

SISTEMA COMPUTACIONAL PARA O PROCESSAMENTO
TEXTUAL DE PATENTES INDUSTRIAIS

Graziella Martins Caputo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS
PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DA UNIVERSIDADE
FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS
NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM
ENGENHARIA CIVIL.

Aprovada por:

Prof. Nelson Francisco Favilla Ebecken, D.Sc

Prof. Alexandre Gonçalves Evsukoff, D.Sc

Prof. Myrian Christina de Aragão Costa, D.Sc

Prof. Adelaide Maria de Souza Antunes, D.Sc

RIO DE JANEIRO, RJ - BRASIL

ABRIL DE 2006

CAPUTO, GRAZIELLA MARTINS

Sistema Computacional para o Processamento Textual de Patentes Industriais [Rio de Janeiro] 2006

X, 132 p. 29,7 cm (COPPE/UFRJ, M. Sc., Engenharia Civil, 2006)

Dissertação - Universidade Federal do Rio de Janeiro, COPPE

1. Mineração de Textos
2. Patentes Industriais
3. Inteligência Competitiva

I. COPPE/UFRJ II. Título (série)

Agradecimentos

Em primeiro lugar, agradeço enormemente aos meus pais, pelo apoio e incentivo que me proporcionaram durante todos os meus percursos percorridos e decisões tomadas, e aos meus irmãos pelo apoio e carinho.

Agradeço ao meu namorado, Alexandre, pela paciência e apoio incondicional, e por ter transformado todo o difícil e longo caminho em algo mais prazeroso de ser vivido. À Auristela e Nina por terem me acolhido como se fosse da família.

Agradeço ao meu orientador, Prof. Nelson Francisco Favilla Ebecken, pela orientação e incentivo mesmo nos momentos que pareciam mais difíceis, fazendo com que todas as dúvidas se esclarecessem.

Agradeço aos amigos Valéria, Renan, Daniel, Guilherme, Ângelo e Estela, por transformarem o trabalho em algo extremamente prazeroso. E ao Carlos Sicsú, pela ajuda indispensável na aquisição das patentes.

E à CAPES pelo suporte financeiro que viabilizou a realização desta dissertação.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

SISTEMA COMPUTACIONAL PARA O PROCESSAMENTO
TEXTUAL DE PATENTES INDUSTRIAIS

Graziella Martins Caputo

Abril /2006

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

O presente trabalho apresenta um estudo relacionado à aplicação de métodos de mineração de textos em patentes industriais brasileiras (banco de dados do INPI) como ferramenta de vantagem competitiva para empresas de tecnologia. O principal objetivo é descobrir, utilizando análise dos resumos das patentes, as principais empresas desenvolvedoras em todas as áreas tecnológicas, auxiliando na identificação de competidores. A mineração de patentes é capaz de descobrir, através de direcionamento das pesquisas, novas tendências tecnológicas que auxiliem a tomada de decisões e a criação de estratégias que antecipem a demanda e forneçam vantagem competitiva através de inovações tecnológicas.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

COMPUTATIONAL SYSTEM FOR THE TEXTUAL PROCESSING
OF INDUSTRIAL PATENTES

Graziella Martins Caputo

April /2006

Advisor: Nelson Francisco Favilla Ebecken

Department: Engenharia Civil

This work presents a study related to the application of text mining methods in Brazilian industrial patents (INPI databases) as a tool of competitive profit for technology companies supplying. The main objective is to discover, using patents abstracts analysis, the main developers companies in all technological areas, helping in the competitors identification. The patents mining are capable to discover, through the research aiming perception, new industrial trends helping the decision taker and create strategies that anticipate the demand and to get a competitive profit through technological innovations.

Índice

Índice	vi
Índice de Figuras	viii
Índice de Tabelas	x
1 Introdução	1
1.1 Motivação	2
2 Prospecção Tecnológica	5
2.1 Inteligência Competitiva.....	7
2.2 Patente	9
2.2.1 Detalhes da Patente.....	10
2.2.2 O Conhecimento e as Patentes.....	13
3 Análise de Patentes	16
3.1 Coleta de dados.....	17
3.2 Análise estatística	23
3.3 Análise de citação	24
3.4 Mineração de textos.....	25
3.4.1 Análise dos Dados	26
3.4.2 Algoritmo de Clustering	28
3.4.3 Aplicação de mineração de texto em patentes industriais	29
3.5 Ferramentas de mineração de textos para patentes.....	30
3.2.1 Wisdomain.....	31
3.2.2 Temis	32
3.2.3 VantagePoint	33
3.2.4 Aureka	35
3.2.5 Anacubis	37
3.2.6 BizInt Smart Charts	38
3.2.7 ClearForest.....	39
3.2.8 Statistica.....	40
4 Sistema de Mineração de Textos nos Documentos de Patentes	42
4.1 Linguagem Java	43
4.2 Captura dos dados.....	43
4.3 Descrição do Sistema.....	46
4.3.1 Análises Estatísticas	47
4.3.2 Pré-Processamento.....	50
4.3.3 Clustering.....	51
5 Estudo de Caso	54
5.1 Base de Dados	54
5.1.1 E21B	55
5.1.2 Petróleo	55
5.2 Análise dos Dados	56
5.2.1 E21B – Classificação.....	56
5.2.2 E21B – Depositantes	60
5.2.3 Petróleo – Classificação.....	61
5.2.4 Petróleo – Depositantes	65
5.3 Pré-processamento.....	67

5.3.1	E21B	68
5.3.2	Petróleo	69
5.4	Clusterização	70
5.4.1	E21B – Clustering	71
5.4.1.1	E21B – Temis	84
5.4.1.2	E21B – Statistica	87
5.4.2	Base Petróleo	91
5.4.2.1	Clustering de Patentes de 1996 a 2000	91
5.4.2.2	Petróleo (1996 a 2000) – Temis	101
5.4.2.3	Petróleo (1996 a 2000) – Statistica.....	103
5.4.2.4	Clustering de Patentes de 2000 a 2005	105
5.4.2.5	Petróleo (2001 a 2005) – Temis	115
5.4.2.6	Petróleo (2001 a 2005) – Statistica.....	118
5.5	Considerações Finais	120
6	Conclusão	122
6.1	Trabalhos Futuros	124
	Referências Bibliográficas.....	126

Índice de Figuras

Figura 2-1 – Ciclo da Inteligência	7
Figura 3-1 – Categorias da Mineração na Web	19
Figura 3-2 – Página de Consulta à Base de Patentes do INPI	21
Figura 3-3 – Página de Detalhes da Patente	22
Figura 3-4 – Módulo de Citação.....	32
Figura 3-5 – Temis, módulo IDC	33
Figura 3-6 – VantagePoint.....	34
Figura 3-7 – ThemeScape.....	36
Figura 3-8 – Citation Analisys.....	36
Figura 3-9 – Anacubis	38
Figura 3-10 – BizInt Smart Chart.....	39
Figura 3-11 – Statistica – Módulo Textual.....	41
Figura 4-1 – Patente capturada pelo Sistema.....	45
Figura 4-2 – Tela Principal.....	46
Figura 4-3 – Tela das Classificações existente no conjunto de Patentes processadas.....	48
Figura 4-4 – Tela dos Depositantes existente no conjunto de Patentes processadas.....	49
Figura 4-5 – Tela de Visualização dos Resultados da Clusterização	52
Figura 5-1 – Subgrupos relacionados à base E21B	57
Figura 5-2 – Subclasses relacionadas à base E21B	59
Figura 5-3 – Depositantes com a classificação E21B.....	60
Figura 5-4 – Quantidade de patentes por depositantes de classificação E21B.....	61
Figura 5-5 – Subclasses presentes na base “Petróleo” entre os anos de 1996 e 2000.....	62
Figura 5-6 – Subclasses presentes na base “Petróleo” entre os anos de 2001 e 2005.....	63
Figura 5-7 – Depositantes de patentes com o termo “petróleo” nos anos entre 1996 e 2000	65
Figura 5-8 – Quantidade de patentes por depositantes nos anos entre 1996 e 2000	66
Figura 5-9 – Depositantes de patentes com o termo “petróleo” nos anos entre 2001 e 2005	66
Figura 5-10 – Quantidade de patentes por depositantes nos anos entre 2001 – 2005	67
Figura 5-11 – Resultado do clustering de E21B.....	71
Figura 5-12 – Gráfico de comparação da ferramenta implementada e Temis para E21B	86
Figura 5-13 – Gráfico de comparação da ferramenta implementada e Statistica para E21B	89
Figura 5-14 – Clusterização da base de dados com o termo “Petróleo” de 1996 a 2000.....	92
Figura 5-15 – Gráfico de comparação da ferramenta implementada e Temis para “petróleo” (1996 a 2000)	102
Figura 5-16 – Gráfico de comparação da ferramenta implementada e Statistica para “petróleo” (1996 a 2000)	105
Figura 5-17 – Clusterização da base de dados com o termo “Petróleo” de 2001 a 2005	106
Figura 5-18 – Gráfico de comparação da ferramenta implementada e Temis para “petróleo” (2001 a 2005)	117

Figura 5-19 – Gráfico de comparação da ferramenta implementada e Statistica para
“petróleo” (2001 a 2005)119

Índice de Tabelas

Tabela 2-1 – Setores da patente	12
Tabela 5-1 - Quantidades de patentes de Petróleo coletadas.....	56
Tabela 5-2 – Documentos e palavras-chave dos clusters de E21B	72
Tabela 5-3 – Resultado da ferramenta Temis para a base E21B	85
Tabela 5-4 - Resultado da ferramenta Statistica para a base E21B.....	88
Tabela 5-5 – Resultado da ferramenta desenvolvida para a base “Petróleo” (1996 a 2000).....	93
Tabela 5-6 – Resultado da ferramenta Temis para a base “Petróleo” (1996 a 2000)....	101
Tabela 5-7 – Resultado da ferramenta Statistica para a base “Petróleo” (1996 a 2000)	104
Tabela 5-8 – Resultado da ferramenta desenvolvida para a base “Petróleo” (2001 a 2005).....	107
Tabela 5-9 – Resultado da ferramenta Temis para a base “Petróleo” (2001 a 2005)....	116
Tabela 5-10 – Resultado da ferramenta Statistica para a base “Petróleo” (2001 a 2005)	118

1 Introdução

Diante da crescente globalização, os avanços tecnológicos e científicos ocorrem rapidamente em diversos setores da indústria, comércio e serviços.

Esse fenômeno, quebrou barreiras comerciais facilitando a obtenção de novos recursos e acelerando o acesso às informações. A internet veio como grande colaboradora do processo, facilitando as vias de comunicação, armazenando e expondo a maior parte das informações utilizadas nos dias de hoje.

Para acompanhar essas mudanças, as empresas necessitam estar em constante renovação para se tornarem competitivas e sobreviverem na atual corrida por busca de consumidores e inovações tecnológicas.

Por isso, as empresas inovadoras e detentoras de conhecimento estão mais hábeis a acompanhar o crescimento tecnológico, entender melhor as necessidades de seus consumidores e realizar estudos para antecipar possíveis mudanças no mercado.

Possuir conhecimento requer, no entanto, que as instituições se atualizem continuamente das tendências do mercado. Isso envolve conhecer as características do mercado no qual a organização está inserida, analisar os fatores que influenciam em seus desenvolvimentos e acompanhar o comportamento dos concorrentes.

Por esse motivo, as organizações – que incluem pequenas e grandes empresas, agências governamentais, associações, centros acadêmicos e outros – buscam entender o ambiente em que operam para organizarem a melhor estratégia e tomarem as melhores decisões, através de diretivas de marketing, econômicas e outras.

Para alcançar tal nível, utilizam técnicas da Inteligência Competitiva, técnicas computacionais e profissionais qualificados, cujas conclusões possam auxiliar na tomada de decisão e obtenção de vantagem competitiva.

Os diversos setores, que estão direta ou indiretamente ligados às organizações, geram, diariamente, uma enorme massa de dados, cujo extenso volume torna difícil sua interpretação e manipulação.

Esse problema vem sendo tratado por técnicas computacionais inteligentes, que auxiliam na busca, seleção e extração de informação. A mineração de textos permite a compreensão das informações existentes nos documentos textuais, e que através de análise e aplicação dos resultados, as organizações se tornam capacitadas para inovarem conforme a demanda.

1.1 Motivação

Todos os dias, novos produtos são inventados, novas idéias surgem e para que seja preservado o direito de propriedade industrial sobre o produto ou idéia, novas patentes são depositadas.

Essas patentes possuem os detalhes dos produtos, e garantem ao depositante o direito sobre qualquer produto que possua as mesmas características especificadas na patente.

Através de uma análise detalhada dessas patentes, é possível visualizar as tendências tecnológicas e entender o ambiente intelectual da organização concorrente e a partir dos resultados, obter um ganho competitivo através de inovações.

Os documentos de patentes são um amplo recurso de conhecimento técnico e comercial em termo de progresso técnico, tendências do mercado e propriedade intelectual, sendo a análise desses documentos alvo de estudos de análise estratégica, prospecção tecnológica, planejamento, gestão, formulação e avaliação de programas e um importante veículo de P&D para as instituições.

O objetivo principal desta dissertação é estudar a importância da utilização de técnicas computacionais de descoberta de conhecimento em documentos de propriedade industrial brasileira, buscando capturar informações relevantes no que diz respeito a tendências e inovações tecnológicas.

O fato das patentes se apresentarem em formato texto, comumente chamados de dados não-estruturados, técnicas tradicionais de mineração de dados não são suficientes para extraírem todo o conhecimento contido nas mesmas. Por tal fato, essa dissertação utiliza a técnica de mineração de texto, visando obter o máximo de informações que podem ser úteis no auxílio do entendimento do mercado competidor.

A mineração de textos faz parte do processo de descoberta de conhecimentos em textos, ou KDT (*Knowledge Discovery from Text*), que busca extrair padrões ou conhecimentos, interessantes e não triviais, a partir de documentos textuais (KOSTOFF, 2004).

A implementação dessa técnica para base de dados de patentes depende de padrões especiais, diferentemente de outros documentos, pois busca não perder as informações referentes às particularidades presentes nos documentos de patentes.

Por isso, foi desenvolvida uma ferramenta de mineração de textos exclusiva para base de patentes, capaz de distinguir e fazer uso de campos específicos da patentes, como CLASSIFICAÇÃO, DEPOSITANTES e outros.

Os resultados obtidos neste estudo com o uso desta ferramenta, serão analisados e comparados com resultados de outras ferramentas de mineração de texto.

Para melhor entendimento da importância da aplicação e melhor aproveitamento dos resultados gerados, o segundo capítulo introduz o conceito de inteligência competitiva e o detalhamento das características dos documentos de patentes.

O capítulo 3 descreve as principais técnicas utilizadas nos dias de hoje, para a manipulação e análise dos campos presentes nas patentes, tanto textuais quanto os categóricos, dando ênfase na metodologia de mineração de textos. Descreve ainda algumas das principais ferramentas existentes de busca por conhecimento em documentos de patentes.

O quarto capítulo apresenta as funcionalidades da ferramenta desenvolvida para manipular as bases de patentes industriais, o processo de mineração de textos ocorrido nos documentos e a busca por outras informações relevantes.

O capítulo 5 apresenta os resultados obtidos pela ferramenta quando aplicada aos estudos de caso. Destaca os pontos mais importantes e os compara com os resultados obtidos a partir de outras ferramentas comerciais.

O capítulo 6 apresenta as conclusões do trabalho, tal como sugestões de outras funcionalidades a serem inseridas no processo de mineração de textos e outras análises de patentes.

2 Prospecção Tecnológica

O mercado passou por uma transição de uma economia tipicamente industrial para uma economia voltada para o setor de serviços, onde a hegemonia econômica e social é exercida por aqueles que administram o conhecimento e a informação (MORAIS, 1999).

Para estar ativa diante do alto nível de concorrência existente para a maioria das empresas, é necessário, acima de tudo que a empresa esteja sempre atualizada e trabalhando de forma inteligente.

Para isso, a mesma precisa possuir as informações necessárias para obter o conhecimento de identificar sentidos, interpretar o seu ambiente, mercados, fornecedores e clientes.

Além disso, as empresas precisam estar sempre atentas para os chamados microambientes e macroambientes. O microambiente é composto por forças próximas a empresa que afetam a sua habilidade para servir aos seus clientes - os canais de marketing, os mercados consumidores, os concorrentes e o público. O macroambiente é composto por forças sociais maiores que afetam todo o microambiente, forças demográficas, econômicas, físicas, tecnológicas, políticas e culturais.

Esses ambientes geram grandes massas de dados e as empresas precisam buscar metodologias de gerenciamento das informações.

Nesse sentido, informação pode ser definida com um conjunto de dados com um determinado significado, sendo o dado um registro de algum determinado evento. Inteligência é a informação devidamente filtrada e analisada.

No contexto técnico-econômico atual, a inteligência tem assumido importância crescente, fazendo com que a empresa necessite desse elemento para o seu processo de inovação tecnológica e para o aumento de sua competitividade.

Métodos de manipulação dessas informações vem sendo estudados e implantados nas organizações em busca de melhor domínio sobre os dados que as mesmas possuem (BRENNER, 2005).

A Inteligência de Negócios (*Business Intelligence* ou BI) é um processo organizacional pelo qual a informação é sistematicamente coletada, analisada e disseminada como inteligência aos usuários que possam tomar ações a partir dela. BI (KUDYBA *et al.*, 2003) é a área que recolhe informações de seus clientes e fornecedores e as analisa, para ajudá-los a identificar oportunidades e criar estratégias que antecipem a demanda. Essa área atende a vários setores da empresa, como a de recursos humanos, marketing, gerenciamento de tecnologia e de especialistas, entre outras, além de utilizar ferramentas para auxiliar nos negócios, como o CRM, *Data Warehouse*, Mineração de Dados e ferramentas OLAP (SULLIVAN, 2001).

Um outro processo que tem sido amplamente explorado por organizações inovadoras é o conceito de Inteligência Competitiva que busca obter informações sobre a concorrência sem recorrer a meios inescrupulosos ou ilegais.

A Inteligência Competitiva (*Competitive Intelligence* ou IC) (TARAPANOFF, 2001) foi definida pela Sociedade de Profissionais de Inteligência Competitiva (*Society of Competitive Intelligence Professionals* ou SCIP em SOCIETY, 2006) como “um programa sistemático e ético para recolhimento, análise e gerenciamento de informações externas que podem afetar os planos, decisões e operações de uma companhia”.

Alguns autores consideram que o termo Inteligência de Negócios é sinônimo de Inteligência Competitiva. Para finalidade dessa dissertação, no entanto, ambos serão tratados de forma diferente, como nas duas principais entidades representativas no assunto: a Sociedade dos Profissionais de Inteligência Competitiva (SCIP), nos EUA, e a Associação Brasileira dos Analistas de Inteligência Competitiva (ABRAIC).

Na próxima seção, a IC e sua importância é brevemente discutida, são descritos os principais detalhes existentes nos documentos de patente, tal como o detalhamento do ganho competitivo obtido quando esses documentos são utilizados pela equipe de IC.

2.1 Inteligência Competitiva

O processo de Inteligência Competitiva tem ganhado importância cada vez maior dentro das empresas, tornando-se ferramenta de apoio indispensável em diversos níveis organizacionais, como planejamento estratégico, marketing, programas de gestão do conhecimento, entre outros (TYSON, 1998; KAHANER, 1996).

Cabe à IC realizar estudos para antecipar possíveis mudanças no mercado, descobrir novos e potenciais concorrentes e se manter atualizado sobre novas tecnologias, produtos e processos, bem como mudanças políticas, legislativas e regulatórias que possam afetar os negócios da empresa

Nesta questão, a organização precisa estar sempre atenta às organizações competidoras, por isso, um dos papéis da inteligência competitiva é unir esses conceitos e entender ligações entre pessoas e companhias.

De um modo geral, um Sistema de Inteligência é descrito através de 4 etapas como mostra a Figura 2.1 .

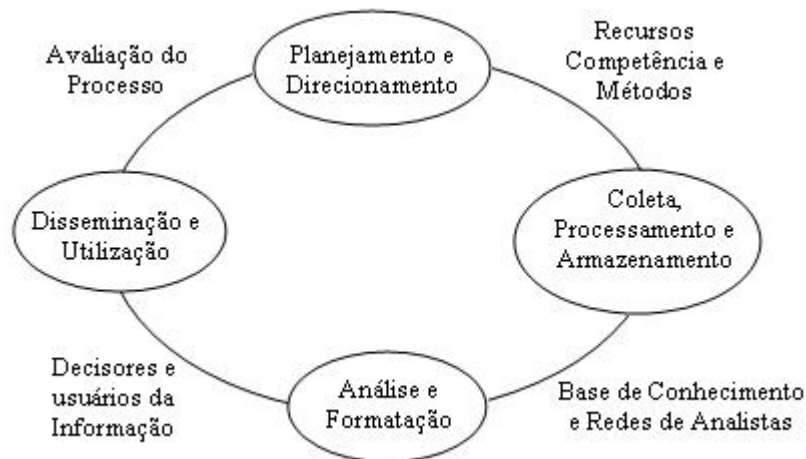


Figura 2-1 – Ciclo da Inteligência

Na primeira fase, ou Planejamento e Direcionamento, a administração deve estar envolvida no processo de definição do tipo de inteligência necessária. É a etapa inicial, mas pode também ser a final, quando o resultado é apresentado para a tomada de

decisões e as ações subseqüentes podem gerar a necessidade de novos processos de inteligência.

A fase de Captura, Processamento e Armazenamento das informações envolve diversas etapas como a determinação das informações necessárias, a identificação das fontes e a coleta de informações.

Para o sucesso desse tipo de sistema, não se deve fazer uso apenas de informações publicadas. A participação de eventos, conversas com consumidores, fornecedores, parceiros, empregados, especialistas na área e até mesmo concorrentes, podem fazer parte do processo de coleta de dados

O conteúdo e as fontes das informações dependerá das técnicas e métodos utilizados para análise, que podem ser diversas: análise SWOT (AZEVEDO *et all.*, 2001), benchmarking, forças de Porter, fatores críticos de sucesso, balanced scored, mineração de dados, perfil dos concorrentes, análise financeira, cenários, jogos de guerra, entre outros. Por isso, a análise é considerada a etapa mais difícil, quando é necessário interpretar, procurar modelos e produzir diferentes cenários.

A disseminação é a etapa da distribuição do produto da inteligência, quando são sugeridas e defendidas possíveis ações a tomar.

Todo esse processo de planejamento e operacionalização do sistema leva em conta aspectos como abrangência do ambiente competitivo e planejamento estratégico da empresa.

Apesar da importância dos Sistemas de IC, a maioria dos autores corroboram que a geração da inteligência é realizada de fato pelos seres humanos, com a ajuda dos sistemas bem estruturados capazes de apresentar métodos e técnicas de análise, bem como proporcionar a coleta e a disseminação da inteligência (KNOWLEDGE, 2006).

É importante lembrar, no entanto que a prática da Inteligência Competitiva guarda as suas próprias peculiaridades éticas. O caráter ético e legal da fase de obtenção da informação, principalmente envolvendo os concorrentes, como pressuposto da IC, foi citada por COTTRILL (1998).

O aumento do uso e coleta de informações sobre o mercado e a concorrência, e o desenvolvimento de técnicas pode gerar métodos cuja aplicação é eticamente questionável. Por esse motivo, é preciso que a organização esteja sempre atenta para

identificar os limites éticos de conduta no desempenho das práticas de Inteligência Competitiva (BOATRIGHT, 2000).

As informações utilizadas na aplicação de métodos de IC numa organização possuem origem em diversas fontes e variados formatos. Nesse contexto, pode-se verificar que vários dados que estão simplesmente armazenados, podem ser utilizados para obter um enorme ganho competitivo, como por exemplo, a utilização de registros de call centers (CAPUTO *et all*, 2006) para a compreensão e melhoria do CRM interno, o monitoramento de páginas web (MARINHO *et all*, 2003) concorrentes visando estar sempre atento às mudanças dos mesmos e de seus planejamento estratégicos (LIU, 2001).

Da mesma forma, qualquer publicação existente relacionada à empresa e às suas concorrentes, pode ser utilizada como recurso para a análise e obtenção de inteligência.

Um outro recurso que tem sido tratado com grande atenção pelos pesquisadores é a utilização de documentos de patentes como recurso da IC (PEREIRA, 2003). Através da análise e aplicação de técnicas computacionais inteligentes, é possível fazer progressões tecnológicas e descobrir o direcionamento estratégico concorrente (LARKEY, 1999).

A seguir, é realizado um estudo aprofundado sobre a utilização das patentes, tal como o detalhamento das características e atributos importantes para a obtenção de conhecimento.

2.2 Patente

Os documentos de patentes, por possuírem grande quantidade de informações sobre detalhes técnicos, são considerados um amplo recurso de conhecimento científico e comercial, como pode ser visto em AHMAD *et all* (2003).

A análise das patentes tem recebido importância nos dias de hoje no processo de inovação, pois o ciclo de inovação tem se tornado menor, e a demanda por *marketing* mais volátil.

Em um nível macro, a análise de patentes tem sido aplicada em diversas áreas, como na geração de indicadores econômicos, medindo a relação entre o desenvolvimento da tecnologia e o crescimento da economia (GRANDSTRAND, 1999; GRILLICHES, 1990; HOLL *et all*, 2000), estimando conhecimento tecnológico e o seu impacto na produtividade (EVENSON *et all*, 1988), ou comparando desempenho inovativo no contexto internacional (PACI *et all*, 1997).

No nível micro, a análise de patentes tem sido usada para evoluir a competitividade (NARIN *et all*, 1987), desenvolver planos tecnológicos (MOGEE, 1991), priorizar investimentos de P&D (HIRSCHEY *et all*, 2001), ou monitorar mudanças tecnológicas internas (ARCHIBUGI *et all*, 1996).

A utilização e aplicabilidade dessa análise é contemplada por uma variedade de áreas de tecnologia, como na gerência de P&D e no meio acadêmico.

Os documentos de patentes são estruturados de acordo com o formato denominado bibliométrico (KARKI, 1999; NARIN, 1994), ou seja, são separados através de vários atributos que identificam e caracterizam as patentes. Esses dados incluem número da patente, tipo de documento, título, resumo, inventor, depositante, requerente, classificação internacional ou PCI, entre outros.

A escolha adequada dos atributos da patente na utilização da Inteligência Competitiva requer a compreensão detalhada de cada um dos mesmos, visando alcançar maior desempenho a partir da técnica de análise utilizada.

2.2.1 Detalhes da Patente

A propriedade intelectual é uma parcela do capital intelectual protegido por legislação específica e engloba patentes, marcas, desenho industrial etc.

Os direitos sobre a propriedade intelectual são tipicamente divididos entre o direito autoral e a propriedade industrial.

O direito autoral, se refere a autores de obras literárias e artísticas, incluindo softwares, artistas intérpretes ou executores, gravadoras de música e órgãos de radiodifusão. Já a propriedade industrial, inclui as invenções, os desenhos e modelos.

As invenções são, tradicionalmente, protegidas através de patentes. Uma patente é um direito, concedido pelo governo, de exploração comercial de uma invenção, durante um determinado limite do tempo. O dono de uma patente, segundo o seu interesse, pode licenciar sua invenção para outros usarem, sob condições de pagamento de "royalties".

As bases de patentes armazenam cerca de 70% da informação tecnológica disponíveis em todo o mundo e por isso, são importantes fontes de informação tecnológica.

Tradicionalmente, uma patente somente tem validade no país onde foi depositada. Para ter validade internacional, a patente deve ser depositada na base de dados de patente internacional, segundo o Tratado de Cooperação em matéria de Patentes - PCT (PCT, 2006), de 1970, onde a patente passa a ter validade nos mais de 100 países signatários deste tratado.

No Brasil, o Instituto Nacional de Propriedade Industrial (INPI, 2006), vinculada ao Ministério do Desenvolvimento, Indústria e Comércio Exterior, é responsável por regulamentar e proteger o direito às propriedades intelectuais de marcas e patentes.

O banco de patentes reúne um volume aproximado de 24 milhões de documentos de patentes armazenados em papel, microformas e em CD-ROM e distribuídos em cerca de 4000 m². Mensalmente são acrescentados a seu acervo cerca de 40 mil novos documentos nacionais e estrangeiros.

Esta documentação é originária dos principais países industrializados e organizações internacionais: Estados Unidos, Grã-Bretanha, França, Holanda, Espanha, Alemanha, Canadá, Austrália, Suíça, Japão (resumos em Inglês), Escritório Europeu de Patentes, Organização Mundial da Propriedade Intelectual (documentação do PCT) e Organização Africana da Propriedade Intelectual, além do Brasil.

As patentes podem ser classificadas em dois tipos diferentes: a) patentes de invenção (PI), que são os avanços do conhecimento técnico que combinam atividade inventiva e aplicação industrial. Esse tipo de patente possui validade de 20 anos a partir da data de depósito; e b) modelo de utilidade (MU), que é uma nova forma ou disposição de objeto de uso prático, com aplicação industrial, e representa melhoria funcional de um produto ou um processo já existente. É caracterizado como um ato inventivo. Possui validade de 15 anos a partir da data de depósito.

Esses documentos possuem um formato padrão internacional, e classificadas de acordo com a Classificação Internacional de Patentes (CIP), que permite a rápida recuperação por área tecnológica, através de indexação.

As patentes podem ser classificadas em 8 (oito) setores principais, com 64000 (sessenta e quatro mil) subdivisões. Cada subdivisão tem um símbolo composto de algarismos arábicos e de letras do alfabeto latino. Os 8 (oito) setores principais são denominados de seções, a saber:

Seção A	Necessidades Humanas
Seção B	Operações de Processamento; Transporte
Seção C	Química e Metalurgia
Seção D	Têxteis e Papel
Seção E	Construções Fixas
Seção F	Eng. Mecânica / Iluminação / Aquecimento
Seção G	Física
Seção H	Eletricidade

Tabela 2-1 – Setores da patente

O símbolo completo da classificação internacional das patentes é constituído por símbolos representando a Seção (conforme anteriormente), Classe (número composto por dois algarismos), Subclasse (letra maiúscula), grupo e Subgrupo, como por exemplo:

A Seção

01 Classe

B Subclasse

1/00 Grupo Principal ou **1/24** Subgrupo

A partir dessa classificação, as patentes são facilmente localizadas tal como o acesso a um determinado tipo de tecnologia.

As patentes brasileiras podem ser consultadas através do site oficial do INPI, onde é possível realizar pesquisa básica ou avançada através de atributos existentes no documento. A pesquisa retorna informações simplificadas das patentes que inclui alguns campos, como o número do pedido, data de depósito, prioridade unionista, classificação, título, resumo, nome do depositante, nome do inventor, nome do procurador e PCT, conforme será descrito no capítulo 4.

Este tipo de consulta on-line é de extrema importância, pois reduz o número de pedidos de patentes inválidas, otimizando o serviço de concessão de novas patentes.

Dessa forma, é de extrema importância para o país que exista um órgão de gerenciamento dos depósitos de patentes, garantindo o direito do inventor, pois essa proteção abrange não apenas empresários e organizações particulares. O aumento no número de patentes de um país pode influenciar diretamente a economia do mesmo, pois através da inovação, o mesmo se torna apto para evoluir em um tecnologia e aumentar a quantidade de produtos comercializados e inovações para a exportação, agregando valor a seus produtos.

2.2.2 O Conhecimento e as Patentes

Necessidades, preocupações, modas, costumes e diversos outros assuntos relacionados a um determinado período de tempo, podem ser facilmente identificados através de análise da frequência em que patentes predominaram naquele período. Por exemplo, a descoberta de uma doença faz com que se desenvolvam produtos para combatê-la, e uma tendência da moda cria vários acessórios, assim como houve períodos de desenvolvimento que marcaram uma época brasileira que podem ser claramente observados nas patentes existentes como a necessidade de higiene, produção de energia meios de comunicação e vários outros.

Neste ponto, podemos observar que o incremento do volume de documentos de patentes numa classificação indica uma tendência tecnológica, ou seja, o direcionamento das pesquisas, a nível mundial, e/ou reflexos posteriores no mercado.

Esta valiosa fonte de informação tecnológica tornou-se um insumo estratégico imprescindível para as empresas que desejam se manter atualizadas sobre o

desenvolvimento tecnológico do seu setor, e assim, possibilitando a elevação do nível técnico, uma vez que nas patentes estão contidos tanto conceitos científicos quanto detalhes do processo (GANGULI, 2004, LARKEY, 1998).

Além disso, através da análise de patentes, é possível identificar efetivamente quais são os seus maiores concorrentes, além de compreender em quais áreas tecnológicas e quais produtos estes pretendem investir, obter informações sobre planos estratégicos, ter uma projeção do risco financeiro ao entrar num mercado altamente competitivo, obter uma análise sobre planos e produtos de seus concorrentes, explicitando uma real consciência dos valores de seus produtos e inovações, através da identificação de aspectos estratégicos do mercado (POYNDER, 1998).

Como resultado, podemos obter um levantamento sobre os mais importantes fornecedores de tecnologia, equipamentos e produtos, ou mesmo se as características de interesse de uma organização já estão obsoletas ou não.

As informações contidas em documentos de patentes oferecem grande utilidade para indústria e empresas através de análise em áreas afins e inovações concorrentes, resultando uma melhora da gestão do conhecimento tecnológico.

A análise de patentes, tal como o seu mapeamento, pode ser executada sobre vários aspectos e podem ser utilizados tanto os documentos de patentes, quanto qualquer literatura a respeito.

No geral, sua análise visa extrair conhecimento processando os dados sob diferentes critérios, dependendo do tipo de informação que se quer ter acesso.

Esse processo pode ser executado a partir de várias perspectivas, como por exemplo, análise estatística, análise de citação, mineração dos dados e mineração de textos. Essa variedade de técnicas computacionais para o processamento são descritas no capítulo 3.

Esses métodos visam capturar padrões existentes nesses documentos e que podem oferecer alguma informação nova e relevante.

Vários estudos são dedicados a alcançar esse propósito, como pode ser verificado em GUPTA *et al.* (2000), onde é aplicada análise bibliométrica em patentes de carbono nanotubo para medir o crescimento da atividade da indústria de carbono de nanotubos e os seus links com a ciência.

Um outro exemplo da aplicação de análise de patente é o estudo realizado por APPLEYARD *et al.* (1999), para detectar o impacto de tecnologia de semicondutores americanos no resto do mundo. E o estudo de KARKI (1997), que utiliza a análise de patentes como uma ferramenta policial.

Nesse sentido, a análise de patentes tem sido grandemente utilizada na identificação de novas tendências tecnológicas e previsão de novas necessidades, em vários setores industriais, como farmacêutico, petrolífero e prestação de serviços, pois antecipando os novos produtos e levando em consideração o seu desenvolvimento industrial ao longo dos anos, é possível fazer uma previsão de como a tecnologia gradativamente evolui.

3 **Análise de Patentes**

A aplicação de técnicas computacionais inteligentes na análise de documentos de patentes como ganho competitivo tem sido amplamente explorada e abordada de diversas maneiras.

Dessa forma, um número de técnicas tem sido utilizadas para a manipulação e análise de dados bibliométricos, que caracterizam os documentos de patentes, por possuírem atributos bem definidos como número da patente, data de depósito, título, inventor, e outros.

A análise bibliométrica nas patentes pode fornecer informação sobre a natureza e crescimento de uma atividade inventiva. Uma série de estudos indicam a utilização desse tipo de técnica para a análise do status e das tendências do desenvolvimento da tecnologia (KARKI, 1999; NARIN, 1994).

No entanto, para que a análise dos documentos seja bem sucedida, algumas questões devem ser respondidas. Um exemplo disso é conhecer bem a necessidade que leva a realizar tal análise. Se a necessidade é apenas conhecer qual a empresa que mais depositou patentes nos últimos 5 anos, uma análise estatística sobre os documentos pode parecer suficiente. Um outro ponto que deve ser levando em conta é a escolha apropriada do domínio de documentos que serão utilizados. A coleta de documentos que não pertencem ao escopo do problema pode prejudicar os resultados.

Além dessas questões, nas próximas seções, serão discutidos os principais assuntos que devem ser levados em conta no procedimento de análise de patentes. Essa discussão inclui o método de captura de patentes comumente utilizado e alguns dos métodos computacionais e estatísticos de análise de patentes. Os objetivos dos respectivos métodos também são apresentados.

No final do capítulo, são descritas algumas das principais ferramentas existentes no mercado capazes de processar e analisar inteligentemente os dados de documentos de patentes.

3.1 Coleta de dados

A utilização da internet para ganho competitivo tem sido amplamente explorada e apresenta resultados bastante animadores devido a grande diversidade e quantidade de dados presentes na World Wide Web atualmente.

A web se tornou um importante canal na condução de negócios devido ao fato de disseminar rápida e eficientemente informações e ser capaz de conectar pessoas no mundo inteiro.

Recuperar informações relevantes na internet, no entanto, é uma tarefa bastante trabalhosa, pois os dados são tipicamente não rotulados, distribuídos, heterogêneos, semi-estruturados e multidimensionais.

As ferramentas de busca existentes podem apresentar baixa precisão. Elas se baseiam na inserção de palavras chaves pelo usuário e apresentam como resposta páginas ordenadas de acordo com a similaridade da consulta. A baixa precisão na relevância dos resultados e inabilidade de indexar todas as informações presentes na web dificultam o encontro de informações relevantes. Uma outra dificuldade é extrair conhecimento potencialmente usual a partir dos dados recuperados (CRAVEN *et al.*, 1998). Outros problemas relacionados a ferramentas de busca podem ser encontrados em LAWRENCE *et al.* (1999). Todos esses fatores irão prejudicar a criação de novos conhecimentos a partir da informação presente na web.

A personalização da informação é um outro problema encontrado na recuperação de páginas. Esse problema está diretamente ligado a forma como a consulta é feita. Diferentes páginas web utilizam termos diferentes para expressarem a mesma entidade, como por exemplo, um sistema de vendas pode utilizar o termo cliente ou consumidor. Da mesma forma, o mesmo termo pode ser utilizado em diversas áreas, tendo um significado diferente em cada uma delas.

Para contornar todos esses problemas, surge a disciplina *web mining*, que busca identificar páginas relevantes na *web* através de análise inteligente, de acordo com a necessidade do usuário e é capaz de estruturar e processar os dados visando obter o maior ganho de informação.

As ferramentas de busca de informações visam rastrear conteúdo dentre as páginas da *web*, aquelas que se assemelham ao esperado pelo usuário. As técnicas de busca por essas informações incluem busca baseada em palavra chave (BRIN *et al.*, 1998), consultas *web* (MENDELZON *et al.*, 1997), e preferências do usuário (UNDERWOOD *et al.*, 1998). Essas técnicas são baseadas em conceitos de Inteligência Artificial, Recuperação de Informação e outras áreas de descoberta de conhecimento.

Os componentes mais importantes de *web mining* incluem recuperação de informação, extração de informação, generalização e análise.

Recuperação de informação (FRAKES *et al.*, 1992) se refere a recuperação automática de documentos relevantes, utilizando indexação de documentos e mecanismos de busca. Extração de informação auxilia na identificação de fragmentos de documentos que constituem a semântica da *web* (COWIE, 1996). Generalização está relacionada a área de reconhecimento de padrões e aprendizado de máquina: *clustering* e mineração de regras de associação. Já a análise corresponde à extração, interpretação, validação e visualização do conhecimento obtido a partir da *Web*.

O processo pode ser efetuado basicamente de três maneiras distintas que incluem Mineração de uso na *Web*, Mineração de Estrutura na *Web* e Mineração de Conteúdo na *Web* (SCIME, 2005).

A figura 3.1 ilustra as diferentes formas de captura de conhecimento na *web*:



Figura 3-1– Categorias da Mineração na Web

A Mineração de Uso visa descobrir padrões de acesso de usuários através de logs em páginas da web, armazenados em bases de dados. A partir desse padrão é possível identificar informações de maior relevância e assim entender as necessidades do usuário, prevendo o seu comportamento nas páginas e melhor auxiliando à organização a tarefa de satisfazer o seu cliente.

A Mineração de Estruturas analisa a organização das páginas web e as ligações existentes entre as mesmas, ou seja, a informação implícita contida dentro dos documentos. Essas ligações são formadas basicamente pelos hiperlinks, e dessa forma, a rede de documentos é tratada como um grafo orientado. A técnica parte do princípio que uma página bastante referenciada por outras possui um grau de importância elevado.

Já a Mineração de Conteúdo utiliza o conteúdo existente dentro dos documentos das páginas web como principal fonte de descoberta de informações relevantes. Esse conteúdo é formado não apenas do texto, mas também de qualquer outro dado presente na página, como áudio, vídeo, símbolos, metadados e hipertextos.

Nessa dissertação, apenas os dados textuais das patentes foram considerados para o descoberta de conhecimentos novos. Por tal fato, a categoria de mineração de conteúdo é a que melhor se destaca no processamento das patentes, pois é capaz de considerar e processar o texto presente no conteúdo e assim trata-lo como um simples problema de mineração de textos.

Dessa forma, a mineração de conteúdo será exemplificada nas próximas seções através do processamento dos dados das patentes recuperadas através de páginas da web e tratadas como mineração de textos.

Especificamente para as patentes brasileiras, uma parcela dos dados encontram-se atualmente disponíveis na página oficial do Instituto Nacional de Propriedade Intelectual. Dessa forma, a mineração de conteúdo realiza a busca a partir dos campos existentes nas patentes.

Essas patentes, tal como os documentos de marcas e desenhos industriais, podem ser consultadas através de um serviço oferecido na página principal do site oficial do INPI.

Esse acesso é garantido a qualquer pessoa que queira buscar algum documento específico na base de dados existente.

Para se ter acesso a essa base, porém, um procedimento inicial é requerido. É necessária uma autenticação através da identificação visual de caracteres expostos no browser através de figura e a digitação dos mesmos em um campo indicado na página de consulta.

O site disponibiliza dois tipos de consulta. A primeira delas, padrão da página é a pesquisa básica, que apresenta opção de consulta através do número do processo e de palavras existentes no TÍTULO, no RESUMO, no NOME DO DEPOSITANTE e no NOME DO INVENTOR. A segunda consulta, além das opções oferecidas pela consulta básica, ainda apresenta busca através de DATAS de DEPÓSITO e de PRIORIDADE, através do PAÍS e do NÚMERO da prioridade, CLASSIFICAÇÃO e número do depósito PCT. A busca através de palavras chaves, na pesquisa avançada possui a possibilidade de manipular operadores lógicos como AND, OR e NOT, como indicado na figura 3.2.

» Consultar por: **Base Patentes** | Pesquisa Básica | Finalizar Sessão

Forneça abaixo as chaves de pesquisa desejadas. *Evite o uso de frases ou palavras genéricas.*

PESQUISA AVANÇADA

(21) Nº do Pedido :	<input type="text"/>	Ex: P/0101161-8; MU6900960-0; M/5500233-1; C10201935-3.
(22) Data Depósito :	16/10/2003 a <input type="text"/>	dd/mm/aaaa" Ex: 10/10/2001.
(31) Nº da Prioridade :	<input type="text"/>	Ex: 392.176
(32) Data da Prioridade :	<input type="text"/> a <input type="text"/>	dd/mm/aaaa" Ex: 10/10/2001
(33) País da Prioridade:	« Clique e escolha »	
(51) Classificação :	<input type="text"/>	Ex: G06F 13/00.
(54) Título :	petróleo	Ex: resfriamento and (líquido or água) and not cruzado.
(57) Resumo :		Ex: milho and herbicida and plantas and not glifosato; carro prox(6) porta.
(86) Número do Depósito Pct:	<input type="text"/>	Ex: US9308239.
(71/73) Nome do Depositante :	petrobras brasileiro)	Ex: petrobras or (petroleo and
(72) Nome Inventor :		Ex: "Antônio Cláudio Corêa"
Nº de Processos por Página :	20	<input type="button" value="pesquisar »"/> <input type="button" value="limpar"/>

Figura 3-2 – Página de Consulta à Base de Patentes do INPI

O resultado apresentado pela consulta consiste de uma lista de links para a descrição simplificada das patentes que obedecem aos atributos requisitados na pesquisa. Essa lista de links é formada pelos respectivos números dos processos, datas de depósito e título das patentes retornadas, e ordenadas de acordo com a data apresentada. Os mesmos levam a uma página contendo uma descrição da patente composta por alguns dos principais atributos presentes nas mesmas. Entre esses atributos, se destacam:

- Número do Pedido;
- Data do Depósito;
- Prioridade Unionista, com o País, Número e Data;
- Classificação da Patente;
- Título;
- Resumo;

- Nome do Titular;
- Nome do Depositante;
- Nome do Inventor;
- Nome do Procurador;
- Início da Fase Nacional;
- PCT;
- W.O.;

A figura 3.3 representa um exemplo da página de patentes oferecida pelo INPI, com alguns dos atributos que a mesma contém.

Consulta à Base de Patentes - Detalhes da Patente

[Pesquisa Base Marcas | Pesquisa Base Desenhos | Ajuda?]

» Consultar por: [Base Patentes](#) | [Finalizar Sessão](#)

Depósito de pedido nacional de Patente

(21) Nº do Pedido:	PI0304916-7
(22) Data do Depósito:	16/10/2003
(51) Classificação:	E21B 43/12
(54) Título:	MÉTODO PARA ATENUAR INSTABILIDADES NA PRODUÇÃO DE UM POÇO SUBMARINO DE PETRÓLEO
(57) Resumo:	"MÉTODO PARA ATENUAR INSTABILIDADES NA PRODUÇÃO DE UM POÇO SUBMARINO DE PETRÓLEO". A presente invenção se refere a um método para atenuar as instabilidades, geradas por um efeito hidrodinâmico de impacto causado por golfadas de líquido e gás, no interior de linhas de coleta de produção em plataformas de petróleo. A estabilidade é obtida pela injeção de ar pressurizado no trecho vertical da linha de coleta, por um tubo flexível e coaxial, em uma altura ideal para minimizar o processo de segregação das fases do petróleo.
(71) Nome do Depositante:	Petroleo Brasileiro S.A. - Petrobras (BR/RJ)
(72) Nome do Inventor:	Fausto Arinos de Almeida Barbuto / Guilherme de Almeida Peixoto
(74) Nome do Procurador:	ANTONIO CLAUDIO CORREA MEYER SANT ANNA

PUBLICAÇÕES			
Nº RPI	Data RPI	Despacho	Complemento do Despacho
1797	14/06/2005	3.1	
1729	25/02/2004	2.1	

Dados atualizados até **06/09/2005** - Nº da Revista: **1809**

Figura 3-3 – Página de Detalhes da Patente

Esses atributos foram coletados da página para a execução da mineração de textos, porém várias restrições foram detectadas a partir da página do INPI.

A primeira restrição ocorre na captura das patentes a partir dos links listados nos resultados. O acesso aos resultados permite uma visualização máxima de 19 patentes. A partir desse número, a página expira e o browser retorna para a página principal.

Além disso, o tempo de permanência na página de consulta dos resultados também pode fazer com que a página expire, retornando, da mesma forma, à página principal.

Por esses motivos, uma ferramenta foi implementada para automatizar o processo de coleta das patentes da página do INPI, segundo as informações descritas na seção 4.2.

3.2 Análise estatística

A análise estatística das patentes consiste em analisar numericamente os atributos existentes nos documentos.

Esse processamento é comumente utilizado para a identificação de frequências presentes nas patentes, como por exemplo, descobrir as empresas que mais depositaram patentes nos últimos 5 anos e que são relacionadas a um determinado produto de interesse da organização que realiza a pesquisa, o que pode auxiliar na identificação dos maiores concorrentes da mesma.

Essa análise é realizada através de consultas (*queries*) nos documentos, podendo ser a pesquisa realizada a partir dos atributos presentes nas patentes, como depositante, data, classificação e outros.

O resultado dessa análise pode ser melhor entendido se for visualizado através de gráficos, onde as frequências são melhor identificadas.

Exemplos de ferramenta que utilizam análise estatística a partir de atributos de patentes serão exemplificados nas seções subsequentes.

3.3 Análise de citação

A manipulação e análise de dados bibliométricos tem sido amplamente utilizado para o estudo de interfaces da ciência e da tecnologia.

A análise de citação de patentes (MICHEL, 2001) é uma das ferramentas mais utilizadas, e é baseada no número de citações de uma patente em suas patentes subsequentes. A quantidade de citações por patentes representa a importância relativa da mesma.

Essas citações são baseadas na ideia de que uma patente irá citar aquelas que a antecedem na criação daquela nova patente, podendo ser uma a evolução da outra, e dessa forma, conhecendo quais patentes foram originadas por algum tipo de tecnologia, ou o contrário, descobrir qual tecnologia gerou alguma do presente, identificando assim, o que muitos autores chamam de família de patentes.

Dessa forma, a análise fornece informações sobre a origem e o crescimento de uma dada atividade inventiva, além dos depositantes ativos na indústria, academia e governo, relacionamento de inventores, ligações com assuntos científicos e tendências tecnológicas.

Basicamente, a metodologia cria uma árvore de conexões entre as patentes a partir das citações que existem nas mesmas, da mesma forma que uma análise de citação científica conecta as referências em uma base de dados de artigos científicos (KARKI, 1997).

Essa metodologia pode produzir vários índices como a quantidade de citações por patentes, as patentes mais citadas, links que não possuem patentes, índice de impacto técnico, tempo de ciclo de uma tecnologia e outros. Esses índices têm sido usados para medir qualidade de vantagens técnicas (HIRSCHEY, 2001), detectar negociações poderosas entre empresas (MOWERY, *et al.*, 1998), descobrir valores econômicos de novas criações em valores de equações (HOLL, *et al.*, 2000) e a quantidade de conhecimento (TIJSSSEN, 2001).

Apesar da facilidade na identificação das citações, alguns problemas são encontrados na análise, como por exemplo, a complexidade das relações existentes entre vários documentos, pois a análise apenas indica a ligação entre duas patentes. Dessa

maneira, o escopo da análise fica restrito à patente que cita e aquela que é citada, limitando assim, o potencial da informação. Um outro problema é que a análise de citação não é capaz de considerar relação interna entre patentes, levando em consideração apenas a existência e frequência das citações. E finalmente, consome um enorme tempo de processamento pois necessita de uma busca exaustiva.

3.4 Mineração de textos

A mineração de dados envolve a extração dos dados nos campos que possuem algum tipo de categoria, por exemplo, quando se deseja encontrar a relação entre procurador e códigos da Classificação Internacional de Patentes para uma área específica de tecnologia (FATTORI, 2003). Através dessa perspectiva, pode-se ter uma idéia dos maiores envolvidos em uma área de tecnologia e em que tipo de trabalho está focado (YOON *et all*, 2003, ZANASI, 2001).

No caso da mineração de textos, o processo envolve a clusterização dos documentos baseados no conceito em que se encontram e os dados utilizados são não estruturados (REZENDE, 2000).

Levando em consideração que grande parte das informações nas patentes encontram-se na forma textual, a aplicação de mineração de textos tem grande utilidade para melhor compreensão do conhecimento existente nesses documentos e consequentemente fornecendo recursos para a aplicação de Inteligência Competitiva.

Um exemplo disso é a análise dos atributos textuais das patentes de um dado procurador. No mapeamento, os conceitos ou assuntos são buscados e clusters dos documentos que possuem os mesmos conceitos são criados. Através da análise desses clusters, que agora encontram-se organizados, pode-se rapidamente obter uma idéia geral dos conceitos envolvidos na organização e como eles se relacionam.

Esses atributos não estruturados são representados pelos campos TITULO e RESUMO, que estão disponíveis nas patentes do INPI. Além desses campos, nas patentes internacionais também estão disponíveis detalhes técnicos que se apresentam sob forma não estruturada no campo DESCRIÇÃO, como no caso da base de dados

USPTO (*United States Patent and Trademark Office*) que pode ser encontrada em UNITED (2006).

Para processamento dos dados textuais é importante identificar atributos ou termos que contenham informações importantes para análises futuras. Em alguns casos, é suficiente encontrar significância em palavras dentro do texto, ou seja, encontrar entre o conjunto total de palavras aquelas que melhor representa o conteúdo do documento, e a partir dessa cadeia de caracteres, coletar estatísticas e usá-las para processamento.

O processo de mineração de textos é composto por várias etapas que podem ser executadas por diferentes técnicas de acordo com a que melhor satisfaça a base de dados a ser processada.

Por isso, a seguir são apresentados alguns detalhes do KDT úteis para o processamento das patentes brasileiras, que requerem técnicas especiais para a adaptação ao idioma português (LOPES, 2004).

3.4.1 Análise dos Dados

Pelo fato da manipulação de arquivos de texto ser de difícil interpretação pelo computador, uma etapa de preparação do texto é necessária no processo de descoberta do conhecimento (ZANASI, 2005).

Um forma bastante comum de representação do conjunto de documentos é sob um modelo geométrico, chamado de Modelo de Espaço Vetorial (ROSS *et al.*, 1997) ou VSM (*Vectorial Space Model*). Nessa representação, os documentos são representados por pontos (ou vetores) em um espaço Euclidiano t-dimensional onde cada dimensão corresponde a um termo do dicionário.

Os termos são representantes das palavras contidas dentro do vocabulário dos documentos. O conjunto total das palavras pertencentes ao espaço vetorial é chamado de dicionário (SALTON, 1989).

Uma grande vantagem na representação pelo VSM é a simplicidade de manipulação dos documentos e a facilidade de visualização. Além dos documentos, as consultas também podem ser interpretadas dentro do modelo, onde cada termo existente na consulta seria representado por uma coordenada dentro do espaço vetorial.

Cada termo possui um peso associado para descrever a sua importância dentro de um documento, o que define a localização deste no espaço vetorial. A distância de dois documentos ou de duas consultas irá definir a similaridade existente entre eles. Essa distância é calculada usando a medida de similaridade chamada de distância do cosseno, que corresponde ao cosseno do ângulo entre dois vetores. Essa medida é representada pela equação a seguir:

$$\cos(q, d) = \frac{\vec{q} \cdot \vec{d}}{\|\vec{q}\| \|\vec{d}\|}$$

ou seja,

$$\cos(q, d) = \frac{\sum_{i=1}^n q_i d_i}{\sqrt{\sum_{i=1}^n q_i^2} \sqrt{\sum_{i=1}^n d_i^2}} \quad (1)$$

onde, $q = (q_1, q_2, \dots, q_n)$ é o vetor da consulta com os pesos de cada termo da consulta q_i e $d = (d_1, d_2, \dots, d_n)$ é o vetor do documento com os respectivos pesos de cada termo d_i (SALTON, 1975).

Os pesos dos termos dos documentos e consultas são atribuídos de várias maneiras, onde os mais comuns deles são: binária, TF e TF*IDF (SALTON, 1983).

Uma análise eficiente do conjunto de documentos deverá detectar palavras com significados iguais e tratá-las como o mesmo termo. Por esse fato, o próximo passo a ser executado ocorre a nível morfológico, onde similaridades de significados e relevância dos termos são identificados, e a dimensão dos documentos pode ser drasticamente reduzida. Esse pré-processamento pode ser realizado através de listas de *stopwords*, onde estão contidos termos que não trazem conhecimento ao texto e por isso podem ser eliminados do processo. Além disso, os termos que apresentam frequência muito alta em todos os documentos também não são considerados discriminatórios na classificação das patentes, por isso, são considerados irrelevantes e acrescentados na lista. Como por exemplo, no caso da consulta ser baseada nos documentos que possuem a palavra “petróleo” no campo resumo, este termo não terá significância discriminatória para a clusterização.

O pré-processamento das patentes também inclui a aplicação de algoritmos de *stemming* para a redução de variações morfológicas para um único radical. Foram adaptados diferentes abordagens desse algoritmo para a língua portuguesa como o caso dos métodos Stemmer S (HARMAN, 1991), Porter (PORTER, 1980) e Lovins (LOVINS, 1968).

3.4.2 Algoritmo de Clustering

Clustering ou agrupamento, é uma técnica de aprendizado não supervisionado, com o objetivo de encontrar uma estrutura em uma coleção de dados que não sejam pré-classificados. Esse procedimento é utilizado para agrupar documentos semelhantes (YANG, *et al.*, 1997). Um sistema de clustering de texto deve desempenhar a tarefa fundamental de descobrir elementos compartilhados por documentos.

O critério de similaridade entre os documentos é dado pela distância existente entre eles. Dois ou mais documentos irão pertencer ao mesmo grupo se estiverem próximos um dos outros no espaço vetorial.

Para o método de clusterização k-means (MACQUEEN, 1967; FABER, 1994), dado um número fixo k , o algoritmo deverá achar k grupos de documentos.

O centro de cada *cluster* é definido como o vetor médio dos dados, ponderado por todos os itens do *cluster*.

O algoritmo inicia inserindo, aleatoriamente, k centróides ao espaço do conjunto. A seguinte iteração é então feita:

- Cada documento é associado ao cluster que possui o centróide mais próximo.
- Calcula-se o centróide de cada *cluster*.

A iteração termina quando não houver mais atualizações ou o número de iterações seja alcançada.

A distância entre os documentos e os centróides é definida como uma função objetivo, onde, nesse caso, a interpretação de documentos textuais é mais eficiente utilizando a medida do co-seno definida na equação (1).

Como os centróides são inseridos aleatoriamente no espaço do conjunto, uma solução diferente é encontrada a cada vez que o algoritmo é executado. Dessa forma, o algoritmo é executado diversas vezes, sendo que cada vez os centróides são iniciados numa posição diferente. A soma das distâncias dos documentos aos seus respectivos centros são armazenadas, e a solução que apresentar menor distância é considerada a melhor solução.

3.4.3 Aplicação de mineração de texto em patentes industriais

A mineração de textos é bastante utilizada no ramo de negócios, oferecendo ganho competitivo e facilidades na compreensão dos dados existentes.

Os resultados obtidos com a análise inteligente das fontes de recursos auxiliam em uma das tarefas mais difíceis existentes dentro de uma organização: a tomada de decisão.

Essas diretivas são amplamente diversificadas e utilizadas de acordo com a necessidade organizacional.

As aplicações de mineração de textos em documentos de patentes industriais permitem a utilização da tomada de decisão gerando impactos comerciais.

Esse processamento se torna indispensável para o entendimento do conteúdo das patentes quando a base de dados consiste de um grande número de documentos (KOSTER *et al*, 2001). Cada patente possui em média 5000 palavras, e milhares de patentes são depositadas por ano.

Sendo assim, a identificação de novos conhecimentos se torna uma tarefa bastante trabalhosa para ser executada por um ser humano. A aplicação de clusterização estrutura os dados e fornece melhor visualização das informações presente nos documentos.

A técnica consiste em criar clusters dos documentos, medindo similaridades entre as patentes e identificando as palavras-chaves desses clusters.

A partir desses clusters e das palavras-chave, é possível identificar diversas vantagens para utilização competitivas, como por exemplo, entender os principais

assuntos relacionados às patentes processadas, gerados por cada cluster, além de reconhecer as principais organizações ativas em cada assunto, assim realizando uma análise da concorrência.

Com essas análises, é possível reconhecer uma segmentação de mercado, agregar valor aos produtos, identificar novos concorrentes e novas tendências, identificar oportunidades e ameaças no mercado, e melhor aproveitar o capital intelectual.

A mineração de textos utilizada para a análise de patentes tem sido amplamente utilizada em diversas áreas de atuação, como por exemplo, empresas de automação, mercado farmacêutico, centros de pesquisa, áreas biológicas, serviços financeiros, telecomunicações, governamentais e outros.

3.5 Ferramentas de mineração de textos para patentes

Muitas ferramentas de mineração de textos têm sido desenvolvidas visando oferecer inteligência técnica competitiva e melhor gerenciamento tecnológico, auxiliando na extração de conhecimento de bases de dados textuais.

Através dessas ferramentas, a mineração dos textos pode ser facilmente aplicada e os resultados interpretados de diferentes formas.

A utilização de patentes como vantagem competitiva cria a necessidade de se possuir ferramentas próprias para o processamento desse tipo de dado. Como já citado anteriormente, as patentes possuem um formato especial e o tratamento das mesmas como simples dados textuais poderia ocasionar numa enorme perda de informações.

Por isso, pode ser notado um crescimento no número de empresas desenvolvedoras de softwares capazes de processar de maneira eficiente os atributos contidos nas patentes.

A Sociedade Internacional de Informações de Patentes é uma organização que busca apoiar o desenvolvimento de sistemas de análise e de pesquisa em informações de patentes (PIUG, 2006).

Existem várias empresas de fornecimento de serviços relacionados a patentes associadas a essas sociedade, como por exemplo, serviços de downloads de patentes, consultoria em busca por patentes específicas, proteção aos direitos de propriedade intelectual, e sendo do interesse dessa dissertação, organizações desenvolvedoras de sistemas de análises de documentos de propriedade intelectual.

Essas empresas oferecem diferentes tipos de sistemas, com diferentes acessos e processamentos dos dados textuais. Algumas das principais empresas podem ser encontradas em (ANALYSIS, 2006), além das citadas a seguir.

3.2.1 Wisdomain

A empresa Wisdomain (WISDOMAIN, 2006), oferece uma ferramenta de suporte à decisão baseada em dados de patentes chamada Focust.

A ferramenta é dividida em três módulos: o Módulo de Busca, o Módulo de Citação e o Módulo de Análise.

O Módulo de Busca fornece um acesso bastante flexível a diversas bases de patentes disponíveis na web, como US, EP, JP, PCT e INPADOC. Além disso, permite a visualização dessas patentes o que ajuda na organização através da interface da ferramenta. A busca pela patente permite a utilização de vários atributos, como busca por palavras-chave, número da patente, ou qualquer outro elemento presente na patente.

Através do Módulo de Citação, é possível criar a genealogia das patentes, ou seja, através de uma árvore, a ferramenta ilustra quais patentes foram precursoras de uma determinada tecnologia e quais patentes seguiram a partir da outra, através de uma análise de citação. Através dessa funcionalidade, é possível descobrir quais foram os principais responsáveis pela evolução da tecnologia, quem predomina na sua evolução, atualmente, além de outras análises.

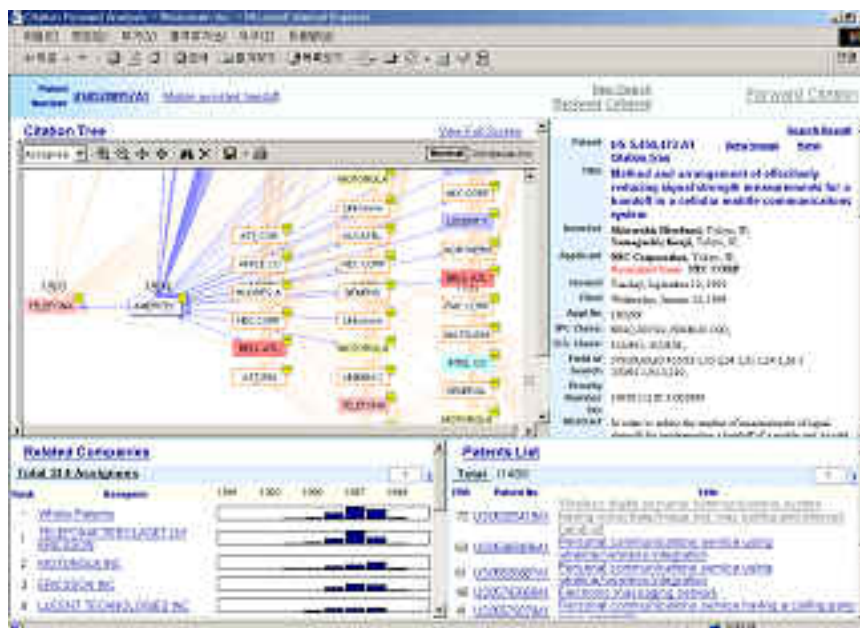


Figura 3-4 – Módulo de Citação

O Módulo de Análise realiza uma estratégia de análise que visa atender ao suporte à decisão através de áreas como inteligência competitiva, tendências de desenvolvimento industrial, valor da patente, ciclo de vida da patente e estratégias de comercialização.

Esse módulo apresenta três funcionalidades: a) análise por mineração de textos, que permite a geração de árvores através de palavras-chave. Os documentos são clusterizados por suas similaridades. b) visualização bi e tridimensional das estatísticas geradas pela ferramenta a partir das patentes, o que auxilia na compreensão dos gráficos e relatórios requeridos pelo usuário. E c) gerenciamento flexível dos documentos, que auxilia na ligação entre os documentos e qualquer atributo e, dessa forma, um conjunto de documentos podem ser visualizados, analisando apenas alguns de seus atributos.

3.2.2 Temis

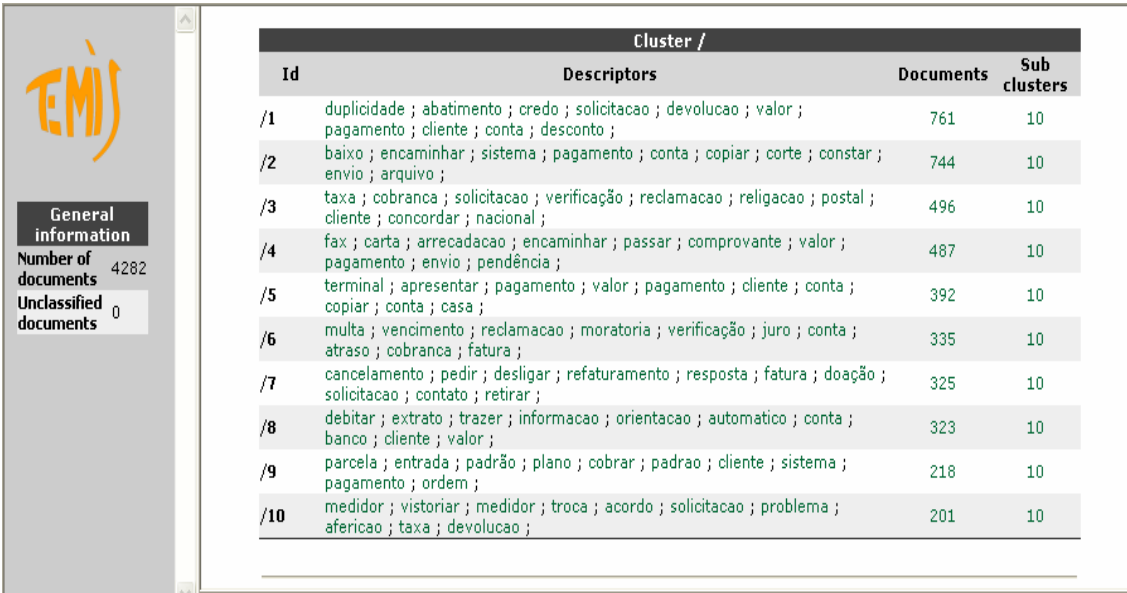
O Insight Discoverer™ Clusterer (IDC, 2006), desenvolvido pela Temis (TEMIS, 2006) é capaz de agrupar e classificar os documentos de acordo com as suas semelhanças semânticas.

O processo de clusterização é baseado em combinação lingüística e análise estatística. O seu pre-processamento é realizado de tal forma a distinguir os documentos através da morfologia e semântica dos seus termos, identificando-os como verbos, adjetivos ou nomes.

Com a utilização de cartuchos específicos para cada idioma, IDC utiliza informações gramaticais para auxiliar na identificação da classe de cada termo.

O usuário tem a possibilidade de determinar o nível que cada cluster pode assumir, e o número máximo de clusters que podem ser gerados.

O resultado é dado através de visualização por arquivos HTML, figura 3.5, onde é possível analisar grupos e sub-grupos encontrados pela ferramenta.



Cluster /			
Id	Descriptors	Documents	Sub clusters
/1	duplicidade ; abatimento ; credo ; solicitacao ; devolucao ; valor ; pagamento ; cliente ; conta ; desconto ;	761	10
/2	baixo ; encaminhar ; sistema ; pagamento ; conta ; copiar ; corte ; constar ; envio ; arquivo ;	744	10
/3	taxa ; cobranca ; solicitacao ; verificacao ; reclamacao ; religacao ; postal ; cliente ; concordar ; nacional ;	496	10
/4	fax ; carta ; arrecadacao ; encaminhar ; passar ; comprovante ; valor ; pagamento ; envio ; pendencia ;	487	10
/5	terminal ; apresentar ; pagamento ; valor ; pagamento ; cliente ; conta ; copiar ; conta ; casa ;	392	10
/6	multa ; vencimento ; reclamacao ; moratoria ; verificacao ; juro ; conta ; atraso ; cobranca ; fatura ;	335	10
/7	cancelamento ; pedir ; desligar ; refaturamento ; resposta ; fatura ; doacao ; solicitacao ; contato ; retirar ;	325	10
/8	debitar ; extrato ; trazer ; informacao ; orientacao ; automatico ; conta ; banco ; cliente ; valor ;	323	10
/9	parcela ; entrada ; padrao ; plano ; cobrar ; padrao ; cliente ; sistema ; pagamento ; ordem ;	218	10
/10	medidor ; vistoriar ; medidor ; troca ; acordo ; solicitacao ; problema ; afericao ; taxa ; devolucao ;	201	10

Figura 3-5 – Temis, módulo IDC

3.2.3 VantagePoint

A VantagePoint (VANTAGEPOINT, 2006) foi desenvolvida para auxiliar gerentes técnicos e profissionais de inteligência técnica competitiva a extraírem conhecimento novo e útil de bases de dados de patentes.

Os atributos são minerados através de busca por padrões, e técnicas baseadas em regras e processamento de linguagem natural.

Essa ferramenta foi desenvolvida para trabalhar com dados bibliográficos, e dessa forma, consegue distinguir os diferentes campos como autor, título, data, país e outros.

Através de matriz de covariância, o usuário pode fazer associações como autor e ano de publicação para identificar tendências de publicações ao longo do tempo.

Além disso, através da análise estatística multidimensional, VantagePoint pode identificar clusters e relações entre conceitos, autores, procuradores e países.

Cada nó da figura 3.6 representa termos combinados de acordo com a frequência que ocorrem juntamente. Os nós da proximidade apresentam a correlação entre os termos.

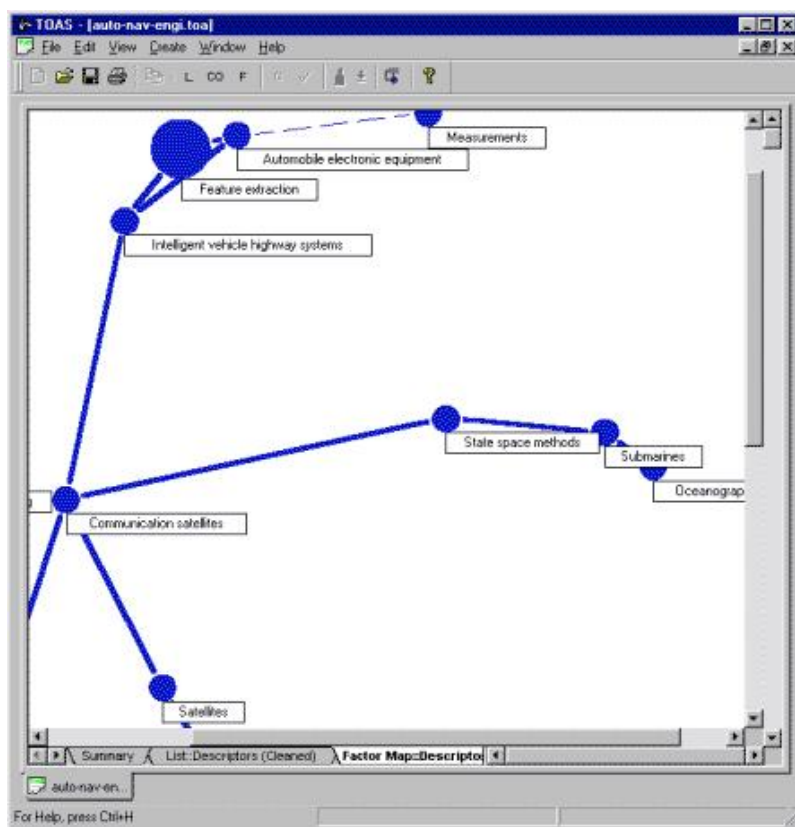


Figura 3-6 – VantagePoint

Através desse mapa, o usuário pode visualizar qualquer atributo do conjunto de documentos, dessa forma, as correlações mostram mais claramente as tendências ao longo do tempo.

Um thesaurus é utilizado pelo VantagePoint onde o usuário possui permissão de edição para melhor especialização e redução da quantidade de termos da base de dados, além de utilizar técnicas de combinação fuzzy para identificar, associar, e reduzir o tamanho total da base.

3.2.4 Aureka

A software Aureka, desenvolvido pela MicroPatent (MICROPATENT, 2006), possui acesso a diversas repositórios de patentes como US, DE, EP, GB, JP (apenas resumos) e autorias de PCT. Além disso, Aureka é capaz de importar documentos como artigos de conferência, artigos de jornal, referências de estado da arte, relatórios técnicos, gráficos, invenções em aberto e outros.

Essas patentes são acessadas através da funcionalidade chamada *PowerBronse*, onde centenas de patentes são facilmente acessadas e com isso, apenas as patentes que o usuário considerar importante serão armazenadas para análise posterior.

Aureka possui um sistema de diretórios, que auxiliam na armazenagem das patentes selecionadas e essas informações são compartilhadas com os diversos usuários da organização.

Apresenta dois modos de visualização dos resultados: o modo *ThemeScape*, representado pela figura 3.7 e o modo *CitationAnalisys* representado pela figura 3.8.

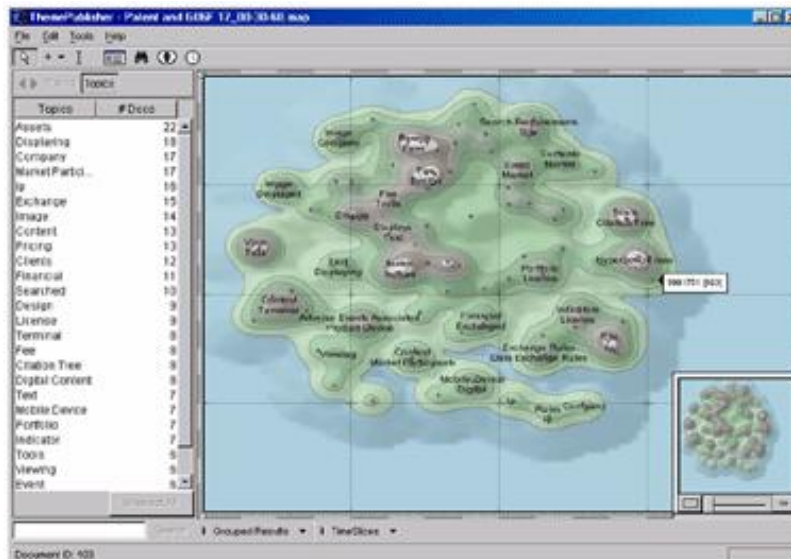


Figura 3-7 – ThemeScape

O modo de visualização através do ThemeScape é uma funcionalidade baseada no mapeamento de conceito. É capaz de analisar estatisticamente quais são as palavras chave ou os tópicos que os documentos possuem em comum.

Os temas, são representados visualmente como contorno no mapa. Os picos representam os assuntos principais. A proximidade dos pontos aos picos representa a relação entre os termos técnicos e o assunto.

Através dessa funcionalidade é possível identificar onde as patentes de uma organização estão relacionadas com as de outras.

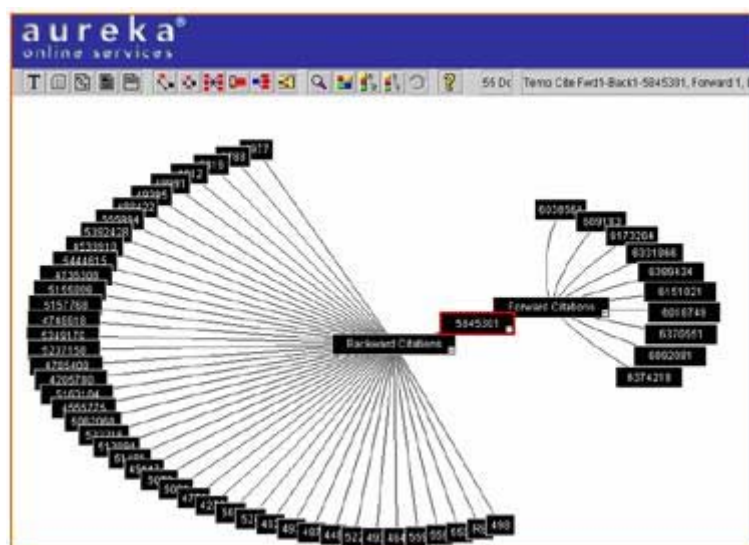


Figura 3-8 – Citation Analysis

CitationAnalisys determina as citações existentes na patente de interesse. Através dessa funcionalidade é possível identificar tendências tecnológicas em uma determinada área.

A representação é dada por uma árvore hiperbólica que demonstra as patentes que referenciam e que são referenciadas por outras. Isso permite encontrar a raiz da tecnologia e entender as direções pela qual aquela tecnologia evoluiu.

3.2.5 Anacubis

O analisador de Propriedade Intelectual da Anacubis (ANACUBIS, 2006) foi desenvolvido visando atender as necessidades de profissionais de patentes. Faz parte da desenvolvedora i2 ChoicePoint's e seu grupo de desenvolvimento. A i2 ChoicePoint's fornece solução de análise investigativa para aplicação de leis em agências governamentais e no setor comercial. Está preparado para ligar diretamente a qualquer sistema de banco de dados e integra com qualquer outro produto i2.

Através do i2 *Analyst's Notebook* é possível apresentar os eventos de acordo com as suas ordens cronológicas, além de revelar estruturas, detectar padrões de crimes e identificar pontos chaves através de análise espacial.

A figura 3.9 mostra o módulo de visualização dos documentos e as conexões que a ferramenta é capaz de efetuar entre pessoas, contas bancárias, organizações e qualquer outro elemento sob investigação. Além disso, os gráficos criados pelos usuários podem facilmente ser compartilhados através do módulo *ChartReader*.

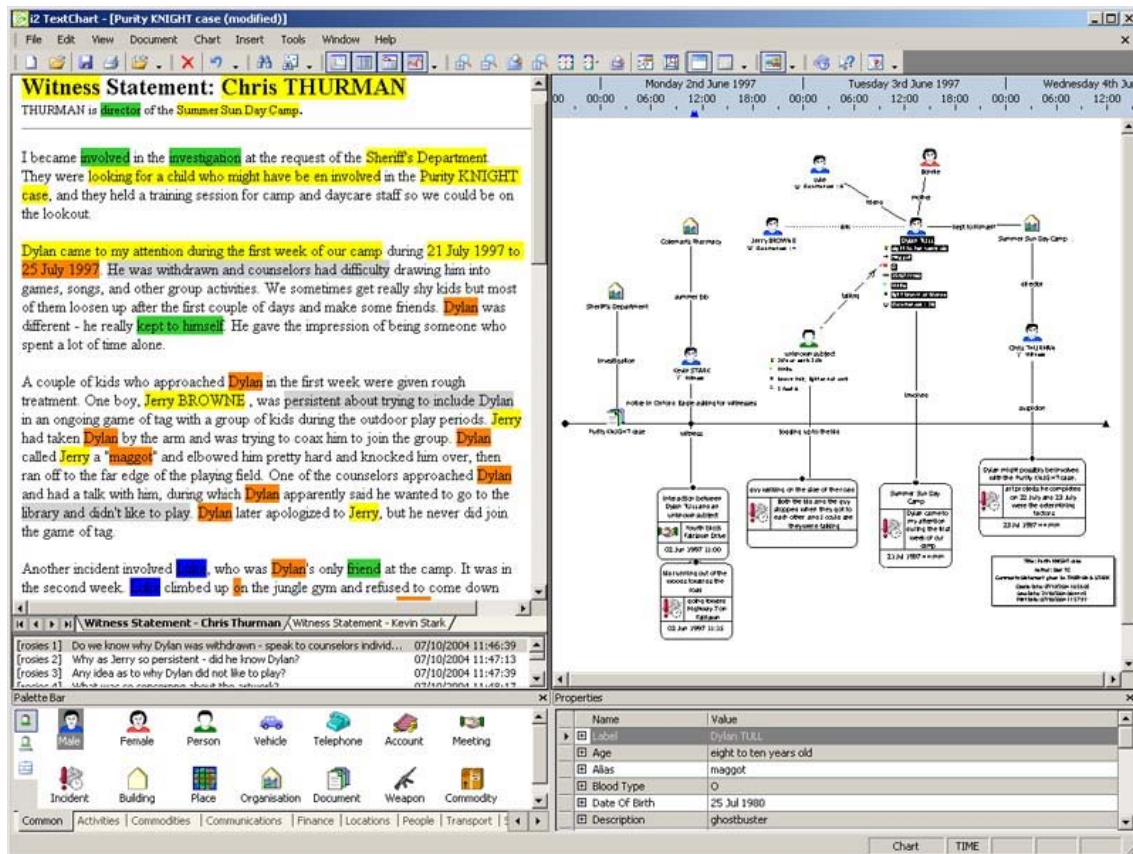


Figura 3-9 – Anacubis

Anacubis tem sido utilizado numa variedade de soluções como na área militar, no combate ao terrorismo, contra o crime organizado, drogas, volume criminal, detecção de fraudes financeiras, prisões, telecomunicações e outros.

3.2.6 BizInt Smart Charts

A BizInt Smart Charts para Patentes (BIZINT, 2006) utiliza como base de dados vários repositórios de patentes, como Derwent World Patents Index, Claims, CA/Caplus, WPI, MicroPatent, Delphion e outros.

A atual versão 3.1 armazena os dados em forma de relatórios tabulares, onde cada coluna representa um atributo da patente, como depositante, título e imagem, e cada linha representa uma patente. Através de duplo clique sobre a linha, uma nova janela é aberta dando acesso à patente completa, e da mesma forma, é possível ter acesso às imagens referentes à patente. Essas tabelas são editáveis e pode-se acrescentar

novas linhas e novas colunas. Esses relatórios podem ser exportados para formato HTML ou Excel, como mostra a figura 3.10.

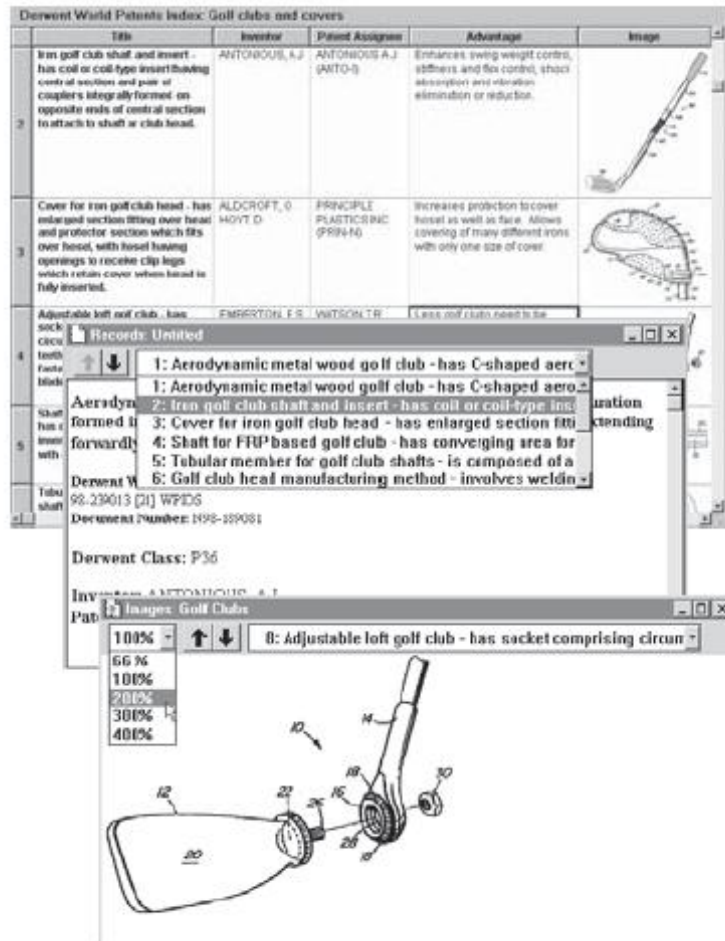


Figura 3-10 – BizInt Smart Chart

É possível também criar os relatórios tabulares importando e combinando resultados de buscas de ferramentas como STN, Questel*Orbit, Dialog, MicroPatent ou Delphion, além de gerar estatísticas de acordo com os dados selecionados.

3.2.7 ClearForest

A ClearForest (CLEARFOREST, 2006) oferece ferramentas para pesquisas intensivas que visam auxiliar em processos de inteligência de negócios.

O módulo ClearTags, através da utilização de tecnologias de tags, semântica avançada, e estatísticas, é capaz de produzir tags estruturais de informações críticas existentes entre os documentos e classificar alguns conceitos básicos como “pessoa”, “companhia” ou “localização”.

Uma vez que os dados estejam separados pelo seu conceito, dado pelo módulo de Tags, o ClearForest Extraction Modules, módulos específicos de extração de conhecimento, é capaz de identificar entidades importantes em uma indústria, e extrair o relacionamento existente entre esses elementos.

Esses módulos atuam sobre bases específicas com diferentes objetivos. O módulo de análise de patentes, desenvolvida para profissionais de propriedade intelectual, age sobre documentos de patentes visando diminuir o tempo do ciclo entre executar e tomar uma decisão, acentuar capacidades de análise competitiva, obter melhor percepção nos esforços de parceiros de P&D e competidores e diminuir o custo de pesquisas.

Os resultados extraídos desse módulo incluem: principais competidores de uma determinada área, patentes mais importantes e gráficos cronológicos.

3.2.8 Statistica

A ferramenta *Statistica*, desenvolvida pela *StatSoft* (STATSOFT, 2006), tem a capacidade de manipular, gerenciar, e visualizar vetores de dados. Possui procedimentos de mineração de dados, tais como classificação, clusterização, predição e técnicas exploratórias.

Apresenta um módulo de conversão de texto para dados numéricos, utilizando algoritmos de frequência dos termos nos documentos, frequência binária, log da frequência e frequência inversa dos documentos.

Além disso, utiliza listas de *StopWords* e algoritmos de *Stemming* (com suporte para vários idiomas, inclusive o Português).

Uma vez que os dados tenham sido convertidos para o formato numérico, a ferramenta trata o problema como uma mineração de dados.

Na figura 3.11 é apresentado o conjunto de passos a serem realizados para o processamento de textos. Este processo inclui o pré-processamento, a alteração dos resultados do pré-processamento, ou seja, junção ou eliminação de outros termos e o resultado da clusterização.

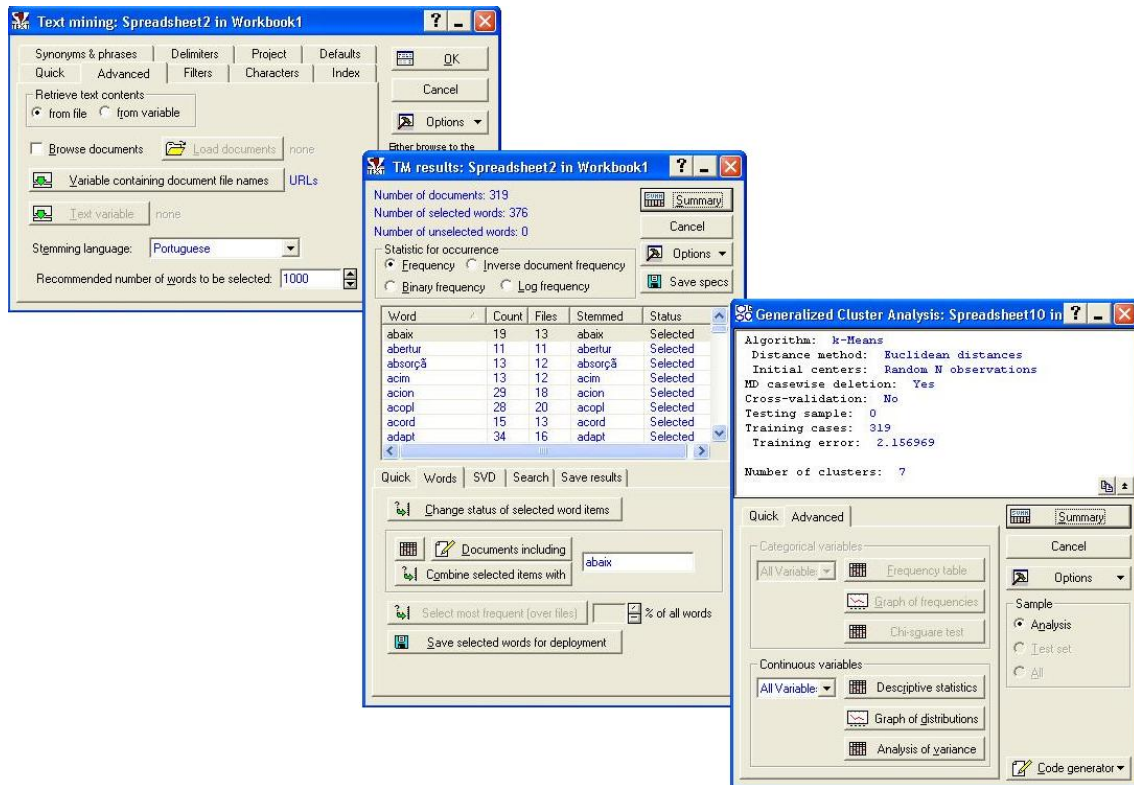


Figura 3-11 – Statística – Módulo Textual

Para utilizar dados de patentes no *Statistica*, apenas os campos textuais são inseridos, tratados como um conjunto de arquivos, cada um contendo um texto. A clusterização retorna como resultado os documentos agrupados pela clusterização e os seus respectivos centróides.

4 Sistema de Mineração de Textos nos Documentos de Patentes

Para a aplicação de métodos computacionais em patentes, como a mineração de textos, são necessários procedimentos especiais, capazes de manipular os atributos presentes em tais documentos, como RESUMO, DATA, e outros, de maneira a fornecer o maior ganho de informação inclusa nos mesmos.

Por esse fato, uma ferramenta de mineração de textos foi implementada e adequadamente modelada para manipular patentes e melhor visualizar os resultados obtidos da mineração de textos.

A ferramenta desenvolvida tem o objetivo de utilizar algumas das técnicas de processamento de documentos, sendo utilizada para a manipulação e entendimento de conteúdo de patentes brasileiras e a partir dos resultados, ser capaz de auxiliar no processo de tomada de decisão.

A mineração de textos foi adotada de acordo com a necessidade de processamento das patentes, visando obter melhores resultados.

Além disso, algumas propriedades estatísticas foram implementadas, visando entender relações existentes entre as patentes, e conseqüentemente, a relação existente entre as áreas atuação das mesmas.

A implementação da fase de pré-processamento foi testada com diversos algoritmos, sendo que apenas os que forneceram melhores resultados para os documentos de patente fazem parte do sistema atual.

A seguir, será descrita a linguagem utilizada para a implementação da ferramenta, os algoritmos utilizados no pré-processamento e para obtenção dos clusters, tal como as opções de processamento dos atributos e as propriedades de análises estatísticas que podem ser utilizadas nas patentes.

4.1 Linguagem Java

A linguagem Java tem ocupado um espaço muito grande entre as ferramentas de desenvolvimento de sistemas por apresentar um grande número de vantagens.

Entre essas vantagens, encontra-se o fato da linguagem possuir diversas bibliotecas que são facilmente anexáveis a qualquer sistema, sendo que algumas podem ser encontradas gratuitamente na internet. Essas bibliotecas são chamadas de APIs.

Dentre essas APIs, encontramos várias bibliotecas gráficas que podem oferecer melhor visualização gráfica. E além disso, o fato de surgir a partir da linguagem C a torna extremamente simples.

A linguagem é disponibilizada gratuitamente através da rede de desenvolvimento Sun (SUN, 2006), e a ferramenta utilizada para desenvolvimento do sistema foi o Eclipse (ECLIPSE, 2006), por ser capaz de manipular java com grande praticidade.

Além disso, o fato da linguagem ser orientada a objetos, é capaz de simular mais facilmente problemas reais. Esse fato é de suma importância, pois oferece portabilidade e facilita a inserção de novos requisitos futuros.

Por esses motivos, o sistema apresentado nesta dissertação foi desenvolvido a partir da linguagem java, visando retirar o melhor proveito, alcançar um melhor desempenho, e dar suporte a futuras necessidades e melhorias a serem realizadas.

4.2 Captura dos dados

A escolha adequada dos dados a ser processados pelo sistema de mineração de textos é um fator de suma importância para a obtenção de resultados consistentes e relevantes. O uso de documentos irrelevantes para uma determinada aplicação reduz o grau de confiabilidade nos resultados e diminui o desempenho do processo.

Nesse sentido, algumas formas de busca de patentes tem sido estudadas, como pode ser visto em DEBOYS (2004), que define a melhor seqüência de decisões a serem tomadas na busca pelos documentos.

Alguns fatores determinam uma boa qualidade das patentes buscadas e os efeitos e riscos associados a decisões comerciais. Alguns pontos podem ser destacados durante a tomada de decisão, como por exemplo:

- Definição daquilo que pode ser buscado
- Definição do que pode ser esperado do resultado da busca levando em conta os recursos disponíveis e suas limitações
- Definição de atributos como palavras-chave, classificações, depositantes, ano, e outros.
- Definição dos atributos relevantes das patentes buscadas e que serão usados na análise.

Todos esses fatores são de extrema importância para a obtenção de resultados consistentes e relevantes. Os detalhes dessa seqüência de passos para a tomada de decisão na busca por patentes pode ser encontrado em FLETCHER (1992).

Pelo fato do site do INPI possuir dificuldades em capturar esses dados, uma ferramenta isolada foi desenvolvida com o objetivo de automatizar essa captura.

O processo de captura dos dados é dividido em duas partes: a) depois de realizada a consulta, os números dos depósitos das patentes retornadas são salvos em um arquivo de formato texto. b) a ferramenta tem a função de ler os 19 primeiros números contidos na lista, espera até que o usuário realize a autenticação manual requerida pela página de consultas do INPI, captura as patentes equivalentes aos números lidos e parte para os 19 próximos números. O processo de autenticação e leitura continua até que todos os números de depósito tenham sido lidos e as patentes armazenadas.

A figura 4.1 representa uma patente recuperada de uma consulta realizada pela ferramenta. As propriedades da patente são preservadas e dessa forma não há perda de informação. Essas propriedades são caracterizadas pelos atributos que a patente possui, e os mesmos são acompanhados dos números que os caracterizam, como no exemplo da figura 4.1, a DATA DE DEPOSITO é identificado pelo número 22.

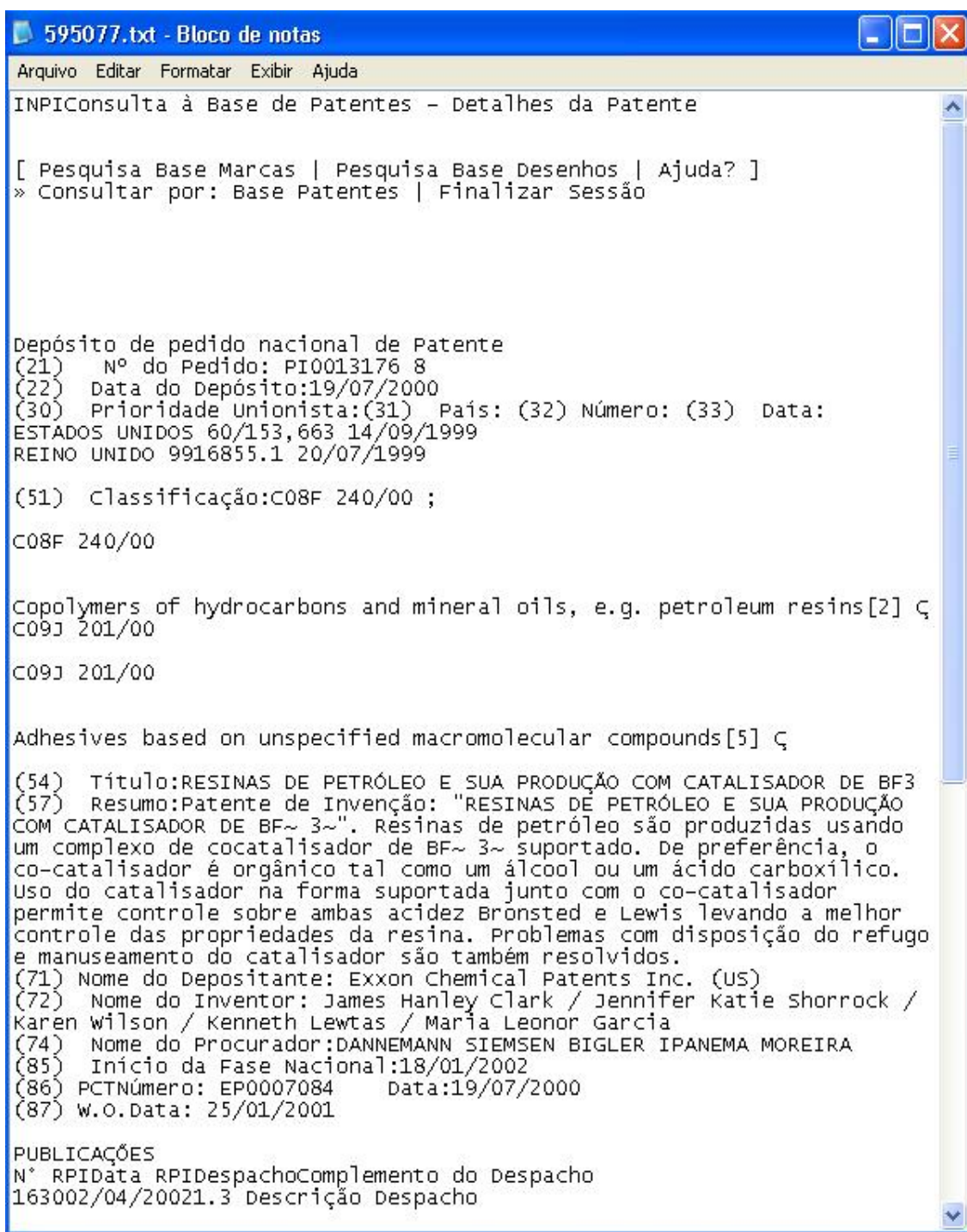


Figura 4-1 – Patente capturada pelo Sistema

4.3 Descrição do Sistema

O sistema de Análise de Patentes foi desenvolvido de tal forma a manipular arquivos de formato texto, representado na figura 4.1, conforme aqueles salvos pela ferramenta descrita anteriormente.

O sistema foi desenvolvido para realizar alguns processamentos de análise das patentes, que podem ser caracterizados pela análise estatística dos atributos CLASSIFICAÇÕES e DEPOSITANTES das patentes, e pela mineração de textos realizado no atributo RESUMO.

A figura 4.2 apresenta a tela inicial do sistema com as opções de processamento e visualizações.

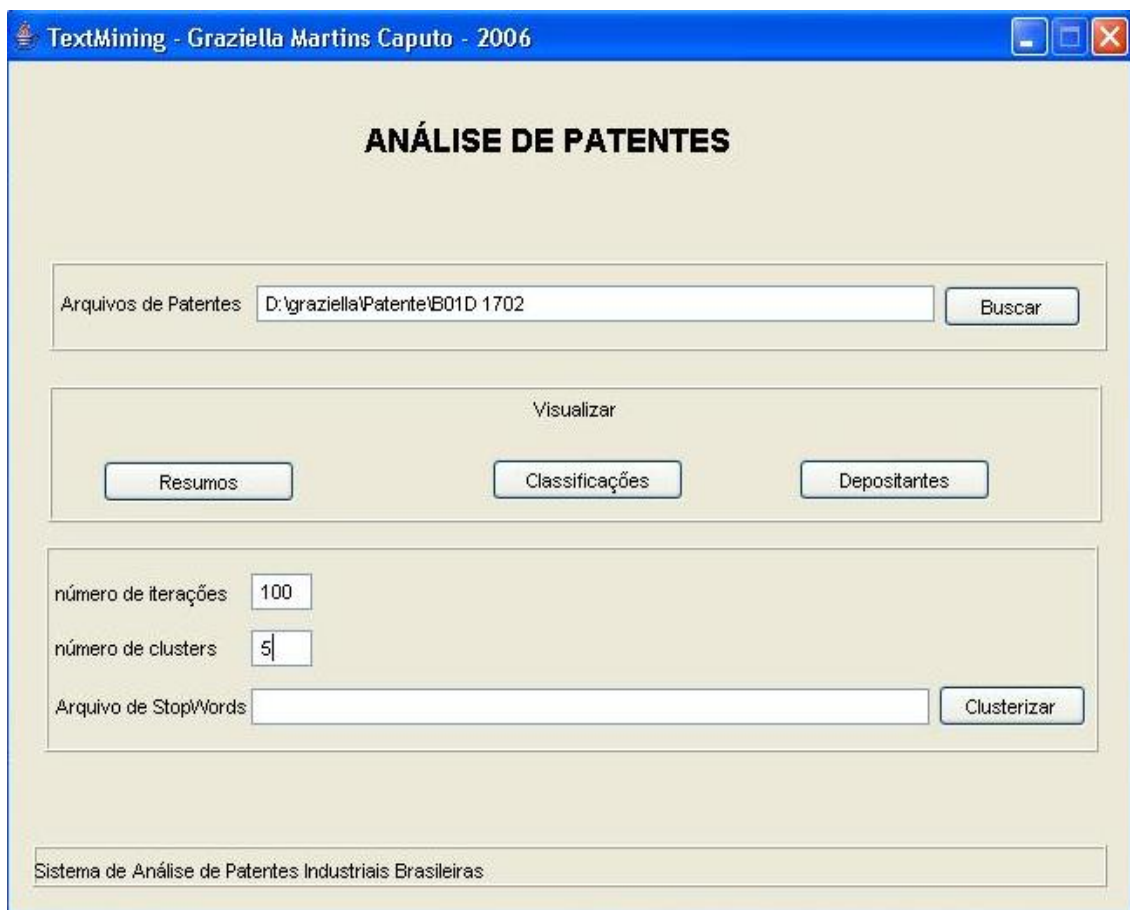


Figura 4-2 – Tela Principal

Os documentos a serem processados são selecionados a partir da opção “Buscar”, onde o diretório selecionado pode ser visualizado no campo “Arquivos de Patentes”. Todas as operações de busca por informação são realizadas sobre esses arquivos.

No geral, os atributos a serem utilizados por cada etapa são identificados pelos números que os caracterizam.

4.3.1 Análises Estatísticas

Das patentes utilizadas no processo de descoberta de conteúdo, é possível identificar determinados valores estatísticos que podem oferecer ganho no que diz respeito a diversificação do conteúdo dos documentos e a relação existente entre as patentes.

Para esse processo de análise estatística, dois atributos foram utilizados no sistema: CLASSIFICAÇÃO e DEPOSITANTES.

Classificação

Dado um conjunto de patentes, obtido a partir da execução de uma consulta qualquer na base de dados de patentes, a frequência com que cada CLASSIFICAÇÃO aparece demonstra os principais setores que se destacam no conjunto, como por exemplo, física ou eletricidade, ou suas classes, subclasses e subgrupos, demonstrando qual o principal assunto referente ao conjunto. A partir dessa informação, pode-se perceber também como um conjunto de patentes retornados por uma mesma consulta abrange diferentes áreas.

A figura 4.3 apresenta a tela presente no sistema cujo papel é listar as classificações existentes no conjunto, tal como a frequência com que cada uma aparece.

Classificação	Frequência
B01D	86
C02F	10
E21B	4
B01J	4
B04C	3
B03B	2
B03D	2
B08B	2
E02B	2
C07K	2
F02M	2
C07C	2
B01F	1
E03C	1
A47B	1
D06L	1
C08F	1
G01F	1
C22B	1
C11B	1
A61K	1

Figura 4-3 – Tela das Classificações existente no conjunto de Patentes processadas

As informações apresentadas na tela de Classificações são obtidas a partir dos documentos recuperados na busca pelas patentes e que possuem o formato apresentado na figura 4.1. O atributo CLASSIFICAÇÕES é obtido a partir do número 51 (cinquenta e um), que é padrão do atributo em todas as patentes internacionais de acordo com o PCT.

Depositantes

A relação dos DEPOSITANTES e o número de patentes industriais depositadas por cada um deles pode auxiliar no processo de Inteligência Competitiva a descobrir quais são os maiores concorrentes em um determinado produto e os maiores detentores de alguma determinada tecnologia.

Além disso, através dessa lista de depositantes, é possível perceber parcerias realizadas entre pesquisadores, empresas ou institutos de pesquisa para o desenvolvimento e pesquisa de um produto ou tecnologia.

A figura 4.4 apresenta a tela no Sistema responsável pelo processamento e visualização dos depositantes das patentes processadas.



Depositantes	Frequência
Petresco International LTD. (GB)	2
José Carlos Dos Santos (BR/SP)	2
Filterwerk Mann & Hummel GmbH (DE)	2
Técnicas Reunidas, S.A. (ES)	1
Tadayoshi Sato (JP)	1
Norsk Hydro ASA (NO)	1
The Procter & Gamble Company (US)	1
Mauro José de Aguiar Bichara Junior (BR/RJ)	1
Henri Bernard Tettelin (BR/SP)	1
Epcon Norge AS (NO)	1
Water Minerals Limited (ZA)	1
ABB Research LTD. (CH)	1
Kvaerner	1
Alfa Laval Corporate AB (SE)	1
Petrobras Distribuidora S.A. (BR/RJ)	1
Alliedsignal INC (US)	1
Shel Internationale Research Maatschappij B ...	1
Kvaerner Process Systems A.S. (NO) / Den ...	1
Johnson Matthey PLC (GB)	1
Fleetguard, INC. (US)	1
Kvaerner Process Systems A.S. (NO)	1

Figura 4-4 – Tela dos Depositantes existente no conjunto de Patentes processadas

O atributo DEPOSITANTE é obtido a partir do número 71 (setenta e um) existente na patente conforme o formato apresentado na figura 4.1. Esse número é padrão conforme determinado pelo PCT.

Um problema bastante comum encontrado nas patentes é a diferente forma de digitalização dos nomes dos depositantes no atributo. O mesmo depositante pode aparecer com pequenas variações no nome, o que impediria o sistema reconhecê-lo como um só depositante. Um exemplo disso é o depositante “Petroleo Brasileiro S/A. - PETROBRÁS (BR/RJ)” e “Petroleo Brasileiro S.A.- Pretrobras (BR/RJ)”, que possui erros de digitação e caracteres diferentes.

Para contornar tal situação, o sistema utiliza uma lista, chamada de Alias, com as variações que podem ser encontradas nos diferentes depositantes e auxilia o sistema a identificar grupos similares de depositantes. Essas relações existentes no arquivo Alias

são digitadas pelo próprio usuário a partir da observação dos depositantes e da necessidade de novos agrupamentos.

4.3.2 Pré-Processamento

Os termos presentes nos documentos são lidos um a um e armazenados em um vetor, chamado de dicionário. Paralelamente, para cada documento, um vetor v é criado (sendo $v = 1 \dots d$, e d o número de documentos), onde a quantidade de cada termo dentro do documento é armazenada, juntamente com o índice do termo dentro do dicionário (SALTON, 1989).

O primeiro processo realizado nos termos é o *case folding*, que facilita a identificação das palavras, convertendo-as para o mesmo tipo de caractere. Nessa implementação o *case folding* utilizado foi o de caracteres minúsculos.

A seguir, uma lista de *StopWords* foi utilizada para retirar as palavras que não apresentavam relevância para a clusterização. Essa lista está contida internamente no sistema e contém termos comumente utilizados para a língua portuguesa, como conjunções e artigos. Além disso, o usuário tem a liberdade de inserir uma nova lista de termos que o mesmo considerar irrelevante para o processo através do campo “Arquivo de StopWords”, presente na tela inicial do sistema, conforme descrito na figura 4.2.

É importante ressaltar ainda que qualquer tipo de caractere não alfabético foi eliminado durante a fase de pré-processamento. Esses caracteres incluem tanto números quanto caracteres especiais como traços, aspas, parênteses e outros.

O algoritmo utilizado para a redução dos termos até os seus radicais foi o *Stemmer Portuguese* ou RSLP (Removedor de Sufixos para a Língua Portuguesa). O algoritmo é composto de uma seqüência de passos que reduz gradativamente os sufixos dos termos, iniciando pela remoção do plural, seguindo da remoção de feminino, redução do advérbio, redução do aumentativo/diminutivo, redução de sufixo de nome, redução de verbo, redução da última vogal, e finalmente, a redução dos acentos. Um melhor detalhamento sobre o algoritmo de *stemming* pode ser encontrado em ORENGO (2001) e CHAVES (2003).

Os termos resultantes do processo de pré-processamento, reduzidos aos seus radicais, são atualizados no vetor de dicionário.

Um passo que desempenhou um papel bastante importante na clusterização, melhorando os resultados do processamento, foi a retirada dos termos que possuíam baixa presença entre os documentos e aqueles que possuíam uma presença muito alta entre os documentos. Constatou-se que palavras que apareciam em apenas alguns documentos, e às vezes em apenas um, não ofereciam ganho discriminatório entre os grupos, tal como aqueles que estavam presente em todos.

Os termos resultantes são submetidos à medida de atribuição de pesos TFxIDF, onde cada termo dentro de cada documento recebe um valor relacionado à importância deste dentro do documento. Esses pesos são armazenados nos vetores dos documentos juntamente com o índice relativo a cada termo dentro do dicionário.

A partir dos vetores de valores obtidos, os dados estão preparados para o processo de clusterização.

4.3.3 Clustering

O *clustering* aplicado aos documentos de patente tem o objetivo de encontrar entre os documentos, aqueles grupos que se assemelham.

A mineração de textos pode ser aplicada em diversos campos de uma patente, como por exemplo, o TITULO, o RESUMO e a DESCRICAO. As patentes utilizadas nessa dissertação, no entanto, possuem apenas os campos TITULO e DESCRICAO.

A opção “Resumos”, implementada no sistema, tem a propriedade de selecionar e armazenar os resumos das patentes selecionadas no campo “Arquivos de Patentes”. Essa funcionalidade pode ser estendida, futuramente, para a clusterização apenas dos títulos das patentes, e para a descrição, caso as patentes a possuam.

O campo RESUMO é obtido a partir do número 57 (cinquenta e sete), conforme indicado na figura 4.1, padrão determinado pelo PCT.

Uma vez aplicado o pré-processamento nos dados o processo de clusterização é iniciado.

O algoritmo kmeans é aplicado aos dados utilizando como medida de distância entre os clusters a medida do cosseno. A quantidade de clusters é predeterminada no campo “número de clusters”, e o algoritmo é executado até que os resultados se estabilizem ou a quantidade de iterações do campo “número de iterações” seja alcançado.

Por configuração, caso o usuário não inserira o número de iterações, o algoritmo irá executar 100 iterações. Além disso, o número de clusters pode ser determinado pelo usuário ou o algoritmo irá supor buscar por dez clusters.

A tela retornada pela clusterização é apresentada na figura 4.5. Para cada cluster, o sistema imprime os 10 termos (radicais das palavras) que mais caracterizam aquele elemento.

The screenshot shows a window titled "TextMining - Graziella Martins Caputo - 2006". The main content area is titled "CLUSTER" and contains a table with the following data:

Cluster	Palavras-Chave
Cluster 0	rosc, extrem, execut, disposic, util, circul, articul, retangul, destin, reduz
Cluster 1	agu, utiliz, form, sist, mont, util, broc, constitu, op, convenc
Cluster 2	perfur, tr, execut, moviment, possu, movel, broc, invenc, perfeit, min
Cluster 3	ferrament, perfur, porc, pass, inclu, relat, execuc, horizont, interi, moviment
Cluster 4	produc, circul, interi, util, inferi, forc, pass, abetur, mol, obt

Below the table is a section titled "Documentos" which contains a list of documents associated with Cluster 1:

- Cluster 1
- 221866.txt
- 221944.txt
- 222561.txt
- 239549.txt
- 246967.txt
- 248600.txt
- 249430.txt

To the right of the document list are two buttons: "Exportar Lista de Documentos" and "Visualizar Documentos".

Figura 4-5 – Tela de Visualização dos Resultados da Clusterização

Ao selecionar uma linha na tabela de cluster, uma tabela imprime os nomes dos documentos de patentes relacionados ao cluster referente àquela linha. Para melhor entendimento do conteúdo do cluster, a opção “Visualizar Documentos” abre um arquivo texto com todas as patentes que pertencem ao cluster selecionado.

A opção “Exportar Lista de Documentos” salva em um arquivo texto os documentos pertencentes àquele cluster. Dessa forma, ao retornar para a tela principal, representada na figura 4.2, caso a seleção do “Arquivo de Patentes” seja esse texto com documentos, as opções de “Classificação” e “Depositantes” irá revelar os principais depositantes e classificações encontradas em cada cluster.

Dessa forma, as opções “Visualizar Documento”, “Classificações” e “Depositantes” irá auxiliar no processo de entendimento de cada cluster e consequentemente, melhorar o processo de tomada de decisão.

5 Estudo de Caso

Neste capítulo, será apresentada a aplicação da ferramenta de mineração de textos desenvolvida para o processamento e busca de conhecimento em patentes e os resultados obtidos das bases de dados utilizadas. As bases foram obtidas a partir do processo descrito anteriormente. Os resultados obtidos da ferramenta de patente serão comparados com os resultados obtidos a partir de outras duas ferramentas de mineração de textos já existentes: o módulo IDC desenvolvido pela Temis e o programa Statistica, desenvolvido pela StatSoft.

5.1 Base de Dados

As bases de dados utilizadas na busca por conhecimento na presente dissertação foram escolhidas através de diversos processamentos anteriormente realizados e testes de bases de dados.

Chegou-se à conclusão que alguns atributos possuíam grande importância na representatividade do atual foco de pesquisa e desenvolvimento.

Os estudos foram baseados nas classificações e palavras-chave que apareciam com grande frequência nas patentes mais atuais disponíveis no site do INPI.

Dentre os atributos que receberam maior destaque, uma palavra-chave e duas classificações foram selecionadas para compor a base de dados a ser utilizada no processamento da presente dissertação.

As patentes foram divididas de acordo com cada atributo e por isso foram geradas duas bases de dados: uma base com documentos de patentes que possuíam a

classificação “E21B” e uma base com documentos de patentes que possuíam o termo “petróleo” inseridos no campo RESUMO.

Essas bases serão descritas a seguir.

5.1.1 E21B

Essa classificação, de acordo com a sétima edição (1999) da Classificação Internacional de Patentes (CIP), volume 5, tem a seguinte descrição:

- Seção E: Construções Fixas.
- Subseção 21 da seção E: Perfuração do Solo; Mineração; Obtenção de Fluidos de Poços.
- Subclasse E21B: Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços.

Foram encontradas 2535 patentes com essa classificação, sendo a mais antiga delas, depositada em janeiro de 1982, e 1412 depositadas nos últimos dez anos.

5.1.2 Petróleo

Para efeito de busca por tendências tecnológicas e industriais desenvolvidas ao longo dos anos, as patentes coletadas com a palavra-chave “petróleo” foram separadas utilizando como critério, o ano do depósito da patente.

Dessa forma, as patentes dos últimos dez anos que possuíam a palavra-chave “petróleo” presente no atributo RESUMO, foram selecionadas para pertencer à base de dados, ou seja, aquelas depositadas entre 1996 e 2005.

No entanto, o número de patentes depositadas em cada ano é muito baixo, e por esse motivo, as patentes foram reagrupadas formando dois grupos de cinco anos cada, um entre os anos de 1996 e 2000 e o outro entre os anos de 2001 e 2005.

A tabela 5.1 apresenta um resumo do número de patentes obtidas na fase de captura dos dados, separadas pelo respectivo conjunto de acordo com o ano em que foi depositada.

Ano	Qtd. de patentes com o termo “Petróleo”
1996 - 2000	226
2001 - 2005	319
Total	545

Tabela 5-1 - Quantidades de patentes de Petróleo coletadas

No total, 545 patentes relacionadas ao termo “Petróleo” foram depositadas nos últimos dez anos.

5.2 Análise dos Dados

As análises das bases de dados foram realizadas com relação às características presentes na ferramenta de manipulação de patentes implementada para a presente dissertação.

A seguir serão apresentados os resultados e conclusões das análises estatísticas realizadas nas bases de patentes, verificando as CLASSIFICAÇÕES e DEPOSITANTES de maior destaque nas bases.

5.2.1 E21B – Classificação

Dentre as patentes que possuem a classificação E21B podem-se observar várias subseções presentes, como mostrado na figura 5.1.

Classificação	Frequência
E21B 43/00	105
E21B 43/12	104
E21B 17/00	83
E21B 17/01	81
E21B 43/01	72
E21B 37/06	62
E21B 47/12	62
E21B 33/13	62
E21B 43/22	60
E21B 43/10	57
E21B 47/00	56
E21B 33/035	55
E21B 33/038	54
E21B 43/25	54
C09K 7/02	52
E21B 17/02	45
E21B 23/00	45
E21B 43/013	43
E21B 19/00	42
E21B 17/042	41
E21B 43/26	40

Figura 5-1 – Subgrupos relacionados à base E21B

Das subseções presentes no conjunto de patentes, pode-se perceber da figura 5.1 que algumas receberam maior destaque por estarem presentes em maior número de patentes, que são:

- E21B 43/00: Perfuração do solo. Métodos ou aparelhos para obter óleo, gás, água, materiais solúveis ou fundíveis ou de lama minerais de poços.
- E21B 43/12: Obtenção de fluidos de poços; método ou aparelho para controlar o fluxo do fluido obtido em poços.
- E21B 17/00: Outros equipamentos ou detalhes para perfuração; Equipamento de poços ou manutenção de poços; Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção
- E21B 17/01: Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção. Tubos ascendentes

- E21B 43/01: especialmente adaptados para a obtenção por meio de instalações subaquáticas
- E21B 37/06: Outros equipamentos ou detalhes para perfuração; Equipamentos de poços ou manutenção de poços; Métodos ou aparelhos para limpar furos de sondagem ou poços utilizando meios químicos para impedir ou limitar a deposição de parafinas ou de substâncias similares.
- E21B 47/12: Provas ou ensaios; Levantamento de furos de sondagem ou de. Meios para transmitir sinais de medição do poço para a superfície, por ex., perfilagem durante a perfuração
- E21B 33/13: Outros equipamentos ou detalhes para perfuração; Equipamentos de poços ou manutenção de poços; Vedação ou obturação de furos de sondagem ou de poços. Métodos ou dispositivos para cimentação, para obturação de furos, fendas ou similares.

Além disso, pode-se observar a presença de um subgrupo com frequência relativamente alta no conjunto de patentes com classificação E21B: C09K 7/02. Que de acordo com a CIP:

- C09K 7/02: Química e Metalurgia. Corantes; Tintas; Polidores; Resinas naturais; Adesivos; Composições diversas; Diversas aplicações de substâncias. Substâncias para aplicações diversas, não incluídas em outro local. Composições para perfuração de poços; Fluídos não aquosos contendo compostos orgânicos ou inorgânicos.

A partir dessa descrição, pode-se perceber a presença de além de materiais de diversas formas de perfuração de solo, mas também compostos que auxiliam nesse processo.

Uma observação macro das classificações relacionadas à classificação E21B nas patentes está indicada na figura 5.2.

Classificação	Frequência
E21B	3429
F16L	116
C09K	105
B63B	42
B01D	39
C04B	38
G01N	29
G01V	28
F16K	25
E02D	22
C10L	21
E21C	15
F17D	15
B01F	15
C10G	14
E21D	14
C08F	13
H01M	11
F04B	11
F15B	10
C07C	10

Figura 5-2 – Subclasses relacionadas à base E21B

Duas classificações receberam grande destaque por se relacionarem com E21B, entre elas, C09K, também destaque nas subseções, e a classificação relacionada F16L, sendo:


- **F16L**: Engenharia Mecânica; Iluminação; Aquecimento; Armas; Explosão. Elementos ou unidades de engenharia; Medidas gerais para assegurar e manter o funcionamento efetivo de máquinas ou instalações; Isolamento térmico em geral. Tubos; Juntas ou acessórios para tubos; Suportes para tubos ou cabos; Meios para isolamento térmico em geral.

A partir das observações presentes nas classificações relacionadas às da base de patente e suas subseções, pode-se avaliar qual conjunto de técnicas, materiais e compostos necessários para a composição dos produtos presentes nos documentos de propriedade intelectual e quais são os geralmente mais utilizados para a realização da perfuração de solo para obtenção de fluidos de poços.

Ou seja, métodos de perfuração de solo para obtenção de óleo, gás, água e outros, estão intimamente ligados a materiais da engenharia como tubos, suportes e meios de isolamento térmico.

5.2.2 E21B – Depositantes

Para a classificação E21B, 585 diferente depositantes foram encontrados. A figura 5.3 ilustra os resultados.



Depositantes	Frequência
Halliburton Company	151
Shel Internationale Research Maatschappij B V (NL)	110
Petrobras Distribuidora S.A. (BR/RJ)	86
Schlumberger Surencó S.A. (PA)	81
Baker Hughes Incorporated (US)	53
ABB Offshore Systems	52
Cooper Cameron Corporation (US)	47
FMC Corporation (US)	43
Halliburton Company (US)	33
Institut Francais du Petrole (FR)	29
Alpha Thames LTD (GB)	20
Sofitech N.V (BE)	20
Weatherford/Lam, INC. (US)	18
Exxonmobil Upstream Research Company (US)	13
Kvaerner	13
Otis Engineering Corporation (US)	12
Sandvik AB (SE)	11
M-I L.L.C. (US)	10
Schlumberger Holdings Limited (GB)	10
National Coupling Company INC. (US)	10
National Oilwell Normay AS (NO)	10

Figura 5-3 – Depositantes com a classificação E21B

Do total de depositantes encontrado, foi observado que 75% empresas depositaram apenas uma patente ao longo dos anos, 11% depositaram duas patentes, e assim sucessivamente, conforme mostrado na figura 5.4.

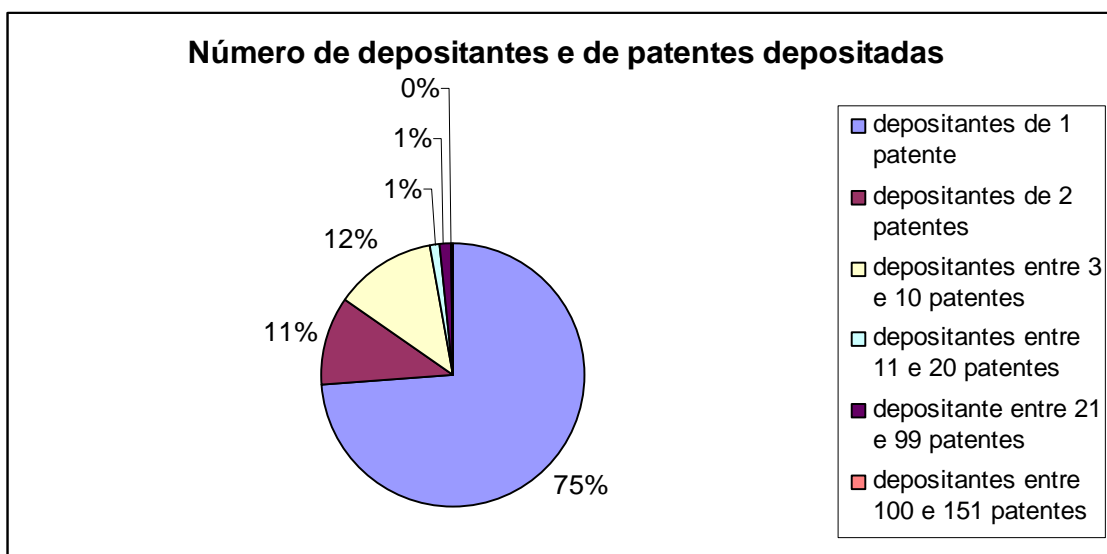


Figura 5-4 – Quantidade de patentes por depositantes de classificação E21B

A empresa que mais tem depositado patentes ao longo dos últimos anos com relação à classificação E21B é a “Halliburton Energy Services, INC (US)” com 151 patentes, sendo a “Shell Internationale Research Maatschappij B.V. (NL)” e a “Petróleo Brasileiro S.A. - Petrobras (BR/RJ)” as maiores depositantes na seqüência, com 110 e 86 patentes, respectivamente.

5.2.3 Petróleo – Classificação

Em busca de mudanças no foco de pesquisa e desenvolvimento de novas tecnologias relacionadas ao componente “Petróleo”, as bases relacionadas ao termo foram analisadas separadamente.

A base que contém as patentes industriais dos anos de 1996 a 2000 possuem um total de 93 diferentes classificações, como pode ser visto na figura 5.5.

Classificação	Frequência
E21B	103
C10G	30
C10L	20
C04B	13
F16L	12
B01J	12
G01N	9
C12P	8
B01D	8
C07C	7
B63B	7
C08L	7
C02F	6
F23Q	6
F17C	6
G01F	6
C08K	5
C10C	5
G01V	5
C09K	5
C06B	5

Figura 5-5 – Subclasses presentes na base “Petróleo” entre os anos de 1996 e 2000

As classificações que mais se destacam na base de dados são E21B, C10G e C10L, definidas de acordo com a CIP:

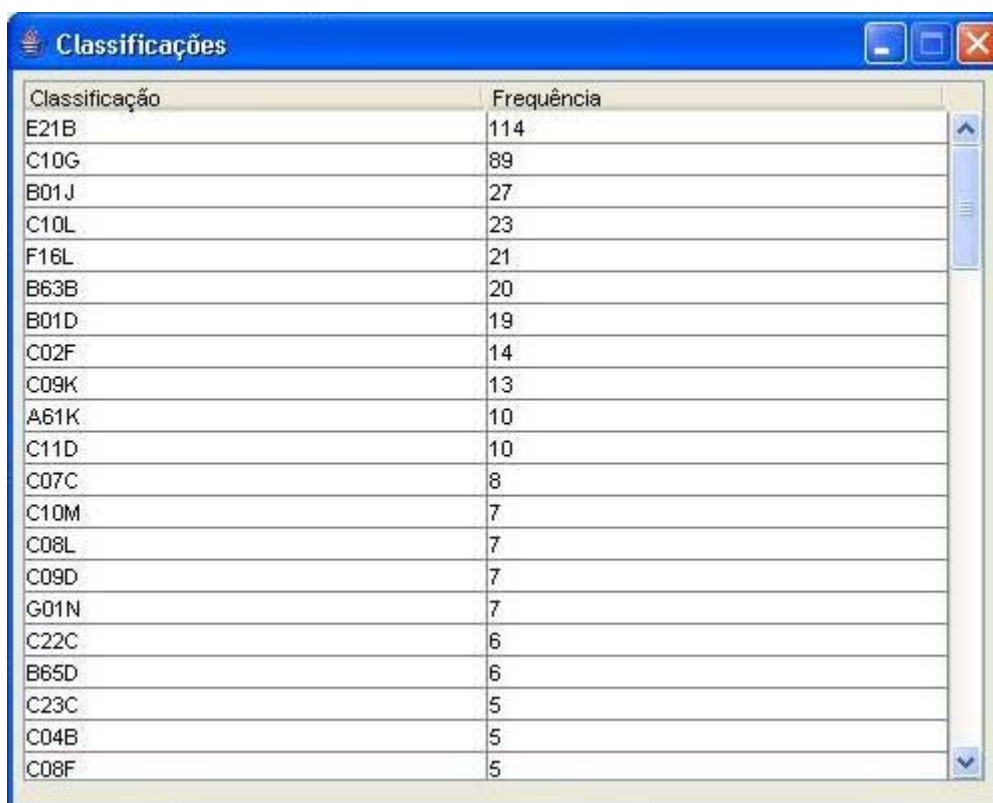
- **E21B:** Construções Fixas. Perfuração do Solo; Mineração; Obtenção de Fluidos de Poços. Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços.
- **C10G:** Indústrias do petróleo; do gás ou do coque; gases técnicos contendo monóxido de carbono; combustíveis; lubrificantes; turfa. Craqueamento de óleos de hidrocarboneto; Produção de misturas líquidas de hidrocarboneto, por ex., hidrogenação destrutiva; oligomerização, polimerização; Recuperação de óleos de hidrocarboneto a partir de xisto etuminoso, arenito oleífero, ou gases; Refinação de misturas constituídas principalmente de hidrocarboneto; Reforma de nafta; Ceras minerais.
- **C10L:** Indústrias do petróleo; do gás ou do coque; gases técnicos contendo monóxido de carbono; combustíveis; lubrificantes; turfa. Combustíveis não

incluídos em outro local; Gás natural; Gás natural de síntese obtido por processos não abrangidos pelas subclasses C10G, K; Gás liquefeito de petróleo; Adição de substâncias a combustíveis ou ao fogo para reduzir fumaça ou depósitos indesejáveis ou para facilitar a remoção de fuligem; Acendedores de fogo.

Sendo que o subgrupo que mais se destaca entre os subgrupos presentes é E21B 37/06:

- E21B 37/06: Outros equipamentos ou detalhes para perfuração; Equipamentos de poços ou manutenção de poços; Métodos ou aparelhos para limpar furos de sondagem ou poços utilizando meios químicos para impedir ou limitar a deposição de parafinas ou de substâncias similares.

Já a base que possui as patentes com termo “Petróleo” depositadas nos anos entre 2001 e 2005 possui 100 diferentes classificações, como se apresenta na figura 5.6.



Classificação	Frequência
E21B	114
C10G	89
B01J	27
C10L	23
F16L	21
B63B	20
B01D	19
C02F	14
C09K	13
A61K	10
C11D	10
C07C	8
C10M	7
C08L	7
C09D	7
G01N	7
C22C	6
B65D	6
C23C	5
C04B	5
C08F	5

Figura 5-6 – Subclasses presentes na base “Petróleo” entre os anos de 2001 e 2005

Do total de classificações encontradas, 3 receberam maior destaque, sendo elas as classificações E21B, C10G e B01J. São definidas segundo a CIP:

- E21B: Construções Fixas. Perfuração do Solo; Mineração; Obtenção de Fluidos de Poços. Perfuração do solo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços.
- C10G: Química e Metalurgia. Indústrias do petróleo; do gás ou do coque; gases técnicos contendo monóxido de carbono; combustíveis; lubrificantes; turfa. Craqueamento de óleos de hidrocarboneto; Produção de misturas líquidas de hidrocarboneto, por ex., hidrogenação destrutiva; oligomerização, polimerização; Recuperação de óleos de hidrocarboneto a partir de xisto etuminoso, arenito oleífero, ou gases; Refinação de misturas constituídas principalmente de hidrocarboneto; Reforma de nafta; Ceras minerais.
- B01J: Operações de Processamento; Transporte. Processos ou aparelhos físicos ou químicos em geral. Processos químicos ou físicos, por ex., catálise, química coloidal; Aparelhos pertinentes aos mesmos.

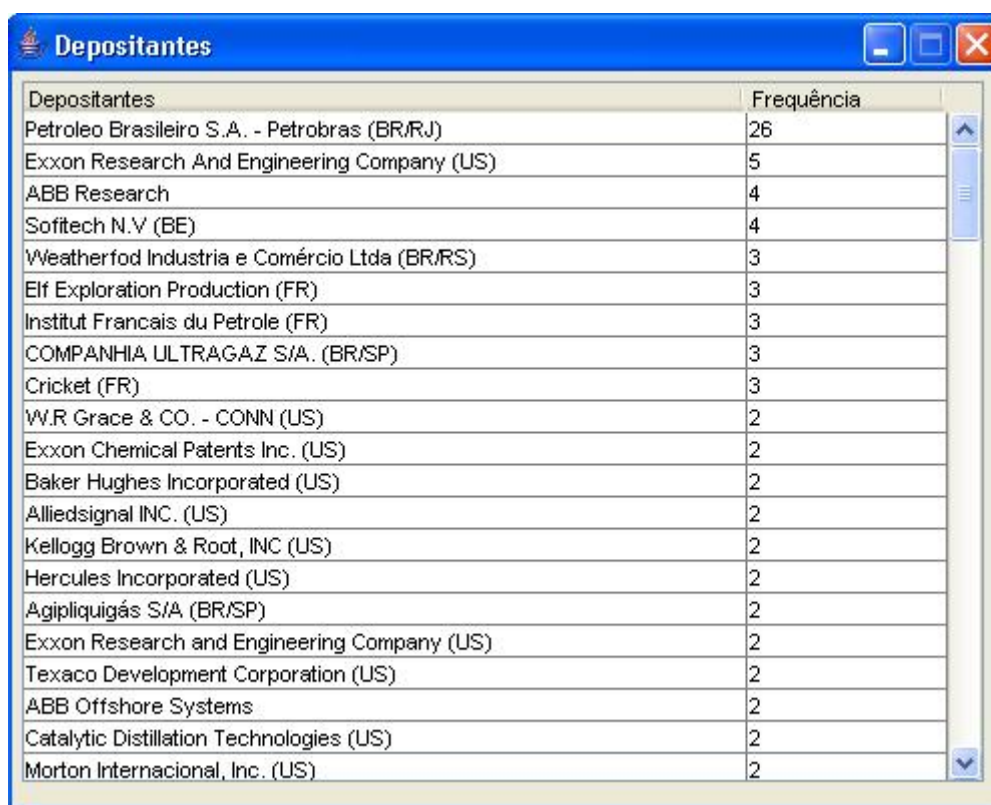
Dentre as subseções, aquelas que receberam maior destaque são C10G 49/00, definida pela CIP como:

- C10G 49/00: Tratamento de óleos de hidrocarboneto, na presença de hidrogênio ou de compostos geradores de hidrogênio

A partir dessas observações, pode-se notar que as classificações que predominam nas patentes relacionadas a petróleo são E21B e C10G, relacionadas a Perfuração de solo e misturas líquidas, respectivamente. Porém, houve uma diferença no foco de pesquisa, ocorridos nos anos determinados pelas bases de estudos, sendo que entre os anos de 1996 e 2000 o estudo foi maior na classificação C10L, combustíveis e gás natural, e entre os anos de 2001 a 2005 o estudo foi maior na classificação B01J, processos químicos ou físicos e aparelhos para os mesmos.

5.2.4 Petróleo – Depositantes

Analisando os depositantes de maior destaque nas bases de dados de patentes selecionadas, pode-se notar que nos anos 1996 a 2000 tiveram 130 diferentes depositantes, incluindo o número de parcerias realizadas. A figura 5.7 apresenta o resultado obtido.



Depositantes	Frequência
Petroleo Brasileiro S.A. - Petrobras (BR/RJ)	26
Exxon Research And Engineering Company (US)	5
ABB Research	4
Sofitech N.V (BE)	4
Weatherfod Industria e Comércio Ltda (BR/RS)	3
Elf Exploration Production (FR)	3
Institut Francais du Petrole (FR)	3
COMPANHIA ULTRAGAZ S/A. (BR/SP)	3
Cricket (FR)	3
W.R Grace & CO. - CONN (US)	2
Exxon Chemical Patents Inc. (US)	2
Baker Hughes Incorporated (US)	2
Alliedsignal INC. (US)	2
Kellogg Brown & Root, INC (US)	2
Hercules Incorporated (US)	2
Agipliquigás S/A (BR/SP)	2
Exxon Research and Engineering Company (US)	2
Texaco Development Corporation (US)	2
ABB Offshore Systems	2
Catalytic Distillation Technologies (US)	2
Morton Internacional, Inc. (US)	2

Figura 5-7 – Depositantes de patentes com o termo “petróleo” nos anos entre 1996 e 2000

O depositante que recebeu maior destaque foi “Petróleo Brasileiro S.A. - Petrobras (BR/RJ)” com 26 patentes depositadas mais uma parceria com “Tecnologia LTDA (BR/RJ)”.

Do total de classificações, pode-se observar que 83% depositantes apenas depositaram 1 patente nos 5 anos analisados, 9% depositaram 2 patentes, e assim por diante, de acordo com a figura 5.8.

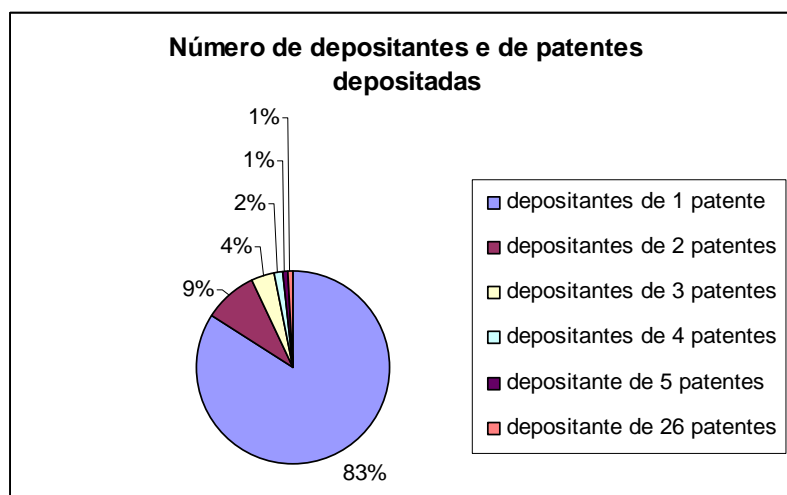


Figura 5-8 – Quantidade de patentes por depositantes nos anos entre 1996 e 2000

Entre os anos de 2001 e 2005, 169 diferentes depositantes foram encontrados, incluindo o número de parcerias realizadas, conforme mostra a figura 5.9.

Depositantes	Frequência
Shell Internationale Research Maatschappij B.V. (NL)	52
Petroleo Brasileiro S.A. - Petrobras (BR/RJ)	42
Rohm And Haas Company (US)	8
Institut Francais du Petrole (FR)	6
Sumitomo Metal Industries LTD. (JP)	5
Chevron U.S.A. Inc. (US)	5
Backer Hughes Incorporated (US)	5
W.R Grace & CO. - CONN (US)	3
Eliberto Eduardo Pinheiro (BR/RN) / Alexandre Azevêdo Borba (BR/BA)	3
José Cássio de Barros Penteado (BR/SP)	3
Atofina (FR)	3
Siderca S.A.I.C. (AR)	3
Halliburton Energy Services, INC (US)	2
Norsk Hydro ASA (NO)	2
Akzo Nobel N.V. (NL)	2
S. C. Johnson & Son, Inc. (US)	2
Brandt Meio Ambiente Indústria Comércio e Serviços Ltda (BR/MG)	2
Statoil Asa (NO)	2
Mineração Curimbaba LTDA (BR/MG)	2
Eni S.P.A (IT)	2
Hercules Incorporated (US)	2

Figura 5-9 – Depositantes de patentes com o termo “petróleo” nos anos entre 2001 e 2005

Pode-se notar que a empresa “Shell Internationale Research Maatschappij B.V. (NL)” depositou, entre os anos 2001 e 2005, 52 patentes, seguindo da empresa “Petróleo Brasileiro S.A. - Petrobras (BR/RJ)”, com 42 patentes depositadas individualmente mais 2 patentes em parceria com as empresas “Albrecht Equipamentos Industriais Ltda. (BR/SC)” e “Akzo Nobel N.V. (NL)”.

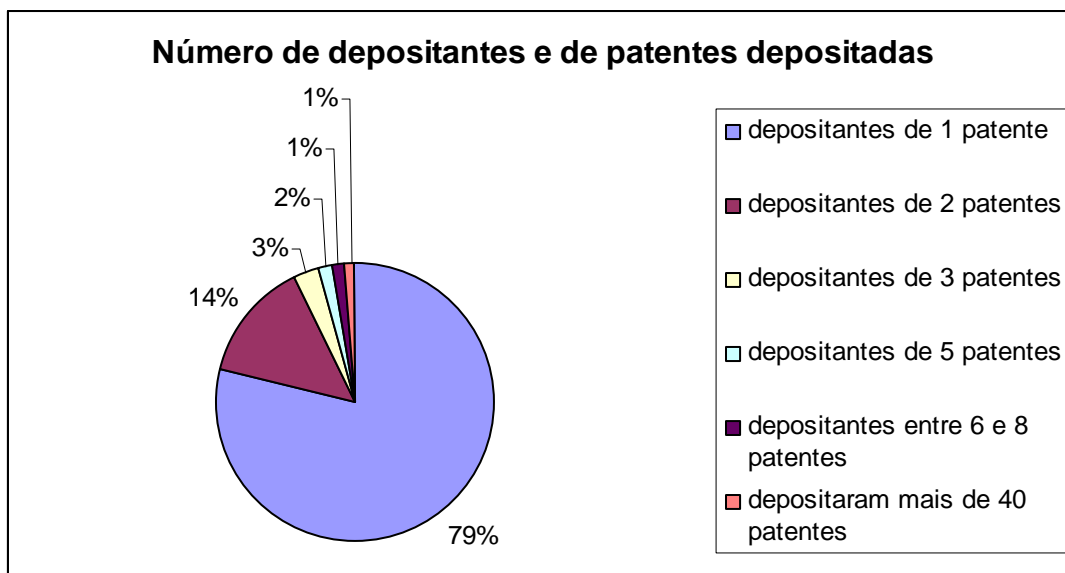


Figura 5-10 – Quantidade de patentes por depositantes nos anos entre 2001 – 2005

Do total dos depositantes, pode-se observar que 79% das empresas depositaram apenas 1 patentes ao longo dos anos, 14% depositou 2 patentes, e assim por diante, conforme mostra a figura 5.10.

5.3 Pré-processamento

A base de dados utilizada para a execução da mineração de textos se reduz aos campos RESUMO nas patentes. Vale destacar que nem todas as patentes possuíam um RESUMO associado, por isso, algumas delas acabaram sendo desconsideradas na análise através da mineração de texto.

Esse campo, tal como alguns outros presentes na patente, possuem alguns erros de digitação, e por esse fato, o trabalho de pré-processamento foi realizado de maneira iterativa, de forma que os termos que resultavam no dicionário e que apareciam de

forma digitalizada incorretamente foram reeditados nos documentos utilizados. Dessa maneira, uma revisão manual dos documentos foi realizada, porém apenas para os termos considerados relevantes para a classificação. Esse processo diminuiu a quantidade total de termos em cerca de 1,4%.

Já o pré-processamento efetuado pela ferramenta para o processamento das patentes constitui da retirada de lista de palavras não relevantes, chamadas de *Stopwords* e da utilização da técnica de *Stemming*.

A lista de *StopWords* é composta, além dos termos que não possuem informação, como conjunção, também alguns termos que não apresentam importância dentro da base de dados sendo processada.

A aplicação do *Stemming* diminui o número total de termos no vetor através do agrupamento de termos semelhantes reduzindo-os aos seus radicais.

Além disso, os termos que possuíam frequência muito alta dentro do conjunto total de documentos e aqueles que possuíam frequência baixa, foram desconsiderados. Por convenção, adotamos que seriam mantidos apenas aqueles termos que apareciam em mais de um documento e menos que todos.

Todo o pré-processamento é realizado em cima do vetor de termos chamado dicionário, e os vetores relativos a cada documento são atualizados a medida que os índices do dicionário, representantes dos termos, são alterados.

Dessa lista de processos a serem aplicados, chegou-se aos resultados que serão apresentados a seguir para cada base de dados.

5.3.1 E21B

A base de dados formada de patentes cuja “CLASSIFICAÇÃO” faz parte do grupo E21B, continha originalmente, 2535 patentes, sendo que apenas 1979 documentos possuem um RESUMO associado.

No total, 301382 termos foram encontrados presentes no campo pré-processado das patentes.

A aplicação da lista de *Stopwords* no vetor reduziu o conjunto total de termos para 13235 termos. A aplicação do algoritmo *Stemming* reduziu esse total para apenas 4726 termos, cerca de 64,3% da base de *Stopwords*.

Com a retirada dos termos mais frequentes e daqueles menos frequentes, a base E21B passou a ter somente 2238 termos no seu vetor de dicionário, reduzindo aproximadamente 52,65% o número de termos, e conseqüentemente, melhorando o tempo de processamento e a qualidade dos resultados.

5.3.2 Petróleo

O pré-processamento da base de dados composta pelas patentes que possuíam o termo “petróleo” no campo RESUMO, foi realizado separadamente para as patentes que foram depositadas entre os anos de 1996 e 2000 e entre os anos de 2001 e 2005.

Como resultado, a primeira base possuía inicialmente um total de 34954 termos sendo esse número reduzido para 4627 com a retirada dos termos presentes na *StopList*. Com a aplicação do algoritmo de *Stemming*, as palavras foram reduzidas aos seus radicais, diminuindo 43% do total de termos presentes no vetor Dicionário da base de dados, compondo assim apenas 1965 termos. A redução dos termos mais frequentes e os menos frequentes resultou em um total de 922 termos considerados relevantes, lembrando que o termo “petróleo”, presente em todos os documentos a serem clustrizados, foi retirado do conjunto de termos por não possuir um alto grau discriminatório dos clusters. Além disso, aqueles termos que estavam presentes em apenas um único documento também foram desconsiderados do processo.

A segunda base de patentes, compreendendo os anos de 2001 a 2005, possuíam originalmente 47457 termos. A retirada dos termos irrelevantes, presentes na lista de *StopWords*, reduziu esse total para 5305 termos e em seguida o algoritmo de *Stemming* reduziu para 2206 radicais. Os termos mais frequentes e menos frequentes no conjunto total de documentos também foram retirados, igualmente ao processo realizado na primeira base, e reduziu para 1122 o número total de termos.

5.4 Clusterização

Pelo fato de não possuir uma classificação prévia dos documentos clusterizados, não é possível obter uma relação do grau de precisão do sistema. No entanto, a partir da análise dos resultados de cada cluster, levando em consideração alguns fatores como o conjunto de documentos agrupados no mesmo cluster, as palavras-chave mais significativas de cada conjunto, as classificações e depositantes predominantes dos resultados dos clusters, algumas conclusões podem ser retiradas e utilizadas como auxílio à IC.

O número de clusters foi escolhido baseado em pré-análises de outras quantidades de clusters. Dessas pré-análises, determinou-se que pela variedade de tópicos existentes dentro do conjunto de patentes, a quantidade ideal de clusters tendia ao maior número possível, o que quer dizer que tendia a criar vários clusters de pequenas quantidades de documentos, com assunto específicos, como por exemplo, clusters separando dispositivos das técnicas de perfuração através de pressão de fluidos aquosos.

Para evitar que assuntos similares fossem separados por pequenos detalhes, o número de clusters foi testado visando também, não perder as informações importantes presentes no conjunto total. Porém, apenas um sistema matemático elaborado poderia determinar com maior precisão o número ideal de clusters.

A seguir, são apresentados os resultados da clusterização realizada pela ferramenta implementada para a dissertação contendo: uma análise de cada cluster criado, uma visão das classificações e dos depositantes predominantes em cada conjunto. São apresentados também os resultados da aplicação das bases em duas ferramentas distintas: o módulo IDC desenvolvido pela Temis e o módulo de mineração de textos do software Statistica de Statsoft.

5.4.1 E21B – Clustering

As patentes de E21B, no total 1979 com o campo RESUMO, foram clusterizadas pela ferramenta da dissertação tendo como base o número de clusters igual a dez.

A figura 5.11 apresenta a tela de saída da clusterização desses documentos, apresentando os radicais considerados mais importantes para cada um dos clusters criados.

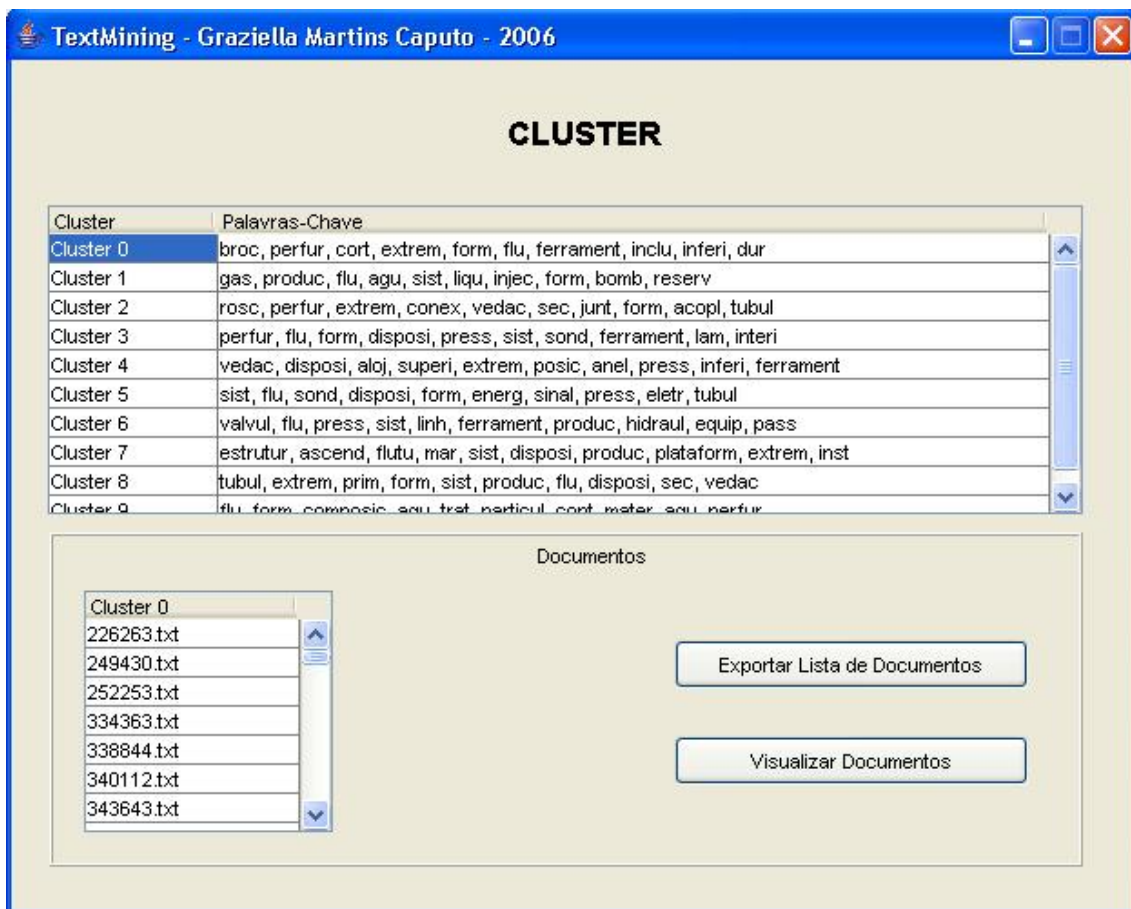


Figura 5-11 – Resultado do clustering de E21B

A tabela 5.2 mostra quantos documentos foram selecionados em cada cluster e as palavras-chaves, que foram consideradas mais representativas do conjunto total de termos presente em cada cluster.

Cluster	Num. de Patentes	Nome	Palavras-chave
1	106	Broca	broca; perfurador; cortador; extremidade; formação; fluido; ferramenta; incluir; inferior; durante.
2	216	Gás	gás; produção; fluidos; água; sistema; líquido; injeção; formação; bombeamento; reservatório.
3	48	Rosca	rosca; perfuração; extremidade; conexão; vedação; seção; junta; formação; acoplamento; tubulares.
4	228	Perfurador	perfurador; fluido; formação; disposição; pressão; sistema; sonda; ferramenta; lama; interior.
5	362	Vedação	vedação; dispositivo; alojamento; superior; extremidade; posição; anel; pressão; inferior; ferramenta.
6	222	Eletricidade	sistema; fluido; sondagem; dispositivo; formação; energia; sinal; pressão; eletricidade; tubulação.
7	214	Válvula	válvula; fluido, pressão; sistema; linha; ferramenta; produção; hidráulica; equipamento.
8	178	Plataforma	estrutura; ascendente; flutuante; marítimo; sistema; disposição; produção; plataforma; extremidade; instrumento.
9	173	Tubulação	tubulação; extremidade; primeiro; formação; sistema; produção; fluido; disposição; segundo; vedação.
10	232	Fluidos	fluido; formação; composição; aquoso; tratamento; partícula; conteúdo; material; água; perfuração.

Tabela 5-2 – Documentos e palavras-chave dos clusters de E21B

Um termo foi selecionado para representar o cluster, como pode ser visto na tabela 5.2. Dessa forma, quando um cluster for citado ao longo do capítulo, o mesmo será referenciado pelo termo escolhido. Esse nome foi escolhido com base nos critérios: principal palavra-chave ou palavra-chave mais representativa ou conjunto de palavras mais representativas.

Partindo desses resultados apresentados, pode-se analisar o significado dessas palavras-chave, levando em consideração o texto presente nos documentos clusterizados.

A seguir, é apresentado uma análise de cada um dos clusters encontrados, juntamente de alguns títulos de patentes escolhidos aleatoriamente dentro dos clusters, como forma de ilustração do conteúdo do mesmo.

As classificações e os depositantes de maiores destaque dentro de cada cluster também foram analisados para efeito de melhor entendimento do resultado da clusterização.

Cluster 1 - Broca

Descrição do cluster:

Broca; perfurador; cortador; extremidade; formação; fluido; ferramenta; incluir; inferior; durante.

Títulos pertencentes ao cluster:

- Título: Dispositivo de guiagem em operações de perfuração de rocha, luva de guiagem e haste de broca
- Título: Cortadores giratórios para brocas de rocha
- Título: Acoplador para circulação contínua de fluidos de perfuração através de um fio de broca, método de adição ou remoção dos tubulares para e de um fio de broca, aparelho para perfuração na terra
- Título: Conepunho acoplável para reaproveitamento das hastes de brocas integrais.
- Título: Broca de diamantes fungíveis para sondas de arrasto

Classificações de Destaque:

- E21B 10/46: Brocas de Perfuração; caracterizadas pelas peças resistentes ao desgaste, por ex., com inserções de diamantes.
- E21D: Poços, Túneis; Galerias; Câmaras subterrâneas grandes

Depositantes de Destaque:

- “Down Hole Technologies PTY Ltd. (AU)”
- “Halliburton Energy Services, INC (US)”

Conclusão da análise:

Broca para perfuração e corte (com formação de hastes inferiores e superiores e com escape de fluidos).

Cluster 2 - Gás

Descrição do cluster:

Gás; produção; fluidos; água; sistema; líquido; injeção; formação; bombeamento; reservatório.

Títulos pertencentes ao cluster:

- Título: Processo para recuperar hidrocarbonetos a partir de uma formação subterrânea por injeção de vapor contendo aditivo.
- Título: Dispositivo de injeção d'água sob pressão de crescente
- Título: Processo para a recuperação de hidrocarbonetos de uma formação subterrânea, na ausência de vapor de água, usando-se um gás não-condensável
- Título: Processo de bombeamento de mistura difásica líquido-gás num poço de extração e dispositivo de aplicação do processo

Classificações de Destaque:

- E21B 43/12: Obtenção de fluidos de poços; método ou aparelho para controlar o fluxo do fluido obtido em poços.
- E21B 43/34: aparelhos separadores. Disposições para separar materiais produzidos pelo poço
- B01D: Processos ou aparelhos físicos ou químicos em geral. Separação.

Depositantes de Destaque:

- “Petroleo Brasileiro S.A. - Petrobras (BR/RJ)“
- “Shell Internationale Research Maatschappij B.V. (NL)“
- “Alpha Thames Ltd (GB)“.

Conclusão da análise:

Sistema de bombeamento para injeção de água e fluidos quentes (vapor e gases) para perfuração, ou injeção de fluidos em poços de reservatório de água e/ou gás.

Cluster 3 - Rosca

Descrição do cluster:

Rosca; perfuração; extremidade; conexão; vedação; seção; junta; formação; acoplamento; tubulares.

Títulos pertencentes ao cluster:

- Título: Processo e dispositivo para realização do aparafusamento de uma junta roscada para tubos de aço
- Título: Disposição em rosca para hastes perfuratrizes
- Título: Conexão roscada e conduto tubular
- Título: Acoplamento de rosca para hastes de perfuração a percussão

Classificações de destaque:

- E21B 17/042: Hastes ou tubos de perfuração; ferramentas flexíveis de perfuração; hastes quadradas (“kellies”); comandos; hastes de sucção; tubulação de revestimento; tubos de produção engates; juntas entre haste e broca ou entre haste e haste roscados.
- F16L: Elementos ou unidades de engenharia; Medidas gerais para assegurar e manter o funcionamento efetivo de máquinas ou instalações; Isolamento térmico em geral. Tubos; Juntas ou acessórios para tubos; Suportes para tubos ou cabos; Meios para isolamento térmico em geral.

Depositantes de Destaque:

- Sandvik AB (SE)
- Petroleo Brasileiro S.A. - Petrobras (BR/RJ)

Conclusão da análise:

Acoplamento tubulares de roscas para perfuração e roscas macho e fêmea para garantir vedação.

Cluster 4 - Perfurador

Descrição do cluster:

Perfurador; fluido; formação; disposição; pressão; sistema; sonda; ferramenta; lama; interior.

Títulos pertencentes ao cluster:

- Título: Dispositivo e método para tratamento de lama de perfuração recebida de um poço, sistema e dispositivo para perfurar um poço
- Título: Processo para a avaliação de fluidos de perfuração baseado na análise termo-gravimétrica (tga)
- Título: Processo para otimizar a remoção de rebocos formados nas paredes de poços horizontais por fluidos de perfuração drill-in
- Título: Método e aparelho para variação de densidade de fluidos de perfuração em aplicações de perfuração de óleo em águas profundas

Classificações de destaque:

- C09K: Corantes; Tintas; Polidores; Resinas naturais; Adesivos; Composições diversas; Diversas aplicações de substâncias. Substâncias para aplicações diversas, não incluídas em outro local.
- E21B 21/00: Métodos ou aparelhos para lavar furos de sondagem, por ex., pela utilização do ar de exaustão do motor

Depositantes de Destaque:

- “Shell Internationale Research Maatschappij B.V. (NL)”
- “Halliburton Energy Services, INC (US)”, sendo uma em parceria com “Commonwealth Scientific And Industrial Research Organisation (AU)”

Conclusão da análise:

Fluidos de perfuração (por pressão) e ferramenta para tratamento de lama formada por perfuração em interiores.

Cluster 5 - Vedação

Descrição do cluster:

Vedação; dispositivo; alojamento; superior; extremidade; posição; anel; pressão; inferior; ferramenta.

Títulos pertencentes ao cluster:

- Título: Conexão giratória com vedação metálica
- Título: Aparelho de ferramenta de sondagem responsiva à pressão no anel e processo para sua operação
- Título: Conjunto de obturação para a vedação entre um mandril interno e um revestimento de poço, e aparelhagem para o assentamento de uma guia de sonda e para a alteração de perfuração através de um revestimento de poço com uma única viagem da coluna de perfuração.
- Título: Dispositivos para vedação de vazamentos em coluna de produção de petróleo ou similares

Classificações de destaque:

- F16L: Elementos ou unidades de engenharia; Medidas gerais para assegurar e manter o funcionamento efetivo de máquinas ou instalações; Isolamento térmico em geral. Tubos; Juntas ou acessórios para tubos; Suportes para tubos ou cabos; Meios para isolamento térmico em geral.
- E21B 33/038: Vedação ou obturação de furos de sondagem ou de poços. Cabeças de poços; Sua fixação. Conectores utilizados sobre cabeças de poços, por ex., para conectar equipamentos preventivos de explosões e tubos ascendentes

Depositantes de Destaque:

- Cooper Cameron Corporation (US)
- Halliburton Energy Services, INC (US)

Conclusão da análise:

Dispositivos de vedação, obturação e sondagem como anéis de borracha e metal.

Cluster 6 - Eletricidade

Descrição do cluster:

Sistema; fluido; sondagem; dispositivo; formação; energia; sinal; pressão; eletricidade; tubulação.

Títulos pertencentes ao cluster:

- Título: sistema para produzir fluidos a partir de duas zonas diferentes no interior de um furo de sondagem, para uso em um ambiente de furo de sondagem, para administração de fluidos com respeito a uma pluralidade de zonas
- Título: processo e sistema para determinação do tamanho de um material penetrado no furo de sondagem e para determinação do tamanho de um furo de sondagem
- Título: dispositivo de furo de sondagem para controlar o fluxo de fluido através de um poço de produção de fluido de hidrocarboneto
- Título: processo de formação de um furo de sondagem e aparelho para realização do processo

Classificações de destaque:

- E21B 47/12: Provas ou ensaios; Levantamento de furos de sondagem ou de. Meios para transmitir sinais de medição do poço para a superfície, por ex., perfilagem durante a perfuração.
- G01V: Física. Medição; Aferição. Meteorologia.

Depositantes de Destaque:

- Shell Internationale Research Maatschappij B.V. (NL)
- Schlumberger Sureco S.A. (PA)

Conclusão da análise:

Dispositivos de furos de sondagem para administração de fluidos (injeção, controle de fluxo e retenção).

Sonda de transmissão, transferência, medição e produção de energias e detecção de sinais de energia (por exemplo, elétrica, eletromagnética e outras).

Cluster 7 - Válvula

Descrição do cluster:

Válvula; fluido, pressão; sistema; linha; ferramenta; produção; hidráulica; equipamento.

Títulos pertencentes ao cluster:

- Título: Válvula para controle do fluxo de um conduto de transmissão de fluido em um poço subterrâneo
- Título: Conjunto de sede de teste rápido, válvula e método para simultaneamente testar no campo a integridade de pressão
- Título: Válvula de pressão aperfeiçoada
- Título: Acionador rápido de válvula e ferramenta que comporta uma válvula
- Título: Suspensor de tubos com válvula esfera no orifício anular

Classificações de destaque:

- E21B 43/06: Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços
- F16K: Engenharia Mecânica; Iluminação; Aquecimento; Armas; Explosão. Elementos ou unidades de engenharia; medidas gerais para assegurar e manter o funcionamento efetivo de máquinas ou instalações; isolamento térmico em geral. Válvulas; torneiras; registros; bóias de acionamento; dispositivos para ventilar ou arejar.

Depositantes de Destaque:

- Petróleo Brasileiro S.A. - Petrobras (BR/RJ)
- Halliburton Energy Services, INC (US)

Conclusão da análise:

Válvula de pressurização de fluidos.

Cluster 8 - Plataforma

Descrição do cluster:

Estrutura; ascendente; flutuante; marítimo; sistema; disposição; produção; plataforma; extremidade; instrumento.

Títulos pertencentes ao cluster:

- Título: Sistema com uma estrutura de guia para tubos ascendentes de produção de petróleo; estrutura de guia para tubos ascendentes; elementos de flutuação de tubos ascendentes e uma plataforma de produção semi-submersível
- Título: Estrutura de torre para plataforma submarina
- Título: Estrutura de torre e construção para sustentar plataforma marítima
- Título: Arranjo em uma unidade de flutuação de suporte, plataforma e processo para fabricar uma plataforma

Classificações de destaque:

- E21B 17/01: Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção. Tubos ascendentes
- B63b: operações de processamento; transporte. Navios ou outras embarcações; Equipamento correlato.

Depositantes de Destaque:

- Petroleo Brasileiro S.A. - Petrobras (BR/RJ)
- Institut Francais du Petrole (FR)

Conclusão da análise:

Estrutura flutuante para plataformas marítimas.

Cluster 9 – Tubulação

Descrição do cluster:

Tubulação; extremidade; primeiro; formação; sistema; produção; fluido; disposição; segundo; vedação.

Títulos pertencentes ao cluster:

- Título: Tubulação para produção de petróleo/gás e processos para coleta de petróleo/gás produzido a partir de pelo menos um poço de produção, para teste da produção de um poço de petróleo/gás e para limpeza por raspagem de um primeiro e um segundo oleodutos de produção de petróleo/gás
- Título: dispositivo de acoplamento para permitir o acoplamento estanque a fluido de extremidades opostas de um primeiro e um segundo elementos substancialmente tubulares, e, processo para acoplar entre si as extremidades opostas de um primeiro e um segundo elementos substancialmente tubulares
- Título: método para bloqueio monitorado por sensor de explosividade para manutenção em tubulações de transporte de fluidos inflamáveis

Classificações de destaque:

- E21B 43/10: Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços. Colocação ou fixação de tubos de revestimento, peneiras (ou filtros) ou tubos auxiliares de revestimento em poços.
- F16L: Elementos ou unidades de engenharia; Medidas gerais para assegurar e manter o funcionamento efetivo de máquinas ou instalações; Isolamento térmico em geral. Tubos; Juntas ou acessórios para tubos; Suportes para tubos ou cabos; Meios para isolamento térmico em geral.

Depositantes de Destaque:

- Shell Internationale Research Maatschappij B.V. (NL)
- Halliburton Energy Services, INC (US)

Conclusão da análise:

Tubulações para transporte e vedação de fluidos.

Cluster 10 – Fluidos

Descrição do cluster:

Fluido; formação; composição; aquoso; tratamento; partícula; conteúdo; material; água; perfuração.

Títulos pertencentes ao cluster:

- Título: Aditivo para adição a um fluido de tratamento de poços, composição de tratamento de poços em microemulsão externa de ácido e/ou água, fluido de fraturamento de espuma e processo para o tratamento de uma formação subterrânea
- Título: fluido viscoso aplicável para tratamento de formações subterrâneas
- Título: Processo para preparar um copolímero adequado para uso como aditivo para fluido de perfuração ou similar; preparar um fluido de perfuração; perfurar um furo de sonda na terra.
- Título: Processo de acabamento ou manutenção de um poço e fluido de perfuração de poço não aquoso isento desólidos

Classificações de destaque:

- C09K: Corantes; Tintas; Polidores; Resinas naturais; Adesivos; Composições diversas; Diversas aplicações de substâncias. Substâncias para aplicações diversas, não incluídas em outro local.
- E21B 37/06: Métodos ou aparelhos para limpar furos de sondagem ou poços utilizando meios químicos para impedir ou limitar a deposição de parafinas ou de substâncias similares.

Depositantes de Destaque:

- Sofitech N.V (BE)
- “Halliburton Energy Services, INC (US)” sendo uma em parceria com “Pinnacle Technologies, INC. (US)” e outra com o “Institut Francais Du Petrole (FR)”

Conclusão da análise:

Fluidos de composição aquosa para tratamento de formações subterrâneas e para perfuração.

Dos resultados apresentados pela ferramenta, pode-se notar que os clusters possuem significados bem definidos o que os diferenciam uns dos outros.

5.4.1.1 E21B – Temis

Tal como foi realizado no teste realizado com a ferramenta anterior, o módulo IDC foi executado criando 10 clusters a partir das 1979 patentes que possuíam o campo RESUMO.

A Tabela 5.3 mostra o número de documentos agrupados em cada clusters, tal como o termo escolhido para representa-lo e o conjunto de palavras mais representativas. O termo representativo foi escolhido com base na palavra que melhor representa cada conjunto, seguido do termo “_Temis”, para diferenciar os clusters semelhantes aos encontrados anteriormente.

Cluster	Núm. de Patentes	Nome	Palavras-chave
1	266	Rosca_Temis	elemento; tubular; vedação; membro; rosca; corpo; anel; interno; superfície; externo.
2	243	Válvula_Temis	válvula; orifício; controle; árvore; pressão; passagem; fluxo; linha; hidráulico; alojamento
3	232	Perfuração_Temis	furo; sondagem; perfuração; ferramenta; poço; fluido; formação; método; coluna; sistema
4	223	Flutuante_Temis	tubo; ascendente; revestimento; flutuante; perfuração; coluna; dispositivo; extremidade; suporte; poço
5	223	Gás_Temis	gás; petróleo; água; produção; bomba; reservatório; óleo; injeção; poço; fase
6	214	Fluidos_Temis	composição; subterrâneo; formação; agente; tratamento; aquoso; ácido; fluido; cimento; água
7	193	Broca_Temis	broca; haste; perfuração; corte; eixo; elemento; rotativo; corpo; extremidade; cortante
8	156	Eletricidade_Temis	tubulação; energia; sinal; elétrico; dado; sistema; poço; cabo; sensor; dispositivo
9	151	Plataforma_Temis	estrutura; conduto; plataforma; mar; submarino; navio; parte; linha; guia; suporte
10	78	Tela_Temis	tela; lama; cascalho; filtro; densidade; intervalo; perfuração; areia; partícula; poço

Tabela 5-3 – Resultado da ferramenta Temis para a base E21B

Comparando os resultados obtidos por Temis e os resultados citados anteriormente pode-se destacar alguns pontos de semelhanças entre os cluster criados.

A figura 5.12 apresenta a proporção de documentos coincidentes em cada cluster. Como por exemplo, o cluster “Broca” possui 0% de documentos semelhantes ao cluster “Rosca_Temis”, 0% do cluster “Válvula_Temis”, 3,7% do cluster “Perfurador_Temis”, e assim sucessivamente, tal como o cluster “Gás” possui 0,4% de documentos semelhantes com o cluster “Rosca_Temis”, etc.

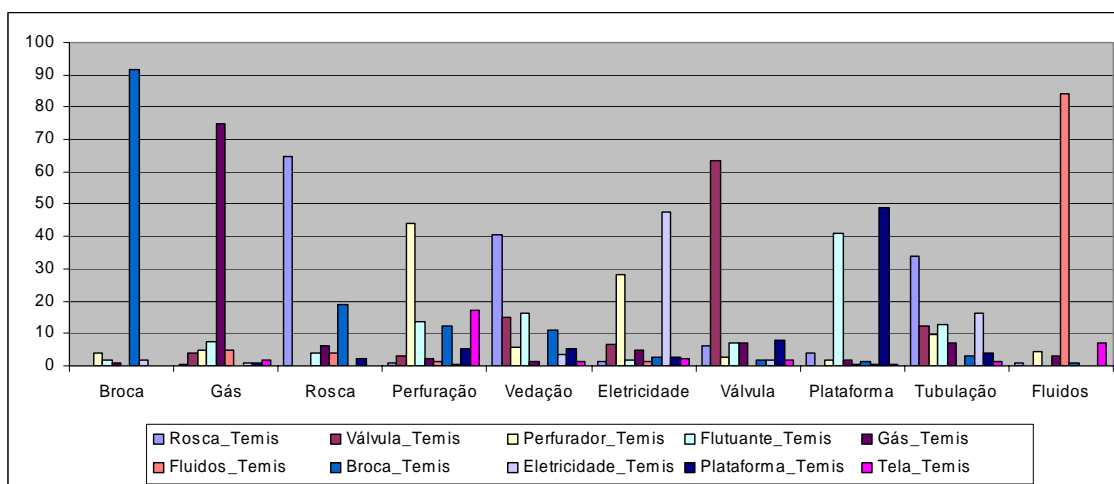


Figura 5-12 – Gráfico de comparação da ferramenta implementada e Temis para E21B

Pode-se observar que:

- O cluster “Broca” encontrado pela ferramenta de mineração de patentes, tem 92% de seus documentos presentes no cluster “Broca_Temis”.
- O cluster “Gás” possui 75% de documentos coincidentes com o cluster “Gás_Temis”.
- O cluster “Rosca” possui 64% de documentos coincidentes com o cluster “Rosca_Temis”.
- O cluster “Perfuração” possui 44% dos seus documentos presentes no cluster “Perfuração_Temis”.
- O cluster “Vedação” possui 40% dos seus documentos presentes no cluster “Rosca_Temis”.

- O cluster “Eletricidade” coincide 48% de seus documentos com o cluster “Eletricidade_Temis”, e em menor proporção, 28% com o cluster “Perfurador_Temis”, indicando o uso de eletricidade para métodos e sistemas de perfuração.
- O cluster “Válvula” coincide 63% dos seus documentos com o cluster “Válvula_Temis”.
- O cluster “Plataforma” coincide 49% de seus documentos com o cluster “Plataforma_Temis”, e 41% com o cluster “Flutuante_Temis”.
- O Cluster “Tubulação” possui maior semelhança com o cluster “Rosca_Temis”. Como o cluster “Rosca_Temis” foi aquele que possui maior número de documentos agrupados, encontra-se distribuído em três clusters diferentes da ferramenta.
- O cluster “Fluidos” coincide 84% com o cluster “Fluidos_Temis”..

O cluster “Tela_Temis” não coincide fortemente com nenhum dos clusters encontrados pela ferramenta.

Apesar de alguns termos encontrados em ambas as clusterizações serem diferentes, o significado dos clusters podem ser considerados semelhantes para aqueles que apresentaram proporção de similaridade alta. Essa diferença se deve ao fato de cada ferramenta possuir uma diferente técnica computacional de busca por palavras-chave.

5.4.1.2 E21B – Statistica

A ferramenta Statistica encontrou um total de 863 termos nos 1979 documentos, utilizando a mesma lista de *StopWords* da ferramenta implementada, e a opção “*Stemming language*” configurada para o idioma português, presente na própria ferramenta Statistica.

A função de frequência inversa dos documentos foi utilizada como cálculo estatístico das ocorrências dos termos.

Cluster	Núm. de Patentes	Nome	Palavras-chave
1	282	Broca_Rosca_Statistica	broca; extremidade; haste; perfurador; elemento; tubo; eixo; rotativo; rosca; interno.
2	193	Plataforma_Statistica	ascendente; mar; tubulação; submarino; estrutura; instalação; plataforma; flutuante; linha; leito.
3	530	Perfuração__Válvula_Statistica	perfuração; sistema; dispositivo; poço; controle; invenção; submarino; válvula; fluido; processo.
4	187	Revestimento_Statistica	furo; revestimento; aparelho; poço; tubulação; método; sondagem; primeiro; perfuração; coluna.
5	164	Gás_Statistica	gás; separador; liquido; produção; reservatório; água; injeção; óleo; petróleo; hidrocarboneto.
6	197	Fluidos_Statistica	composição; ácido; agente; formação; subterrâneo; aquoso; água; tratamento; aditivo; composto.
7	144	Sinal_Statistica	dado; determinar; medida; medição; formação; sinal; sensor; sinal; furo; parâmetro.
8	193	Vedação_Statistica	vedação; válvula; alojamento; anel; corpo; orifício; posição; extremidade; anular; passagem.
9	27	Tela_Statistica	tela; cascalho; areia; intervalo; tubo; fluxo; enchimento; furo; membro; poço.
10	62	Eletricidade_Statistica	energia; elétrica; eletricidade; condutor; cabo; tubulação; petróleo; corrente; comunicação; controle.

Tabela 5-4 - Resultado da ferramenta Statistica para a base E21B

A figura representa o gráfico de documentos coincidentes entre os clusters encontrados pela ferramenta e os do Statistica.

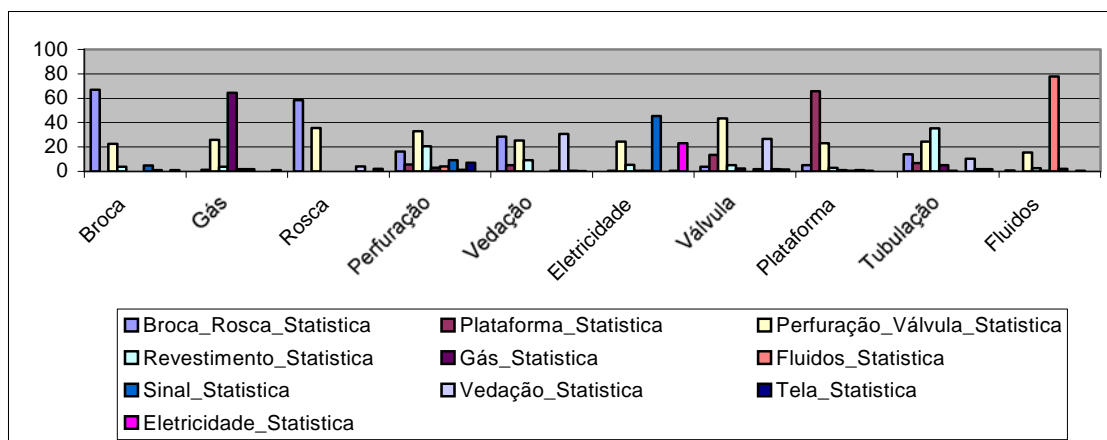


Figura 5-13 – Gráfico de comparação da ferramenta implementada e Statistica para E21B

Pode-se observar:

- O cluster “Broca” possui 67% de seus documentos incluídos no “Broca_Rosca_Statistica”;
- O cluster “Gás” possui 65% de seus documentos incluídos no cluster “Gás_Statistica”;
- O cluster “Rosca” possui 65% de seus documentos incluídos no cluster “Broca_Rosca_Statistica”. O fato do cluster “Broca_Rosca_Statistica” possuir um número alto de documentos, acumula documentos de dois clusters distintos dos encontrados pela primeira ferramenta testada;
- O cluster “Perfuração” possui 33% de seus documentos presentes no cluster “Perfuração_Válvula_Statistica”;
- O cluster “Vedação” possui 30% de seus documentos presentes no cluster “Vedação_Statistica”;
- O cluster “Eletricidade” possui 45% de seus documentos presentes no cluster “Sinal_Statistica”, 24% no cluster “Perfuração_Válvula_Statistica” e apenas 23% no cluster “Eletricidade_Statistica”;

- O cluster “Válvula” possui 43% de seus documentos presentes no cluster “Perfuração_Válvula_Statistica;”
- O cluster “Plataforma” possui 66% de seus documentos presentes no cluster “Plataforma_Statistica”;
- O cluster “Tubulação” possui 35% de seus documentos presentes no cluster “Revestimento_Statistica”;
- O cluster “Fluidos” possui 78% de seus documentos presentes no cluster “Fluidos_Statistica”.

O cluster “Tela_Statistica” não coincide fortemente com nenhum dos clusters comparados, porém apresenta grande semelhança com o cluster “Tela_Temis”, encontrado pelo módulo IDC.

5.4.2 Base Petróleo

Os textos contidos nos campos RESUMO presente nos documentos das patentes que possuem o termo “Petróleo” foram clusterizados em 7 diferentes clusters através da ferramenta desenvolvida para a presente dissertação.

A partir dos resultados obtidos por ambas as bases, que se referem às patentes depositadas entre 1996 e 2000 e às patentes depositadas entre 2001 e 2005, é possível realizar comparações e análises para identificar as tendências e mudanças ocorridas no desenvolvimento industrial e tecnológicos das áreas relacionadas ao tema.

Os resultados de cada uma das duas bases são ainda comparados com aqueles encontrados pelas ferramentas Temis e Statistica, utilizando o mesmo número de clusters.

5.4.2.1 *Clustering de Patentes de 1996 a 2000*

O resultado da clusterização da base de dados que compreende aos anos de 1996 a 2000 pode ser visualizado na Figura 5-14. Na tela de resultados da ferramenta de clusterização de patentes estão disponíveis os nomes dos documentos de patentes agrupados e os radicais das palavras consideradas mais importantes em cada cluster.

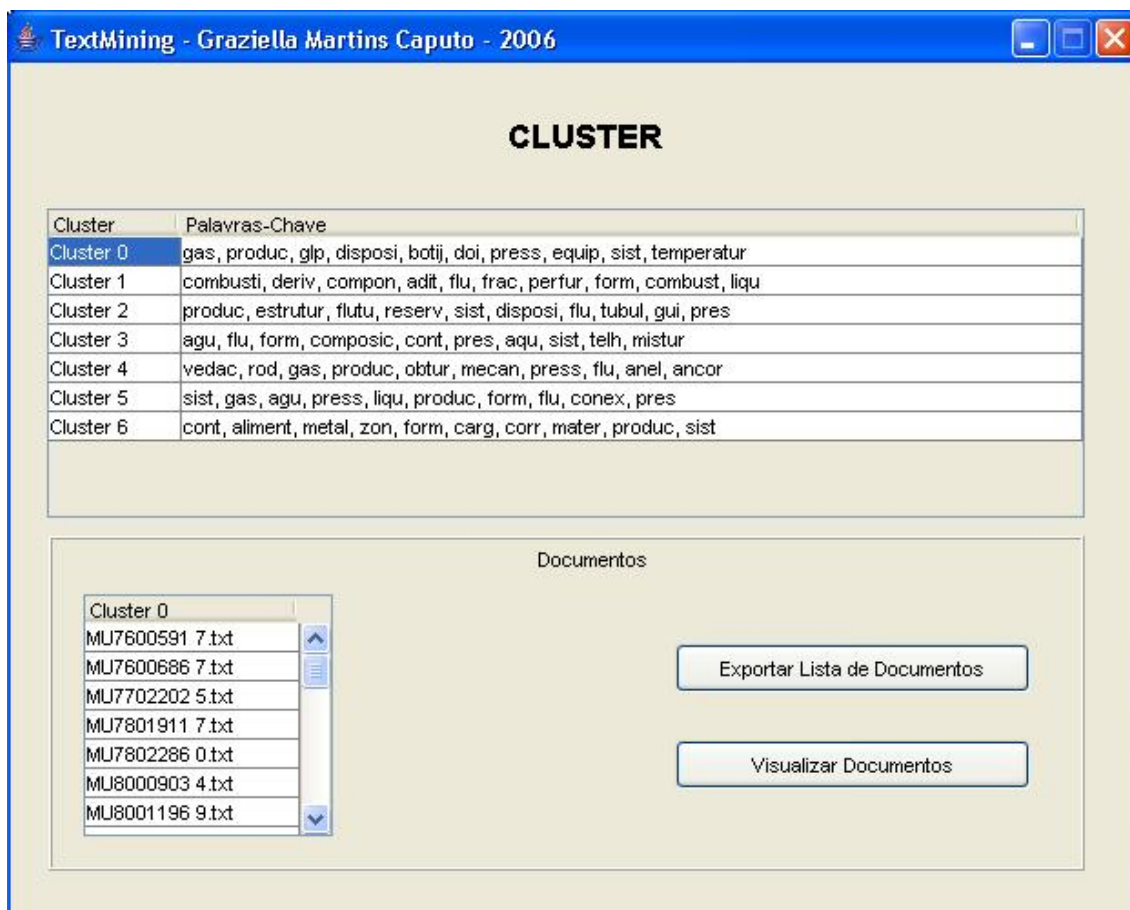


Figura 5-14 – Clusterização da base de dados com o termo “Petróleo” de 1996 a 2000

O número de patentes encontradas em cada cluster e as palavras-chave de cada um deles pode ser visualizado na Tabela 5-5. A cada cluster, foi dado um nome que melhor represente o assunto principal contido nas patentes agrupadas. Esse nome foi dado seguindo algum dos critérios: termo mais significativo ou conjunto de termos que melhor definem o significado do cluster.

Cluster	Núm. de Patentes	Nome	Palavras-chave
1	30	GLP	Gás; produção; glp; disposição; botijão; dois; pressão; equipamento; sistema; temperatura.
2	40	Aditivo	Combustível; derivar; componente; aditivo; fluido; fração; perfuração; formação; combustão; líquido.
3	32	Flutuante	Produção; estrutura; flutuante; reservatório; sistema; disposição; fluido; tubulação; guia; pressão.
4	35	Mistura	Água; fluido; formação; composição; conteúdo; pressão; aquoso; sistema; telha; mistura.
5	20	Vedação	Vedação; roda; gás; produção; obturação; mecanismo; pressão; fluido; anel; ancoragem.
6	44	Gás	Sistema; gás; água; pressão; líquido; produção; formação; fluido; conexão; presente.
7	25	Alimentação	Conteúdo; alimentação; metal; zona; formação; carga; corrente; material; produção; sistema.

Tabela 5-5 – Resultado da ferramenta desenvolvida para a base “Petróleo” (1996 a 2000)

Uma análise em cada um dos clusters resultantes foi realizada, observando os títulos presentes nas patentes agrupadas, as classificações que mais se destacam, os depositantes mais ativos naquela área, e uma análise dos significados das palavras-chave encontradas.

Cluster 1 – GLP

Descrição do cluster:

Gás; produção; GLP; disposição; botijão; dois; pressão; equipamento; sistema; temperatura

Títulos pertencentes ao cluster:

- Título: Medidor de gás liquefeito de petróleo (glp), residencial e industrial
- Título: Vaporizador termo-elétrico de gás liquefeito de petróleo (glp) para uso em instalações de médias e grandes vazões
- Título: Processo aperfeiçoado para o fornecimento de gás liquefeito de petróleo
- Título: Detector de gás
- Título: Equipamento e processo para venda automática de botijões de gás liquefeito de petróleo.

Classificações de destaque:

- E21B 43/12: Obtenção de fluidos de poços. Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços. Métodos ou aparelhos para controlar o fluxo do fluido obtido para ou em poços

Depositantes de Destaque:

- Elf Exploration Production (FR)
- Petroleo Brasileiro S.A. - Petrobras (BR/RJ)

Conclusão da análise:

Produção, medição e vaporização de GLP (Gás Liquefeito de Petróleo).

Cluster 2 – Aditivo

Descrição do cluster:

Combustível; derivar; componente; aditivo; fluido; fração; perfuração; formação; combustão; líquido.

Títulos pertencentes ao cluster:

- Título: Método para separar os elementos componentes de uma dispersão
- Título: Aditivo para gás liquefeito de petróleo (g.l.p.), a usar em motores de combustão interna.
- Título: Aditivo para gás liquefeito de petróleo usado, em fornos cerâmicos, como combustível.
- Título: Controle automático de razão entre vazão em sistema de combustão a oxigênio e gases combustíveis

Classificações de destaque:

- C10L 1/22: Combustíveis não incluídos em outro local gás natural; Gás natural de sintético obtido por processos não abrangidos pelas subclasses c 10 g, k; Gás liquefeito de petróleo; Adição de substâncias a combustíveis ou ao fogo para reduzir fumaça ou depósitos indesejáveis ou para facilitar a remoção de fuligem; Acendedores de fogo. Combustíveis carbonáceos líquidos contendo nitrogênio.

Depositantes de Destaque:

- Petróleo Brasileiro S.A. - Petrobras (BR/RJ)
- COMPANHIA ULTRAGAZ S/A. (BR/SP)

Conclusão da análise:

Técnicas de produção de combustíveis derivados de petróleo e utilização de outros componentes para esse fim, como por exemplo, aditivo.

Cluster 3 – Flutuante

Descrição do cluster:

Produção; estrutura; flutuante; reservatório; sistema; disposição; fluido; tubulação; guia; pressão.

Títulos pertencentes ao cluster:

- Título: Método e dispositivo para estabilização da produção de poços de petróleo
- Título: Método e aparelhagem para escoamento da produção submarina de petróleo
- Título: Estrutura flutuante poligonal para uso no mar alto

Classificações de destaque:

- E21B 43/013: Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços especialmente adaptados para a obtenção por meio de instalações subaquáticas. Ligando uma linha de fluxo de produção a uma cabeça de poço sob água.
- E21B 33/038: Vedação ou obturação de furos de sondagem ou de poços. Cabeças de poços; Sua fixação. Conectores utilizados sobre cabeças de poços, por ex., para conectar equipamentos preventivos de explosões e tubos ascendentes
- E21B 17/01: Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção. Tubos ascendentes
- E21B 34/00: Disposições para válvulas utilizadas em furos de sondagem ou poços

Depositantes de Destaque:

- Petróleo Brasileiro S.A. - Petrobras (BR/RJ)

Conclusão da análise:

Métodos e sistemas para a produção de petróleo através de estruturas flutuantes e utilizando tubulações.

Cluster 4 – Mistura

Descrição do cluster:

Água; fluido; formação; composição; conteúdo; pressão; aquoso; sistema; telha; mistura.

Títulos pertencentes ao cluster:

- Título: Fluidos de poço de escavação de petróleo baseado na água
- Título: Sistema de remoção de sedimentos
- Título: Unidade de separação óleo-água
- Título: Composição química; microemulsão, método para preparar uma composição para limpeza de poço, microemulsificada; método para limpar e umedecer com água cortes removidos a partir de poços de petróleo, bem como métodos para limpar um poço de petróleo

Classificações de destaque:

- E21B 37/06: Métodos ou aparelhos para limpar furos de sondagem ou Poços utilizando meios químicos para impedir ou limitar a deposição de parafinas ou de substâncias similares

Depositantes de Destaque:

- Sofitech N.V (BE)
- Hercules Incorporated (US)

Conclusão da análise:

Trata-se de misturas de conteúdos aquosos e outros fluidos.

Cluster 5 – Vedação

Descrição do cluster:

Vedação; roda; gás; produção; obturação; mecanismo; pressão; fluido; anel; ancoragem.

Títulos pertencentes ao cluster:

- Título: Conjunto de vedação
- Título: Obturador para poços de petróleo
- Título: Obturador removível de segregação de trechos de poços de petróleo para teste seletivo em poços revestidos
- Título: Isqueiro à gás.

Classificações de destaque:

- F23Q 2/46: Engenharia mecânica; iluminação; aquecimento; armas; explosão aparelhos de combustão; processos de combustão. Ignição; Dispositivos extintores Isqueiros contendo combustível, por ex., para cigarros. Rodas de fricção; Disposição das rodas de fricção.
- F23Q 2/16: Engenharia mecânica; iluminação; aquecimento; armas; explosão. Aparelhos de combustão; processos de combustão. Ignição; Dispositivos extintores. Isqueiros contendo combustível, por ex., para cigarros. Isqueiros com combustível gasoso, por ex., sendo o gás armazenado em estado líquido.
- E21B 33/12: Vedação ou obturação de furos de sondagem ou de poços no furo de sondagem. Obturadores; Tampões.

Depositantes de Destaque:

- Cricket (FR)
- Weatherfod Industria e Comércio Ltda (BR/RS)

Conclusão da análise:

Obturador de poços de petróleo e isqueiros à gás.

Cluster 6 – Gás

Descrição do cluster:

Sistema; gás; água; pressão; líquido; produção; formação; fluido; conexão; presente

Títulos pertencentes ao cluster:

- Título: Tratamento de águas residuais contaminadas com petróleo através de misturador/decantador à inversão de fases
- Título: Separador de gás dotado de controle automático de nível
- Título: Equipamento para coleta de óleo e derivados de petróleo da superfície da água
- Título: Separador-bomba centrífugo bifásico de baixa taxa de cisalhamento
- Título: Método e aparelho para filtração, degaseificação, desidratação e remoção de produtos de envelhecimento em óleos de petróleo

Classificações de destaque:

- E21B 43/00: Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços.

Depositantes de Destaque:

- Petróleo Brasileiro S.A. - Petrobras (BR/RJ)

Conclusão da análise:

Sistemas para separar água e gás liquefeito de petróleo.

Cluster 7 – Alimentação

Descrição do cluster:

Conteúdo; alimentação; metal; zona; formação; carga; corrente; material; produção; sistema.

Títulos pertencentes ao cluster:

- Título: Processo para preparar coque de qualidade de anodo a partir de uma carga de alimentação de resíduo de petróleo contendo contaminantes de metal e enxofre
- Título: Processo para converter cargas de alimentação de petróleo ebulindo na faixa de resíduos em produtos de ponto de ebulição mais baixo
- Título: Processo para a redução da quantidade de ácidos carboxílicos em correntes de petróleo
- Título: Processo para preparar coque de qualidade de anodo a partir de uma carga de alimentação de resíduo de petróleo contendo contaminantes de metal e enxofre

Classificações de destaque:

- C07C 7/148: Química Orgânica. Compostos acíclicos ou carbocíclicos. Purificação; Separação; Estabilização; Uso de aditivos por tratamento que origine modificação química de pelo menos um composto
- C08F 240/00: Compostos macromoleculares orgânicos; sua preparação ou seu processamento químico; composições baseadas nos mesmos. Compostos macromoleculares obtidos por reações compreendendo apenas ligações insaturadas carbonocarbono. Copolímeros de compostos tendo uma ou mais ligações triplas carbono-carbono

Depositantes de Destaque:

- Exxon Research And Engineering Company (US)

Conclusão da análise:

Processos e materiais para cargas de alimentação de petróleo e remoção de metal.

5.4.2.2 Petróleo (1996 a 2000) – Temis

A seguir é apresentado na Tabela 5-6 a relação de palavras-chave encontradas pelo módulo IDC para a base de dados composta de patentes depositadas entre os anos 1996 a 2000 e que continham o termo “petróleo” no seu RESUMO. A tabela mostra ainda o número de documentos agrupados em cada cluster, do total de 226 documentos encontrados. O nome dado ao cluster foi escolhido de acordo com o termo que melhor representa o mesmo, seguido do termo “_Temis”.

Cluster	Núm. de Patentes	Nome	Palavras-chave
1	43	Gás_Temis	fase; poço; método; escoamento; gás; produção; hidrocarbonetos; linha; fluido; etapa.
2	39	Alimentação _Temis	ácido; metal; composição; carga; cimento; processo; alimentação; resinar; conter; referir.
3	35	Flutuante _Temis	tubo; conexão; cabo; flutuante; estrutura; navio; submarino; assessor; elemento; interno.
4	30	Vedação _Temis	coluna; bomba; vedação; poço; elemento; bombear; camisa; sistema; duplo; areia.
5	28	GLP_Temis	gás; GLP; recipiente; roda; válvula; liquefazer; botijão; segurança; automático; regulador.
6	26	Proteção _Temis	água; derivar; óleo; tanque; produto; proteção; rocha; alumínio; pena; equipamento.
7	25	Aditivo _Temis	combustível; motor; aditivo; solução; átomo; álcool; forno; veículo; marcador; composto.

Tabela 5-6 – Resultado da ferramenta Temis para a base “Petróleo” (1996 a 2000)

O gráfico da figura 5.15 apresenta a proporção de documentos coincidentemente agrupados nos resultados apresentados pela ferramenta e os resultados de Temis.

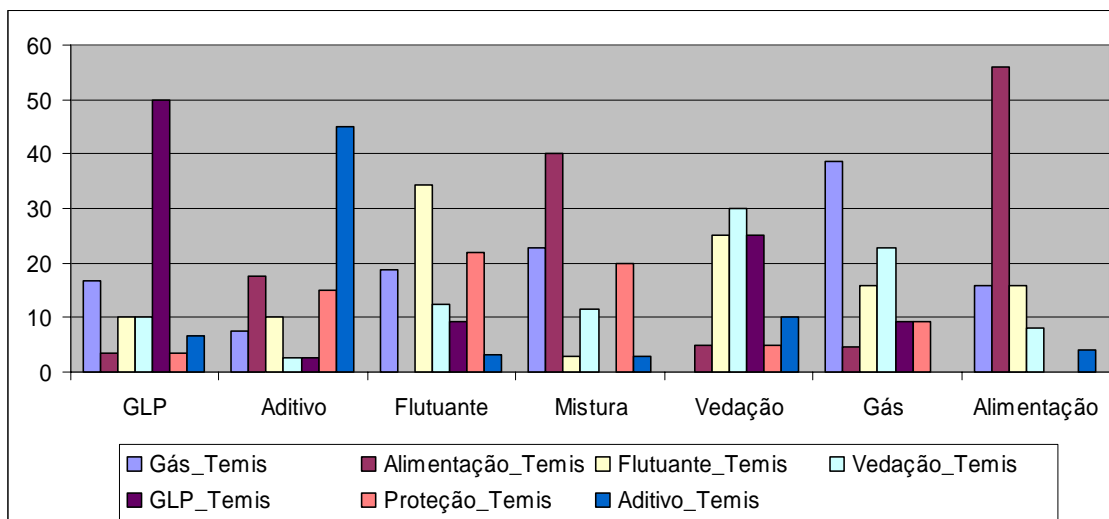


Figura 5-15 – Gráfico de comparação da ferramenta implementada e Temis para “petróleo” (1996 a 2000)

De acordo com o gráfico de comparações dos resultados, pode-se observar que:

- O clusters “GLP” possui 50% de documentos coincidentes com o cluster “GLP_Temis”;
- O cluster “Aditivo” possui 45% de documentos coincidentes com o cluster “Aditivo_Temis”;
- O cluster “Flutuante” possui 34% de documentos coincidentes com o cluster “Flutuante_Temis”;
- O cluster “Mistura” possui 40% de documentos coincidentes com o cluster “Alimentação_Temis”;
- O cluster “Vedação” possui 30% de documentos coincidentes com o cluster “Vedação_Temis” e 25% de documentos coincidentes com os clusters “Flutuante_Temis” e “GLP_Temis”.
- O cluster “Gás” possui 39% de documentos coincidentes com o cluster “Gás_Temis”;

- E o cluster “Alimentação” possui 56% de documentos coincidentes com o cluster “Alimentação_Temis”.

A relação de documentos coincidentes se apresenta com uma proporção alta, formando clusters bem definidos e semelhantes. Porém, o cluster “Proteção_Temis” não se destacou na ferramenta anterior, estando oculto dentro de outros clusters.

5.4.2.3 *Petróleo (1996 a 2000) – Statistica*

A ferramenta Statistica encontrou no seu pré-processamento 712 termos (radicais) relevantes no total de 226 documentos a serem clusterizados, utilizando a mesma lista de *StopWords* utilizada pela ferramenta anterior.

Do total de termos, aqueles que não se encontravam presentes em pelo menos 1% do total de documentos foram retirados do processo de clusterização. O algoritmo de clusterização utilizado foi o k-means, utilizando o mesmo número de clusters usado nos testes anteriores da presente base de dados.

Na tabela 5.7 é apresentado o número de documentos agrupados, o nome dado aos clusters, seguido do termo “_Statistica” e o conjunto de palavras-chave que a ferramenta considerou mais importante para a clusterização.

Cluster	Núm. de Patentes	Nome	Palavras-chave
1	115	Flutuante _Statistica	Poço; tubo; gás; sistema; bombeamento; separador; água; flutuante; fluido; produção.
2	14	Vedação _Statistica	Vedação; válvula; roda; obturação; ancoragem; coluna; produção; poço; fase; injeção.
3	12	Vazamento _Statistica	Fluxo; monitor; sensor; pressão; medição; síntese; fluido; vazamento; diferencial; central.
4	12	Separador _Statistica	Decantar; câmara; funcionamento; leito; condição; separador; orgânico; interface; diesel; mistura.
5	62	Aditivo _Statistica	Decomposição; ácido; contínuo; partícula; peso; processo; aditivo; resíduo; óleo; solução.
6	5	Alimentação _Statistica	Zona; abaixo; corrente; kpa; catalítico; reator; parcial; alimentação; referente; ascendente.
7	15	GLP _Statistica	Botijão; liquefeito; gás; glp; recipiente; segurança; vaporização; desenvolvimento; recebimento; elétrico.

Tabela 5-7 – Resultado da ferramenta Statistica para a base “Petróleo” (1996 a 2000)

Na figura 5.16 é apresentado o gráfico de proporções de documentos clusterizados coincidentemente pela ferramenta implementada e Statistica. O fato do Statistica ter agrupado a maioria dos documentos em um único cluster fez com que vários grupos encontrados anteriormente se distribuíssem ao longo de apenas 2 grupos: “Flutuante_Statistica” e “Aditivo_Statistica”.

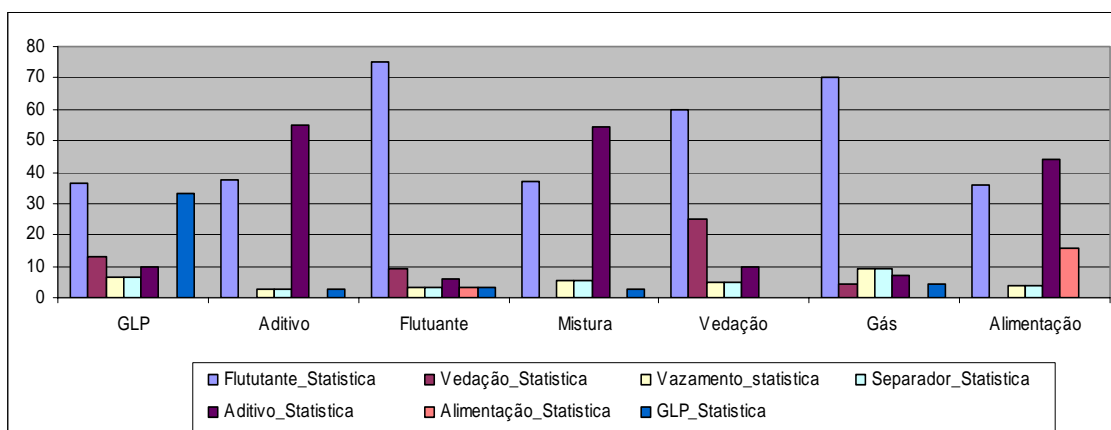


Figura 5-16 – Gráfico de comparação da ferramenta implementada e Statistica para “petróleo” (1996 a 2000)

Pode-se notar que:

- O cluster “GLP” apresentou 33% de seus documentos coincidentes com o cluster “GLP_Statistica”;
- O cluster “Vedação” apresentou 25% de documentos coincidentes com o cluster “Vedação_Statistica”;
- O cluster “Alimentação” apresentou 16% de documentos coincidentes com o cluster “Alimentação_Statistica”.

Os demais clusters não apresentaram tamanha semelhança, pelo fato da distribuição realizada pelo Statistica ocorrer de forma diferente. Essa ferramenta foi capaz de encontrar os dois conjuntos maiores de assuntos relacionados às patentes, ou seja, métodos e aparelhos para flutuação e alguns aditivos, e encontrou ainda algumas patentes que se diferenciam das demais como os clusters “Vazamento” e “Separador”.

5.4.2.4 Clustering de Patentes de 2000 a 2005

A tela de resultados do processamento das patentes depositadas entre os anos de 2001 e 2005 pode ser visualizado na Figura 5-17. A tela apresenta os nomes dos documentos de patentes agrupados e os radicais das palavras consideradas mais importantes em cada cluster.

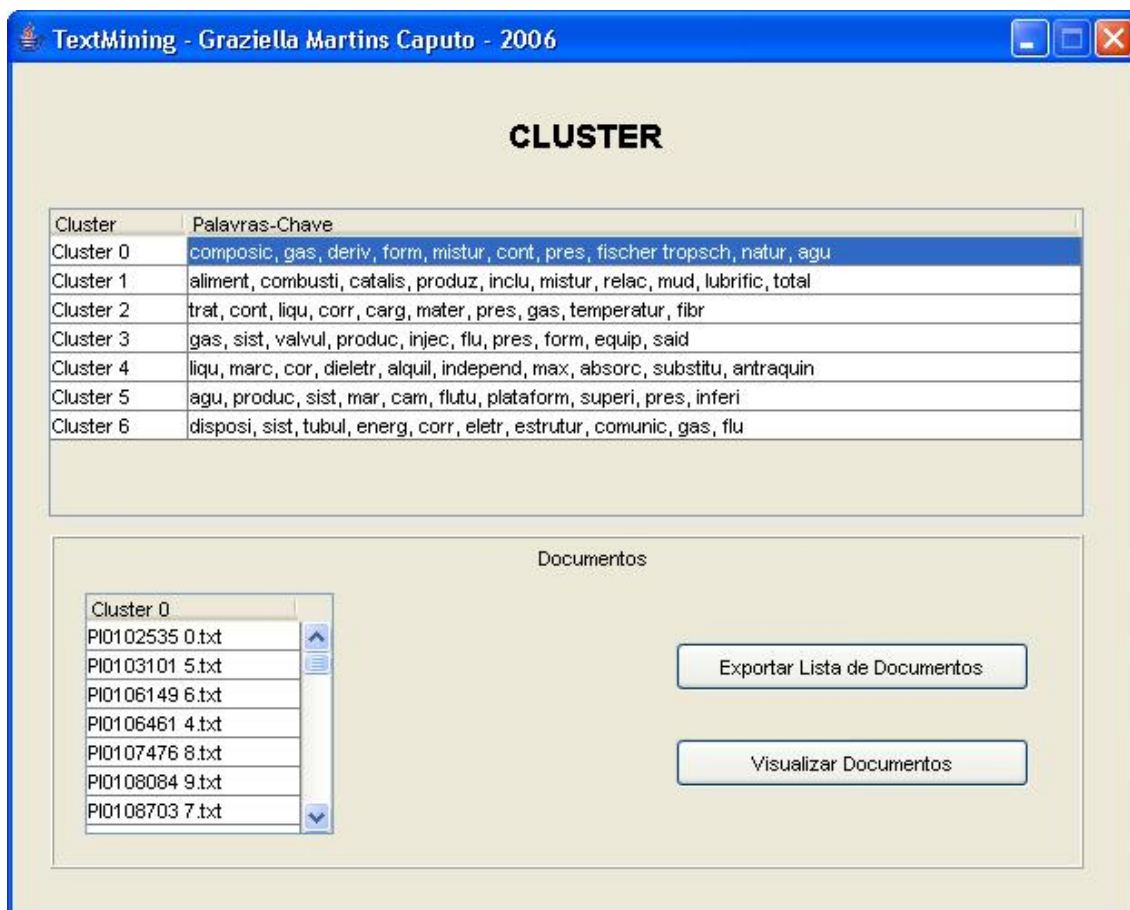


Figura 5-17 – Clusterização da base de dados com o termo “Petróleo” de 2001 a 2005

Na Tabela 5-8 é apresentado o número de documentos agrupados em cada cluster e as palavras-chave que descrevem cada um deles.

Cluster	Núm. de Patentes	Nome	Palavras-chave
1	49	Composição	Composição; gás; derivados; formação; mistura; conteúdo; presença; Fischer-Tropsch; natural; água.
2	43	Alimentação	Alimentação; combustível; catalisador produção; inclusão; mistura; relacionamento; mudança; lubrificantes; total
3	40	Tratamento	Tratamento; conteúdo; líquido; corrente; carga; matéria; pressão; gás; temperatura; fibra
4	56	Gás	Gás; sistema; válvula; produção; injeção; fluido; pressão; formação; equipamento; saída
5	13	Corante	Líquido; marcação; corante; dieletro; alquila; independente; máximo; absorção; substituição; antraquinona
6	77	Plataforma	Água; produção; sistema; mar; camada; flutuante; plataforma; superior; pressão; inferior
7	41	Energia	Dispositivo; sistema; tubulação; energia; corrente; eletricidade; estrutura; comunicação; gás; fluido

Tabela 5-8 – Resultado da ferramenta desenvolvida para a base “Petróleo” (2001 a 2005)

A seguir é apresentada uma descrição do conteúdo presente em cada um dos clusters encontrados, o que inclui as palavras-chave, alguns títulos escolhidos aleatoriamente para a ilustração da consistência dos clusters, as classificações mais relevantes entre as patentes, os depositantes que mais atuam no conjunto de patentes depositadas e uma conclusão resumida sobre o conteúdo das documentos pertencente a cada cluster, obtida através de interpretação dos RESUMOS.

Cluster 1 – Composição

Descrição do cluster: Composição; gás; derivados; formação; mistura; conteúdo; presença; Fischer-Tropsch; natural; água.

Títulos pertencentes ao cluster:

- Título: Material de sustentação de poços de gás e de petróleo, composição e método de formação respectivos e métodos de aumento da permeabilidade de uma fratura de poço de gás ou de petróleo
- Título: Processos para a fabricação de uma matéria-prima básica de lubrificantes e para receber um produto fischer-tropsch de gás natural convertido, e, método para a fabricação de produtos de hidrocarbonetos a partir de campos remotos de gás natural
- Título: Método de inibição da oxidação de um produto de fischer-tropsch, e, produto hidrocarbonáceo misturado
- Título: Composições de fluido de perfuração biodegradável à base de óleo e processo de perfuração de poços de petróleo e gás

Classificações de destaque:

- C09K 7/00: Corantes; tintas; polidores; resinas naturais; adesivos; composições diversas; diversas aplicações de substâncias. Substâncias para aplicações diversas, não incluídas em outro local. Composições para perfuração de poços.

Depositantes de Destaque:

- Petróleo Brasileiro S.A. - Petrobras (BR/RJ)

Conclusão da análise:

Utilização de produtos e derivados de Fischer-Tropsch para formação e mistura de combustível, petróleo e gás natural.

Cluster 2 – Alimentação

Esse cluster representa uma série de patentes depositadas pela empresa “Shell Internationale Research Maatschappij B.V. (NL)”, e se diferenciam, no geral, apenas por algumas classificações.

Descrição do cluster:

Alimentação; combustível; catalisador produção; inclusão; mistura; relacionamento; mudança; lubrificantes; total

Títulos pertencentes ao cluster:

- Título: Métodos de produzir um produto de petróleo bruto e combustível de transporte, combustível de aquecimento, lubrificantes ou substâncias químicas, e, produto de petróleo bruto
- Título: Métodos de produzir um catalisador de sulfeto de metal de transição, um produto de petróleo bruto e combustível de transporte, combustível de aquecimento, lubrificantes ou substâncias químicas, catalisador de sulfeto de metal de transição, e, produto de petróleo bruto

Classificações de destaque:

- C10G: Craqueamento de óleos de hidrocarboneto; Produção de misturas líquidas de hidrocarboneto; Recuperação de óleos de hidrocarboneto a partir de xisto betuminoso, arenito oleífero, ou gases; Refinação de misturas constituídas principalmente de hidrocarboneto; Reforma de nafta; Ceras minerais
- C10G 49/00: Tratamento de óleos de hidrocarboneto, na presença de hidrogênio ou de compostos geradores de hidrogênio.
- C10G 49/26: Controle ou regulagem
- C10G 65/00: Tratamento de óleos hidrocarbonetos apenas por dois ou mais processos de hidrotreatamento

Depositantes de Destaque:

- Shell Internationale Research Maatschappij B.V. (NL)

Conclusão da análise:

Métodos e produtos de produção de combustíveis, lubrificantes, substâncias químicas, produto de petróleo bruto e catalisadores.

Cluster 3 – Tratamento

Descrição do cluster:

Tratamento; conteúdo; líquido; corrente; carga; matéria; pressão; gás; temperatura; fibra

Títulos pertencentes ao cluster:

- Título: Material em aço com superfície tratada, método para sua fabricação, e líquido para tratamento por conversão química
- Título: Processo de tratamento de cargas de hidrocarbonetos
- Título: Processo químico-mecânico para reduzir a contaminação produzida pela combustão de combustíveis fósseis, petróleo e seus derivados
- Título: Processo para tratar um petróleo bruto contendo gás natural
- Título: Método para tratamento de efluentes cáusticos usados, e, método para tratamento de materiais cáusticos

Classificações de destaque:

- **B01J**: Operações de Processamento; Transporte. Processos ou aparelhos físicos ou químicos em geral. Processos químicos ou físicos, por ex., catálise, química coloidal; Aparelhos pertinentes aos mesmos.
- **C10G**: Craqueamento de óleos de hidrocarboneto; Produção de misturas líquidas de hidrocarboneto; Recuperação de óleos de hidrocarboneto a partir de xisto betuminoso, arenito oleífero, ou gases; Refinação de misturas constituídas principalmente de hidrocarboneto; Reforma de nafta; Ceras minerais

Depositantes de Destaque:

- Petróleo Brasileiro S.A - Petrobras (BR/RJ) / Akzo Nobel N.V. (NL)

Conclusão da análise:

Processos de tratamento de cargas de hidrocarboneto, matérias e líquidos derivadas do petróleo com temperatura e pressão.

Cluster 4 – Gás

Descrição do cluster:

Gás; sistema; válvula; produção; injeção; fluido; pressão; formação; equipamento; saída.

Títulos pertencentes ao cluster:

- Título: Válvula de bombeio pneumático com venturi de corpo central
- Título: Disposição introduzida em lacre de segurança para botijão de gás e assemelhados
- Título: sensor para detecção e alarme de vazamento de água, petróleo e seus derivados em tanques ou condutos de carcaça dupla
- Título: sistema de tubulações compostas
- Título: Conferidor manual para gás liquefeito de petróleo

Classificações de destaque:

- E21B: Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços.
- B01D: Processos ou aparelhos físicos ou químicos em geral. Separação.
- F17C: Armazenamento ou distribuição de gases ou líquidos. Vasos para conter ou armazenar gases comprimidos, liquefeitos ou solidificados; Tanques de gás de capacidade fixa; Enchimento ou descarga de vasos com gases comprimidos, liquefeitos ou solidificados.

Depositantes de Destaque:

- Petróleo Brasileiro S.A. - Petrobras (BR/RJ)

Conclusão da análise:

Mecanismos de armazenamento e controle de injeção, bombeio e retenções de GLP e fluidos.

Cluster 5 – Corante

Descrição do cluster:

Líquido; marcação; corante; dieletro; alquila; independente; máximo; absorção; substituição; antraquinona

Títulos pertencentes ao cluster:

- Título: Marcadores moleculares para sistemas de solventes orgânicos
- Título: Método para marcar um hidrocarboneto de petróleo líquido
- Título: Composição, composto, e, método para marcar um produto de petróleo líquido

Classificações de destaque:

- C10L: Indústrias do petróleo; do gás ou do coque; gases técnicos contendo monóxido de carbono; combustíveis; lubrificantes; turfa. Combustíveis não incluídos em outro local; Gás natural; Gás liquefeito de petróleo; Adição de substâncias a combustíveis ou ao fogo para reduzir fumaça ou depósitos indesejáveis ou para facilitar a remoção de fuligem; Acendedores de fogo
- C10L 1/00: Combustíveis carbonáceos líquidos
- C09B: Corantes; tintas; polidores; resinas naturais; adesivos; composições diversas; diversas aplicações de substâncias. Corantes orgânicos ou compostos estreitamente relacionados à produção de corantes; mordentes; Lacas.

Depositantes de Destaque:

- Rohm And Haas Company (US)

Conclusão da análise:

Marcação de produtos de petróleo líquido através de corantes.

Cluster 6 – Plataforma

Descrição do cluster:

Água; produção; sistema; mar; camada; flutuante; plataforma; superior; pressão; inferior.

Títulos pertencentes ao cluster:

- Título: Método para a recuperação secundária de petróleo a partir de uma localização abaixo de um corpo de água salina, e, sistema de tratamento da água do mar
- Título: Aparelho e método para remover matéria em suspensão de um líquido
- Título: Dispositivo de transferência de fluido entre dois suportes flutuantes e instalação de produção petrolífera no mar
- Título: Plataforma marítima semi-submersível de produção

Classificações de destaque:

- E21B: Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços
- B63B 21/50: Amarração; Equipamento para deslocar, rebocar ou empurrar; Ancoragem. Disposições para ancoragem de embarcações especiais, por ex., para plataformas flutuantes de perfuração ou dragas
- C02F: Tratamento de água, de águas residuais, de esgotos, ou de lamas e lodos. Tratamento de água, águas residuais, esgotos, ou de lamas e lodos

Depositantes de Destaque:

- Petróleo Brasileiro S.A. - Petrobras (BR/RJ)

Conclusão da análise:

Sistemas e equipamentos para a manipulação de petróleo em superfícies líquidas.

Cluster 7 – Energia

Descrição do cluster:

Dispositivo; sistema; tubulação; energia; corrente; eletricidade; estrutura; comunicação; gás; fluido

Títulos pertencentes ao cluster:

- Título: Sistema para encaminhar controlavelmente comunicações e energia elétrica tendo uma corrente variável com o tempo através de uma estrutura de tubulação, poço de petróleo para produzir produtos de petróleo e método de produzir produtos de petróleo a partir de um poço de petróleo
- Título: Aquecedor tubular submarino para quebra de hidratos
- Título: Dispositivo de impedância de corrente, e, método para operar um poço de petróleo
- Título: Poço de petróleo para produção de produtos de petróleo e métodos de produzir petróleo a partir de um poço de petróleo e de injetar controlavelmente fluido em uma formação com um poço

Classificações de destaque:

- E21B 47/12: Provas ou ensaios; Levantamento de furos de sondagem ou de. Meios para transmitir sinais de medição do poço para a superfície, por ex., perfilagem durante a perfuração
- E21B 17/00: Outros equipamentos ou detalhes para perfuração; Equipamento de poços ou manutenção de poços; Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção

Depositantes de Destaque:

- Shell Internationale Research Maatschappij B.V. (NL)

Conclusão da análise:

Dispositivos e métodos de operação e comunicação em poços de petróleo, como condução de energia elétrica e injeção de fluidos.

Verificando as principais palavras-chave de cada cluster e analisando superficialmente o significado e conteúdo de cada um deles, pode-se notar uma diferença nos conteúdos das bases de documentos agrupados pelo sistema de mineração de patentes.

Tal fato pode revelar uma mudança no foco de estudos e desenvolvimentos industriais e tecnológicos no que diz respeito a componentes derivados de petróleo, técnicas de utilização do mesmo, dentre outros.

Além disso, o resultado do estudo comprova a viabilidade de executar busca por tendências tecnológicas utilizando os recursos disponíveis nas patentes industriais e ferramentas de mineração de texto.

No entanto, apenas uma análise realizada por especialistas nas áreas em questão poderiam comprovar tal hipótese.

5.4.2.5 Petróleo (2001 a 2005) – Temis

A Tabela 5-9 apresenta o resultado da clusterização realizada pela ferramenta Temis para os anos de 2001 a 2005, com o respectivo número de documentos agrupados em cada cluster. Apresenta ainda o nome dado ao cluster, seguido do termo “_Temis”.

Cluster	Núm. de Patentes	Nome	Palavras-chave
1	68	Alimentação_Temis	bruto; produto; alimentação; propriedade; combustível; produzir; catalisador; bruto; lubrificantes; mudar.
2	62	Energia_Temis	poço; sistema; tubulação; dispositivo; fluido; fundo; injeção; produção; energia; bombear.
3	55	Plataforma_Temis	tubo; plataforma; camada; flutuante; mar; rasgo; superfície; coluna; roscar; linha.
4	39	Tratamento_Hidrocarbonet o _Temis	nitrogênio; processo; aço; peso; carga; hidrocarbonetos; tratamento; fração; síntese; carbono.
5	36	Corante_Comp osição_Temis	composição; R [^] ; grupo; hidrocarboneto; nm; óleo; corante; agente; líquido; marcar.
6	30	Contaminação_Temis	tanque; resinar; massa; derivar; equipamento; modelo; material; plástico; contaminar; água.
7	29	Gás_Temis	combustível; gás; válvula; motor; botijão; combustão; segurança; GLP; circuito; diesel.

Tabela 5-9 – Resultado da ferramenta Temis para a base “Petróleo” (2001 a 2005)

Na Figura 5-18 é representado o gráfico de comparação entre os clusters encontrados na ferramenta implementada e os do módulo IDC, apresentando a proporção de documentos coincidentes nos clusters.

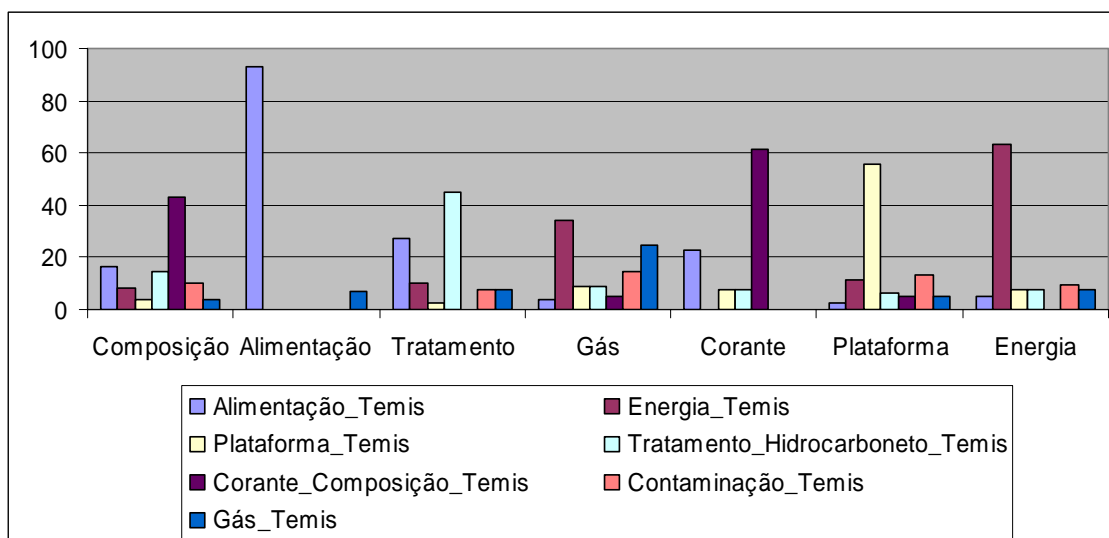


Figura 5-18 – Gráfico de comparação da ferramenta implementada e Temis para “petróleo” (2001 a 2005)

De acordo com o gráfico de comparação, pode-se observar:

- O cluster “Composição” apresenta 43% de documentos coincidentes ao cluster “Corante_Composição_Temis”;
- O cluster “Alimentação” apresenta 93% de documentos coincidentes ao cluster “Alimentação_Temis”;
- O cluster “Tratamento” apresenta 45% de documentos coincidentes ao cluster “Tratamento_Hidrocarboneto_Temis”;
- O cluster “Gás” apresenta 33% de documentos coincidentes ao cluster “Energia_Temis” e 25% ao cluster “Gás_Temis”;
- O cluster “Corante” apresenta 62% de documentos coincidentes ao cluster “Corante_Temis”.
- O cluster “Plataforma” apresenta 56% de documentos coincidentes ao cluster “Plataforma_Temis”.
- O cluster “Energia” apresenta 64% de documentos coincidentes ao cluster “Energia_Temis”.

A quantidade de palavras-chave encontradas nos clusters semelhantes, são, em média, 5, apesar de possuírem conteúdos bastante similares.

5.4.2.6 Petróleo (2001 a 2005) – Statistica

O total de termos utilizado para a clusterização foram 858, utilizando a mesma lista de *StopWords* utilizada pela ferramenta implementada.

A tabela 5.10 apresenta o resultado obtido pela ferramenta Statistica para a base de dados de petróleo, entre os anos de 2001 a 2005.

Cluster	Núm. de Patentes	Nome	Palavras-chave
1	93	Plataforma _Statistica	Tubulação; flutuação; sistema; bombeamento; poço; mar; plataforma; superfície; instalação; disposição.
2	146	Composição _Statistica	Composição; processo; partícula; óleo; derivação; gás; hidrocarboneto; básico; método; fluido.
3	39	Alimentação _Statistica	Bruto; alimentação; produto; propriedade; combustível; produção; catalisador; contato; mpa; mudar.
4	24	Hidrocarbonet o _Statistica	Gás; fração; processo; hidrocarboneto; nitrogênio; corrente; adsorvente; enxofre; natural.
5	12	Energia _Statistica	Comunicação; tubulação; dispositivo; controle; corrente; elétrica; poço; adaptador; energia; sensor.
6	2	Sensor _Statistica	Carcaça; ondas; sinalização; detecção; eletrônico; identificador; vazamento; sensor; bateria; duplo.
7	3	Selante _Statistica	Projeto; seção; sonda; cabeça; segurança; fluido; selante; contemplar; integral; adaptador.

Tabela 5-10 – Resultado da ferramenta Statistica para a base “Petróleo” (2001 a 2005)

A figura 5.19 apresenta a proporção de documentos coincidentes agrupados pela ferramenta Statistica e a ferramenta implementada.

Pode-se observar que dos 319 documentos, 146 foram inseridos em um mesmo cluster. Dessa forma, apenas alguns clusters foram considerados semelhantes daqueles encontrados pela ferramenta comparada.

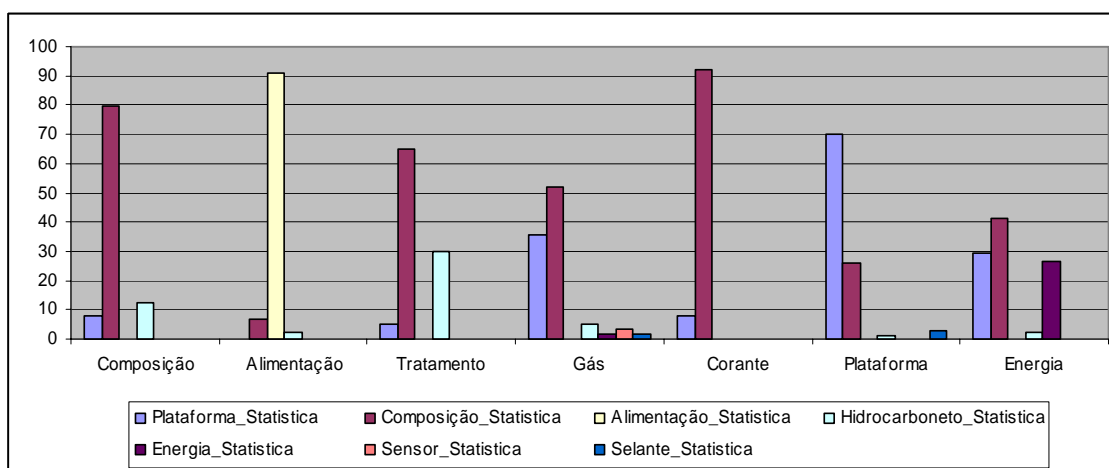


Figura 5-19 – Gráfico de comparação da ferramenta implementada e Statistica para “petróleo” (2001 a 2005)

- O cluster “Alimentação” possui 91% de documentos coincidentes com o cluster “Alimentação_Statistica”;
- O cluster “Plataforma” possui 70% de documentos coincidentes ao cluster “Plataforma_Statistica”;
- O cluster “Energia” possui 27% de documentos coincidentes ao cluster “Energia_Statistica”.

Os demais clusters não apresentam alto grau de semelhança com a ferramenta implementada. Comparando com o resultado de Temis, no entanto, pode-se perceber algumas semelhanças em alguns clusters, como por exemplo, os pares “Plataforma_Statistica” e “Plataforma_Temis”, “Composição_Statistica” e “Corante_Composição_Temis”, dentre outros.

O Statistica encontrou ainda alguns clusters que não foram detectados pelas outras ferramentas por possuírem baixa quantidade de patentes, como o caso dos clusters “Selante_Statistica” e “Sensor_Statistica”.

Desse resultado, conclui-se que o Statistica determinou os tópicos principais presentes no conjunto total de patentes, e encontrou ainda alguns outliers.

5.5 Considerações Finais

A aplicação de uma base de dados em diferentes ferramentas é capaz de gerar diversos resultados, de acordo com o processamento que cada uma realiza.

A análise, citada anteriormente, destacou alguns pontos comuns e outros extremos entre os resultados gerados por cada uma das ferramentas.

As diferenças nos resultados se deve ao fato de cada ferramenta possuir um algoritmo de execução diferente, tanto de pré-processamento, quanto de clusterização.

A ferramenta implementada utilizou lista de *StopWords* iterativamente ajustada para melhor destacar os termos relevantes dos documentos. Foi utilizado como algoritmo de *stemming* o *Stemming Portuguese* e para clusterização, o algoritmo k-means, através de escolha aleatória dos centróides iniciais de cada cluster.

O módulo IDC do Temis utiliza uma lista interna de *StopWords*, além de realizar análise semântica nos termos, reconhecendo-os como substantivo, adjetivo ou verbo. Algumas vezes, a ferramenta pode não reconhecer adequadamente um termo e determinar que o mesmo é verbo e adjetivo ao mesmo tempo, por exemplo. O módulo possui a propriedade de reconhecer diversos idiomas, entre os quais, o português, porém, com menor detalhamento.

O programa Statistica da StatSoft possui um algoritmo de *stemming* e um lista de *StopWords* para a língua portuguesa, tendo a propriedade de alterar essa lista e dessa forma, pôde ser utilizada a mesma lista da ferramenta implementada. Além de possuir diversos algoritmos de mineração de dados, possui também o k-means, com a propriedade de inserir aleatoriamente os centróides no espaço vetorial. A ferramenta tem a propriedade de eliminar as palavras que aparecem em menos de 1% dos

documentos. Isso faz com que vários termos importantes para a discriminação de assuntos sejam retirados da clusterização. Esse fato faz com que a ferramenta destaque apenas o conteúdo principal dos documentos, não considerando o detalhamento das patentes. Dessa forma, a clusterização inseriu vários documentos em poucos clusters, detectando alguns outros outliers.

Essas diferenças entre as ferramentas justifica os resultados encontrados por cada uma delas. Para uma análise completa, a utilização da mineração de textos deve ser comparada em condições iguais.

O sistema implementado tem a propriedade de manipular documentos de patentes de forma adequada, utilizando diversos atributos dos documentos e buscando uma perda mínima de informação. Dessa forma, se mostra eficiente para tal uso, apresentando resultados relevantes para o conhecimento da informação contida na base de dados.

6 Conclusão

Utilizar patentes industriais para geração de Inteligência Competitiva é um estudo que ainda recebe pouca ou nenhuma atenção pelas empresas comparado aos benefícios que pode trazer.

As patentes industriais, por possuírem grande fonte de informação técnica e comercial, armazenam informações preciosas sobre a concorrência e sobre os atuais focos de pesquisa.

O fato de possuírem um órgão de gerenciamento dos depósitos, e seguirem as normas do Tratado Internacional de Patentes (PCT), a Classificação Internacional de Patentes (CIP), as patentes são facilmente localizadas, assim como é possível selecionar aquelas pertencentes a uma determinada tecnologia.

O presente estudo vem afirmar a utilidade e benefícios da implantação de sistemas computacionais inteligentes para a busca de informações relevantes que possam trazer vantagem competitiva para qualquer organização que esteja envolvida com pesquisa e/ou desenvolvimento.

Através de uma ferramenta específica para patentes industriais, a manipulação dos documentos se torna mais fácil, não precisando de mecanismos externos para a separação dos campos a serem utilizados.

A aplicação da mineração de textos prova ser capaz de destacar as informações mais relevantes, agrupando documentos similares, de acordo com uma quantidade pré-determinada de clusters e extrair os seus conceitos principais, através da exibição das palavras-chave.

Além disso, a ferramenta possui a funcionalidade de contabilizar as CLASSIFICAÇÕES presentes em cada grupo, o que complementa e/ou confirma o conteúdo das patentes agrupadas em cada cluster, agregando valor aos resultados

apresentados. A exibição dos DEPOSITANTES das patentes agrupadas é capaz de auxiliar na compreensão dos maiores concorrentes em uma determinada área.

Como o sistema foi implementado em java, possui uma classe própria para a execução do *Stemming*. Isso implica na facilidade de converter o sistema para qualquer idioma, atualizando apenas essa classe. Esse fato é facilitado ainda pela propriedade de direcionar a lista de *StopWords* em tempo de execução. Isso implica na fácil conversão da ferramenta para as patentes internacionais, como por exemplo, USPTO, sendo capaz de capturar conhecimento da mesma forma.

A aplicação da técnica em um número relativamente grande de documentos de patentes comprovou ser de extrema eficiência, revelando clusters bem definidos.

Uma quantidade menor de patentes, no entanto, não foi capaz de apresentar tamanha eficiência, sendo alguns clusters minimamente diferente de outros, ou com documentos com assuntos diversos agrupados num mesmo cluster.

Tais hipóteses, porém, só poderiam ser comprovadas com o auxílio de especialistas da área estudada.

A fragmentação da base de dados pelos períodos de tempo em que as patentes foram depositadas, revelou algumas mudanças de foco de pesquisas ocorridos ao longo dos últimos dez anos. Aprofundando o estudo nesse sentido, uma análise poderia revelar as mudanças exatas ocorridas e provavelmente se tornar fonte de prospecção tecnológica.

Especificamente para as patentes brasileiras, as empresas podem fazer uso para entender quais são os concorrentes multinacionais e os internos, e buscar as principais tecnologias desenvolvidas no país.

Os resultados obtidos podem ser melhor aproveitados se concatenados com outras áreas relacionadas à inteligência competitiva como: mineração de outras fontes de conhecimento como relatórios e páginas web, e ferramentas de informações como: perfil dos competidores, análise financeira, análise SWOT, desenvolvimento de cenários, análise de ganho e perda, jogos de guerra, análise conjunta e simulação/modelagem.

Essas análises são utilizadas em suporte à tomada de decisão, monitoramento do mercado, identificação de oportunidades de mercado, desenvolvimento de planos de mercado, suporte ao marketing e à venda, e diversas outras tarefas.

Com o auxílio de especialistas, os resultados das análises podem ser interpretados e convertidos para conhecimento a ser utilizado pelas empresas de P&D como vantagem competitiva e veículo de inovação.

6.1 Trabalhos Futuros

O sistema desenvolvido nessa dissertação foi implementado para identificar os campos presentes nos documentos de patentes e aplicar técnicas de mineração de textos nos campos RESUMOs.

Algumas outras funcionalidades que poderiam ser inseridas no sistema, agregando valor aos resultados finais, como por exemplo:

- Aplicação de mineração de dados nos campos categóricos existentes nas patentes;
- Aplicação da mineração de textos nos campos DESCRIÇÃO, existente nas patentes. Essa informação não está disponível nos documentos de patentes do site do INPI;
- Utilização de thesaurus para melhoramento do pré-processamento;
- Conversão dos arquivos de patentes para o formato XML, ou para um banco de dados relacional;
- Implementação de um algoritmo para a detecção do número ideal de clusters;
- Inserção de novas formas de visualização dos resultados, como gráficos, por exemplo;
- Implementação de funcionalidades capazes de manipular os campos DATA presentes na patentes, como forma de distinção de tempos e melhor busca por modificações nas tendências ao longo do tempo;

- Implementação de outros como, por exemplo, séries temporais, outros algoritmos de clustering e utilização de lógica fuzzy.

Esses tópicos poderiam ser capazes de otimizar os resultados encontrados pelo sistema, aproveitando melhor a meta informação presente nas patentes.

A aplicação do presente estudo pode ser utilizada em qualquer área que esteja envolvida, direta ou indiretamente com informações presentes em documentos de propriedade intelectual.

Referências Bibliográficas

- AHMAD, K., AL-THUBAITY, AM., 2003, "Can Text Analysis Tell us Something about Technology Progress?". *Workshop on Patent Corpus Processing*. pp.45-65, Sapiro, Japão, Julho.
- ANACUBIS. Disponível em <<http://www.anabubis.com>> Acesso em: 10 jan. 2006.
- ANALYSIS Tools. Disponível em <<http://www.piug.org/vendor.html#bmTools>> Acesso em: 02 jan. 2006.
- ARCHIBUGI, D., PIANTA, M. 1996, "Measuring technological change through patents and innovations surveys. *Technovation*, 16(9), pp. 146-451.
- APPLEYARD, M.M., KALSON, G.A., 1999, "Knowledge diffusion in semiconductor industry". *Journal of Knowledge Management*, v.3, n.4, pp.288-295.
- AZEVEDO, M.C., COSTA, H.G., 2001, "Métodos para a avaliação da postura estratégica", Caderno de pesquisa em Administração, v.08, n.2, abril.
- BOATRIGHT, J. R., 2000, *Ethics and the conduct of business*, ed. 3, Upper Saddle River, NJ7 Prentice Hall.
- BIZINT Smart Charts. Disponível em <<http://www.bizcharts.com/sc4pats/>> Acesso em: 10 jan. 2006.
- BRENNER, M., 2005, "Leveraging Analysis and Collection Techniques", *Competitive Intelligence*, Society of Competitive Intelligence Professionals, v. 8, n. 3 (May/June), pp. 6-19.
- BRIN, S., PAGE, L., 1998, "The anatomy of a large scale hypertextual web search engine", In Proc. WWW7.
- CAPUTO, G.M., BASTOS, V.M., EBECKEN, N.F.F., 2006, Using Text Mining to Understand the call center customers claims, *Data Mining & Information Enginnering*, Prague.

- CHAVES, M.S., *Um estudo e apreciação sobre algoritmos de stemming*, In: IX Jornadas Iberoamericanas de Informática, Cartagena de Índias, Colômbia, Agosto, 2003.
- CLEARFOREST. Disponível em <<http://www.clearforest.com/>> Acesso em: 10 jan. 2006.
- COTTRILL, K., 1998, “Turning Competitive Intelligence into Business Knowledge,” *The Journal of Business Strategy*, v.19, n.4, pp.27-30.
- COWIE, J., LEHNERT, W., 1996, “Information Extraction”, *Communications of the ACM*, v. 39, pp. 81-91.
- CRAVEN, M., DIPASQUO, D., FREITAG, D., *et al.*, *Learning to extract symbolic knowledge from the world wide web*, *Proceeding of the fifteen National Conference on Artificial Intelligence*, pp. 509-516, 1998.
- DEBOYS, J., 2004, Decision pathways in patent searching and analysis, *World Patent Information*, 26, 83-90.
- ECLIPSE. Disponível em <<http://www.eclipse.org>> Acesso em: 10 jan. 2006.
- EVENSON, R., PUTTNAM, J., 1988, *The Yale – Canada patent flow concordance*. New Haven, CT: Economic Growth Center, Yale University.
- FABER, V., Clustering and the Continuous k-Means Algorithm. *Los Alamos Science*, 22:138-144, 1994.
- FATTORI M., PEDRAZZI G., TURRA R., 2003, Text Mining applied to patent mapping: a practical business case.
- FLETCHER, JM. “Quality and risk assessment in patent searching and analysis”. *Proceeding of the 4th International Chemical Information Meeting & Exhibition*, Montreux, pp. 19-21, out, 1992.
- FRAKES, W.B., BAEZA-YATES, R., *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, 1992.
- GANGULI, P., 2004, *Patents and patent information in 1979 and 2004: a perspective from India*, *World Patent Information*, 26, 61-62.
- GRANDSTRAND, O., 1999. *The economics and management of intellectual property: Toward intellectual capitalism*. UK: Edward Elgar.

- GRILLICHES, Z., 1990, "Patent statistics as economic indicators: A survey", *Journal of Economic Literature*, v. 28, pp. 1661-1707.
- GUPTA, V., PANGANNAYA, N., 2000. Carbon Nanotubes: Bibliometric analysis of patent. *World Patent Information*, v. 22, 185-189.
- HARMAN, S.D., "How effective is suffixing?", *Journal of the American Society for Information Science*, v. 42, n. 1, pp. 7-15, 1991.
- HIRSCHEY, M., RICHARDSON, V., 2001. Valuation effects of patent quality: A comparison for Japanese and US firms. *Pacific-Basin Finance Journal*, 9, 65-82.
- HOLL, B., JAFFE, A., TRAJTENBERG, M., 2000, *Market value and patent citation: A first look*. NBER Working Paper Series, Cambridge, MA.
- INMON, W.H., *Building the Operational Data Store*, ed. 2 United States of America, John William & Sons Inc., 1999.
- IDC - Insight Discoverer Clusterer – Developer's Guide, Temis Company, 2002.
- INPI. Disponível em: < <http://www.inpi.gov.br>>. Acesso em: 02 jan. 2006.
- KAHANER, L.1996. *Competitive Intelligence: How to Gather, Analyze, and Use Information to Move your Business to the Top*. New York, Simon and Schuster.
- KARKI, M., 1997. *Patent Citation Analysis: A policy analysis tool*, *World Patent Information*, vol. 19, n. 4, pp.269-272.
- KARKI MMS. Bibliometric analysis of patents: implications for R&D and industry, emerging trends in scientometrics. In: Nagpul PS et al., editor. New Delhi: Allied Publishers, 1999.
- KNOWLEDGE Management and Competitive Intelligence Made Clear. Disponível em: <<http://www.cipher-sys.com/>>. Acesso em: 02 jan. 2006.
- KOSTER, C.H.A, SEUTTER, M., BENEY, J., 2001, *Classifying patent applications with winnow*. In Proceedings Benelearn, Anwerpen.
- KOSTOFF, R.N., 2004, "Text Mining for Global Technology Watch", Office of Naval Research, August.
- KUDYBA, S, HOPTROFF, 2003, *Data Mining and Business Intelligence: A Guide to Productivity*, Idea Group Publishing, 2001.

- LARKEY, L.S., 1998, *Some Issues in the Automatic Classification of U.S. Patents*, Working Notes for the AAAI-98 Workshop on Learning for Text Categorization. 1998.
- LARKEY, L.S., 1999, *A Patent Search and Classification System*, Center for Intelligent Information Retrieval, Massachusetts.
- LAWRENCE, S., GILES, C.L., *Accessibility of Information on the web*. Nature, v. 400 pp.107-109, 1999.
- LOPES, M.C.S, 2004, *Mineração de Dados Textuais Utilizando Técnicas de Clustering para o Idioma Português*. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ.
- LIU, B., MA, Y., YU, P.S., 2001, “Discovering Unexpected Information from Your Competitors’ Web Sites”. In: *International Conference on Knowledge Discovery and Data Mining*, pp. 26-29, San Francisco, USA, Aug.
- LOVINS, J.B., *Development of a stemming algorithm*, Mechanical Translation and Computational Linguistics, vol. 11, pp. 22-31, 1968.
- MACQUEEN, J. B., *Some Methods for classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, vol. 1, pp. 281-297, 1967.
- MARINHO, L., GIRARDI, R., 2003, “Mineração na Web”, *Revista eletrônica de Iniciação Científica da Sociedade Brasileira de Computação*, vol.3, n.2.
- MENDELZON, A., MIHAILA, G., MILO, T., “Querying the World Wide Web”, *Journal of Digital Libraries*, 1(1):68-88, Abril, 1997.
- MICHEL, J., BETTELS, B., 2001, “Patent citation analysis: A closer look at the basic input data from patent search report”, *Scientometrics*, vol. 51, no. 1, pp. 185-201.
- MICROPATENT. Disponível em <<http://www.micropat.com>> Acesso em: 10 jan. 2006.
- MOGEE, M. 1991. Using patent data for technology analysis and planning. *Research-Technology Management*, 34, 43-49.
- MORAIS, Ednalva F. C.. *Inteligência Competitiva: Estratégias para Pequenas Empresas*. Brasília: UnB/CDT, 1999. 60 p.

- MOWERY, D., OXLEY, D., SILVERMAN, B., 1998. *Technological overlap and interfirm cooperation: Implications for the resource-based view of the firm*. *Research Policy*, 27, 507-523.
- NARIN, F., 1994. Patent bibliometrics. *Scientometrics*, 30(1), 147-155.
- NARIN, F., NOMA, E., 1987. Patents as indicators of corporate technological strength. *Research Policy*, 16, 143-155.
- ORENGO, M.V., HUYCK, C.R., 2001, “A Stemming Algorithm for The Portuguese Language”. In: *Proceedings of the SPIRE Conference*, pp. 13-15, Laguna de San Raphael, Chile, Nov.
- PACI, R., SASSU, A., USAI, S., 1997. International patenting and national technological specialization. *Technovation*, 17(1), 25-38.
- PCT - Patent International Treaty. Disponível em <<http://www.wipo.int/pct/es/treaty/about.htm>> Acesso em: 10 jan. 2006.
- PEREIRA, E. C. Monitoramento de normas e patentes como ferramenta para a inteligência competitiva. Curitiba., PR: TECPAR - Instituto de Tecnologia do Paraná, 2003.
- PIUG. Disponível em <<http://www.piug.org>> Acesso em: 02 jan. 2006.
- POYNDER, 1998. “Patent Information on the Internet,” *Business Information Review*, v.15, n.1, pp.58-67.
- PORTER, M.F., *An algorithm for suffix stripping*, Program, vol. 14, pp.130-137, 1980.
- REZENDE, S. O., OLIVEIRA, R.B.T., IMAMURA, C. Y., GONÇALVES, L.S.M. *Text Mining em Documentos de Patentes usando o Sistema Minador*. Proceedings of 21st Iberian Latin-American Congress on Computational Methods in Engineering (Data Mining Workshop), Rio de Janeiro, 2000.
- ROSS, K., SRIVASTAVA, D., 1997, “Fast Computation of sparse datacubes”. In: *Proceedings of 23th International Conference on Very Large Databases (VLDB97)*, pp. 116-125, Athens, Greece, Morgan Kaufmann, Ago.
- SALTON, G., *Automatic Text Processing*, Addison-Wesley, 1989.
- SALTON, G., WONG, A., YANG, C., “A vector space model for automatic indexing”, *Communications of the ACM*, v. 18, pp. 163-620, 1975.

- SCIME, A, *Web Mining: Applications and Techniques*. 2 ed. United States of America, Idea Group Publishing, 2005.
- SOCIETY of Competitive Intelligence Professionals. Disponível em: <<http://www.scip.org>>. Acesso em: 02 jan. 2006.
- STATSOFT - Statistica Software. Disponível em: <<http://www.statsoft.com/>>. Acesso em: 31 jan. 2006.
- SULLIVAN, D., *Document Warehousing and Text Mining*, 1 ed. John Wiley & Sons, New York, 2001.
- SUN Developer NetWork. Disponível em: <<http://java.sun.com>>. Acesso em: 15 jan. 2006.
- TARAPANOFF, k., *Inteligência Organizacional e Competitiva*. Brasília, Editora Universidade Brasília, 2001.
- TEMIS Text Intelligence. Disponível em <<http://www.temis.com/>> Acesso em: 10 jan. 2006.
- TIJSSSEN, R., 2001, *Global and domestic utilization of industrial relevance science: Patent citation analysis of science – technology interactions and knowledge flows*. Research Policy, vol. 30, pp. 35-54.
- TYSON, Kirk W. M. 1998, *The Complete Guide do Competitive Intelligence: gathering, analyzing, and using competitive intelligence*. Kirk Tyson Int. Ltd. Lisle, Chicago.
- UNDERWOOD, G., MAGLIO, P. BARRETT, R., 1998, “User-centered push for timely information delivery”, In Proc: WWW7.
- UNITED States Patent and Trademark Office. Disponível em <<http://www.uspto.gov>> Acesso em: 25 jan. 2006.
- VANTAGEPOINT. Disponível em <<http://www.thevantagepoint.com>> Acesso em: 10 jan. 2006.
- WISDOMAIN. Disponível em <<http://www.wisdomain.com>> Acesso em: 10 jan. 2006.
- YANG, Y., PEDERSEN, J.P., *A Comparative Study on Feature Selection in Text Categorization*, Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412-420, 1997.

YOON, B., PARK, Y, 2003 , “A text-mining-based patent network: Analytical tool for high-technology trend”, *The Journal of High Technology Management Research*, 15, Seoul, South Korea, 37-50, September.

ZANASI, A., 2005, *Text Mining and its Applications to Intelligence, CRM and Knowledge Management*. 1 ed. Great Britain, WIT Press.

ZANASI, A., Text Mining: the new competitive intelligence frontier. *In VST2001 Barcelona Conference Proceedings – IRIT*, Spain, 2001.