UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

JOSÉ CARLOS ALMEIDA PATRÍCIO JÚNIOR

Mining Knowledge TV:
Uma Abordagem de Ambiente de KDD com Ênfase em Mineração de Dados no Ambiente da *Knowledge TV*

JOÃO PESSOA
2012

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

JOSÉ CARLOS ALMEIDA PATRÍCIO JÚNIOR

Mining Knowledge TV: Uma Abordagem de Ambiente de KDD com Ênfase em Mineração de Dados no Ambiente da *Knowledge TV*

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal da Paraíba, como requisito parcial para a obtenção do título Mestre em Informática (Sistemas de Computação).

Linha de Pesquisa: Computação Distribuída

Orientadora: Prof. Dra. Natasha Correia Queiroz Lino.

JOÃO PESSOA

2012

P314m Patrício Júnior, José Carlos Almeida.

Mining Knowledge *TV: uma abordagem de ambiente de KDD com ênfase em mineração de dados no ambiente da* knowledge *TV / José Carlos Almeida Patrício Júnior. - - João Pessoa: [s.n.], 2012.*

117 f. : il.

Orientadora: Natasha Correia Queiroz Lino.
Dissertação (Mestrado) – UFPB/CCEN.

1.Informática. 2.MKTV. 3.TV Digital Interativa. 4.Mineração de dados. 5. Ginga. 6. TV Semântica.

UFPB/BC *CDU: 004(043)*

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **JOSÉ CARLOS A PATRÍCIO JÚNIOR**, candidato ao Título de Mestre em Informática na Área de Sistemas de Computação, realizada em 17 de maio de 2012.

Ao décimo sétimo dia do mês de maio do ano dois mil e doze, às dezessete horas, no auditório do CCEN - da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para examinar o candidato ao grau de Mestre em Informática, na área de "*Sistemas de Computação*", na linha de pesquisa "*Computação Distribuída*", o Sr. **JOSÉ CARLOS A PATRÍCIO JÚNIOR**. A comissão examinadora foi composta pelos professores doutores: NATASHA CORREIA QUEIROZ LINO (PPGI-UFPB), Orientadora e Presidente da Banca Examinadora, CLAUIRTON DE ALBUQUERQUE SIEBRA (PPGI-UFPB) e GUIDO LEMOS DE SOUZA FILHO (PPGI-UFPB), como examinadores internos e VICENTE FERREIRA DE LUCENA JUNIOR (UFAM), como examinador externo. Dando início aos trabalhos, a professora NATASHA CORREIA QUEIROZ LINO, cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra ao candidato para que o mesmo fizesse, oralmente, a exposição do trabalho de dissertação intitulado "UMA ABORDAGEM DE AMBIENTE DE KDD COM ÊNFASE EM MINERAÇÃO DE DADOS NO AMBIENTE DA KNOWLEDGE TV". Concluída a exposição, o candidato foi argüido pela Banca Examinadora que emitiu o seguinte parecer: "*aprovado*". Assim sendo, deve a Universidade Federal da Paraíba expedir o respectivo diploma de Mestre em Informática na forma da lei e, para constar, eu, professor Alisson Vasconcelos de Brito, Vice-coordenador deste Programa, servindo de secretário, lavrei a presente ata que vai assinada por mim mesmo e pelos membros da Banca Examinadora. João Pessoa, 17 de maio de 2012.

Alisson Vasconcelos de Brito

Profa. Dra. Natasha Correia Queiroz Lino
Orientadora (PPGI-UFPB)

Prof. Dr. Clauirton de Albuquerque Siebra
Examinador Interno (PPGI-UFPB)

Prof. Dr. Guido Lemos de Souza Filho
Examinador Interno (PPGI-UFPB)

Prof. Dr. Vicente Ferreira de Lucena Junior
Examinador Externo (UFAM)

Dedico este trabalho aos meus pais, Carlos e Selma, por todo o amor, atenção e ensinamentos passados em minha vida.

AGRADECIMENTO

Eu gostaria de agradecer primeiramente a Deus, pelo dom da vida e pela força para conseguir vencer todos os obstáculos.

A meus pais, Carlos e Selma, pela compreensão em todos os momentos.

À minha orientadora, Natasha Queiroz, por todo o auxílio dado durante toda essa jornada no mestrado.

À minha amada Clara, pela paciência, cumplicidade, companheirismo e apoio em todos os momentos.

A meus irmãos queridos e aos amigos do Encena, Enijesp, Assepe pelo incentivo para vencer barreiras.

Aos companheiros do projeto Knowledge TV, em especial a Jonatas, Ramon, Manoel, Omar e Marcelle pelas sugestões e críticas que, certamente, tornaram mais ricas as contribuições obtidas neste trabalho.

A Capes, pela bolsa de estudos que me permitiu ter mais tranquilidade para desenvolver este projeto.

RESUMO

A TV Digital Interativa Interativa traz inúmeras inovações ao cenário analógico existente, como aumento da qualidade de som, imagem, quantidade de canais, programas e serviços disponíveis ao usuário. Contudo, as informações que representam o conteúdo multimídia se apresentam estruturadas de forma tradicional, em formato de tabelas e arquivos, não provendo interoperabilidade entre as informações, nem possuindo semântica. Desta maneira, o projeto *Knowledge TV - KTV* propõe organizar os dados presentes na TV Digital Interativa semanticamente, utilizando para isso conceitos da WEB Semântica e representação de conhecimento, além de fornecer uma arquitetura para desenvolvimento de aplicações. Um dos principais componentes do KTV é o ambiente de descoberta de conhecimento em base de dados apoiado por conceitos semânticos. Este ambiente tem por objetivo descobrir conhecimentos úteis nos dados, através da Mineração de Dados, organizá-los semanticamente e disponibilizá-los como um serviço. Neste contexto, este trabalho visa especificar e desenvolver esse ambiente que nesta abordagem será chamado de *Mining Knowledge TV* – MKTV.

Palavras-chave: *Mining Knowledge TV*; MKTV; Knowledge TV; KTV; TV Digital Interativa Interativa; Mineração de Dados; Ginga; TV Semântica.

ABSTRACT

The Interactive Digital TV brings many innovations to the existing analog scenario, such as improved sound quality, image, number of channels, programs and services available to the user. However, the information representing the structured multimedia content are presented in a traditional format of tables and files, not providing interoperability between the information, nor having semantics. This way, the project Knowledge TV - KTV proposes to organize the data in the DTV semantically, using this concept of Semantic Web and knowledge representation, and provides an architecture for application development. A major component of the KTV is the environment of knowledge discovery in databases supported by semantic concepts. This environment aims to discover useful knowledge in the data through data mining, semantically organize them and make them available as a service. In this context, this work aims to specify and develop this environment that this approach will be called the Mining Knowledge TV - MKTV.

Keywords: Mining Knowledge TV; MKTV, Knowledge TV, KTV, Interactive Digital TV, Data Mining, Ginga, TV Semantic.

LISTA DE TABELAS

LISTA DE SIGLAS

Sigla	Significado
ACIU	Agente de Captura de Interação do Usuário
ACIPC	Agente de Captura de Informação do Provedor de Conteúdo
API	Aplication Program Interface
ARIB	Association of Radio Industries and Businesses
ATSC	Advanced Television Systems Committee
DVB	Digital Video Broadcast
DM	Data Mining
DMOP	Data Mining Optmization
DW	Data Warehouse
ETL	Extract Transformation and Load
EPG	Electronic Programming Guide
EIT	Event Information Table
HDTV	High Definition TV
ISDB	Integrated Services Digital Broadcasting
IA	Inteligência Artificial
JVM	Java Virtual Machine
KAAS	Knowledge As A Service
KDD	Knowledge Discovery in Database
KTV	Knowledge TV
MKTV	Mining Knowledge TV

NCL	Nested Context Language
NTSC	National Television System Committee
ODBC	Open Database Connectivity
OLAP	Online Analitical Processing
OWL	Ontology Web Language
PAL	Phase Alternating Line
PVR	Personal Video Recorder
SBTVD	Sistema Brasileiro de TV Digital
SDT	Service Description Table
SDTV	Standard Definition Table
SI	Service Information
STB	Set Top Box
TVDI	TV Digital Interativa
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

LISTA DE ILUSTRAÇÕES

SUMÁRIO

Capítulo

1
Introdução

A Televisão Digital Interativa (LEMOS et. al., 2004) (SOUZA FILHO, 2007) é mais uma nova fase vivida pela TV. Estágio esse que prima pela convergência de tecnologias digitais, através da sistêmica substituição de equipamentos analógicos por digitais, produzindo grandes mudanças em toda a cadeia produtiva e principalmente no consumo de mídias. Nesse sentido a visão da Web Semântica (W3C *et al.*, 2011) pode ser relevante nesse processo de transição de tecnologias vivido pela TV.

A visão da Web Semântica tem sido amplamente difundida e apoiada por um grande número de iniciativas tanto da indústria como da academia. A Web Semântica proporciona a compreensão e o gerenciamento dos conteúdos armazenados digitalmente, não importando o formato do arquivo apresentado. Essa pode ser utilizada junto à tecnologia de Descoberta de Conhecimento em Base de Dados (HAN; KAMBER, 2006) para obter padrões que deem maior expressividade ao conteúdo armazenado.

A Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Data Base* - KDD) caracteriza-se como o processo não trivial de extração de padrões nos dados provenientes de grandes bases de dados, representando conhecimento implícito capturado (HAN; KAMBER, 2006). O KDD é composto por fases e sua principal fase é a Mineração de Dados.

A Mineração de Dados (HAN; KAMBER, 2006) é uma área multidisciplinar, que envolve as áreas de Banco de Dados (BD), Inteligência Artificial (IA), Estatística, entre outras. Ela consegue ser aplicada em diversos domínios de conhecimento, traçando padrões e perfis, bem como, buscando um conhecimento implícito e desconhecido sobre os dados armazenados que podem ser úteis.

Devido aos grandes benefícios que pode trazer, a Mineração de Dados vem sendo trabalhada no contexto da TV Digital Interativa em diversas áreas, principalmente no que diz respeito à personalização de conteúdo para o usuário de TVDI.

Os principais sistemas de TV Digital Interativa existentes, são: o ATSC (ATSC, 2010) - sistema americano, DVB (DVB, 2010) - sistema europeu, ISDB (ARIB, 2010) - sistema japonês e o mais recente SBTVDI - sistema brasileiro que tem como *middleware* o Ginga (LAVID;TELEMIDIA, 2008) e será foco desse projeto. Este trabalho está inserido também no contexto do projeto *Knowledge TV*.

O projeto *Knowledge TV* (LINO et al., 2011) propõe ser uma camada semântica na plataforma da TV Digital Interativa Brasileira incorporada ao *middleware* Ginga. A mesma tem a finalidade de fornecer uma rica base de conhecimento contendo descrição de dados, recursos, serviços, aplicações e relacionamentos entre estes. Cada uma destas descrições é feita em uma linguagem padronizada formalmente, capaz de ser processada automaticamente por computadores.

Neste contexto, o presente trabalho tem como objetivo apresentar a definição e implementação do ambiente de KDD, com ênfase na Mineração de Dados, para a camada *Knowledge TV*; enfocando a integração entre a Mineração de Dados e os conceitos semânticos na TV Digital Interativa. Chamamos este modulo, foco desta dissertação de *Mining Knowledge TV*-MKTV.

1.1 OBJETIVOS

1.1.1 Objetivos Gerais

O uso do KDD e da Mineração de Dados na criação de modelos, utilizados só ou em conjunto com outras tecnologias, traz inúmeras vantagens, dentre elas estão: a possibilidade de análise de grandes conjuntos de dados, a possibilidade da descoberta de informação inesperada, possibilidade de utilizar variáveis quantitativas e qualitativas, etc. Na literatura temos diversos exemplos de utilização na prática de técnicas de KDD e Mineração de Dados, por exemplo, na descrição de padrões de comportamento de usuários em sites e lojas de compras, ou então, na identificação de pontos fora da curva (*outliers*) utilizada por empresas de cartão de crédito para detectar fraudes.

No contexto de TV Digital Interativa processos de KDD podem ser utilizados para suportar e resolver problemas relacionados à usabilidade da tecnologia, já que a TVDI é uma tecnologia nova e usuários em diferentes níveis de aptidão enfrentarão dificuldades para a usar. Também para personalização e recomendação, para tratar questões de *overload* de informação, e adicionalmente no contexto de e-*business* pode ser usado, por exemplo, para *marketing* direcionado.

Neste sentido, este trabalho tem como objetivo geral a construção de uma arquitetura de descoberta de conhecimento em base de dados utilizando conceitos semânticos no ambiente de TV Digital Interativa. Esta arquitetura visa prover conhecimentos úteis e não facilmente identificáveis para desenvolvedores de aplicações em TV Digital Interativa. A mesma poderá ser utilizada nas mais diversas áreas de aplicação em TVDI como, por exemplo, personalização e recomendação de conteúdo, elaboração de EPG personalizados, sobrecarga de conteúdo, etc.

1.1.2 Objetivos Específicos

Como objetivos específicos para este trabalho, podemos elencar:

1. Realizar uma revisão bibliográfica sobre o estado da arte em TV Digital Interativa e em Mineração de Dados;
2. Investigar as melhores técnicas de Mineração de Dados para o contexto da TV Digital Interativa;
3. Buscar algoritmos de alta performance e resultados dentro de cada técnica de Mineração de Dados;
4. Levantar requisitos e mapear processos para auxiliar na construção do ambiente de KDD genérico e integrado ao módulo semântico da *Knowledge TV*;
5. Mapear os dados da TV Digital Interativa nos níveis operacional, multidimensional e semântico;
6. Integrar o resultado da Mineração de Dados com o módulo semântico no contexto do projeto *Knowledge TV*;
7. Aplicar a Mineração de Dados em diversos contextos e problemas enfrentados pela TVDI através da elaboração de estudos de caso.

8. Modelar semanticamente o resultado do processo de Mineração de Dados nas informações provenientes do ambiente de TV Digital Interativa.

1.2 METODOLOGIA

A pesquisa desenvolvida neste projeto é do tipo teórico-empírica. Neste sentido foi feito revisão bibliográfica acerca dos temas envolvidos, buscando em livros e anais de congressos pesquisas recentes relacionados ao tema de estudo. Além disso, foram pesquisadas ferramentas de Mineração de Dados para saber da sua utilidade junto à nova plataforma computacional que é a TVDI, juntamente com técnicas e algoritmos medindo sua eficiência quanto à informação descoberta.

Para o desenvolvimento da mesma foram realizadas as seguintes etapas:

- **Etapa I:** Estudo teórico de conceitos das áreas de KDD (DW, DM, OLAP), TV Digital Interativa e Web Semântica/Ontologias – para que se possa entender melhor os conceitos e tecnologias envolvidas.
- **Etapa II**: Análise comparativa de ferramentas, técnicas e algoritmos de Mineração de Dados.
- **Etapa III**: Estudo e levantamento de requisitos para um ambiente de KDD no contexto da emergente plataforma de TV Digital Interativa.
- **Etapa IV:** Engenharia do Conhecimento, onde está contemplada toda a parte de modelagem semântica do projeto MKTV.
- **Etapa V**: Especificação, implementação e integração do MKTV na arquitetura do KTV.
- **Etapa VI**: Estudo de caso usando a arquitetura do MKTV para prover informações úteis na plataforma de TV Digital Interativa.
- **Etapa VII**: Escrita de artigos e dissertação.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho esta organizado em 7 Capítulos. De forma que após esta introdução, no Capítulo 2 são apresentados conceitos a cerca do processo de descoberta de conhecimento em

base de dados – KDD, identificando suas etapas de pré-processamento, Mineração de Dados e pós-processamento.

No Capítulo 3, é dada ênfase a principal fase do processo de KDD, a Mineração de Dados, que é descrita de forma detalhada, apresentado suas tarefas, seus algoritmos e as principais ferramentas de Mineração de Dados.

O Capítulo 4 apresenta conceitos de TVDI relevantes para estre trabalho, apresentando sua contínua evolução até a TVDI, suas diferenças em relação à TV analógica, os principais sistemas de TVDI existentes no mundo e os principais desafios dessa nova tecnologia. Além de apresentar o estado da arte e trabalhos relacionados à Mineração de Dados em TVDI. Com vistas a dar uma visão geral do que tem sido trabalhado em relação às áreas de TVDI e Mineração de Dados, através da análise de publicações nacionais e internacionais acerca do tema.

O Capítulo 5 apresenta a abordagem do *Mining Knowledge* TV (MKTV), um ambiente de Descoberta De Conhecimento em Base De Dados da *Knowledge* TV (KTV). Este trabalho é norteado pelos conceitos de KDD e modelagem semântica na TVDI integrados ao Ginga, *middleware* do SBTVDI.

O Capítulo 6 apresenta os experimentos realizados com o objetivo de validar a aplicabilidade do sistema MKTV.

A conclusão e os trabalhos futuros estão presentes no Capítulo 7, juntamente com as principais contribuições deste trabalho.

CAPÍTULO

2

Descoberta de Conhecimento em Bases de Dados

Diante do cenário existente nos dias atuais, em que a gestão do conhecimento se torna fundamental para o crescimento das organizações, surge a necessidade de explorar a grande quantidade de dados existentes nos bancos de dados para extrair algum conhecimento que ajude a prever, analisar e solucionar problemas.

Para (ROMÃO,2002), conhecimento é "todo o conjunto de dados e informações que as pessoas utilizam na prática para executar ações, a fim de realizar tarefas e criar nova informação".

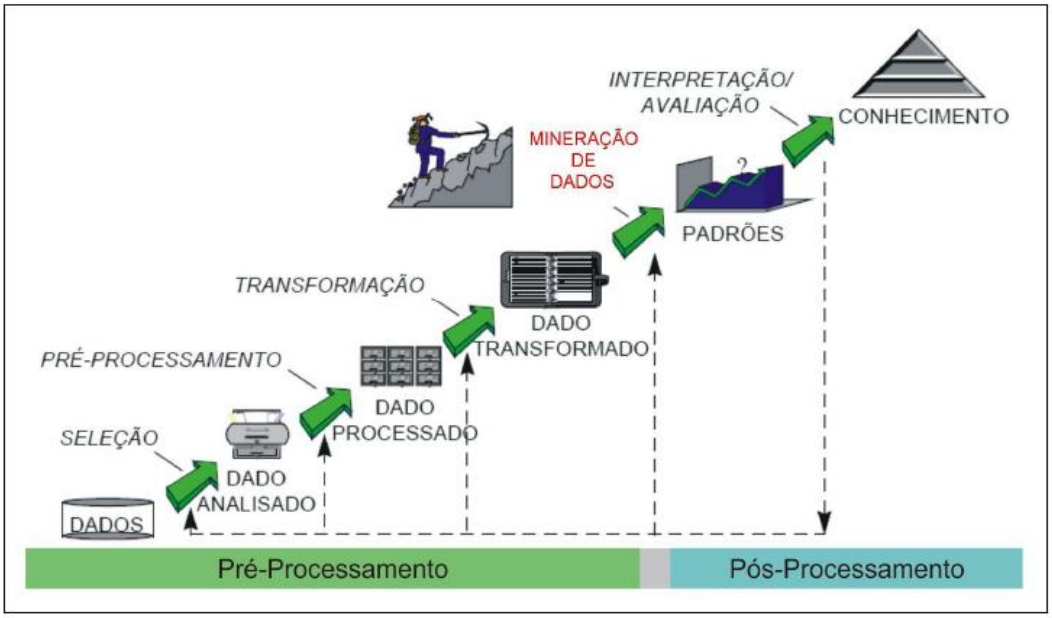
O processo que é capaz de descobrir conhecimento em base de dados é chamado de *Knowledge Discovery in Database* (KDD) (HAN; KAMBER, 2006). O KDD foi proposto em 1989 para referir-se a etapas que produzem conhecimento a partir de um conjunto agrupado de dados, com ênfase na fase de Mineração de Dados (*data mining*), que é a fase na qual são utilizados algoritmos para extração de conhecimento.

De acordo com Fayyad (1996), o processo de KDD pode ser definido como um processo não trivial de identificação de padrões válidos, potencialmente úteis em um conjunto de dados. Esses padrões extraídos devem ser, além de confiáveis, compreensíveis e úteis, podendo ser empregados para tirar proveito do conhecimento adquirido, seja ele científico ou comercial. Em alguns casos, para que sejam mensurados o quão importantes são os padrões encontrados, são estabelecidas métricas do que seria um padrão útil, tais como níveis de confiança, compreensão e utilidade baseados em estimativas estatísticas. Trata-se de uma área dinâmica e evolutiva,

envolvendo integrações com outras áreas de conhecimento como Estatística (ELDER, 1996), Inteligência Artificial (RUSSELL, S.; NORVIG, P., 2009) e Banco de Dados (SILBERSCHATZ, et al.1999).

O processo de descoberta de conhecimento em base de dados é composto por diversas fases que visam: a escolha de uma amostra de dados, o pré-processamento, a transformação dessa amostra, a mineração através da manipulação dos mesmos com a utilização de algoritmos e sua devida interpretação, culminando no conhecimento de uma informação antes não conhecida. De uma maneira simplificada todas as fases do KDD são apresentados na Figura 1.

Figura 1- Fases do Processo de KDD



Fonte: (FAYYAD, 2002) apud (ALMEIDA et. al., 2003, p.2)

2.1 FASES DO KDD

O processo de KDD é interativo e iterativo. O mesmo tem início na análise do entendimento do domínio da aplicação, dos objetivos a serem realizados e na definição do que deve ser feito após ter obtido novos resultados. A partir desta análise, o foco passa a ser o pré-processamento dos dados, para que estes possam ser escolhidos, eliminados os ruídos (dados irrelevantes, incompletos ou inválidos) e padronizados. Em seguida, vem à fase de Mineração de

Dados, onde serão aplicados técnicas e algoritmos específicos no intuito de extrair padrões; os quais na fase de pós-processamento possam ser interpretados quanto a sua qualidade e relevância.

2.1.1 Fase de Pré-Processamento

A fase de Pré-Processamento é iniciada pela etapa de **Seleção** de Atributos, que tem por objetivo: reconhecer entradas que desempenham uma contribuição relevante no processo de definição de saídas, otimizando essa relação em algum modelo. Tem papel importante na eliminação de entradas irrelevantes (FERREIRA, 2005, p.30).

Pode-se considerar que na etapa de seleção, são escolhidos apenas os atributos mais relevantes dentre os existentes no banco de dados, otimizando o tempo de processamento do algoritmo usado na fase de Mineração de Dados, pois ele trabalhará com um subconjunto menor de atributos, diminuindo o tempo de busca.

Ao final desta etapa é importante evidenciar que a base de dados fica completamente diferente da existente antes do processo.

Uma etapa posterior à Seleção, mas não menos importante, é a de **Pré-Processamento** que realiza inúmeras operações nos dados, dentre elas a Limpeza dos Dados (*Data Cleaning*), que tem como objetivo detectar e remover irregularidades, buscando assegurar a qualidade dos dados envolvidos no processo de KDD. Ela realiza operações básicas como a remoção de ruídos, que podem ser, por exemplo, atributos nulos, elementos duplicados, corrompidos ou redundantes. Na execução da limpeza de dados é de fundamental importância a presença de um especialista de domínio, ou seja, alguém que conheça todo o sistema, as regras de negócio, que tenha conhecimento do domínio do problema, pois a devida identificação e correção necessitam de um conhecimento especializado da esfera onde está sendo aplicado o processo.

A etapa seguinte consiste na **Transformação dos Dados**. Nesta etapa, os dados precisam ser armazenados e formatados adequadamente, para que na fase de Mineração de Dados, os algoritmos possam ser aplicados sem falhas. A normalização é uma das transformações mais realizadas e utilizadas, além disso, a fase de pré-processamento pode exigir até 80% do tempo do KDD (MANILLA, 1996).

2.1.2 Fase de Mineração de Dados

Após a realização da fase anterior, a **Mineração de Dados** (*Data Mining*) é iniciada. Esta fase é a mais importante do processo de KDD, sendo realizada através da escolha da técnica e do algoritmo mais compatível com o objetivo da extração, a fim de encontrar padrões nos dados que sirvam de subsídios para descobrir conhecimentos ocultos.

Para Fayyad et al.(1996), Mineração de Dados é um passo no processo de KDD que consiste na aplicação de análise de dados e algoritmos que, sob limitações de eficiência computacional aceitáveis, produz uma relação particular de padrões úteis de dados.

Para obter um bom desempenho na fase de Mineração de Dados, deverá ser escolhido um algoritmo minerador, ou seja, aquele que, diante de uma tarefa especificada, busca retirar conhecimento implícito e útil. Este algoritmo deve ser eficiente e eficaz para os objetivos do processo. Para isso, é demandado um bom entendimento do domínio do problema. Além desta questão, outra problemática que diz respeito à escolha do algoritmo da mineração são as variáveis envolvidas no processo. Neste caso, há a possibilidade do algoritmo ser intolerante a estas variáveis. No Capítulo 3, um aprofundamento desta fase será realizado.

2.1.3 Fase de Pós-Processamento

A Avaliação ou Pós-Processamento é a fase que identifica, entre os padrões extraídos na etapa de Mineração de Dados, os padrões interessantes ao critério estabelecido pelo usuário, ou realiza um novo processamento destes padrões descobertos, podendo voltar à fase inicial para novas iterações.

Segundo Domingues e Rezende (2005), o pós-processamento trata de uma etapa que merece consideração no processo de descoberta de conhecimento em base de dados; no qual o conhecimento extraído deve ser simplificado, avaliado, visualizado ou simplesmente documentado para o usuário final.

Para Aurélio (1999), o principal objetivo dessa fase é melhorar a compreensão do conhecimento descoberto pelo algoritmo minerador, e através da análise de um especialista de domínio e de uma série de medidas de qualidade, validar o processo como um todo.

A fase de pós-processamento tem como principais métodos e procedimentos para sua realização a avaliação, a interpretação e explanação, além da filtragem.

Na avaliação, o conhecimento extraído do conjunto de dados é apreciado através de critérios como, por exemplo, a eficiência, ou seja, tempo de processamento reduzido, compreensão, precisão, entre outros.

Por sua vez, a interpretação e a explanação visam utilizar estratégias como a documentação, a visualização, a modificação e/ou a comparação do conhecimento pré-existente com o conhecimento derivado do conjunto de dados, de forma a torná-lo compreensível ao usuário.

Para Domingues e Rezende, (2005) o procedimento de filtragem consiste em refinar o conhecimento que foi extraído do conjunto de dados, podendo ser realizado por vários mecanismos que variam de acordo com a técnica utilizada.

Ao término da avaliação, o conhecimento descoberto devido ao processo de KDD, deverá ser informado, sempre documentando e publicando os métodos, a fim de apresentar o conhecimento da melhor maneira ao usuário.

2.2 FATORES DECORRENTES DA COMPLEXIDADE DO KDD

Como pôde ser visto ao longo de todo o capítulo, o processo de KDD não é simples, consiste de várias fases, cada uma com suas peculiaridades e detalhes. Alguns fatores dificultam a realização deste método computacional, em sua maioria decorrente da sua complexidade.

Para Fayyad, Piatetsky e Smyth (1996b), dois conjuntos de fatores decorrem da complexidade envolvida no processo de KDD. O primeiro conjunto integra fatores operacionais, pois no processo de descoberta de conhecimento existe dificuldade na hora de integrar algoritmos específicos de mineração. Outra questão relevante neste quesito seria como é custoso manipular grandes conjuntos de dados, além da problemática de tratar os resultados obtidos de maneira ideal. Já o segundo conjunto, diz respeito a fatores de controle, que seriam referentes à dificuldade na hora de gerenciar e direcionar o processo de KDD. Neste caso, pode-se citar a dificuldade na hora de selecionar um algoritmo ideal para aquele objetivo específico.

2.3 CONCLUSÃO

Este capítulo teve o intuito de definir o processo de descoberta de conhecimento em base de dados e descrever todas as suas fases, ilustrando desde o pré-processamento, que é composto pelas etapas de seleção, pré-processamento e transformação, passando pela fase mais importante do KDD, que é a Mineração de Dados; até chegar ao pós-processamento do conhecimento obtido.

Os conceitos de KDD serão aplicados diretamente dentro deste trabalho, pois o mesmo visa à construção de um ambiente completo de KDD no ambiente de convergência de mídias que é a TVDI. Este ambiente deverá encontrar padrões úteis nos dados advindos da TVDI e disponibilizá-los para uso nas mais diversas aplicações.

O Capítulo 3 terá ênfase na Mineração de Dados, onde serão abordados seus conceitos, os algoritmos mais usados, as principais técnicas realizadas, além das ferramentas utilizadas para esse fim.

CAPÍTULO

3
Mineração de Dados

Na literatura, o termo Mineração de Dados tem sido empregado tanto para designar o processo de Descoberta de Conhecimento em Base de Dados, quanto para descrever a própria etapa de Mineração de Dados. Fayyad (1996) foi um dos primeiros a fazer a diferenciação dos termos, enfatizando que a Mineração de Dados especifica unicamente os meios através dos quais são extraídos e numerados uma série de padrões com base nos dados.

Para Berry e Linoff (1997, p.5): "Mineração de Dados é a exploração e a análise, por meio automático ou semiautomático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativos".

A Mineração de Dados é, na verdade, uma descoberta eficiente de padrões válidos e previamente não conhecidos, ou seja, informações não tão simples de serem encontradas em uma grande base de dados.

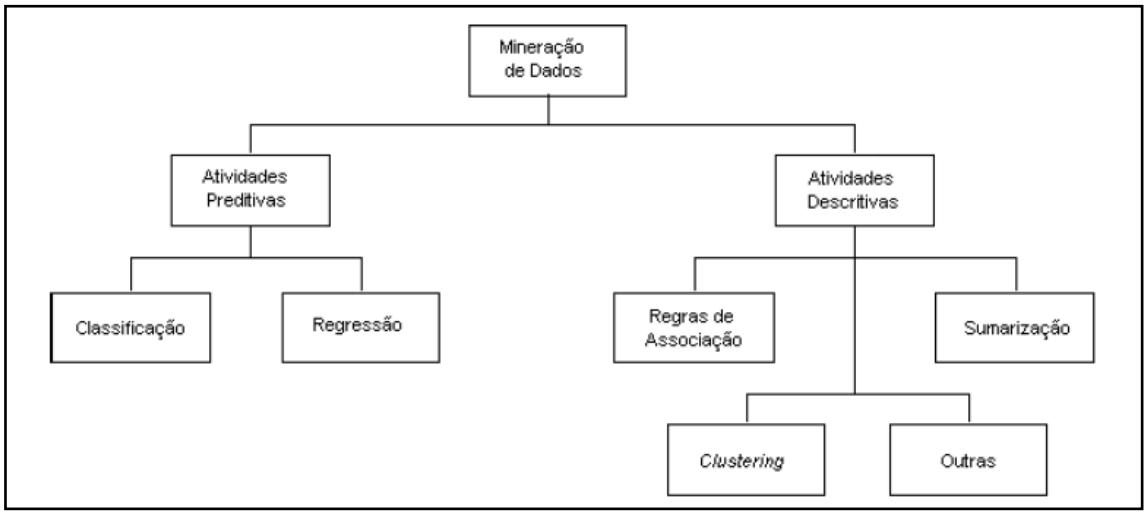
Os principais objetivos da Mineração de Dados são encontrar relacionamentos e padrões entre os dados e fornecer alternativas para que se possa realizar uma previsão de tendências futuras com base em informações pretéritas.

Os resultados gerados pela Mineração de Dados podem ser usados na tomada de decisão, no gerenciamento de processos, no processamento de pedidos, entre outros. Vale ressaltar que a mesma é composta por tarefas que servem para classificar os algoritmos de Mineração de Dados, que deverão ser utilizados no processo. Os algoritmos especificam o que deve ser buscado nos dados, que tipo de padrão poderia ser realmente relevante, qual seria o mais interessante encontrar.

3.1 TAREFAS DE MINERAÇÃO DE DADOS

Com relação às tarefas de Mineração de Dados, podemos considerá-las como sendo métodos de extrair o conhecimento da base de dados. Elas podem ser divididas em dois grupos de atividades: preditivas ou descritivas. Estas têm como principais tarefas: as regras de associação, clusterização e sumarização. Já as preditivas, por sua vez, são compostas das tarefas de classificação e regressão. A Figura 2 mostra os tipos de tarefas de Mineração de Dados, associados as suas tarefas específicas.

Figura 2- Tipos de tarefas de mineração



Fonte: (SCHEIDT; KÖERICH; SANTOS, 2008, p.10)

3.1.1 Atividades Preditivas

As atividades preditivas tomam por base os fatos ocorridos no passado para antecipar e prever se, ou como, os mesmos ocorrerão no futuro, utilizando-se da experiência armazenada na base de dados. Estas tarefas se dividem em Classificação ou Regressão.

3.1.1.1 Classificação

Segundo Amo (2004, p. 4), “Classificação é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar

o modelo para predizer a classe de objetos que ainda não foram classificados". Com a utilização dos algoritmos de Classificação torna-se possível a determinação do valor de um atributo através dos valores de um subconjunto dos demais atributos da base de dados.

Por meio de algoritmos classificadores pode-se realizar inferência de que, por exemplo, clientes do sexo masculino, com renda superior a R$ 2000,00 e com idade superior aos 40 anos compram produtos eletrônicos importados. Neste exemplo, o atributo "compra de produtos eletrônicos importados" é denominado classe, devido este atributo ser o alvo da classificação, que terá como possíveis valores o "sim" ou "não".

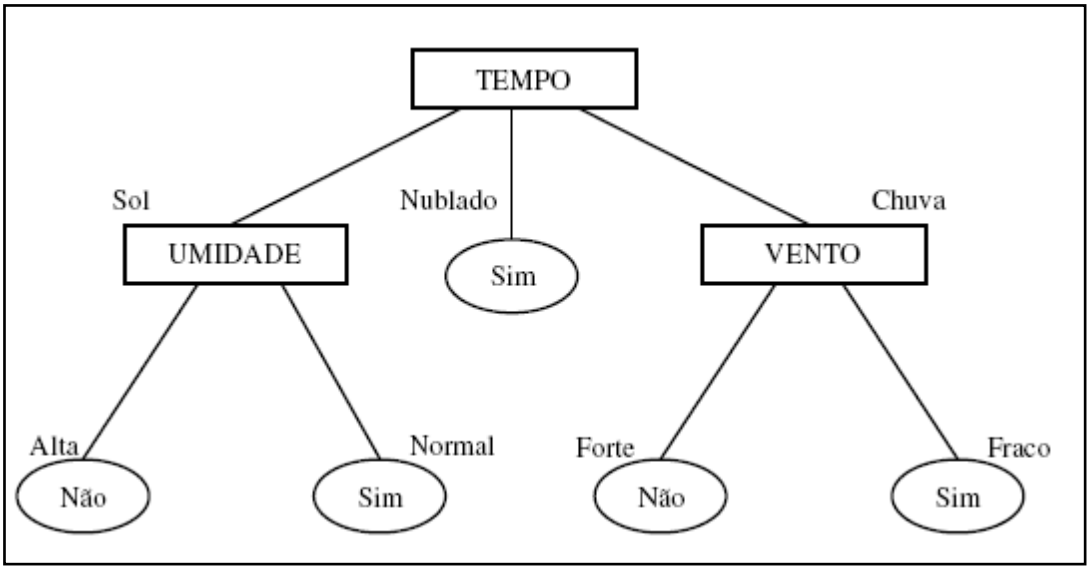
Para representar os conhecimentos descobertos desses algoritmos de classificação é utilizada uma estrutura de regras ou árvores. A Figura 3 ilustra uma árvore de decisão, que representa as condições temporais ideais para jogar tênis, relativas aos dados da Tabela 1. Neste caso, o algoritmo de classificação, após analisar esse conjunto de dados, irá predizer para um novo conjunto de atributos o valor da classe "Joga Tênis" se poderá ou não jogar tênis.

Tabela 1 – Dados das condições climáticas do ambiente de tênis

ATRIBUTOS			**CLASSE**
Tempo	**Umidade**	**Vento**	**Joga Tênis**
Sol	Alta	Fraco	Não
Sol	Alta	Forte	Não
Nublado	Alta	Fraco	Sim
Chuva	Alta	Fraco	Sim
Chuva	Normal	Fraco	Sim
Chuva	Normal	Forte	Não
Nublado	Normal	Forte	Sim
Sol	Alta	Fraco	Não
Sol	Normal	Fraco	Sim
Chuva	Normal	Fraco	Sim
Sol	Normal	Forte	Sim
Nublado	Alta	Forte	Sim
Nublado	Normal	Fraco	Sim
Chuva	Alta	Forte	Não

Fonte: (GUIMARÃES; LIMA, 2006, p.6)

Figura 3 - Árvore de decisão para jogar tênis



Fonte: (GUIMARÃES; LIMA, 2006, p.6)

3.1.1.2 Regressão

As técnicas de regressão são utilizadas para predizer uma informação futura. As mesmas lidam com resultados contínuos, isto é, dados numéricos (ponto flutuante) que se referem a representações em escala, por exemplo: área, peso, velocidade. Diferente da classificação que lida com dados discretos, ou seja, que se referem a contagens. Para ilustrar esta situação, o número de portas de um carro é um dado deste tipo.

Segundo Barbieri (2001, p.204), a Análise de Regressão deverá processar os dados em um conjunto de dados de forma a criar um modelo que represente o relacionamento existente entre as variáveis em estudo. Os principais objetivos da análise de regressão são: sumarização dos dados, predição, controle e estimação.

Como exemplo de situações abrangidas pela regressão, pode-se realizar a estimativa de número de filhos em uma família, determinar a renda total dessa família, entre outros (DIAS, 2001).

3.1.2 Atividades Descritivas

Segundo Xavier (2007 p. 25), nas atividades descritivas “só existe entrada, os algoritmos estabelecem relacionamento entre os dados e o que se deseja encontrar é totalmente desconhecido, sendo determinado pelo algoritmo”.

A tarefa é do tipo descritiva ou não supervisionada, quando a mesma trabalha com um conjunto de dados que não possuem uma classe determinada, tentando identificar padrões comportamentais que sejam comuns a todos os dados. Elas podem ser: regras de associação, segmentação e sumarização.

3.1.2.1 Regras De Associação

O processo de extração de regras de associação foi proposto inicialmente por (AGRAWAL et al.,1993), e representa um padrão ocorrido em combinações de itens com determinada frequência em uma base de dados (GONÇALVES, 2005).

As regras de associação são apresentadas na forma: SE **X** ENTÃO **Y**, onde X e Y são conjuntos de itens antecedentes e consequentes respectivamente, isto é, o X é a premissa da regra e Y é a conclusão a partir da premissa. Considerando as compras feitas em um supermercado (com cesta eletrônica) por um cliente, pode-se obter, por exemplo, a seguinte regra de associação:

SE <PÃO> E <LEITE> ENTÃO <OVOS>

Ou, de forma simplificada:

<PÃO>, <LEITE> → <OVOS>

Nesta situação, se o cliente compra pão e leite, terá grande tendência a comprar ovos também.

Esta tarefa é a mais utilizada em bancos de dados que armazenam muitos itens, como os existentes em grandes redes de supermercados, em que se deseja descobrir associações importantes entre os itens comercializados, de modo que, após ordenar os produtos com alguma relação na prateleira, consiga tirar alguma vantagem desta ordenação, por exemplo, uma maior venda de produtos.

Outra abordagem para o conceito de regras de associação leva em consideração os conceitos de Confiança e Suporte (GYŐRÖDI et al., 2004), no qual a Confiança estima a qualidade da regra, vendo a probabilidade da ocorrência de um conjunto de itens de X que

possuem Y; o Suporte, refere-se à frequência em que os itens contidos numa regra aparecem simultaneamente em outra.

Por exemplo, seja uma Confiança de 85% (0,85) da regra: Compra (empresa, computador) → Compra (empresa, impressora), significa que 85% das empresas que compram computador também compram impressora. E se essa mesma regra fictícia tivesse um Suporte de 3%, quer dizer que de todas as transações comerciais realizadas, 3% são efetuadas por empresas que comprando computador também compram impressoras.

Um conjunto de regras de associação pode ser visto na Tabela 2, que representa uma série de transações em um supermercado, levando em consideração determinados itens. A primeira coluna exibe o identificador da transação, que neste caso é uma compra, e as demais expõem se o determinado item foi ou não comprado.

Tabela 2 – Dados das compras efetuadas no supermercado

ID	LEITE	CAFÉ	CERVEJA	PÃO	MANTEIGA	ARROZ	FEIJÃO
1	Não	Sim	Não	Sim	Sim	Não	Não
2	Sim	Não	Sim	Sim	Sim	Não	Não
3	Não	Sim	Não	Sim	Sim	Não	Não
4	Sim	Sim	Não	Sim	Sim	Não	Não
5	Não	Não	Sim	Não	Não	Não	Não
6	Não	Não	Não	Não	Sim	Não	Não
7	Não	Não	Não	Sim	Não	Não	Não
8	Não	Não	Não	Não	Não	Não	Sim
9	Não	Não	Não	Não	Não	Sim	Sim
10	Não	Não	Não	Não	Não	Sim	Não

Fonte: (FREITAS, 1998 apud AURELIO; VELLASCO; LOPES, 1999)

Considerando a tabela acima, e utilizando o valor, por exemplo, de 0.3 para Suporte e 0.8 para Confiança, se constata no Quadro 1, as regras de associação que seriam considerados padrões interessantes de serem descobertos. Estes padrões foram agrupado pelos conjuntos de itens mais frequentes, para um FSup e Fconf, variáveis que representam o suporte e a confiança, maiores ou iguais aos parâmetros passados como suporte e confiança. O Quadro 1 exibe também conjuntos que tenham dois ou mais itens frequentes.

Qaudro 1 – Regras de associação geradas

Conjunto de itens frequentes: CAFÉ, PÃO. FSup = 0.3 Regra: Se (CAFÉ) então (PÃO). FConf = 1.
Conjunto de itens frequentes: CAFÉ, MANTEIGA. FSup = 0.3 Regra: Se (CAFÉ) então (MANTEIGA). FConf = 1.
Conjunto de itens frequentes: PÃO, MANTEIGA. FSup = 0.4 Regra: Se (PÃO) então (MANTEIGA). FConf = 0.8. Regra: Se (MANTEIGA) então (PÃO). FConf = 0.8
Conjunto de itens frequentes: CAFÉ, PÃO, MANTEIGA. FSup = 0.3 Regra: Se (CAFÉ e PÃO) então (MANTEIGA). FConf = 1. Regra: Se (CAFÉ e MANTEIGA) então (PÃO). FConf = 1 Regra: Se (CAFÉ) então (PÃO e MANTEIGA). FConf = 1

Fonte: (AURELIO;VELLASCO; LOPES 1999)

Para este caso, o algoritmo de regras de associação agrupa-as pelos conjuntos de itens frequentes, dos quais as regras foram geradas, exibindo somente as que têm dois ou mais itens frequentes.

3.1.2.2 Segmentação ou Agrupamento (Clustering)

Chama-se segmentação a tarefa que busca fragmentar blocos heterogêneos de informação em subgrupos ou segmentos (*clusters*) homogêneos. Neste tipo de tarefa, não há classes pré-definidas, característica essa que a distingue da tarefa de classificação. Os registros são agrupados de acordo com o grau de semelhança existente.

Para Barbieri (2001), o objetivo da segmentação é "Identificar a existência de diferentes grupos dentro de um conjunto de dados e, constatada esta existência, agrupar os elementos estudados de acordo com as semelhanças entre si, considerando-se as características analisadas".

Para ilustrar o processo de segmentação, pode-se utilizar os dados provenientes do processo de censo, para formar agrupamentos de domicílios, que podem ser: pela escolaridade dos residentes, pela profissão, grupos de domicílios pela mesma faixa etária, sexo, número de filhos, etc.. Verificando que por não haver classes pré-definidas, poderá se obter em um mesmo grupo, domicílios localizados em lugares opostos geograficamente (PORCARO, 2002).

Gerar um agrupamento de clientes de eletrodomésticos por região do país, ou agrupar o comportamento similar de compras efetuadas por clientes com determinada idade, são exemplo de segmentação.

3.1.2.3 Sumarização

A sumarização tem por objetivo identificar e indicar características semelhantes entre agrupamentos de dados, geralmente é aplicado em *clusters* gerados na tarefa de segmentação. Nesta tarefa, as descrições dos dados são geradas para caracterização resumida dos mesmos.

De acordo com Xavier (2007, p. 30), "Um exemplo de aplicação envolvendo sumarização é identificar as características dos estudantes de escolas públicas e a similaridade entre os conjuntos de dados indica que na maioria são estudantes carentes, cujos pais têm baixo nível de escolaridade".

As tarefas de Mineração de Dados descritas acima são apresentadas de forma resumida na Tabela 3.

Tabela 3 – Tarefas Realizadas por Técnicas de Mineração de Dados

TAREFA	DESCRIÇÃO	EXEMPLOS
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes	• Classificar pedidos de crédito • Esclarecer pedidos de seguros fraudulentos • Identificar a melhor forma de tratamento de um paciente
Estimativa (ou Regressão)	Usada para definir um valor para alguma variável contínua desconhecida	• Estimar o número de filhos ou a renda total de uma família • Estimar o valor em tempo de vida de um cliente • Estimar a probabilidade de que um paciente morrerá baseando-se nos resultados de diagnósticos médicos • Prever a demanda de um consumidor para um novo produto
Associação	Usada para determinar quais itens tendem a co-ocorrerem (serem adquiridos juntos) em uma mesma transação	• Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado
Segmentação (ou *Clustering*)	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos	• Agrupar clientes por região do país • Agrupar clientes com comportamento de compra similar • Agrupar seções de usuários Web para prever comportamento futuro de usuário
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados	• Tabular o significado e desvios padrão para todos os itens de dados • Derivar regras de síntese

Fonte: (DIAS, 2001 , p .11)

3.2 ALGORITMOS DE MINERAÇÃO

Descobrir padrões não conhecidos, regras, informações que aparentavam não existir em uma base de dados é o principal objetivo do KDD. Neste processo, a fase em que o conhecimento, ou seja, em que esses padrões são conhecidos, é na etapa de Mineração de Dados. Nesta etapa os algoritmos de mineração são aplicados na base de dados realizando a extração de conhecimentos.

Existem vários tipos de algoritmos propostos na literatura para cada uma das tarefas de MD citadas anteriormente. Estes algoritmos são baseados em técnicas de algoritmos genéticos[1], baseados em estatística, lógica nebulosa[2], entre outros.

Os algoritmos de associação objetivam encontrar todas as associações em que a presença de um conjunto de itens em uma transação implica em outros itens. Os algoritmos de segmentação visam agrupar o banco de dados em subconjuntos ou grupos menores. Os algoritmos de sumarização identificam quais atributos ou valores melhor representam um grupo de dados.

Segundo (PORCARO, 2002), as regras de associação "Se aplicam nos casos em que se deseja estudar preferências, afinidades, visando principalmente criar oportunidades para formação de "pacotes" para consumidores".

A seguir serão descritos em detalhes os algoritmos de regras de associação *Tertius* e *Predictive Apriori*, que utilizam técnicas da inteligência artificial e fórmulas estatísticas,

---

[1] "Algoritmos Genéticos são algoritmos matemáticos inspirados nos mecanismos de evolução natural e recombinação genética. A técnica de Algoritmos Genéticos fornece um mecanismo de busca adaptativa que se baseia no princípio Darwiniano de reprodução e sobrevivência dos mais aptos". (AURELIO; VELLASCO; LOPES, 1999).

[2] "Lógica Nebulosa (*Fuzzy Logic*) tem por objetivo modelar o modo aproximado de raciocínio humano, visando desenvolver sistemas computacionais capazes de tomar decisões racionais em um ambiente de incerteza e imprecisão". (AURELIO; VELLASCO; LOPES, 1999).

respectivamente, para encontrarem relações entre os dados por eles processados. Estes algoritmos serão aplicados no estudo de caso descrito no próximo capítulo.

3.2.1 Algoritmo *Tertius*

Proposto em 2001 por Flach e Lachiche, o Algoritmo *Tertius* utiliza abordagens de buscas heurísticas, isto é, que tem o caminho mais rápido e mais barato do estado inicial ao final, para que assim consiga encontrar as regras de associação mais eficientes.

O *Tertius* usa um algoritmo de Busca pela Melhor Escolha (*Best-First Search*), que usa fila de prioridades para percorrer uma árvore, com a abordagem *A** (RUSSEL; NORVING, 2009), ou seja, percorrendo todo o espaço de possibilidades das regras de associação através de heurísticas.

Uma busca, como solução do problema, é o processo que gera e analisa sequências de ações para alcançar um objetivo, percorrendo um caminho entre o estado inicial e o estado final; sendo chamado de qualidade da solução, o custo desse caminho. Em uma busca heurística, estima-se qual o melhor nó da fronteira a ser expandido com base em funções heurísticas. Utilizando-se para tal como estratégia de busca a melhor escolha, tendo como direção do estado inicial para o objetivo e do objetivo para o estado inicial.

O Algoritmo *Tertius* busca no espaço de possibilidades das regras de associação, através de uma heurística e uma função de estimativa otimista, isto é, que sempre estima o melhor resultado, selecionar as hipóteses mais coerentes e corretas, a partir dos dados analisados. A primeira coluna do resultado obtido com a aplicação do algoritmo *Tertius* em uma base de dados de compras de clientes, mostrada no Quadro 2, indica o número da regra obtida. A segunda coluna informa o grau de confirmação da regra, enquanto a terceira coluna, a frequência relativa desta mesma regra. Os próximos campos formam a regra encontrada, a qual está inclusa em um domínio de perfil de compras de clientes, e informa que pessoas as quais adquirem produtos do grupo cama e tem idade de 38 a 42 anos pagam mais de 901 reais por compra ou são pessoas do sexo masculino.

Quadro 2 – Exemplo de uma linha do resultado do Algoritmo *Tertius*

7. /* 0,223743 0,000210 */ grupo = CAMA and idade = 38_42 ==> valor = 901_MAX or sexo = M

Fonte: Próprio Autor/2012

É com base nessas informações que um analista especializado pode realizar suas considerações e efetivar suas próximas estratégias. Ao final da execução, o *Tertius* descreve também a quantidade de hipóteses consideradas e a quantidade explorada, e ainda a duração do processamento.

3.2.2 Algoritmo *Predictive Apriori*

Utilizado para encontrar associações de grande valor, conveniência ou interesse entre itens de dados, o Algoritmo *Predictive Apriori* foi definido por (SCHEFFER, 2004) com base no Algoritmo *Apriori* (AGRAWAL et al., 1994), que dá importância ao Suporte e à Confiança na geração de regras associativas.

Para eliminar regras que não se manifestam com frequência nas bases de dados é usado um estimador chamado Suporte, já visto anteriormente na seção 3.1.2.1. Um alto Suporte representa uma maior quantidade de regras, porém, com pouca precisão nas regras geradas em relação aos futuros dados.

Já a Confiança indica o grau de acerto da regra. Uma Confiança alta possibilita a geração de regras confiáveis para poucos registros, podendo apresentar regras otimistas em relação à realidade. Este otimismo pode ser corrigido através do cálculo do Suporte.

O Algoritmo *Predictive Apriori*, através de uma distribuição binomial, busca uma relação entre Suporte e Confiança de maneira a potencializar a geração das melhores regras. Para cada regra, a medida *Predictive Accuracy*, $c(x \rightarrow y)$, servirá para selecionar as *n* principais regras geradas em ordem decrescente. Esta medida fundamenta-se na probabilidade de uma previsão correta para a regra, ajustando Suporte e Confiança e o aumento gradativo do Suporte mínimo para retornar as *n* regras de associação selecionadas.

Um exemplo dos resultados com a execução do *Predictive Apriori* é exibido no Quadro 3, informando no primeiro campo o número da regra obtida. A última coluna informa o valor de precisão da regra (*accuracy*). Entre a primeira e a última coluna está a regra encontrada pelo algoritmo, no qual utiliza o mesmo domínio já mencionado no Quadro 2, e informa que quando um serviço é adquirido por pessoas de 38 a 42 anos, essa pessoa é do sexo masculino.

Quadro 3 – Resultado de uma linha do *Predictive Apriori*

4. grupo= SERVICO idade=38_42 151 ==> sexo=M 151 acc:(0.99487)

Fonte: **Próprio Autor/2012**

3.3 FERRAMENTAS DE MINERAÇÃO DE DADOS

Para se realizar a Mineração de Dados se faz necessário o uso de ferramentas apropriadas a sua execução. Neste sentido, as ferramentas utilizadas atualmente no processo de Descoberta de Conhecimento em Banco de Dados são inúmeras. Existem ferramentas proprietárias e *open source*[3]. Cabe a cada equipe técnica juntamente com a direção de uma organização saber escolher aquela que melhor se aplica à sua realidade.

Na Tabela 4 são apresentadas algumas ferramentas de Mineração de Dados que exploram os dados armazenados em busca de conhecimento, existindo também ferramentas específicas para as diversas etapas do processo de KDD. A seguir a ferramenta WEKA será detalhada devido o seu uso neste trabalho.

Tabela 4 – Ferramentas de Mineração de Dados

Ferramenta	**Tarefas**	**Fabricante e Site de Acesso**
SPSS/Clementine	Classificação, Regras de Associação, Clusterização, Sequência e Detecção de Desvios.	SPSS Inc. www.spss.com
PolyAnalyst	Classificação, Regressão, Regras de Associação, Clusterização, Sumarização e Detecção de Desvios.	Megaputer Intelligence http://www.megaputer.com/polyanalyst.php
Weka	Classificação, Regressão e Regras de Associação.	University of Waikato www.cs.waikato.ac.nz
Intelligent Miner	Classificação, Regras de Associação, Clusterização e Sumarização.	IBM Corp. http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.db2.udb.doc/wareh/getsta06im.htm

[3] Open Source se refere a tecnologia que possui código aberto, este pode ser alterado, distribuído sem ter que pagar qualquer licença. (SILVA; TEIXEIRA; MARINHO, 2005)

WizRule	Sumarização, Classificação e Detecção de Erros.	WizSoft Inc. http://www.wizsoft.com/default.asp?Win=8
Bramining	Classificação, Regras de Associação, Regressão e Clusterização.	Graal Corp. www.graal-corp.com.br
Oracle Data Mining	Classificação, Regressão, Associação, Clusterização e Mineração de Textos	Oracle http://www.oracle.com/technetwork /database/options/odm/index.html
SAS Enterprise Miner	Classificação, Regras de Associação, Regressão e Sumarização	SAS Inc. http://www.sas.com/technologies/analytics/datamining/miner/

Fonte: (BOENTE; OLIVEIRA; ROSA, 2007, p.10)

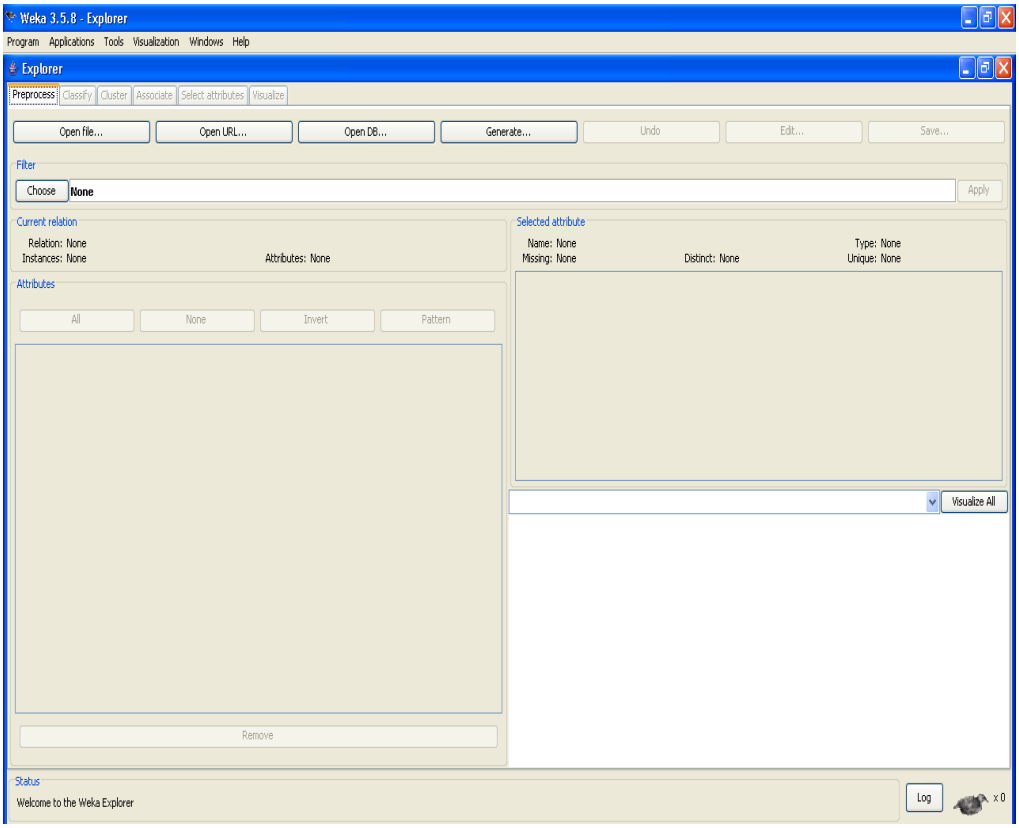
3.3.1 *Weka*

A ferramenta de Mineração de Dados *WEKA* (*Waikato Environment for Knowledge Analysis*), é um conjunto de ferramentas de pré-processamento de dados e algoritmos de aprendizagem de máquina e pós-processamento. Esta ferramenta foi desenvolvida desde 1993 na Universidade de *Waikato* na Nova Zelândia, tendo como métodos implementados em sua API: heurísticas de mineração aplicadas às técnicas de classificação, regressão, agrupamento, regras de associação, etc (ROMAO, 2002).

O *software* é bastante utilizado pois provê interface amigável, fácil interatividade, além de ser um *software* livre, licenciado pela *GPL - General Public License* (GPL, 2011). Por ser desenvolvida em *Java* (DEITEL; DEITEL, 2005), é portável a diversos sistemas operacionais, como *Linux, Windows e Machintosh*. A Figura 4 mostra a interface da ferramenta WEKA no modo Explorer, no qual é possível aplicar algoritmos específicos de mineração.

A ferramenta manipula os dados através de um arquivo texto de entrada de dados no formato".arff`", que deverá ser criado através das informações contidas no banco de dados. Este arquivo é formado por duas partes, respectivamente: o cabeçalho e os dados. Na primeira parte está contido o nome do conjunto de dados, junto com a lista de atributos e seus tipos; na segunda parte estão os registros, ou seja, os dados.

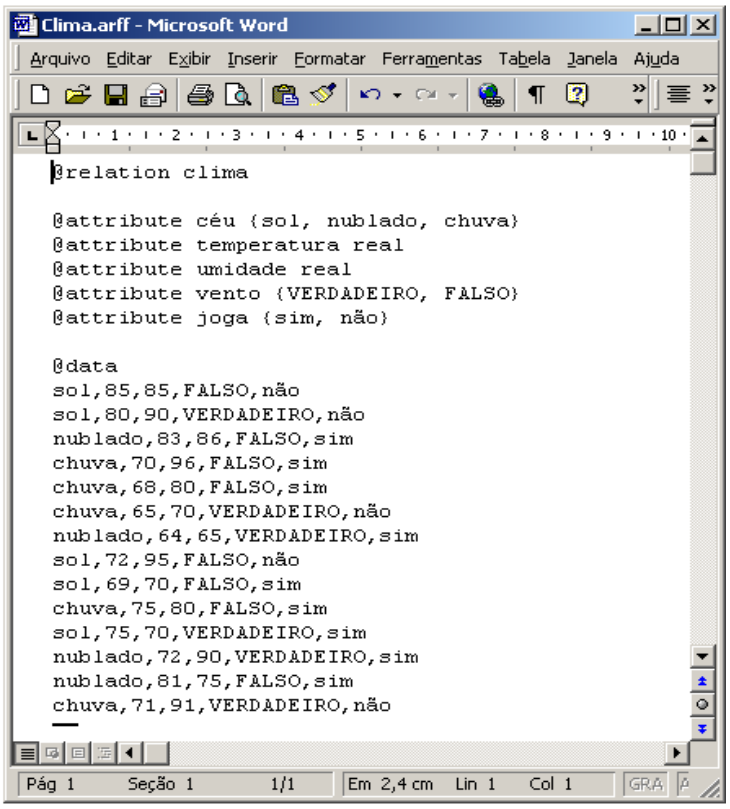
Figura 4 - Weka Modo Explorer



Fonte: Próprio Autor/2008

A Figura 5 mostra um arquivo no formato “.arff”, no qual a base de dados se refere a ocorrências de observações de condições climáticas e idas a prática de vôlei de algum indivíduo, identificado pela palavra *@relation*, e tem como atributos: céu, temperatura, umidade, vento e joga, representado por *@attribute*. Os mesmos só poderão ter os valores ou tipos informados ao lado (deles), como por exemplo, o atributo céu só poderá ter como valores: sol, nublado ou chuva; os atributos que são numéricos são representados pela palavra real. Os dados estão abaixo do identificador @data, onde cada coluna, separada por vírgulas, representa um atributo, por exemplo: na primeira linha - Sol, 85,85, Falso, Não; o item “Sol” representa o atributo céu, o item “85” representa a temperatura, o seguinte item “85” representa a umidade, “Falso” diz respeito à condição do vento, “Não” representa a coluna se haverá jogo o não.

Figura 5 - Arquivo no formato ".arff"



Fonte: Próprio Autor/2008

3.4 CONCLUSÃO

Este Capítulo teve como objetivo descrever a principal fase da Descoberta de Conhecimento em Base De Dados, que é a Mineração de Dados. Foram definidos seus conceitos, as principais tarefas de mineração, e os algoritmos que as aplicam. Também foi realizado um detalhamento de dois algoritmos de regras de associação, assim como a discriminação das principais ferramentas de Mineração de Dados.

Através desse Capítulo pode-se observar o grande potencial existente na tecnologia de Mineração de Dados. Dessa forma utilizaremos Mineração de Dados no nosso projeto em conjunto com outras tecnologias com objetivo descobrir informações úteis e importantes no ambiente de TVDI.

A ferramenta WEKA é o engenho de Mineração de Dados escolhido e implantado para encontrar padrões úteis no contexto do projeto.

O capítulo 4 abordará o ambiente de convergência digital que é a TV Digital Interativa.

CAPÍTULO

4

TV Digital Interativa e o Estado da Arte de Mineração de Dados em TVDI

O contexto histórico da TV Digital Interativa (TVDI) nos mostra a evolução obtida por esse meio de comunicação presente em todas as partes do mundo. Atualmente a TVDI pode ser considerado um ambiente de convergência de mídias digitais, que agrega a internet, aplicações, etc.

A Mineração de Dados é utilizada neste novo ambiente de TV como meio para personalização de conteúdo aos usuários de TV.

Neste Capítulo é apresentado esse novo momento vivido pela TV e como a Mineração de Dados é utilizada para solucionar problemas neste ambiente, juntamente com o estado da arte de TVDI e Mineração de Dados, bem como uma análise comparativa dos trabalhos apresentados.

4.1 TV DIGITAL INTERATIVA

Não tão diferente de outros meios de comunicação, a Televisão vem passando por constantes momentos de evolução desde sua criação. A mudança de visualização em preto e branco para à cores, o desenvolvimento por parte dos alemães de novos padrões, como o PAL (LEMOS et. al., 2004), superando o NTSC (LEMOS et. al., 2004), a chegada do HDTV (LEMOS et. al., 2004) são todos momentos de transformação. A Televisão Digital Interativa é mais uma nova fase vivida pela TV, estágio esse que prima pela convergência de tecnologias digitais, através da sistêmica substituição de equipamentos analógicos por digitais, produzindo grandes

mudanças em toda a cadeia produtiva e principalmente no consumo de mídias (LEMOS et. AL., 2004).

A TV Digital Interativa Interativa (TVDI) apresenta significativas diferenças em relação a televisão convencional ou televisão analógica, dentre elas, a qualidade do som e na definição do vídeo, multiprogramação, portabilidade, mobilidade e interatividade.

Expandindo o formato para 16:9 a TVDI aumenta o formato de tela existente na TV analógica que é de 4:3 tendo um aumento considerável na resolução (BOLAÑOS ET al., 2004). Com relação ao som não seria diferente; o mesmo que era utilizado em formato mono ou estéreo, passado em dois canais, será usado no formato com 5.1 canais, trazendo maior qualidade no que diz respeito ao áudio. Além da grande melhora no que se refere ao áudio e vídeo, outras melhorias com relação à TV analógica dizem respeito a multiprogramação, na qual o canal poderá exibir mais de um programa no mesmo horário; portabilidade e mobilidade pois é permitida a recepção de sinal por dispositivos móveis e em ambientes dinâmicos; já a interatividade será a característica inovadora que permitirá ao expectador de TV responder enquetes, acessar a internet, fazer compras, interagir com o canal e/ou programa assistido.

Como novo meio de comunicação, a TVDI já foi implantada em vários países do mundo, dentre eles os Estados Unidos da América, os países da Europa e o Japão tem sistemas próprios de TVDI, sendo estes: ATSC - sistema americano (ATSC, 2010), DVB - sistema europeu (DVB, 2010), ISDB - sistema japonês (ARIB, 2010). No Brasil, o sistema brasileiro de TV Digital Interativa - SBTVDI (SBTVDI, 2011) esta em fase de implantação e o mesmo fará com que exista uma maior quantidade de canais sendo disponibilizados aos usuários, trazendo uma maior diversidade de programação. Isso implicará também na otimização da cobertura de TV em toda a localidade que implantar a TVDI, e o acesso à internet que é um dos benefícios inerentes a tecnologia, tornará possível uma gama de aplicações interativas englobando diversos tipos de elementos gráficos (SOUZA; OLIVEIRA, 2005).

A otimização da cobertura de TV e o acesso à internet tornam a experiência da TVDI socialmente ainda mais válida, visto que a TV é o meio de comunicação com maior penetração nos lares brasileiros, com cerca de 92%. Essa evolução poderá fazer com que o acesso a educação, a informação e a serviços básicos chegue mais rapidamente em locais de difícil obtenção dos mesmos.

Cada sistema de TV Digital Interativa existente possui um componente chamado *middleware*, cuja função é abstrair detalhes e particularidades do hardware do receptor de TVDI. O *middleware* do SBTVDI é o Ginga (LAVID;TELEMIDIA, 2008) que é composto de dois subsistemas: o Ginga-J (SOUZA FILHO; LEITE; BATISTA; 2007) e o Ginga-NCL (SOARES; RODRIGUES; MORENO; 2007), que são responsáveis pela execução de conteúdo imperativo (Java) e de conteúdo declarativo (NCL) respectivamente.

Nos países que a TV Digital Interativa Interativa já vive uma fase de amadurecimento em sua implantação e utilização, a oferta de conteúdo de TV (programas) tem sido abundante. Fato esse, que tem feito o telespectador ter dificuldade na busca do melhor programa, ou do seu programa favorito. Com o intuito de solucionar esse problema são disponibilizadas informações sobre os programas em revistas especializadas na programação de canais, sites e nos EPG`s (do inglês *Episode Program Guide*) Guias de programação eletrônica. Os mesmos, são ferramentas interativas para visualização da programação de canais disponibilizados pelas emissoras de televisão.

Embora exista o esforço das emissoras em informar de forma eficiente os seus programas aos usuários de TV Digital Interativa, por meio dos supracitados métodos, a informação acaba não chegando de forma adequada, devido a grande quantidade de programas ofertados por inúmeros canais de TV.

Para (EHRMANTRAUT et. Al., 1996), se tornou um desafio obter informação sobre os programas de televisão por telespectadores. Segundo (SILVA, 2005) a quantidade de programas logo excederá os limites do que pode ser razoavelmente observado nos guias de programação dos canais, tornando-se um problema de sobrecarga de informação.

A resolução desse problema de sobrecarga de informação pode ser a personalização, que é a capacidade de fornecer conteúdos adaptados às pessoas sobre as suas preferências e gostos (HAGEN, 1999). Através de métodos de personalização poderão surgir aplicações de *t-commerce* (transações comerciais via TVDI) e de programação personalizada especifica para o usuário, direcionando de acordo com os gostos do espectador, o seu programa favorito.

Alguns pesquisadores vêm aplicando o processo de Mineração de Dados como um dos métodos para se obter guias de programação personalizada, perfis de usuários de TV, entre outros.

O projeto apresentado neste trabalho de mestrado se propõe a dar uma nova contribuição ao estado da arte em TVDI, no que diz respeito à problemática da grande quantidade de informação para o usuário de TV Digital Interativa, uso inteligente da TVDI, entre outros. O projeto MKTV, foco desse trabalho, busca através da integração das tecnologias de Mineração de Dados e Representação de Conhecimento ser uma importante ferramenta para auxiliar usuários de TVDI, que poderão usar ferramentas com conhecimento minerado a seu respeito e, por exemplo, encontrar mais rapidamente um programa que lhe agrade, ou descobrir informações relevantes sobre o programa assistido.

4.2 ESTADO DA ARTE DE MINERAÇÃO DE DADOS EM TVDI

A Mineração de Dados tem sido utilizada em diversas áreas do conhecimento humano, pois a mesmo pode traçar padrões e perfis em bases com grandes volumes de dados. Nesse sentido a Mineração de Dados busca um conhecimento útil, implícito e desconhecido nas informações mineradas.

Devido seu poder computacional, bem como aos benefícios trazidos, como análise de grandes base de dados e descoberta de padrões úteis não conhecidos, a Mineração de Dados vem sendo trabalhada no contexto da TV Digital Interativa através da aplicação de algoritmos estatísticos e com técnicas de IA.

Os temas mais trabalhados no que diz respeito à Mineração de Dados em TV Digital Interativa são os que envolvem a solução ao problema de sobrecarga de informação de programas para os usuários de TV Digital Interativa. Neste sentido eles abordam a criação de arquiteturas, ferramentas ou métodos para personalização de conteúdo; a busca para traçar um perfil de usuários; além de pesquisas voltadas para propaganda direcionada a grupos de telespectadores.

A seguir serão detalhados alguns artigos, dissertações e teses, encontradas na literatura, que trabalham com Mineração de Dados na TV Digital Interativa.

O primeiro a ser descrito será o artigo publicado em 2000, por Gutta S. et. al. Chamado "*TV Content Recomender System"* e que tem por objetivo resolver o problema de sobrecarga de informação para o telespectador através da criação de um sistema de personalização com base nos gostos do usuário de forma implícita, ou seja, sem nenhuma interação do usuário no sentido de fornecer informações.

A proposta é fazer um guia de programação inteligente (do inglês Electronic Programming Guide - EPG), fornecendo a informação de forma intuitiva e usando a estratégia de pontuar os programas mais vistos. O mesmo deverá funcionar como um ranking. Os que têm maior pontuação são exibidos primeiro. O software tem um ambiente para visualização dos dados e utiliza dois algoritmos de Mineração de Dados. Em um primeiro momento usa um classificador Bayesiano (HAN; KAMBER, 2006) para calcular a probabilidade de o telespectador gostar ou não do programa, e em outro momento usa o algoritmo de aprendizagem de maquina C4.5 (HAN; KAMBER, 2006) que tem como característica o uso de árvores de decisão para construir regras para a classificação de um programa. O autor informa que fez um perfil adaptativo do usuário, mas não comenta como o fez.

Em Bozios et al., 2001 é apresentado um módulo do projeto IMEDIA, que tem como objetivo o envio de propaganda direcionada para cada tipo de telespectador, dividindo-os em grupos (*clusters*) de grupos de consumidores alvos; e através de um engenho de Mineração de Dados são definidas quais peças publicitárias vão ser direcionadas para cada *cluster*.

O motor de Mineração de Dados é utilizado para segmentar os telespectadores em subgrupos com perfis de uso da TV semelhantes, e também é usado para gerar regras onde associam preferências, estilo de vida, hábitos de compra, entre outros. São utilizados nesse processo algoritmos de Redes Neurais (HAN; KAMBER, 2006) e Árvores de Decisão (HAN; KAMBER, 2006).

O processo de descoberta de perfil do cliente se dá de forma híbrida, ou seja, buscando informações de maneira explícita e implícita, por meio de um questionário passado antecipadamente que fica armazenado no Set-Top-Box, e a partir da interação com a TV no caso do processo implícito.

Em O'Sullivan et. al (2004) a Mineração de Dados, é usada como abordagem pelos pEPG's (personalised Eletronic Programme Guide) que são apontados como possível solução nos dispositivos de *Personal Vídeo Recorder* (PVR), diferente de outros autores que utilizam o Set-Top-Box como instrumento a ser trabalhado. O método utilizado na Mineração de Dados foi o de Regras de Associação, utilizando o algoritmo *Apriori* (HAN; KAMBER, 2006) que tem como métricas de avaliação de qualidade de padrões descobertos o suporte e a confiança. Através das técnicas de mineração usadas nos dados extraídos da interação com o usuário, busca-se obter conhecimento sobre a relevância de um programa, além de um perfil associado ao usuário.

No artigo em questão foi pesquisada a relevância de técnicas de classificação explícita e implícita, verificando que esta é tão precisa quanto às técnicas explícitas.

O sistema REPTVDI (LUCAS; ZORZO, 2009), foca a descoberta de informação sobre o usuário da TV Digital Interativa através da Mineração de Dados, e sua proposta é utilizar técnicas e algoritmos para refletir a preferência de um grupo de telespectadores de forma totalmente implícita, isto é, sem a participação efetiva do usuário na informação de melhores programas, diferente de alguns sistemas de personalização existentes que necessitam da interação direta do telespectador.

O modelo utilizado se baseia na obtenção de um histórico de uso dos telespectadores que serão armazenados em um arquivo XML (W3C, 2011) e dos metadados provenientes da TV Digital Interativa que descrevem os programas passados. Todos os dados estarão no Set-Top-Box (STB) onde será realizada a Mineração de Dados. Após esse processo as informações geradas são colocadas em outro arquivo onde é feito intersecção com as informações dos metadados e é exibida na tela a recomendação.

Quanto à arquitetura utilizada, o REPTVDI não foi pensado utilizando o canal de retorno. O sistema é dividido em três módulos: metadados, perfil de grupo do usuário, e recomendação. Foi utilizada a abordagem de filtragem baseada em conteúdo, que relaciona os conteúdos semelhantes com as preferências do telespectador; diferente da abordagem de filtragem colaborativa que utiliza a relação entre os perfis de telespectadores distintos com preferências semelhantes.

Como restrição ao trabalho tem a questão de que é necessária uma quantidade mínima de memória no STB, a qual não é especificada, pois o mesmo utiliza o STB como servidor de aplicação.

Em Alves et. Al. (2008) é proposta uma arquitetura, chamada de CollaboraTVware, que busca orientar usuários na busca e escolha de produtos e serviços no ambiente de TV Digital Interativa interativa, através da participação colaborativa de usuários de mesmo contexto e perfil.

A participação colaborativa objetiva obter notas para cada programa ou serviço interativo, e essa avaliação será feita por telespectadores que expressam sua opinião por meio da interação com a TVDI.

As informações referentes ao perfil de usuários são obtidas utilizando a abordagem direta e colaborativa, através das informações explícitas informadas diretamente pelo usuário, tais como dados pessoais, preferências de gênero, programas, etc.

A infraestrutura apresentada utiliza a Mineração de Dados para prover os serviços de predição, responsáveis por extrair os padrões relativos a um conjunto de participações colaborativas informadas pelos telespectadores. A mesma também auxilia, após serem obtidos os padrões, na busca e seleção do programa preferido conforme o contexto e o perfil do usuário.

A tarefa de classificação é realizada no provedor de serviço enquanto a de predição ocorre no dispositivo do usuário. Nas duas tarefas foram implementados algoritmos utilizando técnicas de Árvore de Decisão, Redes Neurais e aprendizagem Bayesiana (HAN; KAMBER, 2006).

O CollaboraTVware utiliza as tecnologias de metadados flexíveis, canal de retorno, e é criada usando a linguagem Java (JAVA, 2012), por isso necessita da maquina virtual Java. O processo de Mineração de Dados se deu utilizando a ferramenta de Mineração de Dados WEKA.

4.2.1 Análise Comparativa Dos Trabalhos

Como visto na seção anterior, a Mineração de Dados tem sido usada no contexto de TV Digital Interativa e convergência de mídias como meio para resolver o problema de sobrecarga de informação, geralmente associada a métodos de recomendação (considerados nos trabalhos (GUTTA et al., 2000) [1] (O'SULLIVAN et el., 2004) [3] (ALVES; BRESSAN, 2008) [4] (LUCAS; ZORZO, 2009) [5] - da Tabela 5 abaixo) usados para trazer a informação de forma coerente ao usuário. A Mineração de Dados também é empregada na criação de perfis de usuários (O'SULLIVAN et el., 2004) [3], propaganda direcionada a públicos específicos (BOZIOS et. Al, 2001) [2] (LEKAKOS; GIAGLIS, 2002) [7], etc.; e usando o Set-Top-Box como servidor de aplicação (como encontrado nos trabalhos (BOZIOS et. Al, 2001) [2], (ALVES; BRESSAN, 2008) [4], (LUCAS; ZORZO, 2009) [5], (AVILA; ZORZO, 2009) [6], (LEKAKOS; GIAGLIS, 2002) [7]).

Neste contexto, as principais tarefas de Mineração de Dados utilizadas encontradas na literatura foram as de Classificação (trabalhos (GUTTA et al., 2000) [1] , (ALVES; BRESSAN, 2008) [4], (LEKAKOS; GIAGLIS, 2002) [7]), Clusterização (trabalho (BOZIOS et. Al, 2001) [2]), e Regras de Associação (trabalhos (O'SULLIVAN et el., 2004) [3] (LUCAS; ZORZO,

2009) [5] (AVILA; ZORZO, 2009) [6]; aplicados em sua maioria pelos algoritmos C4.5 (trabalhos (GUTTA et al., 2000) [1] e (ALVES; BRESSAN, 2008) [4]), *APRIORI* (trabalhos (O'SULLIVAN et el., 2004) [3] (LUCAS; ZORZO, 2009) [5], (AVILA; ZORZO, 2009) [6]), *Back Propagation* (trabalho (BOZIOS et. Al, 2001) [2]), e Classificador Bayesiano (trabalhos (GUTTA et al., 2000) [1] e (ALVES; BRESSAN, 2008) [4]).

Como ambiente para aplicação da Mineração de Dados foi utilizadas: a ferramenta livre criada pela universidade de Waikato - WEKA (WEKA, 2010) que é abordada nos trabalhos (ALVES; BRESSAN, 2008) [4] (LUCAS; ZORZO, 2009) [5]; bem como outro engenho de mineração usado é o DARWIN, software da Oracle (http://www.oracle.com/br/index.html), encontrado no trabalho (BOZIOS et. AL, 2001) [2] (LEKAKOS; GIAGLIS, 2002) [7], além de implementações de engenhos de mineração específicas ao projeto, como no trabalho (AVILA; ZORZO, 2009) [6].

Os trabalhos (ALVES; BRESSAN, 2008) [4] (LUCAS; ZORZO, 2009) [5] desenvolveram suas aplicações para serem executadas sobre o *middleware* existente no Set-Top-Box da TV. Já o trabalho (AVILA; ZORZO, 2009) [6] implementa sua arquitetura acoplada ao *core* do *Middleware* Ginga (LAVID;TELEMIDIA, 2008).

Nenhum trabalho propõe a integração da Mineração de Dados com a Web Semântica no intuito de prover mais expressividade aos dados gerados pela Mineração de Dados. As informações discutidas nesta seção são sumarizadas na Tabela 5 abaixo.

Tabela 5. Características dos trabalhos pesquisados sobre Mineração de Dados em TVDI

CARACTERÍSTICAS	**[1]**	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**	**[7]**
TAREFA DE MINERAÇÃO CLASSIFICAÇÃO	X			X			X
TAREFA DE MINERAÇÃO CLUSTERIZAÇÃO		X					
TAREFA DE MINERAÇÃO REGRAS DE ASSOCIAÇÃO			X		X	X	
FERRAMENTA USADA DARWIN (ORACLE)		X					X
FERRAMENTA USADA WEKA				X	X		
FERRAMENTA USADA IMPLEMENTAÇÃO PRÓPRIA						X	
ALGORITMO USADO C4.5	X	X		X			X
ALGORITMO USADO APRIORI			X		X	X	
ALGORITMO USADO BACK PROPAGATION		X		X			X
ALGORITMO USADO CLASSIFICADOR BAYESIANO	X			X			

OBJETIVO RECOMENDAÇÃO	X		X	X	X	X	
OBJETIVO PROPAGANDA DIRECIONADA		X					X
USA STB COMO SERVIDOR DE APLICAÇÃO		X		X	X	X	X
INTEGRADO AO MIDDLEWARE GINGA						X	
APLICAÇÃO RODANDO SOBRE MIDDLEWARE				X	X		

Fonte: Próprio Autor (2012).

4.3 CONCLUSÃO

Este Capítulo tratou da evolução da televisão e sua nova fase que é a TVDI. Esta nova fase, apresenta uma série de evoluções em relação a sua fase anterior, a analógica, como a melhor qualidade de som e áudio, interatividade, portabilidade, multiprogramação, entre outros.

A TVDI pode ser vista como um ambiente de convergência de mídias, já que compreende as tecnologias de televisão e internet. Neste cenário, a Mineração de Dados pode ser utilizada para auxiliar a resolver problemas que surgem neste novo ambiente de TV. Este capítulo trouxe uma revisão bibliográfica dos principais trabalhos que tem como proposta a inserção da Mineração de Dados na TVDI.

As principais tarefas de Mineração de Dados encontradas nos trabalhos pesquisados foram as de classificação, clusterização e regras de associação. Além dessas, os algoritmos mais utilizados nesses trabalhos foram o apriori, o C4.5, o back propagation e o classificador bayesiano.

A TVDI e a Mineração de Dados tem grande importância neste trabalho, visto que o mesmo desenvolve um ambiente de Mineração de Dados para descobrir conhecimentos das informações provenientes da plataforma de TVDI.

O Capítulo 5 apresenta o Mining Knowledge TV, um ambiente de descoberta de conhecimento nos dados da TVDI, bem como, seus módulos e o projeto KTV, arquitetura na qual o MKTV faz parte.

CAPÍTULO

5
Mining Knowledge TV- MKTV

A Web Semântica (W3C *et al.*, 2009) afirma a intenção de fornecer um framework comum que permita o compartilhamento e reutilização de dados em aplicações, além do processamento automático em computadores. Adicionalmente também provê uma linguagem para fornecimento de conexões semânticas entre os dados e os objetos do mundo real.

Ao analisar as áreas de Inteligência Artificial e TV Digital Interativa pode-se observar diversas oportunidades para estender a visão da Web Semântica à plataforma computacional emergente de TV Digital Interativa (TVDI), com o objetivo de contribuir para a construção da TV Semântica (Semantic TV) seguindo a linha e conceitos da Web Semântica.

Neste contexto o projeto *Knowledge TV- KTV* (LINO et al., 2011), entre outros aspectos, investiga uma arquitetura para prover serviços e aplicações multimídia em TV Digital Interativa baseada em conceitos da Web Semântica. Como exemplos de serviços e aplicações propostos que podem ser fornecidos por tal arquitetura, destacam-se: um serviço de recomendação baseado em modelagem semântica; um ambiente de KDD (*Knowledge Discovery in Databases*); e um serviço de consultas semânticas.

5.1 PROJETO *KNOWLEDGE* TV - KTV

Atualmente, a televisão pode ser considerada como o principal meio de transmitir informação e entretenimento à população. Isto acontece devido a sua ampla área de cobertura e facilidade no acesso a aparelhos de TV, ao longo do último século. Especificamente no Brasil,

pesquisas recentes (IBGE *et al.*, 2008) mostram que o percentual de lares que possuem ao menos um aparelho de TV é superior a 92%. Apesar da expansão contínua da Internet, o número de computadores com acesso a rede continua baixo, especialmente em países onde a maior parte da população possui baixo poder aquisitivo. No Brasil, este percentual é inferior a 35%. Diante destes números, é possível observar o importante papel da TV Digital Interativa na difusão da internet possibilitando mudanças na cultura popular pelo oferecimento de novos serviços e oportunidades.

Além disso, é possível a migração de serviços e aplicações da internet, como e-Bank, e-Gov, e-Learning para o domínio da TV, fazendo com que estes serviços alcancem uma maior parcela da população. A quantidade de programas e serviços no domínio da TV tem crescido por conta da possibilidade do envio de mais serviços no mesmo canal de transmissão. De acordo com (LUGMAYR *et al*., 2004), com todas estas mudanças, o mundo da TV passará a enfrentar os mesmo desafios, como complexidade e volume de informação, já enfrentados por outros meios de transmissão de mídias.

Produção de conteúdo de TV Digital Interativa e internet apresentam diferenças em todos os estágios, indo desde a produção até o consumo do conteúdo. Uma destas principais diferenças está na forma de interação com o usuário. Enquanto na internet o usuário precisa informar "o que ele quer" e quais são as suas preferências, na TV, atualmente, o telespectador apenas recebe o conteúdo transmitido e seleciona suas preferências por meio de mudanças de canal até encontrar algo de seu interesse. Dessa forma, é possível diferenciar e entender os usuários de internet como agentes ativos e os telespectadores como agentes passivos em relação à programação oferecida.

Como essa existem outras diferenças inerentes entre a plataforma computacional de TV Digital Interativa e as outras plataformas computacionais existentes, tais como a plataforma computacional de computadores pessoais ou de dispositivos móveis. Desta forma há uma necessidade de investigação a respeito dos novos requisitos inerentes a esta emergente plataforma computacional para que novos serviços e aplicações sejam conceitualizados e desenvolvidos. O projeto *Knowledge TV* tem como objetivo investigar tais requisitos e contribuir para a construção da TV Semântica, inspirada nos conceitos da Web Semântica (W3C *et al.*, 2009).

5.1.1 Arquitetura Conceitual

A plataforma computacional que dá suporta à TV Digital Interativa tem crescido em ritmo acelerado. Pesquisadores e desenvolvedores concentram-se tanto em questões de *hardware* e *software*, quanto na comunicação entre estes dois aspectos, que é solucionado com o desenvolvimento de uma camada intermediária entre ambos, denominada *middleware*.

No contexto da TV Digital Interativa, *middleware* se refere a uma camada intermediaria de software existente no receptor de TV. O mesmo tem por objetivo abstrair os detalhes de hardware das aplicações.

Muito se tem pesquisado acerca do *middleware* de TV Digital Interativa, em especial sobe o middleware do Sistema Brasileiro de TVDI, principalmente no que diz respeito a novas aplicações utilizando os dados da interação do usuário na TV.

O projeto *Knowledge TV* está inserido neste contexto, e para encontrar soluções para estas lacunas, se baseia em técnicas de Representação do Conhecimento (HAN; KAMBER, 2006), Mineração de Dados, *Data Warehouse*, OLAP. O projeto pretende introduzir uma camada semântica à plataforma computacional de TV Digital Interativa apoiada pela infraestrutura da internet. Esta camada semântica avançará o estado da arte desta tecnologia por fornecer uma rica descrição dos recursos e serviços, obtidos por uma abordagem de modelagem semântica. Permitindo a partir dela, o desenvolvimento de serviços e aplicações mais sofisticadas, bem como a criação e ambientes de convergência entre internet e TV Digital Interativa.

A arquitetura conceitual do projeto *Knowledge TV* na perspectiva da plataforma de TVDI é apresentada na Figura 6, abaixo.

Figura 6 - Arquitetura de camadas na perspectiva da TVDI

Fonte: LINO (2011)

Desta forma a arquitetura conceitual do projeto *Knowledge TV* na perspectiva da plataforma de TVDI possui as seguintes camadas: (1) *hardware*, (2) *middleware*, (3) semântica, e (4) aplicações.

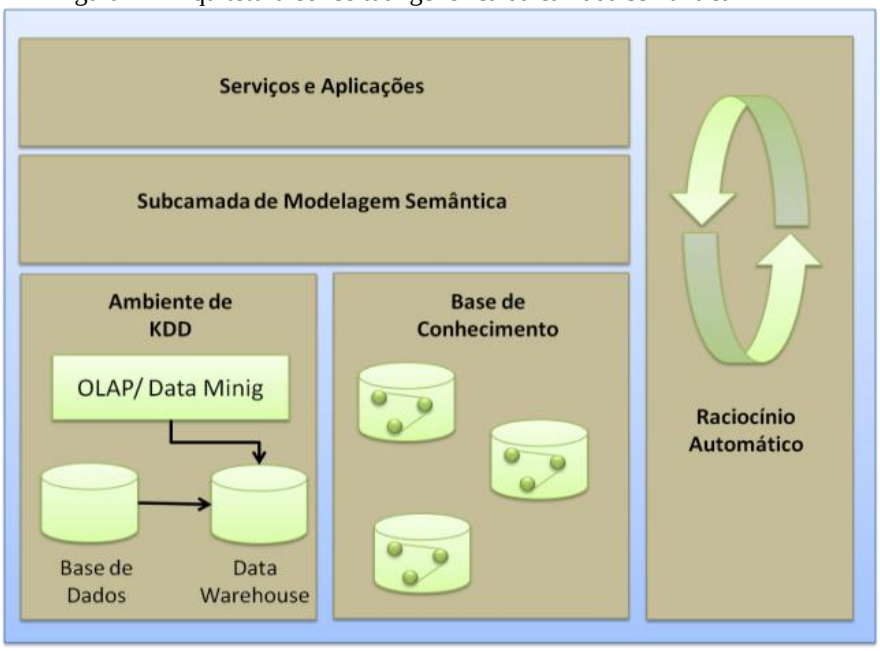
A Camada de *Hardware* concentra todos os aspectos e dispositivos físicos que podem operar em um ambiente de TV Digital Interativa. A Camada de *Middleware* têm a responsabilidade de abstrair as particularidades dos dispositivos de hardware, realizando a comunicação com as camadas superiores, tentando maximizar o desempenho do sistema. A Camada de *Middleware* possui componentes da camada semântica, nela inseridos, com o objetivo de fazer a comunicação com a camada semântica enviando dados necessários para a mesma, além de poder se comunicar diretamente com a camada de aplicação.

A Camada Semântica tem a finalidade de fornecer conhecimento e raciocínio através de uma rica modelagem semântica e base de conhecimento contendo descrição de dados, recursos, serviços, aplicações, e relacionamentos entre estes, etc. Cada uma destas descrições é feita em uma linguagem padronizada formalmente, capaz de ser processada automaticamente por computadores. Na Camada de Aplicação estão aplicações e serviços que utilizam os recursos oferecidos pelas Camadas Semântica e de *Middleware*.

A arquitetura do KTV pode ser concebida de diversas formas. Uma perspectiva mais genérica e que pode ser aplicada a diferentes ambientes, seja ele *Web* ou na TVDI, esta presente na Figura 7. Essa ilustração exibe os principais componentes da visão conceitual genérica da KTV, que são:

- **Ambiente de KDD para TV Digital Interativa:** Aqui estão inseridas às tecnologias relacionadas a um ambiente de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases – KDD*) (HAN; KAMBER, 2006). Este módulo possui: (1) uma base de dados brutos; (2) a modelagem dos dados analíticos multidimensional em um *Data Warehouse – DW* (HAN; KAMBER, 2006), onde serão investigados e levados em conta os requisitos da plataforma computacional de TV Digital Interativa; e (3) um módulo de tecnologias e ferramentas de análise de dados como algoritmos de Mineração de Dados (*Data Mining - DM*) (HAN; KAMBER, 2006) e recursos de *OLAP* (*On-Line Analytical Processing*) (HAN; KAMBER, 2006). Tais tecnologias irão permitir suporte a decisão(SHIM *et al.*, 2002) dentro de um contexto do domínio.

Figura 7 - Arquitetura conceitual genérica da camada semântica



Fonte: LINO (2011)

- **Base de Conhecimento:** Este módulo refere-se às instâncias de armazenamento de dados; modelados de acordo com uma abordagem semântica. Tais instâncias são modeladas de acordo com a especificação contida na subcamada de modelagem semântica. Por exemplo, diferentes ontologias de domínios poderão fornecer conceitos e taxonomias em diferentes níveis e detalhamento.
- **Camada de Modelagem Semântica:** Aqui são encapsulados todos os conhecimentos por meios de métodos e linguagens formais com semântica bem definida que permitirão processamento automático por agentes computacionais. Isto permitirá o desenvolvimento de aplicações e serviços inteligentes e avançados.
- **Módulo de Raciocínio Automático:** Este módulo apoia a todas as operações de Raciocínio Automático (RUSSELL, 2009) efetuadas com base nos modelos de conhecimento fornecidos, e de acordo com os mecanismos de raciocínio planejados.
- **Módulo de Aplicações e Serviços:** Este módulo agrega todas as aplicações e serviços criados no âmbito da arquitetura KTV.

A partir das visões conceituais o projeto KTV está sendo planejado e desenvolvido. Desta forma, os principais objetivos do KTV são: (1) definição de uma arquitetura detalhada da camada

semântica com diversas visões arquiteturais (conceitual, de rede, de integração com *middleware* Ginga, etc.); (2) elaboração de modelagem e descrição semântica de dados operacionais; (3) criação de modelos analíticos semânticos de dados; (4) desenvolvimento de serviço de recomendação semântico; e (5) fornecimento de mecanismos de busca (consultas) de conteúdo de dados que comunicam-se semanticamente.

5.2 MINING KNOWLEDGE TV – MKTV

O objetivo geral desse trabalho de mestrado é a implementação de um ambiente de *KDD* (do inglês *Knowledge Discovery in Databases*) integrado a conceitos semânticos, com ênfase na Mineração de Dados das informações provenientes da TV Digital Interativa. Sendo o mesmo, um dos principais componentes da arquitetura do projeto KTV.

Este ambiente proporcionará aos desenvolvedores de aplicações para TV Digital Interativa diversas facilidades, entre elas, o fato de abstraírem questões de coleta e armazenamento de dados dos usuários de TVDI e de provedores de conteúdo. Além da Mineração de Dados integrada a bases semânticas, que irão organizar semanticamente o conhecimento descoberto através da mineração.

O conhecimento descoberto pela Mineração de Dados poderá ser utilizado na criação de aplicativos que tenham como objetivo solução de problemas nos âmbitos de sobrecarga de informação, personalização, propaganda direcionada, entre outros.

Para o desenvolvimento do projeto MKTV foram observados os seguintes requisitos:

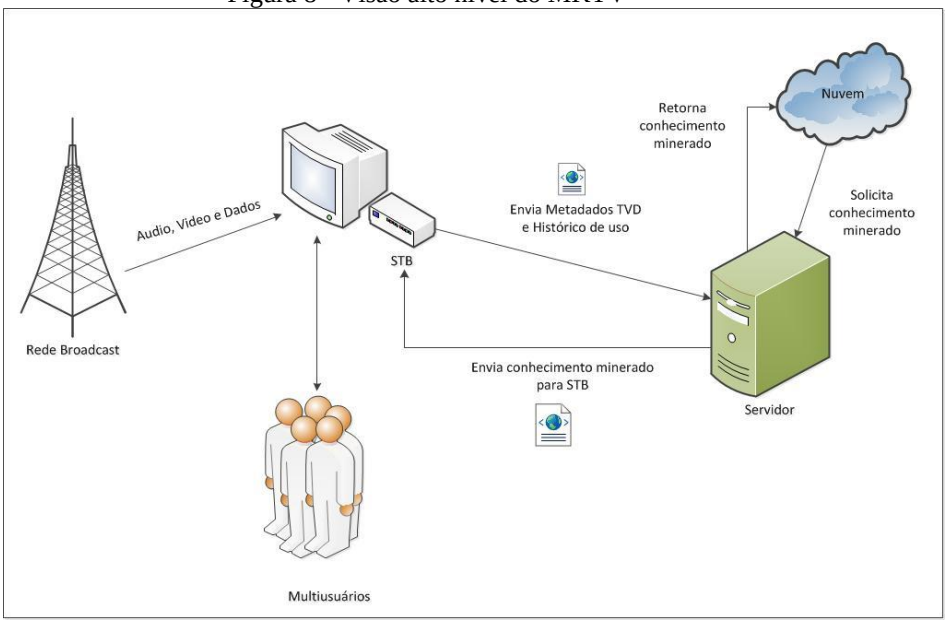
- Ser genérico (de propósito geral), ou seja, preconizar a disponibilização de conhecimento minerado para qualquer sistema, seja ele de recomendação, personalização, entrega de propaganda personalizada, consulta semântica, etc.;
- Disponibilizar conhecimento descoberto após realizar o processo de Mineração de Dados sobre uma base de dados;
- Ser multi usuário, isto é, na perspectiva da TVDI prover descoberta de conhecimento tanto para o usuário final como para o provedor de conteúdo;
- Abstrair o complexo processo de Mineração de Dados que requer uma plataforma física de hardware de grande performance para quem os solicite;

- Ser multi plataforma, podendo ser usado tanto para o uso nos STBs de TV Digital Interativa, quanto para qualquer sistema Web.

Para atender a demanda da TVDI e de sistemas Web, o MKTV, utilizando o conceito de conhecimento como serviço (do inglês KAAS)[4], disponibilizará junto da arquitetura KTV um web service com uma interface de acesso para que os desenvolvedores de sistemas para TVDI possam utilizar o MKTV.

A Figura 8 apresenta uma visão de alto nível do funcionamento básico do MKTV. Em que a rede de transmissão de TVDI envia o áudio, vídeo e os dados para o *Set-Top-Box* (STB) do ambiente de recepção de TVDI. O usuário interage com a TV, por exemplo, zapeando ou usando aplicações. O STB grava localmente e envia os dados do usuário por ele armazenado via canal de retorno (MARGALHO et al., 2007) para o servidor de processamento do MKTV. Em seguida o servidor de processamento do MKTV armazena, descreve e organiza o conteúdo das informações semanticamente através de ontologias, minera os dados, e caso seja solicitado por alguma aplicação, retorna o conhecimento minerado para a mesma que poderá estar em um STB ou na WEB.

Figura 8 - Visão alto nível do MKTV



Fonte: Próprio Autor (2012)

[4] Knowledge as a Service (KaaS) pode ser considerado como um novo paradigma computacional, em que uma determinada informação é requisitada a um provedor de conhecimento e a mesma é entregue como serviço ao consumidor de conhecimento (XU; ZHANG, 2005)

O MKTV abordará a TVDI como um ambiente de convergência digital, isto é, uma ambiente que reúne as características da televisão, internet, telefonia. Neste contexto o sistema comportará requisições web, fornecendo a maior parte da arquitetura como um serviço web, nos quais os serviços e aplicações irão solicitar informações do sistema e receber os conhecimentos minerados pedidos.

A escolha da arquitetura cliente-servidor e de SAAS pela MKTV se deu devido a TVDI no Brasil ser extremamente nova e passível de modificações, já que não está totalmente padronizada. Além disso, os STB em utilização nas TVs no Brasil nos impõem restrições que devemos levar em consideração para a devida modelagem arquitetural. Podemos destacar as seguintes restrições:

- A pequena capacidade de processamento existente no STB;
- Espaço reduzido para persistência de informações;
- Mecanismo de exclusão de aplicações ao mudar de canal, isto é, ao mudar de canal todas as informações sobre uma aplicação será perdida.

Todas essas limitações existentes na arquitetura dos STBs nos levam a utilizar uma arquitetura cliente-servidor. Nessa arquitetura, os componentes da KTV e consequentemente do MKTV que tem maior consumo de processamento e memória estarão executando em um servidor web. Os mesmos se comunicarão com ambientes de TVDI e de convergência em geral via protocolo HTTP (canal de retorno em TVDI mais especificamente, pois é este que implementa o protocolo HTTP em ambiente de TVDI) com os componentes existentes no *middleware*, tornando possível também à conexão entre aplicativos web e o servidor MKTV.

5.2.1 Fontes de dados

Os dados minerados pelo MKTV vem dos metadados contidos nas tabelas de informação de serviços (Service Information - SI) e são enviados via broadcast pelos provedores de conteúdo, chamadas *Event Information Table* - EIT e *Service Description Table SDT* (ABNT; 2007) apêndice I, juntamente com as informações de interação do usuário obtidas através de componentes do KTV junto ao *middleware* Ginga.

Os metadados contidos nas tabelas SI existentes de acordo com o padrão MPEG 2 (ISO; IEC, 1992) usado no SBTVDI (SBTVDI, 2011), são utilizados para representar as informações emitidas pelos provedores de conteúdo de programas de TV, serviços e interação multimídia existentes em toda a plataforma.

Uma das principais tabelas existente é a *Event Information Table* - EIT, que dentre outras informações contém o gênero do programa, subgênero, a data e hora, sinopse do programa, entre outras. Adicionalmente possui estruturas descritoras que servem para introduzir novas informações.

Outra tabela relevante é a *Service Description Table* - SDT que contém dados dos provedores de serviços, popularmente conhecidos como emissoras de TV.

Outra fonte a ser explorada é o comportamento do usuário, ou seja, qual canal assistiu, em que horário e quantos minutos ele consumiu da programação fornecida através da televisão. É importante enfatizar que o projeto KTV provê privacidade aos dados pessoais do usuário, abordagem esta que não faz parte do escopo desta dissertação.

Todas essas informações e fontes de dados de TVDI necessárias ao processo de mineração de dados estão sendo providas ao MKTV, módulo de Descoberta de Conhecimento em Base de Dados do projeto KTV, por meio de uma extensão KTV ao *middleware* Ginga (ARAÚJO, 2011). A mesma agrupa, organiza e envia os dados do *middleware* da TVDI para o servidor, onde esta o MKTV, que os processará.

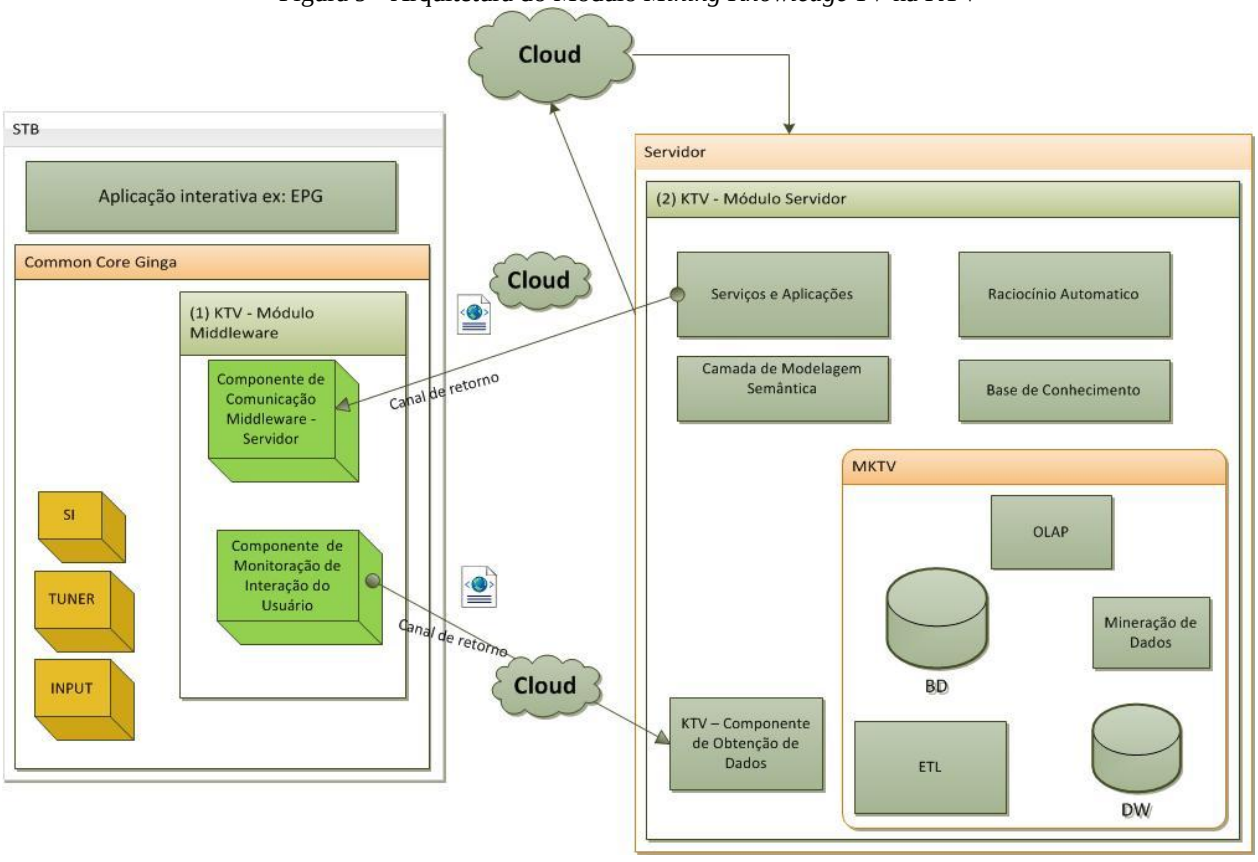
O processo de KDD nessas inúmeras fontes fornece informações novas, que são descritas semanticamente por meio de uma ontologia, possibilitando uma melhor expressividade e explicitando ainda mais o significado presente nesse conjunto de novos dados que foram resultado da Mineração de Dados. Em seguida, o conhecimento descoberto é fornecido como um serviço, que é utilizado principalmente para desenvolvedores de aplicações NCL ou Java para TVDI, dotando-os de informações relevantes para o projeto e desenvolvimento de inúmeras soluções.

Ainda é possível utilizar dados externos, que venham através da internet, como fonte de dados. Estes vem por meio de uma interface do MKTV disponibilizada para recepção de dados externos e poderiam complementar informações da TVDI ou mesmo ser utilizada como fonte principal para descobrir conhecimentos úteis nesses dados.

5.2.2 Descrição da Arquitetura

A arquitetura detalhada do KTV e MKTV mais especificamente, baseada na especificação conceitual, pode ser observada na Figura 9. Nela estão explícitos todos os módulos da arquitetura KTV: módulo (1)*Middleware* – Componente de Comunicação *Middleware*-Servidor, Componente de Monitoração do Usuário; (2) módulo Servidor – Componente Serviços e Aplicações, Componente de Raciocínio Automático, Componente de Base de Conhecimento, Componente de Obtenção de Dados, MKTV, Camada de Modelagem Semântica e OLAP. Além de todos os componentes do núcleo do *middleware* Ginga (LAVID; TELEMIDIA, 2008) que interagem com o KTV: SI, Turner, Input, EPG.

Figura 9 - Arquitetura do Módulo *Mining Knowledge TV* na KTV



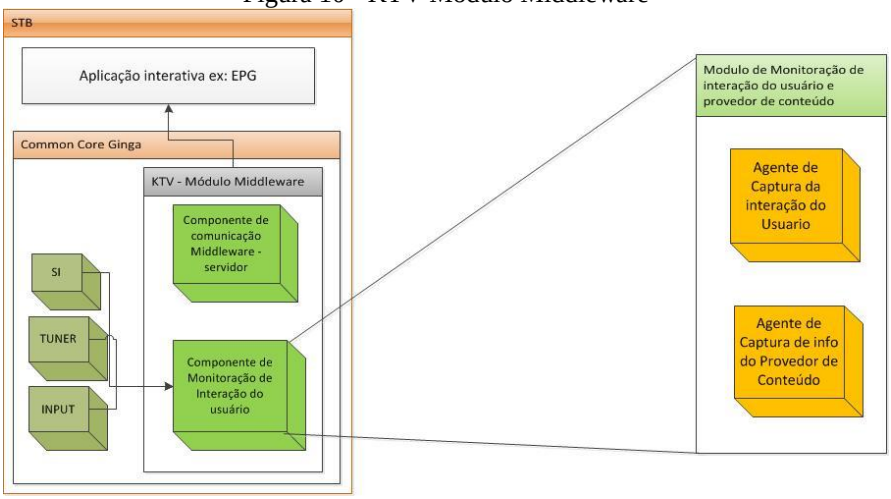
Fonte: Próprio Autor (2012)

O módulo *middleware* (Figura 10) executa no STB, o mesmo, estende o *common core* do *middleware* da TV Digital Interativa Brasileira Ginga (ARAÚJO, 2011). Essa extensão se dá

através do acréscimo dos componentes de Monitoração de Interação do Usuário e de Comunicação Middleware-Servidor ao *common core* do Ginga.

O Componente de Monitoração de Interação do Usuário tem comunicação direta com os componentes: tabelas SI, Turner e Input do *common core* do *middleware*, e é definido por um Agente de Captura de Interação do Usuário – ACIU e por um Agente de Captura de Informação do Provedor de Conteúdo - ACIPC.

Figura 10 - KTV Módulo Middleware



Fonte: Próprio Autor (2012)

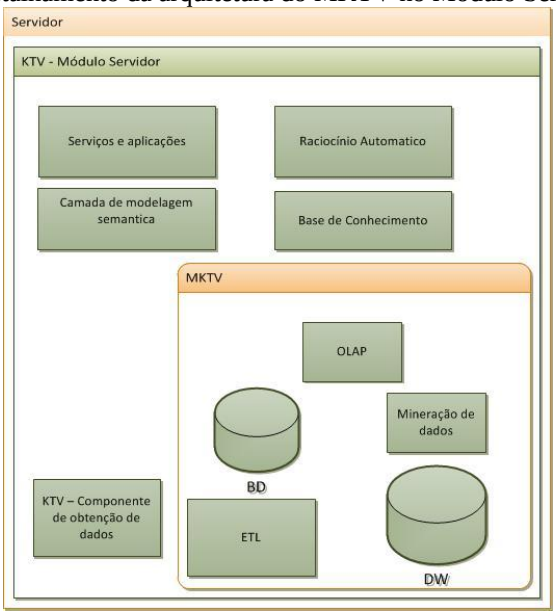
O ACIU é o agente responsável por monitorar e capturar as ações realizadas pelo usuário, assim como o horário, o dia da semana, o provedor de conteúdo (canal de TV) sintonizado, o serviço sintonizado (aplicativo interativo), além das informações do programa escolhido como nome, gênero, subgênero, a faixa etária e a sinopse. Essas ações realizadas pelo usuário, através de mecanismos de interação como o controle remoto, são executadas no *middleware* pelos componentes *Turner* e *Input*. O *Turner* é o componente do *middleware* responsável pela troca de canais, chamada também de ação de *zaping*. O *Input* é responsável por gerenciar ações de entrada vindas do controle remoto. Já as informações relativas aos programas vêm das tabelas de SI, que tem em seu corpo de descritores os dados a cerca da programação a ser exibida. Depois de armazenadas as informações de interação em um arquivo, o mesmo será enviado para o módulo servidor do KTV através do canal de retorno.

O ACIPC é o agente responsável por capturar todas as informações relativas ao provedor de conteúdo existente nas tabelas de SI do *middleware*, armazená-las e enviá-las para o Servidor do KTV via canal de retorno.

O Componente de Comunicação *Middleware*-Servidor é responsável pelo transporte e decodificação das recomendações enviadas pelo ACIPC ao módulo Servidor. Outra de suas atribuições é realizar a comunicação e o abastecimento de informações recomendadas às aplicações residentes no *common core* como, por exemplo, o EPG, componente do *Middleware* responsável pela exibição de um guia de programação na TV. Por fim realiza a comunicação com aplicações interativas, que estejam na camada acima do *common core* do Ginga e que venha a utilizar a interface do KTV.

O MKTV pertence ao ambiente de KDD modelado na arquitetura conceitual e especificado no módulo servidor (Figura 11) que executa em um servidor *Web*. O mesmo é composto por um banco de dados operacional, um componente de Extração, Transformação e Carga (do inglês *Extract, Transformation and Load* - ETL), um *Data Warehouse* (HAN; KAMBER, 2006) e um engenho de Mineração de Dados.

Figura 11 - Detalhamento da arquitetura do MKTV no Modulo Servidor do KTV



Fonte: Próprio Autor (2012)

O banco de dados operacional tem a tarefa de armazenar todos os dados recentes referentes à interação do usuário, provedor de conteúdo e todos os dados que vem do STB, além de dados externos que vem através da internet.

O componente de ETL realiza o processo de extração, transformação e carga dos dados operacionais para o modelo multidimensional de um *Data Warehouse*. O componente de software de ETL tem a função de extrair os dados dos sistemas de TV Digital Interativa, transforma-los de acordo com o modelo do *Data Warehouse* e carregá-los nos *Data Marts.*

O *Data Warehouse* é responsável pelo armazenamento de dados históricos, integrados, e não voláteis advindos do ambiente de TV Digital Interativa. Ele é um banco de dados modelado de forma multidimensional (HAN; KAMBER, 2006) e é muito utilizado no processo de descoberta de conhecimento, pois trabalha com dados históricos possibilitando uma série de análises acerca dos eventos passados, além de maior precisão e rapidez na tomada de decisão. O *Data Warehouse* está, no contexto deste projeto, organizado em *data marts*, que são subconjuntos de dados de um *Data Warehouse*, de acordo com o assunto a ser minerado, por exemplo: personalização, marketing, business, etc.

O engenho de Mineração de Dados é responsável pela descoberta de padrões úteis e desconhecidos através da aplicação de algoritmos baseados em técnicas de inteligência artificial e estatística.

Após a realização da Mineração de Dados o conhecimento descoberto será retornado atualizando as ontologias presentes na base de conhecimento de acordo com a modelagem especificada na camada de modelagem semântica.

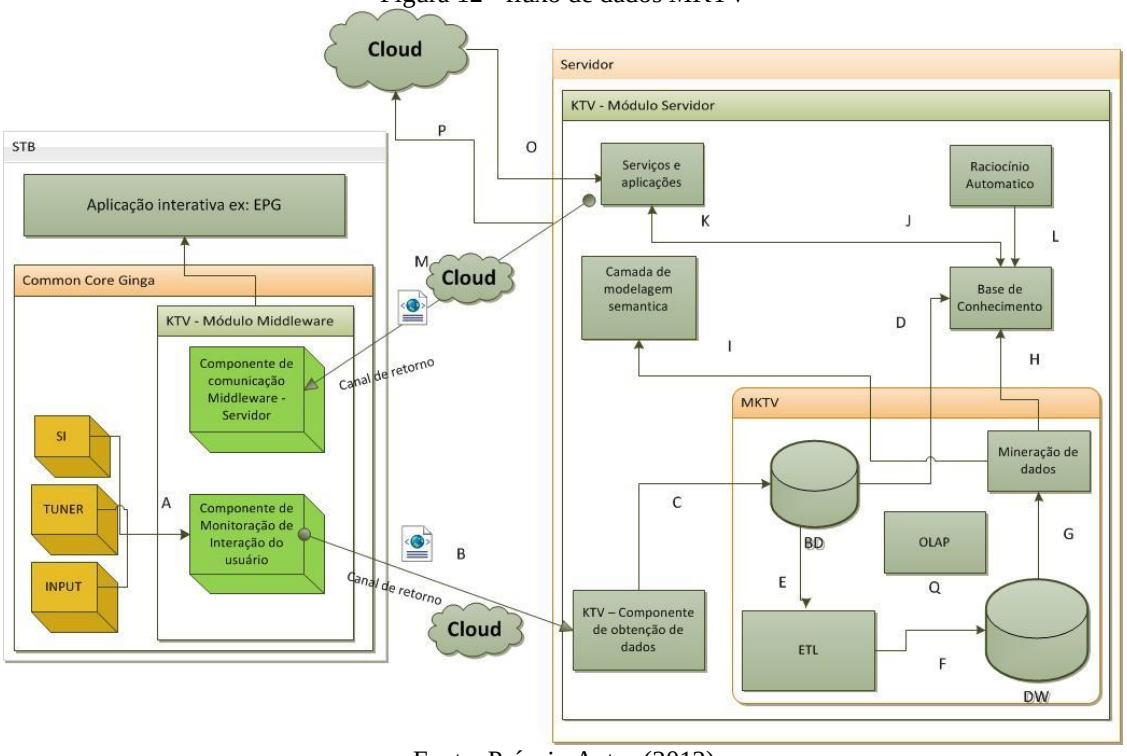
Por ser um projeto acadêmico o MKTV opta por utilizar ferramentas freeware. Neste caso é utilizado banco de dados Postgresql (POSTGRESQL, 2011) para o armazenamento das informações, sendo as mesmas modeladas relacional e multidimensionalmente, para os casos do banco de dados operacionais e *Data Warehouse*, respectivamente.

O software de Mineração de Dados escolhido foi o WEKA (WEKA, 2008). O mesmo é uma ferramenta muito utilizada em pesquisas no meio acadêmico devido a grande quantidade de algoritmos implementados e que dão suporte a diversas tarefas de mineração. Além dessa característica o WEKA é um software livre implementado na linguagem de programação JAVA (JAVA, 2012), o que nos da oportunidade de utilizar uma Aplication Program Interface - API multiplataforma, testada e com fácil documentação de acordo com a necessidade do projeto.

5.2.3 Detalhamento da Arquitetura

Como pode ser observado na Figura 12, os dados referentes às visualizações do usuário, dos programas, etc., são capturados pelo Componente de Monitoração de Interação do Usuário (A). Em seguida, os dados são por ele enviados via canal de retorno para o módulo Servidor do KTV(B). Os dados são recepcionados pelo componente de obtenção de dados e levados para ser armazenados em um banco de dados operacional local (C). Depois de armazenados, os esquemas de dados relacionais são descritos no componente de modelagem semântica e armazenados semanticamente no componente base de conhecimento. (D) Eles ainda sofrem progressivamente o processo de extração, transformação e carga (ETL) das informações (E) para o Data Warehouse – DW (F), onde os dados estão modelados multidimensionalmente em cubos de dados de acordo com o contexto e o propósito que são analisados.

Figura 12 - fluxo de dados MKTV



Fonte: Próprio Autor (2012)

Com os dados organizados no DW, o módulo de Mineração de Dados executa algoritmos, buscando e descobrindo padrões úteis e não conhecidos nas informações existentes do DW(G) de

acordo com o domínio do problema que se quer resolver. Por exemplo, caso o domínio seja entrega de propaganda personalizada, é escolhido à tarefa de Mineração de Dados mais adequada para solucionar o problema em questão, bem como outras primitivas de Mineração de Dados como algoritmo, limites de suporte e confiança, etc.

De posse dos conhecimentos extraídos, o MKTV envia-os para a camada de modelagem semântica (I), onde ficam os modelos (ontologias) do KTV, especificados através de Ontologias OWL (W3C, 2011) e para a base de conhecimento (H), onde ficam as instâncias de acordo com os modelos (ontologias).

O componente de serviços e aplicações tem diversas aplicações, entre elas, uma de recomendação de programação ao usuário de TVDI. Essa aplicação pode se beneficiar do conhecimento descoberto, solicitando através de uma API a base de conhecimento novas informações para serem disponibilizadas para o usuário (J) (K).

Outra possibilidade é solicitar algum serviço do MKTV via protocolo HTTP pela Web, ou seja, algum programa que deseje utilizar de serviços de Mineração de Dados poderá via Web Services utilizar a API do KTV.

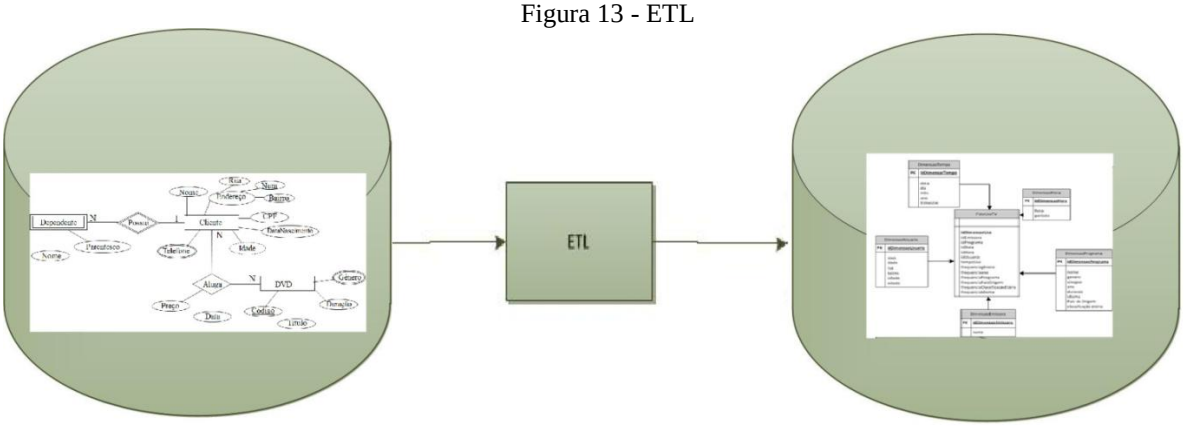
Para o componente OLAP (Q), não foi implementado API ou ferramenta que o execute, ficando como trabalho futuro a implementação de uma ferramenta de OLAP. Apesar de esse componente ser contemplado por meio do modelo multidimensional do DW, que possui em suas análises suporte a consultas OLAP, mesmo não tendo uma ferramenta específica para executá-las.

Os principais conceitos que embasam os módulos do MKTV são: a Extração, Transformação e Carga; o DataWarehouse; Ontologias (OntoMKTV); Mineração de Dados e Algoritmo de mineração de dados. A seguir, é descrito como esses conceitos são utilizados no MKTV.

5.2.3.1 Extração, Transformação e Carga

O processo de ETL é empregado no deslocamento dos dados operacionais localizados no banco de dados relacional de dados operacionais do KTV para o banco de dados multidimensional do projeto, o Data Warehouse. Como ilustrado na Figura 13.

Figura 13 - ETL



Fonte: Próprio Autor (2012)

O projeto KTV tem buscado utilizar em seu desenvolvimento ferramentas de livre utilização, e com o processo de ETL não seria diferente. Para a realização do ETL é utilizado a ferramenta livre Kettle (KETTLE, 2011), que extrai as informações do banco de dados, as transforma e carrega no Data Warehouse do projeto periodicamente.

O software Kettle foi escolhido por possuir atributos importantes ao domínio do projeto, como suporte a XML(W3C (c), 2011), paralelismo de tarefas, acesso ao código fonte, ser uma ferramenta estável e com anos de desenvolvimento, ter documentação disponível, ser multiplataforma; além de possuir clareza nos *logs* e rastreamento de exceções.

5.2.3.2 Data Warehouse - DW

O conceito de *Data Warehouse* - DW (HAN; KAMBER, 2006) surgiu no intuito de suprir a demanda por ferramentas que analisassem grandes bases de dados, identificando tendências e comportamentos. Assim então, acabando por permitir ao gestor de informação tomar suas decisões pautadas no conhecimento contido através de correlação entre os dados.

Para Immon (1995), o *Data Warehouse* disponibiliza uma série de informações utilizadas no apoio à tomada de decisão estratégica, sendo este a implementação física de um modelo de apoio à decisão.

Segundo Kimball (1998), a construção de um DW se dá através do processo de combinar as necessidades de informações de uma comunidade de usuários com os dados que realmente estão disponíveis.

Através dos esquemas multidimensionais os DW ganham a estrutura para prover conhecimento. Desta forma, o banco de dados pode ser observado como um cubo de dados com n dimensões. As técnicas de modelagem relacionais, mais conhecidas e disseminadas, também podem ser usadas na modelagem de esquemas multidimensionais. Neste sentido, os DW's podem ser modelados através de esquemas, que são: estrela, floco de neve ou constelação de fatos.

O projeto KTV tem em seu escopo a criação de um *Data Warehouse* para dar suporte a serviços de recomendação, consulta semântica, Mineração de Dados, entre outros. O esquema usado para modelar o DW foi o esquema estrela(KIMBALL, 1998), que consiste de uma tabela para cada dimensão e uma tabela de fatos relacionada com as mesmas.

Os conceitos utilizados para modelar o domínio de TVDI no DW do MKTV vieram do modelo ontológico CoreKTV (Araújo, 2011), que tomou como base a norma internacional de telecomunicações J200(ITU,2001), J201(ITU,2003), J202 (ITU,2004). Desta forma, o modelo núcleo multidimensional do MKTV é composto pelas seguintes dimensões:

- Usuário – descreve informações relativas ao usuário, como localidade que reside, além de sexo e idade do mesmo;
- Programa – descrimina os dados acerca do conteúdo exibido pelas emissoras, ou seja, os programas de TV;
- Emissora – descreve informações sobre emissora de TV;
- Tempo – dimensão que descreve as informações sobre o tempo em que foi realizada determinada transação no modelo, exemplo: meses, anos, data;
- Hora – determina os dados acerca de hora, como período do dia: manhã, tarde, noite e madrugada.

A partir do detalhamento de cada dimensão, com base no CoreKTV, pode-se encontrar os atributos que constituem as tabelas, estas são a forma física de cada dimensão e tabela de fatos no *Data Warehouse*. Neste sentido, as dimensões e a tabela de fatos são compostas pelos seguintes atributos:

- Dimensão Usuário: sexo, idade, rua, bairro, cidade, estado;
- Dimensão Programa: nome do programa, gênero, subgênero, sinopse, ano da produção, duração, idioma, país de origem, classificação etária, classificação tempo estréia, classificação duração;
- Dimensão Emissora: nome da emissora;

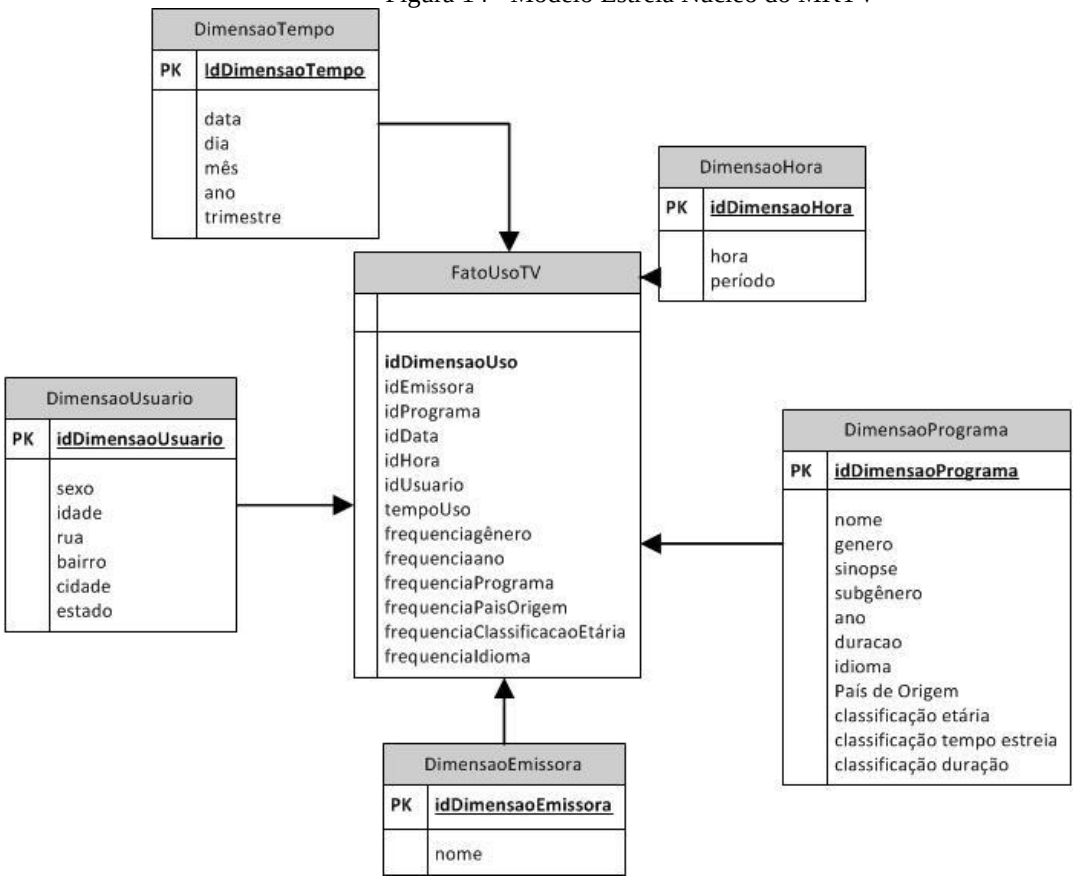
- Dimensão Tempo: data, dia, mês, ano, trimestre;
- Dimensão Hora: hora, turno;
- FatoUsoTV: tempo assistido, frequência gênero, frequência ano, frequência programa, frequência país de origem, frequência classificação etária, frequência idioma.

A Figura 14 ilustra o modelo núcleo do DW do MKTV, podemos chamar de modelo núcleo, pois extensões podem ser propostas.

O DW do MKTV foi modelado para dar suporte a uma série de análises, além de auxiliar os serviços de consulta semântica e de recomendação de conteúdo, serviços estes sendo oferecidos inicialmente pela KTV. As principais consultas (análises) multidimensionais/OLAP suportadas através do modelo multidimensional núcleo do DW do MKTV são:

- Qual o canal com maior índice de audiência em determinada faixa de horário?
- Qual o programa de maior audiência por horário em determinado canal?
- Qual a classificação etária do programa de maior audiência por horário?
- Qual a duração do programa de maior audiência por horário?
- Qual o gênero de programa de maior frequência na programação semanal?
- Quantas aparições semanais determinado programa teve na grade de programação.
- Quais os turnos do dia (manhã, tarde, noite, madrugada) que um programa de classificação maior de 14 anos passa?
- Qual a audiência de programas com a classificação maior de 12,14,16,18; livre por emissora?
- Quanto tempo a emissora da a programas de classificação livre?
- Quantos programas existem com mais de uma hora de duração (media de duração)?
- Quais gêneros de programas com mais de uma hora de duração, com uma hora, com menos de uma hora de duração?
- Qual o turno de maior quantidade de usuários assistindo TV?
- Qual o Turno de maior e menor Uso da TV?

Figura 14 - Modelo Estrela Núcleo do MKTV



Fonte: Próprio Autor (2012)

Todas essas análises são utilizadas em diversas perspectivas, elas são interessantes para o provedor de conteúdo analisar e planejar sua grade de programação, de forma a tornar sua grade mais atrativa para o usuário. Além do provedor de conteúdo, o anunciante de TV também se beneficia com essas informações, podendo direcionar sua propaganda em horários que contemplem um maior público alvo de seus produtos.

Ainda assim, as análises não se limitam a só fornecer conteúdo rico de informações interessantes ao provedor de conteúdo e/ou anunciante de TV, pois o conhecimento revelado através das análises podem auxiliar o usuário, ou serviços (aplicações) automatizados, na busca por programas mais interessantes. Caso o usuário deseje saber qual o programa mais assistido naquele momento; ou então, se ele tiver com pouco tempo para assistir TV e desejar saber o programa de menor duração no momento; ou ainda, se o telespectador tiver restrição quanto a classificação etária do programa poderá encontrar programas de acordo com a classificação do filme.

Esse modelo multidimensional também é útil no que diz respeito aos serviços ofertados pelo KTV. Para o serviço de consulta semântica, o modelo informa os gêneros mais assistidos da programação, e então, através de uma busca poderá ser informado ao usuário, os programas que passarão no horário com o gênero que é o mais popular no momento.

5.2.3.3 Ontologia OntoMKTV

Para Studer (1998), uma ontologia é um entendimento comum de um domínio que pode ser compartilhado entre pessoas e computadores. Neste sentido, uma ontologia pode ser utilizada no contexto do MKTV para estruturar e organizar o conhecimento manipulado nas diversas fases da Mineração de Dados.

Alguns estudos (GOTTGTROY et al.,2004), (NIGRO; CÍSARO; XODO, 2008) (DIAMANTINI ET ALL., 2009A) (DIAMANTINI ET ALL., 2009B) estão sendo realizados unindo as áreas da Mineração de Dados e de ontologia. Eles estão organizados em dois frameworks, da seguinte forma:

- Das ontologias para a Mineração de Dados – As ontologias irão inserir conhecimento ao processo de Mineração de Dados.
- Da Mineração de Dados para as ontologias – O conhecimento minerado é representado em ontologias.

Neste projeto utilizaremos a abordagem da Mineração de Dados para as ontologias, onde o conhecimento proveniente do processo de Mineração de Dados será representado por meio de uma ontologia, chamada OntoMKTV (onto – ontologia e MKTV – projeto em questão), tornando assim o conhecimento descoberto inter operável através de um modelo consolidado - OWL, extensível, podendo agregar novas informações as já descobertas.

A ontologia esta focada na representação do resultado da Mineração de Dados realizada pelos algoritmos da tarefa de regras de associação, feita nos dados do *Data Warehouse* provenientes do ambiente de convergência construído na KTV(TV Digital Interativa e WEB).

As principais vantagens da utilização de uma ontologia para o resultado da Mineração de Dados dizem respeito a:

- Análise das informações mineradas, identificando similaridades e conceitos comuns;

- Reusabilidade do conhecimento encontrado através da mineração de dados aplicável em outros domínios, como o de consultas semânticas, personalização de conteúdo;
- Possibilidade de extensão do conhecimento descoberto através da Mineração de Dados, agregando novas informações;
- Realizar inferências e suposições acerca das informações mineradas;
- Acrescentar um valor semântico ao conhecimento descoberto através da Mineração de Dados, fazendo com que o mesmo deixe de representar um dado de valor unicamente sintático;
- Compartilhar informações acerca da estrutura de informação utilizada para descrição do conhecimento descoberto através das técnicas de Mineração de Dados.

A OntoMKTV encapsula o conhecimento gerado pela tarefa de Mineração de Dados de regras de associação, isto é, para os algoritmos de regras de associação, podendo em um novo trabalho ser estendida para outras tarefas como de clusterização e de classificação.

Seguindo a arquitetura do projeto KTV, a ontologia é especificada no componente de modelagem semântica e mantida (instâncias das regras mineradas) no módulo base de conhecimento, auxiliando em diversos serviços desenvolvidos no projeto.

Para realizar a criação de uma ontologia se faz necessário uma metodologia que guie o seu desenvolvimento (GRUNINGER e FOX, 1995), (USCHOLD e KING, 1995), (FERNANDEZ et al., 1997), (NOY e McGUINNESS, 2001). Neste caso, optou-se pela Metodologia 101 criada por Noy e McGuiness (2001) da Universidade de Stanford por ser uma metodologia já estabelecida e ter sua implementação e documentação baseada na ferramenta Protegé (PROTEGÉ,2012).

A Metodologia 101 prega a construção de ontologias num processo iterativo de sete passos, são ele: determinar o escopo da ontologia, considerar o reuso, listar termos, definir classes, definir propriedades, definir restrições e criar instâncias.

Inicialmente foi definido o escopo da ontologia, que neste caso se trata de representar o conhecimento minerado através das regras de associação. Em seguida foram pesquisadas ontologias sobre o domínio de Mineração de Dados, para não recriar algo que já exista e que possa ser reaproveitado. Desta forma foram encontradas as ontologias: DMOP – *Data Mining Optimization*(http://www.dmo-foundry.org/) (HILARIO et. Al., 2011), KDDOnto

(DIAMANTINI et. Al., 2009) que contêm diversos conceitos sobre o domínio de Mineração de Dados, mas não apresentam conceitos específicos acerca de regras de associação.

Nesse sentido o levantamento dos termos mais relevantes ao seu domínio. Neste caso o OntoMKTV tem como principais conceitos:

- Algoritmo de mineração (Predictive Apriori, Tertius);
- Tarefa de mineração (Regras de associação);
- Regra;
- Termo antecedente;
- Termo consequente;
- Conceito;
- Instância;
- Suporte;
- Confiança.

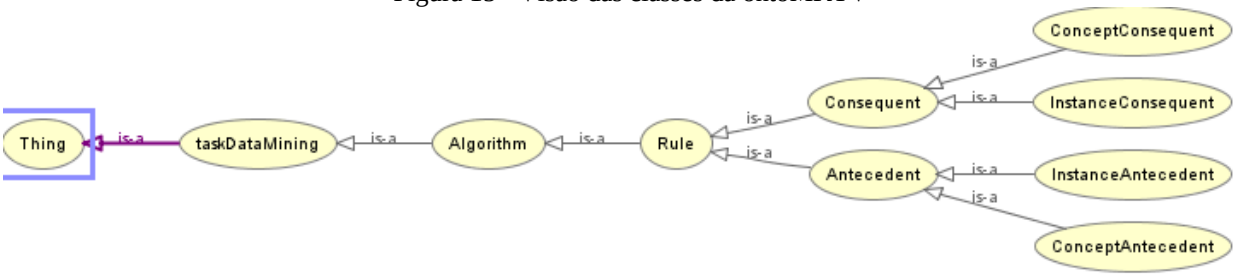
Esses termos são o princípio para modelagem de classes, propriedades e instâncias da OntoMKTV, pois ela busca modelar o contexto de pós-processamento do processo de KDD no que diz respeito às regras geradas. Essas são o resultado da aplicação dos algoritmos de mineração através da tarefa de regras de associação nos dados advindos, no âmbito do projeto KTV, da TV Digital Interativa.

Desta forma as principais classes e grupo de classes modeladas, são (Figura 15 e 16):

- TaskDataMining: Classe que especifica a categoria de padrões que deverão ser encontrados.
- Algorithm: Um algoritmo em geral, é uma sequência bem definida de passos que especifica a forma de resolver um problema ou executar uma tarefa. Normalmente um algoritmo aceita uma entrada e produz uma saída. Um algoritmo de DM é um algoritmo que foi projetado para executar uma das primitivas de DM, como a seleção de recurso, imputação valor em falta, ou modelagem (ou indução).
- Rule: Diz respeito a uma regra de associação criada após a aplicação do algoritmo de mineração de dados. Uma regra é gerada por um algoritmo e é composta por termos antecedente e consequente. Cada regra possui suporte e confiança, que são medidas de corretude da regra.

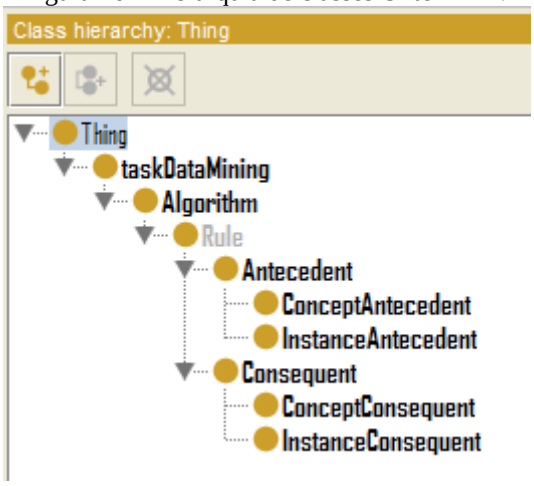
- Antecedent: Descreve o termo antecedente de uma regra no formato X→Y, em que X e Y são formados por um ou mais itens compostos por conceito = instância.
- Consequent: Descreve o termo consequente de uma regra no formato X→Y, em que X e Y são formados por um ou mais itens compostos por conceito = instância.
- Concept: Representa um atributo minerado através da ferramenta de mineração. Este atributo é genérico o suficiente para ser uma classe ou conceito de outra ontologia.
- Instance: Classe que se refere ao valor de um atributo ou a instância de um conceito, pode se remeter também a uma tupla em um banco de dados.

Figura 15 - Visão das classes da ontoMKTV



Fonte: Próprio Autor (2012)

Figura 16 - Hierarquia de classes OntoMKTV



Fonte: Próprio Autor (2012)

Para realizar a conexão entre os conceitos da OntoMKTV foram criadas propriedades, retratando assim as ligações entre as entidades do mundo real, são elas:

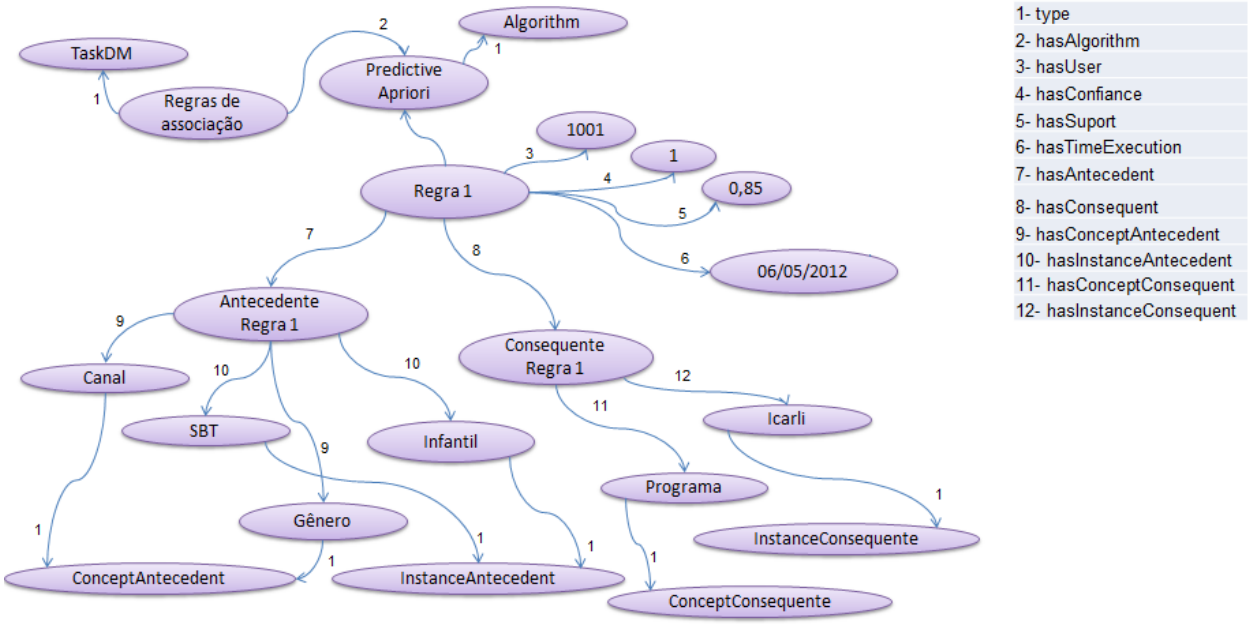
- hasTaskDataMining: possui a classe *Algorithm* como domínio e *TaskDM* como contradomínio.
- hasAlgorithm: possui a classe *TaskDM* como domínio e *Algorithm* como contradomínio.
- hasTimeExecution: possui a classe *Algorithm* como domínio e *TaskDM* como contradomínio.
- hasRule: possui a classe *Algorithm* como domínio e *Rule* como contradomínio.
- isProvidedBy: possui a classe *Rule* como domínio e *Algorithm* como contradomínio.
- hasAntecedent: possui as classes *Rule* e *Consequent* como domínio e *Antecedent* como contradomínio.
- hasConsequent: possui a classe *Rule* e *Antecedent* como domínio e *Consequent* como contradomínio.
- hasConceptAntecedent: possui as classes *Antecedent* e *InstanceAntecedent*, como domínio e *ConceptAntecedent* como contradomínio.
- hasConceptConsequent: possui as classes *Consequent* e *InstanceConsequent* como domínio e *ConceptConsequent* como contradomínio.
- hasInstanceAntecedent: possui as classes *Antecedent* e *ConceptAntecedent* como domínio e *InstanceAntecedent* como contradomínio.
- hasInstanceConsequent: possui as classes *Consequent* e *ConceptConsequent* como domínio e *InstanceConsequent* como contradomínio.
- hasUser: possui a classe *Rule* como domínio e *Algorithm* como contradomínio.
- hasConfiance: possui a classe *Rule* como domínio e *Algorithm* como contradomínio.
- hasSuport: possui a classe *Rule* como domínio e *Algorithm* como contradomínio.

Um exemplo de instância que poderá ser representada pela OntoMKTV é a seguinte (Figura 17):

- Tarefa de Mineração de Dados - Regras de associação
- Algoritmo de Mineração – Predictive Apriori
- Executado dia 06/05/2012

- Encontra como regra numero 1
- Se (conceitoAntecedente) gênero = (instânciaAntecedente) infantil e (conceitoAntecedente) canal = (instânciaAntecedente) SBT → (conceitoAntecedente) programa = (instânciaAntecedente) Icarli
- Usuário número 1001

Figura 17 - Representação do exemplo em forma de ontologia



Fonte: Próprio Autor (2012)

5.2.3.4 Mineração de Dados no MKTV

O processo de Mineração de Dados no projeto MKTV se dá inicialmente nas informações de programa que são: nome, gênero, subgênero, faixa etária e a sinopse, vindas dos metadados contidos nas tabelas SI existentes no Middleware da TV Digital Interativa, juntamente com as informações de interação do usuário na TVDI obtidas pelo ACIU (horário, o dia da semana, o provedor de conteúdo sintonizado, quantidade de tempo sintonizado no provedor de conteúdo).

Após chegarem ao servidor e serem armazenadas no banco de dados, as informações passam pelo processo de ETL e são armazenadas no DW, em seguida elas chegam ao módulo de Mineração de Dados onde são transformadas para o formato ARFF, padrão próprio da ferramenta usada no projeto (WEKA). A mesma também possui uma conexão via drive ODBC (Open

Database Connectivity) para utilização direta com o *Data Warehouse*, mas o projeto se encontra descontinuado, motivo este para a opção pela transformação em ARFF.

Em seguida são usados alguns algoritmos de regra de associação que utilizam técnicas de Inteligência Artificial e da área da Estatística para extrair o conhecimento implícito dos dados. Os principais algoritmos usados nesta primeira versão do MKTV serão: *Predictive Apriori* (SCHEFFER, 2004) *e Tertius* (FLACH; LACHICHE, 2001).

Os conhecimentos que os algoritmos de regras de associação deverão encontrar se concentram especificamente em traçar um perfil de uso do usuário (device set-top-box) na programação da TVDI. Além de identificar preferências e padrões de utilização do usuário de TVDI, como programas adequados aos gostos do usuário.

Esses conhecimentos descobertos ajudarão na implementação de vários produtos de software, como por exemplo, EPG's personalizados para TVDI, aplicações que utilizam consulta semântica, etc..

5.2.3.5 Algoritmos no MKTV

O projeto MKTV utiliza algoritmos de regras de associação para minerar os dados da TVDI, devido sua melhor performance, menor tempo de execução, maior número de regras encontradas, etc., em comparação com algoritmos de outras tarefas como clusterização e classificação (ÁVILA; ZORZO, 2009) no processamento das informações relativas a recomendação em TV e maior corretude na descoberta de regras. Os algoritmos utilizados são: Predictive Apriori e Tertius, com o objetivo de encontrar associações nos dados em que a presença de um conjunto de itens implica em outros itens.

As informações necessárias para a aplicação dos algoritmos são os atributos das dimensões do DW do projeto. Neste sentido os principais atributos utilizados para encontrar conhecimento útil, devido a questões importantes como menor granularidade, maior amplitude do conceito são:

- dia da semana – dimensão Tempo,
- turno – dimensão Hora,
- canal, gênero, subgênero, idioma, país de origem, classificação etária, classificação tempo estréia da produção, classificação duração – dimensão Programa

- tempo assistido, frequência gênero, frequência ano, frequência programa, frequência país de origem, frequência classificação etária, frequência idioma – FatoUsoTV;

Alguns exemplos de regras geradas pelo processo de Mineração de Dados podem ser formuladas/levantadas para ilustrar o potencial do processo no domínio de TVDI.

- Se usuário, ou grupo de usuários, ou região geográfica assiste Canal A, então assiste programas de duração (curta, média ou longa);
- Se usuário assiste Canal B então assiste programas que foram originados ('de 5 até 10 anos ',‘ atual', 'mais 20 anos', 'de 10 até 20 anos‘ , 'até 5 anos');
- Se usuário liga a TV na Quarta-feira (dia da semana) então assiste programas infantis (gênero) durante a noite (período);
- Se usuário liga a TV na Terça-feira (dia da semana) então também liga a TV no Sábado (dia da semana) de madrugada (turno) para assistir programas em inglês (idioma);
- Se usuário assiste comédia (gênero) então o (país de origem do programa) é EUA.

O conhecimento descoberto é tratado de forma semântica por meio de um encapsulamento através da ontologia utilizada no projeto, onde é compartilhado para todos os outros módulos do KTV.

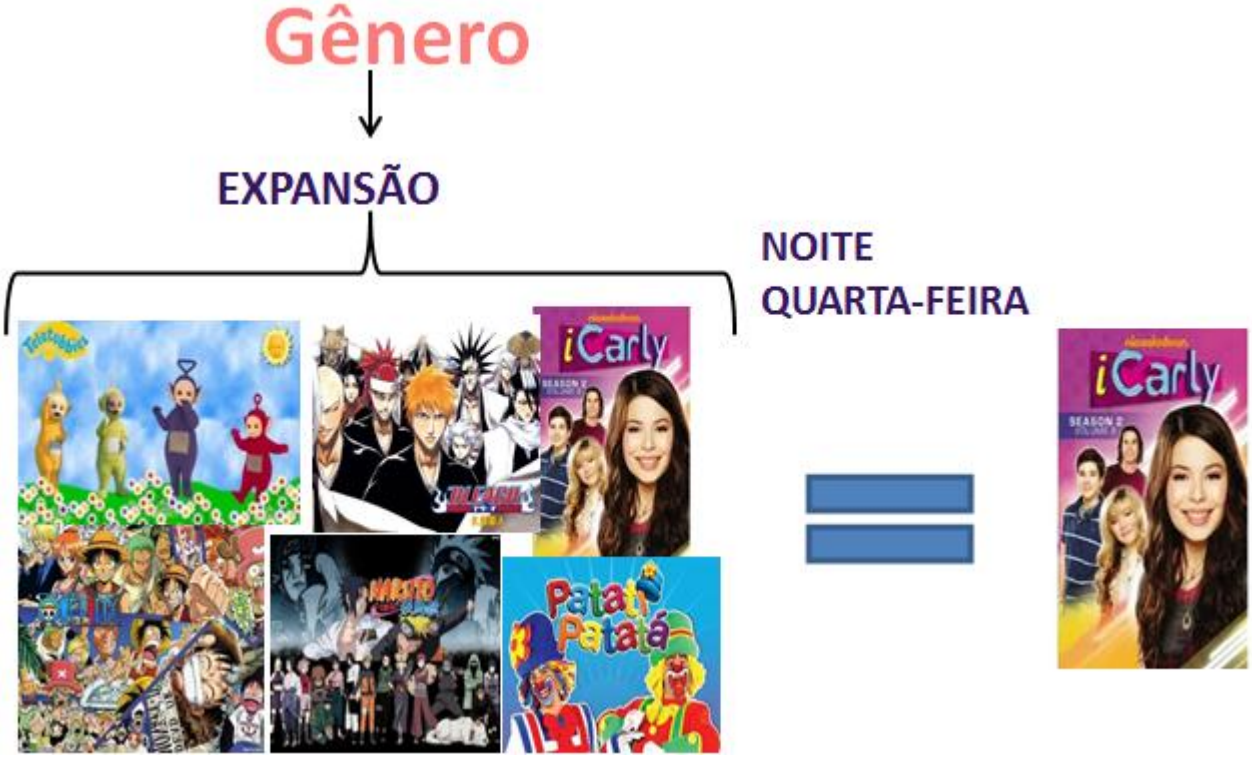
Os padrões são extraídos e encapsulados na ontologia. Desta maneira permitindo: a interoperabilidade do conhecimento descoberto; o reuso das informações; bem como o suporte para a elaboração de serviços e aplicações no KTV, como por exemplo, uma aplicação de consulta semântica suportada pela mineração de dados; além do uso conjunto com outras ontologias do projeto – CoreKTV, etc.

Um provável cenário de uma aplicação que utiliza o conhecimento minerado pelo MKTV aliado a uma consulta semântica (Figura 18) seria:

- Se o usuário liga a TV na Quarta-feira (dia da semana) então assiste programas infantis (gênero) durante a noite (período).
- O conceito de gêneros infantis poderiam ser expandidos segundo a ontologia CoreKTV (ARAÚJO, 2011) para desenhos, animes, programas de auditório, séries teens e em seguida delimitado para os programas desse gênero que passam a noite e que passam na quarta-feira.

- Na figura (18) abaixo, o gênero infantil é expandido para os programas infantis “teletubies”, “anime Bleach”, “anime One Piece”, “ICarly”, “Patati-Patatá”; em seguida é delimitado para o programa que passa na quarta-feira a noite, que no exemplo em questão é a série “ICarly”.

Figura 18 - provável cenário de consulta semântica



Fonte: Próprio Autor (2012)

5.2.2.6 Acesso aos Dados

Para acesso aos conhecimentos descobertos no processo de Mineração de Dados, o MKTV provê um Web Service. A utilização de Web Services tem como vantagem a integração entre aplicações desenvolvidas por tecnologias diferentes. O Web Service do MKTV tem como parâmetros de entrada a data da execução da Mineração de Dados e o número identificador do STB do usuário, ele disponibiliza como resposta a solicitação do requisitante do conhecimento minerado um arquivo XML com as regras de associação geradas.

5.3 CONCLUSÃO

Este Capítulo descreveu as principais idéias relacionados com a uma nova tecnologia para a plataforma de TV Digital, chamada Knowledge TV - KTV. A intensão principal é a inclusão de uma camada semântica que pudesse suportar vários tipos de novos serviços e aplicações semânticas.

O projeto KTV é descrito através de diversos módulos no servidor, como: Ambiente de KDD para TVDI, Base de Conhecimento, Camada de Modelagem Semântica, Módulo de Raciocínio Automático, Módulo de Aplicações e Serviços. Além de um módulo no STB responsável pelo envio das informações para o servidor. Desses o ambiente de KDD para TVDI, chamado de MKTV, é especificado e detalhado, demonstrando sua utilidade para o projeto.

A arquitetura do módulo MKTV é formada por um banco de dados operacional, um componente de ETL, um *Data Warehouse*, um engenho de Mineração de Dados, um componente OLAP e uma ontologia chamada OntoMKTV. Através da utilização desses componentes, o MKTV busca encontrar conhecimentos úteis nos dados advindos da TVDI.

Desta forma, os componentes do MKTV foram implementados da seguinte maneira: o banco de dados operacional utilizado no projeto, juntamente com o *Data Warehouse* foram implantados em um banco de dados PostgreSQL. O componente de ETL foi implementado através da ferramenta Kettle, enquanto o engenho de mineração de dados utilizado foi o WEKA. A ontologia OntoMKTV esta especificada e implementada através do software protegé. O componente OLAP não foi implementado ficando como trabalho futuro ser implementado.

O próximo Capítulo objetiva demonstrar os experimentos realizados através do MKTV.

CAPÍTULO

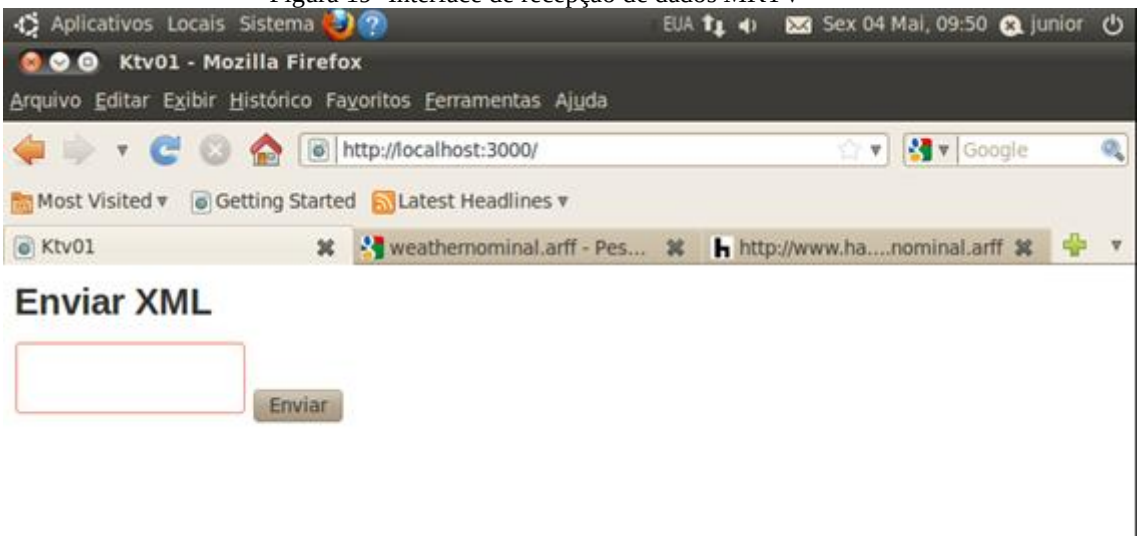
6

Experimentos

Para realizar os experimentos foi desenvolvido um protótipo do MKTV, utilizando a linguagem de programação Java. O mesmo utilizou-se de toda a especificação já mencionada neste trabalho. Desta forma, foram integrados: o banco de dados Postgresql para modelagem relacional, bem como, multidimensional; a engine de Mineração de Dados Weka; a API Jena para trabalhar com ontologias, todos através de um aplicativo com interface web.

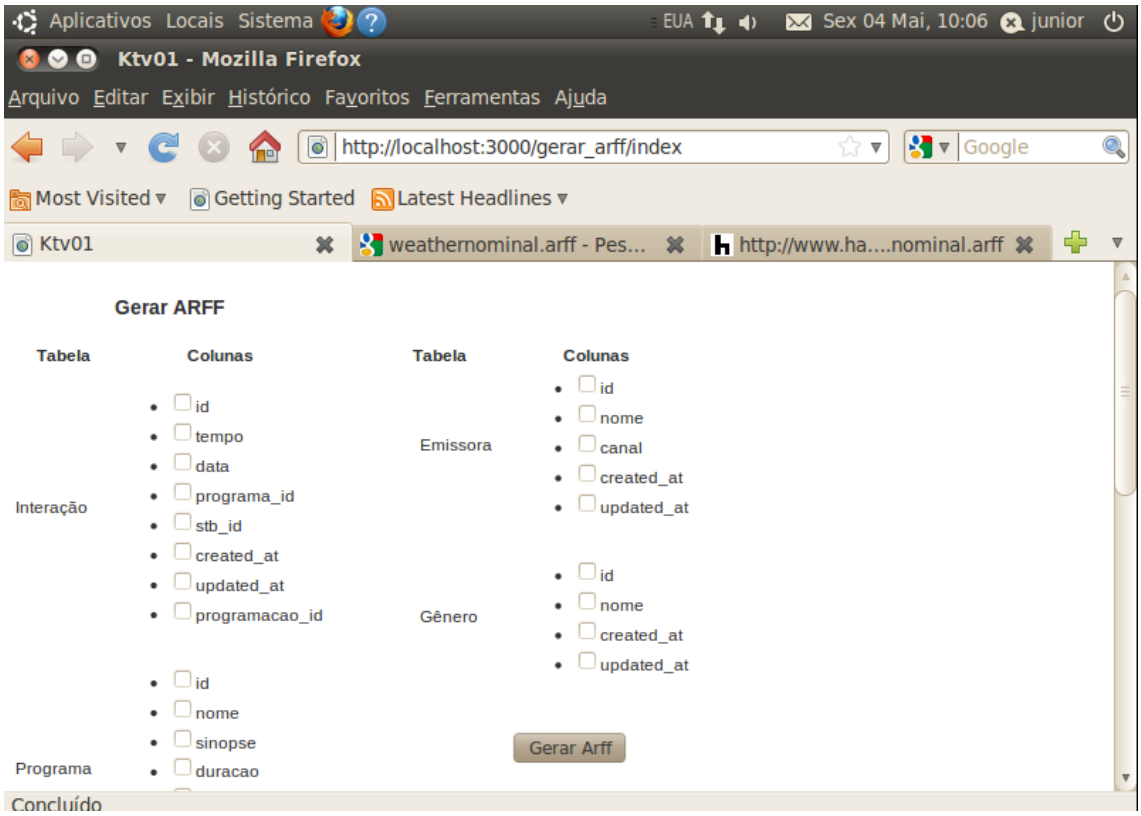
O protótipo do MKTV fornece uma interface web simples para entrada de dados no formato XML (Figura 19), esses dados são inseridos no banco de dados operacional e estão disponíveis para serem minerados (Figura 20).

Figura 19- Interface de recepção de dados MKTV



Fonte: Próprio autor (2012)

Figura 20 - Escolha de atributos para mineração



Fonte: Próprio autor (2012)

As seções 6.1 e 6.2 apresentam os estudos de caso com dados do domínio da TV.

6.1 ESTUDO DE CASO AUDIÊNCIA DE TV

Com o objetivo de validar a aplicabilidade do sistema MKTV, foram realizados experimentos para dar subsídios para avaliar de forma prática o projeto. Nesse sentido o projeto MKTV buscou dados reais de utilização de TVDI, por meio de solicitação a empresa brasileira que faz medição de audiência, IBOPE (IBOPE, 2011) e de redes de TV que pudessem ceder dados de pesquisas realizadas, sendo esta solicitação atendida por meio da rede de televisão local TV Arapuã. Além de pesquisar bases de dados reais na internet, neste sentido, obtendo êxito e encontrando a base de dados de recomendação de filmes da empresa web de aluguel de filmes online NetFlix (www.netflix.com), sendo esta base de dados um subdomínio de TV Digital Interativa, no caso de filmes.

6.1.1 Dados de TV

6.1.1.1 Aquisição e Coleta de Dados para Experimentos

A TV Arapuã (http://www.tvarapuan.com.br/), emissora de TV brasileira com sede em João Pessoa na Paraíba, forneceu para o projeto KTV, uma pesquisa de medição de audiência, desenvolvida pelo instituto brasileiro de pesquisa e medição de audiência (IBOPE), realizada no período de 11 a 17 de setembro de 2009, com alcance em todo o estado da Paraíba. A medição englobou a audiência das principais redes de TV aberta que atuam no estado nordestino, são elas: Globo, Record, SBT, Rede TV, Bandeirantes, TV Educativa.

As informações contempladas no relatório dizem respeito ao percentual de *Share* e *Rate* de cada TV, em intervalos de uma hora, durante uma semana, conforme visto na Tabela 6.

Tabela 6 – Tabela com dados da Pesquisa de audiência

	Média para faixa															
	TOT		BAN		EDU		GLO		RTV		REC		SBT		OUT	
	UNIVERSO JPE		UNIVERSO JPE		UNIVERSO JPE		UNIVERSO JPE		UNIVERSO JPE		UNIVERSO JPE		UNIVERSO JPE		UNIVERSO JPE	
	rat%	shr%	rat%	shr%	rat%	shr%	rat%	shr%	rat%	shr%	rat%	shr%	rat%	shr%	rat%	shr%
06:00:00-07:00:00	4.8	100.0	0.5	10.0	0.0	0.0	2.3	48.7	0.1	2.2	0.7	15.2	0.6	12.1	0.6	11.9
07:00:00-08:00:00	6.9	100.0	0.2	2.2	0.1	1.3	4.1	60.4	0.3	4.9	1.2	18.0	0.5	7.5	0.4	5.9
08:00:00-09:00:00	9.4	100.0	0.2	1.8	0.2	1.9	5.9	62.5	0.2	2.4	1.5	15.6	1.2	12.2	0.4	3.7
09:00:00-10:00:00	13.3	100.0	0.2	1.8	0.2	1.3	7.1	53.7	0.3	2.2	2.3	17.2	1.7	12.9	1.5	10.9
10:00:00-11:00:00	13.5	100.0	0.2	1.1	0.1	0.8	8.0	59.4	0.4	3.0	2.0	15.1	1.1	8.4	1.7	12.2
11:00:00-12:00:00	10.8	100.0	0.1	1.0	0.1	0.8	6.0	55.8	0.4	3.5	1.8	16.5	1.4	12.5	1.1	9.8
12:00:00-13:00:00	13.4	100.0	0.5	3.5	0.0	0.0	6.3	47.0	1.4	10.3	2.2	16.5	2.0	14.7	1.1	8.1
13:00:00-14:00:00	20.5	100.0	0.8	3.9	0.0	0.0	9.5	46.2	1.1	5.6	3.3	16.0	3.2	15.7	2.6	12.6
14:00:00-15:00:00	19.8	100.0	0.5	2.5	0.0	0.0	8.5	42.9	1.0	5.0	3.1	15.7	3.8	19.3	2.9	14.6
15:00:00-16:00:00	16.3	100.0	0.4	2.2	0.1	0.5	6.8	41.8	0.6	3.9	2.8	17.3	3.7	22.5	1.9	11.8
16:00:00-17:00:00	17.0	100.0	0.7	4.0	0.1	0.4	7.9	46.5	1.1	6.3	3.1	18.2	3.0	17.5	1.2	7.1
17:00:00-18:00:00	17.4	100.0	0.9	4.9	0.2	0.9	7.0	40.1	1.3	7.4	3.5	20.0	3.1	17.7	1.6	9.1
18:00:00-19:00:00	21.5	100.0	0.3	1.2	0.0	0.0	9.0	41.9	1.7	7.7	6.3	29.4	3.3	15.3	1.0	4.6
19:00:00-20:00:00	25.6	100.0	0.6	2.3	0.0	0.1	11.1	43.5	1.9	7.3	7.6	29.7	3.2	12.6	1.2	4.5
20:00:00-21:00:00	30.6	100.0	0.3	0.9	0.2	0.7	16.3	53.4	1.9	6.2	6.6	21.5	4.1	13.5	1.1	3.7
21:00:00-22:00:00	30.5	100.0	0.3	1.0	0.0	0.0	15.9	52.3	3.5	11.6	5.7	18.8	4.0	13.0	1.0	3.3
22:00:00-23:00:00	25.4	100.0	0.6	2.3	0.1	0.5	11.3	44.6	4.1	16.0	5.1	20.1	3.1	12.1	1.2	4.6
23:00:00-24:00:00	14.0	100.0	0.5	3.6	0.0	0.0	5.9	42.2	2.7	18.9	2.2	15.5	1.9	13.2	0.9	6.6

Fonte: Próprio Autor (2012)

A estimativa *Share*, diz respeito ao percentual de audiência de um canal/programa em relação a audiência de todas as TVs em um mesmo horário. A Tabela 7 exemplifica a estimativa, nela temos que: num universo que totaliza 1000 espectadores, 35% desses espectadores estão sintonizados no Canal 1, 25% no Canal 2, 7% no Canal 3 e 33% no Canal 4.

Tabela 7 - exemplo de Share

Canal	Universo	
	Qtd	Share (%)
canal 1	350	35
canal 2	250	25
canal 3	70	7
canal 4	330	33
total	1000	100

Fonte: Próprio Autor (2012)

A estimativa *Rate*, refere-se à audiência média de um programa/período de horário. Sendo medido através do cálculo do tempo de permanência do usuário assistindo a determinado programa de TV. Neste caso é levado em consideração todo o universo de usuários que estão com a TV ligada, mesmo que não estejam assistindo ao mesmo programa. Podemos verificar o *Rate* através da Tabela 8. Neste exemplo, um programa de 60 minutos de duração é medido, num universo de 4 indivíduos assistindo TV. O indivíduo A assistiu a 30 minutos do programa dando média de 50% de audiência, o indivíduo B assistiu a 30 minutos do programa dando media de 50% de audiência, o indivíduo C não assistiu o programa e o indivíduo D assistiu a 60 minutos do programa dando média de 100% de audiência. Em seguida, é calculada a média aritmética que nos diz que o programa do exemplo teve um *Rate* ou média de audiência de 50%. Logo o *Rate* do programa seria 100% se todas as televisões estivessem ligadas no programa.

Tabela 8 - exemplo Rate

indivíduo	viu	duração	aud. Media
A	sim	30	50
B	sim	30	50
C	não	0	0
D	sim	60	100
Total		120	200 de 400
Rate %			50%

Fonte: Próprio Autor (2012)

De posse dos dados de audiência por faixa de horário e por canal, ainda foi necessário para compor as informações a serem mineradas no banco de dados, a informação de quais programas eram executados em um determinado horário e canal. Então, como passo seguinte na

aquisição dos dados, foi pesquisado na internet e coletadas as informações de um guia de programação da TV. Este guia apresenta todas as emissoras envolvidas na pesquisa supracitada em formato tabular. Um guia de programação informa, de forma organizada, todos os programas exibidos pelo canal de TV e a faixa de horário que o mesmo é exibido. A Figura 21 ilustra o guia de programação (http://www.sky.com.br/servicos/GuiadaTv/GravacaoDistancia.aspx) utilizado.

Figura 21 - Guia de Programação

Fonte: Próprio Autor (2012)

Nestes termos, os dados para Mineração de Dados estavam quase satisfatórios para um experimento, mas ainda faltavam as informações referentes aos programas, ou seja, o gênero dos programas, o subgênero, a sinopse, o idioma, o ano de produção, o país de origem, etc.. Logo o próximo passo realizado diz respeito à aquisição de dados referentes aos programas. Essa atividade que foi realizada através de consultas aos sites das referidas emissoras, além do site da Dbpedia (dbpedia.com), projeto que visa organizar semanticamente os dados da Wikipédia (wikipedia.com).

Ao fim da atividade de aquisição de informações nas diversas fontes de pesquisa foram criadas duas tabelas com as informações necessárias ao modelo de dados relacional do banco de dados operacional do projeto (Tabela 9 e 10).

Tabela 9 – Dados de Programa

idPrograma	Programa	Canal	Formato	Genero	Sinopse	Classificação	Duração(min)	ano estréia	País de Origem	Idioma Original
1	A Liga	BAN	Programa Jornalístico	Reality-Doc	A Liga é um programa de t	12 anos	60	2010	Brasil	Português
2	A Liga - Reapresentação	BAN	Programa Jornalístico	Reality-Doc	A Liga conta uma história	Livre	60	2010	Brasil	Português
3	Acredite se Quiser	BAN	Programa de Variedades	Entretenimento	Baseado no programa Rip	Livre	60	2009	Brasil	Português
4	Agora é Tarde	BAN	Talk Show	Humorístico	Agora é Tarde é um Talk S	12 anos	45	2011	Brasil	Português
5	Auto +	BAN	TeleJornal	Esportivo	Auto Mais é um programa	Livre	45	2011	Brasil	Português
6	Band Clássicos	BAN	Programa Jornalístico	Esportivo	Sob o comando de Daniel	Livre	30	2009	Brasil	Português
7	Band Esporte Clube	BAN	Programa Jornalístico	Esportivo	Apresentado por Patrícia M	Livre	20	2007	Brasil	Português
8	Band Kids	BAN	Infantil	Infantil	Band Kids é um programa	Livre	75	2000	Brasil	Português
9	Bernie Mac: Um Tio da Pesada	BAN	Seriado	Comédia	Barnie Mac: Um Tio da Pe	Livre	30	2001	EUA	Inglês
10	BrasilCaminhoneiro	BAN	Esportivo	Esportivo	O Programa BrasilCaminh	Livre	30	2009	Brasil	Português
11	BrasilUrgente	BAN	Programa Jornalístico	Jornalístico	BrasilUrgente é um TeleJo	Livre	105	2001	Brasil	Português
12	BrasilUrgente - SP	BAN	Programa Jornalístico	Jornalístico	BrasilUrgente é um TeleJo	Livre	135	2001	Brasil	Português
13	Canal Livre	BAN	Talk Show	Entrevistas	O Canal Livre é um dos pr	Livre	45	2009	Brasil	Português
14	Cine Band	BAN	Filme	Filme	O Cine Band é uma sessã	Livre	120	2009	Brasil	Português
15	Cinema na Madrugada	BAN	Filme	Filme	O Cinema na Madrugada é	Livre	120	2009	Brasil	Português
16	Claquete	BAN	Revista Eletrônica	Entretenimento	Apresentado por Otávio M	Livre	75	2009	Brasil	Português
17	Copa do Mundo de Futebol de Areia	BAN	Esportivo	Esportivo	A Copa do Mundo de Fute	Livre	80	2009	Brasil	Português
18	CQC: Custe o Que Custar	BAN	Programa de Auditório	Humorístico	CQC: Custe o que custar (	12 anos	120	2008	Brasil	Português
19	CQC: Custe o Que Custar - Melhores Mo	BAN	Programa de Auditório	Humorístico	Com humor inteligente, a	12 anos	60	2008	Brasil	Português
20	Dia Dia	BAN	Programa de Auditório	Entretenimento	No Dia Dia a principal alte	Livre	210	2001	Brasil	Português
21	Domingo no Cinema	BAN	Filme	Filme	A faixa de entretenimento	Livre	45	2009	Brasil	Português
22	E24	BAN	Programa Jornalístico	Reality-Doc	O programa E24 tem com	Livre	60	2009	Argentina	Espanhol
23	Espaço Vida Vitoriosa	BAN	Religioso	Religioso	Depois de ouvir mais um n	Livre	120	1980	Brasil	Português
24	Família Dinossauros	BAN	Seriado	Juvenil	Dinosaurs (Família Dinoss	Livre	45	2007	EUA	Inglês
25	Futebol 2011 - Vivo	BAN	Esportivo	Esportivo	Um dos principais torneios	Livre	60	2011	Brasil	Português
26	Futurama	BAN	Seriado	Humorístico	Futurama é uma sitcom ar	Livre	22	1999	EUA	Inglês
27	Horário Político	BAN	Política	Política	Horário Político brasileiro	Livre	10	1965	Brasil	Português
28	Infomercial	BAN	Talk Show	Variedades	Infomerciais são propagan	Livre	240	1984	EUA	Inglês

Fonte: Próprio Autor (2012)

Tabela 10 – Dados de Audiência

Canal	DiaDaSemana	Data	Hora	Programa	Share	Rate
BAN	sexta-feira	9/11/2009	00:00:00-01:00:00	Jornal da Noite	5,6	0,3
BAN	sexta-feira	9/11/2009	01:00:00-02:00:00	Jornal da Noite	0	0
BAN	sexta-feira	9/11/2009	02:00:00-03:00:00	Claquete	0	0
BAN	sexta-feira	9/11/2009	03:00:00-04:00:00	Espaço Vida Vitoriosa	0	0
BAN	sexta-feira	9/11/2009	04:00:00-05:00:00	Espaço Vida Vitoriosa	0	0
BAN	sexta-feira	9/11/2009	05:00:00-06:00:00	Espaço Vida Vitoriosa	0	0
BAN	sexta-feira	9/11/2009	06:00:00-07:00:00	Espaço Vida Vitoriosa	11	0,8
BAN	sexta-feira	9/11/2009	07:00:00-08:00:00	Primeiro Jornal SP	4,2	0,5
BAN	sexta-feira	9/11/2009	08:00:00-09:00:00	Band Kids	2,9	0,4
BAN	sexta-feira	9/11/2009	09:00:00-10:00:00	Quase Anjos	7,7	1,3
BAN	sexta-feira	9/11/2009	10:00:00-11:00:00	Dia Dia	5,5	0,9
BAN	sexta-feira	9/11/2009	11:00:00-12:00:00	Jogo Aberto	8,7	1,4
BAN	sexta-feira	9/11/2009	12:00:00-13:00:00	Jogo Aberto	4,9	1,1
BAN	sexta-feira	9/11/2009	13:00:00-14:00:00	SP Acontece	4,1	0,7
BAN	sexta-feira	9/11/2009	14:00:00-15:00:00	Power Rangers	1,3	0,2
BAN	sexta-feira	9/11/2009	15:00:00-16:00:00	Manual de Sobrevivência Escolar do Ned	2,6	0,6
BAN	sexta-feira	9/11/2009	16:00:00-17:00:00	Videonews	3,8	0,6
BAN	sexta-feira	9/11/2009	17:00:00-18:00:00	Brasil Urgente	6,7	1
BAN	sexta-feira	9/11/2009	18:00:00-19:00:00	Brasil Urgente	4,5	1,3
BAN	sexta-feira	9/11/2009	19:00:00-20:00:00	Jornal da Band	4,8	1,6
BAN	sexta-feira	9/11/2009	20:00:00-21:00:00	Bernie Mac: Um Tio da Pesada	1,7	0,6
BAN	sexta-feira	9/11/2009	20:00:00-21:00:00	Show da Fé	1,7	0,6
BAN	sexta-feira	9/11/2009	21:00:00-22:00:00	NCIS	1,8	0,7
BAN	sexta-feira	9/11/2009	22:00:00-23:00:00	Top Cine	3,1	1,1
BAN	sexta-feira	9/11/2009	23:00:00-24:00:00	Top Cine	5,3	0,9
EDU	sexta-feira	9/11/2009	00:00:00-01:00:00	Dona Joventina	1,6	0,1
EDU	sexta-feira	9/11/2009	01:00:00-02:00:00	Sanhauá	0	0
EDU	sexta-feira	9/11/2009	02:00:00-03:00:00	Sem Censura	0	0

Fonte: Próprio Autor (2012)

Os dados foram inseridos no banco de dados operacional de acordo com o modelo relacional proposto e em seguida sofreram o processo de ETL. Durante o processo de Transformação, alguns atributos que são derivados de outros campos foram preenchidos, de acordo com o modelo multidimensional, como: a classificação de criação de programa, a classificação de duração e o campo turno. A transformação dos campos derivados deu-se da seguinte maneira:

Regra de classificação da criação do programa:

- Se o ano de estreia for igual a 2011 sua classificação quanto ao tempo de estréia será "ATUAL";
- Se o ano de estreia for menor que 2011 e maior igual que 2005 sua classificação será "ATÉ 5 ANOS";
- Se o ano de estreia for menor que 2005 e maior igual que 2000 sua classificação será "DE 5 ATÉ 10 ANOS";
- Se o ano de estreia for menor que 2000 e maior igual que 1990 sua classificação será "DE 10 ATÉ 20 ANOS";
- Se o ano de estreia for menor que 1990 sua classificação será "MAIS 20 ANOS";

Regra de classificação de duração:

- Se a duração do programa for menor igual a 45 sua classificação quanto à duração será "CURTO";
- Se a duração do programa for maior que 45 e menor ou igual a 90 sua classificação quanto à duração será "MÉDIO";
- Se a duração do programa for maior que 90 sua classificação quanto à duração será "LONGO";

Regras de classificação para turno:

- Se a hora do programa for maior igual a 00 e menor igual a 04 sua classificação quanto ao turno será "MADRUGADA";
- Se a hora do programa for maior igual a 05 e menor igual a 11 sua classificação quanto ao turno será "MANHÃ";
- Se a hora do programa for maior igual a 12 e menor igual a 17 sua classificação quanto ao turno será "TARDE";
- Se a hora do programa for maior igual a 18 e menor igual a 24 sua classificação quanto ao turno será "NOITE";

A Figura 22 ilustra todas as variáveis utilizadas no experimento e sua frequência no conjunto de dados.

Figura 22 - Painel de amostragem de dados do experimento

Fonte: Próprio Autor (2012)

6.1.1.2 Mineração de Dados

Como próximo passo, os dados foram carregados no DW, utilizando a ferramenta de ETL Kattle. Logo após, o processo de Mineração de Dados é acionado gerando regras de associação sobre as informações coletadas.

Algumas regras que mostram conhecimento interessante (Apêndice II e III) foram:

- Algoritmo *Predictive Apriori* (Formato saída da engine de mineração):
    - canal=GLO diasemana=sexta-feira periododia=NOITE classificacaoanoestreia=ATUAL tamanhoduracao=MEDIO ==> genero=Telenovela acc:(0.99499)
    - programa=Bom Dia & Cia ==> canal=SBT classificacaoanoestreia=DE 10 ATÉ 20 ANOS acc:(0.99499)
    - diasemana=segunda-feira genero=Filme ==> canal=GLO tamanhoduracao=MEDIO acc:(0.99498)
    - programa=O Astro classificacaoanoestreia=ATUAL ==> tamanhoduracao=CURTO acc:(0.99498)
- Algoritmo *Predictive Apriori* (Formato usando Lógica de 1º ordem):

- SE canal=GLOBO E diasemana=sexta-feira E periododia=NOITE E classificacaoanoestreia=ATUAL E tamanhoduracao=MEDIO ==> ENTÃO genero=Telenovela
- SE programa=Bom Dia & Cia ==> ENTÃO canal=SBT classificacaoanoestreia=DE 10 ATÉ 20 ANOS
- SE diasemana=segunda-feira E genero=Filme ==> ENTÃO canal=GLO tamanhoduracao=MEDIO
- SE programa=O Astro E classificacaoanoestreia=ATUAL ==> ENTÃO tamanhoduracao=CURTO

- Algoritmo *Tertius* (Formato saída da engine de mineração):
  - /* 0,685640 0,026560 */ genero = Jornalístico ==> programa = SPTV or classificacaoanoestreia = MAIS 20 ANOS or canal = REC
  - /* 0,673594 0,118637 */ canal = GLO ==> programa = Sessão da Tarde or genero = Telenovela or classificacaoanoestreia = MAIS 20 ANOS
- Algoritmo *Tertius* (Formato usando Lógica de 1º ordem):
  - SE genero = Jornalístico ==> ENTÃO programa = SPTV OU classificacaoanoestreia = MAIS 20 ANOS OU canal = RECORD
  - SE canal = GLOBO ==> ENTÃO programa = Sessão da Tarde OU genero = Telenovela OU classificacaoanoestreia = MAIS 20 ANOS

Sobre as regras do algoritmo *Predictive Apriori*, podemos inferir que a primeira regra informa que: "quem assiste ao canal Globo, em dias de Sexta-feira à noite, e vê programas de média duração, que estrearam a pouco tempo, assistem telenovela", a mesma tem a medida *predictive accuracy*[5] no valor 0.99499;

A segunda regra diz que: "Quem assiste ao programa Bom Dia e Cia, então assiste ao canal SBT, e gosta de programas que foram criados em um período de 10 a 20 anos atrás";

A terceira regra expressa que: "Quem a assiste filmes as Segundas-feiras, assiste a programas de média duração no canal Globo";

A quarta regra implica que: "Quem assiste ao programa O Astro e a programas que estrearam a pouco tempo assistem programas de curta duração".

---

[5] *predictive accuracy* (SCHEFFER, 2004) é uma combinação das medidas suporte e confiança utilizado pelo algoritmo *predictive apriori* para identificar as melhores regras de associação e classifica-las.

Com relação às regras do algoritmo *Tertius*, podemos inferir que a primeira regra informa que: “Quem assiste a programas do gênero jornalístico, assiste ao programa SPTV, ou a programas com mais de 20 anos de criação, ou assiste o canal Record ”;

A segunda regra traz o seguinte conhecimento descoberto: “Quem assiste ao canal Globo então assiste o programa Sessão da Tarde ou programas do gênero telenovela, ou programas que foram criados, a mais 20 anos”;

Depois de mineradas as regras foram encapsuladas, no formato ontológico, através da OntoMKTV, permitindo que o conhecimento descoberto possa ser utilizado através de outras aplicações, além do conhecimento poder ser reutilizado, estendido.

6.1.2 Dados NetFlix

Os sistemas de recomendação desenvolvem um papel muito importante junto aos sites de comércio eletrônico, bem como em aplicações para TV. Esses sistemas visam sugerir itens de interesse aos usuários. No *e-commerce* de locação de filmes NetFlix (www.netflix.com) não é diferente. O sucesso das vendas ou assinaturas fica a cargo da eficiência da sugestão, pois caso o usuário não encontre o filme que deseja e não se sinta atraído a assistir um novo filme ele não utilizará mais o serviço.

A empresa incentiva seus clientes a expressar opinião sobre o quanto eles gostaram dos filmes que assistiram e de 1998 à 2006 já tinha recolhido cerca 1,9 bilhão de votos de mais de 11,7 milhões de usuários em mais de 85 mil títulos de filmes (BENNETT; LANNING, 2007).

Em outubro de 2006 o NetFlix propõe um desafio para a comunidade científica que trabalha com Mineração de Dados e Aprendizagem de Máquina. Com o intuito de aumentar a precisão de seu sistema de recomendação denominado Cinemarch, o NetFlix disponibilizou uma base de dados de classificação com mais 480 mil classificações de mais de 18 mil títulos de filmes. O mesmo desafiou a comunidade científica a construir algoritmos que sejam mais eficazes na recomendação do que os algoritmos do seu sistema. Este desafio se tornou parte do KDD CUP 2007 que é a competição anual da conferência de descoberta de conhecimento em base de dados SIGKDD (http://www.cs.uic.edu/~liub/KDD-cup-2007/proceedings.html).

A utilização da base de dados disponibilizada pelo NetFlix se dá devido as informações trabalhadas pelo seu sistema. Estas fazem parte de um subdomínio da TVDI que é de filmes, recomendações de usuários e também são reais, não simuladas, agregando um grande valor a pesquisa.

6.1.2.1 Aquisição e Coleta de dados para Experimentos

Para o experimento com os dados de recomendações da internet é necessário uma coleta de dados. A mesma foi realizado através do site do desafio NetFlix (http://www.netflixprize.com/). A base de dados disponibilizada tinha as seguintes informações:

- IDFilme:
  - Inteiro único atribuído arbitrariamente no intervalo [1 ... 17.770].
- CustomerID:
  - Inteiro único atribuído arbitrariamente no intervalo [1 ... 2649429].
- Rating:
  - Número de 'estrelas' atribuídos a um filme por um cliente; um número inteiro de 1 a 5.
- Title:
  - Título em idioma Inglês do filme no site do NetFlix.
- YearOfRelease:
  - Ano que um filme foi lançado no intervalo [1890 ... 2005]. Pode corresponder ao ano de lançamento do DVD correspondente.
- Date:
  - Data de uma classificação na forma AAAA-MM-DD, na faixa de 1998/11/01-2005/12/31.
- NetflixID:
  - Identificador inteiro ID de um filme, como atualmente utilizado na API de desenvolvimento do Netflix (http://developer.netflix.com/)

6.1.2.2 Mineração de Dados

Como próximo passo, os dados foram exportados para tabelas no banco de dados operacional e em seguida para o DW, utilizando a ferramenta de ETL Kattle. Logo após, o processo de Mineração de Dados é acionado gerando regras de associação sobre as informações coletadas.

Algumas regras que mostram conhecimento interessante foram:

- Algoritmo *Predictive Apriori* (Formato saída da engine de mineração):
    - Filme = Lord of the Rings: The Return of the King: Extended Edition: Bonus Material ==> avaliacao=OTIMO acc:(0.30065)
    - mesanorec=jan ==> avaliacao=REGULAR acc:(0.20734)
    - filme= Dinosaur Planet ==> avaliacao=REGULAR acc:(0.20206)
- Algoritmo *Predictive Apriori* (Formato usando Lógica de 1º ordem):
    - SE Filme = Lord of the Rings: The Return of the King: Extended Edition: Bonus Material ==> ENTÃO avaliacao=OTIMO
    - SE mesanorec=janeiro ==> ENTÃO avaliacao=REGULAR
    - SE filme=Dinosaur Planet ==> ENTÃO avaliacao=REGULAR

Podemos interpretar que a primeira regra informa que: "Se o filme é *Lord of the Rings: The Return of the King: Extended Edition: Bonus Material* então tem avaliação positiva";

A segunda regra: "Se o filme é em visto em janeiro então ele é avaliado de maneira regular";

A terceira regra: "Se o filme é *Dinosaur Planet* então ele é avaliado como regular";

- Algoritmo *Tertius* (Formato saída da engine de mineração):
    - /* 0,061504 0,017857 */ filme = Lord of the Rings: The Return of the King: Extended Edition: Bonus Material ==> avaliacao = OTIMO or mesanorec = jun
    - /* 0,048133 0,034226 */ filme = 8 Man ==> avaliacao = PESSIMO or mesanorec = jul
    - (Formato usando Lógica de 1º ordem):
- Algoritmo *Tertius* (Formato usando Lógica de 1º ordem):
    - SE filme = Lord of the Rings: The Return of the King: Extended Edition: Bonus Material ==> ENTÃO avaliacao = OTIMO or mesanorec = junho
    - SE filme = 8 Man ==> ENTÃO avaliacao = PESSIMO or mesanorec = julho

Podemos inferir que a primeira regra informa que: "Quem assiste *Lord of the Rings: The Return of the King: Extended Edition*: Bonus Material geralmente avalia positivamente os filmes no mês de julho";

A segunda regra: "Quem assiste o filme 8 Man avalia filmes com a classificação péssimo".

Depois de mineradas as regras foram encapsuladas, no formato ontológico, através da OntoMKTV, permitindo que o conhecimento descoberto possa ser utilizado através de outras aplicações, além do conhecimento poder ser reutilizado, estendido.

6.2 CONCLUSÕES

Este capítulo buscou demonstrar a aplicabilidade do projeto *Mining Knowledge TV*, ambiente de KDD da arquitetura do KTV, através da utilização de bases de dados reais advindas de uma pesquisa de um instituto de medição de audiência brasileiro e de uma base real proveniente de um portal de aluguéis de filmes sobre demanda, o Netflix.

As bases de dados foram mineradas através da aplicação de algoritmos de associação por meio do engenho de Mineração de Dados utilizado no MKTV, e resultaram em regras de associações com padrões provenientes das informações analisadas.

Há que se considerar que o conhecimento minerado através das informações de audiência de TV reflete um pouco da natureza da base de dados e de suas limitações. Já que as informações mineradas provêm de uma pesquisa que abrange unicamente o estado da Paraíba, com informações de TV aberta, ou seja, 7 canais de TV.

As informações provenientes da base do Netflix obtiveram menos regras úteis devido a pequena quantidade de atributos, além da maioria deles serem de natureza numérica. Uma solução seria realizar uma expansão semântica na base, através de consulta semântica, em que ao passar o título do filme do NetFlix a consulta buscasse as informações de gênero, idioma, país de origem, etc.

Os resultados obtidos demonstram a funcionalidade da MKTV no contexto do Projeto KTV. O próximo capítulo versará sobre as considerações finais e trabalhos futuros acerca desse trabalho.

CAPÍTULO

7
Considerações Finais e Trabalhos Futuros

Este trabalho apresentou a especificação e desenvolvimento do *Mining Knowledge TV*, uma arquitetura de descoberta de conhecimento em base de dados com técnicas semânticas integrada ao projeto *Knowledge TV*. Para a realização do MKTV foi necessário o cumprimento de algumas etapas.

Inicialmente realizou-se um estudo sobre a TV Digital Interativa interativa e sobre a área de descoberta de conhecimento em base de dados e Mineração de Dados. Além disso, foram pesquisados os principais trabalhos que envolviam a utilização conjunta dessas áreas realizando o estado da arte de Mineração de Dados em TVDI.

Em seguida, foram realizadas pesquisas sobre a área de representação do conhecimento e o levantamento de requisitos para a construção do ambiente de descoberta de conhecimento em base de dados do projeto Knowledge TV, o Mining Knowledge TV - MKTV.

Finalizada a pesquisa e o levantamento de requisitos, iniciou-se a modelagem da arquitetura do MKTV, juntamente com a modelagem relacional e multidimensional do banco de dados operacional e do *Data Warehouse* utilizados no projeto. Como passo seguinte foi incorporado ao projeto a engine de mineração de dados WEKA, máquina necessária para a descoberta de conhecimento útil. Outra etapa importante, foi a criação da ontologia OntoMKTV, responsável pela representação semântica do conteúdo minerado.

Para validar o projeto foi necessário a busca por bases de dados reais no contexto da TV Digital Interativa como na pesquisa de audiência e nos dados do NetFlix que pertencem ao domínio da TVDI, neste caso, no subdomínio de Filmes.

Desta forma, o projeto MKTV está inserido no contexto do Sistema Brasileiro de TV Digital Interativa – SBTVDI junto ao *middleware* Ginga, fornecendo informações úteis para serem utilizadas em aplicações e serviços direcionadas ao consumo de usuários e provedores de conteúdo de TVDI.

Podemos atestar a inovação do projeto iniciado, pela pouca quantidade de trabalhos específicos na área da TV Digital Interativa com ênfase na representação de conhecimento obtido através da Mineração de Dados, além da disponibilização desse conteúdo útil para desenvolvedores de aplicações para TVDI. O projeto teve publicações nas seguintes conferências:

- LINO, N. Q.; ARAÚJO, J.; ANABUKI, D.;PATRÍCIO JUNIOR. J. C. A.; BATISTA, M.; NÓBREGA, R.; AMARO, M. ; SIEBRA, C. .Knowledge TV. In: 9th European Conference on Interactive TV and Video – Euro ITV 2011. Lisboa – Portugal. 2011.

- PATRICIO JUNIOR, J. C. A., LINO, N. Q. Mining Knowledge TV: A proposal for Data Integration. In the Knowledge TV Environment" In: 2nd international Workshop on Future of Television – at the EuroITV 2011. Lisboa – Portugal. 2011.

- PATRICIO JUNIOR, J. C. A., LINO, N. Q. Mining Knowledge TV: Uma Proposta de Integração de Dados no Ambiente da Knowledge TV. In: Conferência IADIS Ibero-Americana WWW/Internet 2010, Algarve – Portugal, 2010.

- NOBREGA, R., BATISTA, M., AMARO, M., PATRÍCIO JUNIOR, J. C. A., LINO, N. Q.
  Semantic Queries in Knowledge TV Supported by Knowledge Discovery In: IADIS INTERNATIONAL CONFERENCE WWW/INTERNET 2011, 2011, Rio de Janeiro.

Há que se destacar a complexidade de se implantar uma arquitetura genérica que visa atender ao ambiente emergente de convergência de dados que vem se tornando a TV Digital Interativa, abrigando dados de diversas fontes como a Internet, e consequentemente dotando esse ambiente de uma imensa quantidade de dados.

7.1 CONTRIBUIÇÕES

Através deste projeto podem-se destacar as seguintes contribuições:

1. A análise do domínio, a especificação dos requisitos e o desenvolvimento de um ambiente de descoberta de conhecimento em base de dados, com diversos módulos que contemplam as fases do KDD, em especial o módulo de Mineração de Dados com ênfase semântica na arquitetura do projeto *Knowledge TV*;
2. O levantamento de requisitos e a definição de um modelo de multidimensional de dados para o ambiente do *Knowledge TV* compatível com os metadados da TVDI, especificamente no que se refere ao SBTVDI, constituindo, desta forma um avanço no estado da arte nesta área (Tabela 11 – coluna 8);
3. A definição de um modelo semântico, através da criação da ontologia OntoMKTV, consistindo em um conjunto de classes que encapsulam o conhecimento a respeito da fase de pós-processamento, isto é, o entendimento das regras geradas. A ontologia terá grande importância principalmente pela possibilidade de ser estendida e integrável com as outras ontologias presentes do projeto, podendo a mesma compartilhar o conhecimento descoberto através da Mineração de Dados;
4. O fornecimento de um serviço que dê suporte as mais variadas aplicações no contexto da TVDI provendo conhecimento novo e de extrema relevância;
5. Avanço no estado da arte em termos de métodos para Mineração de Dados auxiliada por representação de conhecimento no contexto da TV Digital Interativa Interativa;
6. Construção de bases de conhecimento que dê suporte ao compartilhamento de informações acerca de conhecimento minerado de dados no contexto da TV Digital Interativa.

Tabela 12 – Tabela de Estado da Arte de TV Digital Interativa e Mineração de Dados

CARACTERÍSTICAS	**[1]**	**[2]**	**[3]**	**[4]**	**[5]**	**[6]**	**[7]**	**[8]**
TAREFA DE MINERAÇÃO CLASSIFICAÇÃO	X			X			X	
TAREFA DE MINERAÇÃO CLUSTERIZAÇÃO		X						
TAREFA DE MINERAÇÃO REGRAS DE ASSOCIAÇÃO			X		X	X		X
FERRAMENTA USADA DARWIN (ORACLE)		X					X	

FERRAMENTA USADA WEKA				X	X			X
FERRAMENTA USADA IMPLEMENTAÇÃO PRÓPRIA						X		
ALGORITMO USADO C4.5	X	X		X			X	
ALGORITMO USADO APRIORI			X		X	X		
ALGORITMO USADO PREDICTIVE APRIORI								X
ALGORITMO USADO TERTIUS								X
ALGORITMO USADO BACK PROPAGATION		X		X			X	
ALGORITMO USADO CLASSIFICADOR BAYESIANO	X			X				
OBJETIVO RECOMENDAÇÃO	X		X	X	X	X		X
OBJETIVO PROPAGANDA DIRECIONADA		X					X	
USA STB COMO SERVIDOR DE APLICAÇÃO		X		X	X	X	X	
INTEGRADO AO MIDDLEWARE GINGA						X		X
APLICAÇÃO RODANDO SOBRE MIDDLEWARE				X	X			

Fonte: Próprio Autor (2012)

7.2 TRABALHOS FUTUROS

Como trabalho futuro, destaco:

- Utilizar outras bases de dados de TV, por exemplo, de TV a Cabo para mais experimentos com o MKTV.
- A necessidade de se implantar algoritmos de diversas tarefas da Mineração de Dados, como por exemplo, classificação, clusterização, entre outros, que tornarão o MKTV mais genérico e robusto, fornecendo solução para problemas de múltiplas áreas. Pois cada tarefa de mineração e apropriada para um tipo de padrão/conhecimento que se quer descobrir..
- Realizar um Pré-processamento na base do NetFlix, utilizando consulta semântica, para que algoritmos de regras de associação tragam resultados uteis.
- Também poderá ser estudado a integração da solução da MKTV para *middlewares* de outros sistemas de TV Digital Interativa seguindo as normas J200, J201 e J202.
- Poderá ser implantado algum aplicativo OLAP para visualização das consultas multidimensionais.

REFERÊNCIAS

ABNT NBR 15603-1:2007. **Televisão Digital Terrestre - Multiplexação e Serviços de Informação (SI) - Parte 1: Serviços de informação do sistema de radiodifusão**. 2007.

AGRAWAL, R.; IMIELINSKI, T.; SRIKANT R. **Mining Association Rules between Sets of Items in Large Databases**. In: International Conference on Management of Data**.** Washington, 1993, p. 207–216.

ALMEIDA, Leandro Maciel et al. **Uma ferramenta para extração de padrões.** In: Revista eletrônica de Iniciação Científica – REIC. Porto Alegre, 2003. Disponível em: < http://www.sbc.org.br/reic/edicoes/2003e4/cientificos/UmaFerramentaParaExtracaoDePadroes.pdf >. Acesso em: Jan. 2012.

ALVES, L. P. G.; SILVA, F. S.; BRESSAN, G.. **CollaboraTVware: Uma proposta de Infraestrutura Ciente de Contexto para Suporte a Participação Colaborativa no Cenário da TV Digital Interativa Interativa**. In: XIV Brazilian Symposium on Multimedia and the Web **–** WebMedia. 2008. Vila Velha – ES.

AMO, Sandra de. **Técnicas de Mineração de Dados**. In: XXIV Congresso da Sociedade Brasileira de Computação. Jornada de Atualização em Informática. Salvador, 2004.

AMORIM, Thiago. **Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados.** Recife, 2006. Monografia;Curso Ciências da Computação) – Universidade Federal de Pernambuco- UFPE.

ARIB - Association of Radio Industries and Business. Disponível em <http://www.arib.or.jp>. Acessado em Jan de 2012.

ATSC - Advanced Television System Committee. Disponível em <http://www.atsc.org>. Acesso em Jan de 2012.

ARAUJO, Jônatas Pereira Cabral De. **CoreKTV - Uma infraestrutura baseada em conhecimento para TV Digital Interativa Interativa**: um estudo de caso para o middleware Ginga. 2011. João Pessoa. Dissertação de Mestrado.

AURÉLIO, Marcos; VELLASCO, Marley; LOPES, Carlos Henrique. **Descoberta de conhecimento e Mineração de Dados**. Rio de Janeiro: PUC-Rio, 1999. Disponível em: < http://www.ica.ele.puc-rio.br/cursos/download/DM-apostila1.pdf >. Acesso em: Jan. 2012.

AVILA, P. M., ZORZO, S.D., 2009. **A personalizad TV Guide System Compliant with Ginga**. In: XV Simpósio Brasileiro de Sistemas Multimídia e Web (Webmedia). Fortaleza - CE.

BARBIERI, C. **BI** – Business Intelligence. Rio de Janeiro. Axcel Books do Brasil Editora, 2001.

BENNETT, James; LANNING, Stan. **The Netflix Prize**. In: KDD Cup 2007. 2007

BERRY, M.J.A.; LINOFF, G. **Data mining techniques**. John Wiley & Sons, Inc. 1997.

BOENTE, Alfredo Nazareno Pereira; OLIVEIRA, Fabiano Saldanha Gomes de ; ROSA, José Luiz dos Anjos. **Utilização de Ferramentas de KDD para Integração de Aprendizagem e Tecnologia em Busca da Gestão Estratégica do Conhecimento na Empresa**. In: Simpósio de Excelência em Gestão e Tecnologia – SEGET. Rio de Janeiro , 2007.

BOLAÑOS, C.; VIEIRA, R. V.. **TV Digital Interativa no Brasil e no mundo**: estado da arte. In: Revista de Economía Política de las Tecnologías de la Información y Comunicación. Vol. VI, n. 2, Mayo – Ago. 2004

BOZIOS, T., LEKAKOS, G., SKOULARIDOU, V., and CHORIANOPOULOS, K. **Advanced techniques for personalized advertising in a digital TV environment**: The imedia system. In Proceedings of the eBusiness and eWork Conference, pp 1025-1031, IOS press, 2001.

DEITEL, H. M., DEITEL, P. J. **Java**: Como Programar. 6ª Ed. Porto Alegre: Bookman, 2005.

DIAMANTINI C., POTENA, D. and STORTI, E. **KDDONTO: an Ontology for Discovery and Composition of KDD Algorithms**. In Proc. of the ECML/PKDD09 Workshop on Third Generation Data Mining: Towards Service-oriented Knowledge Discovery, pages 13-24, Bled, Slovenia, Sep 7-11 2009 A.

DIAMANTINI, C., POTENA, D. and STORTI, E. **Ontology-driven KDD Process Composition**. In N. Adams et al. (Eds.), Proc. of the 8th International Symposium on Intelligent Data Analysis, LNCS, volume 5772, pages 285-296. Springer, 2009B.

DIAS, Maria Madalena. **Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados.** Florianópolis, 2001.Tese (Doutorado em Engenharia de Produção) – Universidade Federal de Santa Catarina. F. 197. Disponível em: < http://teses.eps.ufsc.br/defesa/pdf/3469.pdf >. Acesso em Jan. 2012.

DOMINGUES, Marcos Aurélio; REZENDE, Solange Oliveira. **Pós-processamento de regras de associação usando taxonomias**. In: *Infocomp Journal of Computer Science.. Lavras*: UFLA, 2005.

DVB - Digital Vídeo Broadcasting. Disponível em <http://www.dvb.org>. Acessado em Jan de 2012.

ELDER, J. F. IV ; PREGIBON, D. **A statistical perspective on knowledge discovery in data bases**. In: U. M. Fayyad et al. (Ed.) Advances in Knowledge Discovery and Data Mining, 83-113. AAAI/MIT Press, 1996.

EHRMANTRAUT, M., HARDER, T., WITTIG, H., STEINMETZ, R. **The Personal Electronic Program Guide** - towards the pre-selection of individual TV Programs. In Proc. of CIKM' 96, Rockville, MD, p. 243-50. 1996.

FAYYAD, U. M. et al. **Advances in Knowledge Discovery and Data Mining**. 1996, AAAIPress, The Mit Press.

FAYYAD, U. M.; PIATETSKY, G. Shapiro; SMYTH, P., **Advances in Knowledge Discovery and Data Mining**. 1996b editors AAAI Press / MIT Press, Menlo Park, CA

FERNANDEZ, M.; GOMEZ-PEREZ, A.; JURISTO, H. Methontology: from ontological art towards ontological engineering. 1997. Disponível em: <http://citeseer.ist.psu.edu/context/544607/0/>. Acesso em: Jan 2012.

FERREIRA, Jorge Abrantes. **Mineração de Dados na retenção de clientes em telefonia celular**. Rio de Janeiro, 2005. Dissertação de Mestrado (PUC-Rio)

FLACH, P.; LACHICHE, N. **Confirmation-Guided Discovery of First-Order Rules with Tertius**. In: Journal Machine Learning**.** Kluwer Academic. USA, 2001. pages. 61-95.

GILLMEISTER, Paulo Ricardo Guglieri; CAZELLA, Silvio César. **Uma Análise Comparativa de Algoritmos de Regras de Associação: Minerando Dados da Indústria Automotiva**. IN: Escola Regional de Banco de dados - ERBD**.** Caxias do Sul - RS, 2007.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining:** um guia prático – conceitos, técnicas, ferramentas, orientações e aplicações. Rio de Janeiro: Elsevier, 2005.261p.

GONÇALVES, Eduardo Corrêa. Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas, in: INFOCOMP – Journal of Computer Science. Rio de Janeiro. Universidade Federal Fluminense - UFF.

GOTTGTROY et. al. **An ontology driven approach for knowledge discovery in Biomedicine**. In: Zhang, C., Guesgen, H.W., Yeap, W.K. (Eds.), Lecture notes in artificial intelligence, Vol. 3157. Springer-verlag, Berlin.

GPL - General Public License. ***The GNU General Public License*.** Disponível em: < http://www.gnu.org/licenses/licenses.html#GPL>. Acesso em: Nov. 2012.

GRUNINGER, M.; FOX, M. S. Methodology for the design and evaluation of ontologies. 1995. Disponível em: <http://citeseer.ist.psu.edu/grninger95methodology.html>. Acesso em: Nov. 2011.

GUIMARÃES, Maximiliano Luiz Souza; LIMA, Iremar Nunes de. **Técnicas de Descoberta de Conhecimento em Sistemas de Apoio à Decisão**. 2006. Disponível em: < http://www.marcosmaurilioribeiro.net/websites/iremar/publicacoes/artigo06.pdf >. Acesso em: Set. 2011.

GUTTA, S. et al. **TV Content Recommender System**. In: Seventeenth National Conference on Artificial Intelligence. 2000. Austin, TX, USA, pp. 1121-1122.

GYŐRÖDI, Cornelia et al. **A Comparative Study of Association Rules Mining Algorithms.** in: Hungarian Joint Symposium on Applied Computational Intelligence, Oradea. 2004.

HAGEN P.. **Smart Personalizati**on. In: The Forrester Report, Forrester Research, Cambridge. 1999.

HAN, J. and KAMBER, M. ***Data Mining Concepts and Techniques***. 2a Edição, Editora Elsevier, Reino Unido. 2006.

HILARIO,Melanie; NGUYEN, Phong; DO, Huyen; WOZNICA, Adam; KALOUSIS, Alexandros. Ontology based meta-mining of knowledge discovery workflows. Book chapter in Meta-Learning in Computational Intelligence. Springer 2011

IBOPE. **Página inicial IBOPE**.2011. Disponível em: < www.**ibope**.com.br>. Acesso em: Jul. 2011.

INMON, W.H., HACKARTHORN, R. D. **Como usar o data warehouse**. 1997. Rio de Janeiro: IBPI Press.

IBGE - Instituto Brasileiro de Geografia e Estatística. 2008. **Síntese de Indicadores Sociais**. Brasil, 2008.

ISO; IEC. **Information technology - Generic coding of moving pictures and associated audio information - Part 1**: Systems - MPEG2. ISO/IEC 13818-1.2008.1992.

ITU – International Telecommunication Union. **ITU-T Recommendation J.200**: Worldwide common core – Application environment for digital interactive television services. 2001. (a)

ITU – International Telecommunication Union. **ITU-T Recommendation J.201**: Harmonization of declarative content format for interactive television applications. 2003.(b)

ITU – International Telecommunication Union. **ITU-T Recommendation J.202**: Harmonization of procedural content formats for interactive TV applications. 2004.(c)

JAVA. Disponível em <http://www.oracle.com/technetwork/java/index.html>. Acessado em Abr de 2012.

KETTLE. Disponível em <http://kettle.pentaho.com/>. Acessado em Set de 2011.

KIMBALL, R. **Data Warehouse Toolkit**. 1998. São Paulo: Makron Books.

LAVID; TELEMIDIA. ***Ginga Digital TV Middleware Specification***. Disponível em: < http://www.ginga.org.br/>. Acesso em: Nov 2011.

LEKAKOS, G.; GIAGLIS, G. 2002. **Delivering personalized advertisements** in digital television: A methodology and empirical evaluation.

LEMOS, G.; FERNANDES, J.; ELIAS, G.. **Introdução à Televisão Digital Interativa: Arquitetura, Protocolos, Padrões e Práticas.** In: JAI Jornada de Atualização em informática. Salvador – BA. UFBA. 2004.

LINO, N. Q.; ARAÚJO, J.; ANABUKI, D.; PATRÍCIO JUNIOR, J. C. A.; BATISTA, M.; NOBREGA, R.; AMARO, M.; SIEBRA, C. **Knowledge TV.** In: European Conference on Interactive TV and Video – Euro ITV 2011. Lisboa -Portugal.2011.

LUCAS, Adriano S. ; ZORZO, Sérgio D. . **Personalização para Televisão Digital utilizando a estratégia de Sistema de Recomendação para ambientes multiusuário**. In: XXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos – SBRC. Recife. 2009.

LUGMAYR, A.; NIIRANEN, S. and KALLI, S. (2004). ***Digital Interactive TV and Metadata*** *in: Future Broadcast Multimedia.* Springer-Verlag. New York-USA. 2004.

MANGUEIRA, J. A.; OLIVEIRA, F. S. de; ALVES, K. C.; MEDEIROS, Á. F. de C.; LEMOS, G.. **JCollab**: Uma Ferramenta para Produção e Distribuição de Telejornais no Contexto da Web 2.0. In XXXVI Conferência Latino-Americana de Informatica –CLEI. Assunção – Paraguai. 2010.

MANNILA, H. **Data mining**: machine learning, statistics, and databases. In *International Conference on Scientific and Statistical Database Management*. Stockholm. 1996.

MARGALHO, M.; FRANCÊS, R.; COSTA, J. C. W. A**.. Canal de Retorno para TV Digital Interativa com Interatividade Condicionada por Mecanismo de Sinalização Contínua e Provisionamento de Banda Orientado a QoS**. IEEE LATIN AMERICA TRANSACTIONS - VOL. 5 - NO. 5.2007.

POSTGRESQL. **Pagina oficial POSTGRESQL**. Disponível em: < http://www.postgresql.com/>. Acesso em: 13 Jul. 2011.

NIGRO, H. O.; CISARO S. G.; XODO, D. H**. Data Mining With Ontologies: Implementations, Findings and Frameworks, Information**. Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2007.

NOY, F. N.; GUINNESS, D. L. **Ontology development 101: a guide to create your first ontology**. 2001. Disponível em: <http://ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.doc>. Acesso em: Out 2011.

O'SULLIVAN, D.; SMYTH, B.; WILSON, D. C.; MCDONALD, K.; SMEATON, A.: **Interactive Television Personalization**: From Guides to Programs. In: PERSONALIZED DIGITAL TELEVISION: TARGETING PROGRAMS TO INDIVIDUAL VIEWERS. New York- Unites Estates. Kluwer Academic Publishers, pp 73-91. 2004.

PORCARO, R. M. ; LIFSCHITZ, S. ; CORTES, S. C. . **Mineração de Dados:** funcionalidades, técnicas e abordagens. Monografia em Ciência da Computação, PUC - Rio de Janeiro, 2002. Disponível em: < ftp://ftp.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf>. Acesso em: Set. 2011.

PROTEGÉ. **Welcome To Protégé**. Disponível em: < http://protege.stanford.edu/>. Acesso em: 18 Abril 2012.

ROMÃO, Wesley. **Descoberta de conhecimento relevante em banco de dados sobre ciência e tecnologia.** Florianópolis: UFSC, 2002. Disponível em: <http://teses.eps.ufsc.br/defesa/pdf/3079.pdf>. Acesso em: Ago. 2011.

RUSSELL, S., and NORVIG, P. **Artificial Intelligence**: A Modern Approach. Prentice Hall, 3rd edition. 2009.

SBTVDI – Sistema Brasileiro de TV Digital Interativa. **Especificação Técnica de Referência**. Disponível em <http://sbTVDI.cpqd.com.br/>. Acessado em Fev de 2011.

SCHEIDT, A.; KÖERICH, F. G.; SANTOS, O. dos. **Seminário de Data mining**. Cascavel. 2008. Disponível em: < www.inf.unioeste.br/~olguin/4458-semin/g1-monografia.pdf>. Acesso em: Nov. 2011.

SHIM, J. P.; WARKENTIN, M.; COURTNEY, J.; POWER, D.; SHARDA, R. and CARLSSON, C. ***Past, Present, and Future of Decision Support Technology***. Decision Support Systems, Vol. 33(2) pp. 111-126. 2002.

SILBERSCHATZ, KORTH e SUDARSHAN. **Sistemas de Bancos de Dados**, 3a edição, Makron.1999.

SILVA, Fábio Santos da. **Personalização de Conteúdo na TVDI através de um Sistema de Recomendação Personalizada de Programas de TV (SRPTV)**. In**:** III Fórum de Oportunidades em Televisão Digital Interativa. Poços de Caldas – MG. 2005.

SCHEFFER, T. **Finding association rules that trade support optimally against confidence**. In: European conference on principles of data mining and knowledge discovery. 2004. pages. 424-435.

SOARES, L. F. G., RODRIGUES, R. F. e MORENO, M. F. **Ginga-NCL: the Declarative Environment of the Brazilian Digital TV System**. In: Journal of the Brazilian Computer Society. 2007, Vol. v12, pp. 37-46.

SOUZA, C. T.; OLIVEIRA, C. T. de.. **Especificações de Canal de Retorno em Aplicações para TV Digital Interativa Interativa**. In: XXII Simpósio Brasileiro de Telecomunicações. Campinas – SP. 2005.

SOUZA FILHO, G. L. DE, LEITE, L. E. C. E BATISTA, C. E. C. F. **Ginga-J: The Procedural Middleware for the Brazilian Digital TV System**. In: Journal of the Brazilian Computer Society. 2007, Vol. v12, pp. 47-56.

STUDER, R. et al. **Knowledge engineering: principles and methods**. Data & Knowledge Engineering, v.25, n.1/2, Março de 1998.

USCHOLD, M.; KING, M. Towards a Methodology for Building Ontologies. 1995. Disponível em: <http://citeseer.ist.psu.edu/uschold95toward.html> Acesso em: Nov. 2011.

WEKA. University of Waikato. **Weka 3** – Machine Learning Software in Java. Disponível no site da University of Waikato. 2010. Disponível em: < http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: Nov. 2011.

W3C. World Wide Web Consortium. ***W3C Semantic Web Activity***. Disponível em: < http://www.w3.org/2001/sw/>. Acesso em: Jul. 2011 (a).

W3C. World Wide Web Consortium. **OWL Web Ontology Language – Overview**. Disponível em: < http://www.w3.org/TR/owl-features/> (b). Acesso em: Jul. 2011.

W3C. World Wide Web Consortium. **Extensible Markup Language (XML)**. Disponível em: < http://www.w3.org/XML/> (c). Acesso em: Jul. 2011.

XAVIER, D. S. F.. **Análise de algoritmos de Mineração de Dados.** Monografia (Curso Ciências da Computação) – Centro Universitário de João Pessoa - UNIPÊ. João Pessoa. 2007.

XU, Shouhuai; ZHANG, Weining. **Knowledge as a Service and Knowledge Breaching**. In: SCC '05 Proceedings of the 2005 IEEE International Conference on Services Computing - Volume 01, Pages 87 - 94,IEEE Computer Society Washington, DC, USA ©2005.

APÊNDICE I – TABELA DE METADADOS SERVICE INFORMATION - SI

#	Metadado	Origem	Descrição
1	content_nibble_level_1	EIT	Informação de gênero
2	content_nibble_level_2	EIT	Informação de subgênero
3	country_code	EIT/PMT	Informação de país
4	content_rating	EIT/PMT	Classificação temática do conteúdo
5	age_rating	EIT/PMT	Classificação etária do conteúdo
6	event_name	EIT	Nome do conteúdo (programa)
7	short_description	EIT	Breve descrição do conteúdo
8	sh_language_code	EIT	Indica o idioma da descrição do conteúdo
9	event_id	EIT	Identificador único do evento (programa)
10	start_time	EIT	Horário de início do evento (programa)
11	Duration	EIT	Duração do evento (programa)
12	stream_content	EIT/PMT	Especifica o tipo do fluxo (áudio, video ou dados)
13	component_type	EIT/PMT	Especifica o tipo do componente de áudio, vídeo ou dados
14	component_description	EIT/PMT	Descrição em texto do fluxo do componente
15	cd_language_code	EIT/PMT	Indica o idioma da descrição do componente
16	audio_component_type	EIT	Especifica o tipo do componente de áudio
17	audio_stream_type	EIT	Especifica o tipo do fluxo de áudio
18	audio_multilanguage	EIT	Indica se há mais dois

			idiomas
19	audio_quality_indicator	EIT	Indica modo de qualidade do áudio
20	audio_sample_rate	EIT	Indica a frequencia de amostragem
21	audio_language_1	EIT	Identifica o primeiro idioma
22	audio_language_2	EIT	Identifica o segundo idioma
23	audio_description	EIT	Descrição do componente de áudio
24	video_encode_format	EIT	Indica o formato de codificação de vídeo
25	series_id	EIT	Identificador da série
26	series_repeat_label	EIT	Fornece rótulo de identificação do programa
27	series_program_pattern	EIT	Fornece padrão de transmissão do programa
28	series_ep_number	EIT	Número do episódio da série
29	series_last_ep_number	EIT	Número do último episódio da série
30	series_expire_date	EIT	Data limite do seriado
31	series_name	EIT	Nome da série
32	service_type	SDT	Especifica o tipo do serviço
33	service_provider_name	SDT	Nome do fornecedor do serviço
34	service_name	SDT	Nome do serviço
35	service_id	SDT/EIT	Identificador do serviço
36	service_countries_avability_1	SDT	Lista os países onde o serviço está disponível
37	service_countries_avability_2	SDT	Lista os países onde o serviço não está disponível
38	service_list	NIT/BIT	Lista de serviços transmitidos pela rede/radiodifusor
39	network_name	NIT	Nome da rede
40	network_id	NIT	Identificador da rede
41	state_area_code	NIT/PMT	Estado alvo para

			transmissão de informação de emergência
42	microregion_area_code	NIT/PMT	Microregião alvo para transmissão de informação de emergência
43	signal_level	NIT/PMT	Corresponde ao sinal de alarme de emergência especificado pelos órgãos responsáveis
44	avc_video_profile_idc	PMT	Exibe o perfil do fluxo de vídeo AVC
45	avc_video_level_idc	PMT	Mostra o nível do fluxo de vídeo AVC
46	avc_video_still_present	PMT	Indica se vídeo contém imagens estáticas
47	avc_video_24h_picture	PMT	Indica se o vídeo contém imagens 24 horas
48	aac_audio_type	PMT	Indica o tipo do áudio transmitido
49	program_number	PMT	Identificador de um programa na tabela
50	broadcaster_name	BIT	Nome do radiodifusor
51	broadcaster_id	BIT	Identificador do radiodifusor
52	broadcast_view_property	BIT	Informa se a indicação do usuário para o nome do radiodifusor é apropriado ou não
53	local_time_offset	TOT	Informa a diferença de horário em relação ao UTC-3 na faixa de $\pm$12
54	utc-3_time	TOT	Horário no formato UTC-3

APÊNDICE II – REGRAS GERADAS PELO ALGORITMO PREDICTIVE APRIORI

=== Run information ===

Scheme: weka.associations.PredictiveApriori -N 100 -c -1

Relation: ibope

Instances: 21660

Attributes: 7

canal

diasemana

periododia

programa

genero

classificacaoanoestreia

tamanhoduracao

=== Associator model (full training set) ===

PredictiveApriori

==========

Best rules found:

1. canal=GLO diasemana=sexta-feira periododia=NOITE genero=Telenovela tamanhoduracao=MEDIO 298 ==> classificacaoanoestreia=ATUAL 298 acc:(0.99499)

2. canal=GLO diasemana=sexta-feira periododia=NOITE classificacaoanoestreia=ATUAL tamanhoduracao=MEDIO 298 ==> genero=Telenovela 298 acc:(0.99499)

3. canal=GLO diasemana=sexta-feira genero=Telenovela classificacaoanoestreia=ATUAL tamanhoduracao=MEDIO 298 ==> periododia=NOITE 298 acc:(0.99499)

4. programa=Globo Rural 289 ==> canal=GLO periododia=MANHÃ 289 acc:(0.99499)

5. programa=Globo Rural 289 ==> canal=GLO genero=Jornalístico 289 acc:(0.99499)

6. canal=GLO programa=Globo Rural 289 ==> periododia=MANHÃ tamanhoduracao=CURTO 289 acc:(0.99499)

7. canal=GLO programa=Globo Rural 289 ==> periododia=MANHÃ genero=Jornalístico 289 acc:(0.99499)

8. periododia=MANHÃ programa=Globo Rural 289 ==> canal=GLO tamanhoduracao=CURTO 289 acc:(0.99499)

9. periododia=MANHÃ programa=Globo Rural 289 ==> canal=GLO classificacaoanoestreia=MAIS 20 ANOS 289 acc:(0.99499)

10. programa=Globo Rural genero=Jornalístico 289 ==> canal=GLO tamanhoduracao=CURTO 289 acc:(0.99499)

11. programa=Globo Rural genero=Jornalístico 289 ==> canal=GLO classificacaoanoestreia=MAIS 20 ANOS 289 acc:(0.99499)

12. programa=Globo Rural classificacaoanoestreia=MAIS 20 ANOS 289 ==> canal=GLO tamanhoduracao=CURTO 289 acc:(0.99499)

13. programa=Globo Rural classificacaoanoestreia=MAIS 20 ANOS 289 ==> periododia=MANHÃ genero=Jornalístico 289 acc:(0.99499)

14. programa=Globo Rural tamanhoduracao=CURTO 289 ==> canal=GLO classificacaoanoestreia=MAIS 20 ANOS 289 acc:(0.99499)

15. programa=Globo Rural tamanhoduracao=CURTO 289 ==> periododia=MANHÃ genero=Jornalístico 289 acc:(0.99499)

16. canal=GLO periododia=MANHÃ programa=Globo Rural 289 ==> genero=Jornalístico classificacaoanoestreia=MAIS 20 ANOS 289 acc:(0.99499)

17. canal=GLO periododia=MANHÃ programa=Globo Rural 289 ==> genero=Jornalístico tamanhoduracao=CURTO 289 acc:(0.99499)

18. canal=GLO programa=Globo Rural genero=Jornalístico 289 ==> periododia=MANHÃ classificacaoanoestreia=MAIS 20 ANOS 289 acc:(0.99499)

19. canal=GLO programa=Globo Rural genero=Jornalístico 289 ==> classificacaoanoestreia=MAIS 20 ANOS tamanhoduracao=CURTO 289 acc:(0.99499)

20. canal=GLO programa=Globo Rural classificacaoanoestreia=MAIS 20 ANOS 289 ==> genero=Jornalístico tamanhoduracao=CURTO 289 acc:(0.99499)

21. canal=GLO programa=Globo Rural tamanhoduracao=CURTO 289 ==> periododia=MANHÃ classificacaoanoestreia=MAIS 20 ANOS 289 acc:(0.99499)

22. canal=GLO programa=Globo Rural tamanhoduracao=CURTO 289 ==> genero=Jornalístico classificacaoanoestreia=MAIS 20 ANOS 289 acc:(0.99499)

23. periododia=MANHÃ programa=Globo Rural genero=Jornalístico 289 ==> classificacaoanoestreia=MAIS 20 ANOS tamanhoduracao=CURTO 289 acc:(0.99499)

24. programa=Bom Dia & Cia 283 ==> canal=SBT genero=Infantil 283 acc:(0.99499)

25. programa=Bom Dia & Cia 283 ==> canal=SBT classificacaoanoestreia=DE 10 ATÉ 20 ANOS 283 acc:(0.99499)

26. canal=SBT programa=Bom Dia & Cia 283 ==> genero=Infantil classificacaoanoestreia=DE 10 ATÉ 20 ANOS 283 acc:(0.99499)

27. canal=SBT programa=Bom Dia & Cia 283 ==> genero=Infantil tamanhoduracao=LONGO 283 acc:(0.99499)

28. programa=Bom Dia & Cia genero=Infantil 283 ==> canal=SBT tamanhoduracao=LONGO 283 acc:(0.99499)

29. programa=Bom Dia & Cia genero=Infantil 283 ==> classificacaoanoestreia=DE 10 ATÉ 20 ANOS tamanhoduracao=LONGO 283 acc:(0.99499)

30. programa=Bom Dia & Cia classificacaoanoestreia=DE 10 ATÉ 20 ANOS 283 ==> canal=SBT tamanhoduracao=LONGO 283 acc:(0.99499)

31. programa=Bom Dia & Cia classificacaoanoestreia=DE 10 ATÉ 20 ANOS 283 ==> genero=Infantil tamanhoduracao=LONGO 283 acc:(0.99499)

32. programa=Bom Dia & Cia tamanhoduracao=LONGO 283 ==> genero=Infantil classificacaoanoestreia=DE 10 ATÉ 20 ANOS 283 acc:(0.99499)

33. canal=SBT genero=Infantil classificacaoanoestreia=DE 10 ATÉ 20 ANOS 283 ==> programa=Bom Dia & Cia tamanhoduracao=LONGO 283 acc:(0.99499)

34. canal=OUT periododia=MANHÃ tamanhoduracao=LONGO 269 ==> genero=Musical classificacaoanoestreia=ATUAL 269 acc:(0.99498)

35. programa=Globo Notícia 266 ==> canal=GLO periododia=TARDE 266 acc:(0.99498)

36. programa=Globo Notícia 266 ==> canal=GLO genero=Jornalístico 266 acc:(0.99498)

37. canal=GLO programa=Globo Notícia 266 ==> periododia=TARDE tamanhoduracao=CURTO 266 acc:(0.99498)

38. canal=GLO programa=Globo Notícia 266 ==> periododia=TARDE genero=Jornalístico 266 acc:(0.99498)

39. periododia=TARDE programa=Globo Notícia 266 ==> canal=GLO tamanhoduracao=CURTO 266 acc:(0.99498)

40. periododia=TARDE programa=Globo Notícia 266 ==> canal=GLO classificacaoanoestreia=ATÉ 5 ANOS 266 acc:(0.99498)

41. periododia=TARDE programa=Malhação 266 ==> canal=GLO genero=Telenovela 266 acc:(0.99498)

42. periododia=TARDE programa=Malhação 266 ==> canal=GLO classificacaoanoestreia=DE 10 ATÉ 20 ANOS 266 acc:(0.99498)

43. programa=Globo Notícia genero=Jornalístico 266 ==> canal=GLO tamanhoduracao=CURTO 266 acc:(0.99498)

44. programa=Globo Notícia genero=Jornalístico 266 ==> canal=GLO classificacaoanoestreia=ATÉ 5 ANOS 266 acc:(0.99498)

45. programa=Globo Notícia classificacaoanoestreia=ATÉ 5 ANOS 266 ==> canal=GLO tamanhoduracao=CURTO 266 acc:(0.99498)

46. programa=Globo Notícia classificacaoanoestreia=ATÉ 5 ANOS 266 ==> periododia=TARDE genero=Jornalístico 266 acc:(0.99498)

47. programa=Globo Notícia tamanhoduracao=CURTO 266 ==> canal=GLO classificacaoanoestreia=ATÉ 5 ANOS 266 acc:(0.99498)

48. programa=Globo Notícia tamanhoduracao=CURTO 266 ==> periododia=TARDE genero=Jornalístico 266 acc:(0.99498)

49. diasemana=segunda-feira genero=Filme 260 ==> canal=GLO classificacaoanoestreia=ATUAL 260 acc:(0.99498)

50. diasemana=segunda-feira genero=Filme 260 ==> canal=GLO tamanhoduracao=MEDIO 260 acc:(0.99498)

51. canal=REC genero=Variedades 251 ==> classificacaoanoestreia=ATÉ 5 ANOS tamanhoduracao=LONGO 251 acc:(0.99498)

52. periododia=MANHÃ programa=Bom Dia & Cia 251 ==> canal=SBT tamanhoduracao=LONGO 251 acc:(0.99498)

53. periododia=MANHÃ programa=Bom Dia & Cia 251 ==> genero=Infantil classificacaoanoestreia=DE 10 ATÉ 20 ANOS 251 acc:(0.99498)

54. programa=O Astro 241 ==> canal=GLO periododia=NOITE 241 acc:(0.99498)

55. programa=O Astro 241 ==> canal=GLO genero=Telenovela 241 acc:(0.99498)

56. canal=GLO programa=O Astro 241 ==> periododia=NOITE tamanhoduracao=CURTO 241 acc:(0.99498)

57. canal=GLO programa=O Astro 241 ==> periododia=NOITE genero=Telenovela 241 acc:(0.99498)

58. periododia=NOITE programa=O Astro 241 ==> classificacaoanoestreia=ATUAL 241 acc:(0.99498)

59. programa=O Astro genero=Telenovela 241 ==> classificacaoanoestreia=ATUAL 241 acc:(0.99498)

60. programa=O Astro classificacaoanoestreia=ATUAL 241 ==> tamanhoduracao=CURTO 241 acc:(0.99498)

61. programa=Hoje em Dia 236 ==> canal=REC periododia=MANHÃ 236 acc:(0.99497)

62. programa=Hoje em Dia 236 ==> canal=REC genero=Jornalístico 236 acc:(0.99497)

63. canal=REC programa=Hoje em Dia 236 ==> periododia=MANHÃ tamanhoduracao=LONGO 236 acc:(0.99497)

64. canal=REC programa=Hoje em Dia 236 ==> periododia=MANHÃ genero=Jornalístico 236 acc:(0.99497)

65. periododia=MANHÃ programa=Hoje em Dia 236 ==> canal=REC tamanhoduracao=LONGO 236 acc:(0.99497)

66. periododia=MANHÃ programa=Hoje em Dia 236 ==> canal=REC classificacaoanoestreia=ATÉ 5 ANOS 236 acc:(0.99497)

67. canal=REC programa=Hoje em Dia genero=Jornalístico 236 ==> classificacaoanoestreia=ATÉ 5 ANOS 236 acc:(0.99497)

68. canal=REC programa=Hoje em Dia tamanhoduracao=LONGO 236 ==> classificacaoanoestreia=ATÉ 5 ANOS 236 acc:(0.99497)

69. diasemana=quarta-feira programa=O Clone 231 ==> canal=GLO periododia=TARDE 231 acc:(0.99497)

70. diasemana=quarta-feira programa=O Clone 231 ==> canal=GLO genero=Telenovela 231 acc:(0.99497)

71. diasemana=sábado genero=Telenovela 224 ==> canal=GLO periododia=NOITE 224 acc:(0.99497)

72. diasemana=sábado genero=Telenovela 224 ==> canal=GLO classificacaoanoestreia=ATUAL 224 acc:(0.99497)

73. canal=GLO diasemana=sábado periododia=NOITE genero=Telenovela classificacaoanoestreia=ATUAL 224 ==> tamanhoduracao=MEDIO 224 acc:(0.99497)

74. canal=GLO periododia=MANHÃ genero=Esportivo 223 ==> diasemana=domingo 223 acc:(0.99497)

75. diasemana=quinta-feira programa=Sessão da Tarde 222 ==> canal=GLO periododia=TARDE 222 acc:(0.99497)

76. diasemana=quinta-feira programa=Sessão da Tarde 222 ==> canal=GLO genero=Filme 222 acc:(0.99497)

77. programa=Record Notícias 218 ==> canal=REC periododia=TARDE 218 acc:(0.99497)

78. programa=Record Notícias 218 ==> canal=REC genero=Jornalístico 218 acc:(0.99497)

79. canal=REC programa=Record Notícias 218 ==> periododia=TARDE tamanhoduracao=LONGO 218 acc:(0.99497)

80. canal=REC programa=Record Notícias 218 ==> periododia=TARDE genero=Jornalístico 218 acc:(0.99497)

81. periododia=TARDE programa=Record Notícias 218 ==> classificacaoanoestreia=ATÉ 5 ANOS 218 acc:(0.99497)

82. diasemana=domingo genero=Entretenimento 217 ==> periododia=NOITE 217 acc:(0.99496)

83. programa=Esporte Espetacular 216 ==> canal=GLO diasemana=domingo 216 acc:(0.99496)

84. programa=Esporte Espetacular 216 ==> canal=GLO genero=Esportivo 216 acc:(0.99496)

85. canal=GLO programa=Esporte Espetacular 216 ==> diasemana=domingo tamanhoduracao=LONGO 216 acc:(0.99496)

86. canal=GLO programa=Esporte Espetacular 216 ==> diasemana=domingo genero=Esportivo 216 acc:(0.99496)

87. diasemana=terça-feira programa=O Clone 216 ==> canal=GLO periododia=TARDE 216 acc:(0.99496)

88. diasemana=terça-feira programa=O Clone 216 ==> canal=GLO genero=Telenovela 216 acc:(0.99496)

89. diasemana=domingo programa=Esporte Espetacular 216 ==> classificacaoanoestreia=MAIS 20 ANOS 216 acc:(0.99496)

90. periododia=TARDE programa=SPTV 214 ==> canal=GLO 214 acc:(0.99496)

91. programa=Faixa de Clipes 213 ==> canal=OUT periododia=MANHÃ 213 acc:(0.99496)

92. programa=Faixa de Clipes 213 ==> canal=OUT genero=Musical 213 acc:(0.99496)

93. canal=OUT programa=Faixa de Clipes 213 ==> periododia=MANHÃ tamanhoduracao=LONGO 213 acc:(0.99496)

94. canal=OUT programa=Faixa de Clipes 213 ==> periododia=MANHÃ genero=Musical 213 acc:(0.99496)

95. periododia=MANHÃ programa=Faixa de Clipes 213 ==> classificacaoanoestreia=ATUAL 213 acc:(0.99496)

96. programa=Faixa de Clipes genero=Musical 213 ==> classificacaoanoestreia=ATUAL 213 acc:(0.99496)

97. programa=Faixa de Clipes classificacaoanoestreia=ATUAL 213 ==> tamanhoduracao=LONGO 213 acc:(0.99496)

98. canal=OUT programa=Faixa de Clipes genero=Musical 213 ==> periododia=MANHÃ classificacaoanoestreia=ATUAL 213 acc:(0.99496)

99. canal=OUT programa=Faixa de Clipes tamanhoduracao=LONGO 213 ==> classificacaoanoestreia=ATUAL 213 acc:(0.99496)

100. periododia=MANHÃ programa=Faixa de Clipes genero=Musical 213 ==> tamanhoduracao=LONGO 213 acc:(0.99496)

APÊNDICE III – REGRAS GERADAS PELO ALGORITMO TERTIUS

Scheme: weka.associations.Tertius -K 10 -F 0.0 -N 1.0 -L 4 -G 0 -c 0 -I 0 -P 0

Relation: ibope

Attributes: 7

canal

diasemana

periododia

programa

genero

classificacaoanoestreia

tamanhoduracao

=== Associator model (full training set) ===

Tertius

=======

1. /* 0,685640 0,026560 */ genero = Jornalístico ==> programa = SPTV or classificacaoanoestreia = MAIS 20 ANOS or canal = REC

2. /* 0,673594 0,118637 */ canal = GLO ==> programa = Sessão da Tarde or genero = Telenovela or classificacaoanoestreia = MAIS 20 ANOS

3. /* 0,653587 0,132802 */ canal = GLO ==> programa = SPTV or genero = Telenovela or classificacaoanoestreia = MAIS 20 ANOS

4. /* 0,636866 0,144754 */ canal = GLO ==> programa = Mais Você or genero = Telenovela or classificacaoanoestreia = MAIS 20 ANOS

5. /* 0,631457 0,000000 */ genero = Jornalístico and canal = GLO ==> programa = SPTV or classificacaoanoestreia = MAIS 20 ANOS

6. /* 0,617817 0,158477 */ canal = GLO ==> programa = TV Globinho or genero = Telenovela or classificacaoanoestreia = MAIS 20 ANOS

7. /* 0,614758 0,160691 */ canal = GLO ==> programa = Bem Estar or genero = Telenovela or classificacaoanoestreia = MAIS 20 ANOS

8. /* 0,607607 0,073041 */ canal = GLO ==> programa = SPTV or classificacaoanoestreia = MAIS 20 ANOS or tamanhoduracao = MEDIO

9. /* 0,606287 0,018150 */ classificacaoanoestreia = ATUAL and canal = GLO ==> programa = Sessão da Tarde or genero = Telenovela

10. /* 0,600943 0,064188 */ canal = GLO ==> genero = Telenovela or classificacaoanoestreia = MAIS 20 ANOS or tamanhoduracao = MEDIO

Number of hypotheses considered: 1184210

Number of hypotheses explored: 616481