Holisson Soares da Cunha

Descoberta de Perfis de Consumo a Partir de uma Arquitetura de Data Warehouse Voltada para Dados da Web.

Alegrete

2013

Holisson Soares da Cunha

Descoberta de Perfis de Consumo a Partir de uma Arquitetura de Data Warehouse Voltada para Dados da Web.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal do Pampa

Orientador: Dr. Sérgio Luís Sardi Mergen

Coorientador: Dr. Fábio Natanael Kepler

Alegrete

2013

Holisson Soares da Cunha

Descoberta de Perfis de Consumo A Partir de uma Arquitetura de Data Warehouse Voltada para Dados da Web.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Trabalho de Conclusão de Curso defendido e aprovado em 10 de outubro de 2013

Banca examinadora:

Dr. Sérgio Luís Sardi Mergen
Orientador
UNIPAMPA

Dr. Fábio Natanael Kepler
Coorientador
UNIPAMPA

MSc. Andréa Sabedra Bordin
UNIPAMPA

MSc. Alessandro Bof de Oliveira
UNIPAMPA

Este trabalho é dedicado à meu pai (in memoriam).

Agradecimentos

Em primeiro lugar, gostaria de agradecer à minha Família. Especialmente à meus pais, Theótimo Lopes da Cunha (*in memorian*) e Alaídes Soares da Cunha, pelo seu amor, carinho e dedicação, que permitiram esse sonho se tornar realidade. À minha namorada Luriane Guterres, pelo imenso amor, apoio e paciência nos momentos difíceis. Aos meus irmãos, Roberto, Marta, Jesus e Rosa, pela amizade.

Aos grandes amigos, João Paulo Aires e Jader Saldanha, que tive a feliz oportunidade conhecer através da graduação. Sou grato pelo companheirismo deles, a amizade sincera e por todos os momentos que me proporcionaram durante esse período, seja pelas conversas jogadas fora, as jogatinas, os sorrisos e todas as situações dentro e fora do contexto acadêmico, que fazem deles, pessoas tão especiais em minha vida.

Ao Evaír Severo, pela amizade e paixão pelo Grêmio, que nos levou a comemorar juntos diversas vitórias e lamentar pelos títulos não conquistados. Aos colegas e amigos, Henrique Gressler e Luiza Bagesteiro, pessoas maravilhosas que tive a oportunidade de conhecer durante a graduação. Aos demais colegas que da turma de 2009 da CC, pelas experiências compartilhadas.

À família Balestra, minha segunda família. Agradeço pela grande amizade e por toda ajuda que me deram durante a graduação.

Ao Professor Sérgio L. S. Mergen, pela orientação e dedicação. Não podería deixar de registrar minha admiração e respeito a essa pessoa inteligente e inspiradora, que além de me orientar em diversas frentes de trabalho, foi um grande mestre e amigo.

À Professora Amanda Meincke Melo, pela sua amizade e orientação. Sou grato por todos os ensinamentos e pelo aprendizado que adquiri com ela durante o tempo que atuei em atividades de extensão.

Às professoras Jussara Lipinski, Aline Piccinini e demais colegas do que tive a oportunidade de conhecer através do Projeto Rondon.

À Universidade Federal do Pampa (UNIPAMPA), seus funcionários, direção, e todas as pessoas que trabalham e axiliam no crescimento desse espaço.

Enfim. A TODOS que contribuíram e continuam a contribuir em meu processo de crescimento pessoal e profissional.

Até a próxima.

"Quem senta no fundo do poço para contemplar o céu,
há de acha-lo pequeno". Hanyu

Resumo

Ambientes de Data Warehouse (DW) normalmente são utilizados para manipular dados de bases operacionais concretas. Entretanto, com o crescimento de informações relevantes disponibilizadas na Web sobre os interesses e desejos de consumo de usuários, surge a necessidade de empregar a arquitetura de DW na Web, o que exige que tanto os mecanismos de extração de dados quanto a posterior visualização sejam repensados. Neste trabalho é demonstrada a possibilidade de se utilizar uma arquitetura típica de DW dentro de uma aplicação Web voltada para análise de perfis de consumo. A aplicação é uma rede social de consumo, que visa aproximar pessoas que possuem perfis de compra semelhantes. A extração de dados ocorre através de um formulário de registro de compra, onde os usuários fornecem suas informações de consumo. A visualização é realizada através da geração de comunidades, que permitem a troca de experiências de usuários que tenham perfis de consumo parecidos. Para realizar a descoberta de perfis de consumo, foram avaliados três algoritmos de aprendizado não supervisionado: *K-means, Expectation Maximization e Farthest First*. Através dos experimentos observou-se um comportamento semelhante entre as abordagens quando os perfis de consumo são semelhantes. Com perfis bem definidos, o algoritmo EM obteve melhores resultados.

Palavras-chave: Data Warehouse. Modelagem Multidimensional. Mineração de Dados. Aprendizado de Máquna. Clusterização de Dados.

Abstract

Environments for Data Warehouse (DW) are typically used to manipulate data bases concrete operational. However, with the growth of relevant information available on the Web about the interests and desires of consumer users, the need arises to use the DW architecture of the Web, which requires that both the mechanisms of data extraction as later viewing are rethought. This work demonstrated the possibility of using a DW typical architecture within a Web application oriented analysis consumption profiles. The application is a social network of consumption, which aims to bring together people who have similar buying profiles. Data extraction occurs through a registration form to purchase, where users provide their information consumption. Visualization is performed by generating communities that allow exchange of experiences of users who have similar consumption profiles. To perform discovery of consumption profiles were evaluated three unsupervised learning algorithms: K-means , Expectation Maximization and Farthest First. Through experiments we observed a similar behavior between the approaches when the consumer profiles are similar. With well defined profiles , the EM algorithm achieved better results .

Key-words: Data Warehouse. Multidimensional Modeling. Data Mining. Machine Learning, Data Clustering.

Lista de ilustrações

Lista de siglas

AAD Ambientes de Apoio à Decisão

AM Aprendizado de Máquina

AMNS Aprendizado de Máquina Não Supervisionado

AMS Aprendizado de Máquina Supervisionado

DW Data Warehouse

MD Mineração de Dados

MMD Modelagem Multidimensional de Dados

TCC Trabalho de Conclusão de Curso

Sumário

1 Introdução

A quantidade de pessoas utilizando internet no Brasil aumentou em 143,8% nos últimos seis anos (IBGE, 2013). São milhões de usuários conectados, consumindo e compartilhando informações através da web, tornando-a uma vasta coleção de documentos heterogêneos, com estruturas diferentes e natureza dinâmica, onde diariamente são criadas e indexadas milhares de novas páginas.

O aumento dessas informações trouxe um impacto no atual ambiente de negócios, o tornando cada vez mais complexo e competitivo. As constantes alterações e variações de mercado exigem que as empresas respondam rapidamente a essas novas condições impostas, sendo inovadoras e trazendo diferenciais na forma de operar, visando manter a competitividade e acompanhar as variações no ambiente de negócios.

Um dos segmentos que ganhou força com o aumento da utilização da internet foi o comércio eletrônico. Em 2011 foram gastos R$ 18,7 bilhões com compras pelas internet (IBGE, 2013), o que permite visualizar o aumento da procura de produtos e serviços pelos usuários nesse segmento. Esse crescimento trouxe oportunidades e desafios para as empresas, pois o grande volume de informações de consumo obtidas permite que análises de perfil e segmentação de mercado sejam realizadas, explorando dados fornecidos pelos usuários através de sites de compra, portais de *reviews* e através das mídias sociais, espaço comum onde os usuários indicam seus interesses de consumo e expõem suas opiniões com base em experiências passadas.

Nesse contexto está a descoberta de conhecimento, sendo uma prática cada vez mais explorada por empresas que possuem grandes volumes de dados e buscam conhecer seus clientes, visando oferecer produtos e serviços específicos, baseados no seu perfil de interesse.

Uma das formas de descobrir perfis de interesse é através do agrupamento de usuários. Essa tarefa consiste em agrupar pessoas que possuem interesses de consumo semelhantes, com base em seu histórico de compras, permitindo que empresas direcionem campanhas de marketing para grupos específicos de clientes. No contexto das mídias sociais, o agrupamento de usuários é um forte mecanismo para promover a interação, onde a partir de interesses comuns de compra, usuários podem trocar informações e experiências de consumo que os auxiliem em suas futuras tomadas de decisões.

Devido a possibilidade de extrair conhecimento e informação útil, o uso de ambientes *data warehouse* DW surje como uma abordagem relevante para o problema de agrupamento. Segundo Mannino (2008) "Um DW trata-se de um repositório central de dados, onde dados de bases operacionais e demais fontes heterogêneas são integrados,

limpos e padronizados, servindo principalmente para aplicações de suporte a decisão e análise".

Visando cumprir essas tarefas, ambientes DW utilizam técnicas de coleta/extração de dados que alimentam o repositório de dados. Esse repositório utiliza a Modelagem Multidimensional de Dados MMD, pois a mesma oferece recursos computacionais que permitem a manipulação dos dados de forma flexível, sob diferentes visões. Apoiado pelo uso da MMD, as técnicas de Mineração de Dados (MD) são utilizadas para extrair conhecimento a partir dos dados armazenados em ambientes DW.

Mannino (2008) definiu Mineração de Dados MD como um "processo de descobrir padrões implícitos nos dados e usá-los para obter vantagens de negócio". As técnicas de MD utilizam algoritmos de Reconhecimento de Padrões, Estatística e Inteligência Artificial, sendo comum a adoção de técnicas de aprendizado de máquina, utilizando principalmente métodos não supervionados para extrair conhecimento e padrões consistentes a partir de grandes quantidades de dados, com o objetivo de apoiar na tomadas de decisões. Diversas abordagens envolvem a MD, entre as mais utilizadas nos ambientes de negócios, estão os Sistemas de Recomendação (FILHO; GEUS; ALBUQUERQUE, 2008), Marketing direcionado e Segmentação de mercado.

Entre as técnicas de MD adotadas para problemas de agrupamento, destaca-se o uso de métodos de Aprendizado Não supervisionado. Esses algoritmos utilizam as caracteristicas dos dados, visando encontrar padrões e similaridades entre os mesmos, realizando o agrupamento de forma automática. Entre as diversos métodos utilizadas para problemas de agrupamento, destacamos os que são utilizadas neste trabalho: *Expectation Maximization, Farthest First e K-Means*

Ambientes DW normalmente são utilizados para analisar dados corporativos, extraídos de bases de dados operacionais concretas. Entretanto, com o crescimento de informações relevantes disponibilizadas na Web pelos usuários sobre seus interesses e desejos de consumo, surge a necessidade de empregar a arquitetura de DW na Web. Nesse caso, alguns dos desafios que precisam ser enfrentados envolvem a forma de coletar os dados disponíveis na Web e a disponibilização das análises realizadas sobre os dados na própria Web. Além disso, é importante verificar se algoritmos de clusterização são efetivos para agrupar dados de consumo.

Nesse sentido, este Trabalho de Conclusão de Curso (TCC) possui dois objetivos gerais, conforme destacado abaixo:

- Utilização de uma Arquitetura de *Data warehouse na Web* - Como primeiro objetivo geral, este trabalho visa demonstrar como uma arquitetura de DW pode ser empregada para análise de dados de consumo coletados da web. Abaixo citamos os objetivos específicos:

  - Demonstrar como uma rede social de consumo pode coletar dados de consumo da Web, em vez de utilizar algoritmos de extração e coleta de dados;
  - Demonstrar como uma rede social de consumo pode disponibilizar os agrupamentos identificados para que os usuários da rede possam utilizzá-los;
  - Construir um protótipo de rde social que implemente tanto a coleta dos dados como a disponiblização dos agrupamentos gerados.

- Validar algoritmos não supervisionados – Como segundo objetivo geral, é proposta a validação de algoritmos não supervisionados, utilizados na descoberta de perfis de consumo a partir de dados armazenados na arquitetura de DW. Como objetivos específicos, temos:
  - Criar uma ferramenta que possibilite gerar automaticamente cenários de teste para dados de consumo;
  - Propor métricas de avaliação para verificar a eficiência dos métodos de agrupamento utilizados nesse trabalho;
  - Criar cenários de teste abrangentes, permitindo avaliar o comportamento e as características de cada dos métodos utilizados para descoberta de perfis de consumo;
  - Executar os cenários de teste utilizando o *framework* de mineração de dados WEKA;
  - Analisar o desempenho dos algoritmos definidos para gerar os perfis de consumo: *K-means, Farthest First e Expectation Maximization (EM).*

Este trabalho está organizado da seguinte forma: O capítulo 2 aborda a fundamentação teórica, conceituando as tecnologias e abordagens utilizadas. No capítulo 3, são apresentadas as abordagens relacionadas a esse trabalho. No capítulo 4 é apresentado o ambiente de DW voltado para extração de dados da web. No capítulo 5, são apresentadas as configurações dos experimentos realizados e os resultados obtidos na avaliação dos métodos de agrupamento, e no capítulo 6, sao apresentadas as considerações finais e conclusões do trabalho.

2 Fundamentação Teórica

Este capítulo tem como objetivo conceituar as tecnologias e abordagens aplicadas à esse trabalho. Entre as principais abordagens empregadas, estão: *Data Warehouse*, Mineração de Dados e Aprendizado de Máquina.

2.1 Data Warehouse

O conceito de DW, criado por *William Inmon* em 1990, surgiu da necessidade do domínio de informações estratégicas por parte de empresas, a fim de alcançar respostas rápidas e eficientes, garantindo a competitividade em um mercado de constante mutação (MACHADO, 2010) (ELMASRI et al., 2005). Os avanços tecnológicos e as mudanças organizacionais e estruturais, além da globalização da economia, contribuíram para a absorção da tecnologia *data warehousing* em ambientes corporativos, transformando ambientes de apoio à decisão, em ambientes DW.

"Um DW é um conjunto de dados baseado em assuntos, integrado, não volátil e variável em relação ao tempo, de apoio às decisões gerenciais"Inmon et al. (2002). Trata-se de um repositório central de dados, onde dados de bases operacionais e demais fontes externas, são integrados, limpos e padronizados, servindo principalmente para aplicações de suporte a decisão (Mannino, 2008). A tecnologia DW possibilita a análise e manipulação de grandes volumes de informação, coletados de sistemas transacionais (OLTP).

Por se tratar de um sistema de apoio a decisão, o DW cria as chamadas séries históricas, que possibilitam uma melhor análise de eventos passados, oferecendo suporte à decisões presentes e a previsão de eventos futuros. Por definição, os dados armazenados nesses sistemas não são voláteis, ou seja, eles não são alterados, salvo quando existe a necessidade de realizar correções de dados previamente carregados. Normalmente estando disponíveis somente para leitura e não devendo ser alterados.

O objetivo dessa tecnologia é de fornecer os subsídios necessários para a transformação de uma base de dados transacional e com um conjunto de dados relativamente recente, em uma base que contenha o histórico de todos os dados de interesse em uma empresa, voltados para análise estratégica. Nesse contexto, o DW proporciona aos Ambientes de Apoio à Decisão (AAD) uma sólida e concisa integração da informação armazenada, permitindo que análises gerenciais sejam realizadas (ADELMAN, 1992).

O requisitos de processamento de suporte a decisão foram responsáveis pelo surgimento de quatro principais características na construção de um projeto de DW, sendo eles:

- Orientado por assunto - O DW está organizado de acordo com os principais assuntos de negócio, que podem ser: clientes, pedidos ou produtos. Essa é uma característica que distingue o ambiente DW do ambiente transacional. Pois em ambientes transacionais, o modelo de negócio está mais direcionado a processos diários, enquanto que o DW se preocupa em gerar análises estratégicas temporais.

- Integrado - Os dados provenientes de bases de dados operacionais são extraídos e integrados antes de serem carregados para o DW, esse processo oferece um banco de dados íntegro e unificado para suporte à decisão.

- Variante no tempo - A dimensão de tempo é de suma importância para a identificação de tendências, predição de operações futuras e estabelecimento de objetivos. No DW os dados não são alterados, nem são mantidos sem alterações, guardando um histórico de atividades e processos.

- Não volátil - Em ambientes operacionais, são realizados diversos tipos de consultas sob os dados, tais como alterações/remoções de dados e modificações nas estruturas de armazenamento. No ambiente DW, basicamente são realizados dois tipos de operações: a carga de dados e o acesso aos dados, não existindo atualizações como parte do processo normal.

Nas próximas seções, serão abordados tópicos relacionados a itens que compõe um ambiente DW.

2.1.1 Arquitetura Data Warehouse

A arquitetura de um DW determina como será a organização de seus componentes. A definição de uma arquitetura constitui uma tarefa crucial para o projeto, devido à grande dependência existente entre a implementação dos componentes e sua organização. Uma arquitetura de DW, basicamente está distribuída em três partes: extração, armazenamento e apresentação.

Várias são as arquiteturas descritas na literatura e propostas por empresas que desejam extrair conhecimento de suas bases de dados. Entre as principais, figuram as arquiteturas *Top-Down* e *Bottom-Up*. A diferença entre elas refere-se à forma de implementação dos componentes do DW, de acordo com as necessidade de cada ambiente de negócios.

A arquitetura *Top-down*, introduzida por Inmon(1997) é caracterizada pela existência de um DW central que armazena todas as informações de uma corporação, e uma série de *Data Marts* (MORAIS, 2000) com dados derivados do DW. A arquitetura *Bottom-Up* caracteriza-se pelo armazenamento e extração, a partir da criação incremental

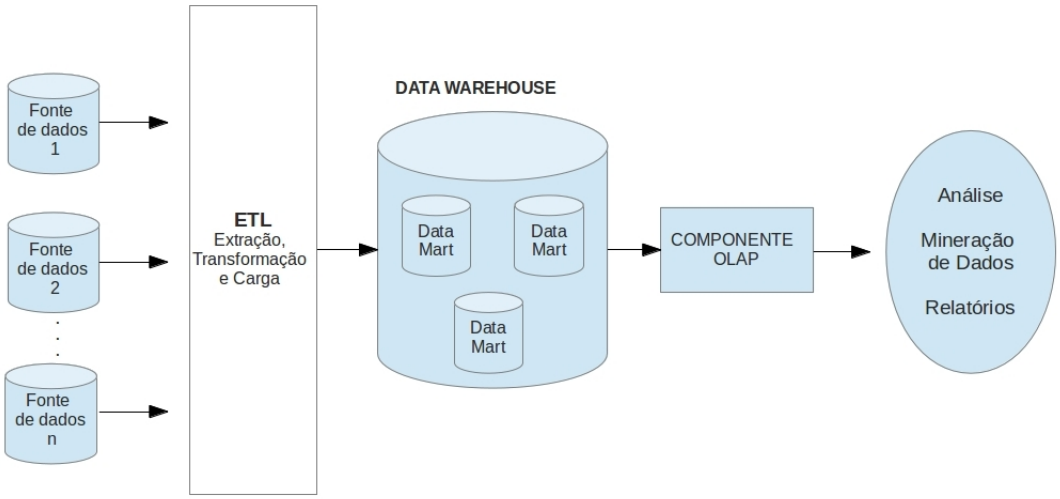

Figura 1 – Arquitetura Genérica de um DW.

de vários *data marts* independentes, com metadados e área de extração individualizadas, que no conjunto forma as fontes de dados que compõe o DW.

A Figura 1 representa a arquitetura genérica de um DW. Nas seções seguintes serão abordados cada um dos itens que compõem essa estrutura.

2.1.2 ETL - Extração, Transformação e Carga

ETL *(Extract, Transform and Load)* é um processo amplo, crítico e talvez um dos mais demorados na construção de um ambiente DW. Consiste na extração de dados oriundos de bases heterogêneas, na transformação e limpeza desses dados e posteriormente no carregamento para o DW. A eficiência das informações de apoio a tomada de decisão está diretamente relacionada ao processo de ETL. Os dados armazenados devem ser tratados, pois falhas nas etapas de extração ou transformação podem afetar na descoberta de padrões e conhecimento, fazendo com que decisões sejam tomadas erroneamente (FERREIRA et al., 2010).

O processo de ETL abrange três etapas principais: extração, transformação e carga de dados. Essas etapas serão vistas nas próximas seções.

2.1.2.1 Extração

A etapa de extração leva em torno de 60% das horas de desenvolvimento de um DW e tem como objetivo, buscar em sistemas e fontes externas, informações que sejam mais relevantes e que estejam em conformidade com o modelo de dados do DW. As fontes de origem desses dados podem ser tanto bases operacionais, informações contidas na web, entre outras fontes heterogêneas. Nesses casos, onde os dados se concentram

em bases distantes, com plataformas diferentes, é necessário construir mecanismos de extração diferentes para cada local (ALMEIDA, 2006).

Após a construção de um DW, é comum que a carga inicial dos dados seja completa, fazendo com que o extrator busque todos os dados da fonte original, entretanto, no decorrer das buscas por dados, é importante que a extração realize apenas a carga incremental dos dados, buscando registros que foram alterados ou inseridos desde a última carga (FERREIRA et al., 2010).

Com foco neste trabalho, existem diversas técnicas automáticas de extração de dados da web, como: *Deep Web Data Extraction, Data Extraction by Example, Wrapper Induction, Regular Expressions, and Natural Language Processing.*

A *Deep web* (HONG, 2010), consiste em extrair conteúdos da web gerados dinâmicamente, que não podem ser encontrados diretamente, sendo encontrados através do preenchimento de formulários, por exemplo. Na *Extract Data by Example* (LAENDER; RIBEIRO-NETO; SILVA, 2002), a extração ocorre através exemplos fornecidos pelo usuário. Esses exemplos são obtidos das bases de dados das quais se deseja extrair objetos, que podem ser páginas da web que contenham informações de produtos e preço. O *wrapper induction* é um mecanismo que permite a extração de dados estruturados e não-estruturados, muitas vezes encontrados através de páginas no formato HTML, onde esses dados estão misturados com outras informações do mesmo contexto (DALVI; KUMAR; SOLIMAN, 2011). O uso de Expressões regulares para extração de dados é um mecanismo utilizado para identificar padrões específicos em um documento, onde se possa retirar informações úteis para posterior análise (BARRERO; CAMACHO; R-MORENO, 2009). Já a extração através do processamento de linguagem natural (PNL) (FRIEDLIN; MCDONALD, 2006), consiste na aplicação de métodos e técnicas com o propósito de extrair semântica de informações disponibilizadas na web, podendo ser utilizada para analisar em rede sociais e páginas de *reviews*, opiniões e experiências de consumo dos usuários a respeito de diversos produtos, servindo como *feedback* para que outros usuários possam realizar uma compra mais adequada.

2.1.2.2 Transformação

Em linhas gerais, a atividade de transformação visa realizar ajustes sobre os dados, com o objetivo de melhorar sua qualidade. Nessa etapa, caso seja necessário, os dados são limpos e consolidados para depois serem carregados no DW. Para solucionar possíveis problemas de limpeza nos dados, a etapa de transformação utiliza técnicas de limpeza de dados (*data cleaning*), que tem como objetivo detectar e remover anomalias dos dados; e Desambiguidade de dados (*Data Disambiguation*)(DREISEITL; VINTERBO; OHNO-MACHADO, 2001), que busca categorizar corretamente esses dados. Através dessas técnicas é possível alcançar unicidade de informação e obter melhores benefícios de suporte

à decisão.

Para garantir a qualidade dos dados, (KIMBALL; MERZ, 2000), apresenta as seguintes características:

a) Unicidade dos dados, evitando assim duplicações de informação; b) Precisão dos dados, os dados não podem perder suas características originais, assim que são carregados para o DW; c) Dados completos, não gerando dados parciais de todo o conjunto relevante às análises e; d) Consistência, ou seja, os fatos devem apresentar consistência com as dimensões que o compõem.

Grandes problemas são encontrados no decorrer deste processo, como a ambiguidade de dados que podendo ter a mesma nomenclatura tem significados diferentes, ou ao contrário, valores diferentes apresentarem o mesmo significa como os valores nulos que podem ser representados com conteúdo vazio ou 0 (zero). Existem problemas com a integridade referencial dos dados, onde dados que farão a composição de um fato podem ser negligenciados ou não encontrados na origem no momento de gerar uma dimensão. A representação de conjunto de caracteres em bases distribuídas e que são configuradas com tabelas de caracteres diferentes também implicam em problemas no momento da transformação. Por fim temos a aplicação das regras de cálculos existente nos metadados para gerar novos valores a partir de valores da origem, como por exemplo sumarizações de vendas.

2.1.2.3 Carga

A carga é a última etapa do processo de ETL, após os dados serem extraídos e transformados, resta armazená-los no DW. Nesse processo, basicamente as dimensões estáticas de modificação lenta e fatos integrantes do modelo de dados são carregados. Por ser uma atividade que demanda alto custo de processamento, muitas vezes não pode ser extensa, devido a utilização contínua do DW (ALMEIDA, 2006).

Em ambientes de DW convencionais, que armazenam dados de bases operacionais, normalmente os dados são carregados durante a noite, em períodos de menos acesso. A frequência da atualização irá depender da regra de negócio aplicada, onde a política de atualização menos frequente, indica que um maior volume de dados será carregado, o que também significa que por um tempo maior, os dados estarão desatualizados. Em casa onde se necessita de dados atualizados em curtos períodos de tempo, essa frequência deve ser maximizada.

Considerando-se dados carregados da web, as políticas de atualização levam em consideração os mesmos requisitos dos ambientes clássicos de DW, entretanto, quando consideramos fontes de dados autônomas ou desconhecidas, podem ocorrer mudanças frequentes em seu conteúdo. A atualização para o DW nesses casos deve ser mais frequente,

pois a extração apenas adquire o dado atual presente naquela fonte, ficando desatualizado logo após que alguma alteração for realizada.

Um desafio para os ambientes DW que extraem dados da web, é determinar o período de atualização, para ter ao tempo todo, dados atualizados para que análises sejam geradas com qualidade e precisão. Nesse contexto é utilizado o reconhecimento de expressões temporais, área do Processamento de Linguagem Natural que utiliza características de tempo, incluídas na fonte de dados para substituir apenas o que for mais atual, apenas completando o banco de dados com novas informações.

2.1.3 Modelo de Dados

A elaboração do modelo de dados concentra-se na observação dos fatos relevantes que ocorrem na realidade, de forma simplificada e com a finalidade de construir um sistema que possa automatizar as necessidades de informação para transações de negócio. Em ambientes de negócio, um modelo de dados bem definido auxilia a visualizar o sistema como ele realmente é, ou como se deseja que ele seja, sendo possível especificar a estrutura e o comportamento do sistema que se quer alcançar. Em linhas gerais, um modelo de dados documenta as decisões que serão tomadas pela corporação, suprindo suas as necessidades de fim estratégico e operacional.

O sucesso no desenvolvimento de um DW depende da escolha correta das estratégias a serem adotadas, de forma que sejam adequadas às características e as necessidades específicas do ambiente onde será implementado. Para o desenvolvimento de um DW, a abordagem utilizada é a Modelagem Multidimensional de Dados, devido a sua flexibilidade e fácil manipulação dos dados. Uma modelagem bem definida leva o usuário a ter uma garantia da confiabilidade dos dados e uma maior qualidade nos resultados obtidos (MONTEIRO; PINTO; COSTA, 2013).

2.1.4 Modelo Multidimensional de Dados

A MMD utilizada em ambientes DW é uma técnica de concepção e visualização de um modelo de dados de um conjunto de medidas que descrevem aspectos comuns de negócios (MACHADO, 2010). É utilizada especialmente para sumarizar e reestruturar dados e apresentá-los em visões que suportem a análise dos valores desses dados. Uma MMD é formada basicamente por três elementos: fatos, dimensões e medidas.

Um fato é uma coleção de itens de dados, composto por medidas que descrevem um assunto central de um negócio. Cada fato representa um item, uma transação ou um evento de um negócio, utilizado para analisar um determinado processo. Entre as características gerais de um fato, podemos citar que é representado por valores numéricos e implementado em tabelas denominadas tabelas fato (SOARES, 1998). Conceitualmente,

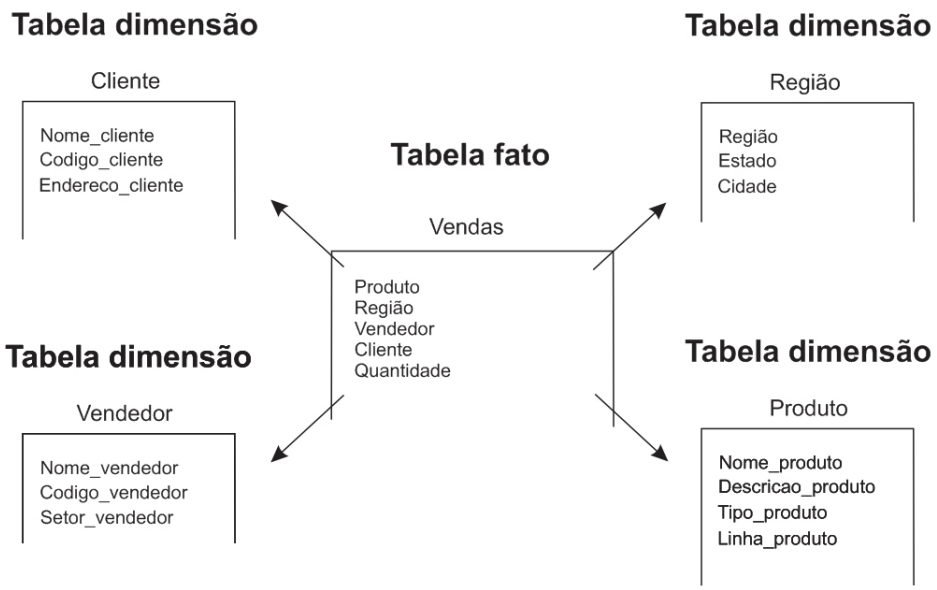

Figura 2 – Esquema Estrela, utilizado na Modelagem Multidimensional de Dados.

dimensão é todo elemento que participa de um fato ou assunto de negócio. As tabelas dimensões são elementos que descrevem o contexto de um assunto e normalmente não possuem valores numéricos, pois são descrições dos elementos que participam de um negócio. Medidas ou variáveis, são os atributos numéricos que representam os fatos.

Na MMD, dois esquemas de dados são muito comuns: o esquema estrela e o esquema floco de neve. Ambos serão abordados nas próximas seções.

2.1.4.1 Esquema Estrela

O esquema estrela é caracterizado por uma tabela central, denominada fato, que se relaciona com entidades menores chamadas dimensões. Esse esquema possui uma estrutura simples, com poucas tabelas, e relacionamentos bem definidos (WAGNER et al., 2012). Sua estrutura assemelha-se ao modelo de negócio, o que facilita a leitura e entendimento não só pelos analistas, como por usuários finais não familiarizados com estruturas de banco de dados.

A modelagem permite a criação de um banco de dados que facilita a execução de consultas complexas, podendo ser realizadas de modo eficiente e intuitivo pelo usuário. Na Figura 2 o centro da estrela é o fato vendas, composta pelas dimensões que participam desse fato.

2.1.4.2 Esquema Floco de Neve

O esquema floco de neve é uma variação do esquema estrela. sendo resultado de uma decomposição de uma ou mais dimensões que possuem hierarquias entre os seus membros (MACHADO, 2010), conforme mostra a Figura 3.

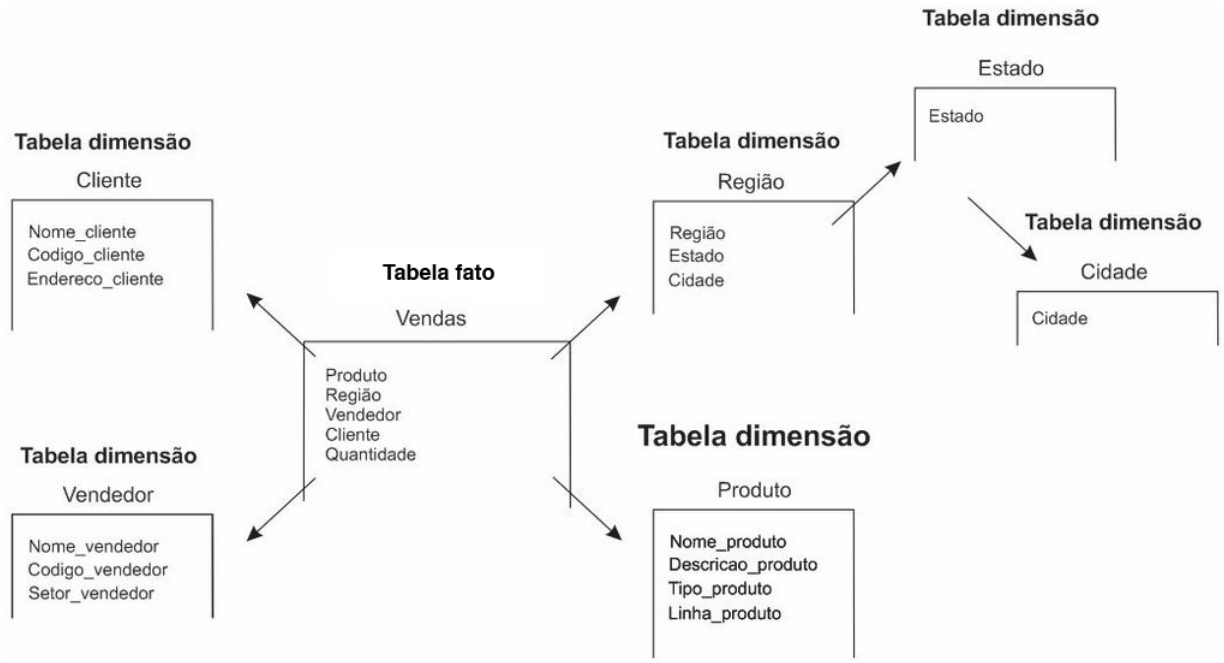

Figura 3 – Esquema Floco de neve, utilizado na Modelagem Multidimensional de Dados.

> Segundo Hokama et al. (2004) "A aplicação deste esquema resulta em uma diminuição na performace nas consultas devido a necessidade de um maior número de joins além do aumento da complexidade da modelagem. Em contrapartida a atualização dos dados no DW será mais rápida devido a normalização das tabelas. Sabendo das diferenças entre os esquemas é que o arquiteto do DW definirá qual será a melhor opção para ser aplicada".

2.1.5 OLAP

Segundo (MACHADO, 2010) OLAP (*Online analytical processing*) é o conjunto de ferramentas que possibilita efetuar a exploração de dados em um ambiente DW. Enquanto o DW tem a finalidade de armazenar dados, as operações OLAP foram desenvolvidas para realizar a recuperação de informação, ambos com o intuito de gerar informação estratégica.

A análise denominada multidimensional representa os dados como dimensões. Realizando a combinação entre essas dimensões, o usuário tem uma visão total sobre os dados, sob diferentes perspectivas. A funcionalidade de uma ferramenta OLAP é caracterizada pela análise dinâmica dos dados. Entre as operações básicas estão: *Slice and Dice* e *Drill*. As operações *Drill*, denominadas *Drill Up*, *Drill Down*, *Drill Throught* e *Drill Cross*, realizam a navegação sobre dados, aumentando ou diminuindo o nível de detalhamento das consultas, enquanto que as operações *Slide and Dice* são utilizadas para criar visões dos dados por meio de sua reorganização, de forma que eles possam ser examinados sob diferentes perspectivas. (MACHADO, 2010).

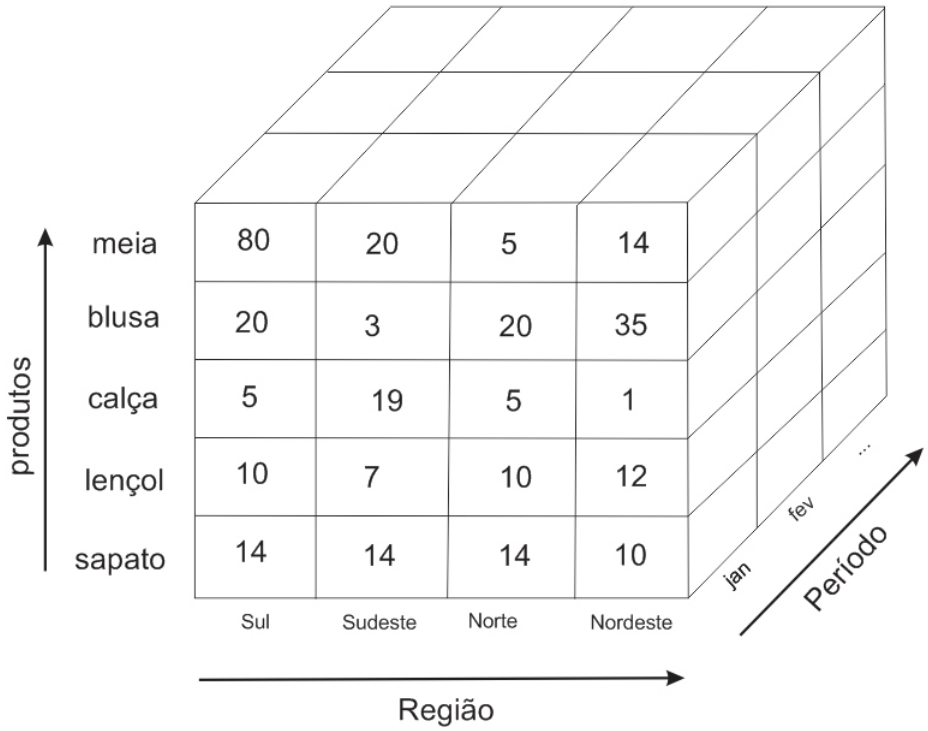

Figura 4 – A Representação dos dados na MMD é feita através de um cubo tridimensional de dados.

2.1.5.1 Cubo Tridimensional de Dados

Modelos Multidimensionais tiram proveito dos relacionamentos inerentes dos dados para preenchê-los em matrizes multidimensionais, chamados cubos de dados (ELMASRI et al., 2005). Um cubo de dados é uma representação que expressa a forma de como os dados se relacionam entre si, sendo formado a partir da entidade central (fato) e pelas entidades dimensão. O cubo armazena todos os dados relacionados a um determinado assunto e permite gerar várias combinações entre eles. Através dessa representação é possível visualizar as várias combinações existentes entre os dados, resultando na extração de várias visões sobre um mesmo assunto.

A Figura 4 mostra um cubo de dados tridimensional que apresenta dados de vendas de produtos por região do Brasil, em um determinado período de tempo. Cada célula contém os dados de produtos específicos, regiões do Brasil e períodos do ano. Dessa forma, é possível analisar todas as ocorrências de vendas de produtos, dividido por região do Brasil e em qualquer período do ano que o usuário desejar. Entre as possibilidades que a MMD oferece, está a mudança de hierarquia (orientação). Nessa técnica o cubo de dados pode ser imaginado realizando giros, para mostrar a orientação sobre diferentes perspectivas, atingindo novas combinações com os mesmos dados.

2.1.6 Mineração de Dados *(Data Mining)*

O ambiente de negócios está se tornando cada vez mais complexo e competitivo. Empresas privadas e públicas se sentem pressionadas, tendo que responder rapidamente a condições do mercado, além de serem inovadoras e trazer diferenciais na forma como operam. Essa necessidade exige das empresas agilidade e eficiência na tomada de suas decisões, sendo fator determinante para manter a competitividade e acompanhar as variações no ambiente de negócios (TURBAN et al., 2009).

Em resposta a essa necessidade, surgiram as técnicas de Mineração de Dados MD. Mannino (2008) definiu MD como um "processo de descobrir padrões implícitos nos dados e usar aqueles padrões para obter vantagens de negócio. Esse processo melhora a habilidade de detectar, compreender e prever padrões". Na literatura, é comum que a mineração de dados seja citada como um sinônimo de *Knowledge Discovery in Databases* (KDD), ou descoberta de conhecimento em banco de dados. KDD é um processo que consiste em tarefas bem definidas, que inclui a mineração de dados como parte desse processo.

A MD, é composta por uma série ferramentas e métodos. Para que a descoberta de conhecimento seja eficiente, é importante que se tenham regras de negócio bem definidas, através de um modelo de dados íntegro. Entre os métodos utilizados na MD, podemos citar a classificação de dados, modelos de relacionamento entre variáveis, análise de agrupamentos, sumarização, modelo de dependência, regras associativas e análise de séries temporais, conforme (FAYYAD et al., 1996). Na maioria desses métodos, são utilizados algoritmos de inteligência artificial, redes neurais e estatística. Esses algoritmos são capazes de explorar um conjunto de dados, encontrando relações existentes, onde através dessa descoberta podem ser gerados gráficos, grafos, árvores de decisão, regras e outras formas para apresentação desse novo conhecimento.

Na próxima seção, serão abordadas técnicas de mineração de dados baseadas em Aprendizado de Maquina, voltadas para o problema de descoberta de conhecimento.

2.2 Aprendizado de Máquina

Aprendizado de máquina Aprendizado de Máquina (AM) é uma área dentro da Inteligência Artificial que tem como objetivo desenvolver técnicas computacionais para construção de algoritmos e sistemas capazes de obter conhecimento de forma automática, através de experiências anteriores bem sucedidas (MONARD; BARANAUSKAS, 2003). O AM faz uso de princípios indutivos com o intuito de alcançar conclusões a partir de um conjunto de exemplos. Por meio desses exemplos, hipóteses são geradas quando efetuadas inferências indutivas.

O AM está dividido em duas técnicas principais: Aprendizado de Máquina Supervisionado Aprendizado de Máquina Supervisionado (AMS) e o Aprendizado de Máquina Não Supervisionado Aprendizado de Máquina Não Supervisionado (AMNS). Nas seções seguintes serão conceituadas as duas áreas.

2.2.1 Aprendizado de Máquina Supervisionado

Técnicas de Aprendizado de Máquina Supervisionado AMS são tipicamente utilizado para treinamentos que envolvem: Redes Neurais Artificial e de Árvores de Decisão. A técnica funciona utilizando um "Supervisor"no ciclo de treinamento, que dirá se os modelos e suas previsões estão corretas ou não.

Uma forma de se implementar o supervisor é através do uso de conjuntos de treinamento, que envolvem exemplos previamente classificados, ou seja, onde o rótulo da classe associada é conhecido. De maneira geral, os exemplos são descritos através de vetores de valores com características ou atributos, e o rótulo da classe associada. O objetivo do algoritmo é construir um classificador que possa determinar corretamente a classe de novos exemplos ainda não rotulados, ou seja, exemplos que não tenham rótulo da classe (SOUTO et al., 2003).

2.2.2 Aprendizado de Máquina Não Supervisionado

O Aprendizado de Máquina Não Supervisionado AMNS é composto por exemplos que não recebem rótulos, isto é, não existe um conjunto de exemplo para treinamento previamente fornecido. São utilizadas apenas entradas disponíveis em seus atributos. Neste caso, algoritmos de AMNS são utilizados com o objetivo de encontrar padrões em dados, baseados em alguma caracterização de regularidade (EVERITT; HOTHORN, 2011). Assim, esses algoritmos consistem em agrupar uma coleção de elementos, baseado em medidas de similaridade existente entre os dados, juntando em grupos (*clusters*) os elementos com maior similaridade entre si (METZ, 2011) .

No AMNS existem diversas abordagens utilizadas para resolução de problemas de agrupamento. Entre alguns métodos, estão aqueles que serão utilizados neste trabalho: *K-Means, Expectation Maximization e Farthest First.*

2.2.2.1 K-Means

K-means é um algoritmo de agrupamento utilizado em técnicas de mineração de dados. Seu objetivo é agrupar elementos com base nos dados de entrada. Esse agrupamento é realizado através da comparação entre os valores numéricos dos dados. Dessa forma, o algoritmo vai determinar os agrupamentos automaticamente, sem a intervenção humana, ou seja, sem nenhum conjunto de treinamento.

É um dos algoritmos clássicos de agrupamento particional. Nele é definido um representante de um agrupamento, chamado centróide, que é um vetor médio calculado a partir de demais vetores que correspondem àquele grupo. Na equação abaixo, podemos visualizar o cálculo do centróide C para um determinado grupo G, onde x representa um usuário pertencente ao grupo G, e o número total de pontos está definido em |G| (REZENDE; MARCACINI; MOURA, 2011).

$$C = \frac{1}{|G|} \sum_{x \in G}^{n} x.$$

Assim, o centróide define o ponto central de um cluster, pois possui a menor distância euclidiana para os demais pontos pertencentes a um determinado grupo. O critério de parada ocorre quando não existem mais alterações nos agrupamentos, ou seja, o centróide gerado na iteração anterior ainda se mantém como centróide, convergindo a clusterização.

O algoritmo *K-means* pode ser descrito pelos passos a seguir, de acordo com (FONTANA; NALDI, 2009):

1 - Atribuem-se valores iniciais para os protótipos seguindo algum critério, por exemplo, sorteio aleatório desses valores dentro dos limites de domínio de cada atributo;

2 - Atribui-se cada objeto ao grupo cujo protótipo possua maior similaridade com o objeto;

3 - Recalcula-se o valor do centróide de cada grupo, como sendo a média dos objetos atuais do grupo;

4 - Repete-se os passos 2 e 3 até que os grupos se estabilizem;

A Figura 5 ilustra um exemplo de execução do algoritmo *K-means*.

(FONTANA; NALDI, 2009)

A complexidade do *k-means* é linear em relação ao número de elementos, o que possibilita uma aplicação eficiente em diversos cenários. No entanto, a necessidade de informar com antecedência o número de grupos pode ser vista como uma desvantagem, pois esse valor geralmente é desconhecido pelos usuários. Além disso, o método apresenta variabilidade nos resultados, pois a seleção dos centróides iniciais, afeta o resultado do agrupamento. Para minimizar esse efeito, o algoritmo é executado diversas vezes, com várias inicializações diferentes, e a solução que apresentar menor valor de erro é selecionada (FONTANA; NALDI, 2009).

2.2.2.2 Expectation Maximization (EM)

O algoritmo *Expectation Maximization* (EM) é uma abordagem iterativa que busca estimar a máxima verossimilhança entre os dados. Consiste na formalização da ideia in-

tuitiva de lidar com dados incompletos que podem ser considerados, por exemplo, como informações não preenchidas em um sistema de registros. Esse método substitui valores inexistentes por valores estimados, buscando atribuir valores consistentes, estimados a partir de amostras coletadas sobre dados completos, para parâmetros inicialmente desconhecidos (LUNA, 2004).

No contexto da mineração de dados, o EM possui um funcionamento um pouco diferente de outras abordagens de agrupamento, pois ele não atribui uma instância a um determinado *cluster*, e sim, calcula a probabilidade de uma instância pertencer a um cluster. Dessa forma, é possível assumir que cada instância pertence ao cluster a que possui a maior probabilidade (GIBSON et al., 2007).

O EM é definido basicamente por dois passos: *Expectation*, onde é calculada a probabilidade de cada instância pertencer a um *cluster*;*Maximization*, onde é calculada a distribuição de cada instância, visando maximizar a distribuição de probabilidade dessas instâncias.

2.2.2.3 Farthest-First

Farthest-First Traversal é um algoritmo de aproximação para o que é chamado de *k-center-problem*, tendo como objetivo agrupar k elementos a partir de uma função de custo, maximizando o raio do *cluster*. O processamento ocorre a partir da escolha de um ponto aleatório em um conjunto de dados. Após isso, é definido um outro ponto, este, afastado do anterior, e logo após outro ponto, distante dos outros dois, até que k pontos sejam obtidos. A distância de um ponto X dentro de um conjunto S é dada por $min\{(x, y) : y \in S\}$. Os pontos definidos são tidos como centróides dos *clusters*, e cada ponto restante é atribuído ao centróide mais próximo (DASGUPTA, 2002).

O algoritmo *Farthest-Firts Travesal* é semelhante ao *K-means*, tendo como principal diferença o cálculo dos centróides. No cálculo usado pelo *K-means*, o objetivo é minimizar a distância do centróide para seus respetivos elementos, conforme descrito na equação abaixo:

$$min|x_i - u_i|^2 \tag{2.1}$$

onde x_i é um elemento do cluster e u_i é o seu respectivo centróide.

Já no cálculo usado pelo *Farthest-First Travesal*, o centróide é definido como sendo o elemento do cluster com valor máximo das distâncias mínimas aos centróides atuais, conforme descrito na equação abaixo:

$$max_i * min_c\, d(x, c) \tag{2.2}$$

onde x é um elemento do cluster e c o respectivo centróide.

A Figura 6 ilustra um exemplo dos centróides obtidos através do processamento dos algoritmos *K-means* e *Farthest-First*.

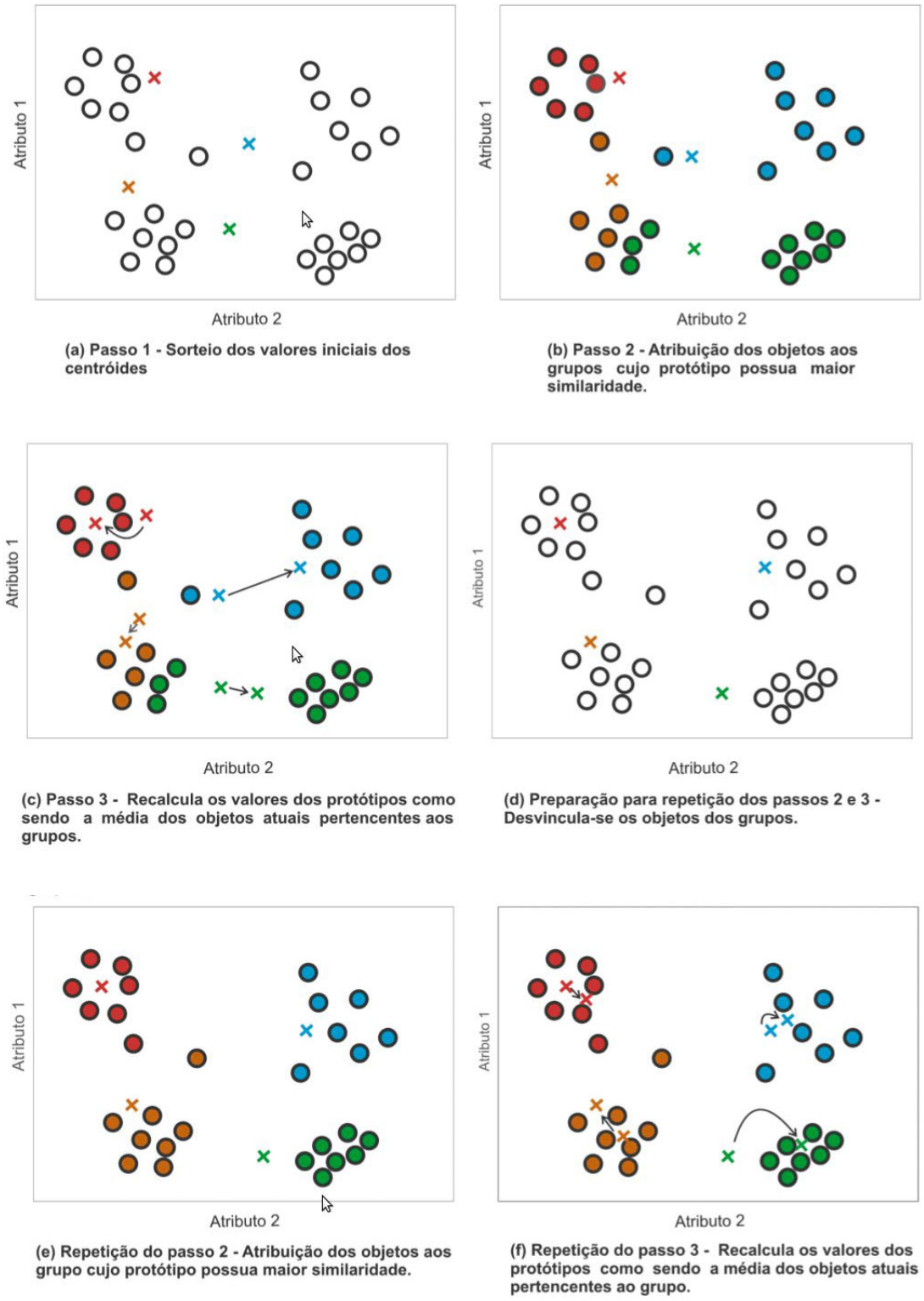

Figura 5 – Passos de execução do algoritmo *K-means*.

Fonte: Fontana e Naldi (2009)

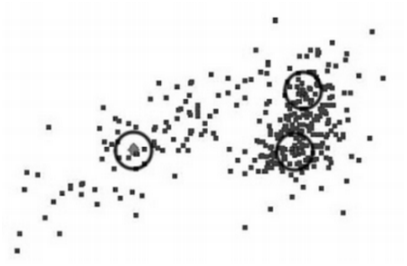

Figura 6 – Centróides formados pelo algoritmo *K-means*.

Fonte: FONSECA, SELECAO e ALGORITMOS (2008)

Figura 7 – Centróides formados pelo algoritmo *Farthest-First*.

Fonte: FONSECA, SELECAO e ALGORITMOS (2008)

3 Abordagens Relacionadas

Na literatura encontramos diversas pesquisas que envolvem a coleta de dados através da web para descoberta de conhecimento. Entre elas podemos citar:

- O uso de *data warehouses* voltados para coleta de dados na web, que diferente dos ambientes clássicos de DW, armazenam somente dados fornecidos pela web para realizar análises;
- Uma Rede Social que classifica pessoas de acordo com seu perfil acadêmico, permitindo a seus usuários conhecer e interagir com pesquisadores de interesses semelhantes.

O objetivo aqui é investigar as técnicas atuais empregadas nesses trabalhos, a fim de definir parâmetros que nos auxiliem a alcançar os objetivos propostos neste trabalho. A seção 3.1 irá apresentar um *framework* dirigido a modelos para o desenvolvimento de um DW voltado para a web, e na seção 3.2 é apresentada uma Rede Social Acadêmica que utiliza algoritmos de agrupamento para aproximar pesquisadores de áreas semelhantes.

3.1 Um ambiente*Data Warehouse* voltado para de dados da web

A análise da navegação na web é o processo de descobrir quais são os interesses de um usuário, com base na análise do seu histórico de navegação. Essas informações são de suma importância para compreensão e descoberta do comportamento, com o objetivo de apoiar no processo de tomada de decisão estratégica e aprimorar a sua experiência de navegação. Para cumprir os requisitos de negócio, ferramentas avançadas de análise na web requerem o desenvolvimento de um ambiente *data warehouse*, estruturando os dados através de uma Modelagem Multidimensional de Dados.

Existem diversas abordagens que definem uma MMD para analisar os registros de navegação dos usuários na web. O uso dessas abordagens permite a utilização de recursos OLAP e técnicas de mineração de dados para analisar esse conteúdo. Contudo, existem dificuldades quanto a uma definição da metodologia apropriada para a definição do modelo de dados. Algumas dessas abordagens permitem que os analistas definam os fatos e dimensões adequadas, enquanto que outras definem o modelo de acordo com o formato específico de arquivos que contenham registros de navegação.

Visando superar esses incovenintes e limitações na definição dos elementos multidimensionais, (HERNÁNDEZ et al., 2011) propõe a criação de um *framework* dirigido à

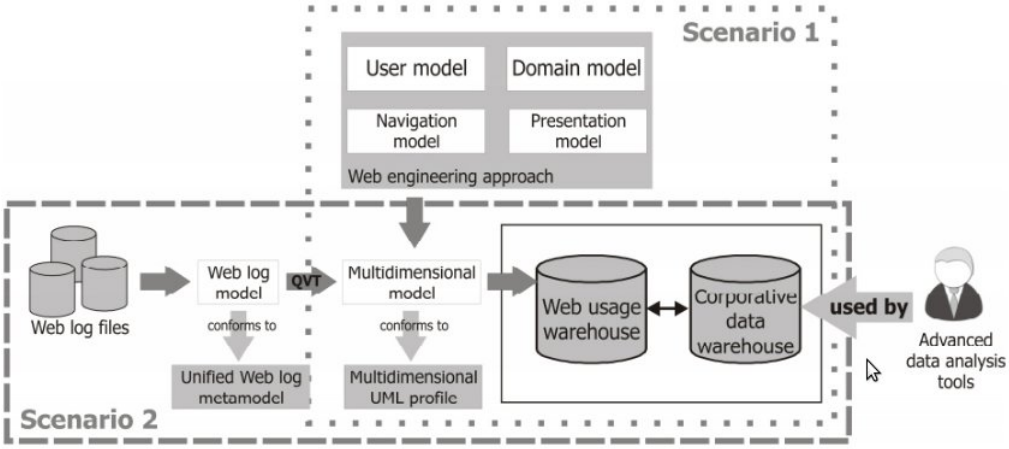

Figura 8 – *Framework* dirigido à modelos para desenvolvimento de um *data webhouse*.

Fonte: (HERNÁNDEZ et al., 2011)

modelos para desenvolvimento de um *data dawarehouse* voltado para web, considerando dois cenários.

O primeiro cenário, como ilustra a Figura 8, envolve o uso de um *data webhouse* dentro do modelo orientado a engenharia na web. Vários modelos conceituais são definidos referentes ao projeto de um site (modelo de dados, modelo de navegação, modelo do usuário, entre outros). Os conceitos multidimensionais (fatos, dimensões, hierarquias) devem ser descobertos dentro do modelo conceitual, a fim de construir um repositório de dados bem estruturado.

Devido ao fato desses modelos as vezes estarem inacessíveis ou desatualizados, é proposto um segundo cenário, que envolve o uso de um *data webhouse* que usa dados de registros na web. Esse *data webhouse* não faz uso dos modelos conceituais, e sim, dos arquivos com registros de navegação. Para isso foi desenvolvido um metamodelo de registro da web que contém elementos semânticos que permitam construir um modelo conceitual a partir arquivos de *log*, o que representa, de forma estática, a interação entre os elementos de dados brutos (ou seja, a localização do usuário remoto no site) e conceitos de uso (usuário da sessão).

O trabalho proposto por (HERNÁNDEZ et al., 2011), em linhas gerais, apresenta uma ferramenta de apoio à tomadas de decisões estratégicas, que reconhece o comportamento dos usuários e armazena o histórico de navegação em um DW, para fins de descoberta de conhecimento. Semelhante ao que é realizado nesse trabalho, o foco além de gerar análises de interesse marcadológico, é satisfazer os clientes, que no contexto da rede social, são usuários que podem interagir e trocar suas experiências de consumo para tomarem boas decisões de compra.

3.2 Classificação de usuários na Rede Social Acadêmica Scientia.Net

O Scientia.Net é uma rede social, com o objetivo de agregar aos seus membros itens de relevância acadêmica relacionados ao seu perfil (MACHADO; LIMA; ARAÚJO, 2012). Através dessa ferramenta, é proposta a construção de um mecanismo que classifica os usuários de acordo com seu perfil acadêmico, permitindo a comunicação entre pesquisadores de áreas de interesse semelhantes.

Neste trabalho foi apresentado um estudo comparativo entre três métodos de aprendizado não supervisionado: *Rede de Kohonen, Cobweb e K-means*.

Os experimentos foram realizados utilizando 2000 usuários e 20 áreas distintas de conhecimento. A geração dos dados se deu através da ferramenta disponível no site generatedata.com, sendo convidadas vinte pessoas de vinte áreas diferentes para auxiliar na geração desses dados. Dessa forma, cada pessoa gerou um total de 200 usuários fictícios, simulando o cadastro de usuários dentro do Scientia.Net.

Os algoritmos propostos para a tarefa de classificação foram implementados utilizando o *framework* WEKA, com exceção do algoritmo Rede de Kohonen, que foi implementado separadamente. Foram analisados a taxa de acerto (em porcentagem) e o tempo de execução destes algoritmos. A taxa de acerto foi verificada observando a homogeneidade dos grupos gerados pelos algoritmos.

A proposta do Sciencia.Net é semelhante ao estudo resalizado neste TCC. Ambos os trabalhos utilizam algoritmos de aprendizado não supervisionado para realizar a classificação de usuários e fazem uso da ferramenta WEKA para executar seus experimentos. Por outro lado, os algoritmos são usados para fins diferentes. No Sciencia.Net os usuários são classificados de acordo com o perfil acadêmico, permitindo-lhes a comunicação com outros alunos e pesquisadores de áreas de pesquisa semelhante, enquanto que neste trabalho o agrupamento está relacionado ao perfil de consumo dos usuários.

4 *Data Warehouse* para Extração de Dados da Web

Baseado em um ambiente clássico de DW, a arquitetura proposta para este trabalho tem por objetivo demonstrar a possibilidade de explorar dados oriundos da web. De modo geral, ela serve para armazenar e visualizar dados de ofertas de consumo (compra), extraídos através dos próprios usuários. Por meio desses dados, é possível gerar análises estratégicas de apoio a tomada de decisão, podendo ser utilizado para diversos fins de mercado. Para este trabalho, é proposta uma aplicação para descoberta de perfis de consumo, onde os usuários são agrupados em comunidades, de acordo com seus interesses de compra.

A arquitetura terá como pano de fundo uma rede social de consumo, que intermedia a interação do usuário com os dados armazenados no DW. Dessa forma, o usuário pode tanto alimentar o DW quanto consumir os dados armazenados, utilizando recursos da própria rede social. Nas próximas seções, os componentes da arquitetura são apresentados de forma detalhada, juntamente com os módulos específicos que foram implementados.

4.1 Fontes de Dados

Os dados de interesse são aqueles relacionados a ofertas de produtos (compras). Esses dados encontram-se distribuídos em diversas fontes externas, que podem ser: bases

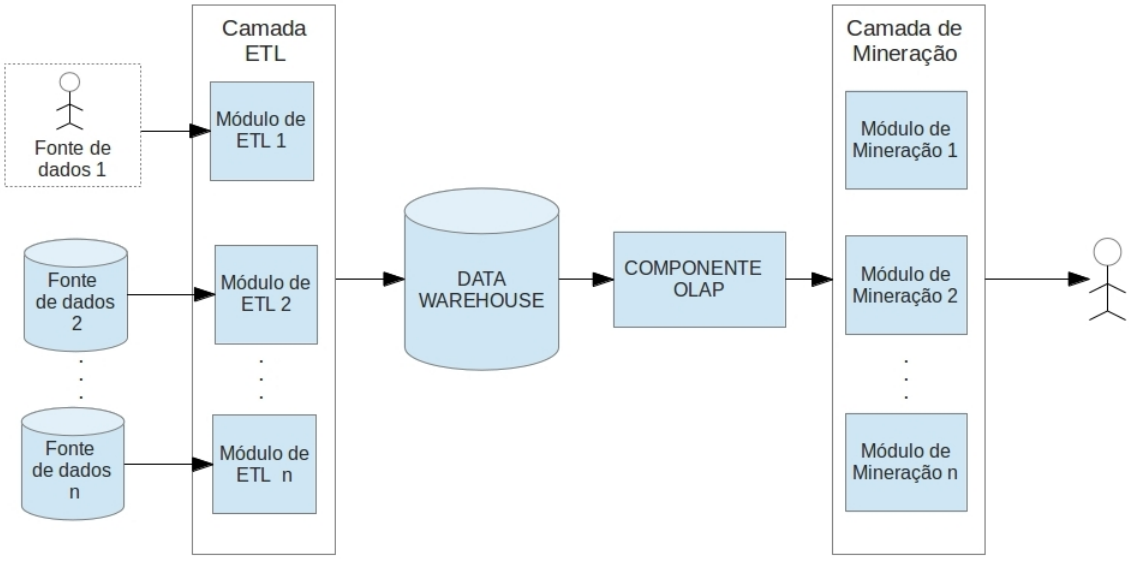

Figura 9 – Arquitetura de DW proposta para descoberta de perfis de consumo.

de dados operacionais; planilhas eletrônicas; documentos XML; sites de vendas pelas internet; através dos usuários.

A arquitetura proposta permite recuperar dados a partir de diferentes fontes, como está ilustrado na Figura 9. Uma possibilidade envolve a extração de dados por meio dos próprios usuários, conforme ilustrado no retângulo tracejado da Figura 9, indicado pela atribuição: "Fonte de dados 1".

4.2 Camada ETL

A carga das fontes de dados para o DW é realizada através da camada de ETL (Extração, Transformação e Carga). Essa camada é composta por módulos ETL, onde cada módulo é responsável pelo carregamento de dados de um tipo específico de fonte. Em ambientes DW convencionais, essas fontes normalmente são bases de dados operacionais, enquanto que a proposta deste trabalho é uma arquitetura que possui n tipos de módulos, permitindo que informações sejam extraídas a partir de fontes heterogêneas, que podem ser: arquivos, planilhas, bases operacionais, XML e através de dados oriundos da web.

Motivado pelo crescimento do comércio eletrônico e pela busca dos usuários por produtos e informações que apoiem nas decisões de compra através da internet, destaca-se neste trabalho a possibilidade de extrair dados diretamente dos usuários, e não de uma fonte de dados concreta já existente. Através desta abordagem, usuários fornecem suas informações de consumo através de um formulário de registro de compra, disponível no ambiente de uma rede social. Essa prática, também chamada de *web mining*(RUSSELL, 2011), tem como intuito coletar dados provenientes da web e armazená-los em um ambiente integrado para (neste caso), gerar análises de consumo, sendo útil para usuários e segmentos de mercado.

Através da Figura 10 é possível visualizar a tela inicial do protótipo de rede social desenvolvida para este trabalho. A aplicação tem como principal objetivo demonstrar o formato de extração de dados através dos usuários e a visualização das comunidades geradas pelos algoritmos de mineração. No contexto da camada de ETL, o formulário de registro de compra (Figura 11), conforme já citado, é o mecanismo escolhido para os usuários disponibilizarem seus históricos de compras, o mesmo estando disponível através da rede social.

Para preencher o formulário, os seguintes campos são fornecidos: Produto comprado, categoria do produto comprado, período da compra, loja onde foi feita a compra e preço do produto comprado. Além de ser uma mecanismo para extração, o formulário é utilizado como um recurso para organização financeira, pois após preencher e cadastrar suas compras, elas estarão visíveis em uma calendário histórico de consumo, no ambiente da rede social.

Figura 10 – Tela Inicial do protótipo da Rede Social de Consumo, chamada My Tag.

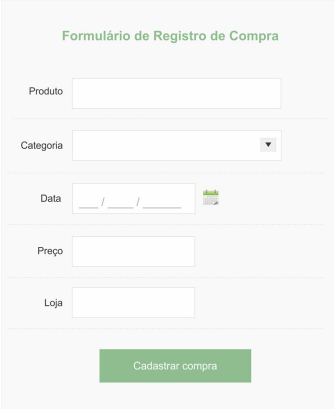

Figura 11 – Formulário de registro de compra, disponível no ambiente da rede social.

4.3 Data Warehouse

O DW é o componente responsável por armazenar os dados de consumo fornecidos pelos usuários através do Módulo ETL. No contexto deste trabalho, o esquema multidimensional definido foi o floco de neve *snowflake*, pois atende as necessidades estruturais e de hierarquia identificadas no projeto. Através da Figura 12 é possível visualizar a estrutura definida para armazenar dados de consumo.

O modelo multidimensional adotado é composto por uma tabela central (fato), chamada oferta, onde são armazenados os registros de compras extraídos através do formulário disponível na rede social. A tabela oferta possui como atributos os valores de cada produto armazenado e identificados para as tabelas dimensão.

Nas tabelas dimensão, são armazenados dados referentes a cada produto adquirido pelos usuários, sendo eles: produto comprado, local da compra, data da compra e usuário que adquiriu o produto. Observe que a dimensão usuário só será preenchida nos casos em

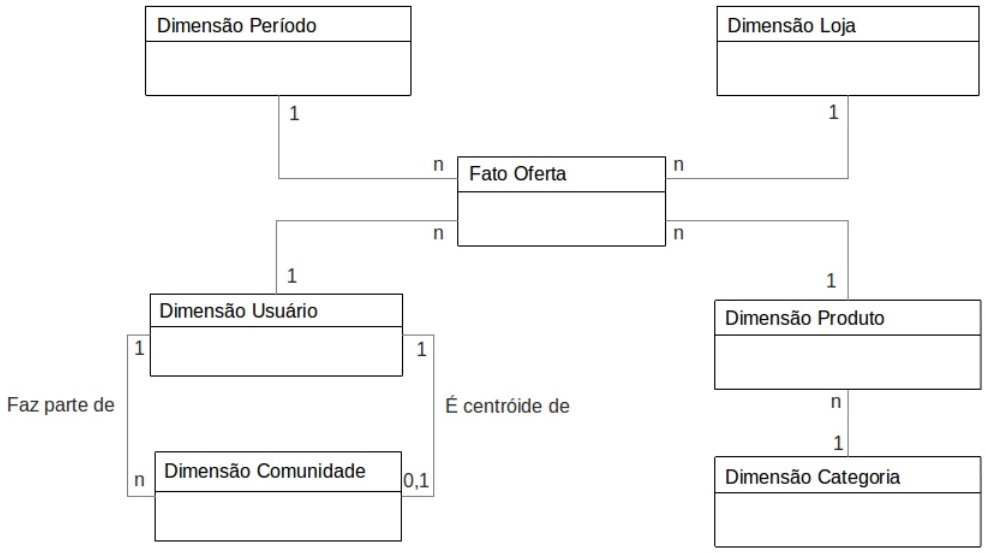

Figura 12 – Modelagem Multidimensional aplicada a arquitetura de descoberta de perfis de consumo.

que a oferta for na verdade um item já adquirido, levando em consideração que o usuário nesse caso, é a pessoa que realizou a compra.

Uma característica interessante da modelagem multidimensional constatada através da utilização do modelo floco de neve, é a eliminação da redundância dos dados quando tabelas são normalizadas. Nas dimensões "produto"e "categoria"é criado um relacionamento que identifica a categoria a que um produto comprado pertence. Já o relacionamento entre as dimensões "usuário"e "comunidades"serve para identificar a comunidade a que o usuário pertence após ser realizado o agrupamento dos dados. A adoção desse método torna a manutenção mais ágil, o modelo mais consistente e permite aumentar o nível de detalhamento entre os dados.

4.4 Componente OLAP

O componente OLAP *(On-line Analytical Processing)* é responsável pelo acesso aos dados armazenados no DW. Esse componente caracteriza-se como um conjunto de ferramentas que permite manipular as dimensões criadas pela modelagem multidimensional, visando explorar os dados relevantes para mineração.

No contexto da descoberta de perfis de consumo, são exploradas duas dimensões: usuário e produto. Essas dimensões contém dados que correspondem a todas as ocorrências individuais de compras dos usuários, informando a quantidade de compras realizadas para cada produto.

Por se tratarem de duas dimensões, se pode visualizar os dados recuperados a partir de uma matriz bidimensional. A Figura 13, ilustra um exemplo dessa matriz,

		Dimensão Categoria			
		Informática e Acessórios	Câmeras Digitais e Filmadoras	Eletrônicos	Esporte e Lazer
Dimensão Usuário	João	10	9	2	5
	Maria		3	2	1
	Alfredo	1	1	2	2
	Paulo	2		1	2

Figura 13 – Matriz bidimensional gerada pelo Módulo OLAP.

onde no eixo y está disposta a dimensão usuário, contendo todos aqueles membros da rede social que cadastraram alguma compra, e no eixo x estao dispostas todas aquelas categorias armazenadas na correspondente dimensão. Em função de a matriz conter todas as categorias e usuários da base de dados, algumas células estarão vazias, pois usuários normalmente se relacionam a um subconjunto de categorias apenas.

O foco neste trabalho é a descoberta de perfis de consumo. Entretanto, através dos dados armazenados no DW, existe uma série de possibilidades que podem ser exploradas utilizando as dimensões existentes. Além de permitir combinar as dimensões existentes sob diferentes perspectivas, existem operadores OLAP que permitem manipular os dados armazenados, permitindo delimitar com precisão aqueles dados que se deseja trabalhar, a fim de fornecer uma resposta eficiente para necessidades existentes.

4.5 Camada de Mineração de Dados

Formada por n módulos, a camada de mineração utiliza as informações de consumo disponibilizadas pelos usuários da rede social para descobrir padrões consistentes e relacionamentos entre os dados armazenados no DW. Através das visões dos dados fornecidos pelo componente OLAP e o uso de técnicas de mineração de dados, é possível gerar análises estratégicas para diversos fins.

Neste trabalho será implementado um módulo específico que utiliza o histórico de compras dos usuários para gerar comunidades com base no seu perfil de consumo, agrupando em comunidades, aqueles que possuem interesses semelhantes.

Por se tratar de um problema que envolve agrupamento de elementos, a construção do módulo de descoberta de perfis fez uso de algoritmos de aprendizado de máquina não supervisionado, devido a essa abordagem permitir agrupar coleções de elementos, com base em medidas de similaridade, sem a necessidade de um conjunto anotado de treinamento.

Dentro do aprendizado não supervisionado existem diversas abordagens que resol-

vem problemas relacionados a agrupamento de dados (SHARMA; BAJPAI; LITORIYA, 2012). Com base nisso, o objetivo aqui é realizar um estudo comparativo entre três abordagens: *K-means, Farthest-First e Expectation Maximization (EM)*, visando encontrar o método mais eficiente para agrupar os usuários.

4.6 Visualização das Comunidades

A Figura 14 ilustra a visualização das comunidades através do ambiente da rede social. Ao acessar o link de comunidades, o usuário terá acesso aos resultados obtidos pela clusterização dos dados, podendo visualizar em ordem de semelhança de perfil, as comunidades que contém os usuários mais próximos de seus interesses de consumo.

A ordem das comunidades está disposta de acordo com o perfil do usuário que estiver logado no sistema, onde a comunidade que estiver classificada no topo é aquela onde deve estar o próprio usuário e aquelas pessoas que possuem maior semelhança com seu perfil. As demais comunidades também se relacionam ao usuário, porém com uma afinidade menor. A forma de cálculo de similaridade entre um usuário e as comunidades pode ser derivada da própria forma como as comunidades são geradas. Por exemplo, supondo que seja utilizado o algoritmo k-means para formar as comunidades, a similaridade entre um usuário e as comunidades pode ser dada pela distância euclidiana entre o usuário e o centróide de cada comunidade.

Através do agrupamento de usuários é possível conhecer as pessoas que possuem perfis semelhantes. Porém, um desafio encontrado é a identificação de cada comunidade, com base nos produtos mais comprados pelos seus membros. O método encontrado para resolver esse problema foi através de uma nuvem de *tags*, contendo o nome dos produtos que mais foram adquiridos dentro de cada comunidade. Através dessa nuvem, é possível que o usuário além de reconhecer os membros do grupo, tenha conhecimento dos itens que mais interessam seus membros.

Figura 14 – Comunidades geradas com base no perfil de consumo dos usuários

5 Experimentos e Resultados Obtidos

Nesse capítulo vamos apresentar a metodologia utilizada para construção dos experimentos propostos para este trabalho, abordando o tipo de dado utilizado para avaliação, o gerador de casos de teste desenvolvido, o cenário proposto para os experimentos, e por fim, os resultados obtidos.

5.1 Gerador Automático de Conjuntos de Dados

A descoberta de conhecimento em grandes bases de dados tornou-se uma tarefa bastante comum no mercado competitivo e também para pesquisadores de mineração de dados. Apesar de existirem excelentes mecanismos para representação do conhecimento e métodos para descobrir esse conhecimento, surge um problema bastante comum: a falta de dados adequados para análise.

Em casos onde não existem dados coletados a partir de um ambiente real, e sabendo do custo alto e das limitações existentes para criar esses conjuntos manualmente, nasce a necessidade de criar uma ferramenta para apoiar essa atividade, com o objetivo de gerar dados automaticamente, que sejam adequados para serem analisados por algoritmos de descoberta de conhecimento.

A partir dessa necessidade, foi desenvolvido um gerador automático de conjunto de dados, com o objetivo de apoiar nos experimentos realizados neste trabalho. O gerador foi construído especificamente para gerar dados referentes a consumo, foco da pesquisa, garantindo que análises abrangentes fossem realizadas a partir das métricas de avaliação definidas.

A Figura 16 mostra a tela inicial da ferramenta desenvolvida. O gerador é uma ferramenta web, composta por cinco campos de entrada que permitem ao usuário especificar as seguintes informações:

- Nome do conjunto de dados;
- Número de agrupamentos;
- Afinidade Interna dos Grupos (AIG);
- Afinidade Externa dos Grupos (AEG);
- Número de usuários em cada grupo.

Figura 15 – Gerador automático de casos de teste para dados de consumo.

Entre os campos disponíveis, merecem destaque "Afinidade Externa dos Grupos (AEG)"e "Afinidade Interna dos Grupos (AIG)". No contexto da análise de consumo, A AIG corresponde ao percentual de categorias de produtos que usuários de um mesmo grupo compraram, enquanto que a AEG corresponde ao percentual de categorias de produtos que usuários de diferentes grupos compraram. Ambos os campos podem recebem valores que variam entre 0 a 100%, definindo assim, seus percentuais de afinidade.

A quantidade de categorias geradas é estabelecida a partir dos percentuais de afinidades definidas pelo usuário, onde são atribuídas a cada grupo, um número mínimo de categorias para atender aos percentuais escolhidos, independentemente da quantidade de usuários presentes em um grupo. O vetor do usuário que identifica as compras que ele realizou é preenchido com valores 1 e 0 para identificar respectivamente, àquelas categorias onde houveram ocorrências de compras e onde não houveram. A faixa de produtos que recebe o valor 1 é predeterminada. Desta faixa, o número de produtos que recebe o valor um é determinada pelo AIG. A escolha de quais produtos da faixa recebe o valor um é feita aleatoriamente.

Para demonstrar como são gerados os casos de teste, é dado um exemplo onde são definidas as seguintes informações:

- 2 grupos;
- 2 usuários por grupo;
- AIG de 75%;
- AIG de 25%.

O número mínimo de categorias para atender os percentuais de AIG e AEG, é definido a partir dos cálculos de Máximo Divisor Comum (MDC) e Mínimo Múltiplo Comum (MMC), que definem respectivamente, número de categorias mínimo para atender separadamente AIG e AEG e o número mínimo de categorias que atenda ambos os percentuais de afinidade.

Para definir as categorias de AIG, inicialmente é realizado o cálculo do MDC, tendo como resultado mdc(100,75) = 25, onde 100 é o valor de afinidade máxima e 75 a afinidade definida para o exemplo. Após calcular o MDC, é definido o total de categorias para AIG, dividindo o percentual máximo de afinidade (100) pelo resultado obtido pelo MDC, tendo (100 / 25) = 4. Já o número de categorias necessárias para atender 75% de AIG é dado pela divisão entre afinidade definida para AIG (75), dividido pelo resultado obtido no MDC, tendo 75 / 25 = 3. Dessa forma, o total de categorias necessárias para atender 75% de AIG é 3, de um total de 4 categorias.

Com a AEG é realizado o mesmo processo, inicialmente calcula-se O MDC, tendo mdc(100,25) = 25, onde 100 é o valor de afinidade máxima e 25 a afinidade definida para AEG. Após calcular o MDC, é definido o total de categorias para AEG, dividindo o percentual máximo de afinidade (100) pelo resultado obtido pelo MDC, tendo (100 / 25) = 4. Já o número de categorias necessárias para atender 25% de AEG é dado pela divisão da afinidade definida para AIG (25), pelo resultado obtido no MDC, tendo 25 / 25 = 1. Dessa forma, o total de categorias necessárias para atender 25% de AEG é 1, de um total de 4 categorias.

Após definir as quantidades de categorias necessárias para atender separadamente AIG e AEG, é possível definir a quantidade de categorias necessárias para atender os dois parâmetros simultaneamente. Essa quantidade é conhecida através do cálculo do Mínimo Múltiplo Comum (MMC), utilizando os totais de categorias para atender AIG (4) e AEG (4). Dessa forma, fazendo mmc(4,4) é obtido o valor 4 como quantidade que indicará o número de categorias mínimas necessárias para atender os 75% de AIG e 25% de AEG.

A Figura 16 ilustra o vetor de compras de cada usuário para o caso de teste do exemplo citado anteriormente. Na imagem existem duas *features*. A *feature* Fc1 indica aquelas categorias adquiridas pelos usuários do *cluster* 1, enquanto que Fc2 representa as categorias do segundo *cluster*. As categorias que foram adquiridas pelos usuários são representadas pelos círculos preenchidos enquanto que aquelas categorias não adquiridas, são representadas pelos círculos vazios.

Em cada *cluster* existe um usuário modelo (Upc1 e Upc2), que demonstra a aquisição de 100% de AIG na *feature* correspondente ao seu *cluster* e 0% de AEG para a *feature* correspondente ao *cluster* que não está associado.

Através do cálculo mostrado anteriormente, vemos que os usuários U1 e U2, que

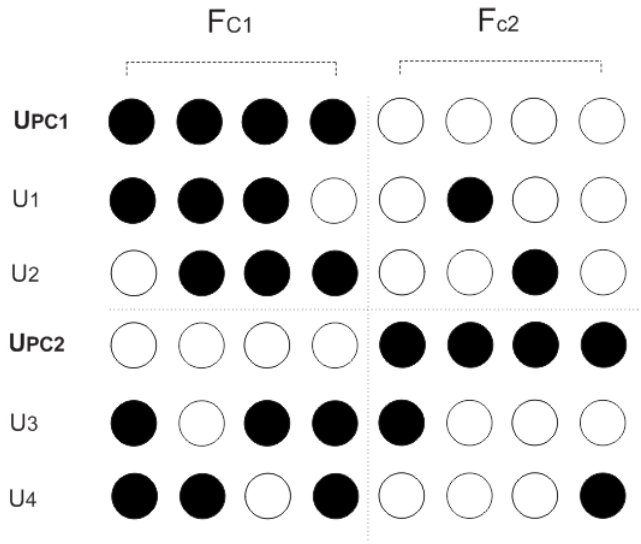

Figura 16 – Exemplo do vetor de usuários gerado pelo gerador de casos de teste

pertencem ao *cluster* 1, são aqueles que adquiriram 75% das categoria dentro da Fc1, tendo 3 categorias preenchidas (compras realizadas), e 1 vazia, onde não foram comprados produtos de nenhuma categoria. Já nas categorias pertencentes a Fc2, vemos que existe apenas 1 categoria que foi comprada, o que representa os 25% de AEG. O mesmo ocorre para os usuários U3 e U4, pertencentes ao *cluster* 2, que neste caso, possuem seus 75% de categorias compradas dentro da Fc2, e 25% concentrado na Fc1, representado respectivamente 3 e 1 categorias adquiridas.

5.2 Experimentos

Um dos objetivos deste trabalho é realizar um estudo comparativo entre algoritmos de aprendizado de máquina não supevisionado, no contexto de uma rede social de consumo. Essa análise irá indicar a abordagem mais eficiente para agrupar os usuários em comunidades, com base nos perfis de compra, tendo como medida os produtos registrados pelos usuários.

Conforme descrito na seção 4.5, foram definidos três abordagens não supervisionadas para avaliação: *K-means, Farthest First e Expectation Maximization (EM).*

Os três métodos utilizados neste trabalho foram implementado pelas bibliotecas da ferramenta de mineração de dados WEKA. O WEKA - *Waikato Environment for Knowledge* (SHARMA; BAJPAI; LITORIYA, 2012) é um conjunto de bibliotecas para descoberta de conhecimento, que possui uma série de algoritmos de aprendizado de máquina, mineração de dados, e validação de resultados. Foi desenvolvido na Universidade de Waikato na Nova Zelândia, sendo um software livre, implementado na linguagem JAVA e com grande apelo no meio acadêmico e comercial.

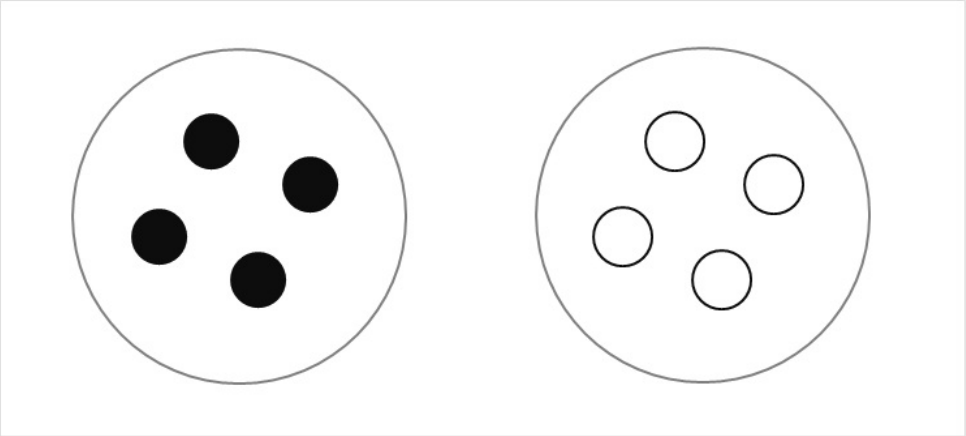

Figura 17 – Grupos antes de executar algoritmo de clusterização.

5.3 Medidas Utilizadas

As medidas de precisão (P), Cobertura (C) e *F-Measure* (F) foram inicialmente propostas para avaliar sistemas de recuperação de informação. No contexto deste trabalho, elas são utilizadas para expressar o quão satisfatórios foram os agrupamentos gerados pelos métodos de clusterização utilizados. A precisão representa a quantidade de agrupamentos corretamente recuperados dentre os agrupamentos recuperados. A cobertura representa a quantidade de agrupamentos corretamente recuperados dentre os agrupamentos corretos gerados pelo caso de teste. Já a *F-Measure* representa a média harmônica entre P e C, equilibrando os valores obtidos entre as duas métricas.

Para exemplificar como funciona o cálculo dessas métricas, considere dois grupos inicialmente definidos cada um com 4 usuários, conforme ilustra Figura 17. Os usuários pertencentes a cada grupo são representados por círculos vazios e preenchidos, onde os usuários do grupo 1 (G1) são identificados por círculos preenchidos e os do grupo 2 (G2), por círculos vazios.

Após a execução do algoritmo de agrupamento, foram gerados 2 novos grupos, visualizados através da Figura 18, onde na nova configuração o grupo 1 contém 2 usuários e o grupo 2, contém 6 usuários.

Com base no exemplo citado, a precisão levará em consideração àqueles usuários de círculos preenchidos e vazios que se mantiveram juntos após a execução do algoritmo de agrupamento(8), dividindo pelo número de agrupamentos recuperados (15), tendo com resultado P = 8 / 15 = 0,53. A cobertura levará em consideração àqueles usuários de círculos vazios ou preenchidos que se manteram juntos após a execução do algoritmo (8), dividindo pelo número de agrupamentos gerados pela configuração correta, conforme apresentado na figura Figura 18(12), tendo com resultado C = 8 / 12 = 0,66. Já a *F-Measure* representa a média harmônica entre P e C, equilibrando os valores obtidos entre

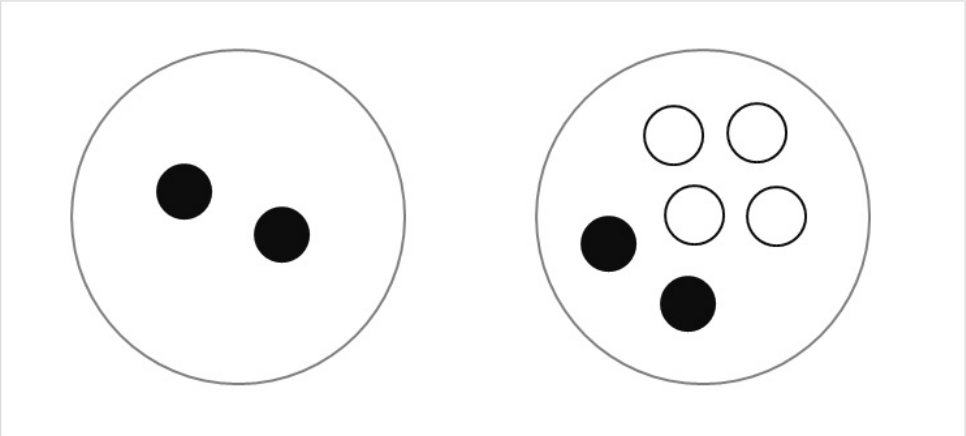

Figura 18 – Grupos gerados após a execução do algoritmo de clusterização.

as outras duas métricas. O Cálculo se dá através da seguinte fórmula: F = (2 * C + P) / (C + P). Obtendo como resultado F = (2*0,53 + 0,66)/(0,53 + 0,66) = 0,611

5.4 Definição dos casos de teste

Foram definidos casos de teste utilizando dois, cinco e dez grupos, cada um contendo cinco e cem usuários. O percentual de Afinidade Interna dos Grupos (AIG) foi definido com o valor fixo de 70%. Enquanto que para a Afinidade Externa dos Grupos (AEG), foi definido um intervalo de percentuais, representado pelo conjunto $CAEG = 10, 20, 30, 40, 50, 60, 70$. Para todas as configurações foram realizadas 10 execuções, visando garantir a acurácia dos dados.

Os casos foram construídos com o objetivo de avaliar o comportamento e a variação dos algoritmos de acordo com a alteração nas afinidades. Inicialmente temos grupos bem definidos, com uma AIG superior a AEG, porém, no decorrer de cada experimento, as afinidades vão se aproximando, criando alterações nos agrupamentos devido a alteração nos perfis dos usuários. A partir dessas variações, podemos aplicar métricas de medição para identificar a eficiência dos algoritmos de agrupamento, comparando os resultados obtidos com os agrupamentos gerados em um ambiente controlado.

5.5 Resultados

Nesta seção são apresentados os resultados obtidos através da análise comparativa entre as abordagens de agrupamento definidas: *K-means, Expectation Maximization* e *Farthest First*.

Conforme visto na 5.4, os experimentos foram criados utilizando o gerador de casos de teste, construído para este trabalho para gerar dados de consumo. O arquivos gerados

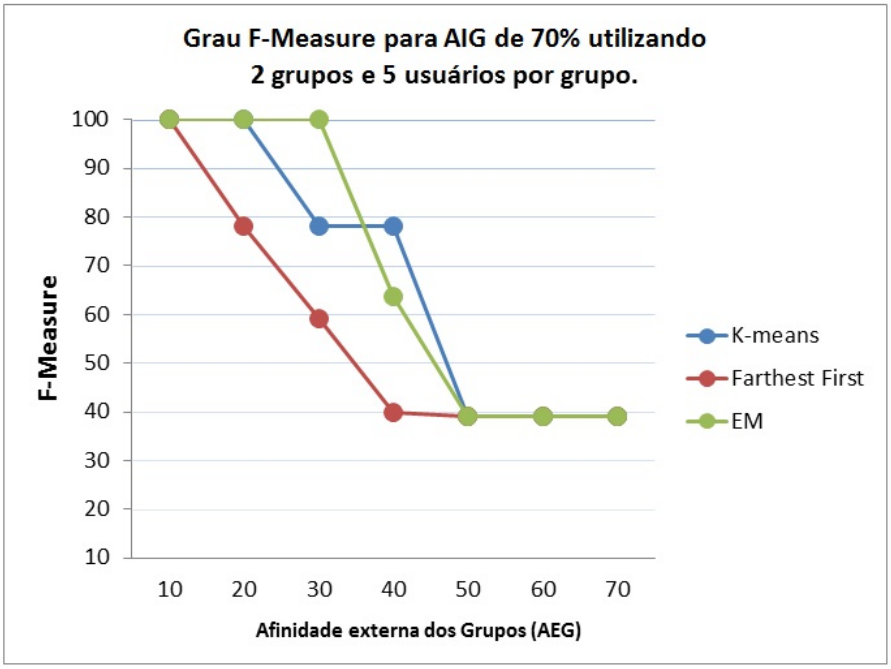

Figura 19 – Percentual de *F-Measure* para AIG de 70%, utilizando 2 grupos e 5 usuários por grupo.

pela ferramenta, são do tipo ARFF, utilizados pelo *framework* WEKA.

Os algoritmos utilizados possuem comportamento estático, ou seja, na entrada dos dados é definida a quantidade de grupos que se deseja obter. Dessa forma, os resultados foram divididos em três cenários: cenário com 2 grupos; cenário com 5 grupos; cenário com 10 grupos. A seguir são apresentados os resultados obtidos.

Cenário 1 – Dois grupos

No primeiro cenário, são apresentados os resultados obtidos através de duas configurações. A primeira, com 5 usuários (Figura 19) e a segunda com 100 usuários (Figura 20).

Criando um comparativo entre as duas configurações, vemos que existe uma diferença bastante significativa no comportamento das técnicas utilizadas. Na configuração com 5 usuários, existe uma queda acentuada no percentual de *F-Measure*, indicando uma desorganização entre os agrupamentos gerados. A exceção é o algoritmo EM, que mantém os usuários corretamente agrupados com até 30% de Afinidade Externa dos Grupos (AEG), enquanto que as demais técnicas, possuem um comportamento inferior, mesmo com os perfis dos usuários bem definidos.

Na segunda configuração, quando existe um número maior de usuários (100), pode-se visualizar que os grupos se mantem parcialmente organizados até o final das execuções, chegando no valor mínimo de 50% de *F-Measure*, mantendo desempenho regular no agrupamento dos usuários. Através dessa configuração, os algoritmos EM e *K-Means*, se

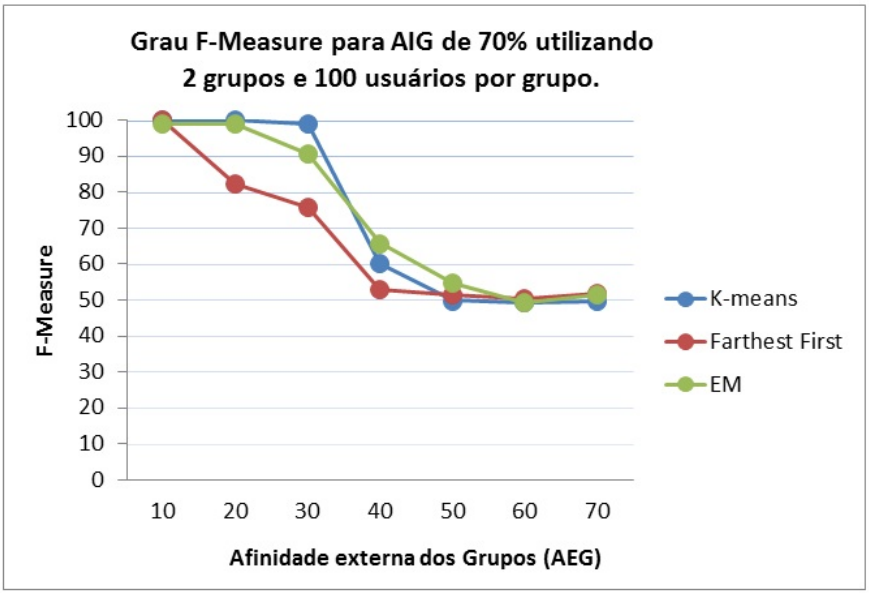

Figura 20 – Percentual de *F-Measure* para AIG de 70%, utilizando 2 grupos e 100 usuários por grupo.

equiparam, mantendo um bom comportamento, enquanto que o *Farthest First*, se mostra pior.

Cenário 2 – Cinco grupos

Igualmente ao primeiro cenário, são apresentados os resultados obtidos através de duas configurações. A primeira, com 5 usuários (Figura 21) e a segunda com 100 usuários (Figura 22), com 5 grupos.

Analisando as duas configurações, novamente pode-se ver observar um baixo rendimento das abordagens com um número menor de usuários e maior número de grupos. Com 5 usuários, observou-se uma queda a partir de 10% de AEG, ou seja, com perfis ainda bem definidos. EM e *Farthest First* se equiparam desta vez. Destaque para o *K-Means*, que parece ser mais sensível a cenários com poucos dados.

Com 100 usuários, vê-se que o algoritmo EM possui um desempenho melhor que os demais, mantendo 100% de *F-Measure* até 20% de AEG. Com baixa AEG, as três técnicas se equiparam, ficando entre 20 e 30% *F-Measure*.

Em comparação ao cenário anterior, quando se utilizou 100 usuários, pode se ver que uma quantidade maior de grupos interfere no resultado, alterando de forma significativa o agrupamento dos usuários. Enquanto que com 2 grupos, a *F-Measure* ficou com mínimo de 50% para todas as abordagens, no cenário atual, caiu para quase 20%.

Cenário 3 – Dez grupos

Neste terceiro e último cenário são apresentados os resultados obtidos através de duas configurações. A primeira, com 5 usuários (Figura 23), e a segunda com 100 usuários

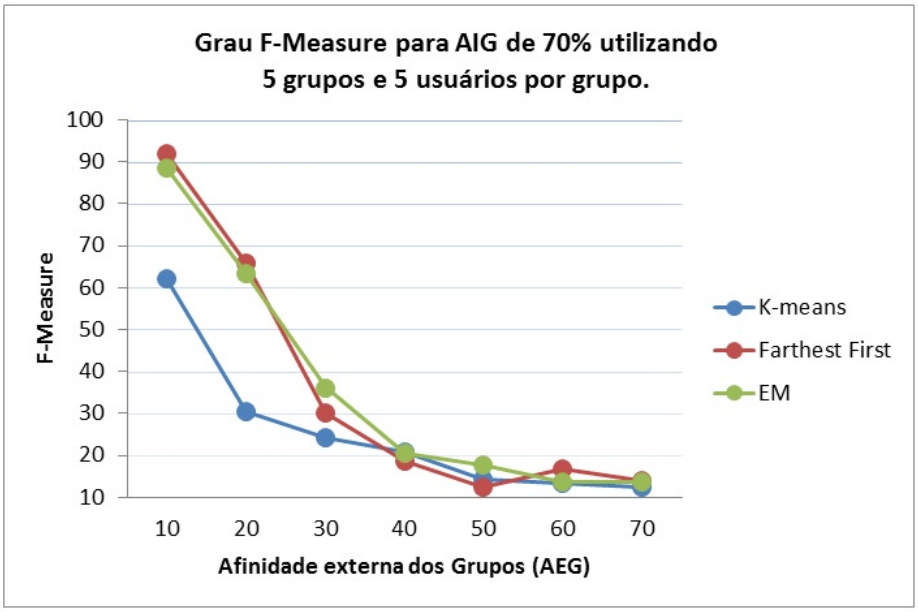

Figura 21 – Percentual de *F-Measure* para AIG de 70%, utilizando 5 grupos e 100 usuários por grupo.

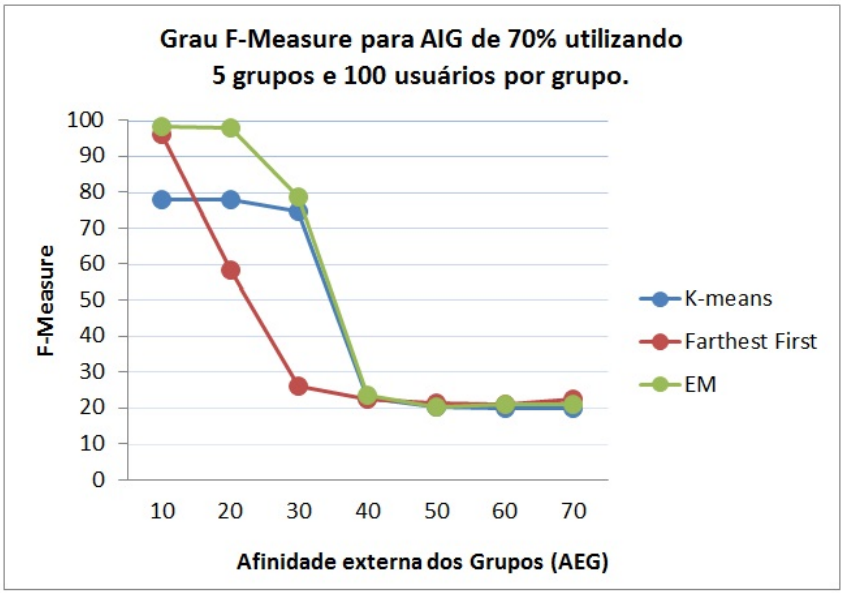

Figura 22 – Percentual de *F-Measure* para AIG de 70%, utilizando 5 grupos e 100 usuários por grupo.

(Figura 24), agora com 10 grupos.

Confirmando o que foi observado nos cenários anteriores, o comportamento das abordagens quando se utiliza 5 usuários é bastante baixa. Isso ocorre devido as alterações que são realizadas nos grupos, mesmo quando os perfis são semelhantes, pois poucos usuários levam a distorções que mudam consideravelmente a configuração dos agrupamentos.

Através da configuração com 5 usuários neste cenário, vemos exatamente isso.

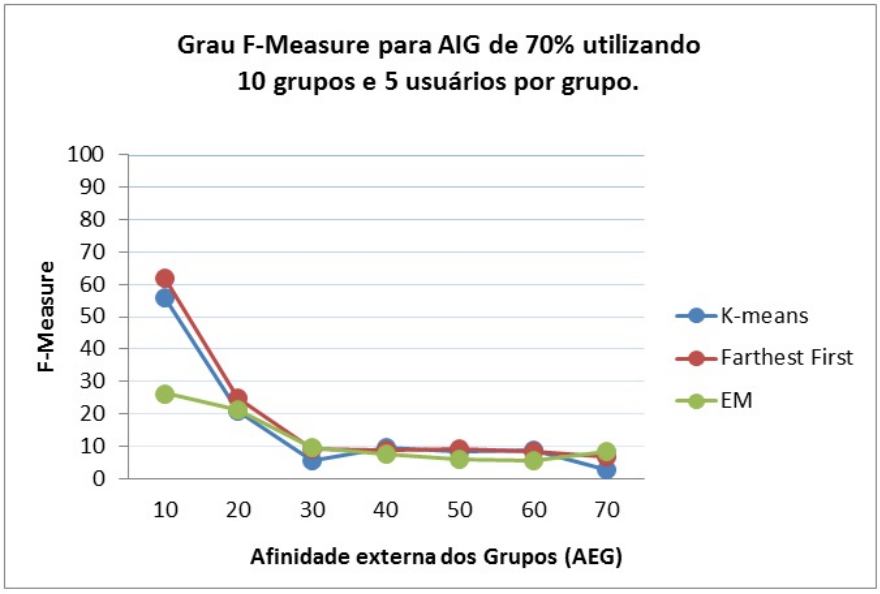

Figura 23 – Percentual de *F-Measure* para AIG de 70%, utilizando 10 grupos e 5 usuários por grupo.

Com usuários ainda com perfis bem definidos, as abordagens EM E *Farthest First* ficam com apenas 60% de *F-Measure*, o que se agrava no decorrer da configuração, ficando abaixo de 10% a partir de 30% de AEG.

Novamente destacamos o algoritmo *K-means*, que com poucos usuários, assim como no cenário anterior, apresenta rendimento abaixo das demais abordagens.

Com 100 usuários o desempenho das três técnicas cai se comparada s com o cenário com 5 grupos e 100 usuários. De 10 a 20% de AGE ainda os grupos se mantém organizados, porém, após ultrapassar essa afinidade, eles caem para 10%, e por lá se equiparam até o final - comportamento semelhante ao Cenário 2. Nessa configuração destacamos o algoritmo EM, que obteve um desempenho melhor , enquanto que o *Farthest First* obteve o pior resultado.

Foi observado através dos resultados que, conforme o número de grupo cresce e o número de usuários decresce, ocorre uma desorganização maior entre os grupos, enquanto que com um número maior de usuários, são apresentados melhores resultados. Esse comportamento aconteceu mesmo quando foram usados 10 grupos e 100 usuários, pois o número de usuários com relação ao número de grupos é pequena, essa situação ficou ainda mais clara quando foram utilizados 5 grupos e 5 usuáris por grupos.

Através dos gráficos, foi possível notar uma consideravel semelhança no comportamento dos algoritmos, em especial, entre o *K-means* e *Expectation Maximization*. Essa semelhança nos resultados dificultou a visualização do desempenho de cada algoritmo apenas pela visualização dos gráficos, pois em algumas execuções, *K-means*, EM e *Farthest First* demonstraram resultados parecidos.

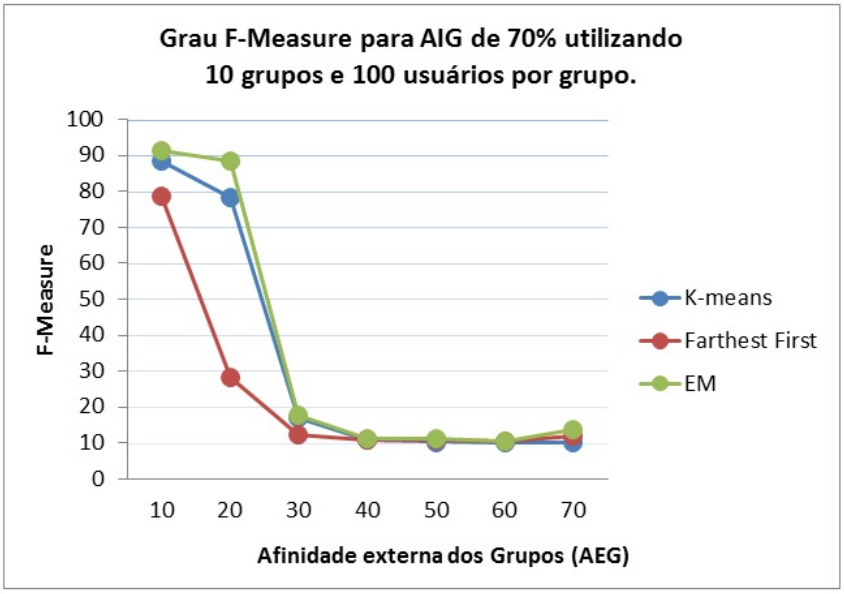

Figura 24 – Grau F-Measure para AIG* de 70% utilizando 10 grupos e 100 usuários por grupo.

	2 grupos	5 grupos	10 grupos
K-Means	67,59	25,42	15,97
EM	68,67	36,22	12,08
Farthest First	56,32	35,65	18,55

Figura 25 – Tabela com as médias de F-Measure para os três algoritmos, utilizando 5 usuários por grupo.

Uma alternativa para apoiar na visualização dos resultados, foi calcular as médias obtidas pelas técnicas para a métrica *F-measure*, identificando o desempenho de cada técnicas para cada cenário definido.

Na Figura 25, podemos visualizar a média do *F-Measure* com 5 usuários por grupo, onde as colunas representam os números de grupos utilizados nos experimentos, onde cada linha é identificada por um método. O resultado nos permite identificar o método EM possui desempenho melhor quando são utilizados configurações com 2 e 5 grupos, enquanto que o *Farthest First* possui média maior para 10 grupos.

Já na Figura 26, vemos a média do *F-Measure* utilizando 100 usuários por grupo. Com essas configurações, o algoritmo EM obteve melhor desempenho na média para os 3 cenários, com *K-Means* e *Farthes First* com resultados inferiores respectivamente.

Os cenários propostos neste trabalho tiveram com objetivo avaliar o comportamento e desempenho das técnicas de agrupamento para diferentes configurações de perfis de consumo. Inicialmente foram utilizados perfis bem comportados nos experimentos, onde foi observado o algoritmo EM tendo um melhor desempenho na maioria das configu-

	2 grupos	5 grupos	10 grupos
K-Means	72,56	44,80	32,05
EM	72,79	51,55	34,80
Farthest First	66,31	38,19	23,30

Figura 26 – Tabela com as médias de F-Measure para os três algoritmos, utilizando 100 usuários por grupo.

rações propostas, já utilizando perfis menos bem definidos, com AEG maior, observou-se um comportamento bastante semelhante entre as técnicas, onde todas obtiveram baixo de desempenho, podendo-se considerar que para este tipo de configuração, é possível utilizar quaisquer das técnicas, não havendo uma mais eficaz com relação a outra, devido a sua semelhança.

6 Conclusão

A descoberta de conhecimento é uma prática cada vez mais explorada por empresas que buscam conhecer melhor seus clientes e oferecer a eles produtos específicos, baseados no seu perfil de interesse. Neste contexto, o uso de ambientes DW permite que grandes volumes de dados de consumo, sejam analisados e manipulados de forma flexível. Para isso, são utilizadas técnicas de mineração de dados, com foco na extração de conhecimento e informação. Através dessas técnicas, os dados são transformados em informações estratégicas, que favorecem à tomada de decisão tanto de segmentos de mercado, como dos usuários.

Neste trabalho, foi demonstrada a utilização de uma arquitetura clássica de DW, voltada para dados oriundos da web, utilizando como fonte de dados os próprios usuários, com o objetivo de realizar análise de oferta de consumo. Por se basear em ambientes tradicionais de DW, a arquitetura proposta é expansível, ou seja, ela permite que os dados sejam coletados através de fontes heterogêneas. Para demonstrar sua aplicabilidade, foi criada uma aplicação no formato de uma rede social, que encapsula o acesso ao DW, oferecendo recursos para realizar a coleta de dados através de um formulário de registro de compra, disponível no ambiente da rede social. Com os dados coletados são geradas comunidades, baseadas no perfil de compra dos usuários, que permite através da rede social, uma interação para troca de experiência, sendo utilizada também para segmentação de mercado.

Para gerar as comunidades, foi realizada uma análise comparativa entre três métodos não supervisionados de agrupamento: *K-means, Expectation Maximization* e *Farthest First*, onde os casos de teste foram criados a partir de um gerador automático, desenvolvido para este trabalho. Os casos de teste gerados foram executados no WEKA, uma ferramenta de mineração de dados empregada em trabalhos acadêmicos.

Os cenários propostos neste trabalho tiveram com objetivo avaliar o comportamento e desempenho das técnicas de agrupamento para diferentes configurações de perfis de consumo. Inicialmente foram utilizados perfis bem comportados nos experimentos, onde foi observado o algoritmo EM tendo um melhor desempenho na maioria das configurações propostas, já utilizando perfis menos bem definidos, com AEG maior, observou-se um comportamento bastante semelhante entre as técnicas, onde todas obtiveram baixo de desempenho, podendo-se considerar que para este tipo de configuração, é possível utilizar quaisquer das técnicas, não havendo uma mais eficaz com relação a outra, devido a sua semelhança.

Durante o desenvolvimento deste trabalho, foram alcançados resultados relaciona-

dos a produção científica, participação em eventos e construção de uma ferramenta para geração de casos de teste:

- Prêmio Ciab FEBRABAN - Participação no Concurso Nacional realizado pela Federação Brasileira de Bancos, que propõe o incentivo a ideias inovadoras que envolvam Tecnologia e o Setor Financeiro. Na oportunidade a Mytag - Rede Social de Consumo ficou com o 3º Lugar no evento, realizado em São Paulo, em Junho de 2012;

- ERBD 2013 - Artigo publicado na IX edição da Escola Regional de Banco de Dados, realizada em Camboriú, em abril de 2013. O trabalho foi intitulado "Uma aplicação de Rede Social de Consumo Baseada em uma Arquitetura de *Data Warehouse*";

- Gerador de Conjunto de Dados - Foi desenvolvido um gerador Automático de conjunto de dados, voltado para dados de consumo, com o objetivo de apoiar nos experimentos realizados neste trabalho e para obtenção de melhores resultados na análise dos métodos de agrupamento.

Como trabalhos futuros, pretende-se aprimorar o módulo ETL existente, para que os dados fornecidos pelos usuários estejam corretos antes de serem armazenados no DW. Um dos métodos a ser estudado envolve o casamento de *String*, a fim de corrigir o nome de algum produto, local de compra ou outra informação, induzindo o usuário ao acerto. Outra melhoria, é implementação de novos módulos de ETL na arquitetura proposta, permitindo que a extração seja realizada a partir de novas fontes de dados. Esses módulos podem ser ativos, acessando fontes de dados da web, ou então passivos, sendo chamados por agentes externos.

Além do módulo ETL, pretende-se construir novos módulos de mineração de dados, com o intuito de explorar novas análises a partir dados armazenados no DW. O objetivo é gerar serviços que apoiem a tomada de decisão dos usuários, e sirvam também, como apoio a descoberta de conhecimento para empresas que buscam conhecer melhor seus clientes.

Outra possibilidade é coleta de novos dados para teste, visando obter melhores resultados e análises sobre o comportamento das abordagens de agrupamento definidos neste trabalho. Para isso, pretende-se estender a ferramenta de geração de conjunto de dados construída, tornando-a mais flexível e precisa para geração de diversos cenários. Uma proposta aqui, é permitir a geração de *outliers*, com o objetivo de construir cenários próximos do que seriam ambientes reais, explorando diversas situações. Além disso, outra proposta é investigar abordagens dinâmicas (dynamic clustering) para os algoritmos de agrupamento, avaliando seu comportamento diante da geração dinâmica dos grupos, realizando um comparativo com a geração estática.

Além do que foi proposto neste trabalho, diversas possibilidades podem ser exploradas no que diz respeito à coleta de dados e descoberta de conhecimento a partir de dados web. Atualmente a massa de informação disponível é gigantesca, e a tendência é que aumente, conforme crescem as vendas pela internet. Essas análises se tornam úteis não apenas para usuários, mas principalmente para empresas de todos os segmentos de mercado, que muitas vezes não conseguem ter controle daquilo que é dito e do que é tendência, de acordo com o que é compartilhado pelos usuários. Com isso, fica o desafio de criar novas ferramentas e formas de descobrir novos conhecimentos então ocultos nessa gigante teia de informação, que é a web.

Referências

ADELMAN, L. *Evaluating decision support and expert systems*. [S.l.]: Wiley-Interscience, 1992. Citado na página 25.

ALMEIDA, A. M. de. *Proposição de indicadores para avaliação técnica de projetos de data warehouse: um estudo de caso no data warerouse da Plataforma Lattes*. Tese (Doutorado) — Universidade Federal de Santa Catarina, 2006. Citado 2 vezes nas páginas 28 e 29.

BARRERO, D. F.; CAMACHO, D.; R-MORENO, M. D. Automatic web data extraction based on genetic algorithms and regular expressions. In: *Data Mining and Multi-agent Integration*. [S.l.]: Springer, 2009. p. 143–154. Citado na página 28.

DALVI, N.; KUMAR, R.; SOLIMAN, M. Automatic wrappers for large scale web extraction. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 4, n. 4, p. 219–230, 2011. Citado na página 28.

DASGUPTA, S. Performance guarantees for hierarchical clustering. In: SPRINGER. *Computational Learning Theory*. [S.l.], 2002. p. 351–363. Citado na página 37.

DREISEITL, S.; VINTERBO, S.; OHNO-MACHADO, L. Disambiguation data: extracting information from anonymized sources. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *Proceedings of the AMIA Symposium*. [S.l.], 2001. p. 144. Citado na página 28.

ELMASRI, R. et al. *Sistemas de banco de dados*. Pearson Addison Wesley, 2005. Disponível em: <http://books.google.com.pe/books?id=lHL7GAAACAAJ>. Citado 2 vezes nas páginas 25 e 33.

EVERITT, B.; HOTHORN, T. Cluster analysis. In: *An Introduction to Applied Multivariate Analysis with R*. [S.l.]: Springer, 2011. p. 163–200. Citado na página 35.

FAYYAD, U. M. et al. *Advances in knowledge discovery and data mining*. [S.l.: s.n.], 1996. Citado na página 34.

FERREIRA, J. et al. O processo etl em sistemas data warehouse. *II Simpósio de Informática*, p. 757–765, 2010. Citado 2 vezes nas páginas 27 e 28.

FILHO, F. M. F.; GEUS, P. L. de; ALBUQUERQUE, J. P. de. Sistemas de recomendação e interação na web social. In: *Proceedings of the 1st Workshop on Human-Computer Interaction Aspects in the Social Web, in conjunction with the VIII Brazilian Symposium of Human Factors on Computer Systems (IHC 08)*. [S.l.: s.n.], 2008. p. 24–27. Citado na página 22.

FONSECA, R. B.; SELECAO, U. E. D. A.; ALGORITMOS, D. D. Ministério da defesa exército brasileiro departamento de ciência e tecnologia instituto militar de engenharia curso de mestrado em sistemas e computacao. 2008. Citado na página 40.

FONTANA, A.; NALDI, M. C. Estudo e comparação de métodos para estimação de números de grupos em problemas de agrupamento de dados. In: ICMC. [S.l.], 2009. Citado 2 vezes nas páginas 36 e 39.

FRIEDLIN, J.; MCDONALD, C. J. Using a natural language processing system to extract and code family history data from admission reports. In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. *AMIA Annual Symposium Proceedings*. [S.l.], 2006. v. 2006, p. 925. Citado na página 28.

GIBSON, J. et al. Ncess project: Data mining for social scientists. *Proceedings of e-Social Science'07*, 2007. Citado na página 37.

HERNÁNDEZ, P. et al. Towards a model-driven framework for web usage warehouse development. *Advances in Conceptual Modeling. Recent Developments and New Directions*, Springer, p. 336–337, 2011. Citado 2 vezes nas páginas 41 e 42.

HOKAMA, D. D. B. et al. A modelagem de dados no ambiente data warehouse. *São Paulo*, 2004. Citado na página 32.

HONG, J. L. Deep web data extraction. In: IEEE. *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*. [S.l.], 2010. p. 3420–3427. Citado na página 28.

IBGE. *IBGE | Sala de imprensa | notícias | PNAD: De 2005 para 2011, número de internautas cresce 143,8% e o de pessoas com celular, 107,2%*. 2013. Disponível em: <http://saladeimprensa.ibge.gov.br/noticias?view=noticia-id=1-busca=1-idnoticia=2382>. Citado na página 21.

INMON, W. H. et al. *Building the data warehouse*. [S.l.: s.n.], 2002. Citado na página 25.

KIMBALL, R.; MERZ, R. *The data webhouse toolkit: building the web-enabled data warehouse*. [S.l.]: Wiley New York, 2000. Citado na página 29.

LAENDER, A. H.; RIBEIRO-NETO, B.; SILVA, A. S. da. Debye–data extraction by example. *Data & Knowledge Engineering*, Elsevier, v. 40, n. 2, p. 121–154, 2002. Citado na página 28.

LUNA, J. E. O. Algoritmos em para aprendizagem de redes bayesianas a partir de dados incompletos. *Universidade Federal de Mato Grosso do Sul*, 2004. Citado na página 37.

MACHADO, F. N. R. *Tecnologia e Projeto Data Warehouse*. 3. ed. [S.l.]: Sao Paulo-SP, 2010. Citado 4 vezes nas páginas 25, 30, 31 e 32.

MACHADO, V. P.; LIMA, B. V. de; ARAÚJO, S. W. Classificaç ao automática de usuários de uma rede social utilizando algoritmos nao-supervisionados. 2012. Citado na página 43.

Mannino, M. V. *Projeto, Desenvolvimento de Aplicações e Administração de Banco de Dados*. 3. ed. [S.l.]: São Paulo-SP, 2008. Citado 4 vezes nas páginas 21, 22, 25 e 34.

METZ, J. Interpretação de clusters gerados por algoritmos de clustering hierárquico. Biblioteca Digital de Teses e Dissertações da USP, 2011. Citado na página 35.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, p. 89–114, 2003. Citado na página 34.

MONTEIRO, A. V. G.; PINTO, M. P. O.; COSTA, R. M. E. M. da. Uma aplicação de data warehouse para apoiar negócios. *Cadernos do IME-Série Informática*, v. 16, p. 48–58, 2013. Citado na página 30.

MORAIS, C. A. d. S. Protótipo de sistemas de informação aplicado a administração de materiais utilizando data warehouse e conceitos de data mart. 2000. Citado na página 26.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F. O uso da mineração de textos para extração e organização não supervisionada de conhecimento. *Revista de Sistemas de Informacao da FSMA n*, v. 7, p. 7–21, 2011. Citado na página 36.

RUSSELL, M. A. *Mining the social web: Analyzing data from Facebook, Twitter, LinkedIn, and other social media sites*. [S.l.]: O'Reilly, 2011. Citado na página 46.

SHARMA, N.; BAJPAI, A.; LITORIYA, M. R. Comparison the various clustering algorithms of weka tools. *facilities*, v. 4, p. 7, 2012. Citado 2 vezes nas páginas 50 e 56.

SOARES, V. J. d. A. Modelagem incremental no ambiente de data warehouse. *Rio de Janeiro*, 1998. Citado na página 30.

SOUTO, M. de et al. Técnicas de aprendizado de máquina para problemas de biologia molecular. *III Jornada de Inteligência Artificial*, 2003. Citado na página 35.

TURBAN, E. et al. *Business Intelligence: Um enfoque gerencial para a inteligência do negócio*. Bookman, 2009. ISBN 9788577804252. Disponível em: <http://books.google-.com.br/books?id=_Uvqyr32hlMC>. Citado na página 34.

WAGNER, C. A. et al. Estudo para implantação de um data warehouse em um ambiente empresarial. Florianópolis, SC, 2012. Citado na página 31.

Índice