O Uso da Descoberta de Conhecimento em Banco de Dados Para Extração e Análise de Informação: Estudo de Caso

André R. J. C. Luiz[1], Daves M. S. Martins[1]

[1]Centro de Ensino Superior de Juiz de Fora (CESJF) - Juiz de Fora – MG - Brasil

andrerjcl@gmail.com, davesmartins@gmail.com

Abstract. *The success of an enterprise is directly connected to knowledge management, requiring a constant demand of decision making about the activities carried out within an organization. One of the tools and theoretical bases used to extract and analyze this information is the Knowledge Discovery Database (KDD), the focus of the study presented here. To a better understanding of the use of KDD, this paper is divided into two stages: the first part is a theoretical survey about the information and characteristics of KDD, data collection and Database analysis; and the second stage presents a case study where such theories are applied in practice to better assess the KDD performance and functionality.*

Resumo. O sucesso de uma empresa está diretamente ligado à gestão de conhecimentos, exigindo uma constante demanda de decisões a serem tomadas pela organização. Uma das ferramentas e bases teóricas utilizadas para extrair e analisar estas informações é a Descoberta de Conhecimento, ou *Knowledge Discovery Database (KDD)*, foco do estudo aqui apresentado. Para entender melhor o uso da Descoberta de Conhecimentos, o presente trabalho se dividiu em duas etapas, sendo a primeira um levantamento teórico acerca das informações e características da KDD e do levantamento e análise de Bancos de Dados, e o segundo um estudo de caso onde tais teorias são aplicadas na prática para melhor avaliação de seu desempenho e funcionalidade.

1. Introdução

Grande parte das operações e atividades que as empresas têm desenvolvido nos últimos anos, tanto no setor privado quanto no público, acabam por ser registrada computacionalmente, gerando um banco de dados amplo. Um dos grandes desafios está em transformar uma grande quantidade de dados em informações úteis e ágeis para a empresa.

No percurso de gerenciamento, recuperação e análise desses bancos de dados, tópico de grande relevância prática para as gestões estratégicas de conhecimento, surge à necessidade de se recorrer a ferramentas de auxílio que permitam otimizar esta busca por informações mais concretas e que sejam de real valia para a análise das atividades da organização.

Segundo Chiavenato (1997), o desenvolvimento das tecnologias dentro das empresas ocorre por meio deste acúmulo de conhecimento. Ademais, seu desenvolvimento se da no *know-how* da empresa, as tarefas da empresa, além de se refletir na manifestação física destes conhecimentos.

A importância das informações de uma empresa é grande, podendo estas serem consideradas um dos ativos mais importantes para os negócios e peça chave na competitividade da empresa, independentemente de seu porte, sendo necessário que sejam sempre informações concretas e que auxiliem na redução de erros durante a tomada de decisão.

Lara (2004) afirma que a tecnologia é a principal responsável pelas grandes mudanças que reduziram as distâncias e fez com que as inovações surgissem de forma rápida e constante, o que promoveu a possibilidade da apresentação de informações em tempo real.

Tal tecnologia se integra com diversas ferramentas utilizadas na obtenção e conhecimento de base de dados de empresas, necessários à gestão estratégica, estando entre elas a ferramenta "Descoberta de Conhecimento", ou *Knowledge Discovery Database (KDD)*, foco deste trabalho e cujos processos não se restringem somente à obtenção de dados, mas também na gestão dos conhecimentos adquiridos através da Base de Dados.

Goldschimdt e Passos (2005), sobre a "Descoberta de Conhecimentos", afirmam que a grande dificuldade por ela apresentada se encontra na percepção e interpretação apropriada dos diversos fatores que se pode observar ao longo dos processos, além da dificuldade de integrar as interpretações de forma que seja possível auxiliar a tomada de decisões em relação a cada contexto e caso apresentado, deixando a cargo do "analista humano" a responsabilidade de orientar e executar os processamentos deste conhecimento a ser repassado para a gestão estratégica.

Pensando no contexto da importância da gestão estratégica de informações e no grande auxílio prestado pela *KDD*, este trabalho se estrutura tomando como objetivo principal as aplicações em seu uso em Bancos de Dados como forma de obtenção e análise de informações que possam ser úteis para o desenvolvimento da organização.

Para tanto, o trabalho toma por base duas metodologias, fazendo, em um primeiro momento, um levantamento teórico acerca da descoberta de conhecimento, suas características e as teorias existentes na literatura especializada sobre este tema, além de uma análise sobre a importância real da obtenção, tratamento, disponibilidade, integridade e segurança informações para as empresas.

Em um segundo momento, a metodologia passa a trabalhar um estudo de caso, onde serão apresentadas as definições de data base, exemplos práticos de aplicação da "Descoberta de Conhecimento" e os resultados obtidos com esta aplicação, desenvolvendo uma base comprovada de informações acerca do tema abordado ao longo da presente pesquisa.

Assim, é possível afirmar que o presente trabalho se justifica pela relevância da temática aqui abordada em decorrência da constante evolução tecnológica e comunicacional que acaba gerando um grande número de dados em um curto espaço de tempo, fazendo com que seja necessário cada dia mais, que tais dados sejam analisados da forma mais eficiente possível para que nenhuma informação essencial à organização se perca, auxiliando as empresas a se manterem competitivas e eficientes.

O trabalho se encerra com as considerações finais acerca das teorias apresentadas, de forma que sirva de base e auxílio para novas pesquisas na área e para

preencher as lacunas acerca da aplicação da Descoberta de Conhecimento em Bancos de Dados.

2. Importância das Informações Para as Empresas.

Conforme as empresas se desenvolvem e ficam mais diversificadas acontece um crescimento da exigência da relevância da informação. A informação passa a ser não apenas proveitosa em grau funcional, mas também em um grau tático e estratégico, ou seja, são aplicáveis também nas estratégias desenvolvidas pela empresa e nas táticas de atuação desta. Nessa esfera, não somente a essência da informação é importante, porém a maneira como a informação é tramitada recebe relevância. A eficiência no trâmite da informação se sujeita a maior parte de como ela é gerenciada e da adequada compreensão de determinadas concepções e relacionamentos, sob a penalidade de disponibilizar ao cliente somente informações sem sentido, prejudicando o processo decisório.

Foram significativas as mudanças acontecidas no século passado, especialmente no que tange às questões tecnológicas: pode-se ressaltar o advento do computador e, por conseguinte, a proposta no mercado dessas mercadorias, sendo perceptível a relevância da informação e do conhecimento entre a população.

A informação virou um bem indispensável para o êxito de qualquer empresa com o gradual e frequente desenvolvimento de distintas tecnologias, como os empreendimentos e comercialização de eletrônicos, computação e telecomunicações e, claro, a internet, isto é, o número de informações acessíveis vem crescendo.

Robredo (2004, p. 43) diz que mesmo com a tecnologia fornecendo resoluções para acomodar elevada quantidade de arquivos, a sistematização da informação neles inserida é uma complicação. Para a bibliografia tem-se uma visão de necessidade de aprofundamento e aprimoramento dos processos de análise da informação bem como na representação da informação, no intuito de se obter maior êxito na recuperação.

Os dados conservados nas empresas são utilizados entre vários sistemas de informação, as resoluções e intervenções são realizadas oriundas destes dados que devem ser exatos, definidos e estarem acessíveis.

Com todo progresso tecnológico que se traçou nos últimos anos, as empresas também estão sendo obrigadas a preverem não apenas a obtenção e usos desses mecanismos informacionais, mas também em como preservá-los.

Entretanto, para que se tenha a constatação da informação enquanto um ativo intangível, faz-se preciso a implantação e administração de domínios nos processos de produção, ingresso e utilização da informação.

Hoje, a informação pode ser reputada como um dos mais importantes patrimônios de uma organização, o qual está sob frequente ameaça (DIAS, 2000). Compreende-se que uma empresa está ameaçada quando suas instabilidades estão sendo conhecidas. Para impedir esta ameaça, uma empresa precisa possuir recursos preparados para apontar suas partes suscetíveis. É preciso possuir métodos para direcionar as atitudes a serem tomadas utilizando mecanismos, métodos e recursos necessários.

Assim, é preciso o domínio, regularização e credibilidade da informação mediada pela administração dos processos e o tratamento dos dados para garantir as

colunas da Segurança da Informação (SI), que são disponibilidade, confiabilidade e integridade.

A administração da informação tem se centrado em métodos nos quesitos de lógica, repetição e padrão de informação através de recursos automáticos. Estes meios precisam assegurar uma estabilidade entre as facilidades postas para garantir a investigação e auditoria tendo em vista o ponto de privacidade e credibilidade.

Deve-se salientar ainda que a administração da informação consiste numa reunião de processos a qual objetiva que seja adquirida uma pesquisa de exigências de informação bem como de fluxograma de dados em vários segmentos da empresa. Referidos segmentos abrangem ações de planejamento, estruturação, condução, disposição e administração em qualquer meio.

Segundo Beal (2008), a administração de informação refere-se ao fluxo de informações seguido da estruturação, onde a determinação de exigências e condições de informação é como um componente fomentador deste fluxo, provocando um ciclo ininterrupto de recolha, tratamento, fornecimento/arquivamento, e utilização com o término da resolução do público interno, com a propagação da informação para o usuário externo à organização. A Figura 1 representa o fluxo de informação sugerido pela bibliografia:

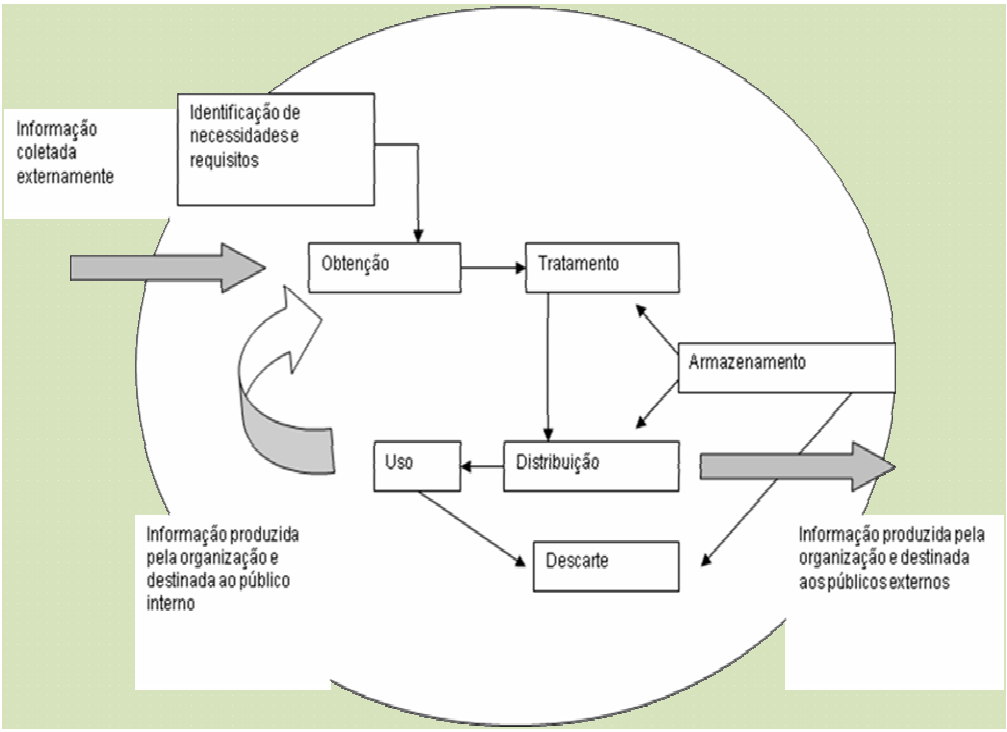


Figura 1: Modelo de representação do fluxo da informação – Beal (2008)

Assim, pode-se dizer pautado na análise da Figura 1 de autoria de Beal (2008), que as fases do ciclo de informação são respectivamente:

1. Identificação das necessidades de informação – observar as informações acessíveis no grupo interior ou exterior, para que se possam propagar os bens e serviços de informação e possuir a perspectiva da exigência de informação.

2. Obtenção: nesta fase, são elaboradas as ações de formação, recebimento ou captação da informação, oriunda de qualquer lugar ou origem, porém que seja de confiança.

3. Tratamento: Para que a informação possa ser utilizada, é necessário que ela passe por um processo de ordenação, adequação, elaboração, categorização, observação, fusão, assimilação e propagação.

4. Uso: Esta é a fase fundamenta da administração de informação porque possuir a informação precisa e acessível não quer dizer que está sendo utilizada, isto é, não é a presença da informação que assegura as melhores consequências, mas sim a utilização que é realizada da mesma.

5. Armazenamento: Existe a necessidade de se preservar as informações proveitosas à organização, autorizando sua utilização ou reutilização. Deve-se ficar atento para qual meio a informação está sendo conservada, fazendo sempre um *backup* e, em situações de informações confidenciais, é preciso elaborar ferramentas de segurança impossibilitando o ingresso de usuários que não estejam permitidos.

6. Descarte: Quando um dado não tem mais importância ou fica ultrapassado para a empresa, ele deve ser eliminado. Isso faz com que se poupem ferramentas de conservação, ampliando a agilidade e eficácia no local de dados precisos.

Declara-se, assim, que a administração de informação no âmbito das empresas precisa operar junto com as políticas da segurança da informação para garantir os ativos informacionais corporativos e ajustar sistemas de informação regulados nas colunas confidencialidade, disponibilidade e integridade (Robredo, 2004).

Ademais, a lei da confidencialidade da informação possui o intuito de assegurar somente que o usuário apropriado possua ingresso à informação.

Cumpre salientar ainda que os dados compartilhados entre as pessoas e organizações nem sempre poderão ser sabidos por todo mundo. Diversas informações fornecidas pelos indivíduos se determinam a um conjunto exclusivo de pessoas e, diversas vezes, a um só indivíduo. Isso quer dizer que essas informações deverão ser sabidas somente por um conjunto moderado de indivíduos, estabelecido pelo agente da informação.

Destaca-se ainda a necessidade de possuir confidencialidade na comunicação é possuir a certeza de que o que foi falado a uma pessoa ou redigido em determinado local será ouvido ou visto por quem tiver permissão para isto. A ausência dessa confidencialidade quer dizer ausência de sigilo. Se um dado for sigiloso, ele será confidencial e deverá ser armazenado com confiança, e não fornecido a indivíduos não autorizados (Dias, 2000).

Complementa-se dizendo que defender a confidencialidade consiste em uma das razões decisivas para a segurança e uma das atividades mais complexas de implantar porque engloba todos os componentes que estão inseridos na comunicação da informação. Parte do emissor, passa pelo percurso e alcança o recebedor. Ademais, informações possuem níveis distintos de confidencialidade, geralmente relacionados a números. Quanto maior for o nível de confidencialidade, maior será o grau de segurança exigente na condição tecnológica e humana que está inserida nesse processo: utilização, tramitação e conservação dos dados (Beal, 2008).

Prosseguindo com o raciocínio construído até aqui com base na leitura de análise do conhecimento de diversos autores que discorrem sobre o tema em comento, pode-se afirmar que o segundo pilar da segurança da informação é a integridade que

possibilita assegurar que a informação não seja modificada de maneira não permitida e, assim, é integral. Em resumo, uma informação completa é uma informação que não foi modificada de maneira inadequada ou não permitida (Beal, 2008).

Para que a informação possa ser usada, ela deve estar intacta. Quando acontece uma modificação não permitida da informação em um documento, isso significa que este não possui mais integridade. A integridade da informação é essencial para o sucesso da comunicação.

Retomando Robredo (2004), o recebedor precisará possuir a segurança de que a informação adquirida, vista ou escutada é rigorosamente a mesma que foi posta ao seu dispor pelo emissor para certo fim. Estar inalterada significa estar em sua condição inicial, sem ter passado por qualquer modificação por uma pessoa que não possua permissão. Se uma informação passa por modificações na sua versão primária, então ela não tem mais integridade, o que pode acarretar falhas e golpes, lesionando a comunicação e o processo de resoluções.

Procurar a integridade é tentar garantir que somente as pessoas ou sistemas permitidos possam realizar modificações no modelo e no empreendimento e na natureza de uma informação, ou que modificações ocasionadas por imprevistos ou falhas de tecnologia não ocorram, assim como no âmbito onde ela é conservada e pela qual tramita em todos os ativos.

Assim, para privilegiar a integridade, é necessário que todos os componentes que fazem parte da base da administração da informação permaneçam em seus padrões iniciais estabelecidos por seus agentes e donos.

Além de se empenhar para que a informação alcance somente as pessoas apropriadas e de maneira completa, precisa-se fazer com esteja acessível na hora propícia. É disso que fala o terceiro pilar da segurança da informação: a disponibilidade.

Diz respeito à disponibilidade da informação e de toda a condição física e tecnológica que possibilita o ingresso, o trâmite e a preservação.

A disponibilidade da informação possibilita segundo Beal (2008) que:

= seja usada quando preciso;

= abranja seus usuários;

= possa ser acessada na hora que se precisar;

Para preservar a disponibilidade da informação, é preciso saber quem são as pessoas que a utilizam e elaborar normas para a sua utilização. A informação dever ser especificada segundo o seu valor para a empresa.

Visto que a disponibilidade é o mais relevante dos princípios da segurança da informação, ela deve atender ao usuário final.

Cada informação do interior da organização tem a indispensabilidade de estar acessível para um indivíduo ou grupo, ao mesmo tempo onde existe a exigência de um domínio que assegure que ela estará, de fato, acessível para os usuários legais.

Assim quando se assegura que a informação está acessível apenas para as pessoas apropriadas, é preciso assegurar que eles façam a utilização correta desses

dados. Para atingir este intuito é preciso que um processo eficaz de especificação da informação seja determinado.

A especificação deve contar com estes argumentos e especialmente o choque nos empreendimentos pela suspensão da disponibilidade, confidencialidade e integridade das mesmas (Dias, 2000).

3. Descoberta de Conhecimentos

A descoberta de conhecimento em base de dados tem como principal característica a possibilidade que apresenta de permitir a extração de conhecimento e informações por meio da ferramenta da KDD, *Knowledge Discovery Database*, o Banco de Dados da Descoberta de conhecimento, que auxilia na busca por informações da gestão de conhecimento estratégico da empresa.

Técnica utilizada desde a década de 1960, e que se aperfeiçoou ao longo dos anos, os processos envolvidos na Descoberta de Conhecimento em bases de dados são, segundo Goldschmidt e Passos (2005), compostos em três etapas: o pré-processamento, compreendendo a captação, organização e o tratamento de dados objetivando preparar os dados para as etapas seguintes, a etapa de Mineração de Dados, que compreende a busca dos conhecimentos uteis do contexto em que o KDD vai ser aplicado, a etapa de pós-processamento, em compreender o tratamento do conhecimento adquirido nas etapas anteriores. Ainda que o KDD se estruture nestas duas etapas, a etapa de pós-processamento nem sempre é necessária.

Assim, como expresso por Goldschmidt e Passos (2005), a grande dificuldade apresentada pela Descoberta de Conhecimento reside na correta interpretação dos dados obtidos durante os processos da organização e na tomada de decisão em cada um dos contextos apresentados.

Boente (2006), ressalta que a descoberta de conhecimento em bases de dados é multidisciplinar, tendo origem em diversas áreas distintas como a estatística, inteligência computacional, reconhecimento de padrões e banco de dados.

O KDD tem suas atividades organizadas em três grandes grupos:

- Desenvolvimento Tecnológico: concepção, aprimoramento e desenvolvimento de algoritmos, ferramentas e tecnologias para a busca de conhecimentos de bases de dados.
- Execução de KDD: atividades de busca de conhecimento em bases de dados.
- Aplicação de Resultados: aplicação incorporada dos resultados no contexto em que o KDD foi realizado.

Com relação às etapas anteriormente mencionadas acerca da Descoberta de Conhecimentos em Bancos de Dados, é necessário ressaltar que cada uma delas possui um conjunto de funções que as caracteriza.

As funções do pré-processamento que, como já mencionado, se ocupa da preparação dos dados que serão usados nos algoritmos da etapa seguinte, a mineração de dados, sendo responsável por captar, organizar e tratar os dados. As principais funções dessa etapa são:

1 – Seleção de Dados ou Redução de Dados: esta função identifica nas bases existentes as informações que são consideradas de relevância efetiva para serem usadas ao longo do processo de KDD.

2 – Limpeza de Dados: é a função que cuida do tratamento dos dados selecionados, assegurando a qualidade e concretude dos fatos.

3 – Codificação de Dados: esta função codifica os dados para um formato que possa ser utilizado como entrada dos algoritmos utilizados nas etapas seguintes.

4 – Enriquecimento de Dados: função que trabalha para obter o máximo de informações que possam ser agregadas aos dados e registros já existentes, deixando-os mais completos para que o processo de Descoberta de Conhecimentos em Bancos de Dados seja mais efetivo.

A etapa de mineração de dados, que pode ser considerada como a principal no processo de KDD, é onde são definidas as técnicas e algoritmos que serão utilizados, fazendo uso, para tanto, de Redes Neurais, Algoritmos Genéricos, modelos Estatísticos e probabilísticos, como afirmam Goldschmidt e Passos (2005). A escolha da técnica utilizada nesta etapa depende da tarefa de KDD que deverá ser executada, e toma por base as classificações abaixo apresentadas:

1 – Descoberta de Associação: trata da ampliação das buscas por itens que tendam a ocorrer simultaneamente em transações de bases de dados.

2 – Classificação: trata de descobrir formas de mapeamento do conjunto de registros em classes categoricamente pré-definidas.

3 – Regressão: trata de mapear os registros buscando por valores reais, se aproximando da função da Classificação.

4 – Clusterização: trata de separar registros da base de dados em subconjuntos de cluster com base em propriedades em comum.

5 – Sumarização: trata de identificar características iguais entre os elementos de um mesmo conjunto de dados.

6 – Detecção de Desvios: trata de localizar registros que não pertençam a um padrão normal para o contexto atual entre os elementos de uma base de dados.

7 – Descoberta de Sequencias: trata de identificar possíveis alterações sazonais de quantidades em remessa de uma peça específica.

A última etapa a ser analisada, o pós-processamento, muitas vezes considerada sem necessidade por objetivar a facilitação de interpretação e avaliação não utilizada em casos de dados mais simples que podem ser interpretados pelo homem, cuida do tratamento do conhecimento conquistado. Ela trabalha de forma a elaborar e organizar os dados, o que inclui, em alguns casos, a simplificação de gráficos e diagramas, entre outros tipos de relatórios.

No que diz respeito ao macro-objetivo desejado, sua classificação de aplicação de KDD pode ser:

1 – Predição: construção de um modelo de conhecimento que possibilite prever valores de alguns atributos em situações diferentes tomando por base um histórico existente.

2 – Descrição: construção de um modelo que descreva as informações já encontradas em um conjunto de dados de forma compreensível.

A Descoberta de Conhecimentos, portanto, utilizando as técnicas de tratamento e processamento de dados, serve como um auxílio às organizações no momento das tomadas de decisão por promoverem a disponibilização dos dados e análises dos processos que compões as atividades das empresas de forma que seja possível avaliar os problemas anteriores e buscar novas soluções para eles, além de auxiliar no desenvolvimento de previsões para projetos futuros, sendo uma base de grande importância para o sucesso de um empreendimento.

4. Experimentos e Resultados

Neste tópico será abordado o estudo de caso no qual demonstra na prática como uma empresa usa seus dados para torná-los em informações relevantes para a tomada de decisão e alcance de melhores resultados.

Para isto, foi utilizada a base de dados dos últimos cinco anos de vendas, são mais de dois milhões de registros de uma empresa do ramo de distribuição de picolés e sorvetes da marca Kibon. A companhia possui mais de quatro mil clientes atuando em parte dos estados de Minas Gerais e Rio de Janeiro, tendo como alguns concorrentes distribuidores das marcas Nestle, Garoto, Sol e Neve e outras várias marcas que atuam na região.

A empresa tem várias necessidades de conhecimento dos dados como a quantidade de estoque necessário para atender as vendas futuras, crescimento da empresa comparado aos anos anteriores, qual o volume vendido em determinados períodos, calcular a meta mensal dos vendedores baseado nos períodos anteriores, determinar o potencial do cliente de acordo com o volume, descobrir quais os potenciais produtos que serão vendidos de acordo com o potencial e localização do cliente, descobrir potenciais locais de crescimento da base de cliente entre outros problemas onde o uso da descoberta de conhecimento irá ajudar a solucionar.

Serão abordados os problemas relacionados ao estoque onde serão usados os dados históricos para gerar uma previsão das vendas e calcular o estoque necessário para que não compre produtos demais e que não falte, outro problema é a visualização do crescimento da empresa comparado aos mesmos períodos dos anos anteriores e com a possibilidade de visualização por grupo de produtos, seguimento de cliente e mês, para que se possam fazer ações onde houve queda nas vendas daquele determinado grupo de produto e seguimento de cliente.

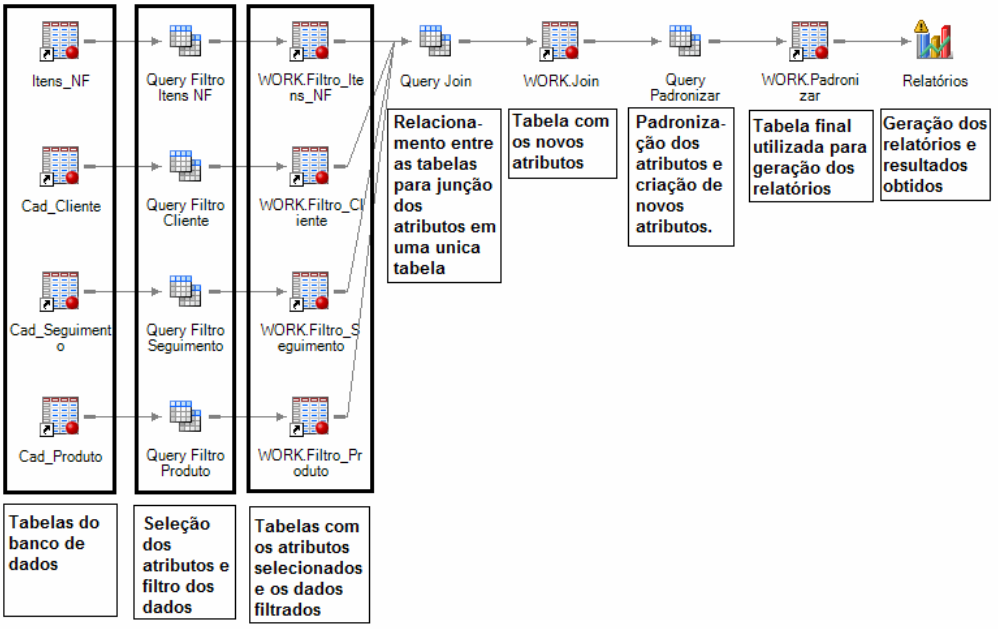


Figura 2. Arvore de processos

4.1. Definição da Base

Como dito anteriormente, é necessário definir os dados a partir dos objetivos da organização, sendo assim, foram selecionadas as tabelas de itens das notas fiscais onde estão às informações referentes à quantidade de itens vendidos, tabela cadastro de cliente para verificar as vendas por seguimento de cliente, tabela de cadastro do produto para converter a quantidade dos produtos vendidos em medidas relevantes para a empresa como litro e litro equivalente, e a tabela seguimento para exibição da descrição do seguimento ao invés do código, segue abaixo a seleção dos atributos.

A Tabela 1 relaciona as tabelas de entrada que serão utilizadas com seus respectivos atributos como base para este estudo de caso. Alguns atributos de identificação serão selecionados para a junção das tabelas na fase de transformação dos dados.

Tabela 1. Seleção dos Atributos das Tabelas do Banco de Dados

Nome da Tabela de Entrada	Atributos
Itens_NF	D2_COD, D2_QUANT, D2_CLIENTE, D2_LOJA, D2_EMISSAO, D2_QTDEDEV, D2_GRUPO
Cad_Cliente	A1_COD, A1_LOJA, A1_NSEGUIM
Cad_Produto	B1_COD, B1_CONV, B1_LTEQ
Cad_Seguimento	ZZH_COD, ZZH_DESC

4.2 Aplicação

A partir da definição de dados descrita na Tabela 1, foram criados filtros de limpeza de dados, onde se selecionou apenas aqueles específicos para geração da informação precisa e descartados os que são irrelevantes, em seguida padronizou-se os dados. Para

os relatórios finais, os dados foram agrupados em conjuntos com o intuito de aprimorar seus resultados.

A escolha da ferramenta foi feita, pois a empresa possui a licença de uso. A preparação dos dados foi feita através do uso da ferramenta *SAS Enterprise Guide,* que é uma das ferramentas da empresa SAS (*Statistical Analysis System*) orientada a projeto, ela possui uma interface gráfica com a facilidade de criar os projetos apontando e clicando, é permitido carregar arquivos em diferentes formatos como banco de dados e arquivos textos e relacioná-los gerando uma tabela de saída onde é possível criar análises dos dados com alta flexibilidade. Com a ferramenta, efetuou-se a leitura das tabelas de entrada, criação de query para selecionar os atributos e filtros para eliminar os dados irrelevantes.

No intuito de atingir os objetivos desejados bem como a criação de novos atributos, foram relacionadas às tabelas de entrada conforme demonstrado abaixo na Figura 3, essa relação foi necessária para a criação de novos atributos gerados a partir de duas tabelas ou para transformação dos atributos gerando assim a tabela de saída.

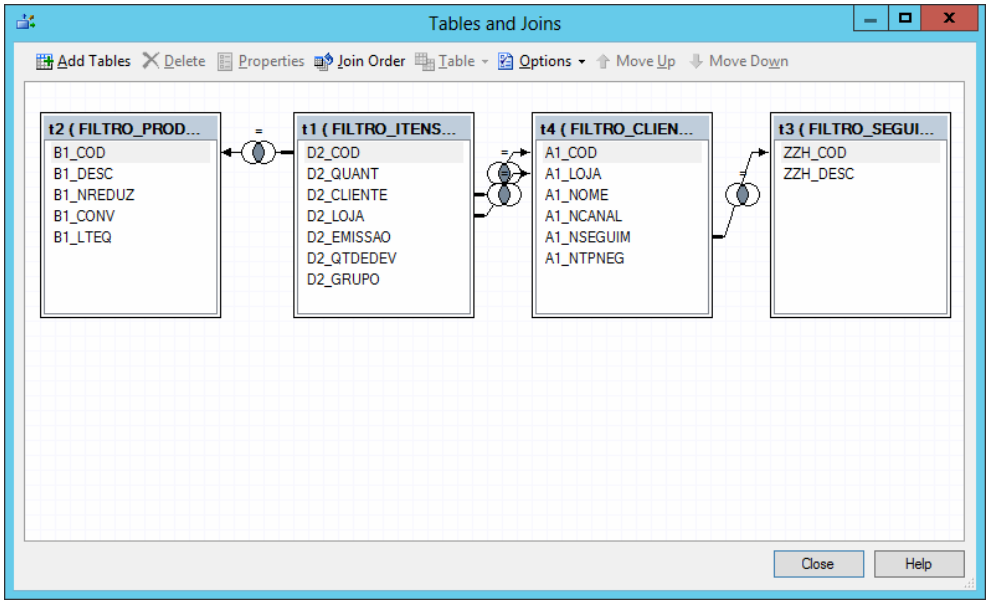


Figura 3. Relacionamento de tabelas de entrada

Após o relacionamento entre as tabelas representado pela Figura 3, novos atributos foram padronizados e criados onde serão usados para a descoberta de conhecimento, estes serão detalhados a seguir e exemplificado na Figura 4, lembrando que os atributos de entrada já foram listados na Tabela 1.

Definição das transformações dos atributos:

– Ano Emissao: Os quatro dígitos referente ao ano de emissão da nota fiscal para agrupar os valores nos relatórios de comparação.

– Qtd Vendida: Quantidade de produtos vendidos subtraindo as devoluções (D2_QUANT – D2_QTDEDEV).

– Litro: Quantidade vendida convertida em litros ((B1_CONV * (D2_QUANT - D2_QTDEDEV)))

– Litro Eq: Quantidade vendida convertida em litros equivalentes ((B1_LTEQ * (D2_QUANT - D2_QTDEDEV)))

– CodCli: O código do cliente é composto por dois atributos, esses atributos foram concatenados em um atributo (t1.A1_COD || "/" || A1_LOJA)

– Seguimento: No relacionamento das tabelas de entrada Seguimento e Cliente, foi utilizado o atributo ZZH_DESC onde contem a descrição do seguimento de mercado do cliente.

– Data de Emissao: A data de emissão da nota fiscal é gravada no banco de dados em formato varchar(8), foi convertida para formato data.

– Grupo: O grupo é uma categoria para os produtos, foi feita uma relação onde caso o conteúdo seja “0001”, passa a ser “Impulso”, caso seja “0002”, passa a ser “Take Home” e caso seja 0003 passa a ser “Scooping”.

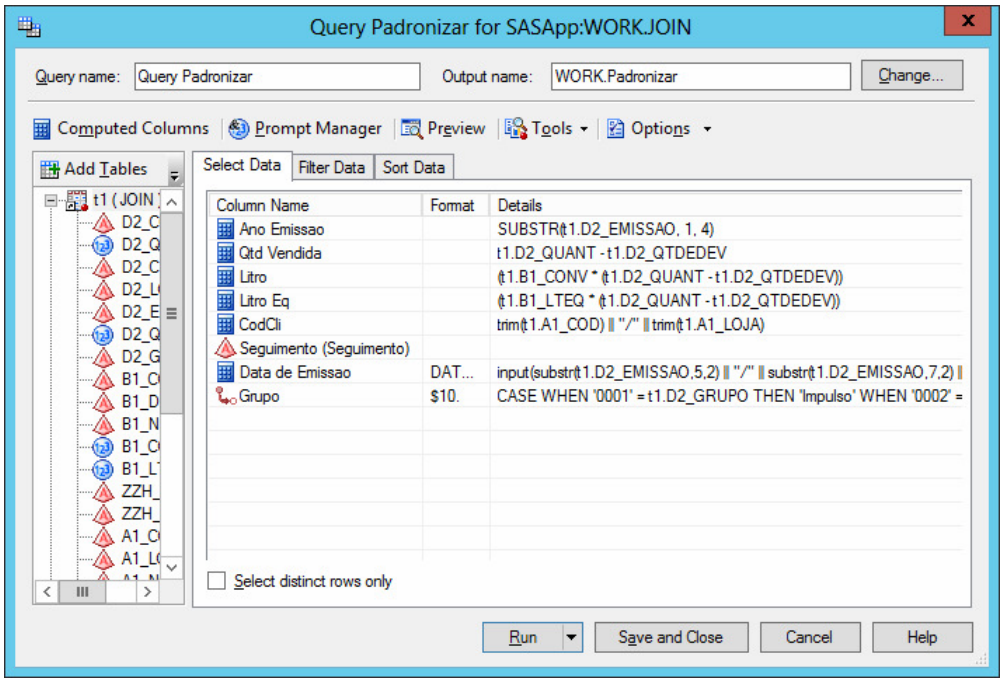


Figura 4. Transformação de dados

Após a transformação dos atributos mostrados acima, foi gerado a tabela de saída Hist_Vendas com os atributos selecionados e transformados, como mostra a Figura 5 onde possui uma amostra dos dados gerados.

	Ano Emissao	Qtd Vendida	Litro	Litro Eq	CodCli	Seguimento	Data de Emissao	Grupo
1	2010	4	3.6	1.22076	1488/04	REGIONAL	12JAN2010	Take Home
2	2010	4	3.6	1.23368	1488/04	REGIONAL	12JAN2010	Take Home
3	2010	4	3.6	1.23368	1488/04	REGIONAL	12JAN2010	Take Home
4	2010	4	3.6	1.23368	1488/04	REGIONAL	12JAN2010	Take Home
5	2010	4	3.6	1.68432	1488/04	REGIONAL	12JAN2010	Take Home
6	2010	12	24	4.61328	1488/04	REGIONAL	12JAN2010	Take Home
7	2010	20	40	7.6888	1488/04	REGIONAL	12JAN2010	Take Home
8	2010	48	96	16.719...	1488/04	REGIONAL	12JAN2010	Take Home
9	2010	80	160	27.8664	1488/04	REGIONAL	12JAN2010	Take Home
10	2010	8	16	2.97832	1488/04	REGIONAL	12JAN2010	Take Home

Figura 5. Tabela de saída Hist_Vendas

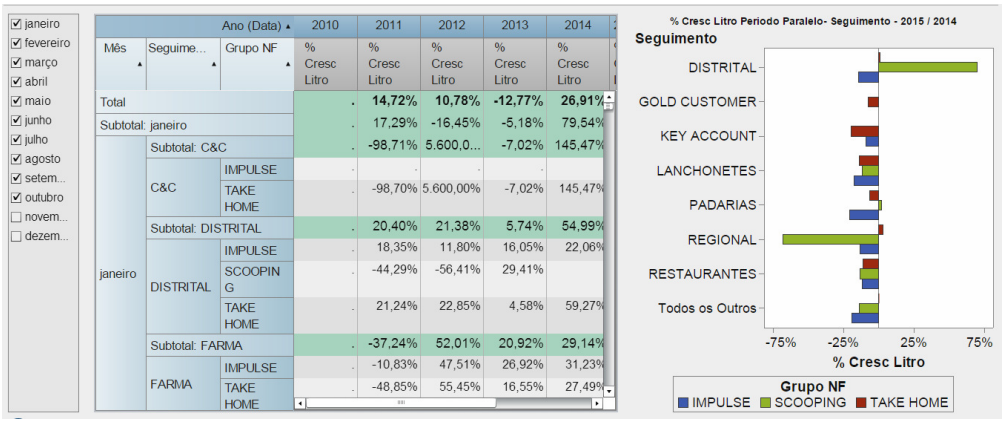
A partir da tabela de saída Hist_Vendas representado pela figura 5 gerada nas etapas anteriores, será usada como fonte dos dados para a geração dos relatórios de análise dos dados.

4.3. Resultados Obtidos

Neste tópico serão abordados os resultados obtidos a partir da análise realizada nos itens anteriores, estes baseados na necessidade organizacional em estudo, sendo assim, foram criados dois relatórios baseados na tabela Hist_Vendas Figura 5. O primeiro relatório será de sumarização dos dados, onde serão demonstradas as vendas comparando-as com o crescimento de um ano para o outro; o segundo relatório será uma previsão das vendas para os próximos seis meses, ambos os relatórios tomarão como base dados dos últimos cinco anos.

Os resultados obtidos servem para que a organização possa identificar seus pontos fracos e fortes, e então planejar um plano estratégico para sua melhoria, podendo destacar seus pontos fortes e/ou melhorar onde está com baixo desempenho, analisar se deve ou não continuar com investimentos ou se ainda vão redirecionar aplicações, ou seja, os resultados são de fundamental importância para uma tomada de decisão organizacional.

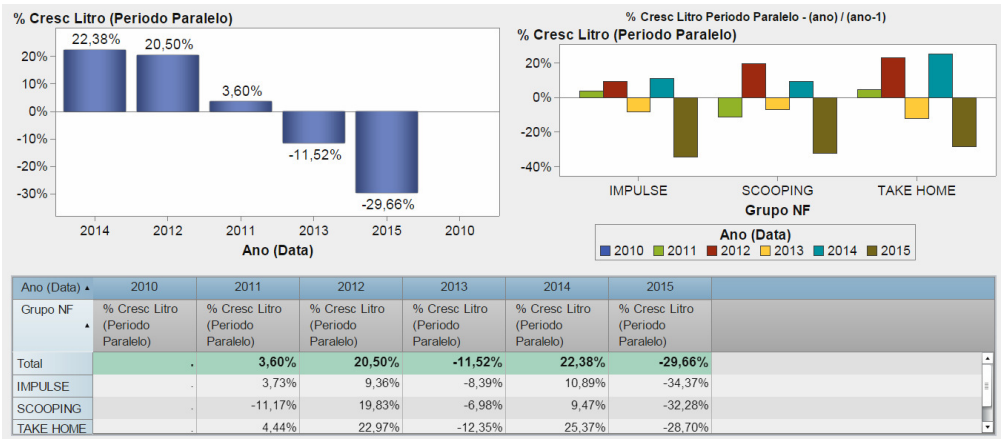
O relatório a seguir, é o resultado sumarizado e comparativo dos últimos cinco anos, como se pode observar na Figura 6, a tabela mostra o comparativo de crescimento referente ao ano anterior, subdividindo por mês, seguimento e grupo, no gráfico de barras é exibido o comparativo do último ano agrupado pelo atributo grupo. Nesse relatório também é possível filtrar os meses que queremos analisar de forma rápida, ao lado esquerdo, estão selecionados os meses que queremos analisar, podendo ser alterado com facilidade e agilidade os resultados comparativos da tabela e gráfico de barras.



Mês	Seguime...	Grupo NF	Ano (Data)	2010	2011	2012	2013	2014
				% Cresc Litro	% Cresc Litro	% Cresc Litro	% Cresc Litro	% Cresc Litro
Total				.	14,72%	10,78%	-12,77%	26,91%
Subtotal: janeiro				.	17,29%	-16,45%	-5,18%	79,54%
janeiro	Subtotal: C&C			.	-98,71%	5.600,0...	-7,02%	145,47%
	C&C	IMPULSE		.	.	.	.	
		TAKE HOME		.	-98,70%	5.600,00%	-7,02%	145,47%
	Subtotal: DISTRITAL			.	20,40%	21,38%	5,74%	54,99%
	DISTRITAL	IMPULSE		.	18,35%	11,80%	16,05%	22,06%
		SCOOPING		.	-44,29%	-56,41%	29,41%	
		TAKE HOME		.	21,24%	22,85%	4,58%	59,27%
	Subtotal: FARMA			.	-37,24%	52,01%	20,92%	29,14%
	FARMA	IMPULSE		.	-10,83%	47,51%	26,92%	31,23%
		TAKE HOME		.	-48,85%	55,45%	16,55%	27,49%

Figura 6. Relatório sumarizado e comparativo I.

Com o mesmo princípio de análise da Figura 6, abaixo segue o relatório com o comparativo da evolução das vendas comparadas ao ano anterior, ao lado esquerdo da Figura 7, o gráfico de barras demonstra de maneira decrescente, qual ano teve o maior crescimento das vendas em relação ao ano anterior, no gráfico de barras da direita, está o crescimento com o mesmo comparativo porém agrupado pelo atributo grupo onde é possível identificar por exemplo que ouve uma queda apenas no grupo *scooping* no ano de 2011 e um crescimento nos outros dois grupos, podendo assim focar as vendas para esse grupo de produtos. A tabela na parte inferior da Figura 7, mostra em percentual, o crescimento anual por grupo.



Ano (Data)	2010	2011	2012	2013	2014	2015
Grupo NF	% Cresc Litro (Periodo Paralelo)	% Cresc Litro (Periodo Paralelo)	% Cresc Litro (Periodo Paralelo)	% Cresc Litro (Periodo Paralelo)	% Cresc Litro (Periodo Paralelo)	% Cresc Litro (Periodo Paralelo)
Total	.	3,60%	20,50%	-11,52%	22,38%	-29,66%
IMPULSE	.	3,73%	9,36%	-8,39%	10,89%	-34,37%
SCOOPING	.	-11,17%	19,83%	-6,98%	9,47%	-32,28%
TAKE HOME	.	4,44%	22,97%	-12,35%	25,37%	-28,70%

Figura 7. Relatório sumarizado e comparativo II.

O relatório a seguir, tem o objetivo de demonstrar uma previsão das vendas para os seis meses futuros, com uma margem de confiança de acerto de 95% considerando a margem de variação demonstrada no gráfico em uma cor sombreada da mesma cor da linha do gráfico. Com essa previsão, é possível projetar a compra de produtos para suprir as vendas, redimencionar o estoque da empresa, verificar se com essa previsão será possível atingir as metas podendo até mesmo alterá-las de acordo com a necessidade.

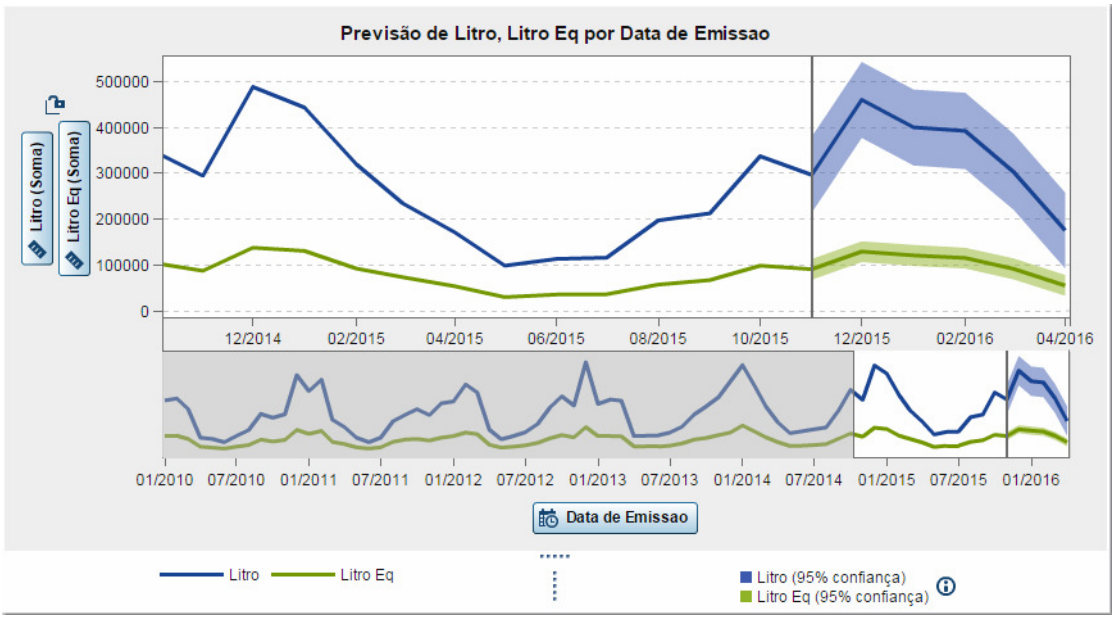


Figura 8. Relatório previsão de vendas

Os dados desses relatórios são atualizados diariamente e os filtros são dinâmicos e podem ser alterados com facilidade para uma melhor análise dos dados.

Com o uso das ferramentas de descoberta de conhecimento, o que eram apenas dados armazenados em banco de dados, passou a ser informações valiosas e descoberta de novos conhecimentos até então desconhecidos pela organização, com o comparativo de mesmo período anterior pode-se ainda ir agrupando por diferentes perspectivas para definir onde houve a maior queda de venda dos produtos, se foi em um seguimento de cliente ou até mesmo uma determinada linha de produtos. A previsão das vendas é uma informação ainda mais valiosa para que não haja desperdício de sobrecarregar os estoques e não ter saída dos produtos, tendo assim um gasto maior de armazenamento dos produtos e produtos parados em estoque, onde representa investimento parado para a companhia.

5. Considerações Finais

Ao longo do trabalho foi descrito a grande importância e o constante aumento da valorização dada às informações internas e bases de dados por parte de empresas, fazendo com que os investimentos para a obtenção e tratamento de dados crescessem, sendo considerada, inclusive, uma ferramenta estratégica no que diz respeito à competitividade de mercado e na gestão de conhecimentos.

Foi considerado, ao longo do segundo capítulo, a importância que o conhecimento concreto dessas informações tem para a empresa, principalmente no que diz respeito à segurança, tratamento, disponibilização para os vários níveis, obtenção e integridade das informações no momento da tomada de decisões por parte da gestão.

Já no terceiro capítulo, a abordagem da ferramenta de Descoberta de Conhecimentos, KDD, demonstrou as formas e metodologias para a obtenção e análise dos dados presentes no Banco de Dados que, geralmente, já existe na empresa, demonstrando o papel de cada uma das etapas para o alcance do objetivo final de

apresentar dados em estruturas mais simples para a interpretação dos responsáveis pela tomada de decisões.

Finalizando a abordagem do tema proposto, o quarto capítulo apresentou um estudo de caso, através do qual foi possível exemplificar a importância e influência de se obter informações de qualidade, quanto às formas de utilização mais apropriadas da Descoberta de Conhecimentos em Bancos de Dados.

Com a analise dos resultados obtidos, a utilização, tratamento e obtenção de dados concretos e de fácil interpretação surtem efeitos positivos na administração empresarial por, como se pretendia demonstrar, permitir a extração e análise mais precisa das informações.

Espera-se que as informações teóricas e práticas apresentadas ao longo deste trabalho possa contribuir com a literatura especializada da área, promovendo a elucidação dos tópicos e fornecendo material para pesquisas, trabalhos e estudos futuros sobre o assunto.

Referências

Beal. A. Segurança da informação: princípios e as melhores práticas para a proteção dos ativos de informações nas organizações. São Paulo: Atlas, 2008.

Boente, A. N. P. Descoberta de Conhecimento em Bases de Dados. Rio de Janeiro, Tese de Doutorado - Departamento de Informática, AWU - Iowa (USA). 2006.

Chiavenato, I. Teoria, Processo e Prática. São Paulo: McGraw Hill, 1997.

Dias, Cláudia. Segurança e Auditoria da Tecnologia da Informação. Rio de Janeiro: Axcel Books, 2000.

Goldschmidt, R. R.; PASSOS, E. Data Mining: Um Guia Prático. Rio de Janeiro: Campus, 2005.

Lara, C. R. D. A Atual Gestão do Conhecimento: A Importância de Avaliar e Identificar o Capital Humano nas Organizações. São Paulo: Nobel, 2004.

Robredo, J. Organização dos documentos ou organização da informação: uma questão de escolha. DataGramaZero - Revista de Ciência da Informação - v.5 n.1 fev/04. Disponível em: <http://www.dgz.org.br/fev04/F_I_art.htm>. Acesso em: 22 out. 2015.