UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIA

PAULO VINÍCIUS MARCONDES CORDEIRO

AVALIAÇÃO DE CONHECIMENTOS EXPLICITADOS EM PATENTES PARA LEVANTAMENTO DE INDÍCIOS DE POSSÍVEIS PARCERIAS EMPRESARIAIS

DISSERTAÇÃO

CURITIBA
2010

PAULO VINÍCIUS MARCONDES CORDEIRO

AVALIAÇÃO DE CONHECIMENTOS EXPLICITADOS EM PATENTES PARA LEVANTAMENTO DE INDÍCIOS DE POSSÍVEIS PARCERIAS EMPRESARIAIS

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Tecnologia. Programa de Pós-Graduação em Tecnologia, Universidade Tecnológica Federal do Paraná. Área de concentração: Tecnologia e Desenvolvimento.

Orientador: Prof. Dr. Dario Eduardo Amaral Dergint

CURITIBA

2010

Dedico este trabalho aos meus pais, Mário e Ros Méri, que me ensinaram que conhecimento é o maior patrimônio que podemos ter. E à minha esposa, Manuella, pelo suporte incondicional na realização de todos os meus sonhos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, fonte da minha esperança. Sei que este mestrado estava previamente em seus sonhos para mim, então me deu a disposição e o ferramental necessário para completar mais esta etapa.

Agradeço à minha esposa Manuella, pelo suporte durante todo o mestrado, primeiramente como namorada e depois como esposa. Sem seu suporte, o peso da jornada poderia ter me derrubado. Você me inspira!

Agradeço aos meus pais Mário e Ros Méri, pelo suporte moral e financeiro. O começo desta meta não seria possível sem sua ajuda.

Obrigado ao professor Dario Dergint pela orientação e por acreditar neste trabalho.

“O homem sábio é poderoso, e quem tem conhecimento aumenta a sua força;... Quem pensa conhecer alguma coisa, ainda não conhece como deveria.”

(Provérbios 24:5; 1Coríntios 8:2)

RESUMO

CORDEIRO, Paulo. Avaliação de conhecimentos explicitados em patentes para levantamento de indícios de possíveis parcerias empresariais. 2010. 136 f. Dissertação (Mestrado em Tecnologia) – Programa de Pós-Graduação em Tecnologia, Universidade Tecnológica Federal do Paraná. Curitiba, 2010.

O modelo de Inovação Aberta é uma boa opção para empresas que não podem arcar com os custos de pesquisa e desenvolvimento (P&D), porém necessitam continuar inovando para permanecer no mercado. Esse modelo possibilita que empresas espalhadas em todo o mundo e em todas as áreas de pesquisa sejam possíveis parceiras em P&D, entretanto as possíveis parcerias não podem ser limitadas ao *know-who* dos gestores de tecnologia. Os documentos de patente podem ser uma rica fonte de informações sobre desenvolvimento tecnológico e inovações de um grande número de empresas, mas a leitura de todos esses documentos, que são criados todos os dias, é uma tarefa enfadonha que os gestores de tecnologia não podem arcar. Esta pesquisa objetiva criar um sistema informatizado de Mineração de Tecnologia que busque por informações sobre possíveis parcerias em P&D para uma empresa geradora de tecnologia. Esse sistema utiliza os registros de patentes como entrada e suas saídas são duas, a saber: (1) regras de interdependência entre tecnologias e (2) agrupamentos de patentes referentes ao mesmo assunto, dentro do quadro de patentes adquiridos. Essas saídas podem ser utilizadas para a busca de complementaridade empresarial. As regras de Associação encontradas pelo sistema e os diferentes grupos os assuntos abordados pelos diferentes conjuntos de patentes apontam a possibilidade de que tecnologias interdependentes podem ser utilizadas para a busca de parcerias de P&D. Os grupos gerados pelo algoritmo de Agrupamento podem aumentar a probabilidade de encontrar invenções de interesse à empresa que busca por parcerias.

Palavras-chave: Mineração de Tecnologia. Complementaridade. Inovação Aberta. Patentes. Mineração de Dados. Prospecção Tecnológica.

ABSTRACT

CORDEIRO, Paulo. Evaluation of Explicit Knowledge in Patents for Survey of Evidence of Possible Business Partnership. 2010. 136 p. Dissertation (Master degree in Technology) – Programa de Pós-Graduação em Tecnologia, Universidade Tecnológica Federal do Paraná. Curitiba, 2010.

The model of Open Innovation is a good option to firms that can not afford research and development (R&D) costs, but intents to continue playing the innovation game. This model offers, as possible partners in R&D, companies spread worldwide and in all research areas. However, the possible partnership can not be restricted in technology manager's know-who. The patent documents can be source of rich information about technical development and innovation from huge amount of firms, nevertheless reading all these documents are daily created must be an herculean task the technology managers can not afford. This research aims to introduce two automated models to seek for match between firms R&D using Technology Mining techniques applied to patents documents datasets. The methodology for this system set the input as patents and the outputs are two kinds: (1) technology interdependence between technologies; and (2) clusters of patents regarding same subjects. These outputs can be used to seek entrepreneurial complementarity. Rules from Association test display the possibility of two companies could knit partnership, due their embodied technology are interdependent. The clusters from Cluster test show a set of patents that can be, probably, in the concerning of company which seeks for technological complementarity.

Key-words: Tech Mining. Complementarities. Open Innovation. Patents. Data Mining. Technological Forecasting.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BD - Banco de Dados

C&T - Ciência e Tecnologia

CT&I - Ciência Tecnologia e Inovação

Co-dev - Co-Desenvolvimento

EPP - Empresas Possíveis Parceiras

ER - Empresa Referência

HTTP - *Hipertext Transfer Protocol* (Protocolo de Transferência de Hipertexto)

IA - Inteligência Artificial

IPC - *International Patent Classifier* (Classificação Internacional de Patentes)

ISI - *Institute for Scientific Information*

KDD - *Knowledge Discovery in Databases*

MD - Mineração de Dados

MT - Mineração de Tecnologia

NLP - *Natural Language Processing* (Processamento de Linguagem Natural)

P&D - Pesquisa e Desenvolvimento

P&G - Procter & Gamble

PI - Propriedade Intelectual

SEQ - Soma do Erro Quadrático

TFIDF - Freqüência do Termo e Freqüência Inversa do Documento

TRIZ - Teoria da Solução Inventiva de Problemas

SUMÁRIO

1 INTRODUÇÃO

A rapidez da difusão do conhecimento é um processo intrínseco à evolução tecnológica dentro da sociedade humana. O conhecimento, inicialmente difundido de modo tácito pelo homem primitivo, ganhou longevidade e tornou-se fiel com a invenção da escrita. Esta se vulgarizou com a invenção da prensa gráfica de Gutenberg, facilitando a troca de opiniões entre os acadêmicos das mais diversas universidades e expandindo o processo de difusão de informações entre eles e os leigos. Atualmente, a difusão da informação é instantânea. Com o auxílio da Internet, o conhecimento codificado em qualquer local do mundo pode ser trocado imediatamente, facilitando pesquisas conjuntas em âmbito internacional.

A capacidade de geração de conhecimento pode ser considerada como o maior ativo de uma empresa tecnológica dependente de inovação para continuar no mercado (ORGANISATION..., 1996). Isso está ligado à emergente economia do aprendizado (JOHNSON e LUNDVALL, 2000), que é o fator principal do aumento da produtividade, da inovação tecnológica e da difusão de todo tipo de informação (ORGANIZATION..., 1996). O ambiente capitalista atual é baseado na rapidez da geração e aproveitamento do conhecimento como sua maior arma de diferencial competitivo (SCHUMPETER, 1984). Porém, sozinhas somente algumas empresas são capazes de gerar todo o conhecimento necessário para seu próprio sustento. Segundo Chesbrough (2003), os modelos que explicam a geração e compartilhamento de conhecimento mostram que uma estrutura centralizada de geração de tecnologia tende a ser mais lenta e ineficiente, que uma estrutura distribuída, e tem potencial de levar a empresa a sucumbir dentro do mercado competitivo. Empresas pequenas, que conseguem ter mais mobilidade em sua estrutura de desenvolvimento e inovação, têm mais agilidade no desenvolvimento de novas tecnologias, ou seja, em inovações. Assim, essas tomam o mercado e acabam líderes em tecnologias que suplantaram as antigas líderes (CHRISTENSEN, 1997).

Ter noção de todo o conhecimento que é gerado tanto no ambiente interno como no externo à corporação é vital para a continuidade das empresas em que a inovação é o maior diferencial competitivo. Um modo de ter domínio sobre esse conhecimento é com as centenas de publicações de material tecnológico existentes atualmente. O conhecimento, quando codificado em artigos científicos ou patentes, torna-se informação que pode voltar a ser conhecimento quando absorvido por um terceiro. O crescimento acelerado da difusão da informação traz um efeito colateral: o custo de tratamento dessa informação torna-se

dispendioso devido à incapacidade humana de processá-la. Isso mina a percepção de qual material pode ser relevante para os negócios e traduz-se na dificuldade de transformar a informação disponível em conhecimento útil.

Tendo em vista o contexto apresentado, a importância do conhecimento dentro da economia atual e a incapacidade humana de trabalhar com tamanha quantidade de informações, infere-se que o desenvolvimento de uma ferramenta que "amplifique" a capacidade humana de procurar por informações relevantes para o seu negócio dentro da enorme quantidade e variedade de publicações existentes atualmente é estratégica para um novo modo de criação de tecnologia e inovação. O campo de Inteligência Artificial é a área de pesquisa que tem grande potencial de uso para a automação na manipulação de grandes quantidades de dados, o desenvolvimento de algoritmos mais avançados possibilita a ampliação do campo de busca das informações relevantes pelos gestores de tecnologia (CANONGIA *et al.*, 2002).

A economia baseada em conhecimento articula-se em hierarquias de redes de conhecimento que são dirigidas pela aceleração da taxa de mudanças e de aprendizado. Essas redes de conhecimento são importantes para a difusão e o uso da informação. A maior interação entre as diferentes hierarquias de redes de conhecimento pode trazer aplicações interdisciplinares que resultariam em inovação (ORGANIZATION..., 1996). Porém, se duas redes, que poderiam atuar conjuntamente, não se conectam em nenhum ponto, ou então não têm uma forte relação uma com a outra, as descobertas ou invenções geradas por uma demorariam a alcançar a outra. Entretanto, conhecimentos de áreas diferentes podem ser correlacionados automaticamente utilizando técnicas computacionais, acelerando o desenvolvimento técnico. Isso é possível uma vez que indícios dos diferentes conhecimentos são disponíveis em patentes, artigos científicos, produtos, currículos entre outros.

Há conhecimentos gerados em áreas não correlacionadas que poderiam resultar em inovações se articulados conjuntamente. March (1991) chama de *exploitation* (exploração) a combinação e re-combinação do conhecimento previamente estabelecido para suprir necessidades dos consumidores. Por exemplo, um conhecimento aparentemente inútil em telecomunicações pode ser útil em agricultura, conhecimentos em mecânica podem ser utilizados em cirurgias plásticas, conhecimentos em engenharia genética podem ser aplicados à construção civil. As possibilidades de correlações são inúmeras. Exemplo cotidiano disso é o uso de ondas eletromagnéticas, inicialmente utilizadas em telecomunicações, para o aquecimento de alimentos nos fornos de microondas. Outro exemplo são as heurísticas computacionais evolucionárias, que derivam das teorias básicas da evolução biológica.

1.1 DELIMITAÇÃO DO TEMA

Para uma empresa manter-se competitiva, somente ter a noção temporal e espacial sobre a geração do conhecimento não é útil. Saber como utilizar essa informação de maneira estratégica pode trazer o diferencial competitivo necessário para a sobrevivência da empresa inovadora. Dentre as possíveis ações que podem ser tomadas pelo gestor de tecnologia após ter ciência de uma informação estratégica, esta pesquisa foca na formação de parcerias como o caminho para chegar ao mercado com inovações. O modelo de Pesquisa e Desenvolvimento (P&D) em Inovação Aberta (*Open Innovation*) (CHESBROUGH, 2003) fornece a premissa de que nenhuma empresa necessita deter todo o conhecimento necessário, relativo à uma determinada área, para realizar inovações em seus produtos. O conhecimento poderia estar espalhado em rede fora dos muros da corporação.

A busca por parcerias pode se restringir à rede de contatos formada no decorrer dos anos de existência da empresa como também pelo *know-who*[1] do gestor de tecnologia. Porém, como conhecer novas possibilidades de parcerias uma vez que as fronteiras nacionais e regionais não podem mais delimitá-las? O conhecimento está espalhado por todo mundo, e a melhor tecnologia que se pode conseguir pode estar escondida em um grupo de pesquisa que ainda está pouco difundido. Uma maneira é a busca pelas informações livre a todos, como: o *website* oficial das empresas, notícias na mídia, portfólio de produtos, portfólio de patentes, o currículo de seus pesquisadores (no caso do Brasil o currículo Lattes).

As atividades científico-tecnológicas (C&T) são de difícil análise direta. O andamento de pesquisas e desenvolvimento somente pode ser mensurado indiretamente. Os principais indicadores de conhecimento apresentados pela ORGANIZATION...(1996) são: 1) gastos com P&D; 2) contratação de engenheiros e pessoal técnico; 3) patentes; e 4) balanços internacionais de pagamento por tecnologia. Este trabalho irá se ater somente à terceira fonte de indício de geração de conhecimento, as patentes, que podem ser consideradas resultados de atividades de P&D. O documento mais rico em informações confiáveis sobre o desenvolvimento de conhecimento em uma organização são os registros de patente (FORAY, 2007).

---

[1] *Know-who* é definido pela ORGANIZATION... (1996) como o conhecimento que envolve a informação de quem sabe alguma coisa e quem sabe fazer alguma coisa. Ele depende da formação de uma rede social de relacionamentos, que possibilita chegar rapidamente aos especialistas em cada assunto.

Muito embora as patentes sejam disponibilizadas livremente em bases de patentes abertas na Internet, sua manipulação e tratamento tornam-se proibitivos, devido ao grande volume existente. Uma alternativa para a busca de informações relevantes nas patentes pode ser a Inteligência Artificial (IA) que engloba tecnologias como Mineração de Dados, Lógica *Fuzzy*, Computação Evolutiva entre outra. Elas podem ser utilizadas como ferramentas para automatizar o tratamento de um grande número de informações. Como esta pesquisa tem o objetivo de buscar informações relevantes dentro de um grande volume de dados, a área de IA que melhor se adéqua a esse quesito é a Mineração de Dados.

1.2 JUSTIFICATIVAS

A utilização de ferramentas computacionais de IA no auxílio de prospecção tecnológica é uma área importante da gestão da tecnologia, não muito difundida no Brasil. Ferramentas que auxiliam a tomada de decisão para direcionamento de investimentos em tecnologia são utilizadas em todo o mundo por empresas desejosas de inovar. Canongia *et al.* (2002) ressaltam a importância da criação de um arcabouço teórico nacional para atuação de ferramentas de prospecção tecnológica e do conhecimento como ponto de apoio para a tomada de decisão em Ciência Tecnologia e Inovação (CT&I), para fortificar a nação em quatro pontos:

- orientar ações onde a ciência e a tecnologia podem intermediar, nas áreas: econômica, ecológica e social;
- nortear a alocação inteligente de recursos financeiros e evitar a possibilidade de "duplicação de esforços pela ação de atores que atuam em âmbitos distintos";
- promover a colaboração de interesses locais com estratégias nacionais ou regionais; e
- promover a articulação entre os principais programas de prospecção no país.

Yang *et al.* (2008) consideram o controle sobre o conhecimento e a construção de ferramentas para prospecção tecnológica, chamada por eles de ferramentas de mineração de texto e visualização, de tamanha importância para a indústria do presente século quanto as máquinas foram para era industrial no século XIX. Fazendo um paralelo, as máquinas a vapor conseguiram amplificar a força muscular dos trabalhadores da época e mecanismos para o controle do conhecimento podem amplificar a capacidade cognitiva e inventiva dos cientistas e engenheiros atuais, ou então, mostrar, com dados empíricos, indícios para investimentos ou parcerias tecnológicas para os gestores.

Desenvolver uma nova tecnologia pode ser uma atividade onerosa para as empresas. Ela pode se traduzir em investimento de recursos que não traz resultados financeiros, já que pode não se ter domínio de todo o processo de desenvolvimento. O modelo tradicional de desenvolvimento, chamado por Chesbrough (2003) como modelo de Inovação Fechada, tem um alto custo operacional, pois todas as atividades de desenvolvimento, desde a concepção da idéia inovadora até a venda do produto, são realizadas dentro da própria estrutura da empresa. Um novo modelo de inovação, Inovação Aberta, surge como uma opção para diminuir os custos da inovação. Nela, o co-desenvolvimento reduz os custos de P&D no desenvolvimento de um novo serviço ou produto. Entretanto, como perceber quem são as potenciais parcerias e que critérios utilizar para sua escolha? O co-desenvolvimento depende da complementaridade entre as empresas, que é uma característica técnica resultante do conhecimento contido pela empresa. A decisão sobre uma parceria pode ser tomada embasada no *know-who* de alguém. Dentro da estrutura da empresa, quando há permeabilidade entre hierarquias para essa informação alcançar os tomadores de decisão, os detentores do *know-who* podem ser desde operários até diretores; fora da empresa, podem ser consultores da área e até mesmo a Internet. Entretanto, esse método não consegue abranger todas as possibilidades. Mesmo que a quantidade de contatos empresariais de um grupo de pessoas seja grande, eles ainda são pequenos quando comparados a todas as possíveis parcerias que podem existir, e mesmo que os sítios de busca da Internet possam ter acesso a uma quantidade imensa de informação, eles não têm uma metodologia apropriada para a busca deste tipo de informação. Quando não se tem acesso a todas as informações disponíveis, não há como buscar pela melhor opção. Portanto, utilizar métodos empíricos com dados reais e confiáveis sistematizaria o processo, podendo trazer a melhor opção de empresa na tomada de decisão em parceria no modelo de Inovação Aberta. Acredita-se que a Mineração de Dados em bancos de patente pode trazer informações que podem ser relevantes na tomada de decisão nos casos de compra de tecnologia, co-desenvolvimento, criação de *joint-ventures*, busca de potenciais clientes para venda de conhecimento, direcionamento tecnológico, entre outras.

1.3 OBJETIVOS

Dado o quadro acima descrito, o objetivo geral desta pesquisa é: **desenvolver um sistema informatizado de avaliação de dados contidos em bases de patentes para obtenção de indicadores de perspectivas de cooperação para inovações tecnológicas no**

contexto da inovação aberta. Delimitando o escopo estabelecido no objetivo geral os objetivos específicos são:

- Compilar um quadro teórico das principais técnicas de manipulação de dados que se adéqüem à prospecção tecnológica.
- Criar uma metodologia para a busca de patentes de empresas que possam ser parceiras em atividades de P&D e uma metodologia para o tratamento computacional destas patentes.
- Programar três módulos principais do sistema, sendo os mesmos: busca de patentes em sítios de base de patentes; extração das informações dos documentos de patentes por meio de expressões regulares[2]; e pré-processamento das patentes.
- Testar o sistema desenvolvido com um exemplo real. Isto é, a metodologia e o sistema computacional serão utilizados neste teste para encontrar possíveis parcerias nas atividades de P&D para uma empresa ativa no mercado.
- Analisar os resultados obtidos pelos testes para verificação da validade e potencial de utilização das metodologias criadas para esta pesquisa.

1.4 NATUREZA DA PESQUISA

Essa pesquisa é classificada como experimental, de natureza aplicada, abordagem qualitativa, e exploratória, segundo os objetivos.

Sua natureza aplicada é devido à abordagem de um problema específico (SILVA e MENEZES, 2001). Sua abordagem é qualitativa, pois ela não quantifica os resultados, mas almeja interpretar os resultados obtidos de forma subjetiva (COHEN e LEVINTHAL, 1990). Segundo seus objetivos, ela é considerada exploratória, pois fornecerá uma visão geral sobre o assunto abordado e é tratada como uma etapa anterior para outras pesquisas (MOREIRA e CALEFFE, 2006). Outra característica de uma pesquisa exploratória é que ela tem o objetivo de construir hipóteses (SILVA e MENEZES, 2001), o que será observado na apresentação dos resultados.

[2] Expressões Regulares são seqüências de caracteres que descrevem um padrão textual. Assim, ele provê para *softwares* de tratamento de texto uma forma flexível de encontrar palavras, números, grupos de caracteres ou padrões de caracteres em um texto.

Esta pesquisa demanda uma abordagem exploratória, pois seu tema não foi explorado previamente. Até o presente momento, não se tem notícia sobre publicações que discutem a utilização de técnicas de Inteligência Artificial visando complementaridade entre as empresas e o fomento da Inovação Aberta. Entre os trabalhos relacionados ao tema pode-se listar: a análise de redes de co-citação (PORTER e CUNNINGHAN, 2005; YOON e PARK, 2004; STERNITZKE, BARTKOWSKI e SCHRAMM, 2008), análise de semelhanças entre patentes por Agrupamentos (FATTORI, PEDRAZZI e TURRA, 2003;) e até mesmo a Classificação automática dos padrões inventivos da TRIZ (CONG e TONG, 2008). Algumas publicações brasileiras que mais se aproximam ao tema apresentado pela dissertação são: Martins (2008) com a Mineração de Dados não estruturados para prospecção tecnológica, em sua dissertação de mestrado; e o artigo de Azarias, Matos e Scandelari (2009) com um exemplo prático de aplicação de Mineração de Dados para a geração de conhecimento. Dentro do Programa de Pós-Graduação em Tecnologia (PPTGE) na UTFPR o trabalho que mais se relaciona com este é a dissertação de Sturm (2005) que utiliza lógica *fuzzy* para automatizar a indicação de competências em currículos.

1.5 ORGANIZAÇÃO DOS CAPÍTULOS

O Capítulo 2, Revisão Teórica, apresenta os marcos teóricos mais importantes para a realização da pesquisa. Entre eles a inovação (seção 2.1); Inovação Aberta (seção 2.1.3), onde são apresentados os principais conceitos definidos por Chesbrough (2003, 2006); Complementaridade (seção 2.2), que é a teoria matemática que define a relação de complemento entre duas empresas; Patentes (seção 2.3), onde são apresentadas as definições desses documentos, sua utilização como informação sobre inovação e as principais informações que são possíveis de serem encontradas em suas linhas; e o Ferramental necessário para a análise de grandes volumes de dados tecnológicos (seção 2.4).

O Capítulo 3, Metodologia, traz o método utilizado para o tratamento dos dados da pesquisa. Ele é iniciado com a conceituação técnica que apresenta os principais quadros teóricos necessários para o entendimento da montagem da metodologia. Em seguida, são apresentadas as ferramentas utilizadas no desenvolvimento da pesquisa, para somente então introduzir os procedimentos para buscar indicadores de perspectivas de cooperação em inovações tecnológicas.

O Capítulo 4, Desenvolvimento do Sistema Computacional e Testes, apresenta o desenvolvimento do sistema informatizado para a obtenção de indicadores de perspectivas de complementaridade entre empresas. Também discorre sobre os dois testes de exemplo que utilizam o sistema computacional. Estes dois testes apresentam os principais resultados que o sistema pode fornecer.

Por fim, o Capítulo 5, Considerações Finais, apresenta as considerações mais importantes sobre a pesquisa, mostra as possibilidades que a presente pesquisa oferece para pesquisas futuras e as principais limitações da mesma.

2 REVISÃO TEÓRICA

Neste capítulo abordam-se os principais pilares teóricos dessa pesquisa, sendo os mesmos articulados em quatro áreas. Para o processo de inovação focou-se no quadro teórico da Inovação Aberta, visto que um de seus pontos fortes é o estudo do processo de inovação distribuído em redes de empresas. Para o estudo de potencial de cooperação de empresas, adotou-se a teoria matemática da Complementaridade, que será utilizada na automação da medida do nível de complementaridade entre empresas distintas. O terceiro tópico visa dos documentos de patentes, visto que esses são uma fonte natural de dados da área tecnológica e que se encontram padronizados e disponíveis em bases de dados via Internet. E, por último, são apresentadas as técnicas de estudo e aquisição de informação sobre campos de P&D, visando correlacionar os elementos levantados nas outras áreas com o objetivo principal desse estudo: "Avaliação de Conhecimentos Explicitados em Patentes para Levantamento de Indícios de Possíveis Parcerias Empresariais" em áreas tecnológicas.

2.1 INOVAÇÃO

O Manual de Oslo (ORGANIZATION..., 1997) define inovação como:

> "Uma **inovação** é a implementação de um produto (bem ou serviço) novo ou significativamente melhorado, ou um processo, ou um novo método de marketing, ou um novo método organizacional nas práticas de negócios, na organização do local de trabalho ou nas relações externas."

Essa definição apresenta algo como inovador somente quando ele é introduzido no mercado, quando a criatividade gera renda para o criador. Para delimitar o processo de inovação são necessárias algumas definições, pois alguns conceitos como descoberta, criatividade, invenção e inovação podem ser mal utilizados. A criatividade é uma capacidade humana de criar, descoberta é o resultado de uma atividade científica, invenção é o resultado da solução de um problema técnico, e a inovação tem seu foco no desempenho econômico atingido pela invenção (CLEMENTE, 2008).

Schumpeter (1934) expõe a inovação como um processo orgânico em que uma nova tecnologia substitui a tecnologia antiga dinamicamente. As tecnologias podem apresentar dois modos de modificação, um por evolução e outro por revolução. O modo evolucionário é a inovação incremental e o modo revolucionário é a inovação radical. A inovação pode ocorrer em cinco frentes: no produto; no processo produtivo; na descoberta de

um novo nicho de mercado; no desenvolvimento de novas fontes provedoras de matérias-primas; e no desenvolvimento de novas estruturas de mercado em uma indústria. Para Schumpeter (1984), a inovação é fruto da concorrência entre empresas, que pode ser considerada como uma característica típica das empresas capitalistas. A concorrência do livre mercado é um disciplinador, que faz com que as empresas tentem tomar o mercado do concorrente, melhorando ou lançando novos produtos que melhorem a margem de lucro na otimização da linha de produção, ou mesmo busquem novos clientes. Enfim, a inovação é resultado direto da necessidade dos empresários em se diferenciarem dos seus concorrentes. Inova quem realmente necessita de uma vantagem competitiva sobre os concorrentes, pois o esforço demandado do processo de inovação pode enfraquecer um negócio estável (DERGINT, 2008). A inovação quebra momentaneamente a concorrência perfeita de mercado, por meio das patentes, rápidas inserções de produtos em mercados e/ou segredos industriais, dando ao inventor os lucros de uma invenção bem aplicada.

2.1.1 Inovação Tecnológica

Para a definição de Inovação Tecnológica deve-se, primeiramente, definir o que é tecnologia. Segundo algumas interpretações (DERGINT, 2008) o termo tecnologia tem sentido se aplicado a estruturas de aprendizado posteriores à estruturação do método científico de Descartes, no século XVII. Antes disso o que existia era a transmissão de conhecimento predominantemente tácito, que é conhecido como técnica. Esse conhecimento era gerado através da experiência diária, e o seu acúmulo acontecia através das gerações sendo passado de pai para filho, de mestre para aprendiz. A estruturação do método de Descartes marca o nascimento da ciência moderna. Assim, a técnica juntou-se à ciência e criou a "Tecnologia", que do grego "*techne*" significa habilidade e "*logos*" significa conhecimento sistematicamente organizado (DERGINT, 2008).

Rocha (1996) define tecnologia como:

> "Cultura simbólica que combina conhecimentos empíricos e técnico-científicos para a produção de bens e serviços para a sociedade; conhecimento organizado e sistematicamente aplicado à produção de bens e serviços e aos seus processos; técnicas de produção fundamentadas em conhecimentos científicos."

Segundo a definição de Rocha, a tecnologia pode ser derivada tanto de conhecimentos empíricos, como de conhecimento científico-tecnológico. A Figura 1 apresenta esta afirmação. Quando um conhecimento técnico é sistematicamente organizado,

como no método cartesiano, ele pode servir para a produção de bens e serviços. Esse conhecimento é chamado de tecnologia.

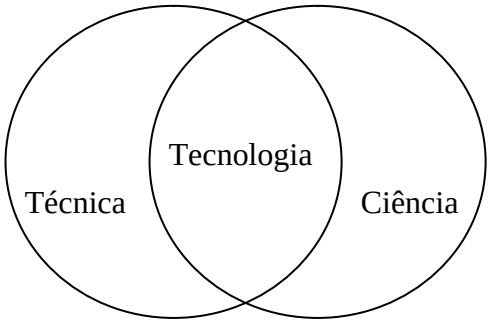


Figura 1- Representação da Tecnologia.
Fonte: Dergint, 2008.

Dado o panorama sobre inovações e tecnologia, pode-se definir a Inovação Tecnológica como desenvolvimentos realizados por meio de pesquisas científicas ou da intuição dos inventores e que alcançam o mercado em novos produtos ou em características inerentes em produtos já existentes, que objetivam diferenciação dos concorrentes no preço ou em características particulares, assim, ganham vantagens no mercado em que estão inseridas, ou inserindo-se em mercados recém abertos (DERGINT, 2008; ORGANIZATION...,1997; SCHUMPETER, 1984).

As inovações tecnológicas podem ser incrementais, com melhorias no desempenho de componentes do sistema tecnológico correntes, ou então radicais, com substituições radicais na arquitetura e no fundamento tecnológico no nível de componente (CHRISTENSEN, 1997). As inovações incrementais são mais freqüentes, pois não demandam muitos esforços em P&D, já as inovações radicais são mais custosas, demandando esforço para a implementação de tecnologias novas que podem substituir tecnologias correntes. As inovações incrementais são percebidas ao longo da evolução de uma determinada tecnologia, ao longo do ciclo de vida da tecnologia, onde se percebem as transformações incrementais de uma tecnologia. Porém em uma inovação radical o ciclo da tecnologia antiga chega ao fim e é substituído por um novo (CHRISTENSEN, 1997).

2.1.2 Inovação Disruptiva

O processo de inovação radical foi abordado por Christensen (1997) como o Dilema do Inovador. Ele percebeu que grandes empresas, líderes em seus segmentos, perdiam

sua posição de mercado para empresas menores. As grandes corporações mantinham-se dominando o mercado com uma determinada tecnologia. Geralmente as inovações feitas por essas empresas eram somente incrementais, pois resultam em menor risco e custo em sua execução. Porém, pequenas empresas surgiam com novas tecnologias, primeiramente tomando nichos de mercado mal atendidos pela tecnologia dominante, e acabavam crescendo e posteriormente "matando" a empresa dominante. A isso Christensen chamou de Inovação Disruptiva (tradução própria, também traduzida como Inovação Interruptiva).

A Figura 2 representa o modelo de Inovação Disruptiva, no qual as inovações sustentadoras, praticadas pelas grandes empresas, dominam o mercado em um primeiro momento. Eles conquistam os clientes mais exigentes, pois a evolução de seus produtos é constante e eles são o *benchmark* da qualidade atualmente existente. As inovações disruptivas modificam totalmente a tecnologia utilizada em um produto ou serviço. Elas aparecem em pequenas empresas e por serem mais baratas, mais simples ou convenientes, conquistam um nicho de mercado menos exigente para si. Essa conquista é suficiente para que os investimentos na nova tecnologia aumentem e ela inicie um processo evolutivo de melhorias. Assim, a nova tecnologia atrai, aos poucos, os clientes mais exigentes e, finalmente, domina o mercado principal da empresa líder, acarretando seu declínio no mercado (CHRISTENSEN e RAYNOR, 2003).

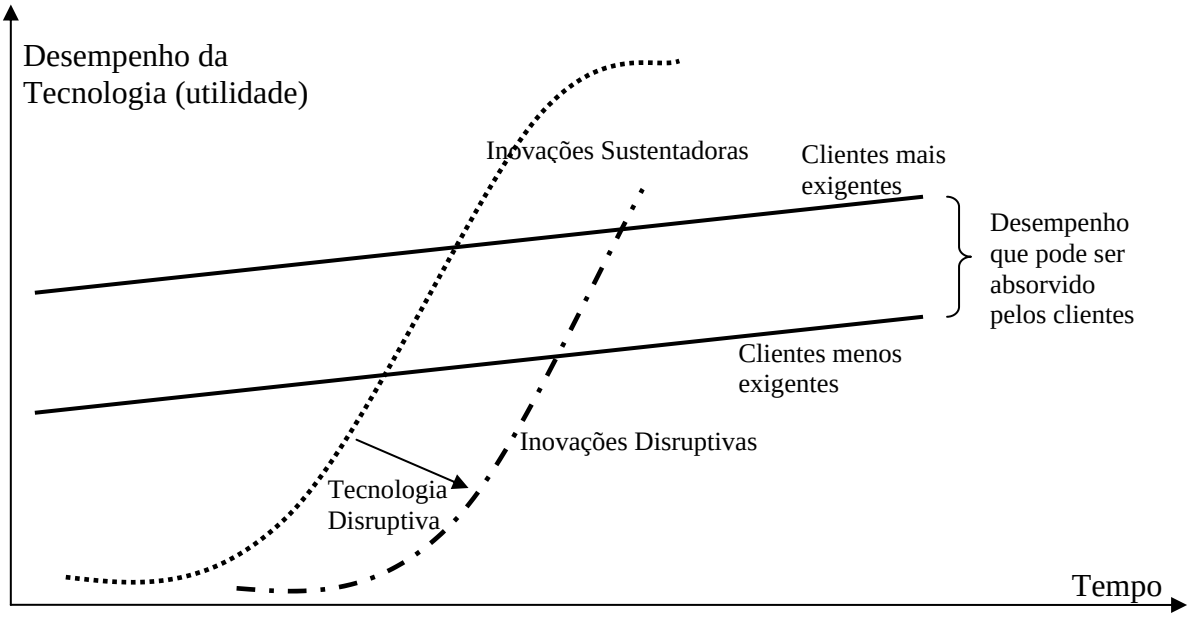


Figura 2 - Modelo de Inovação Disruptiva.
Fonte: Adaptado de Christensen, 1997.

Christensen e Raynor (2003) ainda abordam um elemento crucial no desenvolvimento da inovação disruptiva, o *marketing*. Há uma assimetria motivacional entre as grandes empresas, dominadoras do mercado, e as pequenas, com relação ao nicho pretendido. O esforço financeiro destinado a patrocinar a inovação dentro de uma grande empresa é voltado para o desenvolvimento da sua tecnologia ou produto corrente, pois como o cliente e o produto já são conhecidos, os riscos do investimento são mínimos. Quando uma grande empresa descarta investir recursos de P&D para sanar as necessidades de um determinado grupo de clientes, pois foi decidido que seria uma "aposta" arriscada, é deixado um mercado aberto para as pequenas empresas, que ainda estão desenvolvendo sua nova tecnologia.

2.1.3 Inovação Aberta

Segundo Chesbrough (2003) há dois modelos para produção da inovação: a Inovação Fechada e a Inovação Aberta. A Inovação Fechada foi o modelo de inovação predominante durante a maior parte do século XX. Nele, a ênfase dada para o processo de inovação é tentar, sempre que possível, realizá-lo sozinho como P&D, estudo de mercado, projeto de produto entre outros estágios do desenvolvimento. A Inovação Aberta (*Open Innovation*) é baseada no cenário de conhecimento abundante. Nele, a empresa não precisa esconder seu conhecimento, ele, de algum modo, pode ser compartilhado com outras empresas para que os ganhos gerados possam ser gozados em conjunto.

A filosofia do modelo de inovação fechado foi sumarizada por Chesbrough (2003) como: "Inovações bem sucedidas demandam controle". As companhias deveriam pesquisar o mercado, desenvolver, fabricar, distribuir e acompanhar o cliente por si mesmas. Nesse processo de autoconfiança, altamente centralizado, toda a produção do conhecimento fica confinada dentro da própria estrutura da corporação. As empresas investem mais em P&D interno e contratam os melhores pesquisadores, almejando alcançar o mercado com tecnologias novas antes dos adversários. Assim, há a possibilidade de criar um círculo virtuoso, onde o mercado alcançado antecipadamente pela invenção garante um monopólio que gera mais recursos, que são investidos na P&D interna. Nesse modelo, a propriedade intelectual (PI) é utilizada para monopolizar o mercado e garantir que nenhum competidor possa explorar o mesmo mercado com produtos similares. As patentes que não asseguram a liderança na concorrência não chegam a ser negociadas ou licenciadas e, simplesmente, não

são utilizadas. A difusão do conhecimento entre empresas é vista como prejudicial nesse modelo. A Figura 3 apresenta como o processo de pesquisa e inovação é aproveitado dentro da estrutura das empresas que utilizam o modelo de Inovação Fechada. As saídas da empresa para o mercado são derivadas somente das pesquisas e desenvolvimento interno da empresa.

A partir do final do século XX, a Inovação Fechada começou a definhar, devido à maior mobilidade dos trabalhadores do conhecimento[3] e com o aumento da oferta de capital de risco. Os trabalhadores inovadores, que somente poderiam dar vida a suas idéias dentro de uma estrutura empresarial, conseguiram financiamento para sair de suas antigas empresas e abriram suas próprias, em um processo conhecido como *spin-off*. Isso fez com que as grandes companhias tivessem dificuldade de controlar a propriedade do conhecimento dentro de suas estruturas.

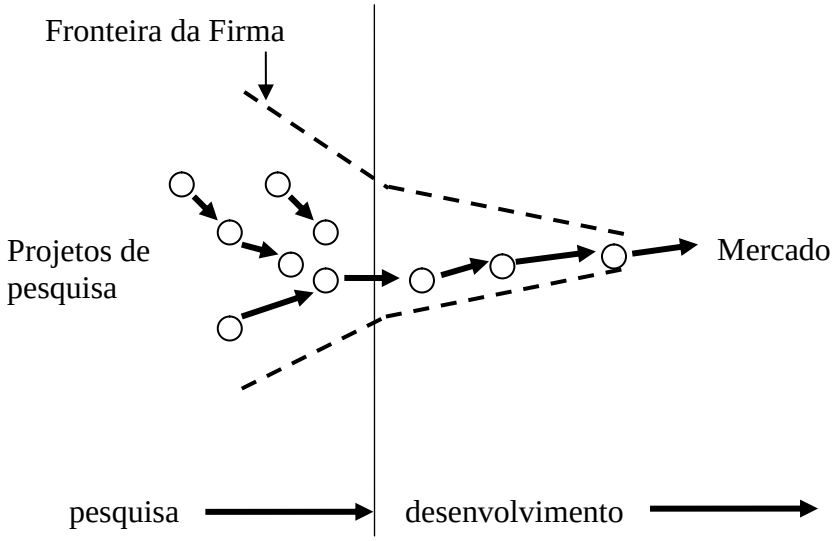


Figura 3 - Aproveitamento da pesquisa dentro do modelo de Inovação Fechada.
Fonte: Adaptado de Chesbrough, 2003.

Na estrutura do modelo de Inovação Aberta, o fluxo de conhecimento é considerado benéfico. A empresa não carece dominar todos os estágios do desenvolvimento de um novo produto. Todo o processo de produção, começando pela ideação do produto até a venda e suporte pós-venda podem ser feitos por empresas diferentes. Porém, desde a empresa que inventa o novo produto até a empresa que viabiliza o produto final têm lucro, pois os

[3] Trabalhadores do conhecimento são todos aqueles trabalhadores que têm o conhecimento como a principal ferramenta de trabalho. Sendo desenvolvendo o conhecimento utilizado na empresa, ou então, utilizando esse conhecimento para o desenvolvimento de produtos.

termos da parceria entre elas são bem estabelecidos previamente. Uma empresa pode utilizar tanto o conhecimento interno como o externo para o desenvolvimento de seu produto, situação essa que é apresentada pela Figura 4. O plano de negócios das empresas torna-se maleável, possibilitando à empresa focar somente no estágio de produção que mais se adéqua ao seu perfil.

Uma importante diferença entre as inovações aberta ou fechada dá-se na forma da utilização das novas idéias. Uma idéia com boas perspectivas de retorno financeiro que, apesar do esforço para seu desenvolvimento, não retorna a expectativa devida é considerada um falso-positivo. Quando uma idéia é descartada, por não ter potencial de retorno financeiro, porém se mostra uma grande oportunidade de negócio, e de algum modo mostra-se viável, essa idéia é considerada um falso-negativo (CHESBROUGH, 2003). Esses erros de avaliação de idéias ocorrem tanto para a Inovação Aberta quanto para a fechada, porém na Inovação Aberta há a capacidade de resgatar os falso-negativos.

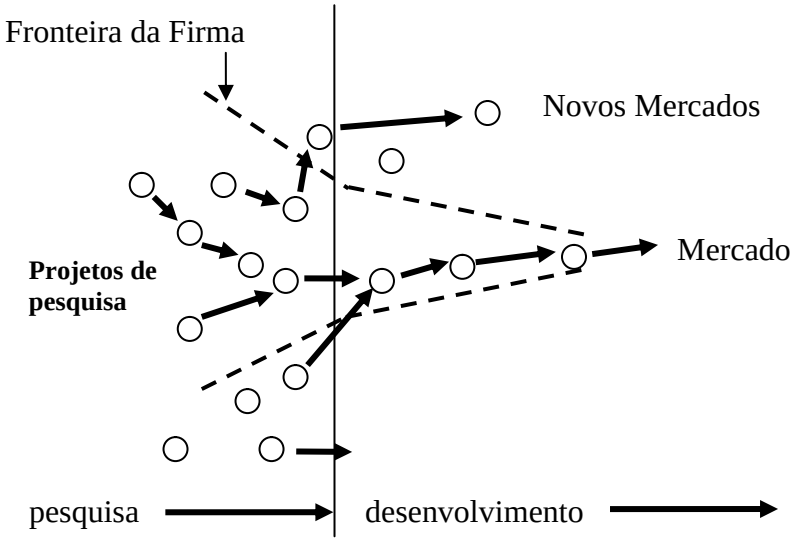


Figura 4 - Aproveitamento da pesquisa dentro do modelo de Inovação Aberta
Fonte: Adaptado de Chesbrough, 2003.

Em empresas de Inovação Fechada, as idéias dos inventores são avaliadas pelos superiores. Se a idéia for boa, há o investimento para o seu desenvolvimento, caso contrário pode sucumbir por falta de apoio. Uma idéia considerada ruim pode não ser amplamente desenvolvida, pois, para a empresa, ela não é aplicável em seu nicho de mercado. Todavia, para a Inovação Aberta sempre há um mercado para o qual a idéia pode ser aplicada e se tornar uma inovação. Como o conhecimento está aberto, várias empresas, de diversas áreas, podem se utilizar dele. Isso abre espaço para a aplicação de uma tecnologia em uma área para

aqual ela não foi inicialmente projetada. Isso é traduzido como a sobrevivência do conhecimento escolhido pelo mercado.

A Inovação Aberta não pode ser confundida com o *outsourcing* do desenvolvimento da tecnologia. Este é a transferência do desenvolvimento de uma tecnologia para centros de baixo custo. Porém, a Inovação Aberta é a busca das boas idéias fora das paredes da empresa e em qualquer lugar, e o uso dessas idéias para a capitalização.

Os custos com P&D utilizando a Inovação Aberta tendem a ser menores, em comparação à Inovação Fechada. A Figura 5 mostra como se dão os custos e os ganhos para a P&D em empresas com modelo de Inovação Fechada e Aberta. Esse modelo de desenvolvimento tem menor custo, pois os laboratórios de pesquisa não são limitados ao espaço físico da empresa e a utilização da tecnologia é através de acordos e licenciamentos. Apesar da receita vinda do próprio mercado continuar igual, outras receitas podem ser geradas através da venda e licenciamento das tecnologias geradas nos próprios laboratórios. Essas tecnologias podem gerar licenciamentos ou vendas das PIs (Propriedade Intelectual), ou até mesmo o *spin-off* em outra empresa que atenda um novo mercado.

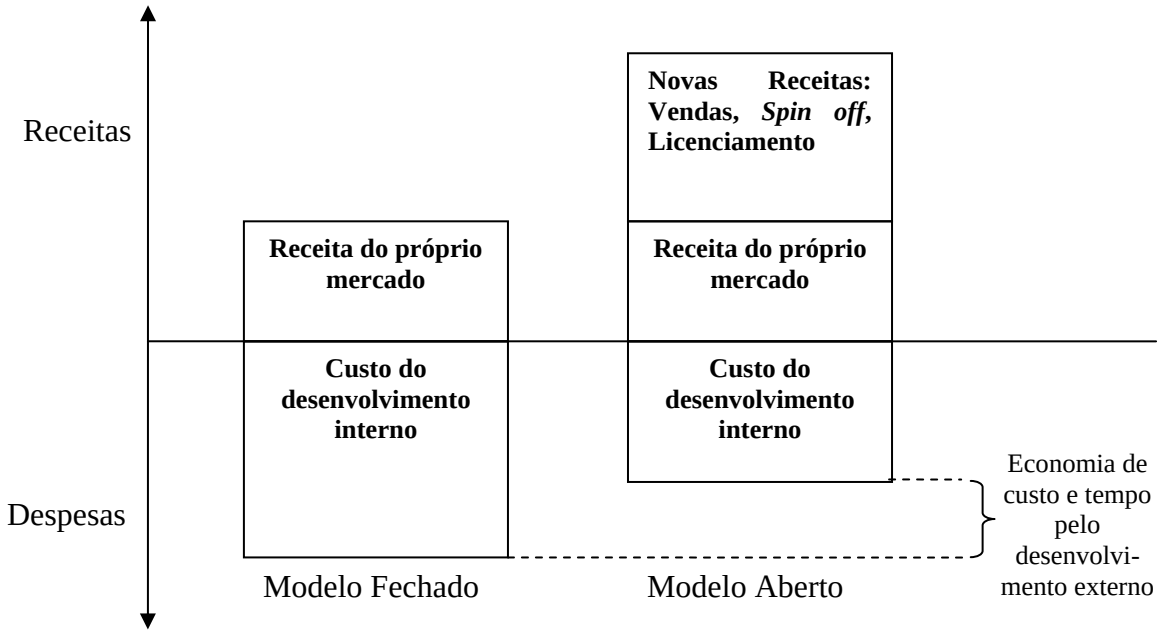


Figura 5 - Custos e Receitas para P&D fechado e o aberto
Fonte: Adaptado de Chesbrough, 2006.

Empresas que continuaram com o modelo de Inovação Fechada até o final do século XX começaram a carregar o fardo de sua PI. Estima-se que menos de 5% do portfólio de patentes das empresas de tecnologia possuem valor significativo; entre 45 e 50% das

patentes são as chamadas patentes defensivas, necessárias para garantir o crescimento futuro das empresas; e 50% das patentes não são relacionadas com os negócios da empresa e acabam não sendo utilizadas (PHELPS, 2006). Para as empresas de alta-tecnologia, as patentes que não geram lucros acabam gerando despesas para manutenção: despesas de PI, advogados, o próprio "desperdício" do investimento prévio em pesquisa entre outros. Exemplo disso foi a IBM, que era uma empresa em que muitos recursos eram destinados à pesquisa e, conseqüentemente, à geração de PI. No entanto, esses ativos eram subutilizados. A partir de 1991, a PI foi transformada em um centro de receitas. A IBM licenciou suas patentes, disponibilizou outras, fez desenvolvimentos conjuntos e, ao longo de uma década, sua PI foi transformada em um negócio de um bilhão de dólares anuais, com margens superiores a 90% (PHELPS, 2006).

Outro exemplo da utilização da Inovação Aberta é a empresa Procter & Gamble (P&G). A empresa começou seu desenvolvimento no modelo fechado. Esse modelo funcionou bem até o começo do século XXI, quando a "explosão" de novas tecnologias começou a elevar os custos de um desenvolvimento totalmente interno. Com os custos de P&D sendo maiores e os retornos financeiros cada vez menores, a diretoria da P&G decidiu apostar na rede de contatos externas às da empresa para a produção de produtos mais rentáveis. Estima-se que para cada pesquisador da P&G há outros 200 cientistas ou engenheiros espalhados pelo mundo, no total de 1,5 milhões de talentos que podem ser utilizados pela empresa. É relatado que os custos de P&D, que eram de 4,8% em 2000, caíram para 3,4% em 2006, porém a produtividade aumentou em perto de 60% (HUSTON e SAKKAB, 2006).

2.1.3.1 Licenciamento de Tecnologia

O licenciamento de uma patente tem muitas vantagens sobre a simples proibição do uso de uma tecnologia por uma empresa concorrente.

Phelps (2006) apresenta a principal vantagem do licenciamento, que é o ganho com *royalties*. Porém, o modo de percepção do licenciamento é importante, pois pode trazer mais ganhos conjuntos. Uma filosofia seria o licenciamento da patente e outra o licenciamento de *know-how*. O licenciamento de patentes é focado no passado, procurando *royalties* dos usos já correntes da tecnologia. Pode ser usado como parte do jogo de proteção intelectual. Por sua vez, o licenciamento de *know-how* é focado no futuro. Isso se traduz no

licenciamento que pode acarretar na diminuição dos custos em P&D e a rápida entrada em mercados para a empresa licenciada. Essa diferença demanda um esforço da empresa licenciadora durante a transferência do *know-how*. Ela deve criar uma estrutura para que a transferência de tecnologia possa acontecer com maior facilidade.

Os ganhos de segunda ordem ao licenciar uma tecnologia, são:

- O uso da tecnologia da empresa licenciadora como padrão da indústria (*standard*).
- O pagamento dos custos de P&D no desenvolvimento das futuras gerações da tecnologia.
- Isso serve como reconhecimento dos desenvolvedores e pesquisadores responsáveis pela criação da tecnologia. Quando uma tecnologia é licenciada eles podem ter acesso a uma parcela do capital gerado por ela. Isso ajuda a mantê-los motivados para a uma inovação continuada.
- Caso não haja o licenciamento, os competidores podem não querer melhorar a própria tecnologia, então o desenvolvimento solitário de uma tecnologia isola o detentor dentro do mercado.

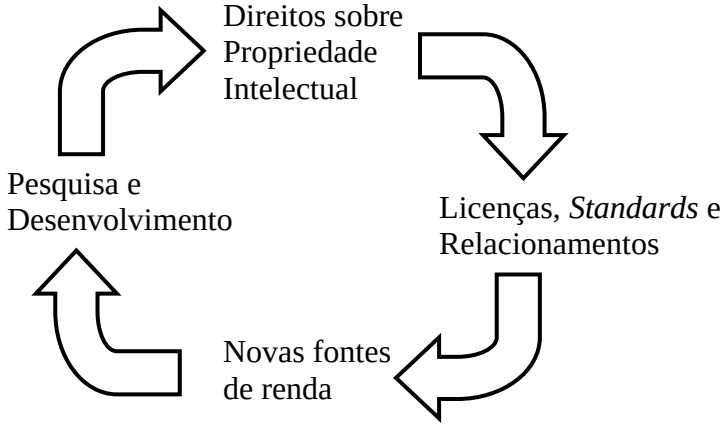


Figura 6 - Círculo Virtuoso gerado pelo Licenciamento de PI.
Fonte: Phelps, 2006.

Segundo Phelps (2006) o licenciamento da tecnologia gera um círculo-virtuoso de desenvolvimento, que é representado pela Figura 6. Por meio do licenciamento de um determinado *know-how* uma companhia pode ajudar a gerar *royalties*, que, por sua vez, definem os padrões da indústria (*standards*), que geram renda extra para a empresa, que financiam mais atividades de P&D. “Estas atividades servem como ‘combustível’ para mais licenciamentos” (PHELPS, 2006).

2.1.3.2 Co-Desenvolvimento

O Co-Desenvolvimento (co-dev) é um modo de inovar no plano de negócios para acentuar a inovação do produto. É um desenvolvimento conjunto entre duas empresas e sua parceria, em co-dev, baixa os custos de P&D de um novo serviço ou produto. Esse modelo abre novos mercados, anteriormente não acessíveis às empresas, separadamente. Ele expande o fluxo de inovações (CHESBROUGH e SCHWARTZ, 2007).

Para a primeira análise do co-dev, é necessário conhecer as necessidades das empresas. Isso depende do seu plano de negócios. Nele, podem-se buscar cinco objetivos: 1) aumentar a lucratividade; 2) diminuir o tempo de lançamento de um produto; 3) acentuar a capacidade de inovação; 4) aumentar a flexibilidade de P&D; e, 5) expandir o acesso ao mercado.

Segundo a análise de Chesbrough e Schwartz (2007) os níveis de comprometimento do co-dev pode ser dividido em: íntima (*core*), crítica e contextual, que são sumarizadas no Quadro 1.

Nível de Comprometimento	**Descrição do Co-desenvolvimento**
Íntima	São realizados com os recursos chave, de alto valor agregado, que a empresa possui. Deve ser gerenciado cuidadosamente e compartilhado eventualmente, pois se dá por meio de um plano de negócios arriscado, que depende de alta confiança em ambos os lados. Preferencialmente, realizar sozinho ou, se necessário, recrutar um fornecedor de P&D.
Crítica	Co-dev vital para finalização de um tipo de produto ou serviço das empresas, porém não é vital para o plano de negócios central de nenhuma as participantes. Realiza-se sozinho somente se não houver opções para parceria.
Contextual	É necessária para completar uma oferta, porém não é suficiente para somar valor agregado ao negócio. Com uma relação de confiança baixa, é de fácil mudança de parceria, caso o parceiro não corresponda às expectativas pode-se mudar rapidamente de parceria.

Quadro 1 - Níveis de comprometimento do co-dev.
Fonte: Adaptado de Chesbrough e Schwartz, 2007.

2.2 COMPLEMENTARIDADE

Complemento é a parte que se junta à outra, para formar um todo completo (PRIBERAM, 2009). A Complementaridade tratada dentro desta pesquisa é a relação matemática entre duas entidades, que permite concluir se uma é o complemento da outra. Em

uma definição formal, é vista como uma interdependência positiva entre pares de prática, também conhecida como teoria da supermodularidade (MILGROM e ROBERTS, 1990; HOLMSTROM e MILGROM, 1994; SIGGELKOW, 2002).

A teoria da supermodularidade tem o ferramental matemático necessário para a determinação da complementaridade entre duas entidades. Neste caso, as entidades são duas empresas. Ela possibilita uma análise empírica da existência de complementaridade. Se há duas atividades A_1 e A_2, cada atividade pode ser realizada por uma firma ($A_i = 1$) ou pode não ser realizada ($A_i = 0$), sendo $i \in \{1,2\}$. A função $\Pi(A_1, A_2)$ é considerada supermodular, e A_1 e A_2 são complementares se:

$$\Pi(1,1) - \Pi(0,1) \geq \Pi(1,0) - \Pi(0,0) \quad (1)$$

Isto é, somando-se uma atividade à outra atividade em andamento obtém-se um efeito incremental no desempenho, maior que a soma de cada atividade isoladamente.

Cassiman e Veugelers (2006) utilizaram a teoria de supermodularidade para verificar a existência direta de complementaridade entre empresas. Nessa pesquisa, foram categorizadas as estratégias de inovação das empresas em quatro níveis: (1) empresas que não criam inovação e somente compram (*NãoFazem&Compram*); (2) empresas que somente criam internamente sua inovação (*SoFazem*); (3) firmas que somente adquirem inovação (*SoCompram*); e, (4) empresas que combinam desenvolvimento e aquisição (*Fazem&Compram*).

O parâmetro utilizado por Cassiman e Veugelers (2006) para medir o desempenho da inovação é o lucro gerado por produtos inovadores ou produtos substancialmente melhorados nos últimos dois anos. Esse trabalho mostrou que empresas que compram e produzem tecnologia (*Fazem&Compram*) têm mais lucros que as demais empresas, apresentadas anteriormente. Outro resultado interessante desse trabalho é o levantamento de algumas variáveis contextuais que afetam a complementaridade.

Vê-se que a teoria de Complementaridade é respaldada pelo quadro teórico da Inovação Aberta. Uma vez que o desenvolvimento conjunto é o foco deste trabalho, o trabalho de Cassiman e Veugelers (2006) mostra que as empresas de maior lucro são aquelas que conseguiram trabalhar conjuntamente a pesquisa interna com externa, não somente um dos lados.

2.3 PATENTES

Os registros de patente são títulos que conferem o monopólio temporal de um determinado conhecimento a uma pessoa física ou jurídica, limitado geograficamente. Elas são a espinha dorsal do sistema de Propriedade Intelectual (PI). Berman (2006) também chama a patente de um “direito negativo”, pois ela não permite, por si só, ao proprietário a prática da invenção patenteada, mas confere a ele o direito de proibir outros de a praticarem.

2.3.1 Sistema de Proteção da Propriedade Intelectual

O conceito de direitos do inventor surgiu nos século XV, mas não foi desenvolvido plenamente até o século XIX. Em 20 de março de 1883 foi assinada em Paris a Convenção para a Proteção da Propriedade Industrial. O propósito da convenção foi interligar e harmonizar as leis de propriedade intelectual (PI) entre onze países, e também criar alguma instituição legal internacional (ORGANIZATION..., 1994).

O sistema de proteção de PI é um elemento importante para o crescimento econômico. Foray (2007) ressalta a importância das patentes como um elemento de equilíbrio dentro da economia do conhecimento. O conhecimento é móvel, e a movimentação de “conhecimento superior” é boa para a melhoria da qualidade da vida humana. Entretanto, Johnson e Lundval (2000) mostram que, dentro da economia do conhecimento, há uma contradição que se refere à criação do conhecimento. Gerar o conhecimento é uma tarefa árdua, que exige das universidades e da indústria investimentos de retorno duvidoso. Porém, uma vez que esse conhecimento se torna evidente, codificado em forma de um produto ou serviço, ele abre as portas aos concorrentes para se apropriarem do conhecimento e lucrarem. Todavia, os concorrentes não investiram seu capital no trabalho de pesquisa. Isso sufoca o criador da nova tecnologia e mata a cadeia criativa (JOHNSON e LUNDVAL, 2000). O patenteamento do conhecimento gerado sana essa contradição, conferindo àquele que gerou a tecnologia o poder de vender o direito de produção, ou de garantir o monopólio de seu uso (FORAY, 2007).

O sistema de proteção do direito intelectual confere ao inventor o direito de usufruir de suas invenções. Os títulos de patente têm o poder de quebrar a concorrência econômica perfeita (SCHUMPETER, 1984). E, além disso, proporcionam a divulgação do estado da arte da tecnologia. Quando algo é patenteado, todo o conhecimento gerado em seu

projeto de pesquisa e desenvolvimento deve ser codificado textualmente em detalhes. Isso facilita o acesso ao conhecimento, pois ajuda a nivelar o conhecimento entre as empresas. Faz o círculo virtuoso da geração constante de novos conhecimentos.

O sistema de proteção da PI gera direitos transferíveis, que são ativos “invisíveis” que dão valores às empresas de pequeno porte. Empresas menores têm em seu capital social a maior parte do seu valor, principalmente em empresas inovadoras. Um modo de assegurar que o conhecimento retido pela empresa tenha valor é por meio dos registros de patente.

2.3.2 Utilização de Patentes como Indício de Inovação Tecnológica

A existência de inovação tecnológica não é condição suficiente para a existência de patentes. Igualmente errado é afirmar que em um ambiente sem patentes não há inovação. Há muitas empresas que sobrevivem no meio corporativo fabricando produtos cuja patente não existe ou já está expirada (POLTORAK, 2006). Discrição no desenvolvimento, rápido lançamento, desenvolvimento de conhecimento intrinsecamente tácito ou baixos preços também podem suplantar o uso das patentes. Elas podem ser desperdício de verbas em áreas de rápido desenvolvimento, pois a invenção facilmente atinge a obsolescência (ORGANIZATION..., 1994). Porém, as patentes podem servir como indício de que a inovação tecnológica está acontecendo, pois seu conteúdo é a explicitação de uma idéia. Isso é o mais perto que se pode chegar da documentação de um conhecimento novo. E sua cobertura é bastante ampla. Em resumo, nem todas as aplicações do conhecimento são patenteadas e nem todas as patentes são relevantes (ORGANIZATION..., 1996).

Muitas ferramentas para análise tecnológica utilizam os registros de patentes como fonte de informação na entrada de seus sistemas. Elas são apresentadas na Seção 2.4.5 (Sistemas Comerciais para Mineração de Texto e Visualização).

Os documentos utilizados para a promoção da mineração de dados podem ser classificados quanto à ordenação de seus dados dentro do documento. Ela pode ter três formas: (1) textos não-estruturados, que não têm padronização, como o encontrado em e-mails, notícias de jornais, apresentação de portfólio, sítios da Internet; (2) textos estruturados, nos quais a apresentação da informação está em campos bem definidos; e, (3) textos híbridos, que é o caso das patentes, cujo texto está estruturado na primeira página, porém há texto não-estruturado na apresentação da invenção (YANG *et al.*, 2008). Portanto, técnicas estatísticas podem ser utilizadas para a observação de um conjunto de patentes, caso os dados utilizados

sejam da primeira página, ou então uma análise textual ou semântica pode ser utilizada para o estudo da descrição da patente.

A existência desses documentos em meio digital facilita a automação da análise dos mesmos. Assim, como é mostrado no APÊNDICE A, documentos datilografados têm potencial de uso em Mineração de Dados. Entretanto, devem ser digitalizados e são tratados como documentos não-estruturados, já que a conversão do meio analógico para o digital pode trazer erros (tecnicamente denominados ruídos) que podem comprometer a análise estruturada.

2.3.3 Campos dos Documentos de Patente

Os documentos de patentes são publicações ricas em informações relevantes. Todos os campos desse documento, caso analisados cuidadosamente, podem fornecer algum nível de informação sobre o andamento de: investimentos em áreas de pesquisa, formação de redes de conhecimento, evolução em um campo de pesquisa, entre outros.

As informações obtidas desses documentos podem ser estruturadas e categorizadas, como no caso do título, data de publicação, citações e autores. Ou podem apresentar informações não-estruturadas no caso das reivindicações, descrição e resumo. As principais informações encontradas no registro de patente são apresentadas nos tópicos abaixo:

- Titulares (*Applicants*): Nome do indivíduo ou empresa que solicita a proteção de uma determinada invenção.
- Inventores (*Inventors*): Nome das pessoas que inventaram a nova tecnologia e desenvolveram a invenção.
- Descrição (*Description*): Explicação clara do conhecimento e tecnologia existente relacionada à nova invenção e a explicação de como essa invenção pode ser aplicada na resolução de problemas que não são contemplados pela tecnologia atual.
- Reivindicações (*Claims*): Definições legais dos tópicos que interessam aos titulares como sua invenção e que tipo de proteção lhe é conferida. Cada reivindicação é uma única sentença, em forma legal, que define uma invenção e uma única característica técnica.

- Prioridade de Submissão (*Priority Filing*): A primeira submissão original na base no qual a posterior submissão nacional, regional ou internacional pode ser realizada no período de um ano.
- Data de Prioridade (*Priority Date*): Data da primeira submissão. A partir dessa data a patente está protegida, caso o pedido seja bem sucedido. E a data a partir da qual passado um ano começa o período de prioridade para posteriores pedidos.
- Data de Submissão (*Filing Date*): Data de submissão de uma patente individual em um escritório de patentes. É a data que a invenção está protegida caso o pedido seja aceito.
- Estado Legal (*Legal Status*): Indica se a patente foi concedida. Se concedida, mostra a quais países ou regiões que a patente é válida. Mostra o estado da patente, caso está válida, ou expirada, ou tenha sido invalidada em algum país ou região em particular.
- Citações e Referências: Podem ser de documentos que sejam patentes ou não. Os documentos incluem referências para tecnologia relacionada que não está coberta pelo requerente (ou pelo examinador durante o processo de concessão).
- Dados Bibliográficos: Como são chamados todos os dados da página inicial da patente. As informações encontradas são: dados de identificação, dados domésticos de submissão, dados de prioridade, dados de publicação, dados de classificação e outros dados referentes ao conteúdo técnico.
- Códigos de tipo (*Kind Codes*): Utilizados para distinguir documentos de patentes de acordo com seu estado.
- Códigos INID: (*International agreed Numbers for the Identification of bibliographic Data*) Tem os requisitos mínimos para a identificação dos diferentes elementos bibliográficos.

Vários campos do registro de patente podem ser utilizados para mineração de dados. Porter e Cunninghan (2005) mostram como é possível a utilização das citações para criar redes de conhecimento dos autores.

Sternitzke, Bartkowski e Schramm (2008) utilizaram estatísticas bibliométricas em informações de co-citação para a análise de redes sociais de desenvolvimento. Os autores apresentaram dois tipos de dificuldade no trato das citações: a primeira é do nome do inventor aparecer com grafia diferente em registros diferentes; e, o segundo, autores homônimos, que é uma situação que tem sua probabilidade aumentada quando o campo de pesquisa é grande, os nomes e sobrenomes dos inventores são populares. Os nomes homônimos são gerados mais facilmente com a abreviação do nome para a primeira letra. Em seu trabalho eles mostram que

essa técnica facilita a visualização das várias redes de cooperação de campos de pesquisa amplos. Os nodos das redes representam os inventores e desenvolvedores, e os laços representam as relações entre eles. Essas análises mostraram-se úteis para gestão de recursos humanos, procura de parceiros em projetos de P&D, análise dos competidores, entre outras análises.

2.4 FERRAMENTAS PARA ANÁLISE DE DADOS TECNOLÓGICOS

A proposta desta seção é apresentar as ferramentas e metodologias que auxiliam na análise de grandes quantidades de comunicação tecnológica publicada. Essas técnicas ajudam o entendimento das publicações em uma visão macro. As técnicas apresentadas serão bibliometria, a cienciometria, a TRIZ e a Mineração de Tecnologia.

A bibliometria e a cienciometria são técnicas similares, baseiam-se em matemática e regras estatísticas para alcançar seus resultados. A Mineração de Tecnologia é uma técnica computacional que vai além da estatística, ela também visa buscar padrões bem definidos dentro dos dados estudados. A TRIZ é uma técnica diferente das mencionadas acima, pois não é utilizada para a análise de grandes quantidades de dados, usualmente sua utilização é para a solução de problemas inventivos, todavia deve ser mencionada porque ela nasceu do estudo criterioso e individual de uma quantidade massiva de registros de patentes, e ele mostra padrões de publicação de patentes durante o ciclo de vida de diferentes tecnologias. A TRIZ mostra uma visão macro de todo o ciclo tecnológico a partir das publicações de patentes.

A última subseção deste capítulo apresentará um resumo das ferramentas computacionais comerciais que são utilizadas para mineração e visualização de grandes quantidades de dados.

2.4.1 Bibliometria

A bibliometria, ou também bibliotecometria, é a área de estudos que utiliza matemática e estatística para estudar um grupo de publicações. Nela o conhecimento escrito pode ser quantificado por meios estatísticos para facilitar sua análise e atribuir valores às publicações (VANTI, 2002). As publicações científico-tecnológicas têm campos bem definidos: autores, título, citações, palavras-chave, data de publicação e outros. Esses campos

são facilmente relacionáveis, o que facilita estatísticas em relação a eles. Esses estudos categorizam o processo de comunicação escrita (GUEDES e BORSCHIVE, 2005).

O campo de estudo da bibliometria foi criado por E. Wyndham Hulme em 1922, porém com o nome de *statistical bibliography* (bibliografia estatística, em tradução literal), primeiramente com o intuito de esclarecer os processos científicos e tecnológicos por meio da contagem de documentos. Porém, somente com Pritchard esse campo de estudos começou a se chamar bibliometria (*Bibliometrics*) com seu artigo *Statistical Bibliography or Bibliometrics* (PAO,1989). Segundo Guedes e Borschive (2005) ela é o conjunto de leis e princípios que contribuem para estabelecer os fundamentos teóricos da Ciência da Informação.

As principais leis bibliométricas são: Lei de Bradford, Lei de Lotka e Leis de Zipf. A primeira é sobre a produtividade de periódicos em relação a determinados assuntos. Com ela pode-se verificar o grau de relevância de um determinado periódico em alguma área de conhecimento. A segunda lei, de Lotka, relaciona a produtividade dos pesquisadores, entre os que produzem mais e os que produzem menos. E a lei de Zipf, que relaciona a freqüência das palavras, o que permite criar uma indexação automática de um tema (GUEDES e BORSCHIVE, 2005).

Análises bibliométricas utilizam diversos campos de estudo como objeto de sua pesquisa, como: *reviews* de campos de pesquisa, meta-análises, avaliação de centros de pesquisa, comparação sistemática entre áreas de pesquisa, atribuição de pesos a publicações, revistas e a pesquisadores. Essas são algumas das informações que podem ser levantadas por meio do uso da bibliometria são:

- colégios invisíveis[4] e frentes de pesquisa;
- relevância de um artigo, através do fator de impacto, que nada mais é que o número de citações de cada artigo publicado pelo periódico em dois anos dividido pelo número total de artigos publicados nesse período;
- grau de ligação de dois artigos, medido pelo número de co-citações idênticas;
- ciclo de vida de uma literatura, que é o tempo de uso de uma literatura;

[4] Colégio Invisível é um grupo de cientistas que trabalha em uma tradição de pesquisa e acaba formando os mesmos valores científicos naquele grupo. Não estão necessariamente alocados no mesmo laboratório, nem tão pouco, na mesma localidade. O grupo tem esse nome, pois o diálogo entre laboratórios ou cientistas não é visto, porque preferem utilizar uma comunicação mais direta que os meios mais lentos da comunicação científica formal.

- elitismo na ciência;
- razão epidêmica de uma idéia, que se fundamenta na teoria de Goffman, de que o desenvolvimento das idéias atua como doenças infecciosas dentro da comunidade científica e que após o crescimento declinaria e se tornaria estável (GOFFMAN & NEWILL, 1964);

2.4.2 Cienciometria

A cienciometria (*Scientometrics*) pode ser definida como a área de pesquisa que estuda quantitativamente os aspectos da comunicação científica, abrangendo as práticas de pesquisa e desenvolvimento (P&D) e as políticas de ciência, tecnologia e sociedade (CTS). Ela é um braço da sociologia da ciência que utiliza a bibliometria, como ferramenta de análise dos dados quantitativos das publicações, no auxílio da tomada de decisões para o desenvolvimento de políticas científicas (MACIAS-CHAPULA, 1998).

A cienciometria recebeu destaque com o início da publicação da revista *Scientometrics*, em 1978 e tem notoriedade até os dias de hoje. Porém foi somente nos anos 80, com a venda da base de dados do *Institute for Scientific Information* (ISI) para diversas instituições de pesquisa, que a cienciometria se tornou uma área de interesse acadêmico.

2.4.3 Descoberta de Conhecimento em Banco de Dados Tecnológicos

A área da computação que pesquisa a descoberta de conhecimento em banco de dados é o KDD (*Knowledge Discovery in Databases*). Ela pode ser definida como um processo não-trivial de identificação de dados novos, válidos, potencialmente úteis, e dispostos em padrões dentro de banco de dados (TAN, STEINBACH e KUMAR, 2006).

Não há um marco de início da utilização de KDD na extração de informação de base de dados científico-tecnológicas para o fomento da inovação. Todavia, o casamento do KDD com as bases de dados científico-tecnológicas pode ser considerado como algo orgânico, que aconteceria naturalmente. Para a aplicação do KDD são necessárias grandes bases de dados, o que é facilmente encontrado nos centros de registros de patente (*e.g.*, WIPO, USPTO, Espacenet, INPI) e bases de artigos científicos (*e.g.*, ISI, ScienceDirect) que têm esse material armazenado, classificado e disposto de modo ordenado.

Porter e Cunninghan (2005) cunharam o termo *Tech Mining* (Mineração de Tecnologia, em tradução literal) para designar a utilização de técnicas computacionais para encontrar informações relevantes em bancos de artigos ou patentes. Eles certamente não foram os primeiros a realizar a Mineração de Dados nesses bancos, há relatos de utilização de técnicas de KDD em bases científico-técnológicas, com finalidade de complementar estatísticas bibliográficas em análises cienciométricas (GLENISSON *et al.*, 2005). Até mesmo o uso da bibliometria pode ser considerado como uma parte do KDD, como será apresentado posteriormente. Porém, a utilização de técnicas computacionais de processamento de dados tecnológicos, para a obtenção de vantagens competitivas é denominada Mineração de Tecnologia (PORTER e CUNNINGHAN, 2005).

Como esta pesquisa é interdisciplinar, é importante realizar a unificação de termos. Para o campo de gestão de tecnologia o termo “conhecimento” tem o significado diferente do que o mesmo termo para o KDD. Para a gestão de tecnologia o conhecimento é um bem, ou um ativo, existente incorporada (*embodied*) dentro das pessoas ou dentro de tecnologias (ORGANIZATION..., 1996). Entretanto, para o KDD o conhecimento é uma informação útil encontrada dentro de um amplo conjunto de dados. Para uniformizar os conceitos dentro desta pesquisa, o termo “conhecimento” do KDD será referido, daqui por diante, como “informação útil”, enquanto que o conceito válido para o termo “conhecimento” será do campo de gestão de tecnologia.

2.4.3.1 Descoberta de Informações Úteis em Banco de Dados

Antes de prosseguir com o uso do KDD como ferramenta para a gestão da tecnologia e inovação, serão apresentados os conceitos gerais do campo de estudos de descobrimento de informações úteis em banco de dados.

O KDD é utilizado como ferramenta de análise em muitas áreas que produzem grande volume de dados, que podem ser ordenados e processados conjuntamente. Alguns exemplos de utilização do KDD são: a de vendas no varejo, buscando informações sobre os padrões de consumo de seus clientes; a indústria o utiliza para detectar pontos de falha na cadeia de produção; geneticistas o utilizam para descobrir os padrões de genes e suas utilidades; os bancos utilizam para determinar o padrão de consumo de seus clientes e poder oferecer-lhes serviços mais adequados, ou então um padrão de consumo diferente do normal, o que já aciona uma verificação de fraude; entre diversas outras aplicações.

O processo de KDD consiste em alguns passos, que vão desde como os dados são acessados até a interpretação e visualização dos resultados. Esses passos são variáveis, dependendo da metodologia utilizada. A metodologia proposta por Fayyad *et al.* (1996) consiste em nove passos, que vão se alternando até a consolidação do padrão descoberto. Essa metodologia não é a única, porém, é muito utilizada no meio acadêmico. Seus passos estão sumarizados abaixo:

1. desenvolver e entender o domínio da aplicação;
2. criar um conjunto de dados alvo, selecionar os atributos e os exemplos;
3. limpar os dados e pré-processamento;
4. redução da dimensionalidade dos dados e projeção;
5. escolher a tarefa de mineração;
6. escolher o algoritmo de mineração mais apropriado;
7. minerar os dados;
8. interpretar dos padrões obtidos;
9. consolidar o padrão descoberto, analisar os dados para a tomada de decisão.

Esses passos são executados seqüencialmente, porém seu sentido não é único. Após a execução de um determinado passo, caso ele não revele um resultado satisfatório, pode-se voltar a um passo anterior para que a nova execução revele novas informações. O processo é iterativo, no qual os passos interagem inúmeras vezes até que o resultado se apresente satisfatório. A Figura 7 apresenta o fluxo das tarefas do KDD.

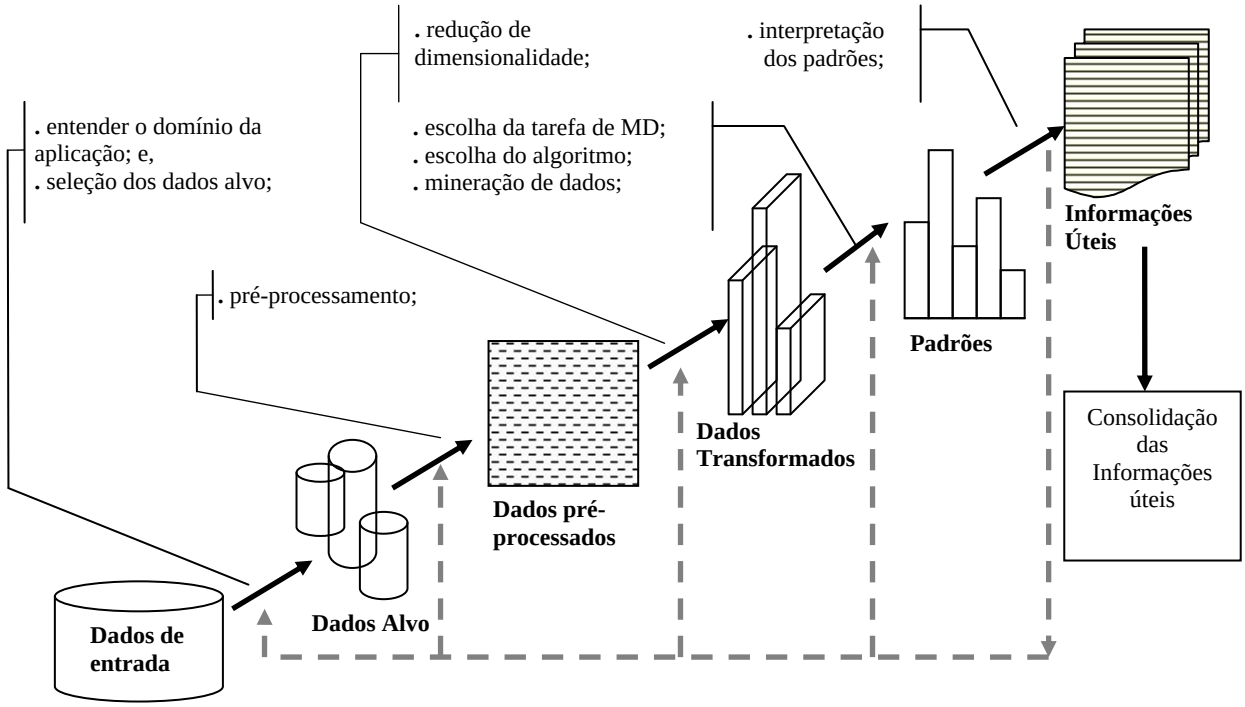


Figura 7 - Passos do KDD e os estados dos dados durante o processo.
Fonte: Adaptado de Fayyad *et al.*, 1996.

2.4.3.2 Mineração de Dados

A mineração de dados (MD) é o sétimo passo dentro da metodologia do KDD (FAYYAD *et al.*, 1996). Ela é baseada em métodos computacionais de aprendizagem de máquina que produzem padrões ou modelos sobre os dados analisado e somente é viável com um grande volume de dados. Caso o quarto passo (redução de dimensionalidade dos dados e projeção) consiga reduzir a dimensionalidade sensivelmente, não se faz viável a análise dos dados através das técnicas de mineração de dados, cabendo perfeitamente uma análise estatística. Se forem análises tecnológicas, cabe o uso de técnicas bibliométricas. A mineração de dados não é realizada somente de um único modo. Ela pode assumir três formas de tarefas: associação, agrupamento e classificação. Cada tarefa tem o potencial de revelar diferentes tipos de informações. Esse é uma decisão importante dentro do KDD, o que explica a existência do quinto passo, em que é decidido o método utilizado:

- **Associação**: correlaciona as características das instâncias processadas, apontando padrões entre eles.

- **Agrupamento**: separa as instâncias processadas em grupos que tenham características em comum. A correspondência entre as características que sobressaltam em cada grupo é dada pela interpretação do pesquisador.
- **Classificação**: rotula cada instância processada em classes pré-estabelecidas.

É um erro pensar que a mineração de dados é somente a extração de estatísticas em grandes quantidades de dados. A mineração de dados vai além disso, ela é uma área interdisciplinar que mistura aprendizagem de máquina, inteligência artificial e estatística (ANDREATTO, 2002) para descobrir novos padrões e relações entre os dados, que não seriam facilmente observados a olho nu.

Apesar haver a confusão entre a MD com o processo de KDD, ou até mesmo, em algumas vezes, os termos serem utilizados como sinônimos, o processo de MD não é o processo mais complicado, nem o mais trabalhoso. Segundo Cios *et al.* (2007), em algumas estimativas, o processo de preparação dos dados, o que compreende o pré-processamento e a diminuição da dimensionalidade, pode levar até 60% do esforço no processo de KDD, enquanto que o processo de MD pode levar de 10 a 18% do esforço no processo total.

2.4.3.3 Mineração de Tecnologia

O campo de estudos de *Tech Mining* (Mineração de Tecnologia) foi definido por Porter e Cunninghan (2005), que o descreveram como: "... a aplicação de ferramentas de mineração de texto às informações científico-tecnológicas, assessorada por entendimento sobre processos de inovação tecnológica.". Porém essa definição não é totalmente correta, ela apresenta uma lacuna conceitual, cabendo uma extensão de seu conceito. Como visto anteriormente, a mineração de dados é somente uma das etapas de um processo maior, o KDD. E, a interpretação dos resultados da execução da mineração de dados também faz parte desse processo. Outro ponto, o conteúdo apresentado em Porter e Cunninghan (2005) é apresentado que a mineração de tecnologia (MT) compreende desde a escolha da base de dados envolvida até a interpretação dos resultados obtidos. Então, pode-se afirmar, estendendo o conceito apresentado por Porter e Cunninghan, que a Mineração de Tecnologia é o próprio processo KDD, porém utilizado exclusivamente para a extração de informações úteis para a gestão de inovação, a partir de bases de dados científico-tecnológicas.

Dois conceitos importantes trabalhados na MT são: *Exploration* e a *Exploitation*. A primeira é o ato de alavancar o conhecimento já existente, na busca de novos

conhecimentos. Um exemplo de aplicação é a TRIZ. A segunda é o ato de juntar conhecimentos de áreas diferentes, combinando e re-combinando para a extração de um novo produto, buscando atender uma necessidade de mercado (MARCH, 1991). Esses conceitos são importantes, pois mostram o potencial de uso da Mineração de Tecnologia.

Na seção 3.1 (Conceituação Técnica) serão abordados mais profundamente os conceitos de Mineração de Tecnologia durante a descrição das atividades utilizadas para alcançar os resultados desta pesquisa.

2.4.4 TRIZ - Teoria da Solução Inventiva de Problemas

TRIZ é a sigla que em russo quer dizer: Teoria da Solução Inventiva de Problemas (de CARVALHO, 2007). Ela foi criação do engenheiro russo G. S. Altshuller, e teve o início de seus estudos na década de 40, quando Altshuller e seus companheiros estudaram centenas de milhares de patentes, eles verificaram que os tipos de problemas e os tipos de soluções se repetiam entre as diferentes áreas técnicas. Inicialmente, buscava-se desenvolver uma metodologia generalizada o suficiente para a resolução de problemas inventivos, porém a TRIZ expandiu-se para além desta perspectiva.

A TRIZ difere das outras ferramentas para a análise de patentes apresentadas anteriormente nesta dissertação, pois ela analisa o micro para entender o macro. Isto é, a análise de várias patentes proporcionou que os problemas resolvidos por elas pudessem ser generalizados, assim, a solução de muitos dos problemas que surgem de novos desenvolvimentos podem ser sistematizados. A TRIZ permite que a criatividade seja controlada. O processo normal de resolução de erros em inventos pode ser representado pela Figura 8A, enquanto o processo de resolução de problemas utilizando TRIZ seria representado pela Figura 8B. No primeiro, os problemas do produto, ou também projeto, são trazidos para a criatividade humana para sua resolução. No segundo, o problema é generalizado e a solução genérica ao problema é aplicada. Então, a solução é particularizada para aquele caso.

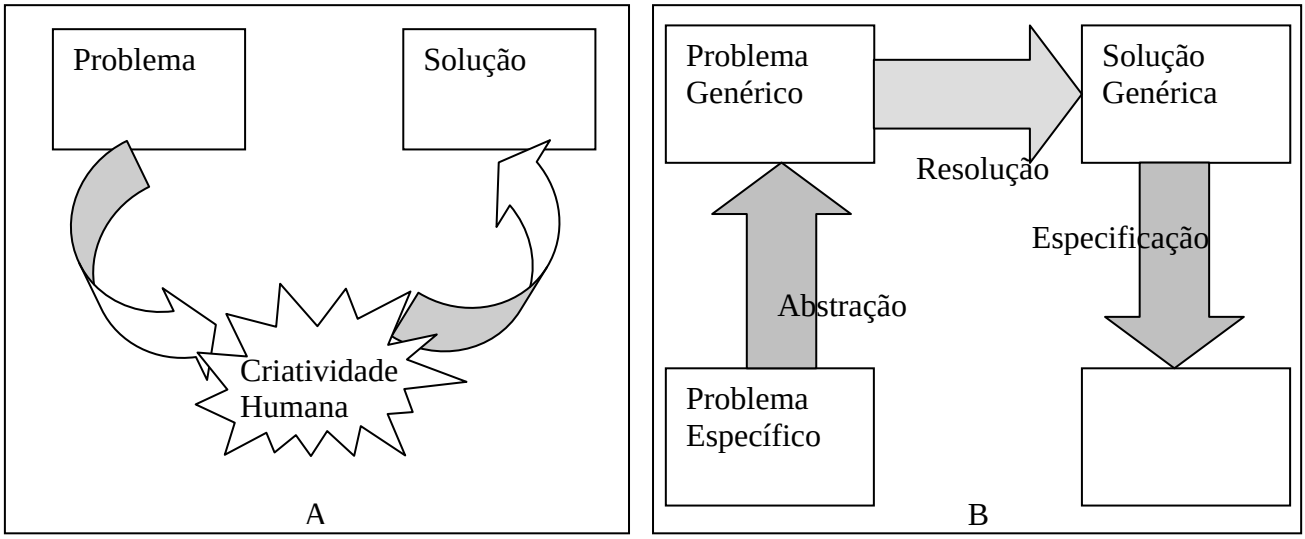


Figura 8 – Modos de resolução de problemas: (A) estratégia normal; e, (B) estratégia utilizando TRIZ.

Fonte: Adaptado de de Carvalho, 2007.

Nessa pesquisa, porém, a utilização da TRIZ é relevante dentro da percepção dos padrões de patenteamento com relação à evolução do ciclo de vida de uma tecnologia, o que é inerente de uma das heurísticas da TRIZ chamada Tendências da Evolução. É importante enfatizar que teoria não fica limitada a essa heurística, há outras ferramentas de ideação que não serão apresentadas, porém merecem ser citadas: o Método dos Princípios Inventivos, as Heurísticas para a Transformação de Sistemas, as 121 Heurísticas, o Método da Separação, os Efeitos Científicos, o Método das Pequeninas Pessoas Espertas, o Método das Partículas ou Método dos Agentes, a Análise Substância-Campo, o Algoritmo para a Solução de Problemas Inventivos (ARIZ), a Hibridização, o Método SIT, o *software* de TRIZ e a IDEATRIZ (de CARVALHO, 2007).

Altshuller (1979 apud de CARVALHO, 2007) mostra que os padrões de patenteamento podem apresentar como um sistema tecnológico pode evoluir com o tempo. A análise desses padrões pode ser uma ferramenta empírica muito ilustrativa. Todavia, esses padrões podem ser de difícil correlação com curvas reais. Algumas estatísticas importantes sobre patenteamento foram apresentadas por Altshuller. Ele apresentou níveis inventivos nas patentes por ele pesquisadas. São cinco níveis, que vão desde o patenteamento de inventos triviais até descobertas científicas. O Quadro 2 apresenta o resumo dos níveis.

O nível 1 de invenções são as invenções triviais, são aquelas limitadas às pequenas modificações no estado da técnica. As invenções de nível 2 necessitaram de mais conhecimento do inventor, porém ficam limitadas dentro da própria área técnica e não passam de uma melhoria da técnica. As invenções de nível 3 trazem elementos estranhos àquela área

da indústria, esse tipo de invenção resolve uma contradição técnica de produto. Uma contradição técnica ocorre quando há duas características incompatíveis ou opostas dentro de um objeto tecnológico. Isto é, a tentativa de melhorar um pode piorar o outro. O quarto nível inventivo apresenta invenções que não estão relacionados com o estado da arte, elas representam um novo paradigma dentro da mesma técnica. E o nível inventivo 5, é o fruto do pioneirismo científico, representa a descoberta de um novo efeito ou fenômeno que abre novos caminhos dentro da ciência.

Nível da Invenção	% do total	Número Estimado de Tentativas	Posição do Problema e dos Meios de Solução
1 – Trivial	32 %	1 a 10	Dentro de uma área de uma profissão
2 – Melhoria	45%	10 a 100	Dentro de uma área de uma indústria
3 – Novidade dentro do paradigma atual	19%	100 a 1000	Em uma área da ciência
4 – Novidade dentro de um novo paradigma	< 4%	1000 a 10000	Fora da área da ciência onde o problema foi originado
5 – Descoberta científica	< 0,3%	> 10000	Fora dos limites da ciência contemporânea

Quadro 2 - Níveis Inventivos das patentes.
Fonte: Adaptado de de Carvalho, 2007.

Por meio da classificação no nível inventivo conseguiu-se traçar alguns padrões de patenteamento nos ciclos tecnológicos. A Figura 9 apresenta os vários aspectos do ciclo tecnológico, a Figura 9A representa o ciclo de vida de uma tecnologia. A Figura 9 B apresenta o número de invenções de uma tecnologia durante o tempo. A Figura 9 C apresenta o nível das invenções, e por fim a Figura 9 D mostra a lucratividade das invenções durante o seu ciclo de vida.

Temporalmente a Figura 9 apresenta os estágios de uma tecnologia. Seu nascimento acontece em T0, quando é reconhecida uma necessidade social, ou quando é fruto de descobertas científicas de alto nível. Geralmente a técnica é primitiva, não-confiável, ineficiente e tem muitos problemas sem solução.

Sua infância, espaço de tempo entre T0 e T1, é caracterizada por ser intensiva em pessoas. São poucas pessoas, devido ao baixo número de patentes apresentadas por B, porém com alta capacitação, que representa o alto nível das invenções em C. Como fase do desenvolvimento tecnológico é de pesquisa básica o campo ainda não dá retorno financeiro, e as finanças são tratadas como investimento que tem potencial de retorno futuramente, capital

de risco, o que é visto em D. O potencial de uso comercial da tecnologia é reconhecido no final da infância, onde os resultados da pesquisa básica começam apresentar um nível mais maduro de invenções (SAVRANSKY, 2001).

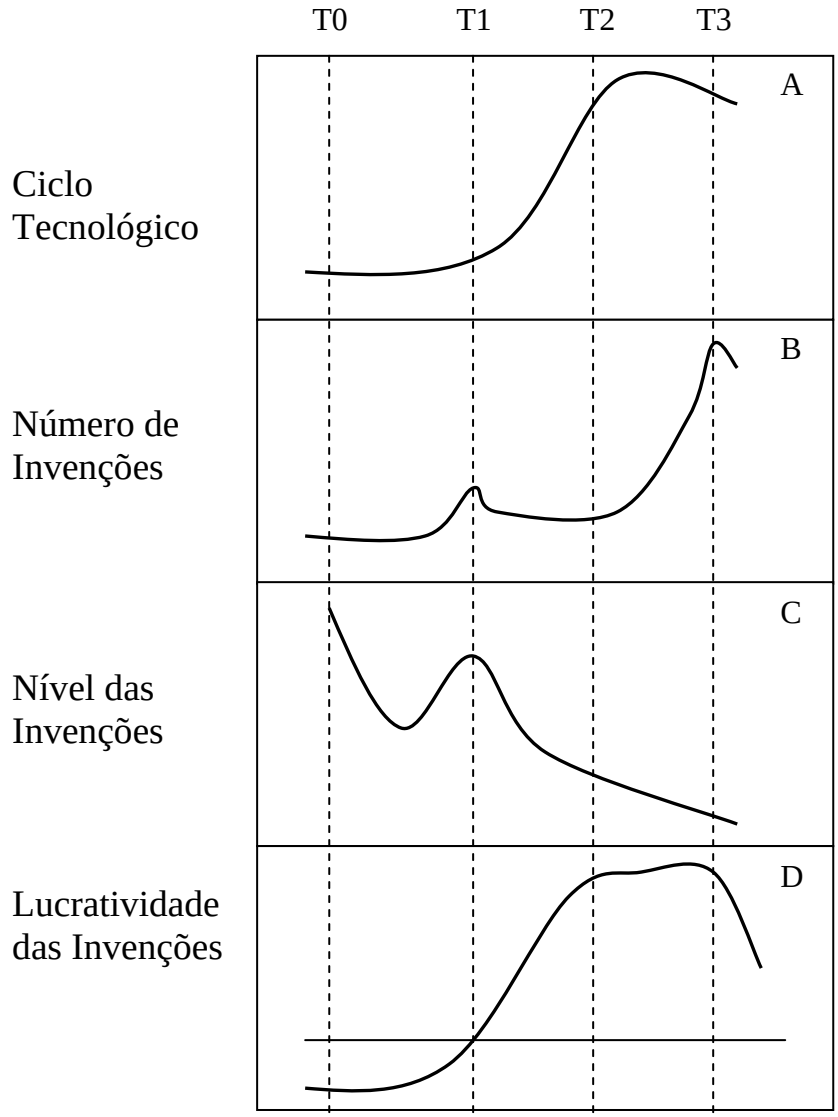


Figura 9 - Padrões das patentes dentro do ciclo tecnológico.
Fonte: Adaptado de Slocum, 1999; de Carvalho, 2007.

Na adolescência (ou crescimento), espaço de tempo de T1 até T2 inicia-se o uso comercial da tecnologia. Vê-se isso com o início da lucratividade, em D. Nesse estágio e o sistema tecnológico original, que é concorrente da tecnologia adolescente, tem quase todos os seus recursos esgotados. Os níveis de invenção, como os que propiciaram a utilização comercial da tecnologia são difíceis de acontecer novamente. Porém, foram suficientes para resolver muitos dos problemas técnicos. O capital humano que estava envolvido nas pesquisas começa trabalhar no desenvolvimento de produtos. Isso explica a queda do nível das

invenções em C e a diminuição do número de patentes em B (SAVRANSKY, 2001; SLOCUM, 1999).

Na maturidade tecnológica, espaço de tempo entre T2 e T3 é onde as invenções deixam de ser científicas para estarem relacionadas à diminuição dos custos da produção. O pequeno vale em B, que representa um mínimo local do número de patentes depositadas representa essa troca. Os lucros estão no estágio máximo e seu impactos social e econômico são alto. O nível das invenções diminui rapidamente (em C), pois os processos para a diminuição de custos podem ser utilizados de outros ciclos tecnológicos. Logo os investimentos financeiros em pesquisa já não apresentam mais o retorno esperado (SAVRANSKY, 2001; SLOCUM, 1999).

A maturidade tecnológica é caracterizada pela substituição da tecnologia dominante. A partir desde ponto, o número de invenções começa a aumentar como nunca antes (em B), pois neste sistema tecnológico há lucratividade, o que está em decadência no sistema dominante.

O ocaso de uma tecnologia acontece quando ela está no auge de seu desempenho e os limites da técnica foram alcançados. Os conhecimentos da área já estão exauridos e pode-se considerar o conhecimento dessa área como ubíquo. Essa situação está representada para tempo > T3. No tempo > T3 a tecnologia tem duas opções, ou morre devido à substituição pela tecnologia mais nova, ou então se estabiliza em nichos mercadológicos específicos, sendo levada por inovações de baixo nível (SAVRANSKY, 2001; SLOCUM, 1999).

2.4.5 Sistemas Comerciais para Mineração de Texto e Visualização

Atualmente existem muitas ferramentas que realizam a mineração de texto e a apresentação gráfica dos resultados. Elas visam encontrar “pepitas” de informações relevantes sobre o desenvolvimento tecnológico. O desenvolvimento dessas ferramentas parte do pressuposto de que é impossível ler e digerir toda informação disponível sobre as diferentes tecnologias. Elas utilizam análise de palavras-chave, estatísticas e análise lingüística para relacionar textos (algumas vezes em diferentes idiomas) e extrair respostas às diferentes questões empresariais.

Yang *et al.* (2008) apresentam uma coletânea de ferramentas que podem ser utilizadas para esse fim. Em um trabalho conjunto com o grupo de análise de patentes da Bristol-Myers Squibb (BMS), eles apresentam suas impressões sobre cada ferramenta

avaliada. O estudo é apresentado como uma visão geral superficial sobre cada ferramenta. Cada sistema foi avaliado através de demonstrações realizadas pelo representante comercial de cada fornecedor. Os principais pontos avaliados foram: tipo de ferramenta, funções, fontes de dados, resultados obtidos, pontos positivos e pontos negativos. O resultado está sumarizado no Quadro 3.

Os sistemas analisados foram divididos em três grupos: (Grupo 1) sistemas mais flexíveis, que funcionam com texto não-estruturado; (Grupo 2) sistemas que funcionam melhor com texto estruturado; e (Grupo 3) sistemas que trabalham com textos híbridos, que são focados em patentes.

Proprietário	Tipo de Ferramenta	Método	Fonte dos Dados	Saídas	Usuários em potencial
Ferramentas do Grupo 1					
ClearForest TextAnalytics	Mineração de texto	Análise semântica/ *Natural Language Processing* (NLP)	Texto estruturado e não estruturado da web, documentos internos, patentes e etc.	Entidades de dados estruturados, ferramentas de visualização – gráficos de tendência, mapas de categorias	Inteligência empresarial (*Business Intelligence*)
Goldfire Innovator	Mineração de texto	Análise Semântica	Texto não estruturado de dados pessoais, dados corporativos, web, patentes, etc.	Resumo e categorização	Cientistas de P&D
Inxight Smart Discovery	Mineração de texto	NLP, extrações contextuais	Texto não estruturado da web, repositórios internos;	Categorização hierárquica	P&D de informação
Omniviz	Texto baseado em visual/ Mineração de dados	Análise estatística	Texto estruturado e não estruturado, numérico, estruturas químicas	Mapas interativos de visualização	Cientistas de P&D
TEMIS	Mineração de Texto	NLP	Texto estruturado e não estruturado da web, documentos internos, patentes, testes médicos, email, bioinformática, etc	Agrupamentos, listas, rankings	Inteligência empresarial, Cientistas de P&D
Ferramentas do Grupo 2					
Quosa	Mineração de texto baseado em extração de conceito/ agrupamento	Análise estatística	Texto estruturado e não estruturado de patentes, documentos internos, PubMed, Google Scholar, Ovid	Coleção de documentos organizados, compartilhamento de time, anotações	Cientistas de P&D
RefViz	Análise textual e visualização de dados	Análise estatística e lingüística	Textos estruturados do ISI Web, PubMed, OCLC	Galáxia e matriz de visualização	Cientistas de P&D, profissionais de informação
STN AnaVist	Provedor de banco de dados e análise textual	Análise estatística	Texto estruturado – CAPLUS, USPatful, DWPI	Quadros, mapas de horizontes de pesquisa	Profissionais da informação / inteligência empresarial, Cientistas de P&D
VantagePoint	Mineração de texto	NLP, procura por	Texto estruturado dentro dos campos de	Listas, sumários, gráficos, mapas,	Profissionais da informação

Proprietário	Tipo de Ferramenta	Método	Fonte dos Dados	Saídas	Usuários em potencial
		padrões	bibliografia	matrizes, quadros	/inteligência empresarial
Thomson Data Analuzer	Mineração de texto	Procura por padrões, NLP	Texto estruturado dos campos bibliográficos	Listas, sumários, quadros, mapas, matrizes, gráficos	Profissionais da informação /inteligência empresarial
Ferramentas do Grupo 3					
Aureka	Mineração de texto e provedor de BD	Palavras chave e análise estatística	Database de patentes MicroPatent	Mapas ThemeScape, árvores de citação hiperbólicas, cluster de texto	Dept. de patentes, Cientistas de P&D, profissionais da informação , planejamento estratégico, inteligência empresarial
M-Cam Doors	Provedor de BD e mineração de texto	Análise semântica lingüística, multi-lingual	Patentes de 88 autoridades de patenteamento e *journals*	Exibição de citações "Compass", visualizador "Magellan", análise competitiva, análise de risco para merge ou aquisição de ações, análise de exclusividade de patente.	Inteligência competitiva, Departamento legal ou patentes, profissional de informação.
Wisdomain	Provedor de BD e mineração de texto	Análise de palavras-chave, mapa de citações.	*Abstracts* de patentes da US, EP, PCT, PAJ, Chinesa e Coreana, Inpadoc legal status.	Árvores genealógicas, mapas de citações, tabelas, quadros	Cientistas de P&D, profissionais de informação
PatAnalyst	Provedor de BD	Não realiza mineração de texto.	Patentes do US, EPO, PCT, PAJ, UK, Alemanha, França e Suíça.	Visualizador com organizador de conjuntos de patentes, palavras chave sublinhadas.	Profissionais de informação , Cientistas de P&D.

Quadro 3 - Empresas de mineração de texto e visualização de dados.
Fonte: Adaptado de Yang *et al.*, 2008.

A conclusão à qual chegaram os autores é que esses sistemas de mineração de dados teriam um resultado melhor caso fossem utilizados em conjunto, pois nenhum deles têm, sozinho, todas as características necessárias para a realização de todas as tarefas de mineração e visualização de texto. Cada ferramenta tem uma aplicação específica. Outra consideração importante apresentada em Yang *et al.* (2008) é de que essas ferramentas têm o potencial de gerar uma revolução na inovação tecnológica, pois podem tratar grandes quantidades de texto e fornecer as informações úteis para o andamento das inovações e a gestão da tecnologia.

2.5 CONSIDERAÇÕES DA REVISÃO TEÓRICA

O quadro teórico articulado nesse capítulo é importante para estruturação, orientação e delimitação do sistema proposto, que será melhor discutido no capítulo subseqüente. Em resumo, o quadro teórico é voltado para a pesquisa sobre empresas que desejam realizar desenvolvimento conjunto. Espera-se que os procedimentos de Mineração de Tecnologia possam facilitar a comparação entre empresas que realizam inovações tecnológicas, através da análise da complementaridade entre elas.

Conduziu-se o quadro teórico de inovação até a inovação tecnológica, então, especializou-o em Inovação Aberta, que é o foco dessa pesquisa, pois essa teoria fornece os subsídios teóricos para compreender o decurso do conhecimento e dos arranjos contratuais entre duas empresas com o propósito de gerar dividendos para ambas. As alternativas apresentadas para as empresas pela Inovação Aberta são: a diminuição dos custos, ou melhorias nos aspectos mercadológicos no processo de desenvolvimento de novas tecnologias.

Tendo o fomento da Inovação Aberta como alvo, o processo utilizado para esse fim é a Mineração de Tecnologia. Como apresentado anteriormente, há maneiras de realizar uma pesquisa macro da tecnologia utilizando outro ferramental, como a bibliometria, ou também as ferramentas computacionais comerciais já existentes para a mineração de texto e visualização. Porém, esse ferramental apresenta-se pouco proveitoso, aquele por não aproveitar a informação textual não estruturada de suas fontes e este por ser comercial e com metodologia desconhecida, o que dificulta sua manipulação para a realização de uma pesquisa mais especializada. A teoria de Mineração de Tecnologia apresentou-se satisfatória para o que se dispõe nessa pesquisa, pois oferece a manipulação de uma quantia vultosa de dados e a maleabilidade de um sistema que pode ser customizado para atender quaisquer tarefas, mesmo as mais especializadas.

As ferramentas comerciais que realizam a extração de informação através de bases de patente estão limitadas dentro do escopo das buscas que se dispõe realizar. Isso encerra o usuário até os limites das ferramentas. A opinião de Yang *et al.* (2008) é de que essas ferramentas podem oferecer um melhor resultado se forem utilizadas em conjunto. Isso mostra que há horizontes de conhecimento escondido que não são alcançados, o que abre mercado para o estudo da Inteligência Artificial aplicado a informações tecnológicas. A apresentação dessas ferramentas ressalta a importância desta pesquisa, pois revela que o

know-how da manipulação dessas importantes informações está confinado nas mãos de poucas empresas.

A matéria-prima para o processo de Mineração de Tecnologia são os registros de patente. Como visto anteriormente, estes documentos não são as únicas fontes de informação sobre inovação, porém são uma rica fonte de dados, pois podem fornecer desde dados para uma visão ampla como para uma visão micro. A publicação das patentes faz com que o conhecimento seja disseminado, o que favorece a inovação.

A TRIZ foi uma teoria que surgiu a partir da análise minuciosa de patentes, e pode mostrar padrões da evolução de uma tecnologia. O conhecimento desses padrões é chave para o investimento financeiro em tecnologia. Apesar desse não ser o foco da pesquisa, os padrões de evolução tecnológica fornecidos pela TRIZ fornecem um embasamento para a análise dos outros padrões que serão gerados por meio do processo de Mineração de Tecnologia. Além de que sua teoria complementa os conhecimentos sobre patenteamento e inovação.

Outra teoria apresentada foi a complementaridade, pois fornece um ponto de julgamento não subjetivo para determinar se os conhecimentos de diferentes empresas podem ser considerados complementares.

O referencial teórico levantado para este capítulo criou o ambiente necessário para a ideação da metodologia. No próximo capítulo os conceitos apresentados neste capítulo são interligados com os procedimentos técnicos que resultam no objetivo geral desta pesquisa.

3 METODOLOGIA

Este capítulo trata da metodologia utilizada no desenvolvimento da pesquisa. Toda informação aqui apresentada deve ser ordenada e com a melhor exatidão possível, para poder promover a repetibilidade da pesquisa. Esta pesquisa baseia-se em procedimentos técnicos para aquisição de conhecimentos que podem ajudar na tomada de decisão a nível empresarial. Os argumentos gerenciais relevantes para a pesquisa foram apresentados no capítulo anterior, entretanto o ferramental utilizado para alcançar o objetivo desta pesquisa foi desenvolvido tecnicamente e está articulado neste capítulo.

Esta metodologia tem dois objetivos, o primeiro é apresentar o modo como serão buscados os dados para a realização dos testes e o segundo é o modo como esses dados serão tratados para, talvez, resultarem em informação útil para o processo de procura por parcerias complementares em P&D. Como a parte técnica esta pesquisa é um procedimento de KDD escolheu-se o modelo de Fayyad *et al.* (1996) para o representar. Durante da descrição da metodologia serão citados os passos do modelo de KDD para a melhor orientação do leitor. O modelo de KDD de Fayyad foi descrito na subseção 2.4.3.1 (p. 41). O modelo descreve a seqüência de procedimentos aplicados aos dados no processo de busca de conhecimento e ele engloba os dois objetivos pretendidos pela metodologia.

O texto da metodologia está dividido em quatro partes, a saber: a seção 3.1 é destinada à conceituação técnica dos principais pontos teóricos que são essenciais para a compreensão das decisões tomadas nos procedimentos, porém não é parte integrante da argumentação central a qual se trata esta dissertação. A seção 3.2 apresenta as ferramentas utilizadas durante a pesquisa, sua validade e sua relevância para este trabalho. A seção 3.3 apresenta os procedimentos para alcançar os dois objetivos da metodologia. Por fim, a seção 3.4 traz as considerações da metodologia, entremeando os principais pontos expostos nos diferentes tópicos e tentando extrair suas melhores características.

3.1 CONCEITUAÇÃO TÉCNICA

Esta seção destina-se apresentar o conteúdo teórico necessário para o entendimento dos procedimentos metodológicos. O conteúdo refere-se ao tratamento dos dados que são matéria prima desta pesquisa: as patentes. A Conceituação Técnica aborda os parâmetros selecionados dentro das patentes, a preparação dos dados e as técnicas

matemáticas utilizadas na busca de conhecimento escondido em padrões dentro de seus parâmetros. Partindo dos registros de patente, são extraídos os dois tipos principais de informação, os IPCs e o texto de descrição da patente.

3.1.1 Parâmetro IPC - Classificação Internacional de Patentes

Cada registro de patente pertence a uma ou mais áreas técnicas. A Classificação Internacional de Patentes (IPC) é um símbolo de caracteres alfa-numérico que representa uma área técnica específica. Assim, geralmente é necessário que cada patente contenha vários IPCs para ser classificada corretamente. O IPC foi desenvolvido para prover uma forma uniforme de classificação dos registros e estabelecer uma ferramenta efetiva para buscas de patentes em diferentes bases. Os principais propósitos da classificação (WORLD ...,2009) são:

- fornecer uma estrutura organizada para facilitar a busca de informações tecnológicas e de estado legal;
- servir de base para a disseminação efetiva de informação para todos os usuários de patente;
- permitir uma base para a investigação do estado da arte nos diferentes campos tecnológicos;
- ser utilizada como base para estatísticas e avaliações sobre o desenvolvimento tecnológico em diversas áreas.

Existem cerca de 70 mil IPCs listados. O símbolo de classificação é montado a partir de uma combinação de símbolos alfa-numéricos, em que a primeira letra denota a seção IPC (*e.g.* H), os dois dígitos seguintes significam a classe IPC (*e.g.* H04), a letra seguinte denota a subclasse (*e.g.* H04L). O número, que pode variar entre um e três dígitos denota o grupo principal (*e.g.* H04L12). Finalmente, o subgrupo do IPC é dentado por uma barra (“/”) seguida de um a três números.

A apresentação do IPC versão 9 nos documentos de patente segue os seguintes padrões:

- símbolos IPC avançados: itálico;
- símbolos IPC centrais: não-itálicos;
- símbolos relacionados à invenção: negrito;
- símbolos adicionais (não relacionados à invenção): não-negrito.

3.1.2 Representação dos Dados

Para o tratamento das patentes selecionadas devem-se dispor suas informações, tanto de IPCs como o texto da descrição, de uma forma matemática. Essa forma são vetores *m*-dimensionais. Neles, cada atributo do conjunto de dados, seja cada IPC da patente ou cada termo contido no texto, é uma dimensão. Quando a mineração de dados é aplicada a textos o processo é conhecido como Mineração de Textos. Os algoritmos de mineração de dados utilizam esses conjuntos de vetores para, por exemplo, calcular a distância vetorial entre um texto e outro, ou a similaridade entre eles.

A Tabela 1 exemplifica o resultado que se espera da preparação dos dados de três patentes (A, B e C). Cada patente é um vetor e suas dimensões são cada uma de suas palavras, cada palavra recebe um "peso" segundo sua importância dentro do conjunto das palavras da própria patente e do conjunto de patentes. Para o tratamento dos IPCs a tabela gerada é similar à Tabela 1, porém o vetor é binário, mostrando a existência ou não de um IPC na patente.

Tabela 1 - Exemplo de representação do texto de três patentes.
Fonte: Autoria Própria.

Patente (vetor)	**Palavras existentes nos textos das patentes (dimensões)**				
	eixo	transmi	resist	. . .	triangul
A	0,01	0	0,201		0,22
B	0	0	1		0,50
C	0,4	0,75	0,12		0,032

Para melhor compreensão desta seção é útil visualizar o processo de tratamento dos dados representado pela Figura 10. Ela apresenta no lado esquerdo as patentes ainda não tratadas, os dois parâmetros que são extraídos delas são os IPCs e o texto do campo de descrição. Na parte central há os procedimentos, sumarizados como Preparação dos Dados, que são utilizados para chegar às tabelas (tecnicamente chamadas de matriz de vetores) que são utilizadas na mineração de dados. A mineração será realizada por duas tarefas diferentes, a saber: Associação e Agrupamento, que serão apresentadas nas subseções subseqüentes.

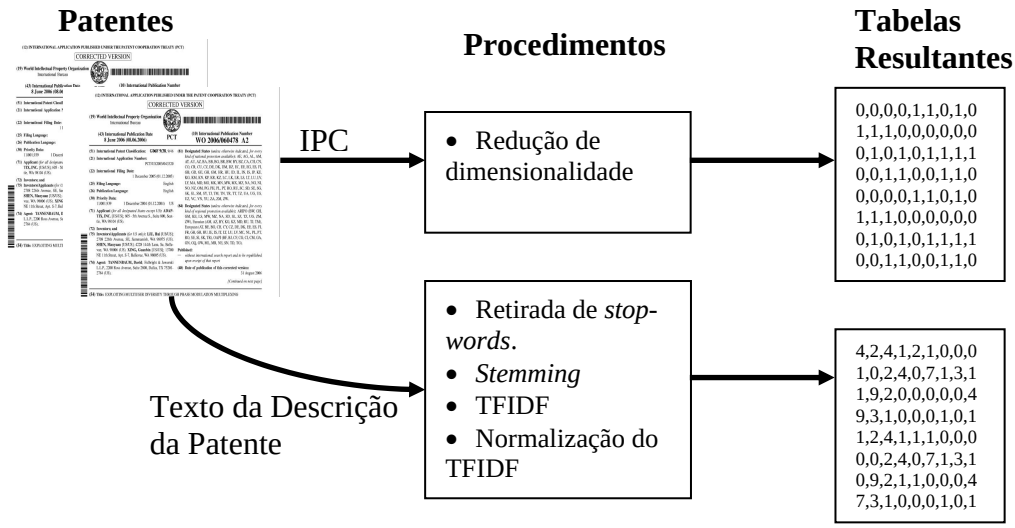


Figura 10 - Exemplificação da aquisição dos dados e tratamento.
Fonte: Autoria própria.

3.1.3 Preparação dos Dados

Nessa seção são abordados todos os procedimentos que os dados sofrem antes de entrarem na tarefa de Mineração de Dados. Os procedimentos abordados englobam o terceiro e quarto passos do modelo de Fayyad *et al.* (1996). Os passos são chamados de “tratamento dos dados e pré-processamento” e “poda dos dados e projeções”, respectivamente. Agora em diante o procedimento que engloba ambos os passos será chamado de “preparação dos dados”.

O passo do KDD denominado **tratamento dos dados e pré-processamento** envolve a remoção de dados errados, a manipulação dos dados com ruído e valores faltantes nos dados. Para o tratamento dos dados descritos por esta metodologia (subseção 3.3.3) foram retirados os registros de patentes publicadas em língua alemã e francesa, como será descrito na subseção 4.2.3.1. Sendo o patenteamento um processo ordenado de publicação de registros, isso reduz os dados faltantes e errados. O que torna desnecessário os dados manipulados passarem pelo processo de “tratamento dos dados e pré-processamento”.

A **poda dos dados e projeções** consiste em encontrar atributos úteis e a aplicação de redução de dimensionalidade, também sendo possível a utilização de métodos de transformação de dados. Busca-se também uma representação invariável dos dados, como a normatização (CIOS *et al.*, 2007).

Como são utilizadas duas técnicas de mineração de dados, o processo de poda e diminuição de dimensionalidade deverá atingir dois objetivos. Um objetivo é a diminuição de

dimensionalidade dos atributos IPC para a tarefa de Associação. Outro objetivo é o tratamento do texto das patentes para a realização da tarefa de Agrupamento.

A união dos vetores *m*-dimensionais resulta em matrizes. A redução da dimensão das matrizes, que são tratadas computacionalmente, é necessária pois o esforço computacional envolvido no tratamento de dados com dimensões maiores tende a ser maior. A redução de dimensões visa diminuir o número de dados a serem computados, porém mantendo a mesma qualidade, ou a mais próxima o possível, no resultado da mineração. Para esse processo há técnicas de tratamento textual, as quais serão apresentadas a seguir: retirada das *stop-words* e extração dos radicais das palavras, que também é conhecido como processo de *stemming*. A técnica de TFIDF atribui "pesos" a cada palavra, como será explicado subseqüentemente.

3.1.3.1 Processo de Retirada de *Stop-Words*

A retirada de *stop-words* é um processo que visa reduzir a dimensionalidade da matriz de vetores, que representa os textos. As *stop-words* são palavras que têm baixa ou nenhuma informação, como artigos, conjunções, preposições, o verbo 'ser', o verbo 'estar' entre outras. Essas palavras estão presentes em qualquer texto, por isso não são determinantes para a diferenciação dele.

Não é redundante lembrar que as *stop-words* podem variar para cada língua. A lista das *stop-words* utilizadas nessa pesquisa está contida no Anexo A. A fonte dessas palavras é a biblioteca **Lingua::StopWords** da linguagem de *script* Perl, e pode ser encontrada no sítio CPAN (CPAN, 2009). Essa biblioteca pode ser encontrada em diferentes línguas, como: inglês, português, alemão, francês, holandês, russo, espanhol, italiano, entre outras.

3.1.3.2 Processo de *Stemming*

O *Stemming* é uma técnica para redução de dimensionalidade que faz cada palavra do texto ser reduzida à sua raiz. Isso possibilita contar os verbos independentemente de sua conjugação e substantivos contados independente de gênero ou número.

O algoritmo de *stemming* mais conhecido, e que foi utilizado nesta pesquisa, é o Stemmer de Porter (PORTER, 1980). Ele é um esquema de regras que processa as palavras mais comuns na língua inglesa. Porém, esse algoritmo pode ser encontrado para diversos

idiomas. Versões já compiladas em bibliotecas executáveis podem ser encontradas para a linguagem de *script* Perl no sítio CPAN (CPAN, 2009) como **Lingua::Stem**.

A implementação do algoritmo de Porter utilizada nessa pesquisa pode ser encontrada tanto para Perl, na referência acima, quanto para C# no sítio Tartarus (2008).

3.1.3.3 Freqüência do Termo e Freqüência Inversa do Documento

A forma mais fácil de codificar um texto em um vetor *m*-dimensional seria em forma binária. Isto é, caso determinada palavra exista no texto o vetor que a representa tem valor um, caso não exista seu valor é zero. Entretanto, este sistema não confere informações relevantes que podem ajudar no tratamento desses vetores. Como, por exemplo, a freqüência que determinada palavra aparece no texto ou sua relevância dentro do conjunto de dados utilizado.

A Freqüência do Termo e Freqüência Inversa do Documento (*Term Frequency and Inverse Document Frequency*) (TFIDF) confere a cada palavra um "peso" em relação a todas as palavras utilizadas em todos os textos do conjunto de dados. Ela pode ser definida como:

$$tfidf(th, dj) = \#(tk, dj).\log\frac{|Tr|}{\#Tr(tk)}, \tag{2}$$

onde #(th,dj) é a freqüência do termo no em seu documento de origem; |Tr| é o tamanho total da coleção de documentos; e #Tr(tk) é o número de documentos onde o termo ocorre. Analisando a fórmula chega-se a duas conclusões: (i) quanto mais freqüente um termo é em um documento, mais representativo ele é nesse documento; (ii) quanto mais um termo ocorre na coleção de documentos, menos discriminatório ele se torna (SEBASTIANI, 2002).

Para que o processo de atribuir "pesos" aos termos dos documentos limite-se a valores contínuos no intervalo entre zero e um é necessário realizar a normalização do valor do TFIDF. Para isso pode-se utilizar a normalização cosseno data por:

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|}(tfidf(t_s, d_j))^2}} \tag{3}$$

3.1.4 Tarefas de Mineração de Dados

Esta subseção apresenta três passos do modelo de KDD de Fayyad *et al.* (1996), a quinta, sexta e sétima. São elas: “escolher a tarefa de mineração”, “escolher o algoritmo mais apropriado” e “minerar os dados”. Na apresentação das tarefas de mineração de dados é descrito um resumo de cada tarefa. Como cada tarefa tem inúmeras implementações, em diferentes algoritmos, serão apresentadas algumas características que podem as diferenciar, como base para o critério de seleção na pesquisa. Para cada tarefa de mineração de dados exposta são apresentados os algoritmos específicos utilizados.

3.1.4.1 Tarefa de Associação

A associação é útil para descobrir relacionamentos interessantes escondidos entre os atributos de grandes quantidades de dados.

Os relacionamentos que a associação apresenta são expressos por meio de regras. Elas são representadas da forma $X \rightarrow Y$, que se lê “se X então Y”[5], onde X e Y são mutuamente disjuntos, isto é $X \cap Y = \phi$. As regras apresentadas pela associação não devem ser interpretados como causalidade, elas apresentam uma co-ocorrência entre itens. Seja o conjunto de todos os IPCs utilizados pelas patentes em uma pesquisa $I = \{i_1, i_2, ..., i_d\}$ e $T = \{t_1, t_2, ..., t_N\}$ o conjunto das vezes que esses IPCs estão relacionados em patentes distintas, e *N* o número total de relacionamentos, então a contagem do suporte de todas as patentes será:

$$\sigma(X) = \left| \{ t_i \mid X \subseteq t_i, t_i \in T \} \right| \quad (4)$$

A qualidade de uma regra de inferência pode ser medida por meio do Suporte e da Confiança. O Suporte (*support*) é a freqüência com que uma regra é aplicada, isto é, quão freqüente a regra $X \rightarrow Y$ está dentro do conjunto de dados. A Confiança (*confidence*) determina qual freqüência um item em *Y* aparece com *X*.

[5] No caso específico desta pesquisa se lê “se o conhecimento técnico X é existente na patente então o conhecimento técnico Y também será”.

Tanto X quanto Y são representações de **conjunto de itens** (*itemset*), que são a coleção de um ou mais itens. Por exemplo, na análise de compras de supermercado pode-se ter: {manteiga, achocolatado, geléia} → {leite, pão}.

$$\text{Confiança, } c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{5}$$

$$\text{Suporte, } s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N} \tag{6}$$

O total de regras possíveis de serem inferidas por um sistema de associação é $R = 3^d - 2^{d+1} + 1$, para 2^d conjunto de itens. Para viabilizar a análise dos dados deve-se limitar o Suporte mínimo e a Confiança mínima. Isso permite a diminuição do número de regras geradas pelo sistema. Esta pesquisa também utilizou a limitação do Suporte máximo, pois as regras geradas por Associação com Suporte alto são banais, e essa técnica auxilia na busca de relações menos triviais dos itens (TAN, STEINBACH e KUMAR, 2006).

Algoritmo Apriori

Caso uma tarefa de associação fosse realizada pelo método de força bruta[6], seu esforço computacional a faria inviável. Um conjunto de dados que contém *k* elementos pode gerar $2^k - 1$ *itemsets*, excluindo o *itemset* vazio. Esses *itemsets* são combinados entre si para gerar as regras de inferência, que, como visto na seção anterior, chega ao número de $3^k - 2^{2k+1} + 1$. A Figura 11 representa a associação entre quatro itens, a saber: A, B, C e D.

Para viabilizar essa tarefa, o algoritmo Apriori utiliza um teorema conhecido como Princípio Apriori, que diz: Se um *itemset* é freqüente, então todos os seus *subsets* são freqüentes (TAN, STEINBACH e KUMAR, 2006).

O princípio apriori se mantém devido à propriedade conhecida como anti-monótona, que é percebido nas medidas de Suporte: $\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$. Isso significa que o Suporte de um conjunto de itens nunca excede o Suporte de seus subconjuntos. Isso é traduzido em uma poda antecipada de regras com *itemset* mais complexos, quando *itemsets* mais básicos, com a mesma composição de alguns itens já não tem Suporte aceitável.

[6] Chama-se de "força bruta" o método que utiliza pouca inteligência computacional e gera todas as combinações possíveis de uma determinada tarefa até achar uma opção que fornece um resultado que é melhor do que os outros.

O algoritmo Apriori utiliza dados discretos para sua realização, caso os dados sejam de fontes contínuas, há a necessidade de discretização para a utilização do método.

Nesta pesquisa utilizar-se-á o algoritmo Apriori, com os dados de entrada sendo os campos IPCs dos registros de patente. E seu objetivo é tentar associar os IPCs mais utilizados dentro de uma área tecnológica, buscando por interdependência entre os diferentes conhecimentos tecnológicos.

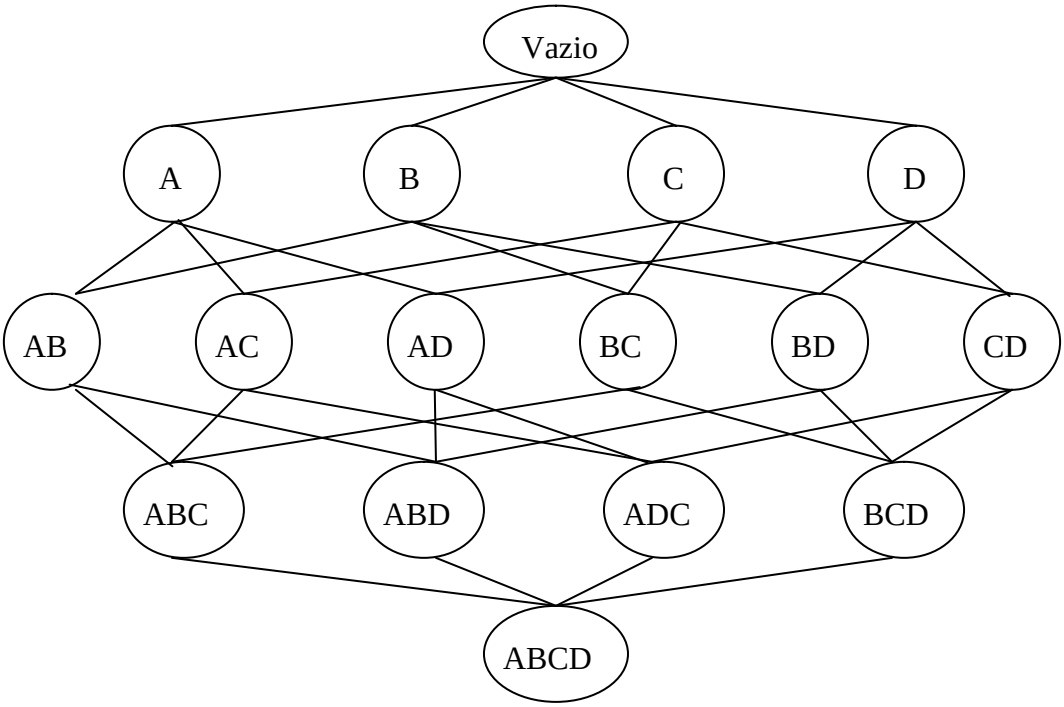


Figura 11 – Associação de quatro itens.
Fonte: Adaptado de Tan, Steinbach e Kumar, 2006.

3.1.4.2 Tarefa de Agrupamento

O agrupamento separa em grupos relevantes os dados fornecidos para análise, levando em consideração as relações entre esses grupos. Ele também é denominado como uma aprendizagem não supervisionada, pois realiza a classificação sem haver definição prévia de quais características comuns delimitarão os grupos. O objetivo de um agrupamento é que objetos colocados dentro de um determinado grupo tenham características similares entre si, que não são encontradas em outros grupos. Deseja-se que quão maiores as diferenças entre os grupos, maiores serão as semelhanças entre um objeto e outro do mesmo grupo. A Figura 12 mostra uma representação bidimensional de um agrupamento, em que três classes distintas são delimitadas.

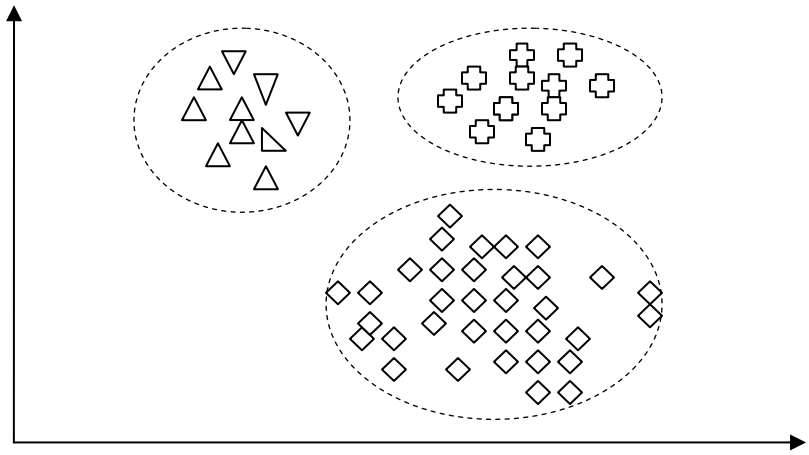


Figura 12 - Representação bidimensional de agrupamento.
Fonte: Adaptado de Tan, Steinbach e Kumar, 2006.

Agrupamento tem utilidade em várias áreas, como: segmentação de imagens, reconhecimento de imagens, busca de informação (*information retrieval*), sumarização de textos, previsão do tempo, psicologia e medicina, entre outras (TAN, STEINBACH e KUMAR, 2006).

O Quadro 4 sumariza as características que podem diferenciar os algoritmos de agrupamento. Essas características ajudam na escolha do melhor algoritmo para cada conjunto de dados (JAIN, MURTY e FLYNN, 1999).

Tipo de Algoritmo	**Características**
Hierárquico Aglomerativo	Toma por pré-suposto que cada elemento é um agrupamento, porém a cada passo do processo há a junção de cada agrupamento menor em um maior.
Hierárquico Divisivo	Toma por pré-suposto que o conjunto de todos os elementos é um agrupamento, porém a cada passo do processo há a divisão desse agrupamento em agrupamentos menores.
Hard	Cada elemento é atribuído a somente um único grupo.
Fuzzy	Cada elemento tem graus de pertinência em cada grupo.
Sobreposto	Um objeto pode pertencer a vários grupos dentro de um data-set.
Determinístico	Técnicas tradicionais baseadas em buscas exaustivas.
Estocástico	Buscas aleatórias no espaço de todos os dados.

Quadro 4 - Sumário das características encontradas nos algoritmos de agrupamento.
Fonte: Adaptado de Tan, Steinbach e Kumar, 2006; Jain, Murty e Flynn, 1999.

Algoritmo *K*-Médio

É uma técnica de agrupamento que tenta encontrar *k* grupos, onde *k* é um número escolhido pelo usuário. Cada grupo é representado por seu centróide, que é a média do grupo de pontos dispostos em um espaço *n*-dimensional. Na prática o centróide é um dos elementos do grupo que representa o seu centro. O *K*-médio é o algoritmo de agrupamento mais velho e mais amplamente utilizado. O funcionamento do *K*-médio está descrito no Quadro 5.

1: Escolha *K* pontos como centróides iniciais
2: **repita**
3: Associe cada ponto ao grupo cujo centróide é o mais próximo
4: Recalcule o centróide de cada grupo
5: **até** não se alterarem os centróides

Quadro 5 - Funcionamento do K-médio.
Fonte: Adaptado de Tan, Steinbach e Kumar, 2006.

A escolha dos centróides iniciais pode ser realizada de diversas maneiras, dentre as quais são destacadas:

- Escolha aleatória;
- escolha aleatória com várias execuções;
- definir os *K* primeiros exemplos como centróides;
- amostragem e agrupamento hierárquico;
- *K*-médio bi-seccional

A escolha aleatória leva a diferentes resultados de grupos cada vez que o algoritmo é executado. Ela mostra-se fraca na distribuição dos grupos. Uma maneira de melhorar os grupos é rodar múltiplas vezes o algoritmo e medir a Soma do Erro Quadrático (SEQ). A rodada que tiver o menos SEQ será a melhor opção. A utilização de amostragem e posterior execução de um agrupamento hierárquico mostram-se capaz de dividir os grupos de maneira satisfatória, no entanto pode ser utilizada quando as amostras são pequenas ou quando *K* é relativamente pequeno comparado às amostras. O *K*-médio bi-seccional tem destaque, pois é menos suscetível às escolhas aleatórias de centróides. Seu funcionamento é simples, para alcançar os *K* grupos o algoritmo parte o conjunto de dados em dois grupos e escolhe um desses grupos para continuar partindo até os *K* grupos serem produzidos.

Para o cálculo do centróide mais próximo pode-se utilizar a distância euclidiana, que é dada por:

$$d_2(x_i, x_j) = (\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2)^{1/2} = \left\| x_i, x_j \right\|_2 \quad (7)$$

Para a função objetivo, utilizando o espaço euclidiano, a medida de qualidade do processo de agrupamento é a Soma do Erro Quadrático:

$$SEQ = \sum_{i=0}^{K} \sum_{x \in Ci} dist(c_i, x)^2 \quad (8)$$

Dados diferentes conjuntos de grupos, pode-se escolher a melhor representação do agrupamento por meio do SEQ.

3.2 FERRAMENTA PARA MINERAÇÃO DE DADOS

A busca e tratamento dos dados para esta pesquisa foram realizadas por *software* específico de implementação própria, porém para a realização da mineração de padrões foi selecionado o *software* Weka na versão 3.6. Ele é uma coletânea de algoritmos para aprendizado de máquina especialmente preparado para a realização de mineração de dados. Ele tem seu formato próprio de arquivos (.arff), que contém todas as informações dos dados de treinamento para os diferentes métodos suportados.

O Weka pode ser adquirido no sítio da Universidade de Waikato na Nova Zelândia (WAIKATO, 2009). Ele foi codificado na plataforma Java, então pode ser instalado sobre qualquer sistema operacional. Sua instalação é de fácil execução e seu uso intuitivo para quem domina os conceitos de KDD. Este *software* é licenciado sob os termos da licença GNU *General Public License* que o libera para utilização em pesquisa científica.

A equipe responsável pelo *software* Weka ganhou o prêmio SIGKDD em 21 de agosto de 2005 em Chicago (SIGKDD, 2009). O SIGKDD *Service Award* é o mais alto reconhecimento por trabalhos na área de mineração de dados e descoberta de conhecimento (SHAPIRO, 2005). O time foi premiado pelo desenvolvimento da ferramenta gratuita e pela documentação gerada para o auxílio de sua utilização. Os principais pontos de apresentados como sucesso do *software*, apontados pela premiação, são: a diversidade de algoritmos para mineração de dados e aprendizado de máquina, a disponibilidade gratuita, a independência de plataforma, utilização amigável, facilidade de criação de *scripts* para seu uso e atualização com os novos algoritmos de mineração de dados. Segundo a premiação, o *software* Weka trouxe contribuição extraordinária (*outstanding contribution*, tradução própria) para o campo da mineração de dados. Do período de 27-04-2000 até 05-06-2009 o sítio sourceforge.net (2009) apresenta que o software foi baixado 1.751.724 vezes.

Figura 13 - Software para Mineração de Dados, Weka.
Fonte: Weka, 2009.

Quando executado o software Weka apresenta uma interface para a escolha da interface gráfica a ser utilizada nas tarefas de mineração de Dados (Figura 13). Essa pesquisa utilizou a interface Explorer, apresentada na Figura 14. Outra maneira de rodar a aplicação é por meio de linhas de comando, onde se aponta o arquivo de entrada e passando os parâmetros de teste como argumentos.

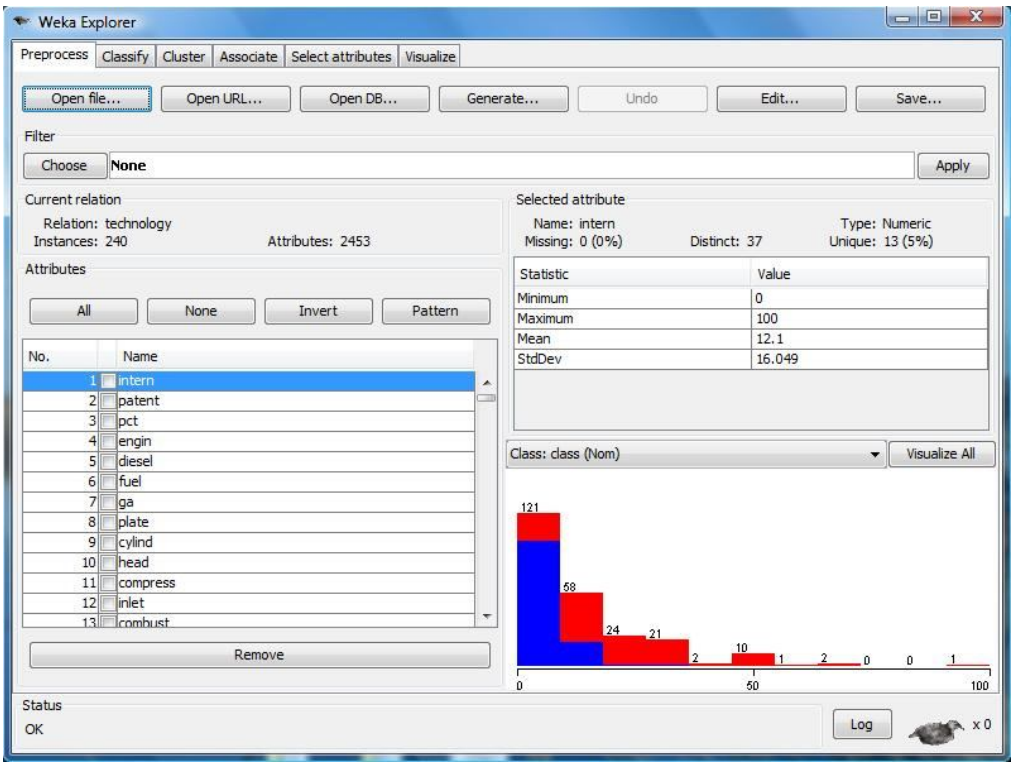


Figura 14 - Weka Explorer.
Fonte: Weka, 2009.

3.3 PROCEDIMENTOS ADOTADOS NA PESQUISA

Esta seção tem o objetivo de descrever os procedimentos utilizados na pesquisa para obtenção de perspectivas de cooperação em inovações tecnológicas entre empresas tendo como ponto de partida seus registros de patente. A primeira subseção inicia com o modelo matemático do problema. A segunda subseção apresenta o resumo dos procedimentos que serão executados durante o desenvolvimento da pesquisa. A terceira subseção apresenta os critérios utilizados na escolha da empresa que servirá de referência para os testes, que patentes serão coletadas para a busca da complementaridade e que base de dados é escolhida para este fim. Por fim, são apresentados os procedimentos para a interpretação dos resultados obtidos dos testes.

3.3.1 Modelo Matemático do Problema

O objetivo deste modelo matemático é demonstrar que o conhecimento tecnológico contido em patentes pode ser utilizado para encontrar indícios de complementaridade entre as P&Ds de empresas de tecnologia. Para tal, procuram-se padrões de interdependência tecnológica nos documentos de patentes. A principal *premissa* é que as patentes podem ser consideradas conjuntos de conhecimentos tecnológicos (ORGANIZATION..., 1994), então, pode-se defini-las como $I = \{t_1, t_2, ..., t_n\}$, onde I é a patente e t os conhecimentos tecnológicos que a compõem.

Uma vez definidas as patentes, as empresas de alta tecnologia, universidades, ou até mesmo cientistas e engenheiros com invenções autônomas podem ser definidos como:

$$E \approx \{\bigcup_{i=k}^{n} I_i\}, \text{ onde } n \text{ é o número de patentes do ator} \qquad (9)$$

Isso é, E é um conjunto de conhecimentos tecnológico resultantes da união das patentes concedidas. Os atores definidos por E (doravante simplificando o termo como “empresas”) são uma aproximação do conjunto de conhecimentos tecnológicos, pois há conhecimentos detidos por eles que não estão protegidos por patentes. As empresas são reduzidas ao conhecimento tecnológico que detém, pois esse é a única característica da empresa que será analisada para encontrar indícios de complementaridade entre os diferentes atores.

A *hipótese* levantada para o desenvolvimento do modelo é que há, no andamento de pesquisas e desenvolvimento tecnológicos, padrões de uso de tecnologias. Uma tecnologia pode ser suporte para o desenvolvimento de outras, ou seja, é necessário um conhecimento tecnológico como base para que outros possam ser construídos sobre ele (DERGINT, 1999). Essa interdependência torna-se evidente quando ocorrem repetidamente essas relações entre tecnologias nas patentes, isso pode ser expresso como:

$$I' = \bigcap_{i=k}^{l} I_i \text{ , onde } \#min \leq l \leq |T| \tag{10}$$

A variável *#min* é o número mínimo de patentes, definidas pelo analisador, que contém o mesmo subconjunto de tecnologias, para considerar *I'* com ocorrência freqüente.

Uma vez que as empresas são definidas, pode-se delimitar o conjunto de empresas que desenvolvem P&D em uma determinada área de pesquisa como $T = \{E_1, E_2, ..., E_n\}$ e o conjunto de padrões de uso das tecnologias como

$$T' = \{I'_i \mid I'_i \subseteq I_x; I'_i \subseteq I_y; ...; I'_i \subseteq I_z\} \text{, sendo } \{I_x, I_y, ..., I_z\} \in T \tag{11}$$

Assume-se que a parceria em complementaridade existe quando a seguinte equação é verdadeira:

$$\exists I'_1 \subseteq E_1; \exists I'_2 \subseteq E_2 \Rightarrow I'_1 \cup I'_2 \in T' \tag{12}$$

Isto significa que existe um subconjunto de tecnologias contidas na Empresa 1, que unida com um subconjunto de tecnologias da Empresa 2 é um subconjunto válido de padrões de uso de tecnologias.

Vê-se que a Equação (12) é compatível com a definição de supermodularidade, Equação (1), relembrando:

$$\Pi(1,1) - \Pi(0,1) \geq \Pi(1,0) - \Pi(0,0)$$

Essa compatibilidade é importante, pois a definição de supermodularidade é um modelo utilizado para a verificação de complementaridade entre empresas, como apresentado na seção 2.2. Como descrito, há complementaridade entre empresas se, somando uma atividade à outra atividade em andamento, resulta em um efeito incremental no desempenho maior que somar cada atividade isoladamente. Nesta pesquisa, caso o conhecimento confinado em I'_1 seja necessário para o desenvolvimento de I'_2, ou vice-versa, pode haver ganhos para a empresa que necessite articular ambos mutuamente e o faça em parceria.

Pode-se ainda afirmar que a complementaridade é a necessidade de um conhecimento, ou tecnologia, existente em outra empresa. A simples presença de um conhecimento exógeno à empresa não pode ser chamada de complementaridade, pois caso

esse não seja necessário para as atividades de P&D ele não se torna complementar. Em resumo, uma empresa somente é complementar à outra quando uma detém o conhecimento que a outra necessita, desde que verificado que possa haver ganhos na produção tecnológica de uma por meio do trabalho colaborativo da outra. Isso é apresentado como resultado da Equação (11), quando foi definido o conjunto *T'* como o conjunto de padrões mais freqüentes de tecnologias utilizadas em P&D, e esses padrões são interpretados como a interdependência de uma tecnologia com outra, ou outras.

3.3.2 Resumo dos Procedimentos

Para melhor compreensão do domínio da aplicação é necessário visualizar, em resumo, todo o processo de KDD. A coleta e análise dos dados são sumarizadas pela Figura 15, que demonstra o processo de como os dados são escolhidos e tratados.

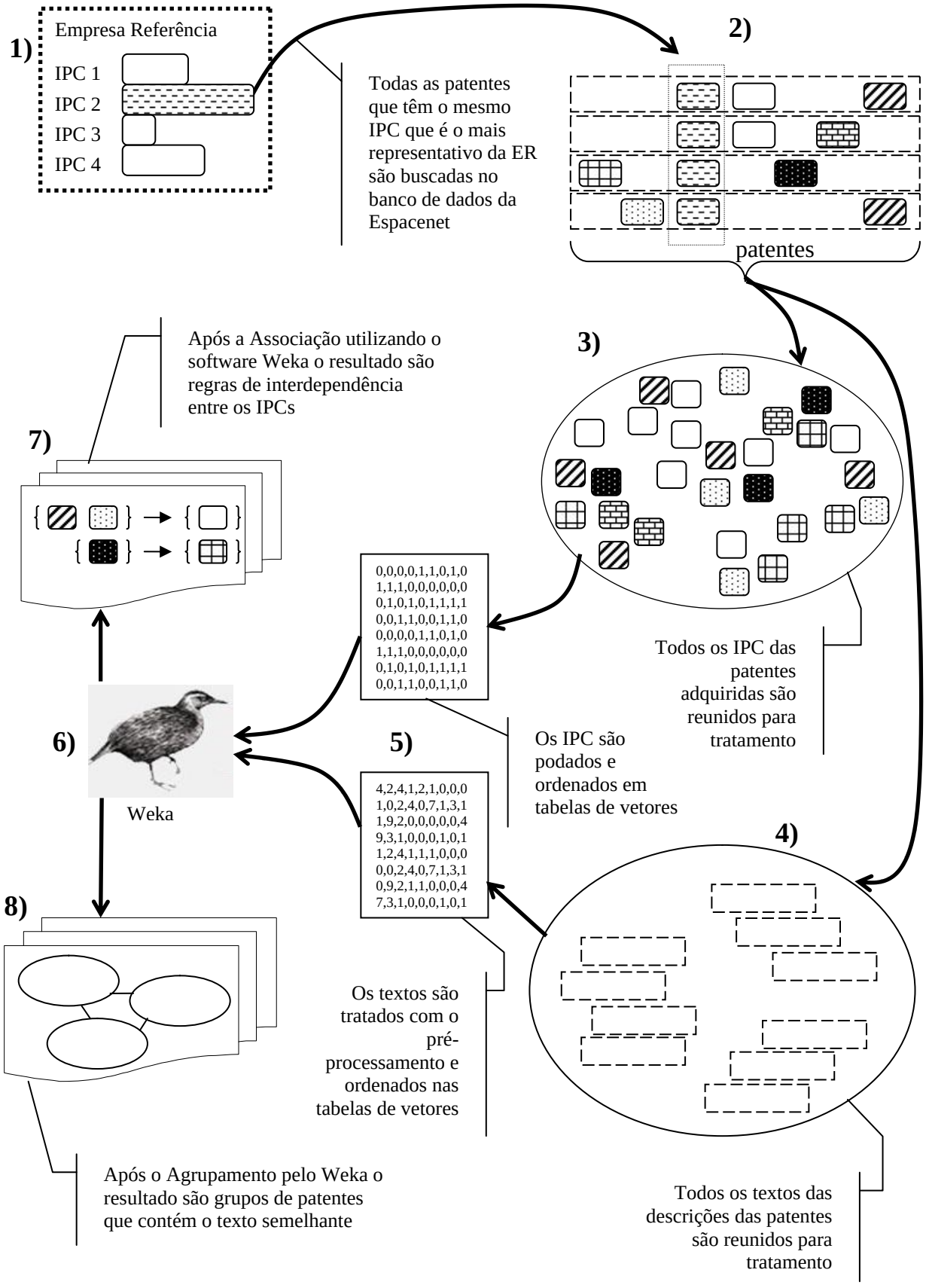


Figura 15 - Resumo do procedimento de coleta e tratamento dos dados.
Fonte: Autoria própria.

Como plano de ação para a pesquisa foi escolhida uma empresa como referência (doravante denominada “Empresa Referência” ou ER) e seu portfólio de patentes foi estudado. Nesse estudo é analisado o campo de IPCs de suas patentes. Essa análise resulta nas informações do bloco 1 da Figura 15, que consiste no conjunto de número de utilizações de um determinado IPC para classificar as patentes da empresa.

O IPC mais freqüente nas patentes da empresa é considerado o mais representativo, então ele é utilizado como parâmetro de busca para as patentes das empresas que têm potencial de colaboração de P&D. Em seguida, é realizada uma busca na base de patentes os documentos que contenham o mesmo IPC, que é o mais representativo na ER. O bloco 2 da Figura 15 representa o conjunto de patentes resultantes dessa busca, essas empresas são denominadas “Empresas Possíveis Parceiras” (EPP). Os retângulos com linhas tracejadas representam cada um dos registros de patente das diferentes empresas e dentro de cada registro há um conjunto de diferentes IPCs, cada campo IPC é representado por um pequeno retângulo dentro dos retângulos tracejados maiores e as diferentes texturas simbolizam as diferentes tecnologias que esses IPCs representam.

Os dados passarão por dois tipos diferentes de processamento e análise, porém os passos da metodologia são os mesmos, eles estão descritos no processo de busca de conhecimento em banco de dados (KDD, apresentado na p. 41). No Bloco 3 é representado o início do processo de Associação dos IPCs. Todos os IPCs das patentes adquiridas são buscados, então podados e ordenados em uma tabela onde cada patente é uma linha e os IPCs são colunas (Bloco 5), a existência do IPC na coluna é marcada com algarismo 1, enquanto que a inexistência é marcada com um zero. O processo de KDD de Agrupamento é marcado pelo Bloco 4, onde os textos de descrição das patentes são utilizados para agrupar patentes com elementos textuais similares. No Bloco 4 o texto das descrições das patentes são reunidos para o pré-processamento, que resulta na tabela localizada inferiormente no Bloco 5 em que cada linha representa uma patente e cada coluna um termo reduzido ao seu radical. Nesse pré-processamento são realizados a retirada das *stop-words*, o *stemming*, o cálculo do TFIDF e a normalização do TFIDF.

O Bloco 6 representa o *software* Weka, que processa as tabelas do Bloco 5 de acordo com o algoritmo selecionado, tanto para Associação quanto para Agrupamento. Isso resulta em relatórios com regras de Associação de IPCs do Bloco 7, ou relatórios contendo os resultados dos Agrupamentos de patentes com texto semelhante do Bloco 8.

3.3.3 Definição dos Dados para as Tarefas de Descoberta

Os dados utilizados para o processo de KDD são como a matéria-prima para a extração das informações úteis. Os dados utilizados na pesquisa são os registros de patentes e a sua fonte são as bases que publicam essas patentes para o mundo. Nessa subseção serão apresentadas: a base de patentes, que forneceu os documentos de patente; as características importantes para a escolha da Empresa Referência; e o subconjunto de documentos que foi escolhido para o andamento da pesquisa.

3.3.3.1 Base de Patente

Na escolha da base de dados que forneceu os registros de patente para a pesquisa buscou-se a base que limitasse minimamente o acesso aos dados. As principais características buscadas nessa base foram:

- facilidade para acesso aos registros de patentes e suas informações;
- documentos em formato HTML;
- livre acesso;
- facilidade na formulação da URL da página para acesso das patentes;
- base ampla.

Dentro dos requisitos supracitados, a base que mais se adequou foi a Espacenet, pois ela cumpre quatro dos cinco requisitos, falhando somente na falta de restrição de acesso. Ela limita as pesquisas em 500 registros e a apresentação de patentes na íntegra em cerca de 150 por dia de acesso. Isso força o *download* das patentes perdurar por alguns dias.

Outras bases foram preteridas, pois limitavam demasiadamente a busca pelas informações das patentes. Exemplo disso são a Derwent, com restrição de acesso; USPTO pela impossibilidade de modificação da URL; WIPO com patentes em formato PDF.

3.3.3.2 Escolha da Empresa Referência

O objetivo de selecionar uma Empresa Referência (ER) é ter um exemplo que permita validar as metodologias desenvolvidas e a implementação do sistema. O setor empresarial escolhido foi o de Telecomunicações, pelo volume de patentes que podem ser encontradas na área e a possibilidade de encontrar empresas nacionais que detém P&D nesse

campo. O ponto mais relevante para a seleção da Empresa Referência foi o número de patentes superior a vinte, porém inferior a cinqüenta. O número de patentes é relevante para a escolha da empresa de referência, pois deve ser suficiente para que possa mostrar em qual área de P&D a empresa está investindo seus recursos. Porém, não podem ser tantos que dificultem a análise, pois ela será realizada manualmente. O critério de escolha da ER deve-se à dinâmica do setor onde ela se encontra e a presença da mesma nas bases de patentes. Outras áreas e empresas que atendessem esses critérios seriam válidas também.

3.3.3.3 Seleção do Subconjunto de Patentes das Empresas Possíveis Parceiras

O número total de patentes das Empresas Passíveis de Parceria (EPP) adquiridas para o estudo não necessita limitação, pois neste segmento da pesquisa o processamento das patentes é realizado computacionalmente, de modo autônomo. Para um processo de KDD, a quantidade de dados é diretamente proporcional à qualidade do processo, porém, a limitação da aquisição de dados foi estabelecida pela base de patentes, que restringe em 500 o número de resultados em uma busca.

As patentes adquiridas não foram escolhidas ao acaso. Um IPC pertencente à ER é eleito como parâmetro de busca. Para sua escolha, é necessária a aquisição de todas as patentes da ER. A disposição dos seus IPCs em gráfico facilita a visualização de forma a denunciar o mais representativo. Nessa avaliação, vê-se qual foi o mais utilizado pela empresa no período em que teve suas patentes publicadas.

3.3.4 Interpretação dos Dados Minerados

Este passo consiste em analisar os padrões descobertos e relacioná-los com o mundo real. Porter e Cunninghan (2005) mostram que esse passo deve ser realizado pelo gestor de tecnologia, para validar um processo de Mineração de Tecnologia.

O guia para a interpretação dos resultados da mineração é seu modelo matemático. Porém, para limitar ainda mais a interpretação, serão apresentadas algumas possíveis tendências equivocadas que não devem ser seguidas durante a interpretação dos dados.

3.3.4.1 Interpretação das Regras Obtidas pelo uso do Apriori

Os resultados fornecidos pelo uso do algoritmo Apriori são uma aplicação direta do modelo matemático apresentado na subseção 3.3.1, a Informação Útil que irá retornar será a interdependência entre tecnologias.

As principais informações fornecidas pela saída do *software* Weka para a análise das regras geradas pelo algoritmo Apriori são: os elementos do lado esquerdo da regra, os elementos do lado direito da regra, Suporte e Confiança. Um Suporte muito baixo pode representar uma regra que acontece por acaso, é pouco para poder determinar um padrão, para determinar uma relação de possível causalidade. A Confiança mede a confiabilidade da inferência feita por uma regra.

Cada regra considerada representativa será confrontada com as informações da ER e das EPP. As informações das EPP serão buscadas diretamente no banco de dados. Os IPCs das regras serão comparados com os da Empresa Referência, o que pode resultar em três diferentes resultados:

- Nenhum dos IPCs encontrados na regra pertence à Empresa Referência. Isso significa que a empresa não detém nenhum dos campos tecnológicos da regra, então a regra não tem informação útil para a empresa.
- Pelo menos um dos IPCs encontrados na regra têm correspondência na Empresa Referência. Isso significa que a empresa tem potencial de firmar parceria com outras empresas, já que há indícios que os campos tecnológicos, representados pelos IPCs, combinam-se em um padrão.
- Todos os IPCs encontrados na regra podem ser encontrados na Empresa Referência. Isso significa que a Empresa Referência detém todos os conhecimentos técnicos que são necessários em uma cadeia de P&D. Nessas áreas ela não necessita de complementaridade, porém ela pode fornecer complementaridade à outras empresas, caso isso seja de seu interesse.

3.3.4.2 Interpretação de Grupos Criados por Algoritmos de Agrupamento

Para exemplificar o que se espera com a utilização dos Agrupamentos a Figura 16 apresenta um resultado obtido por Fattori, Pedrazzi e Turra (2003) com seu primeiro conjunto de dados. Os grupos conseguidos pelos autores são os retângulos e o número dentro deles é o

número de patentes confinados no grupo. Para se conhecer do que se trata cada grupo foi avaliado cada registro. O que resultou na Figura 16.

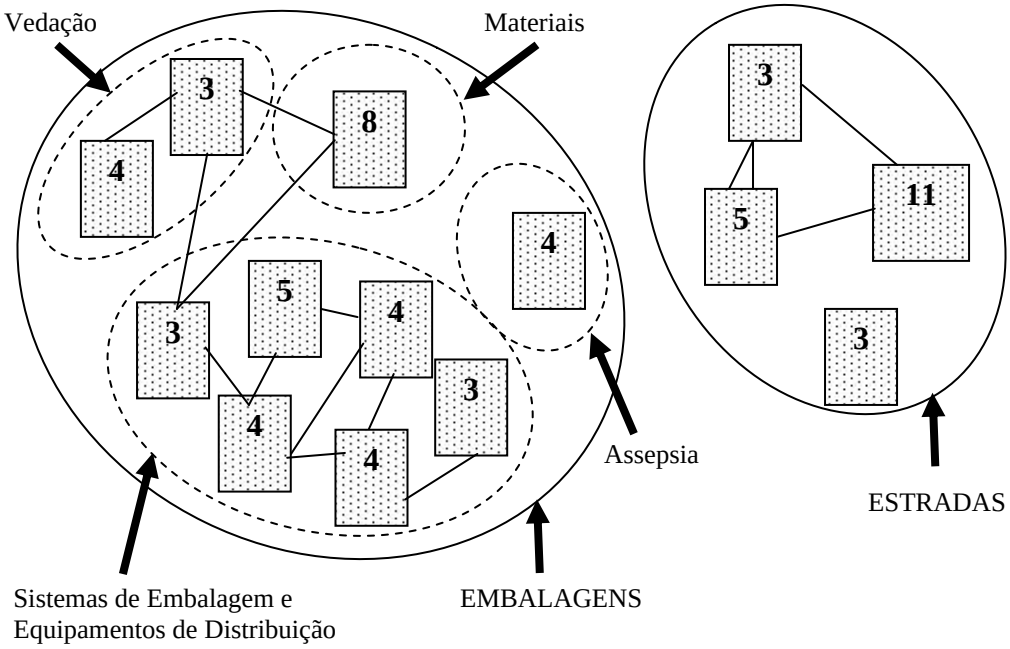


Figura 16 - Resultado obtido por Fattori, Pedrazzi e Turra em seu conjunto de dados.
Fonte: Fattori, Pedrazzi e Turra, 2003.

Em seu experimento, Fattori, Pedrazzi e Turra (2003) analisaram intelectualmente as patentes dos grupos obtidos. Nele foram avaliadas 192 patentes de um grupo industrial especializado em embalagens e outras áreas. Essas patentes foram divididas em dois grupos, um de 86 registros que vão de 1991 até 1995 e outro de 126 que vão de 1996 até 2000. Em sua metodologia eles decidiram que para um agrupamento ser considerado válido pelo menos 50% dos seus registros deveria ser considerado homogêneo.

Para a verificação de complementaridade de P&D de empresas proposto por esta pesquisa é necessário encontrar as variáveis *T'* (conjunto de subconjuntos de tecnologias interdependentes) e *I'* (subconjunto de tecnologias interdependentes) apresentados pela Equação 11 (p. 68).

Utilizando tarefas de Agrupamento não é possível achar a complementaridade diretamente como para Associação, porém probabilisticamente pode-se verificar se a complementaridade pode ser alcançada por empresas que estariam no mesmo grupo ou em grupos diferentes. Assim sendo, deve-se procurar pelo tamanho dos conjuntos de subconjuntos tecnologias interdependentes, *T'*. Então se definem as probabilidades:

$$p(|T'|, in) \tag{13}$$

como a probabilidade do tamanho do conjunto de subconjuntos de tecnologias interdependentes intra-grupo, isto é, dentro do mesmo grupo, e

$$p(|T'|, out) \tag{14}$$

como a probabilidade do tamanho do conjunto de subconjuntos de tecnologia interdependentes inter grupos.

A complementaridade entre empresas, buscado nesta pesquisa, está condicionado à interdependência de tecnologias, ou conhecimentos tecnológicos, entre empresas. Quando empresas desenvolvem uma tecnologia elas podem deter todas as interdependências tecnológicas para isso, e durante a descrição desta tecnologia nas patentes todas essas interdependências podem ser expressas no texto. As ferramentas de mineração de texto podem agrupar os textos que contém as mesmas tecnologias e as mesmas interdependências. Então, presume-se que a probabilidade de se alcançar o número mínimo de patentes (*#min*, na Equação 10) que tenham o mesmo subconjunto de tecnologias interdependentes é maior em patentes similares, isto é, que pertençam ao mesmo grupo do que entre patentes de grupos diferentes. Deste modo, as melhores escolhas para a realização de parcerias em Inovação Aberta estariam nas empresas que pertencem ao mesmo grupo. Portanto:

$$p(|T'|, in) > p(|T'|, out) \tag{15}$$

Para a realização dos testes de Agrupamento, as patentes da ER devem fazer parte do conjunto de patentes que serão submetidas ao processamento computacional, e também farão parte dos resultados. Espera-se que, durante a distribuição dos grupos, as patentes da ER permaneçam agrupadas com patentes similares a elas. Isso possibilita delimitar as empresas que teriam maior probabilidade de sucesso em uma parceria, que seriam as empresas pertencentes ao mesmo grupo que a ER, ou aos mesmos grupos.

A qualidade dos grupos formados pode ser medida por meio de medidas intra e extra-grupos, porém, somente avaliando individualmente os registros pode-se reconhecer se os grupos resultantes são consistentes ou não. Outro motivo para a avaliação individual dos grupos formados é que as técnicas de agrupamento não explicitam as características que ajudaram na formação do grupo. Para saber o porquê de suas uniões o usuário de *software* de mineração de dados deve analisar cada grupo e avaliar quais características formaram o determinado grupo. Para a análise dos grupos serão utilizadas amostras de cada grupo segundo critérios que serão apresentados posteriormente na subseção 4.2.3.2 (Realização do teste de Agrupamento).

3.3.4.3 Limitações na Interpretação dos Resultados Obtidos

Muitas das questões levantadas por essa pesquisa são provenientes de hipóteses que surgirão por meio da interpretação das informações úteis após o processo de mineração de tecnologia. Devido à facilidade de misturar conceitos de P&D com Mercado esta seção objetiva definir o que **não é** para ser interpretado com esses resultados.

As principais limitações de interpretação que podem ser listadas, são:

- Não se pode dizer que o co-desenvolvimento gerará produtos. Todos os resultados deste trabalho devem se restringir à complementação em P&D. Caso produtos bem sucedidos surjam de colaborações, eles são resultados indiretos do trabalho.
- Não se podem utilizar patentes como aproximações para produtos. A não ser que todas as empresas analisadas tenham sua fonte principal de dividendos provenientes de licenciamento de patente. Isto é, as patentes são o produto da empresa. Ou então, caso a empresa contenha somente um produto e uma patente.
- Não se pode analisar o mercado por meio dos resultados. Somente se pode analisar P&D por intermédio dos resultados.

Encontra-se no APÊNDICE B um modelo matemático que descreve a limitação da interpretação dos resultados somente ao P&D e não para o uso no mercado.

3.4 CONSIDERAÇÕES DA METODOLOGIA

O presente capítulo apresentou a metodologia da pesquisa. Ele inicia com a conceituação dos elementos utilizados na metodologia, que são conceitos importantes para o entendimento da pesquisa, porém não fazem parte da argumentação central a que ela se propõe. Em seguida, apresentou as ferramentas utilizadas para a construção do sistema e, por último, descreveu os procedimentos e os critérios para a aquisição dos dados, seu tratamento e a análise dos seus resultados. Por fim, esta seção faz a interligação das seções anteriores com o objetivo de melhorar o entendimento por meio de uma visão geral dos conceitos.

Por meio dedutivo, utilizando a teoria de conjuntos, mostrou-se as aproximações consideradas e hipóteses levantadas, para demarcar os procedimentos que são necessários para a realização da pesquisa. Houve a especialização do conceito de complementaridade para este trabalho, em que: uma empresa somente é considerada complementar à outra quando qualquer uma das duas detém o conhecimento que a outra necessita, desde que verificado que

possa haver ganhos na produção tecnológica de uma por meio do trabalho colaborativo da outra. O modelo matemático mostra-se importante, pois se torna um guia na criação dos métodos utilizados e é um padrão para avaliação da qualidade das ações tomadas e também a qualidade das interpretações dos resultados obtidos da Mineração de Tecnologia.

O alvo da pesquisa, de forma resumida, é verificar a possibilidade de complementaridade entre empresas utilizando as informações fornecidas dentro de registros de patente buscados na Internet. Esse objetivo é realizado por meio de duas tarefas de Mineração de Dados distintas, a saber: Associação e Agrupamento. Sumarizadas abaixo:

- a Associação tem a capacidade de encontrar padrões de tecnologia dentro das patentes utilizadas;
- o Agrupamento tem a capacidade de juntar em grupos as patentes que têm características textuais similares.

As duas tarefas têm metodologia própria de tratamento dos dados, porém a coleta dos dados foi balizada pelo mesmo método. Os registros de patentes utilizados para ambas as tarefas são os mesmos, o que é característico para elas é o campo da patente utilizado durante o processamento da tarefa. O critério para a seleção destas patentes foi o IPC mais representativo da Empresa Referência, todas as patentes que possuem esse mesmo IPC foram adquiridas no sítio da Espacenet (2009). Então, as empresas titulares dessas patentes são reconhecidas como empresas passíveis de parceria de P&D, em Inovação Aberta, com a Empresa Referência.

Ambas as tarefas de Mineração de Dados utilizaram o *software* Weka, que é gratuitamente disponível na Internet, para o processamento de seus dados.

A tarefa de Associação utilizou as classificações técnicas das patentes, que são os IPCs apresentados nos registros. O que se espera desta tarefa é que a combinação dos IPCs possa informar interdependência de conhecimentos tecnológicos envolvidos no processo de produção de tecnologia. Para isso, os IPCs necessitam de tratamento de modo que sua representação em matriz de vetores facilite o processamento computacional. Isso levou a retirada de atributos que estavam abaixo do Suporte esperado e posteriormente retirando as instâncias que não tinham elementos com valor válido. Os resultados da Associação são expressos por meio de regras do tipo $X \rightarrow Y$, onde a existência do elemento da direita da expressão pode ser condicionada a existência do lado esquerdo da mesma.

Espera-se com a tarefa de Agrupamento que os grupos de patentes computados contenham elementos textuais similares, e esses denunciem a aplicação das patentes em áreas diversas. A análise desses grupos pode mostrar quais empresas são mais aptas a trabalharem

em parcerias, no modelo de Inovação Aberta. Resumidamente, o modelo matemático apresentado pretende mostrar que empresas pertencentes aos mesmos grupos têm maior probabilidade de serem complementares entre si que empresas de grupo diferentes. Para possibilitar o tratamento computacional do texto de descrição das patentes por essa tarefa o texto foi representado em forma de matriz de vetores *m*-dimensionais. Onde cada patente é um vetor e cada termo de seu texto é uma dimensão, são utilizadas as técnicas de retirada de *stop-words* e *stemming* para a diminuição de dimensionalidade, e o TFIDF normalizado é utilizado para atribuição de "pesos", com valores entre 0 e 1, para cada termo.

A complementaridade entre empresas pode ser verificada com ambas as técnicas de Mineração de Dados, porém o enfoque de uso para as técnicas é diferente. Enquanto a técnica de Associação possibilita fornecer informações no nível microscópico, a técnica de Agrupamento fornece informações no nível macroscópico. Isto é, a primeira técnica verifica se há padrões não identificados de pesquisa, inferindo que há conhecimentos técnicos que seriam interdependentes e então apresenta diretamente quais empresas podem ser complementares, a segunda técnica infere que conhecimentos técnicos distintos ficam em grupos distintos, então os conhecimentos mais diversos podem ser concentrados no mesmo grupo o que pode ocasionar mais idéias para a diversificação de produtos e tecnologias e sua utilização depende do plano de negócios da empresa.

A metodologia apresentou-se própria para o que se propõe a pesquisa. As técnicas computacionais de Mineração de Dados podem automatizar a busca de complementaridade entre empresas. Como apresentado, tanto a ferramenta para o processamento da mineração, Weka, quanto as fontes de dados, as patentes, podem ser encontradas livremente na Internet, o que pode auxiliar empresas que gostariam de utilizar algum tipo de busca inteligente no auxílio da tomada de decisão, porém não podem arcar com os custos de ferramentas comerciais.

4 DESENVOLVIMENTO DO SISTEMA COMPUTACIONAL E TESTES

Neste capítulo, é relatado o desenvolvimento de cada procedimento executado para alcançar os resultados da pesquisa, eles são embasados nos limites e conceitos apresentados no capítulo anterior. O conteúdo articulado por este capítulo pode ser separado em duas frentes distintas de conteúdo. Primeiramente, é abordada a construção do sistema computacional, em seguida, são descritos o levantamento das patentes e os testes realizados com elas.

O primeiro conteúdo apresentado no capítulo é a construção do sistema computacional. O trabalho realizado durante o desenvolvimento foi escrever o código fonte do sistema, que busca pelos registros de patente e realiza seu tratamento para o processo de Mineração de Dados, porém não é o intuito desta dissertação apresentar esse nível de detalhe, o que também é de pouco proveito. Os conteúdos apresentados do sistema serão sua arquitetura principal e, no decorrer da descrição dos testes são apresentados, os algoritmos e critérios utilizados na delimitação do tratamento dos dados.

A segunda frente de conteúdo deste capítulo são o levantamento dos dados, isto é, a descrição de como foi realizada a busca das patentes no banco de patentes, e a descrição dos testes, primeiro, de Associação e, em seguida, de Agrupamento.

O capítulo encerra-se com as considerações mais relevantes sobre o desenvolvimento do sistema, resultados, e a comparação entre as duas metodologias nos testes.

4.1 CONSTRUÇÃO DO SISTEMA DE MINERAÇÃO DE TECNOLOGIA

Nessa seção, é descrita a implementação do sistema de mineração de tecnologia utilizado nesta pesquisa. Essa descrição apresenta o trabalho realizado durante a pesquisa, porém não atende todos os detalhes técnicos, somente aqueles relevantes para o resultado final do sistema. Detalhes como as classes de objetos do *software*, estrutura de banco de dados comunicação entre os processos entre outros, podem ser realizados de outras maneiras sem afetar a repetibilidade da pesquisa.

4.1.1 Arquitetura básica do sistema

Como determinado nos objetivos gerais da pesquisa, há a necessidade de implementar um *software* para a aquisição e tratamento de dados. Esse foi implementado utilizando a plataforma Visual Studio 2008 da Microsoft, e escrito em linguagem C#.

A Figura 17 apresenta o esquema principal da estrutura do sistema para Mineração de Tecnologia. As entradas do sistema são os registros de patente do sítio da Espacenet (2009). A saída do sistema são os dados processados pelo Weka (2009). Todo o sistema desenvolvido na pesquisa está contido internamente à elipse tracejada.

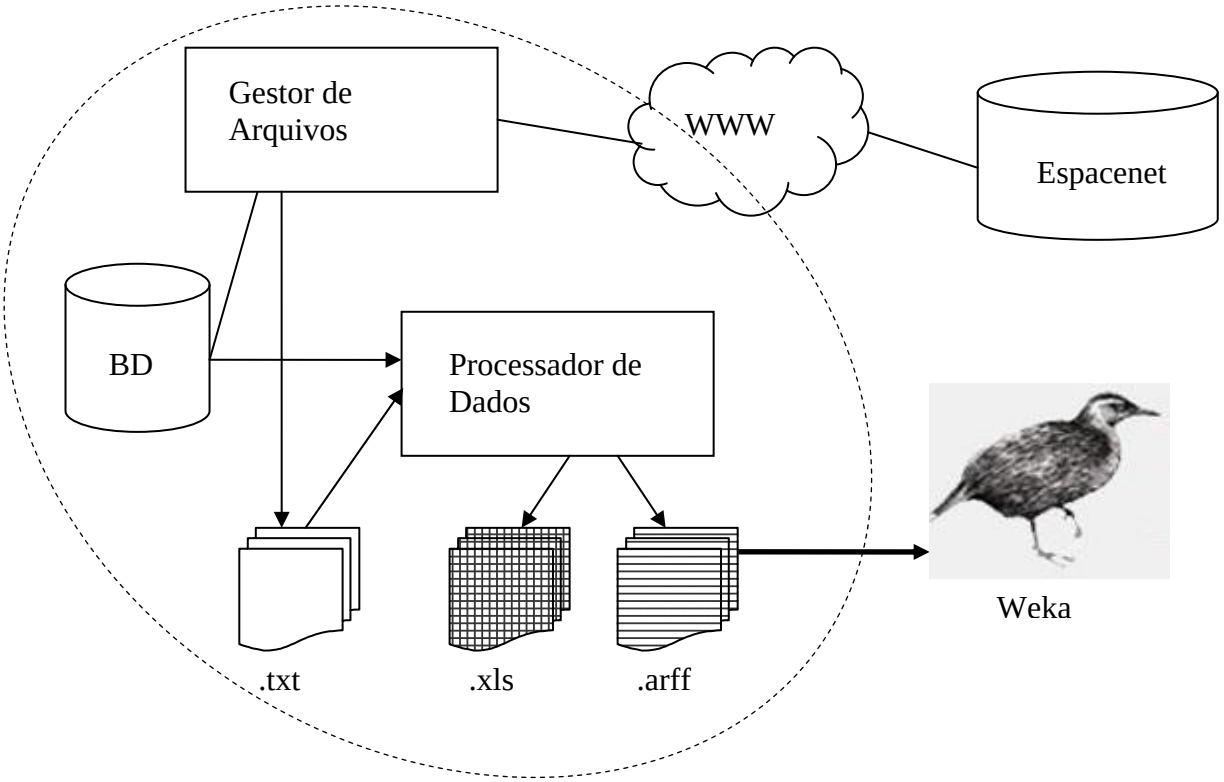


Figura 17 - Estrutura do sistema de Mineração de Tecnologia.
Fonte: Autoria própria.

Os *softwares* para aquisição e tratamento dos dados são representados pelos blocos nomeados “Gestor de Arquivos” e “Processador de Dados”. As saídas do *software* de aquisição e tratamento dos dados são os arquivos nos formatos TXT, XLS (tabelas Excel) e ARFF (arquivos Weka). Os arquivos TXT têm a funcionalidade de *backups* do sistema, os dados adquiridos via Internet são salvos *in natura* em formato TXT para facilitar outra rodada do sistema sem a necessidade de acesso ao banco de dados da Espacenet novamente. Os

arquivos XLS trazem estatísticas para análise da Empresa Referência. Os arquivos ARFF são resultado do que foi chamado de Preparação dos Dados, na subseção 3.1.3.

4.1.1.1 Gestor de Arquivos

O Gestor de Arquivos é responsável por baixar as patentes do sítio da Espacenet por meio de um cliente HTTP e salvá-los para processamento futuro. Ele busca as patentes por intermédio da URL do sítio, que é estruturada em um padrão que permite buscas sem a necessidade de acesso direto a um *browser*[7]. Sua busca realiza-se em duas partes. Primeiramente, com o uso de 20 URLs de requisição modificados e enviados ao servidor da base de patentes. Todos os URL fornecem uma página com informações de 25 patentes, dentro dessas informações encontra-se o *link* principal de cada registro. Utiliza-se esse *link* para baixar todas as informações possíveis de cada registro.

As patentes baixadas são salvas de duas formas: crua e tratada. As patentes são salvas de forma crua, isto é, sem tratamento algum de texto para possibilitar novo pré-processamento caso necessário por incoerência ou falha da construção do sistema. Essas patentes são salvas em hiper-texto, porém em formato TXT. As patentes tratadas são salvas no banco de dados, em campos bem definidos. O tratamento é realizado por meio de expressões regulares e são armazenadas no banco de dados (BD). As expressões regulares são padrões de texto que o sistema procura, elas são necessárias para automatizar a busca dos campos de cada patente. As expressões regulares que filtram os arquivos HTML, resultantes da requisição HTTP ao sítio da Espacenet, são customizadas especificamente para esse sítio, caso fosse necessária a troca de base de patentes também seria necessária a customização de outras expressões regulares.

4.1.1.2 Processador de Dados

O processador de dados é onde os dados são processados para serem apresentados. É ele que realiza a poda de dados, retirada de *stop-words*, stemming, cálculo e normalização

[7] *Browser* é o aplicativo computacional utilizado para acesso à páginas da Internet, como o Internet Explorer, Google Chrome ou Firefox.

do TFIDF. Sua entrada são os dados contidos no BD. Essas informações são buscadas por *queries*[8], ou nos arquivos .txt com o texto da descrição das patentes. Suas principais saídas são:

- arquivos .xls (Microsoft Excel) para relatório de IPC para Empresa Referência; e,
- arquivos .arff de vetores de dados tratados para execução de classificação, associação ou agrupamento.

No relatório de IPCs são listados todos os IPCs da Empresa Referência e suas datas de publicação. Essas informações são importantes para o levantamento de qual IPC é considerado o mais representativo da empresa.

Os arquivos .arff são executados pelo *software* Weka, o que resulta em estatísticas e padrões passíveis de análise.

4.2 BUSCA DOS DADOS E TESTES

O objetivo desta seção é descrever os testes realizados durante a pesquisa. Para alcançar esse alvo, primeiramente deve-se realizar o levantamento dos dados, que são as patentes. Então, logo na seqüência é descrito esse procedimento, que é único para todos os testes realizados. Posteriormente serão apresentados os testes de Associação e depois de Agrupamento.

4.2.1 Busca e Levantamento das Patentes

O processo de busca e levantamento das patentes, tanto para o teste de Associação quanto para o teste de e Agrupamento é o mesmo procedimento. Ele seleciona o mesmo conjunto de dados alvo para os dois testes. Nessa seção são apresentados os procedimentos de busca, primeiramente, das patentes da Empresa Referência (ER) e, posteriormente, das patentes das Empresas Possíveis Parceiras (EPP).

Os critérios para a escolha da ER e o modo de levantamento de seus dados foram previamente apresentados na subseção 3.3.3.2 (Escolha da Empresa Referência). A empresa

[8] *Queries* são linhas de comando com as instruções necessárias para requisição de informações existentes em Banco de Dados.

selecionada como ER foi a Genband LTDA. A empresa define-se como uma das líderes mundiais em *Gateways* IP, *Session Border Controllers* e soluções de segurança FMC (GENBAND, 2009). Ela foi escolhida para ser a ER, pois tinha uma unidade de P&D em Curitiba, localizada na PUC-PR, o que facilitaria o contato com os responsáveis técnicos, e se trata de uma empresa de telecomunicações. Essa unidade foi desativada em julho de 2009, alguns meses após o início dos procedimentos dos testes, porém mesmo assim continuou figurando como ER para a pesquisa.

A ER contava, no dia 10 de fevereiro de 2009, com 26 registros de patentes publicados no sítio da Espacenet (2009). Essas patentes têm a data de sua publicação distribuída de agosto de 2006 até setembro de 2008. A distribuição das áreas técnicas é sumarizada pela Figura 18, que mostra os IPCs utilizados para classificar as publicações da ER versus o número de vezes que foram utilizados para isso. Totalizam-se 21 IPCs que classificam as 26 patentes.

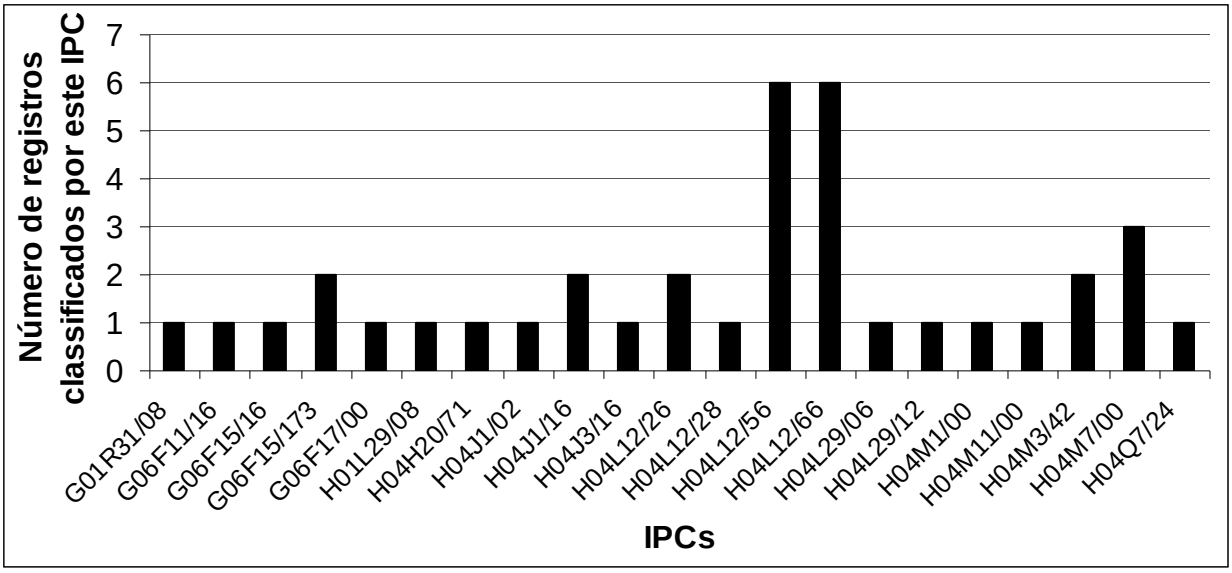


Figura 18 - Relação do número de classificações das patentes pelo IPC.
Fonte: Autoria própria.

O IPC H04L12/66[9] foi eleito como a área de depósito que mais representa a ER, pois há mais publicações de registros com esse IPC, e elas são melhor distribuídas pelo tempo. Não coincidentemente, esse IPC é dedicado aos elementos de rede denominados *Gateways*, que são os principais produtos da empresa.

[9] *Transmission of digital information (H04L) Data switch networks (12/) Arrangements for connecting between networks having differing types of switching systems (e.g. gateways) (66).*

A escolha das Empresas Possíveis Parceiras em desenvolvimento de inovações somente considerou como parâmetro de seleção o IPC mais representativo da ER. As empresas que apresentam esse mesmo IPC dentro da busca realizada pelo servidor de patentes são possíveis candidatas à análise de complementaridade. Espera-se que as EPP tenham alguma área compatível com a ER, por isso todas as empresas que servem como potenciais parceiras devem ter pelo menos uma patente com o IPC que é o mais representativo da ER.

Partindo do IPC H04L12/66 foi realizada a busca na base da Espacenet, que retornou 499 registros de patentes de 274 titulares diferentes com 232 IPCs distintos, co-ocorrentes com o mais representativo, distribuídos de 15 de outubro de 1999 até 12 de fevereiro de 2009.

4.2.2 Execução da Tarefa de Associação

As particularidades encontradas no andamento dos experimentos de Associação e os resultados encontrados pelo uso dessa tarefa são descritas nesta subseção.

4.2.2.1 Passos do Desenvolvimento da Tarefa de Associação

Essa subseção expõe como foi montada a matriz de vetores com as informações de IPC para a tarefa de associação. Os dados inicialmente coletados para essa tarefa poderiam criar uma tabela de 499 vetores com 232 dimensões, onde cada dimensão assume somente dois valores, ‘1’ para marcar a existência do elemento ou ‘0’ para marcar que o elemento não existe. Porém, há dados não necessários dentro desse conjunto, o que pode elevar o custo computacional na mineração de dados. O processo de poda de dados e redução de dimensionalidade resultou em um conjunto de 194 vetores com 103 dimensões.

Para o processo de poda, primeiramente, foi excluído o parâmetro H04L12/66 dentre os IPCs, porque ele não carrega nenhuma informação já que todos os registros o possuem. Depois, foi escolhido um limite mínimo para o número de patentes que possuem o IPC. Isto é, caso o IPC não fosse presente em um número mínimo de patentes, ele seria descartado. Esse limiar deve ser menor que o Suporte, para que diminuição da dimensionalidade não afete o resultado do teste. Nesse teste foi escolhido o limiar de duas patentes que possuíssem o IPC. Após a retirada dos IPCs há registros que ficam com todas as suas dimensões com valor inexistente, então esses registros também foram retirados do

conjunto de dados, o que resultou na economia de 61% dos vetores e 56% nas dimensões, isto é, economia de 61% das patentes e 56% IPCs.

4.2.2.2 Regras Obtidas por Associação

As regras extraídas por mineração de dados estão listadas no Quadro 6. Para a extração dessas regras utilizou-se o Suporte mínimo igual a 2%, Suporte máximo de 30% e confiança mínima de 80%. Limitou-se o Suporte máximo, pois as regras mais raras, portanto mais interessantes, tendem a aparecer em menor freqüência que as mais vulgares.

Dentro do conjunto de regras obtidas, as mais significativas são: 1, 2 e 4. Elas serão analisadas, por que apresentam os três maiores Suportes dentre os resultados, 5,6%, 2,6% e 3%, respectivamente.

Índice	Lado Esquerdo da Regra		Lado Direito da Regra	Suporte	Confiança
1 .	{ H04Q3/64 }	→	{ H04M11/06 }	5,6%	100%
2 .	{ H04L12/12 }	→	{ H04L29/06 }	2,6%	100%
3 .	{ H04L29/06, H04Q3/64 }	→	{ H04M11/06 }	2%	100%
4 .	{ H04L12/56, H04L29/12 }	→	{ H04L29/06 }	3%	83%
5 .	{ H04L12/28, H04M7/00 }	→	{ H04L29/06 }	2,5%	80%
6 .	{ H04M7/00, H04L12/56 }	→	{ H04L29/06 }	2,5%	80%
7 .	{ H04L12/56,H04L29/08 }	→	{ H04L29/06 }	2,5%	80%

Quadro 6 – Regras obtidas com o uso da Associação de IPCs.
Fonte: Autoria própria.

A primeira regra tem um alto Suporte, 5.6% (11 registros), indicando que esse padrão tem uma alta correlação (baixa possibilidade de ser um padrão ao acaso), e alta confiança, 100%, significando que toda vez que o primeiro lado da regra acontece a segunda também acontecerá. A regra pode ser traduzida como: “toda vez que uma invenção estiver

dentro do campo H04Q3/64[10], então essa invenção também utilizará técnica com campo H04M11/06[11]".

A segunda regra não tem Suporte tão alto, 2,6% (5 registros), o que representa uma correlação fraca entre os elementos, porém alta Confiança, que significa que a regra acontece todas as vezes que o primeiro membro da regra é presente em uma patente. Ela pode ser interpretada como: "toda vez que uma invenção estiver dentro do campo H04L12/12[12], então essa invenção também utilizará técnicas descritas por H04L29/06[13]".

A quarta regra apresenta dois elementos no lado esquerdo, representando um *itemset*. Ela, diferente das anteriores, não tem a Confiança absoluta, porém sua confiança é alta (83%) o que significa que 83% das vezes que o *itemset* do lado esquerdo da regra aparecer em uma patente, o elemento do lado direito aparecerá. Seu Suporte é baixo, o que indica que os termos da regra não têm alta correlação, tendo boa probabilidade de acontecer ao acaso. Ela pode ser interpretada como: "83% das vezes que uma invenção for classificada ao mesmo tempo como H04L12/56[14] e H04L29/12[15] pode também ser classificada como H04L29/06[13]".

4.2.2.3 Análise dos Resultados de Associação

Dentre os resultados obtidos por Associação, o mais significativo é a regra {H04L12/12[12]}→{H04L29/06[13]} que, para a empresa referência, representa uma possível parceria. A empresa referência tem uma patente publicada no dia 31/01/2008 que, entre outros IPCs, está enquadrada como H04L29/06[13]. Então, no conjunto de patentes analisado, toda vez

---

[10] *ELECTRIC COMMUNICATION TECHNIQUE (H04) Selecting (Q) Arrangements (3/) Distributing or queuing (64).*
[11] *ELECTRIC COMMUNICATION TECHNIQUE (H04) Telephonic communication (M) systems adapted for combination with other electrical systems (11/) Simultaneous speech and telegraphic or other data transmission over the same conductors (06).*
[12] *ELECTRIC COMMUNICATION TECHNIQUE (H04) Transmission of digital information (L) Data switching networks (12/) Arrangements for remote connection or disconnection of substations or of equipment thereof (12).*
[13] *ELECTRIC COMMUNICATION TECHNIQUE (H04) Transmission of digital information (L) Arrangements, apparatus, circuits or systems, not covered by a single one of groups H04L1/00 to H04L27/00 (29/) Transmission control procedure (06).*
[14] *ELECTRIC COMMUNICATION TECHNIQUE (H04) Transmission of digital information (L) Data switching networks (12/). Packet switching systems (56).*
[15] *ELECTRIC COMMUNICATION TECHNIQUE (H04) Transmission of digital information (L) Arrangements, apparatus, circuits or systems, not covered by a single one of groups H04L1/00 to H04L27/00 (29/) characterized by the data terminal (12).*

que uma patente está em H04L12/12[12] ela também deve estar em H04L29/06[13]. O que se pode inferir desta regra é que o conhecimento necessário para realizar a primeira parte da regra necessita do segundo lado, tendo em mente que a complementaridade é dada por: $\Pi(1,1) - \Pi(0,1) \geq \Pi(1,0) - \Pi(0,0)$, e interpretado em Cassiman e Veugelers (2006) como: "somando-se uma atividade à outra atividade em andamento obtém-se um efeito incremental no desempenho, maior que a soma de cada atividade isoladamente".

São cinco as patentes, dentre as 499 adquiridas para a pesquisa, que contém o IPC H04L12/12, que demonstra uma possível dependência do IPC H04L29/06, presente em uma patente da ER. As empresas que podem ser consideradas potenciais parceiras da ER são: Res in Motion Ltd., Toshiba KK, Alcatel Lucent e Mitel Networks Corp. A empresa Res in Motion Ltd detém duas patentes das cinco que são classificadas pelo IPC H04L12/12.

Outra regra que deve ser analisada, porém não resultou em informação útil nova, é a regra { H04L12/56[14], H04L29/12[15] } → { H04L29/06[13] }. Ela apresenta um padrão de depósito de patentes, que pôde ser constatado na própria empresa referência, corroborando com a hipótese de que há padrões de IPCs dentro de um conjunto de patentes publicadas. Vale ressaltar que as patentes da ER não constam neste teste, então essa regra foi somente a constatação de um padrão de utilização tecnológica corrente também presente em outras empresas.

A última regra, { H04Q3/64[10] } → { H04M11/06[11] }, foi a regra que apresentou o melhor Suporte, porém não apresenta informação útil para a empresa referência, pois nenhum dos dois conhecimentos tecnológicos apresentados está presente nela.

4.2.3 Execução da Tarefa de Agrupamento

As particularidades encontradas no andamento dos experimentos de Agrupamento e os resultados encontrados pelo uso dessa tarefa são descritas nesta subseção.

4.2.3.1 Passos do Desenvolvimento da Tarefa de Agrupamento

Esta subseção apresenta como foi montada a tabela de vetores *m*-dimensionais que representa o texto da descrição de todas as patentes para o sistema computacional. Para este teste foram utilizadas as mesmas patentes utilizadas para o teste de Associação, porém foram somadas as patentes da ER, como descrito anteriormente na subseção 3.3.4.2, e subtraídas as

patentes que apresentam língua de publicação diferente de inglês. O idioma foi um parâmetro que não pode ser selecionado durante a aquisição dos registros, levando a obtenção de patentes com descrição em alemão e francês dentro do conjunto majoritário em inglês. Para a tarefa de Agrupamento o idioma é um parâmetro essencial, pois os algoritmos de retirada de *stop words* e *stemming* são intimamente ligados a ele. Os registros não publicados em inglês foram retirados manualmente do grupo analisado, no total foram três registros em idioma francês e 11 em alemão.

A preparação dos dados para a realização dos testes de Agrupamento seguiu a seguinte ordem de procedimentos:

- filtragem das *stop-words*;
- *stemming* dos termos do texto;
- filtragem das palavras com ocorrência máxima e mínima;
- realização do TFIDF;
- normalização do TFIDF.

O tratamento dos textos resultou em 328 vetores de 2322 dimensões, cada dimensão com valores que podem variar entre zero e um. Isto é, 328 patentes com 2322 termos reduzidos ao seu radical com o valor que expressa o “peso” que o termo tem com relação àquele conjunto de dados. Há discrepância entre o número de patentes utilizadas para este teste com o teste anterior, pois não são todas as patentes adquiridas que têm o texto de descrição, além das 14 patentes retiradas do conjunto por estarem em idioma diferente do inglês e as 26 inseridas da ER. É importante salientar que as patentes da ER entraram neste teste, porque se busca saber quais outras empresas compartilharão o mesmo grupo, ou grupos, que ela.

Para a execução do algoritmo de *K*-médio é necessário apresentar ao algoritmo o número de grupos que ele deve retornar. O número escolhido foi de seis grupos. Essa escolha foi subjetiva, podendo ser alterado devido uma nova demanda gerada pelos próprios resultados. Essa quantidade de grupos não é demasiadamente grande que impossibilite a análise intelectual individual de cada grupo, porém também não é demasiadamente pequena que não permita a invalidação de grupos que não alcançam o número mínimo de patentes em sua composição, os critérios de validade dos grupos são apresentados na próxima subseção. A equação matemática utilizada para o cálculo das distâncias entre as instâncias e seus centróides foi a Distância Euclidiana.

4.2.3.2 Realização do teste de Agrupamento

Na geração dos grupos foi utilizada a técnica de múltiplas rodadas do *software* alterando entre elas a semente aleatória. O critério de escolha da rodada utilizada para a análise considera o menor SEQ e também a quantidade de grupos válidos dentre os apresentados. Nesta pesquisa o critério para a validade do grupo é que ele seja composto por mais de cinco patentes. Na escolha de qual rodada seria utilizada para análise dos grupos, verificou-se, primeiramente, a quantidade de grupos válidos gerados, então, o critério final é o menor valor de SEQ. O SEQ representa a soma das Distâncias Euclidianas entre as instâncias em um grupo com o seu respectivo centróide .

O *software* Weka foi rodado 20 vezes, que resultou na tabela a seguir:

Tabela 2 - Semente aleatória e SEQ de múltiplas rodadas do K-médio.
Fonte: Autoria própria.

Semente Aleatória	Soma do Erro Quadrático (SEQ)
0	4385
1	4381
2	4280
3	4342
4	4342
5	4586
6	4389
7	4343
8	4353
9	4377
10	4388
11	4357
12	4370
13	4328
14	4372
15	4359
16	4392
17	4394
18	4357
19	4354

Dentre as rodadas realizadas, somente as com semente aleatória 3, 8, 11 e 18 apresentaram o maior número de grupos válidos, neste caso quatro. E o grupo com menor SEQ foi o de semente 3, com valor 4342.

O número de patentes dentro de cada grupo está apresentada na Tabela 3, sendo quatro grupos válidos, com mais de cinco elementos, e dois grupos inválidos.

Tabela 3 - Grupos gerados pela tarefa de Agrupamento, número de instâncias e o percentual de instâncias dento do conjunto total de patentes.
Fonte: Autoria própria.

Índice dos Grupos	Número de Instâncias	Percentual do Total
Grupo 0	19	6%
Grupo 1	50	15%
Grupo 2	30	9%
Grupo 3	224	68%
Grupo 4	4	1%
Grupo 5	1	0%

A principal técnica para determinar quais características comuns das patentes que fizeram elas permanecerem no mesmo grupo foi a leitura e avaliação intelectual do título e do resumo de uma amostra aleatória de patentes de cada grupo. Outras duas técnicas foram utilizadas, porém não foram determinantes para a definição dos grupos, então seus resultados foram desconsiderados, porém, elas são apresentadas no APÊNDICE C somente para sua documentação. A avaliação das patentes seguiu o seguinte critério:

- para grupos com mais de 75 elementos é escolhida uma amostra aleatória de 20% patentes para análise;
- para grupos com menos de 75 patentes são amostradas 15 patentes aleatoriamente para análise; e,
- para grupos com menos de 15 patentes todas as patentes são avaliadas.

O critério para definir o quais os principais temas de cada grupo é o mesmo utilizado por Fattori, Pedrazzi e Turra (2003), os temas que forem 50% homogêneos na amostragem do grupo são eleitos como o tema que diferencia o grupo. Caso o grupo não alcance 50% de representatividade em um tema ele é invalidado para a classificação, pois não convergiu em uma área de conhecimento.

4.2.3.3 Resultados do Agrupamento e Critérios Subjetivos

Os temas mais significativos abordados pelas patentes, subjetivamente levantados, são: multimídia, aplicação para clientes finais, interoperabilidade, comunicação de *core*, comunicações por interface ar, conexão de última milha e redes IP. Os temas menos abordados, que serão tratados conjuntamente como “outros” durante a análise são: gerência, qualidade de serviço, segurança e eficiência.

Os sete principais temas foram levantados por interpretação subjetiva do pesquisador e não têm correspondência direta com as classificações definidas com os IPCs.

Os critérios de classificação das patentes estão descritos a seguir, e cada patente pode ser classificada dentro de um ou vários temas. Dentro do tema de **multimídia** foram classificadas as patentes que abordam os assuntos como comunicação em tempo real de dados, sendo voz ou imagens, em qualquer tipo de rede, tanto *packet-switched*[16] quanto *circuit-switched*[17], tanto a transmissão da sinalização utilizada como da própria mídia. O tema **aplicação para clientes finais** apresenta produtos, tanto *software* quanto *hardware*, que são utilizados pelos clientes domésticos e comerciais que não são as operadoras de telecomunicações, porém que podem ser oferecidos pelas operadoras. O tema **interoperabilidade** aborda as interconexões realizadas entre redes que utilizam mídias, sinalizações e transporte diferentes, habilitando interação transparentemente um com o outro. Dentro do tema **comunicação de *core*** foram classificadas as patentes que abordam os assuntos relacionados com a comunicação de alta velocidade que acontece entre os equipamentos das operadoras de telecomunicações, tanto dentro da mesma operadora, como entre operadoras. A comunicação de *core* é a "espinha dorsal" da comunicação das operadoras, podendo ser em meio ótico (fibra ótica), elétrico (cabos coaxiais entre outros) e interface ar. O tema **comunicação em interface ar** aborda todo tipo de comunicação que ocorre pela propagação de ondas eletromagnéticas através do ar, por exemplo, comunicações celulares, *wi-fi*, rádio difusão, comunicações via satélite, entre outras. O tema **comunicação de última milha** aborda toda comunicação que acontece na fronteira entre a rede da operadora de telecomunicações e o cliente final. Por último, o tema **redes IP** aborda toda comunicação que ocorre utilizando o protocolo[18] IP[19], como também os diversos protocolos de roteamento[20].

Os temas que podem classificar os grupos não são disjuntos, isto é um tema pode ter assuntos em comum com o outro. Isso foi permitido, pois as patentes podem não ter um tema principal que as classifiquem, o que ocorre com os grandes temas, pois não há como dissociar totalmente um tema do outro já que todos são interligados de algum modo. Exemplo

---

[16] *Packet-switched* é o paradigma de comunicação que ocorre por troca de pacotes de dados (unidade de transferência de informação) em redes que podem ser compartilhadas por outros. Um exemplo dessa comunicação é o *software* de telefonia Skype.

[17] *Circuit-switched* é o paradigma de comunicação que utiliza alocação permanente de uma via de comunicação. Nesse sistema é necessário um recurso dedicado para suportar a transmissão contínua de informação. Um exemplo dessa comunicação são as linhas telefônicas analógicas e digitais oferecidas pelas operadoras brasileiras.

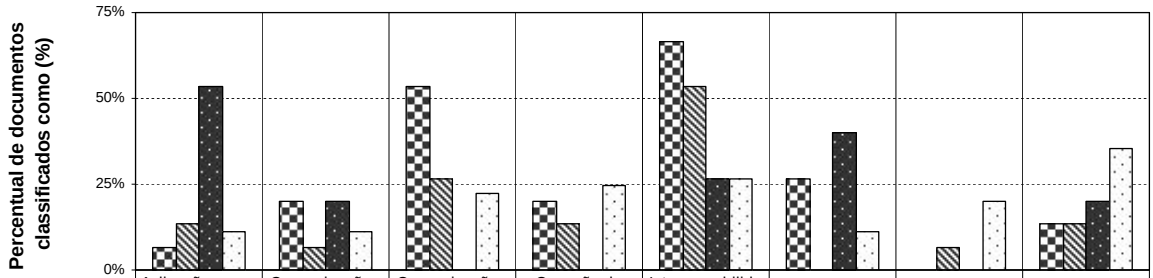
[18] Protocolo é o conjunto bem definido de instruções que dois sistemas computacionais trocam para que realizem operações acordadas entre ambos.

[19] IP é conhecido por ser o protocolo que dirige a Internet.

[20] Roteamento define o caminho (a rota) por onde um pacote de dados irá passar dentro de uma rede.

disso são as comunicações de *core* que podem ser realizadas pela interface ar e também podem ser realizadas por redes IP, ou aplicações de clientes que são para a geração e transmissão de conteúdo multimídia.

Os resultados encontrados pela avaliação subjetiva do título e resumo das patentes estão sumarizados na Figura 19.



	Aplicações para Clientes	Comunicações de Core	Comunicações por Interface Ar	Conexão de última milha	Interoperabilidade	Multimídia	Redes Ip	Outros
Grupo 0	7%	20%	53%	20%	67%	27%	0%	13%
Grupo 1	13%	7%	27%	13%	53%	0%	7%	13%
Grupo 2	53%	20%	0%	0%	27%	40%	0%	20%
Grupo 3	11%	11%	22%	24%	27%	11%	20%	36%

Figura 19 - Resultado da classificação subjetiva dos grupos.
Fonte: Autoria própria.

O Grupo 0 foi representado por barras quadriculadas, na Figura 19. Vê-se que entre os temas abordados os principais são “Comunicações por Interface Ar” e “Interoperabilidade”. Foram avaliadas 15 patentes e 10 delas abordam interoperabilidade e oito abordam interface ar, que foram os temas que ultrapassam o limite de 50% de representatividade para serem eleitos como os temas que definem o grupo.

O Grupo 1 foi representado por barras listradas na diagonal, na Figura 19. Vê-se que o único tema que ultrapassa 50% de representatividade é a Interoperabilidade, que teve 8 registros abordando o tema, dos 15 amostrados no grupo.

O Grupo 2 foi representado por barras pretas com pontos brancos, na Figura 19. O único tema que se ultrapassa o limite de 50% de representatividade para a classificação foi a aplicação para clientes. Para determinação foram amostradas 15 patentes para avaliação e 8 patentes apresentaram este tema.

O Grupo 3 foi representado por barras brancas com pontos pretos na Figura 19. Este foi o único grupo que demandou a análise de uma amostra de 20% de seu total devido a existência de mais de 75 patentes em seu conjunto, que é de 224 patentes. Neste grupo nenhum tema conseguiu atingir o limite mínimo de 50% de representatividade para a classificação. Sendo assim, o grupo não pode ser definido por tema e torna-se inválido para análise. O Grupo 3 foi onde os muitos dos temas menos representativos se encontram, eles,

juntos, representam 36% das classificações individuais das patentes, sendo o tema mais representativo daquele grupo. Esse grupo tem a tendência de ser genérico, ou então considerado como *default*, isto é, todas as patentes que não se enquadraram de algum modo nos temas principais acabaram por serem agrupadas nele.

A representatividade das patentes da ER dentro dos grupos é a informação mais relevante para a busca de complementaridade entre as empresas. Essa informação está representada na Figura 20.

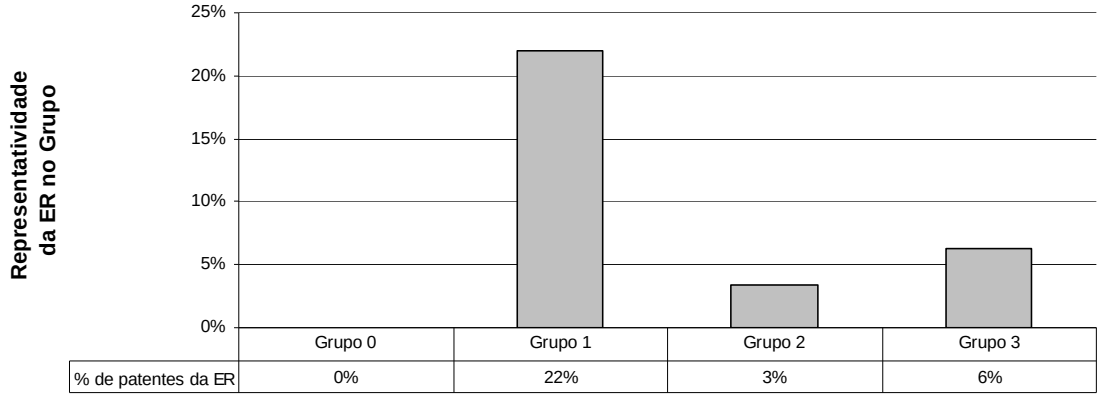


Figura 20 - Representatividade da ER nos grupos.
Fonte: Autoria própria.

O grupo onde a ER é mais presente é o Grupo 1 com 22% de representatividade. Apesar de ser o mais representativo, o Grupo 1 foi o segundo em número de patentes com 11 patentes, das 50 existentes no grupo. O Grupo 3 foi o primeiro em representatividade, porém o primeiro em número de patentes com 14, das 224 existentes no grupo. Para as análises de complementaridade da ER foram utilizados os dados do Grupo 1 uma vez que o Grupo 3 não pôde ser classificado em nenhum tema, o que o invalidou.

4.2.3.4 Análise dos Resultados do Agrupamento

Segundo o modelo matemático da subseção 3.3.4.2, a maior probabilidade de encontrar uma empresa complementar à ER estaria nas empresas que pertencem ao mesmo grupo, ou grupos, onde há grande representatividade da ER.

O Grupo 1 apresentou a maior representatividade das patentes da ER, com 22%. Dentro deste grupo encontram-se as empresas que podem ser as melhores opções como parceiras para a ER. O número de possíveis parcerias fica em 23 empresas e oito inventores independentes. Caso a ER quisesse verificar intelectualmente as patentes das possíveis

parceiras, para identificar alguma patente de seu interesse, o número de patentes diminui de 302 para 39, que é uma diminuição de 87,1% do número de patentes para análise.

A Figura 21 ilustra como foram separados os grupos utilizando a técnica de Agrupamento.

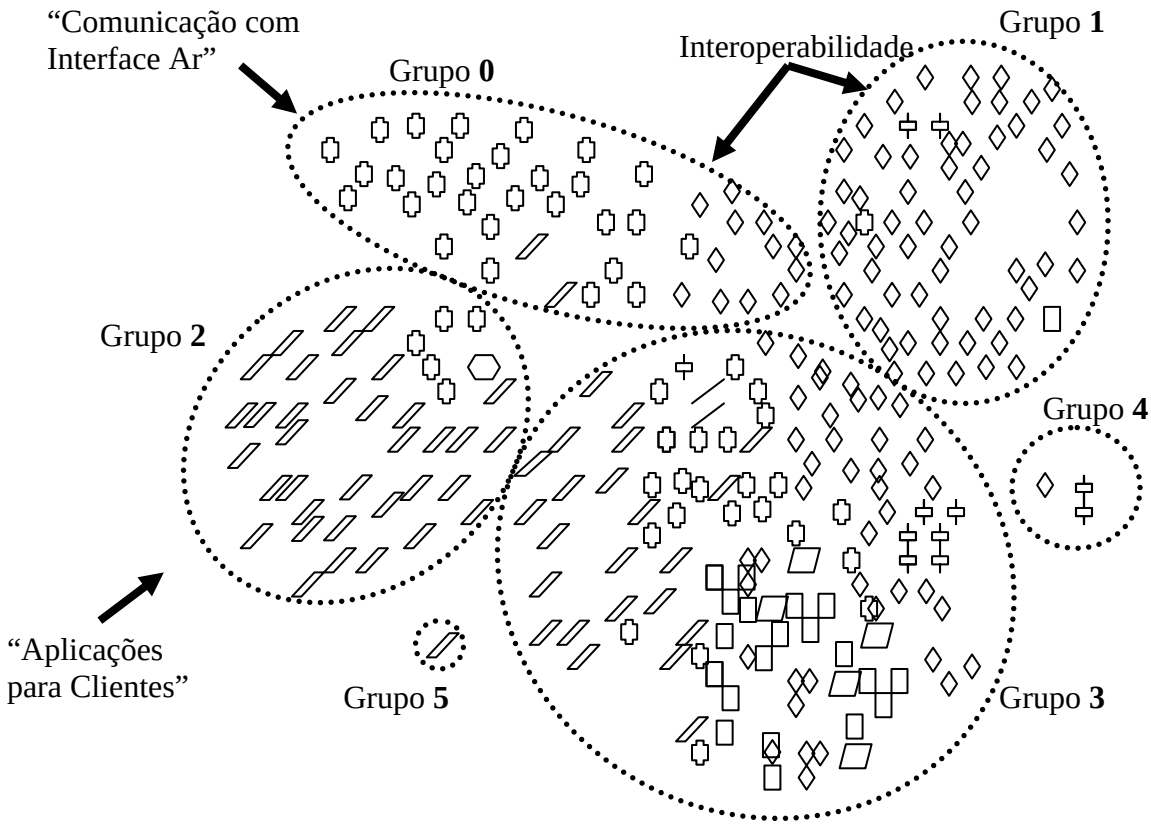


Figura 21 – Ilustração bidimensional do resultado dos agrupamentos.
Fonte: Autoria própria.

4.2.3.5 Considerações do Agrupamento

A análise dos grupos delineados pelo algoritmo *K*-médio foi dispendiosa em tempo. Os resultados apresentam-se plausíveis com a teoria, porém há espaço para melhorias. Dos seis grupos, solicitados ao algoritmo, somente quatro estavam dentro do critério mínimo para análise, e destes quatro somente três estavam dentro do critério mínimo de classificação. Porém, os grupos que puderam ser classificados apresentaram clara tendência nas áreas quais foram dispostos.

Não poder dissociar totalmente os grandes temas entre si para a classificação dos grupos não interfere na validade do método, já que á classificação somente é utilizada para a verificação da convergência dos grupos em um ou mais temas específicos. Isso valida ou invalida um grupo para análise da distribuição das patentes da ER. O mais importante na

metodologia é a distribuição das patentes da ER nos grupos, que são encontradas em grande volume nos Grupos 1 e 3, no total de 11 e 14 unidades, respectivamente, entretanto somente o Grupo 1 pode ser considerado como um grupo válido, pois ele pôde ser classificado em um tema, enquanto que o Grupo 3 foi não conseguiu convergir em nenhum tema.

A técnica de Agrupamento apresenta a utilidade, para a ER, como filtro de patentes que se enquadram dentro do domínio de seu interesse. A metodologia para a busca das patentes, que poderiam interessar à ER, utilizou como parâmetro de filtragem das patentes o código IPC que fosse mais representativo para a empresa, porém esse método retornou 499 patentes, das quais somente 302 continham texto de descrição em inglês, que necessitariam ser verificadas individualmente. Porém, a apartar as patentes em grupos diminuiu o tamanho do universo das patentes de possível interesse para 39, que são as patentes do Grupo 1 que não pertence à ER. Isso diminuiu o grupo para 12,9% do total, de 302 para somente 39 patentes.

Outra possível utilidade desta técnica é para oferecer informações estratégicas para a ER sobre áreas para migração de desenvolvimento. Isto é, caso a ER deseje mudar os rumos de seu P&D devido uma nova estratégia de mercado, mudança de portfólio ou novo plano de negócios. A ER deve localizar a nova área de interesse dentre as áreas apresentadas nos grupos formados, então ela pode firmar parcerias de P&D com empresas que pertençam tanto ao seu grupo, mais representativo, quanto ao grupo da nova área de interesse, e continuar realizando parcerias até adquirir *know-how* suficiente para migrar de área.

4.3 CONSIDERAÇÕES DO DESENVOLVIMENTO E TESTES

Neste capítulo foram apresentados os principais detalhes do levantamento dos dados, a realização dos testes e o desenvolvimento do sistema computacional, que é o objetivo central da pesquisa. Os principais detalhes da pesquisa relatados são os parâmetros utilizados nos diferentes momentos de seu desenrolar e os critérios de decisões relevantes. Essas informações viabilizam a realização de outras pesquisas que necessitem de uma metodologia similar, ou até mesmo igual, da apresentada previamente.

Os documentos para a pesquisa foram coletados via um cliente HTTP e armazenados em base de dados para possibilitar seu rápido acesso quando necessário. Outro fator para a utilização do banco de dados é a estabilidade da aquisição dos dados para o processamento. Uma vez feito, com o máximo de exatidão possível, o processo de busca dos

dados, eles permanecem inalterados para a realização do resto do processo de KDD. Caso os dados devessem ser buscados pela Internet toda vez que uma consulta fosse realizada cada busca poderia retornar um conjunto de registros diferentes. Essa situação não é ideal para a pesquisa, o tratamento de dados é passível de ser reformulado para alcançar a qualidade requerida já que o processo de Mineração de Tecnologia é iterativo.

Os registros de patente foram coletados pelo IPC mais representativo da Empresa Referência. A tarefa de Associação utilizou todos os registros coletados, enquanto que para o Agrupamento foram retiradas as patentes em idioma diferente do inglês e foram adicionadas as patentes da ER do conjunto coletado, para seu processamento.

A preparação dos dados para a Associação resultou na economia de 61% dos vetores e 56% nas dimensões (61% das patentes e 56% nos IPCs). No escopo deste trabalho a redução da dimensionalidade foi essencial no aprendizado na manipulação do *software* Weka, porém poderia ser suprimido já que a quantidade de dados tratados ainda não impossibilitava a busca das regras. Todavia, essa técnica de poda traria ganhos de processamento caso o *itemset* a trabalhar fosse aumentado. Os resultados obtidos pela Associação mostram que é plausível a análise de complementaridade por essa técnica. Um dos resultados corrobora com a busca de padrões de depósito de IPCs e outro apresenta um resultado plausível para as possíveis empresas que podem produzir tecnologias em conjunto com a empresa referência.

A preparação dos dados para a realização do teste utilizando a tarefa de Agrupamento resultou em uma matriz de 328 vetores com 2322 dimensões, isto é, 328 patentes com 2322 termos reduzidos a seu radical, e seu valor sendo o “peso” do termo dentro do conjunto dos dados, variando entre zero e um. Essa tabela de vetores foi processada com o algoritmo *K*-médio, o que resultou em seis grupos, dos quais somente quatro tinham o número mínimo de instâncias para serem considerados válidos para a análise amostral de suas patentes, e dentre esses somente três conseguiram convergir em um tema que atingisse 50% de representatividade entre as patentes. Os temas abordados pelo primeiro grupo foram “Comunicações por Interface Ar” e “Interoperabilidade”, o tema abordado pelo segundo grupo foi, somente, “Interoperabilidade” e o tema abordado pelo terceiro grupo foi “Aplicações para Clientes”. A complementaridade entre as empresas, utilizando a tarefa de Agrupamento, não pode ser verificada diretamente somente probabilisticamente, como apresentado na subseção 3.3.4.2 (Interpretação de Grupos Criados por Algoritmos de Agrupamento, p. 74). Pôde-se observar que a Empresa Referência teve maior representatividade dentro do segundo grupo, o que resulta em 23 empresas e oito inventores independentes como possíveis parceiros em P&D.

A comparação entre os testes de Associação e Agrupamento não convergem para o mesmo resultado. Isto é, nenhuma das empresas que são possíveis parceiras no teste de Associação está contida no conjunto de empresas que podem realizar parceria no resultado de Agrupamento. Porém, isso não invalida o resultado dos testes. A não correlação dos dados pode não ter ocorrido, pois o conjunto de dados utilizado nos testes tornou-se diferente. Inicialmente a intenção da pesquisa era buscar o mesmo conjunto de dados para a realização dos dois testes, e isso foi feito. A metodologia para a busca das patentes e a busca em si foram as mesmas para os dois testes, porém algumas patentes, deste conjunto, não apresentavam a descrição de sua invenção, apesar de apresentarem os IPCs, e isso levou à criação de conjunto de dados com a maior parte do conjunto igual, porém ligeiramente diferentes. Das cinco patentes apontadas pela associação dos IPCs, somente uma patente continha o texto de descrição, e foi agrupada no Grupo 3, que foi descartado da análise por não convergir em um tema. Outro possível motivo para a não convergência dos resultados é que os processos utilizados não são os mesmos, não podendo garantir que o mesmo dicionário de palavras utilizadas para descrever uma patente pudesse classificar essa patente em um mesmo IPC.

As informações resultantes dos testes de Associação e Agrupamento são diferentes, apesar de buscarem o mesmo tipo de Informação Útil. A tarefa de Associação é a aplicação direta do modelo matemático apresentado na subseção 3.3.1, que investiga indícios de interdependência direta entre as tecnologias utilizadas em P&D, o resultado dessas interdependências podem apontar que duas empresas são complementares, desde que uma empresa possua uma tecnologia que seja necessária ou dependente da tecnologia contida na outra empresa. A tarefa de Agrupamento busca por complementaridade probabilisticamente, segundo seu modelo matemático (subseção 3.3.4.2, p. 74) há maior probabilidade de se encontrar empresas complementares entre empresas que têm patentes contidas no mesmo grupo do que entre empresas que têm patentes abrigadas em grupos diferentes. Porém, durante o desenvolvimento desta tarefa nota-se que o Agrupamento pode ser utilizado para filtragem de patentes que podem ser de interesse da Empresa Referência. Patentes que podem ser úteis para os negócios da ER têm maior probabilidade de serem encontradas no grupo, ou grupos, onde a ER tem mais representatividade, que apresentou uma diminuição do universo de busca para 12,9% do montante prévio.

5 CONSIDERAÇÕES FINAIS

O conteúdo exposto por esta dissertação é de cunho interdisciplinar. Os objetivos desta pesquisa são voltados para agregar mais conhecimentos na área de Gestão de Tecnologia, porém o meio para o desenvolvimento do sistema computacional parte de conhecimentos de Ciência da Computação, e para os testes do sistema foram utilizados documentos de patentes da área de Telecomunicações. Essa interdisciplinaridade tornou-se um desafio a mais durante a escrita da dissertação, pois se tentou suavizar os conceitos de todas as áreas envolvidas para que o universo de possíveis interessados pelo texto possa entendê-lo, todavia zelou-se para manter a profundidade teórica mínima no texto de cada área.

Neste capítulo, pretende-se entremear todos os conteúdos anteriormente apresentados, para que os assuntos apresentados, discriminados segundo suas áreas, possam ser tratados como um só. Aqui, faz-se a intersecção de todos os capítulos anteriores, resultando na descrição do caminho que a pesquisa percorreu, desde os estudos mais relevantes da revisão bibliográfica que apoiaram a construção da metodologia e a utilização da metodologia para alcançar os resultados. Posteriormente são apresentadas algumas limitações da pesquisa juntamente com pontos de melhorias. Por fim, são apresentadas algumas sugestões de projetos futuros que poderiam se utilizar desta pesquisa como base.

5.1 RETOMADA DAS CONSIDERAÇÕES DOS CAPÍTULOS E OBJETIVOS

Como apresentado anteriormente, a grande quantidade de informação técnica, necessária para gerar inovações de maior qualidade ou mais rapidamente, impossibilita seu tratamento individualizado. As diversas fontes de informação como revistas especializadas, fóruns de discussão, portfólios de empresas diversas, bancos de patentes entre outros, são importantes fontes de conhecimento que não devem ser deliberadamente desprezadas, pois, apesar de haver em grande quantidade, esses documentos podem trazer conhecimentos necessários para o próximo passo na evolução de uma tecnologia ou de um produto.

O quadro da Inovação Aberta mostra que a P&D não necessita estar encerrada nos muros de uma corporação. Quando necessário, a tecnologia, como também o *know-how* necessário o desenvolvimento de produtos que a utilizam, podem ser comprados, vendidos ou divididos com outras empresas quando a necessidade de inovar é maior que o desejo de defender o próprio *know-how*. Quando uma empresa deseja inovar utilizando o *modus*

operandi proposto pelo quadro da Inovação Aberta, a busca por potenciais parceiros tem o raio de cobertura maior que a rede de clientes, fornecedores e concorrentes, enfim, o *know-who* do gestor de tecnologia. Em teoria, qualquer empresa, cientista ou engenheiro que faz uma pesquisa, independente da área de estudos, é um potencial parceiro de desenvolvimento. Porém, como conseguir informações confiáveis desses pesquisadores, ou de suas tecnologias já que a empresa interessada na tecnologia pode nem, ao menos, saber de sua existência? Ou o caso contrário, quando o desenvolvedor não imagina que seu estudo pode ser utilizado em outra área diferente da sua? Indícios de inovação são apresentados pelas patentes. Esses documentos podem revelar em que uma empresa está trabalhando realmente, onde ela está aplicando seu capital e, também, quais os conhecimentos detidos pela equipe de P&D.

Sendo as patentes uma matéria prima para descobrir empresas com potencial de parceria em P&D, é necessário um método computacional para o tratamento das informações contidas nelas. Para utilizar a computação na busca da complementaridade empresarial foram abordados alguns quadros teóricos. A TRIZ apresentou-se uma opção, já que sua origem vem do estudo individual de inúmeras patentes e tem o objetivo de solucionar problemas inventivos. Supunha-se, inicialmente, que dela pudesse vir uma metodologia que correlacionasse automaticamente o problema inventivo de uma empresa com a tecnologia existente de outra empresa, porém essa abordagem não se mostrou viável dentro do cronograma estipulado para a pesquisa. Outra abordagem seria utilizar uma das diversas ferramentas computacionais de Mineração de Texto e visualização, porém elas são comerciais e não poderiam ter seu código fonte avaliado nem alterado para satisfazer os objetivos desta pesquisa. A bibliometria e a cientometria fornecem ferramentas estatísticas para análise macroscópica de campos científicos, porém não aplicáveis no âmbito desta pesquisa. A Mineração de Tecnologia mostrou-se um quadro bastante flexível para o desenvolvimento de uma metodologia própria e, juntamente, com o quadro da teoria de Complementaridade forneceu as bases para o desenvolvimento da pesquisa. A Mineração de Tecnologia tem a capacidade de expor relações entre informações não triviais e que somente podem ser capturadas computacionalmente.

O quadro teórico descrito acima relata resumidamente o primeiro objetivo específico da pesquisa, que foi compilar um quadro teórico das principais técnicas de manipulação de dados que se adéqüem à prospecção tecnológica. As teorias e ferramentas estudadas e relacionadas na dissertação completam com sucesso esse primeiro objetivo.

Nesta pesquisa, a complementaridade entre empresas é analisada considerando-se os relacionamentos de dependência entre tecnologias, isto é, uma tecnologia para ser criada

depende da existência prévia de outra tecnologia que dá o suporte para a criação da primeira. Assim, atividades que seriam presentes em ambas as tecnologias podem dar sinergia durante o processo de P&D. A Mineração de Tecnologia criou as bases de dois modos para encontrar indícios de possibilidade de parcerias. O primeiro modo é por meio da análise de relações entre os IPCs das patentes das empresas, que é realizada com a tarefa de Associação. O segundo modo fornece as patentes com maior probabilidade de interesse para a Empresa Referência, e ele utiliza a tarefa de Agrupamento das patentes que têm o texto de descrição similar, as patentes de interesse para a referida empresa tendem a ficar dentro de grupos que contém patentes dela com boa representatividade. Esses dois modos de busca automática por complementaridade empresarial foram implementados baseados em uma metodologia que, primeiramente, delimita as patentes as quais serão adquiridas e, em seguida, utilizadas no processamento computacional para a busca de informações úteis, isto é, para a Mineração de Tecnologia. A metodologia desenvolvida realiza o segundo objetivo específico da pesquisa que é de criar uma metodologia para a busca de patentes de empresas que possam ser parceiras em atividades de P&D e uma metodologia para o tratamento computacional dessas patentes.

Com o desenvolvimento da metodologia foi possível a construção do sistema computacional, que é o objetivo final desta pesquisa. O primeiro módulo construído foi o de busca automática de patentes, ele realizava a busca das patentes necessárias para a pesquisa no sítio da Espacenet. Para a extração das informações dessas patentes foi necessário a criação do segundo módulo de *software* que retira essas informações por meio de expressões regulares, assim, para facilitar sua utilização as informações das patentes foram dispostas de forma estruturada em um banco de dados. O último módulo de *software* desenvolvido é responsável por realizar a preparação dos dados, traduzindo eles de informações brutas para tabelas que podem ser tratadas por meio de algoritmos computacionais. Os algoritmos computacionais estão presentes no *software* Weka, livremente disponível na Internet. Com o desenvolvimento desses três módulos de *software* foi alcançado o terceiro objetivo específico desta pesquisa. Os módulos fazem parte do sistema que é o objetivo final da pesquisa, e ele não foi construído limitado a buscar somente os campos de interesse desta pesquisa, mas busca toda informação útil dentro de um registro de patente. O sistema foi projetado modularmente para propiciar a junção de novos módulos, isso deixa um caminho facilitado para outras pesquisas poderem evoluir a partir deste *framework*, até mesmo utilizando novas metodologias de coleta e tratamento de dados. Não é proveitoso apresentar o código fonte do

sistema desenvolvido para retratar seu desenvolvimento, porém, seus resultados de sua operação refletem-se nos testes realizados.

Para a realização dos testes selecionou-se uma empresa ativa em P&D, com um bom portfólio de patentes, e para essa empresa buscou-se empresas que pudessem ser suas parceiras em P&D. Com a empresa escolhida, verificou-se qual a área, definida por um IPC, de maior depósito de suas patentes, então, outras patentes pertencentes à mesma área foram selecionadas, o que resultou em 499 patentes de 274 titulares distintos. Foram realizados dois testes com este conjunto de dados, cada um com suas particularidades de métodos e algoritmos. O primeiro teste realizou a associação dos IPCs das patentes das empresas, para a verificação de existência de dependência entre as tecnologias, e resultou para a Empresa Referência quatro empresas, do universo inicial de 274, como possíveis parceiras em atividades de P&D. O segundo teste reuniu as patentes que continham texto com características similares em grupos distintos, durante o teste, 328 patentes formaram seis grupos, dos quais somente três grupos foram considerados válidos para a análise de complementaridade. Dois destes grupos foram desqualificados, pois não continham o número mínimo de cinco patentes e o outro retirado da análise porque não convergiu em um tema que o classificasse. Nos grupos válidos, a ER tinha maior representatividade em somente um grupo, então as patentes pertencentes a esse grupo, que não são da ER, são consideradas com maior probabilidade de ser de interesse para ela. Apesar das duas técnicas serem oriundas da mesma metodologia, elas não convergiram em um mesmo resultado, pois, apesar das patentes utilizadas para os testes serem da mesma fonte, a informação de algumas patentes resumia-se à primeira página do documento, que tem somente a informação necessária para o primeiro teste, enquanto que algumas apresentavam patente completa, incluindo o texto de descrição que é necessário para o segundo teste, o que resultou em conjuntos de dados diferentes para os dois testes. Com a realização dos testes, acima descritos, o quarto objetivo específico da pesquisa, realização de teste com exemplo real, foi finalizado com sucesso.

Vê-se que a técnica de Associação avalia a complementaridade empresarial no nível microscópico, enquanto que a técnica de Agrupamento avalia a complementaridade no nível macroscópico. As técnicas devem ser utilizadas focando meios diferentes de aplicação. A técnica Associativa pode ser utilizada para as empresas que não imaginam para qual tecnologia elas podem ser complementares, na prática se resume que a empresa tem algumas tecnologias e gostaria de saber quem são os possíveis compradores ou parceiros que necessitam daquela tecnologia. A técnica de Agrupamento caberia melhor para empresas que já tem a idéia de um novo produto ou pesquisa, porém gostariam de vislumbrar quem seriam

os melhores parceiros para a compra de tecnologia, parceria em desenvolvimento, licenciamento de tecnologia, entre outros. Outro uso para o Agrupamento é para a filtragem de patentes que seriam de seu interesse, o que no teste reduziu o número de patentes para análise para 12,9% do total. Com a avaliação dos testes o quinto objetivo específico da pesquisa foi finalizado com sucesso. Os testes foram importantes para a avaliação da consistência da metodologia e a qualidade da implementação do sistema. Seus resultados mostram-se plausíveis com o modelo matemático e com a teoria apresentada. O escopo do teste se resume na utilização do sistema para o levantamento de critérios que são somente possíveis com a experimentação prática utilizando dados reais. Entretanto, não se podem garantir resultados positivos na aplicação desta metodologia na vida real, pois ela não foi validada junto à ER, assegurando, assim, que os resultados obtidos seriam realmente as melhores ou, pelo menos, boas opções de parcerias em P&D.

Por fim, com os cinco objetivos específicos finalizados, o objetivo geral da pesquisa é alcançado naturalmente. Relembrando-o: **desenvolver um sistema informatizado de avaliação de dados contidos em bases de patentes para obtenção de indicadores de perspectivas de cooperação para inovações tecnológicas no contexto da Inovação Aberta**. Os resultados obtidos mostram que a metodologia é promissora para a busca de informações úteis no campo da gestão de tecnologia e que a automação de verificação de complementaridade empresarial é viável. A metodologia foi desenvolvida tentando-se limitar as variáveis e guiá-las por critérios, o que pode facilitar uma nova aplicação desta em um novo conjunto de dados, possibilitando a repetibilidade dos experimentos.

Esta pesquisa tem potencial de oferecer às empresas inovadoras ferramentas que auxiliam na busca por parcerias que melhor se adéquam às suas pesquisas. Empresas que buscam inovar dentro do contexto da Inovação Aberta são as empresas que podem lucrar mais com a utilização dos resultados desta pesquisa. Dentro da sociedade da informação, o domínio de técnicas de busca automática de informações úteis em grande quantidade de dados pode ser o diferencial competitivo decisivo para o sucesso de uma empresa, principalmente para as desenvolvedoras de tecnologia. Para elas, a oportunidade de novas aplicações para suas pesquisas podem gerar os dividendos que ajudam no aprimoramento da tecnologia nos momentos mais delicados do início de seu ciclo de vida.

5.2 LIMITAÇÕES DA PESQUISA E PONTOS DE MELHORIAS

A pesquisa teve em seu decorrer algumas limitações que a restringiram em algum ponto, ou também são pontos que podem ser melhorados, porém não foram realizados por limitações de cronograma. Tanto as limitações quanto as melhorias são apresentadas nessa seção para serem um relato das dificuldades, explicando o porquê de algumas ações não serem tomadas e também um alerta para as pesquisas subseqüentes sobre pontos de dificuldade.

5.2.1 Limitações da Pesquisa

As principais limitações estão relacionadas nos tópicos abaixo:

- A impossibilidade de adquirir documentos de patente inteiramente em língua inglesa. Isso é devido o portal da Espacenet não suportar, pelo menos durante o período de coleta de dados, a língua de publicação como um parâmetro na busca.
- A limitação de acessos por parte do portal da Espacenet. O acesso ao sítio é interrompido após algumas requisições de HTTP, o que é um obstáculo para a aquisição de mais dados e atualização dos já existentes no Banco de Dados.
- A fragilidade do gestor de arquivos, percebida durante a busca dos dados no portal da Espacenet. Foi necessário supervisionar seu funcionamento para garantir a qualidade das buscas. O que, também, limita a aquisição de dados.
- A ausência de P&D da Empresa Referência no Brasil. No começo da pesquisa a ER possuía laboratório de P&D no Brasil, porém ele foi encerrado em junho de 2009. Isso não limita esta pesquisa, pois a aplicação de questionário para a empresa não estava no cronograma inicial, porém, isso limita uma segunda pesquisa que poderia validar os resultados apresentados anteriormente.
- A incapacidade de estender a interpretação dos resultados focando diretamente o mercado. Com a utilização de patentes não é possível buscar informações úteis do mercado, para isso deve-se utilizar informações de tecnologia que existem em produtos.

5.2.2 Pontos de Melhoria para a Pesquisa

Os principais pontos de melhora para a pesquisa são:

- Aumentar o número de patentes utilizadas no sistema. A pesquisa limitou-se a 499 patentes de possíveis parceiras. O ideal seria buscar todas as patentes de todas as empresas que tivessem qualquer vínculo com qualquer IPC da ER. Entretanto, essa abordagem tornar-se-ia custosa para o Gestor de Arquivos, por ainda ser frágil. Somente o IPC mais representativo da empresa resultou em 274 titulares diferentes, então os outros 21 IPCs da empresa referência poderiam fornecer mais inúmeros titulares distintos que detém inúmeras patentes. O ideal seria investir mais tempo em melhorias no Gestor de Arquivos para torná-lo mais robusto.
- Utilizar as redes de co-citação das patentes na criação de um *ranking* das redes de conhecimento que melhor serviriam a empresa referência. O sistema desenvolvido já tem o ferramental para separação das citações, então essa abordagem não seria de difícil desenvolvimento. Porém, isso aumenta a dificuldade na metodologia e análise dos dados.
- Limitar os dados do sistema para utilizar somente patentes que apresentem tanto o IPC como o texto de descrição da invenção. A possível convergência de resultados entre os testes de Associação e Agrupamento não pôde ser verificada devido alguns dos registros que fizeram parte do conjunto de dados e apresentavam a informação dos IPCs, porém não continham texto de descrição. Isso fez com que somente algumas patentes não aparecessem em ambos os testes, apesar das patentes serem provenientes da mesma busca.
- Buscar por uma teoria matemática que forneça um parâmetro sobre a representatividade de uma empresa em uma determinada área, que é uma falta que pode ocorrer na análise dos resultados do Agrupamento. Quando foi escolhido um grupo para que a Empresa Referência pudesse assumir as outras patentes dentro do grupo como de interesse para a ela, o que foi apresentado na Figura 20 (p. 94). Poderia acontecer, por exemplo, de um grupo apresentar 50% de suas patentes como a ER, porém essas mesmas patentes representam somente 1% da empresa, certamente as outras patentes deste grupo não podem ser consideradas mais interessantes para a empresa do que um grupo que contém 10% de patentes da empresa, porém essas representem 50% da empresa. Nesta pesquisa essa questão não interfere no resultado,

pois o grupo onde a empresa foi mais representativa também foi onde havia mais patentes da empresa.

5.3 SUGESTÕES PARA TRABALHOS FUTUROS

Esta pesquisa é exploratória, o que a torna uma fonte natural de futuras pesquisas. Os trabalhos que são propostos abaixo são ponto de melhoria para a técnica proposta por esta dissertação. Abaixo são listadas algumas sugestões de trabalhos que podem ser realizados tendo esta pesquisa como ponto de partida:

- Validação da proposta realizada por esta pesquisa utilizando-se de questionários aplicados a Gestores de Tecnologias de empresas de P&D para avaliação da busca automática por parcerias. Isso pode ser realizado selecionando-se uma nova empresa como Empresa Referência, e então, utilizaria o sistema computacional para a busca das possíveis parcerias. Esses resultados podem ser validados por meio de questionários a serem respondidos pelos responsáveis pela gestão de tecnologia ou responsáveis pela tomada de decisão sobre parcerias na empresa.
- Utilização de outras fontes de dado para a avaliação de P&D, como: artigos científicos, portfólio de produtos, manuais técnicos dos produtos. Caso isso seja factível, possibilitaria a verificação de mercados e potenciais parceiros no desenvolvimento de produtos em conjunto, não somente atividades de P&D como apresentado nesta pesquisa.
- Utilização de outros algoritmos de Associação e Agrupamento na busca de novos resultados ou na melhoria no desempenho dos resultados já alcançados.
- Criação de uma nova metodologia para Associação de IPCs baseada na hierarquia desses campos, pode-se utilizar os vários níveis da hierarquia do IPC na busca por parcerias. Isto é, procurar a interdependência tecnológica os níveis superiores da hierarquia do IPC. Ainda, outra variante da metodologia pode ser criada se considerar a marcação do IPC como: “avançado”, “central”, “relacionado à invenção” ou “não relacionado”.
- Utilização das experimentações de análise de grupos apontadas no APÊNDICE C como forma de automatizar totalmente a metodologia de testes de Agrupamento.

- Utilização das técnicas de *Natural Language Processing* (NPL) e/ou Mineração Semântica para potencializar a criação e diferenciação de grupos na utilização da tarefa de Agrupamentos.

REFERÊNCIAS

ALTSHULLER, G. S.; **Creativity as An Exact Science** – The Theory of The Solution of Inventive Problems. 1a. ed. Luxemburg: Gordon & Breach, 1984 (1a ed. russa, 1979).

ANPROTEC, Associação Nacional de Entidades Promotoras de Empreendimentos de Tecnologias Avançadas SEBRAE - Serviço Brasileiro de Apoio às Micro e Pequenas Empresas. **Glossário dinâmico de termos na área de Tecnópolis, Parques Tecnológicos e Incubadoras de Empresas**. Brasília, setembro de 2002. Disponível em: http://www.anprotec.org.br/ArquivosDin/GLOSSARIO_pdf_12.pdf. Acessado em 1 de abril de 2010.

AZARIAS P.; MATOS S. N.; SCANDELARI L.; **Aplicação da mineração de dados para geração de conhecimento**: um experimento prático. In: V Congresso Nacional de Excelência em Gestão. Niterói, RJ, Brasil, 2, 3 e 4 de julho de 2009. ISSN 1984-9354.

BERMAN, Bruce. **Making Innovation Pay**: People Who Turn IP into Shareholder Value. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

CANONGIA, C.L. ; SANTOS, D. M. ; SANTOS, M. M. ; ZACKIEWICZ, M.; **Integração entre Inteligência Competitiva, Gestão do Conhecimento e Visão do Futuro**: Reflexão Sobre o Sistema de Prospecção Tecnológica e do Conhecimento para o Setor de Ciência, Tecnologia e Inovação Brasileiro. In: 3o. Workshop Brasileiro de Inteligência Competitiva e Gestão do Conhecimento, 2002, São Paulo. 3o. Workshop Brasileiro de Inteligência Competitiva e Gestão do Conhecimento, 2002.

de CARVALHO, Marco Aurélio. **Metodologia IDEATRIZ para a Ideação de Novos Produtos**. 2007. 232 f. Tese (Doutorado) – Programa de Pós-Graduação em Engenharia da Produção, Universidade Federal de Santa Catarina, Florianópolis, 2007.

CASSIMAN, Bruno; VEUGELERS, Reinhilde. **In Search of Complementarity in Innovation Strategy**: Internal R&D and External Knowledge Acquisition. Management Science. Vol. 52, No. 1, January 2006, pp. 68–82.

CIOS, Krzysztof *et al.* **Data Mining:** A Knowledge Discovery Approach. 4a. ed, Los Angeles. 500p. Hardcover. 2007. ISBN: 978-0-387-33333-5.

CHESBROUGH, Henry W. **The Era of Open Innovation**. MIT Sloan Management Review. Vol. 44, No.3, Spring 2003, pp. 35-41.

CHESBROUGH, Henry W. **Open Business Models**: How to Thrive in the New Innovation Landscape, Boston: Harvard Business School Press, 2006, ISBN: 1-42210-427-3.

CHESBROUGH, Henry W.; SCHWARTZ, Kevin. **Innovating Business Models with Co-Development Partnerships**. Research Technology Management. Vol. 50, No. 1, January/February 2007. pp. 55-59.

CHRISTENSEN, Clayton. Patterns in the Evolution of Product Competition. European Management Journal Vol. 15, No 2, pp. 117-127, April 1997. Elsevier Science Ltd. Great Britain.

CHRISTENSEN, Clayton. **The Innovator's Dilemma**: When New Technologies Cause Great Firms to Fail. Harvard Business School Press. Boston, Massachusetts. 1997. ISBN 0-87584-585-1.

CHRISTENSEN Clayton, RAYNOR Michael. **The Innovator's Solution**: Creating and Sustaining Successful Growth. Boston: Harvard Business School Press, 2003. 301p.

CLEMENTE, Rafael. **Gestão da Inovação**. In: Curso Avançado de Propriedade Intelectual para NITs, 25-03-2008, Curitiba.

CONG, He; TONG Loh Han. **Grouping of TRIZ Inventive Principles to facilitate automatic patent classification**. Expert Systems with Applications v. 34, p. 788-795. 2008.

CPAN. Disponível em: http://www.cpan.org/ . Acessado última vez dia 09/junho/2009.

DASH, M.; LIU, H. **Feature Selection for Classification**, In: Intelligent Data Analysis. 1997, March. Singapore: Vol 1, p. 131-156.

DERGINT, Dario Eduardo Amaral. **Apprentissage Colletif et Milieux Innovateurs:** Étude de Cas à Grenoble et Simulations Multi-Agents. 1999. 567 f. Tese (Doutorado) em Economia. Université de Technologie de Compiègne. Unité d'accueil COSTECH. Compiègne, 1999.

DERGINT, Dario Eduardo. Notas de aula, Tecnologia e Sustentabilidade. 2008.

DIETTERICH, Thomas. **Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms**. Corvallis: Oregon State University, 1997.

ESPACENET. Disponível em: http://gb.espacenet.com/search97cgi/s97_cgi.exe?Action=FormGen&Template=gb/en/advanced.hts . Acesso em 4 de março de 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **The KDD process for extracting useful knowledge from volumes of data**. Communications of the ACM, 1996. v. 39(11) p. 27–34.

FATTORI, Michele; PEDRAZZI, Giorgio; TURRA, Roberta. **Text Mining Applied to Patent Mapping: a practical business case**. Word Patent Information v. 25, p. 335-342, 2003.

FORAY, Dominique. **Patents and Development in the Knowledge Economy,** In: Views on the Future of the Intellectual Property System. 2007, June. Selected Issue Briefs No. 1, ICTSD Programme on Intellectual Property Rights and Sustainable Development, International Centre for Trade and Sustainable Development, Geneva, Switzerland.

GENBAND. Disponível em: http://www.genband.com/Home/About/At-a-Glance.aspx. Acesso em 17 de outubro de 2009.

GHOLAMIAN, M.R.; GHOMI, Fatemi. **Meta Knowledge of intelligent manufacturing:** An overview of state-of-the-art. Applied Soft Computing, v. 7, 2007 p. 1-16.

GLENISSON, Patrick; GLÄNZEL, Wolfgang; JANSSENS, Frizo; De MOOR, Bart. **Combining full text and bibliometric in mapping scientific disciplines**. Information Processing and Management. 2005, v. 41, p. 1548-1572.

GOFFMAN, W. **Mathematical approach to the spread of scientifics ideas:** the history of mast cell research. Nature, [s.1], v. 212, p.449-452, Oct. 1966.

GOFFMAN, W.; NEWILL, V. A. **Generalization of epidemic theory:** an application to the transmission of ideas. Nature, [s.l.], v. 204, n. 4955, p. 225-228, Oct. 1964.

GUEDES, V. L. da S. ; Borschiver . **Bibliometria:** uma revisão da literatura dessa ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. In: VI CINFORM, 2005, Salvador. Anais Eletrônico do VI CINFORM. Salvador -BA : UFBA, 2005.

HOLMSTROM, Bengt; MILGROM, Paul. **The firm as an incentive system**, American Economic Review, 1994. v. 84, p. 972-991.

HUSTON, Larry; SAKKAB, Nabil. **Connect and Develop**: Inside Procter & Gamble's New Model for Innovation. Harvard Business Review, March 2006.

JAIN A. K.; MURTY M. N.; FLYNN P. J.. **Data Clustering: A Review**. ACM Computing Surveys, v 31, no 3, Setembro 1999.

JOHNSON, Björn.; LUNDVALL, Bengt.-Åke. **Promoting Innovation Systems as a Response to the Globalising Learning Economy**. Paris: Organisation for Economic Co-Operation and Development, 2000.

KOHAVI, Ron. **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. In Internation Join Conference of Artificial Inteligence (IJCAI), 1995.

MACIAS-CHAPULA, C. A. **O papel da informetria e da cienciometria e sua perspectiva nacional e internacional**. Ciência da Informação, Brasília, v. 27, n. 2, p. 134-140, maio/ago. 1998.

MARCH, J.G.; **Exploration and Exploitation in Organizational Learning**, Organization Science. 1991,Vol. 2, No. 1, 71–81.

MARTINS, C. J. M.; **Aplicação de ferramentas computacionais para prospecção tecnológica por Mineração de Dados não-estruturados sobre patentes industriais em idioma inglês**. 2008. 191 p. Dissertação (Mestrado em Engenharia Civil) Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, 2008.

MILGROM, P., ROBERTS, J. **The Economics of modern manufacturing:** Technology, strategy, and organization. Americam Economic Review, 1990, v. 80 p. 511-528.

ORGANISATION FOR ECONOMIC CO-OPERATIONS AND DEVELOPMENT. **Manual de Oslo: Diretrizes para Coleta e Interpretação de Dados sobre Inovação**, Terceira Edição, 1997.

ORGANISATION FOR ECONOMIC CO-OPERATIONS AND DEVELOPMENT. **Patent Manual: Using Patent Data as Scientific and Technology Indicators**. 1994, The Organisation for Economic Co-Operation and Development, Paris.

ORGANISATION FOR ECONOMIC CO-OPERATIONS AND DEVELOPMENT. **The Knowledge-Based Economy**. In: Science, Technology and Industry Outlook 1996. 1996, The Organisation for Economic Co-Operation and Development, Paris.

PAO, M. L. **Concepts of information retrieval**. Englewood, Colorado:Libraries Unlimited, Inc., 1989. 285 p.

PHELPS, Marshall. **Turning a Patent Portfolio into a Profit Center**, In: Making Innovation Pay: People Who Turn IP into Shareholder Value. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

POLTORAK, Marshall. **On Patent Trolls and Other Myths**, In: Making Innovation Pay: People Who Turn IP into Shareholder Value. John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

PORTER, Alan; CUNNINGHAM, Scott W. **Tech Mining**: Exploiting New Technologies for Competitive Advantage, New Jersey: Wiley Interscience, Hoboken, 2005.

PORTER, . **An Algorithm for Suffix Stripping Program**, Vol. 14 1980, no. 3 , pp 130-137.

PRIBERAM. Endereço eletrônico: http://www.priberam.pt/dlpo/. Acesso em: 9 de junho de 2009.

ROCHA, Ivan. **Ciência, Tecnologia e Inovação: Conceitos Básicos**. Brasília: SEBRAE, 1996.

RULEQUEST. http://www.rulequest.com/see5-info.html. Acesso em 10 de junho de 2008.

SAVRANSKY, Semyon. **Engineering of Creativity:** Introduction to TRIZ Methodology of Inventive Problem Solving. CRC Press LLC, Corporate Blvd., Boca Raton, Florida. 2001. ISBN 0-8493-2255-3.

SILVA, Edna L.; MENEZES, Estera. **Metodologia da pesquisa e elaboração de dissertação.** Terceira edição. Laboratório de ensino a distância da UFSC. Florianópolis: 2001.

SHAPIRO, Gregory P. **Winner of SIGKDD Data Mining and Knowledge Discovery Service Award**. 28-06-2005. Endereço eletrônico http://www.kdnuggets.com/news/2005/n13/2i.html, acessado em 29/12/2009.

SEBASTIANI, Fabrizio. **Machine Learning in Automated Text Categorization**. In: ACM Computing Surveys, Vol. 34, No. 1, March 2002. pp 1-47.

SCHUMPETER, Joseph A. **The Theory of Economic Development**, Harvard University Press, 1934. Cambridge, Massachusetts.

SCHUMPETER, Joseph A. **Capitalismo, Socialismo e Democracia**. Edição Brasileira Zahar Editores S.A., Rio de Janeiro. 1984.

SIGKDD. Endereço eletrônico http://www.sigkdd.org/. Acessado em 27/12/2009.

SIGGELKOW, N. **Misperceiving Interactions among Complements and Substitutes**: Organizational Consequences. Management Science, 2002, v. 48, p. 900-916.

SLOCUM, Michael. **Technology Maturity Using S-curve Descriptors**. TRIZ Journal, 1999. April. Vol 4.

SLOCUM, Michael; LUNDBERG, Catherine O. **Technology Forecasting**: From Emotional to Empirical. Creativity and Innovation Management, Vol. 10, Number 2, June 2001.

SOURCEFORGE.NET. Endereço eletrônico http://www.sourceforge.net, acessado em 27/12/2009.

STERNITZKE, Christian; BARTKOWSKI, Adam; SCHRAMM, Reinhard. **Visualizing patent statistics by means of social network analysis tools.** World Patent Information, 2008. Vol 30 p 115-131.

STURN, Wilerson; **Avaliação do potencial de uso da lógica fuzzy para a identificação de indicadores de competências no currículo Lattes**. 2005. 103 f. Dissertação (Mestrado em Tecnologia) – Programa de Pós-Graduação em Tecnologia. Centro Federal de Educação Tecnológica do Paraná, Curitiba, 2005.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining.** Addison-Wesley. Cloth, 769 f, ISBN-10: 0321321367, 2006.

TARTARUS. Endereço eletrônico http://tartarus.org/~martin/PorterStemmer/, acessado em 13/12/2008.

TESSERACT. Endereço eletrônico http://code.google.com/p/tesseract-ocr/. Acessado em 07/06/2008.

VANTI, Nadia Aurora Peres. **Da Bibliometria à Webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro de informação e a difusão do conhecimento**. Ci. Inf., Brasília, v. 31, n. 2, p. 152-162, maio/ago. 2002.

VARGAS, Milton. **Metodologia da Pesquisa Tecnológica**. 1985, Rio de Janeiro: Globo.

YANG Y *et al.*, **Text mining and visualization tools – Impressions of Emerging Capabilities**, World Patent Informat (2008), doi:10.1016/j.wpi.2008.01.007

YOON, Byungun; PARK, Yongtae. **A text-ming patent network**: Analitical tool for high-technology trend. Journal of High Technology Management Research, v. 15, p. 37-50. 2004.

WAIKATO, Universidade de. Endereço eletrônico: http://www.cs.waikato.ac.nz/~ml/weka/ . Acesso em 09/06/2009.

WEKA, Waikato Environment for Knowledge Analisys, Version 3.6.1. The University of Waikato Hamilton, New Zealand, 2009.

WIKIPÉDIA. Endereço eletrônico: http://pt.wikipedia.org/ e http://en.wikipedia.org. Acesso em: 01 de abril de 2010.

WORLD INTELLECTUAL PROPERTY ORGANIZATION. Endereço eletrônico: www.wipo.int. Acesso em: 4 setembro 2008.

WORLD INTELLECTUAL PROPERTY ORGANIZATION. **International Patent Classification**: (Version 2009), Endereço eletrônico: http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc_2009.pdf . Acesso em: 8 setembro 2009.

GLOSSÁRIO

Agrupamento – É a tarefa de Mineração de Dados que separa em grupos relevantes os dados fornecidos para análise, levando em consideração as relações entre esses grupos.

Apartar – Neste texto tem o significado de separar por algum critério.

Associação – É a tarefa de Mineração de Dados que é útil para descobrir relacionamentos interessantes escondidos entre os atributos de grandes quantidades de dados.

Atividades de Ciência e Tecnologia (C&T) - Qualquer trabalho relacionado com a geração, o avanço, a difusão e aplicação de conhecimento científico e técnico em todos os campos da atividade humana. Constitui atividades de ciência e tecnologia a pesquisa básica, científica, aplicada ou tecnológica; o desenvolvimento experimental de produto ou processo; a administração da pesquisa científica e tecnológica; os eventos técnicos e científicos; a educação e o treinamento em ciência e tecnologia e os serviços de apoio à pesquisa (ANPROTEC, 2002).

Ativos - Bens, direitos e valores pertencentes a uma pessoa ou empresa (ANPROTEC, 2002).

Banco de dados - Acervo de informações e dados coletados de pesquisa, planilhas, relatórios e publicações, reunidos em arquivo manual ou eletrônico para uso da organização em estudos e tomada de decisões (ANPROTEC, 2002).

Base de conhecimento - Acervo de informações, experiências e conhecimento utilizados como a principal fonte de recursos para o desenvolvimento da empresa (ANPROTEC, 2002).

Booleano – ver lógica booleana.

Capital de risco (*Venture capital*) - Investimento temporário em empresas emergentes com evidente potencial de crescimento: participação direta no capital social da empresa por aquisição de ações ou debêntures conversíveis em ações visando rentabilidade acima das alternativas disponíveis no mercado financeiro (ANPROTEC, 2002).

Centro de pesquisa e desenvolvimento (P&D) ou instituto de P&D - Organização que abriga atividades de estudos empíricos e laboratórios (ANPROTEC, 2002).

Cliente - Indivíduo ou organização que se beneficia de serviço ou produto ofertado (ANPROTEC, 2002).

Complementaridade – Relação matemática entre duas entidades, que permite concluir se uma é o complemento da outra. Complemento é a parte que se junta à outra, para formar um todo completo (PRIBERAM, 2009).

Computação Evolucionária - É um ramo da ciência da computação que tem por base os mecanismos evolutivos encontrados na natureza. Esses mecanismos estão diretamente relacionados com a teoria da evolução de Darwin (WIKIPEDIA, 2010).

Confiança – É o atributo da tarefa de associação que determina qual freqüência um item do lado direito de uma regra é freqüente juntamente com o item, ou itens, do lado direito da regra.

Conhecimento científico - Competência que se adquire através da pesquisa ou investigação científica, seguindo as etapas da metodologia científica e que dão origem a teorias explicativas dos fenômenos estudados (ANPROTEC, 2002).

Contrato - Instrumento jurídico celebrado entre pessoas físicas com fins de aquisição, modificação ou extinção de direitos ou estabelecimento de obrigações recíprocas (ANPROTEC, 2002).

Cooperação tecnológica - Forma de colaboração entre empresas e Instituições de Ensino e Pesquisa para o desenvolvimento de produtos e processos quando a tecnologia usada não pode ser efetivamente transferida através da venda do direito de utilização ou da simples transferência de informações. Implica melhoria das condições de trabalho, do meio ambiente, da assistência técnica e da reciclagem (ANPROTEC, 2002).

Default – Quando alguma instância de um grupo não pode ser enquadrada em qualquer das formas pré-existentes, porém deve ser enquadrado em algo, este algo é o *default*.

Discretização – Processo de transformar por aproximações infinitos valores contínuos em finitos valores individuais.

Economia do Aprendizado – As economias modernas são caracterizadas como economia do aprendizado, onde o conhecimento é um recurso crucial e o aprendizado é o mais importante processo para sua aquisição. Nesta economia, tanto o aprendizado quanto o "esquecimento" são processos importantes para a diferenciação das empresas.

Empresas Possíveis Parceiras (EPP) – São empresas quais têm alguma patente que está classificada com o IPC igual ao mais representativo da Empresa Referência. Durantes os testes da pesquisa elas são as possíveis parceiras para a Empresa Referência.

Empresa Referência (ER) – Empresa escolhida para a realização dos testes da pesquisa. Para ela são encontradas possíveis parcerias em P&D.

Expressões Regulares - São seqüências de caracteres que descrevem um padrão textual. Assim, ele provê para *softwares* de tratamento de texto uma forma flexível de encontrar palavras, números, grupos de caracteres ou padrões de caracteres em um texto.

Fomento - Aplicação de recursos orçamentários governamentais em atividades diversas relacionadas à pesquisa científica e tecnológica (ANPROTEC, 2002).

Força Bruta - Chama-se de "força bruta" o método computacional que utiliza pouca inteligência e gera a maior parte, senão todas, as combinações possíveis de uma determinada tarefa até achar uma opção que fornece um resultado que é melhor do que os outros.

Gestão - (a) Ato de gerir; administração; gerenciamento; (b) planejamento, organização, liderança e controle das pessoas que compõem uma empresa e das tarefas e atividades por elas realizadas (ANPROTEC, 2002).

Gestão da inovação tecnológica - (a) Conjunto de atividades da função gerencial que coordena esforços para apoiar a criatividade dos seus membros e prover contextos de pesquisa e desenvolvimento para que eles gerem novos produtos e processos; (b) integração dos princípios e métodos de administração, avaliação, economia, engenharia, informática e matemática aplicada ao processo de inovação tecnológica (ANPROTEC, 2002).

Informações Úteis – No campo de mineração de dados e busca de informações, as informações úteis são conhecidas como "conhecimento". Para o campo de gestão de tecnologia elas são informações que podem trazer vantagens competitivas para uma empresa, podendo trazer ganhos no P&D.

Inovação - Introdução no mercado de produtos, processos, métodos ou sistemas não existentes anteriormente, ou com alguma característica nova e diferente daquela até então em vigor, com fortes repercussões socioeconômicas (ANPROTEC, 2002).

Inovação de produtos e processos tecnológicos (PPT) - Adoção de métodos de produção e colocação no mercado de produtos novos ou aprimorados, resultantes do uso de novo conhecimento, mudanças de equipamento e/ ou de organização da produção (ANPROTEC, 2002).

Inovação incremental - Introdução em uma empresa, sem alteração da sua estrutura industrial, de qualquer tipo de melhoria em produto, processo ou organização da produção (ANPROTEC, 2002).

Inovação organizacional - Renovação de procedimentos e métodos de organizar empresas, fornecedores, produção e comercialização de bens e serviços (ANPROTEC, 2002).

Inovação radical - Introdução de novo produto ou processo ou renovação da forma de organização da produção que pode resultar em ruptura estrutural com o padrão tecnológico até então utilizado, dar origem a novas indústrias, setores ou mercados (ANPROTEC, 2002).

Inovação tecnológica - Introdução de produtos e processos tecnologicamente novos ou aprimorados (ANPROTEC, 2002).

Inteligência Artificial (IA) - é uma área de pesquisa da ciência da computação dedicada a buscar métodos ou dispositivos computacionais que possuam ou simulem a capacidade humana de resolver problemas, pensar ou, de forma ampla, ser inteligente (WIKIPEDIA, 2010).

Interdependência Tecnológica ou **Interdependência de tecnologias** – É o relacionamento existente entre duas tecnologias onde para existir uma é necessário que a outra exista previamente ou, então, mutuamente.

IPC (Classificação Internacional de Patentes) – Codificação hierarquizada que tem finalidade de classificar as tecnologias encontradas em uma patente.

Joint.venture - Forma de aliança inter.empresarial que objetiva a criação de novo negócio, para atuação em mercados conjugados na comercialização de produtos ou na complementa ção de projetos de desenvolvimento de produtos. É normalmente estabelecida entre uma empresa com capital necessário ao financiamento do projeto, e outra que domina as competências técnicas, os contatos comerciais, ou ambos. Nesse sentido, a franquia pode ser considerada como uma espécie de *joint-venture* (ANPROTEC, 2002).

Knowledge Discovery in Databases (KDD) - Processo não-trivial de identificação de dados novos, válidos, potencialmente úteis, e dispostos em padrões dentro de banco de dados (TAN, STEINBACH e KUMAR, 2006).

***Know-how* –** É a habilidade ou a capacidade de fazer algo. Como por exemplo, a capacidade de um empresário analisar as perspectivas de mercado para um novo produto, ou então, a capacidade de um operário operar uma maquinaria complicada. No caso específico do texto o *know-how* é aplicado ao conhecimento raro, que não é de fácil obtenção, que um cientista ou engenheiro detém sobre a operação de uma determinada tecnologia. Esse tipo de conhecimento é uma das mais importantes razões da formação de redes industriais, pois as empresas necessitam dividir e combinar elementos de *know-how* (ORGANIZATION..., 2006).

***Know-who* -** É o conhecimento que envolve a informação de quem sabe alguma coisa e quem sabe fazer alguma coisa. Ele depende da formação de uma rede social de relacionamentos, que possibilita chegar rapidamente aos especialistas em cada assunto (ORGANIZATION..., 1996).

Licenciamento de tecnologia - Acordo contratual pelo qual uma organização vende a outra empresa os direitos de uso de tecnologia de sua propriedade, sob a forma de patentes, processos e/ou know-how técnico e pelo qual recebe pagamentos de royalties e/ou outra forma de compensação (ANPROTEC, 2002).

Lógica Booleana – Também conhecida como lógica binária, é a lógica matemática onde os elementos são definidos somente como 0 ou 1, verdadeiro ou falso, existente ou não existente.

Lógica Fuzzy ou lógica difusa – É uma forma de lógica imprecisa. Diferentemente da lógica booleana, a lógica *fuzzy* pode variar entre 0 e 1 podendo ser um valor impreciso.

Mineração de Tecnologia (MT) – É o processo de KDD utilizado exclusivamente para a extração de informações úteis para a gestão de inovação, a partir de bases de dados científico-tecnológicas.

Nicho de mercado - (a) segmento de mercado especializado, com características próprias e que oferecem oportunidades de negócios para empreendimentos específicos; (b) Segmento específico de mercado que a empresa se propõe atender com produtos ou serviços extremamente ajustados às suas necessidades; (c) Oportunidade de negócio originário de segmentos privilegiados de mercado (ANPROTEC, 2002).

Normalização – É o processo computacional de apresentar os valores dos atributos de um conjunto de dados de forma que fiquem entre 0 e 1, para que o processamento conjunto, com os outros atributos dos dados, não apresentem interferência devido a grandeza das variáveis.

***Outsourcing* -** Terceirização: forma de transferir para outras empresas a realização de tarefas e/ou serviços, ou a fabricação de produtos de que uma empresa necessita (ANPROTEC, 2002).

Pacote de Dados – O pacote de dados é a menor unidade de informação que pode ser transmitida pela Internet, ou por qualquer comunicação em redes. Toda informação carregada por um pacote deve ser dividida na origem e remontada no destino. Um pacote típico contém até 1500 bytes de informação. Eles são formados basicamente por três estruturas: cabeçalho (onde carrega a informação de origem, destino, tamanho e etc) , corpo (onde carrega a informação) e o rodapé (que contém os bits que avisam o término do pacote).

Patente - Os registros de patente são títulos que conferem o monopólio temporal de um determinado conhecimento a uma pessoa física ou jurídica, limitado geograficamente.

Pesquisa e Desenvolvimento (P&D) - Normalmente se refere a atividades de longo prazo e/ou orientadas ao futuro, relacionadas a ciência ou tecnologia, usando técnicas similares ao método científico sem que hajam resultados pré-determinados mas com previsões gerais de algum benefício comercial (WIKIPEDIA, 2010).

Portfólio - Conjunto dos investimentos feitos por um mesmo indivíduo ou organização (ANPROTEC, 2002).

Processo de inovação - Procedimento interativo para o qual contribuem vários agentes econômicos e sociais e que consiste na conjugação de oportunidades técnicas com as necessidades de um empreendimento. Tem por finalidade a introdução ou modificação de produtos ou processos para comercialização (ANPROTEC, 2002).

Produtividade - (a) Maximização dos resultados da empresa através da otimização dos recursos utilizados; (b) medida da eficiência de uma empresa ou organização na utilização de recursos, calculada através da divisão da produção física obtida numa unidade de tempo por um dos fatores de produção (trabalho, bens, capital) (ANPROTEC, 2002).

Propriedade industrial - Conjunto de direitos relacionados com atividades industriais ou comerciais do indivíduo ou da empresa relativos a marcas e patentes (ANPROTEC, 2002).

Propriedade intelectual - Toda espécie de propriedade que provenha de concepção ou produto da inteligência para exprimir um conjunto de direitos que competem ao intelectual (escritor, artista ou inventor) como autor de obra imaginada, elaborada ou inventada. No sentido lato, o poder irrestrito de autor ou criador sobre bem imaterial. Torna-se restrita, se condicionada a prerrogativas de tempo e espaço. O título de propriedade intelectual pode ser concedido nas categorias: artística, técnica, e científica (ANPROTEC, 2002).

Prospecção tecnológica - (a) Tentativas sistemáticas para observar, no longo prazo, o futuro da ciência, da tecnologia, da economia e da sociedade, com o propósito de identificar tecnologias emergentes que possam produzir benefícios econômicos e/ou sociais; (b) . Estudos sobre tendências tecnológicas em setores industriais específicos, utilizando principalmente informações contidas em documentos de patentes nacionais ou estrangeiros (ANPROTEC, 2002).

Protocolo - É o conjunto bem definido de instruções que dois sistemas computacionais trocam para que realizem operações acordadas entre ambos.

Queries - São linhas de comando com as instruções necessárias para requisição de informações existentes em Banco de Dados.

Rede de empresas (*network*) - empresas que interagem entre si - como fornecedores, clientes, ou parceiros na transferência de tecnologia - e/ou com centros de pesquisa, centros técnicos, universidades e outras entidades públicas ou privadas a fim de aumentar a sua competitividade, resolver problemas, entrar em novos mercados, desenvolver e produzir bens e serviços (ANPROTEC, 2002).

***Royalties*-** a importância paga ao detentor ou proprietário ou um território, recurso natural, produto, marca, patente de produto, processo de produção, ou obra original, pelos direitos de exploração, uso, distribuição ou comercialização do referido produto ou tecnologia (WIKIPEDIA, 2010).

Semente Aleatória – É um número qualquer fornecido para inicializar um programa gerador de números pseudo-aleatórios.

Sistema Computacional – Nesta dissertação um sistema computacional é um conjunto de programas de computador que conseguem alcançar um objetivo computacional, isto é, processar um conjunto de dados somente se utilizados em conjunto.

Sociedade da Informação ou Sociedade do Conhecimento - Configuração de padrão sócio-técnico-econômico em que as atividades humanas estão baseadas e organizadas em torno de atividades de geração, recuperação e uso de informação e conhecimento. Na sociedade da informação, o sucesso das empresas está relacionado ao volume dos seus ativos intangíveis (ANPROTEC, 2002).

Software – Conhecido em português como programa de computador.

Spin-off - Empresa oriunda de laboratório e resultante de pesquisa acadêmica ou industrial (ANPROTEC, 2002).

Suporte – É o atributo da tarefa de associação que define a freqüência com que uma regra é aplicada dentro do conjunto de dados.

Tecnologia - (a) Método para transformar inputs em outputs; (b) aplicação dos resultados de pesquisa científica à produção de bens e serviços; (c) tipo específico de conhecimento, processo ou técnica exigido para fins práticos; (d) conhecimentos de que uma sociedade dispõe sobre ciências e artes industriais, incluindo os fenômenos sociais e físicos, e sua aplicação à produção de bens e serviços. Identificam-se duas grandes categorias de tecnologia: tecnologia de produto: componentes tangíveis e facilmente identificáveis e tecnologia de processo: técnicas, métodos e procedimentos (ANPROTEC, 2002).

Transferência de tecnologia - Intercâmbio de conhecimento e habilidades tecnológicas entre instituições de ensino superior e/ou centros de pesquisa e empresas. Faz-se na forma de contratos de pesquisa e desenvolvimento, serviços de consultoria, formação profissional, inicial e continuada, venda de patentes, marcas e processos industriais, publicação na mídia científica, apresentação em congressos, migração de especialistas, programas de assistência técnica, espionagem industrial e atuação de empresas multinacionais (ANPROTEC, 2002).

URL (*Uniform Resource Locator*) – É o localizador de um recurso dentro de uma rede. Isto é, pode ser um arquivo, impressora, banco de dados entre outros. Sua estrutura é `<protocolo>://<máquina>/<caminho>/<recurso>`.

Valor agregado - (a) Procedimento através do qual uma empresa adquire e melhora produto ou serviço antes de oferecê-lo a seus clientes; (b) conhecimento embutido num produto, serviço ou processo (ANPROTEC, 2002).

Vantagem competitiva - Conjunto de fatores fundamentais que influem na diferenciação de produtos e processos num ambiente de concorrência econômica (ANPROTEC, 2002).

APÊNDICE A - Descrição Detalhada do Teste Piloto

Para fazer o teste piloto foi escolhida a tarefa de Mineração de Dados de classificação de patentes de duas áreas distintas, uma vez que elas já são inerentemente organizadas segundo a área de atuação. O percentual de acertos na classificação correta mostra a qualidade do sistema e seu potencial de uso. Um modo de realizar essa prova de conceito foi a implementação de dois *scripts* na linguagem Perl. Um para transformar os textos de registros de patente do formato PDF para o formato TXT, e outro para utilizar esses registros em TXT e classificá-los segundo sua origem.

Caso o pequeno sistema conseguisse ler o texto das patentes em formato PDF, o pior caso de fonte de patentes, e conseguisse classificar com uma pequena margem de erro as patentes, ele se mostraria robusto o suficiente para o tratamento de outras fontes de dados. E também se mostraria robusto o suficiente para fazer outras tarefas de mineração de dados, como agrupamentos e associações.

Durante a apresentação do teste piloto primeiramente será introduzido o referencial teórico do processo de Classificação, na seção A.1. A ferramenta para a execução do teste é apresentada na seção A.2. Então o primeiro *script* para a realização dos testes é apresentado, que é a conversão dos arquivos PDF para TXT é apresentado na seção A.3. A realização do teste de Classificação é apresentado na seção A.4 e os resultados são apresentados na seção A.5. Finalmente as conclusões são apresentada na seção A.6.

A.1 CLASSIFICAÇÃO

A Classificação é a tarefa de Mineração de Dados que tem o objetivo de classificar, ou “apartar”, automaticamente os elementos do conjunto de dados de acordo com as definições de tipos, ou rótulos, previamente fornecidos. Porém para a realização desta tarefa deve-se fornecer um conjunto de dados, previamente rotulados, para que o algoritmo de classificação possa “aprender” quais características podem distinguir os diferentes tipos das instâncias do conjunto. Como a função chave de um algoritmo de classificação é aprender uma função, eles também são chamados de algoritmos de aprendizagem supervisionada.

Classificação também pode ser utilizada como um modelo de previsão para classificar registros desconhecidos. O objetivo de um algoritmo de aprendizagem é construir modelos com boa capacidade de generalização. Um conjunto de treinamento (*trainning set*) é

utilizado para construir o modelo de classificação e o conjunto de teste (*test set*) é utilizado para validar o modelo. O conjunto de testes nunca pode ser utilizado para treinar o modelo.

A qualidade do desempenho de um modelo de classificação pode ser dado pela matriz de confusão, que é uma tabela baseada na contagem dos registros corretamente preditos e os registros incorretamente preditos. A Tabela 4 apresenta um exemplo de Matriz de Confusão para classificação de duas classes. A contagem *f11* representa os exemplos com Classe = 1 que foram corretamente classificadas como Classe = 1. A contagem *f10* indica quantos exemplos com Classe = 1 que foram erroneamente classificadas como Classe = 0. As funções *f00* e *f01* são as correspondentes para a Classe = 0.

Tabela 4 - Exemplo de Matriz de Confusão.
Fonte: Adaptado de Tan, Steinbach e Kumar, 2006.

		Classes Preditas	
		Classe = 1	Classe = 0
Classe Atual	Classe = 1	f_{11}	f_{10}
	Classe = 0	f_{01}	f_{00}

Outras métricas de desempenho são a precisão (*accuracy*) e a taxa de erro (*error rate*). O total de exemplos classificados corretamente é dado por *f11 + f00*, e o total de exemplos classificados erroneamente são dados por *f01 + f10*, o que nos leva a:

$$\text{Precisão} = \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{10} + f_{01}} \tag{16}$$

$$\text{Taxa de Erro} = \frac{f_{01} + f_{10}}{f_{11} + f_{00} + f_{10} + f_{01}} = 1 - \text{Precisão} \tag{17}$$

Uma característica relevante para a escolha do algoritmo é a sua simplicidade. Caso dois algoritmos tenham o mesmo desempenho, deve-se escolher o mais simples, porque apresenta, provavelmente, maior poder de generalização.

A.1.1 Algoritmo Naïve Bayes

O método de aprendizado de máquina de Naïve Bayes é uma simplificação do método Ótimo de Bayes. Este método fornece uma abordagem probabilística para a aprendizagem e está baseado na suposição de que as quantidades de interesse são reguladas por distribuição de probabilidade. Teoricamente para ocorrer a simplificação do método

Ótimo de Bayes para o Naïve Bayes há a obrigação de que os dados analisados sejam condicionalmente independentes, porém na prática a simplificação do método com algumas correlações fracas entre os atributos não atrapalham o processamento do algoritmo.

A.1.2 Árvores de Decisão

Os algoritmos de árvore de decisão classificam as instâncias de um conjunto de dados por meio de uma árvore de opções de valores das características das instâncias, onde cada opção de valores torna-se um ramo. Essa árvore é construída por meio do conjunto de dados chamados de dados de treinamento, que “treinam” o algoritmo para reconhecer novas instâncias e classificá-las. Então, para classificar uma instância desconhecida, segue-se a árvore comparando os seus valores com o da instância, então quando se chega ao final dos ramos há uma classificação para a instância.

O Quadro 7 apresenta uma árvore de decisão para a “apartar” patentes relacionadas à área de mecânica com a área de telecomunicações. Os termos foram tratados com *stemming*, por isso estão em seu radical, e seus valores são resultado da aplicação do TFIDF no conjunto de patentes. Ao verificar os termos de uma patente, após a realização de *stemming* e o TFIDF, pode-se classificá-lo ou como pertencente às patentes de mecânica, ou telecomunicações

```
combust > 0: mecânica
combust <= 0:
|___pressur > 8.23: mecânica
    pressur <= 8.23:
    |___shaft > 0: mecânica
        shaft <= 0:
        |___air > 7.12: mecânica
            air <= 7.12:
            |___rotat > 0:
                |___pair <= 0: mecânica
                |   pair > 0: telecomunicações
                rotat <= 0:
                |___mechan > 12.22:
                    |___direct <= 15.1: mecânica
                    |   direct > 15.1: telecomunicações
                    mechan <= 12.22:
                    |___open <= 8.42: telecomunicações
                        open > 8.42:
                        |___character <= 0: telecomunicações
                            character > 0: mecânica
```

Quadro 7 - Árvore de decisão para classificação entre patentes de mecânica ou telecomunicações.

Fonte: Autoria própria.

A técnica de aprendizagem de árvores de decisão é, para aprendizagem indutiva, um dos métodos mais utilizados. É indicado para valores discretos, mas também pode ser aplicada a elementos contínuos. Para aquele se utiliza o algoritmo ID3 e para este utiliza-se sua extensão o algoritmo C4.5.

A.2 *SOFTWARE* SEE 5.0

Além do software Weka, já descrito anteriormente, o outro software utilizado para o teste piloto foi o See5.0 (RULEQUEST, 2008), ele é uma ferramenta para plataforma Windows desenvolvida pela *Rulequest Research data mining tools*™ para o treinamento de árvores de decisão. Sua versão para uso livre tem limitação para 400 exemplos de treinamento, o que torna essa ferramenta passível de uso nesse trabalho, uma vez que somente o teste piloto não contém uma quantidade massiva de instâncias. No total são 240 instâncias que estão disponíveis para treinamento da árvore de decisão. Para a utilização deste *software* é necessário a criação de dois tipos de arquivos, um arquivo que contém o nome dos atributos com seus possíveis valores e extensão .names, e outro arquivo, com extensão .data, que contém os dados de cada instância separado por “,” (vírgulas) e a classe a que pertence cada objeto.

A.3 CONVERSÃO DOS ARQUIVOS PDF PARA TXT

A base de dados para o experimento de mineração de dados são 240 arquivos de registro de patente em formato PDF publicados em língua inglesa. Os arquivos foram coletados no sítio da *World Intellectual Property Organization* (WORLD..., 2009). Esses arquivos foram divididos igualmente em duas partes de 120 patentes de duas classes: telecomunicações e mecânica.

Todos os arquivos de telecomunicações têm o código de classificação internacional **H04B-1/00**[21]. As patentes de engenharia mecânica não foram pegas da mesma

[21] *ELECTRIC COMMUNICATION TECHNIQUE* (H04) *Transmission* (B) *Details of transmission systems not characterized by the medium used for transmission* (1/00).

classe internacional, porém foram pegas da mesma seção **F**[22]. Assim sendo, 75 patentes IPC **F02B**[23], 17 de IPC **F01C**[24], 15 de IPC **F02C**[25] e 13 IPC **F03C**[26]. Vê-se que as classes não têm muitas características em comum.

É relevante mostrar nesses registros de patente que o intervalo de tempo na publicação das patentes de telecomunicações vai desde 18/05/2006 até 15/05/2008, aproximadamente dois anos de depósitos. O que pode significar que a linguagem utilizada nessas patentes pode não ser tão diferente entre si quanto a linguagem utilizada nas patentes de engenharia mecânica, que têm suas publicações mais espalhadas no tempo, a qual vai de 21/01/1982 até 28/02/2008, são mais de 26 anos de depósitos. Isto somado ao fato de que essas patentes somente são da mesma seção, mas não de mesma classe internacional, faz surgir a idéia de que esse conjunto de dados está mais espalhado que o conjunto de telecomunicações, e é mais propenso a falhas quanto a sua classificação. Isso será discutido posteriormente na seção A.6 (Considerações do Teste Piloto).

Cada arquivo PDF teve que passar por um tratamento para poder ser testado. Era necessário que o arquivo PDF fosse transformado para o formato TXT. Assim, o sistema poderia tratar a informação contida em cada documento. Porém, durante a implementação dos testes verificou-se que a extração de texto dos arquivos em PDF não seria acessível, pois os registros eram figuras, como se o texto tivesse sido escaneado[27]. Porém, com a utilização do *software* de reconhecimento de caracteres Tesseract (TESSERACT, 2009) o texto pode ser extraído. O caminho seguido pelo *script* em Perl para a transformação do PDF em TXT está representado pela Figura 22, e listado passo-a-passo abaixo:

- Abrir o arquivo .pdf;
- desmembrar suas páginas em arquivos individuais;
- converter essas páginas do formato .pbm para .tiff;
- converter as imagens .tiff em arquivos texto;
- agrupar todos os arquivos texto individuais em um só;

---

[22] *MECHANICAL ENGINEERING; LIGHTING; HEATING; WEAPONS; BLASTING*.
[23] *Internal-combustion piston engines; Combustion engines in general*.
[24] *Rotary-piston or oscillating-piston machines or engines.*
[25] *Gás-turbine plants; air intakes for jet-propulsion plants; controlling fuel supply in air breathing jet-propulsion plants.*
[26] *Positive-displacement engines driven by liquids.*
[27] alguns registros, principalmente os mais antigos, eram datilografados e foram escaneados para estar dentro dos arquivos PDF.

Após sucessivas conversões os dados apresentam muito ruído. As duas fontes principais de ruído são as figuras inseridas no documento, que acabam sendo reconhecidas como texto, e erro de reconhecimento de caracteres. O exemplo mais comum disso é o reconhecimento da letra “e” como “c”. Muito desse ruído é filtrado durante o pré-processamento dos dados, como veremos a seguir. Entretanto, algum ruído pode passar pelo pré-processamento. Portanto, o método de classificação deve ser robusto o suficiente para suportar esses ruídos.

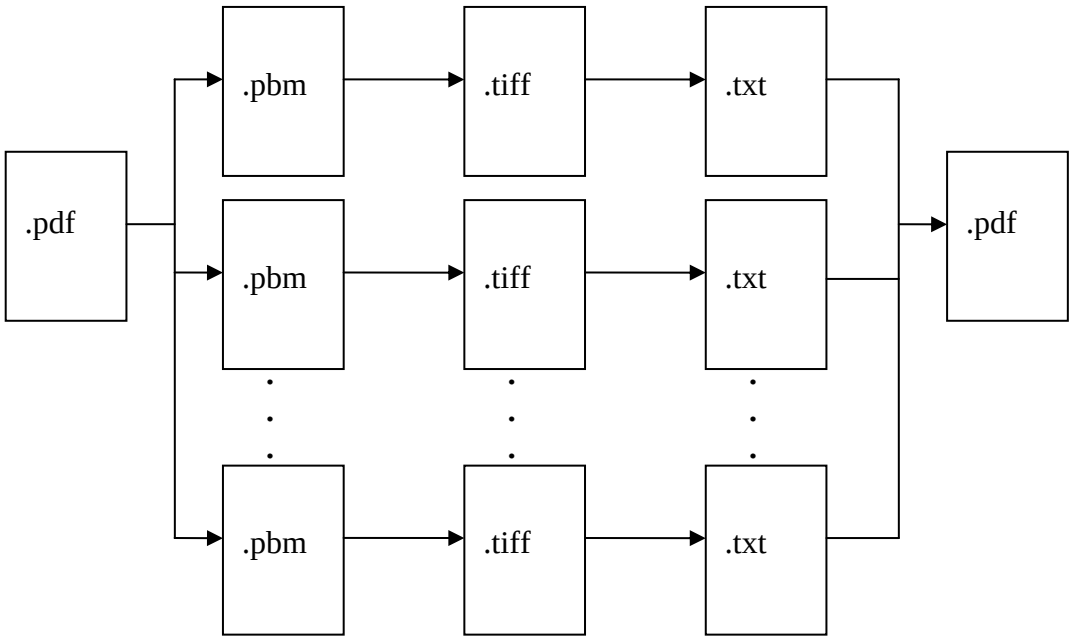


Figura 22 - Esquema de transformação de arquivo .pdf para .txt

A.4 DESENVOLVIMENTO DA CLASSIFICAÇÃO

Essa seção apresentará de forma resumida todo o processo de KDD. Não é intuito apresentar o processo detalhadamente, pois a metodologia é a mesma descrita no capítulo 3. O *script* para a realização da classificação foi escrito em Perl. Seu trabalho foi realizar a poda de dados e organização dos mesmos para a entrada do software de mineração, neste caso Weka e See5.0.

Os passos que o *script* executa são:

- Filtragem das palavras com menos de três caracteres;
- filtragem das *stop-words*;
- *stemming*;
- filtragem das palavras com ocorrência máxima e mínima;

- contagem das palavras;
- realização do TFIDF;
- criação das tabelas de nome e dados.

A filtragem das palavras tem o objetivo de retirar palavras sem tem significância dentro do contexto da mineração. Porém, diferentemente do conceito de *stop words*, acaba fazendo a filtragem de muito do ruído que acontece na conversão de imagem para texto.

As palavras, no âmbito deste trabalho, foram definidas como o ajuntamento de três ou mais caracteres de “A” até “Z”, sendo maiúsculas ou minúsculas, podendo ser unida por apóstrofes. Então, toda palavra que contenha número, seja menor que três caracteres, ou seja, separada por hífen não é reconhecida, neste caso cada palavra será reconhecida separadamente.

Outras palavras que são filtradas do conjunto de atributos são as palavras que têm ocorrência menor que dez vezes ou maior que 120 vezes. Essas palavras não contribuem na diferenciação entre as classes, pois contém pouca informação, devido sua raridade ou sua vulgarização. Esses valores foram escolhidos empiricamente por meio de sucessivos testes e análise subjetiva. A filtragem das palavras com pouca representação é importante para filtrar os ruídos da conversão dos arquivos de imagens para arquivos de texto. Quando há este processo os textos formados são inteligíveis, contendo caracteres não alfanuméricos e probabilisticamente difíceis de ocorrerem mais de uma vez. Portanto, esses textos são filtrados, devido à sua baixa representatividade.

Depois da filtragem, foram realizados os processos de retirada de *stop-words*, *stemming*, e o cálculo do TFIDF de cada palavra. Esses procedimentos já foram descritos na metodologia, não necessitando de mais aprofundamento em seus conceitos.

O resultado desse processamento são 240 vetores com 2494 dimensões cada um. Isso é, materializado em uma tabela de 240 linhas e 2494 colunas, onde cada linha representa um registro de patente e cada coluna representa um radical de palavra utilizada no texto e o valor do eixo é o TFIDF de cada palavra, ou o “peso” dessa palavra com relação às outras dentro do conjunto de obras.

A mineração de dados foi realizada com os *softwares* Weka e See5.0. Os algoritmos utilizados em cada *software* foi Naïve Bayes e ID3, respectivamente. Os testes foi utilizada a validação cruzada para auxílio na avaliação da exatidão dos métodos.

A.1.3 Validação Cruzada Para comparação de Classificações

A tarefa de Mineração de Dados de Classificação, dentro do escopo dessa pesquisa, será utilizada para verificar a viabilidade da implementação do sistema com o Teste piloto. Para tal, ela será interpretada por meio da estatística de seus erros e acertos. A técnica utilizada para esse fim é a validação cruzada.

A validação cruzada é uma forma de avaliar a exatidão e comparar métodos de classificação, principalmente quando o conjunto de dados para treinamento dos métodos não é grande suficiente para dispersar qualquer tipo de propensão dos dados. Na validação cruzada de *k*-partes o conjunto de dados D é dividido aleatoriamente em *k* conjuntos exclusivos *(D1, D2, ..., Dk)* com aproximadamente o mesmo tamanho. A cada tempo $t \in \{1,2,3,...,k\}$ é "treinado" o conjunto de dados D sem o subconjunto D*t*, posteriormente testado em D*t*. A exatidão da validação cruzada é o número de classificações corretas divididas pelo número de instâncias no conjunto de dados. O *k* escolhido para a validação cruzada dos testes, do teste piloto, foi de tamanho moderado, 20, pois reduz a variância enquanto aumenta a tendência (KOHAVI, 1995).

A.5 APRESENTAÇÃO DE RESULTADOS DO TESTE PILOTO

Apesar de diferentes, a apresentação dos resultados tanto do Weka como no See5.0, podem ser comparadas sem maiores problemas.

O grupo de patentes de telecomunicações foi retirado de um intervalo de tempo relativamente pequeno (2 anos) em relação ao das patentes de engenharia mecânica (aproximadamente 26 anos). Os registros de telecomunicações são da mesma seção e classe, contrastando com os de mecânica que são somente de mesma seção. Isso faz pensar que os dados de telecomunicações podem estar mais aglomerados e os de engenharia mecânica estão dispersos. Todavia, as matrizes de confusão, ilustradas na Tabela 6 e Tabela 9, mostram que a classe Telecom teve classificação errada em três casos para o classificador de Naïve Bayes, enquanto que a classe Mechanical não tem nenhuma classificação errônea. Para a árvore de decisão feita pelo *software* See5.0 a classificação foi de oito casos para ambas.

Tabela 5 - Erros médio e quadrático utilizando See5.0.
Fonte: Autoria própria.

Erro médio	6.70%
SE	1.60%

Tabela 6 - Matriz de confusão para See5.0.
Fonte: Autoria própria.

Classificado como →	Telecom	Mechanical
Telecom	112	8
Mechanical	8	112

A Tabela 5 apresenta o erro médio, de 6,7% e o erro quadrático de 1,6%, a Tabela 6 apresenta a matriz de confusão para a execução do *software* See5.0, em que das 120 patentes de telecomunicações 112 foram classificadas corretamente, enquanto 8 foram classificadas como de mecânica. E das 120 patentes de mecânica, 112 foram classificadas corretamente, enquanto 8 foram classificadas erroneamente como de telecomunicações.

Tabela 7 - Sumário de classificação para Naïve Bayes.
Fonte: Weka, 2009.

Correctly Classified Instances	237	98.75%
Incorrectly Classified Instances	3	1.25%
Kappa statistic	0.975	
Mean absolute error	0.0125	
Root mean squared error	0.1118	
Relative absolute error	2.50%	
Root relative squared error	22.36%	
Total Number of Instances	240	

Tabela 8 - Exatidão detalhada por classe para Naïve Bayes.
Fonte: Weka, 2009.

Class	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F-Measure*
Telecom	0.975	0	1	0.975	0.987
Mechanical	1	0.025	0.976	1	0.988

Tabela 9 - Matriz de confusão para Naïve Bayes.
Fonte: Autoria própria.

Classificado como →	Telecom	Mechanical
Telecom	117	3
Mechanical	0	120

A Tabela 7 apresenta o sumário da classificação realizada por meio do *software* Weka. Nela há o número total de classificações corretas, 237, e incorretas, três. Sendo que as classificações incorretas são inteiramente das patentes de telecomunicações. Todas as patentes relacionadas à área de mecânica foram classificadas corretamente, como apresentado pela matriz de confusão (Tabela 9).

O método de Naïve Bayes resultou em melhores resultados, se contrastado com Árvores de Decisão. O erro médio absoluto daquele foi de 1,25%, enquanto que este foi de 6,7%, e o tempo de execução foi 5,27 vezes menor.

A.6 CONSIDERAÇÕES DO TESTE PILOTO

A estratégia de desenvolvimento da pesquisa inicia com a implementação do teste piloto, que agregou *know-how* suficiente para o desenvolvimento do sistema de Mineração de Tecnologia. O teste piloto apresentou-se satisfatório, demonstrando que patentes poderiam ser utilizadas como matéria-prima para a pesquisa, mesmo que oriundas de antigos documentos datilografados posteriormente escaneados.

O processo de KDD teve seu esforço muito concentrado nas fases de preparação e poda dos dados. O pré-processamento e a poda geram arquivos de alta dimensionalidade, 2494 atributos que são formados dos radicais das palavras encontradas no universo dos textos. A grande dimensão dos vetores não foi um problema no processamento. O treinamento das árvores e a validação cruzada foram rápidas, não mais que alguns segundos. Mas isso se deve ao pequeno número de instâncias para treinamento dos métodos.

O fato das patentes estarem em formato PDF gerado a partir de imagem não impossibilitou a realização do KDD, porém foi o único gargalo dentro do processo. A implementação tem uma lógica complexa, e dificuldade média. Entretanto, o código foi bem implementado, o que frutificou nos bons resultados. Para todo esse processo foi utilizada a linguagem de *script* Perl, que se mostrou muito eficaz para o tratamento de textos. Primeiro, porque é intrínseco dela e bem desenvolvido o tratamento de expressões regulares, o que traz vantagens na seleção de texto. Outra facilidade apresentada pela linguagem são as várias bibliotecas para tratamento de texto que foram desenvolvidas especificamente para mineração de textos, as quais podem ser encontradas na página da organização *Comprehensive Perl Archieve Network* (CPAN, 2008).

Pela análise mais aprofundada das várias árvores criadas pelo *software* See5.0, onde se encontra o exemplo abaixo, percebe-se que os nós das árvores são fieis a cada tipo de classe classificada. Os termos que são apontados por cada nó correspondem à terminologia de cada área. Isso traz mais credibilidade aos *scripts* de pré-tratamento das informações.

Por meio de uma análise subjetiva o teste piloto pode ser considerado bem sucedido, porque apresentou de forma clara que é possível realizar com sucesso a mineração

de textos em registros de patentes. As taxas de erro médio absoluto de 1,25% e 6,7% para os algoritmos de Naïve Bayes e ID3, respectivamente, são satisfatórias para o teste, e são resultado suficiente para motivar a implementação de um sistema mais amplo e complexo.

APÊNDICE B - Modelo Matemático para Limitação na Interpretação dos Resultados

O seguinte modelo matemático será utilizado para demonstrar a validade dessas proposições de limitações.

Premissa: Segundo Dergint (1999) todo produto tecnológico é constituído de uma ou mais tecnologias que o compõe. Dado um produto tecnológico *p* e para tecnologias *t*, podemos afirmar que: $p = \{t_1, t_2, ...t_n\}, n \in \mathrm{N}$.

Caso o produto em questão não fosse tecnológico, sendo ele realizado por tarefas derivadas de conhecimento ubíquo, ele poderia ser reduzido às atividades envolvidas no processo de fabricação e o tempo de trabalho em cada uma. Porém, como o objeto em questão é tecnológico a maior parte do seu valor deriva da tecnologia envolvida em sua fabricação, reduze-o à tecnologia que ele utiliza ou é utilizada em sua fabricação. Nessa pesquisa o único motivo que levaria uma empresa a querer realizar parceria é por que outra empresa detém o conhecimento (tecnologia) que ela não possui. Outros motivos de parceria como entrada em novos mercados, ajuda financeira, ajuda empresarial entre outros, estão fora do escopo desta pesquisa.

O mercado *M* é definido como o conjunto de produtos *p* pertencentes a este mercado. Então, o mercado de produtos tecnológicos é representado por: $M = \{p_1, p_2, ..., p_m\}, m \in \mathrm{N}$.

Pode-se definir uma empresa tecnológica pela sua tecnologia encontrada em seus produtos. O que significa que a tecnologia detida por uma empresa é igual à tecnologia detida em seus produtos. Nessa definição não é considerado o fato da empresa comprar tecnologia de outrem: $E \approx \bigcup_{i=0}^{n} pi = \{t_1, t_2, ..., t_l\}$.

Teorema: Tendo este quadro, e sendo *Tn* subconjuntos de tecnologias pode-se afirmar que uma Parceria em desenvolvimento quando a Equação 2:

$$\exists T_1 \subseteq E_1 : \exists T_2 \subseteq E_2 \Rightarrow P\{T_1 \cup T_2\} \in M \tag{18}$$

Existe um subconjunto de tecnologias da primeira empresa (T_1) é unido a um subconjunto de tecnologias da segunda empresa (T_2), gerando um produto, que está no mercado.

Esse modelo não leva em conta a qualidade do produto oferecido, nem a quantidade que uma tecnologia influencia no valor final. Esse modelo também não oferece

nenhuma possibilidade de sucesso da tecnologia. Porém, um modelo tentando considerar o sucesso da tecnologia será modelado nas próximas linhas.

Hipótese: Considerando o mercado (*M*), é plausível pensar que a combinação de tecnologias que está gerando a maior parte dos produtos de sucesso, dentro desse mercado, é utilizada mais de uma vez. Então, será chamado de *M'* o mercado de produtos de com potencial sucesso:

$$M' = \{p'_i \mid p'_i \subseteq p_x, p'_i \subseteq p_y, ..., p'_i \subseteq p_m, (p_x, p_y, ..., p_m) \in M\}, 0 \leq x < y \leq m \qquad (19)$$

Onde, *p'* é um subconjunto dos produtos *p*, porém para existir ele deve estar presente em outros produtos encontrados no mercado *M*. O número mínimo de produtos em que o subconjunto deve estar presente é determinado pelo observador do conjunto.

Para utilização de registros de patente na análise de produtos deve-se tentar aproximar, ou encontrar uma relação entre as patentes e os produtos. Denomina-se o conjunto de patentes dentro do mercado de $I=\{t_1,t_2,...,t_n\}$, onde cada patente *i* é formada por um sub-conjunto de tecnologias.

Premissa: Pode-se afirmar uma patente pode ser utilizada em vários produtos, e um produto pode se utilizar de várias patentes. Como ilustrado pela Figura 23. Que resulta na equação abaixo:

$$p = \bigcup_{i=k}^{n} i_i = \{t_1, t_2, ..., t_m\} \qquad (20)$$

Caso a Equação 20 fosse substituída na Equação 19 o resultado de *M'* não poderia ser obtido a não ser se conhecesse quais patentes *i* fazem parte de cada produto *p*.

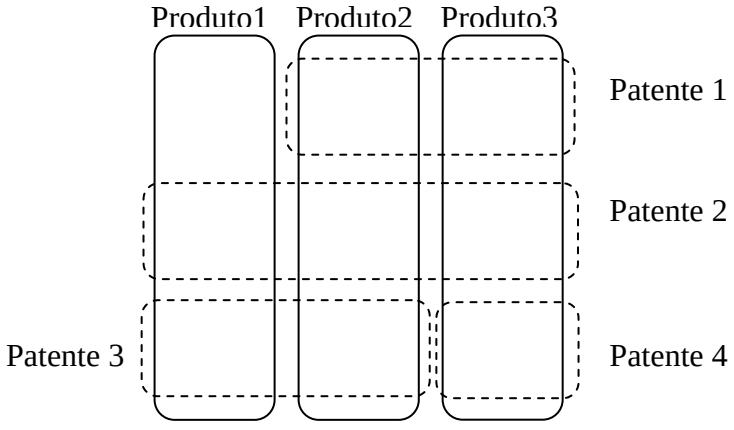


Figura 23 - Relação entre Produtos e Patentes

APÊNDICE C – Outras Formas de Análise dos Grupos

Para a análise dos grupos conseguidos nos testes com Agrupamento foi utilizado o método de leitura individual das patentes e a verificação subjetiva de qual o tema principal de cada patente, que foi utilizado por Fattori, Pedrazzi e Turra (2003). Adicionalmente, foram utilizadas outras duas técnicas para a análise, que foram aplicadas, porém seus resultados não foram considerados no andamento da pesquisa para manter a metodologia concisa e consistente. Entretanto, a experimentação é apresentada aqui para enriquecer novas idéias para a automação do processo de agrupamento.

IPCs mais significativos de cada grupo

A segunda técnica utilizada para tentar validar automaticamente cada grupo foi a localização dos IPCs mais significativos contidos em cada grupo. Para tal, foi criado um módulo no *software*, no sistema, que automaticamente retorna em qual grupo cada patente estava e então apresentava os IPCs contidos em cada grupo. O resultado obtido foi analisado e o IPC que fosse presente pelo menos duas vezes a mais que qualquer outro grupo era considerado mais significativo daquele grupo. Os resultados obtidos estão resumidos na Tabela 10.

Tabela 10 - IPCs mais significativos de cada grupo.
Fonte: Autoria própria.

Grupo 0	Grupo 1	Grupo 2
H04B1/38	H04J1/02	G06F3/00
H04M3/533	H04J3/00	G06F3/048
H04N1/00	H04W36/38	G07B15/02
H04W28/22	G06F15/173	H04L12/00
H04W4/16		H04L12/06
H04W40/02		H04M3/38
H04W8/08		H04N7/14
H04W8/18		
H04W8/22		
H04W88/16		
H04W92/02		

O Grupo 0 apresentou a grande maioria dos IPCs dentro da área H04W (*Wireless Communication Networks*). Segundo a avaliação intelectual individual da amostragem do grupo, o Grupo 0 foi classificado como “Comunicação com Interface Ar” e

“Interopearabilidade”, o que confere com o resultado da análise de significância do IPC dentro do grupo.

O Grupo 1 apresentou dois IPCs dentro da área H04J (*Multiplex Communication*), um IPC na área H04W (*Wireless Communication Networks*), e um em G06F (*Electric Digital Data Processing*). Na análise intelectual individual da amostragem de patentes do grupo ele foi classificado como “Interoperabilidade”, o que condiz com os IPCs H04J e G06F.

O Grupo 2 apresentou dois IPCs na área G06F (*Electric Digital Data Processing*), um na área G07B (*Tichet-Issuing Apparatus*), dois em H04L (*Transmission of Digital Information*), um em H04M (*Telephonic Communication*) e um em H04N (Pictorial Communication). Segundo a análise intelectual individual da amostra do grupo, ele foi classificado como “Aplicações para Clientes”. Os resultados não foram condizentes.

O Grupo 3 não foi representado na Tabela 11, pois apresentou 97 IPCs mais significativos, o que condiz com sua situação de grupo *default.*

Palavras mais significativas

A primeira técnica criada para tentar descobrir qual o tema do grupo foi a utilização de 20 palavras mais significativas do grupo. Isto é, quais os termos que fizeram o grupo se diferenciar dos outros. Como foi utilizado o TFIDF normalizado há como medir os termos que mais “pesaram” para diferir um grupo do outro. Assim, um script em Perl foi criado com dois parâmetros, o primeiro foi o Delta, que é a diferença de tamanho do TFIDF do grupo analisado com os outros grupos; a outra variável é a Lina de Corte, que apresentava qual o valor mínimo do TFIDF que seria considerado de cada grupo. Os resultados alcançados com essa técnica estão sumarizados abaixo na Tabela 11.

Na filtragem das palavras mais significativas o Delta utilizado foi 10. Isto é, o valor do TFIDF de cada termo deve ser no mínimo dez vezes maior que dos outros para ser filtrado.

Tabela 11 - Palavras mais significativas de cada grupo.
Fonte: Autoria própria.

Grupo 0	Grupo 1	Grupo 2	Grupo 3
achiev	box	amount	algorithm
bss	camp	bob	anchor
categori	child	click	atm
constraint	hairpin	desktop	confer
datagram	isim	disk	cpe
engag	jump	drive	dslam
hold	matter	holder	handset
hss	metric	html	headset
invoc	neighbor	imvoc	ipv
jack	qsig	interpret	item
magcf	quarantin	onlin	mac
ned	residenti	partner	multi
ongo	rsvp	referenc	pdsn
pw	spa	seller	physic
restart	stp	sender	pot
rnc	subject	socket	standbi
surviv	superfram	spr	test
umt	top	stun	uplink
utran	topolog	track	vlan
www	tree	va	wtru

A análise das palavras apresentadas pela Tabela 11 não revelam para quais temas cada grupo está convergindo. Nota-se que cada palavra está reduzida ao seu radical, o que é resultado do processo de *stemming*.

Conclusão

Das duas técnicas apresentadas, para facilitar a análise dos grupos e possibilitar a automação da validação dos grupos, somente a técnica de verificação de IPCs mais significativos se mostrou mais apta. Apesar de alguns IPCs não serem condizentes com a o tema atribuído ao grupo pela avaliação individual de patentes, ela pode mostrar que o Grupo 3 não convergiu, acumulando vários IPCs. Isso já seria suficiente para invalidar o Grupo 3 e faria o resultado da pesquisa convergir para o mesmo. A técnica de palavras mais significativas não apresentou sucesso, pois as palavras não apontaram para alguma interpretação.

ANEXO A - Lista das *Stop-Words*

Lista das *Stop-Words* utilizadas para redução de dimensionalidade do problema de mineração de dados:

"i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself",”yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itself", "they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that", "these", "those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having", "do", "does", "did", "doing", "would", "should", "could", "ought", "i'm", "you're", "he's", "she's", "it's", "we're", "they're", "i've", "you've", "we've", "they've", "i'd", "you'd", "he'd", "she'd", "we'd", "they'd", "i'll", "you'll", "he'll", "she'll", "we'll", "they'll", "isn't", "aren't", "wasn't", "weren't", "hasn't", "haven't", "hadn't", "doesn't", "don't", "didn't", "won't", "wouldn't", "shan't", "shouldn't", "can't", "cannot", "couldn't", "mustn't", "let's", "that's", "who's", "what's", "here's", "there's", "when's", "where's", "why's", "how's", "a", "an", "the", "and", "but", "if", "or", "because", "as", "until", "while", "of", "at", "by", "for", "with", "about", "against", "between", "into", "through", "during", "before", "after", "above", "below", "to", "from", "up", "down", "in",”out", "on", "off", "over", "under", "again", "further", "then", "once", "here", "there", "when", "where", "why", "how", "all", "any", "both", "each", "few", "more", "most", "other", "some", "such", "no", "nor", "not", "only", "own", "same", "so", "than", "too", "very".

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo