

5º CONGRESSO BRASILEIRO DE PESQUISA E DESENVOLVIMENTO EM PETRÓLEO E GÁS



TÍTULO DO TRABALHO:

Sistema de Monitoramento e Controle da Qualidade de Combustíveis

AUTORES:

Delano Brandes Marques¹, Paulo José Melo Gomes Correa³, Glene H. R. Cavalcante²,
Allan Kardec D. Barros³, Sofiane Labidi¹

INSTITUIÇÃO:

Universidade Federal do Maranhão – UFMA, CCET: DETQI e DEE

¹Laboratório de Sistemas inteligentes – LSI

²Lab.de Análises e Pesquisa em Química Analítica de Petróleo e Biocombustíveis – LAPQAP

³Laboratório de Processamento da Informação Biológica – PIB

Este Trabalho foi preparado para apresentação no 5º Congresso Brasileiro de Pesquisa e Desenvolvimento em Petróleo e Gás- 5º PDPETRO, realizado pela Associação Brasileira de P&D em Petróleo e Gás-ABPG, no período de 15 a 22 de outubro de 2009, em Fortaleza-CE. Esse Trabalho foi selecionado pelo Comitê Científico do evento para apresentação, seguindo as informações contidas no documento submetido pelo(s) autor(es). O conteúdo do Trabalho, como apresentado, não foi revisado pela ABPG. Os organizadores não irão traduzir ou corrigir os textos recebidos. O material conforme, apresentado, não necessariamente reflete as opiniões da Associação Brasileira de P&D em Petróleo e Gás. O(s) autor(es) tem conhecimento e aprovação de que este Trabalho seja publicado nos Anais do 5ºPDPETRO.

Sistema de Monitoramento e Controle da Qualidade de Combustíveis

Abstract

To guarantee fuels' quality and normalization it is essential the development of efficient tools to allow their monitoring from any point and for any kind of fuel. Considering the variety of criteria, a decision making must be based on the evaluation of the most varied types of space and non-space data. This paper presents studies which aim the implementation of an integrated system that allows, besides a better monitoring, practicality and efficiency, enables the control and optimization of problems related to the petroleum industry. For this, it is used the process of Knowledge Discovery in Databases. The proposed system aims to cover several areas of fuel monitoring. From the survey and analysis of different informations used in the Brazilian National Agency of Petroleum, Natural Gas and Biofuels' (ANP) databases a data warehouse model was created. Following data mining techniques (Principal Components Analysis, Cluster Analysis and Regression) were applied aiming the discovery of knowledge, identification of trends, patterns which can assist decision making regarding the monitoring of fuel quality.

Key words: Fuel analysis, KDD process, Datawarehouse, Data Mining, Geographic information system.

Introdução

O grande aumento no volume de dados, em diversas áreas do conhecimento, tem levado ao crescente acúmulo de informações, em grande velocidade e quantidade, decorrente dos avanços tecnológicos [Li(2004), Küçüksille (2009)]. Estas grandes massas de dados são superiores à capacidade humana de processá-las, levando o homem a recorrer aos computadores (que a princípio são também causadores desta sobrecarga) por uma solução.

Muitos esforços têm sido feitos na tentativa de adaptar, modificar ou melhorar as técnicas convencionais usadas em descoberta do conhecimento com base na experiência da base de dados (KDD) para projetar modelos apropriados e novos de mineração [Compieta (2007)].

Neste cenário há cada vez mais uma necessidade de ferramentas computacionais que possam auxiliar o homem na extração de informações úteis e conhecimento deste crescente volume de dados digitais. Entretanto, estas informações não estão caracterizadas explicitamente. O processo capaz de descobrir este conhecimento em bancos de dados chama-se KDD (Knowledge Discovery in Databases), o qual foi proposto em 1989 para referir-se às etapas que produzem conhecimentos a partir dos dados e, principalmente, à etapa de mineração dos dados, que é a fase que transforma dados em informações (Fayyad et al., 1996).

Além da sobrecarga de dados em que vivemos, e que tende a crescer, outro fator que dificulta o acesso e análise desses dados é o fato de serem oriundos de diversas fontes heterogêneas. A tecnologia de *Data Warehouse* vem como uma solução para esses e outros problemas. De acordo com INMON (1997) um *data warehouse* pode ser definido como um conjunto de dados baseado em assuntos, integrados, não voláteis, e variáveis em relação ao tempo, para dar suporte ao processo gerencial de tomada de decisão. Esta tecnologia pode fazer parte do processo KDD no chamado “pré-processamento” que consiste na preparação dos dados para a etapa de mineração de dados. É neste contexto que se baseia a pesquisa proposta neste trabalho.

O Laboratório de Análises e Pesquisas em Química Analítica de Petróleo e Biocombustíveis – LAPQAP, da UFMA, foi implantado no ano 2000, com recursos do CTPETRO (Programa de Apoio à Ciência e Tecnologia voltado para o setor de Petróleo e Gás Natural). Este laboratório tem como principal atividade o monitoramento de combustíveis do Estado do Maranhão, através do Programa de Monitoramento da Qualidade de Combustíveis – PMQC, coordenado pela Superintendência de Biocombustíveis e de Qualidade de Produtos (SBQ) da Agência Nacional do Petróleo (ANP). O PMQC foi instituído em 1998, e é oficialmente regulamentado pela Resolução ANP nº 29, de 26 de outubro de 2006 (ANP, 2009) tendo alcançado todas as unidades da federação em 2005.

A UFMA, a exemplo de muitas universidades, através do LAPQAP desenvolve atividades de pesquisa e monitoramento da qualidade de combustíveis (LAPQAP, 2009) a cerca de 9 anos e, ao longo deste tempo, um grande volume de dados foi acumulado. As análises físico-químicas são realizadas com base nas normas e especificações vigentes, determinadas pela ANP. Os dados são divididos em 4 regiões geográficas do Estado do Maranhão e compreendem cerca de 870 postos de combustíveis. Mensalmente 380 amostras são analisadas, sendo 160 de gasolina, 160 de diesel e 80 de etanol combustível. Dentre os dados armazenados, existem resultados de diversos tipos de análises físico-químicas que são realizadas nos diferentes tipos de combustíveis (www.anp.gov.br).

O presente trabalho tem como objetivo principal colaborar com o programa de monitoramento de combustíveis aplicando o processo de Descoberta do Conhecimento em Bancos de Dados (KDD) na base de dados do LAPQAP visando o apoio à tomada de decisão.

Metodologia

As etapas do processo KDD aplicado são apresentadas na Figura 1 e incluem a seleção dos dados, limpeza de amostras, transformação dos dados, dedução de padrões, modelos das amostras através de algoritmos de *Data Mining* e a estimativa da qualidade do conhecimento extraído.

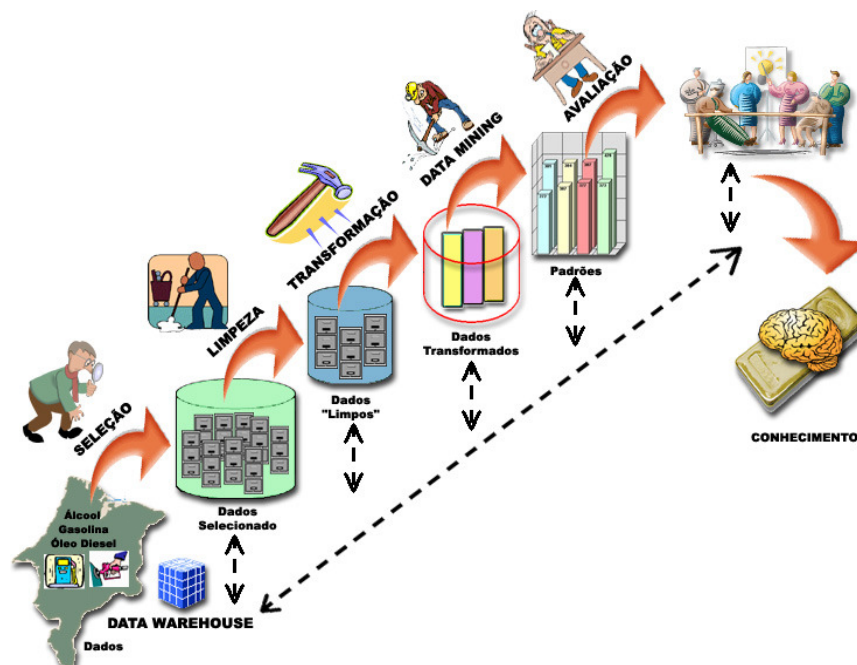


Figura 01 – Etapas do Processo de Descoberta de Conhecimento em Banco de Dados

No processo KDD, cada fase pode possuir uma interseção com as demais. Desse modo, os resultados produzidos em uma fase podem ser utilizados para melhorar os resultados das fases seguintes.

Esse cenário indica que o processo é interativo e iterativo (Berry, 2000), ou seja, depende da constante interferência do usuário (técnico especialista) e se repete de acordo com a necessidade.

Nas três primeiras etapas ocorreram vários passos para a construção de uma base de dados consistente. Conforme é sabido, estas etapas são responsáveis por consolidar as informações relevantes para o algoritmo minerador buscando reduzir a complexidade do problema. A execução destas fases não necessita de uma ordem, podendo ser executadas durante a construção de um Data Warehouse.

Abaixo a descrição das etapas do processo, aplicadas no presente estudo:

Seleção dos Dados: Nesta fase é feita a seleção dos dados apropriados para a análise. Seu principal objetivo é a escolha dos atributos relevantes do conjunto existente no banco de dados e que deverão ser extraídos para a mineração. A finalidade dessa seleção é otimizar o tempo de processamento da etapa de mineração, visto que ela apenas trabalha com um subconjunto de atributos, desse modo diminuindo o seu espaço de busca.

Limpeza dos Dados: A limpeza dos dados envolve uma verificação da consistência das informações, a correção de possíveis erros e o preenchimento ou a eliminação de valores nulos e redundantes. A execução dessa fase corrige a base de dados eliminando consultas desnecessárias que seriam executadas pela etapa de *data mining* e que afetariam o seu processamento.

Transformação dos Dados: Existem vários tipos de técnicas de *data mining*, sendo que cada uma necessita de uma entrada específica. A finalidade desta fase foi transformar os dados, de modo que eles se tornassem compatíveis com as entradas dessas diversas técnicas. Em resumo essa fase converteu os dados na forma mais adequada para a construção e interpretação do modelo.

Data Mining: é uma das fases mais importantes do processo KDD (Fayyad et al., 1996). Através desta fase foi feita a aplicação dos algoritmos de mineração de dados, que diante da tarefa especificada foi capaz de extrair eficientemente conhecimento implícito e útil dos dados. A tecnologia de mineração de dados possui a vantagem de extrair informação que não seria possível de ser obtida através de consultas tradicionais em bancos de dados.

No presente trabalho foram aplicadas as seguintes técnicas:

- **Análise de Componentes Principais** - técnica utilizada para redução do número de variáveis em uma base multidimensional de dados. Uma vez reduzida a dimensionalidade deste conjunto de variáveis fica mais fácil a interpretação de interdependências existentes entre elas. Os componentes principais são combinações lineares das variáveis originais cuja consideração de seus valores, permite um melhor entendimento do conjunto de dados. Ou seja, a análise de componentes principais é um mecanismo de simplificação de dados através da redução do número de variáveis.
- **Análise de Agrupamento** – seu principal objetivo é classificar os valores de uma matriz de dados sob estudo em grupos discretos. Ela consiste em agrupar tipos similares de dados ou identificar exceções de acordo com valores em comum. Quando esta técnica é aplicada, ainda não é conhecida nenhuma classe, sendo que o objetivo da função de análise de agrupamento é produzir uma segmentação do conjunto de registros de entrada de acordo com algum critério. Este critério é estabelecido por uma ferramenta de análise de agrupamento e as funções podem produzir descrições implícitas e explícitas.
- **Regressão Múltipla** - tem como principal objetivo a previsão. Na análise de regressão objetiva-se o desenvolvimento de um modelo estatístico que possa ser utilizado para prever os valores de uma variável dependente ou variável de resposta, com base nos valores de pelo menos uma variável independente ou explicativa. Tem como objetivo a previsão de uma variável de resposta em relação a um conjunto de variáveis.

Avaliação: Nesta etapa é avaliado se o conhecimento gerado é relevante e acionável. Se a resposta não for satisfatória, então é necessário repetir todo ou parte do processo KDD.

Quando um padrão é identificado na etapa de *data mining*, ele deve ser examinado para se determinar se ele é novo, relevante e “correto” por uma medida padrão. Quando o padrão é julgado relevante e útil, ele pode ser considerado conhecimento. O conhecimento então deve ser armazenado na base de conhecimentos para subseqüentes interações.

Esses conhecimentos serão consolidados em forma de relatórios demonstrativos com a documentação e explicação das informações relevantes ocorridas em cada etapa do processo KDD.

Resultados e Discussão

A seguir são apresentados os resultados da aplicação das técnicas de mineração de dados usadas no presente estudo.

Análise de Componentes Principais

A análise de dados utilizando a técnica de Análise de Componentes Principais foi bastante proveitosa em relação à redução do número de variáveis. Dentre os quatro tipos de combustíveis analisados (AEH – Álcool Etílico Hidratado Comum; GCA – Gasolina C Aditivada; GCC – Gasolina C Comum; e OBC – Óleo Diesel B Comum.), apenas em um (OBC) foram encontradas dificuldades. A dificuldade encontrada reside no fato de haverem poucas variáveis disponíveis, apenas três. Neste caso, apesar de uma componente principal representar bem as demais variáveis, optou-se por duas, uma vez que não é possível o gráfico das componentes somente com uma variável. De qualquer forma, a segunda componente tem uma representatividade razoável em relação às amostras.

Em geral, foi possível a redução de variáveis em componentes principais que representam bem o modelo. Analisando as componentes principais formadas, e as variáveis por elas representadas, tem-se uma visão mais clara das relações existentes entre as variáveis. No Contexto das amostras a existência de alguns outliers pôde ajudar na Análise de Agrupamentos, feita posteriormente, como também mostrou uma tendência centralizada das amostras.

Na tabela 01, podemos ver um resumo da aplicação desta técnica aos dados referentes às amostras coletadas e aos tipos de combustíveis analisados.

Tabela 01. Resumo da aplicação da Análise de Componentes Principais.

Combustível	Nº de Variáveis	Nº de Componentes Principais
AEH	4	2
GCC	13	5
GCA	12	4
OBC	3	2

Análise de Agrupamento

Foram realizadas análises de agrupamento tanto no universo das amostras como no das variáveis de cada tipo de combustível. Em ambas foram gerados dendrogramas. Onde foi possível a identificação de grupos, formados devido a similaridades encontradas, tanto nas amostras como nas variáveis. Este tipo de análise é muito subjetiva, e no caso deste trabalho, poderiam ser formados mais grupos, entretanto menos significativos. Por esta razão, foi-se bem rígido quanto ao grau de similaridade, entre as amostras e variáveis, visando a formação de agrupamentos melhores ajustados.

De modo geral a técnica foi bastante eficiente, possibilitando futuros estudos das amostras e variáveis estudadas, como também facilitando na análise de Hipóteses e Tomada de Decisão.

A tabela 02 mostra um resumo da aplicação desta técnica aos dados referentes às amostras coletadas e aos tipos de combustíveis analisados.

Tabela 02. Resumo da aplicação da Análise de Agrupamento.

Combustível	Nº de Amostras	Nº de Grupos Formados	Nº de Variáveis	Nº de Grupos Formados
AEH	154	6	4	2
GCC	413	13	13	5
GCA	102	13	12	5
OBC	387	8	3	2

Regressão Múltipla

A análise de regressão tem como principal objetivo prever uma variável de resposta em relação às demais variáveis de um determinado conjunto de dados através de um modelo matemático. O critério utilizado neste trabalho para escolha da variável resposta foi o nível de variação das variáveis ao longo das amostras. Foi escolhida a variável com maior variação por representar melhor o conjunto. A tabela 03 mostra um resumo da aplicação desta técnica aos dados. A variável resposta de cada tipo de combustível, juntamente com suas respectivas variáveis independentes.

Tabela 03. Resumo da aplicação da Regressão.

Combustível	Variável Resposta	Variáveis Independentes
AEH	7 (Condutividade Elétrica)	8 (Massa Específica); 9 (pH) ; 10 (Teor Alcoólico).
GCC	95 (Destilação-PFE)	13
GCA	68 (Destilação-PFE)	54 (AEAC); 58 (Destilação-10%); 62 (Destilação-50%); 66 (Destilação-90%); 69 (Resíduo); 70 (MON); 71 (IAD); 75 (Saturados-iv ou ir); 76 (Olefinas-iv); 77 (Aromáticos-iv); 78 (RON-iv).
OBC	28 (Destilação-85%)	17 (Índice de Cetano); 24 (Destilação-50%).

Conclusões

Com base nos resultados acima, espera-se que os resultados obtidos neste trabalho possam contribuir com o processo de análise de combustíveis, tendo em vista a grande quantidade de dados acumulados. Os resultados poderão auxiliar na interpretação de resultados, bem como na predição de comportamentos funcionais dos diversos parâmetros físico-químicos que compõem o programa de monitoramento da ANP.

Estes resultados motivam a continuidade da pesquisa, possibilitando a descoberta de novos padrões relevantes. Dentre as futuras atividades podemos destacar a integração entre as etapas do processo KDD: Pré-Processamento (Data Warehouse); Data Mining; e Avaliação, por meio de um sistema multiagentes, possibilitando maior automação no processo. E a implantação de SIG (Sistema de Informações Geográficas) possibilitando uma análise da distribuição espacial dos postos de combustíveis de acordo com a qualidade do combustível fornecido, dentre outros aspectos.

Agradecimentos

CNPq (projeto CIMCO), FAPEMA, ANP, FINEP

Referências Bibliográficas

ANP <http://www.anp.gov.br/petro/programa_monitoramento.asp> 2009. Acessado em: 13 de junho de 2009.

BERRY, M. and LINOFF, G., **Mastering Data Mining á Art and Science of Customer Relationship Management**, Ed. Wiley, 2000.

COMPIETA, P., MARTINO,S.D.; BERTOLOTTO, M.; FERRUCCI, F.; KECHADI, T. **Exploratory spatio-temporal data mining and visualization**. J. Vis. Lang. and Comput., 2007, 18, 255–279.

FAYYAD, U. (Ed.); PIATETSKI-SHAPIO (Ed.). **Advances in Knowledge Discovery in Databases and Data Mining**, Massachusetts: AAAI Press, The MIT Press, 1996. 611 p.

INMON, W.H & HACKATHORN, R. D. **Como Usar o Data Warehouse**. Rio de Janeiro: Editora Infobook, 1997, 277p.

KÜÇÜKSILLE, E.U; SELBAS, R.; SENCAN, A. **Data mining techniques for thermophysical properties of refrigerants**. Energy Conv. and Manag., 2009, 50, 399–412.

LAPQAP <www.quimica.ufma.br/lapqap> 2009. Acessado em: 14 de junho de 2009.

Li ST, SHUE LY. **Data mining to aid policy making in air pollution management**. Expert Syst Appl 2004, 27:331–40.