web-based Architecture for e-Gov Application Development. ICE-B: Proceedings of the International Conference on e-Business, 2008.

[22] Purificação, M. C. S. Construção de uma solução de Business Intelligence como suporte à tomada de decisões gerenciais na UFBA. Trabalho de conclusão de curso, Instituto de Matemática, Departamento de Ciência da Computação, Universidade Federal da Bahia, 2009.

[23] Dayal, U.; Castellanos, M.; Simitsis, A.; Wilkinson, K. Data Integration Flows for Business Intelligence. ACM SIGMOD Record, v. 360 n. 1, 2009.

[24] Inmon, W. H. Como Construir o Data Warehouse, 2 ed. Rio de Janeiro, Campus, 1997.

[25] Inmon, W. H.; Hackarthorn, R. D. Como usar o data warehouse, Rio de Janeiro: IBPI Press, 1997.

[26] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. Artificial Intelligence Magazine, 1996.

[27] Chaki, N.; Sarkar, B. B. Virtual Data Warehouse Modeling Using Petri Nets for Distributed Decision Making. University of Calcutta, Kolkata, India, 2010.

[28] Uba, D. M.; Dutra, L. V. Seleção de Candidatos: Uma Estratégia para Incorporação da Distância de Mahalanobis no Algoritmo K-Médias. 7th Brazilian Conference on Dynamics, Control and Applications, Unesp at Presidente Prudente, SP, Brasil, 2008.

[29] Tarek, K. M.; Sofiane, K.; Farouk, B. Kohonen Maps Combined to K-means in a Tow Level Strategy for Time Series Clustering Application to Meteorological and Electricity Load Data. University Badji Mokhtar of Annaba, Algeria, 2010.

[30] Takahashi, A.; Bedregal, BRC.; Lyra, A. Uma versão intervalar do método de segmentação de imagens utilizando o k-means. Tendências em Matemática Aplicada e Computacional. vol. 6 nr. 2, 2011.

[31] Vesanto, J.; Alhoniemi, E. Clustering of the Self-Organizing Map. IEEE, Transactions on Neural Networks, v. 11, n. 3, 2000.

[32] Kohonen, T.; Kaski, S. Self-Organized Formation of Various Invariant-Feature Filters in the Adaptive-Subspace SOM. Neural Computation. v. 9, nº 6. p. 1321-1344, 1997.

[33] Jackson, J. Data Mining: A Conceptual Overview. Comunications of the Association for Information Systems, v. 8, p. 267-296, 2002.

[34] Kohonen, T. Self Organizing Maps. Berlin, Springer, 2001.

[35] Palote, V. A.; Pachghare, V. K.; Kulkarni, P. Self Organizing Maps to Build Intrusion Detection System. International Journal of Computer Applications, v. 1 Nr. 8, 2010.

[36] Zuchini, M. H. Aplicação de Mapas Auto Organizáveis em Mineração de Dados e Recuperação de Informação. Dissertação de Mestrado. Universidade Estadual de Campinas, 2003.

[37] Silva, M. A. S. Mapas Auto-Organizáveis na Análise Exploratória de Dados Geoespeciais Multivariados. Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, Instituto Nacional de Pesquisas Espaciais, 2005.

[38] Ultsch, A. Knowledge extraction from self-organizing neural networks. Opitz, O. ed. Information and Classification. Berlin: Springer, 1993.

[39] Ministério do Desenvolvimento Social http://www.mds.gov.br/programabolsafamilia/noticias/aplicativo-do-indice-de-desenvolvimento-da-familia-ja-esta-disponivel/, 2011. Acessado em 09 de Janeiro de 2012.

[40] Caixa Econômica Federal, http://www1.caixa.gov.br/gov/gov_social/estadual/distribuicao_servicos_cidadao/cadastramento_unico/saiba_mais.asp, 2011. Acessado em 09 de Janeiro de 2012.