Universidade Federal de Santa Catarina
Programa de Pós-Graduação em
Engenharia de Produção

UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS EM BASES DE C&T: UMA ANÁLISE DOS GRUPOS DE PESQUISA NO BRASIL

Alexandre Leopoldo Gonçalves

Dissertação apresentada ao
Programa de Pós-Graduação em
Engenharia de Produção da
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do título de Mestre em
Engenharia de Produção

Florianópolis
2000

Alexandre Leopoldo Gonçalves

UTILIZAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS EM BASES DE C&T: UMA ANÁLISE DOS GRUPOS DE PESQUISA NO BRASIL

Esta dissertação foi julgada e aprovada para a obtenção do título de **Mestre em Engenharia de Produção** no **Programa de Pós-Graduação em Engenharia de Produção** da Universidade Federal de Santa Catarina

Florianópolis, 14 de agosto de 2000.

Prof. Ricardo Miranda Barcia, Ph.D.
Coordenador do Curso

BANCA EXAMINADORA

_____

Roberto C. S. Pacheco, Dr.
Orientador

_____

José Leomar Todesco, Dr.

_____

Aran Bey Tcholakian Morales, Dr.

_____

Ricardo S. Lourenço, Esp.
Consultor CNPq

Agradecimentos

Ao Senhor pela força, iluminação, saúde e imensa proteção. Por ter me colocado em contato com pessoas fantásticas, e por me agraciar com a oportunidade de participar de algo tão belo e que por vezes nos esquecemos tão facilmente, encontrar amigos em um ambiente de conhecimento.

Agradecimento em especial aos professores Roberto C. S. Pacheco, Aran B. T. Morales, José Leomar Todesco, Alejandro Martins e Ricardo S. Lourenço, que contribuíram com opiniões e um apoio inestimável.

Gostaria de agradecer ao Programa de Pós-Graduação em Engenharia de Produção, pela oportunidade em participar de um prestigiado curso. Ao *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), pela autorização na utilização dos dados referentes ao Diretório dos Grupos de Pesquisa no Brasil, ao Laboratório Stela, onde foi possível a realização deste trabalho, e à empresa Maximiza Ltda., pelo incentivo na fase inicial do trabalho.

Obrigado ao professor e amigo Jomi F. Hübner e ao professor Paulo de T. Luna, pela confiança depositada.

Também gostaria de agradecer aos amigos que de alguma maneira participaram desta fase. Sou muito grato a Sandro Rautenberg, Alexander R. Valdameri, Carlos P. Niederauer, Marcos Scremin e aos amigos do Grupo Stela.

Aos meus pais, José e Izelda Gonçalves, e familiares pelo incentivo, confiança e exemplos de força e dedicação.

Finalmente, gostaria de agradecer à minha esposa por quem tenho admiração e tem estado sempre ao meu lado. Obrigado Ednara M. K. Gonçalves.

“A sabedoria é a coisa principal; adquire pois a sabedoria, emprega tudo o que possues na aquisição de entendimento”

Provérbios 4:7

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Gráficos

Lista de Reduções

Siglas

ART	Adaptive Resonance Theory
BAM	Bidirecional Associative Memory
C&T	Ciência e Tecnologia
CAs	Comitês Assessores
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
COAV	Coordenação de Planejamento, Acompanhamento e Avaliação
CTC	Conselho Técnico-Científico
DEA	Análise Envoltória de Dados
DNA	Ácido Desoxirribonucleico
ECG	Eletrocardiograma
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
FBR	Função de Base Radial
FINEP	Financiadora de Estudos e Projetos
GPAs	Grupos de Planejamento e Avaliação
GPGEs	Grupos de Planejamento e Gestão Estratégica dos Programas Cooperativos
GTC	Grupo Técnico de Coordenação,
IES	Instituição de Ensino Superior
KDD	Knowledge Discovery in Database
MCT	Ministério da Ciência e Tecnologia
MD	Mineração de Dados
MLP	Multi-Layer Perceptron
NSF	National Science Foundation
P&D	Pesquisa e Desenvolvimento
PADCT	Programa de Apoio ao Desenvolvimento Científico e Tecnológico
RNAs	Redes Neurais Artificiais
SCG	Secretária de Coordenação e Gerenciamento
SIN	Superintendência de Informática
SNPG	Sistema Nacional de Pós-Graduação
SOM	Self-Organizing Maps
SUP	Superintendência de Planejamento

Resumo

Os avanços na Ciência e na Tecnologia (C&T) de uma nação cumprem papel decisivo no desenvolvimento de sua sociedade e na melhoria da qualidade de vida de sua população. Nas últimas décadas, o produto desses desenvolvimentos, aliado aos próprios avanços na área de Tecnologia da Informação (TI), permitiram a formação de grandes bases de dados em C&T. Nesse contexto, a área de TI viu o desenvolvimento da área de *Extração de Conhecimento* e de métodos de *Mineração de Dados* (MD), visando transformar massas de dados em conhecimento. O objetivo principal desta dissertação é demonstrar a viabilidade e a utilidade de técnicas de MD na elucidação de conhecimento a partir de informações constantes em bases de C&T. Para tal, aplicam-se técnicas de MD em processo de avaliação de grupos de pesquisa. Por meio de agrupamento não supervisionado (*Rede Neural Kohonen*), o trabalho classificou grupos de pesquisa em cinco classes, segundo critérios utilizados em algoritmo de avaliação parametrizado. O resultado aponta a coerência de classificação e, devido à natureza da técnica neural utilizada, permitiu a elucidação de regras exploratórias (*Regras de Associação*) que descrevem textualmente as características das classes encontradas pelo algoritmo neural. Como principal contribuição, a dissertação oferece meios tanto para trabalhos exploratórios de bases de C&T como amplia o conjunto de ferramentas atualmente disponível em processos de avaliação e gestão de C&T.

Abstract

Advances in a nation's Science and Technology (S&T) play a decisive role in the development of its society and the improvement of the quality of life among its population. Over the last few decades the products of these developments, along with advances in the area of Information Technology (IT) in themselves, have provided for the formation of larger S&T databases. In this context, the area of IT witnessed the development of the area of *Extraction of Knowledge* and *Data Mining* (DM) methods, with the aim of transforming masses of data into knowledge. The chief objective of this dissertation is to demonstrate the viability and utility of DM techniques in the elucidation of knowledge through the information found in S&T bases. To this end, DM techniques are applied in the process of evaluating research groups. By means of non-supervised grouping (*Kohonen Neural Network*), this study has divided research groups into five classes, according to criteria utilized in an evaluation algorithm with parameters. The result points to coherence of classification and, due to the nature of the neural technique utilized, it was possible to elucidate the exploratory rules (*Association Rules*) that textually describe the characteristics of the classes found by the neural algorithm. As the principal contribution, the dissertation offers means both for exploratory work based on S&T and for enhancing the set of tools currently available in processes of S&T evaluation and management.

1 INTRODUÇÃO

"A tecnologia ensinou uma lição à humanidade: nada é impossível"

Lewis Mumford

Cada vez mais Ciência e Tecnologia (C&T) demonstram a necessidade de investimentos para que sejam alcançados resultados futuros que promovam o crescimento da sociedade em seu conjunto e, principalmente, que elevem a capacitação técnico-científica das sociedades menos desenvolvidas. Esse desenvolvimento se faz necessário diante do atual quadro em que se intensificam os inter-relacionamentos entre as sociedades, sejam elas produtoras ou consumidoras.

O atual estágio mostra, acima de tudo, um grande amadurecimento, visto que até o século passado pouca teoria podia ser transformada realmente em tecnologia aplicada. Como afirmam Schwartzman et al. (1993), ciência e tecnologia são elementos de transformação cruciais para uma nação, quando se trata de elevar o padrão de vida da população, consolidar uma economia moderna e participar com plenitude em um mundo cada vez mais globalizado. Nesse contexto, para que um país atinja níveis adequados de desenvolvimento social, cultural e tecnológico, deve investir na formação de indivíduos capazes de competirem em um ambiente cada vez mais dinâmico.

Levando-se em consideração que investimentos em C&T visam posicionar uma nação entre as geradoras de ciência, as agências de fomento podem ter um papel de grande importância nos rumos de desenvolvimento de um país. As principais agências de fomento são, no Brasil, o *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), a *Financiadora de Estudos e Projetos* (FINEP), a *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), as Fundações Estaduais de Apoio à Pesquisa (com destaque para a paulista FAPESP), e os Institutos de Pesquisa.

Embora já se possa apontar exemplos concretos do modelo brasileiro, fruto do esforço de agências, instituições de ensino e pesquisa e de pesquisadores na construção da base científica no país, um longo caminho ainda deve ser percorrido. Um dos avanços necessários é a construção de instrumentos que dêem maior suporte às

decisões de avaliadores e técnicos das agências, a fim de racionalizar o fomento e permitir flexibilidade tanto na definição de políticas quanto na aplicação de recursos públicos em projetos de C&T. Em particular, é necessário aplicar diferentes indicadores e sistemas de avaliação em níveis qualitativos e quantitativos. É nesse contexto que se insere a presente dissertação. O trabalho consiste na utilização de técnicas de Mineração de Dados (MD), na identificação e extração de conhecimento, bem como na utilização de tais técnicas na hierarquização dos Grupos de Pesquisa no Brasil, tomando-se como fonte de informação o conjunto de dados fornecido pelo CNPq.

1.1 JUSTIFICATIVA

A avaliação de indicadores de C&T é um processo complexo e que necessita de um amplo entendimento dos responsáveis pela análise. Devido ao grande volume e à heterogeneidade dos dados disponíveis, este trabalho torna-se por muitas vezes de difícil realização. Tais obstáculos dificultam a avaliação e a própria tomada de decisão na gestão de C&T.

Nesse contexto, ferramentas, métodos e algoritmos que auxiliem nos procedimentos de análise e de tomada de decisão tornam-se de grande relevância, principalmente quando se aumentam a quantidade e a complexidade dos dados armazenados. Entre os métodos que procuram auxiliar nesse processo, encontram-se a análise de agrupamentos ou análise de categorias (*clustering*), e regras de associação.

A análise de agrupamentos visa identificar as relações entre dados, agrupando-os segundo algum critério de similaridade. Esse tipo de análise fornece informações tanto conhecidas quanto desconhecidas. Pode, também, validar conhecimentos intuitivos, por meio de um processo mais simplificado. O segundo método identifica regras entre itens de um determinado conjunto de dados, apresentando a existência de relacionamentos. Essas regras podem descrever conhecimentos ou inter-relações, sem que o analista necessite extraí-las através de complexas buscas nos dados. A aplicação desses métodos na descoberta de conhecimento fornece uma importante ferramenta no processo de análise, avaliação e mensuração de C&T.

1.2 OBJETIVO

Este trabalho tem como objetivo principal demonstrar a viabilidade e utilidade da aplicação de técnicas de Mineração de Dados no processo de descoberta de conhecimento em bases de C&T, bem como utilizar tais técnicas no processo de avaliação de C&T no país.

Para alcançar esse objetivo, o trabalho tratará ainda de metas específicas, descritas a seguir.

1.2.1 OBJETIVOS ESPECÍFICOS

- realizar um levantamento do sistema brasileiro de ciência & tecnologia, identificando os principais mecanismos de avaliação e fomento à C&T no país;
- estudar e implementar as técnicas de descobrimento de conhecimento a partir de bases de dados;
- aplicar as técnicas implementadas na base de dados do *Diretório dos Grupos de Pesquisa no Brasil*, visando comparar o conhecimento elucidado com os resultados obtidos pelas técnicas utilizadas atualmente pelo CNPq (*Algoritmo de Hierarquização dos Grupos de Pesquisa no Brasil*) (Guimarães et al., 1999);

1.3 METODOLOGIA

Para efetivar os objetivos, o trabalho fundamenta-se em três etapas: (a) levantamento do sistema brasileiro de C&T; (b) estudo das principais técnicas de Mineração de Dados (MD); e (c) aplicação dos algoritmos e análise dos resultados em uma base de dados do sistema nacional de C&T. A Figura 1.1 apresenta uma visão esquemática da metodologia de construção do trabalho.

A primeira etapa divide-se em três pontos principais, entre eles:

- *análise do sistema brasileiro de C&T*: são levantados aspectos de C&T no país, tais como seu contexto histórico, seu perfil demográfico, distribuição espacial (formação, produtividade, etc.). São descritos indicadores e o processo de fomento e avaliação de C&T nas principais agências do país e universidades;
- *identificação das principais Bases de Dados de C&T do Brasil*: aqui são relacionadas algumas das principais bases de dados de C&T no país, incluindo a base do Diretório de Grupos de Pesquisa no Brasil (CNPq), a base de dados curriculares do CNPq (Currículo Lattes), e a base Coleta do Sistema de Pós-graduação da CAPES;
- *detalhamento do Projeto Diretório dos Grupos de Pesquisa no Brasil*: por ser este o alvo da aplicação do trabalho, serão identificados os principais itens que compõem essa base, tais como, indicadores, linhas de pesquisa, titulações, etc.

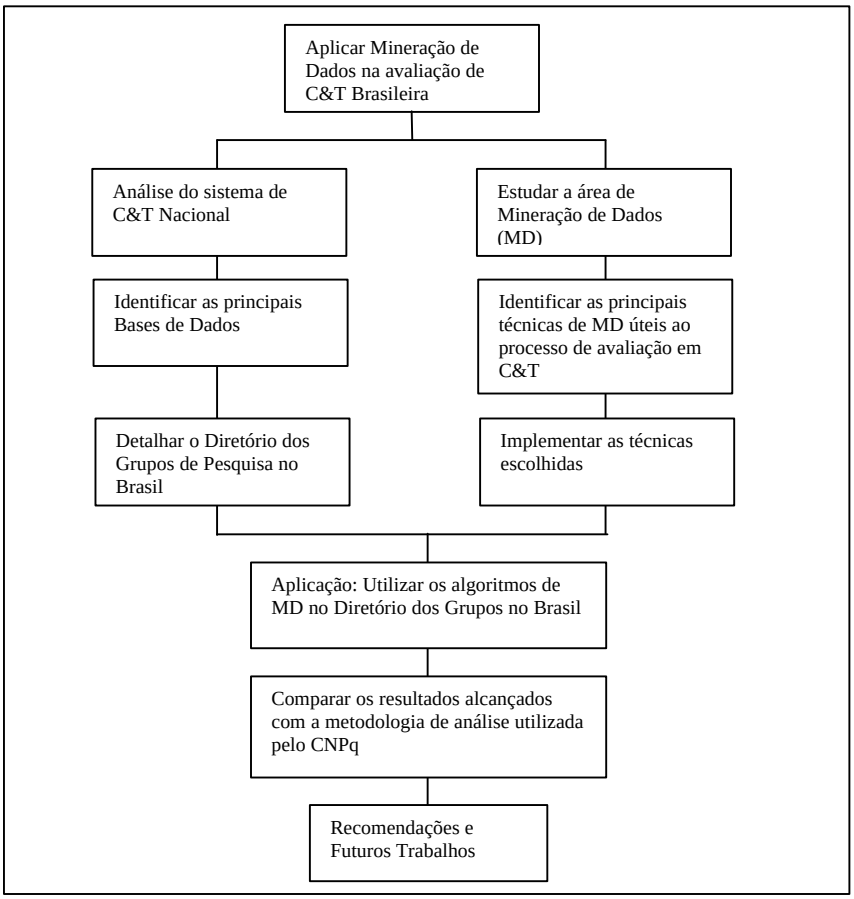


Figura 1.1 - Metodologia do trabalho

A segunda etapa divide-se em três pontos principais, entre eles:

- *estudo da área de Mineração de Dados*: constitui-se no estudo das diversas técnicas, tais como técnicas estatísticas, de inteligência artificial e o algoritmo de regras de associação;
- *identificação das principais técnicas de Mineração de Dados*: procura identificar as principais técnicas de MD úteis ao processo de análise em bases de C&T;
- implementação das técnicas escolhidas durante a fase de identificação.

A última etapa procura validar as técnicas utilizadas na extração de conhecimento e apresentar os resultados obtidos por meio destas. Esta divide-se em três pontos principais:

- *aplicação dos algoritmos escolhidos*: são aplicados os algoritmos de MD escolhidos e identificados como relevantes ao processo (redes neurais e regras de associação), na base do Diretório dos Grupos de Pesquisa no Brasil;
- *apresentação e comparação dos resultados*: são apresentados os resultados alcançados com a utilização das técnicas citadas anteriormente, comparando-os com os resultados já consolidados pelo CNPq;
- *recomendações e futuros desenvolvimentos*: são apresentadas recomendações visando a implementação de novos algoritmos para a avaliação de C&T, bem como uma proposta para trabalhos futuros que forneça uma ferramenta de análise de C&T que possa ser utilizada pela agência (CNPq).

1.4 ESTRUTURA DO TRABALHO

Este trabalho compreende outros quatro capítulos além do presente. Estes capítulos objetivam introduzir e demonstrar a viabilidade da aplicação de técnicas de mineração de dados em problemas de C&T. A estrutura do trabalho apresenta-se da seguinte maneira:

- *Capítulo 2: Ciência & Tecnologia* - é apresentada uma pequena introdução sobre ciência e tecnologia no Brasil, demonstrando a sua relevância na formação de uma memória nacional em C&T. Também é abordada uma visão histórica de

C&T no país, bem como uma visão do processo de fomento e avaliação das principais agências;

- *Capítulo 3: Mineração de Dados* - são abordadas as técnicas de MD estudadas e utilizadas no trabalho, os algoritmos e as possíveis aplicações dessas técnicas;
- *Capítulo 4: Análise de C&T utilizando técnicas de Mineração de Dados* – são abordados aspectos mais detalhados do modelo proposto na descoberta de conhecimento e avaliação de C&T, os esforços na elaboração do trabalho e seus resultados e limitações;
- *Capítulo 5: Conclusões e Recomendações* – são relatadas as conclusões do trabalho, assim como recomendações e projetos para trabalhos futuros.

2 CIÊNCIA & TECNOLOGIA

"Não há capital que dê melhores frutos a uma nação do que aquele que é posto à disposição dos jovens estudiosos e dos homens que, com inteligência, amor e liberdade, se dedicam à pesquisa científica"

Guilherme Guinle

2.1 INTRODUÇÃO

O século XX apresenta-se para a humanidade como o século da ciência e tecnologia. Novos desafios intrigam cientistas e pesquisadores, novos paradigmas são criados, a sociedade almeja novas explicações e necessidades, criando-se assim premissas para que novas teorias sejam formuladas ou explicadas. Esta tem sido a válvula propulsora para descobrimentos e invenções que revolucionaram e continuam a impulsionar a humanidade. É o caso da teoria da relatividade (Albert Einstein), dos estudos sobre a estrutura do DNA (James Watson e Francis Crick), da descoberta da Penicilina (Alexander Fleming), e da aviação (Alberto Santos-Dumont). Estes e inúmeros outros desenvolvimentos ocorreram em meio a incertezas, proporcionaram quebras de paradigmas e colocaram as novas descobertas aos olhos da sociedade.

Pode-se verificar, assim, a maturidade que a ciência atingiu durante os últimos séculos, principalmente em seus métodos científicos. Esses métodos são reflexos das nossas necessidades e possibilidades materiais (Andery et al., 1994), aumentando a complexidade das inter-relações entre ciência e sociedade. No século XVII, os investigadores eram amadores, possuíam recursos escassos e a ciência era tida como uma atividade periférica. Em apenas três séculos, tornou-se poderosa e movida a objetivos econômicos com um caráter transformador perante a sociedade e o próprio Estado (Morin, 1999). A ciência promove a tecnologia, que, conforme Ferné (1993), é a junção da técnica e da pesquisa.

Visto que o atual estágio tecnológico promove maior democratização da informação, torna-se cada vez mais necessário disponibilizar informações e serviços às atividades científicas, a fim de que estas sejam instrumentos transformadores dos

padrões qualitativos de vida do ser humano. Assim sendo, ciência e tecnologia tornam-se um único elo e são conceitos-chave para o crescimento das sociedades. Isto é verificado na crescente importância do conhecimento em relação ao trabalho e ao capital (Silva, 1994).

> O conhecimento científico pode ser entendido, abstratamente, como um conjunto de informações ou dados, cujo valor independe dos homens que o produziram. No entanto, talvez o principal resultado desta pesquisa tenha sido a constatação de que ciência é, acima de tudo, uma comunidade de pessoas bem-formadas, trabalhando com entusiasmo no ápice de suas inteligência e criatividade. O resultado deste trabalho – artigos, informações, aplicações tecnológicas, dados – não passa da ponta do iceberg de valor precário, temporário, e que não tem como se sustentar sem a base que lhe dá existência, que são os homens que o produziram. (Schwartzman, 1979).

O aumento crescente na relevância e interferência da ciência e tecnologia no cotidiano pode ser verificado na afirmação de Price (1976), que diz que a cada 15 anos há duplicação nos indicadores científicos, isso em termos de número de cientistas, número de periódicos e referências bibliográficas. Esse padrão de crescimento é em geral quebrado devido a grandes mudanças, como por exemplo, a Segunda Guerra Mundial, sendo essas mudanças agentes de transformação no contexto científico. Levando-se em consideração um contexto histórico, notam-se períodos de estagnação e outros de ascensão. Atualmente a curva é exponencial e, considerando que pontos de estagnação devam ocorrer em determinados períodos, essa curva tende a uma curva logística (Figura 2.1) (Price, 1976). Contudo, como o próprio autor cita, os atuais estágios de crescimento científico resistem à desaceleração.

Essas questões, mesmo com suas contribuições inegáveis ao estilo de vida moderno e, principalmente, ao suporte e manutenção do futuro tecnológico da humanidade, promovem reações favoráveis, mas também contrárias. Críticas e crescentes preocupações florescem, principalmente no campo da ética, em questões quanto ao uso adequado da ciência. Existe uma relação contraditória entre a ciência e a ética, pois estas pertencem a diferentes campos – o campo do “ser” e o campo do “deve ser” (UNESCO, 1999). Portanto, medidas cooperativas entre as sociedades devem ser criadas, de modo que estas se beneficiem.

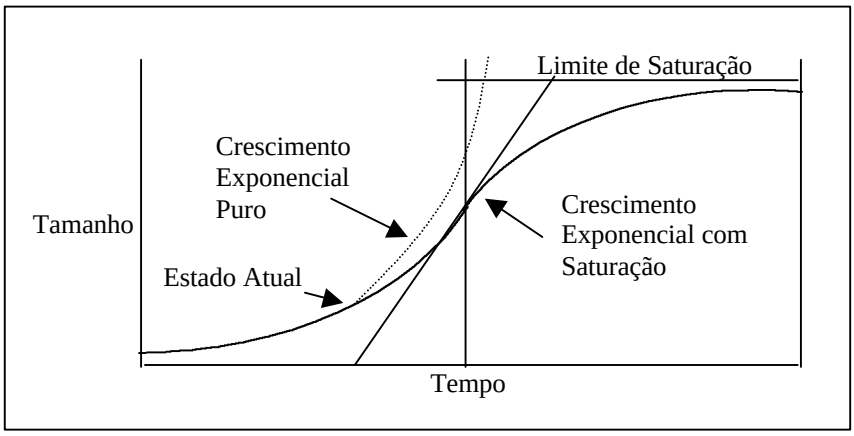


Figura 2.1 - Forma geral da curva logística (Adaptado de Price, 1976)

Como se pode ver, o desenvolvimento científico e tecnológico tem pressões de contexto (sociedades e governos buscando novos patamares de desenvolvimento sócio-econômico) e de imersão de informações (fruto do grande volume de produção científica, tecnológica e artística/cultural). Do ponto de vista da gestão e fomento à C&T, as exigências de sociedade e governo traduzem-se em necessidade de tomadas de decisão rápidas e eficientes, enquanto o volume da produção dificulta a análise de cenários, comparações e avaliações quantitativas e qualitativas.

Este trabalho tem por objetivo introduzir a aplicação de uma das mais recentes áreas de suporte à decisão e descoberta de conhecimento perante os grandes volumes de dados no contexto de C&T. Para tal, neste capítulo apresenta-se uma visão geral de C&T no país, as principais instituições de fomento e as bases de dados e plataformas que promovem suporte às operações dessas organizações.

2.2 HISTÓRICO

“No passado, ciência e tecnologia não foram somente atividades separadas e com objetivos diferentes, mas atividades realizadas por pessoas totalmente distintas, de classes sociais diversas e com pouca comunicação entre si.” (Schwartzman, 1979).

Analisando-se esta relação, pode-se vislumbrar a ciência como um complexo conjunto de conhecimentos modificáveis ao longo do tempo, devido a novas descobertas, e à quebra de paradigmas. A ciência baseia-se em um modelo distribuído,

uma vez que o conhecimento não pertence a um único indivíduo, sendo este conhecimento multidisciplinar e por vezes complexo.

Essa complexidade pode ser verificada nos últimos séculos em que a atividade científica migrou de um movimento individualizado para um atividade científica institucionalizada. Isso possuiu maior ênfase a partir do século XVII, com a crescente evolução da ciência na Europa e EUA, em instituições como a *Royal Society* (1660), na Inglaterra, a *École Polytechnique* (1795), na França, a *Sheffield Scientific School* (1847) e o *Massachusetts Institute of Technology* (1865), nos EUA, (Meis, 1994).

A institucionalização da ciência promove cada vez mais o crescimento da pesquisa e desenvolvimento (P&D), com conseqüências diretas sobre a sociedade. Dois exemplos bastante elucidativos são apresentados por Meis (1994), considerando os seguintes pontos:

- *transportes*: em 1600 a.C., o veículo mais rápido era a biga, que atingia 8km/h. Em 1840, surgiu a diligência, que podia alcançar 30 km/h. Atualmente, naves espaciais podem atingir até 28 mil km/h em velocidades orbitais;
- *crescimento populacional*: no início da Era Cristã a população mundial era de 300 milhões de habitantes, alcançou 900 milhões no século XIX e atualmente chega a mais de 6 bilhões de habitantes. Esse grande aumento nos últimos dois séculos e principalmente no século atual deve-se aos avanços da medicina (*e.g.*, aumento da expectativa de vida média de 35 anos no início do século passando para 75 anos nos dias atuais em países desenvolvidos), à expansão da agricultura e à educação em geral.

Isso indica o crescente aumento do conhecimento e, em contrapartida, a dificuldade de assimilação da informação gerada por esse conhecimento. Basta citar que até poucos séculos atrás, o conhecimento podia ser resumido em algumas centenas de livros, disponíveis nas principais bibliotecas do mundo. Atualmente, milhares de artigos são publicados anualmente em revistas científicas indexadas pelo SCI (*Scientific Citation Index*). Existem atualmente 5.722 periódicos (ciências naturais, matemática e tecnologia), indexados no SCI, dos quais 16 são periódicos brasileiros (Tuffani, 1999). Por outro lado, isso demonstra uma grande dependência em relação à

produção de conhecimento externo, visto que países em desenvolvimento, como o Brasil, são grandes consumidores de conhecimento e a pequena comunidade científica desses países possui a responsabilidade de levar até aos jovens a grande massa de conhecimento científico (Meis, 1994).

Sob um contexto histórico nota-se que mesmo uma ciência menos desenvolvida não impede o avanço e o progresso tecnológico, bem como um ambiente científico organizado não garante um desenvolvimento tecnológico apropriado. Exemplos desses fatos podem ser verificados no século XIX nos Estados Unidos, que, por meio de processos adaptativos, conseguiram assimilar e acelerar seu desenvolvimento tecnológico diante de países como Inglaterra e Alemanha, que na época possuíam uma cultura científica mais organizada e estruturada. Com relação ao segundo ponto, ressalta-se mais recentemente a Índia, que, mesmo com uma comunidade científica bem desenvolvida, não possui um desenvolvimento tecnológico adequado (Schwartzman, 1979). Case ressaltar que atualmente a Índia é um país com excelência no desenvolvimento de software.

Um ponto crítico na relação ciência/tecnologia é a necessidade de resultados rápidos, principalmente nos países em desenvolvimento, levando estes a buscarem facilidades na obtenção de tecnologia, normalmente conseguida através da absorção e importação.

Pode-se, assim, configurar dois hemisférios: o primeiro, com ênfase na obtenção de lucro e o segundo, mais negativo, que consiste na dependência tecnológica, indo desde a manutenção e atualização até a capacitação técnica de recursos humanos. Esse tipo de fenômeno promove um ponto de desequilíbrio gerando o que pode ser chamado de uma subciência ou uma ciência mais elementar.

No Brasil, especificamente, o desenvolvimento tecnológico poucas vezes conseguiu fazer frente ao *know-how* estrangeiro. Para que ocorram mudanças nesse quadro, uma política de desenvolvimento científico e tecnológico deve estar baseada no fomento vinculado a uma avaliação consistente juntamente com investimentos em educação básica. Essas iniciativas não podem basear-se em um curto prazo, embora uma política adequada de investimentos em C&T deva produzir resultados concretos nesse período.

> O conhecimento científico e tecnológico emerge como a base de produção da riqueza nacional, em todos os níveis. É justamente nos países com patrimônios tecnológicos mais avançados que se verifica uma melhor qualidade de vida e maior participação no cenário político e econômico mundiais. (Sardenberg, 1999).

Entretanto, nos últimos anos tornou-se possível verificar um desenvolvimento tecnológico no país em diversas áreas com resultados comparáveis aos padrões internacionais.

Nas próximas seções serão apresentados uma visão geral e um perfil da C&T no Brasil, o processo de fomento e de avaliação utilizados pelas agências nacionais e as principais bases de dados organizadas no país.

2.3 CIÊNCIA, PESQUISA E DESENVOLVIMENTO NO BRASIL

Os esforços para promover a C&T no Brasil nas últimas décadas podem ser consideráveis, mas insuficientes para fazer frente ao processo de globalização que torna países em desenvolvimento cada vez mais vulneráveis às mudanças nos mercados internacionais. Nessas últimas três décadas, muitas fases podem ser evidenciadas no desenvolvimento científico e tecnológico do país, entre elas o crescimento econômico dos anos 70, a recessão dos anos 80 e o avanço nas técnicas de gestão e produção dos anos 90.

Nos anos 70, o Brasil registrou altas taxas de crescimento econômico, mas não conseguiu ajustar-se às mudanças que ocorreram na década de 80, entrando num período de recessão. Essa instabilidade promoveu a redução dos recursos destinados à programas de C&T e reduziu a confiança da opinião pública em relação à importância da pesquisa científica. Diante dos constantes avanços científicos e tecnológicos torna-se necessária uma política de C&T capaz de promover o desenvolvimento da ciência nacional, que deve estar baseada na liberdade e na criatividade de seus pesquisadores e no vínculo destes com a própria sociedade, com objetivos claros e com uma visão a longo prazo (Schwartzman et al., 1995).

Segundo Schwartzman et al. (1995), para que esses objetivos sejam alcançados, são recomendados:

- utilização e desenvolvimento de tecnologia e ciência aplicadas, visando atender às novas realidades econômicas e sociais, estimular e fortalecer os vínculos dos grupos de pesquisa com o setor produtivo e promover maior participação do setor privado em investimentos de C&T;
- apoio à ciência básica, preservação da capacitação científica já instalada e maior envolvimento das instituições de pesquisa no ensino técnico e na formação profissional;
- ampliação e manutenção dos canais de cooperação internacional entre o Brasil e as instituições, agências e a comunidade científica internacional;
- disseminação da informação e do conhecimento visando sua transferência e incorporação ao processo industrial;
- ampla reforma institucional, nos ministérios, agências de fomento e nas universidades; e
- adoção de uma política de projetos dirigidos, buscando tanto o fortalecimento de áreas estratégicas como a redução de outras.

A ciência aplicada brasileira obteve um grande impulso nas primeiras décadas deste século, principalmente na área da agricultura, pecuária e na área de recursos minerais. Esse impulso pode ser verificado mais claramente a partir da década de 50, durante o governo Dutra, com a criação do *Conselho Nacional de Pesquisas* (CNPq) (atual *Conselho Nacional de Desenvolvimento Científico e Tecnológico*), tendo como maior incumbência representar a comunidade científica e fomentar a pesquisa e o desenvolvimento tecnológico no país.

Isso representou um marco fundamental na participação do Estado no processo de desenvolvimento científico e tecnológico do Brasil e atendeu a uma antiga reivindicação da comunidade científica (Barbieri, 1993). Tais reivindicações tornam-se de grande relevância, quando se observa a implantação de órgãos com o objetivo de investir nas atividades de pesquisa e na formação de recursos humanos em instituições estrangeiras, tais como o *Conseil Supérieur de la Recherche Scientifique* na França, e o *National Science Foundation* (NSF), nos Estados Unidos.

Contudo, a limitação de recursos fez com que a agência (CNPq) nem sempre conseguisse desempenhar seu objetivo principal plenamente, provocando certa

insatisfação na comunidade científica. Essa insatisfação era condicionada a critérios de rentabilidade econômica e resultados imediatos, obrigando muitas vezes os pesquisadores a buscar alternativas no fomento às suas pesquisas. Mesmo quando instituições não proporcionavam aos cientistas maiores condições ou incentivos, foi possível levar adiante trabalhos de pesquisa (Schwartzman, 1979).

Outro aspecto a considerar é o processo de industrialização fortemente baseado no capital estrangeiro, que contribuiu para inibir o desenvolvimento da ciência brasileira no início do século. De acordo com Schwartzman (1979), existem alguns pontos a serem destacados: primeiro, em algumas áreas, a ciência brasileira fundamentou-se essencialmente na ciência européia, principalmente história natural, astronomia, medicina bacteriana e química tradicional; segundo, a ciência praticada no país dependeu de pesquisadores vindos de outros países que aqui se radicavam, ou de brasileiros que se dirigiam ao exterior para obter formação; e terceiro, havia poucos lugares onde esta educação importada ou adquirida no exterior pudesse ser aplicada. Esses pontos relatam as primeiras tentativas, por vezes pouco satisfatórias, de se implantar uma ciência moderna no Brasil.

A evolução do desenvolvimento científico brasileiro, a partir da década de 80, e principalmente nos anos 90, apresenta um caráter muito mais adaptativo ao novo contexto de globalização mundial do que o fortalecimento da ciência nacional. Entretanto, os esforços do governo, agências de pesquisa (CNPq, CAPES, FINEP), a iniciativa de se criarem agências estaduais baseadas na experiência da FAPESP e o papel das instituições de ensino na formação de mão-de-obra, analisados a longo prazo, tendem a formar uma comunidade científica mais consciente de seu papel perante a sociedade brasileira. Esses esforços vão desde os investimentos necessários à formação de recursos humanos até à formação de bases de dados destinadas a registrar a memória da ciência nacional.

Um dos principais levantamentos sobre a ciência nacional ocorreu na primeira metade da década de 90, objetivando criar um mapa da atividade dos grupos de pesquisa brasileiros identificando fatores como a distribuição geográfica, as áreas de atuação e a produção C&T desses grupos. Esse trabalho realizou-se, inicialmente, em 1992 (Guimarães et al., 1999) e encontra-se na versão 4.0. O conceito de grupos de

pesquisa não é novo e tem como um dos trabalhos iniciais um estudo realizado pela UNESCO, em meados da década de 70. Esse estudo, segundo Schwartzman e Castro (1986), visava corrigir a tendência das análises que, em um extremo, preocupava-se com a atividade de pesquisa individual e no outro, com sistemas nacionais de ciência e tecnologia.

2.3.1 O PERFIL DA C&T BRASILEIRA

Apesar de recentes, os esforços para reunir dados confiáveis para a análise de C&T permitem gerar indicadores e traçar um perfil da C&T brasileira. Esses indicadores tendem a se tornar mais confiáveis nos próximos anos, em virtude da recente plataforma de captura de dados curriculares do CNPq, à integração das bases do CNPq e CAPES e à nova versão do Diretório dos Grupos de Pesquisa no Brasil, integrado à Plataforma Lattes.

Considerando esses aspectos, serão identificados os principais pontos de análise visando demonstrar um perfil de C&T relativo a grupos de pesquisa, pesquisadores (utilizando a base do Diretório dos Grupos de Pesquisa no Brasil mantida pelo CNPq) e cursos de pós-graduação (utilizando a base mantida pela CAPES). Essas bases permitem apresentar um perfil de C&T mostrando indicadores importantes quanto a: (a) distribuição por áreas do conhecimento (pesquisadores, estudantes, pós-graduação, produção C&T), e (b) distribuição demográfica dos grupos de pesquisa, pesquisadores e avaliação dos cursos de pós-graduação por região do país.

a) DISTRIBUIÇÃO POR ÁREAS DO CONHECIMENTO

O perfil por áreas do conhecimento permite visualizar a distribuição das diferentes unidades no processo de C&T num contexto histórico, entre elas: grupos de pesquisa, linhas de pesquisa, estudantes vinculados a grupos de pesquisa, alunos da pós-graduação (mestrado e doutorado) e produção C&T. As grandes áreas de conhecimento analisadas são: Ciências Agrárias, Ciências da Saúde, Ciências

Humanas, Lingüística, Letras e Artes, Ciências Sociais Aplicadas, Engenharias e Ciência da Computação, Ciências Exatas e da Terra e Ciências Biológicas.

Os grupos de pesquisa estão distribuídos de forma relativamente homogênea, sendo que Ciências da Vida (Ciências Biológicas, Ciências da Saúde, Ciências Agrárias) possuem o maior número de grupos, com 43%, seguidos pelas Ciências da Natureza (Ciências Exatas e da Terra, Engenharias e Computação) e Humanidades (Ciências Humanas, Ciências Sociais Aplicadas, Letras, Lingüística e Artes), com 31% e 26%, respectivamente (Gráfico 2.1, Anexo II). As linhas de pesquisa estão mais concentradas em três das oito Grandes Áreas do Conhecimento, sendo que Engenharias e Ciência da Computação, Ciências da Saúde e Ciências Humanas possuem juntas 52% (Gráfico 2.2, Anexo II).

Quanto aos estudantes e estagiários com atuação direta nas atividades de pesquisa dos grupos, estes estão mais concentrados nas áreas de Ciências da Saúde, Engenharias e Ciência da Computação, Ciências Humanas e Ciências Biológicas, com 66% (Gráfico 2.3, Anexo II). Quanto à distribuição dos alunos de pós-graduação, as áreas de Ciências Humanas e Ciências da Saúde possuem a maior concentração, tanto no mestrado, com 33% (Gráfico 2.4, Anexo II), quanto no doutorado, com 38% (Gráfico 2.5, Anexo II).

Por último, a produção científica pode ser dividida em itens relacionados a grupos de pesquisa e Programas de Pós-graduação. Quanto aos grupos de pesquisa, a distribuição das produções por área do conhecimento apresenta-se de maneira relativamente homogênea, como mostra o Gráfico 2.6, Anexo II. Na pós-graduação ocorre uma grande concentração no número de produções na área de Ciências da Saúde, com 36%, seguida pela área de Ciências Humanas, com 14% (Gráfico 2.7, Anexo II).

Para melhor visualizar as informações descritas no Anexo II é apresentada a Tabela 2.1, que demonstra as variáveis/unidades de análise divididas por Grandes Áreas do Conhecimento. Analisando-se a distribuição dos estudantes de mestrado no CNPq e CAPES, nota-se uma homogeneidade nas Grandes Áreas, diferindo basicamente nas Ciências Biológicas (12% para o CNPq e 9% para a CAPES), nas Engenharias e Ciências da Computação (21% para o CNPq e 15% para a CAPES) e nas

Ciências Sociais Aplicadas (8% para o CNPq e 11% para a CAPES). Para os estudantes de doutorado, a diferença pode ser evidenciada em quatro Grandes Áreas com percentuais mais expressivos. Nas Ciências Biológicas, 17% e 13%; nas Ciências da Saúde, 15% e 20%; nas Engenharias e Ciências da Computação, 19% e 14%; e nas Ciências Humanas, 14% e 18%, respectivamente para CNPq e CAPES.

Tabela 2.1 - Distribuição das unidades de análise segundo as Grandes Áreas do Conhecimento

	Ciências da Vida			Ciências da Natureza		Humanidades			
Unidade de Análise / Grande Área	Ciências Biológicas	Ciências da Saúde	Ciências Agrárias	Ciências Exatas e da Terra	Engenharias e C. da Computação	Ciências Humanas	Ciências Sociais Aplicadas	Letras, Lingüística e Artes	Multidis-ciplinar
Grupos de Pesquisa (CNPq)	16%	17%	11%	16%	16%	14%	7%	5%	0%
Pesquisadores em Grupos (CNPq)	14%	16%	12%	16%	16%	14%	7%	5%	0%
Estudantes em Grupos (Mestrado – CNPq)	12%	15%	12%	11%	21%	16%	8%	5%	0%
Estudantes em Grupos (Doutorado – CNPq)	17%	15%	11%	16%	19%	14%	4%	4%	0%
Estudantes em Pós-graduação (Mestrado – CAPES)	9%	15%	12%	12%	15%	18%	11%	6%	2%
Estudantes em Pós-graduação (Doutorado – CAPES)	13%	20%	10%	15%	14%	18%	5%	5%	0%
Linhas de Pesquisa dos Grupos (CNPq)	15%	17%	13%	17%	18%	10%	6%	4%	0%
Produção C&T (CNPq)	17%	19%	13%	16%	17%	18%			0%

b) DEMOGRAFIA DOS PESQUISADORES

A distribuição demográfica dos pesquisadores no país apresenta sua maior concentração na região Sudeste, devido ao maior número de pesquisadores, grupos de pesquisa e programas de pós-graduação.

Com relação à distribuição de grupos de pesquisa, verifica-se um maior número dos grupos na região Sudeste (Gráfico 2.8, Anexo II), com 66%, seguido pela região Sul, com 16%. O mesmo acontece com os pesquisadores vinculados a esses grupos, que estão em sua maioria na região Sudeste, com 64% (Gráfico 2.9, Anexo II).

Na avaliação dos programas de pós-graduação, ocorre uma maior concentração destes na região Sudeste, com 62% (Gráfico 2.10, Anexo II) – quase a totalidade dos programas com nota 7 (nota máxima na avaliação da CAPES) (Gráfico 2.11, Anexo II).

Depois da região Sudeste, seguem as regiões Sul, Nordeste, Centro-Oeste e Norte, sendo que destas, somente a Sul e a Sudeste possuem curso com avaliação 7. Constata-se, também, que a maior concentração de notas fica entre 3 e 5, constituindo mais de 80% dos cursos avaliados. As Tabelas 2.2 e 2.3 apresentam as informações descritas no Anexo II para a distribuição dos pesquisadores por região, e dos programas de pós-graduação por região e notas.

Tabela 2.2 - Distribuição das unidades de análise segundo as regiões do país

Unidade de Análise / Região	**Sudeste (%)**	**Sul (%)**	**Nordeste (%)**	**Centro-Oeste (%)**	**Norte (%)**
Grupos de Pesquisa (CNPq)	66	16	12	4	2
Pesquisadores em Grupos (CNPq)	64	17	12	5	2
Cursos de Pós-graduação (CAPES)	62	17	14	5	2

Tabela 2.3 - Distribuição dos programas de pós-graduação por região segundo a nota de avaliação

Região / Nota	**Nota 7 (%)**	**Nota 6 (%)**	**Nota 5 (%)**	**Nota 4 (%)**	**Nota 3 (%)**	**Nota 2 (%)**	**Nota 1 (%)**	**Sem avaliação (%)**
Sudeste	1,70	6,43	16,65	20,22	13,48	1,94	0,62	0,31
Sul	0,08	1,01	3,41	5,65	6,12	0,46	0,08	0,00
Nordeste	0,00	0,23	1,94	6,04	5,42	0,54	0,15	0,00
Centro-Oeste	0,00	0,39	0,70	1,47	2,48	0,23	0,00	0,08
Norte	0,00	0,00	0,15	0,70	1,01	0,15	0,15	0,00

2.3.2 O FOMENTO EM C&T NO BRASIL

No contexto da C&T, fomento pode ser entendido como um mecanismo que visa aumentar a qualificação, o aperfeiçoamento, a especialização e a formação de pesquisadores ou pessoal ligado à ciência e tecnologia, por meio da concessão de recursos financeiros, materiais ou logísticos. Esse mecanismo é essencial para as atividades de pesquisa e desenvolvimento (P&D) e para a formação de uma ciência básica forte que seja capaz de oferecer condições ao desenvolvimento de projetos, bens e serviços. Também possui forte ligação com a avaliação da pesquisa, sendo que, nas últimas décadas, montantes crescentes de recursos destinados à ciência tiveram sua distribuição vinculada ao resultado de avaliações (Schwartzman e Castro, 1986). A questão da avaliação será discutida na seção 2.3.3.

No Brasil, os investimentos em C&T são quase totalmente financiados pelo Governo Federal, sendo CNPq e CAPES as instituições especializadas que mais investem nessa atividade (MCT, 1998a). Cabe também ressaltar a atuação da FINEP, através do FNDCT, responsável pela infra-estrutura laboratorial de P&D, bem como as agências estaduais (com destaque para a FAPESP), e as empresas estatais.

Existem, porém, alguns pontos importantes a serem considerados, entre eles a baixa participação da iniciativa privada nos investimentos em C&T e o aumento do número de pesquisadores no país através do financiamento das agências de fomento.

A participação da iniciativa privada na pesquisa e desenvolvimento é de grande importância, pois esses investimentos fornecem subsídios para a formação e o aumento da capacitação tecnológica no país. Contudo, em muitos países, principalmente nos países em desenvolvimento como o Brasil, os investimentos em pesquisa e desenvolvimento são realizados com recursos públicos. Isso pode ser melhor entendido se comparados os investimentos do setor privado em C&T, que ficam em torno de 18% do montante investido (MCT, 1998a) no país, enquanto que nos países industrializados podem chegar a 50% (Krieger e Galembeck, 1996) e mesmo mais de 70% no caso do Japão (Arruda, 1994). Entretanto, essa interação não deve ser exclusiva, ou seja, a iniciativa privada também deve investir em P&D, criando seus centros de pesquisa e contribuindo para que as profissões de C&T migrem de acadêmicas à geradoras de PIB (Cruz, 1997). A importância desses investimentos cresce quando se leva em consideração que o custo da ciência tem aumentado proporcionalmente ao quadrado do número de cientistas (Price, 1976).

O atendimento das demandas de C&T, tanto no MCT quanto nas agências de fomento, tem sofrido com a instabilidade da economia, promovendo uma redução dos investimentos. Esses investimentos tiveram o seu maior ápice em meados da década de 70, sofrendo graves cortes nas décadas de 80 e 90, mantendo-se posteriormente estáveis, ressaltando contudo a queda nos investimentos em 1992 (MCT, 1998a). Outro ponto importante a ser considerado é a concessão de bolsas nas suas diversas modalidades fomentadas pelo CNPq. Após a queda de 1992, esse tipo de fomento tem-se mantido estável, tanto em valores (com exceção de 1994) (Tabela 1a, Anexo I), quanto na quantidade de bolsas oferecidas (Tabela 1b, Anexo I).

Em um contexto mais amplo, a redução dos investimentos em C&T apresenta fortes aspectos políticos. Segundo Guimarães (1994), podem ser visualizados quatro motivos para a "deterioração" das políticas de C&T: (1) desequilíbrio nas contas do governo, em virtude do choque do petróleo e do aumento das taxas de juros no mercado internacional, ocorridos na virada das décadas de 70-80; (2) dificuldade na obtenção de recursos com credores externos na década de 80; (3) aprofundamento da crise fiscal ao final da década de 80 e início de 90; e (4) crise na principal base institucional de ciência e tecnologia do país, as universidades públicas.

Assim sendo, uma política de investimentos em desenvolvimento de C&T visando projeções futuras tanto em níveis internos quanto perante à comunidade científica internacional, deve, sobretudo, estar baseada na definição de áreas prioritárias ou tidas como ciências básicas, na manutenção e no aumento do vínculo entre as universidades e a iniciativa privada, além do aumento no número de pesquisadores no país.

Em seguida, a expansão da base científica do país depende de alguns fatores, em especial da estrutura ou modelo atual de ensino superior implantado no país, e do nível de recursos destinados à formação de futuros pesquisadores. Isso pode ser verificado na população matriculada no ensino superior – cerca de 1% contra 3 a 5% nos países industrializados, sendo 60% da rede privada, e 70% dos cursos estão nas áreas sociais e humanas (Krieger e Galembeck, 1996).

> Conseguir passar da ciência para a tecnologia e depois ser capaz de transferi-la para as linhas de montagem representa o nível máximo de amadurecimento industrial de um país. A característica da indústria de ponta é ser caudatária de um sólido aparato de P&D que, por sua vez, se inspira na ciência que avança. (Castro et al., 1992).

Na visão de Krieger e Galembeck (1996), para o aumento do número de pesquisadores no país são necessárias algumas políticas específicas, com destaque para:

- eficiência do sistema educacional, possibilitando maior ingresso de alunos à universidade;
- estímulo aos alunos visando a orientação desses para profissões ligadas à C&T;

- parcerias entre as universidades e a iniciativa privada, objetivando o desenvolvimento de inovações científicas e tecnológicas e promovendo uma indústria nacional competitiva;
- recuperação da universidade pública como base da ciência nacional e incentivos às universidades privadas.

Além de uma política em C&T adequada, o Brasil somente conseguirá fortalecer o seu desenvolvimento se intensificar os investimentos em P&D, visando a capacitação e formação de recursos humanos. Nesse sentido, a meta para 1999 era de 1,5% do PIB, sendo 50% do setor público (15% oriundo dos estados), 40% do setor privado e 10% de financiamentos externos, valores estes bem mais expressivos em relação aos 0,6% a 0,8% dos últimos anos (MCT, 1998a), mas ainda inferiores aos 2% a 3% dos países desenvolvidos. Contudo, vale registrar o fato de que esse percentual é algo significativo e por vezes mal computado, pois estes 2% a 3% consideram investimentos totais, ou seja, realizados pelo governo e setor privado. Outro ponto a ser mencionado refere-se ao PIB brasileiro que foi multiplicado por 13 nos últimos 50 anos, sendo o 0,7% (geralmente referenciado como o percentual-padrão), também multiplicado por 13. Esse desempenho foi superado somente por alguns países, entre eles o Japão, que cresceu dezenove vezes (Dias, 1997). Sob esse enfoque, os investimentos em C&T são consideráveis e promovem, na medida do possível, o desenvolvimento da ciência no país.

2.3.3 AVALIAÇÃO DE C&T NO BRASIL

A avaliação de C&T é um processo vinculado ao fomento, ou seja, a análise visa produzir subsídios para a tomada de decisão determinando-se os níveis de recursos e quem deve recebê-los. De um modo geral, para se determinar o sucesso científico, utiliza-se o número de publicações feitas por cada autor em periódicos de aceitação geral (Price, 1976). Contudo, como o próprio autor justifica, esse critério mostra-se insuficiente, porque induz a uma aceitação extremamente quantitativa.

Critérios mais elaborados são aplicados pelas agências, levando-se em consideração a natureza do financiamento. As agências em geral possuem metodologias de avaliação tradicional e também metodologias próprias.

Seguindo o modelo de classificação do número de trabalhos/autores apresentado por Price (1976), a Tabela 2.4 demonstra a distribuição dos autores em relação à quantidade de trabalhos produzidos (artigos, trabalhos apresentados em eventos, livros e capítulos de livros, produtos e processos e orientações de teses e dissertações), estando estes divididos em vinte classes. Pode-se notar que os autores com baixa produção (primeira e segunda classe), representam 26,18% do total de autores, correspondendo a 3,84% do total de produções, enquanto que os autores que possuem uma maior produção (últimas duas classes) representam 1,56% do total de autores e correspondem a 11,45% do total de produções. Esses dados, apesar de superficiais, induzem a uma análise mais detalhada das classes extremas, podendo ser agregadas informações que possibilitem uma visão qualitativa.

De um modo geral, o processo de avaliação pode ser visto como uma metodologia de avaliação de pesquisa, sendo esta dividida geralmente em dois grupos (Kostoff, 1997a): qualitativo (*peer review*) e quantitativo (bibliometria ou cientometria e índices econométricos), ou pela combinação dos dois métodos (Schwartzman e Castro, 1986).

Um dos modelos mais utilizados na avaliação de C&T refere-se à avaliação qualitativa ou avaliação por pares *peer review*, sendo esta amplamente utilizada no mundo inteiro (Kostoff, 1997a). O autor propõe uma visão bastante abrangente desse modelo de avaliação, definindo-o como um processo realizado por uma pessoa ou grupo de pessoas em relação à avaliação de outros de mesma categoria. O termo categoria pode ser visto como uma determinada área do conhecimento, em que os integrantes da avaliação são pessoas de respeito acadêmico, com conhecimento acumulado em determinada área. A avaliação leva em consideração alguns critérios, tais como: a qualidade e unicidade do trabalho, o impacto científico e tecnológico em determinada área do conhecimento, a distinção entre o aspecto revolucionário e evolucionário de determinado trabalho, etc.

Tabela 2.4 - Relação de trabalhos por autor produzidos no Diretório 3.0 entre janeiro de 1995 e junho de 1997

Trabalhos/Autor	Autores	%	Trabalhos	%
1	5230	15,13	5.230	1,56
2	3818	11,05	7.636	2,28
3	3067	8,87	9.201	2,75
4	2648	7,66	10.592	3,16
5	2180	6,31	10.900	3,25
5(+)-10	7204	20,84	55.684	16,62
10(+)-15	4423	12,80	57.933	17,29
16(+)-20	1692	4,90	30.456	9,09
20(+)-25	1462	4,23	32.651	9,75
25(+)-33	1267	3,67	35.898	10,71
33(+)-50	1030	2,98	40.513	12,09
50(+)-100	485	1,40	32.657	9,75
100(+)	57	0,16	5.700	1,70
Total	**34563**	**100,00**	**335.051**	**100,00**
Média de Trabalhos/Autor	**9,7**			

Quanto ao modelo quantitativo, este se preocupa com o desempenho e com os resultados da pesquisa em C&T, medidos pelo número de publicações ou número de citações. É, portanto, um processo de contagem de itens que já foram qualitativamente julgados (Schwartzman e Castro, 1986). Os principais meios utilizados para a avaliação são índices bibliométricos e econométricos. O primeiro item leva em consideração o total de publicações, patentes, índice de citações, conferências, capítulos em livros, etc., sendo que esses indicadores servem para medir o desempenho da realização científica e tecnológica. O segundo item baseia-se em medidas econômicas na avaliação do retorno do investimento em pesquisa, levando-se em conta um modelo voltado à análise de retorno de investimento aplicado no meio industrial. Segundo Kostoff (1997b), isso promove algumas falhas quando aplicado em C&T, pois tem-se dificuldades de implementação, tanto técnicas quanto organizacionais.

No Brasil, cada agência de C&T possui suas particularidades e algumas semelhanças no processo de avaliação, que serão discutidas mais adiante. De acordo com Guimarães (1994), a avaliação pode ser contextualizada quanto a agente avaliador, modo, universo e tempo.

Com relação ao avaliador, observa-se, principalmente, a avaliação efetuada por pares ou a utilização de grupos de especialistas não pesquisadores, dentro ou fora das agências. O modo dessas análises pode ser quantitativo ou qualitativo, sendo o processo qualitativo muito utilizado. O universo de conhecimento refere-se à unidade a ser

analisada (grupo de pesquisa, pesquisador, programa de pós-graduação, instituição, etc.).

O processo de avaliação de C&T brasileira ocorre principalmente nas agências de fomento (CNPq e FINEP, no âmbito do Ministério da Ciência e Tecnologia (MCT); CAPES, no Ministério da Educação (MEC); e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), esta a mais importante das agências estaduais). As agências cumprem um importante papel, pois são responsáveis pela maior parte do financiamento das atividades de C&T no país. Esse financiamento ocorre na forma de bolsas de iniciação científica, de pós-graduação ou na forma de auxílios à pesquisa, a congressos, entre outros.

a) CNPq

O sistema de avaliação do CNPq visa o acompanhamento e a avaliação de diversos subsistemas, tais como os de formação de recursos humanos (bolsas no País e no exterior), o de fomento à pesquisa, o de execução direta de pesquisas (Institutos), e o próprio desempenho institucional (CNPq, 1998). Esse sistema inclui a avaliação pelos Comitês Assessores (CAs), pelos pares *peer review system*, o uso de indicadores bibliométricos e a utilização de consultores *ad hoc*.

Os CAs são responsáveis pela avaliação do mérito dos pedidos de bolsas e auxílios dos programas de demanda espontânea. Esses processos, após passarem por uma pré-análise, são examinados por um consultor *ad hoc* que elabora um parecer quanto à validade da solicitação. Sendo assim, a tomada de decisão é predominantemente do comitê, e os pareceres servem como uma base para a tomada de decisão (Guimarães, 1994).

A avaliação por pares é o procedimento mais utilizado pela agência. Esse sistema de avaliação utiliza membros conceituados da comunidade científica que ficam encarregados das análises prévias de uma determinada área de conhecimento. Contudo, esse sistema recebe críticas, sendo a principal a questão de favorecimentos a pesquisadores mais conhecidos, em relação a novos pesquisadores, bem como às instituições mais tradicionais em relação às mais novas (Barbieri, 1993). Para Barbieri, a avaliação por pares, apesar de sofrer críticas, mostra-se mais adequada ao uso de indicadores quantitativos, uma vez que estes podem ser altamente

tendenciosos. Isso pode ocorrer em razão de autocitações e citações entre pesquisadores de uma mesma instituição. Esses indicadores devem atuar como um instrumento auxiliar ou complementar à avaliação dos pares.

Para Barbieri (1993), a junção da opinião do consultor *ad hoc* com a dos CAs representa um aspecto amplamente positivo. No intuito de melhorar o processo de avaliação durante a gestão de 1995-1998 foi proposto um conjunto de ações (CNPq, 1998), a saber:

- avaliar as atividades de apoio à formação de recursos humanos, sobretudo às modalidades de Doutorado, Doutorado Sanduíche e Pós-Doutorado. Esse processo visa avaliar e medir os resultados desses investimentos, tanto em termos de qualidade quanto de relevância para o desenvolvimento científico e tecnológico;
- utilizar a base de informação do Diretório dos Grupos de Pesquisa no Brasil objetivando desenvolver e implantar mecanismos de avaliação da pesquisa no país. Esse mecanismo é viabilizado através do conceito da *hierarquização dos grupos de pesquisa*;
- criar procedimentos adequados de acompanhamento e avaliação de projetos de pesquisa;
- estimular e promover processos de avaliação do desempenho institucional.

a.1) HIERARQUIZAÇÃO DOS GRUPOS DE PESQUISA

A hierarquização dos grupos de pesquisa tem como proposta fornecer um *ranking* dos grupos de pesquisa, atribuindo um grau de qualificação. Esses grupos são classificados em cinco estratos de "A" até "E". O estrato "A" determina os grupos de excelência, com alta participação de seus pesquisadores no sistema de bolsa de pesquisa do CNPq e no sistema de avaliação dos programas de pós-graduação da CAPES. À medida que se evolui nos estratos, diminui-se o número de pesquisadores qualificados. Para o cálculo dos estratos, utiliza-se o índice de qualificação (Q), sendo este uma estatística padronizada, com média e desvio-padrão iguais a 50 e 20, respectivamente (Guimarães et al., 1999).

Entre as variáveis de análise, pode-se citar, para o índice (Q), o número de bolsistas de pesquisa do CNPq, classificados nas categorias 1A, 1B, 1C, 2A, 2B, 2C, e o número de docentes vinculados a programas de pós-graduação e avaliados pela CAPES com grau igual ou superior a 3. Com relação ao índice (P), as variáveis de análise referem-se à produção C&T, sendo compostas por artigos publicados em periódicos, trabalhos em eventos, livros e capítulos, produção tecnológica, e teses e dissertações.

b) CAPES

As atividades do sistema de avaliação da pós-graduação foram iniciadas na década de 70, mais precisamente no ano de 76, e esse sistema é hoje considerado um modelo reconhecido entre os sistemas de avaliação existentes no país, em que a unidade de análise é o próprio Programa de Pós-graduação *stricto sensu*.

A avaliação envolve alguns pontos importantes no que diz respeito à pontuação, às escalas e à unidade de avaliação. A unidade de avaliação é o programa de pós-graduação e a escala de classificação é numérica entre 1 e 7, sendo que a nota superior a 5 somente é atribuída a programas que mantenham doutorado e mestrado. Essa avaliação fundamenta-se em quatro momentos (CAPES, 1999a):

- análise das Comissões de Áreas dos programas de pós-graduação e enquadramento destes em um dos cinco primeiros níveis;
- análise das Comissões de Áreas dos programas enquadrados anteriormente como nível 5, sendo a estes candidatos atribuídos os níveis 6 e 7, estabelecidos como “Perfil de um Curso A”. Nessa análise são levados em consideração (1) o desempenho na produção científica, cultural, artística ou tecnológica; (2) a competitividade com programas similares situados no exterior; e (3) a representatividade de seu corpo docente perante sua respectiva comunidade;
- apreciação pelo Conselho Técnico-Científico (CTC) dos resultados de cada área e posterior homologação dos resultados da avaliação;
- divulgação dos resultados.

Mesmo considerando a importância do sistema, existem pontos a serem melhorados. Isso pode ser verificado nos últimos estudos realizados pela agência, no sentido de projetar uma nova política (CAPES, 1999b) envolvendo objetivos, dos quais se podem destacar:

- projetar a pós-graduação nacional no contexto internacional da produção de conhecimento científico;
- ampliar a competência do sistema de pós-graduação, visando qualificar um número maior de profissionais em um espaço de tempo menor, respeitando as peculiaridades de cada área;
- reduzir as disparidades entre as regiões do país e entre as áreas do conhecimento;
- diversificar o sistema de pós-graduação, atendendo assim à crescente demanda de profissionais altamente qualificados, tanto no meio acadêmico quanto no não-acadêmico;
- diminuir a rigidez na relação seqüencial estabelecida entre o mestrado e o doutorado, permitindo maior independência;
- aumentar o grau de interação da pós-graduação com o sistema de terceiro grau, objetivando aumentar a qualidade desse sistema e reduzir o efeito corretivo das deficiências da graduação.

Portanto, a proposta pretende avaliar a capacidade de formação de recursos humanos dos programas de pós-graduação, em vez de avaliar os cursos por eles oferecidos. Essa avaliação permite à CAPES ter um mapa da atividade de pós-graduação no país, e sua distribuição quanto aos seus níveis (mestrado e doutorado), sendo que o processo na sua totalidade promove a consolidação, a qualidade e o desenvolvimento da pós-graduação nacional e a projeção desta em níveis internacionais.

c) FINEP

A Financiadora de Estudos e Projetos (FINEP) possui um sistema de avaliação de projetos que pode ser considerado estável e criterioso, com uma análise por vezes mais detalhada, em que dependendo do projeto, visitas ao local de desenvolvimento são realizadas pelo corpo técnico. A base desse sistema é formada pela interação entre o corpo técnico da agência e a comunidade científica no papel de consultora *ad hoc*, não

existindo, portanto, comitês com caráter permanente (Guimarães, 1994). Além disso, esse sistema baseia-se na análise de projetos de pesquisa científica ou tecnológica, e procura fornecer subsídios à tomada de decisão sobre para quais projetos devem ser despendidos recursos financeiros. A importância dessas avaliações pode ser verificada no nível dos projetos apoiados pela agência, entre eles: o avião Tucano da Embraer, o melhoramento genético de animais e alimentos, o AZT nacional, entre outros (FINEP, 1999).

O processo de solicitação de recursos e aprovação de determinado projeto, em linhas gerais, acompanha os seguintes passos: primeiramente, o postulante apresenta uma solicitação de financiamento. Sobre essa solicitação é realizada a avaliação, que deve estar previamente justificada tanto em nível técnico quanto financeiro. Em seguida, as solicitações são enviadas a consultores que emitem um parecer sobre a relevância do projeto. De posse desses pareceres, o processo pode ser encerrado, podendo ou não serem efetuadas visitas ao local de desenvolvimento do projeto. Esse método é considerado uma ferramenta mais eficiente do que simplesmente os pareceres (FINEP, 1999).

Em linhas gerais, o sistema mostra-se eficiente quanto à análise, principalmente às visitas, permitindo que se obtenha uma real contextualização do projeto, bem como sua relevância. Contudo, existem problemas no processo de avaliação da FINEP. De acordo com Guimarães (1994), o mais importante é a falta de normas e de uma memória institucional formalizada, informações estas coletadas com técnicos da agência.

d) FAPESP

O sistema de avaliação da FAPESP está sedimentado na análise de propostas que se enquadram em um de seus programas, por meio dos quais estas propostas são avaliadas segundo sua relevância científica ou tecnológica. A avaliação é realizada por assessores escolhidos entre cientistas de reconhecida competência (FAPESP, 1999), sendo estes pesquisadores atuantes no Estado de São Paulo (mais de 6 mil assessores), e em outras regiões do país e exterior.

O sistema de avaliação da FAPESP baseia-se na revisão por pares e pode ser configurado em três etapas principais. A primeira é efetuada pelo próprio solicitante, que enquadra o pedido em uma das áreas de conhecimento (Ciências Biológicas, Ciências da Saúde, Ciências Exatas e da Terra, Engenharias, Ciências Agrárias, Ciências Sociais Aplicadas, Ciências Humanas, Lingüística, Letras e Artes) e encaminha a solicitação à coordenação da área. Em seguida, a coordenação da área escolhe, para cada projeto, um pesquisador que avalia o mérito científico e tecnológico através de um parecer *ad hoc*. Esse parecer retorna à coordenação da área, que pode ou não aceitá-lo. Em caso de aceite, o projeto retorna à diretoria científica, que o agrupa em três grandes áreas de conhecimento (Ciências da Vida, Exatas e da Natureza, e Humanas e Sociais), encaminhando-o para outro grupo de pares (Guimarães, 1994). Essa etapa tem por objetivo compatibilizar divergências ou minimizar incoerências.

Como afirma Guimarães (1994), as principais deficiências desse sistema estão relacionadas à escolha dos consultores, uma vez que a comunidade científica é relativamente pequena, o que torna restrito o número de consultores qualificados de determinada área e dificulta um acompanhamento que compatibilize o *gap* entre o tempo da produção de conhecimento e a duração dos auxílios e bolsas.

e) MINISTÉRIO DE C&T (MCT)

O Ministério de C&T (MCT) tem concentrado esforços na captação e disponibilização de recursos destinados ao desenvolvimento científico e tecnológico do país. Um desses esforços é Programa de Apoio ao Desenvolvimento Científico e Tecnológico (PADCT). Essa é uma iniciativa para fomentar a C&T nacional, mas é também um processo de avaliação e de acompanhamento dos projetos aos quais serão destinados recursos. O PADCT é um programa do governo brasileiro, administrado pelo Ministério de Ciência e Tecnologia e operacionalizado pelas agências executoras FINEP, CNPq e CAPES, além de outras Agências (MCT, 1998b). Esse programa tem por objetivos (1) contribuir para a ampliação da capacidade tecnológica nacional, (2) promover a capacitação dos recursos humanos e (3) contribuir para o melhor desempenho global do setor de ciência e tecnologia (MCT, 1998b).

O processo de acompanhamento e avaliação do PADCT será coordenado pela Secretária de Coordenação e Gerenciamento (SCG) em três níveis: Projeto, Subprograma e Programa (MCT, 1998b).

1. *Nível de Projeto*

O coordenador do projeto deve preparar um relatório semestral, devidamente padronizado, que servirá de referência para a avaliação das agências. As agências executoras, juntamente com consultores *ad hoc*, são responsáveis pela avaliação, cabendo à agência a responsabilidade de acompanhamento dos projetos, e aos consultores, a avaliação do projeto.

O processo envolvendo as agências resulta no acompanhamento da aplicação dos recursos, na aquisição dos equipamentos, no envolvimento dos recursos humanos e na identificação de possíveis correções na execução dos projetos sobre aspectos financeiros. Para tal, as agências utilizam um relatório semestral, fornecido pelo coordenador do projeto, e o fluxo financeiro. Por outro lado, é de responsabilidade dos consultores o acompanhamento do cumprimento das metas e objetivos do projeto, a identificação de eventuais correções no âmbito financeiro e a elaboração do relatório semestral de avaliação. Para a realização dessas tarefas, os consultores utilizam-se de subsídios das agências e visitas técnicas anuais, realizadas em conjunto com representantes das próprias agências.

Os resultados apurados na avaliação devem ser consolidados em relatórios trimestrais encaminhados à SCG.

2. *Nível dos Subprogramas*

Neste nível são avaliados os componentes (Ciência e Tecnologia, Desenvolvimento Tecnológico, Suporte Setorial) e subcomponentes (Subprogramas do Componente de Ciência e Tecnologia) do PADCT, sendo estes de responsabilidade dos Grupos de Planejamento e Avaliação (GPAs), Grupos de Planejamento e Gestão Estratégica dos Programas Cooperativos (GPGEs), e Grupo Técnico de Coordenação (GTC), abrangendo os seguintes aspectos:

- acompanhar e realizar as metas e objetivos constantes nos Documentos Básicos dos Subprogramas elaborados e aprovados para o PADCT III, juntamente com os indicadores de desempenho estabelecidos no Documento Básico do Programa;
- acompanhar a execução orçamentária e financeira dos subcomponentes;
- avaliar os resultados obtidos;
- identificar eventuais alterações nas diretrizes dos subcomponentes.

Os grupos avaliadores utilizam relatórios semestrais fornecidos pelas agências executoras, e como resultado do processo são elaborados relatórios anuais, encaminhados à Secretária de Coordenação e Gerenciamento do PADCT.

3. *Nível de Programas*

A Secretaria de Coordenação e Gerenciamento (SCG) do PADCT, em termos globais, é responsável pelo acompanhamento e pela avaliação do programa, consolidando as avaliações das agências e do colegiado, bem como a utilização de outros instrumentos, tais como seminários, reuniões técnicas e consultorias especializadas, abrangendo os seguintes aspectos:

- acompanhamento dos objetivos e metas gerais do PADCT;
- acompanhamento e execução orçamentária;
- avaliação do desempenho do programa;
- avaliação da contribuição do programa para a realização dos objetivos do setor de C&T;
- avaliação dos impactos e inovações em comparação à produção científica e tecnológica nacional.

f) UNIVERSIDADES

Um sistema de avaliação para universidades pode ser entendido como uma ferramenta que auxilie na tomada de decisão e avaliação de departamentos acadêmicos, medindo tanto a produtividade quanto a qualidade dos serviços prestados e, em níveis gerais, que forneça um dimensionamento dos recursos destinados a cada departamento. Como exemplo, pode-se citar um modelo de produtividade e qualidade de

departamentos acadêmicos proposto por Lopes (1998). Em cursos de pós-graduação, instrumentos de avaliação podem configurar-se em itens estratégicos para o desempenho do Programa, principalmente considerando o modelo nacional estabelecido pela CAPES. Isso já tem sido adotado em programas de pós-graduação do país. Como exemplo, pode-se citar o sistema para a determinação de *ranking*, possibilitando a distribuição de vagas ao corpo docente do Programa de Pós-graduação em Engenharia de Produção da Universidade Federal de Santa Catarina, chamado SisCoor (PPGEP, 1999).

O modelo de avaliação cruzada da produtividade proposto por Lopes (1998) mede o desempenho entre departamentos, utilizando para isso a *Análise Envoltória de Dados* (DEA). Nessa avaliação é considerado um conjunto de indicadores de desempenho e qualidade. Esses indicadores são formados por uma estrutura de insumo (número de docentes em tempo integral) e produtos (indicadores de produtividade em ensino, pesquisa, extensão e qualidade), baseados em pesos definidos pelos departamentos. Como resultado final, o modelo apresenta uma medida de excelência de cada departamento e fornece subsídios para que o órgão central da Universidade avalie cada departamento, ou seja, identifique quais departamentos precisam de maior ou menor atenção.

O SisCoor (PPGEP, 1999) é um sistema destinado a estabelecer o *ranking* no Programa de Pós-Graduação de Engenharia de Produção da UFSC. Esse sistema divide-se em duas etapas principais: a etapa de coleta das informações e a etapa de cálculo e publicação dos resultados. Na primeira etapa, os dados considerados pelo SisCoor são originados do sistema de Currículo Lattes e Pró-Coleta Professor, ou seja, o professor deve primeiramente preencher seus dados curriculares e específicos para a pós-graduação e enviá-los à coordenação do curso. De posse desses dados, o sistema calcula o *ranking* levando em consideração a produção C&T (por tipo de produção), as defesas de mestrado e doutorado realizadas no período de análise, os cargos administrativos exercidos e um bônus adicional para os recém-professores[1]. Após esta análise, é determinada a quantidade de vagas para mestrado e doutorado de cada professor no programa.

[1] Atribui-se a mediana da pontuação obtida pelo conjunto de professores durante o período de avaliação do recém-professor.

2.4 BASES DE CIÊNCIA E TECNOLOGIA NO BRASIL

As bases de C&T demonstram os esforços das agências de fomento em manter um registro da memória de C&T no Brasil. As principais bases de dados hoje em operação descrevem a situação da ciência no país em três unidades: indivíduo com a base de currículos, integrante da Plataforma Lattes (CNPq), grupos de pesquisa com a base do Diretório dos Grupos de Pesquisa no Brasil (CNPq) e programa de pós-graduação, com a base da Plataforma Coleta (CAPES).

2.4.1 DIRETÓRIO DOS GRUPOS DE PESQUISA NO BRASIL

O Projeto Diretório é uma base que descreve o comportamento dos grupos de pesquisa no Brasil, tendo iniciado em 1992. Essa base é constituída pelos principais grupos de pesquisa no país, por seus integrantes e pela produção científica desses grupos. Em suma, o Diretório preocupa-se com a captação dos dados relativos aos grupos de pesquisa no país. O grupo de pesquisa pode ser visto como a união de um ou mais indivíduos, organizados hierarquicamente, com objetivos em determinadas linhas de pesquisa e sob a liderança de um ou dois líderes (Guimarães et al., 1999).

O Diretório possui três finalidades principais (Guimarães et al., 1999):

- ser um eficiente instrumento para o intercâmbio e a troca de informações, permitindo identificar com rapidez a posição do indivíduo dentro do grupo;
- ser uma ferramenta poderosa destinada ao planejamento e gestão das atividades de C&T;
- possuir um papel importante na preservação histórica da ciência e tecnologia no Brasil.

A base do Diretório pode ser analisada segundo algumas abordagens, entre elas: o próprio grupo de pesquisa, as linhas de pesquisa, os pesquisadores, estudantes, técnicos, áreas do conhecimento e a produção C&T. Nesse primeiro ponto pode-se verificar as principais características dos grupos, tais como a instituição à qual estes pertencem, a área do conhecimento à qual estão vinculados e a produção científica de seus integrantes. As linhas de pesquisa identificam a atuação desses grupos. Quanto aos

indivíduos, pode-se verificar principalmente a formação acadêmica, bem como sua produção científica.

Desde a sua primeira versão, ou o primeiro censo, o número de grupos duplicou. Sendo inicialmente 4.241 grupos em 99 instituições, esse número atingiu 7.271 grupos em 158 instituições na versão 2.0 (1995), e 8.544 grupos em 181 instituições na versão 3.0, resultando num aumento de 17,5% em relação à versão anterior (Guimarães et al., 1999).

Atualmente, o projeto encontra-se na versão 4.0, sendo que a principal diferença em relação às versões anteriores está no armazenamento da produção científica, que agora faz parte da base de currículos do CNPq (Plataforma Lattes). A Tabela 2.5 apresenta um histórico do projeto, desde a sua primeira versão.

Tabela 2.5 - Evolução histórica do projeto Diretório dos Grupos de Pesquisa no Brasil

Unidade / Ano	1993	1995	1997	2000
Instituições	99	158	181	224
Grupos	4.404	7.271	8.632	11.760
Linhas de Pesquisa	15.854	21.523	25.483	41.539
Pesquisadores	21.541	26.779	34.040	48.781
Estudantes	33.565	61.345	115.696	60.225

Fonte: CNPq

2.4.2 PLATAFORMA LATTES

Com a decisão de ampliar o projeto de integração dos sistemas de informações surgiu a Plataforma Lattes, resultado do esforço conjunto de vários órgãos, entre eles MCT, CNPq, FINEP e CAPES/MEC.

A Plataforma Lattes procurou atender a alguns objetivos básicos:

- implantar alterações e ajustes no sistema de currículos, solicitados pelos consultores *testers*, que entre março e abril de 1999 avaliaram o produto cujo conteúdo ajudaram a definir no ano anterior;
- ampliar o conjunto de informações e adaptá-las de forma a permitir a adesão da CAPES ao projeto de integração dos sistemas, prevista para o segundo semestre de 1999;

- estabelecer critérios de avaliação da qualidade das informações coletadas junto à comunidade científica do país. Historicamente, os sistemas eletrônicos de currículo não apresentavam essa preocupação, permitindo o cadastro livre de uma série de informações essenciais à análise (ex.: autoria, palavras-chave, etc.). Além disso, o conteúdo solicitado por esses sistemas atendia mais a uma demanda operacional das agências do que às solicitações dos comitês assessores;
- integrar as bases de dados, melhorando o fluxo de informações para pesquisadores, instituições, agências de fomento e órgãos do governo.

Para alcançar esses objetivos, foi necessário integrar os esforços da Superintendência de Informática (SIN), da Superintendência de Planejamento (SUP), da Coordenação de Planejamento, Acompanhamento e Avaliação (COAV) e da Direção de Administração do CNPq, que articularam com as demais áreas da agência a visão integrada do conjunto de informações. Além disso, foi preciso interagir com a CAPES para tornar possível a compatibilização entre a Plataforma Lattes e os sistemas da CAPES.

A Plataforma Lattes compõe-se de um conjunto de sistemas que promove suporte à captação e manutenção dos dados curriculares dos pesquisadores no país, dividindo-se em vários sistemas responsáveis, desde o preenchimento dos dados curriculares pelo pesquisador, por meio do Sistema de Currículo Lattes, passando pelos sistemas de recepção dos dados e os sistemas de controle dentro da agência (CNPq).

O Sistema de Currículo Lattes é o formulário eletrônico responsável pela coleta das informações que servem de apoio na descrição da pesquisa no país ao nível do indivíduo. Essas informações são, em geral, originadas de pesquisadores ou de usuários do CNPq, que requisitam recursos como bolsas ou auxílios para projetos de pesquisa. De um modo geral, a plataforma fornece subsídios ao incremento e manutenção da base de dados curriculares do CNPq.

Portanto, este projeto tem como uma de suas principais finalidades efetivar o potencial fornecido através da estruturação da infra-estrutura dos currículos no CNPq, de forma a possibilitar a construção de um repositório comum de currículos a todos os agentes institucionais de pesquisa e desenvolvimento de C&T no Brasil.

2.4.3 PLATAFORMA COLETA/CAPES

A Plataforma Coleta tem por finalidade captar dados oriundos dos programas de pós-graduação (*stricto sensu*) do país, possibilitando assim a avaliação e posterior classificação destes. Por programa de pós-graduação entende-se a denominação atribuída aos cursos de mestrado acadêmico, mestrado profissionalizante e de doutorado em uma Instituição de Ensino Superior (IES) atuantes numa mesma área do conhecimento (CAPES, 1999c). Os dados coletados nesse processo são necessários à avaliação dos programas de pós-graduação, constituindo as informações consolidadas sobre o Sistema Nacional de Pós-Graduação (SNPG).

O modelo que comporta essa base de dados é um dos mais completos entre as bases de C&T e possui uma riqueza nas informações que permitem análises importantes sobre o perfil de C&T no país, embora, como a própria agência menciona, esses dados não possuem um caráter censitário.

Assim sendo, a base de dados na sua totalidade é o suporte para um conjunto de aplicativos computacionais, que visam desde o suporte ou a captação dos dados junto aos programas até a avaliação dos cursos. A plataforma de sistemas é composta pelas seguintes aplicações (CAPES, 1999c):

- uma aplicação responsável pela coleta de dados junto aos programas (Coleta);
- uma aplicação fornecida à pró-reitoria que permite a utilização dos dados captados pela IES, realizando consultas e emitindo relatórios (Coleta Reitoria);
- uma aplicação para integração dos dados captados pelas IESs na CAPES, possibilitando consultas e emissão de relatórios;
- uma aplicação que possibilita classificar os veículos de divulgação e eventos para fins de avaliação;
- uma aplicação que consolida e sintetiza as informações necessárias à avaliação dos programas;
- uma aplicação de acompanhamento, permitindo administrar o processo de avaliação;

- e uma aplicação para avaliação de cursos, utilizada pelas comissões que permite a atribuição dos conceitos aos programas.

Nos últimos anos, a evolução tecnológica dos ambientes computacionais colocou em xeque a interface utilizada na Plataforma Coleta e, principalmente, a lógica de captura de informações com os professores dos programas de pós-graduação. Esse modelo pode configurar-se de duas maneiras (Figura 2.2), descritas a seguir:

(a) os professores preenchem os dados em papel (formulário gerado pelo sistema Coleta) e entregam ao coordenador para digitação;

(b) ou cada professor registra suas informações no sistema Coleta. O coordenador aguarda o final da atualização e confere o preenchimento.

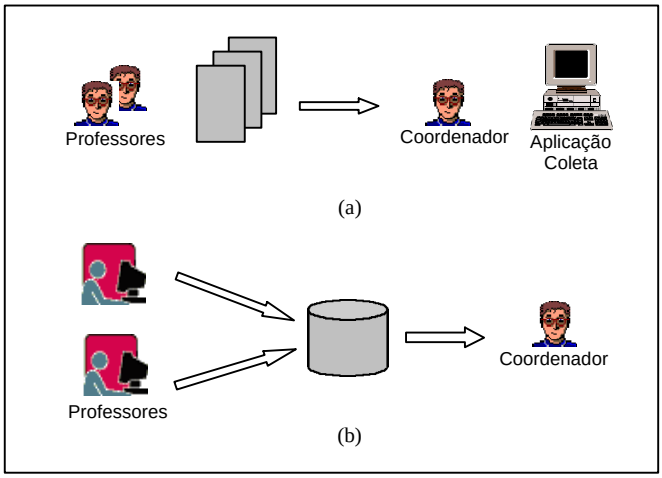


Figura 2.2 - Modelo de captura de informações nos programas de pós-graduação

Ambas as alternativas têm sido muito criticadas pela comunidade científica, devido à carga operacional que causa aos programas de pós-graduação, surgindo a necessidade clara de uma integração entre a visão do professor e a visão do coordenador.

2.5 INTEGRAÇÃO DAS PLATAFORMAS COLETA E LATTES

A integração das plataformas Lattes e Coleta representa um primeiro passo na unificação das bases de C&T do país e tem por objetivo maior permitir a troca de informações, disponibilizando tais informações de maneira unificada, utilizando-se de

um ambiente semelhante em qualquer ponto, ou seja, em qualquer uma das agências (CNPq, CAPES, FINEP, etc.).

Essa interação é uma antiga solicitação da comunidade científica, tornando-se uma realidade pela vontade política e administrativa do Ministério da Ciência & Tecnologia (MCT) e principalmente pela colaboração das agências, entre elas CNPq, FINEP e CAPES. O resultado é a disponibilização de informações de maneira integrada, e de um veículo que tem o forte potencial de tornar-se definitivo em sua intenção de ser uma fonte única de captura de dados junto aos agentes individuais da pesquisa e do desenvolvimento científico e tecnológico do país (Pacheco e Barcia, 1999).

A Figura 2.3 demonstra o fluxo de informações e os processos que afetam a relação entre os agentes de desenvolvimento científico-tecnológico em termos de programa de pós-graduação. Tomando-se a pós-graduação como centro desse relacionamento, pode-se notar a solicitação pela integração vinda da comunidade acadêmica do país. Coordenadores de programas de pós-graduação em especial vêem-se na obrigatoriedade de atender a diversas solicitações da instituição de ensino e das agências do governo. Ao mesmo tempo, essas agências (em especial CNPq e Fundações de Apoio à Pesquisa) solicitam informações dos integrantes (docentes, pesquisadores e discentes) das pós-graduações, em sistemas diferentes. O coordenador, no entanto, não pode utilizar os arquivos formados pelos integrantes da pós-graduação e duplica o pedido, com novos formulários (no caso da CAPES, em papel). O resultado leva a uma significativa perda de recursos, redundância nas informações e impossibilidade de comparações entre as diversas fontes de informação.

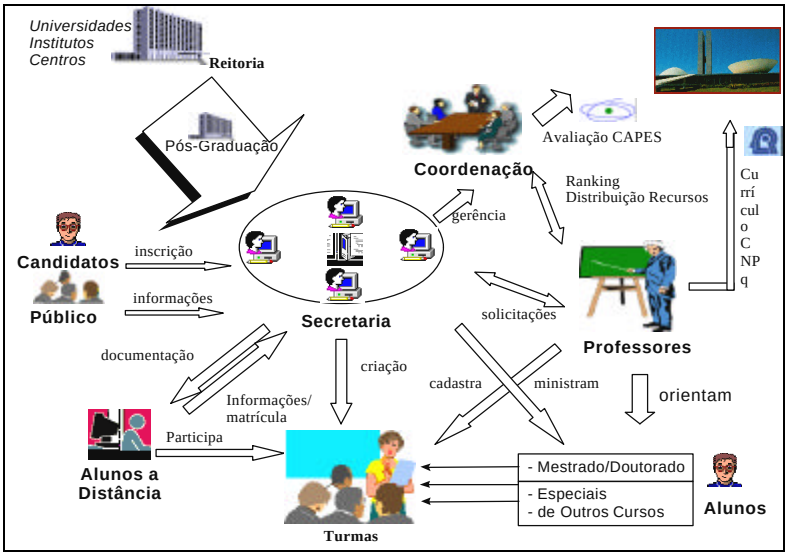


Figura 2.3 - Fluxo de informações e processos entre agências, instituições e usuários clientes (Pacheco e Barcia, 1999)

A integração constitui um primeiro passo no intuito de reunir duas das principais bases de C&T do país e possibilitar o desenvolvimento de um ambiente comum às principais agências de C&T. O *gateway* entre as duas plataformas é realizado pelos sistemas Pró-Coleta (Professor e Coordenador). O sistema Pró-Coleta Professor captura as informações adicionais do professor, que não são requeridas pelo CNPq, mas são necessárias à avaliação dos cursos de pós-graduação. A partir dessas informações o segundo sistema, Pró-Coleta Coordenador, integra todos os dados à plataforma Coleta. Nesse mesmo sistema é possível gerar um currículo individual servindo novamente de interface com o Currículo Lattes. Isso produz um fluxo comum de informações, permitindo que as plataformas (Lattes e Coleta) comuniquem-se e disponibilizem informações de maneira consistente e confiável.

A perspectiva de integração dos instrumentos das agências amplia potencialmente o conjunto de aplicações de extração de conhecimento a partir de bases de dados de C&T. Neste trabalho, a compatibilização de modelos entre a Plataforma Lattes e os sistemas Coleta da CAPES permitiram comparar dados da pesquisa brasileira organizada por grupos de pesquisa com a pós-graduação nacional (como será apresentado no capítulo 4).

A experiência de construção dos sistemas Pró-Coleta e a estruturação da base de currículos do CNPq como um superconjunto das diversas visões individuais de

informação fornece os recursos para diversas análises exploratórias das bases em C&T nacionais.

2.6 CONSIDERAÇÕES FINAIS

Este capítulo abordou aspectos da Ciência e Tecnologia de interesse para os objetivos do trabalho. O estudo mostra a formação histórica da C&T brasileira fundamentada no fomento público, baseado em agências federais (CAPES, FINEP, CNPq) e estaduais (ex.: FAPESP), e apresenta as principais bases de C&T em operação no país, formadas nos últimos 50 anos. Essas bases (pesquisadores, grupos de pesquisa e programas de pós-graduação) representam todas as áreas do conhecimento, e se concentram em grande parte na região Sudeste.

O processo de auxílio conduzido pelas agências levou à constituição de sistemas de avaliação e diferentes plataformas computacionais. O principal mecanismo de avaliação tem sido a avaliação por pares e, dependendo da agência, com algum auxílio de ferramentas cientométricas quantitativas. A multiplicidade de sistemas, embora tenha criado ricas bases de dados, tem prejudicado a qualidade da informação e gerado descontentamento na comunidade científica pela falta de racionalização de recursos.

Mais recentemente, surgiram esforços de plataformas integradas procurando minimizar o esforço de captura dos dados necessários, tanto na avaliação quanto no fomento. Essas plataformas estão formando bases de dados que permitem ampliar o conjunto de ferramentas de análises utilizadas atualmente, e que é o principal objetivo deste trabalho. Para tal, no próximo capítulo aprofundam-se as técnicas de Mineração de Dados e extração de conhecimento em bases de dados.

3 MINERAÇÃO DE DADOS

“A inteligência é uma espécie de paladar que nos dá a capacidade de saborear idéias”

Susan Sontag

3.1 INTRODUÇÃO

Os avanços da chamada “era da informação” têm colocado como desafio a implementação de técnicas que consigam mensurar e descobrir padrões relevantes na crescente massa de dados, resultante, sobretudo, do aumento da complexidade das tarefas operacionais e decisórias. Essa explosão nos dados pode determinar a sobrevivência ou não de uma organização, desde que esta consiga extrair informações úteis à tomada de decisão e à melhoria nos processos operacionais.

Essa visão da análise de dados baseia-se em duas novas tendências: uma avalanche de informações e um questionamento sobre esses dados (Hair et al., 1998). Nesse enfoque, uma nova perspectiva é apresentada, em que a análise de dados é vista com um caráter exploratório. De maneira mais abrangente, encontra-se *Knowledge Discovery in Database* (KDD), sendo esta a designação para o processo que envolve a seleção, o pré-processamento e a transformação dos dados, bem como a aplicação de algoritmos, a interpretação dos resultados e a geração de conhecimento (Figura 3.1 Fayyad, et al., 1996a).

Dentro desse processo encontra-se *Data Mining,* ou Mineração de Dados (MD), que é uma etapa no processo de KDD, responsável pela aplicação dos algoritmos com a finalidade de identificar padrões *Ej* sobre uma base de dados *F* (Fayyad, 1996b), ou gerar um conjunto de regras que descrevam o comportamento de uma base de dados.

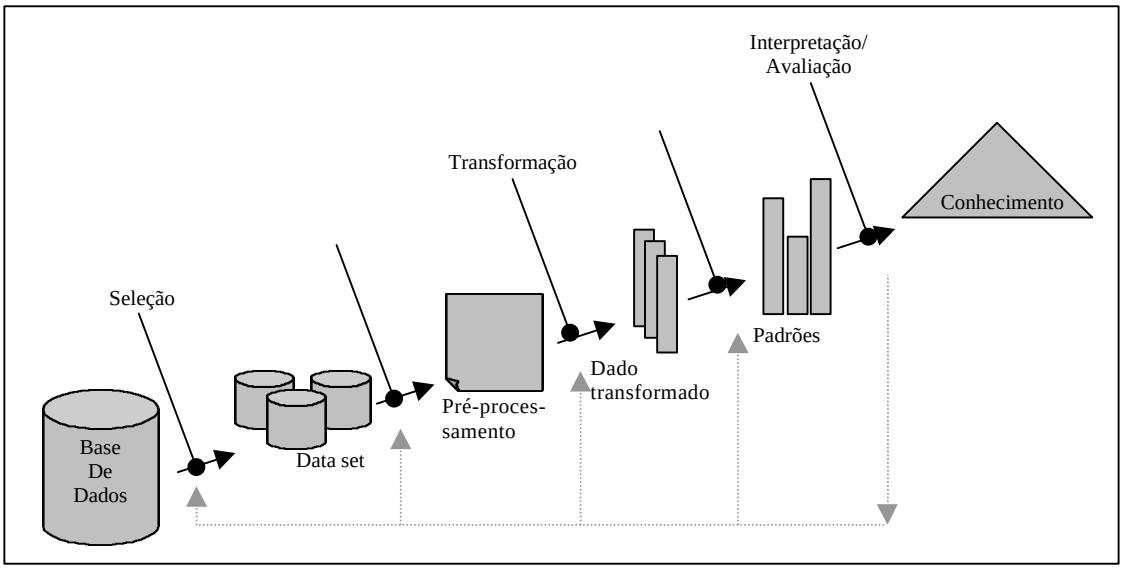


Figura 3.1 - Processo de KDD. Adaptado de Fayyad et al. (1996a)

A proposta de MD é proporcionar uma perspectiva nova ou, mais precisamente, uma evolução nos processos de análise, permitindo a descoberta de novos padrões ou a validação de padrões conhecidos. Tais análises são geralmente efetuadas em grandes quantidades de dados. Segundo Harrison (1998), MD é a exploração e análise, por meios automáticos ou semi-automáticos, de grandes quantidades de dados para descobrir modelos e regras significativas.

Na visão de Berry e Linoff (1997), existem metas primárias para um sistema de mineração de dados:

- previsão: envolve a utilização de algumas variáveis (atributos da base de dados) para prever valores desconhecidos ou futuros de outras variáveis de interesse;
- descrição: procura por padrões que descrevam os dados e que sejam interpretáveis.

Levando-se em conta tais metas, pode-se observar a real importância desse processo dentro de uma organização, sendo que essas metas devem prever algumas fases. Harrison (1998) identifica quatro fases, entre elas:

- identificar problemas e áreas para as quais a análise de dados pode fornecer valor;

- transformar dados em informações acionáveis, usando técnicas de mineração de dados;
- agir sobre a informação e, a partir dela, melhorar os processos que regem o relacionamento da empresa com seus consumidores e fornecedores; e
- medir os resultados dos esforços para fornecer idéias sobre como explorar os dados. Esta fase proporciona o *feedback* para o aumento constante na qualidade dos resultados.

O processo de descoberta de conhecimento envolve algumas etapas, entre elas a definição do domínio da aplicação, a limpeza e o pré-processamento dos dados, a representação dos dados, a mineração de dados e a interpretação dos resultados.

Primeiramente, deve-se definir o problema e as metas desejadas pelo usuário. Tal fase deve preocupar-se com critérios de desempenho, gargalos no domínio da aplicação e a interoperabilidade com o usuário final.

Na etapa seguinte, são realizados o pré-processamento e a limpeza dos dados pela remoção de ruídos ou dados inválidos que atrapalhem o processamento, bem como são adotadas estratégias para manusear campos que apresentem dados perdidos.

Em seguida, a representação dos dados procura modelá-los de maneira que possam ser utilizados por algum algoritmo de extração de conhecimento. Como exemplo, pode-se citar a transformação de valores lingüísticos em valores numéricos dentro de um domínio ou a transformação de valores contínuos para discretos.

A etapa de mineração de dados propriamente dita constitui-se na busca em uma base de dados por informações relevantes e que não sejam facilmente identificáveis. De acordo com Fayyad (1996b), a busca é realizada em três etapas: primeiramente, decide-se se o processo será de classificação, agrupamento ou sumarização; em seguida, escolhe-se um dos métodos a serem utilizados na busca por padrões; e, por último, efetua-se o processo de busca ou a mineração dos dados. Uma grande variedade de técnicas analíticas tem sido utilizada em mineração de dados, técnicas que vão desde as tradicionais da estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos. A Tabela 3.1 demonstra as principais funções da mineração de

dados, algoritmos utilizados e exemplos de aplicações. A escolha de uma ou outra função depende essencialmente do negócio, da aplicação e da quantidade e qualidade dos dados disponíveis.

Tabela 3.1 - Funções da mineração de dados (Bigus, 1996)

Funções	Algoritmos	Aplicações
Associação	Estatística, teoria dos conjuntos	Análise de mercados
Classificação	Árvores de decisão, redes neurais, algoritmos genéticos	Controle de qualidade, avaliação de riscos
Agrupamentos	Redes neurais, estatística	Segmentação de mercado
Modelagem	Regressão linear e não linear, redes neurais	*Ranking* de clientes, controle de processos, modelos de preços
Previsão de séries temporais	Estatística, redes neurais	Previsão de vendas, controle de inventário
Padrões seqüenciais	Estatística, teoria dos conjuntos	Análise de mercado sobre o tempo.

Por último, a etapa de interpretação dos dados verifica a validação do conhecimento extraído da base de dados, apresentando este conhecimento de maneira mais simplificada por meio de gráficos, tabelas e regras. O conhecimento extraído pode ser validado através de métodos estatísticos ou pelo parecer de um especialista.

3.2 TÉCNICAS DE MINERAÇÃO DE DADOS

Tem-se percebido um grande crescimento, tanto na elaboração e aperfeiçoamento das técnicas de mineração de dados quanto na utilização destas nas mais variadas áreas. A classificação pode ser representada pelas técnicas utilizadas ou, de maneira mais abrangente, como proposta por Chen et al. (1996), um sistema de mineração de dados pode ser classificado de acordo com os seguintes critérios:

- *tipos de base de dados*: os sistemas de mineração de dados podem ser classificados segundo o tipo da base de dados em que estão sendo executados, ou seja, se esse sistema é um minerador de dados relacional

quando executado sobre uma base de dados relacional, ou um minerador de dados orientado a objetos se executado sobre uma base orientada a objetos;

- *tipos de conhecimento*: existem dois modelos a considerar. Um preocupado com o conhecimento propriamente dito, incluindo regras de associação, regras de classificação, agrupamentos, e outro preocupado com o nível de abstração do conhecimento descoberto;
- *tipos de técnicas*: a escolha da técnica está fortemente relacionada com o tipo de conhecimento que se deseja extrair ou com os dados nos quais se aplicam tais técnicas. Entretanto, nota-se uma visão mais genérica, em que as técnicas são caracterizadas em mineração baseada na generalização, em padrões e na estatística.

De maneira geral, é comum visualizar a mineração de dados pelas técnicas de extração de conhecimento. Entre essas técnicas pode-se citar: regras de associação, árvore de decisão, redes neurais, lógica difusa, algoritmos genéticos, estatística multivariada e mesmo consultas *ad hoc*. Contudo, no presente trabalho serão abordadas somente as técnicas utilizadas, regras de associação e redes neurais. A fim de promover um maior entendimento da rede utilizada (Kohonen), será feita uma breve introdução sobre essa técnica de Inteligência Artificial.

3.2.1 REDES NEURAIS ARTIFICIAIS

Na mesma década em que foram criados os computadores, iniciaram-se os primeiros estudos sobre como emular a inteligência humana. Duas correntes destacaram-se nos anos 50: Inteligência Artificial Simbólica e Inteligência Artificial Conexionista. A IA Simbólica preocupava-se em simular o raciocínio dedutivo no computador, ou seja, o objetivo de implementar a manifestação de inteligência. Por outro lado, a IA Conexionista (Redes Neurais Artificiais - RNAs) nasceu das pesquisas por modelos que simulassem o funcionamento fisiológico do cérebro, para reproduzirem a inteligência.

Após mais de dez anos de descrédito, por estarem limitadas a problemas lineares (Minsky e Papert, 1969), ressurgiram no início dos anos 80 e, desde então, têm sido

utilizadas em uma ampla variedade de problemas. Mais recentemente, com o crescimento da utilização de técnicas de extração de conhecimento a partir de bases de dados, as RNAs têm sido empregadas na elucidação de informações contidas em bancos de dados.

Na avaliação de C&T, o uso de RNAs é pouco difundido, em grande parte pela tradição da aplicação de métodos estatísticos e algoritmos específicos. Contudo, a utilização de RNAs e, principalmente, de arquiteturas que permitam determinar agrupamentos mostram-se úteis no processo de descoberta de conhecimento necessário à tomada de decisão e avaliação de C&T.

a) CARACTERÍSTICAS

As Redes Neurais Artificiais (RNAs) são uma das técnicas de IA que procuram simular a inteligência humana. Segundo Fausset (1994), RNAs são modelos computacionais implementados em software ou hardware, que visam simular o comportamento dos neurônios biológicos por meio de um grande número de elementos de processamento interconectados – os neurônios artificiais.

Essa técnica possui algumas características relevantes na concepção de uma grande variedade de aplicações (Pandya et al., 1995), entre elas:

- *adaptabilidade*: algoritmos de aprendizagem e regras auto-organizáveis permitem a adaptação em ambientes dinâmicos;
- *processamento não-linear*: habilidade de executar tarefas que envolvam relacionamentos não-lineares e tolerantes a ruídos tornam as RNAs uma boa técnica para classificação, predição e agrupamentos;
- *processamento paralelo*: o grande número de unidades de processamento promove vantagens para o armazenamento de informações distribuídas, bem como para o processamento paralelo.

Além das RNAs, outras técnicas de IA possuem forte influência da biologia e do comportamento humano. Entre elas, pode-se citar os Algoritmos Genéticos, sendo esta uma técnica de busca baseada nos mecanismos de seleção e genética natural (Goldberg,

1989), e a teoria dos Conjuntos Difusos, definida por Zadeh (1965), que visa modelar conceitos incertos e vagos pertinentes às decisões humanas.

b) O NEURÔNIO BIOLÓGICO E O ARTIFICIAL

O sistema nervoso central é composto por mais de 100 bilhões de neurônios e um total estimado de 100 trilhões de conexões (sinapses). Levando-se em conta essas afirmações, pode-se verificar a complexidade do cérebro humano e, em conseqüência, a eficiência que essa estrutura pode alcançar, visto que ainda não existem mecanismos que processem eventos à taxa de 10^{-3} milissegundos (Pandya et al., 1995).

Na visão individual, encontram-se os neurônios que são formados pelo corpo celular ou soma, o axônio e os dendritos (Figura 3.2a). O corpo celular ou soma possui o núcleo celular e pode variar de forma e tamanho. Alguns podem ser gigantescos com um diâmetro de 0,5 mm ou reduzido com um diâmetro de 2 mícrons (Kovács, 1997). O axônio é um filamento que funciona como um canal de saída dos estímulos (informações) para as várias partes do sistema nervoso e do organismo. A terceira parte é composta pelos dendritos. Estes se encontram organizados em complexas árvores dendritais e possuem amplas ramificações para a recepção de informações através das sinapses. As sinapses ocorrem na chamada fenda sináptica, que é uma região entre a membrana pré-sináptica e a membrana pós-sináptica, através dos neurotransmissores que propagam a informação.

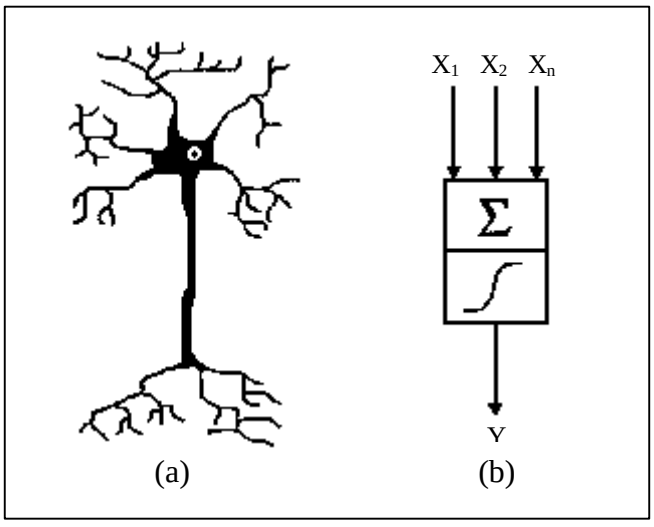


Figura 3.2 - Modelo de neurônio biológico (a) e artificial (b) (Medler, 1998)

O neurônio artificial (Figura 3.2b) tem sua inspiração no modelo natural, tentando imitar o seu funcionamento por meio de funções matemáticas e algoritmos computacionais. O neurônio artificial é um modelo simplificado, visto a complexidade e a não compreensão de todo o sistema nervoso. De uma maneira análoga, os dendritos do neurônio artificial são representados pelo conjunto de conexões interligando os diversos neurônios artificiais. O soma, no neurônio artificial, é representado pelo somatório das entradas e por uma função de transferência que é ativada toda vez que este somatório ultrapassa um limiar. Por último, a informação que trafega pelo axônio no modelo biológico, no modelo artificial é o resultado da função de ativação. Isso pode ser melhor verificado na Tabela 3.2.

Tabela 3.2 - Comparativo entre os modelos biológico e artificial de neurônios

Neurônio biológico	**Neurônio artificial**
dendritos	conexões com a camada anterior
corpo ou soma	somatório das entradas e função de transferência
axônio	resultado da função de transferência

De um modo geral, o funcionamento do neurônio artificial segue o modelo biológico. Cada neurônio recebe informações de outros neurônios conectados à camada anterior. Após isso, é feita a ponderação entre os sinais de entrada e os pesos das conexões. O resultado é aplicado a uma função de transferência que, dependendo do limiar, promove a ativação do neurônio ou a saída da informação.

c) ARQUITETURA DE RNAs

Arquiteturas de RNAs determinam o modo pelo qual os neurônios artificiais estão organizados, existindo diversos modelos. Uma arquitetura apresenta algumas características, entre elas:

- o número de camadas que constituem uma RNA;
- o número de neurônios em cada camada;
- o modo pelo qual os neurônios estão conectados;
- as funções de transferência.

Um exemplo de arquitetura para RNA multicamada é apresentado na Figura 3.3. A rede é composta por três níveis de ativação: unidades de entrada, unidades ocultas e unidades de saída. Os pesos entre as unidades de entrada e as unidades ocultas são representadas por $w1_{ij}$, enquanto que os pesos entre as unidades ocultas e as unidades de saída são representados por $w2_{ij}$. Cada unidade de determinada camada é conectada a todas as unidades da camada seguinte, formando uma conexão total.

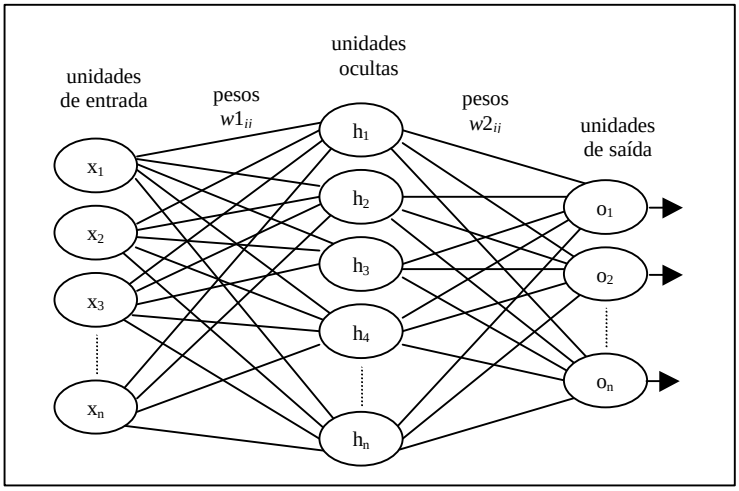


Figura 3.3 - Exemplo de RNA multicamada

Essas características necessárias à escolha de uma arquitetura são definidas em função da natureza do problema, dos dados e dos resultados esperados. A escolha de uma arquitetura de RNA para a resolução de algum problema é o ponto inicial no processo de utilização dessa técnica.

As arquiteturas de RNAs possuem propriedades particulares, sendo que cada uma pode atender a um conjunto maior ou mais restrito de aplicações. A Tabela 3.3 apresenta alguns modelos de RNAs existentes, bem como suas principais aplicações (Loesch, 1996).

Tabela 3.3 - Exemplos de modelos de RNAs (Loesch, 1996)

RNA	Ano	Principais Aplicações
Adaptive Resonance Theory (ART)	1983	Reconhecimento de padrões, classificação e mineração de dados
Multi-Layer Perceptron (MLP)	1974-1986	Reconhecimento de padrões, controle, aproximação de funções, classificação, compressão de dados
Redes com Funções de Base Radial	1987-1988	Reconhecimento de padrões, aproximação de funções e classificação
Bidirecional Associative Memory (BAM)	1987	Associação de padrões
Hopfield	1982	Evocação auto-associativa e otimização
Neocognitron	1975-1982	Reconhecimento de padrões
Mapas auto-organizáveis	1981	Reconhecimento de padrões, compressão de dados, otimização e mineração de dados

d) FASES DE IMPLEMENTAÇÂO DE UMA RNA

A construção de uma RNA possui algumas fases para que possa ser utilizada na solução de determinado problema, entre elas a representação dos dados ou pré-processamento, a determinação dos parâmetros, os treinamento e os testes.

Na primeira etapa, procura-se projetar uma representação dos dados visando maximizar o desempenho da RNA quando apresentado um padrão de entrada para que esta produza uma saída correta. Essa fase possui grande importância devido a fatores de desempenho, bem como ao correto mapeamento dos padrões apresentados, uma vez que diferentes representações podem produzir diferentes resultados.

Em seguida, a escolha adequada dos parâmetros de determinada arquitetura de RNA pode promover o sucesso ou a falha na obtenção dos resultados desejados. Existem diversos parâmetros e entre os mais gerais estão: o número de entradas, o número de saídas, o número de camadas intermediárias, o número de neurônios em cada camada, as funções de transferência, o número de iterações ou épocas, a taxa de aprendizagem para a fase de treinamento e a taxa de erro para a fase de testes.

A fase de treinamento constitui-se na apresentação de um vetor de treinamento, ou padrão, na qual a rede produz um mapeamento que permite determinar uma resposta em relação ao padrão apresentado. Nessa fase são utilizados os algoritmos de aprendizagem, sendo esses procedimentos responsáveis pela modificação dos pesos

sinápticos da rede (Pandya et al., 1995). O treinamento pode ser *supervisionado* ou *não supervisionado*. No treinamento supervisionado utiliza-se um conjunto de valores de entrada, sendo que a saída ou resposta já é previamente conhecida ou classificada. Na essência, o aprendizado supervisionado ou ativo é a disponibilidade de um conhecimento prévio do ambiente que está sendo representado (Haykin, 1994). Ao contrário do treinamento supervisionado, o treinamento não supervisionado ou auto-organizável não possui uma saída previamente conhecida, ou seja, este deve encontrar suas próprias classes, nas quais uma seqüência de vetores de entrada é apresentada, mas nenhum alvo é especificado (Fausett, 1994). O resultado de um aprendizado não supervisionado é um conjunto de descrições de classes que mapeiam todos os objetos inseridos em um determinado ambiente (Holsheimer e Siebes, 1994). Nesse modelo, os pesos são ajustados em função da similaridade dos padrões de entrada que são atribuídos a um determinado agrupamento.

Por último, a fase de testes tem por objetivo validar o grau ou poder de generalização da rede. Para isso, os exemplos que não foram apresentados na fase de treinamento são agora utilizados. O processo consiste na apresentação dos padrões ainda não vistos pela rede, na qual o sinal é propagado até a camada de saída. As saídas geradas são comparadas com seus respectivos alvos considerando uma taxa de erro. Caso a rede obtenha sucesso, ou seja, o erro produzido esteja abaixo do valor especificado, a rede encontra-se apta ao uso, caso contrário, deve-se alterar os parâmetros ou modificar a representação dos dados apresentados à RNA. Dependendo da arquitetura, não existem classes a serem comparadas, sendo necessário analisar o mapa organizado durante a fase de treinamento para verificar se a rede conseguiu atingir uma taxa de aprendizagem satisfatória.

e) APLICAÇÕES DE RNAs

As RNAs são utilizadas em uma grande variedade de aplicações. Essas aplicações podem ser divididas em alguns grupos, a saber: reconhecimento de padrões, classificação, previsão, controle, entre outros. A seguir, serão descritas algumas aplicações para cada um desses grupos.

e.1) RECONHECIMENTO DE PADRÕES

Reconhecimento de padrões pode ser definido como um processo de busca por estrutura nos dados e classificação dessas estruturas em categorias, sendo que o grau de associação é alto entre estruturas de mesma categoria e baixo entre estruturas de diferentes categorias (Klir and Yuan, 1995).

Um sistema de reconhecimento de padrões pode ser dividido em algumas tarefas (Pandya et al., 1995), tais como: particionamento da imagem em objetos isolados, extração de características e análise de contexto.

Entre as várias aplicações na área de reconhecimento de padrões pode-se citar:

- **Hugo (1995)**: utilizou uma RNA Kohonen com o objetivo de demonstrar a viabilidade de sistemas baseados no reconhecimento da fala. O estudo resultou em um Sistema Gerenciador de Central de Informações de Frete;
- **Tafner (1996)**: utilizou uma RNA Kohonen para reconhecer palavras faladas isoladamente.

e.2) CLASSIFICAÇÃO

A classificação é uma categoria de problemas cujos padrões (entradas) pertencem ou não a uma ou várias classes previamente definidas (Fausett, 1994). Essa categoria de rede neural produz saídas que representam um conjunto de características apresentadas através de um padrão de entrada. Abaixo são descritas algumas aplicações que utilizam classificação:

- **Todesco (1995):** utilizou uma rede de Função de Base Radial (FBR) para a classificação de cromossomos humanos. Nesse trabalho, é confrontada a solução baseada em uma rede neural em relação aos métodos mais tradicionais, tais como classificadores paramétricos e não-paramétricos utilizados pela estatística e probabilidade;
- **Frenkel e Nadal (1999):** utilizaram uma rede *feedforward* (*backpropagation*) para detecção de mudanças em segmentos isoelétricos, entre a onda S e a

onda T (segmento ST) em um eletrocardiograma (ECG), relacionados a episódios isquêmicos do coração;

- **Melo et al. (1999):** esse trabalho apresenta uma variação da rede de Kohonen para suportar treinamento supervisionado, buscando detectar arritmias cardíacas a partir de eletrocardiogramas (ECG).

e.3) PREVISÃO

Previsões são uma das aplicações em que redes neurais artificiais têm sido utilizadas com bastante sucesso, principalmente sobre dados financeiros e econômicos. A grande relevância das previsões é que estas oferecem subsídios à tomada de decisão.

Entre as várias aplicações visando previsões e predições pode-se citar algumas, relacionadas a seguir.

- **Paz e Borges (1999)**: utilizaram a rede neural RBF na previsão de consumo de energia elétrica. A previsão, primeiramente, baseou-se em dados de 1956 até 1995, visando prever o consumo de 1996;
- **Rautenberg (1998):** construiu uma ferramenta baseada em uma rede RBF, de modo a auxiliar na predição de receitas de cores em uma estamparia, objetivando reduzir o desperdício de materiais e o retrabalho.

e.4) CONTROLE

A utilização de RNAs no processo de controle mostra-se como uma das técnicas mais eficientes (Kovacs, 1996). Isso ocorre principalmente devido à rápida expansão de controle em tempo real, sendo que a principal vantagem dessa abordagem em relação aos sistemas industriais de controle tradicionais está na habilidade de se aprender através de experiências. Abaixo são apresentadas algumas aplicações que utilizam RNAs em processos de controle:

- **Kawato et al. (1992)**: utilizou a rede RNA na elaboração de um modelo visando movimentos voluntários em aplicações para robótica. Esse modelo é utilizado no controle de um manipulador robótico industrial.

- **Sciavicco e Siciliano (1996)**: utilizaram uma RNA no controle de manipuladores robóticos em relação aos controladores clássicos, por estes não apresentarem boas respostas a pequenas variações em parâmetros como o atrito nas junções.

3.2.2 REGRAS DE ASSOCIAÇÃO

Regras de associação consistem na descoberta de relacionamentos existentes entre variáveis. Esses relacionamentos são descobertos efetuando-se múltiplos passos iterativos sobre a base de dados. A cada iteração é levado em consideração o conjunto de itens gerados no passo anterior, chamado de conjunto de itens candidatos. Seja $I = \{i_1, i_2,\ldots,i_m\}$ um conjunto de itens, uma regra de associação é uma implicação na forma $X \Rightarrow Y$, onde $X \subseteq I$, $Y \subseteq I$ e $X \cap Y = \emptyset$. Uma regra de associação possui dois parâmetros básicos: um suporte e uma confiança. O suporte é caracterizado pelo número mínimo de ocorrências de $X \cup Y$, ou seja, a união de itens no conseqüente e antecedente da regra está presente em um suporte mínimo **s%** na base (Agrawal, 1993), enquanto que a confiança é um percentual das transações na base de dados que satisfazem o antecedente da regra (X) e também o conseqüente da regra (Y), (Cheung et al., 1996).

O problema de regras de associação pode ser decomposto em três passos principais (Agrawal, 1993):

- *gerar todas as combinações de itens*;
- *descobrir conjuntos de itens*: Este passo consiste em gerar um conjunto com todas as combinações de itens obedecendo a um limiar, chamado *suporte mínimo*. As combinações que satisfazem essa condição são chamadas de conjunto de itens grandes, enquanto que as que não satisfazem são chamadas de conjunto de itens pequenos;
- *gerar as regras de associação para a base de dados*: Após o conjunto de itens finais ter sido produzido, deve-se gerar as regras de associação de um conjunto de itens $Y = I_1,I_2,,\ldots,I_k$, sendo $k \geq 2$. O antecedente da regra será um conjunto X de Y, tal que X possua k-1 itens e o conseqüente seja Y-X. Para

verificar a validade de uma regra, a confiança da regra (suporte(Y)/ suporte(X)) deve satisfazer o valor mínimo de confiança informado.

Existem algumas variações do algoritmo básico de geração de regras. A grande maioria é uma extensão do algoritmo *Apriori*. Entre essas derivações, pode-se citar *AprioriTid*, DHP e *Partition*. Na seção seguinte será descrito o algoritmo *Apriori*.

a) ALGORITMO Apriori

O algoritmo *Apriori* (Figura 3.4) é um dos algoritmos mais conhecidos e referenciados na geração de regras de associação e tem como objetivo encontrar os conjuntos de itens mais freqüentes (L_k). Este algoritmo baseia-se em duas funções: a primeira chamada *Apriori-gen*, que gera os itens candidatos levando em consideração o valor do suporte (percentual indicado que fornece a ocorrência mínima de determinada combinada de itens na base de dados), e uma segunda função, chamada *Genrules*, que gera as regras de associação considerando o parâmetro de confiança informado.

A primeira etapa do algoritmo consiste na contagem dos itens para determinar o *itemset* inicial L_1. Em seguida, utilizando (k-1) passos, são gerados os potenciais itens candidatos através da função **apriori-gen**, composta de dois passos. No primeiro passo, é realizada a junção dos itens, sendo o conjunto C_k gerado a partir de L_k-1 com L_k-1. Em seguida, no passo de poda, são eliminados todos os *itemsets* $c \in C_k$, tal que um dado (k-1)-*subset* de c não esteja em L_k-1.

```
L1 = {1-itemsets};
Para (k=2; Lk-1 ≠ 0; k++) faça {
  Ck = apriori_gen(Lk-1);
  Para toda transação t ∈ D faça {
    Ct = subset(Ck, t);
    Para todo candidato c ∈ Ct faça
      c.count++
  }
  Lk = {c ∈ Ck | c.count ≥ minsup}
}
Retorna = Lk;
```

Figura 3.4 - Algoritmo apriori básico (Agrawal et al., 1996)

A função *Apriori-gen* (Figura 3.5) divide-se em duas etapas, sendo uma etapa de união e uma de poda dos itens. Na primeira etapa, a função retorna um conjunto de *k-itemsets* candidatos unindo L_k-1 com L_k-1. No passo seguinte são eliminados todos os *itemsets* $c \in C_k$ tal que um dado (*k*-1)-*subset* de *c* não esteja em L_k-1. Seja L_k-1 {{1,2,3}, {1,2,5}, {1,3,5},{1,4,5},{1,4,6},{2,3,5},{3,1,2},{3,1,3},{3,2,3}}, o conjunto C_k será {{1,2,3,5}, {1,4,5,6}, {3,1,2,3}}. Na fase de poda o item {1,4,5,6} será eliminado, pois o *itemset* {1,5,6} não se encontra em L_k-1.

```
Passo 1
   insere em Ck
     seleciona p.item1, p.item2,…,p.itemk-1, q.itemk-1
     de Lk-1 p, Lk-1 q
     onde p.item1 = q.item1,…,p.itemk-2 = q.itemk-2, p.itemk-1 < q.itemk-1;

Passo 2
  para todo itemset c ∈ Ck faça
    para todo (K-1)-subset s de c faça
      se (s ∉ Lk-1) então
         elimina c de Ck;
```

Figura 3.5 - Função apriori-gen (Agrawal et al., 1996)

Na última fase do algoritmo, são geradas todas as regras para todos os *subsets* não vazios de *l* através do procedimento g*enrules* (Figura 3.6). Para todo *subset a*, é gerada uma regra $a \Rightarrow (l - a)$, se a confiança da regra (suporte(*l*) / suporte(*a*)) é menor ou igual à mínima confiança especificada. Para gerar uma regra que possua vários conseqüentes, consideram-se todos os *subsets* de um determinado *itemset*, ou seja, tomando como exemplo o *itemset* ABCD, considera-se o primeiro *subset* ABC, então AB, etc. Para evitar a geração desnecessária de regras, caso $ABC \Rightarrow C$ não alcance a confiança mínima, não é necessário verificar se $AB \Rightarrow CD$ a possui.

```
Procedimento genrules(lk: k-itemset, am: m-itemset)
  A = {m-1)-itemsets am-1 | am-1 ⊂ am};
  para todo am-1 ∈ A faça
     conf = suporte(lk)/suporte(am-1);
     se (conf ≥ minconf) então
        apresenta regra am-1 ⇒ (lk - am-1)
        se (m –1 > 1) então
           chama genrules(lk , am-1);
        fim se
     fim se
  fim para
fim
```

Figura 3.6 - Procedimento *genrules* (Agrawal et al., 1996)

3.2.3 ANÁLISE DE AGRUPAMENTOS

A análise de agrupamentos pode ser definida como uma técnica que agrupa um conjunto de itens, indivíduos ou objetos, sendo que os objetos incluídos em um mesmo agrupamento são os mais similares entre si e menos similares em relação aos objetos que estão em outros agrupamentos (Hair et al., 1998). Nessa técnica, objetiva-se a descoberta de um conjunto de classes, ou seja, deve-se achar *N* agrupamentos (*A1*, *A2*, ...,*An*) que descrevam o comportamento dos dados. Agrupamentos baseiam-se principalmente na maximização da similaridade intraclasses e na minimização da similaridade interclasses (Chen et al., 1996).

Os agrupamentos são realizados por meio de uma distância de similaridade (dissimilaridade) (Johnson e Wichern, 1998). Dessa maneira, a pessoa que realiza a análise deve possuir conhecimento suficiente sobre o problema, visando distinguir grupos úteis, necessários à realização de consultas. A análise de agrupamentos é uma metodologia objetiva para quantificar uma característica estrutural de um conjunto de observações (Hair et al., 1998), e apresenta três desafios básicos:

- *como medir a similaridade entre os itens*. Neste item, torna-se necessário a adoção de um parâmetro de qualificação dos itens;
- *como formar os agrupamentos*. Aqui deve-se determinar quais variáveis fazem parte da geração de determinados agrupamentos;
- *quantos grupos devem ser formados*. Existem basicamente duas abordagens, uma que define o número de agrupamentos desejados, e outra que usa um critério, tal como um raio de abrangência do agrupamento.

A análise de agrupamentos consiste na descoberta natural de agrupamentos de itens (Johnson e Wichern, 1998), em que a complexidade é determinada pelo número de meios de classificar *N* tuplas em *k* agrupamentos não vazios, dado por: $P(N,k) = \frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}(-1)^j(k-j)^N$, e o número de maneiras que uma base de dados pode ser agrupada, denotado por *C*(*N*) (Holsheimer e Siebes, 1994) é: $C(N) = \sum_{k=1}^{N} P(N,k) = \sum_{k=1}^{N}\frac{1}{k!}\sum_{j=0}^{k}\binom{k}{j}(-1)^j(k-j)^N$.

O exemplo abaixo possibilita uma melhor compreensão, possuindo oito itens para serem agrupados em um total de oito agrupamentos (Holsheimer e Siebes, 1994). O resultado final apresenta o total de maneiras possíveis de se efetuarem esses agrupamentos.

K	1	2	3	4	5	6	7	8	Total
P(8,*k*)	1	127	966	1701	1050	266	28	1	4140

Existem várias técnicas e algoritmos que procuram encontrar grupos potencialmente úteis, entre os quais estão incluídos os métodos estatísticos e as redes neurais. Tal tarefa constitui um passo importante durante uma análise, visto que nem sempre se possui, a priori, um conhecimento completo dos dados ou de seus relacionamentos. Após a formação desses agrupamentos, pode-se então derivar um conjunto de regras. Essas regras podem descrever ou facilitar a formação de uma taxonomia que identifique os agrupamentos, facilitando a visualização e as análises posteriores.

a) ANÁLISE DE AGRUPAMENTOS NO PROCESSO DE DECISÃO

A análise de agrupamentos pode ser vista como um processo que se divide em cinco etapas (Hair et al., 1998). São elas:

- *objetivos da análise de agrupamentos*: constituem a meta primária deste processo, em que um conjunto de objetos é dividido em dois ou mais grupos baseados em alguma medida de similaridade entre estes objetos, levando-se em conta suas características. Nessa fase, espera-se atingir alguns objetivos, tais como: descrição taxonômica, cuja formação é conseguida através de uma classificação empírica; simplificação dos dados, permitindo que a análise de um objeto seja efetuada em termos mais gerais; e, por último, a identificação de relacionamentos, que geralmente não são possíveis se avaliados sob uma perspectiva individual;
- *projeto*: nesta etapa deve-se considerar a detecção de *outliers* e as possíveis estratégias para manipular essas exceções, bem como a similaridade entre os objetos e questões pertinentes quanto à normalização dos dados;

- *algoritmos utilizados e número de agrupamentos*: esta fase visa determinar quais algoritmos e quantos agrupamentos serão utilizados. Esses algoritmos devem, portanto, agrupar itens de modo a maximizar as diferenças entre os agrupamentos e maximizar as semelhanças dos objetos pertencentes a determinado agrupamento. Os métodos utilizados na categorização podem ser hierárquicos e não hierárquicos. Quanto à definição do número de agrupamentos, esta tende a ser um processo empírico. Nesse passo pode-se, contudo, utilizar-se de alguma regra estatística empírica, tal como $1 + 3.3 \log_{10} N$ (Pacitti e Atkinson, 1977), onde N é o tamanho da amostra. Entretanto, o que geralmente se verifica é que o conhecimento e a experiência do analista de informações assumem um importante papel na determinação do número de classes relevantes ao problema;
- *interpretação dos agrupamentos*: nesta etapa, procura-se examinar cada agrupamento, de modo a se extrairem conhecimentos que descrevam o comportamento destes agrupamentos. A identificação de padrões válidos é efetuada nesta fase, na qual se pode identificar perfis distintos para todos os agrupamentos ou similaridades entre alguns em relação a outros, sendo esta fase de grande importância para o processo de descoberta de conhecimento;
- *Validação*: nesta etapa deve-se tomar um certo cuidado para garantir a praticidade da solução final. Essa medida pode ser realizada através da aplicação de métodos complementares, permitindo fazer comparações entre os resultados alcançados, ou mesmo incrementar estes resultados.

b) KOHONEN (SELF-ORGANIZING MAPS)

As redes auto-organizáveis são uma classe de RNAs em que a aprendizagem ocorre de modo não supervisionado. Esses sistemas, de maneira semelhante a sistemas biológicos, podem descobrir padrões e características relevantes. A rede de Kohonen (SOM) possui forte influência biológica, sendo estas estruturas encontradas no córtex auditivo e visual (Pandya et al., 1995).

O uso desses mapas computacionais oferece as seguintes vantagens (Knudsen et al., 1987, in Haykin, 1994):

- *processamento de informação eficiente*: o sistema nervoso é capaz de analisar uma grande quantidade de informações e efetuar ações em ambientes complexos, devido ao seu paralelismo. Mapas computacionais simulam esse comportamento através de *arrays* de processamento paralelo, provendo uma simplificação quando são processadas informações complexas;
- *simplicidade de acesso para informação processada*: o uso de mapas computacionais simplifica o *schema* de conectividade requerido para utilizar e recuperar informações;
- *forma comum de representação*: computacionalmente permite uma representação de como o sistema nervoso constrói estratégias para manipulação de informações.

A arquitetura da rede de Kohonen é composta de duas camadas totalmente conectadas, sendo a primeira a camada de entrada, e a segunda, a camada de Kohonen (Figura 3.7), podendo ser de uma ou mais dimensões.

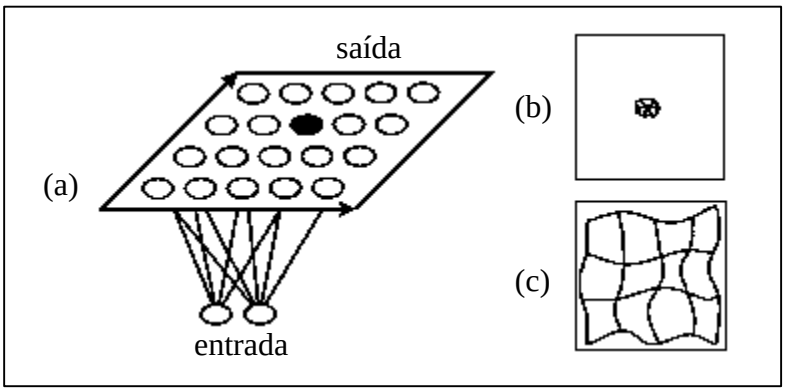


Figura 3.7 - (a) Rede de Kohonen com duas entradas mapeadas em um grid de 4x5, (b) estrutura de pesos antes do treinamento, (c) estrutura de pesos após o treinamento (Medler, 1998)

Esse modelo de rede baseia-se na aprendizagem competitiva, na qual os neurônios da camada de saída competem entre si para determinar o vencedor em relação ao padrão de entrada. O vencedor possui seus pesos atualizados e, possivelmente, a sua vizinhança, dependendo do raio de abrangência. Nessa arquitetura, os neurônios são arranjados geralmente em nodos de uma ou duas dimensões, sendo possível, contudo, mapas com dimensões maiores. À medida que são apresentados os padrões de entrada, os neurônios de saída tornam-se ajustados, de acordo com a seguinte equação $W_{ij}^{novo} = W_{ij}^{velho} + \eta\left(x_i - W_{ij}^{velho}\right)$, ou variações, onde η é a taxa de aprendizagem.

A rede de Kohonen cria um mapa ajustando os pesos a partir dos nodos de entrada comuns para M nodos de saída, arranjados em uma matriz geralmente bi-dimensional (Figura 3.7). Em relação às unidades de saída, estas são criadas a partir da apresentação seqüencial de padrões de entrada, não especificando uma saída desejada. Ao final dessa fase, resulta em um conjunto de centros dos vetores, que mapeiam o espaço de entrada.

A inicialização dos pesos da rede possui algumas maneiras de ser efetuada (Kohonen, 1995), entre elas:

- *inicialização aleatória*: esta abordagem é mais rápida e tem demonstrado que a organização do vetor inicialmente desordenado ocorre ao longo da execução, geralmente com poucas centenas de épocas;
- *um exemplo inicial*: utilizar exemplos de entrada x(t) selecionados de maneira aleatória.

Um dos métodos utilizados para determinar o neurônio vencedor é a mínima distância Euclidiana de um vetor de entrada $d_j = ||x_i - w_{ij}||$, onde x_i é o nodo de entrada e w_{ij} é o vetor de pesos no nodo de entrada $_i$ para o nodo de saída $_j$. O uso da distância Euclidiana para determinar o vencedor, dependendo da aplicação, pode ser vantajoso levando-se em consideração que esta não requer que os pesos ou o vetor de entrada estejam normalizados (Pandya et al., 1995).

Outra característica na arquitetura é a definição da topologia da vizinhança, que permite que os pesos sejam atualizados levando-se em consideração um raio de abrangência, ou seja, quando um vencedor é determinado, os pesos deste são atualizados, juntamente com os da sua vizinhança. A topologia pode ser retangular, hexagonal ou implementada utilizando-se uma função Gaussiana.

Dois outros parâmetros devem ser considerados: o raio da vizinhança *Nc* e o fator da taxa de aprendizagem η. Ambos podem ser considerados em função do tempo e freqüentemente possuem um decréscimo durante o processo de organização. A taxa de aprendizagem deve apresentar a relação $0 < \eta(t) < 1$, enquanto que *Nc* pode ser maior do que a metade do diâmetro da rede (Kohonen, 1995).

b.1) O ALGORITMO SOM

Existem muitas variações do algoritmo, em geral efetuadas para atender a determinado conjunto de aplicações. Entretanto, o algoritmo SOM básico pode ser implementado utilizando-se os seguintes passos:

- passo 1- inicialização: Inicializar o vetor de pesos w_j aleatoriamente, sendo comum que os pesos possuam uma magnitude pequena, geralmente entre 0 e 1. Inicializar a taxa de aprendizagem η e a função de avaliação de vizinhança $\Lambda_i(x)$. Ambos, taxa de aprendizagem e função de vizinhança sofrem um decréscimo durante a fase de aprendizado;
- passo 2: apresentar os vetores de entrada *x*, que representam os padrões que a rede deve aprender;
- passo 3: verificar a similaridade, identificando-se o neurônio vencedor através de uma medida de similaridade, como a mínima distância Euclidiana:

$$d_j = \sum_{i=0}^{N-1} \left(x_i(t) - w_{ij}(t)\right)^2, \quad j = 1,2,\ldots,N$$

- passo 4: atualizar o vetor de pesos para todos os neurônios, utilizando-se a fórmula:

$$w_j(n+1) = \begin{cases} w_j(n) + \eta(n)\left[x(n) - w_j(n)\right], & j \in \Lambda_{i(X)}(n) \\ w_j(n), & \textit{caso contrário} \end{cases}$$

- passo 5: repetir os passos 2, 3 e 4, até que não sejam mais observadas mudanças nos pesos ou que seja atingido um número máximo de épocas.

b.2) INICIALIZAÇÃO DOS PARÂMETROS

O processo de aprendizagem na computação do algoritmo é estocástico por natureza (Kohonen, 1995). O sucesso na formação do mapa depende principalmente da inicialização da taxa de aprendizagem e da função de vizinhança. Contudo, não existe uma base teórica para a inicialização desses parâmetros, sendo estes determinados na tentativa e erro (Haykin, 1994).

Alguns pontos podem ser úteis na determinação desses parâmetros. O parâmetro da taxa de aprendizagem h deve variar em função do tempo. Durante as primeiras 1000 iterações, h deve decrescer gradualmente, porém deve ficar acima de 0.1, sendo que essa variação pode ser linear ou exponencial. Com relação à função de vizinhança $\Lambda_i(x)$, esta pode assumir algumas formas, sendo as mais comuns a retangular e a hexagonal (Figura 3.8). Na fase de ordenação, $\Lambda_i(x)$ sofre uma redução linear em função do tempo e, durante a fase de convergência, $\Lambda_i(x)$ deve considerar somente a vizinhança mais próxima em relação ao neurônio vencedor, sendo este parâmetro 1 ou 0. Uma forma mais apurada para a atualização dos pesos é a utilização de uma função Gaussiana, por meio da qual a influência da atualização diminui proporcionalmente à distância da vizinhança em relação ao agrupamento vencedor.

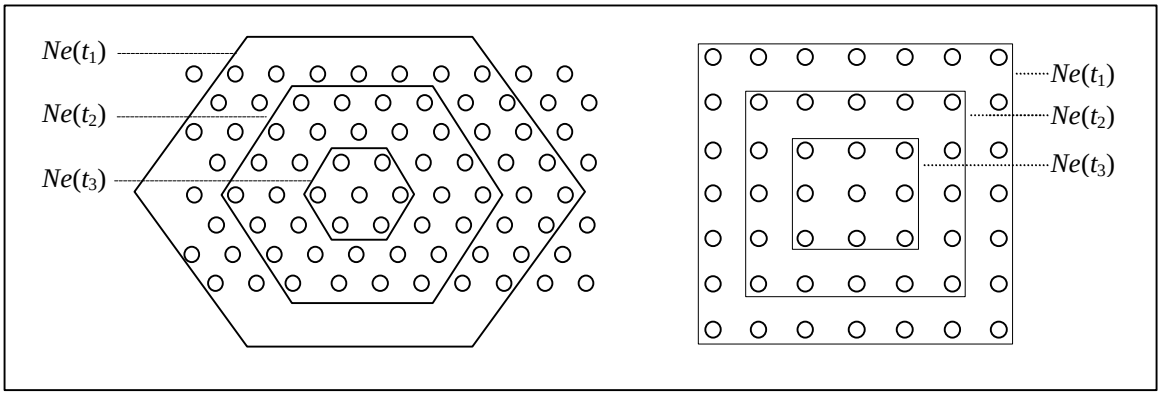


Figura 3.8 - Topologia da vizinhança Ne variando em função do tempo t1 < t2 < t3

b.3) APLICAÇÕES DA REDE DE KOHONEN

A rede de Kohonen é amplamente utilizada em aplicações que envolvam a geração de agrupamentos. Isso pode ser verificado em Kaskit et al. (1998), em que são apresentados mais de 3000 trabalhos produzidos entre 1981 e 1997 que utilizaram essa arquitetura. Na área de mineração de dados e C&T, pode-se identificar alguns trabalhos realizados, entre eles:

- **Back et al. (1996)**: utilizou a RNA Kohonen para efetuar um *benchmarking* competitivo de processos internos de empresas, em que o desempenho de uma companhia é medido em relação a outras companhias. Para isso, foram levadas em consideração variáveis de medida de desempenho, tais como: lucro operacional, investimentos e respectivo retorno, solidez, etc.;

- **Campanario** (**1995**): produziu um mapa da ciência utilizando citações de publicações, no qual procurava determinar um conjunto de autocitações fornecendo uma estrutura dinâmica sobre a ciência ao nível de inter-relacionamento entre os periódicos analisados;
- **Kohonen (1998):** demonstra a utilização de SOM sobre uma base de um milhão de documentos, estando cada documento mapeado em um determinado agrupamento. Quando uma consulta é realizada por meio de palavras-chave, o sistema apresenta os documentos mais relevantes, partindo sempre do mais similar ao valor apresentado e, em seguida, dos seus vizinhos.

3.3 CONSIDERAÇÕES FINAIS

Este capítulo abordou as técnicas utilizadas na análise e extração de conhecimento de uma base de dados. Uma pequena introdução sobre RNAs foi apresentada visando contextualizar a técnica e promover um melhor entendimento da arquitetura de RNA utilizada no trabalho. Maior ênfase foi dispensada às técnicas de Mineração de Dados (MD), que fazem parte do processo de *Knowledge Discovery in Database* (KDD).

Essas técnicas têm promovido um novo paradigma na análise e descoberta de informações úteis e necessárias à tomada de decisões. Esse aspecto mostra-se cada vez relevante, pois a quantidade de dados armazenados atualmente em grandes empresas, agências de pesquisa e universidades torna por vezes complexa a tarefa de análise dessas informações.

O avanço nessa área, seja no nível de técnicas, visualização ou interpretação dos dados, tende a fornecer ferramentas cada vez mais poderosas às pessoas que possuem a função de traçar a atuação de uma grande empresa, de analisar possíveis fraudes, ou de analisar tanto o nível de investimentos como o retorno desses investimentos. Atualmente, diversas técnicas têm sido implementadas em soluções comerciais, além das técnicas já citadas, entre elas: Algoritmos Genéticos, *Machine Learning*, Redes Bayesianas, Sistemas Difusos, Árvores de Decisão, *Link Analysis*, etc.

As técnicas utilizadas – Rede Neural (Kohonen-SOM) e Regras de Associação – promovem suporte na obtenção dos objetivos do trabalho. Esses objetivos estão relacionados à extração de conhecimento e à apresentação de um método que possa ser comparado em relação ao algoritmo de hierarquização dos grupos, atualmente em fase de discussão no CNPq. Para tal, no próximo capítulo são apresentados os resultados obtidos em função da utilização dessas técnicas.

4 ANÁLISE DE C&T UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

"Há três séculos, o conhecimento científico não faz mais do que provar suas virtudes de verificação e de descoberta em relação a todos os outros modos de conhecimento"

Edgar Morin

4.1 INTRODUÇÃO

Todo trabalho realizado na preparação, análise e extração de conhecimento de uma base de dados requer esforço e tempo. Este trabalho por vezes torna-se penoso, devido, sobretudo, ao número de variáveis envolvidas e à quantidade de dados existentes.

O analista que deseja extrair algum tipo de conhecimento de uma base de dados deve sempre estar preocupado com a limpeza desses dados, que se constitui numa das fases do pré-processamento, com a recuperação dos dados, utilizando consultas submetidas a uma base de dados, com a tabulação dos dados, através de programas específicos ou planilhas de cálculo e, por último, com a apresentação dos resultados. Observando-se essas fases, nota-se a complexidade do processo.

A utilização de ferramentas integradas de análise visa amenizar o trabalho realizado durante o processo de extração de conhecimento. Neste trabalho, implementa-se a fase de mineração de dados, utilizando-se um algoritmo de categorização e um algoritmo de geração de regras.

Como análise final, o trabalho efetua comparações em relação ao algoritmo de hierarquização dos grupos de pesquisa (Guimarães et al., 1999). Nessa etapa, é apresentada uma visão não parametrizada, ou seja, sem a utilização das tabelas de ponderações empregadas pelo algoritmo tradicional, e uma visão parametrizada que utiliza essas tabelas. Ainda são apresentadas comparações dos resultados alcançados

com o modelo em relação ao algoritmo de hierarquização, referente à qualificação dos grupos de pesquisa.

4.2 METODOLOGIA

Para desenvolver o sistema de extração de conhecimento, empregou-se a metodologia de Berry e Linoff (1997), com exceção à etapa de comparação com o modelo do CNPq. A Figura 4.1 apresenta o esquema metodológico.

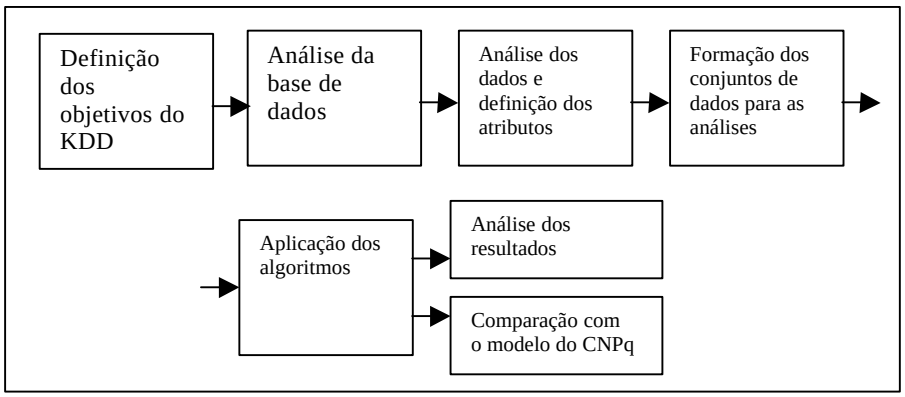


Figura 4.1 - Metodologia utilizada na extração de conhecimento

A análise da base de dados consiste na etapa de estudo do ambiente tecnológico dos dados. O trabalho utilizou a base de dados do *Diretório dos Grupos de Pesquisa no Brasil*, em sua versão 3.0[2].

Após a análise da base, verificou-se a natureza dos dados envolvidos, a fim de formar os conjuntos para as análises e aplicação dos algoritmos. A estes conjuntos aplicaram-se os algoritmos de categorização e as regras de associação. Na etapa de análise são apresentados os resultados obtidos pela utilização dos algoritmos de MD e os resultados do modelo proposto de avaliação dos grupos de pesquisa em relação ao algoritmo de hierarquização tradicional. Cada uma dessas etapas é descrita nas próximas seções.

[2] A base de dados do *Diretório dos Grupos de Pesquisa no Brasil* descreve o comportamento dos grupos de pesquisa no país, sendo composta por informações de identificação dos grupos, suas linhas de pesquisa, seus integrantes e sua produção C&T.

4.3 A BASE DE DADOS DO DIRETÓRIO

A base de dados utilizada do Diretório dos Grupos de Pesquisa no Brasil, versão 3.0, encontra-se sobre um banco de dados Oracle®. Essa base foi formada pelo CNPq a partir do censo realizado com 8.632 grupos, em 181 instituições de pesquisa do país, com informações captadas entre setembro de 1997 e março de 1998, compreendendo a produção C&T entre o período de 1° de janeiro de 1995 e 30 de julho de 1997 (Guimarães et al., 1999). A Tabela 4.1 demonstra o total de registros na base.

Tabela 4.1- Indicadores macros da base do Diretório

Item	Total
Total de grupos	8.632
Total de instituições	181
Total de pesquisadores	34.040
Total de itens de produção C&T	222.989

Para a aplicação dos algoritmos de mineração de dados no trabalho, foram utilizadas as seguintes tabelas do Diretório:

- tabela de dados cadastrais dos Grupos de Pesquisa;
- tabelas de Áreas do Conhecimento em que atuam os pesquisadores dos Grupos de Pesquisa;
- tabelas de detalhamento e contexto da Produção Científica, Tecnológica e Artística/Cultural dos Grupos de Pesquisa.

Além das tabelas do Diretório dos Grupos de Pesquisa, o trabalho considera ainda as seguintes informações:

- informações sobre o Programa de Pós-graduação em que atuam pesquisadores do grupo (dados da CAPES);
- informações sobre a participação de bolsistas de pesquisa no grupo (dados complementares do CNPq).

4.4 DEFINIÇÃO DOS OBJETIVOS

A primeira tarefa realizada no processo de KDD implantado consistiu em estabelecer os objetivos da análise. O projeto Diretório dos Grupos de Pesquisa no

Brasil, além do censo das atividades dos grupos e de seus pesquisadores, possui o algoritmo de hierarquização que estratifica os grupos de pesquisa em cinco categorias, de acordo com o grau de qualificação da pós-graduação em que atuam os pesquisadores e do grau de qualificação dos bolsistas de pesquisa vinculados ao grupo.

Assim, optou-se por estabelecer como objetivos a aplicação de técnicas de MD na classificação dos grupos de pesquisa segundo o mesmo conjunto de informações utilizado pelo algoritmo de hierarquização. O objetivo é verificar se a classificação dos grupos de pesquisa utilizando MD apresenta correspondência com os resultados fornecidos pelo algoritmo de estratificação.

Além da comparação de métodos de classificação, a análise pretende investigar relações entre as variáveis componentes de um grupo de pesquisa. Nas próximas seções, estão sendo elucidadas as seguintes relações:

• Natureza do Veículo de Produção C,T&A do grupo	Titulação máxima dos pesquisadores do grupo	(Primeiro conjunto de dados)
• Natureza do Veículo de Produção C,T&A do grupo	Dados de identificação dos pesquisadores do grupo	(Segundo conjunto de dados)
• Sistema de classificação de bolsistas do CNPq	Sistema de avaliação da Pós-graduação da CAPES	(Terceiro conjunto de dados)

4.5 DEFINIÇÃO DOS ATRIBUTOS

A análise e a definição dos atributos têm por objetivo selecionar os tipos de informações disponíveis mais afins à natureza do conhecimento que se pretende descobrir. Neste trabalho, considerando-se os objetivos do processo de KDD que se deseja implementar, os atributos relevantes referem-se aos seguintes conjuntos de informações (Tabela 4.2 e 4.3):

- *dados dos pesquisadores*: identificação e formação (atributos: nacionalidade, idade, sexo e titulação máxima); participação no grupo de pesquisa (atributo: tempo de dedicação ao grupo); qualificação do programa de pós-graduação em que atuam (atributo: nota do programa na avaliação da CAPES); qualificação de bolsa de pesquisa no CNPq (atributo: nível do bolsista);

- *dados de produção científica, tecnológica e artística/cultural*: quantidade e classificação da produção segundo a árvore de tipos do CNPq.

Tabela 4.2 - Variáveis de análise referentes aos pesquisadores integrantes dos grupos de pesquisa

Variável	Descrição	Valores	Base
BPQ	Pesquisadores com ao menos uma bolsa de pesquisa do CNPq durante o período de janeiro de 1997 a agosto de 1998	1A, 1B, 1C, 2A, 2B e 2C	SIGEF (CNPq)
DOC	Docentes com vínculo em cursos de Pós-graduação avaliados pela CAPES	D7, D6, D5, D4, D3	Coleta (CAPES)
DED	Tipo de dedicação do pesquisador	I = dedicação integral ou exclusiva P = dedicação em tempo parcial	Diretório
NAC	Nacionalidade do pesquisador	B = nacionalidade brasileira E = nacionalidade estrangeira	Diretório
TIT	Titulação máxima do pesquisador	1 = nível de graduação (GRD) 2 = nível de especialização (ESP) 3 = nível de mestrado (MDO) 4 = nível de doutorado (DDO)	Diretório
ID	Idade do pesquisador	Idade entre 15 e 88 anos	Diretório
SX	Sexo do pesquisador	M = masculino F = feminino	Diretório

Tabela 4.3 - Variáveis de análise quantitativas referentes aos grupos de pesquisa e seus integrantes

Variável/ Natureza	Descrição	Tipos	Base
ANE	Artigos publicados em periódicos especializados, nacionais e estrangeiros, com corpo editorial e sistema de *referees*	a) An = artigos publicados em periódicos nacionais b) Ae = artigos publicados em periódicos estrangeiros	Diretório
TER	Trabalhos em eventos, completos e resumos, e revistas não incluídas anteriormente	a) Tc = trabalhos completos publicados em eventos b) Ar = artigos publicados em periódicos sem corpo editorial e sem sistema de *referees* c) Rv = trabalhos publicados em revistas técnico-científicas d) Rs = resumos publicados de trabalhos apresentados em eventos técnico-científicos	Diretório
LC	Livros e Capítulos de Livros publicados	a) Li = livros publicados b) CI = capítulos de livros publicados	Diretório
PTC	Produção Tecnológica desenvolvida	a) Sf = softwares desenvolvidos b) Pd = produtos tecnológicos desenvolvidos c) Pc = processos tecnológicos desenvolvidos	Diretório
TD	Teses e Dissertações orientadas	a) Te = teses de doutorado b) Di = dissertações de mestrado	Diretório

4.6 RECUPERAÇÃO E PRÉ-PROCESSAMENTO DOS DADOS

Esta etapa envolveu primeiramente o desenvolvimento de uma rotina de recuperação e geração de uma tabela, com os totais para cada tipo de produção C&T e com as informações referentes aos dados individuais dos pesquisadores. Os dados referentes aos bolsistas do CNPq e docentes vinculados a cursos de pós-graduação avaliados pela CAPES já estavam tabulados, sendo estes fornecidos pelo CNPq. Após esse passo, foram elaboradas consultas para a recuperação dos conjuntos de dados, bem como para a normalização destes dados.

4.6.1 CONJUNTOS DE DADOS UTILIZADOS

Os conjuntos de dados utilizados foram extraídos procurando descrever o conteúdo da base de dados do Diretório dos Grupos de Pesquisa no Brasil. Esses conjuntos têm como objetivo básico identificar possíveis padrões nos relacionamentos entre Grandes Áreas do Conhecimento, informações gerais dos pesquisadores e dos grupos de pesquisa, bem como o seu relacionamento com o nível de formação dos integrantes dos grupos e sua produção C&T.

O primeiro conjunto de dados (Tabela 4.4) é composto pelas variáveis de natureza da produção C&T de cada grupo, e pela identificação do nível de formação dos integrantes do grupo. Esse conjunto procura por padrões iniciais de relacionamento da produção C&T, vinculada ao nível de formação dos pesquisadores nos grupos.

Tabela 4.4 - Estrutura do primeiro conjunto de dados

Colunas	Composição
Vx1 = Artigos publicados	ANE(An)+ANE(Ae)
Vx2 = Trabalhos em eventos	TER(Tc)+TER(Ar)+TER(Rv)+TER(Rs)
Vx3 = Livros e Capítulos	LC(Li)+LC(Cl)
Vx4 = Produção Técnica	PTC(Sf)+PTC(Pd)+PTC(Pc)
Vx5 = Teses e Dissertações	TD(Te)+TD(Di)
Vy1 = Graduação	GRD
Vy2 = Especialização	ESP
Vy3 = Mestrado	MDO
Vy4 = Doutorado	DDO

O segundo conjunto de dados (Tabela 4.5) é composto pelas variáveis *sexo*, *nível de formação*, *idade*, *nacionalidade*, *tipo de dedicação*, e pelas variáveis de natureza da produção C&T. Esse conjunto busca identificar padrões entre os dados gerais de cada pesquisador e os veículos de publicação do grupo.

Tabela 4.5 - Estrutura do segundo conjunto de dados

Colunas	Composição
Vz1 = Sexo	1 = Feminino, e 2 = Masculino
Vz2 = Nível de formação	1 = Graduação (GRD), 2 = Especialização (ESP), 3 = Mestrado (MDO) e 4 = Doutorado (DDO)
Vz3 = Idade	Valores entre 20 e 88 anos
Vz4 = Nacionalidade	1 = Estrangeiro, e 2 = Brasileiro
Vz5 = Dedicação	1 = Parcial, e 2 = Integral
Vx1 = Artigos publicados	ANE(An)+ANE(Ae)
Vx2 = Trabalhos em eventos	TER(Tc)+TER(Ar)+TER(Rv)+TER(Rs)
Vx3 = Livros e Capítulos	LC(Li)+LC(Cl)
Vx4 = Produção Técnica	PTC(Sf)+PTC(Pd)+PTC(Pc)
Vx5 = Teses e Dissertações	TD(Te)+TD(Di)

O terceiro e último conjunto de dados (Tabela 4.6) é composto pelas variáveis de análise qualitativa. Nesse conjunto, procura-se identificar relações entre as variáveis do sistema de avaliação de bolsas de pesquisa do CNPq e o sistema de avaliação de pós-graduação da CAPES. As variáveis são formadas pelos níveis das bolsas 1A, 1B, 1C, 2A, 2B, 2C e pelo conceito do programa de pós-graduação em que atua o pesquisador: D7, D6, D5, D4, D3.

A classificação dos bolsistas do CNPq é atribuída pelos CAs, com base na avaliação dos projetos de pesquisa e dos currículos dos pesquisadores candidatos às bolsas. Quanto ao conceito obtido pela avaliação da CAPES, este obedece a uma escala numérica que varia de 1 a 7, na qual 7 é o ápice, 5 é a nota máxima admitida para programas que ofereçam somente mestrado, e 3 é a nota equivalente ao padrão mínimo para que sejam validados os diplomas expedidos pelos programas (CAPES(b), 1998).

Tabela 4.6 - Estrutura do terceiro conjunto de dados

Colunas	Composição
Vx1 = 1A	Bolsista categoria 1A
Vx2 = 1B	Bolsista categoria 1B
Vx3 = 1C	Bolsista categoria 1C
Vx4 = 2A	Bolsista categoria 2A
Vx5 = 2B	Bolsista categoria 2B
Vx6 = 2C	Bolsista categoria 2C
Vy1 = D7	Docente vinculado a programa com grau 7
Vy2 = D6	Docente vinculado a programa com grau 6
Vy3 = D5	Docente vinculado a programa com grau 5
Vy4 = D4	Docente vinculado a programa com grau 4
Vy5 = D3	Docente vinculado a programa com grau 3

4.6.2 NORMALIZAÇÃO DOS CONJUNTOS DE DADOS

Para a fase de normalização[3] dos dados para Kohonen (SOM) foram realizados vários testes, alterando-se as equações de tratamento dos conjuntos de análise (Equação 4.1, 4.2, 4.3, 4.4 (Rautenberg, 1998)) (Tabela 4.7).

Contudo, as equações 4.5 e 4.6 (Tabela 4.7) apresentaram melhores resultados na primeira fase de análise, que consistia na descoberta de conhecimento, onde *x* e y são vetores de entrada e *i* o índice da variável de entrada. Essas equações empregadas para cada conjunto de dados tendem a relativizar todas as variáveis de análise qualitativas e quantitativas. No primeiro conjunto (Tabela 4.4) utilizou-se a equação 4.5 para as variáveis de produção (Vx1,..,Vxn), e a equação 4.6 para as variáveis de formação acadêmica (Vy1,..,Vyn).

[3] Normalização ou padronização dos dados pode ser entendida como uma transformação nos dados originais, de modo que os dados transformados apresentem valores equivalentes entre as dimensões a serem analisadas.

Tabela 4.7 - Lista de equações utilizadas na normalização da rede

$x_i = \sqrt{\frac{x_i}{\max(x)}}$	(4.1)
$x_i = (x_i - \min(x)/(\max(x) - \min(x))$	(4.2)
$x_i = x_i / \max(x)$	(4.3)
$x_i = \sqrt[4]{\frac{x_i}{\max(x)}}$	(4.4)
$x_i = \frac{Vxi}{\sum Vx}$	(4.5)
$y_i = \frac{Vyi}{\sum Vy}$	(4.6)

No segundo conjunto, referente à tabela 4.5, para as variáveis de produção (Vx1,..,Vxn), a equação 4.5 foi considerada, e para as demais variáveis, foi utilizada a normalização apresentada na Tabela 4.8.

Tabela 4.8 - Normalização utilizada para as variáveis individuais

Variável	Valores
Vz1 = Sexo	0 = Feminino e 1 = Masculino
Vz2 = Nível de formação	0 = Graduação, 0.33 = Especialização, 0.66 = Mestrado e 1 = Doutorado
Vz3 = Idade	Utilizada a equação 4.4
Vz4 = Nacionalidade	0 = Estrangeiro e 1 = Brasileiro
Vz5 = Dedicação	0 = Parcial e 1 = Integral

Com relação ao último conjunto (Tabela 4.6), empregou-se a equação 4.5 para as variáveis de qualificação atribuídas pelo CNPq (Vx1,..,Vxn), e a equação 4.6 para as variáveis de qualificação atribuídas pela CAPES (Vy1,..,Vyn).

Quanto à normalização para a geração das regras de associação, foram realizadas consultas utilizando-se as equações apresentadas na Tabela 4.9, onde cada coluna representa o percentual relativo em relação ao total de produções e formação do grupo (Tabela 4.4), total de produções do pesquisador (Tabela 4.5) e total de bolsistas e docentes vinculados a programas de pós-graduação (Tabela 4.6). A equação 4.7 é utilizada nas variáveis (Vx1,..,Vxn) (Tabela 4.4, 4.5 e 4.6), e a equação 4.8 nas variáveis (Vy1,..,Vyn) (Tabela 4.4 e 4.6).

Tabela 4.9 - Lista de equações utilizadas nas consultas para a geração das regras

$x_i = \left(\frac{Vxi}{\sum Vx} \right) * 100$	(4.7)
$y_i = \left(\frac{Vyi}{\sum Vy} \right) * 100$	(4.8)

De posse dessas consultas, foram gerados os conjuntos utilizados na extração das regras. Os dados foram categorizados utilizando-se as especificações apresentadas na Tabela 4.10, onde cada variável de entrada é dividida em cinco faixas. Se um valor pertence a uma das faixas, recebe o conteúdo 1, caso contrário, é atribuído o valor 0. Para as variáveis individuais referentes à Tabela 4.5, são utilizadas as normalizações da Tabela 4.11.

Tabela 4.10 - Normalização utilizada na categorização das variáveis percentuais

Variável	Faixas Percentuais
Primeiro conjunto de dados (Tabela 4.4) (Produção C,T&A e Titulação dos pesquisadores)	Faixa 1 = (maior que 0 e menor que 9) Faixa 2 = (maior igual a 9 e menor 25) Faixa 3 = (maior igual a 25 e menor 50) Faixa 4 = (maior igual a 50 e menor 85) Faixa 5 = (maior igual a 85)
Segundo conjunto de dados. (Tabela 4.5). Variáveis Vx1, Vx2, Vx3, Vx4, Vx5 (Produção C,T&A)	Faixa 1 = (maior que 0 e menor que 10) Faixa 2 = (maior igual a 10 e menor 25) Faixa 3 = (maior igual a 25 e menor 50) Faixa 4 = (maior igual a 50 e menor 85) Faixa 5 = (maior igual a 85)
Terceiro conjunto de dados (Tabela 4.6) (Nível da Bolsa de Pesquisa e Conceito do Programa de Pós-graduação)	Faixa 1 = (maior que 0 e menor que 9) Faixa 2 = (maior igual a 9 e menor 25) Faixa 3 = (maior igual a 25 e menor 50) Faixa 4 = (maior igual a 50 e menor 85) Faixa 5 = (maior igual a 85)

Tabela 4.11 - Normalização utilizada na categorização das variáveis individuais

Variável	Faixas de Categorias
Vz1 = Sexo	Faixa 1 = (Masculino) e Faixa 2 = (Feminino)
Vz2 = Nível de formação	Faixa 1 = (Graduação), Faixa 2 = (Especialização), Faixa 3 = (Mestrado) e Faixa 4 = (Doutorado)
Vz3 = Idade	Faixa 1 = (menor que 35 anos) Faixa 2 = (maior igual a 35 anos e menor igual a 49 anos) Faixa 3 = (maior que 49 anos)
Vz4 = Nacionalidade	Faixa 1 = (Brasileiro) e Faixa 2 = (Estrangeiro)
Vz5 = Dedicação	Faixa 1 = (Integral) e Faixa 2 = (Parcial)

4.7 ANÁLISE DOS DADOS

A análise dos dados é realizada em duas etapas. Primeiramente, são identificados potenciais agrupamentos entre os cinco agrupamentos definidos para cada

análise (C1,..,C5), que através de seus centros em relação às variáveis de entrada (V1,..,Vn) permitem uma visualização gráfica do comportamento dos dados. Em uma segunda etapa são geradas as regras desses agrupamentos escolhidos, a fim de mapear o conhecimento extraído de uma maneira mais compreensível.

Os gráficos são gerados a partir dos centros de cada agrupamento, ou seja, são empregadas as matrizes de pesos de cada análise. As análises efetuadas são divididas por Grandes Áreas do Conhecimento. Para efeitos de validação do trabalho, as análises concentraram-se na grande área de Engenharias e Ciências da Computação, sendo as demais áreas analisadas de maneira mais geral. Esse procedimento foi adotado tanto na fase de extração de conhecimento e aplicação das técnicas de MD quanto na fase de geração e comparação dos índices de qualificação e produtividade dos grupos de pesquisa, em relação ao algoritmo utilizado pelo CNPq. Os centros dos agrupamentos utilizados na geração dos gráficos e os desvios-padrão são apresentados no Anexo III.

Após isso, são identificados os principais agrupamentos para análise e geração de regras. Para cada regra são fornecidos (1) um suporte e uma confiança individual, considerando-se somente os dados que compõem o agrupamento, (2) um suporte e uma confiança geral, sendo que todos os dados da grande área Engenharias e Ciências da Computação são considerados, e (3) a composição de cada regra ao nível dos estratos (classificação esta atribuída pelo algoritmo de hierarquização). Os estratos “A” e “B” indicam grupos *Consolidados,* “C” e “D” *Em Consolidação* e “E” *Em Formação* (Guimarães et al., 1999).

4.7.1 PRIMEIRO CONJUNTO DE DADOS

O primeiro conjunto de dados agrupa as variáveis de classificação da Produção C,T&A e o nível de titulação máxima dos pesquisadores do grupo de pesquisa. O objetivo é verificar a relação entre a natureza dos veículos de publicação e a titulação de seus autores. Além disso, procura-se identificar a variação no comportamento dessas relações entre as Grandes Áreas do Conhecimento. Os resultados podem ser vistos na Figura 4.2.

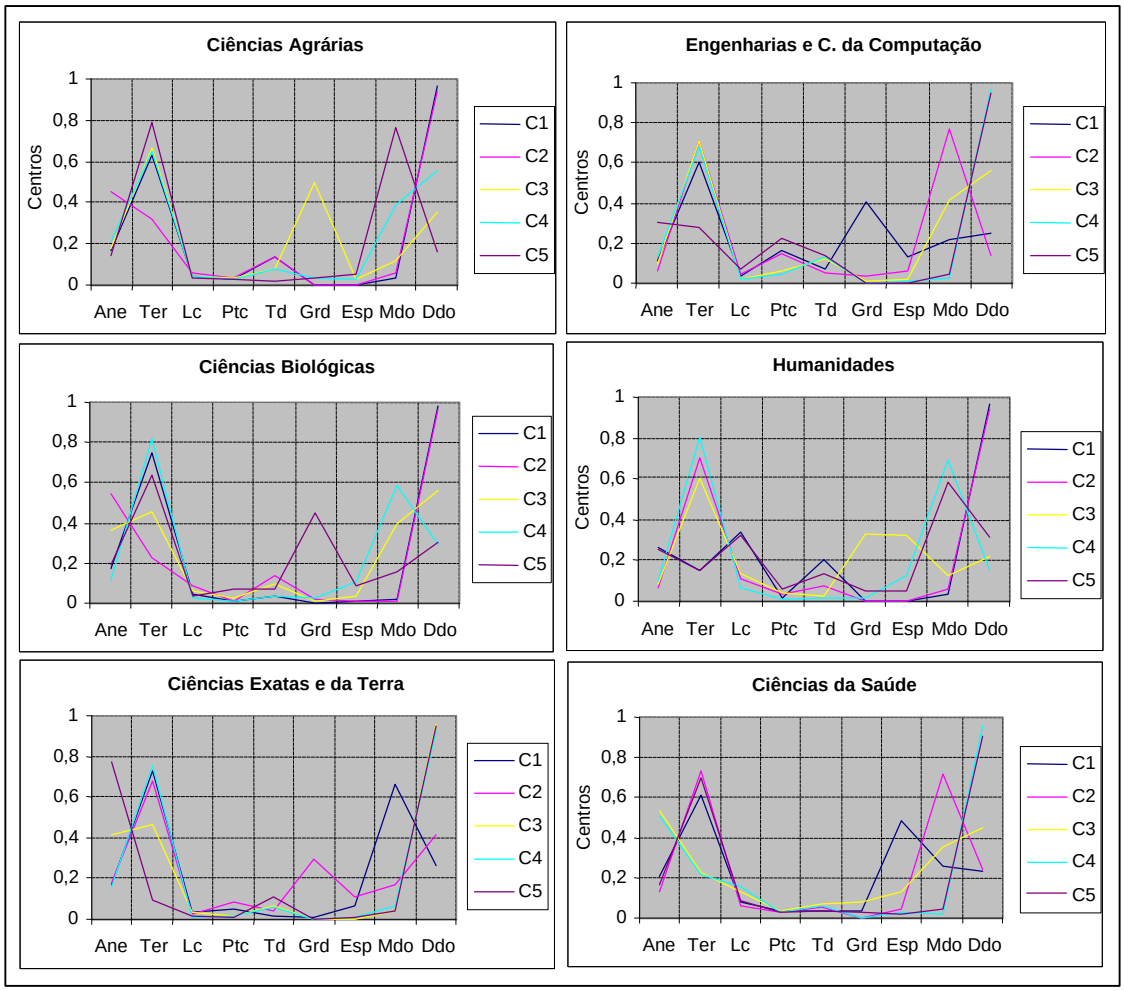


Figura 4.2 - Evolução dos agrupamentos para o primeiro conjunto de dados

Pode-se verificar nos gráficos apresentados uma maior concentração nas variáveis de produção (ANE, TER) e nas variáveis de formação (MDO e DDO). As variáveis de formação demonstram ser as que promovem a alocação dos itens em cada agrupamento, sendo verificados grupos distintos para doutores, mestres, especialistas e graduados.

Os agrupamentos, apesar de não apresentarem a mesma ordem entre as grandes áreas, promovem uma evolução semelhante. Contudo, em algumas grandes áreas são identificados padrões específicos, como nas Engenharias e Ciências da Computação, nas quais pode ser observado um grande número de produções técnicas (PTC), isto para os agrupamentos C1, C2 e C5. Já nas Humanidades, existe um grande volume de Livros e Capítulos (LC) para os agrupamentos C1 e C5.

Para a análise das Engenharias e Ciências da Computação, três agrupamentos são considerados (C1, C2 e C5). Esses agrupamentos são separados levando-se em conta a concentração por nível de formação, ou seja, o agrupamento C1 apresenta uma maior concentração de pesquisadores da graduação, o agrupamento C2 apresenta maior concentração dos pesquisadores mestres, e o agrupamento C5 maior concentração de pesquisadores doutores.

Em uma análise mais geral, os gráficos foram divididos em seis critérios, conforme apresentado na Tabela 4.12. Os critérios auxiliam na análise de como as variáveis envolvidas promovem a separação dos agrupamentos.

Tabela 4.12 - Análise dos critérios de separação dos agrupamentos do primeiro conjunto

Critérios	**Comportamentos**	**Significado**
Artigos (ANE) e Trabalhos em Eventos (TER)	• nas Agrárias, ANE e TER separam o agrupamento C2 dos demais; • nas Engenharias e C. da Computação, ANE e TER separam o agrupamento C5 dos demais; • nas Biológicas, os agrupamentos C2, C3 e os demais estão separados; • nas Humanidades, os agrupamentos C1 e C5 se separam dos demais; • nas Exatas, os agrupamentos C3, C5 e os demais estão separados; • nas Ciências da Saúde separam os agrupamentos C3 e C4 dos demais.	Em geral, existem um ou dois agrupamentos que se reúnem com base em artigos, enquanto que os demais se agrupam com base nos trabalhos. ANE e TER contribuem para a separação dos agrupamentos.
Livros e Capítulos (LC)	• Nas Agrárias, Eng. e C. Comp., Biológicas e Exatas, a variável LC praticamente não promove a separação de agrupamentos; • Nas Humanas, LC separa-se em dois conjuntos distintos; • Na Saúde, a separação em dois conjuntos ocorre, mas é pouco relevante.	A variável LC possui maior relevância nas Humanidades separando os agrupamentos em dois conjuntos. Já na Saúde ocorre mas com menos importância. Isso demonstra a natureza dessas áreas nas quais se produzem em geral mais livros que as demais.
Produção Técnica (PTC)	• Nas Agrárias, Biológicas, Humanidades, Exatas e Saúde, essa variável não promove a separação dos agrupamentos; • Nas Eng. e C. da Comp. PTC, separam os agrupamentos C3 e C4 dos demais.	Isso pode ser verificado pela natureza da produção nas áreas de nas Eng. e C. da Comp. que possuem grande ênfase na produção técnica (software, produto e processo). Essa variável é importante para a separação de agrupamentos na área.
Teses e Dissertações (TD)	• Nas Agrárias, Biológicas e Exatas, pode-se observar a divisão em três conjuntos; • Na Saúde, um conjunto contendo os agrupamentos é formado; • Nas Humanidades, os agrupamentos estão todos separados; • Nas Eng. e C. Comp., dois conjuntos são formados.	Normalmente teses e dissertações não promovem a separação dos agrupamentos, pois em geral compõem de maneira uniforme as produção C&T em todas as grandes áreas, com exceção para a área de Humanidades, que possui um número maior de teses e dissertações.
Titulação (Graduação-GRD e Especialização-ESP)	• Nas agrárias, Eng. e C. Comp., Biológicas, Humanidades e Exatas, existe um agrupamento que se separa dos demais; • Na Saúde, a divisão dos agrupamentos ocorre no nível de especialização.	Pode-se verificar que, na maioria das Grandes Áreas, GRD é uma variável relevante na determinação dos agrupamentos. Quanto à variável ESP, esta possui mais importância na Saúde, pois o número de pesquisadores especialistas é maior que nas outras Grandes Áreas.
Titulação (Mestrado-MDO e Doutorado-DDO)	• Em geral, os agrupamentos se dividem em quatro conjuntos, com exceção das Exatas, que se divide em três conjuntos.	As variáveis MDO e DDO são determinantes na geração dos agrupamentos. Isso ocorre devido ao maior número de integrantes com essas titulações.

A partir dos agrupamentos identificados para a análise (C1, C2 e C5) são apresentadas as regras (Tabela 4.13). A primeira regra do agrupamento C1 afirma que se um grupo possui mais de 35% de trabalhos em eventos, então este possui mais de 9% de pesquisadores com graduação. O suporte e a confiança dessa regra em termos de agrupamento é alto, 80,55% e 95,16%, respectivamente, enquanto que o suporte geral é de apenas 9,16%, e a confiança geral é de 17,38%. A regra é formada em sua maioria por grupos em consolidação (estratos C e D), com 44,82%. A segunda regra do agrupamento C1 apresenta a mesma relação em termos percentuais para artigos. Entretanto, possui suporte e confiança mais baixos, tanto individual com 29,16% e 94,29%, quanto geral com 3,7% e 14,52%. Essa regra é formada em sua maioria por grupos em consolidação e em formação (D e E) com 76,2%.

A regra do agrupamento C2 indica que, se um grupo possui mais de 53% de produções em anais de eventos, então este possui mais de 12% de pesquisadores mestres. Essa regra possui suporte e confiança individual de 83,62% e 100% e suporte e confiança geral de 38,09 e 47,96%. Constitui-se, em sua maioria, por grupos em formação (E) com 57,73%. Por último, a regra do agrupamento C5 apresenta uma relação em que um grupo que possua mais de 35% de produções em eventos possui mais de 60% de pesquisadores doutores. A regra possui suporte individual de 42,38%, suporte geral de 52,51%, confiança individual e geral de 100% e 69,98%, respectivamente, sendo formada principalmente por grupos consolidados (A e B), 54,69% e em consolidação (C) com 31,25%.

Tabela 4.13 - Regras geradas para o primeiro conjunto de dados nas Engenharias e Ciências da Computação

Agrp	Regras	Suporte (%)		Confiança (%)		Estratos (%)				
		S	SG	C	CG	A	B	C	D	E
C1	Se % de TER no grupo > 35% então % de graduados no grupo > 9%	80,55	9,16	95,16	17,38	8,62	18,97	22,41	22,41	27,59
	Se % de ANE no grupo > 35% então % de graduados no grupo > 9%	29,16	3,70	94,29	14,52	0,00	14,29	9,52	38,10	38,10
C2	Se % de TER no grupo > 53% então % de mestres no grupo > 12%	83,62	38,09	100,00	47,96	4,12	13,40	14,43	10,31	57,73
C5	Se % de TER no grupo > 35% então % de doutores no grupo > 60%	42,38	52,51	100,00	69,98	15,63	39,06	31,25	10,94	3,13

A Figura 4.3 apresenta uma representação das regras, deduzidas entre a titulação máxima dos pesquisadores do grupo, segundo o percentual de participação e o tipo de veículo de publicação de produção C&T. Para o primeiro agrupamento (C1), o processo

demonstrou que grupos com mais de 9% de graduados tendem a ter mais de 35% de sua produção concentrada em trabalhos em eventos e artigos. O agrupamento C2 é mais ilustrativo na relação, pois indica que grupos com mais de 12% de mestres concentram mais de 53% de sua produção em trabalhos em eventos. Por último, o agrupamento C3 apresenta que mais de 60% dos doutores possui uma produção concentrada nos trabalhos em eventos em relação aos demais tipos de produção. Isso pode ser visto como uma característica das Engenharias e Ciências da Computação, em que o ciclo do conhecimento exige veículos de rápida divulgação.

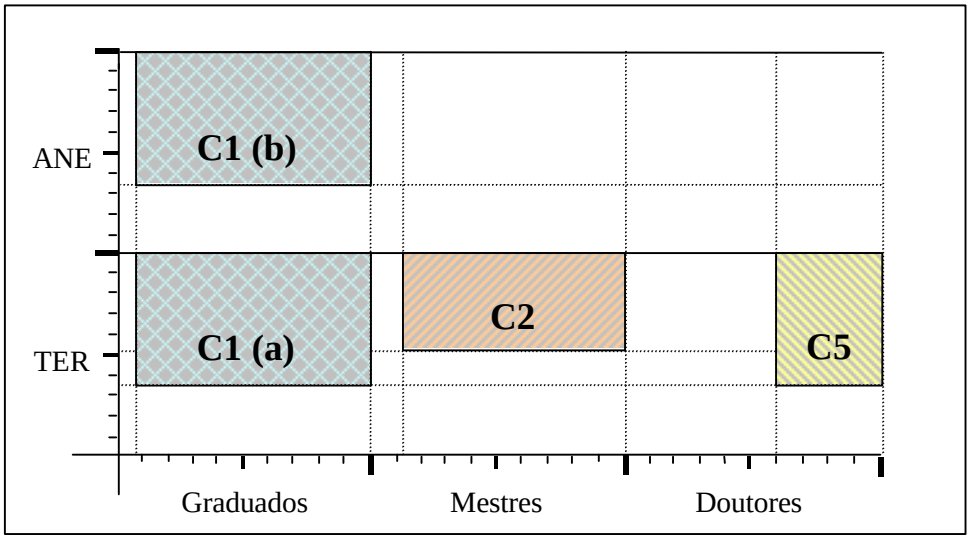


Figura 4.3 - Representação gráfica das regras geradas no primeiro conjunto de dados

4.7.2 SEGUNDO CONJUNTO DE DADOS

O segundo conjunto de dados procura estabelecer uma relação entre os dados de perfil dos pesquisadores do grupo e o tipo de veículo de publicação dos itens de produção do grupo.

As análises associadas a este conjunto de dados apresentam a relação entre os dados individuais dos pesquisadores e a sua produção C&T. A Figura 4.4 demonstra esses relacionamentos.

Em geral, a evolução dos gráficos nas áreas segue o mesmo padrão quanto às variáveis individuais. Com relação às variáveis de produção, trabalhos e artigos (TER e ANE) apresentam uma maior concentração, enquanto que as variáveis individuais (SX)

e (DED) determinam o agrupamento dos itens. Podem ser verificados grupos de pesquisadores por sexo e dedicação (exclusiva ou parcial) bem distintos.

Em algumas áreas podem ser identificados padrões específicos, como nas Engenharias e Ciências Humanas. Nas Engenharias, pode ser observado um número mais elevado de produções técnicas (PTC) para o agrupamento C1, enquanto que nas Humanidades existe um grande volume de Livros e Capítulos (LC).

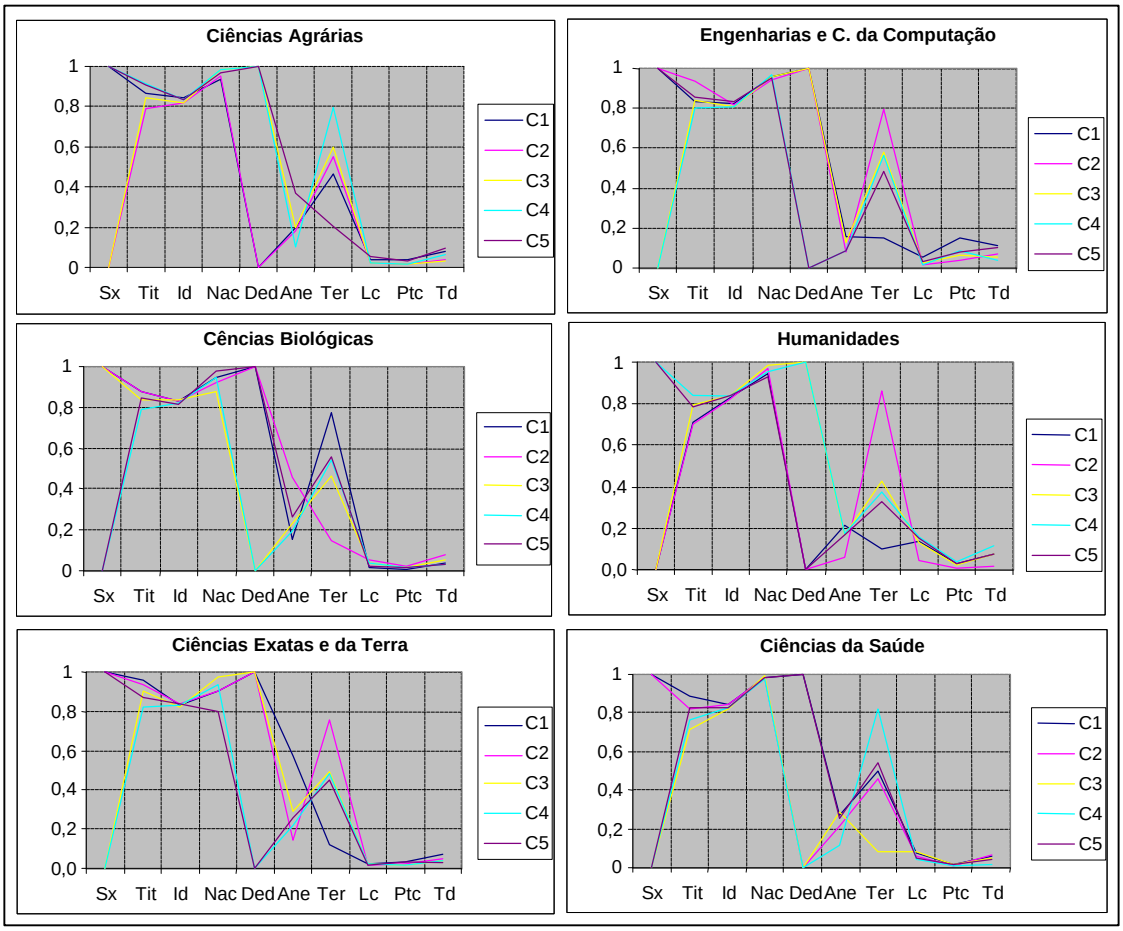


Figura 4.4 - Evolução dos agrupamentos para o segundo conjunto de dados

O gráfico das Engenharias e Ciências da Computação permite destacar três agrupamentos (C1, C2 e C3). Esses agrupamentos foram separados levando-se em conta a concentração por tipo de produção C&T e sexo, ou seja, o agrupamento C1 apresenta concentração em produção técnica (PTC) e pesquisadores masculinos, o agrupamento C2 apresenta maior concentração para trabalhos em eventos (TER) e pesquisadores masculinos, e o agrupamento C3 maior concentração nos trabalhos em eventos (TER) e pesquisadores femininos.

Para a análise conjunta das Grandes Áreas do Conhecimento, os gráficos foram divididos em cinco critérios, conforme apresentado na Tabela 4.14. Os critérios auxiliam na análise de como as variáveis envolvidas promovem a separação dos agrupamentos.

Tabela 4.14 - Análise dos critérios de separação dos agrupamentos do segundo conjunto

Critérios	**Comportamentos**	**Significado**
Sexo (SX)	• Em todas as Grandes Áreas, SX separa os agrupamentos em dois conjuntos.	A variável SX divide os agrupamentos em masculino e feminino, mas provê pouca informação. Uma abordagem seria utilizar essa variável em uma análise posterior ou mesmo modificar sua normalização.
Titulação Máxima (TIT), Idade (ID) e Nacionalidade (NAC)	• As variáveis TIT, ID e NAC em todas as Grandes Áreas agrupam-se formando três novos grupos.	Esse comportamento uniforme pode ser explicado pela concentração da produção C&T de pesquisadores na faixa do 30 aos 50, mestre e doutores brasileiros.
Dedicação (DED)	• Em todas as Grandes Áreas, DED separa os agrupamentos em dois conjuntos.	A variável DED contribui na separação dos agrupamentos.
Trabalhos em Eventos (TER)	• Em todas as Grandes Áreas, TER separa os agrupamentos em três conjuntos.	Esta variável possui grande importância na separação dos agrupamentos. Como mencionado anteriormente, este é o veículo de divulgação científica mais utilizado pelos pesquisadores.
Artigos (ANE), Livros e Capítulos (LC), Produção Técnica (PTC) e Teses e Dissertações (TD)	• Em geral, essas variáveis têm menor influência na separação dos agrupamentos. • A variável ANE promove a separação em dois conjuntos, com exceção às Eng. E C. Comp., na qual a variável PTC possui maior representatividade. • Nas Humanidades, a variável LC contribui na separação dos agrupamentos. • Já a variável TD contribui de maneira homogênea em todas as Grandes Áreas.	A variável ANE possui a maior importância para a separação dos agrupamentos considerando-se esses veículos de publicação. Pode-se verificar uma maior influência da variável PTC nas Eng. e C. Comp e da variável LC nas Humanidades.

A partir dos agrupamentos identificados para a análise (C1, C2 e C3) são apresentadas as regras na Tabela 4.15. A primeira regra do agrupamento C1 mostra que pesquisadores doutores do sexo masculino, com idade entre 35 e 49 anos, brasileiros e com dedicação exclusiva possuem mais de 25% de trabalhos em eventos. Esta possui um suporte e confiança individual de 17,98% e 53,38%, e um suporte e confiança geral de 21,66% e 88,96%, respectivamente. A segunda regra do agrupamento C1 apresenta a mesma relação em termos percentuais para produção técnica, com suporte e confiança individual de 11,05% e 19,47%, e suporte e confiança geral de 5,75% e 15,47%.

Tabela 4.15 - Regras geradas para o segundo conjunto de dados nas Engenharias e Ciências da Computação

Agrp	Regras	Suporte (%)		Confiança (%)		Pertencem a Grupos dos Estratos (%)				
		S	SG	C	CG	A	B	C	D	E
C1	Se pesquisador (masculino, doutorado, idade entre 35 e 49 anos, brasileiro, dedicação exclusiva) então % de trabalhos em eventos do pesquisador > 25%	17,98	21,66	53,38	88,96	12,18	33,97	30,13	16,67	7,05
	Se pesquisador (masculino, doutor, idade entre 35 e 49 anos, brasileiro, dedicação exclusiva) então % de produção técnica do pesquisador > 25%	11,05	5,75	19,47	15,47	11,76	35,29	29,41	15,69	7,84
C2	Se pesquisador (masculino, doutor, idade entre 35 e 49 anos, brasileiro, dedicação exclusiva) então % de trabalhos em eventos do pesquisador > 50%	51,06	18,40	99,00	75,55	7,46	31,96	30,85	18,75	10,99
	Se pesquisador (masculino, doutor, idade entre 35 e 49 anos, brasileiro, dedicação exclusiva) então % de teses e dissertações do pesquisador > 10%	20,28	7,31	39,32	40,90	10,15	40,61	32,49	13,96	2,79
	Se pesquisador (masculino, doutor, brasileiro, dedicação exclusiva) então % de trabalhos do pesquisador >= 50% e % de teses e dissertações do pesquisador > 10%	27,79	10,01	36,49	26,95	12,22	40,56	29,63	13,89	3,70
C3	Se pesquisador (feminino, doutor, idade entre 35 e 49 anos, brasileiro, dedicação exclusiva) então % de trabalhos em eventos do pesquisador > 50%	35,39	4,04	78,99	78,99	3,67	25,69	38,53	17,43	14,68

O agrupamento C2 apresenta três regras, sendo a primeira formada por pesquisadores doutores, sexo masculino, com idade entre 35 e 49 anos, brasileiros com dedicação exclusiva que possuem mais de 50% de trabalhos em eventos. Essa regra possui suporte e confiança individual de 51,06% e 99%, e suporte e confiança geral de 18,40% e 75,55%, respectivamente. A segunda regra mostra que pesquisadores doutores, sexo masculino, com idade entre 35 e 49 anos, brasileiros, com dedicação exclusiva, possuem mais de 10% de teses e dissertações, com suporte e confiança individual de 20,28% e 39,32%, e suporte e confiança geral de 7,31% e 40,90%. Para a última regra do agrupamento C2, retirando-se da premissa a variável idade, há indicação de que mais de 50% da produção individual é de trabalhos em eventos e mais de 10% são de teses e dissertações. Isso ocorre com suporte e confiança individual de 27,79% e 36,49% e suporte e confiança geral de 10,01% e 26,95%.

As regras do agrupamento C3 apresentam os pesquisadores brasileiros doutores, sexo feminino, idade entre 35 e 49 anos, com dedicação exclusiva. Na primeira regra, mais de 50% são de trabalhos em eventos, com suporte individual de 35,39% e suporte geral de 4,04%. A confiança da regra tanto individual quanto geral é de 78,99%. Para segunda regra, também formada pelos trabalhos em eventos dos pesquisadores, elevando-se o percentual para mais de 85%, esta apresenta um suporte individual de 11,04%, um suporte geral de 1,26% e uma confiança individual e geral de 24,64%. Em todas as regras dos agrupamentos (C1, C2 e C3), pode-se observar uma concentração nos grupos consolidados (B) e em consolidação (C).

A comparação das regras geradas com a classificação que o algoritmo de hierarquização fornece para os grupos dos pesquisadores permite observar que a veracidade das regras é diferente para cada classificação dos grupos em que atuam esses pesquisadores. Um estudo adicional poderia incluir o estrato do grupos aos antecedentes das regras, para confirmar essa observação.

4.7.3 TERCEIRO CONJUNTO DE DADOS

As análises realizadas no terceiro conjunto de dados apresentam a relação entre o sistema de qualificação de bolsistas do CNPq e o sistema de avaliação da pós-graduação da CAPES. A Figura 4.5 mostra essa relação para as Grandes Áreas do Conhecimento.

A evolução dos gráficos demonstra uma concentração na variável de análise (2C) referente aos bolsistas CNPq, e uma distribuição mais homogênea nas variáveis de análise referente aos docentes vinculados a programas de pós-graduação avaliados pela CAPES. Contudo, nas variáveis de bolsistas (CNPq), a área de Ciências Exatas e da Terra apresenta uma concentração nos níveis (2A) e (2C), e a área de Ciências da Saúde, uma concentração elevada no nível (2C). Nas variáveis dos docentes vinculados a cursos de pós-graduação (CAPES), verifica-se uma concentração no nível D7 para Ciências Biológicas e Ciências Exatas e da Terra, bem como a baixa concentração no nível D6 para Ciências da Saúde.

No gráfico de Engenharia e Ciência da Computação, três agrupamentos são analisados (C1, C4 e C5). Esses agrupamentos foram separados levando-se em conta a concentração das variáveis de avaliação da CAPES, ou seja, o agrupamento C1 apresenta uma maior concentração para D6 (grau 6), o agrupamento C4 apresenta uma maior concentração para D5 (grau 5), e o agrupamento C5 uma maior concentração para D4 (grau 4). As concentrações representam os centros dos agrupamentos para cada dimensão. O que se verifica é uma evolução mais acentuada das variáveis que representam os docentes, devido ao número mais expressivo destes em relação aos bolsistas de pesquisa.

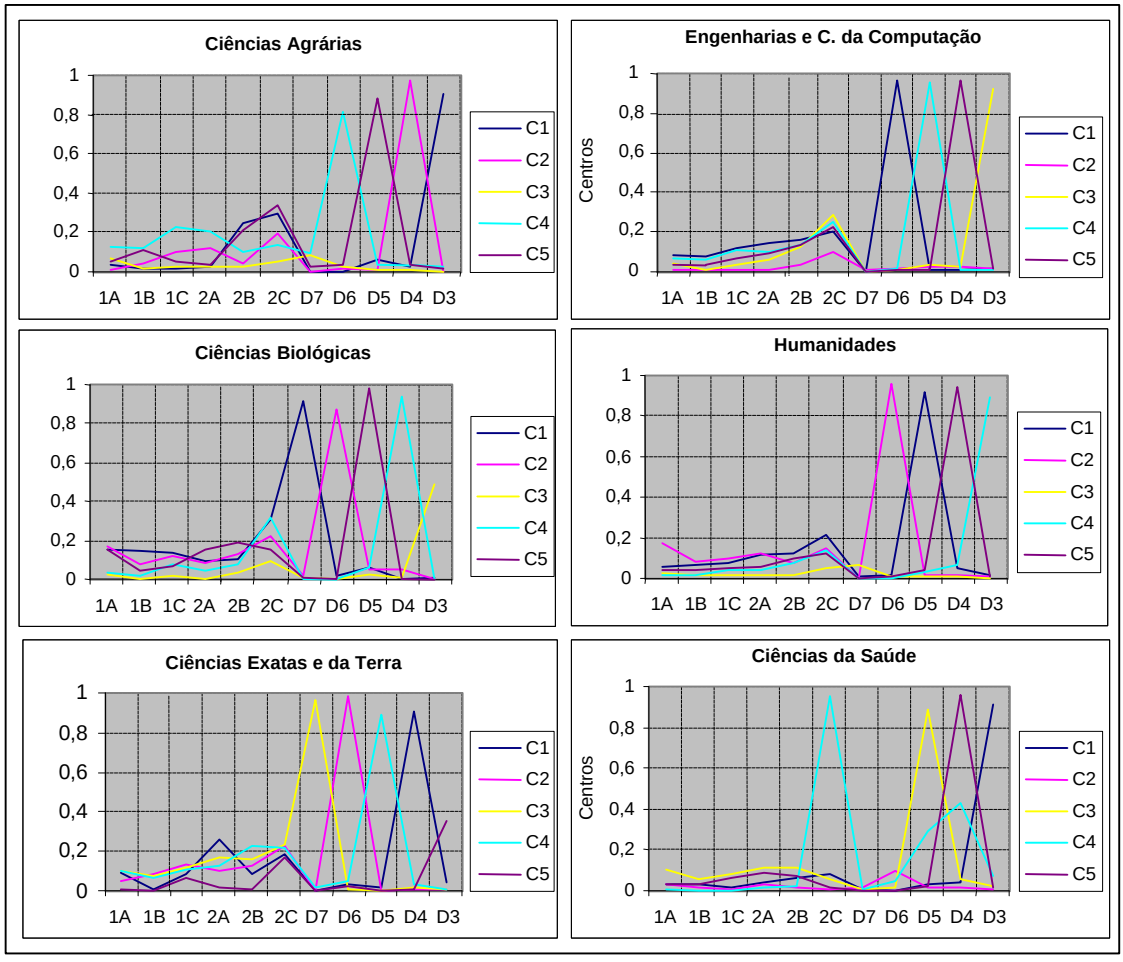


Figura 4.5 - Evolução dos agrupamentos para o terceiro conjunto de dados

De uma maneira geral, os gráficos demonstram não existir uma relação direta entre os docentes participantes de programas de pós-graduação com desempenho mais elevado e os bolsistas de pesquisa de maior nível. Isso pode ser evidenciado analisando-se as variáveis 2A, 2B e 2C, que representam a maior parcela do total de bolsistas. Um caso mais particular ocorre nas Ciências da Saúde, em que a variável 2C

separa o agrupamento C4 dos demais. Nesse agrupamento, estão incluídos os grupos que possuem em sua maioria bolsistas 2C e docentes que atuam em cursos com conceitos 6 e 7.

Para a análise conjunta das Grandes Áreas do Conhecimento, os gráficos foram divididos em cinco critérios, conforme apresentado na Tabela 4.16. Esses critérios auxiliam na análise de separação dos agrupamentos.

Tabela 4.16 - Análise dos critérios de separação dos agrupamentos do terceiro conjunto

Critérios	**Comportamentos**	**Significado**
Bolsistas (1A, 1B, 2A, 2B)	• Essas variáveis em geral não promovem a separação dos agrupamentos. Contudo, nas Agrárias, as variáveis 1C, 2A e 2B separam os agrupamentos em três conjuntos. • Nas Exatas, a variável 2A separa os agrupamentos em três conjuntos.	Essas variáveis possuem pouca representatividade na separação dos agrupamentos, visto que a maior concentração é de bolsistas 2C. Nas Agrárias, 1C, 2A e 2B possuem alguma contribuição na formação dos agrupamentos, enquanto que nas Exatas, 2A e 2B contribuem para a separação dos agrupamentos. Pode-se ainda verificar uma relação entre as variáveis 2A e D4 para o agrupamento C1, que reúne os grupos formados em sua maioria por docentes vinculados a cursos com grau 4 e bolsistas 2A.
Bolsista 2C	• Nas Agrárias, Biológicas e Humanidades, 2C separa o agrupamento C3 dos demais, enquanto que nas Eng. e C. Comp. separa o agrupamento C2 dos demais. • Nas Exatas, não existe separação, e na Saúde separa o agrupamento C4 dos demais.	Esta variável contribui para a separação dos agrupamentos principalmente na Saúde, na qual pode ser identificado o agrupamento C4, que possui na sua maioria bolsistas 2C e docentes vinculados a cursos com graus 5 e 4.
Docentes vinculados a cursos com grau 7 (D7)	• Nas Biológicas e Exatas, D7 divide os agrupamentos em dois conjuntos. • Nas demais Áreas, a variável D7 possui pouca ou nenhuma participação na separação dos agrupamentos.	A variável D7 nas Biológicas reúne grupos que possuem em sua maioria docentes vinculados a cursos com grau 7 e bolsistas 2C, e as Exatas reúnem grupos que possuem em sua maioria docentes de cursos com grau 7 e bolsistas distribuídos nas diversas categorias. Essas duas áreas possuem o maior número de docentes em cursos com essa avaliação.
Docentes vinculados a cursos com grau 6 (D6)	• Em geral divide o agrupamento em dois conjuntos, com exceção da Saúde, que possui pouca representatividade.	A variável D6, com exceção da Saúde, contribui para a separação dos agrupamentos. Reúne grupos que em sua maioria são formados por docentes vinculados a cursos com grau 6.
Docentes vinculados a cursos com grau 5 (D5), 4 (D4) e 3 (D3)	• Individualmente separam os agrupamentos em dois conjuntos.	Cada variável contribui para a separação dos agrupamentos reunindo os grupos que possuem em sua maioria docentes em cursos com graus 5, 4 e 3.

Após a análise inicial dos agrupamentos (C1, C4 e C5) são apresentadas as regras geradas para cada agrupamento (Tabela 4.17). O agrupamento C1 apresenta duas regras com uma concentração na variável D6 (grau 6). A primeira regra afirma que, se um grupo possui mais de 8% de bolsistas com nível 2B, então mais de 40% de seus pesquisadores estão vinculados a cursos com grau 6. Essa regra possui suporte e confiança individual de 28,36% e 100% e suporte e confiança geral de 8,31% e 39,67%, respectivamente. A segunda regra apresenta uma relação inversa, em que um

grupo formado por mais de 30% de bolsistas 2C possui mais de 9% de seus pesquisadores vinculados a cursos com grau 6, com suporte e confiança individual de 31,94% e 100%, e suporte e confiança geral de 9% e 31,61%, respectivamente. As regras geradas nesse agrupamento são formadas por grupos consolidados (A e B), com 89,48% para a primeira regra, e com 79,44% para a segunda regra.

As regras geradas para o agrupamento C4 mostram a relação entre bolsistas 2B e 2C perante os pesquisadores vinculados a cursos com grau 5. A primeira regra afirma que, se um grupo possui mais de 13% de bolsistas com nível 2B, então o percentual de docentes vinculados a cursos com grau 5 é maior que 50%. Essa regra possui suporte e confiança individual de 22,93% e 98,39%, e suporte e confiança geral de 5,28% e 25,31%. A segunda regra apresenta uma relação em que um grupo formado por mais de 15% de bolsistas 2C possui mais de 50% de docentes vinculados a cursos com grau 6. Essa regra possui suporte e confiança individual de 38,35% e 95,33% e suporte e confiança geral de 8,83% e 28,65%. As regras geradas nesse agrupamento são formadas por grupos consolidados (B), com 68,85% para a primeira regra, e 63,73% para a segunda regra.

Por último, o agrupamento C5 apresenta uma regra em que o percentual de docentes vinculados a cursos com grau 4 é maior que 45%, para um grupo que possua mais de 9% de bolsistas com nível 2C. Os suportes individual e geral são de 30,45% e 6,15%%, respectivamente. Quanto à confiança, essa é de 100% para a individual, e de 19,89% para a geral, sendo o agrupamento formado em sua maioria por grupos em consolidação (C), com 64,18%.

Tabela 4.17 - Regras geradas para o terceiro conjunto de dados nas Engenharias e Ciências da Computação

Agrp	Regras	Suporte (%)		Confiança (%)		Estratos (%)				
		S	SG	C	CG	A	B	C	D	E
C1	Se % de 2B no grupo > 8% então % de D6 no grupo > 41%	28,36	8,31	100,00	39,67	30,53	58,95	9,47	1,05	0,00
	Se % de 2C no grupo > 30% então % de D6 no grupo > 9%	31,94	9,00	100,00	31,61	20,56	58,88	15,89	4,67	0,00
C4	Se % de 2B no grupo > 13% então % de D5 no grupo > 50%	22,93	5,28	98,39	25,31	9,84	68,85	19,67	1,64	0,00
	Se % de 2C no grupo > 15% então % de D6 no grupo > 50%	38,35	8,83	95,33	28,65	6,86	63,73	23,53	5,88	0,00
C5	Se % de 2C no grupo > 9% então % de D4 no grupo > 45%	30,45	6,15	100,00	19,89	0,00	14,93	64,18	16,42	4,48

A Figura 4.6 apresenta uma representação das regras geradas a partir da relação entre os bolsistas CNPq e docentes vinculados a programas de pós-graduação avaliados pela CAPES, segundo o percentual de participação desses pesquisadores em grupos de pesquisa.

Nota-se na representação gráfica que não existem intersecções entre bolsistas de níveis mais elevados com os docentes vinculados a programas de pós-graduação. Isso pode ser explicado da seguinte maneira: primeiro, pelo suporte sugerido para a extração das regras superior a 20% e confiança superior a 90%; segundo, pela própria natureza das Engenharias e Ciências da Computação, em que o principal veículo de publicação são os trabalhos em congressos.

Também sugere um foco maior no desenvolvimento de projetos, muitas vezes diretamente com a iniciativa privada, sem requisições de bolsa de pesquisa e, conseqüentemente, sem a avaliação dos Comitês Assessores. Do total de bolsistas das Engenharias e Ciências da Computação, 35% são bolsistas 2C e 21% bolsistas 2B. Outro ponto importante a salientar é a participação destes pesquisadores em grupos de pesquisa consolidados (estratos A e B) para os agrupamentos C1 e C4, e programas de pós-graduação com conceitos 5 e 6 para os mesmos agrupamentos, demonstrando que existe uma grande correlação entre a classificação atribuída pela hierarquização (discutida na próxima seção) e a avaliação efetuada pela CAPES..

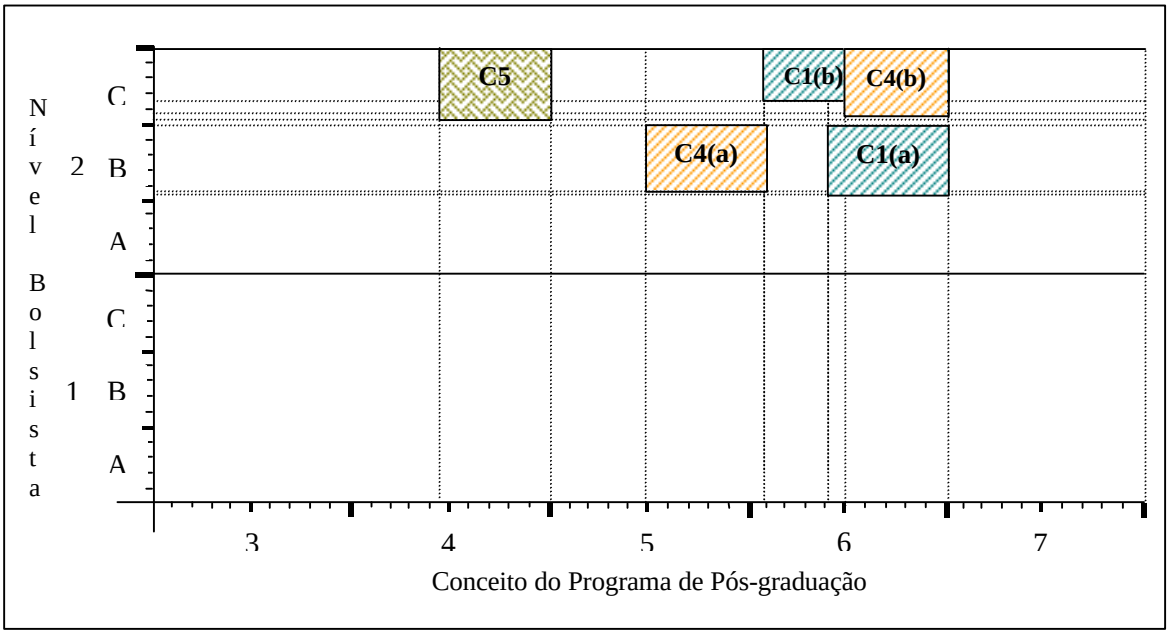


Figura 4.6 - Representação gráfica das regras geradas no segundo conjunto de dados

4.7.4 COMPARAÇÃO COM O ALGORITMO DE HIERARQUIZAÇÃO DOS GRUPOS DE PESQUISA

As seções anteriores mostraram os resultados da aplicação dos algoritmo SOM e as regras de associação como métodos de mineração de dados aplicados em bases de gestão de C&T. A combinação de variáveis qualitativas e quantitativas permite estabelecer uma base comparativa entre os algoritmos de mineração de dados e o algoritmo de hierarquização (Guimarães et al., 1999) para a estratificação dos grupos de pesquisa brasileiros. O objetivo é verificar a consistência das classes identificadas pelo algoritmo SOM, quando comparadas à classificação parametrizada do algoritmo de hierarquização.

Para tal, divide-se o estudo em duas etapas. Na primeira etapa, utilizou-se uma visão não parametrizada, ou seja, sem as tabelas de ponderações empregadas pelo algoritmo de hierarquização. A seguir, apresenta-se uma visão parametrizada utilizando-se as tabelas de ponderações. As Tabelas 4.18 e 4.19 apresentam as ponderações necessárias ao cálculo do índice (Q), e a Tabela 4.20 apresenta as ponderações para os itens de produção C&T, necessários à geração do índice (P). Em uma segunda etapa, são mostradas a distribuição dos grupos de pesquisa por estrato nas três abordagens e a relação de semelhança entre essas abordagens.

Tabela 4.18 - Tabela de ponderações por categoria/nível de bolsistas do CNPq (Guimarães et al., 1999)

Categoria/Nível	1A	1B	1C	2A	2B	2C
w_h	1	0,8	0,65	0,55	0,45	0,35

Tabela 4.19 - Tabela de ponderações por conceito atribuído pela CAPES aos programas de pós-graduação (1997-1998) (Guimarães et al., 1999)

Conceito	7	6	5	4	3
v_g	1,00	0,70	0,50	0,35	0,25

Tabela 4.20 - Tabela de ponderações por natureza de produção C&T (Guimarães et al., 1999)

Subconjunto	YI	Natureza	Tipo	Ponderação (v)
1	Y1	ANE	An	0,3
			Ae	0,7
2	Y2	TER	Tc	0,5
			Ar	0,2
			Rv	0,2
			Rs	0,1
3	Y3	LCL	Li	0,7
			Cl	0,3
4	Y4	PTC	Sf	0,33
			Pd	0,33
			Pc	0,33
5	Y5	T&D	Te	0,7
			Di	0,3

Para as comparações, são utilizados os dados da Grande Área de Conhecimento de Engenharias e Ciências da Computação, composta por 1155 grupos de pesquisa. Essas comparações foram realizadas empregando-se duas redes neurais, ou seja, uma rede para a geração do índice de qualificação (Q) e outra para a geração do índice de produtividade (P).

A primeira rede utiliza parte do primeiro conjunto de dados (Tabela 4.4), composta pelas variáveis de análise quantitativas (ANE, TER, LC, PTC, TD). Na normalização dos dados de entrada foram realizados testes com algumas equações, sendo que a equação $Y_i = \sqrt{\log\left(1+\frac{\sum T_j * v}{n}\right)}$ (utilizada no algoritmo de hierarquização) apresentou melhores resultados, onde T representa os tipos de produção C&T de mesma natureza, n o número de doutores pertencentes ao grupo de pesquisa, e v as ponderações atribuídas a cada tipo de trabalho de mesma natureza (não parametrizada v=1 e parametrizada, utilizando-se a Tabela 4.18). Esses dados servem de entrada para a rede, que produz uma matriz de pesos utilizada na determinação do índice de produtividade.

Após o processamento da rede, o índice inicial é determinado por $U_i = \frac{\sum Y_{ij} + w_{ij}}{x}$, onde Y_{ij} é a entrada da rede, w_{ij} o peso da matriz SOM para o índice i e j, e x o número de variáveis de entradas na rede. Em uma segunda etapa determina-se

o índice *z*, em um intervalo entre –2.5 e 2.5, através da equação $zi = \frac{U_i - \mu(U_s)}{\sigma(U_s)}$ (Guimarães et al., 1999), sendo $\mu(U_s)$ a média aritmética e $\sigma(U_s)$ o desvio-padrão. Por último, o índice de produtividade é gerado utilizando-se a equação $P_i = 50 + 20z_i$ (Guimarães et al., 1999).

A segunda rede está baseada no terceiro conjunto de dados (Tabela 4.6), composta pelas variáveis de análise qualitativas (BPQ, DOC) mais o número de doutores no grupo de pesquisa. Na normalização utilizou-se a equação $N_i = \sqrt[4]{\frac{n_j}{\max(n)}}$ para a primeira entrada, e as equações $B_i = \frac{b_i * v}{n}$, e $D_i = \frac{d_i * v}{n}$ para as demais entradas, onde *bj* representa os bolsistas doutores avaliados pelo CNPq, *dj* representa os pesquisadores doutores vinculados a programas de pós-graduação avaliados pela CAPES, com grau igual ou superior a 3, *n* o número de doutores pertencentes a cada grupo de pesquisa, e *v* as ponderações (não parametrizada *v*=1 e parametrizada através das Tabelas 4.16 e 4.17).

Após a fase de agrupamento da rede, o índice inicial é determinado por $X_i = \frac{\sum Y_{ij} + w_{ij}}{x}$, onde Y_{ij} é a entrada da rede, w_{ij} o peso da matriz SOM e *x* o número de entrada da rede. Na etapa seguinte, calcula-se o índice *z* utilizando-se a equação $zi = \frac{Xi - \mu(X_s)}{\sigma(X_s)}$ (Guimarães et al., 1999), sendo $\mu(X_s)$ a média aritmética e $\sigma(U_s)$ o desvio-padrão. O índice *z* deve ficar em um intervalo entre –2.5 e 2.5. Por último, o índice de qualificação é produzido utilizando-se a equação $Q_i = 50 + 20z_i$ (Guimarães et al., 1999).

A Figura 4.7 mostra a disposição dos grupos em classes para as duas abordagens utilizadas, uma parametrizada e outra não parametrizada (ambas utilizando o algoritmo SOM), em relação ao índice de produtividade e de qualificação. Essas classes seguem as definições originais do algoritmo de hierarquização, possuindo uma distribuição de freqüência com amplitude igual a 5, sendo a primeira classe formada pelos grupos com

(Q) inferior a 20, a segunda classe entre 20 e 25, e assim sucessivamente. A última classe é formada pelos grupos com índice (Q) igual ou superior a 85.

De modo geral, as análises demonstram a relação entre qualidade e produtividade, sendo que, quanto maior a qualificação de um grupo, maior sua produtividade. Contudo, pode-se notar que a análise não parametrizada apresenta classes que não obedecem a essa tendência. Isso deve-se ao fato de a rede ser invariável ao grau de importância (pesos) atribuído aos diferentes tipos de produção ou à classificação atribuída pelos sistemas de avaliação das agências. Observa-se também que a rede parametrizada aproxima-se mais dos resultados do algoritmo de hierarquização.

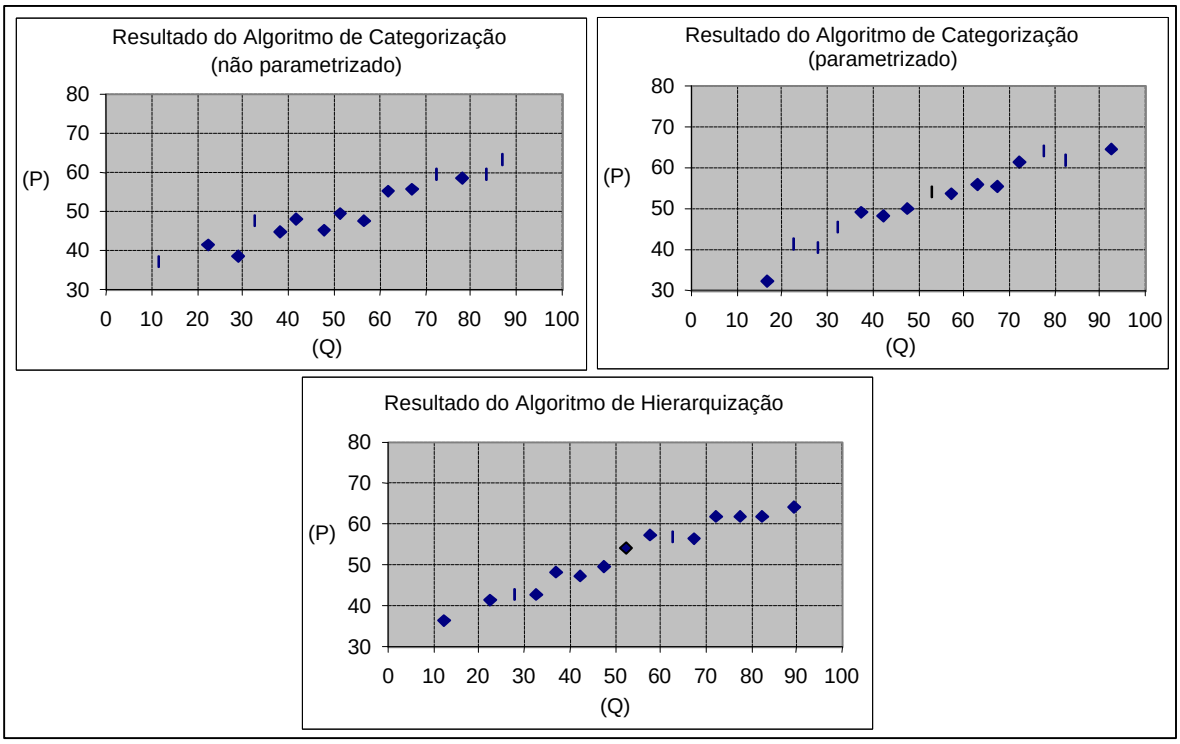


Figura 4.7 - Correlação entre as médias dos índices de qualificação (Q) e produtividade (P), para o algoritmo de hierarquização e o algoritmo de categorização

Como última análise, são apresentados os resultados em níveis de estratos para o algoritmo de categorização (parametrizado e não parametrizado), e para o algoritmo de hierarquização. Os estratos variam entre "A" e "E", sendo que essa classificação é obtida utilizando-se o índice de qualificação (Q). De posse desse índice, são calculados os decis, distribuídos conforme a Tabela 4.21.

Tabela 4.21 - Tabela de intervalos (decis) utilizados na determinação dos estratos (Guimarães et al., 1999).

Estratos	A	B	C	D	E
Intervalos (decis)	$\geq D_9$	D_6 | -- D_9	D_3 | -- D_6	D_1 | --D_3	$< D_1$

Calculando-se os decis para o índice (Q) final, são atribuídos os estratos correspondentes a cada grupo. Na Tabela 4.22 apresenta-se a distribuição dos grupos por estrato em cada uma das abordagens. Após a estratificação, os grupos foram avaliados com relação à semelhança entre os resultados fornecidos pela utilização do algoritmo de categorização, em comparação aos resultados fornecidos pela hierarquização.

Tabela 4.22 - Relação de grupos de pesquisa por estrato para as Engenharias e Ciências da Computação

Estratos	Categorização (Não parametrizada)	Relação CnP/H	Categorização (Parametrizada)	Relação CP/H	Hierarquização
A	113	1.11	102	1	102
B	277	0.91	304	1	304
C	333	1.04	315	0.99	319
D	187	0.97	195	1.02	192
E	245	1.03	239	1	238
Total	**1155**		**1155**		**1155**

A Tabela 4.23 apresenta a relação que utiliza a visão não parametrizada, ou seja, não foram atribuídos pesos às diferentes classificações para as variáveis de entrada. O percentual de 66% mostra, sobretudo, que essa abordagem permite um estudo preliminar, podendo ser obtidos, a partir desses resultados, os pesos iniciais. Esses pesos poderiam sofrer ajustes, a fim de melhorar o desempenho na classificação dos grupos. Por outro lado, pode ser evidenciado que o algoritmo de categorização, quando são apresentados os valores de entrada semelhantes ao algoritmo de hierarquização, fornece um resultado final similar em 88% para os grupos analisados (Tabela 4.24). Nos casos restantes, verificou-se que em grande parte o índice (Q) posiciona-se na fronteira dos intervalos, ocorrendo uma classificação diferente em relação ao algoritmo de hierarquização.

Para melhor compreensão das Tabelas 4.22, 4.23 e 4.24, analisa-se o estrato A. Na visão não parametrizada, dos 113 grupos classificados nesse estrato, 71 grupos também são classificados de maneira semelhante em relação ao algoritmo de hierarquização, que possui 102 grupos no estrato A, com uma relação de 69,6%. Para a visão parametrizada, o número de grupos no estrato A é o mesmo do algoritmo de

hierarquização, mas destes, 87 são classificados de maneira semelhante, ou seja, pertencem aos dois conjuntos analisados, perfazendo uma relação de 85,3%.

Tabela 4.23 - Percentual de grupos classificados de maneira semelhante para as Engenharias e Ciências da Computação (Categorização não parametrizada x Hierarquização)

Estratos	Categorização (Não parametrizada)	Hierarquização	Relação (%)
A	71	102	69,6%
B	213	304	70,1%
C	217	319	68,0%
D	75	192	39,1%
E	194	238	81,5%
Total	**770**	**1155**	**66,7%**

Tabela 4.24 - Percentual de grupos classificados de maneira semelhante para as Engenharias e Ciências da Computação (Categorização parametrizada x Hierarquização)

Estratos	Categorização (Parametrizada)	Hierarquização	Relação(%)
A	87	102	85,3%
B	263	304	86,5%
C	274	319	85,9%
D	165	192	85,9%
E	227	238	95,4%
Total	**1016**	**1155**	**88,0%**

Com referência aos valores 66,7% e 88%, pode-se verificar ainda, para cada uma das visões (parametrizada e não parametrizada), a distribuição dos estratos classificados de maneira diferente em relação ao algoritmo de hierarquização. A Tabela 4.25 apresenta esses valores para a visão não parametrizada, e a Tabela 4.26 para a parametrizada. Utilizando-se como exemplo o estrato C da Tabela 4.25, pode-se verificar que, dos 333 grupos atribuídos ao estrato C pelo algoritmo de categorização, 217 pertencem à mesma classe atribuída pelo algoritmo de hierarquização, e os outros 116 grupos pertencem a diferentes classes. Desses 116, dois pertencem ao estrato A, 33 ao estrato B e 81 ao estrato D. Sob outro enfoque, pode-se verificar que, dos 319 grupos atribuídos pelo algoritmo de hierarquização ao estrato C, 217 pertencem à mesma classe atribuída pelo algoritmo de categorização, e o restante está representado nos demais estratos (35 para o B, 52 para o D e 15 para o E).

Tabela 4.25 - Total de grupos estratificados pelo algoritmo de Categorização (C) e Hierarquização (H) para as Engenharias e Ciências da Computação.
(Categorização não parametrizada x Hierarquização)

C \ H	A	B	C	D	E	Total
A	71	42	0	0	0	**113**
B	29	213	35	0	0	**277**
C	2	33	217	81	0	**333**
D	0	16	52	75	44	**187**
E	0	0	15	36	194	**245**
Total	**102**	**304**	**319**	**192**	**238**	**1155**

Tabela 4.26 - Total de grupos estratificados pelo algoritmo de Categorização (C) e Hierarquização (H) para as Engenharias e Ciências da Computação.
(Categorização parametrizada x Hierarquização)

C \ H	A	B	C	D	E	Total
A	87	15	0	0	0	**102**
B	15	263	26	0	0	**304**
C	0	26	274	15	0	**315**
D	0	0	19	165	11	**195**
E	0	0	0	12	227	**239**
Total	**102**	**304**	**319**	**192**	**238**	**1155**

4.8 CONSIDERAÇÕES FINAIS

Neste capítulo, foram apresentadas diferentes aplicações de MD em bases de C&T. Os resultados demonstram a viabilidade na aplicação de técnicas de mineração de dados na busca de conhecimento em bases de C&T, bem como a utilização de técnicas de análise de categorização (*clustering*) para elaboração de um *ranking* de qualidade perante a produtividade, como um modelo complementar.

O primeiro conjunto de aplicações permite traçar relações entre dados de identificação e dados de qualificação de unidades de análise. Estudos dessa natureza permitem conhecer, por exemplo, a relação entre o nível de formação de autores e seu veículo predileto de divulgação (e.g., nas Engenharias e Ciências da Computação, o aumento na participação de mestres promove o aumento na produção de trabalhos em eventos).

A segunda área de aplicação de MD neste trabalho corresponde à categorização e à classificação das unidades de análise. O trabalho mostra que o algoritmo SOM praticamente reflete os valores encontrados pelo algoritmo de hierarquização

tradicional de avaliação C&T e que, quando liberado (sem a influência das ponderações), permite uma análise investigativa das diferenças.

Cabe ressaltar que a utilização dessas técnicas em processos de avaliação de C&T é algo bastante recente, sendo ainda necessário estudos para a determinação de equações mais adequadas, bem como a determinação de novos algoritmos visando melhorar o processo puramente não parametrizado.

4.9 LIMITAÇÕES DO TRABALHO

Embora a solução proposta apresente-se como uma ferramenta útil e mesmo como uma ferramenta adicional à descoberta de conhecimento em bases de C&T e hierarquização dos grupos de pesquisa, deve-se destacar alguns pontos limitantes na realização e nos resultados do trabalho, tais como:

- *Grande Área do Conhecimento:* as análises foram concentradas na Grande Área de Engenharias e Ciências da Computação, a fim de demonstrar a viabilidade do trabalho proposto. Um estudo mais detalhado deve contemplar todas as grandes áreas do conhecimento;
- *hierarquização*: os resultados obtidos são preliminares, sendo necessário estudos adicionais que busquem melhorar o resultado fornecido pelo algoritmo de categorização (SOM);
- *algoritmos utilizados*: os algoritmos utilizados no trabalho são alguns dos vários algoritmos utilizados em MD. Pode-se citar árvores de decisão, algoritmos genéticos, lógica difusa, tecnologia de agentes, *link* análise, entre outros. A implementação e a utilização de outros algoritmos visam incrementar os resultados obtidos na extração de conhecimento.

5 CONCLUSÕES E RECOMENDAÇÕES

"A descoberta de que o universo está em expansão foi uma das grandes revoluções intelectuais do século XX. Depois dela torna-se fácil perguntar por que ninguém pensou nisso antes."

Stephen W. Hawking

A construção, a disseminação e o acesso às diversas plataformas de informações em C&T no país permitem vislumbrar um conjunto de aplicações voltadas à extração de conhecimento. Em particular, a aplicação de Mineração de Dados pode contribuir para elucidar relações entre as diversas unidades de análise nas bases de dados, ou mesmo para a formação de categorias ou classes de unidades, segundo critérios utilizados na avaliação e no planejamento de C&T.

Para tal, foram utilizadas técnicas e conceitos da área de extração de conhecimento na base de dados do Diretório dos Grupos de Pesquisa no Brasil (CNPq) e na relação com a avaliação da pós-graduação realizada pela CAPES perante a classificação dos bolsistas de pesquisa do CNPq.

Neste capítulo, são apresentadas as conclusões do estudo que possibilitaram a elaboração deste trabalho, bem como idéias para trabalhos futuros.

5.1 CONCLUSÕES

A pesquisa científica e tecnológica adquire cada vez mais importância e impacto perante a sociedade, sendo realizada cada vez mais por grupos de pesquisa. Nesse contexto, a base dos grupos de pesquisa mantida pelo CNPq, juntamente com a base de currículos e informações provenientes da base de pós-graduação, produz um importante mapa da pesquisa realizada em nível nacional.

A primeira e principal conclusão demonstra a viabilidade de serem utilizadas técnicas de Mineração de Dados na extração de conhecimento em bases de dados de

C&T. Particularmente o trabalho mostra a utilidade da técnica de categorização na hierarquização dos Grupos de Pesquisa no Brasil como um modelo de suporte à definição dos pesos utilizados na visão parametrizada.

Com relação às técnicas utilizadas (Redes Neurais e Regras de Associação), estas demonstram ser úteis na resolução de diversos problemas, e quanto à utilização dessas técnicas no problema proposto, observa-se que elas ampliam o conjunto de ferramentas que podem prover subsídios na avaliação em C&T. Levando-se em consideração um caráter mais exploratório dos dados, essas ferramentas podem validar conhecimentos explícitos, ou produzir novos padrões, auxiliando na tomada de decisão.

Os resultados mostram que é possível aplicar MD tanto no estudo de relações entre as unidades de análise (identificação, titulação, produção C&T) como na categorização ou classificação dos grupos de pesquisa. Para o primeiro caso, mostrou-se (1) a relação entre o nível de formação dos integrantes dos grupos de pesquisa e a produção C&T, (2) os dados de identificação em relação à produção C&T e (3) a relação entre os sistemas de avaliação de bolsistas (CNPq) e pós-graduação (CAPES). No segundo caso, pode-se observar uma potencial utilização do algoritmo SOM na categorização dos grupos de pesquisa.

Por outro lado, os resultados, ainda que satisfatórios, são preliminares, sendo necessários novos estudos, que devem ser concentrados tanto em questões de normalização dos dados, métricas de similaridade e avaliação dos dados quanto na determinação de novas técnicas de avaliação em C&T.

5.2 TRABALHOS FUTUROS

Quanto à continuidade do trabalho, deve-se enfatizar o estudo e a implementação de novas técnicas estatísticas e de IA. Isso permitiria a elaboração de uma ferramenta composta por um conjunto maior de técnicas, possibilitando assim diversas visões e maneiras de extração de conhecimento e identificação de padrões, aumentando o suporte às decisões gerenciais.

Essas pretensões, surgidas durante o desenvolvimento do trabalho e principalmente na fase de análise dos resultados, são descritas a seguir, levando-se em consideração os seguintes aspectos:

- *bases de dados de C&T*: as bases de C&T, nesse atual estágio, encontram-se em processo de integração. Contudo, estudos devem ser feitos para se produzir um modelo integrado, que comporte visões de pesquisadores, de grupos de pesquisa e de pós-graduação. Utilizando-se um modelo integrado, aumentam as possibilidades de análises e de cruzamento de informações.
- *algoritmos adicionais*: a utilização de outros algoritmos, entre eles Árvores de Decisão e Raciocínio Baseado em Casos, aplicados sob um modelo integrado de C&T, permitiriam a ampliação na ferramenta visando o suporte à avaliação em C&T. Cabe ressaltar o estudo de novos algoritmos com enfoque específico em C&T, levando-se em consideração que algoritmos dessa natureza são utilizados ou encontram-se em análise nas agências de fomento.
- *criação de uma ferramenta de Mineração de Dados em C&T*: a ferramenta utilizada como protótipo para o trabalho comporta apenas dois algoritmos, não possuindo um módulo de pré-processamento e um módulo de análises finais. A implementação desses módulos e novos algoritmos, aplicados sobre uma base integrada de C&T, forneceriam uma ferramenta útil aos processos de análise realizados pelas agências de C&T no país.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, Rakesh; IMICLINSKI, Tomasz; SWAMI, Arun. Mining association rules between sets of items in large database. In: ***Proceedings of the ACM SIGMOD Conference***, Washington, D.C., May 1993. Disponível em. <http://www.cs.bham.ac.uk/~anp/bibtex/kdd.bib.html>. Acesso em: 28 abr. 1999.

AGRAWAL, Rakesh; MANNILA, Heikki; SRIKANT, Ramakrishnan; TOIVONEN, Hannu; VERKAMO, A. Inkeri. Fast discovery of association rules. In: ***Advances in Knowledge Discovery and Data Mining***, AAAI Press / The MIT Press, MIT, Cambridge, Massachusetts, and London, England, 1996, p. 307-328.

ANDERY, Maria Amália; MICHELETTO, Nilza; SÉRIO, Tereza Maria Pires et al. ***Para compreender a ciência: uma perspectiva histórica***. 5. ed. Rio de Janeiro: Espaço e Tempo, 1988. 446 p.

ARBID, M. A. ***Brains, machines and mathematics***, 2. ed., Springer-Verlag, Berlin, 1987.

ARRUDA, Mauro Fernando Maria. A indústria e o desenvolvimento tecnológico nacional. In: ***Ciência & Tecnologia: Alicerces do Desenvolvimento,*** São Paulo: Cobram, 1994. 164 p. 1. ed. p. 23-44.

BACK, Barbro; IRJALA, Mikko; SERE, Kaisa; VANHARANTA, Hannu. ***Managing complexity in large data bases using self-organizing maps***. Turku Center for Computer Science, Technical Report, No. 48, set. 1996.

BARBIERI, José Carlos. ***Ciência e tecnologia no Brasil: Uma nova política para um mundo global***. Escola de Administração de Empresas de São Paulo da Fundação Getúlio Vargas, 1993.

BERRY, Michel J. A., LINOFF, Gordon. ***Data mining techniques - for marketing, sales, and customer support***. John Wiley & Sons, New York, 1997.

BIGUS, Joseph P. ***Data mining with neural networks: Solving business problems from application development to decision support***. Computing McGraw-Hill, New York, NY, 1996.

CAMPANARIO, J. M. ***Using neural networks to study networks of scientific journals. In: Scientometrics***, Vol. 33, No.1, 1995. p. 23-40.

CAPES(a). ***Avaliação da pós-graduação: Procedimentos básicos***, 1999. Disponível em: <http://www.capes.gov.br>. Acesso em: 5 nov. 1999.

CAPES(b). ***Reformulação do sistema de avaliação da pós-graduação: O modelo a ser implantado na avaliação de 1998***. Disponível em: <http://www.capes.gov.br>. Acesso em: 5 nov 1999.

CAPES(c.) ***Sistema de avaliação: Coleta de Dados 5.0 - Manual do usuário***. Fundação CAPES/MEC e SCIRE-COPPE/UFRJ, jul. 1999.

CASTRO, Cláudio de Moura; OLIVEIRA, João Batista de Araújo e. ***Os recursos humanos para a ciência e tecnologia***. Fundação Getúlio Vargas, 1992.

CHEN, Ming-Syan; HAN, Jiawei; YU, Philip S. Data mining: an overview from a database perspective. In: ***IEEE Transactions on Knowledge and Data Engineering***, v. 8, n.6, p. 866-883, dez. 1996.

CHEUNG, David W.; NG, Vicent T.; FU, W.; FU; Yongjian. Efficient mining of association rules in distributed databases. In: ***IEEE Transactions on Knowledge and Data Engineering***, v. 8, no. 6, dez. 1996.

CNPq. ***Construindo o futuro: Propostas e Realizações da Gestão 95-98***. Brasília, 1998.

CRUZ, Carlos Henrique de Brito. In: ***Atraindo a Inteligência : o início de um processo: reflexões e debates da I Conferência Brasileira de Ciência e Tecnologia (1. : 1997 : Boston)***. Brasília : Ministério das Relações Exteriores, Departamento de Cooperação Científica, Técnica e Tecnológica, 1997. p. 275-291.

DIAS, Lindolpho de Carvalho. Panorama atual da ciência e tecnologia no Brasil. In: ***Atraindo a Inteligência : o início de um processo: reflexões e debates da I Conferência Brasileira de Ciência e Tecnologia (1. : 1997 : Boston)***. Brasília : Ministério das Relações Exteriores, Departamento de Cooperação Científica, Técnica e Tecnológica, 1997. p. 39-49.

FAPESP. ***Processo de avaliação***. 1999. Disponível em: <http://www.fapesp.br>. Acesso em: 9 nov. 1999.

FAYYAD, Usama M. Data mining and knowledge discovery: making sense out of data, ***IEEE Exper****t*, 1996a.

FAYYAD, Usama, PIATETSKY-SHAPIRO, Gregory, SMYTH, Padhraic. From data mining to knowledge discovery: An overview. In: ***Advances in Knowledge Discovery and Data Mining***, AAAI Press / The MIT Press, MIT, Cambridge, Massachusetts, and London, England, 1996b, p.1-34.

FERNÉ, Georges**.** Science & Technology in the new world order. In: ***Organization for Economic Cooperation and Development (OECD)***, Paris, 1993.

FINEP. ***O que é a FINEP***. 1999. Disponível em: <http://www.finep.gov.br/Scripts/siteCf.exe/cf_tab>. Acesso em: 9 nov. 1999.

FRENKEL, Daniel; NADAL, Jurandir. Detecção de eventos isquêmicos do eletrocardiograma utilizando redes neurais. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 19, Rio de Janeiro. ***Anais do XIX Congresso Nacional da Sociedade Brasileira de Computação***, v.4 – Rio de Janeiro : EntreLugar, 1999. 584 p. p. 65-77.

GOLDBERG, David E. ***Genetic algorithms in search, optimization & machine learning***, University of Alabama, 1989.

GUIMARÃES, Reinaldo; GALVÃO, Gerson; COSAC, Silvana M.; LOURENÇO, Ricardo S., et al. ***A pesquisa no Brasil – Perfil da pesquisa no Brasil e hierarquização dos grupos de pesquisa a partir dos dados do Diretório dos Grupos de Pesquisa no Brasil***, 1999.

GUIMARÃES, Reinaldo. ***Avaliação e fomento de C&T no Brasil: Propostas para os anos 90***. Brasília: MCT/CNPq, 1994. 178 p.

HAIR Jr., Joseph F.; ANDERSON, Rolph E.; TATHAM, Ronald L.; BLACK, William C. ***Multivariate data analysis***. Prentice-Hall, Upper Saddle River, 5. ed., New Jersey, 1998.

HARRISON, Thomas H. ***Intranet data warehouse*****.** São Paulo: Berkeley Brasil, 1998.

HAYKIN, Simon. ***Neural networks: A comprehensive foundation***. Macmillan Publishing Company, New Jersey, 1994.

HOLSHEIMER, Marcel; SIEBES, Arno. ***Data mining: The search for knowledge in databases***. Computer Science/Department of Algorithmics and Architecture, Amsterdam, 1994. Disponível em: <http://www.cwi.nl/cwi/publications/index.html>. Acesso em: 17 ago. 1999.

HUGO, Marcel. ***Uma interface de reconhecimento de voz para o sistema de gerenciamento de central de informação de fretes***. Florianópolis, 1995. Dissertação (Mestrado em Engenharia de Produção) – Engenharia de Produção e Sistemas, UFSC.

JOHNSON, Richard A.; WICHERN, Dean W. ***Applied multivariate statistical analysis***. 4. ed., New Jersey: Prentice-Hall, 1998.

KASKI, Samuel; KANGAS, Jari; KOHONEN, Tuevo. Bibliography of Self-Organizing Map (SOM) papers: 1981-1997. In: ***Neural computing surveys 1***, p. 102-350, 1998. Disponível em: <http://www.icsi.berkeley.edu/~jagota/NCS/vol1.html>. Acesso em: 2 fev. 2000.

KAWATO, Mitsuo; UNO, Yoji; ISOBE, Michiaki; SUZUKI, Ryoji. Hierarquical neural network model for voluntary movement with application to robotics. In ***Artificial Neural Networks: Concepts and Control Application***, IEEE Computer Society Press, 1992. 584 p. p. 497-505.

KLIR, George J.; YUAN, Bo. ***Fuzzy sets and fuzzy logic: Theory and applications***. Prentice Hall PTR, Upper Saddle River, NJ, 1995.

KOHONEN, Teuvo. ***Self-Organizing maps***. Springer Series in Information Sciences, Heidelberg, Germany, 1995.

KOSTOFF, Ronald N. ***Research program peer review: Principles, practices, protocols***. 1997a. Disponível em: <http://www.dtic.mil/dtic/kostoff/index.html>. Acesso em: 2 fev. 2000.

KOSTOFF, Ronald N. ***The handbook of research of impact assessment***, ed. 7, 1997b. Disponível em: <http://www.dtic.mil/dtic/kostoff/index.html>. Acesso em: 2 fev. 2000.

KOVÁCS. Zsolt L. ***O Cérebro e a sua mente: Uma introdução à neurociência computacional***, São Paulo: Edição Acadêmica, 1997.

KRIEGER, Eduardo M., GALEMBECK, Fernando. A capacitação brasileira para a pesquisa. In: Simon Schwartzman. ***Ciência e tecnologia no Brasil: a capacitação para a pesquisa científica e tecnológica***, Rio de Janeiro: Editora Fundação Getulio Vargas, 1996, 370 p. v.3, p.1-18.

LOESCH, Claudio. ***Redes neurais artificiais: Fundamentos e modelos***. Blumenau: Ed. da Furb, 1996.

LOPES, Ana Lúcia Miranda. ***Avaliação cruzada da produtividade e qualidade de departamentos acadêmicos de uma universidade com um modelo de análise envoltória de dados e conjuntos difusos***. Florianópolis, 1998. Tese (Doutorado em Engenharia de Produção) - Engenharia de Produção e Sistemas, UFSC.

MCT. ***Ciência e tecnologia no governo federal 1997***. Brasília, 1998a. Disponível em: <http://www.mct.gov.br/publi/rel97.htm>. Acesso em: 21 out. 1999.

MCT. ***Programa de apoio ao desenvolvimento científico e tecnológico – PADCT III (Manual Operativo)***, ago. 1998b. Disponível em: <http://www.mct.gov.br>. Acesso em: 8 fev 2000.

MEDLER, David A. A brief history of connectionism. Department of Psychology, University of Alberta, Alberta, Canada, In: ***Neural Computing Survey 1***, p. 61-101, 1998. Disponível em: <http://www.icsi.berkeley.edu/~jagota/NCS/vol1.html>. Acesso em: 2 fev. 2000.

MEIS, Leopoldo de. Os cientistas e as implicações sócio-econômicas da distribuição da ciência e recursos humanos no planeta. In: ***Ciência & Tecnologia: Alicerces do Desenvolvimento,*** São Paulo: Cobram, 1994. p. 13-21. 164 p.

MELO, Saul Luiz de; CALOBA, Luiz Pereira; NADAL, Jurandir. Classificação de batimentos cardíacos utilizando rede neural com treinamento competitivo supervisionado. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 19, Rio de Janeiro. ***Anais do XIX Congresso Nacional da Sociedade Brasileira de Computação***, v.4 – Rio de Janeiro: EntreLugar, 1999. 584 p. p. 1-12.

MINSKY, M., PAPERT, S. ***Perceptrons***, Cambridge: MIT Press, 1969.

MORIN, Edgar. ***Ciência com consciência***. 3. ed. Rio de Janeiro: Bertrand Brasil, 1999. 336 p.

PACHECO, R. C. S.; BARCIA, R. M. Plataforma Lattes: ***Desenvolvimento de sistemas de informações gerenciais, integrados ao novo modelo de gestão do CNPq***, 1999. Projeto de pesquisa.

PACITTI, Tércio; ATKINSON, Cyril P. ***Programação e métodos computacionais***. 2. ed., v. 2, Livros Técnicos e Científicos Editora S/A, Rio de Janeiro, 1977.

PANDYA, Abhijit S.; MACY, Robert B. ***Pattern recognition with neural networks in C++***. CRC Press, Florida Atlantic University, Boca Raton, Florida, 1995.

PAZ, Salete do Bonfim; BORGES, Díbio Leandro. Uso de redes neuronais artificiais de função de base radial em previsão do consumo de energia elétrica. In: SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 19, Rio de Janeiro. ***Anais do XIX Congresso Nacional da Sociedade Brasileira de Computação***. Rio de Janeiro: EntreLugar, 1999. 584 p. v. 4, p. 481-493.

PRICE, Derek J. de Solla. ***O desenvolvimento da ciência: análise histórica, filosófica, sociológica e econômica***. Rio de Janeiro: Livros Técnicos e Científicos, 1976.

RAUTENBERG, Sandro. ***Predição de receitas de cores na estamparia têxtil através de redes neurais com função de base radial.***. Florianópolis, 1998. Dissertação (Mestrado em Engenharia de Produção) – Engenharia de Produção e Sistemas, UFSC.

SARDENBERG, Ronaldo Mota. ***Em Busca de novas conquistas***. Folha de S. Paulo, 10 de outubro de 1999, p. 13.

SCHWARTZMAN, Simon; CASTRO, Cláudio de Moura. ***Pesquisa universitária em questão***. Rio de Janeiro: Ed. Ícone, 1986.

SCHWARTZMAN, Simon; KRIEGER, Eduardo; GALEMBECK, Fernando; GUIMARÃES, Eduardo A.; BERTERO, Carlos Osmar. ***Ciência e tecnologia no Brasil: Uma nova política para um mundo global***, 1993.

SCHWARTZMAN, Simon; KRIEGER, Eduardo; GALEMBECK, Fernando; GUIMARÃES, Eduardo A.; BERTERO, Carlos Osmar. Ciência e tecnologia no Brasil: Uma nova política para um mundo global. In: ***Ciência e tecnologia no Brasil: Política Industrial, Mercado de Trabalho e Instituições de Apoio***, Rio de Janeiro: Editora da Fundação Getúlio Vargas, v. 2, 1995, 384 p, p.1-59.

SCHWARTZMAN, Simon. ***Formação da comunidade científica no Brasil***. São Paulo: Ed. Nacional; Rio de Janeiro : FINEP – Financiadora de Estudos e Projetos, 1979.

SCIAVICCO, L.; SICILIANO, B. ***Modeling and control of robot manipulators.*** McGraw-Hill, USA, 1996.

SILVA, Roberto Leal Lobo. Diagnóstico da ciência e tecnologia. In: ***Ciência & Tecnologia: Alicerces do Desenvolvimento,*** 1. ed. São Paulo: Cobram, 1994. 164 p. p. 13-21.

TAFNER, Malcon Anderson. ***Reconhecimento de palavras faladas isoladas usando redes neurais artificiais***. Florianópolis, 1996. Dissertação (Mestrado em Engenharia de Produção) – Engenharia de Produção e Sistemas, UFSC.

TODESCO, José L. ***Reconhecimento de padrões usando rede neuronal artificial com uma função de base radial: uma aplicação na classificação de cromossomos humanos***. Florianópolis, 1995. Tese (Doutorado em Engenharia de Produção) - Engenharia de Produção e Sistemas, UFSC.

TUFFANI, Maurício. ***Ranking da ciência***. Folha de S. Paulo, 12 setembro de 1999.

UNESCO/ICSU. Em defesa da ciência. ***Conferência Mundial sobre a Ciência: Ciência para o Século XXI: Um Novo Compromisso***, Budapeste, 1999. Disponível em: <http://www.mct.gov.br/sobre/galeria/bresser/conferencias/defesa.htm>. Acesso em: 28 out. 1999.

ZADEH, Lofti A. Fuzzy sets. ***Information and Control***, v. 8, 1965, p. 338-353.

Anexo I

Tabela 1a - Dispêndio segundo os instrumentos de fomento em R$ 1.000,00

Período	Bolsas no País		Bolsas no Exterior		Total de Bolsas		Apoio à Pesquisa		Total Geral	
Anual	Valor	Var %	Valor	Var %	Valor	Var %	Valor	Var %	Valor	Var %
1994	182.816,30		34.733,10		217.549,40		21.639,80		239.189,20	
1995	416.044,90	127,6	48.960,20	41,0	465.005,10	113,7	37.291,10	72,3	502.296,20	110,0
1996	422.877,60	1,6	41.944,00	-14,3	464.821,60	0,0	50.013,00	34,1	514.834,60	2,5
1997	415.284,10	-1,8	32.106,90	-23,5	447.391,00	-3,7	65.724,80	31,4	513.115,80	-0,3
1998	360.112,30	-13,3	26.894,80	-16,2	387.007,10	-13,5	38.882,90	-40,8	425.890,00	-17,0
1999 (Prev.)	**414.435,80**	**15,1**	**43.113,20**	**60,3**	**457.549,00**	**18,2**	**31.123,50**	**-20,0**	**488.672,50**	**14,7**

Fonte: http://www.mct.gov.br/estat/secav/disp_p_inst.htm em 27 de outubro de 1999

Tabela 1b - Dispêndio em bolsas segundo modalidade em R$ 1.000,00

Período Anual	Iniciação Científica	Aperfeiç./ Estágio/ Especial.	Mestrado	Doutorado	Iniciação Tecn. e Industrial	Pós-Doutorado	Produti-vidade em Pesquisa
1994	15.131	2.142	9.417	4.012	1.523	59	6.871
1995	18.790	2.397	10.960	4.965	1.710	89	7.188
1996	18.761	1.990	9.618	4.584	2.368	82	7.263
1997	18.856	1.896	7.764	5.032	2.597	55	7.394
1998	17.533	1.274	6.256	5.205	2.342	45	7.386
1999 (Prev.)	**18.650**	**920**	**5.938**	**5.875**	**1.850**	**40**	**7.350**

Período Anual	Pesq./ Especialista Visitante	Pesq. Associado	Desenv. Cient. Regional	Recém-Doutor	Desenv. Tecn. e Industrial	Apoio Técnico	Total
1994	228	118	128	404	1.255	714	42.002
1995	293	150	129	570	1.563	1.105	49.909
1996	350	127	183	693	1.985	1.309	49.313
1997	292	58	219	484	2.162	1.403	48.212
1998	258	13	242	230	2.162	1.530	44.476
1999 (Prev.)	**235**	**12**	**221**	**210**	**1.972**	**1.550**	**44.823**

Fonte: http://www.mct.gov.br/estat/secav/bol_p_modalid.htm em 28 de outubro de 1999

Anexo II

Gráfico 2.1 - Distribuição dos Grupos de Pesquisa
por Grande Área do Conhecimento em 1997 (Guimarães et al., 1999)

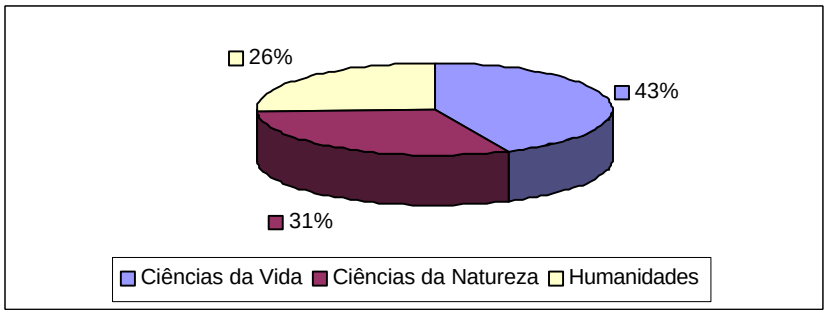


Gráfico 2.2 - Distribuição das Linhas de Pesquisa
por Grande Área do Conhecimento em 1997 (Guimarães et al., 1999)

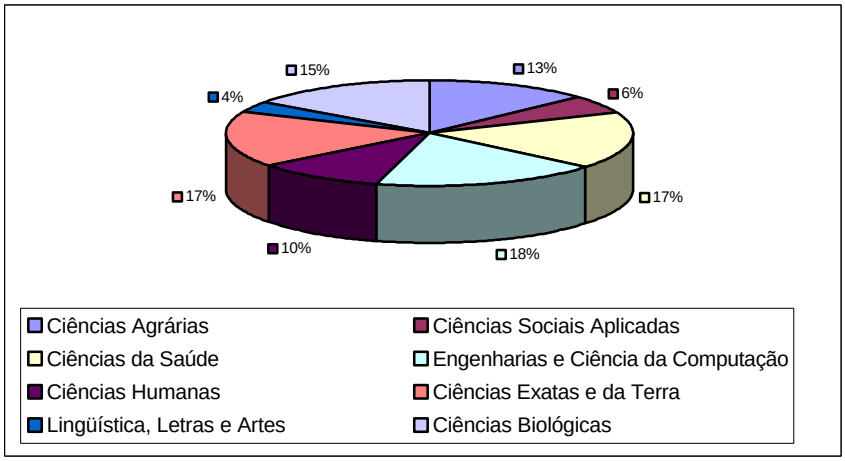


Gráfico 2.3 - Estudantes e Estagiários por Grande Área do Conhecimento
em 1997 (Guimarães et al., 1999)

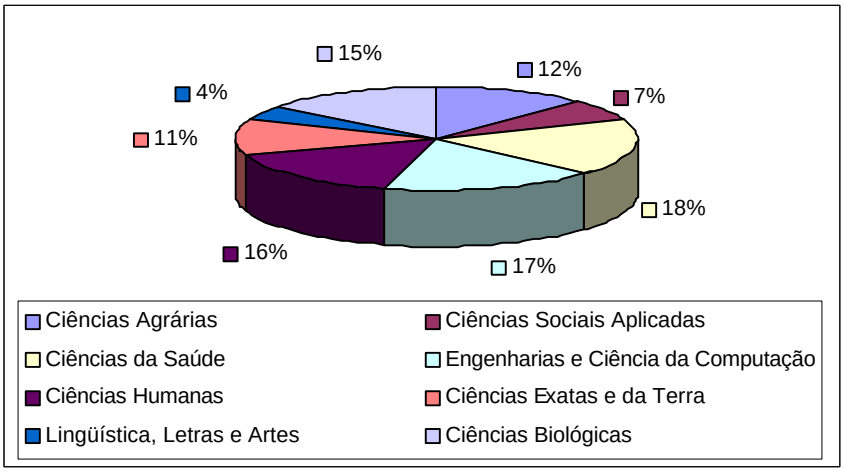


Gráfico 2.4 - Distribuição de Alunos de Mestrado
por Grande Área do Conhecimento em 1997
(Fonte: http://www.capes.gov.br em 17/11/1999)

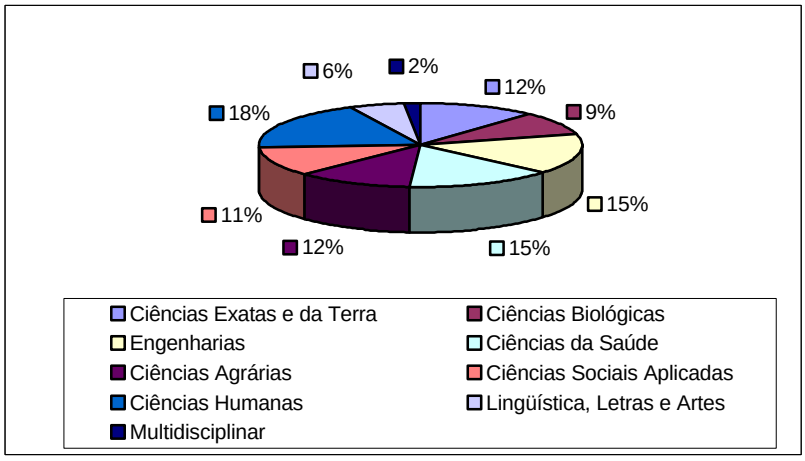


Gráfico 2.5 - Distribuição de Alunos de Doutorado
por Grande Área do Conhecimento em 1997
(Fonte: http://www.capes.gov.br em 17/11/1999)

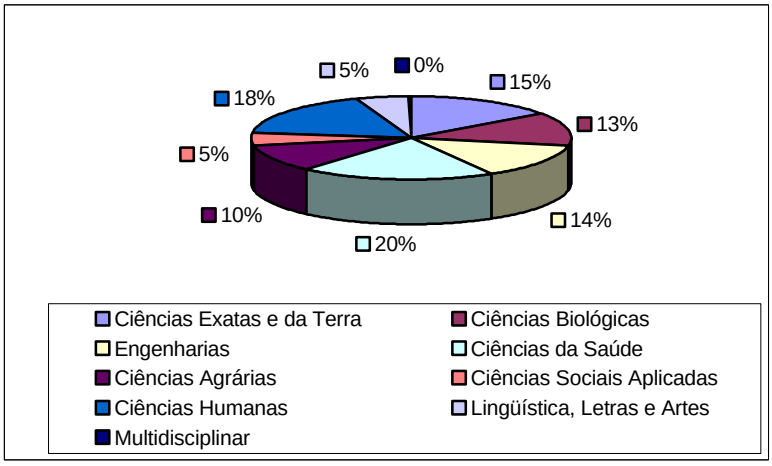


Gráfico 2.6 - Distribuição de Produção C&T
por Grande Área do Conhecimento em 1997 (Guimarães, et al., 1999)

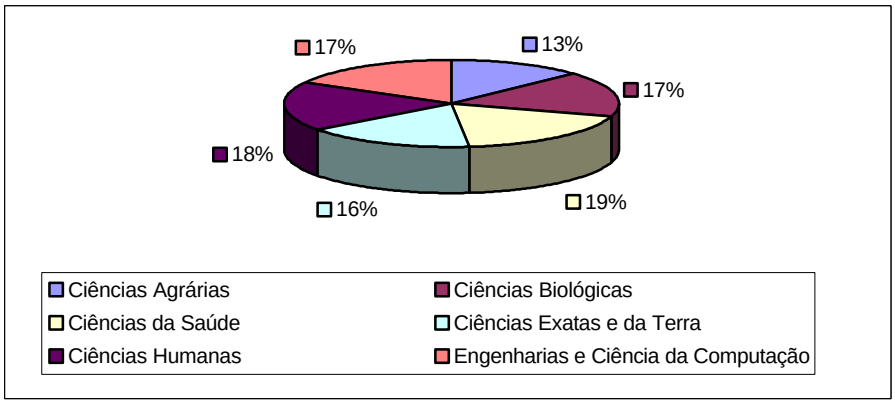


Gráfico 2.7 - Distribuição de Produção C&T (Artigos em Revistas Científicas, Capítulos de Livros e Trabalhos Completos em Anais no País e Exterior) por Grande Área do Conhecimento em 1997
(Fonte: http://www.capes.gov.br em 17/11/1999)

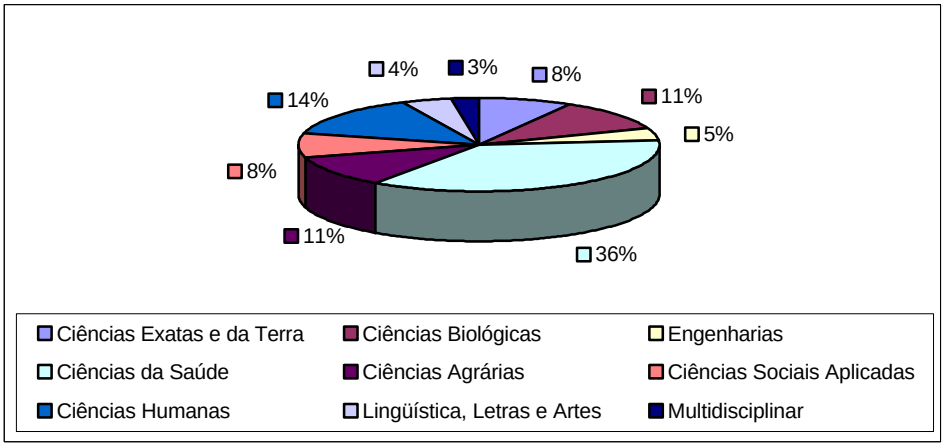


Gráfico 2.8 - Distribuição dos Grupos de Pesquisa por Regiões do País em 1997 (Guimarães et al., 1999)

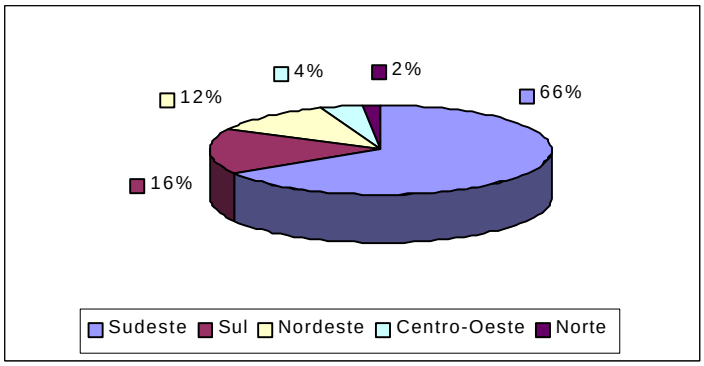


Gráfico 2.9 - Distribuição dos Pesquisadores por Regiões do País em 1997 (Guimarães et al., 1999)

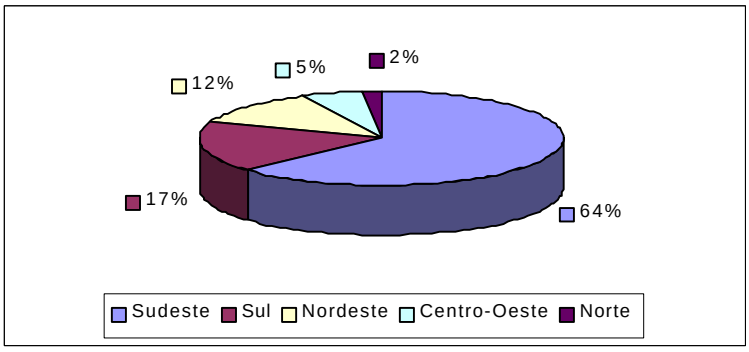


Gráfico 2.10 - Distribuição dos Cursos de Pós-graduação
por Regiões do País em 1997
(Fonte: http://www.capes.gov.br em 17/11/1999)

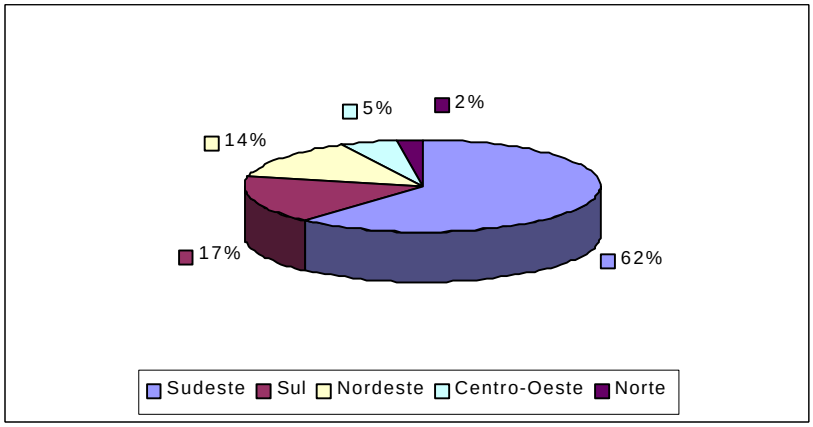
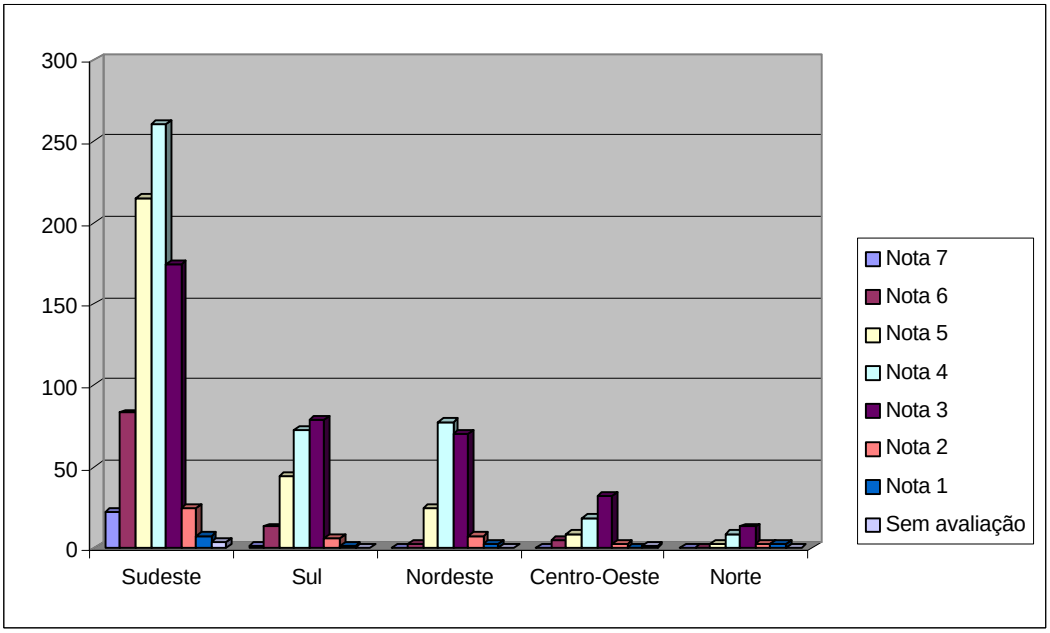


Gráfico 2.11 - Distribuição dos Cursos de Pós-graduação
por Regiões do País, segundo a Avaliação em 1998
(Fonte: http://www.capes.gov.br em 17/11/1999)



Anexo III

Nesta seção serão relacionados todos os centros dos agrupamentos (C_i) e todos os desvios-padrão (D_i) de cada agrupamento para as seis Grandes Áreas (Ciências Agrárias, Ciências Biológicas, Ciências Exatas e da Terra, Engenharias e Ciências da Computação, Humanidades e Ciências da Saúde) nos três conjuntos de dados analisados.

1. Primeiro conjunto de dados

Legenda:	Ane	Artigo publicado em periódico especializado, nacional e estrangeiro
	Ter	Trabalho em eventos completos e resumos
	Lc	Livros e capítulos de livros publicados
	Ptc	Produção tecnológica
	Td	Teses e dissertações orientadas
	Grd	Graduação
	Esp	Especialização
	Mdo	Mestrado
	Ddo	Doutorado

Tabela 4a - Centros e desvios-padrão dos agrupamentos nas Ciências Agrárias

	Ane	Ter	Lc	Ptc	Td	Grd	Esp	Mdo	Ddo
C1	0,1707	0,6344	0,0386	0,0243	0,1321	0,0002	0,0006	0,0313	0,9679
D1	0,1078	0,1360	0,0701	0,0598	0,0958	0,0288	0,0137	0,0657	0,0734
C2	0,4573	0,3222	0,0562	0,0331	0,1311	0,0011	0,0003	0,0548	0,9438
D2	0,1915	0,1485	0,1153	0,0778	0,1146	0,0114	0,0194	0,1019	0,1029
C3	0,1738	0,6670	0,0429	0,0378	0,0785	0,4984	0,0264	0,1200	0,3552
D3	0,1139	0,1617	0,0728	0,0696	0,0852	0,1516	0,0792	0,1303	0,1824
C4	0,2138	0,6457	0,0393	0,0283	0,0729	0,0324	0,0279	0,3827	0,5570
D4	0,1517	0,2133	0,1142	0,0721	0,1021	0,0549	0,0945	0,1238	0,1252
C5	0,1392	0,7858	0,0330	0,0260	0,0161	0,0302	0,0496	0,7630	0,1572
D5	0,2189	0,2533	0,1053	0,0677	0,0728	0,0655	0,1059	0,1685	0,1330

Tabela 4b - Centros e desvios-padrão dos agrupamentos nas Ciências Biológicas

	Ane	Ter	Lc	Ptc	Td	Grd	Esp	Mdo	Ddo
C1	0,1693	0,7461	0,0403	0,0093	0,0350	0,0000	0,0044	0,0131	0,9826
D1	0,1096	0,1358	0,0452	0,0449	0,0744	0,0256	0,0376	0,0618	0,0775
C2	0,5433	0,2216	0,0897	0,0116	0,1338	0,0186	0,0060	0,0071	0,9682
D2	0,2145	0,1853	0,1137	0,0451	0,1684	0,0468	0,0451	0,0652	0,0932
C3	0,3641	0,4549	0,0633	0,0237	0,0940	0,0093	0,0320	0,3953	0,5634
D3	0,1806	0,1955	0,0854	0,0748	0,0957	0,0514	0,1128	0,1374	0,1569
C4	0,1145	0,8177	0,0295	0,0065	0,0318	0,0217	0,1034	0,5844	0,2906
D4	0,1005	0,1486	0,0941	0,0495	0,0690	0,0653	0,1250	0,1990	0,1982
C5	0,1901	0,6361	0,0384	0,0687	0,0667	0,4515	0,0890	0,1540	0,3055
D5	0,1229	0,1998	0,0826	0,1544	0,0876	0,2113	0,2239	0,1357	0,1674

Tabela 4b - Centros e desvios-padrão dos agrupamentos nas Ciências Exatas e da Terra

	Ane	Ter	Lc	Ptc	Td	Grd	Esp	Mdo	Ddo
C1	0,1691	0,7329	0,0319	0,0482	0,0180	0,0089	0,0666	0,6615	0,2631
D1	0,1756	0,2291	0,0707	0,1529	0,0822	0,0520	0,1243	0,1752	0,1958
C2	0,1781	0,6748	0,0217	0,0821	0,0426	0,2985	0,1142	0,1724	0,4149
D2	0,1499	0,1917	0,0429	0,1386	0,0634	0,2231	0,1783	0,1662	0,1716
C3	0,4191	0,4645	0,0300	0,0202	0,0662	0,0000	0,0000	0,0397	0,9602
D3	0,1299	0,1305	0,0624	0,0888	0,1527	0,0168	0,0111	0,0926	0,0959
C4	0,1619	0,7570	0,0113	0,0138	0,0559	0,0001	0,0096	0,0685	0,9219
D4	0,0949	0,1117	0,0444	0,0528	0,0824	0,0252	0,0218	0,1109	0,1147
C5	0,7710	0,0947	0,0150	0,0057	0,1136	0,0014	0,0071	0,0460	0,9455
D5	0,1700	0,0936	0,0634	0,0356	0,1512	0,0344	0,0213	0,1000	0,1174

Tabela 4d - Centros e desvios-padrão dos agrupamentos nas Engenharias e C. da Computação

	Ane	Ter	Lc	Ptc	Td	Grd	Esp	Mdo	Ddo
C1	0,1126	0,5996	0,0303	0,1677	0,0655	0,4042	0,1323	0,2139	0,2496
D1	0,1228	0,2593	0,0896	0,2542	0,0975	0,2375	0,2046	0,1602	0,1718
C2	0,0590	0,7042	0,0447	0,1434	0,0487	0,0311	0,0639	0,7648	0,1403
D2	0,1323	0,2644	0,0869	0,2106	0,0894	0,0674	0,1281	0,1694	0,1412
C3	0,0967	0,7035	0,0185	0,0641	0,1172	0,0107	0,0139	0,4114	0,5640
D3	0,0996	0,1779	0,0587	0,1057	0,1255	0,0553	0,0627	0,1098	0,1164
C4	0,1323	0,6833	0,0147	0,0443	0,1254	0,0036	0,0051	0,0290	0,9623
D4	0,0980	0,1345	0,0476	0,0759	0,1158	0,0328	0,0243	0,0568	0,0708
C5	0,3029	0,2718	0,0651	0,2241	0,1360	0,0011	0,0034	0,0473	0,9483
D5	0,2416	0,1622	0,1370	0,2668	0,1933	0,0328	0,0519	0,1241	0,1331

Tabela 4e - Centros e desvios-padrão dos agrupamentos nas Humanidades

	Ane	Ter	Lc	Ptc	Td	Grd	Esp	Mdo	Ddo
C1	0,2626	0,1519	0,3351	0,0163	0,2022	0,0034	0,0002	0,0318	0,9647
D1	0,2215	0,1413	0,2306	0,0903	0,2393	0,0721	0,0258	0,0895	0,1164
C2	0,0693	0,7029	0,1117	0,0376	0,0785	0,0011	0,0009	0,0598	0,9382
D2	0,1112	0,1701	0,1128	0,0789	0,1172	0,0446	0,0332	0,1307	0,1449
C3	0,0883	0,6027	0,1378	0,0385	0,0283	0,3281	0,3220	0,1300	0,2198
D3	0,1546	0,2781	0,1662	0,1251	0,0991	0,2857	0,2678	0,1432	0,1676
C4	0,1058	0,8015	0,0670	0,0102	0,0154	0,0208	0,1238	0,6989	0,1565
D4	0,1176	0,1625	0,1049	0,0800	0,0655	0,0853	0,1144	0,2046	0,1884
C5	0,2560	0,1558	0,3238	0,0598	0,1371	0,0521	0,0478	0,5884	0,3117
D5	0,2322	0,1653	0,2283	0,1503	0,2079	0,1279	0,1009	0,1941	0,1911

Tabela 4f - Centros e desvios-padrão dos agrupamentos nas Ciências da Saúde

	Ane	Ter	Lc	Ptc	Td	Grd	Esp	Mdo	Ddo
C1	0,2100	0,6137	0,0786	0,0310	0,0382	0,0321	0,4800	0,2570	0,2309
D1	0,1550	0,2100	0,1282	0,0622	0,0788	0,1620	0,2258	0,1623	0,1581
C2	0,1332	0,7334	0,0590	0,0263	0,0482	0,0032	0,0419	0,7117	0,2432
D2	0,1231	0,1650	0,1046	0,0612	0,0828	0,0967	0,0774	0,1909	0,1893
C3	0,5387	0,2243	0,1329	0,0338	0,0703	0,0737	0,1316	0,3507	0,4440
D3	0,2266	0,1756	0,1672	0,1469	0,1345	0,1763	0,1571	0,1983	0,1506
C4	0,5088	0,2173	0,1563	0,0222	0,0573	0,0000	0,0276	0,0159	0,9565
D4	0,2458	0,1673	0,1649	0,0717	0,1578	0,0240	0,0530	0,0633	0,0846
C5	0,1595	0,6950	0,0878	0,0238	0,0339	0,0297	0,0163	0,0473	0,9067
D5	0,1187	0,1524	0,0926	0,0485	0,0960	0,0880	0,0500	0,1121	0,1486

2. Segundo conjunto de dados

Legenda:

Sx	Sexo do pesquisador
Tit	Titulação máxima do pesquisador
Id	Idade do pesquisador
Nac	Nacionalidade do pesquisador
Ded	Tipo de dedicação do pesquisador
Ane	Artigo publicado em periódico especializado, nacional e estrangeiro
Ter	Trabalho em eventos completos e resumos
Lc	Livros e capítulos de livros publicados
Ptc	Produção tecnológica
Td	Teses e dissertações orientadas

Tabela 4g - Centros e desvios-padrão dos agrupamentos nas Ciências Agrárias

	Sx	Tit	Id	Nac	Ded	Ane	Ter	Lc	Ptc	Td
C1	1,000	0,864	0,841	0,934	0,000	0,196	0,463	0,038	0,041	0,078
D1	0,000	0,222	0,043	0,208	0,000	0,303	0,390	0,147	0,113	0,170
C2	0,000	0,788	0,822	0,954	0,000	0,178	0,553	0,021	0,019	0,043
D2	0,000	0,296	0,042	0,207	0,000	0,286	0,404	0,077	0,081	0,129
C3	0,000	0,845	0,820	0,981	1,000	0,198	0,599	0,027	0,020	0,035
D3	0,000	0,226	0,041	0,111	0,000	0,276	0,372	0,084	0,102	0,091
C4	1,000	0,915	0,836	0,983	1,000	0,100	0,796	0,023	0,017	0,064
D4	0,000	0,162	0,040	0,130	0,000	0,121	0,163	0,069	0,063	0,091
C5	1,000	0,906	0,839	0,969	1,000	0,374	0,207	0,058	0,033	0,097
D5	0,000	0,185	0,043	0,169	0,000	0,335	0,205	0,152	0,156	0,205

Tabela 4h - Centros e desvios-padrão dos agrupamentos nas Ciências Biológicas

	Sx	Tit	Id	Nac	Ded	Ane	Ter	Lc	Ptc	Td
C1	1,0000	0,8783	0,8290	0,9483	1,0000	0,1547	0,7750	0,0189	0,0116	0,0399
D1	0,0000	0,2212	0,0481	0,2167	0,0000	0,1342	0,1662	0,0571	0,0603	0,0745
C2	1,0000	0,8787	0,8324	0,9244	1,0000	0,4576	0,1434	0,0516	0,0266	0,0751
D2	0,0000	0,2389	0,0535	0,2531	0,0000	0,3655	0,1849	0,1490	0,1098	0,1814
C3	1,0000	0,8366	0,8354	0,8745	0,0000	0,2363	0,4659	0,0376	0,0177	0,0574
D3	0,0000	0,2541	0,0543	0,3231	0,0000	0,3256	0,4011	0,1481	0,1076	0,1633
C4	0,0000	0,7916	0,8200	0,9447	0,0000	0,1928	0,5398	0,0376	0,0172	0,0302
D4	0,0000	0,2919	0,0457	0,2066	0,0000	0,3000	0,4139	0,1379	0,1449	0,1311
C5	0,0000	0,8427	0,8166	0,9773	1,0000	0,2650	0,5587	0,0232	0,0143	0,0331
D5	0,0000	0,2508	0,0471	0,1343	0,0000	0,2841	0,3628	0,0838	0,0697	0,1145

Tabela 4i - Centros e desvios-padrão dos agrupamentos nas Ciências Exatas e da Terra

	Sx	Tit	Id	Nac	Ded	Ane	Ter	Lc	Ptc	Td
C1	1,0000	0,9622	0,8346	0,9000	1,0000	0,5746	0,1222	0,0270	0,0291	0,0762
D1	0,0000	0,1671	0,0448	0,3174	0,0000	0,3665	0,1751	0,0973	0,1275	0,1774
C2	1,0000	0,9334	0,8365	0,9044	1,0000	0,1459	0,7616	0,0173	0,0241	0,0511
D2	0,0000	0,1745	0,0425	0,2671	0,0000	0,1392	0,1866	0,0668	0,0961	0,0998
C3	0,0000	0,9048	0,8249	0,9761	1,0000	0,2864	0,4947	0,0228	0,0127	0,0432
D3	0,0000	0,2119	0,0415	0,1804	0,0000	0,3044	0,3719	0,0761	0,0817	0,1157
C4	0,0000	0,8253	0,8290	0,9373	0,0000	0,2154	0,4838	0,0225	0,0158	0,0374
D4	0,0000	0,2579	0,0431	0,2638	0,0000	0,2957	0,4073	0,0997	0,1082	0,1340
C5	1,0000	0,8722	0,8410	0,7959	0,0000	0,2616	0,4545	0,0199	0,0349	0,0326
D5	0,0000	0,2451	0,0470	0,3743	0,0000	0,3259	0,3843	0,1085	0,1358	0,1470

Tabela 4j - Centros e desvios-padrão dos agrupamentos nas Engenharias e C. da Computação

	Sx	Tit	Id	Nac	Ded	Ane	Ter	Lc	Ptc	Td
C1	1,0000	0,8343	0,8229	0,9602	1,0000	0,1599	0,1501	0,0524	0,1477	0,1133
D1	0,0000	0,2846	0,0441	0,2243	0,0000	0,2638	0,1842	0,1395	0,3032	0,2263
C2	1,0000	0,9405	0,8206	0,9470	1,0000	0,0780	0,7924	0,0182	0,0399	0,0715
D2	0,0000	0,1564	0,0382	0,2046	0,0000	0,1163	0,1774	0,0606	0,0954	0,1164
C3	0,0000	0,8405	0,8107	0,9581	1,0000	0,1271	0,5795	0,0244	0,0634	0,0542
D3	0,0000	0,2379	0,0380	0,1857	0,0000	0,2256	0,3709	0,1034	0,1905	0,1172
C4	0,0000	0,7999	0,8084	0,9709	0,0000	0,0851	0,5600	0,0183	0,0866	0,0432
D4	0,0000	0,2712	0,0418	0,1632	0,0000	0,1994	0,4130	0,1129	0,2333	0,1512
C5	1,0000	0,8611	0,8301	0,9560	0,0000	0,0861	0,4854	0,0306	0,0816	0,1047
D5	0,0000	0,2658	0,0445	0,2137	0,0000	0,2020	0,3929	0,1015	0,2162	0,1962

Tabela 4l - Centros e desvios-padrão dos agrupamentos nas Humanidades

	Sx	Tit	Id	Nac	Ded	Ane	Ter	Lc	Ptc	Td
C1	0,0000	0,7092	0,8324	0,9487	0,0000	0,2147	0,1001	0,1380	0,0325	0,0791
D1	0,0000	0,3273	0,0532	0,1988	0,0000	0,2899	0,1588	0,3051	0,1464	0,1969
C2	0,0000	0,6991	0,8270	0,9711	0,0000	0,0616	0,8632	0,0480	0,0112	0,0160
D2	0,0000	0,3019	0,0499	0,1599	0,0000	0,0994	0,1778	0,1247	0,0703	0,0676
C3	0,0000	0,7953	0,8388	0,9824	1,0000	0,1700	0,4307	0,1332	0,0220	0,0783
D3	0,0000	0,2905	0,0492	0,1553	0,0000	0,2381	0,3810	0,2423	0,1160	0,1818
C4	1,0000	0,8426	0,8391	0,9579	1,0000	0,1769	0,3751	0,1591	0,0381	0,1132
D4	0,0000	0,2755	0,0498	0,2154	0,0000	0,2755	0,3651	0,2462	0,1253	0,2129
C5	1,0000	0,7887	0,8361	0,9293	0,0000	0,1706	0,3259	0,1550	0,0295	0,0753
D5	0,0000	0,2968	0,0506	0,2451	0,0000	0,2625	0,3852	0,2857	0,1451	0,1974

Tabela 4m - Centros e desvios-padrão dos agrupamentos nas Ciências da Saúde

	Sx	Tit	Id	Nac	Ded	Ane	Ter	Lc	Ptc	Td
C1	1,000	0,888	0,845	0,980	1,000	0,267	0,501	0,077	0,013	0,058
D1	0,000	0,224	0,044	0,148	0,000	0,304	0,363	0,157	0,081	0,136
C2	1,000	0,818	0,840	0,981	0,000	0,215	0,459	0,061	0,011	0,062
D2	0,000	0,284	0,049	0,172	0,000	0,303	0,393	0,203	0,070	0,174
C3	0,000	0,714	0,821	0,992	0,000	0,285	0,084	0,080	0,019	0,046
D3	0,000	0,311	0,047	0,084	0,000	0,380	0,164	0,223	0,117	0,172
C4	0,000	0,765	0,825	0,973	0,000	0,111	0,820	0,040	0,011	0,018
D4	0,000	0,264	0,041	0,165	0,000	0,133	0,176	0,101	0,044	0,063
C5	0,000	0,826	0,828	0,987	1,000	0,255	0,540	0,050	0,014	0,038
D5	0,000	0,252	0,040	0,125	0,000	0,291	0,379	0,149	0,102	0,111

3. Terceiro conjunto de dados

Legenda:
- 1A Bolsista categoria 1A
- 1B Bolsista categoria 1B
- 1C Bolsista categoria 1C
- 2A Bolsista categoria 2A
- 2B Bolsista categoria 2B
- 2C Bolsista categoria 2C
- D7 Docente vinculado a programa com grau 7
- D6 Docente vinculado a programa com grau 6
- D5 Docente vinculado a programa com grau 5
- D4 Docente vinculado a programa com grau 4
- D3 Docente vinculado a programa com grau 3

Tabela 4n - Centros e desvios-padrão dos agrupamentos nas Ciências Agrárias

	1A	1B	1C	2A	2B	2C	D7	D6	D5	D4	D3
C1	0,0346	0,0165	0,0208	0,0254	0,2420	0,2945	0,0000	0,0041	0,0594	0,0292	0,9073
D1	0,0783	0,1220	0,1394	0,1326	0,3254	0,4245	0,0000	0,0783	0,1223	0,1079	0,1687
C2	0,0077	0,0455	0,1002	0,1198	0,0452	0,1943	0,0001	0,0186	0,0093	0,9717	0,0003
D2	0,1343	0,2096	0,2155	0,3034	0,3216	0,3708	0,0216	0,0924	0,0950	0,1349	0,0458
C3	0,0678	0,0187	0,0218	0,0276	0,0228	0,0501	0,0887	0,0253	0,0116	0,0081	0,0021
D3	0,2443	0,1014	0,1248	0,2052	0,1164	0,2759	0,3844	0,1101	0,1327	0,0914	0,0697
C4	0,1299	0,1209	0,2310	0,2045	0,1012	0,1322	0,0918	0,8156	0,0309	0,0360	0,0257
D4	0,2195	0,2436	0,3332	0,2546	0,2345	0,2759	0,1309	0,2084	0,1248	0,1214	0,0754
C5	0,0488	0,1101	0,0537	0,0362	0,2114	0,3350	0,0257	0,0376	0,8815	0,0380	0,0173
D5	0,1990	0,2384	0,2446	0,2606	0,3234	0,3566	0,0394	0,0905	0,1540	0,1078	0,0492

Tabela 4o - Centros e desvios-padrão dos agrupamentos nas Ciências Biológicas

	1A	1B	1C	2A	2B	2C	D7	D6	D5	D4	D3
C1	0,1497	0,1483	0,1396	0,0957	0,0997	0,3036	0,9120	0,0173	0,0617	0,0036	0,0054
D1	0,3361	0,2771	0,2987	0,2520	0,2952	0,3386	0,1400	0,0818	0,1096	0,0292	0,0292
C2	0,1716	0,0786	0,1169	0,0869	0,1309	0,2188	0,0209	0,8760	0,0472	0,0555	0,0004
D2	0,3183	0,2153	0,2847	0,3365	0,2457	0,3660	0,0759	0,2061	0,1463	0,1010	0,0455
C3	0,0286	0,0001	0,0148	0,0013	0,0318	0,0961	0,0109	0,0000	0,0236	0,0081	0,4888
D3	0,1530	0,1544	0,1778	0,1547	0,1818	0,3431	0,0490	0,0595	0,1255	0,1148	0,4454
C4	0,0340	0,0137	0,0753	0,0463	0,0785	0,3197	0,0002	0,0025	0,0587	0,9374	0,0012
D4	0,2288	0,1993	0,2243	0,1770	0,3384	0,4125	0,0373	0,0707	0,1484	0,1809	0,0683
C5	0,1531	0,0399	0,0657	0,1506	0,1910	0,1529	0,0079	0,0027	0,9842	0,0026	0,0025
D5	0,2892	0,2016	0,2123	0,2785	0,3362	0,3674	0,0639	0,0497	0,1542	0,1213	0,0544

Tabela 4p - Centros e desvios-padrão dos agrupamentos nas Ciências Exatas e da Terra

	1A	1B	1C	2A	2B	2C	D7	D6	D5	D4	D3
C1	0,0914	0,0057	0,0867	0,2610	0,0866	0,1866	0,0000	0,0318	0,0183	0,9099	0,0400
D1	0,1655	0,1638	0,2165	0,2845	0,3096	0,3730	0,0204	0,0830	0,0569	0,1342	0,0779
C2	0,0480	0,0813	0,1341	0,1029	0,1271	0,2253	0,0019	0,9860	0,0003	0,0075	0,0044
D2	0,2428	0,1929	0,2541	0,2884	0,3009	0,2956	0,0200	0,0925	0,0402	0,0752	0,0317
C3	0,1016	0,0754	0,1146	0,1723	0,1632	0,2352	0,9700	0,0045	0,0025	0,0166	0,0063
D3	0,2426	0,2198	0,2098	0,2526	0,2473	0,2872	0,1073	0,0222	0,0255	0,0739	0,0214
C4	0,1046	0,0687	0,1082	0,1247	0,2228	0,2204	0,0148	0,0474	0,8934	0,0335	0,0108
D4	0,2472	0,2337	0,2638	0,2968	0,2897	0,3638	0,0443	0,0592	0,1299	0,0991	0,0394
C5	0,0096	0,0002	0,0694	0,0193	0,0063	0,1650	0,0000	0,0222	0,0018	0,0064	0,3565
D5	0,1718	0,0370	0,1354	0,1794	0,2530	0,3841	0,0351	0,0764	0,0807	0,1173	0,4402

Tabela 4q - Centros e desvios-padrão dos agrupamentos nas Engenharias e C. da Computação

	1A	1B	1C	2A	2B	2C	D7	D6	D5	D4	D3
C1	0,0815	0,0760	0,1140	0,1411	0,1577	0,1993	0,0006	0,9645	0,0071	0,0119	0,0159
D1	0,2242	0,2146	0,2612	0,2896	0,2945	0,3348	0,0109	0,1083	0,0567	0,0619	0,0706
C2	0,0053	0,0046	0,0083	0,0095	0,0332	0,1028	0,0054	0,0143	0,0251	0,0253	0,0181
D2	0,0443	0,0424	0,0793	0,0812	0,1600	0,2965	0,0750	0,0702	0,1016	0,1018	0,0825
C3	0,0444	0,0108	0,0356	0,0620	0,1264	0,2849	0,0000	0,0101	0,0320	0,0294	0,9285
D3	0,1858	0,0719	0,1544	0,2125	0,2869	0,4141	0,0000	0,0594	0,1026	0,0905	0,1495
C4	0,0674	0,0575	0,1106	0,0976	0,1307	0,2539	0,0023	0,0171	0,9565	0,0124	0,0117
D4	0,1993	0,1842	0,2627	0,2447	0,2727	0,3507	0,0238	0,0690	0,1193	0,0700	0,0561
C5	0,0303	0,0325	0,0694	0,0961	0,1338	0,2283	0,0000	0,0108	0,0095	0,9658	0,0140
D5	0,1510	0,1517	0,2374	0,2516	0,3056	0,3807	0,0000	0,0559	0,0503	0,1003	0,0668

Tabela 4r - Centros e desvios-padrão dos agrupamentos nas Humanidades

	1A	1B	1C	2A	2B	2C	D7	D6	D5	D4	D3
C1	0,0541	0,0660	0,0756	0,1166	0,1241	0,2126	0,0046	0,0163	0,9193	0,0452	0,0147
D1	0,2083	0,2075	0,2266	0,2892	0,2941	0,3434	0,0493	0,0551	0,1634	0,1228	0,0761
C2	0,1721	0,0804	0,0971	0,1261	0,0756	0,1439	0,0037	0,9603	0,0140	0,0147	0,0074
D2	0,3401	0,2451	0,2817	0,2961	0,2382	0,3151	0,0407	0,1311	0,0749	0,0790	0,0520
C3	0,0218	0,0132	0,0131	0,0181	0,0168	0,0454	0,0639	0,0058	0,0065	0,0086	0,0027
D3	0,1669	0,1303	0,1448	0,1536	0,1314	0,2192	0,3155	0,0725	0,0799	0,0625	0,0347
C4	0,0177	0,0158	0,0407	0,0418	0,0764	0,1359	0,0000	0,0022	0,0362	0,0675	0,8941
D4	0,1310	0,1222	0,1788	0,1887	0,2644	0,3165	0,0000	0,0273	0,1024	0,1527	0,1787
C5	0,0392	0,0370	0,0472	0,0589	0,0964	0,1245	0,0017	0,0080	0,0424	0,9415	0,0064
D5	0,1941	0,1717	0,1930	0,2405	0,2737	0,3138	0,0463	0,0592	0,1174	0,1427	0,0421

Tabela 4s - Centros e desvios-padrão dos agrupamentos nas Ciências da Saúde

	1A	1B	1C	2A	2B	2C	D7	D6	D5	D4	D3
C1	0,0349	0,0313	0,0201	0,0435	0,0639	0,0836	0,0093	0,0037	0,0363	0,0416	0,9090
D1	0,1646	0,1653	0,1342	0,2083	0,2286	0,2714	0,0573	0,0425	0,1104	0,1258	0,1751
C2	0,0330	0,0141	0,0098	0,0292	0,0200	0,0093	0,0167	0,0963	0,0146	0,0125	0,0064
D2	0,1966	0,1268	0,0761	0,1791	0,1497	0,0598	0,1336	0,3225	0,0747	0,0793	0,0525
C3	0,1020	0,0600	0,0814	0,1098	0,1099	0,0564	0,0087	0,0170	0,8893	0,0602	0,0248
D3	0,2550	0,2178	0,2328	0,2756	0,2547	0,1492	0,0537	0,0850	0,1844	0,1444	0,0913
C4	0,0053	0,0000	0,0000	0,0169	0,0274	0,9503	0,0066	0,0467	0,2893	0,4341	0,0753
D4	0,0429	0,0000	0,0000	0,0796	0,1063	0,1348	0,0913	0,1718	0,3844	0,4177	0,1872
C5	0,0365	0,0341	0,0618	0,0891	0,0736	0,0179	0,0022	0,0032	0,0243	0,9564	0,0139
D5	0,1847	0,1468	0,2114	0,2597	0,2476	0,0932	0,0290	0,0377	0,0776	0,1132	0,0674