APLICAÇÃO DE FERRAMENTAS COMPUTACIONAIS PARA PROSPECÇÃO TECNOLÓGICA POR MINERAÇÃO DE DADOS NÃO-ESTRUTURADOS SOBRE PATENTES INDUSTRIAIS EM IDIOMAS INGLÊS

Cristiano José Mariotti Martins

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. (a) Beatriz de Souza Leite Pires Lima, D.Sc.

_____
Prof. Luís Landau, D.Sc.

_____
Prof. (a) Myrian Christina de Aragão Costa, D.Sc.

RIO DE JANEIRO, RJ - BRASIL
ABRIL DE 2008

MARTINS, CRISTIANO JOSÉ MARIOTTI

Aplicação de ferramentas computacionais para prospecção tecnológica por Mineração de Dados não-estruturados sobre patentes industriais em idioma inglês [Rio de Janeiro] 2008

XII, 191 p. 29,7cm (COPPE/UFRJ, M.Sc., Engenharia Civil, 2008)

Dissertação – Universidade Federal do Rio de Janeiro, COPPE

1. Mineração de Textos
2. Patentes Industriais
3. Inteligência Competitiva
4. Prospecção Tecnológica

I. COPPE/UFRJ II. Título (série)

AGRADECIMENTOS

Primeiramente, agradeço a DEUS, na certeza de que todas as nossas conquistas, glórias e vitórias em nossas vidas são graças, por ELE, concedidas.

Agradeço à minha mãe, Léa Tremontino Mariotti Martins que sempre me passou muito amor, muita força, muita tranqüilidade e muito incentivo, mesmo nos momentos em que pensava serem mais difíceis. Agradeço ao meu pai, Manuel Joaquim Pinto Martins, e aos meus avós paternos, José Joaquim Martins e Leonor Pinto, que em vida, sempre me deram amor, apoio e o carinho necessário para que eu pudesse evoluir como homem.

Agradeço à minha noiva e futura esposa com quem constituirei uma nova família no futuro, Andréia Balbi Lourenço, dentre outras coisas, por todo seu amor e compreensão nos momentos em que mais precisei.

Agradeço aos meus demais familiares e amigos não citados pelo fato de não querer cometer nenhuma injustiça com ninguém, mas que sabem, em seu interior, que contribuíram para meu enfrentamento desse estimulante desafio em minha vida.

Agradeço aos meus professores de Mestrado, cito Alexandre Gonçalves Evsukoff, Beatriz de Souza Leite Pires Lima, Luís Pereira Calôba, Myrian Christina de Aragão Costa, por todo o conhecimento transmitido a mim aula após aula e na certeza de que muito aprendi; e em especial, ao meu professor-orientador, Nélson Francisco Favilla Ebecken, pela paciência, tranqüilidade, confiança e conhecimentos transmitidos a mim durante os momentos em que mais precisei de seu auxílio; e ao Professor Luís Landau, pela atenção, orientação e receptividade dada a mim no momento de definição do trabalho e que precisei de seu auxílio.

Agradeço aos meus companheiros da jornada de dois anos de Mestrado, pois foram eles que me passaram força e perseverança para que pudéssemos chegar aos nossos

objetivos, além do auxílio oferecido em momentos de dúvidas, cito Adriana Aparício, Ângelo, Aretha, Bruno Vilela, Cláudio Czura, Daniel, Graziella Caputo, Ingrid, Leonardo Falcão, Luís Fernando e Renan Souza.

Agradeço, também, aos demais funcionários da Secretaria Acadêmica do Programa de Engenharia Civil e do Laboratório de In ard Inm(a)3.74(l)-2.16557874(m)-2.45995.17

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

APLICAÇÃO DE FERRAMENTAS COMPUTACIONAIS PARA PROSPECÇÃO TECNOLÓGICA POR MINERAÇÃO DE DADOS NÃO-ESTRUTURADOS SOBRE PATENTES INDUSTRIAIS EM IDIOMAS INGLÊS

Cristiano José Mariotti Martins

Abril / 2008

Orientador: Nelson Francisco Favilla Ebecken

Programa: Engenharia Civil

A presente pesquisa apresenta estudos com relação à exploração de patentes industriais norte-americanas ligadas a seis áreas da Indústria de Petróleo e Gás (em fase de crescimento no mercado) por técnicas de mineração de dados não-estruturados. Buscaram-se e recuperaram-se tais documentos no site da instituição *USPTO* através de uma aplicação computacional implementada para devidos fins. Como objetivos principais, a dissertação propõe descobrir padrões interessantes sobre o campo *ABSTRACT* das patentes utilizando-se duas diferentes plataformas desenvolvidas para geração de Inteligência Competitiva – *RapidMiner / YALE* e *PolyAnalyst*; e avaliar o desempenho desses dois *softwares* empregados ao final das análises. A pesquisa elaborada mostra ainda que a prospecção tecnológica por mineração de patentes, apesar de não ser tratada com a devida importância que lhe é merecida, torna-se de grande valia para as instituições, norteando-as para novas tendências em tecnologia, de forma a provê-las suporte para a tomada de decisões; além da elaboração de estratégias competitivas que permitam à instituição monitorar suas concorrentes e antever-se a elas nas inovações tecnológicas, garantindo sua vantagem competitiva.

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

APPLICATION OF COMPUTATIONAL TOOLS FOR TECHNOLOGICAL PROSPECTION BY NOT – STRUCTURALIZED DATA MINING ON INDUSTRIAL PATENTS IN ENGLISH LANGUAGES

Cristiano José Mariotti Martins

April / 2008

Advisors: Nelson Francisco Favilla Ebecken

Department: Civil Engineering

This research presents studies related to the exploration on North American industrial patents the six areas of the Industry of Oil and Gas (in phase of growth in the market) by not-structuralized data mining techniques. They had searched and they recovered such documents in the institution USPTO site using an implemented computational tool. As objective main, the dissertation considers to discover interesting standards on ABSTRACT patents field using itself two different platforms developed for generation of Competitive Intelligence - RapidMiner/YALE and PolyAnalyst; and to evaluate performance of these softwares to the end of the analyses. Moreover, the elaborated research shows that the technological prospection applied on patents mining, although not to be dealt with the had importance that is deserved to it, becomes of great value for the institutions, guiding them for new trends in technology, providing them support for helping the decisions; beyond the elaboration of competitive strategies that allow the institution to monitor its competitors and to be foreseen it in the technological innovations, guaranteeing its competitive advantage.

INDICE:

INDICE DE FIGURAS:

INDICE DE TABELAS:

1. Introdução

1.1. A sociedade da informação: Conceitos e Características

A sociedade da informação, em tempos contemporâneos, traz paradigmas ligados à economia, tais como produtividade, rapidez e qualidade, cria novos recursos para o desenvolvimento e exige de todos os envolvidos uma nova postura diante das mudanças sociais. Por conseqüência, as instituições (grandes, médias e pequenas empresas, organizações governamentais, associações, centros educacionais etc.) devem estar atentas, exigindo-se das mesmas uma competência suficiente no enfrentamento dessas mudanças.

Obter, gerar e aplicar conhecimentos inteligentes sobre seus negócios é fundamental para que a instituição se mantenha em competitiva evidência perante os concorrentes nos dias atuais. A sociedade da informação é caracterizada pela economia alicerçada sobre a informação e telemática[1] e assim se tornou devido à necessidade criada pelos constantes e acelerados avanços da tecnologia em função do fenômeno da globalização. Esses avanços tecnológicos se refletem em todos os três setores de nossa economia.

Nos dias atuais, compreende-se informação como "matéria-prima", ou seja, o insumo básico que desencadeia todo o processo de extração de conhecimentos que sustentam a competitividade de uma organização perante o mercado; comunicação como veículo responsável pela disseminação da informação; e tecnologias como infra-estruturas responsáveis por armazenar, processar e acessar a informação de maneira objetiva, rápida e eficaz.

[1] Telemática consiste no tratamento da informação suportada pelas novas tecnologias emergentes e aplicada em favor da comunicação (telefonia fixa e móvel, transmissão de sinais de TV e rádio, provedores de acesso à *Internet* entre outras aplicações) a fim de tratá-la com mais qualidade, rapidez e eficiência.

1.2. Globalização, *Internet* e obtenção de conhecimentos

A globalização, que tem a *Internet* como forte aliado de cooperação para esse crescimento desenfreado, rompeu as fronteiras, acelerou os acessos às informações e facilitou o estudo sobre novas tendências de mercado. Em contrapartida, exigiu das instituições uma postura cada vez mais organizada e proativa. Isso fez com que a empresa não somente estivesse atualizada às mudanças e inovações que acontecem na sociedade sob a qual ela está inserida, como também estivesse determinada à descoberta de novos caminhos que gerassem lucro e competitividade, de forma a analisar fatores críticos e aproveitar-se de suas concorrentes no antever das ações que pudessem acontecer.

Tudo isso exigiu das instituições uma compreensão mais apurada sobre seus concorrentes, sobre o que, como e quando investir, e a proverem-se de estratégias capazes de transformar a informação encontrada em uma rica fonte de conhecimento, de forma precisa, rápida e qualitativa, e com a garantia do retorno tangível e intangível de investimento.

1.3. O Porquê da Inteligência Competitiva

Socialmente, consideram-se a tecnologia de informação, de comunicação e a *Internet* os marcos divisores desta nova realidade, ou seja, quem a elas possui acesso tem maiores chances de levar uma grande vantagem sobre um segundo grupo que fica relegado à condição de excluídos e à condição de dominados digitalmente.

Sabe-se, porém, que não basta somente ter acesso às informações disponibilizadas na grande rede mundial. Considera-se que a maior parte das informações geradas e publicadas na *Internet* é de caráter não-produtivo, e nada acrescenta nas decisões de

negócios de uma empresa, além de tornar o processo de análise, interpretação e processamento dos dados lento, trabalhoso e difícil, o que não é interessante para quem busca sempre estar à frente de seus concorrentes.

Portanto, torna-se necessário a cada dia mais um tratamento inteligente sobre as informações e sobre os documentos eletrônicos publicados, de forma a se desprezar os conteúdos inócuos e a se preservar a essência, transformando as informações, realmente, em conhecimentos válidos e produtivos a um especialista, para que esse conhecimento gerado seja capaz de indicar novas tendências de investimento de mercado ao longo do tempo e prover suporte à tomada de decisões de forma racional e concisa.

Dessa forma, a instituição tende a ganhar respeitabilidade e competitividade, de forma a valer-se do conhecimento extraído para aplicá-lo sobre novos produtos, novos serviços, melhorando o panorama organizacional interno e externo, dentre outros benefícios tangíveis e intangíveis.

Com isso, as técnicas computacionais são empregadas para que se haja a busca, recuperação, seleção e extração da informação de maneira inteligente. Justamente a partir das dificuldades encontradas no enfrentamento de questões ligadas à extração de conhecimentos, surgiu a área de Descoberta de Conhecimentos em Bases de Dados (*Knowledge Discovery in Database – KDD*) e, conseqüentemente, para os documentos eletrônicos sobre os quais se necessita da essência do conhecimento implícito, surgiu a área de Descoberta de Conhecimentos em Textos (*Knowledge Discovery in Texts – KDT*). A debruçar-se sobre a composição de ambas, verifica-se que essas áreas possuem aparato tecnológico interligado com disciplinas de ciências exatas, como a Probabilidade, a Estatística e a Inteligência Artificial e cujos recursos gerados podem ser aplicados na prospecção tecnológica e na obtenção de vantagens competitivas de uma instituição.

1.4. Motivação

A partir da Revolução Industrial, a aplicação bem-sucedida de conhecimentos científicos para produção de tecnologias passou a ocorrer em grande escala. Desde então, o conhecimento científico deixou de ser um bem permanente cultural e se tornou um insumo para o sucesso econômico. Surgiu, então, o conceito de “Propriedade Intelectual”.

Os objetos de Propriedade Intelectual são as criações oriundas da mente humana. As patentes são as formas de proteção legal e temporária de uma invenção em formato de documento escrito, e são concedidas ao inventor ou ao seu titular, dando-lhe o direito único de exploração e impedindo terceiros de explorarem seus benefícios.

Com o crescimento tecnológico alavancado em virtude das razões já comentadas nas seções anteriores, surgem novos produtos, marcas, idéias e, conseqüentemente, novas patentes são depositadas a cada ano, com detalhes sobre os produtos, de forma a garantir o depositante de usufruir o direito de exploração de sua invenção ou de qualquer outro produto com características similares detalhadas na patente.

A patente não é só uma proteção legal contra crimes de violação à Propriedade Intelectual. Trata-se de um bem econômico e uma rica fonte de informação tecnológica que deve ser utilizada para solucionar problemas técnicos e na realização de pesquisas comerciais sobre o mercado, garantido uma ótima Prospecção Tecnológica, caso sejam bem exploradas. Tal afirmação é justificável ao se analisar os documentos de patentes e se descobrir que nelas estão contidas as informações detalhadas e mais recentes em relação ao estado da técnica de diversas áreas do desenvolvimento humano.

Uma pesquisa bem elaborada por conhecimentos sobre uma base de patentes evita que esforços sejam colocados sobre o desenvolvimento de tecnologias já existentes; permite identificar novas tecnologias emergentes ou alternativas; fornece embasamento

para aplicações comerciais indicando, por exemplo, melhores caminhos para compra de tecnologia; permite verificação da disponibilidade de tecnologia em um país; permite o monitoramento de tecnologias concorrentes; entre outras vantagens competitivas.

Neste caminho, os recursos tecnológicos se tornam elementares para o tratamento de informações contidas em uma base de dados de patentes sobre um determinado assunto, de forma a transformar essas informações encontradas em conhecimentos valiosos para a abertura de novos horizontes comerciais que facilitem à tomada de decisão de negócios de forma inteligente.

A considerar a *Internet* como ampla para depósito de informações das mais diversas, porém não tão rico em conhecimentos realmente relevantes, podem-se encontrar domínios de instituições que concedem patentes e que as publicam em formato de documento eletrônico (arquivo de texto) ou, até mesmo, em formato de hipertexto em seu ambiente, e sobre os quais podem ser extraídas, de forma a utilizar-se, para tanto, de aplicativos bastante úteis para Mineração de Textos.

Esta metodologia impulsionou o estudo e a pesquisa sobre processos de pré-processamento e transformação de dados não-estruturados, bem como a implementação de novas ferramentas capazes de manipulá-los. Utilizam-se, com isso, as técnicas e os algoritmos já antes conhecidos em Mineração de Dados a fim de que se possam extrair conhecimentos interessantes sobre seus conteúdos pré-processados.

Dessa forma, essas novas ferramentas computacionais oriundas da Mineração de Textos podem auxiliar o processo de garimpagem para Prospecção Tecnológica.

1.5. Proposta de Pesquisa Acadêmica

A presente dissertação propõe a utilização de recursos computacionais para prospecção tecnológica por extração de conhecimentos de grande valia sobre documentos de propriedade intelectual no idioma inglês, de forma a verificar-se a estrutura do documento e elaborar pesquisas sobre um de seus depósitos de dados hospedados na *Internet*. Consideram-se os dados recuperados nas bases como documentos eletrônicos não-estruturados e, portanto, este trabalho de pesquisa aplica técnicas advindas da metodologia de Mineração de Textos (ou de Dados Não-Estruturados).

A Mineração de Textos possui vários *softwares* desenvolvidos no mercado para a execução de suas tarefas provenientes. Portanto, para essa pesquisa, propõe-se a utilização de dois *softwares* distintos encontrados no mercado para tal finalidade. Um desses *softwares* é o *RapidMiner* / *YALE* (*MIERSWA et al*., 2006), um *framework* em versão mais atualizada e desenvolvida sob a plataforma da linguagem de programação Java. Outro software escolhido é o *PolyAnalyst*, desenvolvido e comercializado pela empresa *Megaputer Intelligence*. Propõe-se, com isso, realizar as tarefas pertinentes às etapas de pré-processamento e de processamento em si de Mineração de Textos com cada aplicativo, confrontar os resultados obtidos e, por fim, comparar o desempenho de cada um deles.

Haja vista que as patentes industriais possuem uma estrutura própria com subdivisões específicas ao longo de seu documento textual, a proposta é realizar os experimentos decorrentes da tarefa de extração de conhecimentos proposta sobre o campo *ABTRACT* (RESUMO), extraindo-se tecnologias implícitas em seu corpo textual.

1.6. Estrutura da dissertação

Este capítulo teve por objetivo fazer uma introdução sobre os conceitos de sociedade da informação e o que nos levou a chegar ao estágio de desenvolvimento tecnológico no qual nos encontramos hoje, além de abordar conceitos introdutórios e pertinentes sobre Inteligência Competitiva, Mineração de Textos e Patentes Industriais. Procurou-se transmitir, além desses conceitos, a importância das patentes nos processos de Inteligência de Negócios, justificando-se, dessa forma, a motivação para se realizar uma pesquisa cientifica sobre esses tipos de documentos.

Para se ter um melhor entendimento a respeito da relevância da pesquisa, disserta-se sobre Inteligência Competitiva, seus conceitos e sua fundamental importância para os negócios dentro da sociedade da informação no Capítulo 2.

O Capítulo 3 apresenta uma visão mais profunda sobre as patentes industriais nos Estados Unidos (EUA), assim como uma análise sobre sua estrutura e sobre seus campos específicos. O capítulo apresenta uma descrição sobre Mineração de Texto, seus conceitos, etapas e algoritmos utilizados. Além disso, este capítulo comenta sobre o depósito de patentes hospedado na *Internet* escolhido para a pesquisa, apresenta um elo entre Mineração de Textos e Patentes Industriais dos EUA, e disserta de que forma a Mineração de Textos pode ser empregada para a extração de conhecimentos sobre as patentes.

O Capítulo 4 apresenta uma breve descrição sobre algumas das principais ferramentas existentes e que podem ser utilizadas para KDT, mais precisamente, para a extração de conhecimentos sobre documentos de patentes. Aborda, de maneira mais enfática, o ambiente computacional utilizado para esta pesquisa.

O Capítulo 5 disserta sobre o detalhamento dos estudos de casos em torno do processo de mineração de textos, no qual submetem-se as bases com as respectivas

patentes. Uma vez que os experimentos foram realizados com a utilização das ferramentas escolhidas, de forma complementar, este capítulo disserta sobre os resultados obtidos por ambas, de forma a confrontá-los entre si, e destacando os pontos essenciais sobre as tarefas executadas.

Finalmente, o capítulo 6 apresenta as considerações finais e recomendações para futuros trabalhos de pesquisa envolvendo patentes industriais e as ferramentas empregadas - *RapidMiner / YALE* e *PolyAnalyst* – deixadas como herança pela presente dissertação.

2. Inteligência Competitiva nas Organizações

2.1. O Ambiente de negócios: Panorama Atual e Tendências

Haja vista que o mundo contemporâneo passa por uma fase de grandes transformações econômicas, políticas e tecnológicas que influenciam todos os níveis da sociedade, é evidente que as organizações buscam alternativas para poder sobreviver neste competitivo mercado de negócios. A globalização fez com que as organizações se preocupassem em colocar no mercado produtos e serviços com o máximo de qualidade e rapidez possíveis a preços acessíveis para o consumidor, que por sua vez, tornou-se mais exigente no retorno sobre seus investimentos. “Devi

decisões, bem como indique novos caminhos a fim de que possa antever uma tecnologia ou um produto novo, de modo a obter ganhos sobre a severa concorrência.

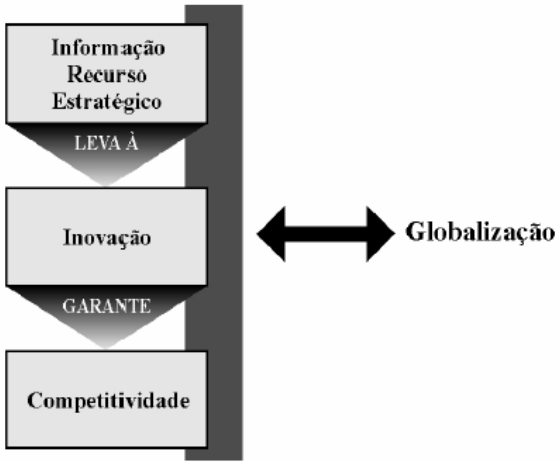

Figura 2-1 – Síntese do novo paradigma organizacional [2]

Para tanto, as organizações precisam, em si, possuir no mercado de negócios uma vantagem competitiva que seja enxergada pelos clientes. Um modelo conceitual de Inteligência Competitiva diz que para se tirar proveito dentre os consumidores, uma organização precisa estar em permanente vigília sobre o fluxo de processos que passam por ela (GOMES e BRAGA, 2004).

Os pesquisadores da área de Gestão e Inteligência Estratégica da Informação comungam da teoria de que estar sob vigília significa que a organização mantém-se em constante e permanente processo de busca por informações pertinentes que lhe permitam a decisão sobre a adoção de estratégias competitivas para seu negócio através da análise em torno do ambiente.

Segundo Fachinelli (2004), a vigília tecnológica define-se como a observação e a análise do ambiente científico (pesquisas em laboratórios, dados teóricos), técnico (patentes), tecnológico (processos e montagens de unidades), técnico-econômico (capacidades-resultados) e econômico (estatísticas setoriais ou macroeconômicas), seguida

[2] Fonte: (SILVA, 2003, p.116)

da difusão bem direcionada aos responsáveis, e das informações selecionadas e tratadas, que sejam úteis na tomada de decisões estratégicas por parte da organização.

Ainda segundo a autora, a vigília tecnológica foi raiz da vigília estratégica e de outros tipos de vigília mais específicos, como a vigília científica ou a vigília concorrencial, sob a qual a organização mantém-se. Ainda a basear-se sobre seus estudos, Fachinelli (2004) afirma que “a vigília tecnológica situa-se no movimento de observação de diversos ambientes da empresa e também no ambiente constituído pela defesa do patrimônio industrial, científico e técnico”. Dessa forma, a tecnologia tornou-se um fator de mudanças sociais, e as organizações são obrigadas, hoje para sobreviverem, a buscar e a dominar as novas tecnologias.

Neste atual panorama e a respaldar-se sobre a citação anterior, pode-se afirmar que com a globalização e os recursos de comunicação cada vez mais expansivos, as organizações aumentam suas chances de se colocarem no mercado, pois podem romper fronteiras, se colocar em diversos lugares ou países, terem acesso a diversificados hábitos e valores culturais, dentre outros parâmetros que, até há décadas atrás, eram considerados como utopias. O desafio maior é lidar com incertezas, instabilidades e turbulência que a atravessam, de forma inteligente e segura de sua tomada de decisões.

Segundo Gomes e Braga (2004), pode-se dizer que, além de parâmetros de satisfação como redução de custos e diferenciação de produtos e serviços, as organizações precisam estar sempre sob vigília de seus ambientes externos e internos que a compõem.

O ambiente interno à organização é composto por seus pontos fortes e pontos fracos, ao passo que o ambiente externo à organização é composto pelas oportunidades e ameaças. O ambiente interno pode ser controlado pelos dirigentes da organização, já que ele é o resultado de estratégias de atuação definidas por eles mesmos. Desta forma, quando se percebe um ponto forte na análise do negócio da organização, deve-se ressaltá-lo ainda

mais. Em contrapartida, quando se percebe um ponto fraco, deve-se atuar para controlá-lo ou, pelo menos, minimizar seu efeito.

Já o ambiente externo está totalmente fora do controle da organização. Isso não significa que não seja útil conhecê-lo, muito pelo contrário. Apesar de não poder controlá-lo, pode-se monitorá-lo e procurar aproveitar as oportunidades da maneira mais ágil e evitar as ameaças enquanto for possível. Diversos fatores externos à organização podem afetar seu desempenho. E as mudanças no ambiente externo podem representar oportunidades ou ameaças ao desenvolvimento do plano estratégico de qualquer organização. A avaliação do ambiente externo costuma ser desmembrada em duas partes: os fatores macroambientais, entre os quais podemos citar questões demográficas, econômicas, tecnológicas, políticas, legais, etc.; e os fatores microambientais – entre os quais podemos citar os beneficiários, suas famílias, as organizações congêneres, os principais parceiros econômicos etc.

Neste contexto apresentado, Kotler (2002) afirma:

> Para se garantir uma vantagem competitiva sobre os concorrentes apesar das constantes mudanças no mercado, as empresas precisam se antecipar às mudanças, enxergar as oportunidades e observar com olhos críticos o panorama sócio-econômico. Para se fazer isso, devem-se monitorar os fluxos de informações de negócios que envolvem a organização. (p.28)

Sabe-se, contudo, que para monitorar todo esse fluxo de informação de negócios, devem-se analisar o ambiente interno e o ambiente externo da organização, e conseqüentemente, interagir com todos os fatores e variáveis que a afetam diretamente. Possuir um bom modelo de gestão estratégica é uma fonte vital para a empresa nos dias atuais, assim como afirma Dou (1995).

Sabe-se que uma análise estratégica dentro de uma organização demanda uma atividade trabalhosa. Tal fato justifica-se à medida que se pensa que os ambientes externos e internos da organização, por si só, geram grande carga de dados, e as organizações,

precisam tratá-las de forma a aproveitar somente o verdadeiro conhecimento, em síntese, o que servirá de suporte para as tomadas de decisões estratégicas.

2.2. Dados, Informação e Conhecimento

Antes de se comentar sobre o processo de Inteligência Competitiva, sua procedência e sua relevância, é interessante que se saiba distinguir os significados de cada termo que compõe o trinômio “dados, informação e conhecimento”, uma vez que no mundo dos negócios analisados de forma inteligente, possuir muitos dados à disposição não significa que o cliente ou o especialista, por exemplo, estarão bem providos de

conhecimentos estruturados, estruturáveis e não-estruturados para o negócio são ações que contribuem diretamente para o desenvolvimento do processo de Inteligência Competitiva dentro das organizações.

Para Miranda (1999, p.285), dados são “um conjunto de registros qualitativos ou quantitativos conhecidos que, quando organizado, agrupado, categorizado e padronizado

tocante às suas significâncias, informação e conhecimento possuem suas diferenças na prática, assim cita Lastres e Albagli (1999):

> Informação e conhecimento estão correlacionados mas não são sinônimos. Também é necessário distinguir dois tipos de conhecimentos: os conhecimentos codificáveis - que, transformados em informações, podem ser reproduzidos, estocados, transferidos, adquiridos, comercializados etc. - e os conhecimentos tácitos. Para estes a transformação em sinais ou códigos é extremamente difícil já que sua natureza está associada a processos de aprendizado, totalmente dependentes de contextos e formas de interação sociais específicas (p.30).

Ao contrário de Lastres e Albagli (1999), Miranda (1999) é um pouco mais detalhista em suas definições de conhecimento, de forma em que ele o distingue em três diferentes tipos: conhecimento explícito, que segundo o autor, é o conjunto de informações lícitas em algum suporte (livros, documento etc.), de forma a caracterizar o saber disponível sobre tema específico; conhecimento tácito que segundo o autor, é o acúmulo de saber prático sobre um determinado assunto e que agrega convicções, crenças, sentimentos, emoções e outros fatores ligados à experiência e à personalidade de quem detêm; e conhecimento estratégico que segundo o autor, é a combinação de conhecimento explícito e tácito formado a partir das informações de acompanhamento, agregando-se o conhecimento de especialistas.

2.3. Inteligência Competitiva

A Inteligência Competitiva (IC) é uma área que busca suprir as necessidades de informação estratégica de uma empresa, de forma a garantir sua efetividade decisória operacional e sua competitividade ao longo do tempo, através do incremento da capacidade de implementação de otimização de produtos, processos e do ambiente da organização em si. De acordo com Miller (1997), denomina-se IC “um processo de coleta, análise e

disseminação da inteligência relevante, específica, no momento adequado". Giesbrecht (2000) afirma que IC é um: "radar que proporciona à organização o conhecimento das oportunidades e das ameaças identificadas no ambiente, que poderão instruir suas tomadas de decisões, visando à conquista de vantagem competitiva". Por sua vez, Kahaner (1996) define-a como "um processo de coleta sistemática e ética de informações sobre as atividades de seus concorrentes e sobre as tendências gerais dos ambientes de negócios".

De forma a contemplar-se todas as definições apresentadas, Gomes e Braga (2004) define IC como "um processo ético de identificação, coleta, tratamento, análise e disseminação da informação estratégica para a organização, viabilizando seu uso no processo decisório". O processo de Inteligência Competitiva possui como intuito reverter um quadro desfavorável que ainda paira em grande parte das organizações, que trabalham com uma carga bruta e em demasia de dados irrelevantes, pequena quantidade de informações com valores relevantes agregados e muito pouca inteligência contida para tomadas de decisões. Com isso e tendo a necessidade de uma redução maior em custo e tempo de análises, segundo Tyson (1998; KAHANER 1996 apud CAPUTO, 2006), o processo de IC ganha importância cada vez maior dentro das organizações, de forma a tornar-se uma ferramenta de suporte indispensável em diversos níveis organizacionais, tais como planejamento estratégico, marketing, gestão de conhecimentos entre outros.

2.3.1. Presença das T.I no Processo de Inteligência Competitiva

No âmbito de desenvolvimento atual, diversas tecnologias de informação (TI) úteis existem para auxiliar o processo de IC dentro de uma organização. Dentre as quais, podemos citar, além dos sistemas de informação dos mais diversos tipos, a GED (Gestão

Eletrônica de Documentos), as tecnologias de integração em redes de computadores (*Internet*, Intranet e Extranet), dentre outras.

Pode-se dizer que a TI apóia todas as etapas de um processo relativo à IC, desde a fase de identificação das necessidades de informação até a avaliação dos produtos, de forma a organizar o fluxo de informações contidos neste ciclo e os auxilia nos principais objetivos de um Sistema de Inteligência Competitiva (SIC), que segundo Gomes e Braga (2004) são: mencionar possíveis oportunidades e ameaças, suportar as tomadas de decisões estratégicas, monitorar e debruçar-se sobre os desempenhos dos concorrentes, as tendências políticas, econômicas e sociais, além de apoiar o planejamento e os processos de gestão estratégica. A gestão estratégica, por sua vez, permite que o tomador de decisões dentro da empresa monitore as tendências temporais de mercado e à evolução dos concorrentes, de forma a antecipar-se às mesmas.

O ciclo de inteligência resume-se, basicamente, a partir de quatro fases básicas que podem ser representadas na figura 2-3:

1. **Planejamento e Direcionamento:** é a fase inicial do ciclo de inteligência, de

e necessárias para todo o ciclo, posteriormente a busca pelas origens de tais informações, seus recolhimentos nas fontes e o posterior armazenamento. Inicialmente, concentra-se a busca pela informação nas fontes externas, apesar de boa parte das informações que se encontram no ambiente externo à organização já se concentrarem dentro da própria. A partir de recursos textuais tais como e-mail, páginas *web*, listas de discussão entre outros*;* acessa-se uma quantidade grande de informações sobre o ambiente competitivo. Na subfase da coletânea de dados e valendo-se dos recursos que a *Internet* propicia, vale a busca rápida de forma a se utilizar *sites* e ferramentas de buscas rápidas existentes, que pode se dizer que é a forma mais adequada de se coletar informações, pelo menos em princípio. Ainda assim, mesmo com a melhor tecnologia em posse, essas buscam requerem habilidade, percepção, paciência, perseverança dentre outras virtudes que um profissional de informação ou um pesquisador precisa. Vale ainda o diálogo com consumidores, fornecedores e agregados, e participações em congressos e eventos pertinentes;

3. **Análise da Informação e Formatação:** As TI a cada dia melhoram o desempenho de seus produtos lançados no mercado e decaem os preços de algumas ferramentas conforme as novidades e os *releases* são lançados, o que deixa o custo de aquisição mais acessível. Tornam-se padrões os *softwares* que suportam a análise de tendências e prospecção tecnológica. Tais ferramentas necessitam de entrada e alimentação de informações em nível bem específico, e fazem parte com freqüência de uma rede colaborativa de trabalho entre analistas de informações. Caputo (2006) afirma que o conteúdo e as fontes das informações dependerão das técnicas e métodos utilizados para análise, que podem ser das mais diversas, dentre elas, benchmarking, fatores

críticos de sucesso, mineração de dados estruturados e não-estruturados, perfil dos concorrentes, análise financeira dentre outros. Gomes e Braga (2004) ressaltam que uma ferramenta somente terá utilidade dentro de um processo de análise caso ofereça facilidades que complementem ou que suportem a análise de informação. Dentre todos estes motivos levantados, pode-se concluir que a análise de informações é a etapa mais trabalhosa de um ciclo de inteligência, e esse nível de dificuldade tende a aumentar caso seja necessário a busca e interpretação de modelos e a produção de diferentes cenários;

4. **Disseminação e Utilização:** Nessa etapa, visa-se apoiar a distribuição dos produtos gerados pela equipe de inteligência, tais como avisos, boletins e perfis. Pode ser feito através da distribuição simples dos mesmos anexados a uma mensagem eletrônica (*e-mail*) ou acessando-os pela *Internet*. Nessa etapa, a ferramenta é responsável pela entrega da informação sem, no entanto, exercer o poder de convencimento sobre o tomador de decisão dentro da empresa.

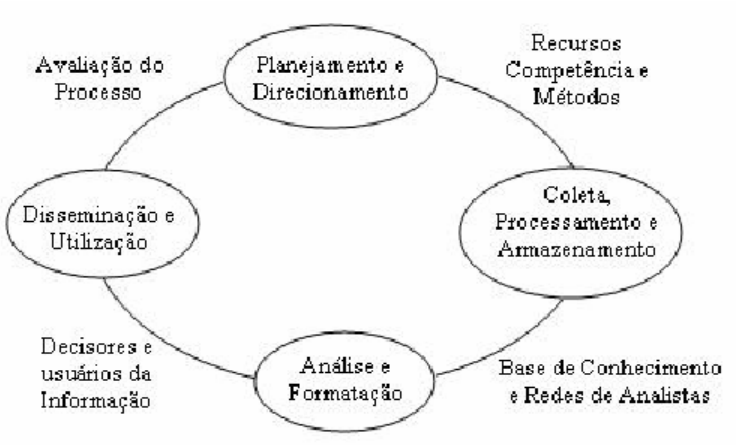

Figura 2-3 – Ciclo de vida de Inteligência Competitiva[3]

[3] FONTE: Caputo (2006, p. 7)

2.3.2. Tópicos Complementares sobre Inteligência Competitiva

Segundo Knowledge (2006), a geração de inteligência é realizada pelos seres humanos de fato, com auxílio de sistemas bem estruturados e capazes de apresentar métodos e técnicas, bem como proporcionar a coleta e a disseminação da inteligência.

O aumento da competitividade no mercado transita a uma velocidade muito grande em todas as fases do ciclo de inteligência (já citadas na seção anterior) para as tomadas de decisões. Isso leva ao aumento das demandas por SIC, e isso leva o papel das TI a tornarem-se fundamentais para o sucesso dos projetos. Além disso, aumenta a exigência de seus usuários, e isso faz concluir que as TI para um SIC tendem ao crescimento de importância constante e em um ritmo acelerado.

No entanto, afirma-se que o ponto central para o sucesso de um sistema é a qualidade dos profissionais envolvidos no processo, de forma que os sistemas em si são os suportes tecnológicos necessários.

2.3.3. Ética aplicada à Inteligência Competitiva

É muito importante que os profissionais envolvidos em Sistemas de Inteligência Competitiva estejam conscientes de que sua prática seja realizada da forma mais ética e transparente possível. Haja vista que a maioria das informações necessárias é de caráter público, sua obtenção deve ser feita com certas precauções que vise resguardar a moral e a ética da organização. Portanto, Gomes e Braga (2004) citam alguns exemplos de medidas que devem ser adotadas no tocante à utilização de dados ou informações pertencentes a outras fontes:

- O não uso de dados ou informações confidencias e sigilosas, ou sem a devida autorização do seu responsável legal;
- A não obtenção de informações confidenciais sob alegações de caráter duvidoso ou mentiroso;
- A não obtenção de informações que coloquem em risco o caráter privado de uma pessoa ou organização;
- O não uso de dados ou informações que violem leis antitrustes.

Assim sendo, a preocupação com a ética é fundamental, pois um deslize pode colocar em risco a imagem de uma organização ou profissional. Logo, a definição de padrões éticos de comportamento para os funcionários de uma organização é fundamental para o ativo de uma corporação e com isso, avaliam-se suas credibilidade e conduta dentro do mercado.

3. Patentes como instrumentos de proteção à Propriedade Intelectual

3.1. Introdução à Propriedade Intelectual

O conceito de Propriedade Intelectual surgiu a partir da Revolução Industrial, época em que a aplicação bem-sucedida de conhecimentos de natureza científica para produção de tecnologia em grande escala na indústria tomou grandes proporções. Segundo Goulart *et al.* (2005), estimam-se mais de 80% dos conhecimentos científicos e tecnológicos gerados em âmbito mundial foram gerados após a Segunda Grande Guerra Mundial, e de forma a se manter tais dimensões de crescimento, 50% dos conhecimentos convertidos em bens materiais que estaremos utilizando no mundo nos próximos dez anos ainda sequer foram inventados.

A Propriedade Intelectual agrega a proteção das criações advindas da inteligência humana, sejam elas de natureza material ou bens não-materiais apropriáveis. Segundo a classificação tradicional, a Propriedade Intelectual agrupa dois ramos distintos: a Propriedade Industrial, que protege as criações inventivas orientadas para a indústria como marcas e patentes; e os Direitos Autorais, de forma a tutelar a atividade criativa orientada para o ambiente cultural.

3.2. Sobre as Patentes

Silva e Carvalho (2004, apud GOULART *et al.*, 2005) definem patentes como tipos de proteção legal de caráter temporário concedido pelo Estado ao inventor ou ao seu titular, de forma que somente os beneficiários pela concessão sejam os únicos legalmente reconhecidos com o direito de usufruir os retornos tangíveis ou intangíveis a partir das

idéias registradas e lavradas por escrito em seus conteúdos. Dessa forma, vedam-se quaisquer direitos de pessoas não - legalmente competentes se utilizarem, produzirem ou realizarem quaisquer atividades com o bem protegido sem sua devida autorização.

Existem dois tipos de patentes: as Patentes de Invenção, que são concedidas a um bem tecnológico, seja ele um produto ou um processo de produção, encarado no mercado como uma novidade em relação à transformação do estado da técnica em termos de qualidade e aplicação às atividades industriais; e as patentes de Modelo de Utilidade, que são concedidas ao objeto de aplicação prática (ou à sua parte) que apresenta novo formato, disposição e que obtenha como resultado a melhoria funcional em seu uso ou fabricação.

Segundo consta no Artigo Nº. 40 da Lei de Propriedade Intelectual Nº. 9.279, "a Patente de Invenção vigorará pelo prazo de vinte anos e a patente de Modelo de Utilidade pelo prazo de quinze anos a serem contados a partir da data do depósito dos documentos". Uma patente possui validade somente no país no qual foi feito o depósito. Assim, para que se obtenha validade internacional, a patente deve ser depositada na base de patentes internacionais, e isso é o que reza o Tratado de Cooperação de Patentes – PCT (2006), firmado em 19 de junho de 1970 em *Washington* (EUA), com o intuito de desenvolver o sistema de patentes e de transferência de tecnologia. O acordo prevê, basicamente, uma rede de cooperação entre os países industrializados e os países em desenvolvimento. Dessa forma, a patente passa a valer em todos os mais de cem países signatários ao tratado.

3.2.1. Importância das Patentes para Prospecção Tecnológica

Os documentos de patentes podem ser considerados como um amplo recurso de conhecimento de natureza científica e comercial, portanto, são grandes fontes de informações a respeito de detalhes de natureza tecnológica. Comprova-se tal afirmação por

Oliveira *et al.*, (2005), pois afirma: “uma das formas de se medir o desenvolvimento de um país está diretamente ligada ao número de patentes concedidas aos seus nacionais em outros países, especialmente os desenvolvidos”. Ainda em seus estudos, os autores citam algumas vantagens consideráveis que os documentos de patentes possuem em comparação às outras fontes de propriedade intelectual. Dentre elas, o fato de que os documentos de patentes divulgam informações de forma mais rápida em comparação às outras fontes, pois na maioria dos países os documentos são publicados antes de sua concessão, e dessa forma, a essência do que se tem de mais novo em termos de tecnologia chega ao conhecimento geral de forma mais rápida. Além disso, as patentes são detentoras de informações de publicação exclusiva a ela somente, segundo um levantamento feito por Alfred Marmor (1979) e publicado na revista *World Patent Information*, o que a torna o instrumento principal para extração de conhecimentos com relação ao estado da arte da tecnologia, pesquisas e comprovações científicas.

Além de ser uma proteção legal à propriedade intelectual e um bem econômico, a patente é um uma fonte de informação tecnológica e que pode ser utilizada para solucionar problemas técnicos e nas áreas de pesquisas acadêmicas (OLIVEIRA *et al*., 2005).

No Brasil, as discussões em torno da importância das atividades de pesquisa científica e tecnológica concentram-se no campo acadêmico principalmente, de forma que uma parte expressiva das atividades no tocante às pesquisas e desenvolvimento ocorre em instituições do governo, com participação ainda diminuta do setor de produção. Portanto, há ainda muito que fazer em pró do avanço tecnológico, tendo como prova uma pesquisa[4] feita pela ONU e publicada em 2001, que mostra o Brasil qualificado na 43ª posição em termos de Índice de Avanço Tecnológico, atrás de países vizinhos como Argentina e Chile, por exemplo.

[4] Fonte: Revista Veja publicada em 18/07/2001 e disponível em http://veja.abril.com.br/180701

Outra fonte de pesquisa que exemplifica a necessidade de uma exploração ainda mais acelerada em termos de avanço de prospecção de tecnologia em território nacional é o relatório[5] exibido pelo Projeto de Digitalização Acelerada da Federação das Indústrias do

3.2.2. Sistemas de Classificação de Patentes

A classificação das patentes surgiu da necessidade de tratar de uma grande quantidade de documentos de patentes. A classificação é usada para duas tarefas diferentes:

a. **A classificação de um documento de patente por seu conteúdo técnico**: devido à grande quantidade de informações de patentes que são armazenadas nos escritórios de patentes, o sistema de classificação foi introduzido para tornar o depósito mais fácil e para garantir um acesso rápido aos documentos. As invenções são classificadas de acordo com os campos da indústria, da técnica ou da atividade humana em relação às quais são relevantes caracteristicamente. Esse enfoque é o comumente designado de orientação industrial, e a antiga Classificação de Patentes Alemã, que exerceu alguma influência no Sistema de Classificação Internacional de Patentes, utilizou este enfoque.

b. **Para sua utilização como uma ferramenta de pesquisa**: os sistemas de classificação das patentes são usados para pesquisas, tais como pesquisas do estado da arte, pesquisas de contrafação, pesquisas de novidades, pesquisas de validade entre outras. Em virtude da necessidade em se facilitar ao máximo o processo de busca e de recuperação da informação, as invenções são classificadas de acordo com as funções para as quais são pertinentes caracteristicamente. Esse enfoque é comumente designado de orientação segundo a função e tem como exemplo o Sistema de Classificação de Patentes dos Estados Unidos da América (*United States Patents Classification - USPC*).

3.2.2.1. Sistema de Classificação Internacional de Patentes (CIP)

A Classificação Internacional de Patentes (CIP) é uma classificação especial utilizada internacionalmente para indexação de documentos de Patentes de Invenção e Modelo de Utilidade. É um instrumento que possibilita a busca e recuperação das informações tecnológicas e legais relativas ao conteúdo da patente de forma mais rápida e objetiva. A CIP é utilizada por cerca de 70 (setenta) países e 3 (três) Administrações Regionais e pela Secretaria Internacional da Organização da Propriedade Intelectual. Atualmente, a CIP está em sua edição 2007.1 desde 01/01/2007, e é revisada periodicamente pelo menos a cada cinco anos pelos países membros da Organização Mundial de Propriedade Intelectual, e publicada em CDROM, podendo também ser acessada nos sites referentes ao Instituto Nacional de Propriedade Industrial (INPI) e ao *World Intelectual Property Organization (WIPO).*

Atualmente, a Classificação Internacional de Patentes divide hierarquicamente a técnica em 8 (oito) setores principais, com 64.000 (sessenta e quatro mil) subdivisões. Cada subdivisão tem um símbolo composto de algarismos arábicos e de letras do alfabeto latino. Este sistema de classificação ainda conta com 21 subseções, 120 classes e 628 subclasses. Os 8 (oito) setores principais são denominados de seções, conforme mostra a tabela 3-1.

O símbolo completo desse sistema de classificação é composto por símbolos da Seção (conforme tabela 3-1), Classe (número constituído por dois algarismos indo - arábicos), Subclasse (representada por letra maiúscula de nosso alfabeto), Grupo e mais o Subgrupo. A tabela 3-2 mostra uma análise do significado da classificação internacional da patente C07D 401/12 como exemplo.

SEÇÕES	DESCRIÇÃO
A	Necessidades Humanas
B	Operações de Processamento; Transporte
C	Química e Metalurgia
D	Têxteis e Papel
E	Construções Fixas
F	Engenharia Mecânica, Iluminação, Aquecimento
G	Física
H	Eletricidade

Tabela 3-1 Principais Setores das Patentes

Seção	C	Química e Metalurgia
Subseção	C07	Química Orgânica
Classe	C07D	Compostos heterocíclicos
Subclasse	C07D 401	Compostos heterocíclicos contendo dois ou mais heteroanéis, tendo átomos de nitrogênio como os únicos heteroátomos do anel, pelo menos um dos anéis sendo que um de seis membros, com apenas um átomo de nitrogênio.
Grupo	C07D 401/12	Compostos heterocíclicos ligados por uma cadeia contendo heteroátomos como elos da cadeia.

Tabela 3-2 Exemplo de Análise de Patente pelo Sistema CIP

3.2.2.2. Sistema de Classificação Europeu (EC)

O sistema de Classificação Europeu (EC) é um sistema de classificação interno do Escritório de Patentes Europeu e baseado nos símbolos da Classificação Internacional de Patentes. O Escritório de Patentes Europeu (*European Patents Office* ou EPO) atribuiu os símbolos EC às patentes na publicação, a fim de fazer a classificação mais precisa e mais fácil de usar. Para isto, subgrupos EC são adicionados ao símbolo CIP.

Diferentemente do CIP que permanece fixo por, no máximo, cinco anos, o sistema de Classificação EC se desenvolve constantemente a fim de se adaptar às necessidades do desenvolvimento tecnológico. Contudo, somente existe uma versão em qualquer época. Além disto, se o sistema de classificação for alterado no sistema EC, todas as patentes são

reclassificadas de acordo com o novo sistema. Isto se difere do sistema CIP, onde todas patentes são classificadas conforme a edição em vigor na data de depósito da patente. Quando a próxima edição estiver em vigor cinco anos mais tarde, as patentes classificadas sob a edição anterior não são reclassificadas. Isto remete a uma situação em que os documentos de patentes têm um símbolo de classificação diferente do que elas deveriam ter conforme a edição atualmente em vigor.

3.2.2.3. Sistema de Classificação dos Estados Unidos da América (USPC)

O *United States Patents Classification - USPC* - é um sistema de classificação de patentes utilizado pela *United States Patent and Trademark Office (USPTO)* para organizar todos os originais das patentes dos Estados Unidos e muitos outros originais técnicos em coleções relativamente pequenas baseadas em matéria comum. Cada divisão da matéria no *USPC* inclui um componente principal chamado classe e um componente menor chamado subclasse. Uma classe delineia geralmente uma tecnologia de outra. As subclasses delineiam processos, características estruturais, e características funcionais da matéria sujeita abrangidas dentro do espaço de uma classe. Cada classe tem um identificador

associa o original à classe e à subclasse identificada pela classificação. Os originais estão classificados em uma subclasse se uma classificação que corresponde à subclasse original lhe for atribuída. Um original pode ser um membro de mais de uma coleção, isto é, ele pode ter mais de uma classificação atribuída a ela. As classificações são atribuídas aos originais baseados na divulgação no original. Há mais de 450 classes no sistema de classificação USPC, e mais de 150.000 subclasses.

Como pontos positivos, o sistema *USPC* facilita a recuperação plena e rápida de originais técnicos relacionados e distribui aplicações de patente dentro do *USPTO* para a examinação. Periodicamente, o sistema *USPC* é emendado para cobrir tecnologias novas ou para cobri-las em tecnologias mais finas e detalhadas sobre as já existentes na base de dados. As revisões ao sistema *USPC* que requerem uma redistribuição dos originais entre coleções são feitas com os projetos de reclassificação dos documentos originais das patentes depositadas.

3.2.3. Bancos de Dados de Acesso Gratuito

Existem diversos repositórios de dados disponíveis na *Internet* de forma a permitir o acesso gratuito a várias bases de dados – bases estas que possuem uma grande gama de informações a respeito das patentes depositadas em instituições competentes e responsáveis pela concessão dos mesmos instrumentos ao seu responsável legal. Pode-se dizer que cada banco de dados possui características e particularidades próprias, bem como seus respectivos sistemas de busca e recuperação de conteúdos relativos às patentes solicitadas. As bases de dados passíveis de acessos variam de um banco de dados para outro.

Além de gratuitos, os bancos de dados citados a seguir são extremamente completos e contêm as coleções de patentes dos países de maior significância tecnológica. A maior desvantagem deles em relação aos bancos de dados comerciais com acessos mediante pagamentos está relacionada às limitações quanto aos procedimentos de busca. Para uma busca bem-sucedida, é fundamental a escolha correta das palavras-chave.

3.2.3.1. Banco de Dados do INPI

O Instituto Nacional da Propriedade Industrial (INPI) é o órgão responsável pela disponibilização da documentação das patentes em território brasileiro. Dentre as formas de disponibilização, citam-se a forma impressa, em microfilmes e em CDROM. Pode também ser acessada por meios eletrônicos via *Internet*. No site da instituição, encontra-se disponível um banco de dados com todos os pedidos de patentes brasileiras publicadas a partir de agosto de 1992, com mais de 50 mil registros. Os documentos completos de patentes podem ser solicitados ao INPI. O próprio domínio da instituição disponibiliza links que permite aos interessados acessar os bancos de dados gratuitos de diversos outros escritórios de patentes no mundo.

3.2.3.2. Banco de Dados de Pedidos PCT

Oriunda do projeto de Biblioteca Digital de Propriedade Intelectual da Organização Mundial de Propriedade Intelectual *(Intelectual Property Digital Library of World Intelectual Property Organization – IPDL / WIPO)*, a base de dados contém os documentos completos dos pedidos de patentes depositados segundo o Tratado de Cooperação de Patentes.

3.2.3.3. Banco de Dados do Escritório Europeu de Patentes

O Escritório Europeu de Patentes (EPO) hospeda em seu site uma base de dados que permite a pesquisa nos dados bibliográficos de patentes de diversos países, no texto e nas reivindicações. Quando selecionadas, algumas das patentes podem ser vistas na sua forma integral, inclusive imagens, com possibilidade de se obter suas patentes correspondentes.

3.2.3.4. Banco de Dados do Japão

O Escritório Japonês de Patentes hospeda em seu site uma base de dados permite a pesquisa nos dados bibliográficos dos pedidos de patentes no Japão com a possibilidade de se obter cópia dos documentos originais japoneses.

3.2.3.5. Banco de Dados da *USPTO*

A Organização Norte-Americana de Marcas e Patentes (*United States Patents and Trademarks Organization* ou *USPTO*) é uma agência federal do Departamento de Comércio dos Estados Unidos da América, ocupando um espaço de cinco edifícios interconectados e localizados em Alexandria, Virgínia. Emprega aproximadamente sete mil funcionários para executar suas funções principais: a examinação e a emissão das patentes e a examinação e o registro das marcas registradas. Como serviços primários da agência, processam e disseminam aplicações das patentes e das marcas registradas, bem como das informações nelas contidas.

A *USPTO* disponibiliza *on line* uma base que contém todos os documentos de patente americana desde o número um. As buscas em texto completo das patentes só podem ser realizadas em documentos publicados a partir de janeiro de 1976. Os documentos anteriores a esta data só podem ser pesquisados por número ou pela classificação americana. Existem ainda as bases de busca por número de documento e por pedido publicado.

3.3. Análise de Patentes

Pelas razões já levantadas até o momento, pode-se afirmar que a análise textual de patentes em busca de conhecimentos relevantes para o processo de Inteligência Competitiva obtém resultados bastante vantajosos para os analistas de negócios interessados e para os responsáveis por tomadas de decisão estratégicas, visando não só manipular seu próprio negócio, como também antever-se no aquilo que seus concorrentes pensam e procuram dentro do mercado comercial. Não basta para a organização explorar somente o seu próprio horizonte: mais do que isso, o fluxo de informações que transitam pelo contexto das patentes permite uma análise temporal a respeito de tendências de negócios ao longo do tempo sobre um determinado produto ou tecnologia; permite um processo de extração de informações sobre os depositantes de patentes, suas preferências e possíveis caminhos que poderão seguir; permite a verificação de tecnologias emergentes e pouco exploradas dentre outras informações de suma

formas. A análise bibliométrica sobre o conteúdo das patentes pode fornecer informações em torno da natureza e crescimento de uma atividade inventiva. Os dados bibliométricos caracterizam os documentos das patentes pelo fato dos mesmos possuírem uma estrutura bem definida de campos. Portanto, para que sua análise obtenha êxito, deve-se ter plena consciência da natureza da questão a ser investigada para se eleger o tipo de análise a ser empregado (KARKI, 1999; NARIN, 1994 apud CAPUTO, 2006).

Outra questão a ser considerada é quanto à formatação concisa e escolha consciente dos documentos de irão compor a base de dados que será submetida às técnicas computacionais. Documentos que fogem à fronteira do problema devem ser desconsiderados, sob pena de distorção do resultado final.

Uma vez conhecidas as principais características das patentes, suas importâncias para os processos de Inteligência Competitiva e sabendo-se que existem bases de patentes completas disponibilizadas na *Internet* de forma gratuita, torna-se uma tarefa interessante uma análise mais detalhada sobre uma base de patentes, de forma a submetê-las aos processos computacionais de extração de conhecimentos relevantes. Nesta seção, debater-se-ão o processo de coleta de patentes, bem como alguns métodos de análise computacional sobre as mesmas e seus devidos objetivos. Apresentar-se-ão, ainda, assuntos que devem ser considerados para os procedimentos de análise, tais como Mineração na *Web*, por conseguinte, Mineração de Textos e suas principais etapas.

3.3.1. Coleta de Dados

A coleta dos dados é uma etapa de grande importância para a obtenção de resultados dignos de relevância ao final dos processos relativos à Descoberta de Conhecimentos. Utilizar documentos com conteúdos não-relevantes a um determinado

assunto prejudica o resultado final da análise, de forma que se reduz sua confiabilidade, além de onerar o custo computacional do processo.

Tendo em vista o mundo contemporâneo, a tecnologia de busca e recuperação de dados e informações pela *Internet* ganhou uma significância muito grande no mercado de negócios, devido à rapidez no retorno dos resultados desejados.

O valor sofisticado do serviço de informação *on line*, atualmente, encontra-se no uso de bases de dados (motivo pelos quais algumas bases de dados de patentes completas e de acesso gratuito foram citadas anteriormente) não somente para recuperar informações, mas também para analisar e validar os resultados, de forma a combiná-los com outras informações.

Compreende-se a técnica de busca *on-line* como um processo de agregação de valores no tocante aos procedimentos de seleção e refinamento realizados com base em estratégias de pesquisas inteligentes. Para prover esse tipo de serviço e usufruir dele com o máximo de qualidade possível, o profissional da informação precisa ter habilidades analíticas, além de estar familiarizado com as técnicas avançadas de busca *on line* (*WORMEL*, 1998).

Para se aproveitar o conteúdo de seu escopo de forma inteligente, é necessário, primeiramente, que se saiba fazer uso das ferramentas de busca, objetivando a essência do que se deseja procurar. A busca se baseia na inserção por parte do usuário de palavras-chaves e retorna como resposta uma lista de páginas organizadas de acordo com a semelhança das palavras inseridas com os documentos eletrônicos pesquisados.

Todavia e segundo cita Caputo (2006), as ferramentas de busca podem apresentar baixa precisão no encontro de informações que realmente interessem e que seja base para a extração de conhecimentos realmente relevantes, e isso se deve a alguns fatores, citando-se

a baixa precisão em termos de relevância de resultados apresentados e de inabilidade de indexar todas as informações pertencentes à grande rede mundial de computadores.

A recuperação de informações é outro processo trabalhoso devido ao fato de que são dados semi-estruturados, não rotulados, distribuídos, heterogêneos e multidimensionais. Devido a estes fatores, geram-se dificuldades de se extrair os conhecimentos a partir dos dados recuperados.

O domínio da instituição norte-americana *USPTO* hospedado na *Internet* apresenta uma seção reservada somente para o tratamento das patentes, conforme pode ser visualizado na figura 3-1.

Encontram-se as patentes emitidas a partir do ano de 1976 com seus conteúdos de texto completos, e são pesquisáveis por qualquer um de seus campos, enquanto que as patentes entre os anos de 1790 e de 1975 são igualmente disponibilizadas, porém, somente são pesquisáveis pelos campos Data de Emissão, Número da Patente e Classificação Americana Atual.

O sistema de busca do site (figura 3-2) disponibiliza três rótulos de consulta: o primeiro trata-se de um rótulo de pesquisa rápida e simples (figura 3-3), no qual se utiliza somente dois termos-chaves digitados pelo usuário, e que são digitados em duas *textboxes* diferenciadas. Nesse processo, o usuário, também, seleciona os campos das patentes no qual a pesquisa será feita (os campos estão contidos em duas *comboboxes*, uma para cada termo-chave digitado para pesquisa) e o operador lógico a ser considerado (*AND*, *OR* ou *ANDNOT*), obtendo como resposta o retorno de uma lista de patentes (em forma de *links*) que contém aquelas palavras-chaves específicas nos campos desejados para busca.

O segundo rótulo de pesquisa (figura 3-4) é mais avançado e possibilita a utilização de *queries* para um retorno mais refinado, restringindo a pesquisa e minimizando a possibilidade de se retornar um documento não-pertencente ao escopo do tema – razão

pelo qual foi utilizado em presente dissertação. Esse processo permite ao usuário buscar documentos por mais de um termo-chave presentes em seus conteúdos. Além disso, e tal como o primeiro rótulo de pesquisa já comentado, permite diferenciar os campos de busca, de modo que o usuário pode, numa mesma pesquisa, querer o retorno de uma lista de documentos que contenham um termo-chave em um determinado campo *AND / OR* e outro termo-chave em outro determinado campo da patente. Nesse rótulo, é importante a utilização de operadores lógicos para junção das *queries*, tais como os operadores *AND / OR* já citados.

Finalmente, pelo rótulo de pesquisa pelo número da patente, basta que o usuário digite o número da patente no qual ele deseja visualizar (figura 3-5).

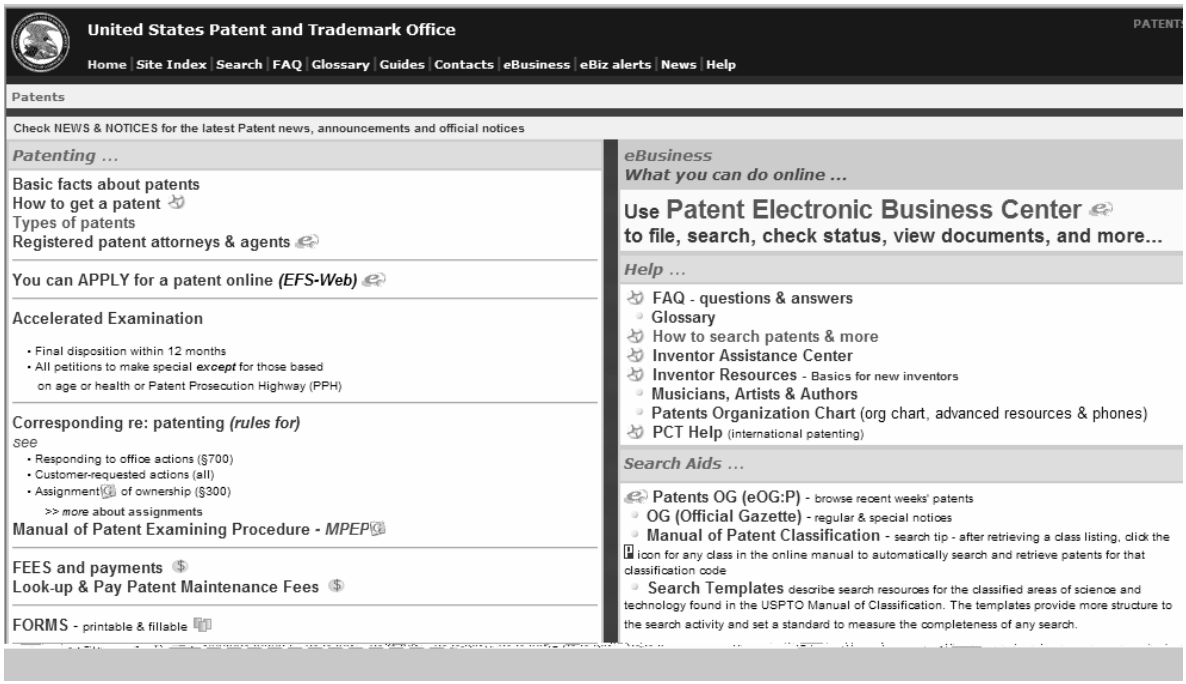

Figura 3-1 Seção sobre Patentes disponibilizada no site da *USPTO*

Figura 3-2 Rótulos de pesquisa disponibilizados no site da *USPTO*

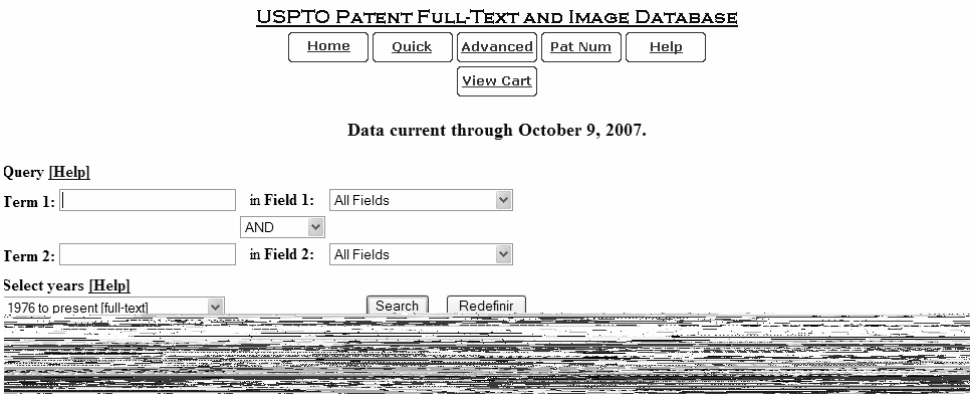

Figura 3-3 Modo de Pesquisa Rápida disponibilizada no site da *USPTO*

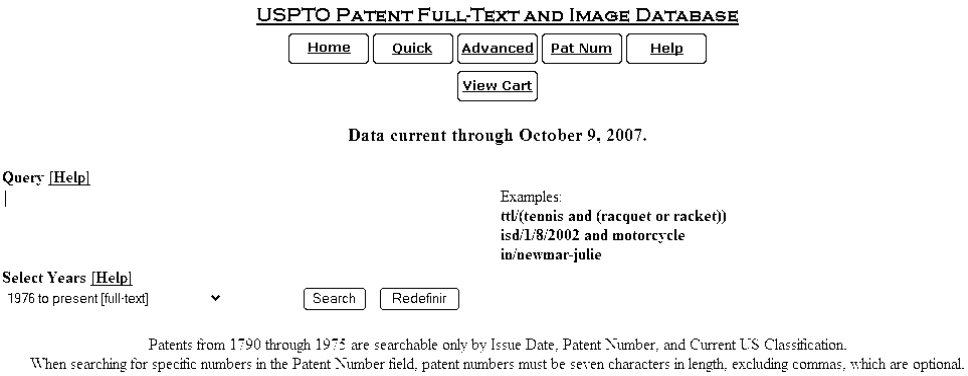

Figura 3-4 Modo de Pesquisa Avançada disponibilizada no site da *USPTO*

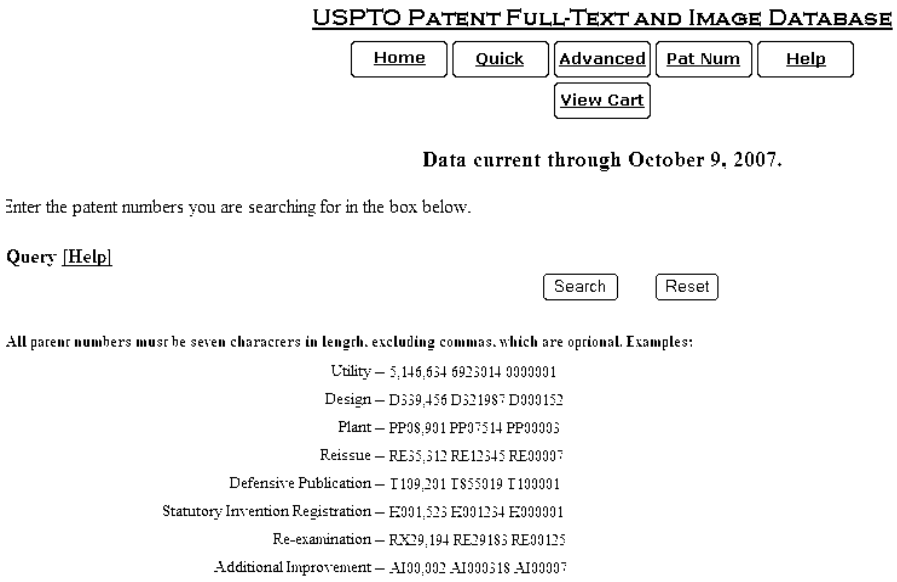

Figura 3-5 Modo de Pesquisa por número de patente disponibilizada no site da *USPTO*

O sistema exibe uma tela com uma lista de links em duas colunas para a visualização da descrição completa das patentes, bem como de seus campos, como retorno do resultado da pesquisa solicitada. Nessa lista de links retornados, constam as patentes que cumprem aos atributos que o usuário requisitou para pesquisa. A lista de links da esquerda da tela é composta por links indexados pelos números, ao passo que a lista de links da direita é composta por links indexados pelos títulos das patentes (figura 3-6). Uma vez que um link respectivo a uma determinada patente é acessado, o sistema remete-nos à sua visualização completa, com todos os campos e conteúdo da patente indexada.

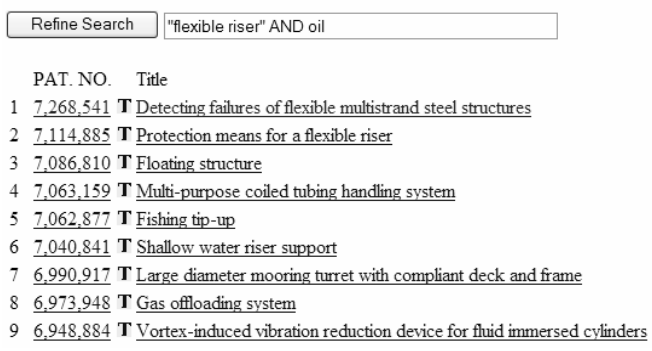

Figura 3-6 Exemplo de um resultado de uma pesquisa por patentes no site da *USPTO*

Quanto à sua estrutura, uma patente americana é composta dos seguintes e principais campos em destaque:

- PATENT SUMMARY (Sumário):
  - PATENT NUMBER (Número da Patente);
  - FILING DATE (Data de Preenchimento);
  - ISSUE DATE (Data de Emissão);
  - INVENTORS (Inventores);
  - ASSIGNEE (Cessionário);
  - CURRENT US CLASSIFICATION (Classificação Americana Atual);
  - INTERNATIONAL CLASSIFICATION (Classificação Internacional);
  - APPL Nº. (Número da Aplicação);
  - FILED (Data de Preenchimento)
- CITATIONS (Citações)
- REFERENCED BY (Referenciada por);
- SECTIONS PATENT (Seções da Patente):
  - ABSTRACT (Resumo);
  - DRAWING (Imagem);
  - DESCRIPTIONS (Descrição);
  - CLAIMS (Reivindicações).

Uma vez que se compreendeu o procedimento para coleta de dados sobre o banco de dados da *USPTO* escolhido para este trabalho, a próxima etapa foi definir os critérios de pesquisas para recuperação da informação disponível. Esta etapa foi muito importante para a elaboração dos estudos de casos, pois seguiram-se critérios para que as patentes pesquisadas fossem realmente pertinentes e delimitados nas fronteiras dos temas propostos, a fim de que os resultados da mineração não fossem distorcidos. Os temas propostos para

os estudos de casos estão ligados ao Setor da Indústria de Petróleo e Gás, e serão comentados com maiores detalhes no Capítulo 5 desta presente dissertação.

A etapa seguinte consiste em descrever o processo de Mineração na *Web*, uma vez que os textos a serem analisados encontram-se presentes no site da *USPTO* hospedado na grande rede mundial.

3.3.2. Mineração na *Web*

3.3.2.1. Justificativa

A *Internet* é uma rica fonte de armazenamento de dados eletrônicos nos dias atuais inegavelmente. Nesse sentido, tornou-se um importante e dinâmico canal de divulgação de marcas, de serviços, de compras e vendas, de comunicação rápida e até mesmo em tempo real e com respostas imediatas. É capaz de conectar pessoas pelo mundo inteiro, rompendo fronteiras e colocando a informação à disposição de quem a ela possui acesso. Por outro lado, milhões de documentos eletrônicos em seu escopo surgem e desaparecem diariamente, o que gera dúvidas com relação ao seu desempenho pleno como ferramenta mais compreensível e eficaz.

Devido a todas estas características, evidencia-se com clareza o porquê de sua utilização perspicaz acarretar uma vantagem competitiva a quem souber explorar o seu conteúdo de forma inteligente. Segundo Marinho e Girardi (2007), mais de um bilhão de páginas são indexadas pelos mecanismos de busca, e achar a informação desejada pode algumas vezes se tornar uma tarefa das mais trabalhosas. Essa grande quantidade de informações e recursos instigou a necessidade do desenvolvimento de ferramentas automáticas de mineração e descoberta de informações na *Web*.

3.3.2.2. Teoria

Utilizar e compreender os dados disponíveis na *Web* é uma tarefa complexa, e isso se deve ao fato de que esses dados são muito mais sofisticados e dinâmicos do que os sistemas de gerenciamento de bancos de dados tradicionais. Enquanto os bancos de dados tradicionais utilizam estruturas de armazenamento bem definidas, a *Web* não possui nenhum tipo de controle sobre a estrutura ou o tipo dos documentos que armazena. Outro aspecto que diferencia a mineração de dados da mineração na *Web* é a existência de vínculos de hipertexto entre os seus documentos. O hipertexto é uma rica fonte de informações a ser explorada, pois dentre outras coisas, ajudam no processo de classificação de páginas pelos mecanismos de busca e na identificação de pequenas comunidades na *Web*.

A mineração na *Web* ou *Web Mining* pode ser conceituada como a descoberta e análise inteligente de informações úteis da *Web* (COOLEY *et al*., 1997). Seus princípios advêm da mineração de dados, que consiste na descoberta de padrões válidos e úteis sobre uma base de dados (*KDD – Knowledge Discovery in Database*). Enxerga-se a mineração na *Web* como a utilização de técnicas de mineração de dados para a recuperação automática, extração e avaliação de informação para a descoberta de conhecimentos inteligentes sobre documentos e serviços da *Web*. Surgiu como uma disciplina atenuante aos problemas gerados pela necessidade em se explorar os dados contidos na *Internet* de uma forma mais inteligente e buscando sempre a vantagem competitiva sobre as organizações concorrentes.

Como técnica, a mineração na *web* vem sendo citada e estudada desde meados de 1996, e tem realmente ganho importância nestes últimos anos, e tal crescimento pode ser justificado devido ao aumento das transações comerciais na *Web;* e também devido ao

desenvolvimento da *Web* Semântica (DECKER *et al*., 2000) e da tecnologia dos agentes da informação (SYCARA, 1998), de forma que se utilizam as técnicas de mineração na *Web*.

3.3.2.3. Etapas do Processo

O processo de mineração na *Web* é dividido em quatro tarefas subordinadas análogas às fases do processo KDD (ETIZIONE, 1996). Tais fases podem ser visualizadas na figura 3-7.

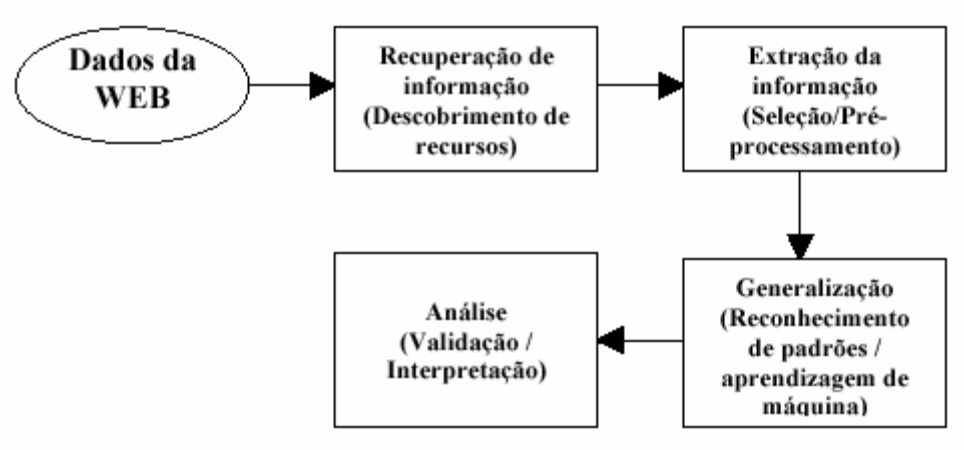

Figura 3-7 Tarefas da Mineração na *Web*

A tarefa de recuperação de informação ou descobrimento de recursos trata da automatização do processo de recuperação de documentos relevantes, que inclui, principalmente, representação, indexação e busca por documentos.

A indexação é um processo de classificação onde é realizada uma análise conceitual do documento ou elemento de informação. Existem várias técnicas de indexação, porém, independentemente da técnica empregada, o desafio maior é facilitar ao máximo possível a indexação de páginas da *web*, haja vista que indexá-la por inteira é uma tarefa utópica devido ao seu grande dinamismo, enorme quantidade e periodização constantes em atualizações. Nas técnicas de indexação baseadas no modelo do espaço

vetorial, a indexação envolve a atribuição de elementos de informação a certas classes, onde uma classe é o conjunto de todos os elementos de informação para o qual um termo de indexação (ou palavra-chave), em particular, tem sido atribuído. Os elementos de informação podem fazer parte de várias classes. Algumas técnicas atribuem pesos aos termos de indexação de um elemento de informação de forma a refletir sua relativa relevância (GIRARDI, 1998). Nas técnicas baseadas no modelo estatístico, os termos de indexação são extraídos a partir de uma análise de freqüência das palavras ou frases em cada documento e em toda a fonte de informação. Nas técnicas lingüísticas, os termos de indexação são extraídos utilizando técnicas de processamento da linguagem natural, por exemplo, análise morfológica, lexical, sintática e semântica (GIRARDI, 1995).

A Extração de informação é uma tarefa que visa identificar fragmentos específicos que constituem o núcleo semântico de um documento em particular, além de construir modelos de representação da informação (conhecimento). É a etapa posterior à recuperação dos documentos e nela aplicam-se algoritmos de mineração de dados e aprendizagem de máquinas de forma efetiva, pré-processando, dessa forma, os documentos eletrônicos resgatados. Seus métodos, geralmente, envolvem a escrita de código específico responsáveis pelo mapeamento do documento para algum modelo de representação do conhecimento. O entrave desta etapa é que para cada documento da *Web* é necessário que se escreva um código específico. Como os documentos da *Web* não possuem uma semântica agregada às informações que contém, e nem um padrão de como apresentar essas informações ao usuário, é necessário que se aprenda acerca da estrutura de cada documento para que se escreva o código respectivo ao mesmo, o que gera uma dificuldade para que se entenda ou se generalize este mesmo código escrito para os demais documentos.

Torna-se importante salientar a diferencia entre as fases de recuperação e extração de informação. As técnicas de extração de informação buscam derivar conhecimento de documentos resgatados segundo sua estrutura, ao passo que as técnicas de recuperação de informação visualizam o documento apenas como um conjunto de palavras.

A terceira etapa é a utilização de técnicas de mineração de dados e aprendizagem de máquinas após a extração das informações e da construção de um modelo de representação das informações. Nesta etapa, visa-se a descoberta de novos conhecimentos a partir do que já existe por parte do emprego dos algoritmos de mineração. O maior problema é a falta de marcação semântica das informações. Muitos algoritmos de mineração de dados requerem como entrada exemplos positivos ou negativos de algum conceito para que o mesmo possa tomar como base. Por exemplo, se existisse um conjunto de páginas da *Web* marcadas como exemplo positivo e negativo, seria fácil modelar um algoritmo classificatório para a classificação automática de novas páginas subseqüentes. Embora a *Web* atual dificulte o processamento das suas informações por parte das máquinas, a *Web* Semântica, segundo Lee (2001), oferece uma solução para este problema.

A quarta etapa consiste na aplicação de técnicas computacionais pelos analistas, de forma a se entender, visualizar, interpretar e validar os padrões descobertos na etapa anterior. Muitos sistemas aplicam técnicas oriundas das ferramentas *OLAP* (*On Line Analytical Processing* ou Processamento de Análise em tempo real), citando Girardi (1998), Han (2000) dentre outros.

3.3.2.4. Categorização

No tocante à sua categorização, a mineração na *Web* se divide em três, de acordo com a parte da *Web* a ser minerada.

Figura 3-8 Categorização da Mineração na *Web*

A figura 3-8 ilustra a categorização da Mineração na *Web*. A mineração de conteúdo aborda a mineração dos dados contidos dentro dos documentos da *Web* como fonte central de descoberta de informações relevantes. A grande quantidade de formatos que os dados podem assumir (textos comuns, páginas HTML, imagens, áudio, vídeo dentre outras) guia as técnicas de mineração a serem utilizadas.

A mineração de estrutura aborda a mineração das informações contidas entre os documentos da *Web*. Os documentos da *Web* se relacionam basicamente através de vínculos de hipertexto, e esses vínculos escondem informações valiosas e interessantes não só sobre a topologia da *Web*, mas também sobre como os documentos se relacionam entre si, tratando a rede interligada de documentos como um grafo orientado.

A mineração de uso aborda a mineração das informações sobre como o usuário interage com a *Web*. Nessa categoria são tratadas questões como personalização, interfaces adaptativas e aprendizado de perfis de usuários. Esse padrão possibilita a identificação de informações mais relevantes e o entendimento das reais necessidades do usuário. Isso facilita a organização no seu tratamento ao cliente, de forma a ter uma base para satisfazê-lo ao máximo possível.

3.3.3. Bibliometria e Análise de Citação

Define-se a Bibliometria como um ramo da ciência da informação capaz de exercer a função de inferir sobre a produção bibliográfica de um autor, de forma a mensurar a produtividade de cada autor e, a partir destes estudos, criar métodos de comparação entre vários autores de diversos estilos.

Outro método utilizável pode ser mensurado pelo numero de citações feitas do artigo original, de forma a incidir sobre a qualidade do contexto publicado. A hipótese básica deste método afirma que qualquer ato que cite o autor do periódico anterior é sempre significativo, pois se considera que a quantidade de citações que o mesmo autor tenha em outros artigos afins é um forte indicativo de sua qualidade e do impacto de sua produção científica naquilo que ele se propõe a escrever e publicar.

Nesse sentido, um impulso considerável na Bibliometria foi dado pela introdução de técnicas para Análise de Citação em documentos, que reza, exatamente, este princípio citado anteriormente. Pode-se afirmar que os indexadores de citações encontrados na literatura bibliométrica estabelecem a sistematização da produção científica, além de mensurar o grau de sua importância na confecção de obras de pesquisas subseqüentes publicadas.

A proposta da Análise de Citação surgiu por Garfield (1995) e deu origem à criação do *Science Citation Index* (*SCI*), publicado pelo *Institute for Scientific Information* (ISI), na Filadélfia, Estados Unidos. Desde esta época, vem sendo desenvolvida, posteriormente, em análises fundamentadas nas citações, também utilizadas por sociólogos, historiadores e pesquisadores da ciência em geral (CALLON, COURTIAL, PENAN, 1995).

Para avaliar e determinar a influência de um único escritor ou para descrever o relacionamento entre dois ou mais escritores podem-se usar seus métodos, e um caminho comum para condução dessa pesquisa é o uso do *Social Science Citation Index*, do *Science Citation Index* ou do *Arts and Humanities Citation Index* (BUFREM e PRATES, 2005).

Nas patentes, a análise de citação está entre as ferramentas que são mais utilizadas. A quantidade de citações que uma patente sofre em outras subseqüentes indica o grau de sua importância relativa. Além disso, as citações auxiliam na identificação das famílias das patentes, ou seja: uma nova patente que cita outras patentes anteriores a ela pode representar a evolução de um determinado produto ou tecnologia ou até mesmo uma nova tecnologia, de forma que esta indexação com as outras patentes anteriores permite identificar a partir de quais patentes houve o estudo para que a nova tecnologia ou a evolução de uma tecnologia anterior fosse descoberta, conseqüentemente, que a nova patente fosse publicada.

A família de patentes, com isso, estabelece a conexão entre as patentes existentes e permite a identificação de informações preciosas, como origem, crescimento de um determinado invento, depositantes, relacionamentos entre inventores, ligações com assuntos inventivos e tendências no mercado de tecnologia ao longo do tempo.

Em seus estudos, Caputo (2006) cita que a metodologia das análises pode gerar algumas informações pertinentes ao processo de geração de Inteligência Competitiva, dentre elas a quantidade de citações por patentes, patentes mais citadas, índices de impacto tecnológico, ciclo de vida de uma tecnologia entre outros indicativos, que têm sido utilizados para mensurar vantagem tecnológica, descobrir valores mensuráveis econômicos de novos inventos, natureza e quantidade de conhecimentos entre outros.

Dentre as desvantagens encontradas neste tipo de análise, podem-se citar:

- Grau de complexidade das relações existentes entre mais de dois documentos;

- Limitação da busca pelo valor potencial da informação disponível em seus conteúdos;
- Incapacidade de consideração da relação no conteúdo das patentes em si mesmas;
- Consumo de largo tempo de processamento no processo de busca e recuperação das patentes para estabelecimento das conexões entre elas e verificação da pertinência entre seus conteúdos.

3.3.4. Mineração de Textos

Em presente dissertação, consideraram-se somente os dados textuais das patentes para o processo de descoberta de conhecimentos relevantes. Para tanto, pôde-se eleger a categoria de mineração de conteúdo como a melhor para o processo de prospecção tecnológica por ser capaz de considerar e processar seu texto presente em conteúdo e tratá-lo como um problema simples de mineração de textos. Desta forma, ir-se-á considerar a mineração de conteúdo nas seções seguintes. A partir dos dados processados das patentes industriais recuperados da *web*, o processo a ser aplicado para prospecção tecnológica será tratado como mineração de textos.

A tecnologia de Mineração de Textos advém das técnicas de recuperação de informações, aprendizado de máquinas (que é um segmento do estudo de sistemas de Informação inteligentes que por sua vez é uma das aplicações notáveis da Inteligência Artificial), e da descoberta tradicional de informações estruturadas através do uso de bancos de dados e de procedimentos estatísticos.

Afirma-se, com isso, que Mineração de Textos é um conjunto de métodos empregados para navegar, organizar, encontrar e descobrir informação sobre bases

textuais. Pode ser vista como uma extensão da área de Mineração de Dados, com enfoque na análise de textos. Também é chamada de *Knowledge Discovery in Texts* ou, simplesmente, *KDT*.

O uso dessa tecnologia permite recuperar informações, extrair dados, resumir documentos, descobrir padrões, associações e regras e realizar análises qualitativas ou quantitativas em documentos de texto (ARANHA e PASSOS, 2006).

Diferencia-se Mineração de Textos de mecanismo de busca: na busca, o usuário já sabe o que quer encontrar. Em contrapartida, a tecnologia empregada em Mineração de Textos auxilia o usuário a descobrir informações. Além disso, Mineração de Textos é diferente de análise de constituintes, pois não é necessário formalizar toda a construção sintática do texto; é diferente de Robôs de Conversação, pois não se pretende simular o comportamento humano; é diferente de Mineração de Dados, pois, apesar de utilizar algoritmos e tecnologia semelhantes para as etapas de processamento tais como Classificação Supervisionada e Não-Supervisionada, trabalha com textos ao invés de dados estruturados em uma base de valores definidos.

Segundo Chen (2001), 80% das informações disponibilizados na *Internet* estão em formato de dados não-estruturados, assim como 80% dos conteúdos informativos das organizações. Em função dessa realidade, motivou-se o desenvolvimento das técnicas e ferramentas computacionais de extrações de padrões úteis para ganhos em vantagem competitiva pelos analistas e programadores interados com esta rica área de exploração.

Normalmente, as técnicas de mineração de textos são aplicadas em documentos tais como: *e-mails*, textos livres obtidos por resultados de pesquisas, arquivos eletrônicos gerados por editores de textos, páginas da *Internet*, campos textuais em bancos de dados, documentos eletrônicos, digitalizados a partir de papéis entre outros. As patentes não fogem a essa regra: são ricas em informações, como já foi dito anteriormente, e podem ser

valiosas quando bem exploradas em seus campos específicos de forma estratégica. Como exemplo, detalhes técnicos valiosos podem ser descobertos em campos não-estruturados como *ABSTRACT*, por exemplo, que estão disponíveis nas patentes internacionais contida no banco de dados da *USPTO*. Nesse mesmo banco de dados, nas patentes pertencentes encontram-se detalhes técnicos que podem ser muito úteis uma vez que submetidos aos processos de extração de conhecimentos nos campos de dados não-estruturados como *CLAIMS* (Reivindicações) e *DESCRIPITIONS* (Descrições).

Em síntese, o processo de Mineração de Textos, apesar de eficaz quando bem explorado, é trabalhoso, e a maior parte dos esforços demandam-se na fase de pré-processamento, que envolve a eliminação de palavras que nada acrescentam na descoberta de conhecimentos sobre os textos, redução das palavras contidas nos mesmos em radicais, eliminação de radicais excedentes com o mesmo significado entre outras medidas tomadas (e que são debatidas com maiores riquezas de detalhes na seção a seguir) para que se reduza a carga de dados, a fim de aperfeiçoar o processo como um todo, de forma a deixá-lo mais rápido e menos custoso computacionalmente. Cada etapa da Mineração de Textos pode empregar um processo computacional diferenciado, e isso vai de acordo com a que melhor satisfizer a base de dados a ser submetida ao processo.

Em razão disso tudo debatido, apresentam-se, nas próximas seções, alguns detalhes muito úteis para o Processo de Descoberta de Conhecimentos sobre Textos e enfocado no processamento das patentes internacionais.

3.3.4.1. Análise dos Dados Textuais e Pré-Processamento

A estruturação dos dados é a primeira etapa no processo de Descoberta de Conhecimentos sobre Textos. Manipular arquivos de textos é uma tarefa computacional de

difícil interpretação. Nessa etapa antecedente aos processos que realmente irão objetivar a extração de conhecimentos, preparam-se os textos de forma a transformar os seus dados para as etapas seguintes. A lógica dessa transformação está presente no próprio texto, através de padrões lingüísticos que devem ser considerados. Para se entender a forma como se processa a sintaxe lingüística de um determinado idioma, é necessário entender os modelos do significado de uma palavra. E a forma como se compreende essa palavra dentro do texto pode ser bastante útil na resolução de problemas na área. Comprova-se tal afirmação pelo estudo promovido por Passos (2006), no qual o autor afirma:

> Apesar de que todas as línguas que trataremos, são faladas por humanos, e por isso deveriam seguir o mesmo sistema lógico, as variações culturais podem ser bastante drásticas. Um único padrão essencial para entender todas às línguas ainda é utópico. Sendo assim, qualquer sistema de processamento de textos tem parâmetros específicos para cada língua, até os mais genéricos (como o sistema de busca Google) devem armazenar uma lista de palavras sem poder de discriminação de documentos para cada língua (p.3).

Uma forma bastante utilizada pelos analistas da área para representação da base de documentos e recuperação de informação é a aplicação pelo método tradicional do Modelo de Espaço Vetorial (*Vectorial Space Model* ou VSM), que pode ser definida como a representação geométrica dos documentos por pontos ou vetores contidos em um espaço Euclidiano t - dimensional, de forma que cada dimensão corresponde a uma palavra contida no dicionário. Segundo Salton (1989), nomeia-se dicionário o conjunto de total de palavras pertencentes a este espaço.

No Modelo de Espaço Vetorial, cada palavra tem um peso associado para descrever sua significância, que pode ser sua freqüência em um documento, ou uma função dela. A similaridade entre dois documentos é definida ou como a distância entre os pontos ou como o ângulo entre os vetores, desconsiderando o comprimento do documento. O comprimento é desconsiderado para levar em conta documentos de tamanhos diferentes. Assim, cada documento é normalizado de forma que fique com comprimento unitário.

A simplicidade de manipulação dos documentos e a facilidade de visualização tornam vantajoso o emprego da representação de documentos textuais por Modelo de Espaço Vetorial (VSM) e suas variantes. Uma explicação para isso é que operações de vetores podem ser executadas muito rapidamente e existem algoritmos padronizados para realizar a seleção do modelo, a redução da dimensão e visualização de espaços de vetores. Em parte por estas razões, o modelo de espaço vetorial e suas variações têm persistido em avaliações de qualidade, no campo da recuperação de informação segundo Baeza-Yates e Ribeiro Neto (1999).

Um problema óbvio com o modelo de espaço vetorial é a dimensionalidade alta: o número de palavras diferentes em uma coleção de documentos facilmente atinge centenas de milhares. Outro problema conhecido é composto por variações de estilo de escrita, eventuais erros etc. Além disso, quaisquer duas palavras são consideradas por definição não-relacionadas. Entretanto, é difícil obter uma informação exata das relações semânticas apenas a partir das informações textuais, automaticamente.

Na variante de Atribuição de Pesos, cada documento é representado como um vetor cujas dimensões são os termos presentes na coleção de documentos inicial a ser minerada. Cada coordenada do vetor é um termo e tem um valor numérico que representa sua relevância para o documento. Normalmente, valores maiores implicam em relevâncias maiores. É o processo de enfatizar os termos mais importantes. Existem várias medidas de atribuição de pesos, entre as quais podemos citar três mais utilizadas que são: Binária, TF e TF*IDF.

O esquema binário usa os valores 1 e 0 para revelar se um termo existe em um documento ou não, respectivamente.

A freqüência do termo (*Term Frequency* - TF) conta as ocorrências de um termo em um documento e usa este contador como medida numérica. Normalmente, as medidas

são normalizadas para valores no intervalo [0,1]. Isto é feito independentemente para cada documento, dividindo-se cada medida de coordenada pela medida de coordenada mais alta do documento considerado. Este procedimento ajuda a resolver problemas associados com o tamanho (comprimento) do documento. Sem a normalização, um termo pode ter uma medida maior num certo vetor-documento simplesmente porque o documento correspondente é muito grande.

O terceiro esquema é o TF*IDF (*Term Frequency – Inverse Document Frequency*) onde se multiplica a medida de coordenada oriunda de um esquema TF por seu peso global. Sendo assim, a medida total de um termo se torna a combinação de sua medida local (medida TF) e global (medida IDF). A medida IDF para o termo t é definida como log (N / Nt) onde N é o número total de documentos e Nt é o número total de documentos contendo t. A IDF aumenta conforme a singularidade do termo entre os documentos aumenta – isto é conforme sua existência diminui – dando assim ao termo um peso maior. Termos que ocorrem muito num certo documento (contagem local alta) e termos que ocorrem em poucos documentos (contagem global alta) são ditos como tendo alto poder de decisão e recebem pesos altos. O peso de um termo em um documento é uma combinação de suas contagens local e global. Pelas mesmas razões previamente expostas, a normalização é, também, muito utilizada neste processo. Para normalizar medidas baseadas no esquema TF*IDF, emprega-se a normalização do cosseno. . Ela é calculada segundo a fórmula apresentada:

$$Nt_k d_j = \frac{TF * IDF(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} (TF * IDF(t_s, d_j))(TF * IDF(t_s, d_j))}} \quad (1)$$

Em referência à fórmula apresentada acima, consideram-se tk e dj o termo e o documento levados em consideração, respectivamente; TF*IDF (ts,dj) é a medida da coordenada de ts em dj; e |T| é o número total de termos no espaço de termos.

O Modelo do Espaço Vetorial usa as estatísticas dos termos para comparar as necessidades de informação e os documentos. Cada documento é representado como um vetor de n dimensões onde cada dimensão é um termo proveniente do conjunto de termos usados para identificar o conteúdo de todos os documentos e consultas. A cada termo é então dado um peso dependendo da sua importância em revelar o conteúdo de seu documento associado ou consulta. Freqüências de termos podem ser binárias (simples) ou números (mais exatas). Um esquema típico usado para associar pesos a termos é o TF*IDF explicado previamente. Para comparar um documento dado com alguma consulta ou mesmo com outro documento (ambos representados no formato de vetor), o cosseno do ângulo entre dois vetores é medido como:

$$Cos(I,D) = I.D / (\|I\| \|D\|) \qquad (2)$$

onde I é ou o vetor de consulta; D é um vetor documento; I.D é o produto escalar de I e D; e ||I|| é a raiz quadrada do produto escalar do vetor I por ele mesmo.

Este método tem as vantagens de requerer uma intervenção de usuário mínima e ser robusto; entretanto, ele ainda sofre de efeitos de homonímia e sinonímia nos termos. Devido a este fato ocorrente, o passo seguinte consiste em executar uma análise de morfologia, a fim de se detectar e eliminar palavras de significados similares e relevância das mesmas, de forma a minimizar a dimensão dos documentos. Tal etapa de pré-processamento engloba a aplicação de métodos, tais como:

- Elaboração de uma lista contendo palavras a serem descartadas: este conjunto de palavras é chamado de *Stopwords* (conhecido ainda como *Stoplist*). *Stopwords* são palavras que não tem conteúdo semântico significante no contexto em que ela existe e são palavras consideradas não relevantes na análise de textos. É um dos primeiros passos no processo de preparação dos dados e visa à identificação do que pode ser desconsiderado nos passos posteriores do processamento dos dados. Normalmente isso acontece por se tratarem de palavras auxiliares ou conectivas (e, para, a, eles) e que não fornecem nenhuma informação discriminativa na expressão do conteúdo dos textos. Na construção de uma lista de *stopwords* incluem-se palavras como preposições, pronomes, artigos e outras classes de palavras auxiliares. *Stopwords* também podem ser palavras que apresentam uma incidência muito alta em uma coleção de documentos, portanto, não são consideradas discriminatórias;
- Redução das palavras em seus radicais, extraindo-se seus prefixos e / ou sufixos. Este processo denomina-se *Stemming*. Algoritmos de *stemming* empregam fatores lingüísticos e são dependentes do idioma. Os algoritmos de *stemming* atuais não costumam usar informações do contexto para determinar o sentido correto de cada palavra, e realmente essa abordagem parece não ajudar muito. Casos em que o contexto melhora o processo de *stemming* não são muito freqüentes, e a maioria das palavras pode ser considerada como apresentando um significado único. Os erros resultantes de uma análise de sentido imprecisa das palavras, em geral, não compensam os ganhos que possam ser obtidos pelo aumento de precisão do processo de *stemming*. Existem, porém, outros tipos de erros que devem ser observados e controlados durante a execução do *stemming*, *d*entre os quais, o *Overstemming*, que acontece quando a cadeia de caracteres

removida não era um sufixo, mas parte do *stem*, e que pode resultar na junção de termos não-relacionados; e o *Understemming*, que pode acontecer quando um sufixo não é removido e isto, geralmente, ocasiona uma falha na junção de palavras relacionadas.

Em decorrência de presente pesquisa debruçar-se sobre as patentes industriais internacionais, especificamente editadas em idioma inglês, apresentam-se, na seção a seguir, alguns métodos de *stemming*, com o objetivo de mostrar as diferentes abordagens utilizadas pelos algoritmos existentes. Estes métodos foram desenvolvidos, justamente, para o idioma inglês, embora sejam encontradas adaptações de alguns deles para diversos idiomas.

3.3.4.1.1. Métodos de *Stemming*

Um método de *stemming* simples é o *stemmer* S (HARMAN, 1991), no qual apenas uns poucos finais de palavras da língua inglesa são removidos: “ies”, “es”, e “s” (com exceções). Embora o *stemmer* S não descubra muitas variações, alguns sistemas práticos o usam, pois ele é conservador e raramente surpreende o usuário negativamente.

Outro método para redução de palavras aos seus respectivos radicais é o processo de *stemming* de Porter (1980), que consiste da identificação das diferentes inflexões referentes à mesma palavra e sua substituição por um mesmo *stem*. O intuito é conseguir agregar a importância de um termo pela identificação de suas possíveis variações. Outro ponto importante é o aumento de desempenho de um sistema quando ocorre a substituição de grupos de termos por seu *stem*. Isto acontece por causa da remoção dos diferentes sufixos, -AR, -ADO, -AÇÃO, -AÇÕES, deixando apenas o radical comum. Sendo assim, o

algoritmo de Porter remove 60 sufixos diferentes em uma abordagem multifásica. Cada fase remove sucessivamente sufixos e promove alguma transformação no *stem*.

Finalmente, pode-se achar na literatura, também, o método de *Lovins* (1968), que é constituído de um único passo, sensível ao contexto e que utiliza um algoritmo da combinação mais longa para remover cerca de 250 sufixos diferentes. Este método remove no máximo um sufixo por palavra, retirando o sufixo mais longo conectado à palavra. O algoritmo foi desenvolvido a partir de um conjunto de exemplos de palavras na língua inglesa, usadas para formar a lista de regras de *Lovins*. No entanto, vários sufixos não foram contemplados por esta lista, o que não o impede de ser considerado o mais agressivo método dos três algoritmos de *stemming*.

3.3.4.2. Análise de Agrupamentos (*Clustering*)

Clustering ou Agrupamento se refere à separação ou agrupamento de objetos ou elementos similares em classes, geralmente de modo automático, com ferramentas automatizadas. A Clusterização é uma técnica de aprendizado não-supervisionado, que visa à descoberta de estruturas em uma coleção de dados que sejam de natureza desconhecida. Diferencia-se da Classificação Supervisionada: a Clusterização (ou Classificação Não-Supervisionada) visa à criação de classes através da organização dos elementos considerados com semelhanças entre si, enquanto que a Classificação Supervisionada procura alocar elementos com seus rótulos já conhecidos em classes pré-definidas.

A separação dos elementos é feita com base numa avaliação de similaridade entre os mesmos, de forma a colocar os elementos mais similares na mesma classe. Para tanto, devem ser escolhidas as características que irão representar os objetos. No caso da Mineração de Textos, adotou-se como critério de similaridade a distância existente entre

eles, de forma que dois documentos ou mais poderão ser alocados em uma mesma classe se estiverem próximos entre si no espaço vetorial.

De uma forma geral, o processo de *clustering* requisita ao usuário a definição de qual será o número de agrupamentos a ser considerado, e a partir da definição deste número, separam-se os registros de dados por critérios de similaridades entre si, de forma a maximizar a similaridade *intracluster* e minimizar a similaridade *intercluster*. Quando separados os registros e definidos os grupos, pode-se, então, debruçar-se sobre os elementos que os compõem, de forma a identificar as características comuns aos seus elementos, portanto, criar um identificador representativo para cada agrupamento.

Para que um algoritmo de *clustering* possa executar o que ele propõe, torna-se necessária a utilização de uma estrutura de dados capaz de guardar os objetos a serem processados ou as informações que dizem respeito às relações entre os mesmos. De uma forma geral, destacam-se duas estruturas principais de dados, que são a matriz de dados e a matriz de similaridade.

Na matriz de dados, as linhas representam cada um dos objetos a serem agrupados e as colunas representam os atributos de cada objeto. Na matriz de similaridade, cada elemento componente representa a distância entre pares de objetos. Segundo Han e Kamber (2000, apud GOLDSMITH e PASSOS, 2005), quando um algoritmo que trabalha utilizando matriz de similaridade recebe uma matriz de dados como entrada, o primeiro procedimento que o algoritmo executa é transformar a matriz de dados em uma matriz de similaridade, para em um segundo momento, submetê-la ao processo de *clustering*.

No tocante aos métodos e técnicas de *clustering*, existem várias presentes ma literatura técnica, dentre os quais pode-se citar o *K-Means*, cujo método utiliza-se da média dos objetos pertencentes ao agrupamento em questão (centro de gravidade). Define-se o centro de gravidade do agrupamento como o vetor médio dos dados que pondera-se por

todos os itens do agrupamento. O intuito deste processo algorítmico é encontrar k-grupos de documentos a partir de um número natural fixo k, e inicia-se com a inserção aleatória de k centróides ao espaço vetorial do conjunto. Associa-se cada documento ao agrupamento que possuir o centróide mais próximo, e em etapa posterior, calcula-se o centróide de cada agrupamento. O processo algorítmico encerra quando não houver mais atualizações ou quando alcança-se número máximo de iterações do algoritmo.

3.3.4.3. Extração de Conhecimentos por Mineração de Textos sobre Patentes Industriais em Indústria de Petróleo e Gás

A descoberta de conhecimentos relevantes em uma base de documentos textuais oferece ganhos competitivos e oportunidades, para as organizações, de anteverem-se às mudanças dentro de um mercado de negócios cada vez mais exigente. Os resultados advindos da utilização de processos computacionais que objetivem a análise inteligente de dados oferecem suporte às tomadas de decisões por parte das organizações. É considerada, dentro do processo de extração de conhecimentos, como uma etapa pós-processamento: uma vez que as informações são extraídas, transformando-se em conhecimentos de grande importância, o responsável pela tomada de decisões poderá prover-se dos resultados para direcionar os rumos de seus negócios, de forma a considerar, também, a necessidade de investimentos de sua organização.

A Mineração de Textos sobre documentos textuais de patentes ganha, com isso, grande valor e importância de mercado, à medida que considera-se uma patente como uma fonte detentora de ricas informações à respeito de tecnologias de mercado emergentes e precursoras de outras tecnologias. Sua utilização com eficácia pode gerar grande impacto comercial quando bem utilizados seus resultados pelo responsável pelas tomadas de decisões. É evidente que a extração de conhecimentos em patentes, por serem documentos bem estruturados e subdivididos em diversos campos, torna-se um processo trabalhoso para a extração de conhecimentos.

No entanto, pode-se minimizar esta dificuldade, uma vez que os documentos de patentes extraídos da *Internet* em uma base específica disponibilizada *on line* são tratados como textos simples, portanto, aplicam-se sobre eles as etapas já descritas para o processo

de Mineração de Textos. Milhares de patentes são depositadas a cada ano, portanto, esse processamento torna-se indispensável para a compreensão do contexto das patentes em si.

Sempre que um produto novo ou uma técnica inovadora perante as técnicas já existentes surgem no mercado, é interessante que suas patentes sejam alvo de exploração. Na indústria, o panorama não pode ser diferente: explorar novas tecnologias emergentes, descobrir os seus campos de aplicação, descobrir detalhes tecnológicos interessantes inovadores ou novos detalhes aperfeiçoados de tecnologias aplicadas anteriormente e, que serviram que ponto de partida para as tecnologias subseqüentes advindas de documentos textuais de natureza intelectual dentre outras aplicações possíveis, são fatores que impulsionam a indústria e estimula a concorrência entre grandes empresas. Com isso, a tendência é que o mercado, em uma visão geral, ganhe com esse processo evolutivo, e nessa linha de raciocínio tende a ganhar maiores vantagens competitivas as empresas que souberem enxergar primeiro toda essa barganha de conhecimentos inteligentes e de grande valor para os negócios.

No mercado de Petróleo e Gás essa perspectiva é semelhante: a concorrência entre marcas pela exploração das tecnologias de extração dessas ricas fontes de minério *offshore*, por exemplo, ganhou grandes proporções nas últimas décadas, e isso porque é o segmento *offshore* que o setor industrial de Petróleo responde mais alto: 81% do que se produz vem do subsolo marinho; 64% das áreas em concessão estão em águas profundas e ultra-profundas; e 90% das reservas comprovadas de petróleo estão no mar, segundo dados divulgados pela Petrobras.

Sendo assim, o intuito da pesquisa é a exploração das patentes referidas a assuntos que estão em fase de emergência dentro da indústria petrolífera utilizando duas plataformas distintas desenvolvidas para extração de conhecimentos. As bases das patentes industriais foram delimitadas dentro de seis temas e que serão citados no Capítulo 5 deste trabalho.

Procurou-se utilizar as patentes contidas no banco de dados textuais da *USPTO* como fonte de busca para seis assuntos emergentes, o que remete à exploração de seis bases de dados diferentes e, conseqüentemente, a seis estudos de casos para esta pesquisa. Por serem tecnologias emergentes, algumas bases de dados foram delimitadas com poucas patentes industriais sobre o assunto determinado, mas que nem assim diminui a importância da aplicação dos processos de prospecção tecnológica.

A criação de agrupamentos é feita de forma a mensurar as similaridades entre as patentes, e nesse processo, identificam-se as palavras-chaves para cada agrupamento formado. A partir disso, torna-se possível a identificação das vantagens para utilizações competitivas, tal como o entendimento dos principais assuntos relacionados às patentes industriais processadas para cada estudo de caso feito. Estes assuntos podem ser visualizados através da geração dos agrupamentos, e fornecem estruturas de dados que simplificam a compreensão das informações contidas nos documentos textuais submetidos à análise.

4. Ambiente Computacional

Em presente capítulo, serão comentados os *softwares* utilizados em todas as fases da pesquisa, desde a busca e recuperação de documentos eletrônicos no *site* da instituição norte-americana *USPTO* até a visualização dos resultados oriundos das tarefas de Mineração de Textos.

4.1. Coleta de Dados – Utilização do *Web Crawling / Parsing*

Nessa fase, obedeceram-se aos critérios de pesquisas por termos-chaves previamente definidos e utilizou-se o sistema do *site* da *USPTO* para busca por patentes. Para que os documentos pertencentes à lista retornada de cada pesquisa elaborada fossem capturados da *Internet* e copiados em formato de documento de texto simples para o computador local, utilizou-se um mecanismo implementado para automatizar o processo de busca e recuperação de conteúdos. Esse recurso computacional é adotado pela ferramenta BIGUÁ – *software* de código livre para mineração de textos cujo projeto é desenvolvido por profissionais competentes na COPPE/UFRJ. Como melhoramento a esse recurso computacional, acrescentou-se uma interface gráfica, que pode ser visualizada na figura 4-1.

A ferramenta possui como vantagem o mecanismo de busca em profundidade por páginas indexadas ao *link* solicitado para pesquisa e extração de informações. Possui, além disso, a funcionalidade de extrair o código HTML na medida em que os documentos eletrônicos são recuperados, transformando-os em documentos de textos simples para exploração posterior em etapas futuras.

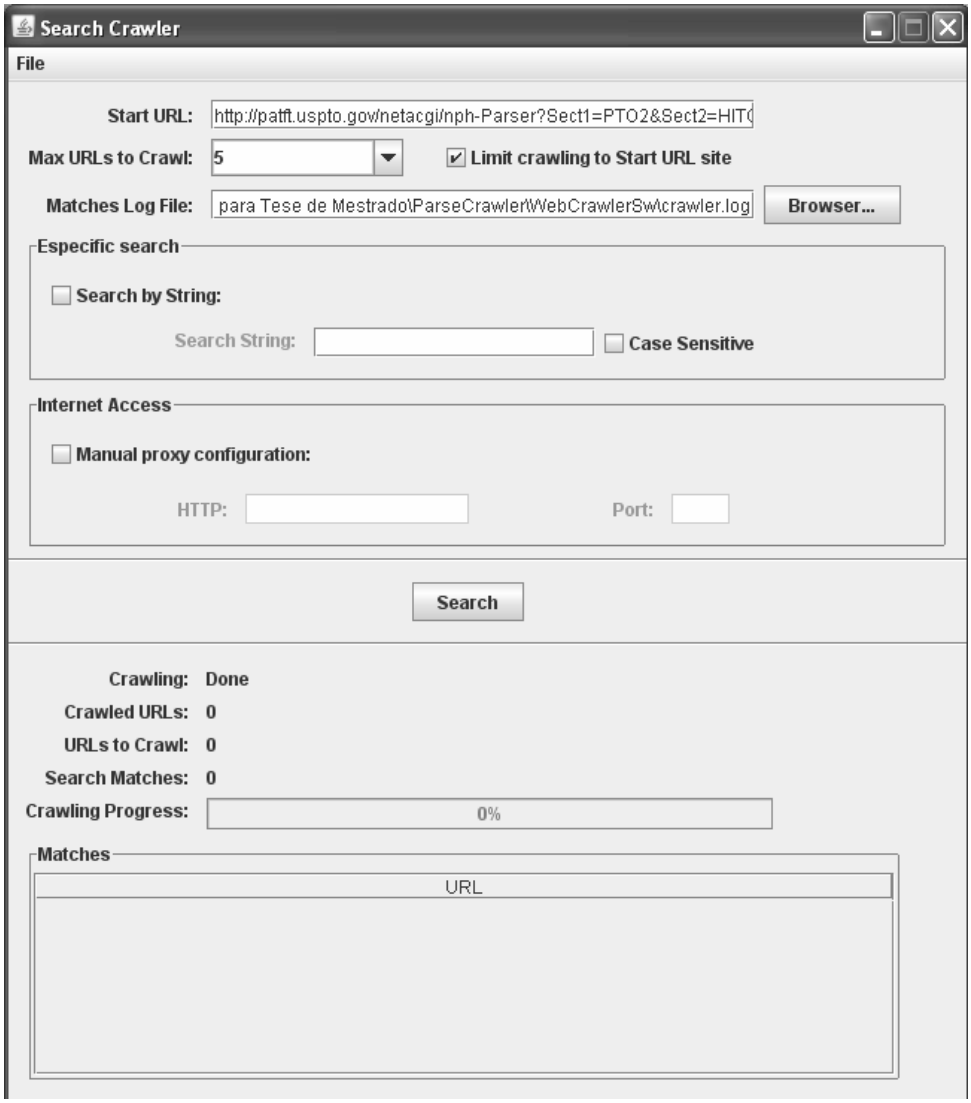

Figura 4-1 Interface gráfica do *Web Crawling* utilizado

Através da opção *"Start URL"*, o usuário pode digitar o *link* desejado para extração dos documentos da *Internet* (*download*). Tal *link* representa a lista refinada de documentos já pesquisados no site da instituição norte-americana em etapa anterior. A opção *"Max URLs to Crawl"* define a profundidade máxima de busca e recuperação pela informação, ou seja, até que nível hierárquico da árvore de documentos eletrônicos submissos à *URL* requerida pelo usuário tal pesquisa deverá ser feita. Caso o usuário desejasse que a pesquisa fosse feita até a "raiz" da árvore, sem definir qualquer valor, bastaria selecionar a opção *"Limit Crawling to start URL site"*. O botão *"Search"*

acionado pelo usuário determina o início da pesquisa. A cada documento eletrônico pesquisado, o recurso realiza o seu *download* e converte-o para arquivo eletrônico com extensão de documento de texto simples (“.txt”). O processo termina, automaticamente, quando o último documento eletrônico é processado.

Em procedimento posterior, os documentos recuperados foram preparados para submissão às etapas inerentes ao pré-processamento de dados. O campo *ABSTRACT* presente nas patentes coletadas foi escolhido para a submissão aos processos de mineração de textos pelas duas plataformas que serão citadas em seções a seguir. Por esta razão, mantiveram-se, somente, os conteúdos do campo *ABSTRACT* presente em cada uma das patentes coletadas, e eliminaram-se os demais campos dos documentos.

4.2. Análise Estatística - *OpenOffice.org Calc*

OpenOffice.org Calc é um aplicativo capaz de trabalhar com planilhas eletrônicas, manipular fórmulas e funções matemáticas das mais variadas, além de gerar gráficos com eficiência para melhor visualização dos resultados obtidos. É uma ferramenta que se pode executar na plataforma *Microsoft*, porém possui como vantagem ser um *software* de distribuição gratuita dentro do mercado.

Cada planilha formatada contém, além dos registros que são os documentos que pertencem à base e que estão dispostos em linhas, os principais atributos de cada patente, que são seus campos escolhidos para representarem os documentos (nome da patente e número de registro) e para serem analisados estatisticamente (Classificações Internacionais das Patentes, datas de registros e cessionários / inventores) e estão dispostos em colunas. Cada planilha formada representa um estudo de caso diferente.

4.3. Algoritmo de Validação de *Clusters*

Haja vista que o algoritmo de *Clustering K-Means* empregado na plataforma *RapidMiner / YALE* – que será descrita no próximo tópico – necessita de uma entrada de valor inteiro *K* que corresponde ao número de agrupamentos a serem formados pelo método sobre uma base de dados e que tal plataforma não oferece uma métrica de validação usual tal como o índice *PBM* (PAKHIRA *et al*., 2004), alguns critérios tiveram que ser adotados para que houvesse a definição consciente de qual valor *K* seria adotado para cada estudo de caso proposto nesse trabalho.

Um dos critérios adotados para validação do número de agrupamentos foi a utilização do algoritmo de validação de *clusters* – implementado em plataforma de programação *Matlab* e disponibilizado para o corpo discente da COPPE / UFRJ para fins acadêmicos. Tal algoritmo oferece como vantagens o fato de observar o número de *clusters* sugeridos segundo quatro métricas de validação distintas: *Calinski Harabasz* (*CALINSKI* and *HARABASZ*, 1974), *Discriminante* de *Fisher* (1996), *PBM* (*PAKHIRA* et al., 2004) e *Xie-Beni* (*XIE* and *BENI*, 1991); além de gerar gráficos de convergência do algoritmo, facilitando dessa forma a visualização do resultado de cada índice.

Como entrada, o algoritmo solicita ao usuário que digite o nome de arquivo em formato de texto (extensão ".dat", por exemplo) e o número de atributos do arquivo correspondente à base de dados a ser examinada, além do número máximo de *clusters* a serem testados para validação. Vale ressaltar que todos os arquivos de entrada correspondentes às bases de dados dos estudos de casos foram extraídos da plataforma *RapidMiner / YALE* após as etapas inerentes ao pré-processamento de dados em formato de arquivo digital "*.xls*", e trabalhados no *software OpenOffice.org Calc* para formatação dos valores e conversão dos mesmos arquivos ao formato ".dat".

Mesmo nos casos em que, após diversas iterações, houve divergências quanto aos resultados finais oferecidos por cada índice de validação, elegeu-se como resultado o valor oferecido pela maioria dos índices. Pode-se afirmar que a utilização desse algoritmo foi fundamental e bem considerada para a tomada de decisões quanto ao número de agrupamentos considerados em cada estudo de caso distinto. Um exemplo de visualização de gráficos gerados ao final da iteração do algoritmo pode se visualizar na figura 4-2.

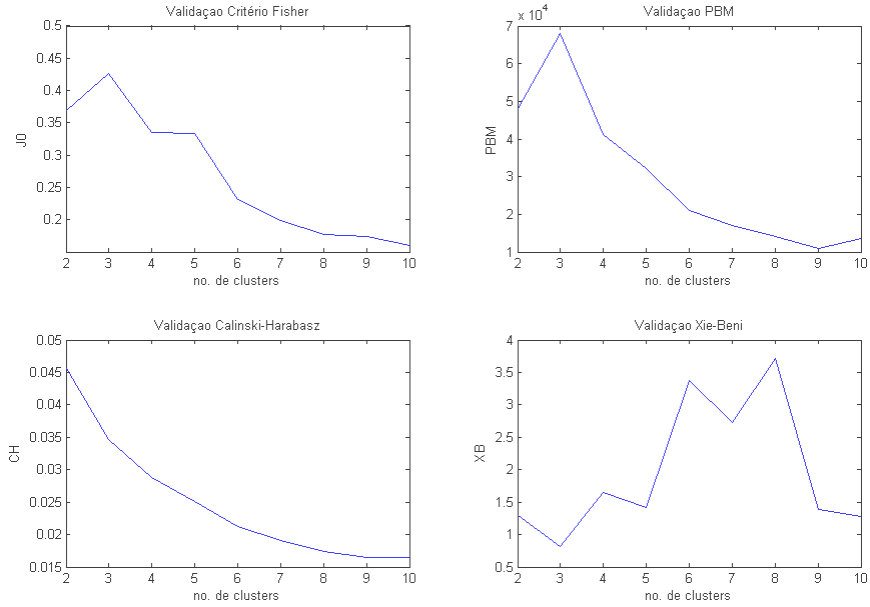

Figura 4-2 Exemplo de gráficos gerados pelo algoritmo de validação de agrupamentos

4.4. *RapidMiner / YALE*

4.4.1. Breve Histórico

RapidMiner (anteriormente denominado *YALE*, em sua primeira versão) é uma plataforma de código-fonte aberto e implementado em linguagem de programação *Java* que foi projetada para as tarefas inerentes à geração de Inteligência Competitiva. O desenvolvimento de grande parte de seus conceitos começou em 2001 na unidade de

Inteligência Artificial da Universidade de *Dortmund*, sendo que sua primeira versão foi lançada em 2002. Nos dias atuais, a plataforma encontra-se na versão 4.1 *beta*, e pode ser instalada e executada em ambientes *MS Windows* e *Linux,* e seu código-fonte aberto encontra-se hospedado pela *SourceForge* para que colaboradores possam realizar seu *download* para testes e melhorias em favor da plataforma, que permanece em constante evolução. Sua versão de código-aberto e uma versão de código não-aberto de *RapidMiner / YALE* são mantidas pela organização *Rapid-Intelligence* e encontram-se hospedadas em seu domínio na *Internet.*

4.4.2. Ambiente Operacional

A plataforma oferece tarefas inerentes à Mineração de Dados, com diversos algoritmos implementados para as fases de pré-processamento e de processamento de dados em forma de operador. O conceito de módulo do operador permite o projeto de correntes aninhadas complexas do operador para um vasto número de problemas de aprendizagem. A manipulação de dados faz-se de forma transparente aos operadores, ficando para o usuário a tarefa de organizá-los de forma coerente, estruturando-os em uma forma de árvore de processamento, do qual a plataforma executa-a em ordem descendente seqüencial. Um exemplo de árvore de processamento para descoberta de conhecimentos sobre textos *(Knowledge Discovery in Texts* ou *KDT)* pode ser visualizado na figura 4-3.

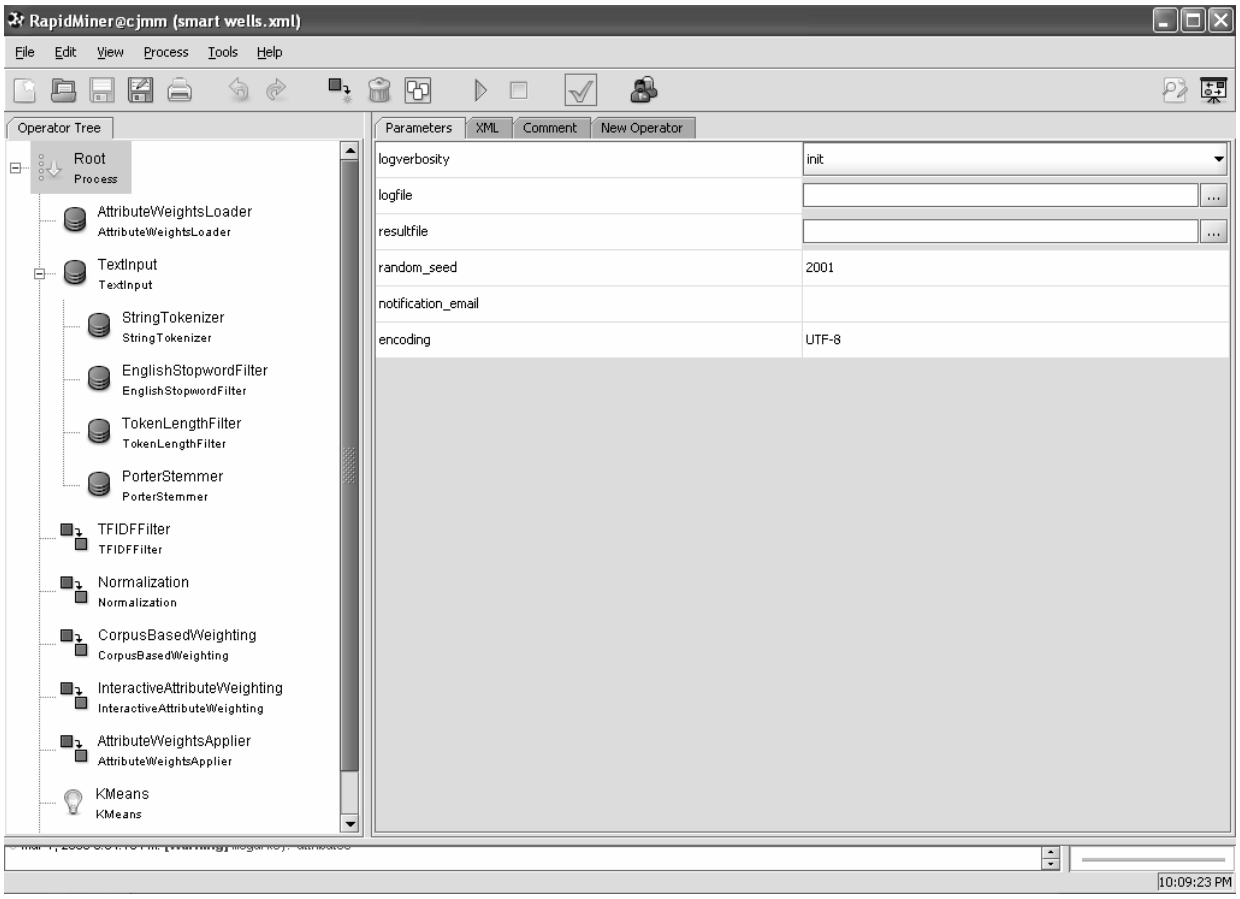

Figura 4-3 Ambiente Operacional e exemplo de árvore de processamento

Dentre os recursos que a ferramenta computacional disponibiliza, podem-se citar: métodos para tratamento da entrada de dados; métodos de análise inteligente de dados (matriz de correlação, por exemplo); técnicas para seleção de atributos (análise de componentes principais e algoritmos genéticos, por exemplo); técnicas e algoritmos implementados para Classificação Supervisionada e Não-Supervisionada de Dados; comandos para utilização de ferramentas *OLAP* e geração de gráficos para melhor compreensão dos resultados. Compondo seu ambiente interativo, a plataforma *RapidMiner* oferece total integração com a biblioteca de aprendizado de máquina implementados na ferramenta *WEKA* (também programada em linguagem Java); *plugins* simples e uma larga variedade de *plugins* que já existem para as tarefas de Mineração de Textos, simulação de

conceitos sobre modelos de séries-temporais, Mineração de Dados Distribuída e *Data Stream*.

Em presente pesquisa, executou-se o *plugin Word Vector* (vetor de palavras) e utilizaram-se seus recursos para Pré-Processamento e Processamento de Dados Não-Estruturados (Mineração de Textos). Na seção seguinte, descrever-se-ão um pouco mais sobre os comandos utilizados.

4.4.3. Descoberta de Conhecimentos em Textos utilizando

redução ao seu próprio radical; e *Porter Stemmer*, que elimina os prefixos e os sufixos das palavras consideradas utilizando o algoritmo de *Porter*.

O operador *TFIDFFilter* possui a finalidade de calcular os pesos *TFIDF* de cada termo remanescente do pré-processamento. O *operador Normalization* realiza a normalização dos termos para uma faixa de valores definidos pelo usuário; no caso de presente dissertação, definiu-se esta faixa entre os valores zero e um. O *operador CorpusBasedWeighting* usa um rótulo dos exemplos para caracterizar uma única classe ajustando-se os pesos – em cada estudo de caso, o rótulo definido foi o próprio nome do tema abordado. O operador *InteractiveAttributeWeighting* mostra uma janela com pesos inerentes a cada termo considerado e permite que o usuário os modifiquem, caso queira. A figura 4-4 permite a visualização da janela exibida pelo comando durante sua execução. Aproveitando-se de seus resultados, o operador *AttributeWeightsApplier* elimina os termos com peso zero e recalcula os demais pesos dos termos remanescentes no vetor.

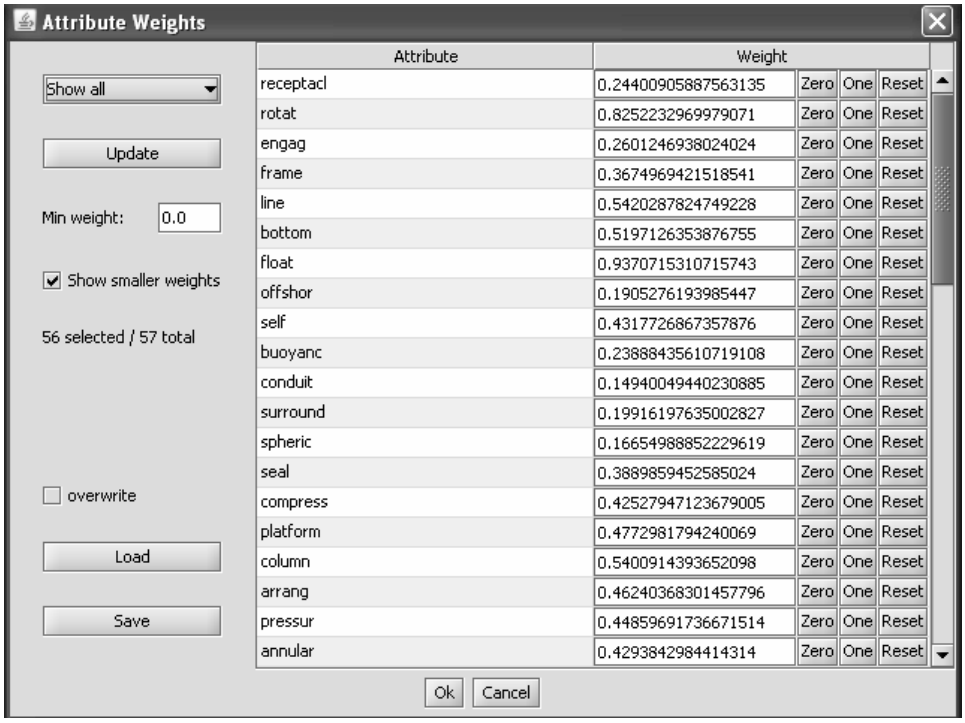

Figura 4-4 Janela de execução do comando *InteractiveAttributeWeighting*

Após todos esses operadores citados serem executados na fase de pré-processamento, executou-se o operador do algoritmo de *Clustering K-Means* sobre o vetor de termos criado. Contudo, antes de sua execução, a tabela de valores do vetor de termos criado na etapa de pré-processamento é extraída da plataforma e gravada em um arquivo de extensão “dat”, através de um recurso de exportação presente na tela de resultados do vetor e disponibilizada sob acionamento do botão “*Save*...”. Conforme já foi dito, o tratamento da tabela é feito pelo *software Office.org Calc*, e sua submissão ao algoritmo de validação de *clusters* ajuda a determinação de quantos *clusters* serão formados. A figura 4-5 permite visualizar a tela de resultados do vetor de palavras.

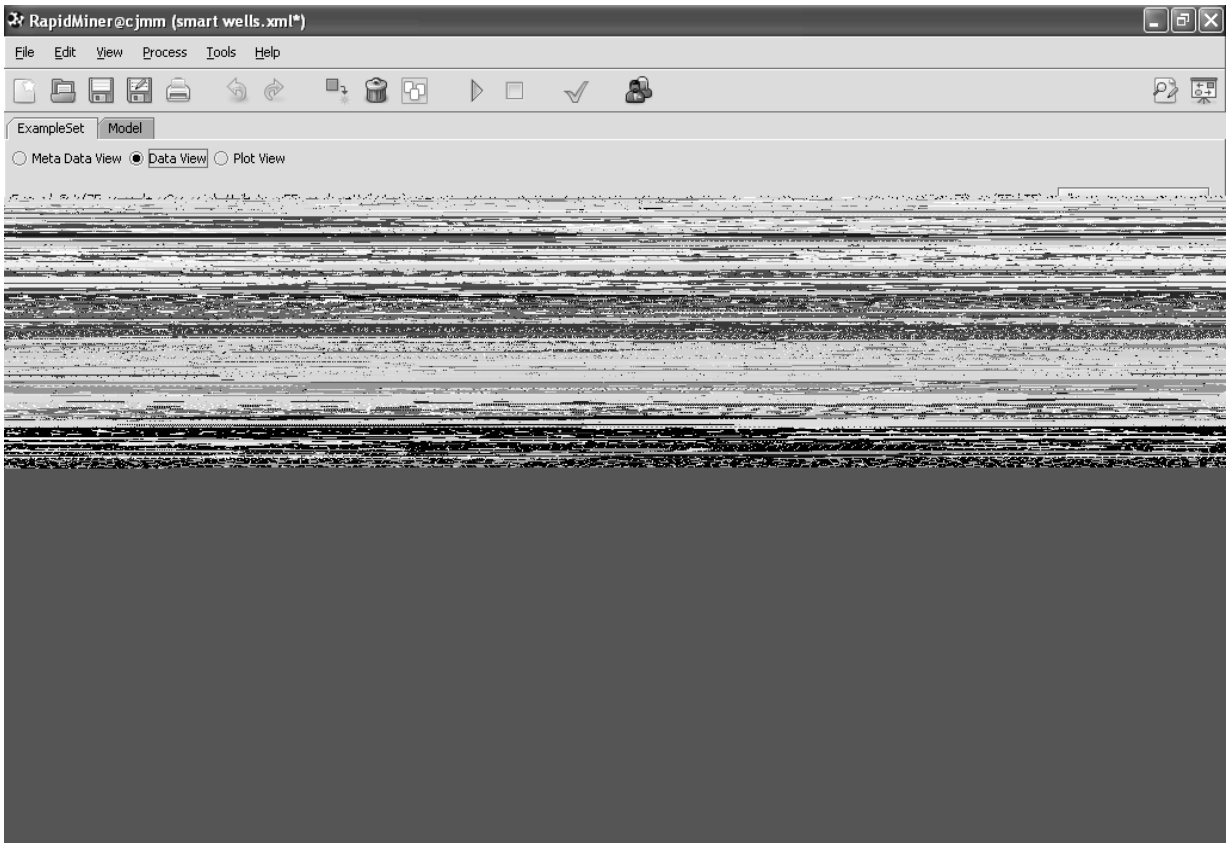

Figura 4-5 Tela dos resultados do vetor de termos formado pelas etapas de pré-processamento

Por fim, executam-se, novamente, os operadores inerentes à Mineração de Textos, ativando-se o operador *K-Means* e inserindo o número de *clusters* desejados para formação do algoritmo pertinente.

O operador *AttributeWeightsLoader* permite a recuperação e leitura do resultado mais recente do pré-processamento sobre a mesma base textual de dados, provendo ao usuário a opção de modificar ou eliminar um termo que o próprio considerar como não-interessante ao processo. Além do valor k (número de clusters), o operador *K-Means* oferecem parâmetros para o usuário definir o número de iterações mínima e máxima que o algoritmo executará. Para os estudos de casos presentes no trabalho, definiram-se como dez e cem os valores mínimos e máximos, respectivamente.

Os processos, ao final, podem ser gravados em arquivos de extensão *XML*, *(eXtensable Modeling Language)* e reaproveitados para execuções futuras. A tela dos resultados finais permite quatro tipos de visualização: a visualização dos termos mais relevantes de cada cluster formado (*Text View*, que pode se visualizar na figura 4-6); a visualização *Folder View*, que forma as pastas que representam os *clusters*, de forma que cada pasta contém os documentos presentes em cada uma delas; *Graph View*, que permite a visualização dos clusters formados em gráficos ilustrativos; e a opção *Centroid Plot View*, que permite visualizar o gráfico dos centróides dos documentos de cada *cluster* formado (figura 4-7). A parte superior do gráfico de centróides representa a legenda com cores diferenciadas de tracejados. Cada tracejado do gráfico representa a colocação dos centróides de seus respectivos *clusters*. Quanto maior a distância entre os centróides, compreende-se que mais bem definidos são os *clusters*, e assim, pôde-se utilizar o gráfico como suporte ao número de partições para a etapa de processamento.

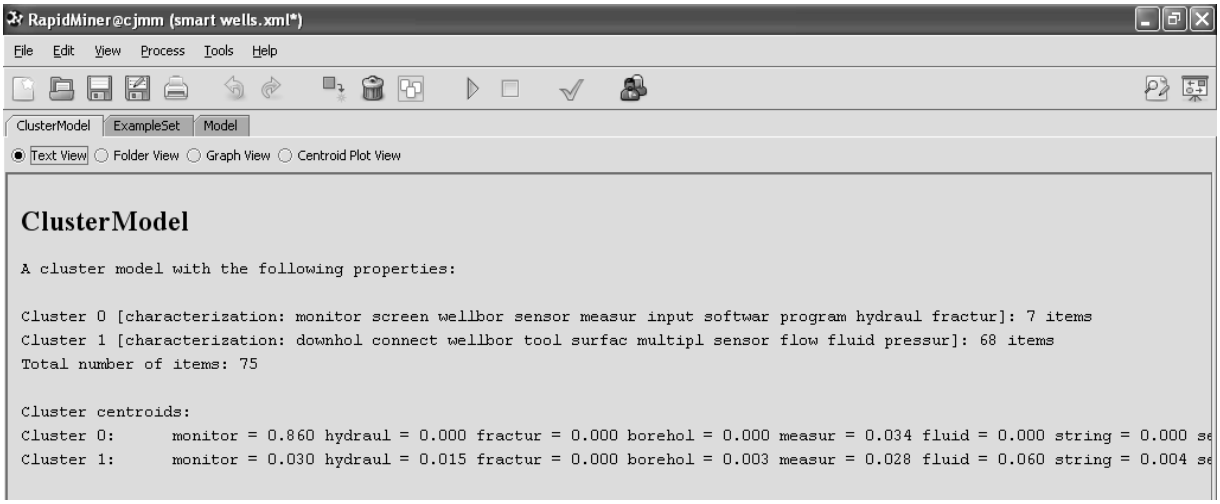

Figura 4-6 Tela dos resultados finais: Visualização *Text View*

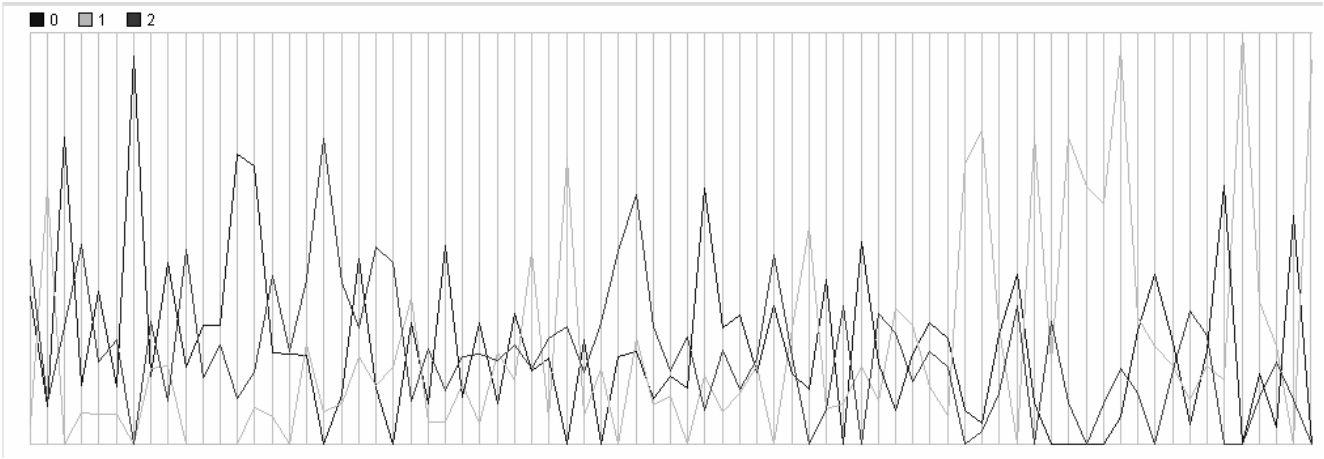

Figura 4-7 Exemplo de Gráfico de centróides dos documentos

4.5. *PolyAnalyst*

A plataforma *PolyAnalyst* possui como detentora de seus direitos de exploração e comercialização a empresa *Megaputer Intelligence* e encontra-se, atualmente, na versão 6.0. É um sistema que possui os processos necessários para mineração de dados, orientados para geração de Inteligência Competitiva e que executa uma larga variedade de métodos de forma mútua, complementando-se entre si para a análise de dados automática.

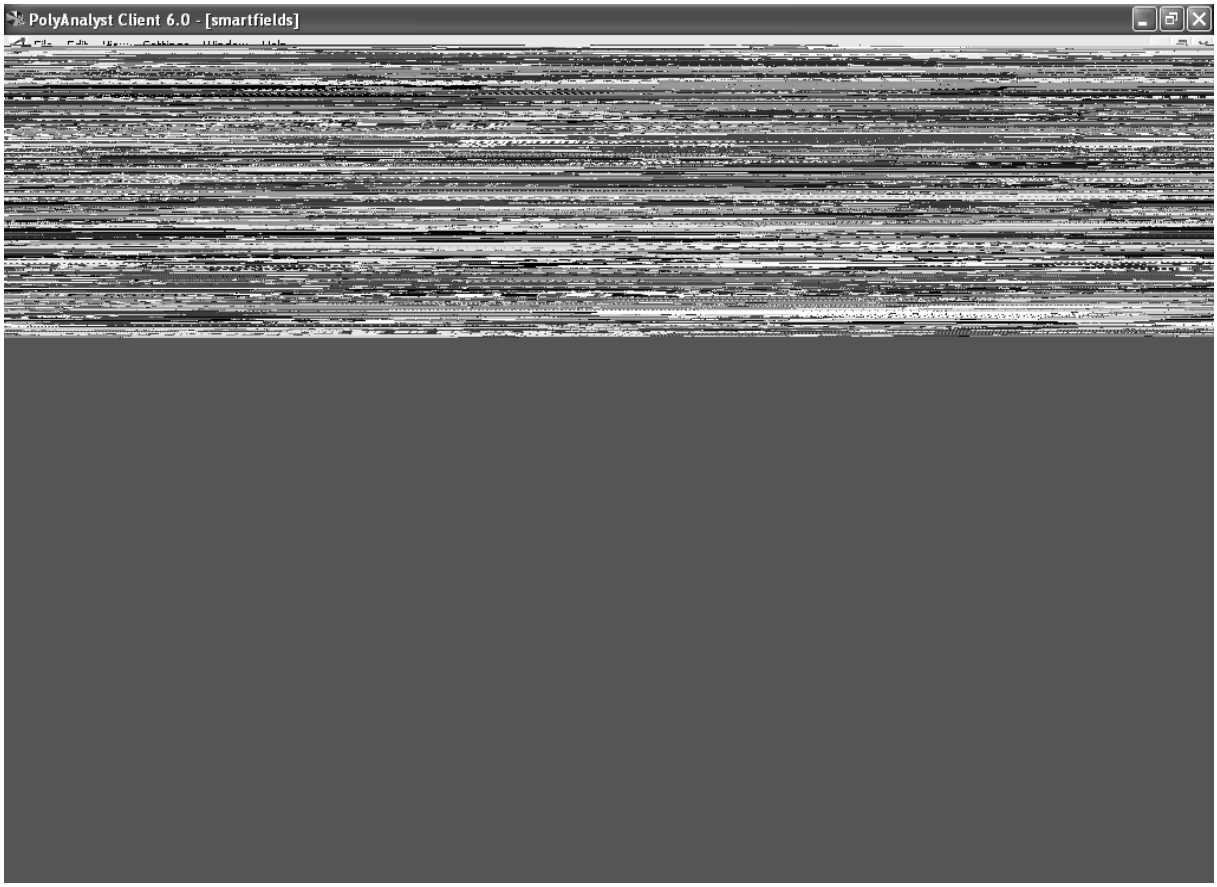

Figura 4-8 Área de trabalho de *PolyAnalyst* e exemplo de fluxograma para Mineração de Textos

4.5.1. Descoberta de Conhecimentos em Textos utilizando *PolyAnalyst*

O operador *Files* possui como função disponibilizar ao usuário a entrada de dados a serem submetidos aos processos de extração de conhecimentos. Sua saída é entrada do operador *Text Clustering*, que é responsável pelos processos inerentes às fases de pré-processamento – extração de palavras irrelevantes *(Stopwords)*, redução de palavras ao seu próprio radical *(Stemming)* e de extração dos símbolos e pontuações *(Tokening)*, bem como a utilização de *Thesaurus* para aprimoramento – e de processamento de textos – em nosso caso, Classificação Não-Supervisionada. Dessa forma, evidencia-se dizer que o fluxograma é organizado com o operador *Text Clustering* como operador central de todo o processo de Mineração de Textos, sendo as entradas dos demais operadores coligados em sua saída.

A plataforma disponibiliza um dicionário já elaborado de palavras que não agregam conhecimentos – o dicionário *WordNet*. Caso queira, o usuário pode alterá-lo, adicionando ou removendo palavras de seu escopo e salvando um clone modificado do dicionário, renomeando-o. Para gerenciar o dicionário, existe a opção *Dictionary Manager*, cuja tela pode se visualizar na figura 4-9. A figura 4-10 exibe a tela do dicionário explorado com algumas palavras presentes em seu escopo.

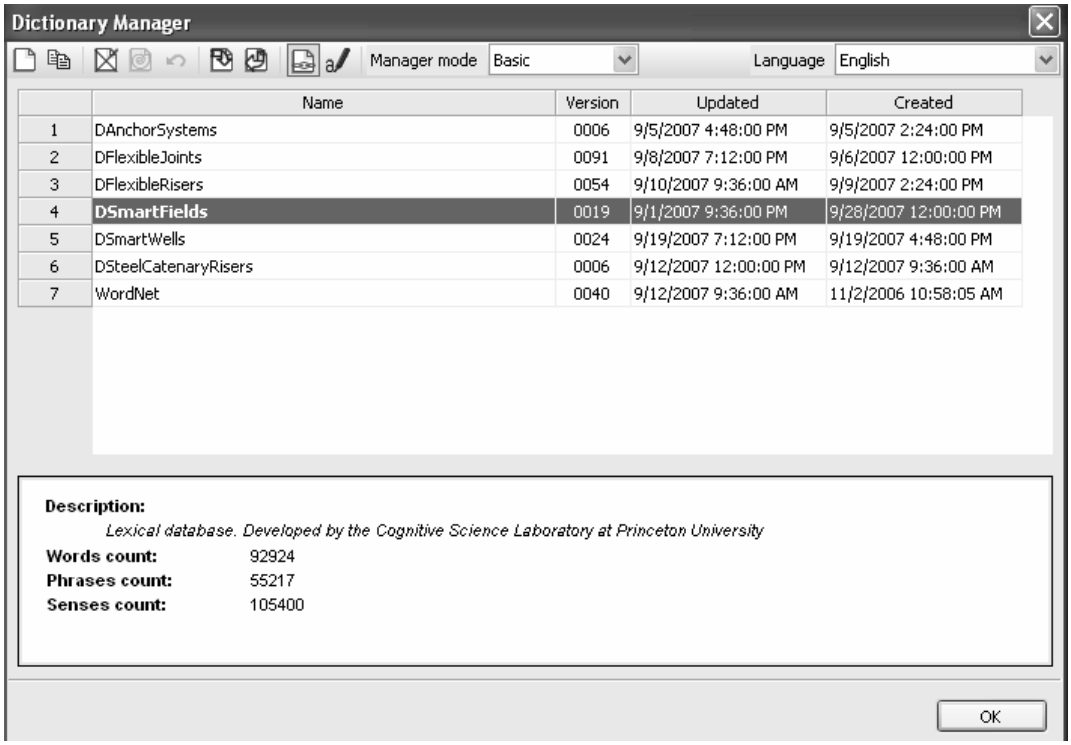

Figura 4-9 Tela do Gerenciador do Dicionário da plataforma *PolyAnalyst*

A debruçar-se sobre o processo de Classificação Não-Supervisionada, a plataforma utiliza a implementação do algoritmo *Suffix Tree Clustering – STC*, descrito por *Weiner* (1973). Seu funcionamento é descrito em três passos básicos.

No primeiro passo, o método constrói uma árvore de sufixos, contendo todos os sufixos em formato *string* de dados. Cada borda da árvore é rotulada com uma entrada não-vazia contendo uma parte da *string* de sufixos formada. O critério de compactação das

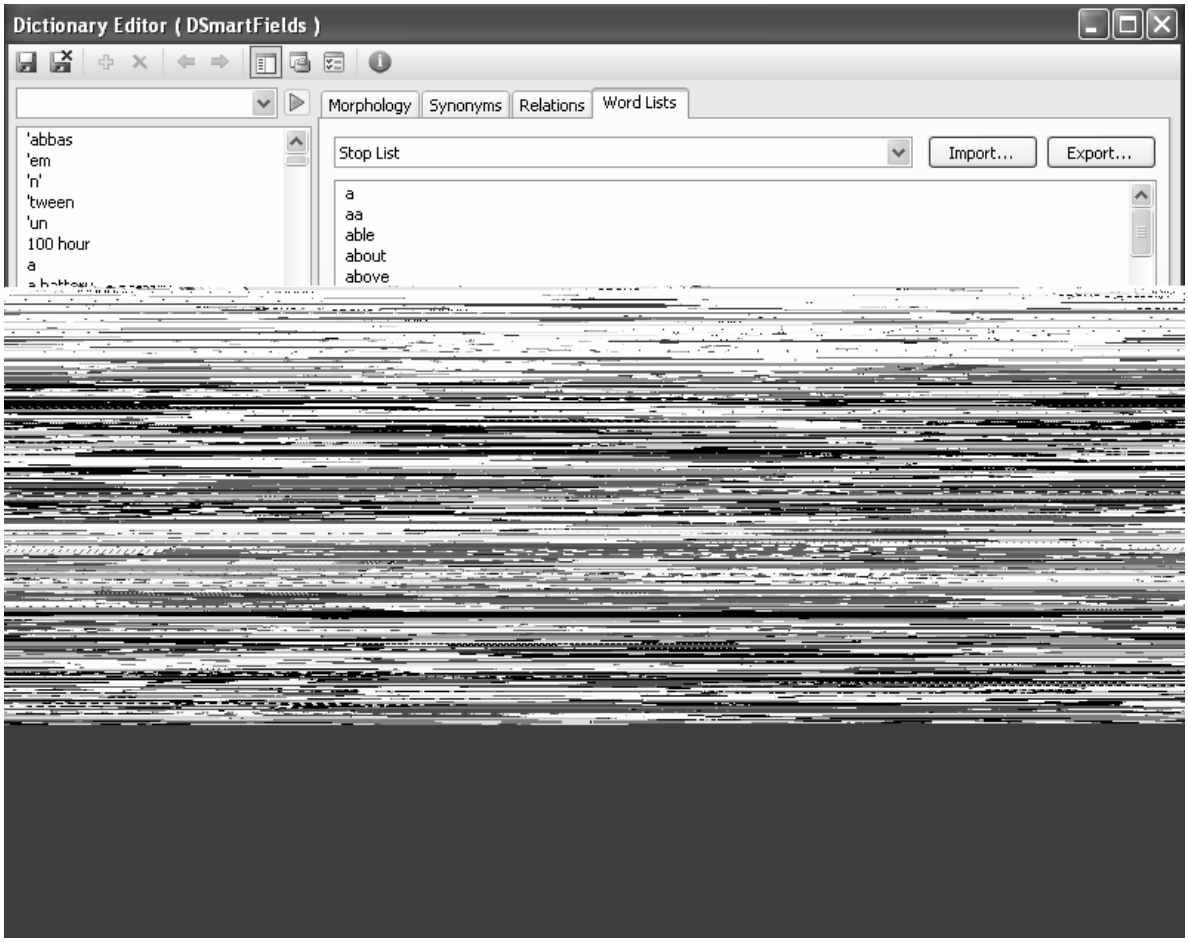

Figura 4-10 Tela do Dicionário da plataforma *PolyAnalyst* explorado

substrings diz que nenhuma dentre duas bordas fora do mesmo nó deve conter rótulos de borda que comecem com a mesma palavra. Para cada sufixo Q presente na *string* S, há um nó com etiqueta Q. Cada nó para um sufixo Q da *string* Z presente na *string* total S é etiquetado com índice da posição da *substring* Z, que é a *string* formada pela junção das demais substrings adjacentes, e com a posição de início do sufixo Q presente na *substring* formada Z. A partir desse primeiro passo, constrói-se a "árvore de sufixos" em um tempo de processamento linear. Um exemplo de árvore de sufixos pode se visualizar na figura 4-11.

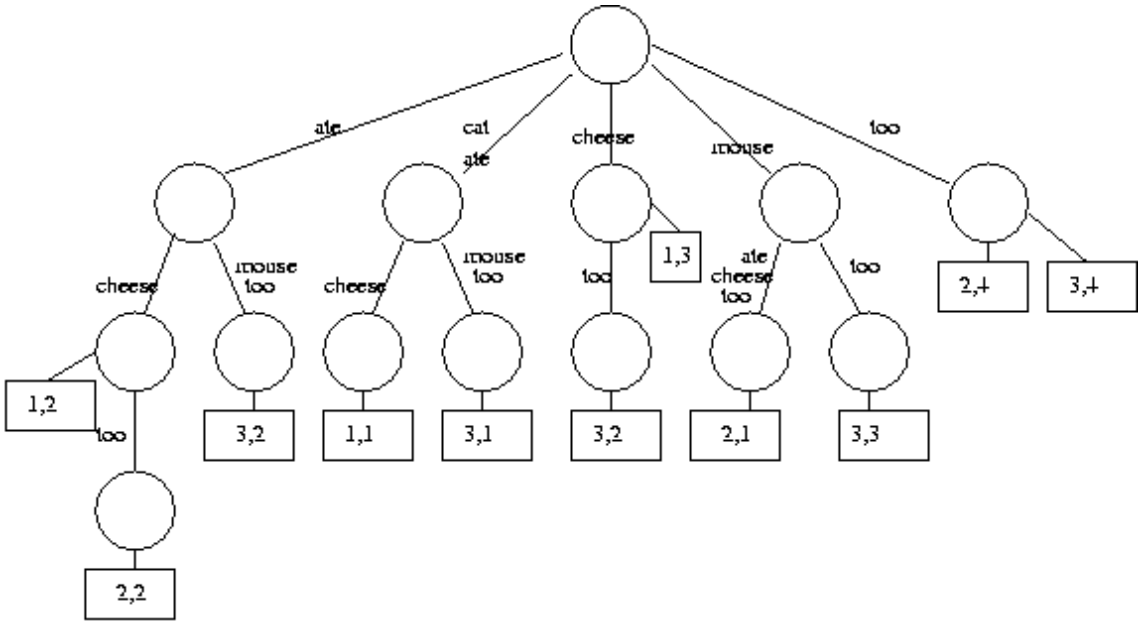

Figura 4-11 Exemplo de árvore de sufixo

No passo dois, para cada nó N presente na árvore de sufixos, o algoritmo forma e rotula como D(N) uma seleção de documentos na sub-árvore formada por cada nó N. Com isso, o processo forma e rotula uma frase com as junções dos nós N, denominando-a como P(N). Após esse processo, o coeficiente de N, que denomina-se c(N), é definido através da fórmula algébrica c(N) = |D(N)| * f(|P(N)|). f(1); considerando K o número de agrupamentos a serem formados; f(K) = K para K = 2 até 6; f(K) = 6 para K > 6.

No terceiro passo, constrói-se um grafo não-direcionado cujos vértices sejam nós da árvore do sufixo. Há um arco de N1 ao N2 se ambos os nós N1 e N2 estiverem dentre os 500 nós mais bem classificados segundo a pontuação obtida pela fórmula do coeficiente no passo dois; e se a proposição | D(N1) intersecção D(N2) | / (Maximo (|N1|,|N2|)) > 0,5 for satisfeita. Cada componente conectado desse grafo é um agrupamento. A pontuação do agrupamento varia de acordo com o cômputo das pontuações nos nós pertencentes ao mesmo, e o algoritmo retorna os dez agrupamentos de melhor pontuação. Com isso, a descrição do grupamento pode ser feita usando frases de seus nós ST.

No que se refere à sua parametrização, pode-se afirmar que existem parâmetros que o operador considera de suma importância para a obtenção de êxito ao final dos processos de *Clustering*.

O primeiro passo é a configuração do método que será empregado: o usuário pode escolher dentre os métodos de *Base Clustering, Exclusive Clustering ou Hierarchical Clustering*. Base Clustering é um método que desenvolve uma seleção de descrições de agrupamentos não-exclusivos utilizando o algoritmo *STC* padrão.

O método *Exclusive Clustering* compreende-se como uma extensão do método básico pelo fato de marcar uma segunda passada pelos agrupamentos já descobertos, de forma a verificar seus registros mais significativos. Se dois ou mais agrupamentos marcarem um registro mais significativo, o método exclusivo assinala o maior valor para o agrupamento com maior relevância sobre os demais. Caso tal condição não se satisfaça, o número de agrupamentos é podado. Alguns agrupamentos que são tão pequenos para encontrar o percentual mínimo de registros significativos são removidos e o texto pertencente para tais registros é realocado no próximo agrupamento de maior relevância. Pela sua consistência e pelos seus cuidados preservados a fim de indicar sempre os resultados realmente significativos ao final de cada convergência do algoritmo, elegeu-se tal método para os estudos de casos inerentes à pesquisa.

Por fim, o método *Hierarchical Clustering* principia que o algoritmo *STC* é não-exclusivo. Com isso, os agrupamentos são examinados para verificar se um agrupamento contém certa percentagem de registros de significância que caracterizem outro agrupamento. Se o ponto de partida for passado, o conjunto parcialmente ou completamente contido é representado como uma criança do conjunto contido dentro de uma hierarquia. Repete-se, então, o processo para todos os agrupamentos encontrados.

Tendo em vista que o método de *Clustering* fora escolhido, o passo seguinte foi

compreender os parâmetros requisitados pelo operador *Text Clustering* para submissão da base textual aos cuidados das etapas inerentes ao processo de mineração e extração de conhecimentos. Tais parâmetros podem se visualizar na figura 4-12.

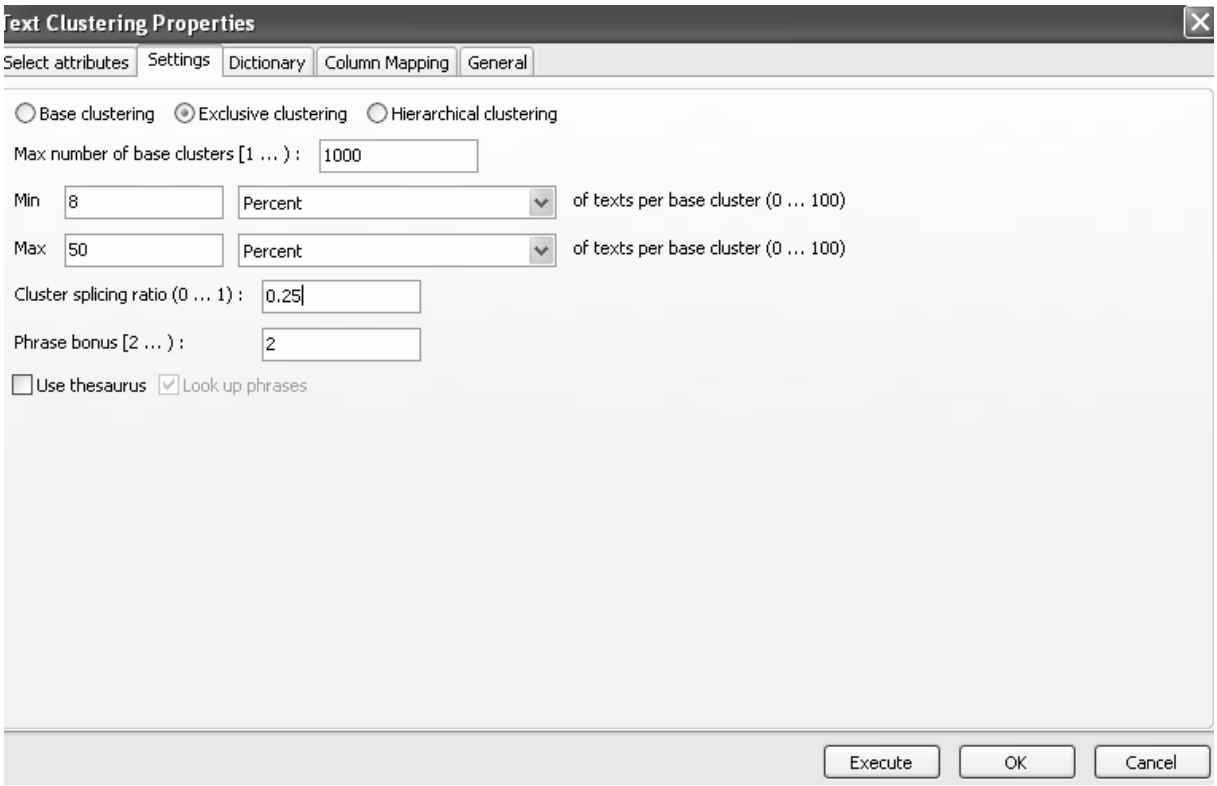

Figura 4-12 Tela dos parâmetros requisitados pelo operador *Text Clustering*

O parâmetro *"Max number of base clusters"* refere-se ao máximo número de frases individuais e palavras que são descobertas na primeira etapa do algoritmo *STC*. Abaixo dessa opção, o operador solicita que seja informado qual será o percentual mínimo e máximo de textos a serem considerados em cada agrupamento a ser descoberto. Evidencia-se, com isso, que se aumentar o valor mínimo, o número final de agrupamentos descobertos será menor, e vice-versa.

O parâmetro *"Cluster Splicing Ratio"* refere-se a um coeficiente (a) calculado segundo a equação (1) durante a fusão dos agrupamentos:

$$a = ||(C1) \cap (C2)|| / ||(C1) \cup (C2)|| \quad (1)$$

Considerando-se C1 como o conjunto de registros que contém uma determinada frase ou palavra-chave, e C2 como outro conjunto de registros que contém outra frase ou palavra, o operador realiza o devido cálculo inserindo tais valores na fórmula (1) exibida acima. Se o resultado do cálculo for maior do que o ponto inicial informado pelo usuário, as duas frases são fundidas dentro de um único agrupamento. Isso remete ao raciocínio de que valores baixos informados pelo usuário resultam em um número maior de agrupamentos ao final do processo, e vice-versa.

O parâmetro *Phrase Bonus* permite ao usuário informar ao operador se ele deseja que as frases de maior significância extraídas da base textual tenham o mesmo peso das palavras-chaves igualmente extraídas. Para isso, basta que o usuário informe valores elevados, de forma que as maiores frases serão favorecidas mais freqüentemente.

Por fim, a plataforma oferece um dicionário padrão de relacionamentos e de sinônimos entre termos, de modo que o usuário pode utilizá-lo caso deixe assinalada a opção *Use Thesaurus.*

De forma a se encontrar os resultados mais relevantes possíveis, os parâmetros tiveram seus valores diferenciados entre os estudos de casos que compõem a pesquisa. No entanto, alguns parâmetros foram comuns à todos, como, por exemplo, o parâmetro *"Max number of base clusters"*, que foi considerado com o valor 1000 igualmente para todos, e a utilização do thesaurus, tendo em vista melhorar o pré-processamento textual. Após o processo de *Clustering*, os resultados são enviados do operador *Text Clustering* para os demais operadores. O operador *Search Query* tem a função de realizar uma busca na base formada com os principais dados de cada documento (tais como títulos, palavras mais

representativas etc.) por um determinado termo que o usuário desejar. Os operadores *Phrase Extraction* e *Keyword Extraction* têm como funções extrair, respectivamente, as frases e as palavras mais significantes de cada cluster formado em cada base textual submetida aos processos, ao passo que o operador *Spell Check* realiza a checagem de cada palavra não-reconhecida pelo dicionário de idioma da plataforma, permitindo ao usuário as opções de substituí-la por termos ou expressões de significado semelhante e sugerido pelo *software*, ou simplesmente eliminá-la do processo.

Os operadores *Dataset Filter* e *Distinct* possuem funções semelhantes, que é a de realizar a filtragem sobre a base de dados formada após o processo de *Clustering*, eliminando linhas repetidas da tabela e exibindo, dentre outros atributos, o título de cada documento, seus termos mais significativos e a identificação do agrupamento no qual o documento pertence e gerando, inclusive, um gráfico para ilustração dos resultados.

5. Estudos de Casos

Em presente capítulo, serão apresentados as análises estatísticas e os estudos de casos relativos à aplicação do *software RapidMiner / YALE* nas etapas de pré-processamento e de processamento de dados textuais das patentes, tendo como finalidade a obtenção de conhecimentos válidos para prospecção tecnológica. Em etapa posterior, submeter-se-ão as mesmas bases de dados textuais aos processos de extração de conhecimentos com o *software PolyAnalyst version 6.0* para Mineração de Textos. Por fim, comparar-se-ão os resultados apresentados por cada um em cada base de dados textuais, de forma a mensurar seus desempenhos na tarefa de Mineração de Textos.

5.1. Bases de Dados Textuais

As bases de dados textuais utilizadas em presente dissertação foram, criteriosamente, definidas através de indicações feitas por profissionais ligados às pesquisas advindos da área de exploração de Petróleo e Gás. Consideraram-se temas indicados interessantes e em crescimento nos últimos anos dentro desse ramo da tecnologia. Com isso, cada tema pesquisado constituiu-se numa base diferenciada que recebeu, cada uma, o próprio nome do tema eleito para pesquisa, num total de seis temas diferenciados entre si, o que constituíram-se em seis estudos de casos.

Em cada caso, após a coleta dos documentos eletrônicos, fez-se uma leitura sobre cada patente retornada, a fim de se comprovar se sua natureza era inerente ou não ao tema pesquisado, evitando assim, a distorção dos resultados ao final e todas as etapas de Mineração de Textos.

Nos próximos subitens subseqüentes, serão apresentados, para cada base de dados, os critérios utilizados para os processos de busca e recuperação das patentes no *site* da *USPTO* e suas informações sobre as estatísticas primárias (número de documentos eletrônicos encontrados).

5.1.1. Base de Dados *Anchoring Systems*

Como termos-chaves, utilizaram-se *offshore*, *oil*, *vessel*, em conjunto com a expressão *anchor system (*ou *anchoring system)* presente em seu campo *ABSTRACT*. Definiu-se a composição da base textual de dados com 74 patentes. Encontrou-se a patente mais antiga registrada em 03 de fevereiro de 1976, e a mais recente, registrada desde 09 de outubro de 2007.

5.1.2. Base de Dados *Flexible Joints*

Como critérios para busca e coleta dos documentos, consideraram-se os termos-chaves *flexible joint* ou *higue* presentes no campo *ABSTRACT* das patentes contidas dentro do banco de dados da instituição norte-americana, associados à presença dos termos *platform* ou *vessel*; *oil* ou *petroleum* ou *gas* e *platform*; e *offshore* presentes em seus escopos. Definiu-se a composição da base textual de dados com 80 patentes. Encontrou-se a patente mais antiga registrada em 30 de novembro de 1976, e a mais recente, registrada desde 17 de julho de 2007.

5.1.3. Base de Dados *Flexible Risers*

Como termos-chaves para consulta, utilizaram-se *oil* ou *petroleum* e a expressão *flexible riser*. Foram encontradas 127 patentes com este tema, sendo a patente mais antiga registrada em 27 de abril de 1976, e a mais recente, registrada desde 30 de outubro de 2007.

5.1.4. Base de Dados *Smart Fields*

Para a pesquisa, utilizou-se a presença de termos-chaves contidos nos campos *ABSTRACT* e *DESCRIPICTION* dos documentos eletrônicos hospedados no site da instituição norte-americana *USPTO* tais como: *"smart field"*, *"intelligent field"*, *"artificial rise"*, *"production"*, *"controlling"*, *oil* ou *petroleum*. Como resultado, resumiu-se a base composta em um número total de 195 patentes coletadas, sendo a mais antiga registrada em 07 de junho de 1983, e a mais recente, registrad

5.1.6. Base de Dados *Steel Catenary Risers*

Um total de 27 patentes foi encontrado na pesquisa feita no site da instituição, obedecendo ao termo-chave "*steel catenary riser*". Considerou-se a presença destes termos nos campos *ABSTRACT, DESCRIPTION e TITLE* das patentes. A patente encontrada mais antiga está registrada em 14 de dezembro de 1993, e a mais recente está registrada em 27 de março de 2007.

5.2. Análise Estatística

Nesta seção, procurou-se aprofundar os estudos no tocante aos levantamentos estatísticos sobre as bases de patentes coletadas, de modo a verificar suas CLASSIFICAÇÕES e seus INVENTORES / CESSIONÁRIOS de destaque. Para estes estudos, algumas considerações preliminares são válidas e aplicáveis a todos:

1. Consideraram-se as patentes cujos registros foram efetuados no site da instituição a partir do ano de 1976 até 30 de novembro de 2007. Portanto, vale essa ressalva, considerando-se o fato de que, devido ao dinamismo da *Internet*, possa haver alguma patente registrada hoje – além dessa data – que obedeça aos critérios de pesquisa e que não esteja agregada ao escopo das bases de dados textuais formadas inerentes às pesquisas elaboradas;
2. A data de registro que consta no documento eletrônico da patente contida no site da instituição *USPTO* é, na verdade, a data em que o instrumento encontra-se em vigor, podendo ser o primeiro registro ou uma renovação de direitos e de validade sobre o mesmo;

3. A grande maioria das patentes pesquisadas possui os campos *INVENTOR* (INVENTOR) e *ASSIGNEE* (CESSIONÁRIO) preenchidos, respectivamente, com o(s) nome(s) da(s) pessoa(s) física(s) que realizaram o depósito e o nome da empresa, instituição ou organização (pessoa jurídica) que ganhou os direitos exclusivos de exploração do texto que reza o documento. Neste caso, considerou-se, para os levantamentos estatísticos, o nome da instituição (pessoa jurídica) contida no campo *ASSIGNEE* como representante legal do documento depositado;
4. Para a minoria dos casos em que nas patentes não consta nome algum de pessoa jurídica no campo *ASSIGNEE*, adotou-se, para a pesquisa estatística, o nome do inventor – contido no campo *INVENTOR* - como representante legal do instrumento depositado;
5. Na maior parte dos casos, as patentes coletadas apresentam mais de uma classificação internacional em seu campo CIP. De modo a não prejudicar os resultados das análises, todas as classificações internacionais presentes foram consideradas;
6. As descrições de cada classificação internacional de maior incidência presentes nas patentes em cada base textual de dados ao final dos levantamentos estatísticos sobre os campos CIP permitem a previsão sobre tecnologias e materiais que podem estar presentes em seus registros documentais.

Apresentar-se-ão, nos próximos subitens, os resultados das análises estatísticas para cada base textual de dados nos próximos itens subseqüentes.

5.2.1. *Anchoring Systems*

Em um estudo estatístico preliminar elaborado sobre a base textual de dados definida seguindo os critérios de pesquisa já comentados para este presente tema, averiguou-se que os anos de 2006 e 2007 apresentaram os maiores índices de patentes registradas no banco de dados da instituição (*USPTO*): 19 e 16 respectivamente. Este fato corresponde a um índice de 46,05% do total de patentes do conjunto estudado. O gráfico relativo ao estudo pode ser visualizado na figura 5-1.

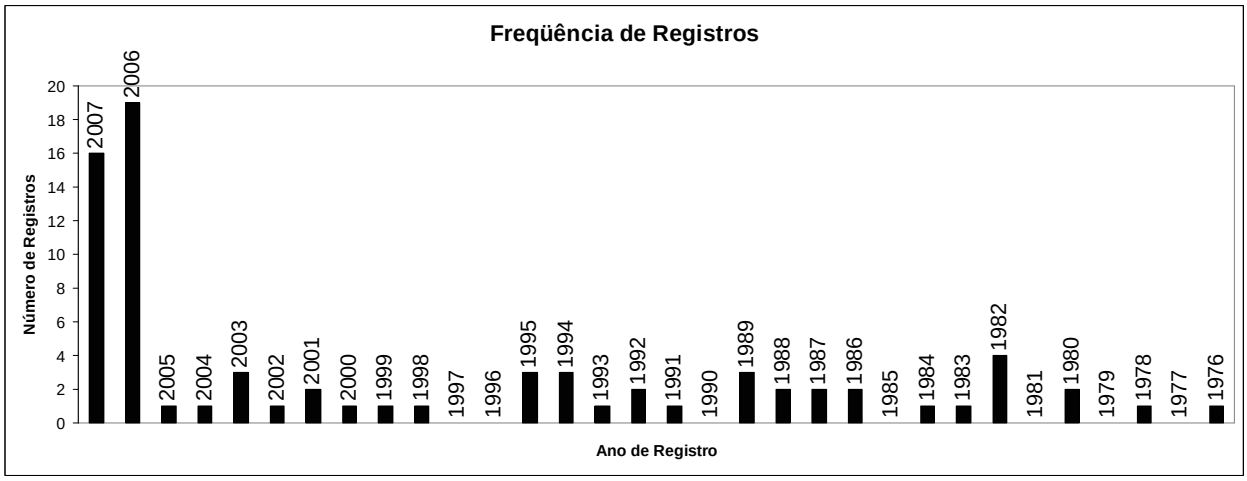

Figura 5-1 Gráfico de Freqüência de Registros das patentes da BD *Anchoring System*

5.2.1.1. Classificações

Os resultados da análise das classificações internacionais sobre as patentes contidas na base indicam a existência de várias subclasses dentro das classes B63B e E21B. O gráfico apresentado na figura 5-2 permite a visualização de cada classificação encontrada com o número de registros relativos a cada classificação.

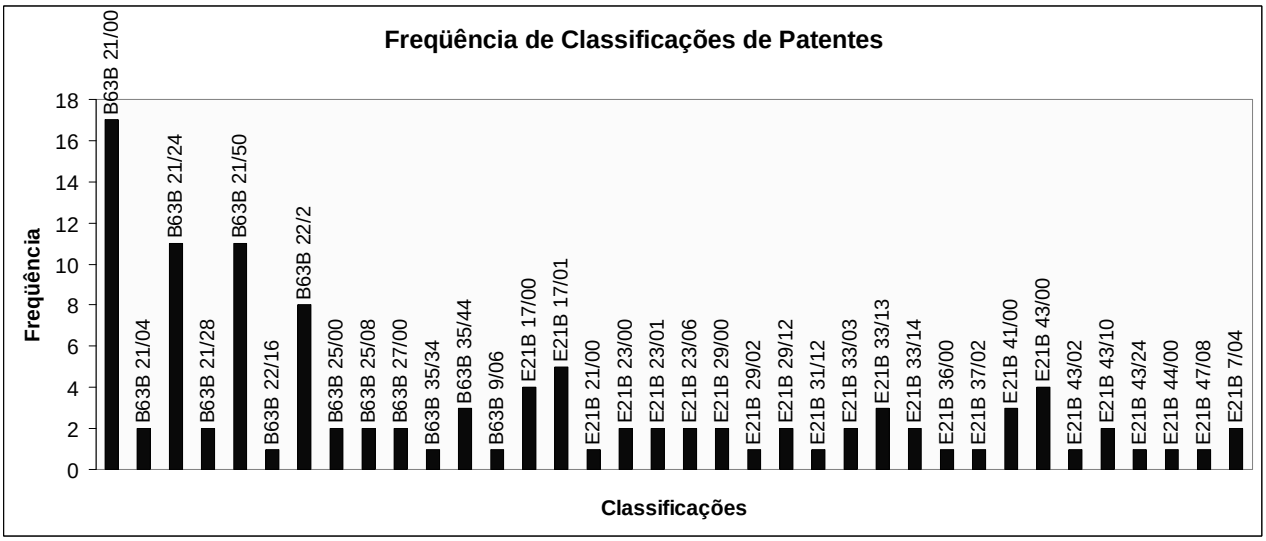

Figura 5-2 Gráfico de Freqüência de Classificações das patentes da BD *Anchoring System*

A partir da análise do gráfico apresentado, algumas considerações importantes são passíveis de serem comentadas:

- As classificações B63B 21/00, B63B 21/24, B63B 21/50 e B63B 22/02 apresentaram os maiores números de registros perante as demais encontradas para esta base de dados. Suas descrições seguem abaixo para melhor compreensão dos resultados no tocante à análise de contexto sobre os dados e as informações úteis que estão contidas neles:
    - Seção B: Operações de processamento; transporte;
    - Subseção 63: Navios ou outras embarcações;
    - Classe B: Equipamento para navegação; Amarração;
        - → Subclasse B63B 21/00: Equipamento para deslocar, rebocar ou empurrar; Ancoragem;
        - → Subclasse B63B 21/24: Âncoras;
        - → Subclasse B63B 21/50: Disposições para ancoragem de embarcações especiais, por exemplo, para plataformas flutuantes de perfuração ou dragas;

→ Subclasse B63B 22/02: Bóias; especialmente adaptadas para amarração de embarcações;

- Comprovou-se, além disso, que as classificações E21B apresentaram um equilíbrio, com relação ao número de patentes registradas, em seus subgrupos encontrados, de forma que as descrições dos subgrupos mais freqüentes presentes na base de dados explorada encontram-se abaixo:
    - Seção E: Construções Fixas;
    - Subseção 21: Perfuração do solo; Mineração;
    - Classe B: Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços;
        → Subclasse E21B 17/00: Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção (engates de hastes em geral F16D; tubos ou acoplamentos de tubos em geral F16L);
        → Subclasse E21B 17/01: Tubos ascendentes (conectores de tubos ascendentes);
        → Subclasse E21B 43/00: Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços (aplicáveis somente à água E03B; obtenção de jazidas petrolíferas ou de matérias solúveis ou fundíveis por técnicas de mineração E21C 41/00; bombas F04);

- Além disso, e para uma melhor compreensão, averiguou-se o significado das classificações E03B, E21C 41/00, F04; F16D E F16L encontradas em algumas das descrições das classificações citadas acima:
    - E03B: Instalações ou Métodos para obter, coletar ou distribuir água (perfuração de poços, obtenção de líquidos em Geral de poços E21B; Sistemas de Canalizações em geral F17D – Sistemas de tubulação; Dutos);
    - Subclasse E21C 41/00: Mineração ou exploração de pedreiras; Métodos de mineração subterrânea ou de superfície;
    - F04: Engenharia Mecânica, Iluminação; Aquecimento; Armas; Explosão; Máquinas de deslocamento positivo a líquidos; bombas para líquidos ou fluídos elásticos;
    - F16D: Acoplamentos para transmissão de rotação; embreagens; freios;
    - F16L: Tubos; Juntas ou acessórios para tubos; Suportes para tubos, cabos ou tubulação de proteção; meios para isolamento térmico em geral.

5.2.1.2. Inventores / Cessionários

Em uma análise geral, encontraram-se 52 inventores / cessionários diferentes de patentes sobre *Anchoring Systems*, concedidas pela instituição *USPTO*. Dentre eles, comprovou-se que a maioria – 78,85% do número geral são representantes legais de somente uma patente nos registros do respectivo banco de dados. O gráfico apresentado na figura 5-3 ilustra os resultados encontrados no tocante a tal estudo elaborado.

Dentre as instituições que possuem um número significante de patentes concedidas, destacam-se a *Secretary of the Army (Washington, DC)* do governo dos Estados Unidos da América, com 06 patentes, e a *Baker Hughes Incorporated (Houston,*

TX), com 05 patentes. Dentre outras empresas com um número de concessões obtidas acima de duas, estão: *Brown & Root, Inc. (Houston, TX), Halliburton Energy Services, Inc. (Houston, TX), Shell Offshore Inc. (Houston, TX) e Weatherford/Lamb (Houston, TX)*, cada uma com 03 patentes cedidas pela *USPTO* e registradas em seu banco de dados.

A pessoa física *Korsgaard; Jens (Princeton Junction, NJ)* aparece na análise estatística como representante legal de 03 patentes. As instituições *Diamant Boart France-SA-Division Petrole (Serres Castet, FR), Exxon Production Research Co. (Houston, TX),* e a empresa brasileira Petróleo Brasileiro S/A – Petrobrás (BR) aparecem na relação com 02 patentes cedidas pela *USPTO* a cada uma.

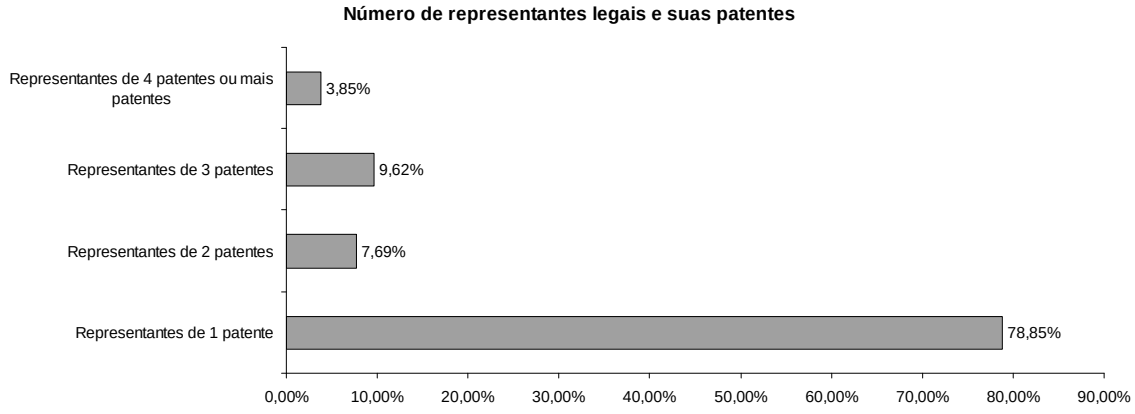

Figura 5-3 Gráfico de Inventores / Cessionários das patentes da BD *Anchoring Systems*

5.2.2. *Flexible Joints*

A partir do levantamento estatístico preliminar elaborado sobre os documentos eletrônicos encontrados e que constituem a base textual de dados em questão, gerou-se um gráfico (figura 5-4) que apresenta os resultados dos registros das patentes no banco de dados da instituição norte-americana ano após ano, desde 1976 até 2007.

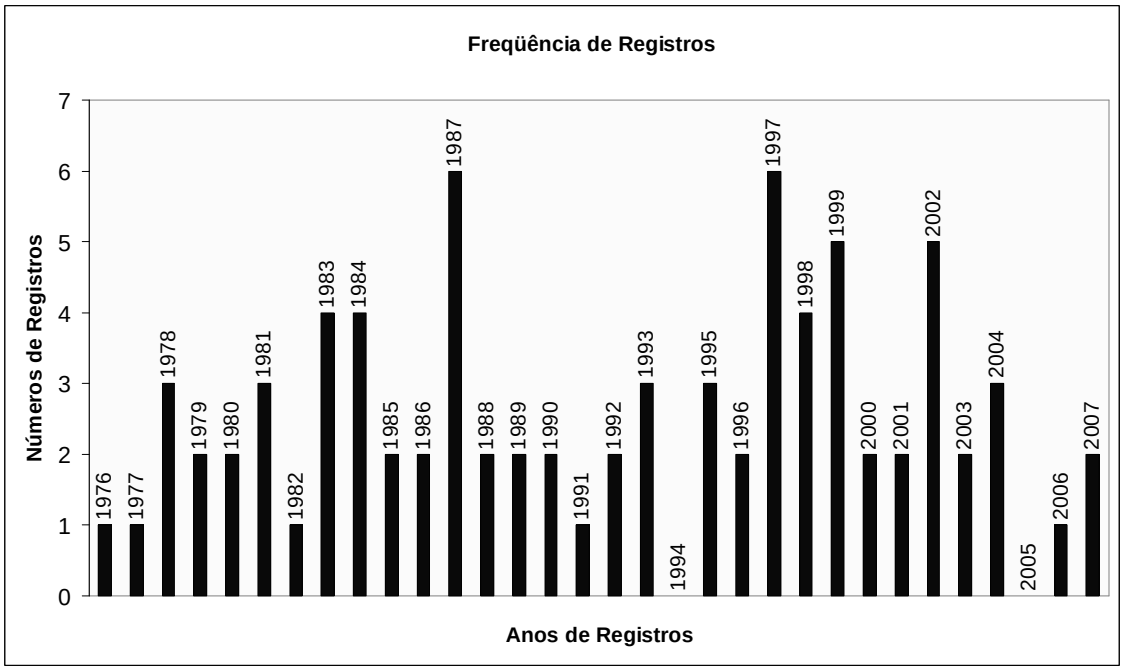

Figura 5-4 Gráfico de Freqüência de Registros das patentes da BD *Flexible Joints*

A debruçar-se sobre seus resultados, algumas constatações, em princípio, podem ser feitas e que refletem o equilíbrio das freqüências anuais das concessões das patentes relacionadas:

- No triênio 1997 a 1999, a instituição concedeu um total de quinze patentes, o que representa um índice de 18,75% do total de patentes presentes na base;
- De 1993 até 2007, foram encontradas quarenta patentes concedidas pela instituição, o que representa 50% do total de patentes contidas na base formada concedidas pela instituição nos últimos quinze anos;
- A maioria dos anos – 13 de um total de 32 – registrou um total de duas patentes concedidas pela instituição americana;
- Os anos 1987 e 1997 apresentaram seis patentes registradas em cada um e foram os anos com maior número de registros de patentes relativo a esse tema, ao passo que os anos de 1994 e de 2005 não apresentaram nenhum novo registro no banco de dados da instituição norte-americana.

5.2.2.1. Classificações

Os resultados de cada classificação internacional encontrada para a base de dados em questão com seus respectivos números de patentes registradas podem ser visualizados no gráfico gerado (figura 5-5).

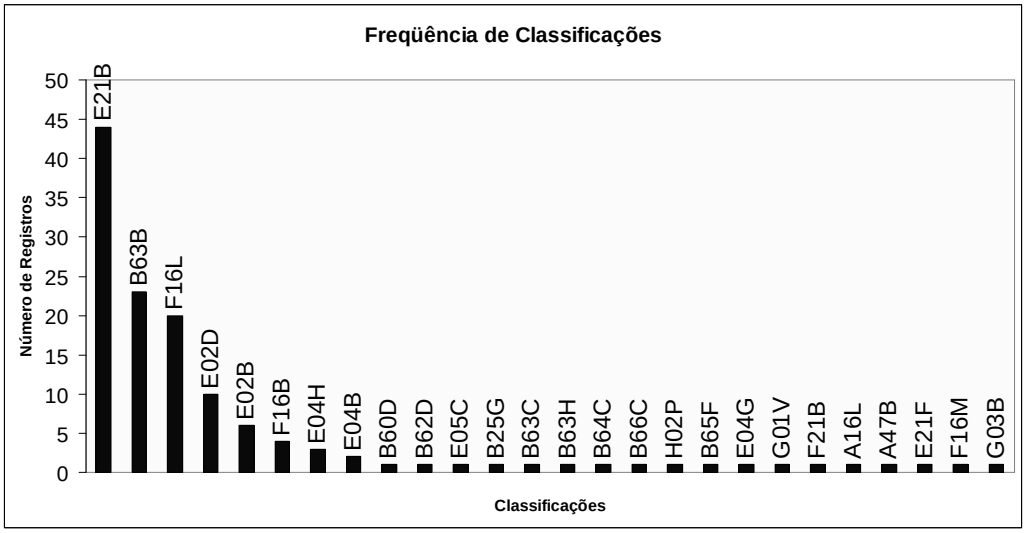

Figura 5-5 Gráfico de Freqüência de Classificações das patentes da BD *Flexible Risers*

Do número total de 26 diferenciadas seções de classificações internacionais encontradas nas patentes da base, 18 delas possuem somente uma única patente registrada, o que corresponde a um índice de 69,23% do número total de seções presentes. Três dessas seções merecem destaque em relação às demais, pela alta freqüência de patentes registradas: E21B, B63B e F16L, com 43, 23 e 20 registros, respectivamente. As seções E21B e B63B já foram descritas anteriormente (subitem 5.2.1.1), bem como algumas subclasses a elas pertencentes. Desse modo, descrever-se-ão, a seguir, a seção F16L e sua subclasse mais freqüente, além da subclasse B63B 22/00:

- Seção F: Tubos; Juntas ou acessórios para tubos;
    - Subseção 16: Suportes para tubos, cabos ou tubulação de proteção;

- Classe L: Meios para isolamento térmico em geral;
  - → Subclasse F16L 27/02: Juntas universais, isto é, com ligação mecânica permitindo movimento angular ou ajustagem dos eixos da peça em qualquer direção.

- Seção B: Operações de processamento; transporte;
  - Subseção 63: Navios ou outras embarcações;
  - Classe B: Equipamento para navegação; Amarração;
    - → Subclasse B63B 22/00: Bóias; meios para indicar a localização de objetos submersos.

5.2.2.2. Inventores / Cessionários

Em um levantamento estatístico sobre os inventores / cessionários, encontraram-se 55 representantes legais diferentes de patentes inerentes ao tema *Flexible Joints*, concedidas pela instituição norte-americana. Dentre eles, averiguou-se que a maioria – 78,18% do total – é representante de somente uma patente em nos registros do respectivo banco de dados. Já o percentual de representantes de duas patentes sobre o mesmo assunto é de 14,55%. O gráfico 5-6 exibe os resultados obtidos deste levantamento estatístico.

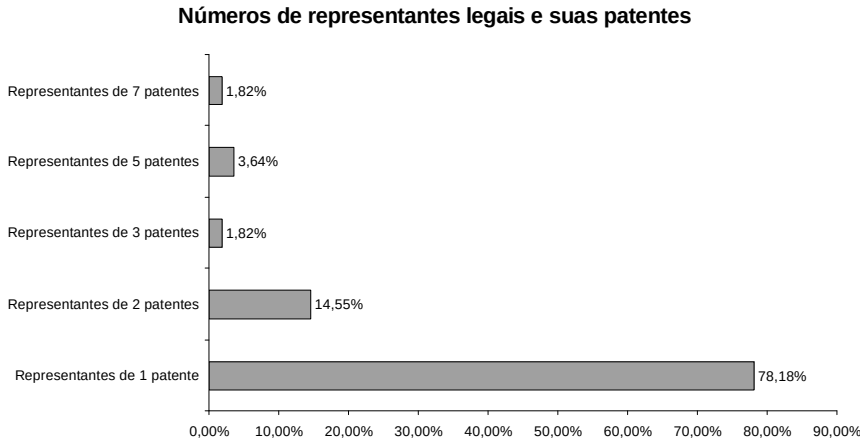

Figura 5-6 Gráfico de Inventores / Cessionários das patentes da BD *Flexible Joints*

Dentre as instituições que possuem um número significante de patentes concedidas, destacam-se as empresas *Vetco Offshore Industries, Inc. (Ventura, CA)*, com 07 patentes; e as empresas *Shell Oil Company (Houston, TX)* e *Hydril Company (Houston, TX)*, com 05 patentes concedidas a cada uma.

5.2.3. *Flexible Risers*

Seguindo os critérios de pesquisa já comentados para este presente tema, elaborou-se um estudo estatístico inicial sobre as patentes coletadas a respeito do tema proposto nesta seção. A figura 5-7 apresenta o gráfico representante dos registros feitos no banco de dados da instituição.

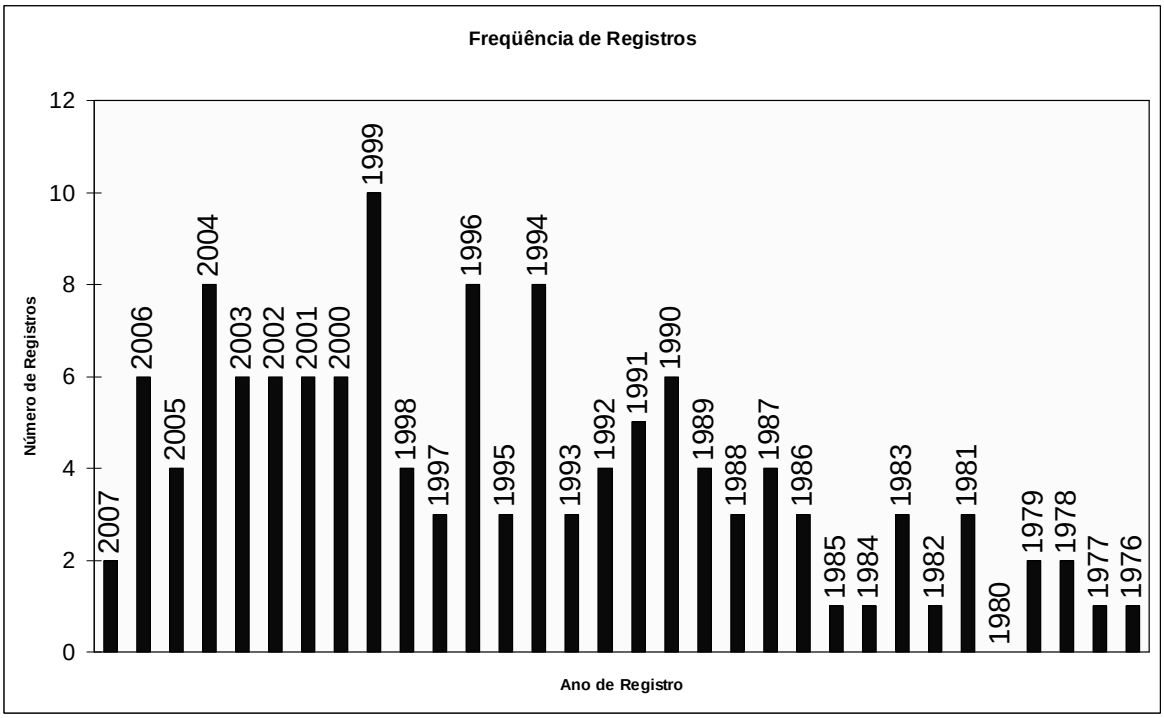

Figura 5-7 Gráfico de Freqüência de Registros das patentes da BD *Flexible Risers*

A debruçar-se sobre seus resultados, podem-se constatar tais dados estatísticos primários:

- Das 127 patentes relativas ao assunto, 61 foram registradas nos últimos dez anos, o que equivale a um índice de 48,03% do total dos documentos encontrados;
- A década de 1980 teve 23 patentes relativas ao assunto registradas (18,11% do número total dos documentos pesquisados), contra 52 patentes da década de 1990 (40,94% do número total dos documentos pesquisados). Porém, a considerar somente do ano 2000 até o presente momento, 44 patentes relativas ao assunto foram registradas (34,64% do número total das patentes);
- Esta tendência de crescimento detectou-se sob os estudos estatísticos na medida em que consideramos o fato de que o ano de 1999 – último ano da década de 1990 – foi o que mais recebeu patentes registradas no banco de dados da instituição e relacionadas ao assunto em questão. Naquele ano, constatou-se um total de 10 patentes registradas: o maior índice de registro (7,87% do número total das patentes pesquisadas), não somente da década de 1990, como também de todos os anos considerados.

5.2.3.1. Classificações

Para ilustrar os resultados, gerou-se um gráfico (que pode ser visualizado na figura 5-8) que confronta cada classificação encontrada com o seu número de registros.

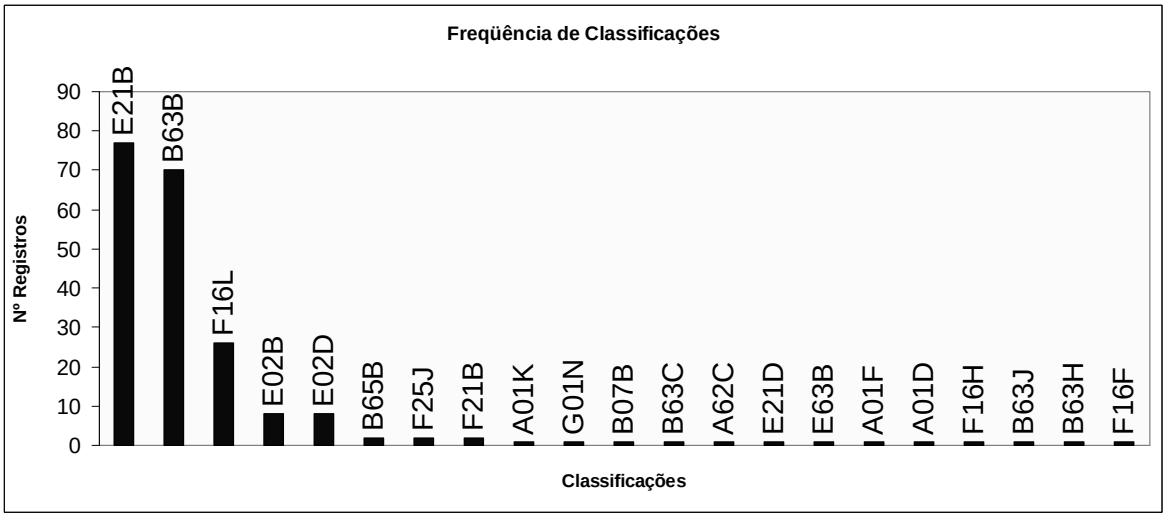

Figura 5-8 Gráfico de Freqüência de Classificações das patentes da BD *Flexible Risers*

Das classificações internacionais presentes nas patentes desta base de dados, podem-se destacar como as três de maior freqüência:

- E21B: presente em 60,63% do total de patentes coletadas para esta base;
- B63B: presente em 55,12% do total de patentes coletadas para esta base;
- F16L: presente em 20,47% do total de patentes coletadas para esta base.

As classificações E21B, B63B e F16L encontram-se descritas em subitens anteriores (5.2.1.1 e 5.2.2.1). O restante das classificações encontradas oferece índices menores. Muitas delas a citar: as classificações A01K, G01N, B07B, B63C, A62C, E21D, E63B, A01F, A01D, F16H, B63J, B63H e F16F, oferecem uma freqüência de registro de apenas 0,79%, das patentes totais encontradas, e por isso suas respectivas descrições foram desconsideradas.

5.2.3.2. Inventores / Cessionários

Em uma análise geral, encontraram-se 73 inventores / cessionários diferentes de patentes sobre *Flexible Risers*, concedidas pela instituição *USPTO*. Dentre eles,

comprovou-se que a maioria – 68,49% do número geral – é representante legal de somente uma patente nos registros do respectivo banco de dados. Já o percentual de inventores / cessionários de duas patentes sobre o mesmo assunto é de 20,47%. A figura 5-9 exibe um gráfico que representa os resultados obtidos em virtude do levantamento estatístico.

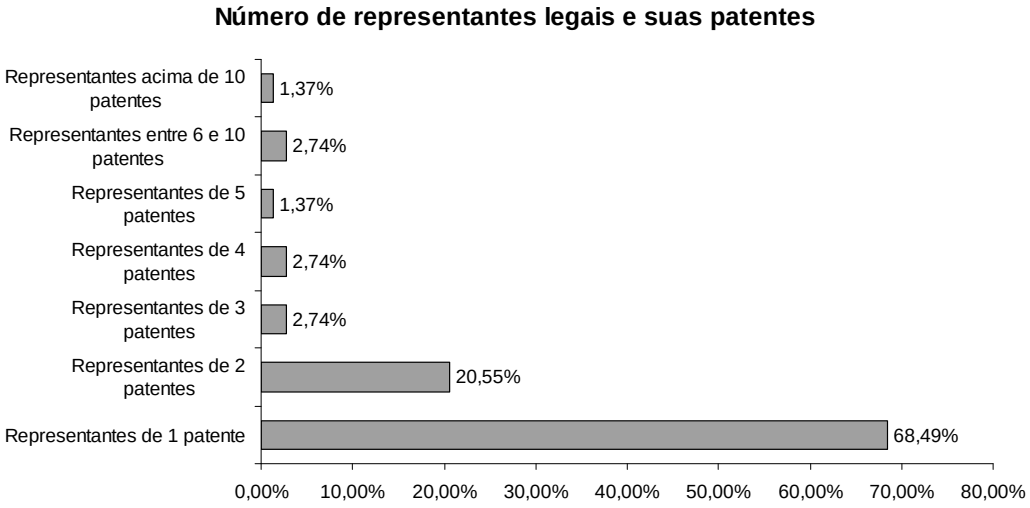

Figura 5-9 Gráfico de Inventores / Cessionários das patentes da BD *Flexible Risers*

Dentre as instituições que possuem um número significante de patentes concedidas, destacam-se as empresas *Den Norske Stats Oleselskap A.S* da Noruega, com 11 patentes, e a *Institut Francais du Petrole* da França, com 09 patentes. Dentre outras empresas com um número de concessões obtidas acima de duas, estão: *SOFEC, Inc. (Houston, TX)* com 06 patentes; *Single Buoy Moorings Inc. (Marly, CH)* com 05 patentes; *FMC Corporation (Chicago, IL)* com 04 patentes; e a *Shell Oil Company (Houston, TX)* com 03 patentes. Duas pessoas físicas aparecem na relação como representantes legais de patentes: *Headworth; Colin Stuart (Houston, TX)* com 04 patentes; e *Head; Philip (London, NW10 7XR, GB)*, com 03 patentes.

A empresa brasileira Petróleo Brasileiro S/A – Petrobrás (BR) aparece na relação como a única de nosso país a obter a concessão de patentes (03 no total) da *USPTO* sobre o tema tratado em presente item.

5.2.4. *Smart Fields*

A análise estatística elaborada para presente estudo de caso indica que 175 das 193 patentes coletadas para este tema específico foram registradas de janeiro de 1998 até novembro de 2007, o que indica um índice representativo de 90,67% dos registros textuais presentes na base. A debruçar-se sobre tal índice, evidencia-se, claramente, o crescimento acentuado desse assunto perante a indústria de exploração de Petróleo e Gás. Os anos que apresentaram os maiores números de patentes registradas foram os de 2006 e 2007 com, respectivamente, 41 e 31 registros, que somados, representam 40,57% do conteúdo total registrado nos últimos dez anos, ratificando a importância de sua exploração. Para representar de forma visual estes dados, elaborou-se um gráfico que pode ser visualizado na figura 5-10.

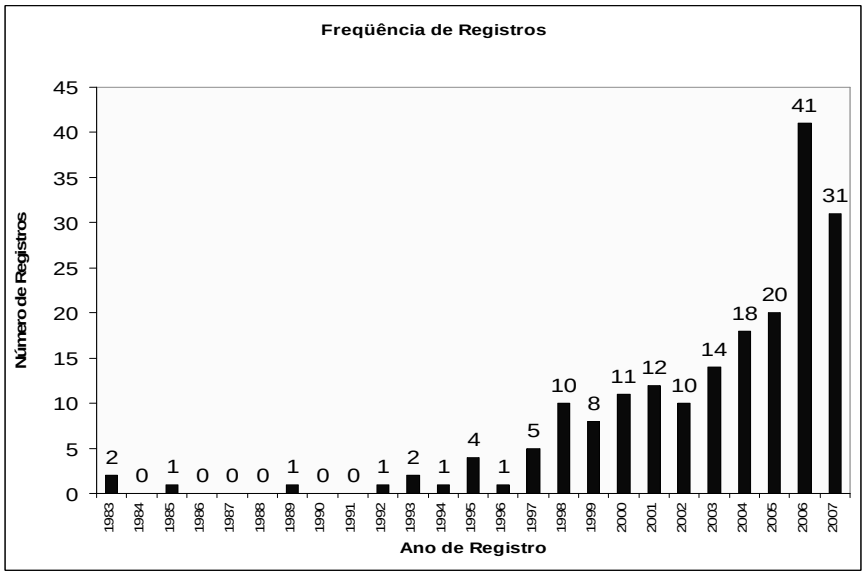

Figura 5-10 Gráfico de Freqüência de Registros das patentes da BD *Smart Fields*

5.2.4.1. Classificações

De forma a contemplar-se os resultados dos levantamentos estatísticos sobre as classificações internacionais presentes, gerou-se um gráfico ilustrativo que pode ser visualizado na figura 5-11.

A debruçar-se sobre o gráfico, evidencia-se que as classificações internacionais G06F e G05B são as mais freqüentes, de forma que a soma da quantidade de registros das duas representam 65,76% do total da base, merecendo destaque em relação às demais. Procurando-se compreender tais classificações de destaque de uma maneira mais profunda, de forma a buscar conhecimento sobre tipos de atividades e material a que cada uma se refere, realizou-se uma pesquisa sobre o manual atualizado com a nova versão da Classificação Internacional de Patentes (CIP) e relataram-se suas descrições:

- Seção G: Física;
    - Subseção 05: Controle; Regulagem;
    - Classe B: Sistemas de controle ou regulagem em geral; elementos funcionais de tais sistemas; disposições de monitoração ou de teste para tais sistemas ou elementos (acionadores movidos por pressão de fluido ou sistemas que atuam por meio de fluídos em geral; válvulas; caracterizados por elementos mecânicos; elementos sensíveis; elementos de correção);
- Seção G: Física;
    - Subseção 06: Cômputo; Cálculo; Contagem (computadores de contagem para jogos; combinações de instrumentos de escrita com dispositivos de computação);

- Classe F: Processamento elétrico de dados digitais (computadores em que parte da computação é efetuada hidráulica ou pneumaticamente; equipamentos periféricos independentes de entrada e saída; redes de impedância utilizando técnicas digitais).

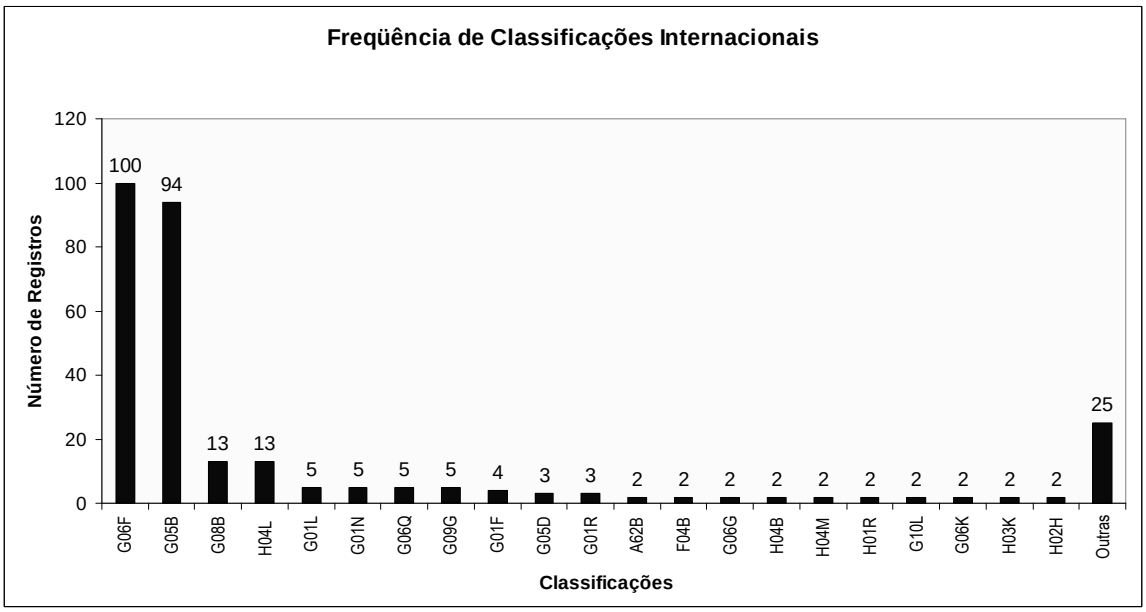

Figura 5-11 Gráfico de Freqüência de Classificações das patentes da BD *Smart Fields*

5.2.4.2. Inventores / Cessionários

Em virtude do levantamento estatístico sobre o campo específico pertencente às patentes concedidas pela *USPTO* e que compõem a base textual de dados formada, constatou-se um número total de 42 inventores / cessionários diferentes para o tema *"Smart Fields"*. Dentre os cessionários, verificou-se que a maioria – 57,14% do número geral – são representantes legais de somente uma patente. Já o percentual de inventores / cessionários de duas patentes sobre o mesmo assunto é de 14,29%.

Alguns cessionários, no entanto, merecem destaque perante os demais: a instituição *Fischer Rosemount Systems, Inc. (Austin, TX)*, por exemplo, foi a que obteve o maior número de patentes concedidas pela instituição norte-americana: no total, foram

encontrados 88 registros de documentos eletrônicos tendo seu nome como representante legal. As instituições *Rosemount Inc. (Eden Prairie, MN)* com 14 registros, *Invensys Systems, INC (Foxboro, MA) e Smar Research Corporation (Holbrook, NY)* com 12 registros cada uma, também são merecedoras de destaque, pois foram as únicas que também conseguiram obter mais de dez registros de documentos em seus nomes, perante todos os inventores / cessionários encontrados na base explorada em questão.

Abaixo do número de dez patentes registradas, podem-se citar as instituições *Siemens Aktiengesellschaft (Berlin and Munich, DE),* com 06 patentes; e as instituições *CIDRA Corporation (Wallingford, CT)* e *Metso Automation Oy (Helsinki, FI),* com 05 patentes concedidas a cada uma. Para ilustração dos resultados percentuais obtidos, gerou-se um gráfico comparativo que pode ser visualizado na figura 5-12.

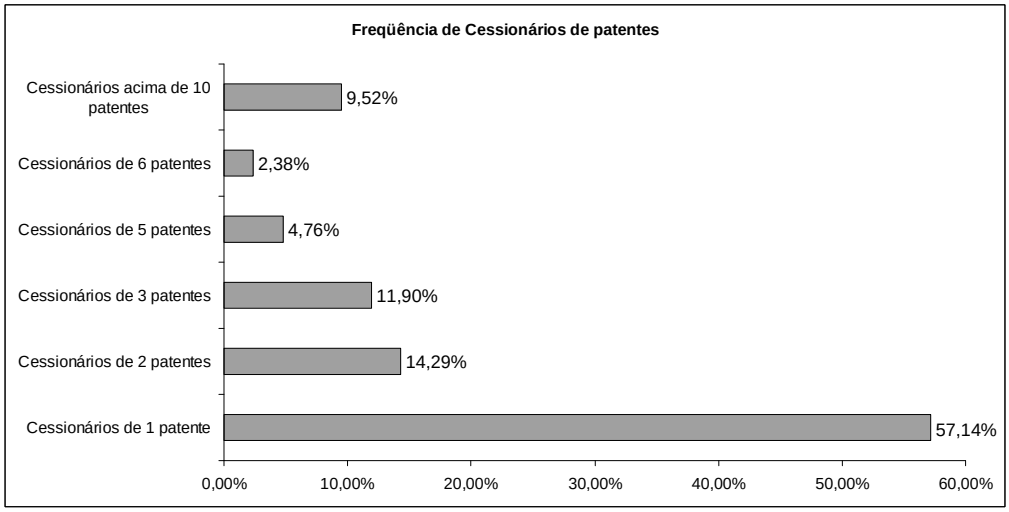

Figura 5-12 Gráfico de Inventores / Cessionários das patentes da BD *Smart Fields*

5.2.5. *Smart Wells*

A partir do levantamento estatístico preliminar elaborado sobre os documentos eletrônicos que compõem a base de dados, averiguou-se que 77 das patentes analisadas foram registradas no banco de dados da instituição entre os anos 2002 e 2007, de forma a

corresponder a um índice de 87,50% do total. Ente os anos 1989 - 1992 e 1998 – 1999, encontrou-se uma patente registrada por ano, e entre os anos 1993-1997, não se encontraram quaisquer registros de documentos pertinentes a tal assunto. Outra boa constatação que pode-se fazer a partir da visualização do gráfico é quanto ao número de registros por ano das patentes relativas ao tema tratado: desde 2004 - à exceção do ano 2005, quando averiguou-se um número de 14 patentes registradas – registram-se 15 patentes por ano. No último quatriênio, o número de patentes total registrado corresponde a 67,04% do total de registros contidos na base textual composta. Considerando-se somente os anos depois de 2000, constata-se uma média anual acima de 11 patentes. O gráfico relativo ao levantamento pode ser visualizado na figura 5-13.

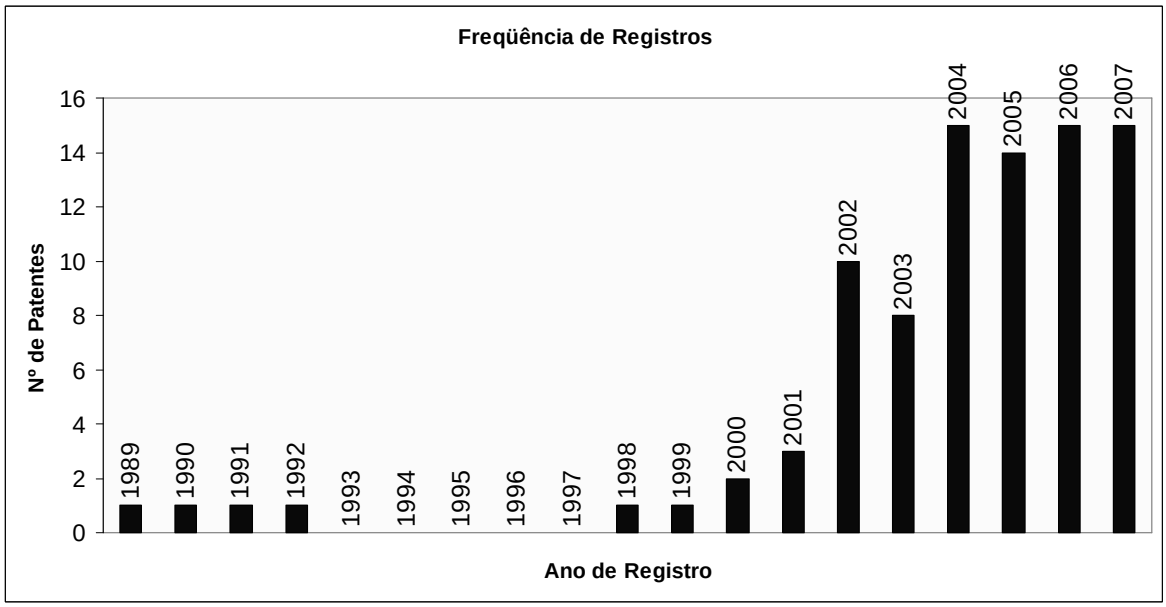

Figura 5-13 Gráfico de Freqüência de Registros das patentes da BD *Smart Wells*

5.2.5.1. Classificações

Dentre as patentes analisadas que constituem a base de dados em questão, verificou-se a presença de 31 classificações internacionais diferentes. O grupo E21B aparece como grande destaque do levantamento, com 51 patentes que contém essa

- Classe E21B: Perfuração do Solo, por exemplo, perfuração profunda; Obtenção de Óleo, Gás, Água, materiais solúveis ou fundíveis ou uma lama de minerais de poços;
  - → Subclasse E21B 43/00: Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis, fundíveis ou de lama, minerais de poços;
  - → Subclasse E21B 43/10: Colocação ou fixação de tubos de revestimento, peneiras (ou filtros), ou tubos auxiliares de revestimento em poços (empencando ou forçando tubos de

investigação ou análise de materiais terrestres através da determinação de suas propriedades químicas ou físicas G01N; medição de variáveis elétricas ou magnéticas em geral, outras além da direção ou a amplitude do campo magnético da terra G01R; combinações de ressonância magnética em geral G01R 33/20);

- Classe G01N: Investigação ou análise dos materiais pela determinação de suas propriedades químicas ou físicas (separação de componentes de materiais em geral B01D, B03, B07; aparelhos totalmente abrangidos por uma outra única subclasse, ver a subclasse apropriada, por ex., B01L; processos de medição ou teste, outros que não ensaios imunológicos, envolvendo enzimas ou microorganismos C12M; investigação do solo para fundações E02D 1/00; dispositivos para monitoramento ou diagnóstico para aparelhagem de tratamento de exaustão de gás F01N 11/00; indicação das variações de umidade para compensação de medições de outras variáveis ou para compensação de leituras de instrumentos devido a variações da umidade, ver G01D ou a subclasse apropriada para a variável medida; teste ou determinação das propriedades das estruturas G01M; medição ou teste das propriedades elétricas ou magnéticas dos materiais G01R; sistemas ou métodos em geral para determinar distância, velocidade ou pressão, utilizando a recepção ou emissão de ondas de rádio ou outras ondas baseada nos efeitos da propagação, por ex. efeito Doppler, tempo de propagação, direção da propagação, G01S; determinação da sensibilidade, granulosidade, ou densidade de materiais fotográficos G03C 5/02; teste das peças componentes de um reator nuclear G21C 17/00);
- Classe A45C: Necessidades Humanas; Artigos portáteis ou de viagem; Bolsas; sacos ou cestas de viagem; valises (recipientes em geral B 65 D, por ex., recipientes

flexíveis portáteis B65D 27/00 a 37/00; fabricação de artigos de couro, lona ou similares B 68 F).

5.2.5.2. Inventores / Cessionários

Sobre os representantes legais, um estudo estatístico apontou a empresa *Schlumberger Technology Corporation (Sugar Land, TX)* como a principal cessionária das patentes coletadas que compõem a base textual de dados analisada, com um número total de 24 patentes registradas. A empresa *Weatherford/Lamb, Inc. (Houston, TX)* aparece logo a seguir, com 10 patentes, seguida da empresa *Halliburton Energy Services, Inc. (Dallas, TX),* com 07 patentes. As empresas *FMC Corporation (Chicago, IL)* e *Baker Hughes Incorporated (Houston, TX)* aparecem na relação com 06 patentes cedidas pela instituição norte-americana a cada uma.

A pesquisa ainda mostra que 68,75% do total dos 32 diferentes representantes legais das patentes que compõem a base textual possuem apenas 01 patente cedida em seus nomes. O gráfico elaborado que confronta os representantes legais das patentes com suas relativas freqüências de registros pode ser visualizado na figura 5-15.

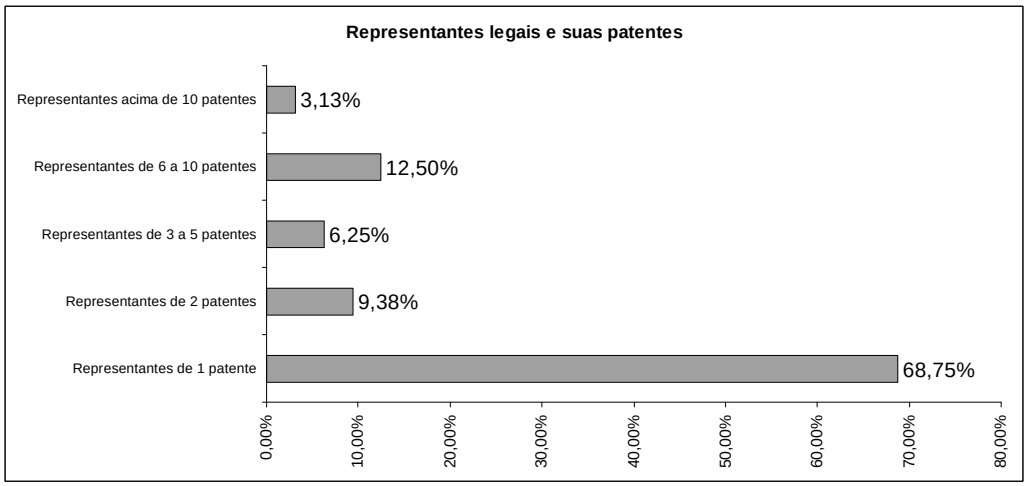

Figura 5-15 Gráfico de Inventores / Cessionários das patentes da BD *Smart Wells*

5.2.6. *Steel Catenary Risers*

Através do estudo estatístico preliminar elaborado e seguindo os critérios de pesquisa já comentados para este presente tema, verificou-se que 22 das 27 patentes coletadas no banco de dados da instituição americana foram registradas a partir do ano de 2000, o que representa um índice expressivo de 81,48% do total de patentes da base textual de dados formada. A primeira patente encontrada foi registrada em 14 de dezembro de 1993, e a mais recente, em 27 de março de 2007. O ano que registra o maior índice de patentes cedidas é 2004, seguido de 2006, com seis e cinco patentes, respectivamente. O gráfico formado para tal estudo pode-se visualizar na figura 5-16.

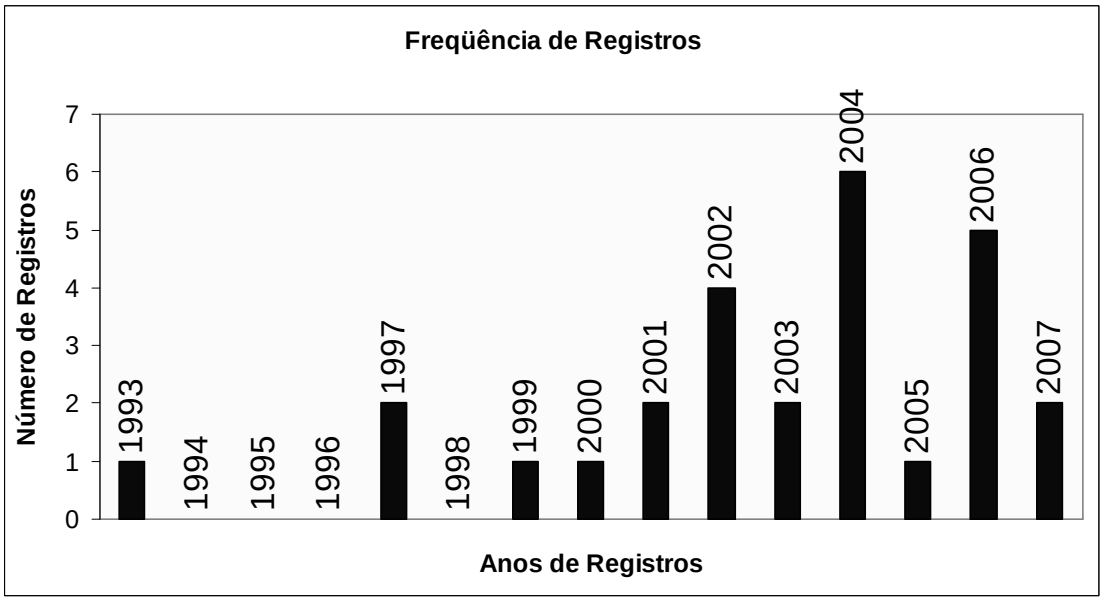

Figura 5-16 Gráfico de Freqüência de Registros das patentes da BD *Steel Catenary Risers*

5.2.6.1. Classificações

Ao realizar um estudo estatístico sobre as classificações internacionais inerentes ao tema tratado nesse tópico, verificou-se que as patentes coletadas, de um modo geral, apresentam oito classificações internacionais presentes em ser relativo campo.

Debruçando-se sobre elas, constatou-se que as classificações E21B, F16L e B63B apresentam a maior freqüência de registros de patentes com, respectivamente, dezesseis, treze e dez registros. A figura 5-17 permite a visualização do gráfico gerado, de forma a confrontar cada classificação internacional encontrada com seus números de registros relativos.

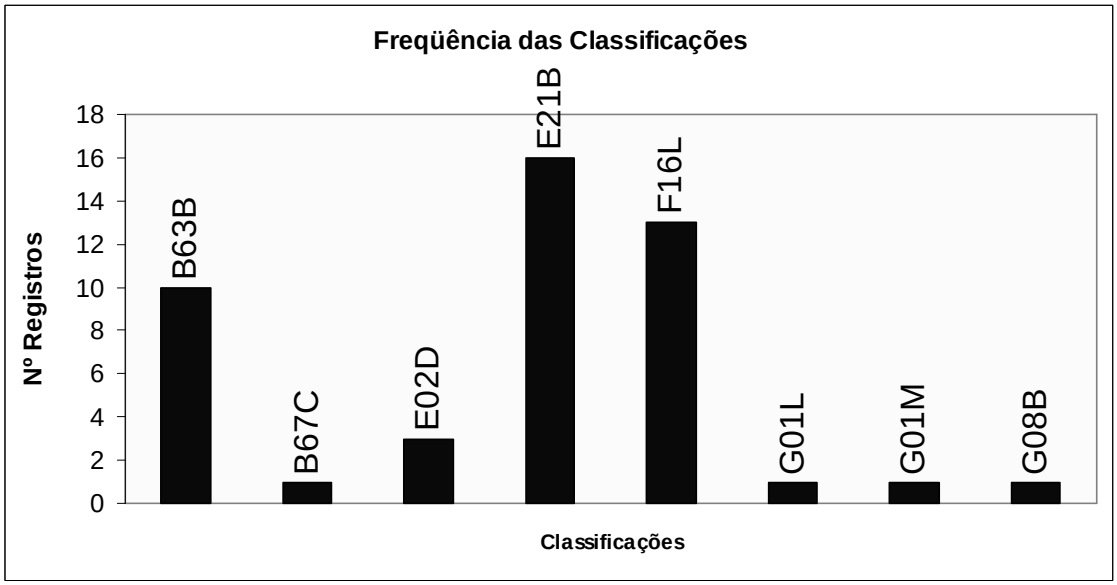

Figura 5-17 Gráfico de Freqüência de Classificações das patentes da BD *Steel Catenary Risers*

Dentre os grupos e subgrupos relativos a cada uma das classificações mais freqüentes, podem-se exemplificar alguns como os de maior predominância:

- Para a classificação internacional E21B, as subclasses que mais se destacam são E21B 17/00, E21B 17/01 (ambas já descritas nos subitens 5.2.1.1 e 5.2.2.1 desta dissertação) e E21B 19/00, cuja descrição apresenta-se a seguir:
    - E21B 19/00: Perfuração do solo; mineração; Perfuração profunda (mineração, exploração de pedreiras E21C; escavação de Poços, abertura de galerias ou túneis E21D); obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços; Manuseio de hastes de produção, tubos de revestimento ou similares

fora do furo de sondagem, por ex., na torre de perfuração (acionamentos superficiais 1/02, 3/02);

- Para a classificação internacional F16L, podem-se citar como subclasses de maior freqüência F16L 1/12 e F16L 1/20, cujas descrições apresentam-se a seguir:
    - F16L 1/12: Tubos; juntas ou acessórios para tubos; suportes para tubos, cabos ou tubulação de proteção; meios para isolamento térmico em geral; Assentamento ou recuperação de tubos sobre ou debaixo d'água (mangueiras flutuantes 11/133);
    - F16L 1/20: Tubos; juntas ou acessórios para tubos; suportes para tubos, cabos ou tubulação de proteção; meios para isolamento térmico em geral; Assentamento ou recuperação de tubos; Montagem ou recuperação de tubos sobre ou debaixo d'água; Acessórios para esse fim, por ex., flutuadores, pesos (bóias B63B 22/00);
- Para a classificação internacional B63B, as subclasses B63B 21/00, B63B 21/50 e B63B 22/00 são as mais freqüentes. Suas descrições já foram feitas em subitens anteriores (5.2.1.1 e 5.2.2.1) desta dissertação.

5.2.6.2. Inventores / Cessionários

A partir da análise feita sobre os inventores / cessionários registrados em relativos campos presentes nos documentos eletrônicos analisados, encontraram-se dezessete instituições diferentes que tiveram patentes cedidas pela instituição americana estudada. Ainda a debruçar-se sobre esse estudo, constatou-se que dez instituições são cessionárias de somente uma patente, o que representa um índice de 58,82% do total de cessionários;

cinco instituições são cessionárias de duas patentes diferenciadas, o que representa um índice de percentual de 29,41% do total; uma instituição é cessionária de três patentes diferenciadas, e uma instituição é cessionária de quatro patentes diferenciadas, o que representa, para cada uma, um índice percentual de 5,88% sobre o número de cessionários total. Vale ressaltar que nesse levantamento estatístico, todos os campos *ASSIGNEE* (CESSIONÁRIO) das patentes presentes na base de dados analisada estavam devidamente preenchidos com o nome da instituição que adquiriu os direitos legais sobre o que reza nos documentos. A figura 5-18 ilustra o gráfico resultante desse estudo.

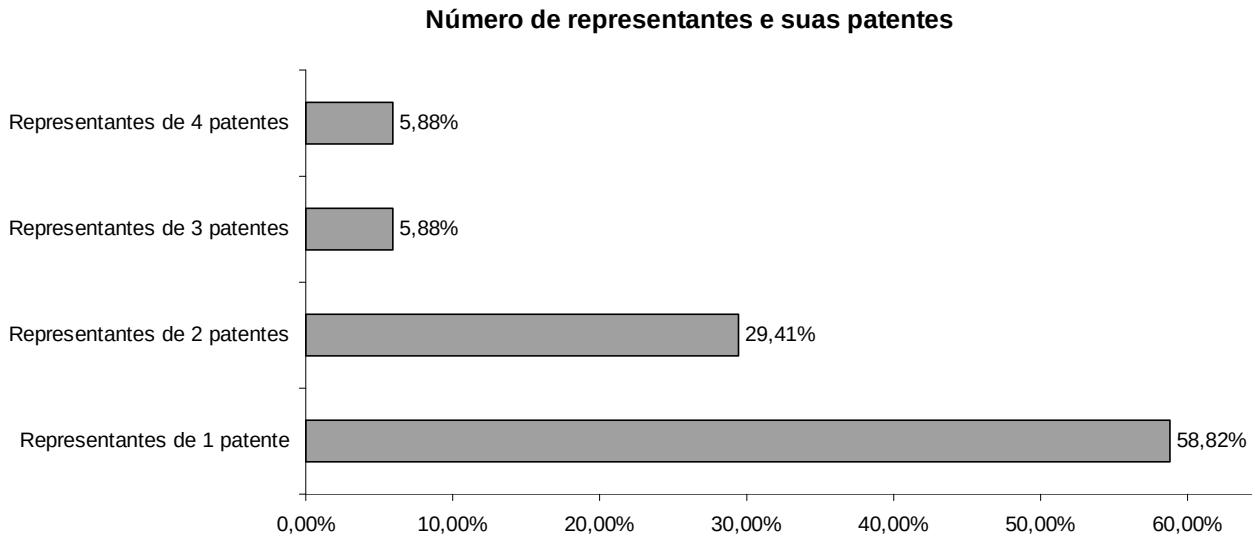

Figura 5-18 Gráfico de Inventores / Cessionários das patentes da BD *Steel Catenary Risers*

Dentre as instituições que tiveram maior número de patentes cedidas, destacam-se a empresa brasileira Petróleo Brasileiro S.A (Petrobras, BR), com quatro patentes cedidas; e a empresa americana *FMC Corporation (Chicago, IL),* com três patentes cedidas.

5.3. Pré-Processamento Textual

Antes da realização das etapas de pré-processamento, revisaram-se de forma manual os conteúdos das patentes a fim de que pudessem se detectar erros de digitação que pudessem vir a prejudicar o desempenho das atividades de mineração. Vale salientar, no entanto, que a revisão ocorreu somente nos termos considerados significantes para as etapas subseqüentes.

Ainda em fase anterior ao pré-processamento, verificou-se, para cada base textual de dados analisada, se suas patentes continham os conteúdos de seu campo *ABSTRACT* completamente iguais entre si, o que poderia ser fruto de uma reedição do documento por conseqüência de um possível aperfeiçoamento sobre um método, técnica ou produto mais antigo. Por coerência, nos casos em que se detectaram tais redundâncias, optou-se por eliminar as patentes com datas de registros mais antigas e manter-se somente a patente com a data de registro mais recente.

O processo de retiradas de termos não-relevantes para a classificação dos documentos eletrônicos constitui de uma lista composta de termos que não traduzem informação para a descoberta final de conhecimentos, tais como artigos, conjunções, interjeições, preposições entre outros, além de outros termos que demonstram-se descartáveis dentro da base de dados textuais utilizada pelo fato de se mostrarem demasiadamente repetitivos, o que os caracterizam como não-bons discriminadores (rotuladores) de possíveis classes de documentos a serem pesquisadas e descobertas. Esta lista, conhecida na literatura como *Stoplist,* é armazenada em arquivo digital, de modo que programa-se o *software RapidMiner / YALE* para considerá-lo durante esta etapa de pré-processamento.

A aplicação do algoritmo de *Stemming* resume ainda mais o número total de termos inseridos no vetor criado pelo programa *RapidMiner / YALE*, através de uma redução sobre os mesmos, extraindo-se os respectivos radicais e desconsiderando os prefixos e sufixos de cada termo para a classificação. Dentre os algoritmos oferecidos pelo mesmo programa, convencionou-se o algoritmo de *Porter* – próprio para idioma inglês que é o idioma no qual estão escritas as patentes tratadas em presente estudo – para aplicação em todas as bases de dados textuais desta pesquisa.

Uma vez aplicadas estas duas etapas, visou-se ainda mais diminuir a carga de trabalho sobre o algoritmo de *Clustering*. Optou-se, portanto, em eliminar os termos de freqüências muito alta e de freqüências muito baixas, de forma a manter-se, somente, os termos que encontravam-se em mais de um documento e menos que todos. Por conseqüência, este índice de redução representa um ganho computacional significante na etapa de processamento, de forma a resumir o tempo e a melhorar a qualidade do trabalho.

Além disso, por razões de dimensionalidade, procurou-se resumir os vetores de termos criados, de forma que o número de variáveis fosse igual ou menor que o número de registros contidos na base de dados. Vale a lembrança que os registros de cada base representam seus respectivos documentos eletrônicos coletados, e as variáveis representam os termos de maior significância remanescentes após todas as etapas de pré-processamento já comentadas.

Toda a etapa de pré-processamento textual é realizada sobre o vetor de termos *(Word Vector)* criado pelo *software* em questão (*RapidMiner / YALE),* e os vetores de termos, bem como seus respectivos pesos, são atualizados a cada iteração do programa sobre a base submetida ao processo.

A partir da descrição da etapa de pré-processamento textual e com o suporte do *software RapidMiner / YALE*, obtiveram-se os resultados que serão apresentados nos

próximos itens a seguir. Cada item representa um resultado para cada base textual de dados diferente.

5.3.1. *Anchoring Systems*

Em princípio, definiu-se a base textual de dados sobre o tema tratado com 74 patentes. Contudo, verificou-se que duas patentes tinham os conteúdos de seu campo *ABSTRACT* completamente iguais entre si. Eliminou-se, com isso, a patente com conteúdo repetido. Desta eliminação, o número de patentes que compuseram a base caiu para 73 (setenta e três).

Um total aproximado de 6000 termos encontravam-se nos campos *ABSTRACT* das patentes antes da submissão da base às etapas anteriores à etapa de processamento de dados. Com a criação do vetor pelo *software RapidMiner / YALE* e aplicação da *Stoplist* e do algoritmo de *Stemming* sobre o mesmo criado, reduziram-se os termos para aproximadamente 250, o que representa uma redução de carga computacional de 95,83% da base textual de dados bruta.

A retirada dos termos mais freqüentes e menos freqüentes reduziu o número total de termos para 75, ou seja, um índice expressivo de somente 1,25% do total de termos da base textual de dados bruta.

5.3.2. *Flexible Joints*

Em um primeiro momento, definiu-se a base textual de dados com 80 patentes. Todavia, verificou-se que dez patentes tinham os conteúdos de seu campo *ABSTRACT* completamente iguais entre si. Com a eliminação de nove documentos e a manutenção de

um, o número de patentes presentes na base textual de dados resumiu-se em 71 (setenta e uma).

Antes da submissão da base textual às etapas anteriores à etapa de processamento de dados, averiguou-se que 5400 termos, aproximadamente, encontravam-se presentes nos campos *ABSTRACT* das patentes. Com a criação do vetor pelo *software RapidMiner / YALE* e aplicação da *Stoplist* e do algoritmo de *Stemming* sobre o mesmo criado, reduziram-se os termos para aproximadamente 462, o que representa uma redução de carga computacional de 91,44% da base textual de dados bruta.

A retirada dos termos mais freqüentes e menos freqüentes reduziu o número total de termos para 57, ou seja, um índice expressivo de somente 1,05% do total de termos da base textual de dados bruta, facilitando demais a convergência do algoritmo de *Clustering* no processo subseqüente ao pré-processamento.

5.3.3. *Flexible Risers*

Inicialmente, a base textual de dados tratada em presente item foi formada com 127 patentes. Constatou-se que dez documentos tinham os conteúdos de seus *ABSTRACT*S completamente iguais entre si. Com a eliminação das patentes com conteúdos repetidos, o número de registros que compuseram a base textual caiu para 118 (cento e dezoito).

Um total aproximado de 7500 termos encontrava-se nos campos *ABSTRACT* das patentes antes da submissão ao pré-processamento. Com a criação do vetor pelo *software RapidMiner / YALE* e aplicação da *Stoplist* e do algoritmo de *Stemming* sobre o mesmo criado, reduziram-se os termos para aproximadamente 500 termos, o que representa uma redução de carga computacional de aproximadamente 93,33% da base de dados textuais bruta.

A retirada dos termos mais freqüentes e menos freqüentes reduziu o número total de termos para 103, ou seja, um índice representativo de somente 1,36% do total de termos da base de dados textuais bruta.

5.3.4. *Smart Fields*

A base textual de dados em questão foi composta, em princípio, com um número total de 193 patentes. Constatou-se que 24 patentes deveriam ser eliminadas, pois seus conteúdos textuais presentes no campo *ABSTRACT* eram muito semelhantes entre si, o que poderia prejudicar a análise processual dos dados, principalmente em etapa posterior ao pré-processamento. Dessa forma, resumiu-se em 169 o número total de registros que foram submetidos aos processos de mineração de textos.

Constatou-se que um número total aproximado de 16300 termos encontrava-se nos campos *ABSTRACTS* dos documentos eletrônicos antes da submissão ao pré-processamento. Realizou-se a criação do vetor de termos pelo *software RapidMiner / YALE,* e com a aplicação da *Stoplist* e do algoritmo de *Stemming*, reduziu-se o conjunto para aproximadamente 740 termos, o que representa uma redução de carga computacional de aproximadamente 95,46% da base textual bruta.

Adotando-se o critério de retirarem-se os termos de maiores e de menores freqüências, reduziu-se o número total de termos para 170, ou seja, um índice representativo de aproximadamente 1,04% do total de termos da base textual bruta de dados.

5.3.5. *Smart Wells*

A pesquisa sobre o tema tratado retornou como resultado um número total de 88 documentos eletrônicos. Posteriormente, ao submetê-las aos processos de verificação, constatou-se a necessidade de eliminação de 13 patentes, por ter os seus conteúdos no campo *ABSTRACTS* completamente iguais aos de outras patentes coletadas, o que poderia prejudicar o processamento textual, além de serem textos redundantes e que nada mais acrescentariam de informações a serem exploradas. Dessa forma, o conjunto de dados textual resumiu-se à presença de 75 documentos eletrônicos.

Inicialmente, encontrou-se presente nos campos *ABSTRACT* um número total aproximado de 5475 termos. Com a criação do vetor de palavras pelo *software RapidMiner / YALE* e das aplicações da *Stoplist* e do algoritmo de *Stemming* sobre o vetor criado, reduziu-se o vetor para aproximadamente 435 termos, o que representa uma redução de carga computacional de aproximadamente 92,06% do conjunto de dados textual bruto.

Seguindo o critério de eliminar-se da pesquisa os termos de maiores e menores freqüências, reduziu-se o vetor de palavras para um número total de 70 - um índice representativo de somente 1,28% do total de termos do conjunto de dados textual bruto.

5.3.6. *Steel Catenary Risers*

Para as patentes que descrevem métodos, técnicas e materiais a respeito de *SCR* (*Steel Catenary Risers*), em princípio, encontraram-se 27 documentos eletrônicos. Verificou-se que, ao contrário de algumas bases de dados textuais dos outros temas pesquisados, nenhum documento eletrônico pertencente a esse escopo teve os valores

textuais de seus *ABSTRACTS* completamente repetidos, o que permitiu que todas vinte e sete patentes encontradas fossem consideradas para os estudos posteriores.

Constatou-se que, aproximadamente, 2650 termos encontravam-se presentes nos campos *ABSTRACTS* das patentes antes da submissão da base às etapas anteriores à etapa de processamento de dados. Com a criação do vetor pelo *software RapidMiner / YALE* e aplicação da *Stoplist* e do algoritmo de *Stemming* sobre o mesmo criado, reduziu-se o vetor para 192 termos, o que representa uma redução de carga computacional de 92,75% da base textual de dados bruta.

Com a retirada dos termos mais freqüentes e menos freqüentes, reduziu-se o número total de termos para 22, o que representa um índice expressivo de somente 0,83% do total de termos da base bruta.

5.4. Processamento

Uma vez que o vetor de termos-chaves foi criado durante a fase anterior ao processamento, a etapa seguinte foi a submissão do mesmo vetor de termos-chaves criado, para cada estudo de caso, visando a descoberta de grupos *(clusters)* de dados correlatos entre si, utilizando-se o algoritmo de *Clustering K-Means*.

No entanto, conforme já fora comentado no Capítulo 04 dessa presente dissertação, o *software* em questão não possui uma métrica de validação que evidencie um número de *clusters* a ser sugerido para formação, que gere respectivos gráficos de convergência do algoritmo, ou que mensure uma relação do grau de precisão do sistema de maneira clara e objetiva. A fim de que tal entrave pudesse ser resolvido, em primeira instância recorreu-se a uma análise sobre cada resultado obtido ao final de cada chamada do algoritmo *K-Means*, de modo que consideraram-se alguns aspectos que possibilitassem uma decisão concisa sobre o número mais próximo possível de agrupamentos (segmentos) a serem considerados, tais como: coordenadas do centróide de cada documento e distâncias intra - *clusters* dos mesmos; termos-chaves de maior significância para cada grupo; e as classificações que exercessem uma hegemonia sobre os resultados dos grupos.

Em uma segunda instância, recorreu-se ao algoritmo de validação de agrupamentos implementado utilizando-se o *software Matlab* e já comentado no Capítulo 04. Dessa forma e para cada estudo de caso, uma vez que definiu-se o vetor de termos-chaves, formou-se a tabela de dados, exportou-se a mesma do *software RapidMiner / YALE* em formato de arquivo com extensão ".dat" para o *software* de edição de planilhas eletrônicas. Uma vez que cada tabela foi exportada, trabalharam-se cada uma das planinhas de valores formadas, utilizando-se o *software* específico. Após a formatação, submeteu-se a tabela de valores dos termos-chaves à execução do algoritmo de validação de

agrupamentos, com o intuito de detectar o número ideal dos agrupamentos de documentos eletrônicos a serem formados para esta etapa de processamento.

Vale a ressalva de que cada tabela de dados formada possui:

- Como atributos: os termos-chaves considerados presentes na base textual de dados;
- Como registros: cada documento eletrônico pertencente; e
- Como valores numéricos: os índices TF-IDF de cada termo em cada documento.

A partir dos resultados obtidos em comunhão com os demais aspectos considerados em primeira instância, puderam-se tomar decisões sobre o número de agrupamentos a ser considerado para os resultados finais.

Ao final de muitas iterações do algoritmo *K-Means* - próprio para os processos de *Clustering* e implementado no *software RapidMiner / YALE* - para cada estudo de caso elaborou-se uma tabela-resultado com a intenção de exibir os resultados de cada agrupamento obtido. Por convenção, adotou-se um número indo-arábico para cada agrupamento encontrado e que estão contidos na primeira coluna de cada tabela formatada. A segunda coluna de cada tabela-resultado é composta do número de patentes encontradas para cada agrupamento. A terceira coluna da tabela-resultado é composta por um termo designado para representar o agrupamento encontrado. Por critério, convencionou-se que o termo representante do agrupamento seria o seu termo mais representativo, de forma a expressar a idéia principal dos documentos presentes em seu escopo. Por fim, a quarta e última coluna de cada tabela-resultado representa os termos mais representativos de cada agrupamento formado.

Com o intuito de aprofundar ainda mais a pesquisa, promoveu-se para cada estudo de caso, após a formação e exibição de cada tabela-resultado, uma análise de caráter mais

exploratório sobre os rótulos dos agrupamentos formados. Exemplificaram-se alguns títulos de patentes pertencentes aos seus conjuntos, alguns de seus inventores / cessionários de maior significância, seus termos-chaves encontrados, suas classificações de maior freqüência dentre as patentes presentes em cada agrupamento, além do campo *ABSTRACT* de um dos documentos pertencentes ao agrupamento.

Em última instância e com o objetivo de mensurar os resultados obtidos pela ferramenta computacional *RapidMiner / YALE* em seus processos de agrupamentos de dados não-estruturados para cada estudo de caso, submeteram-se as mesmas bases às tarefas de análise de dados não-estruturados utilizando-se o *software PolyAnalyst 6.0* e criando-se tabelas com os resultados gerados. As tabelas geradas com os resultados extraídos da ferramenta *PolyAnalyst* seguem o mesmo padrão que as tabelas geradas para visualização dos resultados gerados pela ferramenta *RapidMiner / YALE*.

Posteriormente, comentou-se a respeito da comparação dos resultados visualizáveis em cada tabela formada em cada estudo de caso. As próximas subseções do presente capítulo apresentarão os resultados obtidos desses processos de análise comentados.

5.4.1. *Anchoring Systems*

Três das quatro métricas empregadas para o algoritmo, após muitas iterações do mesmo, nortearam o resultado para um número sugerido de três agrupamentos. De forma a considerar os resultados obtidos com as métricas de validação em conjunto com outros fatores já comentados no item 5.4, concluiu-se que o valor norteado e sugerido pelo algoritmo de validação seria adotado como número de agrupamentos a serem formados.

A tabela 5-1 exibe os resultados dos agrupamentos formados pelo algoritmo de *Clustering.* Após sua exibição, apresentar-se-á a análise mais detalhada de cada um dos agrupamentos.

Classificação Não-Supervisionada sobre a Base *Anchoring Systems*			
Programa: *RapidMiner / YALE*		**Algoritmo empregado: *K-Means***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chave**
1	24	*Pipe*	***Pipe;*** *vessel; depth; vertical; riser; support; buoyant; section; flexible; bottom.*
2	20	*Wellbore*	***Wellbore***; tube; s*tring*; *fluid*; *valve*; *movable*; *assemble*; *pressure*; *drill*; *hydraulic*.
3	29	*Attach*	***Attach; device****; surface; direct; offshore; engage; rotate; float; adapt; body.*

Tabela 5-1 – Resultados da classificação não-supervisionada para a BD *Anchoring Systems*

Cluster 1 – *Pipe*

Termos- chaves do agrupamento:

Pipe; *vessel*; *depth*; *vertical*; *riser*; *support*; *buoyant*; *section*; *flexible*; *bottom*.

Títulos que pertencem ao agrupamento:

- *Anchor system for the transfer of fluids*;
- *Anchoring structure for marine riser assembly;*
- *Method and device for linking surface to the seabed for a submarine pipeline installed at great depth;*
- *Seafloor/surface connecting installation for a submarine pipeline which is connected to a riser by means of at least one elbow pipe element that is supported by a base*.

Principais Classificações:

- B63B 21/00: Operações de processamento; transporte; Navios ou outras embarcações; Equipamento para navegação; Amarração; Equipamento para deslocar, rebocar ou empurrar; Ancoragem;
- B63B 21/50: Operações de processamento; transporte; Navios ou outras embarcações; Equipamento para navegação; Amarração; Disposições para ancoragem de embarcações especiais, por exemplo, para plataformas flutuantes de perfuração ou dragas.

Principal Cessionário:

Korsgaard; Jens (Princeton Junction, NJ).

ABSTRACT que represente o agrupamento:

The present inventions relate to improved apparatus and methods for radially expanding tubulars, such as tubing, casing and sand-control screen assemblies in a subterranean oil or gas well, and more specifically, to a variable diameter expansion tool for expanding downhole tubulars to varying diameters. In general, the inventions provide apparatus and methods for radially expanding a tubular, such as pipe, tubing, screen or screen assembly, deployed in a subterranean well by moving an expansion tool axially through the well. An automatically infinitely variable-diameter expansion cone tool is provided. A variable-diameter cone is provided, movable between an expanded position and a retracted position. The cone is enlarged to its expanded position and advanced through expandable components until a restriction is reached. At the restriction, the variable-diameter cone automatically retracts enough to allow the tool to continue advancing through the wellbore. When the restriction is past, the variable-diameter cone enlarges again to its expanded position.

Cluster 2 – *Wellbore*

Termos- chaves do agrupamento:

Wellbore; *tube*; *string*; *fluid*; *valve*; *movable*; *assemble*; *pressure*; *drill*; *hydraulic*.

Títulos que pertencem ao agrupamento:

- *Hydraulic and mechanical noise isolation for improved formation testing*;
- *System for lining a section of a wellbore;*
- *Wellbore isolation apparatus, and method for tripping pipe during underbalanced drilling;*
- *Method and apparatus for a monodiameter wellbore, monodiameter casing, monobore, and/or monowell;*

Principais Classificações:

- E21B 17/00: Construções Fixas; Perfuração do solo; Mineração; Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais 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 43/00: Construções Fixas; Perfuração do solo; Mineração; Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços; Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços.

Principais Cessionários:

- *Baker Hughes Incorporated (Houston, TX);*
- *Weatherford/Lamb (Houston, TX).*

ABSTRACT que represente o agrupamento:

The present invention relates to an apparatus and method for isolating a wellbore condition such as formation pressure during a wellbore operation. The invention has particular application in connection with underbalanced drilling. In one arrangement, a formation isolation apparatus is provided that serves as a selectively actuatable plug. The plug in one aspect is selectively set and released by a setting/releasing tool. The setting/releasing tool includes a system for setting the plug in the wellbore, and a system for releasing the plug from the wellbore. The setting/releasing tool is releasably connected to the plug. Thus, after the plug has been set, the setting/releasing tool may be removed from the wellbore. The plug includes a flapper valve that is restrained in its open position by the setting/releasing tool. Removal of the setting/releasing tool from the wellbore allows the flapper valve to close, thereby isolating pressures in the wellbore below the flapper valve. The plug is wireline retrievable. In another aspect, a formation isolation apparatus is provided for use during sidetrack drilling operations. The sealing element is movable from a first released position below the lateral wellbore, to a set position above the lateral wellbore.

Cluster 3 – *Attach*

Termos-chaves do agrupamento:

A*ttach*; *device*; *engage*; *adapt*; *surface*; *direct*; *offshore*; *system*; *float*; *body*.

Títulos que pertencem ao agrupamento:

- *Off-shore mooring device for a large-sized floating body;*
- *Vessel mooring system and vessel equipped for the system;*
- *Method and system for anchoring a buoy via a screw-type anchor;*
- *Offshore mooring and fluid transfer system*;

Principais Classificações:

- B63B 21/00: Operações de processamento; transporte; Navios ou outras embarcações; Equipamento para navegação; Amarração; Equipamento para deslocar, rebocar ou empurrar; Ancoragem;
- B63B 21/24: Operações de processamento; transporte; Navios ou outras embarcações; Equipamento para navegação; Amarração; Âncoras.

Principal Cessionário:

- *The United States of America as represented by the Secretary of the Navy (Washington, DC).*

ABSTRACT que represente o agrupamento:

An anchoring tool to position an attached tool, such as a chemical cutter or other tools, in a predetermined position within the bore of a tubular. After the desired location for cutting

the tubular or performing other functions is determined, the anchoring tool, along with the chemical cutter or other tools, is lowered into the bore of the tubular until the predetermined position is reached. Thereafter, an ignitor expands pressure propellants, positioned within the body of the anchoring tool, and produces pressure which causes anchoring slips to anchor the tool and any associated equipment within the interior of the tubular.

5.4.1.1. Resultado de *Anchoring Systems* obtido pelo *software PolyAnalyst*

Nesta etapa, submeteu-se a base textual de dados definida às tarefas de pré-processamento e de processamento textual, executadas pela ferramenta *PolyAnalyst* em seu módulo de Mineração de Textos. A tabela 5-2 permite visualizar os resultados obtidos, em um número total de 04 agrupamentos sugerido pelo algoritmo de *Clustering* utilizado pela plataforma para execução do processo.

Classificação Não-Supervisionada sobre a Base *Anchoring Systems*			
Programa: *PolyAnalyst 6.0*		**Algoritmo empregado: *Suffix Tree Clustering***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chave**
1	30	*Attach*	*Attach; vessel moor; vessel connect*; *surface; submerge; buoyant moor; vessel seal; offshore structure; seal.*
2	14	*Cable*	*Cable; suspend.*
3	09	*Oil*	*Oil, tubular.*
4	20	*Wellbore*	*Borehole*; *wellbore*; *downhole*; *subsea*; *drill*; *string*; *well*; *vertical*.

Tabela 5-2 – *PolyAnalyst*: Resultados da Mineração de Textos para a BD *Anchoring Systems*

Ao comparar os termos-chaves obtidos e visualizáveis na tabela acima com os termos-chaves obtidos pelo *K-Means* implementado na ferramenta *RapidMiner / YALE* (tabela 5-2), nota-se uma proximidade de significados (sinônimos) entre as respectivas palavras e expressões encontradas por ambos os algoritmos para os respectivos agrupamentos formados. Contudo, vale ressalvar que as configurações oferecidas pela

ferramenta *RapidMiner / YALE* proporcionou a vantagem de se solicitar um número estipulado de palavras mais significativas para cada agrupamento formado, o que não ocorreu para a plataforma *PolyAnalyst*. Em contrapartida, essa mesma ferramenta proporcionou uma busca mais abrangente por expressões e frases correlatas entre os textos, ao passo que a segmentação de dados do algoritmo *K-Means* proporcionou buscas simples somente por palavras, sem a extração de expressões ou frases relevantes dos textos analisados.

A fim de que pudesse se realizar uma comparação entre os resultados obtidos pelas duas técnicas de *Clustering* empregadas para a base de dados em questão, gerou-se um gráfico (figura 5-19 abaixo) com o intuito de se mensurar as similaridades entre os agrupamentos formados por cada plataforma.

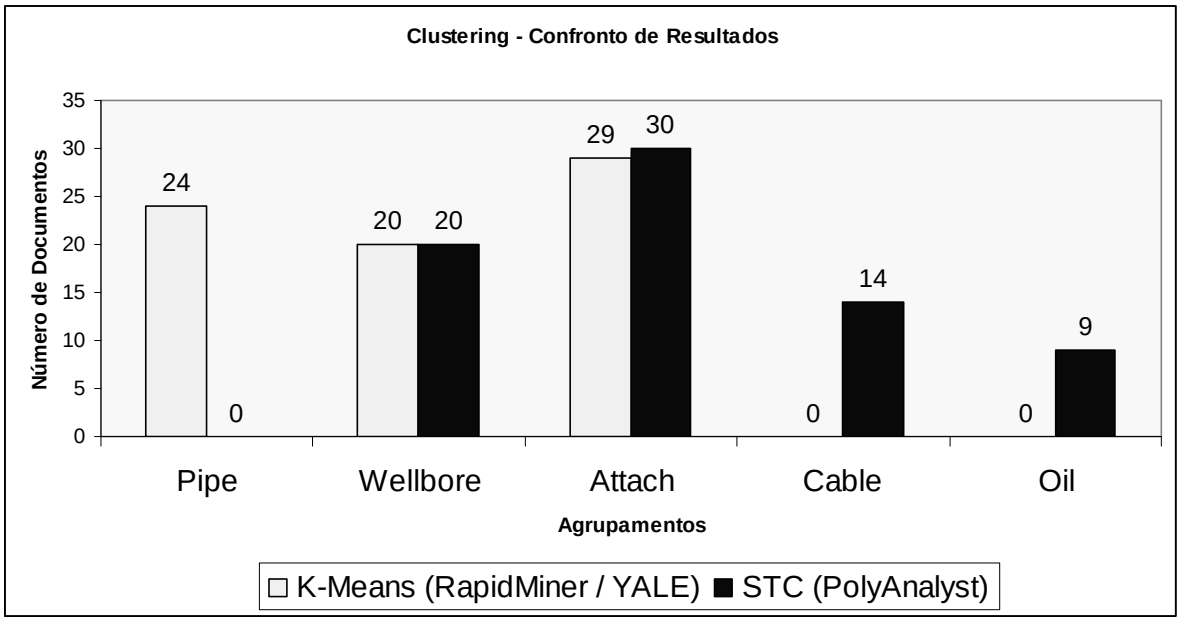

Figura 5-19 Comparação entre os resultados de *Clustering* para a BD *Anchoring Systems*

A debruçar-se sobre o gráfico, algumas considerações tornam-se interessantes: evidencia-se que as técnicas computacionais empregadas coincidem seus resultados na formação de dois agrupamentos de documentos, que como pôde-se visualizar sobre o gráfico, apresentam-se bastante semelhantes entre si. O agrupamento *Wellbore* formado

pelo algoritmo *K-Means (RapidMiner / YALE)* possui 100% de número de documentos coincidentes com o número de documentos pertencentes ao mesmo agrupamento formado pelo algoritmo *STC (PolyAnalyst)*. O agrupamento *Attach* formado pelo algoritmo *STC* possui 29 dos 30 documentos pertencentes ao mesmo agrupamento formado pelo algoritmo *K-Means,* o que significa um índice expressivo de 96,66%.

Já o número de documentos pertencentes ao agrupamento *Pipe* formado pelo algoritmo *K-Means* foi subdividido pelo algoritmo STC em dois segmentos, nomeando-os como *Cable,* com 14 documentos, e *Oil*, com 09 documentos. Ao analisarem-se os termos-chaves de cada um, nota-se uma evidente similaridade com relação aos termos-chaves inerentes ao agrupamento *Pipe*. Dessa forma, torna-se relevante a constatação de que os parâmetros estipulados para convergência do algoritmo *STC* proporcionou-o um maior poder de especificação e que incidiu diretamente na formação de um quarto agrupamento, ao passo que o algoritmo *K-Means* generalizou os registros inerentes, por considerar a distância entre os centróides dos registros muito próximos entre si.

Em síntese: apesar de alguns termos encontrados em ambas as análises serem diferentes entre si, bem como a diferença entre o número de agrupamentos conforme já fora comentado, o significado dos agrupamentos pode ser considerado semelhante para aqueles que apresentaram proporção de similaridade alta. Conforme as configurações (em termos de valores para execução) dos parâmetros de cada algoritmo, bem como suas respectivas lógicas para alcance dos resultados, essas diferenças tornam-se normais. Além disso, cada ferramenta possui uma diferente técnica computacional de busca por palavras-chave, o que também implica diretamente no resultado final.

5.4.2. *Flexible Joints*

Após muitas iterações do mesmo, duas das quatro métricas nortearam o resultado para um número sugerido de três agrupamentos. Todavia, vale a ressalva de que o algoritmo não chegou a uma conclusão segura de qual número de agrupamentos seria considerado ideal para a base analisada, de forma que os resultados de suas métricas, a cada iteração, oscilaram sempre seus resultados entre dois e seis agrupamentos. Considerou-se então, os gráficos referentes aos centróides dos documentos eletrônicos e suas distâncias intra - *clusters* gerados após a submissão da base ao algoritmo de *Clustering K-Means* para cada caso, variando o número de partições *K* da base entre duas e seis. Ao final, concluiu-se que a alternativa de maior coerência seria considerar como três o número ideal de agrupamentos a serem formados, pois foi a quantidade de partições que apresentou a melhor distância entre os centróides.

Classificação Não-Supervisionada sobre a Base *Flexible Joints*			
Programa: *RapidMiner / YALE*		**Algoritmo empregado: *K-Means***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chave**
1	38	*Attacher*	***Attacher*; *assembly*; *adjust*;** *column*; *pipe*; *platform*; *secure*; *self*; *anchor*; *tension*; *frame*; *clamp*.
2	23	*Protection*	***Protection***; *pressure*; *compress*; *spring*; *seal*; *control*; *annular*; *angular*; *mount*; *rotate*; *hollow*; *inner*.
3	10	*Drill*	***Drill***; ***apparatus*;** *borehole*; *rigid*; *down hole*; *fold*; *mechanic*; *string*; *degree*; *link*; *lock*; *pipe*.

Tabela 5-3 Resultados da classificação não-supervisionada para a BD *Flexible Joints*

Apresentar-se-á, após a tabela de resultados, uma análise mais detalhada sobre os agrupamentos formados para a base textual de dados em questão.

Cluster 1 – *Attacher*

Termos- chaves do agrupamento:

Attacher; assembly; ***adjust;*** *column*; *pipe*; *platform*; *secure*; *self*; *anchor*; *tension*; *frame*; *clamp*.

Títulos que pertencem ao agrupamento:

- *Underwater chain stopper and fair lead apparatus for anchoring offshore structures;*
- *Apparatus for securing a tubular structure to an anchor*;
-

Principais Cessionários:

- *Shell Oil Company (Houston, TX);*
- *Vetco Offshore Industries, Inc. (Ventura, CA).*

ABSTRACT que represente o agrupamento:

An underwater chain stopper and fairlead apparatus for offshore structures, drilling platforms, ships or other vessels. The apparatus comprises a mounting member, a fairlead member and a chain stopper member. The mounting member is attached to an underwater surface of the offshore structure or vessel and includes a bracket for coupling the fairlead member. The bracket may comprise a hinge allowing the fairlead member to pivot in an approximately horizontal plane. The chain stopper member is coupled to the fairlead member through a hinge which allows the chain stopper member to pivot with respect to the fairlead member in an approximately vertical plane. The chain stopper member includes a chain stopper flapper having a horseshoe shaped opening at one end. The other end of the flapper is connected to the chain stopper member through a hinge which allows the flapper to swing between an open position and a closed position. In the open position, the chain links for the anchor chain are allowed to pass by the horseshoe shaped opening on the flapper. The horseshoe shaped opening also includes a chain link seat which stops movement of the anchor chain through the chain stopper member when the flapper is in the closed position. The chain stopper flapper moves to the closed position under the force of gravity to provide a self-locking chain stopper. A latch mechanism is provided for latching the chain stopper flapper in an open position.

Cluster 2 – *Protection*

Termos- chaves do agrupamento:

Protection; *pressure*; *compress*; *spring*; *seal*; *control*; *annular*; *angular*; *mount*; *rotate*; *hollow*; *inner*.

Títulos que pertencem ao agrupamento:

- *Aerial location self actuating emergency sea surface marker for capsized vessels;*
- *Slip joint intervention riser with pressure seals and method of using the same;*
- *Method for connecting a slender structure to a reference body and for suppressing the vibration of such slender structures;*
- *Flexible sealing joint.*

Principais Classificações:

- E21B 17/00: Construções Fixas; Perfuração do solo; Mineração; Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços; Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção (engates de hastes em geral F16D; tubos ou acoplamentos de tubos em geral F16L);
- E21B 33/06: Vedação ou obturação de furos de sondagens ou de poços; Equipamentos preventivos de (ou para evitar) explosões;
- F16L 27/02: Tubos; Juntas ou acessórios para tubos; Suportes para tubos, cabos ou tubulação de proteção; meios para isolamento térmico em geral; Juntas universais, isto

é, com ligação mecânica permitindo movimento angular ou ajustagem dos eixos da peça em qualquer direção.

Principal Cessionário:

- *Hydril Company (Houston, TX).*

ABSTRACT que represente o agrupamento:

A removable companionway, for use in conjunction with a manufacturer-supplied door as desired depending on weather and user preference, having a frame with adjustable corner brackets to accommodate companionways of different dimensions, a cover for protection from wind, rain, and objects, and a hinge means for moving said companionway about an axis.

Cluster 3 – *Drill*

Termos- chaves do agrupamento:

Drill; ***apparatus*****;** *borehole*; *rigid*; *downhole*; *fold*; *mechanic*; *string*; *degree*; *link*; *lock*; *pipe.*

Títulos que pertencem ao agrupamento:

- *Elevator for supporting an elongate member such as a drill pipe;*
- *Curved drilling apparatus;*
- *Apparatus for drilling a curved subterranean bore hole*;
- *Flexible electrical submersible motor pump system for deviated wells.*

Principais Classificações:

- E21B 1700: Construções Fixas; Perfuração do solo; Mineração; Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços; Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção (engates de hastes em geral F16D; tubos ou acoplamentos de tubos em geral F16L).

Principal Cessionário:

- *Amoco Corporation (Chicago, IL).*

ABSTRACT que represente o agrupamento:

A device for controlling the drilling direction of drills includes a cylinder-type housing 6, a first ring-formed component 11 which is located on an inner peripheral surface that is eccentric with respect to the cylinder-type housing 6, a second ring-formed component 12 which is located on the inner peripheral surface that is eccentric with respect to the circular inner surface of the first ring-formed component 11, and hollow-type harmonized reduction gears 13,14 which rotate the first and second ring-formed components 11,12 relatively around their respective centers. A resolver is positioned between the first and second ring-formed components 11,12 and the hollow-type harmonized reduction gears 13,14 to detect the rotating angular position of the first and second ring-formed components 11,12. A fulcrum bearing 8 of the rotating shaft 2 is located at a midpoint between the drill bit and the first and second ring formed components 11,12. A flexible joint 3 is located at the upper portion of the first and second ring-formed components 11,12 and a bearing 15 is further mounted on the flexible joint in order to support the rotating shaft 1.

5.4.2.1. Resultado de *Flexible Joints* obtido pelo *software PolyAnalyst*

Uma vez conhecidos os resultados do algoritmo *K-Means* executado pelo *software RapidMiner / YALE*, o procedimento posterior foi submeter tal base textual de dados em questão à execução dos processos de mineração de textos pela ferramenta *PolyAnalyst* em seu módulo de Mineração de Textos. Seus resultados podem ser visualizados na tabela 5-4. Os termos em destaque negrito representam o termo mais representativo em cada *cluster*.

Classificação Não-Supervisionada sobre a Base *Flexible Joints*			
Programa: *PolyAnalyst 6.0*		**Algoritmo empregado: *Suffix Tree Clustering***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chave**
1	21	*Protection*	*Attacher; control*; ***protection***; *conduit*; *diameter; flange; fluid; maintain; pressure*; *seal.*
2	12	*Drill*	*Borehole; downhole*; ***drill****; rigid*; rotate; *rotate; well.*
3	19	*Fix*	*Anchor;* ***fix****; float; marine; offshore; platform; riser; string; tension; vessel.*
4	19	*Mount*	*Assembly*; *connector*; *link; lock; mechanism;* ***mount****; pipe; pivot; spring* .

Tabela 5-4 – *PolyAnalyst*: Resultados da Mineração de Textos para a BD *Flexible Joints*

De acordo com os resultados obtidos e comparando-os com os resultados do algoritmo *K-Means*, alguns comentários relevantes tornam-se possíveis.

Ao analisar-se, em primeira instância, o agrupamento *Drill* (número 02) gerado pelo algoritmo em questão em comparação com o mesmo agrupamento gerado pelo algoritmo *K-Means*, verificou-se que grande parte dos termos-chaves é comum a ambos os respectivos resultados. Verificou-se, ainda para tal agrupamento, que dez dos doze

documentos presentes nos resultados do algoritmo *STC* estão presentes nos resultados do algoritmo *K-Means*.

Os resultados do agrupamento *Protection* (número 01), em comparação ao mesmo agrupamento formado em virtude da convergência do algoritmo *K-Means* são semelhantes, apesar de algumas palavras não-coincidentes que existem entre os mesmos. Todavia, isto não representa qualquer tipo de problema, pois por se tratar de métodos algorítmicos que trabalham com eliminação de palavras e / ou com dicionários de sinônimos e de relação entre termos semelhantes (*thesaurus*), a eliminação de alguns termos ou substituição por termos de significados semelhantes torna-se normal dentro do pré-processo. Ainda assim, mesmo que não estejam listadas dentre as palavras mais significativas, somente suas presenças no escopo dos textos durante o processo já contribui para convergência do algoritmo.

Ao analisar o agrupamento *Attacher* (número 01) sugerido pelo algoritmo *K-Means* e compará-lo aos agrupamentos *Fix* (número 03) e *Mount* (número 04) do algoritmo *STC*, é possível a percepção de que todos são muitos semelhantes entre si. No tocante ao desempenho de ambas as técnicas algorítmicas de *Clustering*, comprova-se, assim como ocorrera no estudo de caso anterior presente nesse trabalho, uma tendência do algoritmo *STC* em segmentar os documentos analisados em um maior número de agrupamentos possíveis. Pode-se comprovar, ainda, a presença de cem por cento do número de documentos que pertencem ao agrupamento *Attacher* (número 01) formado pelo algoritmo *K-Means* pertencerem, também, aos dois agrupamentos *Fix* (número 03) e *Mount* (número 04) do algoritmo *STC*, subdividindo-os entre esses dois (dezenove para cada um).

Para ilustrar os resultados comparativos entre as duas técnicas de *Clustering* empregadas, cada uma implementada em sua respectiva plataforma (*RapidMiner / YALE* e

PolyAnalyst), gerou-se um gráfico comparativo de seus resultados, que pode ser representado na figura 5-20.

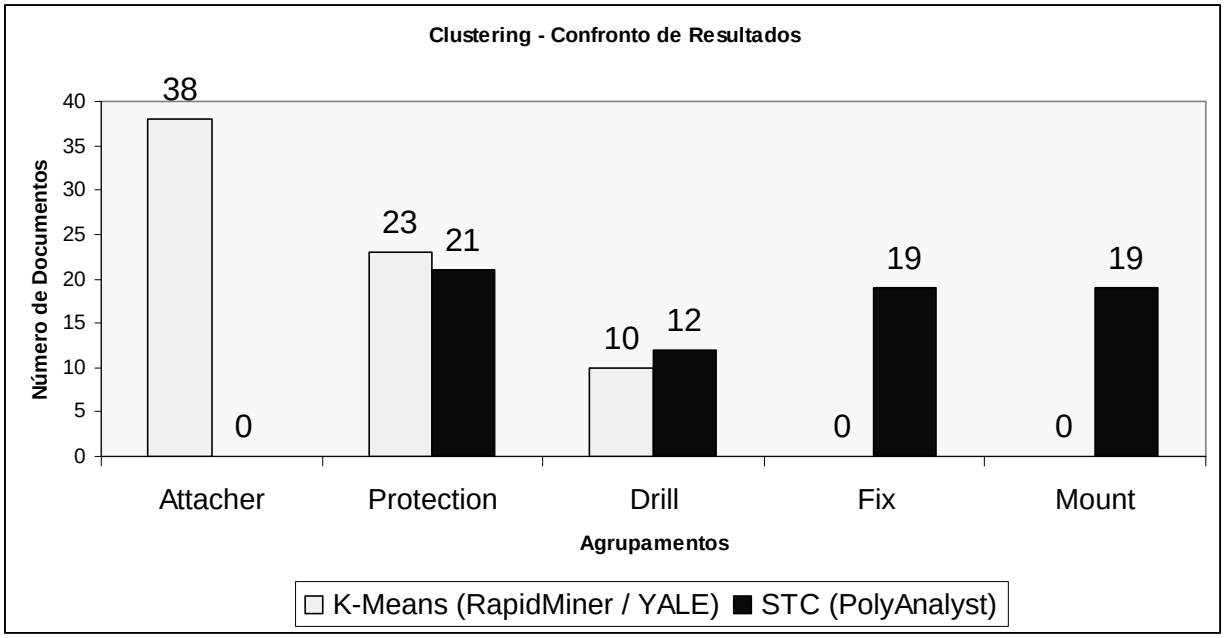

Figura 5-20 Comparação entre os resultados de *Clustering* para a BD *Flexible Joints*

5.4.3. *Flexible Risers*

Ao ter a tabela de dados de valores numéricos resultante do pré-processamento formatada pelo editor de planilhas eletrônicas e submetida ao algoritmo de validação de agrupamentos, as métricas de validação convergiram, depois de muitas iterações do algoritmo, para o resultado ideal o número de agrupamentos igual a dois. Consideraram-se tais resultados obtidos pelas métricas em comunhão com outros fatores já comentados no item 5.4, e chegou-se à conclusão de que as patentes, contidas em presente base de dados textuais seriam classificadas de modo não-supervisionado considerando o número de agrupamentos igual a dois. A tabela 5-5 exibe os resultados dos agrupamentos formados pelo algoritmo de *Clustering.* Após sua exibição, apresentar-se-á a análise mais detalhada sobre cada um dos agrupamentos formados.

Classificação Não-Supervisionada sobre a Base *Flexible Risers*			
Programa: *RapidMiner / YALE*		**Algoritmo empregado: *K-Means***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chave**
1	59	*Pipe*	***Pipe**; install; float; structure; flow; platform; oper; support; valve; mount.*
2	59	*Buoy*	***Buoy***; *line*; *moor*; *release*; *fluid*; *anchor*; *hydrocarbon; transfer*; *buoyant*; *submerse*.

Tabela 5-5 Resultados da classificação não-supervisionada para a BD *Flexible Risers*

Cluster 1 – *Pipe*

Termos- chaves do agrupamento:

Pipe*; install; float; structure; flow; platform; operate; support; valve; mount.*

Títulos que pertencem ao agrupamento:

- *Support and connection device for flexible riser;*
- *Device for suspending flexible and semi-flexible pipes on structures at sea;*
- *System for accessing oil wells with compliant guide and coiled tubing*;
- *Non-rigid marine platform with surface wellheads*.

Principais Classificações:

- E21B 17/00: Equipamentos ou detalhes para perfuração; Equipamento de poços ou de manutenção de poços; Hastes ou tubos de perfuração; Ferramentas flexíveis para perfuração; Hastes quadradas (*Kellies*); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de proteção;
- B63B 21/00: Operações de processamento; transporte; Navios ou outras embarcações; equipamento para navegação; Amarração; Equipamento para deslocar, rebocar ou empurrar; Ancoragem.

Principais Cessionários:

- *Institut Francais Du Petrole (Cedex, FR);*
- Petroleo Brasileiro S.A. - Petrobras (Rio de Janeiro, BR);

ABSTRACT que represente o agrupamento:

An improved method of and structure for mooring for example supertankers offshore which combines the mooring process and the connections of cargo transfer lines into one simple operation and includes a mooring buoy capable of fluid transfer which is particularly adaptable to a super port type operation. The buoy includes a relatively long flotation caisson structure having a relatively small diameter, such proportion insuring greater stability in rough sea conditions. A flow line is located down the center of the buoyant caisson and extends above the bow of the ship it will service. A fitting is located on the top end of the flow line which is capable of securely attaching to a mooring pedestal located on the bow of the ship. This arrangement allows fluid flow between the riser structure and the ship as well as forming a substantial mooring linkage between the ship and the anchor lines which are attached to the bottom of the buoy. The system also includes the use of a mooring winch on the ship. The line from the winch is conducted through the hawse pipe in the center of the mooring pedestal and is attached to a hang line, which extends down from the center of the attaching head of the riser structure. The winch will pull the attaching head down around the mooring pedestal where it is automatically latched. This simple operation moors the ship as well as connects the fluid lines.

Cluster 2 – *Buoy*

Termos-chaves do Agrupamento:

Buoy; *line*; *moor*; *release*; *fluid*; *anchor*; *hydrocarbon*; *transfer*; *buoyant*; *submerse*.

Títulos que pertencem ao agrupamento:

- *Vessel with a disconnectable riser supporting buoy;*
- *Arrangement in a loading/unloading buoy for use in shallow waters;*
- *Vessel mooring system and method for its installation*;
- *Cantenary anchor leg mooring buoy*.

Principais Classificações:

- B63B 21/00: Amarração; Equipamento para deslocar, rebocar ou empurrar; Ancoragem;
- B63B 22/00: Bóias; meios para indicar a localização de objetos submersos;
- B63B 22/02: Bóias; especialmente adaptadas para amarração de embarcações.

Principais Cessionários:

- ***Den Norske Stats Oleselskap A.S. (Stavanger, NO)*- Principal Cessionário*;***
- *FMC Corporation (Chicago, IL);*
- *Head; Philip (London, NW10 7XR, GB);*
- *Headworth; Colin Stuart (Houston, TX);*
- *Institut Francais Du Petrole (Rueil-Malmaison, FR);*
- *Single Buoy Moorings Inc. (Marly, CH);*
- *Sofec, Inc. (Houston, TX).*

ABSTRACT que represente o agrupamento:

The present invention relates to equipment installed in an intermediary region between a floating production unit at the ocean surface and a wellhead of an oil reservoir on the ocean floor. The equipment, known as a subsurface buoy, is designed to support rigid and flexible production tubes used for the transport of reservoir fluids from the oil well and/or fluids used in support systems for the oil reservoir. Four cylindrical bodies 1, 2, 3, and 4, connected at their extremities, form the body of the subsurface buoy. In each of the vertices of the subsurface buoy, components of a tying and dynamically stabilizing system 13 are rigidly connected. This tying and dynamically stabilizing system 13 is configured to control the positioning of the subsurface buoy, the tension of installation chains 15 and anchoring tendons 21, thereby promoting a stabilization of the entire assembly against large-amplitude rotations or angular changes even after rigid tubes 10 and flexible tubes 11 are coupled to the body of the subsurface buoy. In addition, a method to install the subsurface buoy is also presented.

respectivos números de documentos encontrados por cada processo de *Clustering* mencionado.

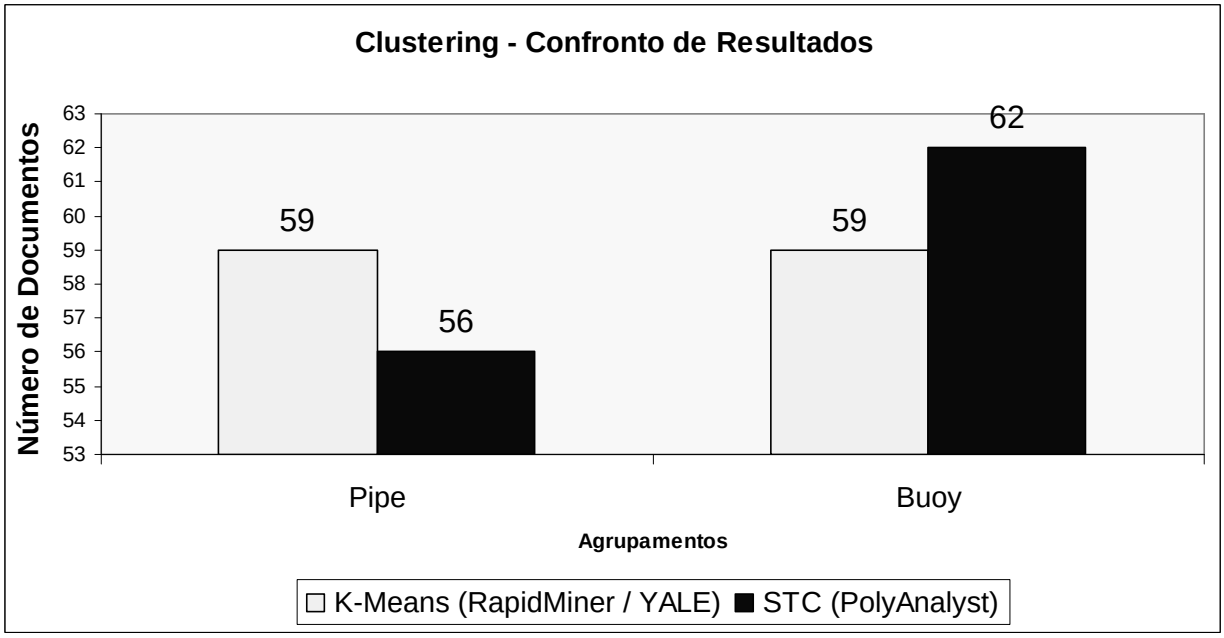

Figura 5-21 Comparação entre os resultados de *Clustering* para a BD *Flexible Risers*

Contudo, essa divergência com relação à quantidade de registros por agrupamento não afetou o contexto da análise: evidencia-se que para a base de dados em questão, esses são os dois temas fortes presentes em seus registros textuais. Além disso, essa diferença de números se deve também às etapas de pré-processo, haja vista que a boa convergência do algoritmo de *Clustering*, bem como seus resultados, depende de um pré-processamento bem executado, eliminando-se de fato todos os dados textuais (termos) não-relevantes; fazendo-se uso do processo de extração de radicais dentre outros procedimentos.

Ao comparar as duas etapas, verificou-se - para este estudo de caso - que o algoritmo de *Stemmer* empregado pelo *software RapidMiner / YALE* atuou de forma mais abrangente que o algoritmo empregado para a plataforma *PolyAnalyst.* Verificou-se, além disso, a necessidade de seu empregar um dicionário de sinônimos e de relacionamento de

termos (*thesaurus*) - oferecido por essa plataforma - capaz de fazer a associação entre palavras sinônimas ou expressões lingüísticas correlacionadas a fim de que se apurassem, com maior precisão, os termos realmente relevantes e a conduzi-los ao processamento em etapa posterior.

5.4.4. *Smart Fields*

Após diversas iterações, a maior parte das métricas de validação convergiu para o resultado de dois *clusters* a serem formados. A considerar tais resultados em conjunto com outros fatores já comentados no item 5.4, concluiu-se que as patentes inclusas no estudo de caso em questão seriam classificadas de modo não-supervisionado considerando o número de agrupamentos igual a dois. A tabela 5-7 exibe os resultados dos agrupamentos formados pelo algoritmo de *Clustering.* Após sua exibição, apresentar-se-á a análise mais detalhada sobre cada um dos agrupamentos formados.

Classificação Não-Supervisionada sobre a Base *Smart Fields*			
Programa: *RapidMiner / YALE*		**Algoritmo empregado: *K-Means***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chave**
1	105	*Configure*	***Configure**; network; interface; connect; associate; provide; display; monitor; module; diagnostic; access; memory; integrate; communicate; manage.*
2	64	*Operate; Maintenance*	***Oper**; **maintenance**; generate; compare; collect; assembly; conduct; monitor; require; produce; sensor; detect; report; electric; voltage.*

Tabela 5-7 Resultados da classificação não-supervisionada para a BD *Smart Fields*

em geral; válvulas; caracterizados por elementos mecânicos; elementos sensíveis; elementos de correção).

Principais Cessionários:

- *Fisher Rosemount Systems, Inc. (Austin, TX).*

ABSTRACT que represente o agrupamento:

A control studio object system is disclosed which allows a process control environment to be easily and quickly configured or modified. The control studio object system includes a stencil portion having stencil items conforming to algorithms and a diagram portion to which the stencil items may be copied via a drag and drop operation. Because the stencil items are objects which contain all of the information required by a diagram portion to create an object that contains all of the information necessary to program the process control environment, the completed diagram portion reflects the actual configuration of the process control environment. Additionally, providing the stencil items as objects allows the diagrammed environment to be installed directly to nodes without requiring the diagram to be compiled or rewritten in a language conforming to the node.

Cluster 2 – *Operate; Maintenance*

Termos-chaves do Agrupamento:

Oper; ***maintenance***; *generate*; *compare*; *collect*; *assembly*; *conduct*; *monitor*; *require*; *produce*; *sensor*; *detect*; *report*; *electric*; *voltage*.

Títulos que pertencem ao agrupamento:

- *Assembly for remote control and/or remote operation of a field device by means of a controller via a field bus;*
- *Generation of data indicative of machine operational condition;*
- *Method of providing maintenance services;*
- *System and method for condition-based maintenance;*
- *Electric actuator for fluid control valves.*

Principais Classificações:

De todas as subseções pertencentes a esse agrupamento, a grande maioria pertence ao subgrupo G05B, que significa:

- Classe G: Física;
- Seção 05: Controle; Regulagem;
- Subclasse B: Sistemas de controle ou regulagem em geral; elementos funcionais de tais sistemas; disposições de monitoração ou de teste para tais sistemas ou elementos (acionadores movidos por pressão de fluido ou sistemas que atuam por meio de fluídos em geral; válvulas; caracterizados por elementos mecânicos; elementos sensíveis; elementos de correção).

Principais Cessionários:

- *Fisher Rosemount Systems, Inc. (Austin, TX).*

ABSTRACT que represente o agrupamento:

The present invention relates to an interface for a maintenance system used in conjunction with a process instrumentation system. More specifically, the invention relates to an interface used for maintaining and configuring smart devices. Even more specifically, the invention relates to an interface that may safely be used to maintain and configure smart devices where such smart devices are located in hazardous areas. The interface, which may be removably mounted to a termination board, has a control section, a port replacement section, a permanent storage section, a temporary storage section, an address/data bus, a UART, a standard clock pulse generation device, an option select device, a modem, a channel selection decoder, a wave shaping device, and at least one multiplexer. The interface and a termination board for use in process instrumentation systems that require intrinsic safety are explained by themselves and in system and function contexts.

5.4.4.1. Resultado de *Smart Fields* obtido pelo *software PolyAnalyst*

A submissão da base textual de dados tratada em presente estudo de caso aos processos de Mineração de Textos pelo *software PolyAnalyst* nortearam os resultados para um número sugerido de três agrupamentos, que podem ser visualizados na tabela 5-8 .

Classificação Não-Supervisionada sobre a Base *Smart Fields*	
Programa: *PolyAnalyst*	**Algoritmo empregado:**

foi subdividido pelo algoritmo *STC* da plataforma *PolyAnalyst* em dois agrupamentos distintos entre si, que foram denominados *Maintenance* e *Detection*, com 35 e 29 registros textuais respectivamente. Há de se ressaltar, porém, que existe a semelhança entre os termos-chaves desses dois agrupamentos unidos e os termos-chaves do agrupamento denominado [*Operation; Maintenance*], de forma a se supor que o algoritmo *STC* subdividiu-o, mantendo, no entanto, a coerência e a confiabilidade nos resultados encontrados entre ambos os processos de *Clustering* empregados.

Para ilustração dos resultados, gerou-se um gráfico comparativo que pode ser observado na figura 5-22.

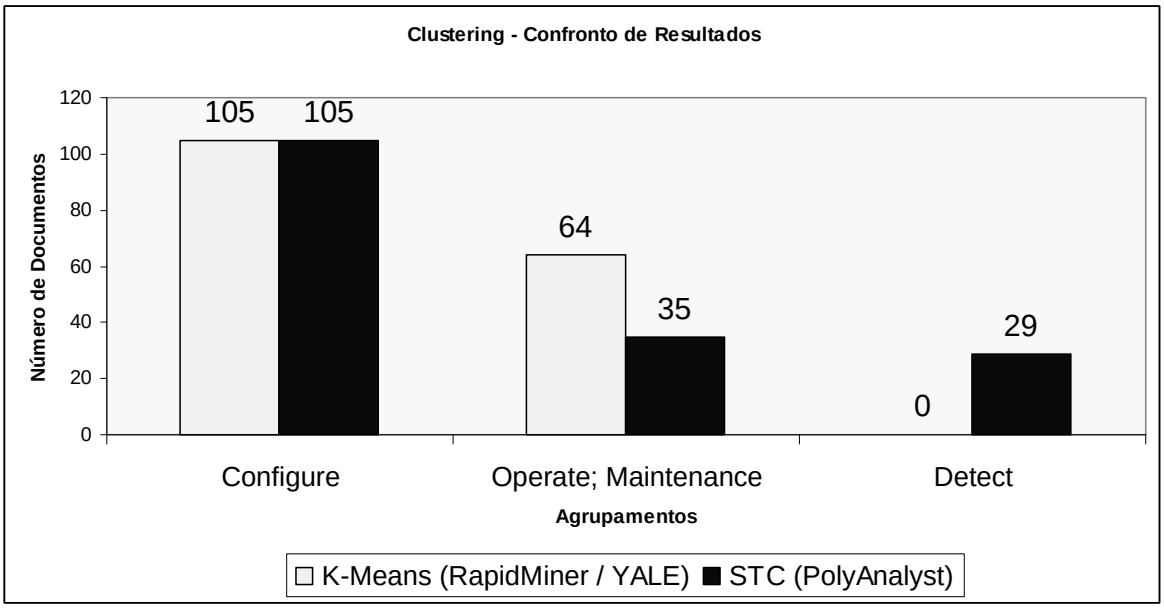

Figura 5-22 Comparação entre os resultados de *Clustering* para a BD *Smart Fields*

5.4.5. *Smart Wells*

Após diversas iterações do algoritmo de validação sobre a tabela de dados formada, concluiu-se que as métricas não conseguiram retornar um resultado preciso de qual seria o número ideal de agrupamentos para o conjunto textual de dados em questão. Todas as quatro métricas oscilaram seus valores mais prováveis de indicação entre dois e

cinco agrupamentos. Dessa forma, analisaram-se os gráficos de centróides gerados pelo algoritmo *K-Means* (utilizando o *software RapidMiner / YALE*) ao final de suas execuções, variando o valor da variável de particionamento K entre dois e cinco. A análise sobre cada gráfico norteou o resultado mais provável para um número de dois agrupamentos, devido à maior distância inter-*cluster* entre seus respectivos centróides. A tabela 5-9 permite a visualização dos resultados dos agrupamentos gerados pelo algoritmo de *Clustering*. Suas análises mais detalhadas para a base de dados em questão serão apresentadas logo a seguir.

Classificação Não-Supervisionada sobre a Base *Smart Wells*			
Programa: *RapidMiner / YALE*		**Algoritmo empregado: *K-Means***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chave**
1	61	*Control*	***Control;*** *monitor**;** sensor; connector; wellbore; flow; multiplex; fluid; screen; electric.*
2	14	*Downhole*	***Downhole****; surface; wellbore; connect; flow; channel; multiplex; mount; tube; drill.*

Tabela 5-9 Resultados da classificação não-supervisionada para a BD *Smart Wells*

Cluster 1 – *Control*

Termos-chaves do agrupamento:

Control*; *monitor; *sensor*; *connector*; *wellbore*; *flow*; *multiplex*; *fluid*; *screen*; *electric*.

Títulos que pertencem ao agrupamento:

- *Wells communication system;*
- *System for use in controlling a hydrocarbon production well;*
- *Apparatus for sensing fluid in a pipe*;
- *Method for measuring properties of flowing fluids, and a metering device and a sensor used for performing this method.*

Principais Classificações:

A grande maioria dos documentos desse agrupamento possui classificação internacional E21B nas seguintes subclasses:

- E21B 43/00: Construções Fixas; Perfuração do solo; Mineração; Perfuração do solo, por exemplo, perfuração profunda; Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços (aplicáveis somente à água E03B; obtenção de jazidas petrolíferas ou de matérias solúveis ou fundíveis por técnicas de mineração E21C 41/00; bombas F04);
    - 43/12: Métodos ou aparelhos para controlar o fluxo de fluidos obtido para ou em poços (43/25 tem prioridade; disposições de válvulas 34/00);
    - 43/14: Extração de um poço de zonas múltiplas;
- E21B 47/00: Levantamento de furos de sondagem ou de poços (controle de pressão ou do fluxo de fluidos de perfuração 21/08; perfilagem geofísica G01V).

Principais Cessionários:

- *Schlumberger Technology Corporation (Sugar Land, TX).*

ABSTRACT que represente o agrupamento:

An air monitoring system is disclosed having an air monitoring unit with at least one sensor for measuring data of an air quality parameter and a computer for storing the air quality parameter data received from the sensor. The air monitoring unit may use an installed or a portable system, or a combination of both, for measuring the air quality parameters of interest. A remote data center may be provided, and the data may be uploaded to the data center from the unit by a communications media such as the Internet. Information or instructions may also be downloaded from the data center to the unit via the communications media for controlling or modifying the function of the unit. An expert system may be provided with the air monitoring system for controlling the unit. The information or instructions downloaded to the unit may be generated by the expert system.

Cluster 2 – *Downhole*

Termos- chaves do agrupamento:

Downhole; *surface; wellbore; connect; flow; channel; multiplex; mount; tube*; *drill*.

Títulos que pertencem ao agrupamento:

- *Apparatus for receiving downhole acoustic signals;*
- *Downhole multiplexer and related methods;*
- *Downhole flow control tool;*
- *Method and system for wireless communications for downhole applications.*

Principais Classificações:

A grande maioria dos documentos eletrônicos presentes nesse agrupamento pertencem à classificação E21B nas seguintes subclasses:

- E21B 43/00: Construções Fixas; Perfuração do solo; Mineração; Perfuração do solo, por exemplo, perfuração profunda; Métodos ou aparelhos para obter óleo, gás, água, matérias solúveis ou fundíveis ou de lama minerais de poços (aplicáveis somente à água E03B; obtenção de jazidas petrolíferas ou de matérias solúveis ou fundíveis por técnicas de mineração E21C 41/00; bombas F04);
    - 43/02: Filtragem do subsolo;
    - 43/04: guarnição de poços com tubos de cascalho;
    - 43/08: Filtros ou camisas;
    - 43/11: Perfuradores; Penetradores;
- E21B 47/00: Levantamento de furos de sondagem ou de poços (controle de pressão ou do fluxo de fluidos de perfuração 21/08; perfilagem geofísica G01V).

- 47/12: Meios para transmitir sinais de medição do poço para a superfície, por exemplo, perfilagem durante a perfuração.

Principais Cessionários:

- *Schlumberger Technology Corporation (Sugar Land, TX).*

ABSTRACT que represente o agrupamento:

The present invention provides a downhole method and apparatus using a flexural mechanical resonator, for example, a tuning fork to provide real-time direct measurements and estimates of the viscosity, density and dielectric constant of formation fluid or filtrate in a hydrocarbon producing well. The present invention additionally provides a method and apparatus for monitoring cleanup from a leveling off of viscosity or density over time, measuring or estimating bubble point for formation fluid, measuring or estimating dew point for formation fluid, and determining the onset of asphaltene precipitation. The present invention also provides for intercalibration of plural pressure gauges used to determine a pressure differential downhole. A hard or inorganic coating is placed on the flexural mechanical resonator (such as a tuning fork) to reduce the effects of abrasion from sand particles suspended in the flowing fluid in which the flexural mechanical resonator is immersed.

5.4.5.1. Resultado de *Smart Wells* obtido pelo *software PolyAnalyst*

Através dos ajustes de valores sobre os parâmetros de *Clustering* oferecidos pelo algoritmo *STC* empregado pela plataforma *PolyAnalyst* para o processo de Mineração de Textos, a base textual de dados tratada em questão foi particionada em dois agrupamentos concisos e distintos entre si – tal como já ocorrera com a aplicação do algoritmo *K-Means* empregado pelo *software RapidMiner / YALE*. Dentre os particionamentos gerados, observa-se a existência de uma semelhança significativa entre eles: divergem-se quanto ao número de documentos presentes em cada agrupamento obtido por cada algoritmo, contudo, coincidem entre si com relação à semelhança dos termos-chaves presentes em cada um deles. Portanto, conclui-se que os dois agrupamentos encontrados são os assuntos mais relevantes com relação ao assunto tratado em presente estudo de caso. A tabela 5-10 ilustra os resultados obtidos pela ferramenta *PolyAnalyst*. A figura 5-23 permite visualizar os resultados obtidos por cada um dos algoritmos empregados.

Classificação Não-Supervisionada sobre a Base *Smart Wells*			
Programa: *PolyAnalyst*		**Algoritmo empregado: *Suffix Tree Clustering***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chave**
1	63	*Control*	***Control**; monitor; fluid; flow; connect; surface; sensor; tube; pressure; attach.*
2	12	*Downhole*	***Downhole**; wellbore; hole; leave; fluid; migrate; channel; encapsule; tube; drill.*

Tabela 5-10 *PolyAnalyst*: Resultados da Mineração de Textos para a BD *Smart Wells*

thesaurus no pré-processo elaborado pela plataforma *PolyAnalyst* para que houvesse a redução dos termos a fim de facilitar o processo computacional de mineração dos dados.

5.4.6. *Steel Catenary Risers*

Como resultado da submissão da tabela de valores formada aos cuidados do algoritmo de validação de agrupamentos após muitas iterações, três das quatro métricas empregadas convergiram o resultado para um número sugerido de dois agrupamentos. A

Cluster 1 – *Riser*

Termos- chaves do agrupamento:

Riser; *flexible*; *joint*; *pipe*; *float*; *connect*; *assembly*; *vessel*; *structure*; *tension*; *surface*; *bottom*.

Títulos que pertencem ao agrupamento:

- *Marine bottomed tensioned riser and method;*
- *Top tensioned riser;*
- *Tethered buoyant support for risers to a floating production vessel;*
- *Hybrid riser for deep water.*

Principais Classificações:

- E21B 17/00: Construções Fixas; Perfuração do solo; Mineração; Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços; Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção (engates de hastes em geral F16D; tubos ou acoplamentos de tubos em geral F16L);
- E21B 17/01: Tubos ascendentes (conectores de tubos ascendentes);
- F16L 1/20: Tubos; juntas ou acessórios para tubos; suportes para tubos, cabos ou tubulação de proteção; meios para isolamento térmico em geral; Acessórios para esse fim, por exemplo; flutuadores, pesos (bóias B63B 22/00).

Principais Cessionários:

- *FMC Corporation (Chicago, IL).*

ABSTRACT que represente o agrupamento:

A new marine oil production riser system for use in deepwater applications is disclosed. An efficient means for accommodating movements of the host facility, while maintaining riser top tension within the limits for long-term riser performance. Long riser stroke lengths can be accommodated without requiring complex interfacing with the topsides. The riser assembly comprises: a generally extendable substantially non-vertical section having an upper end adapted to be in flow communication with a generally vertical marine riser carried by a facility floating on the surface of a body of water, and having a lower end adapted to be in flow communication with a fluid source on the seafloor; and tensioning means, mechanically connecting the upper end of the marine riser with the lower end of the marine riser, for biasing said ends towards each other. The tensioning means comprises: a cylinder having one end open to sea pressure, having an opposite end sealed from sea pressure, and connected to one end of the marine riser; a piston within the cylinder disposed for movement within the cylinder; and a piston rod passing through the opposite end of the cylinder and having one end connected to the other end of the marine riser.

Cluster 2 – *Tension*

Termos- chaves do agrupamento:

***Tension**; monitor; stress; assembly; connect; control; buoy; chain; attach; pipe; support; marine.*

Títulos que pertencem ao agrupamento:

- *Constant tension steel catenary riser system;*
- *Subsurface buoy and methods of installing, tying and dynamically stabilizing the same;*
- *Apparatuses and methods for monitoring stress in steel catenary risers;*
- *Stress limiting device for offshore oil reservoir production pipe;*

Principais Classificações:

- B63B 21/00: Operações de processamento; transporte; Navios ou outras embarcações; Equipamento para navegação; Amarração; Equipamento para deslocar, rebocar ou empurrar; Ancoragem;
- E21B 17/00: Construções Fixas; Perfuração do solo; Mineração; Perfuração do solo, por exemplo, perfuração profunda; Obtenção de óleo, gás, água, materiais solúveis ou fundíveis ou uma lama de minerais de poços; Hastes ou tubos de perfuração; Ferramentas flexíveis de perfuração; Hastes quadradas (“Kellies”); Comandos; Hastes de sucção; Tubulação de revestimento; Tubos de produção (engates de hastes em geral F16D; tubos ou acoplamentos de tubos em geral F16L).

Principais Cessionários:

- *Mobil Oil Corporation (Fairfax, VA); e Shell Oil Company (Houston, TX).*

ABSTRACT que represente o agrupamento:

A steel catenary riser (SCR) system includes a tensioning mechanism on a floating facility that controllably applies a substantially constant tension to an SCR that is fluidly coupled to the facility by a flexible jumper conduit. More specifically, the system includes a tensioning device located on the floating facility; and an SCR having an upper portion, which, in the preferred embodiment, extends above the surface of the body of water. The upper portion of the SCR is connected to the tensioning device by a connection element, such as a cable, chain, rope, or wire, whereby tension is controllably applied from the tensioning device to the SCR. A flexible jumper conduit is fluidly connected between the upper portion of the SCR and the floating facility for conducting fluid from the SCR to the floating facility. In a preferred embodiment, the connection element is attached to the upper portion of the SCR at an attachment point, and the flexible jumper conduit is fluidly coupled to the SCR near the attachment point.

5.4.6.1. Resultado de *Steel Catenary Risers* obtido pelo *software PolyAnalyst*

Tal como obtido pelo *software RapidMiner / YALE*, a submissão da respectiva base textual de dados às tarefas inerentes ao pré-processamento e ao processamento de dados pelo *software PolyAnalyst* norteou o resultado para um número sugerido de dois agrupamentos. Seus resultados podem ser visualizados na tabela 5-12.

Classificação Não-Supervisionada sobre a Base *Steel Catenary Risers*			
Programa: *PolyAnalyst*		**Algoritmo empregado: *Suffix Tree Clustering***	
Agrupamento	**Número de Patentes**	**Nome**	**Palavras-chaves**
1	19	*Riser*	***Riser***; *tube*; *rigid*; *bottom*; *attach*; *fluidly connect*; *curve*; *flexible*; *chain*; *pipe string*.
2	08	*Tension*	*Surface*; *connect*; *float*; *support*; ***tension***; *fluid*; *structure*; *assembly*; *control*; *pipe*.

Tabela 5-12 *PolyAnalyst*: Resultados da Mineração de Textos para a BD *Steel Catenary Risers*

Apesar das diferenças nos processos computacionais entre os algoritmos de *Clustering* empregados – *K-Means* e *STC* – nesse caso (bem como já houvera em outros casos presentes nesse trabalho) não houve divergência quanto ao número sugerido de agrupamentos formados. Os termos-chaves entre os agrupamentos *Riser* e *Tension* formados pelos métodos apresentaram-se muito semelhantes entre si, assim como o número de documentos presentes nos agrupamentos formados. O gráfico que confronta os resultados obtidos pode ser visualizado na figura 5-24.

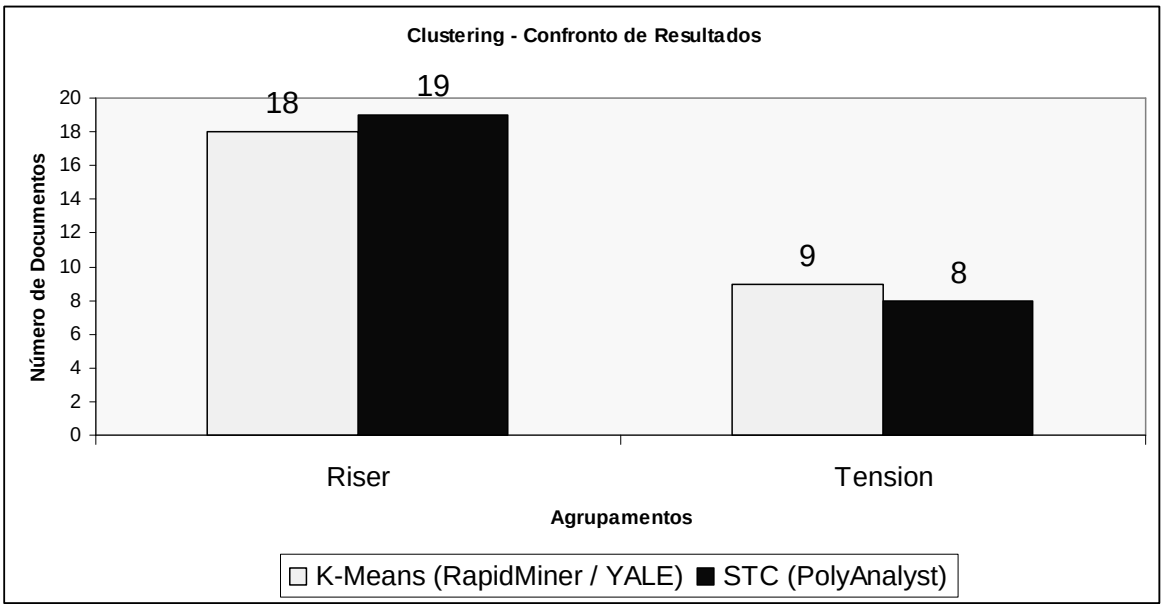

Figura 5-24 Comparação entre os resultados de *Clustering* para a BD *Steel Catenary Risers*

5.5. Conclusões

A aplicação das etapas provenientes de Mineração de Textos sobre o campo *ABSTRACT* (*RESUMO*) das patentes industriais norte-americanas coletadas demonstrou o quão importante podem ser tais documentos na obtenção de conhecimentos e geração de Inteligência Competitiva. Em cumprimento à sua proposta, a pesquisa destacou as palavras mais significativas de cada estudo de casos analisado ao final dos processos, e dessas palavras pode ser ter uma idéia de tecnologias e materiais interessantes e de grande valia que podem estar implícitos em seus campos textuais analisados.

Contudo, somente a análise de um especialista da indústria de Petróleo e Gás é capaz de validar os resultados, o que não é de competência de presente pesquisa. Assim como é de competência do próprio analisar e investigar se há possibilidades ou não de correlação entre si de dois ou mais temas abordados.

O aperfeiçoamento sobre a ferramenta *Web Crawling* facilitou os processos de busca e recuperação das informações, pois foi capaz de extrair os conteúdos de real interesse pertencentes aos documentos eletrônicos relativos a cada estudo de caso planejado para esse trabalho de pesquisa, além de transformá-los em documentos de textos simples prontos para submissão às etapas de mineração de dados de cada uma das plataformas utilizadas.

A elaboração de análises estatísticas sobre cada um desses estudos no campo específico e destinado às classificações internacionais das patentes (CIP) relacionadas verificou a natureza das atividades das patentes industriais coletadas e seus principais materiais presentes. Tal averiguação pode prover suporte a um especialista da área, dando-lhe consistência e confiabilidade em suas conclusões em torno dos resultados oriundos do processo de mineração de textos.

Já a análise estatística elaborada sobre o campo específico das patentes destinado aos cessionários / inventores permitiu o conhecimento das principais organizações que exploram essas tecnologias descobertas em favor da indústria de extração de Petróleo e Gás. Dessa forma, essas organizações merecem destaque por suas atitudes pró-ativas diante de uma fatia do mercado que a cada dia mais acentua o seu crescimento, segundo as comprovações feitas pelas análises estatísticas elaboradas em presente pesquisa sobre o número de patentes industriais registradas ano após ano na instituição norte-americana.

A submissão das seis bases textuais formadas em duas plataformas diferentes de extração de conhecimentos sobre dados não-estruturados foi capaz de gerar, após diversas iterações dos algoritmos de *Clustering*, divergentes resultados em três dos seis estudos de caso, no tocante ao número de agrupamentos indicados pelo processo de Classificação Não – Supervisionada. De uma forma geral e baseando-se em outros estudos anteriores, quando dois algoritmos apresentam estimativas de números de agrupamentos divergentes para uma mesma base de dados, pode-se avaliar tal divergência em três etapas: na escolha do algoritmo que será pontuado; no critério de pontuação atribuído pelo algoritmo ao índice; ou na medida de distância ou similaridade utilizada pelo índice.

Porém, no tocante aos conhecimentos extraídos por cada uma das duas plataformas de execução em cada uma das seis bases formadas, pôde-se verificar a semelhança entre os conhecimentos obtidos. Tal comprovação respalda-se na análise comparativa sobre os termos-chaves de cada um dos *clusters* formados para cada base. Além disso, nos três casos específicos em que se comprovou divergência quanto ao número de agrupamentos, verificou-se que um agrupamento formado pelo algoritmo *K-Means (RapidMiner / YALE)* decompôs-se em dois formados pelo algoritmo *STC (PolyAnalyst)*, de forma que os dois conjuntos de termos-chaves juntos formados por esse segundo algoritmo representa o conhecimento do agrupamento formado pelo primeiro

algoritmo citado. Em síntese: os conhecimentos foram extraídos, cada qual pela forma determinada pela plataforma no qual os algoritmos de *Clustering* estão implementados. No caso da plataforma *RapidMiner / YALE*, verificou-se uma tendência em generalizar os dados, ao passo que na plataforma *PolyAnalyst*, houve uma especificação mais detalhada. Isso se deve, também, ao fato de que cada plataforma possui um algoritmo diferente, tanto para as etapas de pré-processamento, quanto de validação de números de agrupamentos e de *Clustering* também.

A debruçar-se sobre o desempenho da plataforma *RapidMiner / YALE* para o processamento textual de patentes, pode-se afirmar que sua utilização apresentou resultados valiosos em todos os estudos de casos elaborados. Para utilização de tal ferramenta computacional proposta, no entanto, foi preciso se realizar um preparo sobre os documentos eletrônicos coletados e isso se deve ao fato de que as patentes contêm suas informações estruturadas em campos específicos que delimitam, em cada campo, suas pertinentes informações. No tocante à configuração de suas tarefas a serem executadas, constatou-se a presença de seus comandos oferecidos em forma de operadores. Esse fato permite a estruturação dos processos em formato de árvore, de maneira funcional ao usuário, facilitando assim, a compreensão de cada etapa de forma isolada. A configuração dos parâmetros de cada operador é bem simples e agradável ao usuário; basta ao mesmo ser conhecedor dos princípios de Mineração de Textos para poder configurá-los de forma coerente.

Para os presentes estudos de casos realizados com a plataforma *RapidMiner / YALE*, a inserção de uma lista de *StopWords*, além da relação de palavras não-significativas que já faz parte do filtro da plataforma tais como advérbios, interjeições, conjunções, preposições entre outras, reduziu consideravelmente o tempo de execução do algoritmo *K-Means* para *Clustering*, pois eliminou toda a carga não-necessária de termos

que em nada contribuiriam para a geração de resultados que expressassem conhecimentos relevantes. As eliminações dos termos de maiores e menores freqüências de forma manual também foram de grande valia, pois permitiu que a plataforma equilibrasse os resultados finais obtidos e manipulasse os documentos presentes de forma adequada. Com a adoção desse critério para cada base, a ferramenta utilizou os diferentes atributos dos documentos presentes, de modo a não considerar somente os seus temas centrais, generalizando-os, e a não especificar demais os seus temas tratados, desviando o foco e eliminando a essência da análise.

Os algoritmos de *Stemmer* de *Porter* e de *Tokenizer* empregados pelo *RapidMiner / YALE* na etapa de pré-processamento obtiveram resultados de grande valia, assim como o algoritmo de *Clustering K-Means*, que inicializou os centróides iniciais de cada agrupamento através de um processo de escolha aleatória, e teve seu processamento facilitado pelo ótimo ganho computacional obtido pela fase de pré-processamento.

Por ser uma plataforma em constante aperfeiçoamento, a plataforma *RapidMiner / YALE* ainda não foi capaz de oferecer um algoritmo de validação de agrupamentos de forma clara e objetiva ao pesquisador (usuário) para presente pesquisa. Tal problema pôde ser minimizado com a adoção do algoritmo de validação de *clusters* e com o gráfico de distância dos centróides gerado pela plataforma computacional em questão.

Em relação à plataforma *PolyAnalyst*, pode-se afirmar que seus resultados também foram muito interessantes, tal como a plataforma descrita anteriormente, porém mais específicos na geração de agrupamentos que traduzissem conhecimentos sobre as bases textuais de dados analisadas. Seus operadores fazem parte de uma lista, de forma que o usuário possa escolhê-los para a área de execução e estruturar o processo em formato de fluxograma, ligando-os entre si. Tal como a plataforma *RapidMiner / YALE*, os comandos são de fácil configuração, e assim o usuário realiza de forma isolada a cada um.

Com relação às suas etapas pertencentes à fase de pré-processamento, uma lista de *StopWords* baseada no dicionário *WordNet* já vem pronta na plataforma *PolyAnalyst*, com os termos básicos compostos por palavras gramaticais que nada representam de informação importante sobre os textos analisados. Essa mesma lista permite a inserção de palavras novas a critério do usuário, e é muito fácil de manipulá-la. O processo de *Stemmer* de tal plataforma, contudo, não é tão fácil de manipulação como na plataforma *RapidMiner / YALE*, pois exige que o usuário insira manualmente as palavras primitivas que deseja, para que o algoritmo procure e processe as suas palavras derivadas. Evidencia-se, com isso, que o trabalho manual do usuário é maior nessa etapa. Em contrapartida, a utilização de seu *Thesaurus* oferecido melhorou consideravelmente o pré-processamento, assim como a substituição de alguns termos lingüísticos não identificados pelo algoritmo de *Stemmer*, pois provavelmente são termos específicos criados para uma tecnologia específica ou, simplesmente, são gírias, razões pelas quais não constam no dicionário de língua inglesa.

No tocante ao processo de *Clustering*, por executar um algoritmo diferente da plataforma *RapidMiner / YALE* (que executa o tradicional *K-Means* conforme já fora comentado), alguns parâmetros merecem uma maior atenção para a aferição de valores, de forma com que os resultados não se tornem incoerentes entre si. Ao contrário do algoritmo *K-Means* – que basta definir somente o número de *clusters* a ser gerado – o algoritmo *STC* empregado para a plataforma *PolyAnalyst* requer a definição de outros parâmetros diferentes. Por exemplo: número mínimo e número máximo de documentos em cada agrupamento, coeficiente de significância de uma expressão dentro de um texto em comparação à significância de uma única palavra dentro desse mesmo texto (isso faz com que uma expressão seja tratada como se ela toda fosse um termo-chave) entre outros parâmetros. Dessa forma, esse tipo de algoritmo empregado requer um estudo mais elaborado por parte do pesquisador, além de sua coerente percepção com relação ao

número final de *clusters* e ao percentual de documentos realmente processados na base submetida – tudo isso em favor da obtenção de resultados confiáveis ao final do processo de mineração de textos.

6. Considerações Finais

As patentes industriais são de fato, nos dias atuais, um instrumento de exploração de grande valia para a descoberta de competências que provêem suporte ao tomador de decisões dentro de uma organização. Trata-se de documentos dotados de informações técnicas e comerciais que, quando bem explorados, viabilizam a descoberta de interessantes conhecimentos sobre produtos, tecnologias e materiais utilizados nos empregos de tais tecnologias, podendo, assim, nortearem-se tendências de investimentos dentro de um mercado consumidor cada vez mais concorrido.

A utilização das patentes industriais para geração de Inteligência Competitiva tende a crescer à medida que o empresário conscientizar-se de que tais informações de grande preciosidade presentes em seus escopos são capazes de retornar benefícios em favor de seu próprio negócio, além das facilidades na captura dessas informações através de tecnologias computacionais de alto desempenho.

Pois nos dias atuais, com o processo de globalização que intensificou a expansão das Tecnologias de Informação e Comunicação, podem-se encontrar patentes industriais hospedadas em *sites* de instituições espalhadas ao redor do mundo. Além disso, existem *sites* com mecanismos de pesquisas orientados para as buscas por patentes, tal como *Free Patents On Line* e *Google*, que facilitam esse processo de busca e retorno de informação, permitem retornar uma lista ordenada de documentos por ordem de significância com relação aos termos-chaves digitados para pesquisa, e selecionar os documentos pertencentes a uma determinada tecnologia que se deseja consultar e explorar.

Em presente dissertação, os esforços concentraram-se na necessidade de demonstração das reais importâncias dos documentos de propriedade intelectual para geração de Inteligência Competitiva com adoção de tecnologias computacionais de alto

desempenho. Elegeram-se, então, as patentes industriais norte-americanas hospedadas no *site* da *USPTO*, e seis assuntos em fase de emergência dentro da indústria de exploração de Petróleo e Gás, por ser uma área reconhecidamente em expansão, não somente no Brasil como em países que investem em pesquisas nas áreas tecnológicas espalhados pelo mundo. Procurou-se definir quais os seis assuntos que seriam abordados através de investigações e conversas com professores e pesquisadores que se interessam por essa área da indústria.

A definição da exploração das patentes industriais hospedadas no *site* da instituição norte-americana *USPTO* deveu-se ao fato de que os Estados Unidos da América são um país investidor em pesquisas por tecnologia, além do intuito de se explorar as patentes industriais editadas em idioma inglês. No caso da *USPTO*, o próprio *site* da instituição forneceu um rápido – e de fácil manuseio – mecanismo de pesquisa por palavras-chaves, que facilitou a coleta dos documentos eletrônicos.

Elaborar análises estatísticas sobre determinados campos das patentes, tais como Classificação Internacional e cessionários / inventores conforme se abordaram em presente dissertação, permitem aos analistas de negócios visualizarem, dentre outros benefícios mais, se um produto ou tecnologia está em fase de emergência ou em declínio dentro do mercado; quais são as tecnologias e materiais possíveis de serem utilizados para sua devida exploração ou aperfeiçoamento; e quais são as organizações que obtiveram concessões sobre suas invenções – posicionando-as como concorrentes frente a uma determinada tecnologia ou produto que a organização queira investir.

No caso da proposta de se extrair Inteligência Competitiva sobre os campos *ABSTRACT* (RESUMO) de cada patente, a aplicação dos processos de Mineração de Textos propiciou a descoberta de resultados interessantes em cada uma das seis bases textuais de dados definidas, através do processo de agrupamento de documentos similares

entre si e da extração de seus assuntos principais presentes nas principais palavras-chaves encontradas em cada agrupamento formado em cada estudo de caso.

Procurou-se elaborar estudos diferenciados para cada base textual, aplicando-se duas plataformas computacionais diferenciadas desenvolvidas para geração de Inteligência Competitiva, dentre muitas outras que existem no mercado para os mesmos propósitos. Cada plataforma aplicada utilizou um diferente algoritmo de Segmentação de Dados *(Clustering).* Em ambas as aplicações para as bases textuais de dados, revelaram-se agrupamentos bem definidos, comprovando que a tecnologia pode ser de grande eficiência. Assim sendo, por essa experiência demonstrou-se que existem diferentes métodos de extração de conhecimentos que apresentam resultados bem interessantes que podem ser utilizados para determinados fins.

No entanto, algumas considerações tornam-se de grande importância. O sucesso na obtenção dos resultados com um tempo menor de convergência do algoritmo empregado para segmentação de dados não-estruturados depende muito dos esforços que são dedicados nas etapas de pré-processamento. A análise sobre tais resultados obtidos em cada método é de competência do profissional responsável pela extração de conhecimento, e para total comprovação a respeito da relevância sobre os conhecimentos descobertos e as hipóteses levantadas, torna-se necessário consultar-se especialistas sobre os assuntos referentes à área pesquisada. Com auxílio desses especialistas, os resultados provenientes dos processos de mineração podem ser validados e convertidos em conhecimentos para que as organizações interessadas possam usufruir seus benefícios em favor de seu caráter mantenedor e competitivo perante o mercado.

6.1. Heranças da dissertação para futuras pesquisas

Como herança, a pesquisa desenvolvida deixa algumas propostas interessantes para trabalhos futuros.

Com relação à Indústria de exploração de Petróleo e Gás, pode-se realizar uma pesquisa sobre outros temas interessantes e que podem ser alvos de prospecção tecnológica, tais como foram os temas abordados nessa dissertação, e submetê-los aos processos de extração de conhecimentos utilizando-se outras ferramentas computacionais presentes no mercado para geração de Inteligência Competitiva, tais como: *Focus* desenvolvido pela empresa *Wisdomain* (WISDOMAIN, 2006), *Aureka* desenvolvido pela empresa *MicroPatent* (MICROPATENT, 2006*), Insight Discoverer Clusterer* (IDC, 2006) desenvolvido pela empresa *Temis* (TEMIS, 2006) e *Statistica* desenvolvida pela *StatSoft* (STATSOFT, 2006).

Sobre os ambientes computacionais empregados, por se tratarem de plataformas que provêem suporte, também, aos processos de extração de conhecimentos em dados (KDD), podem-se realizar novos experimentos com os mesmos *softwares* utilizados em presente pesquisa para exploração de bases de dados estruturadas. Em se tratando de patentes, poder-se-ia aplicar tais ferramentas em processos de mineração sobre os campos categóricos presentes nos documentos.

Ainda com relação aos ambientes computacionais e no tocante à extração de conhecimentos sobre os campos presentes nas patentes, os mesmos *softwares* empregados em presente dissertação – bem como outros *softwares* existentes para os mesmos fins e alguns deles, inclusive, já citados – poderiam ser aplicados sobre o campo *DESCRIPTION* (DESCRIÇÃO) de cada patente, em uma tentativa de se descobrir informações mais abrangentes correlatas a outros temas interessantes de exploração.

Com relação à plataforma *RapidMiner / YALE*, por ser um *software* de código-fonte livre, alguns melhoramentos sobre sua plataforma podem ser implementados, como por exemplo, um algoritmo de validação baseados em métricas mais confiáveis – como o índice PBM – capaz de nortear o usuário a conclusões mais claras sobre o número ideal de agrupamentos a seres adotados dentro de um processo; geração de novos gráficos de apoio para os resultados de *Clustering*; implementação de um *thesaurus* em sua plataforma para otimização do pré-processamento; e melhoramento sobre a função de *Crawler* já existente em seu ambiente. No tocante aos trabalhos, pode-se testá-lo com outros algoritmos de *Clustering* já disponibilizados pela plataforma e em trabalhos futuros que requeiram a análise sobre modelos de séries-temporais.

A plataforma *PolyAnalyst* trabalha com um algoritmo de agrupamentos que produziu bons resultados para presente dissertação e cujo sucesso nos mesmos depende da aferição coerente dos seus parâmetros. Torna-se um trabalho interessante aplica-la em favor da descoberta de conhecimentos sobre dados estruturados e para outras tarefas futuras, como análise de modelos séries-temporais. Como sugestão, a sua funcionalidade para a etapa de pré-processamento de *Stemmer* poderia ser melhorada se os seus detentores dos direitos de exploração e comercialização do *software* programassem um operador exclusivo com suas funções já definidas, tal como ocorre com a plataforma *RapidMiner / YALE*. Dessa forma, ficaria mais fácil ao usuário ao invés de ter que inserir os termos para o processo via linha de comando, tal como ocorre na versão atual.

No tocante à questão de desenvolvimento de *software* especialista, ainda que existam muitas ferramentas no mercado capazes de manipularem o campo textual das patentes caso a preparação dos documentos, em uma etapa anterior ao pré-processamento seja bem feita antes da submissão dos mesmos aos processos exploratórios – conforme ocorrera em presente estudo, a presente pesquisa deixa como proposta a implementação de

uma ferramenta especialista em mineração sobre patentes em idioma inglês e uma possibilidade de mensurar o seu desempenho com outros *softwares* já existentes, conforme já ocorrera no estudo elaborado por (CAPUTO, 2006).

Por fim, outras bases de patentes existentes (já citadas no Capítulo 04) e hospedadas na *Internet* poderiam ser utilizadas para pesquisas por patentes relativas a determinados assuntos, tal como fora feito em presente dissertação.

REFERÊNCIAS BIBLIOGRÁFICAS

ARANHA, C., PASSOS, E. P. L., 2006, "A Tecnologia de Mineração de Textos". *RESI - Revista Eletrônica de Sistemas de Informação*, no. 2, pp. 2.

BAEZA-YATES, R., RIBEIRO NETO, B., 1999, *Modern Information Retrieval*. Harlow, Addison Wesley.

BERNERS – LEE, T., HENDLER, J., LASSILA, O., 2001, "The Semantic Web". *Scientific American*, May.

BUFREM. L.; PRATES, Y., 2005, "O saber científico registrado e as práticas de mensuração da informação". *Ciência da Informação*, Brasília, v. 34, n. 2 (Maio - Ago), pp. 9-25.

CALINSKI, R. B., HARABASZ, J., 1974, "A dendrite method for cluster analysis". In: *Communication in statistics*, vol. 3, pp. 1–27.

CALLON, M., COURTIAL, J. P., PENAN, H., 1995, *Cienciometria: el estudio cuantitativo de la actividad científica: de la bibliometria a la vigilancia tecnológica*. Ediciones Trea, Gijón.

CAPUTO, G.M., 2006, *Sistema Computacional para o processamento textual de patentes industriais*. Dissertação de M.Sc., COPPE / UFRJ, Rio de Janeiro, RJ, Brasil.

CHEN, Z., 2001, *Data Mining and Uncertain Reasoning - an Integrated Approach*. New York, John Willey and Sons, Inc.

COOLEY, R., 2000, *Web usage mining: Discovery and application of Interesting Patterns from Web data*. PhD thesis, Dept. of Computer Science, University of Minnesota, Minnesota, USA.

COOLEY, R., MOBASHER, B., SRIVASTAVA, J., 1997, “Web Mining: Information and Pattern Discovery on the World Wide Web”. In: *9th IEEE International Conference on Tool with Artificial Inteligence*, pp 558-567, Newport Beach, Nov.

COOLEY, R., MOBASHER, B., SRIVASTAVA, J., 1999, "Data preparation for mining world wide Web browsing patterns". *Journal of Knowledge and Information Systems*, vol.1, Feb., pp.5-32.

DECKER, S., MELNICK, S., HARMELEN, F. V., FENSEL, D., KLEIN, M., BROEKSTRA, J., ERDMANN, M., HORROCKS, I., 2000, “THE SEMANTIC WEB: The Roles of XML and RDF”. *IEEE Internet Computing*, vol. 13, no. 5 (Oct), pp. 63-74.

DOU, H., 1995, *Veille technologique et competitive*. 1 ed. Paris, Dunod.

ETZIONE, O., 1996, “The World Wide Web Quagmire or gold mine”. *Communications of the ACM*, vol.39, no.11 (Nov), pp. 65-68.

FACHINELLI, A. C., 2004, “Elementos Metodológicos de Vigília e de Inteligência Econômica para o processamento de informações organizacionais”., *Conexão - Comunicação e Cultura*, v. 2, n. 1(Jan), pp. 153-162.

FISHER, N. I., HALL, P., JING, B. Y., WOOD, A. T. A., 1996, “Improved pivotal methods for constructing confidence regions with directional data”, *Journal of the American Statistical Association*, pp. 1062–1070.

GARFIELD, E., 1995, “Quantitative analysis of the scientific literature and its implications for science policymaking in latin america and the caribbean”. *Bulletin of the Pan American Health Organization*. v. 29, no. 1 (Jan), pp. 87-95.

GIESBRECHT, H. O., 2000, *Inteligência tecnológica: estudo das práticas de dois institutos de pesquisa tecnológica no Brasil*. Dissertação de Mestrado em Ciência da

Informação - Programa de Pós-Graduação em Ciência da Informação, Universidade de Brasília, Brasília, DF, Brasil.

GIRARDI, M. R., 1995, *Classification and Retrieval of Software through their Descriptions in Natural Language*, Ph.D. dissertation, University of Geneva, Geneva, Switzerland.

GIRARDI, M. R., 1998, "Main Approaches to Software Classification and Retrieval". In: *Ingeniería del Software y reutilización: Aspectos Dinámicos y Generación Automática*. Editores J. L. Barros y A. Domínguez, Universidad de Vigo, Ourense, del 6 al 10 de julio, Julio.

GIRARDI, M. R., MARINHO, L. B., 2007, "A Domain Model of Web Recommender Systems based on Usage Mining and Collaborative Filtering". *Requirements Engineering (London)*, Aug. 2006 (On-line first), v. 12, pp. 23-40.

GOMES, E., BRAGA.F., 2004, *Inteligência Competitiva: como transformar informação em um negócio lucrativo*. 2 ed. Rio de Janeiro, Elsevier.

HAN, J., 2000, "OLAP Mining: An integration of OLAP with Data Mining", *School of Computing Science*, Simon Fraser University, British Columbia, Canada.

HAN, J., KAMBER, M., 2006, *Data Mining: Concepts and Techniques*. 2 ed. New York, Morgan Kaufmann Publishers.

HARMAN, D. K., 1991, "How effective is suffixing?" *Journal of the American Society for Information Science*, v. 47, no. 1, pp. 70-84.

IDC - Insight Discoverer Clusterer – Developer's Guide, Temis Company, 2002.

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., 1999, *Bibliometric analysis of patents: implications for R&D and industry, emerging trends in scientometrics*. In: *NAGPAUL P.S. et al*., editor. New Delhi, Allied Publishers.

KARKI, M., 1997. Patent Citation Analysis: A policy analysis tool, World Patent Information, vol. 19, n. 4, pp. 269-272.

KNOWLEDGE Management and Competitive Intelligence Made Clear. Disponível em: <http://www.cipher-sys.com/>. Acesso em: 12 jan. 2008.

KOTLER, P., 2002, *Administração de Marketing: a edição do novo milênio*. 1 ed. São Paulo, Prentice Hall.

LEMOS, C., 1999, "Inovação na Era do Conhecimento". In: LASTRES, H.M.M. & ALBAGLI, S (orgs), *Informação e globalização na era do conhecimento*. 1 ed., Cap. 5, Rio de Janeiro, Editora Campus.

LOVINS, J. B. *Development of a stemming algorithm. Mechanical Translation and Computational Linguistics*, v. 11, p. 22-31, 1968.

MICROPATENT. Disponível em <http://www.micropat.com> Acesso em: 12 jan. 2008.

MIERSWA, I., WRUST, M., KLINKENBERG, R., SCHOLZ, M., EULER, T., 2006, "YALE: Rapid Prototyping for Complex Data Mining Tasks", *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 935-939, Philadelphia, Pennsylvania, USA, Aug.

MIRANDA, R. C. R., 1999, "O uso da informação na formulação de ações estratégicas pelas empresas". *Ciência da Informação*. Brasília, v. 28, n. 3, p. 284-290. Set./Dez.

NARIN, F., 1994. "Patent bibliometrics". *Scientometic*s, v. 30, ed. 1 (May), pp. 147-155.

NARIN, F., NOMA, E., 1987. "Patents as indicators of corporate technological strength". *Journal Research Policy*, v. 16, n. 2-4 (Aug) pp. 143-155.

OLIVEIRA, L. G., SUSTER, R. P., ANGELO, C., Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-04220050007000007&lng=en&nrm=iso> Acesso em: 08 jan. 2008.

PAKHIRA, M. K., BANDYOPADHYAY, S., MAULIK, U., 2004, "Validity index for crisp and fuzzy clusters", *Pattern Recognition*, vol. 37, no. 3 (Mar), pp. 487–501.

PASSOS, E., GOLDSMITH, R., 2005, *Data Mining: um guia prático*. Rio de Janeiro, Elsevier.

PCT - Patent International Treaty. Disponível em <http://www.wipo.int/pct/es/treaty/about.htm> Acesso em: 10 jan. 2008.

PONJUÁN DANTE, G., 1998, *Gestión de información en las organizaciones: principios, conceptos y aplicaciones*. 1 ed. Santiago, CECAPI.

PORTER, M. E., 1990, *Competitve Advantage of Nations*. 1 ed. New York., Free Press.

PORTER, M. F. *An algorithm for suffix stripping. Program*, v. 14, p. 130-137, 1980.

SALTON, G., *Automatic Text Processing – The Transformation, Analysis and Retrieval of Information by Computer*. Addison – Wesley, 1989.

SILVA, L. F., CARVALHO, M. B., 2004, "Aspectos Gerais da Propriedade Intelectual nas Instituições de Ensino e Pesquisa", *Cadernos REPICT (Rede de Tecnologia do Rio de Janeiro)*, E-papers: Rio de Janeiro, vol.1, pp. 43.

STATSOFT - Statistica Software. Disponível em: <http://www.statsoft.com/>. Acesso em: 16 jan. 2008.

SYCARA, K., 1998, "Multiagent Systems", *AI Magazine*, Vol. 10, No. 2 (Feb), pp. 79-93.

TEMIS Text Intelligence. Disponível em <http://www.temis.com/> Acesso em: 12 jan. 2008.

TYSON, Kirk W. M. 1998, *The Complete Guide do Competitive Intelligence: gathering, analyzing, and using competitive intelligence*. Kirk Tyson Int. Ltd. Lisle, Chicago.

WEINER, P., 1973, "Linear pattern matching algorithm". *14th Annual IEEE Symposium on Switching and Automata Theory*, pp. 1-11, New York, USA, Jul.

WEINER, P., 1973, “Linear pattern matching algorithms”. In: *Proceedings of the 14th IEEE Symposium on Switching and Automata Theory*, pp. 1–11, Washington DC.

WISDOMAIN. Disponível em <http://www.wisdomain.com> Acesso em: 08 jan. 2008.

WORMELL, I., 1998, *Informetrics: exploring databases as analytical tools*. Database, v. 21, n. 5, pp. 25-30, Oct. / Nov.

WURMAN, R. S., 1995, *Ansiedade de informação: como transformar informação em compreensão*. 5.ed. São Paulo, Cultura Editores.

XIE, X., BENI, G., 1991, "A validity measure for fuzzy clustering." In: IEEE Transactions Pattern Analysis and Machine Intelligence, vol. 13, n. 8, pp. 841-847.

Livros Grátis

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

Milhares de Livros para Download:

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