UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UMA ABORDAGEM PARA IDENTIFICAÇÃO DE DOMÍNIOS DE APLICAÇÃO EM AMBIENTE DE CONVERGÊNCIA DIGITAL

AMANDA DRIELLY DE SOUZA PIRES

JOÃO PESSOA-PB
JULHO-2013

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UMA ABORDAGEM PARA IDENTIFICAÇÃO DE DOMÍNIOS DE APLICAÇÃO EM AMBIENTE DE CONVERGÊNCIA DIGITAL

AMANDA DRIELLY DE SOUZA PIRES

JOÃO PESSOA-PB
JULHO-2013

AMANDA DRIELLY DE SOUZA PIRES

UMA ABORDAGEM PARA IDENTIFICAÇÃO DE DOMÍNIOS DE APLICAÇÃO EM AMBIENTE DE CONVERGÊNCIA DIGITAL

DISSERTAÇÃO APRESENTADA AO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DA PARAÍBA, COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE EM INFORMÁTICA (SISTEMAS DE COMPUTAÇÃO).

Orientador: Prof. Dr. Ed Porto Bezerra
Co-Orientadora: Profa. Dra. Natasha Correia Queiroz Lino

JOÃO PESSOA-PB
JULHO-2013

P667u Pires, Amanda Drielly de Souza.
Uma abordagem para identificação de domínios de aplicação em ambiente de convergência digital / Amanda Drielly de Souza Pires.-- João Pessoa, 2013.
114f. : il.
Orientador: Ed Porto Bezerra
Coorientadora: Natasha Correia Queiroz Lino
Dissertação (Mestrado) – UFPB/CI
1. Informática. 2. Sistemas de computação. 3. Televisão digital interativa. 4. Web semântica. 5. Análise de similaridade léxica e semântica.

UFPB/BC CDU: 004(043)

Ata da Sessão Pública de Defesa de Dissertação de Mestrado de **Amanda Drielly de Souza Pires**, candidata ao Título de Mestre em Informática na Área de Sistemas de Computação, realizada em 23 de Julho de 2013.

Ao vigésimo terceiro dia do mês de Julho do ano dois mil e treze, às dez horas, no auditório do CCEN- da Universidade Federal da Paraíba, reuniram-se os membros da Banca Examinadora constituída para examinar a candidata ao grau de Mestre em Informática, na área de "*Sistemas de Computação*", na linha de pesquisa "*Computação Distribuída*", a Sra. **Amanda Drielly de Souza Pires**. A comissão examinadora foi composta pelos professores doutores: ED PORTO BEZERRA (PPGI-UFPB), Orientador e Presidente da Banca, CLAUIRTON DE ALBUQUERQUE SIEBRA (PPGI-UFPB), examinador interno, NATASHA CORREIA QUEIROZ LINO (PPGI-UFPB), examinadora interna e FRANKLIN DE SOUZA RAMALHO (UFCG) como examinador externo. Dando início aos trabalhos, o professor ED PORTO BEZERRA cumprimentou os presentes, comunicou aos mesmos a finalidade da reunião e passou a palavra à candidata para que a mesma fizesse, oralmente, a exposição do trabalho de dissertação intitulado "*Uma Abordagem para Identificação Semântica de Domínios de Aplicação em Ambientes de Convergência Digital*". Concluída a exposição, a candidata foi arguida pela Banca Examinadora que emitiu o seguinte parecer: "*Aprovada*". Assim sendo, deve a Universidade Federal da Paraíba expedir o respectivo diploma de Mestre em Informática na forma da lei e, para constar, eu, Micherlon Xavier Bezerra, Assistente em Administração, servindo de secretário, lavrei a presente ata que vai assinada por mim e pelos membros da Banca Examinadora. João Pessoa, 23 de Julho de 2013.

Micherlon Xavier Bezerra

Prof. Dr. Ed Porto Bezerra
Orientador (PPGI-UFPB)

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

Prof. Dra. Natasha Correia Queiroz Lino
Examinadora Interna (PPGI-UFPB)

Prof. Dr. Franklin de Souza Ramalho
Examinador Externo (UFCG)

Aos meus pais e ao meu marido, por toda dedicação.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida e por essa vitória. Aos meus pais, Erinaldo e Doracy, que apesar da distância me incentivaram e me apoiaram nos momentos difíceis.

Aos meus irmãos, Carla e Bruno, pelo carinho e torcida. Ao meu marido Radams Venceslau, que soube compreender minha ausência e superar a distância. Obrigado pelo incentivo e parceria.

Aos demais familiares, em especial meu avô Sabino Targino (in memoriam), pelo apoio aos meus estudos, incentivando-me sempre. A minha prima Tatiana Veras, seu marido Flávio Sousa e seu filho João Gabriel, pelo acolhimento em sua casa e pela preocupação constante com meu bem estar. A minha amiga Fernanda Moura pela ajuda e incentivo indispensável nessa conquista. Ao CNPQ, pelo auxílio financeiro que permitiu a concretização deste trabalho.

A meu orientador Ed Porto, pela amizade, confiança e atenção a mim dispensada. A minha co-orientadora Natasha Queiroz, pelos ensinamentos, apoio e disponibilidade, fundamentais nessa etapa.

Aos professores do PPGI, pelo comprometimento que contribuiu muito em meu crescimento profissional. Aos meus amigos de mestrado, em especial a Sttiwe Washington, Lisieux Andrade, Priscilla Vieira, Mainara Nóbrega e Wanderson Souza, pelos inúmeros auxílios e contribuições, em diferentes momentos, sempre prestativos. Aos colegas do projeto Knowledge TV, Dalia Maria e Manoel Amaro, pela disponibilidade e compreensão.

RESUMO

O surgimento da Televisão Digital Interativa proporciona além de ganho de qualidade na transmissão, a adição de novos recursos e serviços disponíveis ao usuário. Com o advento da convergência digital entre as plataformas de TV e *Web*, novas propostas de organização semântica de conteúdo estão sendo desenvolvidas. Além disso, foi possível introduzir conceitos da *Web* Semântica e de representação do conhecimento que permitem descrever semanticamente os metadados de conteúdo através de ontologias. Nesse contexto, esse trabalho propõe uma abordagem para identificação de domínios de aplicação no ambiente de convergência digital baseada em conceitos da *Web* Semântica e nas análises de similaridade léxica e semântica. Um componente integrado a plataforma Knowledge TV, foi implementado para validar a abordagem.

Palavras-chave: Televisão Digital Interativa; *Web* Semântica; Ontologia; Análise de Similaridade Léxica e Análise de Similaridade Semântica.

ABSTRACT

The emergence of the Interactive Digital Television provided, as well as advantages gain quality and optimization of the transmission, the addition of new features and services available to the user. With the advent of digital convergence between TV and Web platforms, new proposals of semantic organization of content are developed. Moreover, it was possible to introduce concepts of the Semantic Web and knowledge representation that allow semantically describe the metadata of content through ontologies. In this context, this work proposes an approach to identifying of domain of application in digital convergence environment based on the Semantic Web concepts and analysis of lexical and semantic similarity. One component integrated with Knowledge TV platform, was implemented to validate the approach.

Keywords*: Interactive Digital TV; Web Semantic, Ontology, Analysis of Lexical Similarity and Analysis of Semantic Similarity.*

LISTA DE ILUSTRAÇÕES

Figura 1 - Camadas do Sistema Brasileiro de TV Digital 23
Figura 2 - Arquitetura conceitual do middleware Ginga 25
Figura 3 - Camadas lógicas da Web Semântica 30
Figura 4 - Classificação de ontologias 34
Figura 5 - Representação de uma declaração em RDF 36
Figura 6 - Exemplo de sentença RDF 36
Figura 7 - Relação entre os diferentes tipos de OWL 37
Figura 8 - Definição da função de *Levenshtein* 40
Figura 9 - Fluxograma do algoritmo de radicalização *Porter* 42
Figura 10 - Fluxograma do algoritmo de radicalização *Lovins* 43
Figura 11 - Fluxograma do algoritmo de radicalização *Paice/Husk* 44
Figura 12 - Exemplo de uma taxonomia simples 46
Figura 13 - Fragmento da hierarquia *is-a* da *WordNet* 48
Figura 14 - *WordNet* visual para o termo "*void*" 49
Figura 15 - Arquitetura geral do *Knowledge* TV 53
Figura 16 - Arquitetura conceitual do *Knowledge* TV 54
Figura 17 - DIKTV integrado a arquitetura KTV 56
Figura 18 - Arquitetura conceitual do DIKTV 59
Figura 19 - Diagrama de Atividades DIKTV 61
Figura 20 - Visão Geral dos Conceitos da *Movie Ontology* 70
Figura 21 - Indivíduos da superclasse *Genre* da *Movie Ontology* 71
Figura 22 - Diagrama *Sport Ontology* aplicada ao ciclismo olímpico 72
Figura 23 - Nuvem de dados interligados *Linked Data* 73
Figura 24 - Cenário de validação SQTV e DIKTV 74
Figura 25 - Arquitetura geral do DIKTV e SQTV 75
Figura 26 - Exemplo da métrica *Reciprocal Ranking* 78
Figura 27 - Gráfico de avaliação da Análise de Similaridade 79
Figura 28 - Exemplo de arquivo XML encontrado na base de dados 80
Figura 29 - Exemplo de consulta semântica por domínio 81

LISTA DE TABELAS

Tabela 1 - Tabelas XML definidas na especificação de metadados TV-*Anytime* 29
Tabela 2 - Exemplo da matriz de *Levenshtein* 40
Tabela 3 - Metadados utilizados na validação 68
Tabela 4 - Comparativo entre os resultados das consultas 82
Tabela 5 - Comparativo entre os resultados das expansões 83
Tabela 6 - Comparativo entre os trabalhos apresentados 90

LISTA DE SIGLAS

Sigla	Significado
ABNT	Associação Brasileira de Normas Técnicas
API	*Application Programming Interface*
CRID	*Content Reference Identifier*
CSS	*Cascading Style Sheets*
DAML	*Darpa Agent Markup Language*
DIKTV	*Domain Identifier Knowledge* TV
GPL	*General Public License*
HE-AAC	*High Eficiency - Advanced Audio Coding*
HTML	*HyperText Markup Language*
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
ISDB	*Integrated Services Digital Broadcasting*
JSON	*JavaScript Object Notation*
JVM	*Java Virtual Machine*
JWLN	*Java WordNet Library*
KDD	*Knowledge Discovery in Databases*
KTV	*Knowledge* TV
LAVID	Laboratório de Aplicações de Vídeo Digital
MO	*Movie Ontology*
MPEG	*Moving Pictures Expert Group*
MRR	*Mean Reciprocal Ranking*
NLP	*Natural Language Processing*
OWL	*Web Ontology Language*
PSI	*Program Specific Information*
RBC	Raciocínio Baseado em Casos
RDF	*Resource Description Language*
RI	Recuperação de Informação
RR	*Reciprocal Ranking*
SI	*Service Information*
SPARQL	*Query Language for RDF*
SBTVD	Sistema Brasileiro de TV Digital
SBTVD-T	Sistema Brasileiro de TV Digital Terrestre
SQTV	*Semantic Query* TV

STB	*Set Top Box*
TVA	TV-*Anytime*
TVD	TV Digital
TVDI	TV Digital Interativa
URI	*Uniform Resource Identifier*
W3C	*World Wide Web Consortium*
XML	*Extensible Markup Language*
XSD	*XML Schema Definition*

SUMÁRIO

RESUMO ix
ABSTRACT x
LISTA DE ILUSTRAÇÕES xi
LISTA DE TABELAS xii
LISTA DE SIGLAS xiii
SUMÁRIO xv
1 Introdução 17
1.1 Motivação 18
1.2 Objetivos 20
1.2.1 Objetivo Geral 20
1.2.2 Objetivos Específicos 20
1.3 Organização da Dissertação 21
2 Fundamentação Teórica 22
2.1 TV Digital Interativa 22
2.1.1 Sistema Brasileiro de TV Digital 23
2.1.1.1 *Middleware* Ginga 25
2.1.2 Padrões de Metadados para TV Digital 26
2.1.2.1 MPEG-2 PSI/SI 27
2.1.2.2 TV-*Anytime* 27
2.2 Web Semântica 30
2.2.1 Ontologias 32
2.2.1.1 Classificação de Ontologias 33
2.2.1.2 Linguagens para Representação de Ontologias 35
2.3 Análise de Similaridade 38
2.3.1 Análise de Similaridade Léxica 39
2.3.1.1 Algoritmo de *Levenshtein Distance* 39
2.3.1.2 Técnica de Radicalização 41
2.3.2 Análise de Similaridade Semântica 45
2.3.2.1 *WordNet* 48
2.4 Conclusão 50
3 O Componente DIKTV 52
3.1 *Knowledge TV* 52

3.2 Componentes Cliente e Servidor KTV 53
3.3 *Domain Identifier Knowledge* TV - DIKTV 55
3.3.1 Bases de Conhecimento 58
3.3.2 Arquitetura Conceitual DIKTV 59
3.3.3 Tecnologias Empregadas no DIKTV 62
3.4 Cenário Motivacional 64
3.5 Conclusão 66
4 Validação do DIKTV 67
4.1 Fontes de Dados 67
4.2 Ontologias de Domínio 68
4.3 Aplicações do DIKTV no módulo SQTV 72
4.4 Expansão de Consulta Semântica 75
4.5 Métricas Utilizadas na Avaliação 76
4.6 Experimentos e Avaliações dos Resultados 78
4.6.1 Aplicação da Análise de Similaridade 78
4.6.2 Consulta por Domínio 80
4.6.3 Expansão por termo genérico e específico 83
4.7 Conclusão 84
5 Trabalhos Relacionados 85
5.1 Correspondências entre esquemas 85
5.1.1 Madhavan et. al. 85
5.1.2 Noll et. al. 86
5.1.3 COMA 3.0 86
5.2 Cenários de Identificação de Domínios de Aplicação 87
5.2.1 Saccol 87
5.2.2 Moro et. al. 88
5.3 Análise Comparativa 89
5.4 Conclusão 90
6 Considerações Finais 92
6.1 Principais Contribuições 92
6.2 Trabalhos Futuros 93
Referências 95
ANEXO I 104
ANEXO II 108

1 Introdução

A Televisão Digital Interativa (TVDI) aliada a *Web* provê uma modelagem própria de interatividade entre pessoas, imagens e dados (Fantauzzi, 2009). Essa modelagem permitiu a inserção de conceitos como a *Web* Semântica (Berners-Lee et. al., 2001) e de mecanismos de Representação do Conhecimento.

A *Web* Semântica provê mecanismos de gerenciamento de conteúdo com a finalidade de agregar significado às informações. Contudo, às informações disponíveis necessitam ficar de um formato padronizado, visando uma posterior recuperação e compartilhamento.

Ontologias representam de forma explícita a semântica contida nos dados (Dziekaniak & Kirinus, 2004), especificando conceitos e as relações entre esses conceitos dentro de um domínio de aplicação.

As ontologias de domínio fornecem um universo de conceitos relacionados a um domínio genérico e dispõem de um vocabulário controlado que distingue as palavras pelo seu significado. Por exemplo, ao realizar uma consulta em algum site de compras com a palavra “cartão”, podemos encontrar significados distintos. A ontologia de domínio de dispositivos eletrônicos pode inferir o seu significado como um “cartão de memória”; e a ontologia de domínio de serviço postal poderá entender que se trata de um “cartão romântico”.

As heterogeneidades estruturais e semânticas encontradas nos metadados podem ser solucionadas através da adoção de ontologias para a integração dos dados. Através de análises de similaridade é possível investigar

correspondências entre metadados e bases de conhecimento (como ontologias, tesauros[1] etc).

A análise de similaridade léxica é responsável por efetuar a correspondência entre as sequências de caracteres, pontuando cada similaridade encontrada. Na análise de similaridade semântica são investigados os significados dos termos, usualmente empregando tesauros nessa correspondência.

Nesse trabalho propomos uma abordagem para identificação de domínios de aplicação, elegendo as ontologias-padrão para cada domínio. Para validar essa abordagem criamos um componente, chamado *Domain Identifier Knowledge TV* (DIKTV), pra restringir o espaço de busca, recuperar dados de forma mais rápida, prover representação semântica através das ontologias de domínio, expandir consultas semânticas e possibilitar a aquisição de novas ontologias-padrão e novas métricas de similaridade.

1.1 Motivação

A convergência digital entre as mídias TV e *Web* permitiu o aumento e a diversificação dos serviços e das aplicações interativas disponíveis aos usuários. No entanto, isso trouxe alguns problemas existentes no ambiente Web. Um desses problemas é a manipulação do elevado número de informações pelas aplicações e serviços, que necessitam utilizar mecanismos eficientes de recuperação e organização.

Outro problema se refere à integração dos dados, devido à natureza divergente dos ambientes (TV e *Web*), são necessários mecanismos que integrem o vocabulário heterogêneo dos dados *Web*, permitindo sua utilização pelos serviços e aplicações. A integração e o tratamento dos dados

[1] Tesauro é um conjunto de termos semântica e genericamente relacionados, cobrindo uma área específica do conhecimento, sendo ainda este um instrumento de indexação e recuperação de informações (Gomes, 2012).

potencializa a recuperação das informações, ou seja, soluciona o problema mencionado anteriormente.

Para tornar a recuperação das informações eficaz, são necessários mecanismos que permitam a identificação dos domínios de aplicação, possibilitando a restrição do espaço de busca. Isto diminuirá o tempo gasto na busca por dados pertencentes a domínios diferentes.

Esses mecanismos devem ser capazes de tratar os metadados advindos dos padrões de TVDI, como também devem lidar com a heterogeneidade dos dados advindos da *Web*, para proporcionar a integração dos mesmos.

O ambiente da TVDI fornece informações escassas ou insuficientes sobre os programas de TV, causando muitas vezes o desinteresse dos usuários que buscam informações prévias sobre os mesmos. Para o enriquecimento dessas informações, são necessárias aplicações de consultas semânticas, surgindo à necessidade de utilização da abordagem proposta neste trabalho.

Inserido no ambiente de convergência digital e do *middleware* Ginga, nossa abordagem é especificada e implementada através de um componente que provê raciocínio automático fazendo uso de domínios presentes na TV, como filmes, séries, eventos esportivos etc, para facilitar o acesso a informações sobre programas de TV.

O componente utiliza mecanismos da *Web* Semântica (Berners-Lee et. al., 2001) e as análises de similaridade, tanto na perspectiva léxica quanto na semântica, objetivando a correspondência entre as estruturas de metadados e o domínio a que pertencem. Para tanto, utilizamos ontologias de domínio, algoritmos de radicalização e um *tesauro.*

Em resumo, a identificação dos domínios de aplicação pode prover: (i) A restrição do espaço de busca, proporcionando maior eficácia e agilidade à mesma, (ii) Integração dos dados, consolidando os metadados de TV e *Web*, e a (iii) Representação semântica dos dados, através da adoção das ontologias de domínio.

1.2 Objetivos

Nas próximas seções, serão definidos o objetivo geral e os específicos para a realização deste trabalho.

1.2.1 Objetivo Geral

A convergência digital entre *Web* e a TV permitiu a inserção de conceitos e práticas de modelagem semântica no cenário da TV. Dentre alguns conceitos empregados nessa modelagem, encontram-se a Web Semântica e a representação do conhecimento através das ontologias.

O objetivo geral desse trabalho é propor uma abordagem para identificação de domínios de aplicação em ambientes de convergência digital, empregando ontologias de domínio e métodos de análise de similaridade léxica e semântica. Para tanto, foi especificada, implementada e testada à abordagem através de um componente, chamado *Domain Identifier Knowledge TV* (DIKTV), que está inserido na plataforma *Knowledge* TV (LINO et. al., 2011), provendo serviços como o raciocínio automático.

1.2.2 Objetivos Específicos

Como objetivos específicos, podemos destacar:

- Estudo e aplicação das ontologias de domínio em ambientes de convergência digital;
- Descrição das técnicas de análise de similaridade, em sua perspectiva léxica e semântica;
- Levantamento dos requisitos e mapeamento dos metadados do padrão SBTVD, no contexto da plataforma *Knowledge* TV;
- Investigar e eleger algoritmos e ferramentas para implantação dos requisitos identificados;
- Criação de um componente genérico para TV Digital Interativa que atenda aos requisitos estudados;

- Validação do componente através de sua implementação e aplicação a um estudo de caso.

1.3 Organização da Dissertação

Este trabalho é constituído de seis capítulos, incluindo esta Introdução. Os capítulos restantes são organizados da seguinte forma:

- No capítulo 2 é apresentada a fundamentação teórica do trabalho, através da descrição dos conceitos que fornecem a base para o desenvolvimento deste trabalho.
- No capítulo 3 apresentamos a abordagem utilizada para a identificação de domínios de aplicação e o componente desenvolvido para aplicá-la.
- No capítulo 4 descrevemos os mecanismos necessários para a validação da abordagem proposta através do componente DIKTV.
- O capítulo 5 apresenta alguns trabalhos relacionados e as principais semelhanças e divergências entre eles e a abordagem proposta nesta dissertação.
- No capítulo 6 descrevemos as considerações finais desse trabalho, incluindo as principais contribuições e os trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo são apresentados os fundamentos da TVDI, tratando aspectos como o Sistema Brasileiro de TV Digital, o middleware Ginga e os padrões de metadados de TV utilizados nesse trabalho. Em seguida abordamos os conceitos relacionados à *Web* Semântica, enfatizando o uso das ontologias e por fim, abordamos os conceitos sobre a análise de similaridade tanto na perspectiva léxica quanto na semântica.

2.1 TV Digital Interativa

Inaugurada oficialmente no Brasil em dezembro de 2007, a TV Digital cobre 46,8% da população do país segundo dados da Agência Nacional de Telecomunicações (ANATEL, 2013).

A interatividade surgiu como um diferencial importante na evolução da TV Digital, proporcionando inúmeras possibilidades de criação e disponibilização de conteúdo. Dessa forma a TV Digital Interativa tornou-se propulsora da convergência de diversas mídias, incluindo a convergência entre TV e *Web*, que permitiu a aquisição de novos recursos e serviços.

Uma vez definida a estrutura da TV Digital com a transmissão de áudio, vídeo e dados e agregando serviços de internet, surgem recursos de interatividade e os usuários tornam-se mais ativo diante da TV, participando da construção do conteúdo através de um canal de retorno (Kulesza et. al., 2006).

Para que essa transmissão de conteúdo seja realizada, a TV Digital requer a adoção de normas e padrões de metadados de TV, sendo estes necessários para que haja compatibilidade entre os sistemas. A escolha pelos modelos e padrões que serão empregados é decorrente de questões de cunho comercial e políticos (Nascimento, 2011).

No Brasil foi implantado o Sistema Brasileiro de TV Digital Terrestre (SBTVD-T), sendo este regulamentado pela Associação Brasileira de Normas Técnicas – ABNT (ABNT NBR 15606-1, 2008), que define entre outras especificações, a codificação, transmissão e o middleware utilizado.

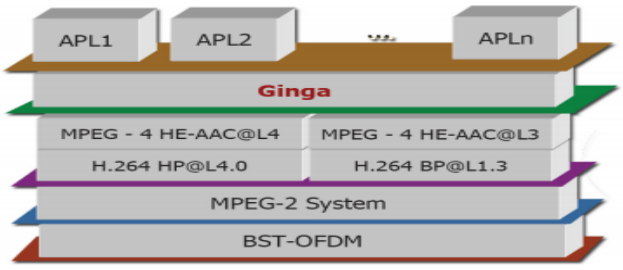
Neste cenário também é necessário escolher padrões de metadados para o conteúdo multimídia, sendo estes responsáveis pela indexação e organização do conteúdo transmitido por *broadcast* ou canal de retorno. Nos tópicos seguintes discutiremos o modelo adotado pelo SBTVD e os padrões de metadados para TVDI.

2.1.1 Sistema Brasileiro de TV Digital

O Sistema Brasileiro de TV Digital (SBTVD) adota o padrão Japonês (ISDB) para transmissão terrestre, devido as suas características técnicas de mobilidade em detrimento dos demais padrões. O padrão brasileiro já foi aceito por diversos países e o principal fator de adesão do mesmo é referente à interatividade provida.

O componente fundamental para o desenvolvimento de serviços interativos é o *middleware*, denominado Ginga, cuja especificação de referência foi definida pelo LAVID/UFPB (Soares & Lemos, 2007). Assim como os demais padrões, ele é capaz de processar tanto aplicações declarativas, pelo ambiente Ginga-NCL, quanto imperativas, pelo Ginga-J. Conforme pode ser visto na Figura 1, o SBTVD segue a tendência de outros *middlewares* e seu funcionamento pode ser observado em camadas.

Figura 1 - Camadas do Sistema Brasileiro de TV Digital



Fonte: (Souza & Soares, 2013)

Conforme Brackmann (2008), podemos detalhar as camadas do SBTVD (Sistema Brasileiro de TV Digital) como:

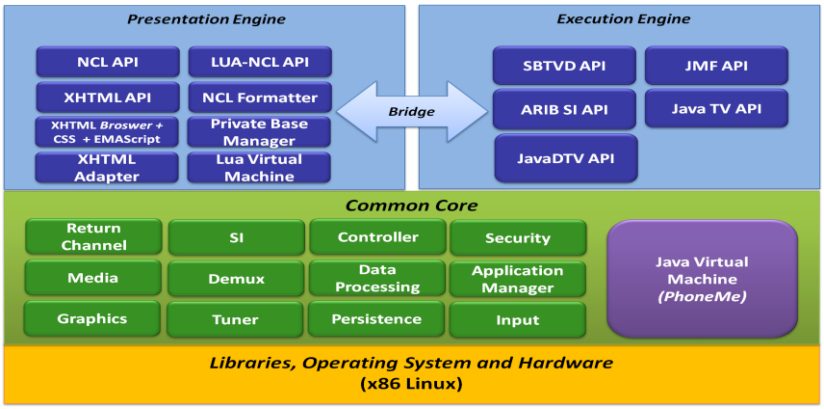
- BST-OFDM (*Band Segmented Transmission - Orthogonal Frequency Division Multiplexing*): Na modulação OFDM, divide-se o canal em diversas sub-portadoras e transmite essas sub-portadoras paralelamente. Isto permite que mesmo com interferências, uma pequena parte da informação transmitida seja perdida e por isso tecnologia OFDM é mais imune a interferências do ambiente.
- MPEG-2 *System* (*Moving Pictures Expert Group*): O padrão MPEG-2 foi adotado na camada de transporte, sendo responsável por conduzir o áudio e vídeo.
- H.264 - Camada de Vídeo: Para que seja possível a visualização da imagem em telas de alta resolução e também em aparelhos portáteis. Vale salientar, que o uso do *codec* H.264, ao invés do MPEG-2, é um dos principais diferenciais entre o padrão brasileiro e o padrão Japonês (ISDB).
- MPEG-4 HE-AAC (*High Eficiency - Advanced Audio Coding*) - Camada de Áudio: foram especificados nos padrões MPEG-2 e MPEG-4. Esta tecnologia leva em conta o modelo psicoacústico humano.
- Ginga - Camada de Middleware: desenvolvido pela PUC-RIO e UFPB. É responsável pela interatividade e interface da aplicação. Será visto com mais detalhes na próxima seção.
- Aplicações: onde se encontram as aplicações enviadas pelo canal de dados. As aplicações são executadas de acordo com as necessidades e interesse do telespectador.

2.1.1.1 *Middleware* Ginga

No contexto da TVDI, o *middleware* é uma camada de *software* intermediária que está presente no receptor e sua função principal é abstrair as particularidades do sistema para aplicações e usuários (Leite et. al., 2005). Seu objetivo é mascarar a heterogeneidade e complexidade do *hardware* e do *software*, facilitando o desenvolvimento dos aplicativos por parte dos programadores.

O SBTVD adotou o *middleware* desenvolvido no próprio país, denominado Ginga (ABNT NBR 15606-2, 2008). Esse *midlleware* surgiu da junção de dois *middlewares*, um desenvolvido pela UFPB, chamado *FlexTV* (Leite et. al., 2005) e o outro desenvolvido pela PUC-RJ, conhecido como Maestro (SOARES; 2006) é está subdividido em dois subsistemas, também chamados de Máquina de Execução (Ginga-J) e Máquina de Apresentação (Ginga-NCL), como podemos observar na Figura 2.

Figura 2 - Arquitetura conceitual do middleware Ginga



Fonte: (Araújo, 2011)

Estes subsistemas são interligados através de uma “ponte”, que permite a intercomunicação dos mesmos, ou seja, aplicações imperativas acessam serviços das aplicações declarativas e vice-versa. Dessa forma o desenvolvimento de aplicações fica mais amplo, pois existe a escolha pelo paradigma de programação que seja mais adequado aos requisitos. Na Figura

2 podemos observar em detalhes a arquitetura conceitual do middleware Ginga, estas podem ser descritas como:

- Ginga-NCL: realiza o processamento de documentos escritos da linguagem declarativa NCL e inclui interpretadores CSS (*Cascading Style Sheets*) e ECMAScript além da máquina de apresentação Lua responsável por interpretar scripts Lua (ABNT NBR 15606-2, 2008).

- Ginga-J: é responsável por processar aplicações Java para TV. Além de ser um ambiente de aplicação procedural, o Ginga-J ainda possui uma máquina de execução procedural que é composta pela Máquina Virtual Java – JVM (ABNT NBR 15606-2, 2008).

- Common Core: local onde são definidas as funcionalidades dos sistemas em TVDI. Algumas destas funcionalidades são: armazenamento, sintonização de canais e informações de serviço, estas funcionalidades dão suporte aos ambientes de apresentação e execução do Ginga.

2.1.2 Padrões de Metadados para TV Digital

O conteúdo multimídia pode ser indexado e estruturado por metadados, e estes podem seguir dois modelos: os rígidos e flexíveis. Os metadados rígidos não permitem a extensão dos descritores, ou seja, os descritores previamente definidos não interpretam novos descritores, diferentemente dos metadados flexíveis que possuem a função de criação de novos descritores a partir dos existentes.

Os principais sistemas de TVD, inclusive o SBTVD, adotam os metadados rígidos e utilizam as Tabelas de Informação de Serviço (SI) (ABNT NBR 15603-1, 2007). O pradrão MPEG-2 é um exemplo dessa adoção e será discutido na seção 2.1.2.1.

Alguns sistemas possuem serviços que necessitam de informações adicionais e bem detalhadas sobre o conteúdo, o que impede a utilização das tabelas SI, pelas limitações que estas apresentam. Para diminuir as limitações

apresentadas por essas tabelas, são utilizados metadados flexíveis. Já existem pesquisas e recomendações de um padrão flexível orientado a broadcast chamado de TV-*Anytime* (TVA, 2013), esse padrão também será discutido na seção 2.1.2.2.

2.1.2.1 MPEG-2 PSI/SI

O MPEG-2 é um padrão amplamente adotado pelos sistemas de TVD. Esse padrão utiliza descritores de baixo nível como as tabelas PSI/SI. Apesar da fácil manipulação, os descritores desse padrão são rígidos e não podem ser estendidos, limitando assim, a oferta de serviços mais avançados.

O conjunto de tabelas chamado PSI (*Program Specific Information*), é responsável pela disponibilização das informações necessárias para que o receptor desmultiplexe o fluxo de transporte e decodifique seus fluxos elementares. No entanto, outros dados são necessários para que o receptor identifique os serviços existentes no fluxo de transporte, sendo essa a função das tabelas SI (*Service Information*). Além dessa função, as tabelas SI fornecem informações sobre o conteúdo disponível em cada um dos serviços (Oliveira, 2010).

Os sistemas de TVDI adotam estruturas de dados como as tabelas PSI/SI para transmissão do conteúdo multimídia, contudo, além dessas tabelas, cada sistema de TVDI especifica o seu conjunto de tabelas. A tabela dos metadados do padrão MPEG-2 podem ser encontradas no Anexo I desta dissertação.

2.1.2.2 TV-*Anytime*

O TV-*Anytime* (TVA) foi criado em 1999 através de um fórum que contou com a participação de cem organizações e empresas. Esse padrão é aberto e provê um *framework* de descrição semântica dos programas e serviços que serão consumidos pelos usuários.

O principal diferencial do TV-*Anytime* em relação aos outros padrões é a separação entre a referência do conteúdo e as informações de acesso ao mesmo.

O TV-*Anytime* dispõe de um CRID (*Content Reference Identifier*), que contém os metadados de referência única do conteúdo independente da localização e disponibilidade. A estrutura de um CRID é formada por um campo destinado a (i) descrição de conteúdo, (ii) descrição de instância, (iii) metadados do usuário e (iv) segmentação. Cada uma desses campos pode ser descrito como (Giglio et. al., 2011):

i) Descrição de conteúdo: informações globais referentes ao conteúdo. Por exemplo, informações referentes ao título do programa, descrição, gênero, etc.

ii) Descrição de Instância: descrevem informações para dar suporte a mecanismos de busca de programas, parâmetros de entrega (*pay-per-view*) e regras de uso (programas ao vivo, horário de início e fim). Através desta categoria o usuário consegue recuperar a listagem de locais onde o conteúdo pode ser recuperado (internet, guia de programação eletrônico, etc.).

iii) Metadados do usuário (*Log entry*): descreve as informações relacionadas a tudo que é visualizado pelo usuário, possibilitando o monitoramento das ações do usuário enquanto este consome o conteúdo. Exemplos destas ações são as operações de *play*, *pause*, *fast-foward* e gravar.

iv) Segmentação: refere-se à capacidade de manipular intervalos temporais (pedaços de conteúdo multimídia), permitindo o uso de novas formas de consumo, navegação e reconfiguração do conteúdo.

O padrão TV-*Anytime* adota o XML (*Extensible Markup Language*) (W3C, 2012) como formato para representação dos metadados e o XML *Schema* para definição formal da estrutura de representação.

Os elementos XML normalmente são agrupados em tabelas, que são classificadas em três tipos distintos: *content description metatada* (metadados de descrição de conteúdo), *instance description metadata* (metadados de descrição de instância) e *consumer metadata* (metadados do consumidor) (Oliveira, 2010). A Tabela 1 ilustra as tabelas XML separada em seus respectivos tipos.

Tabela 1 - Tabelas XML definidas na especificação de metadados TV-*Anytime*

Content Description Metadata	**Instance Description Metadata**	**Consumer Metadata**
ProgramInformationTable	ServiceInformationTable	UserPreferences
GroupInformationTable	ProgramLocationTable	UsageHistory
CreditsInformationTable		
ProgramReviewTable		
SegmentInformationTable		

Fonte: (Oliveira, 2010)

As tabelas referentes aos metadados de descrição de conteúdo descrevem as informações que de determinado conteúdo, independente da sua localização. Exemplos desses metadados são: nome do programa, sinopse e classificação indicativa.

As tabelas referentes aos metadados de descrição de instâncias fornecem informações de um determinado conteúdo em uma localização especifica. Exemplos desses metadados são: nome do provedor de serviço e o horário de início do programa.

As tabelas referentes aos metadados do consumidor fornecem informações de acesso dos usuários, como por exemplo, histórico e suas preferências. A tabela dos metadados do padrão TV-*Anytime* podem ser encontradas no Anexo II desta dissertação.

2.2 Web Semântica

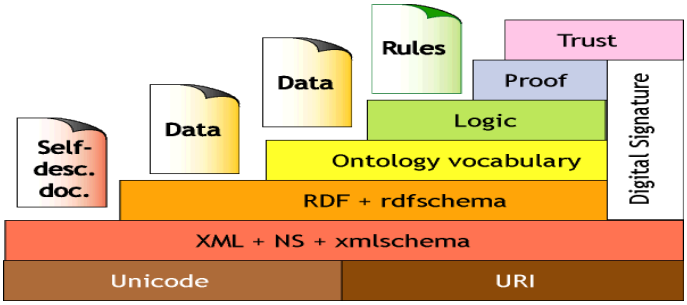
A *Web* Semântica surgiu em 2001, quando Berners-Lee et. al. (2001) propôs em seu artigo uma nova forma de manipular conteúdo na *Web*, através de uma estruturação semântica que consequentemente viabilizaria a tarefa de processamento das informações por parte dos agentes inteligentes.

A identificação do conteúdo é mais sofisticada, de maneira que a busca é realizada pelo significado de uma palavra e não somente pela forma sintática em que se apresenta. “A *Web* Semântica não é uma *Web* separada, mas uma extensão da atual. Nela a informação é dada com um significado bem definido, permitindo melhor interação entre os computadores e as pessoas” (Berners-Lee et. al., 2001).

Souza & Alvarenga (2004) definem que a essência da *Web* Semântica é a criação e implantação de padrões, linguagens e metadados que podem ser utilizados a partir de regras que viabilizam o armazenamento e compartilhamento das informações descritas pelos próprios usuários.

Em 2000, Berners-Lee desenvolveu um diagrama que exemplifica as camadas e protocolos que compõem a *Web* Semântica. Este diagrama é conhecido como "bolo de aniversário" e pode ser observado na Figura 3.

Figura 3 - Camadas lógicas da Web Semântica



Fonte: (Berners-Lee, 2000)

Conforme Berners-Lee (2000) as camadas podem ser descritas como:

- Unicode e URI: responsável por definir os caracteres e mecanismos de referencia de objetos a partir do endereço.

- XML+NS+*xmlschema*: essas linguagens de marcação, permitem a autodescrição dos documentos.
- RDF + rdfschema: fornece a criação e organização de objetos Web em hierarquias.
- *Ontology Vocabulary*: permite à construção de ontologias a partir da linguagem OWL.
- *Logic*: utiliza as ontologias para a escrita de aplicações específicas baseadas em conhecimento.
- *Proof*: envolve meios de representação e validação de provas em linguagem *Web.*
- *Trust*: utiliza os agentes de confiança ou agências de classificação e certificação para o provimento do conhecimento.

Existem “tecnologias que foram desenvolvidas para a *Web* Semântica, tais como o XML, linguagem de marcação que permite aos usuários criarem *tags* personalizadas sobre o documento, diferentemente do HTML, que possui estrutura de *tags* fixas, impedindo a criação de novos tipos de descritores” (Dziekaniak & Kirinus, 2004).

Com base na linguagem XML, foram criadas outras linguagens definidas para o contexto da *Web* Semântica, como o RDF e o OWL. A integração dessas e outras tecnologias facilita a criação de serviços *Web* interoperáveis.

Segundo Berners-Lee et. al. (2001), os computadores necessitam de acesso às informações estruturadas e às regras de inferência que auxiliam no processo de raciocínio automatizado.

As regras de inferência são especificadas pelas ontologias, possibilitando o acesso à semântica dos dados. Na próxima seção apresentamos conceitos e linguagens de representação de ontologias.

2.2.1 Ontologias

Analisando o contexto em que se encontram os sistemas de gestão do conhecimento é possível perceber a necessidade da utilização de conceitos que pertençam ao mesmo domínio, como forma de facilitar a recuperação, reuso e compartilhamento das informações.

As ontologias surgem como recurso para organizar e gerenciar o conhecimento em domínios de aplicações específicas, proporcionando dentre outras vantagens o acesso eficiente a grandes volumes de informação.

Encontra-se na literatura definições divergentes sobre ontologia, porém algumas são apenas complementares e reforçam conceitos. A definição de uso mais frequente é a descrita por Gruber (1993), que define a ontologia como sendo “uma especificação formal e explícita de uma conceituação compartilhada”.

Com objetivo de exemplificar os conceitos tratados nesta definição, alguns autores utilizam a descrição de Ding & Foo (2002) para obter um significado mais amplo dos termos da sua definição:

- Conceituação – “modelo abstrato de um fenômeno no mundo”;
- Explicita – “os tipos de conceitos usados e suas restrições devem estar explicitamente definidos”;
- Formal - “a ontologia deve ser processada por máquina”;
- Compartilhada –“a ontologia deve capturar o conhecimento aceito por consenso pelas comunidades que dela faz uso”.

Diversos pesquisadores têm retratado o uso de ontologias como mecanismo de troca de informações por meio de um vocabulário.

De acordo com Guarino (1998), “na sua utilização mais predominante em IA (Inteligência Artificial), uma ontologia refere-se a um artefato de engenharia, constituído por um vocabulário específico usado para descrever

certa realidade, além de um conjunto de suposições explícitas em relação ao significado pretendido das palavras do vocabulário".

O uso de ontologias vem sendo empregado em diversas áreas com sucesso e seu estudo ainda se encontra dirigido em muitos casos para a *Web*, onde se verifica um problema ainda muito comum: a recuperação imprecisa de informações.

A solução dessa problemática tem atraído o interesse de profissionais das áreas de Computação, que idealizam a melhoria do nível de precisão das informações recuperadas, proporcionando mais semântica ao conteúdo das páginas *Web* (Sales, 2006).

São inúmeras as vantagens obtidas através da utilização de ontologias, Hinz (2007) cita algumas como:

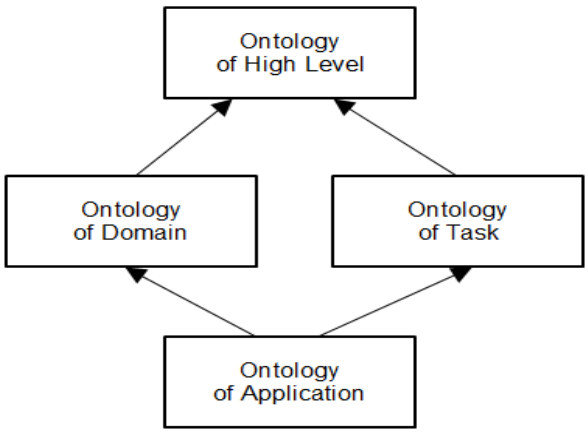
- Dispõem de um vocabulário para representação do conhecimento, regido por regras que evitam interpretações ambíguas do mesmo.
- Permitem o compartilhamento de informações, pois, caso exista uma ontologia para certo domínio, a mesma pode ser compartilhada por pessoas que desenvolvam aplicações dentro desse domínio.
- Fornece uma linguagem formal e explicita, facilitando a interpretação e o intercâmbio de informações.
- Possibilita a tradução da linguagem da ontologia sem alterar sua conceitualização, permitindo assim, expressá-la em várias línguas.
- Oferece a possibilidade de extensão de ontologia de domínio genérico para um domínio específico.

2.2.1.1 Classificação de Ontologias

Encontra-se na literatura descrições divergentes sobre os tipos de classificação das ontologias. Alguns autores como (Guarino, 1998; Maedche, 2002) utilizam a conceitualização como critério de categorização.

Apesar dos autores apresentarem descrições complementares as suas definições, a utilizada por (Maedche & Staab, 2002) nos parece ser a mais completa. A Figura 4 representa o grau de classificação das ontologias.

Figura 4 - Classificação de ontologias



Fonte: (Guarino, 1997)

Conforme Maedche & Staab (2002), as ontologias podem ser classificadas em:

- Ontologias de alto-nível (*Ontology of High Level*) – descrevem conceitos muito gerais que não estão condicionados a um domínio específico. Sendo assim, é bem razoável ter-se uma ontologia de alto-nível compartilhada por grandes comunidades de usuários.
- Ontologias de domínio (*Ontology of Domain*) – descrevem o vocabulário relacionado a um domínio genérico, estendendo os conceitos introduzidos nas ontologias de alto-nível. São exemplos de ontologia de domínio, ontologias de veículos, documentos, etc.
- Ontologias de tarefa (*Ontology of Task*) – descrevem um vocabulário relacionado a uma tarefa ou atividade genérica, permitindo o uso de conceitos pertencentes às ontologias de alto-nível. Um exemplo de ontologias de tarefa são as utilizadas em engenharia de software para modelagem de diagramas.

- Ontologias de aplicação (*Ontology of Application*) – são as ontologias mais específicas por serem utilizadas dentro das aplicações. Um exemplo é uma ontologia para uma aplicação que trabalhe com carros de luxo, essa ontologia especializará conceito da ontologia de veículos (que é uma ontologia de domínio).

As ontologias de domínio são a ênfase desse trabalho pelo fato de ser composto por um domínio genérico, agregando um universo de palavras menos restrito.

Diante de vários exemplos do uso de ontologias de domínio, os autores WU; TSAI, HSU, (2003), destacam a possibilidade de elaboração de um índice com significado nos documentos, auxiliando a compreensão textual, visto que os conceitos encontrados estão em conformidade com suas respectivas definições.

2.2.1.2 Linguagens para Representação de Ontologias

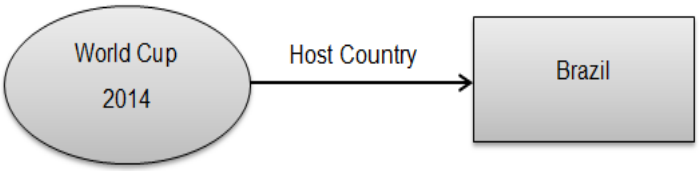
As ontologias dão suporte à estruturação semântica das informações, viabilizando o processamento destas informações pelas máquinas. Para construir e representar as ontologias, foram criadas diversas linguagens que de acordo com seu propósito possuem mecanismos formais e particulares para a interpretação das informações. Nessa seção detalhamos duas linguagens recomendadas pela W3C: (i) RDF (RDF, 2012) e (ii) OWL (OWL, 2012).

I. RDF

RDF (*Resource Description Framework*, ou modelo de descrição de recursos) “é uma linguagem baseada em XML para descrever a informação contida em um recurso” (Deitel et. al., 2003). Um recurso pode ser, por exemplo, uma página *Web*. O XML é utilizado para definir a estrutura, e o RDF expressa à semântica contida nos dados.

O modelo de dados RDF pode ser definido como: (i) recursos, que são os objetos que se quer representar, (ii) propriedades, descrevem as características dos objetos, (iii) declarações, que é uma tripla objeto-atributo-valor (ou sujeito-predicado-valor), consistindo de um recurso, uma propriedade e um valor, respectivamente. Os recursos que representam os objetos das sentenças devem utilizar identificadores no padrão URI, pois representam um endereço único para cada recurso na *Web*. As sentenças RDF podem ser representadas na forma de grafo, conforme é apresentada na Figura 5.

Figura 5 - Representação de uma declaração em RDF



Fonte: Próprio Autor (2013)

Na Figura 5, o recurso é "*World Cup* 2014", a propriedade é "*Host Country*" e o valor é "*Brazil*". Na forma de grafo as triplas são representadas como ligações nó-arco-nó, com a propriedade definindo o relacionamento entre recurso e o valor.

Em XML, as sentenças provem o intercambio entre as máquinas, facilitando a interpretação das aplicações e serviços. Na Figura 6 podemos observar um exemplo de sentença RDF.

Figura 6 - Exemplo de sentença RDF

```
<rdf:RDF>
      <rdf:Description rdf: about: "World Cup 2014">
            <prop: hostcountry>
                        Brazil
            </prop: hostcountry>
      </rdf:Description>
</rdf:RDF>
```

Fonte: Próprio Autor (2013)

Nesta representação, o documento RDF utiliza-se de um elemento XML com a *tag* rdf: RDF, o conteúdo desse elemento é um conjunto de descrições

que utilizam a etiqueta rdf: *Description*, onde cada descrição refere-se a um recurso, através deste é possível acessar a propriedade e encontrar o valor atribuído.

II. OWL

A OWL (*Web Ontology Language*) (OWL, 2012) representa uma família de linguagens com a finalidade de representar conhecimento. Essa linguagem é responsável por inserir significado ao conteúdo *Web* proporcionando o acesso eficaz do conteúdo pelos agentes inteligentes. A linguagem OWL é baseada em XML, podendo ser construída em cima da sintaxe do padrão RDF (*Resource Description Framework*) (RDF, 2012). O OWL é dividido em três sublinguagens que podem ser observadas na Figura 7.

Figura 7 - Relação entre os diferentes tipos de OWL

Fonte: (Hahn, 2011)

Na Figura 7 podemos observar a organização das sublinguagens de acordo com sua capacidade de representatividade e computabilidade. Uma descrição detalhada das sublinguagens segundo Hahn (2011) e OWL (2004):

- OWL *Lite* – suporta a criação de ontologias com uma hierarquia de classificação e características de restrição simples. Essa sublinguagem suporta restrições de cardinalidade e fornece um caminho de migração mais rápida para enciclopédias e outras taxonomias.
- OWL DL – representa um meio termo entre a sublinguagem OWL *Lite* e a OWL *Full*, pois permite uma maior expressividade e garante a

computabilidade. Além disso, inclui restrições como o tipo de separação (a classe não pode ser um indivíduo ou a propriedade, e a propriedade também não pode ser um indivíduo ou classe).

- OWL *Full* – provê uma maior expressividade na representação do conhecimento do que a OWL *Lite*, porém não garante que os resultados obtidos a partir da ontologia são sempre computáveis. A OWL *Full* ainda permite que a ontologia aumente o significado do vocabulário pré-definido (RDF ou OWL).

As linguagens menos expressivas (OWL *Lite* e DL) estão contidas dentro das mais expressivas (OWL DL e *Full*), de maneira que uma ontologia definida numa linguagem menos expressiva é aceita por uma linguagem mais expressiva. Toda ontologia OWL *Lite* válida é uma ontologia DL válida e toda OWL DL válida é uma ontologia OWL *Full* válida.

2.3 Análise de Similaridade

A correspondência ou *matching* (Madhavan et. al., 2001) consiste da manipulação de operações executadas a partir de dois esquemas de entrada com objetivo de retornar um mapeamento que identifique se existem elementos correspondentes nos dois esquemas.

Podemos citar o emprego das técnicas de similaridade em aplicações e-*business*, mapeamento entre bancos de dados heterogêneos, recuperação da informação (RI), Mineração de Texto, Raciocínio Baseado em Casos (RBC), dentre outros contextos.

O termo análise de similaridade vem em uma crescente constante de pesquisa, abrangendo inúmeras técnicas, algoritmos e variações de algoritmos específicos. Conforme Madhavan et. al. (2001), seu desenvolvimento tem ocorrido em torno de três passos principais:

- Normalização: ocorre quando termos semanticamente equivalentes apresentam nomes diferentes em esquemas distintos. Esse passo

sugere o uso de um tesauro que descreva termos de uma linguagem comum ou referentes a um domínio especifico.

- Categorização: esse passo consiste na separação dos termos em classes, para reduzir as comparações entre termos diferentes.
- Comparação: é definido um *score* de similaridade computado entre os termos em suas respectivas categorias.

A determinação de similaridade pode ocorrer sobre duas perspectivas: através da comparação de termos, chamada de análise de similaridade léxica e através da comparação entre os significados dos termos, chamada de análise de similaridade semântica.

2.3.1 Análise de Similaridade Léxica

Do ponto de vista matemático, a análise de similaridade léxica pode ser entendida como a medida da similaridade entre sequências de caracteres. Cada métrica empregada na análise de similaridade possui uma maneira de pontuar a correspondência apresentada nas sequências de caracteres. Esta pontuação é comumente representada por um valor entre 0 (nenhuma similaridade) e 1 (equivalência total) (Hahn, 2011).

Um dos principais algoritmos de comparação de sequências de caracteres utilizado no meio acadêmico é o algoritmo de *Levenshtein Distance*, na próxima seção apresentamos esse algoritmo.

2.3.1.1 Algoritmo de *Levenshtein Distance*

O algoritmo *Levenshtein Distance* ou *Edit Distance* (distância de edição) (McCallum, 2006) foi criado em 1965. Esse algoritmo é considerado um dos primeiros algoritmos de comparação de sequências de caracteres, sendo até hoje um dos mais utilizados. Sua execução é realizada através do cálculo de distância entre duas cadeias, essa distância é obtida através do numero

mínimo de operações de inserção, deleção e substituição necessárias para que se encontre a equivalência entre os termos em questão.

Por exemplo, se temos duas sequências como, "*Andrew*" e "*Amdrewz*", a primeira operação executada é a substituição de "m" por "n" e a segunda operação é a deleção do "z", o que corresponde a uma distância igual a 2 (dois). Portanto, ao menor custo de operações apresentado podemos determinar a semelhança entre os pares de termos. Na figura 8, podemos verificar a definição da função de *Levenshtein* utilizada nesse algoritmo.

Figura 8 - Definição da função de *Levenshtein*

$$M(i, j) = Max\{M(i-1, j) - 1, \quad //insertion$$
$$M(i-1, j-1) + p(i, j), \quad //equality\ or\ replacement$$
$$M(i, j-1) - 1\} \quad //removal$$

$$Onde\ p(i, j) = +2\ se\ Xi = Yj \quad //equality$$
$$-1\ se\ Xi \neq Yj \quad //replacement$$

Fonte: (Souza, 2012)

Segundo Souza (2012), M é a matriz de *Levenshtein* e a função p(i, j) é utilizada para determinar se houve igualdade entre os termos comparados ou não (substituição). X e Y são as *strings* que estão sendo comparadas, sendo "i" e "j" respectivamente as posições dos caracteres. Na Tabela 2 um exemplo de aplicação da função de *Levenshtein*.

Tabela 2 - Exemplo da matriz de *Levenshtein*

	$	T	E	C	H	N	O	L	O	G	Y
$	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
T	-1	2	1	0	-1	-2	-3	-4	-5	-6	-7
E	-2	1	4	3	2	1	0	-1	-2	-3	-4
C	-3	0	3	6	5	4	3	2	1	0	-1
H	-4	-1	2	5	8	7	6	5	4	3	2
N	-5	-2	1	4	7	10	9	8	7	6	5
I	-6	-3	0	3	6	9	9	8	7	6	5
C	-7	-4	-1	2	5	8	8	8	7	6	5
A	-8	-5	-2	1	4	7	7	7	7	6	5
L	-9	-6	-3	0	3	6	6	9	8	7	6

Fonte: (Souza, 2012)

Na Tabela 2 podemos verificar que a pontuação obtida na comparação das sequências de caracteres “TECHNOLOGY” e “TECHNICAL” foi 6 (seis). Quanto ao desempenho, o tempo gasto por este algoritmo é da ordem de O(|x| |y|), enquanto o espaço requerido é de apenas O(min{|x|, |y|}).

2.3.1.2 Técnica de Radicalização

Nesta seção são apresentados os três principais algoritmos de radicalização: *Porter*, *Lovins* e *Paice*/*Husk*.

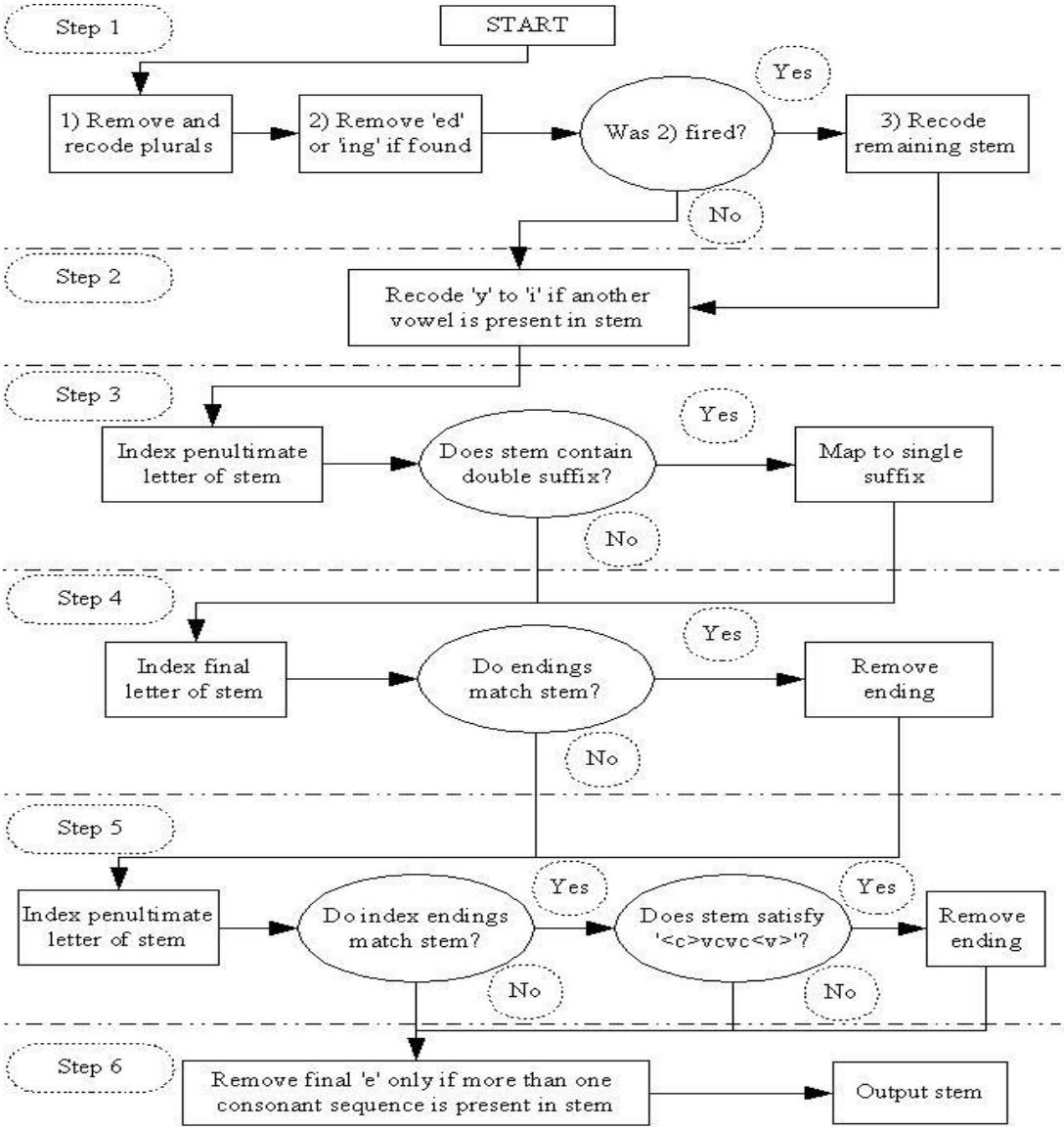
A radicalização (*stemming*), “visa reduzir variações de uma mesma raiz vocabular com a finalidade de recuperar palavras correlatas” (Viera & Virgil, 2007). Essa redução de palavras ao seu elemento raiz é comumente utilizada na correspondência entre palavras com formatos diferentes. Por exemplo, as palavras “estudar”, “estudante” e “estudando” podem ser reduzidas a palavra estudo, evitando ambiguidades e facilitando o emprego dos algoritmos de comparação de sequências de caracteres.

Existem diferentes tipos de algoritmos de radicalização de acordo com a língua em questão, visto que sua execução depende de um vocabulário (Viera & Virgil, 2007).

Porter

Segundo Stemming (2012), *Porter* é um algoritmo de remoção de sufixo sensível ao contexto que possui implementações em vários idiomas, inclusive em português. Além disso, é considerado o mais utilizado entre os algoritmos de radicalização. Na Figura 9 podemos observar os passos executados pelo algoritmo.

Figura 9 - Fluxograma do algoritmo de radicalização *Porter*



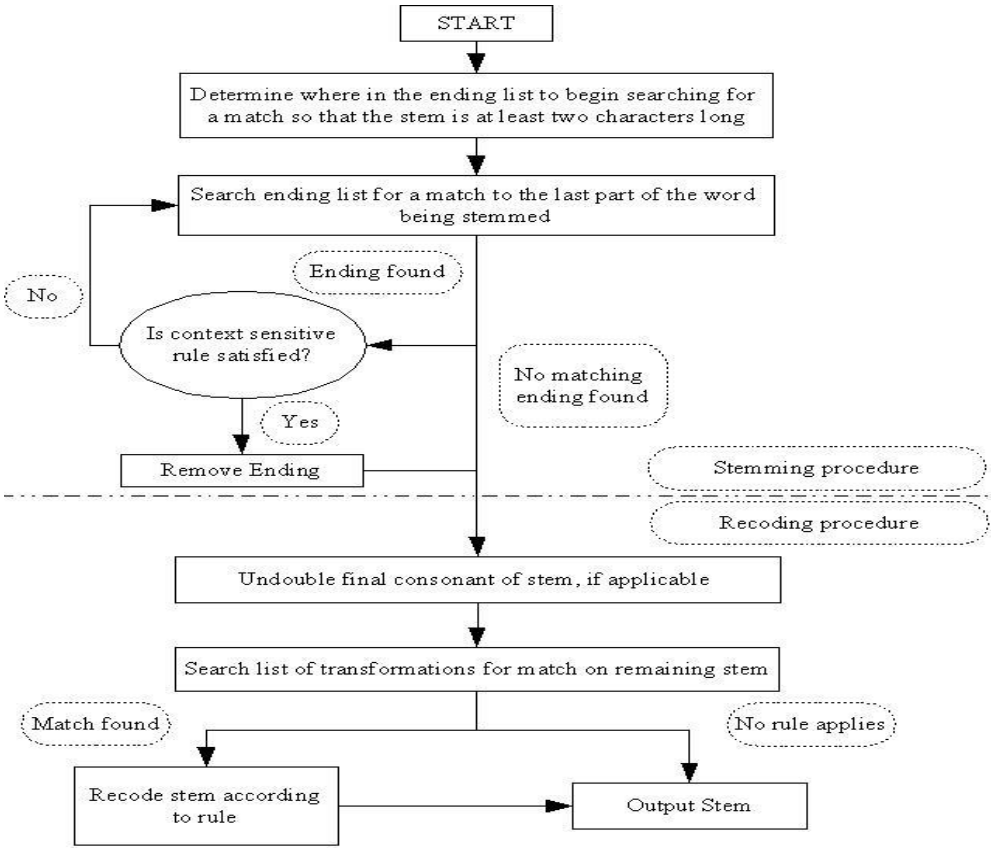
Fonte: (Stemming, 2012)

O algoritmo é dividido em cinco ou seis passos lineares, no primeiro passo o algoritmo é projetado para lidar com particípios passados e plurais, essa etapa é considerada a mais complexa e por essa razão é separada em três partes na definição original, 1A, 1B e 1C. As etapas restantes são relativamente simples e contém regras para lidar com classes diferentes, inicialmente ocorre à transformação de sufixos duplos para um único sufixo e a remoção de sufixos para que as condições necessárias sejam atendidas.

Lovins

O algoritmo *Lovins* remove as terminações dos radicais de acordo com um padrão, evitando a produção de radicais ambíguos. Existem exceções na ortografia devido às diferenças no padrão de linguagem britânico e americano que causam a duplicação de certas consoantes quando um sufixo é adicionado, essas exceções são chamadas de correspondência parcial e recodificação. Esse algoritmo é executado em duas fases, como ilustrado na Figura 10.

Figura 10 - Fluxograma do algoritmo de radicalização *Lovins*



Fonte: (Stemming, 2012)

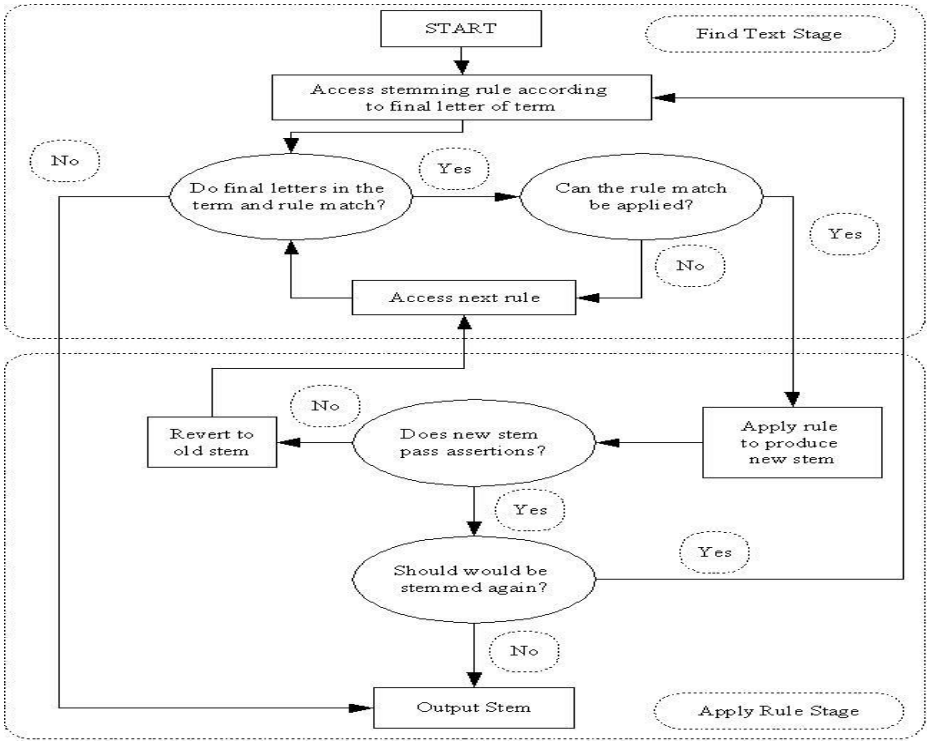
Como citado anteriormente, o algoritmo é composto por duas fases, à primeira inclui a remoção dos terminais e a análise de exceções associadas entre outras etapas e a segunda fase realiza a recodificação.

Paice/Husk

O algoritmo *Paice/Husk* foi desenvolvido por *Chris Paice* com o auxílio de *Husk Gareth* (Stemming, 2012). Esse algoritmo utiliza uma única tabela de regras onde podem ser encontradas operações de remoção ou substituição de sufixos, a substituição é realizada para evitar os eventuais problemas de ortografia que podem ser encontrados caso simplesmente removêssemos os sufixos das palavras.

A execução desse algoritmo é dividida em quatro passos principais, conforme ilustrado na Figura 11.

Figura 11 - Fluxograma do algoritmo de radicalização *Paice/Husk*



Fonte: (Stemming, 2012)

Conforme Figura 11, o algoritmo executa no primeiro passo a investigação do sufixo do termo de acordo com a tabela de regras. No passo 2, verifica-se à aplicabilidade da regra, ou seja, investiga-se se o sufixo do termo não corresponde a regra ou definições são violadas, senão, efetua-se o passo

4. O passo 3 é equivalente à aplicação da regra, onde remove-se o sufixo e verifica-se o símbolo de terminação. Após isso se encerra o processo ou volta para ao passo 1. Finalizando, o passo 4 corresponde à procura por outra regra, prosseguindo assim para a próxima regra da tabela e em seguida o processo é encerrado, senão volta para o passo 2.

Os algoritmos de radicalização podem ser comparados quanto as suas vantagens e desvantagens. Segundo Jivani (2011), o algoritmo Porter produz a melhor saída em comparação com os algoritmos *Lovins* e *Paice*/*Husk* e sua principal desvantagem é a lentidão, que é consequência dos cinco passos efetuados pelo mesmo.

O algoritmo *Lovins* possui uma lista de terminações maior que o algoritmo Porter, contudo da forma que é utilizado torna-se mais rápido, além disso, o *Lovins* pode lidar com a remoção de letras duplas e muitos plurais irregulares. A principal desvantagem do *Lovins* são os muitos sufixos não disponíveis na tabela de terminações. O *Paice*/*Husk* é bem simples e cada interação trata da exclusão e substituição de acordo com a regra aplicada. A desvantagem é que é um algoritmo lento.

2.3.2 Análise de Similaridade Semântica

Nesta seção discutimos as quatro abordagens que podem ser utilizadas na análise de similaridade semântica: Abordagem baseada em ontologias, no índice de informações compartilhadas, em características e híbrida.

A análise de similaridade semântica trata da correspondência de termos conceitualmente semelhantes, ou seja, com significados equivalentes.

A similaridade léxica dos termos aplicada de forma isolada não fornece possibilidades eficazes de correspondência, tornando indispensável o emprego de abordagens que tratem do aspecto semântico (Santos, 2010).

A informação, em geral, pode ser adquirida a partir de fontes (esquemas) com terminologias diferentes, o que faz necessário o uso de meios que forneçam a correspondência semântica entre essas fontes. Mas “em

praticamente todo o processo de integração de esquemas se faz necessário o uso de uma medida de similaridade semântica entre termos" (Silva, 2008).

A principal motivação de medir semelhança semântica vem das aplicações de processamento de linguagem natural (NLP), tais como desambiguação, sumarização e anotação de texto, extração e recuperação de informação, indexação automática e seleção lexical (Budanikst, 1999).

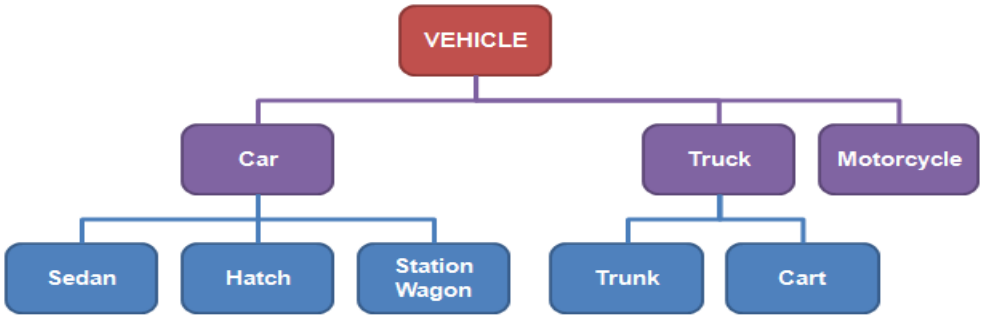
Conforme Wang (2005), Petrakis et. al. (2006) apud Silva (2008), as medidas de similaridade semântica são classificadas dentro de quatro categorias: (i) abordagem baseada em ontologias, (ii) abordagem baseada no índice de informações compartilhadas, (iii) abordagem baseada em características e (iv) abordagem híbrida.

I. Abordagem baseada em ontologias:

Nesta abordagem estão inclusas o uso de recursos e bases de conhecimento (como ontologias, dicionários e vocabulários) para melhorar o cálculo do grau de similaridade semântica entre os termos.

Essa abordagem é geralmente baseada em redes ou estruturas de grafos e usualmente utiliza-se de relacionamentos do tipo é um (*is-a*) para definir relações de subclasses e superclasses entre os conceitos presentes na hierarquia da ontologia. A Figura 12 ilustra o exemplo de uma taxonomia simples.

Figura 12 - Exemplo de uma taxonomia simples



Fonte: (Silva, 2008)

A taxonomia de veículos apresentada na Figura 12 descreve em seu primeiro nível a superclasse “*Vehicle”* e classifica os conceitos presentes na subclasse através dos relacionamentos “é um”, como no exemplo temos que o “*Car*” é um “*Vehicle”*, assim como, o “*Truck*” e a “*Motorcycle*”. Em seu último nível encontram-se os conceitos mais específicos da hierarquia, essa descendência continua utilizando os relacionamentos do tipo “é um” para definir as relações.

Na abordagem baseada em ontologias também estão inclusos outros tipos de cálculo do grau de similaridade usando o *WordNet* (2012) e outras redes semânticas disponíveis, o *Wordnet* é detalhado na seção 2.3.2.1.

II. Abordagem baseada no índice de informações compartilhadas:

Essa abordagem compreende todas as técnicas que basicamente utilizam o cálculo de similaridade semântica entre dois termos através do grau de informações que eles têm em comum, ou seja, o grau de informações que elas compartilham (Resnik, 1995 apud Silva, 2008).

Palavras que co-ocorrem bastante próximas de outra palavra específica são consideradas como sendo “características” ou “propriedades” desta palavra. Portanto, um conjunto de classes de palavras pode ser extraído através do mapeamento das classes da taxonomia para descobrir os níveis hierárquicos, classes e subclasses das quais o termo pertence.

III. Abordagem baseada em características:

Essa abordagem considera o conjunto de informações referentes à palavra desejada, ou seja, quanto mais características os termos têm comum, mas similares eles são. Este método estabelece que duas palavras sejam semanticamente relacionadas considerando a combinação de características em comum que elas possuem (ou vice-versa).

IV. Abordagem híbrida:

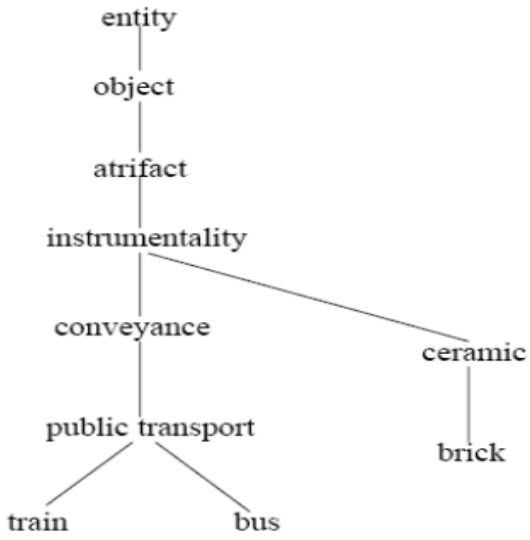
Essa abordagem se caracteriza pela combinação de algumas das abordagens descritas anteriormente. Dentre essas abordagens foram encontrados trabalhos que utilizam estruturas de ontologias e informações compartilhadas em suas métricas (Nguyen& Al-Mubaid, 2006).

Existem técnicas e algoritmos apropriados para o problema em questão, neste sentido, a escolha de um ou mais mecanismos deve ser feita através da investigação e validação dos mesmos no contexto pretendido.

2.3.2.1 *WordNet*

O *WordNet* é um sistema *on-line* de referência lexical desenvolvido na Universidade de Princeton, cujo objetivo é modelar o conhecimento lexical utilizando a língua inglesa. Além disso, ele pode ser visto como uma ontologia que contém cerca de 100.000 termos organizados em hierarquias taxonômicas (Hliaoutakis et. al., 2006). Nas Figuras 13 podem ser observados fragmentos da *WordNet*.

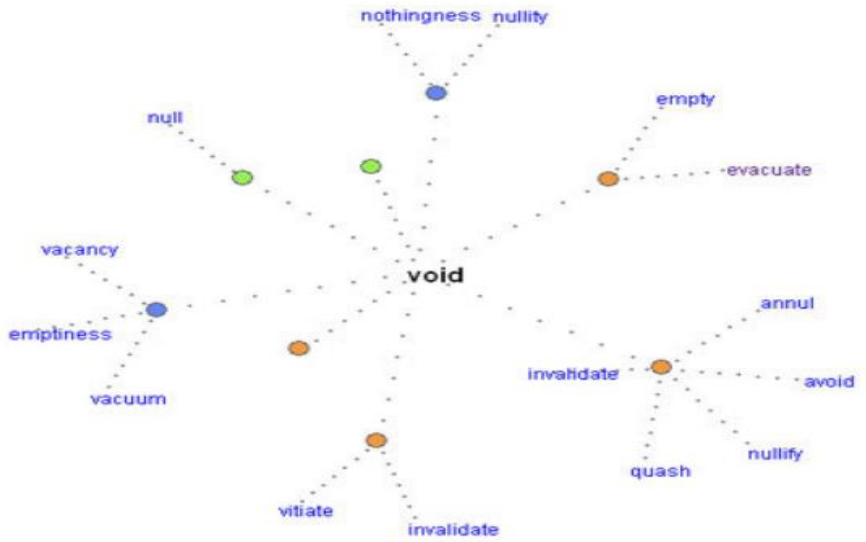
Figura 13 - Fragmento da hierarquia *is-a* da *WordNet*.



Fonte: (Hliaoutakis et. al., 2006)

Na figura 13 encontramos uma hierarquia de conceitos da *WordNet* que utiliza o relacionamento do tipo “é um” para definir relações de subclasses e superclasses. Nessa hierarquia do tipo “é um”, existem nove conceitos, entre substantivos e verbo (Hliaoutakis et. al., 2006). Um exemplo visual de busca semântica por um termo no *Wordnet* encontra-se na Figura 14.

Figura 14 - *WordNet* visual para o termo “*void*”



Fonte: (Ajaxian, 2012)

O nó que possui o menor caminho entre outro nó, é considerado mais similar a ele (Resnik, 1995 apud Silva, 2008). Portanto, o tamanho do caminho entre os termos (nós), determina o grau de similaridade semântica entre eles.

O *Wordnet* pode ser utilizado como um *tesauro* para o mapeamento dos relacionamentos terminológicos entre os esquemas, e fornece um “sistema organizado de palavras por conceito e relacionamentos semânticos” (Saccol, 2008).

A base do *WordNet* é uma rede de conceitos, onde cada conceito corresponde a um conjunto de palavras que são sinónimos (*synsets*) entre si. Algumas das relações semântica de conceitos existentes no *WordNet* (2012) são as seguintes (Miller et. al.,1990 apud Oliveira et. al., 2007):

I. Sinonímia: É a relação mais importante do *WordNet*, ocorre quando duas palavras possuem o mesmo conceito (significado) dentro de um contexto linguístico, a substituição de uma palavra pela outra não afeta seu valor. O *WordNet* organiza os conceitos em categorias (nomes, verbos, adjetivos e advérbios) para evitar que conceitos de diferentes categorias sejam considerados sinónimos.

II. Antonímia: Segundo os autores a antonímia é uma relação difícil de definir, pois apesar de um conjunto de sinônimos (*synset*) conterem uma palavra antónima de uma palavra de outro *synset*, pode haver outras palavras dentro dos dois *synset* que não fazem parte dessa relação.

III. Hiponímia/hiperonímia: Também conhecida pela relação é-um (*is-a*), o hipónimo de um conceito esta relacionado às características mais específicas deste conceito e herda essas características do conceito mais genérico (hiperónimo), possuindo ao menos uma característica que o diferencia dos demais hipónimos desse conceito.

IV. Meronímia/Holonímia: Também conhecida como parte-de (*part-of*). O merónimo de um conceito X é um conceito Y que faz parte do primeiro. Em outras palavras, Y é uma parte de X.

2.4 Conclusão

Neste capítulo, foram apresentados os principais temas ligados à proposta, conforme suas definições, mecanismos e aplicações. Inicialmente discutimos sobre os assuntos ligados à TVDI e como a interatividade impulsionou a convergência das mídias. Também retratamos a implantação do SBTVD e a especificação do middleware Ginga, responsável pela criação e disponibilização dos serviços e aplicações interativas.

Em seguida, discutimos acerca dos padrões de metadados utilizados no ambiente de TV, descrevendo seus modelos estruturais e como tem ocorrido a adoção dos mesmos em decorrência de suas limitações. O MPEG-2 PSI/SI e o TV-*AnyTime* foram os padrões discutidos, pois os mesmos são base desse trabalho, o MPEG-2 por ter sido o padrão adotado pelo SBTVD e o TVA por ser o padrão de referência utilizado pelo SBTVD.

Apresentamos conceitos relacionados à *Web* Semântica, suas camadas, linguagens e tecnologias associadas, enfatizando o uso das ontologias de domínio. O estudo acerca das ontologias incluiu seus principais aspectos como, vantagens, classificações e representação, especialmente no que trata das ontologias de domínio necessárias para a implementação desse trabalho.

No intuito de identificar as correspondências entre os elementos foram definidas abordagens e algoritmos de análise de similaridade léxica e semântica. Foram descritas a modelagem e os algoritmos que serão utilizados em casa fase.

Por fim, destacamos assuntos pertinentes ao desenvolvimento do trabalho, como metadados de TV Digital, ontologias e seus meios de inferência semântica e a forma escolhida para a correspondência entre os mesmos.

3 O Componente DIKTV

Neste capítulo descrevemos a abordagem utilizada para a identificação de domínios de aplicação. Primeiramente, contextualizamos a plataforma *Knowledge* TV e a aplicação do nosso componente nessa plataforma, em seguida, descrevemos os componentes *Knowledge* TV, Cliente e Servidor e apresentamos o componente proposto, denominado *Domain Identifier Knowledge* TV (DIKTV), discutindo em detalhes o componente e as tecnologias utilizadas em sua implementação.

3.1 *Knowledge TV*

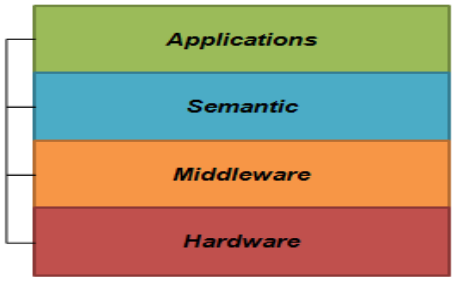
Ao longo do tempo, a televisão tem se destacado por ser um meio de transmissão de informações em massa, atingindo uma grande parte da população.

Pesquisas recentes mostram que 96% dos lares brasileiros possuem ao menos um aparelho de TV (IBGE, 2012). O grande consumo de informações nesse ambiente despertou o interesse pelo desenvolvimento de serviços e aplicações interativas.

Nesse contexto a plataforma *Knownledge* TV- KTV (LINO et. al., 2011) investiga uma arquitetura capaz de fornecer serviços e aplicações multimídia em TV Digital baseada em conceitos da *Web* Semântica. A *Web* Semântica fornece meios de interoperabilidade, comunicação entre os dados e os objetos do mundo real, desenvolvimento de serviços que melhoramos resultados de consultas, além de automatizar tarefas.

No ambiente de TVDI, o *Knowledge* TV está inserido, conceitualmente, entre as camadas do middleware e a camada de aplicações e serviços, conforme ilustrado na Figura 15.

Figura 15 - Arquitetura geral do *Knowledge* TV



Fonte: (Lino et. al., 2011)

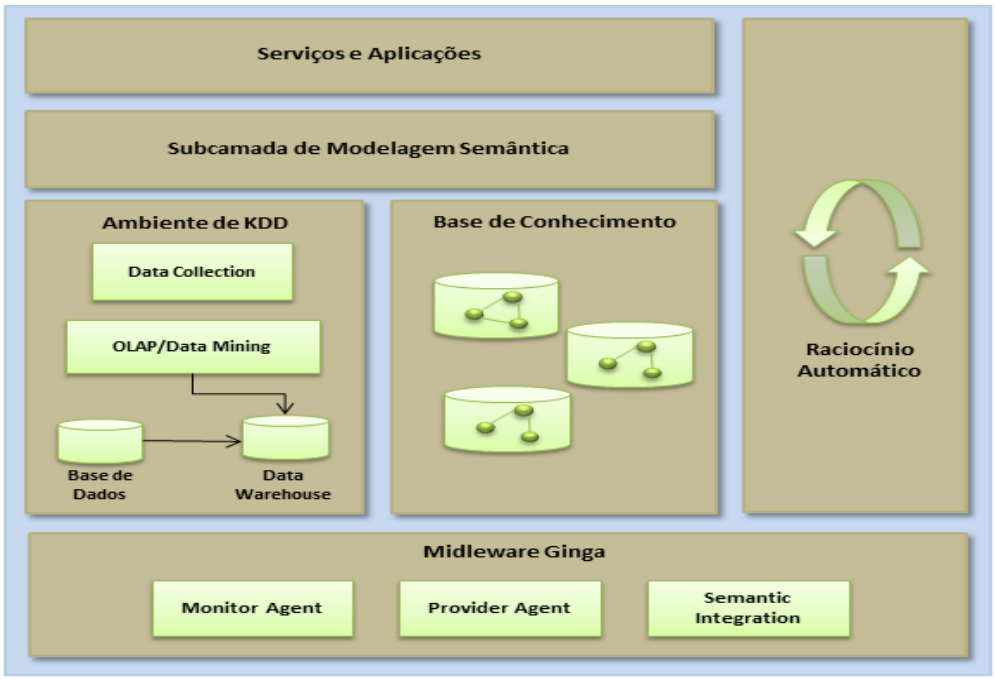
Nesta visão simplificada da arquitetura de um ambiente de TVDI são apresentadas as seguintes camadas: (i) Aplicações (*Applications*), utilizam serviços fornecidos pela camada semântica. (ii) Semântica (*Semantic*), provê serviços semânticos e modelagem de dados. (iii) *Middleware*, responsável por abstrair detalhes dos dispositivos de hardware, facilitando o intercambio das informações entre o *hardware* e as camadas superiores e o (iv) *Hardware*, composto por todos os componentes físicos que operam em um ambiente de TVDI.

Na próxima seção descrevemos os componentes Cliente e Servidor da plataforma KTV e suas funcionalidades específicas.

3.2 Componentes Cliente e Servidor KTV

A plataforma *Knowledge* TV está dividida em Componente Cliente, responsável manipular as informações advindas do STB do usuário e Componente Servidor, local onde essas informações são processadas pelos módulos e componentes do KTV. Na Figura 16, podemos observar a modelagem conceitual da plataforma KTV em detalhes.

Figura 16 - Arquitetura conceitual do *Knowledge* TV



Fonte: (Próprio autor, 2013)

De acordo com a Figura 16, O componente Cliente do KTV encontra-se no *middleware* Ginga e possui os agentes: (i) monitor, responsável por acompanhar o comportamento dos usuários do STB em relação ao conteúdo exibido e enviá-los ao agente provedor, (ii) provedor, encarregado de capturar as informações enviadas pelo agente monitor e obter os metadados contidos nas tabelas de informação de serviços (*Service Information* - SI), enviando essas informações ao *Semantic Integration*, cuja função é estruturar e gerenciar em um arquivo XML as informações obtidas pelos agentes monitor e provedor. Estas informações são enviadas para o módulo *Data Collection*, que se encontra no lado servidor do KTV, via canal de retorno.

O lado servidor do KTV esta dividido em 5 entidades, conforme Figura 17, onde cada entidade possui componentes e suas funções podem ser definidas como:

- Ambiente de KDD: Integra tecnologias em um ambiente de Descoberta do Conhecimento em Base de Dados.

- Bases de conhecimento: Refere-se ao armazenamento de dados, advindos da subcamada de modelagem semântica.
- Subcamada de modelagem semântica: Modela semanticamente o conhecimento por intermédio de linguagens formais que permitem o processamento automático por agentes computacionais.
- Raciocínio Automático: Responsável por todas as operações de raciocínio automático efetuado sobre a subcamada de modelagem semântica.
- Serviços e Aplicações: Fornece serviços e aplicações baseadas em representação de conhecimento, raciocínio automático e KDD.

3.3 *Domain Identifier Knowledge* TV - DIKTV

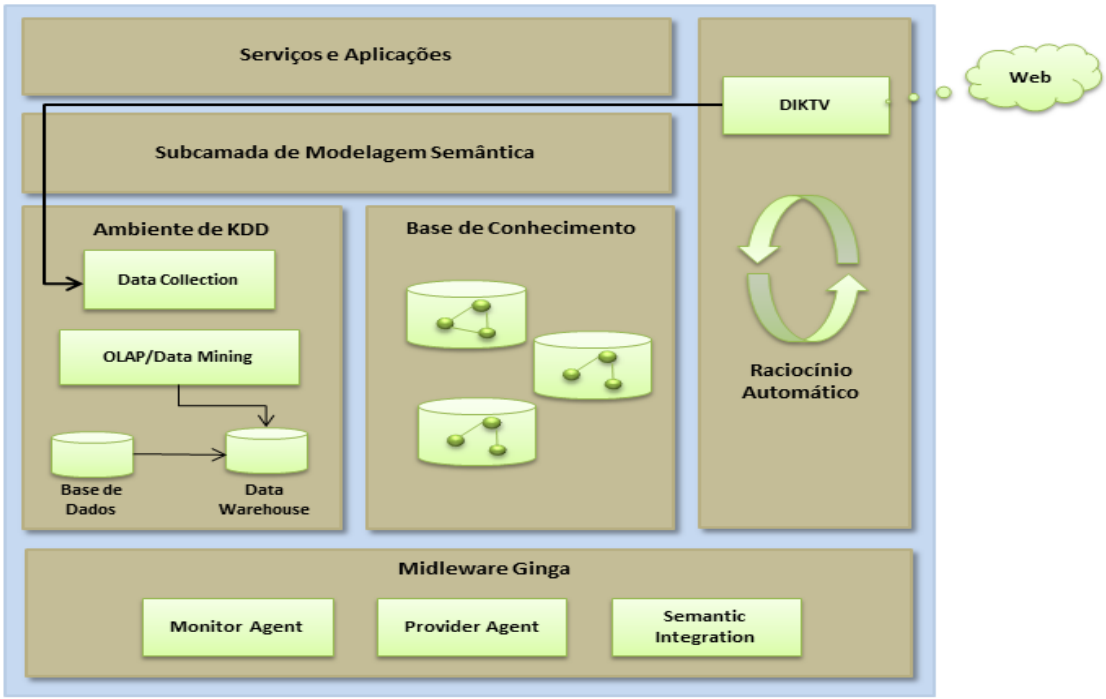
O advento da TVDI proporcionou o aumento no número de serviços e conteúdos multimídia disponíveis. Esse elevado número de informações tornou o ambiente de TV mais complexo exigindo a utilização de tecnologias que as gerencie mais eficientemente. Para isso, metadados vêm sendo utilizados para descrever e gerenciar conteúdos multimídia.

O compartilhamento de informações na TV Digital (TVD) é baseado em definições de metadados enviados por broadcast. No entanto, esses metadados possuem três fatores que desafiam a identificação do domínio dificultando sua utilização: (i) Pouca ou nenhuma semântica nos metadados, devido à sua manipulação XML, (ii) Divergências entre conceitos usados para descrever a mesma informação, isso referente à inexistência de um padrão na *Web* e (iii) conceitos semelhantes que tratam de domínios divergentes, nesse caso, isso se deve à falta de semântica nos dados. Estes fatores dificultam a utilização dos metadados pelos serviços e são considerados neste trabalho. Por isso, tratamos esses metadados, fornecendo a semântica necessária para identificação do domínio através do DIKTV, como forma de atender à demanda de novos serviços.

A plataforma *Knowledge* TV propõe uma camada semântica que provê serviços na plataforma da TV Digital. Alguns desses serviços são os de consulta semântica e de recomendação de conteúdo.

O DIKTV é responsável pela identificação de domínios de aplicação, a partir dos metadados de TV e *Web*. A captura de metadados via *Web* deve ser realizada por busca em sites especializados em divulgar a programação dos canais. Esta abordagem é necessária por não se ter sempre todos os metadados necessários por broadcast. Nosso trabalho instancia a arquitetura conceitual do *Knowledge* TV, conforme Figura 17.

Figura 17 - DIKTV integrado a arquitetura KTV



Fonte: (Próprio autor, 2013)

O componente DIKTV encontra-se na entidade chamada de Raciocínio Automático, pois permiti o raciocínio automático utilizando ontologias. O raciocinador utilizado no mecanismo é da própria API *Jena*, descrita em detalhes na seção 4.3.

O DIKTV acessa o módulo Coleta de Dados (*Data Collection*), que é responsável pela captura dos metadados e verifica a necessidade de enriquecimento ou não dos metadados através de uma busca que é realizada no arquivo XML que identifica se todos os campos foram preenchidos. Se essa necessidade existir, ocorrerá a busca *Web* pelos metadados essenciais para o funcionamento do DIKTV.

Em seguida, através de uma interface de acesso, o componente disponibiliza suas funcionalidades, para que possam ser utilizadas por serviços e aplicações, tais como: obter a identificação do domínio e a(s) ontologia(s) de domínio candidata(s), verificar a ontologia de domínio candidata, escolher qual algoritmo de radicalização utilizar no processo de identificação e expandir a consulta semântica.

O DIKTV emprega tecnologias baseadas na Web Semântica, chamadas de ontologias, para representar o domínio dos metadados de TV e utiliza mecanismos de análise de similaridade que possui métodos heurísticos para lidar com as divergências estruturais e terminológicas encontradas nas fontes de dados.

Há três tipos de programas de TV a serem identificados pelo DIKTV: filmes, esportes e outros. Se o candidato não pertence a nenhum dos primeiros dois tipos então é classificado como outro. Os tipos identificados são:

- Filme: Os filmes contêm uma lista de gêneros equivalentes e são detectados através da duração do programa. Se forem detectados gêneros de filme e se a duração for superior a 80 minutos então o programa será classificado como "filme".
- Esporte: a duração não é tão fixa como em "filme", por esse motivo a pesquisa deve ser realizada em toda gama de duração.

A seguir discutimos aspectos do DIKTV como bases de conhecimento (seção 3.3.1), arquitetura conceitual (seção 3.3.2) e tecnologias empregadas (seção 3.3.3).

3.3.1 Bases de Conhecimento

A implantação do componente DIKTV requer o uso de bases de conhecimento, que tornaram possível a análise de similaridade entre termos ligados a um domínio específico e os metadados obtidos da TV. Nesta dissertação, são utilizadas as seguintes bases de conhecimento:

I. Dicionário de gêneros TV-*Anytime*: O fórum TV- *Antime* (TVA, 2012) especifica as normas de utilização dos metadados por parte das emissoras, disponibilizando em um apêndice da norma um dicionário de gêneros de TV. Esse dicionário é utilizado, assim como as outras bases de conhecimento, na análise de similaridade léxica e semântica para identificar correspondências entre os metadados e os termos de domínio encontrados nas mesmas.

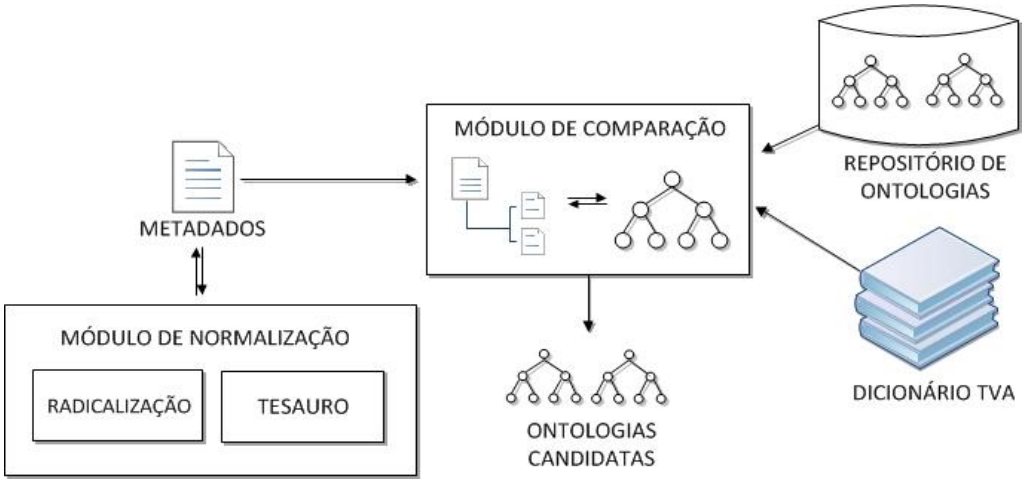
II. Ontologias de Domínio: Neste estudo, assumimos as ontologias de domínio como fonte de informação, essas ontologias são utilizadas como uma forma de vocabulário controlado, devido a sua expressividade e representação precisa de um determinado domínio. Além de utilizar as instancias dessas ontologias para identificação, ainda as utilizamos para expandir as consultas através de suas superclasses e subclasses. Utilizamos a ontologias de filme *Movie Ontology* e a de esporte *Sport Ontology*.

III. *WordNet*: O *WordNet* é o *tesauro* utilizado neste trabalho, onde através das suas relações semânticas de sinonímia, antonímia, hiponímia, hiperonímia, meronímia e holonímia, é possível obter a semântica dos termos e utilizar seu significado para encontrar termos correspondentes.

3.3.2 Arquitetura Conceitual DIKTV

Nesta dissertação utilizamos as bases de conhecimento apresentadas anteriormente em uma abordagem que realiza o processamento das mesmas através do uso das técnicas e algoritmos de análise de similaridade, tanto na perspectiva léxica quanto na semântica. Na Figura 18, apresentamos a arquitetura conceitual do DIKTV.

Figura 18 - Arquitetura conceitual do DIKTV



Fonte: (Próprio Autor, 2013)

Cada módulo e artefato possuem internamente as seguintes especificações:

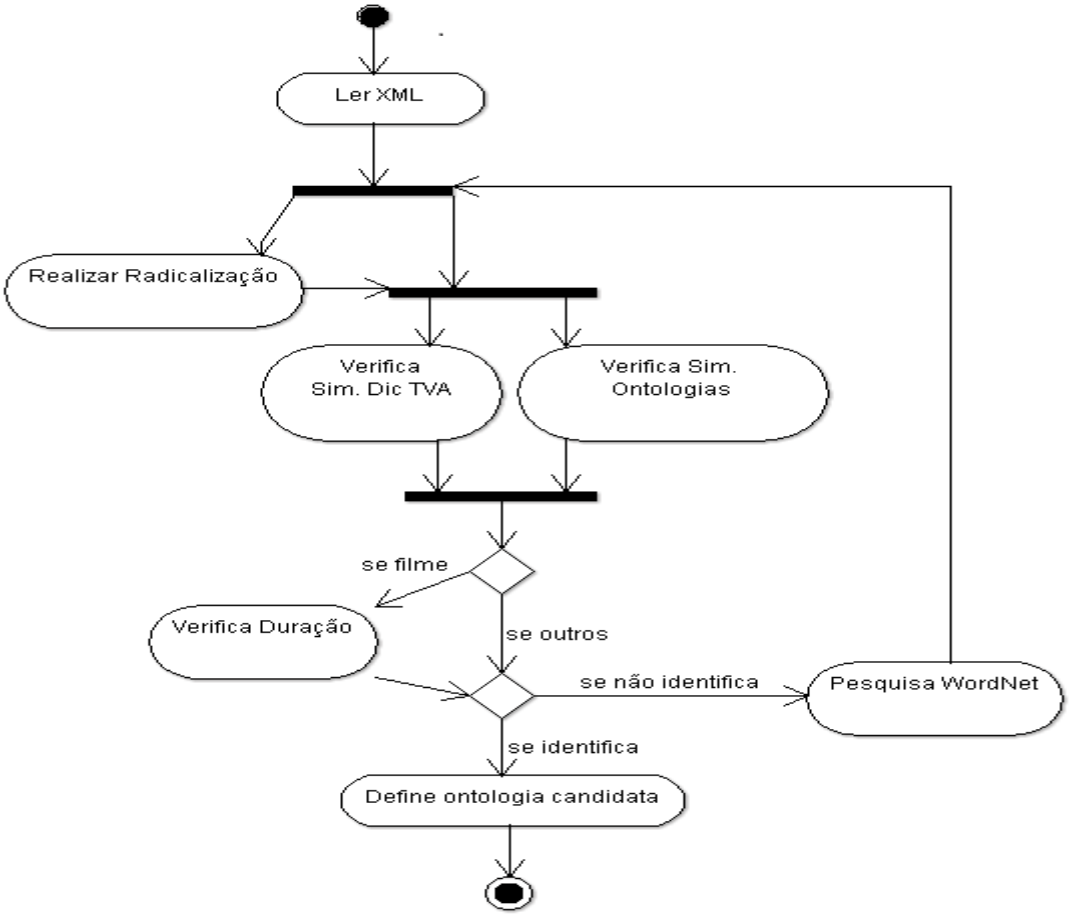
- Módulo de Normalização: Esse módulo é responsável pela normalização e categorização dos metadados através do uso de mecanismos como a radicalização e um tesauro. Na radicalização os metadados são utilizados como entrada para os algoritmos, que realizam seu processamento e retornam o radical relativo ao metadados de entrada. Os metadados também podem ser utilizados pelo tesauro *WordNet* para expansão dos termos. Os metadados de saída deste módulo são então comparados aos termos encontrados no repositório de ontologias e dicionário TVA. Por exemplo, utilizando o metadado de genêro

"*sporting*", utilizamos a radicalização para transformar essa palavra no radical "*sport*" que será comparado no módulo de comparação, se não existe similaridade entre essa palavra e os termos do repositório e do dicionário, a palavra "*sporting*" pode ser expandida através dos relacionamentos semânticos do tesauro.

- Módulo de Comparação: Esse módulo é responsável por comparar os metadados aos termos encontrados no repositório de ontologias e no dicionário TVA. Essa comparação é realizada pelo algoritmo de similaridade *Levenshtein Distance*, que dispõe de um *score* que possibilitará a escolha de uma ontologia de domínio candidata que é enviada a aplicação ou serviço que requisitou esta funcionalidade. Nesse módulo, os metadados podem ser comparados diretamente, ou seja, da forma em que foram recuperados no módulo *Data Collection*, e se não foi encontrada nenhum domínio similar, passam pelo processo de normalização e categorização do módulo de Normalização.
- Repositório de Ontologias: Contém uma ontologia-padrão para cada domínio, inicialmente em nosso estudo de caso, consideramos as ontologias no domínio de filme e esporte. Também definimos uma estrutura genérica para aquisição de novas ontologias de domínio.
- Dicionário TVA (TV-*Anytime*): Contêm as informações acerca dos gêneros de TV que foi disponibilizado em um apêndice da norma que rege o padrão TV-*Anytime*, consideramos as informações correspondentes a filme e esporte.

O diagrama de atividades (Figura 19) ilustra em detalhes como é realizada a identificação dos domínios de aplicação.

Figura 19 - Diagrama de Atividades DIKTV



Fonte: (Próprio Autor, 2013)

O diagrama de atividades DIKTV ilustra o processamento da abordagem proposta, na finalidade de identificar os domínios, inicialmente o arquivo XML é carregado, podendo então ser utilizado pelos mecanismos do DIKTV. A partir disso, os termos encontrados no arquivo XML são comparados utilizando o algoritmo léxico para identificar termos correspondentes no dicionário TVA e nas ontologias de domínio.

Se encontrado algum termo similar ao domínio de Filme, é importante verificar a duração, distinguindo assim, por exemplo, das series que contem gêneros semelhantes. Se os termos não pertencem ao domínio de filme ou esporte, é classificado como outro. Também na análise léxica pode ocorrer pela utilização dos algoritmos de Radicalização, onde a escolha pelo algoritmo

a ser utilizado pode ser feita pelo usuário. Se o domínio ainda não pode ser identificado utiliza-se o tesauro *WordNet* e suas relações semânticas, onde os termos relacionados são expandido e utilizados na verificação de similaridade e na radicalização com objetivo de fornecer a identificação do domínio.

3.3.3 Tecnologias Empregadas no DIKTV

O componente DIKTV foi desenvolvido na linguagem Java e sua implementação requer o uso de algoritmos e bibliotecas compatíveis com essa linguagem, nesta seção definimos as principais tecnologias utilizadas no desenvolvimento do componente.

Os principais algoritmos utilizados são: (i) *Levenshtein Distance*, (ii) *Porter*, *Lovins* e *Paice*/*Husk*.

I. *Levenshtein Distance*: Esse algoritmo foi escolhido para a comparação entre sequências de caracteres, ou seja, para medir a similaridade através de um coeficiente.

Esse algoritmo permite a execução do maior número de operações sendo o modelo mais geral onde as outras funções de distância podem ser obtidas com poucas alterações. Devido a isto, esta é a função de distância mais estudada na literatura (Fonseca, 2003). Além disso, o algoritmo pode ser utilizado em qualquer língua, sem alterações e informa um número determinando quão semelhantes são duas palavras.

II. *Porter*, *Lovins* e *Paice/Husk*: Esses são os principais algoritmos de Radicalização e a ferramenta *Setemmer Evaluation* (Stemming, 2012) desenvolvida na linguagem *Java* implementa os três. De acordo com as necessidades do componente DIKTV os algoritmos da ferramenta foram adaptados, podendo ser escolhidos conforme o resultado satisfaça as necessidades.

A implementação do componente também requer o uso de bibliotecas que são um conjunto de funções desenvolvidas para resolver determinados

problemas. Através das APIs, ou seja, das descrições dessas funções, é possível utilizá-las em outras aplicações. As bibliotecas utilizadas são: (i) *SimMetrics* (SimMetrics, 2013) (ii) *Jena* (Jena, 2013), (iii) RiTaWN (RiTa, 2013), (iv) *XStream* (XStream, 2013) e (v) *Joda-Time* (Joda, 2013).

I. *SimMetrics*: É uma biblioteca *Java* de licença livre GPL (*General Public License*) que implementa algoritmos de similaridades entre *strings* (Crocco, 2010). Utiliza métricas e resultados normalizados, que podem ser em ponto flutuante entre 0 e 1, onde 0 é a desigualdade e 1 a igualdade, para facilitar comparações entre os algoritmos.

II. *Apache Jena*: é um *framework Java* que fornece um conjunto de ferramentas e bibliotecas para construção de aplicações da *Web* Semântica (Jena, 2013), através desse *framework* manipulam-se formatos RDF, RDFa, OWL e SPARQL de acordo com as recomendações do W3C.

Além disso, o *Jena* inclui funções como: (i) uma API para leitura, processamento e gravação de dados RDF em XML, N-triplas e formatos *Turtle*, (ii) uma API para manipulação de ontologias em formato OWL e RDF, (iii) um motor de inferência baseado em regras para o raciocínio com fontes de dados RDF e OWL e (iv) um mecanismo de consulta compatível com a mais recente especificação SPARQL (Jena, 2013).

As aplicações normalmente acessam o mecanismo de inferência através do *ModelFactory* para associar um conjunto de dados com algum raciocinador que é responsável pelo processamento das regras de inferência. Dessa forma é possível acessar recursos e declarações presentes no modelo original ou utilizando regras de inferência (DILLI et. al., 2009).

O Jena possui raciocinadores pré-definidos, como: (i) *Transitive reasoner*, (ii) RDF *rule reasoner*, (iii) OWL, OWL Mini, OWL *Micro Reasoners*, (iv) DAML *micro reasoner* e (v) *Generic rule reasoner.* Em nosso trabalho

utilizamos especificamente o raciocinador de regras genérico que provê suporte as sublinguagens OWL *Lite*, DL e *Full*.

III. RiTaWN: Fornece um acesso simples ao *tesauro WordNet*, incluindo a obtenção de sinônimos, antônimos, hiperónimos, hipônimos, holonímos, merônimos, termos correlacionados, similares, normalizações, grupos de verbo e termos derivados (RiTa, 2013). Esta biblioteca é coberta pela licença GPL e é baseado em código de JWNL[2] e *Jawbone*[3].

IV. XStream: é uma biblioteca para serializar objetos para XML e vice-versa (XStream, 2013). Velocidade e baixo consumo de memória são características dessa biblioteca, tornando-o adequado para grandes gráficos de objetos ou sistemas com alta taxa de transferência de mensagens.

A partir das anotações disponíveis nessa biblioteca é possível utilizar as *tags* do arquivo XML para obter o conteúdo e mapeá-lo diretamente para um atributo correspondente na classe Java.

V. *Joda-Time*: é uma biblioteca que prove a manipulação de classes de data e tempo em *Java*. O projeto permite a múltiplos sistemas de calendário, enquanto continua a fornecer uma API simples. O calendário "padrão" é a norma ISO8601 que é usado pelo XML (Joda, 2013).

3.4 Cenário Motivacional

O SBTVD adota o padrão de metadados MPEG-2, esses metadados são rígidos e muitas vezes não possuem a semântica necessária para a utilização em aplicações e serviços na TVDI, fornecendo informações insuficientes

[2] http://sourceforge.net/projects/jwordnet

[3] http://mfwallace.googlepages.com/jawbone.html

acerca dos programas de TV. Essas informações são valiosas e podem atrair o interesse ou não do telespectador por determinada programação.

Com objetivo fornecer uma solução para este problema a plataforma Knowledge TV propõe uma camada semântica que provê serviços na plataforma da TV Digital. Inserido no ambiente de convergência digital (TV e Web), o Knowledge TV inclui mecanismos como a Web Semântica e Ontologias em sua plataforma, e através destes fornecem informações semanticamente enriquecidas. No entanto, alguns dos serviços desenvolvidos nessa plataforma precisam lidar com a grande quantidade de conteúdo semântico disponível por estes mecanismos, surgindo assim, a necessidade de investigar uma solução que possibilitasse identificar o domínio associado aos metadados de TV.

Para solucionar este problema fornecemos uma abordagem para identificação dos domínios de aplicação, baseada em conceitos da Web Semântica e nas análises de similaridade léxica e semântica, com objetivo de prover raciocínio automático acerca de domínios presentes na TV, como filmes, séries, eventos esportivos, etc. Essa abordagem visa principalmente à restrição no espaço de consulta e a agilidade na recuperação do conteúdo a ser disponibilizado aos telespectadores.

Aplicado na plataforma Knowledge TV essa abordagem pode auxiliar no processamento de serviços e aplicações, como por exemplo, na consulta semântica (SQTV). O serviço de consulta semântica enriquece as informações acerca do conteúdo transmitido, fornecendo informações adicionais ao usuário. Porém, como na Web existe uma grande quantidade de informações, é necessário um mecanismo que restrinja o espaço de busca, retorne os resultados das consultas de forma rápida e diminua a possibilidade de retorno de informações irrelevantes. Com isso os telespectadores podem obter mais informações e conteúdo relevante.

A abordagem de identificação de domínios de aplicação será validada através de um componente no capítulo 4.

3.5 Conclusão

Neste capítulo foi apresentada uma visão geral da abordagem proposta por esse trabalho de mestrado, onde inicialmente foi contextualizada a plataforma KTV, apresentando suas principais características e arquitetura conceitual.

Em seguida foram detalhados os componentes Cliente e Servidor do KTV, explanando suas funções e como ocorre a comunicação dos mesmos dentro do middleware. É importante que ocorra o entendimento acerca da plataforma KTV e de seus componentes para a compreensão do componente implementado nesse trabalho.

O componente DIKTV instancia a arquitetura conceitual do KTV e nesse ambiente é implementada e testada a abordagem proposta. Inicialmente descrevemos requisitos básicos do componente como o uso de bases de conhecimento, que tornam possível a análise de similaridade entre termos ligados a um domínio específico e os metadados obtidos da TV. Também foi realizado o detalhamento da arquitetura proposta, apresentando todos os seus componentes e modos de interação, incluindo um diagrama que viabiliza o entendimento da abordagem proposta.

Por fim, foi descrito em detalhes os principais algoritmos e bibliotecas utilizados na implantação da abordagem e onde os mesmos podem ser encontrados. No próximo capítulo são discutidos os experimentos que validaram a abordagem proposta e discutidos os resultados alcançados.

4 Validação do DIKTV

Esse capítulo descreve a validação da abordagem proposta e os requisitos para aplicação da mesma, tais como, fontes de dados utilizadas, ontologias de domínio escolhidas, a integração do DIKTV ao módulo de consulta semântica (SQTV), a expansão da consulta semântica, as métricas utilizadas na validação e a discussão dos resultados obtidos.

4.1 Fontes de Dados

Conceitualmente a fonte de dados utilizada no componente DIKTV é obtida através dos metadados situados no Coleta de dados (*Data Collection*), que armazena os dados de programação da emissora. Esses metadados baseados nos metadados obrigatórios do SBTVD representam as informações acerca do nome do programa, gênero do programa, emissora, data e horário em que o usuário começou e terminou de assistir ao programa etc.

Os arquivos contidos no *Data Collection* podem ser estruturados de formas divergentes e atualmente estão disponíveis no formato JSON (*JavaScript Object Notation*) (JSON, 2012) e XML (*Extensible Markup Language*) (XML, 2012). A Tabela 3 representa os metadados contidos no Data Collection que são utilizados pelo DIKTV nessa validação.

Tabela 3 - Metadados utilizados na validação

Metadados	Função
SBT-id	Identificar o Set-Top-Box
Program_name	Identificar o nome do programa
Channel	Identificar o canal de TV
Program_genre	Identificar o gênero do programa
Star_time	Data e horário de início do programa
End_time	Data e horário de término do programa

Fonte: (ABNT NBR 15603-1, 2007)

O formato de arquivo escolhido é o XML, pois permite que a origem e a forma de transmissão da mensagem sejam abstraídas, possibilitando assim a comunicação com o *middleware*, com outros serviços da *Web* e com as aplicações que enviem arquivos a serem armazenados e utilizados pelos componentes específicos do KTV (Araújo, 2011).

Segundo as normas do SBTVD (ABNT NBR 15603-1, 2007), o envio de alguns descritores é obrigatório ou opcional. Mas, como alguns desses descritores opcionais possuem informações necessárias à identificação do domínio por parte do componente DIKTV, é necessário utilizar a busca por esses metadados na Web.

Para a validação utilizamos o site XMLTV (XMLTV, 2013), que disponibiliza metadados de programação de TV de canais como SKY, NET, VIVO, OiTV, ClaroTV, CTBC, etc. Programação de mais de 600 canais de TV, 5 dias de programação por canal e atualizações diárias de 2ª a 6ª feira. Num formato compatível com a maioria dos programas de computador existentes na atualidade, ou seja, o formato XML.

4.2 Ontologias de Domínio

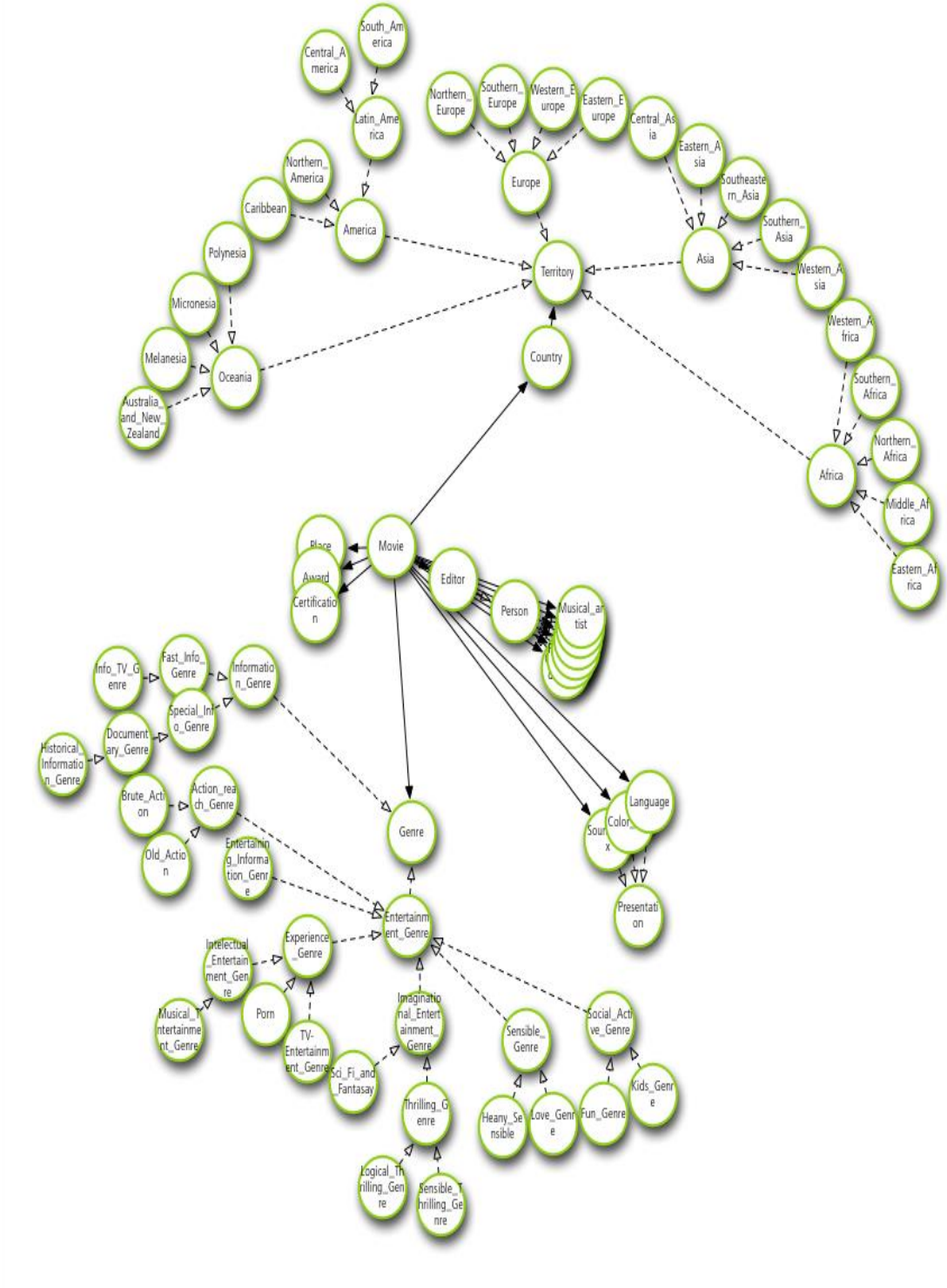
Para a validação do DIKTV foram escolhidas duas ontologias de domínio, uma de domínio de filmes e outra no domínio de esportes, devido à

enorme quantidade disponível de conteúdo de programas gerados a respeito desses domínios.

A ontologia escolhida para descrever o domínio de filmes é a *Movie Ontology* (Bouza, 2012), que descrever conceitos relacionados a filmes tais como o nome do filme, gênero, diretor, ator e indivíduos. Por exemplo, "Idade do Gelo", "Drama", "*Steven Spielberg*" ou "*Johnny Depp*".

“As ontologias de filme existentes definem indivíduos de um conceito superficialmente uma ontologia deve fornecer em uma hierarquia de conceitos e um conjunto suficiente de indivíduos que podem ser usados para descrever filmes” (Bouza, 2012). A MO (*Movie Ontology*) dispõe de infraestrutura para uma descrição em maiores detalhes e está representada em formato OWL. A visão geral dos conceitos modelados por essa ontologia é apresentada na Figura 20.

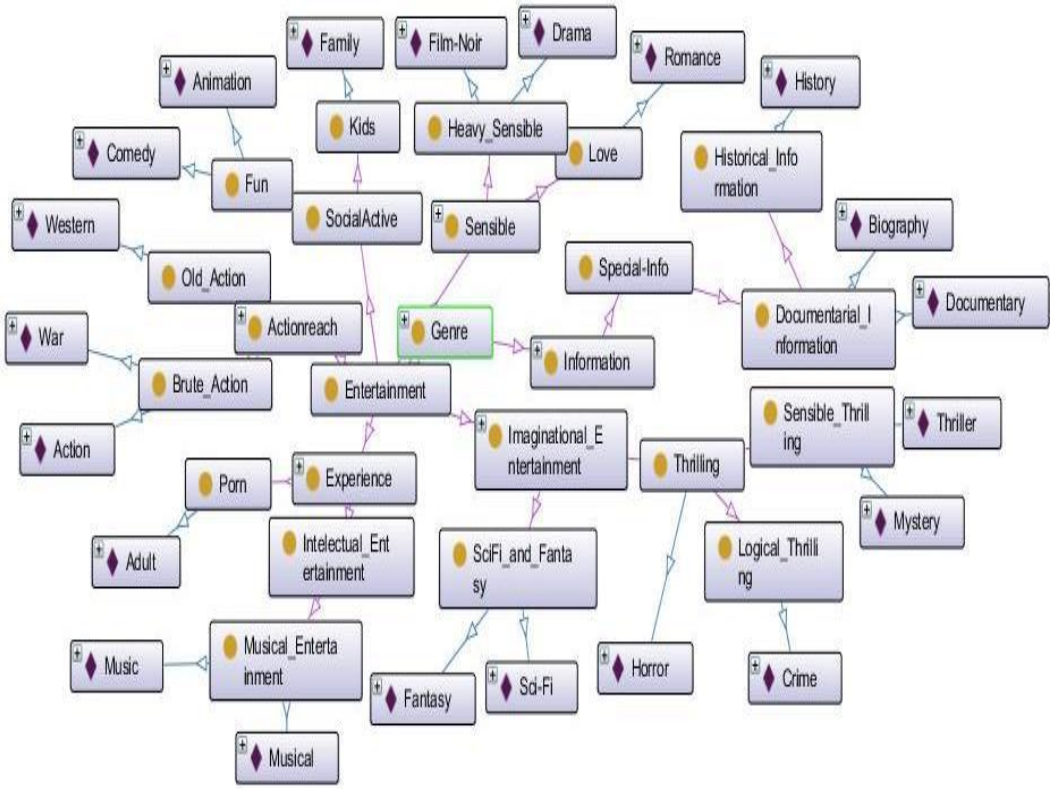
Figura 20 - Visão Geral dos Conceitos da *Movie Ontology*



Fonte: (Bouza, 2012)

Na Figura 21, ilustramos a superclasse Gênero, suas subclasses e individuo membros dessas subclasses.

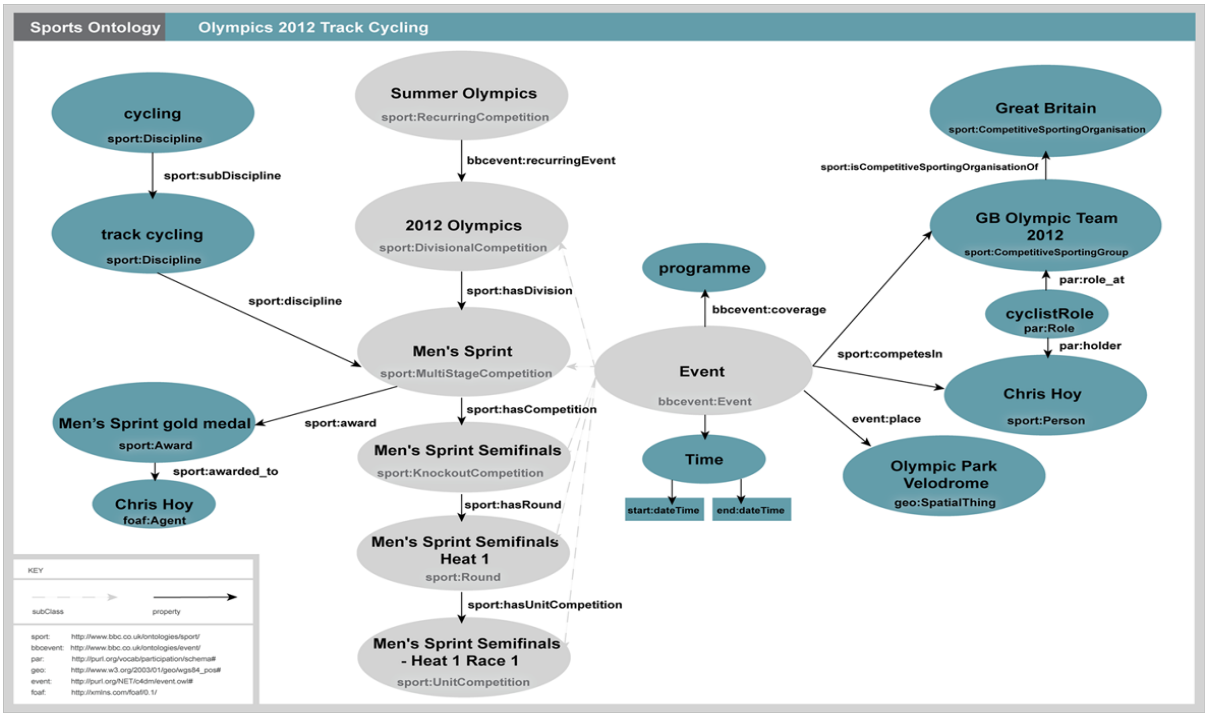
Figura 21 - Indivíduos da superclasse *Genre* da *Movie Ontology*



Fonte: (Bouza, 2012)

Durante a pesquisa foi identificado um modelo ontológico, feito pela BBC, que descreve competições esportivas de um modo geral, a ontologia de *Sport Ontology* (Sport Ontology, 2012). Na Figura 22 é apresentado um diagrama que ilustra as relações entre as classes chave da ontologia aplicada ao ciclismo olímpico.

Figura 22 - Diagrama *Sport Ontology* aplicada ao ciclismo olímpico



Fonte: (Sport Ontology, 2012)

A ontologia *Sport Ontology* permite a publicação dos dados sobre uma estrutura de torneios desportivos com um prêmio associado à competição. Além disso, é uma ontologia aplicável a uma grande variedade de eventos e permite a interoperabilidade com outras ontologias e encontra-se disponível em formato RDF (Sport Ontology, 2012).

4.3 Aplicações do DIKTV no módulo SQTV

O módulo *Semantic Query TV* (SQTV), parte do projeto *Knowledge* TV, é responsável pelas consultas semânticas baseadas em *Linked Data*[4]. O termo *Linked Data* é usado para definir uma prática usada para a disponibilização e compartilhamento das informações na *Web* Semântica. O objetivo do módulo SQTV é utilizar o *Linked Data* para enriquecer semanticamente os metadados

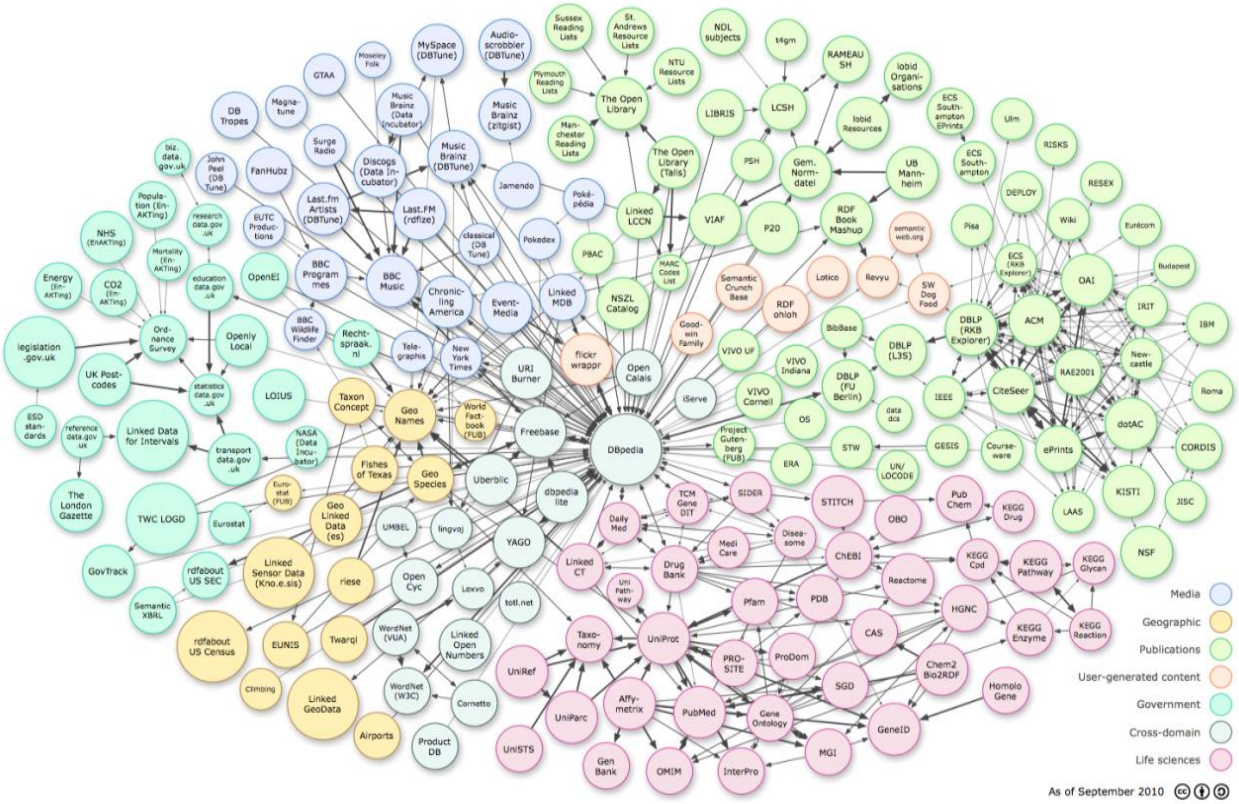
[4] http://linkeddata.org/

advindos das emissoras e retornar informações mais detalhadas, precisas e relevantes aos usuários de TV.

Porém, o *Linked Data*, representado na Figura 23, possui um enorme acervo de dados e consequentemente um diversificado número de domínios, como por exemplo, os representados na Figura 23 pelas cores. Isto dificulta uma consulta feita pelo módulo SQTV. Daí, a necessidade de utilização de um mecanismo que restrinja do espaço de busca apenas ao domínio que representa os metadados utilizados.

Um mecanismo de identificação do domínio além de possibilitar a restrição no espaço de busca ainda aumenta a eficácia do processamento de consulta e diminui o tempo gasto na consulta por dados pertencentes a domínios diferentes.

Figura 23 - Nuvem de dados interligados *Linked Data*

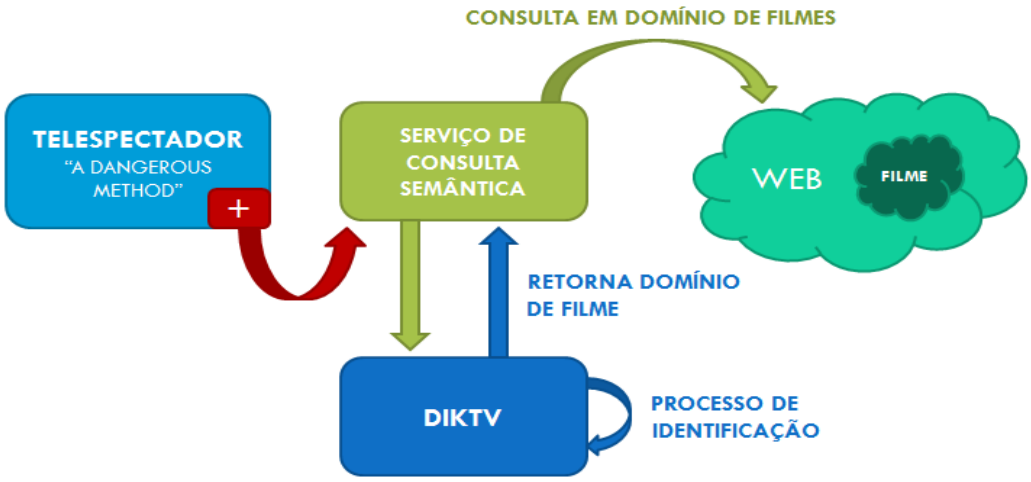


Fonte: (Heath & Bizer, 2011)

A Figura 23 é referente à interligação dos dados na *Web* disponível através das práticas do *Linked Data*. Como exemplo, se previamente é identificado o domínio da consulta, restringimos o espaço de busca apenas no que é relativo a esse domínio, e se esse mecanismo não é utilizado à busca acontecerá em um volume muito maior de dados.

Na Figura 24, demonstramos o cenário de validação do DIKTV através do módulo de consulta semântica.

Figura 24 - Cenário de validação SQTV e DIKTV

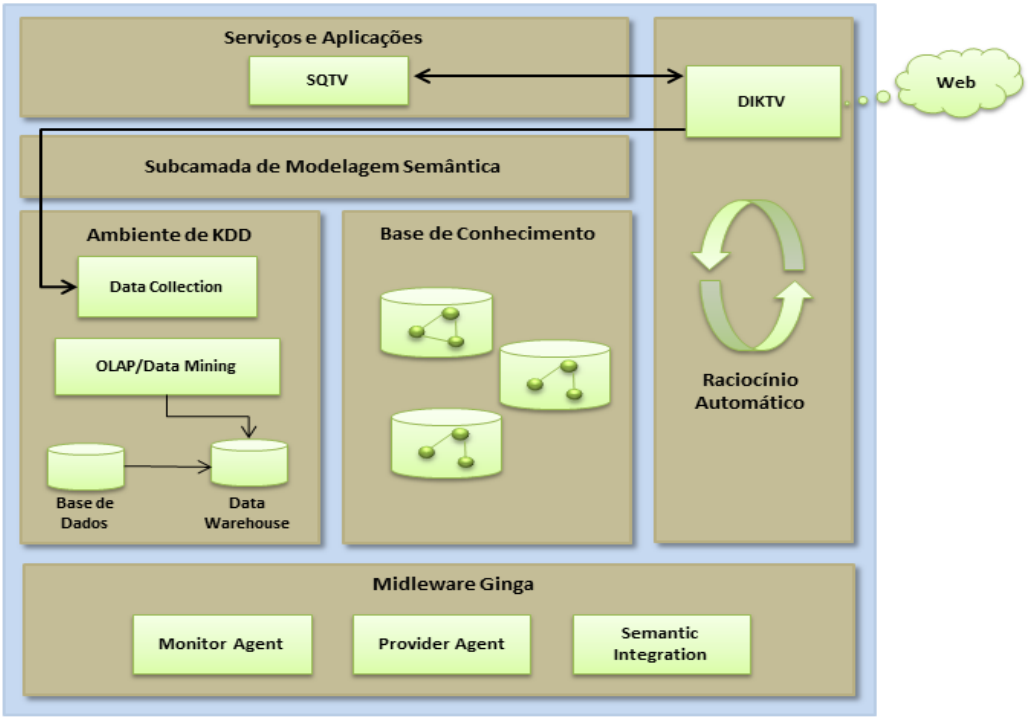


Fonte: (Próprio Autor, 2013)

Na figura 24, O telespectador assiste ao programa "*A Dangerous Method*" e deseja obter mais informações a respeito desse tipo de programação. Através de uma interface com a TV, o serviço de consulta semântica é acionado e realiza a consulta a palavra chave "*A Dangerous Method*" em toda a *Web*. Frequentemente, esse serviço não encontra informações relevantes, necessitando de mecanismos de identificação de domínio de aplicação. Nosso componente (DIKTV) é acessado pelo serviço de consulta semântica e através dos mecanismos da nossa abordagem, identifica domínios candidatos, neste caso o domínio de filme. Através do domínio, o módulo de consulta semântica realiza a consulta apenas no domínio de filmes.

O componente DIKTV provê os mecanismos necessários à identificação dos domínios de aplicação e é validado por meio da sua utilização conjunto com o módulo SQTV, um dos serviços do projeto *Knowledge* TV. A Figura 25 ilustra o processamento do componente.

Figura 25 - Arquitetura geral do DIKTV e SQTV



Fonte: (Próprio autor, 2013)

Na Figura 25, o componente SQTV acessa o DIKTV, que inicia seu processamento acessando a Coleta de Dados (*Data Collection*) e a *Web*, ao final desse processamento os resultados são retornados ao SQTV de acordo com a função escolhida. Através de uma interface de acesso, o SQTV poderá requisitar serviços disponíveis no DIKTV.

4.4 Expansão de Consulta Semântica

A expansão de consulta semântica é uma técnica utilizada em diferentes abordagens. Essa técnica consiste da adição de termos utilizados na consulta através de uma base de conhecimento léxico-semântico (na forma de

taxonomias, *tesauros* ou ontologias) ou então uma base de conhecimento léxico-estatístico geralmente construída através da medida da co-ocorrência dos termos nos documentos contidos no acervo (Bechara, 2010).

Essa técnica pode ser processada de forma automática, inserindo os termos da base de conhecimento na consulta ou semi-automática, quando os termos são apresentados ao usuário para que este escolha os termos que serão adicionados.

Nesta perspectiva, o presente trabalho visa validar o componente desenvolvido pela expansão de consultas baseada em ontologias de domínio, assumindo as ontologias de domínio como fonte de informação.

A escolha de uma taxonomia permite que sejam avaliadas as expansões semânticas baseadas nas relações de generalização-especialização. Na relação de generalização, ou expansão por termo genérico, são utilizados os conceitos de subclasses da ontologia e na relação de especialização, ou expansão por termo específico, são utilizados os conceitos de superclasses da ontologia. Por exemplo, termos encontrados na ontologia de filmes, como “diretor” e “linguagem” podem ser utilizados na expansão semântica e a consulta se divide em termos encontrados na subclasse e na superclasse da ontologia.

Na seção 4.6.3, verificamos a aplicação da técnica de expansão de consulta na validação do DIKTV.

4.5 Métricas Utilizadas na Avaliação

Nesta seção exploramos as métricas de avaliação utilizadas na validação deste trabalho: (i) Precisão (P) e (ii) *Mean Reciprocal Rank* (MRR).

I. Precisão (P): Bastante difundida na área de Recuperação de Informações (RI), a medida de precisão é expressa em um intervalo de 0 a 1, quando o resultado é igual a 1, significa que todos os objetos recuperados são relevantes e nenhum objeto irrelevante foi retornado na consulta.

Na área de RI, o cálculo de precisão corresponde ao número de elementos relevantes recuperados (A), dividido pelo número total de elementos recuperados (R) (Baeza-Yates & Neto, 1999), como ilustra a equação 1.

$$\text{Precisão (P)} = \sum (A) / \sum R \quad (1)$$

II. *Mean Reciprocal Rank* (MRR): A média de classificação recíproca é uma medida estatística que tem objetivo de avaliar qualquer processo que produz uma lista de possíveis respostas a uma amostra de consultas, ordenados por probabilidade de acerto (Voorhees, 1999).

A relevância das respostas encontradas nessa lista é atribuída pelo usuário, ou seja, o usuário deve julgar a adequação e relevância de cada documento retornado em satisfazer sua necessidade de informação.

O MRR é uma derivação da métrica *Reciprocal Rank* (RR). A RR tem como equação um *score* por consulta individual i que é igual à classificação recíproca ri, que corresponde à colocação em que a primeira resposta correta apareceu (Teufel, 2006). Ilustramos o RR na equação 2.

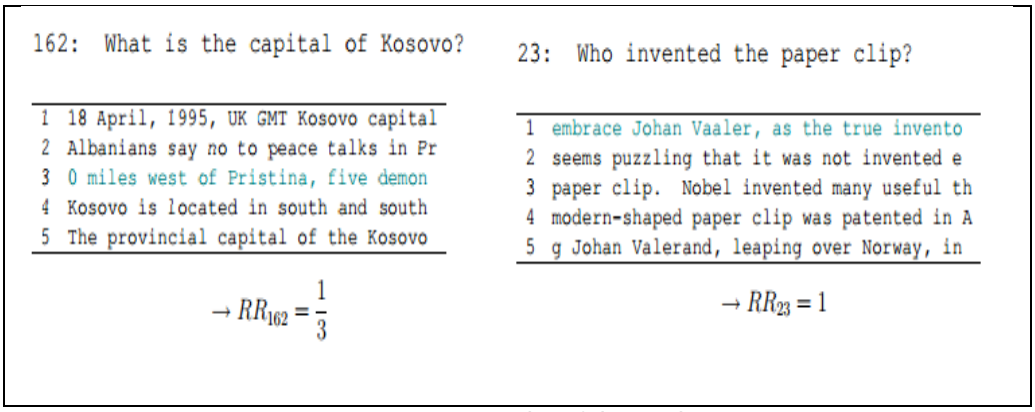
$$RR_i = \frac{1}{r_i} \quad (2)$$

A partir da equação da métrica RR pode-se definir a equação de MRR efetuando a equação MRR sobre n consultas, temos a equação 3.

$$MRR = \frac{1}{n} \sum_{i=1}^{n} RR_i \quad (3)$$

Na Figura 26, temos como exemplo duas consultas. A partir da resposta dessas consultas calculamos o RR e posteriormente o MRR.

Figura 26 - Exemplo da métrica *Reciprocal Ranking*



Fonte: (Teufel, 2006)

Conforme Figura 26, na consulta 162 a resposta mais relevante se encontra na terceira colocação e por esse motivo a equação RR é igual a 1/3, o mesmo ocorre com a consulta 23, no entanto, a resposta correta está na primeira colocação. O MRR é então calculado através da soma das consultas individuais (RR), dessa forma, temos como resposta: (1/3 + 1) / 2 = 0,66.

4.6 Experimentos e Avaliações dos Resultados

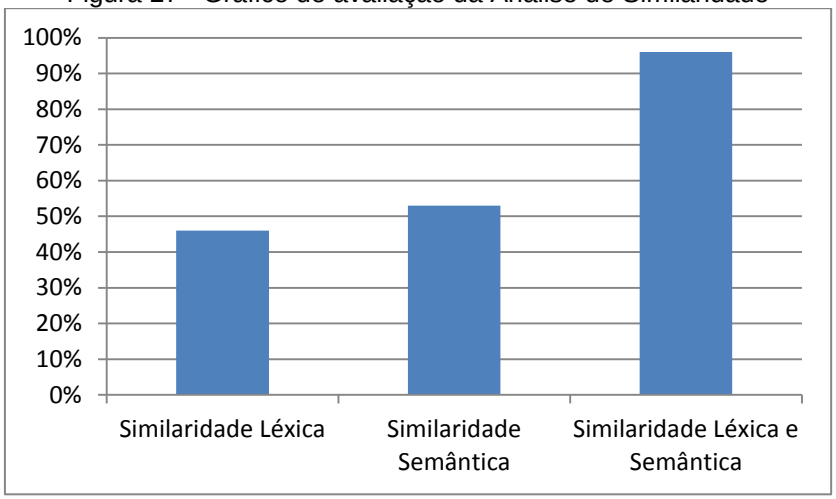
Experimentos foram realizados para avaliar de forma prática o componente DIKTV, objetivando validar sua aplicabilidade. Inicialmente investigamos o processamento do DIKTV acerca dos resultados obtidos quanto à utilização das análises de similaridade léxica e semântica (seção 4.6.1). Posteriormente foram realizados experimentos aplicando o DIKTV ao módulo de Consulta Semântica do KTV (SQTV) (seções 4.6.2 e 4.6.3).

4.6.1 Aplicação da Análise de Similaridade

Nesse experimento utilizamos uma base de dados contendo 84 itens (arquivos XML) e geramos os resultados que foram avaliados pela medida de

Precisão. Através da métrica de Precisão avaliamos a execução do DIKTV em relação à base de dados utilizada, e geramos o gráfico conforme Figura 27.

Figura 27 - Gráfico de avaliação da Análise de Similaridade



Fonte: (Próprio Autor, 2013)

O gráfico apresentado na Figura 27, foi gerado de acordo com os seguintes passos: inicialmente, utilizamos a abordagem proposta que define a utilização das análises de similaridade léxica e semântica no contexto dos domínios utilizados nesse trabalho. Os itens que correspondiam aos domínios especificados são definidos como corretos e a soma total é dividida pela base de 84 itens utilizados nas consultas, o resultado de precisão encontrado é de 96%. Os 4% resultantes são considerados itens incorretos, ocorrem por erro semântico, ou seja, palavras consideradas semanticamente equivalentes a um domínio não pertencem ao mesmo, tornando o resultado ambíguo.

Entendemos que seria essencial demonstrar a partir dos itens corretos da base, como são os resultados das perspectivas léxicas e semânticas individualmente. Na análise individual obtivemos uma precisão de 46% acerca da utilização da análise léxica e 53% na perspectiva semântica. Esse resultado demonstra que a combinação das duas análises é necessária e resulta em ganho de precisão.

4.6.2 Consulta por Domínio

Disponibilizamos o componente DIKTV em formato de API para que fosse integrado ao SQTV e utilizamos uma base de dados contendo 81 itens de consulta (arquivos XML), do qual 19 foram retornados pelo SQTV para comparação.

Através da interface de acesso (Figura 25) escolhemos a função de identificação de domínio, que inicia seu processamento acessando o módulo *Data Collection* através da função de coleta e se necessário a *Web*. Como retorno da função de coleta tem-se um arquivo do tipo XML (Figura 28), que é utilizado como entrada na função de identificação de domínio.

Figura 28 - Exemplo de arquivo XML encontrado na base de dados

```
<?xml version="1.0" encoding="ISO-8859-1"?>
    <SBT-id>7634732812</SBT-id>
    <program_name>A Dangerous Method</program_name>
    <channel>Telecine  HD</channel>
    <program_genre>Biography</program_genre>
    <start_time>20130418155500</start_time>
    <end_time>20130418174500</end_time>
```

Fonte: (Próprio Autor, 2013)

O DIKTV carrega as instâncias do Dicionário TVA (Figura 18) e das ontologias de domínio encontradas no repositório de ontologias (Figura 18) para posteriormente carregar os metadados de entrada. Realizada a análise de similaridade léxica e semântica, o algoritmo encontra similaridade total com a instância "*Biography*" da ontologia de domínio *Movie Ontology* (Figura 21). O resultado *Movie* é então retornado como domínio-candidato ao SQTV que realiza o processo de consulta semântica utilizando o domínio de filme. Através da API *Jena* (seção 3.3.3) realiza-se a consulta Sparql disponível nesta API (Figura 29).

Figura 29 - Exemplo de consulta semântica por domínio

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbpprop: <http://dbpedia.org/property/>
    SELECT distinct ?abstract
        WHERE {
          {?busca rdf:type dbpedia-owl:Film.}
          {?busca dbpedia-owl:abstract ?abstract.}
           filter (regex(?busca,'" + metadados.getNomePrograma() + "'))}
```

Fonte: (Próprio Autor, 2013)

A consulta SQTV original não realiza a consulta por um domínio. Portanto, através do DIKTV é possível realizar a consulta por domínio como destacado na Figura 29. Em consultas SPARQL o uso do "*PREFIX*" permite que o usuário abrevie URI[5] declarando prefixos, facilitando o acesso as informações.

Através da cláusula "*SELECT*" é possível extrair os valores brutos como o "*abstract*": os resultados são retornados em um formato de tabela onde utilizamos o parâmetro "*distinct*" para restringir o retorno de resultados iguais. Utilizando a cláusula "*WHERE*" definimos os parâmetros que são utilizados para encontrar uma correspondência no conjunto de dados de consulta. Na Figura 29, utilizamos os prefixos para acessar o domínio de "*Film*" com o objetivo de retornar o abstract referente ao nome do programa utilizado.

Avaliamos a aplicação do DIKTV ao módulo de Consulta Semântica do KTV (SQTV). Através das métricas de Precisão e MRR obtivemos os resultados apresentados na Tabela 4.

[5] URI do inglês (Uniform Resource Identifier) é uma cadeia de caracteres compacta usada para identificar ou denominar um recurso na internet.

Tabela 4 - Comparativo entre os resultados das consultas

Consultas/Métricas	**Precisão**	**MRR**
Consulta Original	46%	53%
Consulta por Domínio	60%	73%

Fonte: (Próprio Autor, 2013)

Como exemplo de consultas utilizadas nesse experimento, o filme "*The Perks of Being a Wallflowe*r" e a programação de esporte "*The Ultimate Fighter* 14".

Os percentuais de precisão são calculados da seguinte forma: na consulta original pelo nome do filme "*The Perks of Being a Wallflower*" a precisão é calculada onde Precisão = 19/27, resultando em 0.70%, ou seja, dos 27 resultados retornados apenas 19 estão relacionados ao filme. Na consulta por domínio, a precisão é calculada com Precisão = 7/7, resultando em 1.0, ou seja, dos 7 retornos os 7 estavam relacionados com o filme em questão.

Na consulta original por "*The Ultimate Fighter* 14" a precisão é calculada onde Precisão = ¾, resultando em 0.75%. Na consulta por domínio, a precisão é calculada onde Precisão = 4/4, resultando em 1.0.

Realizadas todas as consultas os resultados são somados e divididos pela quantidade de consultas resultando em 46% de precisão por consulta original e 60% na consulta por domínio.

Os percentuais de MRR são calculados da seguinte forma: Na consulta original e na consulta por domínio esse filme obteve como resultado 1, pois a resposta mais relevante está na primeira posição. Na consulta original por "The Ultimate Fighter 14" o retorno resultou em ½ e na consulta por domínio o resultado é 1.

Realizadas todas as consultas os resultados são somados e divididos pela quantidade de consultas resultando em 53% de MRR por consulta original e 73% na consulta por domínio. Logo, obtivemos uma melhoria de 14% na medida de precisão total, que era 46% na consulta original e passou para 60%

na consulta por domínio, e de 20% na média de classificação recíproca, que era 53% na consulta original e passou para 73% na consulta por domínio.

4.6.3 Expansão por termo genérico e específico

Com base no experimento anterior, as consultas foram realizadas utilizando termos genéricos e específicos. Através das ontologias de domínio utilizadas nesse trabalho investigamos as expansões semânticas baseadas nas relações de generalização-especialização.

O SQTV realiza a chamada ao método de expansão semântica de consulta através de uma interface de acesso (Figura 25). O DIKTV utiliza a ontologia-candidata de domínio para expandir os termos da ontologia e retorna os resultados ao SQTV. Dentre as consultas que retornaram resultados aos termos encontrados nas ontologias, verificamos a quantidade de termos utilizados em detrimento aos existentes nas ontologias. Utilizando a métrica de Precisão obtivemos os resultados que são apresentados na Tabela 5.

Tabela 5 - Comparativo entre os resultados das expansões

Ontologias/Expansão	**Expansão por Termos Genéricos**	**Expansão por Termos Específicos**
Movie Ontology	12%	5%
Sport Ontology	8%	0%

Fonte: (Próprio Autor, 2013)

Apesar de um baixo percentual as ontologias podem servir como modelo para a elaboração de uma consulta semântica, utilizando termos relevantes como os que obtiverem retornos na maioria das consultas desse experimento, por exemplo, “*country*”, “*language*”, “*writer*”, “*producer*” e “*round*”. Nesse experimento não utilizamos a métrica MRR, pois nosso objetivo é investigar a utilização dos termos das ontologias no processo de consulta. Investigamos a possibilidade de utilizar os termos encontrados nas ontologias como modelo para consulta.

4.7 Conclusão

Esse capítulo retrata a validação da abordagem proposta através do componente DIKTV. Para a validação do componente foi necessário especificar a fonte de dados utilizada e quais foram às ontologias de domínio escolhidas.

Em seguida, contextualizamos o módulo SQTV e definimos como e com que finalidade o SQTV aplicará o módulo DIKTV. Instanciando a arquitetura conceitual do KTV, podemos determinar como é feita a integração do componente e como ocorre o fluxo de informações.

Também definimos o conceito de expansão semântica e apresentamos sua aplicação no contexto da consulta semântica, destacando os benefícios do uso dos conceitos obtidos nas ontologias de domínio.

Por fim, apresentamos as métricas utilizadas e os experimentos e resultados obtidos na validação, detalhando os resultados através de um comparativo entre as análises de similaridade, a consulta original e a consulta por domínio, além de demonstrar os resultados de avaliação por expansão de termo genérico e por expansão de termo específico. Em resumo, obtivemos resultados satisfatórios quanto à aplicação da abordagem, além de contribuições válidas a consulta semântica.

5 Trabalhos Relacionados

Este capítulo apresenta alguns trabalhos encontrados na literatura que abordam estratégias de correspondências entre esquemas e alguns trabalhos relacionados aos cenários de identificação dos domínios de aplicação. A partir da análise desses trabalhos descrevemos as principais semelhanças e divergências entre eles e a abordagem proposta nesta dissertação.

5.1 Correspondências entre esquemas

Nesta seção apresentamos três trabalhos relacionados sobre correspondência de esquemas. O foco desses trabalhos são as abordagens léxicas e semânticas entre estruturas na finalidade de determinar sua similaridade.

5.1.1 Madhavan et. al.

Madhavan et. al. (2001) apresenta um algoritmo genérico, denominado CUPID, que realiza a combinação de esquemas (XML *schemas*, ontologias, entre outros) através de técnicas linguísticas como tesauro e radicalização.

Nessa abordagem os esquemas são transformados em árvores e sua execução compreende três fases. Na primeira fase efetua-se uma combinação linguística composta por três passos, conforme (Freitas, 2007): a normalização, onde é realizada a decomposição dos termos, além de utilizar um *tesauro* que pode conter tanto termos de uma linguagem comum quanto referências de domínio específico: a categorização e a comparação, onde se define o coeficiente de similaridade.

Tal abordagem é similar à proposta desta dissertação, pois consideram os três passos utilizados na análise de similaridade, e utilizam a técnica de radicalização e um *tesauro* para fornecer semântica aos termos. Entretanto, a

abordagem linguística do CUPID utiliza apenas as relações de sinonímia e hiperonímia fornecida pelo *tesauro*, enquanto que o DIKTV faz uso destas e de mais relações como antonímia e meronímia para expandir o significado dos termos. Isso reflete nos resultados, pois manipulamos mais termos e consequentemente mais possibilidades de encontrar equivalências.

5.1.2 Noll et. al.

No trabalho de Noll et. al. (2007), intitulado “Uma proposta para análise de similaridade entre documentos XML e ontologias em OWL”, é apresentada a ferramenta *The Matcher*, desenvolvida utilizando a linguagem de programação *Java*, que tem por finalidade avaliar a similaridade léxica e semântica entre documentos XML e ontologias.

Segundo os autores, “A ferramenta *The Matcher* foi incorporada ao *framework* DetVX, desenvolvido por Saccol (2008), com propósito de auxiliar na etapa de descoberta do domínio de conhecimento (ontologia) que descreve um conjunto de documentos XML”. Após escolha do arquivo XML e de uma das ontologias de domínio, o grau de similaridade é apresentado.

Alguns aspectos dessa ferramenta são semelhantes as do DIKTV, porém o processamento é diferente, considerando que os ambientes são diferentes. O *The Matcher*, assim como, o DIKTV, utiliza a Radicalização e o *WordNet* em seu processamento, entretanto, o *The Matcher* não dá suporte ao formato de ontologias RDF e utiliza o algoritmo de sobreposição de taxonomias para realizar a correspondência entre os esquemas. Esse algoritmo é bastante utilizado quando através das estruturas dos esquemas é possível identificar semelhanças e prover a comparação. Como no ambiente de TV o XML gerado não possui relações de hierarquia bem definidas, o uso desse algoritmo é inviável, e por esse motivo o DIKTV não faz uso do mesmo.

5.1.3 COMA 3.0

O COMA 3.0, sucessor do COMA ++, foi desenvolvido pela Universidade de Leipzig. Seu objetivo é realizar o casamento entre esquemas e ontologias,

através de uma ferramenta genérica e personalizável, que possui interface gráfica abrangente, além de prover suporte aos modelos XML *Schema*, SQL, W3C XSD e OWL (Aumueller et. al., 2005).

Essa ferramenta identifica as correspondências semânticas entre as estruturas ou modelos, na finalidade de prover a interoperabilidade de serviços e a integração de dados em vários domínios de aplicação de forma automática.

De acordo com Saccol et. al. (2008), a similaridade entre dois modelos é determinada por uma função de similaridade entre dois elementos pertencentes à taxonomia (descritos pela ontologia), no entanto, os autores não demonstram como feito o cálculo da similaridade entre os modelos e a taxonomia. O COMA utiliza um algoritmo básico com base na análise de rótulos que podem detectar as relações entre conceitos, esta abordagem pode ser estendida usando, por exemplo, *WordNet* (Massmann et. al., 2011).

A análise de similaridade semântica e o uso do tesauro *WordNet* são os pontos similares entre a abordagem desta dissertação e o COMA, porém está ferramenta é bastante direcionada à interoperabilidade de ontologias, sendo aplicada somente a similaridade semântica. Em contrapartida, o DIKTV aborda aspectos léxicos e semânticos requeridos pelo problema que o DIKTV soluciona.

5.2 Cenários de Identificação de Domínios de Aplicação

Nesta seção são apresentados dois trabalhos relacionados sobre identificação de domínios de aplicação em contextos diferentes. Esses trabalhos utilizam mecanismos semelhantes à proposta neste trabalho para determinar o domínio

5.2.1 Saccol

No trabalho intitulado “Detecção, Gerenciamento e Consulta a Réplicas e a Versões de Documentos XML”, a autora desenvolveu um *framework* chamado DetVX (Detector de Réplicas e de Versões de Documentos XML). Um

dos componentes principais do *framework* é o gerenciador de ontologias que é responsável pela categorização dos arquivos em seus respectivos domínios de aplicação. Ele assume que cada arquivo corresponde a um domínio descrito por uma ontologia.

Nesse *framework* as buscas por réplicas e versões de arquivos aplicam-se somente aos arquivos específicos de seu domínio, trazendo benefícios como a restrição no espaço de buscas e o aumento da eficiência no processamento de consultas.

Conforme Saccol (2008), o arquivo XML é analisado pelo componente gerenciador de ontologias responsável por identificar seu domínio e posteriormente registrado. Após esse processo, o ambiente verifica se estes arquivos referem-se a versões ou réplicas de documentos, tarefa realizada pelo gerenciador de réplicas e de versões.

Como estudo de caso, o *framework* foi aplicado em sistemas *peer-to-peer* com a finalidade agrupar os arquivos relacionados e realizar o processamento de consultas somente em um subconjunto da rede. O *framework* proposto por Sacool permite a restrição das consultas e viabiliza o processamento dos documentos XML, além de utilizar as ontologias de domínio como fonte de informação, diferenciando-se da nossa abordagem no cenário de aplicação.

5.2.2 Moro et. al.

Em seu trabalho, Moro et. al. (2009) apresentam a Tricot (*TRIple Content-based OnTology*), uma proposta baseada em ontologias para disseminação de documentos XML. O objetivo dessa abordagem é agrupar as consultas (documentos) de acordo com um mesmo domínio, para isso é realizado o casamento entre os documentos e as ontologias. Este trabalho, assim como o trabalho de Saccol (2008), é aplicado a sistemas *peer-to-peer* e seus processamentos são similares.

Incialmente o documento é recebido pelo *super peer* que realiza o casamento com o seu conjunto de ontologias e identifica o domínio da consulta, em seguida o *super peer* encaminha a consulta para todos os *peers* de mesmo domínio. Os *peers* contêm apenas os conjuntos de consultas que se referem aos mesmos domínios dos seus documentos, então o número de consultas a serem avaliadas é consideravelmente reduzido. Dessa forma, não existe a necessidade de processar todo o conjunto de consultas em cada documento e os *peers* processam apenas o subconjunto de consultas do mesmo domínio do documento. Similar à abordagem empregada nesta dissertação, às aplicações *peer-to-peer* necessitam utilizar um mecanismo que forneça a identificação do domínio, possibilitando assim a utilização de *peers* de um determinado domínio e as consultas referentes ao mesmo.

5.3 Análise Comparativa

Os trabalhos apresentados foram comparados de acordo com algumas características. Os principais aspectos observados foram:

1. Similaridade Léxica: através de algoritmos léxicos de comparação de sequências de caracteres retornam um percentual indicando a semelhança entre os termos utilizados.
2. Similaridade Semântica: faz uso de bases de conhecimento como tesauros para identificar correspondências semânticas entre os termos.
3. Técnica de Radicalização: através de algoritmos que empregam essa técnica é possível obter radicais de palavras e utiliza-los na comparação dos termos.
4. Tesauro: modela o conhecimento lexical através de relações semânticas.
5. Algoritmo *Levenshtein*: algoritmo léxico, aplicado em casos onde não há análise de grafia.
6. Ontologias de Domínio: estrutura semântica de representação de conhecimento que possui conceitos referentes a um determinado domínio.

7. Suporte XML, OWL e RDF: faz uso de linguagens de estruturação e representação de conteúdo.
8. Normalização, Categorização e Comparação: contém passos utilizados no emprego da análise de similaridade.

A Tabela 6 apresenta as características utilizadas por cada um dos trabalhos discutidos conforme a correspondência entre esquemas e os cenários de identificação de domínios de aplicação. Os campos sem marcação indicam que o trabalho não apresenta a característica correspondente.

Tabela 6 - Comparativo entre os trabalhos apresentados

Características	Correspondências entre esquemas			Cenários de Identificação de Domínios de Aplicação		DIKTV
	Madhavan et.al.	Noll et. al.	COMA 3.0	Saccol	Moro et. al.	
Similaridade Léxica		X		X		X
Similaridade Semântica	X	X	X	X		X
Técnica de Radicalização		X		X		X
Tesauro	X	X	X	X		X
Algoritmo *Levenshtein*						X
Ontologias de Domínio	X	X	X	X	X	X
Suporte XML, OWL e RDF			X			X
Normalização,Categorização e Comparação	X					X

Fonte: Próprio Autor (2013)

5.4 Conclusão

Neste capítulo foram descritos trabalhos que utilizam aspectos de correspondência entre esquemas e cenários de identificação de domínios de

aplicação. Através de uma análise comparativa verificamos as características atendidas por cada trabalho e o cenário de aplicação das mesmas.

Exceto pelo algoritmo de *Levenshtein* utilizado em nossa abordagem, todos os trabalhos compartilham de alguma característica definida. Devido à busca Web realizada pelo componente DIKTV, não podemos afirmar que os resultados da busca contêm ou não erros de grafia e por esse motivo elegemos o referido algoritmo. Por exemplo, o conteúdo da *Wikipedia*, onde os usuários adicionam conteúdo sendo que não existe verificação ortográfica do mesmo.

Encontramos cenários de identificação de domínios de aplicação implantados em sistemas *peer-to-peer* que identificam os domínios dos documentos contidos nesse sistema com objetivo de separá-los e dessa forma facilitar a recuperação dos mesmos. Não encontramos trabalhos inseridos no contexto de TVDI, o que demonstra que nosso trabalho é relativamente novo nesse ambiente.

Alguns dos trabalhos não dão suporte as principais linguagens de representação como XML, OWL e RDF, ficando limitados ao uso de determinadas ferramentas. Em nosso trabalho provemos o suporte e aplicação dessas linguagens tanto para manipulação dos metadados quanto para o uso das ontologias.

Outra característica compartilhada apenas por um trabalho relacionado foi o modelo de aplicação da análise de similaridade. Neste trabalhou adequamos o modelo aos requisitos levantados e essa estrutura facilitou a compreensão das etapas da arquitetura conceitual.

6 Considerações Finais

Neste capítulo são discutidas conclusões acerca deste trabalho, incluindo as principais contribuições do mesmo e alguns trabalhos futuros.

6.1 Principais Contribuições

Inicialmente, focamos nossos estudos na aplicação das ontologias de domínio em ambiente de convergência digital, para tanto, foi necessário o estudo acerca da *Web* Semântica integrada a TV Digital Interativa e como o emprego das ontologias de domínio apoia a investigação dos domínios tratados nesse ambiente. Esse emprego de ontologias é um tema relativamente novo neste domínio de convergência digital, e os trabalhos relacionados são reduzidos.

Após um estudo detalhado sobre as ontologias e o processo de convergência digital, foi possível investigar métodos de implantação desse trabalho. Analisando as necessidades existentes elegemos a análise de similaridade como técnica para aplicação da abordagem proposta, utilizando sua perspectiva léxica e semântica com a finalidade de apoiar a investigação dos domínios de aplicação apresentados.

O emprego das duas perspectivas é extremamente necessário, como pode ser visto na validação de desempenho da análise de similaridade nesse trabalho. Acerca desse tema também foram alvo de estudos um algoritmo de correspondência de caracteres, algoritmos de radicalização e um tesauro.

Essa fundamentação proporcionou o levantamento dos requisitos utilizados na implantação da abordagem proposta, especialmente incluindo o mapeamento dos metadados utilizados, conforme normas e padrões adotados. Os requisitos foram então implantados em um componente que está integrado a plataforma Knowledge TV. Utilizamos um cenário de uso no qual integramos

o componente ao módulo de Consulta Semântica como forma de validação do componente desenvolvido neste trabalho. Nessa validação, obtivemos resultados satisfatórios e contribuições validas relativas à precisão das informações.

Apresentamos uma abordagem para identificação de domínios de aplicação em TVDI que através do componente, denominado DIKTV, foi validada. Através dessa abordagem, foi criado um processo automatizado e genérico de identificação de domínios almejando algumas contribuições para a área de pesquisa, tais como: apresentação de uma abordagem para identificação de domínios de aplicação que permite a restrição do espaço de busca e o aumento da eficácia no processo de recuperação das informações; representação semântica de conteúdo, através das ontologias de domínio; uma infraestrutura genérica, possibilitando o acesso por diversos serviços; expansão de consultas semânticas e a aquisição de novas ontologias de domínio e métricas de similaridade.

6.2 Trabalhos Futuros

Como trabalhos futuros, identificamos incialmente a possibilidade de integração do componente DIKTV com demais serviços que serão desenvolvidos na plataforma Knowledge TV, como também a utilização do DIKTV por outros serviços independentes. Dessa forma é possível agregar novos cenários de uso para nosso componente, avaliando assim suas contribuições. Algumas perspectivas de implantação da abordagem são as seguintes:

- Aplicação de outros mecanismos de análise de similaridade;
- Inclusão de novas ontologias de domínio;
- Inclusão de outros formatos de metadados, visando à integração dos mesmos;
- Utilização de outros repositórios da *Web* como fonte de recursos;

- Criar uma interface em que o usuário possa escolher a partir dos termos encontrados nas ontologias, quais os termos que o mesmo deseja utilizar na consulta.
- Aplicação da expansão por sinonímia a partir do tesauro *WordNet*.
- Aplicação dos mecanismos de radicalização e do tesauro *WordNet* nos termos das ontologias de domínio e no termos do dicionário TVA.

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.

ABNT NBR 15606-1. 2008. Televisão digital terrestre — Codificação de dados e especificações de transmissão para radiodifusão digital – Parte 1: Codificação de dados. 2008

ABNT NBR 15606-2. 2008. Televisão digital terrestre — Codificação de dados e especificações de transmissão para radiodifusão digital – Parte 2: Ginga-NCL para receptores fixos e móveis – Linguagem de aplicação XML para codificação de aplicações. 2008.

AJAXIAN. Trying to generate more hype than Rails. Disponível em <http://ajaxian.com/archives/visual-wordnet>. Acesso em 15 de Novembro 2012.

ANATEL. Disponível em <http://www.anatel.gov.br>. Acesso em 21 de maio de 2013

ARAÚJO, J. P. C. CoreKTV - Uma infraestrutura baseada em conhecimento para TV Digital Interativa: um estudo de caso para o middleware Ginga. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática. Universidade Federal da Paraíba. João Pessoa, PB: 2011 p.139.

AUMUELLER, D., Do, H., MASSMANN, S. E RAHM, E. Schema and Ontology Matching with COMA++. Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland. p. 906-908, 2005.

BAEZA-YATES, Ricardo B.; NETO, Berthier R. Modern Information Retrieval. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

BECHARA, A. Expansão semântica de consultas baseada em esquemas terminológicos: uma experimentação no domínio biomédico. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática. Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ: 2010.

BERNERS-LEE, T. Semantic Web - XML2000. [2000]. Disponível em: <http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html>. Acesso em: 25 de Julho de 2013.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, New York, May, 2001.

BOUZA, A. The Movie Ontology. Disponível em: <http://movieontology.org>. Acesso em: 22 de Novembro de 2012.

BUDANITSKY, A.; Lexical Semantic Relatedness and Its Application in Natural Language Processing, Processing Computer Systems Research Group, University of Toronto, 1999.

BRACKMANN, C. P. Sistema Brasileiro de TV Digital. Programa de Pós Graduação em Informática. Universidade Católica de Pelotas. Pelotas, RS. 2008.

COMA. COMA 3.0: Schema and Ontology Matching with COMA 3.0. Disponível em< http://dbs.uni-leipzig.de/de/Research/coma.html>. Acesso em 15 de Outubro de 2012.

CROCCO, G. Agrupamento de Textos por Similaridade para Sistema de Clipping Web: ClusterClipping. Dissertação (Mestrado em Ciência da Computação) - Universidade Luterana do Brasil. Canoas - RS. 2010.

DEITEL, H. M; DEITEL, P. J; NIETO, T. R. et al. XML: How to program. Porto Alegre: Bookman, 2003.

DILLI, R. N; YAMIN, A. C; PALAZZO, L. A. M. Em Direção a Descoberta de Recursos Baseada em Matching Semântico para UBICOMP. Centro Politécnico. Universidade Católica de Pelotas (UCPel), Pelotas, RS. 2009.

DING, Y; FOO, S. Ontology research and development. Part 1 - a review of ontology mapping and evolving. Journal of Information Science*, v.28, n.10, 2002.*

DZIEKANIAK, G. V; KIRINUS, J. B. Web Semântica. Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, n.18, p. 20-39, 2004. Disponível em < http://www.periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2004v9n18p20>. Acesso em 15 de Outubro de 2012.

FANTAUZZI, E. Cultura da Convergência e a TV Digital Interativa: Novos Desafios para o Design Instrucional de Cursos a Distância Mediados pelas TICS. LECOTEC - II Simpósio de Comunicação, Tecnologia e Educação Cidadã. Bauru, SP. 2009.

FERNANDO FILHO, W. B. H.; LÓSCIO, B. F. Web Semântica: Conceitos e Tecnologias. Disponível em <http://www.ufpi.br/subsiteFiles/ercemapi/arquivos/files/minicurso/mc9.pdf>. Acesso em 21de Novembro de 2012.

FERNANDO FILHO, W. B. H.; LÓSCIO, B. F.; MACEDO, J. A. F. Geração incremental de correspondências entre ontologias. IX Workshop de Teses e Dissertações em Banco de Dados. Belo Horizonte, MG. 2010.

FONSECA, P. G. S. Índices Completos para Casamentos de Padrões e Inferência em Mo-tifs. 2003. 118f. Dissertação (Mestrado em Ciência da Computação) - Centro de Informática, Universidade Federal de Pernambuco, Recife, 2003.

FREITAS, J. B. SiSe: Medida de Similaridade Semântica entre ontologias em português. Dissertação (Mestrado) – Programa de Pós-Graduação em Ciência da Computação. Pontifícia Universidade Católica do Rio Grande do Sul. Porto Alegre, RS: 2007 p.119.

FREITAS, F. L. G.. Ontologias e a Web Semântica. Anais do XXIII Congresso da Sociedade Brasileira de Computação. Volume 8: Jornada de Mini-Cursos em Inteligência Artificial. Campinas: SBC, 2003, v. 8, p. 1-52

GIGLIO, K.; VEIRAS, A. F.; SOUZA, M. V.; SPANHOL, F. J. Metadados como Viabilidade Para Organização e Gerenciamento De Conteúdo Multimídia Interativo No Ambiente Digital Televisivo. Intercom – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação. XXXIV Congresso Brasileiro de Ciências da Comunicação. Recife, PE. 2011.

GOMES, H. E. Classificação, Tesauro e terminologia: fundamentos comuns. Disponível em: <http://www.conexaorio.com/biti/tertulia/tertulia.htm>. Acesso em: 10 de Setembro de 2012.

GRUBER, T. A translation approach to portable ontology specifications. *Knowledge*
Acquisition, v.5, p. 199-220, 1993.

GRUBER, T. What is an ontology? [S.l.:s. n.], 1996. Disponível em < http://www-ksl.stanford.edu/kst/what-is-an-ontology.html >. Acesso em 10 de setembro de 2012.

GUARINO, N. Formal Ontology and Information Systems. In: PROCS. OF FOIS, 1998. Anais...IOS Press, 1998.

GUIZZARD, G. Desenvolvimento para e com reuso: Um estudo de caso no domínio de vídeo sob demanda. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática. Universidade Federal do Espírito Santo. Vitória, ES: 2000 p.153.

HAHN, R. M. Sim: Uma Arquitetura para Determinação de Similaridade entre Trilhas. Dissertação (Mestrado) – Programa Interdisciplinar de Pós-Graduação em Computação Aplicada. Universidade do Vale do Rio dos Sinos. São Leopoldo, RS: 2011 p.68.

HINZ, V. T. Algoritmos para interoperabilidade entre ontologias. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática. Universidade Católica de Pelotas. Pelotas, RS: 2007 p.90.

HLIAOUTAKIS, A., VARELAS, G., VOUTSAKIS, E., PETRAKIS, E. G. M., MILIOS, E. Information Retrieval by Semantic Similarity. International Journal on Semantic Web and Information Systems (IJSWIS), Special Issue of Multimedia Semantics, Vol. 3, No. 3, July/September, 2006, pp. 55-73.

IBGE. Pesquisa Nacional por Amostra de Domicílios, Síntese de Indicadores 2009. Disponível em: < http://www.ibge.gov.br/home/estatistica/populacao/trabalhoerendimento/pnad2009/pnad_sintese_2009.pdf >. Acesso em 21 de Novembro de 2012.

JENA. Disponível em: <http://jena.apache.org>. Acesso em 14 de Maio de 2013.

JIVANI, A. G. “A Comparative Study of Stemming Algorithms,” IJCTA, Vol. 2, No. 6, pp. 1930-1938, 2011. Disponível em: < http://www.ijcta.com/documents/volumes/vol2issue6/ijcta2011020632.pdf>. Acesso em 25 de Julho de 2013.

JODA. Disponível em: <http://joda-time.sourceforge.net/>. Acesso em 14 de Maio de 2013.

JOSÉ JÚNIOR, C. A. P. Mining *Knowledge* TV: Uma Abordagem de Ambiente de KDD com Ênfase em Mineração de Dados no Ambiente da *Knowledge* TV. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática. Universidade Federal da Paraíba. João Pessoa, PB: 2012 p.115.

JSON. Introducing JSON. Disponível em: <http://www.json.org/>. Acesso em: 23 de Novembro de 2012.

KANTROWITZ M.; MOHIT B.; MITTAL V. Stemming and its effects on TFIDF ranking. In: Nicholas J, Peter I, Mun-Kew L, eds. Proc. of the 23rd Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2000. 357-359.

KULESZA, R.; ALVES, L.G.P.; SILVA, F.S.; JUCA, P.; BRESSAN, G. Análise Comparativa de Metadados em TV Digital. Anais do XXIV Simpósio Brasileiro de Redes de Computadores/ II Workshop de TV Digital, Curitiba-PR, pp. 87-98, 2006.

LEITE, L. E. C., et al. FlexTV – Uma Proposta de Arquitetura de Middleware para o Sistema Brasileiro de TV Digital. Revista de Engenharia de Computação e Sistemas Digitais. 2005, Vol. 2, pp. 29-50.

HEATH. T; BIZER. C. Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. 2011.

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.

MADHAVAN, J.; BERNSTEIN, P. A.; RAHM, E. Generic Schema Matching with Cupid. In: International Conference on Very Large Data Bases, 27., 2001. Proceedings…New York: ACM 2001. P. 48-58.

MAEDCHE, A.; STAAB, S. “Measuring similarity between ontologies”. In: Proceedings of the European Conference on *Knowledge* Acquisition, Modeling and Management- EKAW, 2002.

MARTINS, M. B. Organização de Dados Operacionais, Analíticos, Semânticos e Linkados no Projeto *Knowledge* TV. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática. Universidade Federal da Paraíba. João Pessoa, PB. 2012.

MASSMANN, S., Raunich, S., Aumüller, D., Arnold, P., & Rahm, E. Evolution of the COMA match system. In The Sixth International Workshop on Ontology Matching. Germany, 2011.

McCALLUM, A. String Edit Distance (and intro to dynamic programming). Computational Linguistics ,CMPSCI 591N, Spring 2006. University of Massachusetts Amherst. Disponível em <

http://people.cs.umass.edu/~mccallum/courses/cl2006/lect4-stredit.pdf >. Acesso em 23 de Outubro de 2012.

MILLER, G. A.; BECKWITH, R.; FELLBAUM, C.; GROSS, D.; MILLER, K. J. Introductio to Wordnet: An On-line Lexical Database. Int J Lexicography, 3(4):235-244, January 1990.

MORO, M.; GALANTE, R. ; SACOOL, D. ; LÓSCIO, B. F. . Disseminação de Conteúdo XML Baseada em Ontologias. In: SEMISH - XXXIV Seminário Integrado de Hardware e Software, 2009, Bento Gonçalves. Anais do XXXIV Seminário Integrado de Hardware e Software. Porto Alegre: SBC, 2009.

NASCIMENTO, F. F. AVANTV: Uma Aborgagem para Personalização do Conteúdo de Aplicações de TV Digital Interativa Sensível ao Contexto. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática. Universidade Federal da Paraíba. João Pessoa- Paraiba, 2011.

NGUYEN, H., AL-MUBAID, H.: New Ontology-based Semantic Similarity Measure for the Biomedical Domain. IEEE conference on Granular Computing GrC-2006, pp. 623-628, 2006.

NOLL, R.; SACCOL, D. B., EDELWEISS, N. Uma proposta para análise de similaridade entre documentos XML e ontologias em OWL. In: Sessão de Pôsteres, em conjunto com Simpósio Brasileiro de Banco de Dados (SBBD), João Pessoa, Brasil, 2007.

OLIVEIRA, F. N, B. Aplicação Adaptativa de Guia Eletrônico utilizando o Ginga-NCL. Dissertação (Mestrado) - Programa de Pós-Graduação em Informática. Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, RJ: 2010.

OLIVEIRA, H. G.; GOMES P.; SANTOS, D. Papel – Trabalho Relacionado e Relações Semânticas em Recursos Semelhantes. Departamento de Engenharia Informática. 2007.

OWL. OWL Web Ontology Language Guide [2004]. Disponível em: < http://www.w3.org/TR/owl-guide/>. Acesso em 25 de Julho de 2013.

OWL. Disponível em: <http://www.w3.org/OWL/>. Acesso em 12 de Outubro de 2012.

PETRAKIS, E., VARELAS, G., HLIAOUTAKIS, A.,RAFTOPOULOU, P. : Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies. 4th Workshop on Multimedia Semantics (WMS'06), pp. 44-52, 2006.

PIRES, A. D. S.; BEZERRA, E. P.; LINO, N. C. Q. (Submetido)

RESNIK, P.: Using information content to evaluate semantic similarity. In proceedings of the 14th International Joint Conference on Artificial Intelligence, pp 448-453, Canada, 1995.

RITA. Disponível em: <http://www.rednoise.org/rita/wordnet/documentation/>. Acesso em Fevereiro de 2013.

RDF. Disponível em: <http://www.w3.org/RDF/>. Acesso em 22 de Outubro de 2012.

SACCOL, D. D. B. Detecção, Gerenciamento e Consulta a Réplicas e a Versões de Documentos XML. Dissertação (Mestrado) - Programa de Pós-Graduação em Computação. Universidade Federal do Rio Grande do Sul. Porto Alegre, RS: 2008 p.150.

SACCOL, D. B. ; EDELWEISS, N. ; GALANTE, R. M. ; MELLO, M.R.. Gerenciamento de Domínios de Aplicação através do Uso de Ontologias. Em: IV Escola Regional de Banco de Dados - ERBD, 2008, Florianópolis. Anais da ERBD 2008.

SALES, F. L. Ontologias de domínio: estudo das relações conceituais e sua aplicação. Dissertação (Mestrado) - Programa de Pós-Graduação em Ciência da Informação. Universidade Federal Fluminense. Niteroi, RJ: 2006 p.139.

SANTOS, R. N; LIRA, M. A; LINO, N. C. Q. Consulta Semântica na TV: Uma Abordagem de Consulta Semântica em Ambientes de Convergência (Web e TV). Workshop de Pós-Graduação (WPG) - I Escola Paraibana de Informática (EPI 2011), João Pessoa, PB, Brasil, 2011.

SANTOS, V. Buscas Semânticas na identificação de similaridades entre conceitos para Integração Semântica de Informações. Universidade Federal do Estado do Rio de Janeiro. Simpósio Brasileiro de Sistemas de Informação (SBSI), 2010.

SILVA, D. F. Estudo de Funções de Similaridade Semântica de Termos Aplicadas a um Domínio. Monografia (Graduação) - Graduação em Ciência da Computação. Universidade Federal de Pernambuco. Recife, PE: 2008 p.45.

SIMMETRICS. Disponível em: < http://sourceforge.net/projects/simmetrics/>. Acesso em 23 de Fevereiro de 2013.

SOARES, L. F. G. MAESTRO: The Declarative Middleware Proposal for the SBTVD. Proceedings of the 4th European Interactive TV Conference. 2006.

SOARES, L. F.; LEMOS, G. (2007) Interactive Television in Brazil: System Software and the Digital Divide. In European Interactive TV Conference - EuroITV2007. Amsterdam, 2007.

SOUZA, G. L. SOARES, L. F. G. Interactive Television in Brazil. Disponível em: <http://www.ncl.org.br/documentos/EuroiTV2007S.pdf >. Acesso em 08 de Maio de 2013.

SOUZA, L. A. C. Uma Metodologia para detecção e previsão de fluxo de tráfego, com determinação de percursos alternativos em regiões monitoradas. Dissertação (Mestrado) – Mestrado Profissional em Computação Aplicada. Universidade Estadual do Ceará. Fortaleza, CE. 2012.

SOUZA, R. R.; ALVARENGA, L. A web semântica e suas contribuições para a ciência da informação. Ciência da Informação, Brasília, v. 33, n. 1, p. 132-141, jan./abr. 2004.

SPORT ONTOLOGY. Disponível em: <http://www.bbc.co.uk/ontologies/sport/2011-02-17.shtml>. Acesso em 22 de Novembro de 2012.

STEMMING. The Lancaster Stemming Algorithm. Disponível em < http://www.comp.lancs.ac.uk/computing/research/stemming/>. Acesso em 23 de Outubro de 2012.

TEUFEL, S. Information Retrieval. Lecture 7: Question Answering. University of Cambridge, 2006. Disponível em: <http://www.cl.cam.ac.uk/teaching/0607/InfoRtrv/lec7.2.pdf>. Acesso em 08 de Junho de 2013.

TVA – TV-Anytime Forum. Disponível em: <http://www.tv-anytime.org>. Acesso em 08 de Maio de 2013.

VIEIRA, A. F. G; VIRGIL, J. Uma revisão dos algoritmos de radicalização da língua portuguesa. Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de Santa Catarina, Florianópolis, Brasil, 2007.

VOORHEES, E.M. "Proceedings of the 8th Text Retrieval Conference". TREC-8 Question Answering Track Report. pp. 77–82, 1999.

WANG, Y.: An Empirical Evaluation of Semantic Similarity Measures Using the WordNet and UMLS Ontologies. Master of Computer Science thesis, Miami University, Oxford, Ohio, 2005.

WORDNET. WordNet, A lexical database for English. Disponível em:< http://wordnet.princeton.edu/>. Acesso em 15 de Novembro de 2012.

WU, S.H.; TSAI, T.H.; HSU, W.L. Domain event extraction and representation with domain ontology. In: IJCAI-03 Workshop On Information Integration On The Web. Acapulco, 2003. p. 33-38.

W3C. World Wide Web Consortium. W3C Semantic Web Activity. Disponível em: <http://www.w3.org/2001/sw/>. Acesso em: 15 de Setembro de 2012.

XStream. Disponível em: <http://xstream.codehaus.org/>. Acesso em Março de 2013.

XML: eXtensible Markup Language. Disponível em:<http://www.w3.org/XML/>. Acesso em: 23 de Novembro de 2012.

XMLTV. Disponível em: < http://www.xmltv.com.br>. Acesso em 23 de Março de 2013.

ANEXO I

Tabela de metadados MPEG-2 PSI/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

ANEXO II

Tabela de metadados *TV-Anytime*

#	Metadado	Origem	Descrição
1	TimePoint	TVATimeType	Designa um ponto no tempo
2	Duration	TVATimeType	Designa um período no tempo
3	PersonName	TVAAgentType	Especifica o nome de uma pessoa
4	PersonNameIDRef	TVAAgentType	Elemento usado para indentificar um PersonName na CIT
5	OrganizationName	TVAAgentType	Especifica o nome de uma organização
6	OrganizationName IDRef	TVAAgentType	Elemento usado para indentificar um OrganizationName na CIT
7	Keyword	KeywordType	Define uma palavra-chave para o conteúdo
8	KeywordType	KeywordType	Indica a ordem de importância da palavra-chave que descreve o conteúdo (principal/secundário/outro)
9	Genre	GenreType	Define um gênero para o conteúdo

10	GenreType	GenreType	Indica a ordem de importância do gênero que descreve o conteúdo (principal/secunário/outro)
11	Synopsis	SynopsisType	Define o tamanho da sinopse (curta/média/longa)
12	SynopsisType	SynopsisType	Define a sinopse do conteúdo
13	RelatedMaterial	RelatedMaterialType	Referencia outras propriedades de mídia relacionadas ao conteúdo AV descrito
14	HowRelated	RelatedMaterialType	Natureza do relacionamento entre conteúdo e propriedade AV
15	Format	RelatedMaterialType	Define tipo do arquivo da propriedade relacionada
16	MediaLocator	RelatedMaterialType	Define localização da propriedade de mídia
17	PromotionalText	RelatedMaterialType	Fornece texto promocional sobre a ligação (usado como atrativo adicional)
18	PromotionalMedia	RelatedMaterialType	Possibilita uso de informação não-textual, como logotipos, gráficos, etc.

19	SourceMediaLocator	RelatedMaterialType	Define localização da mídia cuja descrição é associada
20	Character	CreditsItem	Especifica o nome do personagem representado pelo artista (usado em conjunto com TVAAgentType).
21	CategoryAward	AwardType	Especifica a categoria na qual o conteúdo foi premiado
22	NomineeAward	AwardType	Especifica o candidato que venceu na categoria
23	RecipientAward	AwardType	Especifica a pessoa/organização promotora do prêmio
24	TitleAward	AwardsListItemType	Nome/Título do prêmio
25	YearAward	AwardsListItemType	Ano da premiação
26	Award	AwardsListItemType	Informação detalhada do prêmio
27	ShortTitleType	ShortTitleType	Define um título de tamanho máximo de 80 caracteres
28	CreationDate	CreationCoordinatesType	Define a data de criação de um conteúdo (TVATimeType)

29	CreationLocation	CreationCoordinatesType	Define a região onde foi criado o conteúdo (‘region code’)
30	DepictedDate	DepictedCoordinatesType	Define a data retratada no conteúdo
31	DepictedLocation	DepictedCoordinatesType	Define a localização retratada no conteúdo
32	ReleasedDate	ReleasedDateType	Data (dia, mês e ano) de lançamento do conteúdo
33	ReleasedYear	ReleasedDateType	Ano de lançamento do conteúdo
34	ReleasedLocation	ReleasedInformationType	País onde o conteúdo foi lançado
35	Title	BasicContentDescriptionType	Define o título do conteúdo (*full*)
36	MediaTitle	BasicContentDescriptionType	Define uma propriedade de mídia como título do conteúdo (imagem, por exemplo)
37	TVAParentalGuidance	BasicContentDescriptionType	Contém a classificação de pais para o programa
38	Language	BasicContentDescriptionType	Define o idioma do conteúdo

39	CaptionLanguage	BasicContentDescriptionType	Define a linguagem de informação de *caption* incluída no conteúdo
40	SignLanguage	BasicContentDescriptionType	Define a linguagem de sinais incluída no conteúdo
41	BitRate	BitRateType	Indica a taxa de bits para um conteúdo
42	AudioLanguage	AudioLanguageType	Indica o idioma do áudio
43	AudioCoding	AudioAttributesType	Formato de codificação do áudio
44	NumChannels	AudioAttributesType	Indica o número de canais de áudio (Mono/Stéreo/Multi)
45	AudioMixType	AudioAttributesType	Tipo da mixagem do áudio
46	FrameRate	VideoAttributesType	Indica a taxa de quadros do vídeo
47	VideoCoding	VideoAttributesType	Formato de codificação do vídeo
48	VideoScan	VideoAttributesType	Indica o vídeo scan
49	VideoHorizontalSize	VideoAttributesType	Dimensão horizontal do vídeo (em pixels)
50	VideoVerticalSize	VideoAttributesType	Dimensão vertical do vídeo (em pixels)
51	VideoAspectRation	VideoAttributesType	Indica o vídeo aspect

52	VideoColor	VideoAttributesType	Indica o formato de cores do video (preto e branco, colorido, etc)
53	CaptioningCoding	CaptioningAttributesType	Indica o formato de *captioning*
54	FileFormat	AVAttributesType	Define o formato da instância do arquivo
55	FileSize	AVAttributesType	Define o tamanho do arquivo
56	Ratio	RationType	Especifica a taxa no formato “h:v”
57	BasicDescription	ProgramInformation/GroupInformation	Descrição básica do conteúdo/grupo de conteúdo
58	AVAttributes	ProgramInformation	Descrição dos atributos de áudio e vídeo do conteúdo
59	NumOfItens	GroupInformation	Informa a quantidade total de elementos do grupo de conteúdo
60	MediaReview	MediaReviewType	Descreve uma crítica sobre o conteúdo
61	ServiceInformationName	ServiceInformation	Informa o nome do serviço
62	ValidFrom	ServiceInformation	Data e horário a partir do qual o serviço é válido
63	ValidTo	ServiceInformation	Data e horário limite da validade do serviço
64	ServiceInformationOwner	ServiceInformation	Informa nome do produtor do conteúdo

65	ServiceGenre	ServiceInformation	Informa o gênero do serviço
66	ServiceLanguage	ServiceInformation	Informa o idioma falado no serviço
67	Logo	ServiceInformation	Logotipo da rede, como uma imagem ou *jingle*
68	ServiceRelatedMaterial	ServiceInformation	Referência a outro material relacionado ao serviço
69	PublishedStartTime	ScheduleEvent	Horário anunciado para início do programa
70	PublishedEndTime	ScheduleEvent	Horário anunciado para término do programa
71	PublishedDuration	ScheduleEvent	Duração anunciada para o programa
72	Live	ScheduleEvent	Indica que a transmissão é ao vivo
73	Repeat	ScheduleEvent	Indica que é uma repetição
74	FirstShowing	ScheduleEvent	Indica que é a primeira exibição
75	LastShowing	ScheduleEvent	Indica que é a última exibição
76	InstanceDescription	ScheduleEvent/BroadcastEvent	Descreve uma instância do conteúdo
77	ServiceIDRef	BroadcastEvent	Identifica o serviço ao qual um evento de broadcast é transmitido