CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
FUNDAÇÃO DE ENSINO "EURÍPIDES SOARES DA ROCHA"
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Agente Semântico de Extração Informacional no Contexto de Big Data

Caio Saraiva Coneglian

Marília, 2014

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
FUNDAÇÃO DE ENSINO "EURÍPIDES SOARES DA ROCHA"
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Agente Semântico de Extração Informacional no Contexto de Big Data

Monografia apresentada ao Centro Universitário Eurípides de Marília como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciência da Computação
Orientador: Prof. Dr. Elvis Fusco

Marília, 2014

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO – AVALIAÇÃO FINAL

Caio Saraiva Coneglian

Agente Semântico de Extração Informacional no Contexto de Big Data

Banca examinadora da monografia apresentada ao Curso de Bacharelado em Ciência da Computação do UNIVEM/F.E.E.S.R., para obtenção do Título de Bacharel em Ciência da Computação.

Nota: 10 (dez)

Orientador: Elvis Fusco

1°. Examinador: Leonardo Castro Botega

2°. Examinador: Ricardo José Sabatine

Marília, 02 de dezembro de 2014.

AGRADECIMENTOS

A Deus, a meu pai, José Artur, minha mãe, Ana Maria, meu irmão Fernando, também meus tios e meus fiadores durante a graduação, Domingos e Isabel.

Aos meus amigos, em especial Marcelo, Felipe, Sílvio, Lucas, Anderson, Marianas Cristina, Mariana Regina, Natalia e Adriana.

Aos todos os meus colegas de sala, em especial Victor, Lucas, Alexandre, Danilo, Maycon e Luana.

Ao meu orientador Prof. Dr. Elvis Fusco, pelo auxílio e orientação durante a produção do trabalho.

A todos os professores que lecionaram durante minha graduação em Bacharelado em Ciência da Computação.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

RI ..... Recuperação de Informação
OWL..... Web Ontology Language
TI..... Tecnologia da Informação
XML..... Extensible Markup Language
RDF..... Resource Description Framework
NoSQL ..... Not Only SQL
RDF..... Resource Description Framework

CONEGLIAN, Caio Saraiva. **Agente Semântico De Extração Informacional No Contexto De Big Data**. 2014. f. Trabalho de curso. (Bacharelado em Ciência da Computação) - Centro Universitário Eurípides de Marília, Fundação de Ensino "Eurípides Soares da Rocha", Marília, 2014.

Resumo

O grande aumento da produção e disseminação de dados na Internet pode oferecer informações de alto valor agregado às organizações. Estas informações podem estar em bases distintas e heterogêneas e em fontes que antes não eram consideradas relevantes, como mídias sociais, blogs e outros. Se as organizações conseguirem utilizar destas fontes, podem fazer com que haja uma nova visão de gestão conhecida como Inteligência Competitiva. No contexto de uma arquitetura de Recuperação da Informação, esta pesquisa tem como objetivo a implementação de um agente de extração semântica no contexto da Web que permita a localização, armazenamento, tratamento e recuperação de informações no contexto do Big Data nas mais variadas fontes informacionais na Internet que sirva de base para a implementação de ambientes informacionais que auxiliem o processo de Recuperação da Informação, utilizando de ontologia para agregar semântica ao processo de recuperação e apresentação dos resultados obtidos aos usuários, conseguindo desta forma atender suas necessidades informacionais.

Palavras-chave: Ontologia, Recuperação da Informação, Big Data, Web Semântica, Agente de Extração

CONEGLIAN, Caio Saraiva. **Agente Semântico De Extração Informacional No Contexto De Big Data**. 2014. f. Trabalho de curso. (Bacharelado em Ciência da Computação) - Centro Universitário Eurípides de Marília, Fundação de Ensino "Eurípides Soares da Rocha", Marília, 2014.

Abstract

The large increase in the production and dissemination of data on the Internet can offer information of high-earned value to organizations. This information may be on different bases and heterogeneous and supplies that were not considered relevant as social media, blogs, and more. If organizations get used these sources can make a new management vision known as Competitive Intelligence. In the context of an architecture of Information Retrieval, this research aims implementing a semantic extraction agent in the context of the Web allowing the location, storage, processing and retrieval of information like Big Data in various informational sources on the Internet serving as a base for the implementation of information environments the process of Information. Using Ontology to add semantics to the recovery process and presentation of results to the users, thus being able to meet their informational needs.

Keywords: Ontology, Information Recovery, Big Data, Semantic Web, Extraction Agent

Introdução

A explosão de geração massiva de dados está testando a capacidade das mais avançadas tecnologias de armazenamento, tratamento, transformação e análise de informações. As áreas do tratamento e da recuperação da informação estão sendo desafiadas pelo volume, variedade e velocidade de uma inundação de dados semiestruturados e não estruturados de natureza complexa, que também oferece às organizações excelentes oportunidades de terem um aprofundamento no conhecimento mais preciso de seus negócios.

Neste contexto, surgem inúmeras oportunidades em agregar valor ao negócio com base nessas informações que são geradas tanto no ambiente interno quanto no externo, porém há a necessidade de uma nova abordagem na estrutura de TI (Tecnologia da Informação) das empresas em transformar esses dados em conhecimento para as organizações, que causará impacto de longo alcance.

Para agregar e utilizar as informações que estão espalhadas nos ambientes internos e externos das organizações, surge o conceito da Inteligência Competitiva que segundo ABRAIC (Associação Brasileira dos Analistas de Inteligência Competitiva), é um processo informacional proativo que conduz à melhor tomada de decisão, seja ela estratégica ou operacional, visando descobrir as forças que regem os negócios, reduzir o risco e conduzir o tomador de decisão a agir antecipadamente, bem como proteger o conhecimento gerado (BRASILIANO, 2002).

No cenário atual destas informações geradas nos ambientes organizacionais, principalmente nos que tem a Internet como plataforma, encontram-se dados que devido às suas características, atualmente classificam-se como Big Data.

Dentre estas características destacam-se: volume - enormes conjuntos de dados que são de magnitude maior do que os dados mantidos em sistemas de armazenamento tradicional; variedade - dados heterogêneos, complexos e variáveis que são gerados em formatos diversos que tem como fonte: e-mails, mídias sociais, vídeos, imagens, blogs e bases da web; velocidade - os dados são gerados em fluxo constante com consultas em tempo real de informações significativas para tomada de decisão; valor - esses dados são potenciais para geração de conhecimentos significativos que oferecem análises preditivas para futuras tendências e padrões, que vão além dos resultados tradicionais de consultas e relatórios de sistemas de informação transacionais.

Na publicação do Journal of Science (GRAHAN-ROWE, 2008) Big Data é

definido como a representação do andamento dos processos cognitivos humanos, que geralmente inclui conjuntos de dados com tamanhos além da capacidade da tecnologia atual, métodos e teorias para capturar, gerenciar e processar os dados dentro de um tempo determinado. Beyer e Laney (2012) define Big Data como o alto volume, alta velocidade e/ou alta variedade de informações que requerem novas formas de processamento para permitir melhor tomada de decisão, nova descoberta do conhecimento e otimização de processos.

Nos ambientes de Big Data apenas o uso de bancos de dados relacionais não é adequado para a persistência, processamento e recuperação dos dados em ambientes escaláveis e heterogêneos. Para tentar resolver esta questão no âmbito da persistência da informação surgem novos conceitos nas tecnologias de banco de dados, como o NoSQL (Not Only SQL) que para De Diana e Gerosa (2010) veio representar soluções alternativas ao modelo relacional, oferecendo maior escalabilidade e velocidade no armazenamento dos dados surgindo como uma opção mais eficaz e barata.

O uso de conceitos de Business Intelligence e Inteligência Competitiva e tecnologias como Data Warehouse, OLAP, Analytics, Datamining, NoSQL e robôs de busca semântica representam abordagens para capturar, gerenciar e analisar cenários de Big Data. A necessidade da utilização dessas tecnologias no tratamento desses dados massivos e complexos estão causando uma mudança de paradigma que está levando as organizações a reexaminar sua infraestrutura de TI e sua capacidade de análise e gestão corporativa da informação.

A gestão eficaz e a análise de dados em larga escala representam um interessante, mas crítico desafio, pois os modelos de gestão baseados na Inteligência Competitiva estão sendo influenciados por esse universo complexo de informações geradas com o conceito de Big Data e novas investigações são necessárias para dar solução a esse desafio de uso eficiente das informações no processo de gestão.

No processo de busca da informação em cenários da Inteligência Competitiva e Big Data são utilizados robôs de extração de dados na Internet, que segundo Deters e Adaime (2003) são sistemas que coletam os dados da Web e montam uma base de dados que é processada para aumentar a rapidez na recuperação de informação e que segundo Silva (2003), a extração de informações relevantes pode classificar uma página segundo um contexto de domínio e também retirar informações estruturando-as e armazenando-as em bases de dados.

Com o propósito de adicionar significado aos conteúdos buscados em domínio

específico associam-se aos robôs de busca na Web conceitos semânticos, que permitem realizar a procura não mais por palavras chaves num processo de busca textual, mas sim por significado e valor, extraindo das páginas e serviços da Web informações de real relevância, descartando aquilo que é desnecessário. A partir disto, a ontologia aparece como solução na busca de inserir semântica neste processo.

A ontologia, no contexto filosófico, é definida por Silva (2003) como a parte da ciência que estuda o ser e seus relacionamentos e neste sentido, o uso de ontologias é essencial no processo de desenvolvimento dos robôs de busca semântica, sendo aplicada na Ciência da Computação e na Ciência da Informação para possibilitar uma busca de maneira mais inteligente e mais próxima do funcionamento do processo cognitivo do usuário de forma que a extração de dados se torne muito mais relevante.

Atualmente vivencia-se uma nova disrupção tecnológica pela convergência da colaboração, mobilidade e grande volume de dados (Big Data). O grande desafio para a pesquisa de sistemas computacionais e para a forma de uso das informações nas organizações está em promover a integração destas tecnologias para balancear as necessidades de geração, acesso e controle destas informações, bem como as oportunidades deste comportamento emergente e suas inovações.

Objetivos

Esta pesquisa tem como objetivo criar uma plataforma semântica de Recuperação de Informação na Web que permita a localização, armazenamento, tratamento e recuperação de informações inseridos em um contexto de Big Data, nas mais variadas fontes informacionais na Internet que sirvam de base para uma arquitetura computacional que transforme a informação desagregada em um ambiente de conhecimento estratégico, relevante, preciso e utilizável para permitir aos usuários o acesso as informações com maior valor agregado, que consiga satisfazer as necessidades informacionais do usuário, aderindo uma semântica ao processo de Recuperação da Informação.

Tem como objetivos específicos:

- Definir uma estrutura ontológica de representação do domínio de instituições de ensino superior;
- Projetar uma estrutura de representação informacional conceitual, lógica e de persistência do domínio utilizando tecnologias relacionais e NoSQL;
- Criar um robô de busca semântica na Web baseado na estrutura semântica e nas fontes informacionais do domínio;
- Desenvolver um protótipo de validação do agente computacional que implemente um ambiente informacional de processamento, fusão, recuperação e representação visual e interativa da informação, de forma a promover um raciocínio analítico, preditivo e prescritivo, visando contemplar processos analíticos e de tomada de decisão de instituições de ensino superior. Este protótipo possibilitará a análise dos resultados de extração semântica do agente proposto.

Metodologia

O projeto foi dividido em três partes principais:

- Levantamento bibliográfico e pesquisa de trabalhos correlatos e tecnologias: Foi realizada a busca bibliográfica sobre temas como: ontologia, Big Data, Inteligência Competitiva, Recuperação da Informação, robô de busca, entre outros. Também foi procurado tecnologias trabalhos correlatos e tecnologias relacionadas com as utilizadas durante a construção do projeto.
- Construção da ontologia: A etapa da construção da ontologia se baseia na definição da estrutura ontológica, e do estudo acerca de como se relaciona os objetos dentro desta ontologia.
- Implementação do robô de busca e integração entre a ontologia e o robô de busca: Durante a implementação do robô, é realizado a análise de como é feita a extração dos dados das páginas HTML, e posteriormente, como que as informações extraídas passarão pelo processo de validação e verificação através do uso da ontologia.

Trabalhos Correlatos

Arquiteturas de Recuperação de Informação com o uso de agentes foram propostos por outras pesquisas, onde realizam a extração da informação para o uso posterior em algum cenário.

Desta forma Beppler (2005), propôs uma Arquitetura de Recuperação de Informação. Esta recuperação ocorre apenas com a análise de documentos e armazenamento de informação, sem observar o contexto existente, sendo que esta análise ocorre de forma sintáticas. Esta proposta é interessante pois é possível extrair informação de uma maneira eficiente, mas é limitado, pois a busca é sintática, diminuindo assim, a eficiência desta arquitetura.

Já Wisner (2008) propôs uma solução semântica para este problema. Esta proposta reúne uma arquitetura que usa uma solução onde a semântica ocorre através do uso de ontologias para ter uma base de integração de conhecimento, utilizando um agente que realiza associações e integrações do conhecimento. Esta pesquisa pode realizar boas associações para cada tipo de conhecimento, mas a semântica é limitada porque apenas faz associações de informações, não tratando como deve ser representada e apresentada as informações ao usuário, sendo desta maneira uma pesquisa que consegue realizar parte do processo, mas não as aplica de fato na representação da informação, outra questão, é se de fato aquelas informações terão real valor para um domínio particular.

1. Recuperação de Informação

A recuperação da informação tem se tornado alvo de muitos estudos, devido à grande quantidade de informações que hoje se encontram espalhados pela rede.

A recuperação da informação lida com a representação, armazenamento, organização e acesso as informações. Devendo prover ao usuário aquilo que ele necessita de uma maneira facilitada (BAENZA-YATES E RIBEIRO-NETO, 1999).

O conceito de recuperação de informação é diferente de recuperação de dados. A recuperação de dados consiste em extrair de um banco de dados qualquer documento que contém uma expressão regular ou os termos ali contidos. Sendo que a recuperação da informação vai além, levando em conta a sintaxe e a semântica daquela informação, buscando satisfazer o que o usuário está pesquisando (BAENZA-YATES E RIBEIRO-NETO, 1999).

Desta maneira a recuperação da informação tem assumido um papel diferenciado na Ciência da Informação e na Ciência da Computação, pois aparece como elo final na busca pela apresentação da informação mais adequada ao usuário no menor tempo possível.

O processo de recuperação da informação não consiste apenas em técnicas e métodos que envolvem o armazenamento e os algoritmos de recuperação, mas também em adaptar os sistemas no comportamento do usuário, entendendo desta maneira, como é a construção da informação e das instruções para a recuperação da informação (SANTAREM SEGUNDO, 2010).

Com o surgimento da Web houve grande aumento no volume das informações eletrônicas, que trouxeram muitas vantagens quanto à possibilidade de troca, difusão e transferência de dados. Entretanto, este crescimento trouxe muitos problemas relacionados ao acesso, busca e recuperação das informações de real valor imerso em grandes volumes de dados (MODESTO, 2013).

Assim, um dos desafios da recuperação da informação é conseguir fazer com os Ambientes Informacionais Digitais entendam o que o usuário está necessitando, de forma que os resultados vindos da busca possam ser de real valor e importância para o usuário.

1.1. Definição

O termo Recuperação da Informação foi trazido pela primeira vez em 1951, por

Mooers (1951), quando definiu os problemas que seriam tratados por esta nova disciplina. Desta maneira a Recuperação da Informação trata dos aspectos da descrição e especificação das buscas da informação. Tratando também de qualquer sistema, técnicas e máquinas utilizadas no processo de recuperação da informação.

Desta maneira o processo de Recuperação da Informação, consiste em encontrar em um conjunto de documentos de um sistema, quais são os que atendem às necessidades informacionais do usuário. Assim, o usuário não está interessado em recuperar dados, nem achar documentos que satisfaçam sua expressão de busca, e sim em encontrar a informação sobre um determinado assunto (FERNEDA, 2003).

Assim os sistemas de Recuperação de Informação devem representar os documentos e apresenta-los aos usuários de maneira que, o usuário através daqueles documentos recuperados consigam satisfazer total ou parcialmente as suas necessidades informacionais (FERNEDA, 2003).

1.2. Modelos de Recuperação da Informação

O principal desafio durante o processo da Recuperação da Informação é conseguir atender as necessidades dos usuários de forma que, consiga-se atender exatamente aquilo que ele busca. Isto se torna muito complexo, pela tarefa do computador ter uma linguagem diferente daquela que o usuário possui, de forma que o usuário precisa passar aquilo que ele necessita, e o computador necessita entender isto.

Desta maneira, vários autores sugeriram muitos modelos para a realização da recuperação da informação. Abaixo os modelos foram divididos em uma classificação básica, a de modelos quantitativos e de modelos dinâmicos.

1.2.1. Modelos Quantitativos

Os modelos quantitativos são modelos construídos em cima de conceitos de lógica, estatística e teoria dos conjuntos. Sendo que estes modelos foram construídos nas décadas de 60 e 70, mas até hoje, estão presentes na maioria dos sistemas de recuperação de informação.

Neste tipo de modelo, os documentos são representados por um conjunto de

termos de indexação. Um termo de indexação representa um significado ou um conceito de um documento. A questão gira em torno de qual será a representatividade destes termos para aquele documento, ou seja, se aquele termo, de fato representará aquele documento e conseguirá dar real representatividade àquele documento. Portanto decidir qual termo será utilizado como índice, não é fácil, pois deve levar em consideração diversos aspectos. Desta maneira, cada termo de indexação possui diversos graus de relevância, de acordo com os documentos e os sistemas de informação (FERNEDA, 2003).

A seguir é relatado alguns dos modelos mais utilizados dentro dos métodos quantitativos.

1.2.1.1. Modelo booleano

O modelo booleano se baseia na lógica como base. A álgebra booleana é um sistema binário, onde os dados podem assumir somente dois estados, 0 ou 1, falso ou verdadeiro.

O modelo booleano se encontra em quase todos os sistemas de buscas de informação, pois este é a principal maneira de realizar expressões de busca. E quando apresenta uma quantidade muito grande ou muito pequena de dados, é possível ir aumentando ou diminuindo o número de documentos, até atingir a quantidade desejada.

Porém o modelo booleano apresenta a desvantagem de não conseguir ordenar os documentos resultantes de uma busca. E hoje esse modelo não seria o mais adequado para os modernos sistemas de busca de texto integral, como motores de buscas da Web, onde o ordenamento dos documentos é fundamental frente a enorme quantidade de dados que são recuperados (FERNEDA, 2003).

1.2.1.2. Modelo Vetorial

O modelo vetorial foi criado por Salton (1988) em 1968, motivado pelas limitações que apresentavam o modelo booleano.

Segundo Santarem Segundo (2010, p. 32 e 33)

> "[...] Esse modelo tem como premissa considerar a similaridade parcial entre os termos, representando-os através de um vetor numérico, onde cada elemento do vetor representa

> um termo de consulta e a este e atribuído um peso que indica tamanho e direção do vetor de representação. São esses pesos que possibilitam a proximidade de consulta e o cálculo da similaridade parcial entre os termos da consulta e os documentos, possibilitando que os resultados sejam grau de similaridade entre o termo na expressão de busca e o documento recuperado. O cálculo de proximidade entre os vetores e realizado de acordo com o angulo do vetor, e dessa forma e calculado o grau de similaridade. [...]
> [...] No modelo vetorial, a consulta e realizada em busca dos termos designados, e a classificação apresentada como resultado baseia-se na frequência dos termos no documento em relação ao peso atribuído a cada termo, utilizando-se o grau de similaridade calculado. [...]"

Desta maneira, o modelo vetorial vai utilizar pesos tanto para os termos de indexação quanto para os termos de expressão de busca, conseguindo desta maneira ter um valor que representa a relevância de um documento perante a expressão de busca (FERNEDA, 2003).

O modelo vetorial tem como desvantagem o não uso de expressões booleanas, que em alguns momentos podem ter uma grande valia, além disso, este modelo se caracteriza por aproximar muito as combinações, podendo encontrar relações, que não tenham de fato relação.

1.2.1.3. Modelo Probabilístico

A teoria probabilística dentro da matemática, efetua o cálculo da chance de ocorrência de um número em um determinado experimento aleatório. Por exemplo um sorteio de uma loteria ou um lançamento de um dado.

O Modelo Probabilístico traz a classificação de documentos pela probabilidade em relação aos termos aplicados na busca, verificando a relação de relevância da expressão de busca para cada documento, para assim investigar a probabilidade de relevância entre eles, supondo que exista um conjunto ideal de documentos que atenda as consultas realizadas (SANTAREM SEGUNDO, 2010).

E conforme forem sendo feitas buscas, o usuário dá um feedback, para que este sistema possa ser aperfeiçoado, e consiga desta maneira determinar quais são os documentos mais relevantes deste conjunto (SANTAREM SEGUNDO, 2010).

1.2.2. Modelos Dinâmicos

Os modelos dinâmicos para recuperação da informação surgiram a partir do momento que os modelos quantitativos apresentaram um certo esgotamento quanto as funções e fórmulas matemáticas, além de que os modelos quantitativos acabam não tendo uma participação efetiva do usuário na representação dos documentos.

Bentlet (2002) relata alguns modelos computacionais que se baseiam em processos biológicos, como neurônios e a genética.

1.2.2.1. Redes Neurais

O cérebro humano é composto por uma quantidade muito grande de neurônios. O processamento paralelo e distribuído das redes de neurônios são os responsáveis pelo ser humano ter a capacidade de aprender.

As redes neurais artificiais é uma maneira de realização de processamento de informações, onde busca-se implementar modelos matemáticos que simulem o funcionamento do cérebro humano, onde os neurônios realizam ligações com outros neurônios que simulam as ligações sinápticas (SANTAREM SEGUNDO, 2010).

Dentro de Sistemas de Recuperação de Informação, as redes neurais artificiais se destacam pela capacidade que estes sistemas têm de aprender com as características do usuário, e assim utilizar este aprendizado para conseguir oferecer ao usuário resultados que tenham maior relação com o que aquele usuário necessita (FERNEDA, 2003).

1.2.2.2. Algoritmos Genéticos

Os algoritmos genéticos têm como base a genética, que afirma que todo ser herda características de seu pai e sua mãe, sendo que pode herdar mais ou menos características de seus genitores.

Este modelo é interessante no uso da Recuperação da Informação, pois interage diretamente com o usuário, de modo que o comportamento do usuário irá influir diretamente nas próximas buscas realizadas.

A cada iteração (geração) que existe no sistema, um novo conjunto de estruturas são criadas, que utilizam as informações provenientes das gerações anteriores, e esses

conjuntos irão se adaptando ao ambiente, até um ponto que as estruturas criadas estão muito próximo de uma solução ótima (FERNDEDA, 2009)

Ferneda (2009) afirma que os algoritmos genéticos aplicados aos sistemas de Recuperação da Informação representam uma nova maneira de ver este processo, pois a representação dos documentos, será alterado conforme o que o sistema for aprendendo com o usuário.

1.3. Recuperação da Informação na WEB

Com o grande aumento na Web, ultimamente o foco de pesquisas relacionadas a Recuperação da Informação tem sido como conseguir recuperar os dados da Web.

O grande desafio da recuperação da informação na Web é o fato que esta foi construída de maneira descentralizada, de forma que muitas estratégias de buscas citadas a cima, não conseguem ter um bom funcionamento.

Segundo Santarem Segundo (2010, p. 39)

> "[...] Dentro de uma nova dimensão como a Internet, fica visível o esgotamento de alternativas com relação a esses modelos já conhecidos, visto que existe uma clara mudança do corpus de consulta. Com a introdução da Internet no contexto do usuário, passa-se a ter um depósito de informações muito mais amplo, que carrega consigo a ligação de documentos e informações através de links, criando uma interligação entre os documentos armazenados e disponíveis na rede[...]".

Um dos métodos mais utilizados ultimamente para realizar a busca da informação na Web, é o método Page Ranking. Este método foi proposto pelo Google, e funciona de maneira que verifica-se a importância de um site, através da quantidade de vezes que este site é citado por outros, ou seja, quanto mais vezes aparecer o link de uma página em outras páginas, indicam go grau de importância. De forma que os mecanismos de busca indexam, e ordenam os sites pela sua importância, que é definida pelo algoritmo de Page Ranking (SANTAREM SEGUNDO, 2010).

Verifica-se portanto a necessidade de buscar novas maneiras de realizar a recuperação da informação, neste novo ambiente, chamado de Web, onde as informações são dos mais variáveis tipos, onde os motores de busca, apresentam uma quantidade muito grande de links e páginas para que o usuário possa encontrar o que atende a sua necessidade.

No terceiro capítulo será abordado o tema da ontologia, onde neste trabalho, faz-se uso de ontologias para poder aprimorar o processo de Recuperação da Informação neste ambiente da Web.

1.4. Recuperação da Informação em Big Data

Com o crescimento exponencial das informações contidas dentro da Web, o processo da Recuperação de Informação se depara com um novo desafio: como conseguir recuperar informações de forma eficiente e desta maneira resgatar as informações que apresentam real valor que estão imersos a tantos outros dados. Para entender este processo, é necessário visualizar a questão do Big Data, e como este processo está mudando a maneira como se vê as informações dentro da Web.

2. Big Data

Este capítulo tratará conceitos relacionados à Big Data, e como este se organiza e pode ser utilizado em pesquisas e em empresas.

2.1. Definições

Atualmente vive-se um momento de grande geração e uso das informações geradas de forma online. Esses dados são gerados por e-mails, compartilhamento de informações por redes sociais, transações online, celulares, GPS, entre vários outros meios.

Para verificar este movimento, foi verificado que até 2003 toda a humanidade criou cerca de 5 exabytes de informações, atualmente esta quantidade de informação é gerada em menos de dois dias (SAGIROGLU E SINANC, 2013).

Zikopoulos e Eaton (2011) define de forma sintática que Big Data se aplica a informações que não podem ser processadas ou analisadas com as ferramentas e os métodos tradicionais. E diz ainda que a era do Big Data é resultado das mudanças que tem ocorrido no mundo, onde através dos avanços das tecnologias, foi possível que várias pessoas e programas se intercomunicasse não somente num intervalo de tempo, mas sim durante todo o tempo.

O termo Big Data não diz respeito somente ao aspecto de armazenamento de dado, mas também a outros aspectos como a velocidade em que os dados podem ser capturados e processados, quase que em tempo real, dando assim, vantagens competitivas as organizações (MCAFEE, 2012). Na figura 3 é ilustrada a relação entre os três aspectos que McAfee considera principais neste processo que são o volume, a velocidade e a variedade:

- Volume: O número de dados gerados todos dias na web ultrapassam 2,5 exabytes, e este número tem dobrado a cada quarenta meses, em breve a unidade de medida de dados será o zetabytes. Algo que faz com que este volume cresça de forma exponencial, é que a própria Web fornece possibilidades para uma criação de novas informações, como as redes sociais, onde o usuário acaba gerando cada vez mais dados (ZIKOPOULOS E EATON, 2011). Todas as informações geradas são armazenadas, como dados financeiros, médicos, compras realizadas na internet, conversas realizados pelos mensageiros, gerando assim um volume extremamente grande de dados.

- Velocidade: Muitas vezes a velocidade com que o dado é criado e processado é fundamental, pois estes dados podem ser utilizados em tempo real. Esta velocidade não está ligada somente a entrada de dados, mas também a velocidade do fluxo de dados. Ou seja, ter velocidade em conseguir acompanhar a geração e a demanda das requisições das informações.
- Variedade: as formas que os dados estão armazenados são cada vez mais diversas, entre elas textos, músicas, vídeos, imagens. E isto promove que não exista um padrão com que os dados são gerados e/ou armazenados. Segundo Kakhani (2013) os dados podem ser não-estruturados, semiestruturados ou estruturados, sendo de uma natureza heterogênea, pois os dados podem vir de mídias sociais, de blogs, sendo desde textos não-estruturados, à vídeos e fotos, não tendo, assim, uma estrutura fixa e definida.

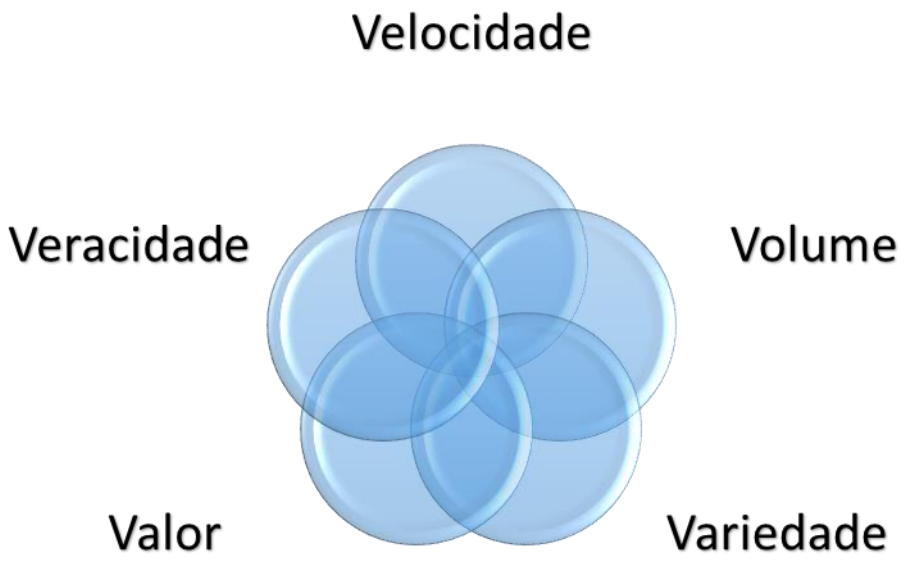

Figura 1: 5 V's do Big Data

Posteriormente a McAfee, alguns autores (KAKHANI, 2013) (KATAL, 2013) também incluíram outros dois conceitos que tem importância, para a definição de Big Data, que é a veracidade e o valor:

- Veracidade: todos os dados presentes neste universo, podem ser das mais diversas naturezas, portanto é necessário que se tenha dados que sejam verdadeiros, para não trazer informações equivocadas, ao final de um estudo (KAKHANI, 2013).

- Valor: a partir de dados e informações que foram fornecidas ou adquiridas pelos sistemas, pode se chegar a resultados de muito valor, pois podem demostrar tendências do mercado, que pode levar aos administradores das empresas a tomarem medidas para mudar ou readequar as estratégias comerciais (KATAL, 2013).

O processo do Big Data aparece pelo grande crescimento do uso e da geração da informação, onde a mudança quantitativa (grande crescimento de dados) trouxe uma mudança qualitativa das informações (informações cada vez mais precisas) (MAYER-SCHÖNBERGER E CUKIER, 2013).

Esta mudança quantitativa, significa, fazer as análises das informações levando em consideração todo o banco de dados existente. Pois antes do atual momento da tecnologia, as análises realizadas, eram feitos quase que totalmente em cima de amostras, o que pode trazer um grande risco, de se ter informações que não são reais, por maior o cuidado que exista para esta amostra representar uma população real, isto é um processo que pode ter falhas.

Esta técnica de amostragem, é fruto de um momento, onde não existia tecnologias para conseguir processar e analisar todas as informações existentes. Hoje este conceito já não faz tanto sentido, pois atualmente é possível processar quantidades enormes de informações em segundos, podendo assim, usar todas as informações do banco de dados, não necessitando pegar uma amostra deste.

Como consequência disto, os resultados destas analises, que agora utiliza o todo, passaram a ser muito mais precisas, e oferecerem dados e informações que antes seria muito difícil de obter.

Verificando todo este movimento, algumas ciências, como a astronomia e a genômica, passaram a utilizar disto e deram o nome de Big Data para este processo.

O uso do Big Data pode ser visualizado, quando necessita-se trabalhar com grandes escalas de dados para se extrair novas ideias e criar novas formas de valor que alterem mercados, governos, organizações, entre outros (MAYER-SCHÖNBERGER E CUKIER, 2013).

Este processo pode ser percebido nas grandes corporações da internet, como Google que processa mais de 24 pentabytes de dados por dia, e o Facebook que recebe mais

de dez milhões de fotos a cada hora (MAYER-SCHÖNBERGER E CUKIER, 2013).

2.2. Armazenamento das Informações

A quantidade de informações que são gerados tem sido um grande desafio, pois cada vez mais o número de dados crescem e as médias de informações armazenadas também aumentam.

O armazenamento das informações geradas é um grande desafio, pois atualmente um disco consegue armazenar por volta de poucos terabytes. E os números da web giram em torno de exabytes, ou seja, necessita-se de muitos discos para conseguir fazer este armazenamento. Na tabela 1 é possível verificar os números da web atualmente (KAISLER, 2003).

Tabela 1: Quantidade de Dados Gerais (Kaisler, 2003)

Domínio/ Conjunto de Dados	**Descrição**
Grande Colisor de Hádrons - CERN	13-15 petabytes em 2010
Internet Communications (Cisco)	667 exabytes em 2013
Mídias Sociais	12+ Tbytes de tweets todos os dias. Média de retweets são de 144 por tweet.
Human Digital Universe	1.7 Zbytes (2011) -> 7.9 Zbytes em 2015
British Library UK Website Crawl	~ 110 TBytes por domínio de rastreamento à ser arquivado
Outros	RFIDS, medidores elétricos inteligentes, 4.6 bilhões de câmeras de celular com GPS

2.3. Valor dos Dados

Mayer-Schönberguer e Cukier (2013) diz que antigamente, os dados eram utilizados como subprodutos das vendas, e não como produto propriamente dito. Por mais que sempre essas informações foram valorizadas, nunca isto aconteceu como agora na época do

Big Data, onde os dados viraram o produto, onde as empresas perceberam que dados antes desprezados podem ter um valor muito grande, como por exemplo, as buscas realizadas em um motor de busca, os caminhos indicados pelo GPS, ou quais foram os produtos pesquisados antes do consumidor fechar uma compra.

Todos esses dados, a partir de um momento podem ser reaproveitados para publicidade, ou para sugerir uma busca mais adequada àquele usuário, e assim ter um reaproveitamento das informações, tendo um alto valor agregado.

Este fenômeno também é causado pelo fato de que antes, não era possível coletar, armazenar e analisar tais dados, e hoje não existem mais essas limitações para fazer isto. Sendo que é possível captar quantidades enormes de informações e as armazenar de uma maneira barata, pois o discos de armazenamento, hoje tem um custo muito menor do que anteriormente. Também é possível registrar uma quantidade muito grande de informações, como em um site de vendas, que consegue guardar cada clique dos usuários para oferecer os produtos mais adequados para aquele cliente e uma fábrica que consegue controlar tudo o que está acontecendo dentro de sua linha de produção.

Neste contexto, muito além de se utilizar os dados apenas como o valor apresentado naquele momento, os dados têm um valor que aparece de maneira oculta, e pode ser utilizado de forma que não tem um relacionamento direto com as informações que aquele dado está apresentando. Um exemplo disto é saber a incidência de doenças apenas pelas buscas realizadas em um motor de buscas.

Neste sentido, o valor que os dados podem ter, é muito grande, e Mayer-Schönberguer e Cukier (2013) define três modos principais de se extrair os valores dos dados: a reutilização básica, a fusão de banco de dados e a utilização de um mesmo dado em diversos cenários.

2.3.1. Reutilização Básica

Quando se analisa as informações a um primeiro momento, apenas analisa-se os dados de uma maneira única, sem levar em questão o que aqueles dados estão mostrando e o que pode-se concluir levando em consideração alguns aspectos.

Desta maneira, algumas empresas, perceberam que muitos destes dados se agregadas com outras informações, ou se reutilizarem estes dados em outro momento, para outros fins, existe então, uma fonte de valor imensurável, pois, é possível determinar

comportamentos e tendências de consumidores e mercados, que sem fazer esta análise, é muito impreciso (MAYER-SCHÖNBERGER E CUKIER, 2013).

Exemplos disto, são os principais motores de buscas, que utilizam das pesquisas realizadas pelos usuários, para traçar um perfil destes, e conseguir assim, oferecer propagandas e publicidades que tenham uma relação maior com este usuário, além de utilizar informações de outros usuários, para conseguir ter um melhor perfil acerca de um grupo de pessoas, e trazer melhores resultados de pesquisa para um usuário pertencente àquele grupo.

Outro exemplo, são as telefonias, que tem informações de grande valor, ao saber o local que seus clientes estão usando os seus serviços, o deslocamento destes clientes, e várias outras informações. E estas empresas neste momento, tem buscado maneiras de ganhar dinheiro em cima deste negócio, pois estas informações podem ser de grande valor, por exemplo, para uma empresa de outdoor, que deseja saber o fluxo de pessoas que passam por determinada rodovia, e as telefonias conseguirão fornecer informações sobre isto utilizando o deslocamento realizado por seus clientes.

2.3.2. Fusão de Bancos de Dados

Quando se realiza a fusão de dois ou mais banco de dados, é possível que consiga-se chegar a conclusões acerca de padrões e conseguir concluir se existe relação entre dois comportamentos, ou dois fatos distintos, e conseguir assim chegar a informações como por exemplo, se o uso de tal aparelho aumenta ou não a probabilidade de se desenvolver alguma doença.

Isto só é possível pois ao unir vários bancos de dados distintos, é possível analisar todas as informações inter-relacionadas. Antigamente era muito utilizado o esquema de amostras para conseguir realizar tais pesquisas, pois era inviável analisar todas as informações existentes. Mas na era do Big Data, isto é possível, e é muito mais adequado, pois utilizando como amostra o total dos dados existentes, as conclusões resultantes destas analises, são muito mais precisas e relatam informações que antes não era possível concluir (MAYER-SCHÖNBERGER E CUKIER, 2013).

2.3.3. Utilização de um dado em diversos cenários

Uma forma de conseguir reutilizar os dados, é fazer com que os dados sejam coletados já pensando na utilização destes para outras funções, ou seja, é realizar mecanismos que no momento da extração dos dados, consiga-se retirar ou utilizar os dados, de uma maneira que estes possam ser uteis para outras necessidades.

Um exemplo disto, seria de varejistas, onde muitos tem utilizados as câmeras de vídeo, além de fazer a segurança, ou seja, para verificar se alguém levou algum produto de maneira irregular, mas também para verificar o movimento de pessoas na loja, e os momentos de maiores fluxos no dia, ou na semana (MAYER-SCHÖNBERGER E CUKIER, 2013).

2.4. Aplicações do Uso de Big Data

Existem diversas possibilidades de se usar as informações provenientes do Big Data. Como em redes sociais, armazenamento de logs em sistemas de informação, analises de riscos, entre outros.

Katal (2013) traz algumas destas aplicações como:

- Armazenamento de Logs em indústrias de TI: as indústrias de TI, armazenam logs de erros e avisos de seus produtos, para conseguir tratar e consertar isto. Mas estes logs são em grande quantidade, trazendo grandes problemas para o armazenamento. A análise desses dados é de grande importância, para conseguir descobrir pontos de falhas, além de aumentar a longevidade das informações extraídas destes dados.
- Dados de sensores: a grande quantidade de informação resultante dos sensores, também é um grande problema para o Big Data. Pois estas informações são muito grandes, e apenas uma parte delas são utilizadas. Desta maneira, deve-se utilizar esta grande quantidade de dados, buscando encontrar maneiras de trata-las de uma forma a trazer lucros, e que as análises resultantes tenham um valor agregado alto.
- Analises de Riscos: é algo importante, por exemplo, para instituições financeiras, para que elas possam modelar os dados de maneira a deixar os riscos a níveis aceitáveis. E uma grande quantidade de dados consegue determinar os padrões de riscos com mais precisão.

- Mídias Sociais: uma grande parte do uso do Big Data é voltado para as mídias sociais, como quais são os sentimentos dos clientes pelos produtos das empresas. Portanto estar atento, ao que os clientes estão falando a respeito das empresas, é uma informação muito importante, podendo modificar decisões e estratégias.

2.5. Inteligência Competitiva

Saber tomar as decisões corretas, em cima de bases e números e análises realizadas, é fundamental para a manutenção e o desenvolvimento de uma instituição. Pois as empresas, necessitam estar pautada em cima de dados realmente confiáveis, e que agregaram valor a organização.

Assim Inteligência Competitiva (IC) é definida como o processo de saber o que o seu concorrente está fazendo e ficar um passo à frente dele. Adquirindo informação sobre os concorrentes, e aplicando assim, estas informações para o planejamento estratégico (TEO E CHOO, 2001).

Outra definição é a de Prescott (1995), dizendo que a Inteligência Competitiva é o processo de desenvolvimento de uma previsão a partir de questões da própria empresa, como o crescimento da mesma, dos fornecedores, dos clientes, dos possíveis competidores e dos fatores fora do mercado, como regulamentos governamentais, taxas, juros. E se todos estes fatores, serem bem estudados podem ser utilizados para dar vantagens competitivas, a quem utiliza-las.

Desta maneira, Prescott, diz que o domínio da IC, é muito amplo, afirmando que o movimento da inteligência competitiva, observa além da varredura tradicional da empresa, e de pesquisa de mercado, todos os aspectos do ambiente da empresa (competitivos, tecnológicos, políticos, econômicos e sociais) e em vários níveis da empresa (à distância, na indústria e operacional). Sendo que a IC delineia entre a informação e a sua análise, a fim de que produza inteligência, enfatizando assim, a importância da inteligência no processo decisório.

Neste sentido, é importante ter a clareza que a IC, não é espionagem, e uma das bases da IC, diz que, 90% das informações que uma empresa necessita para fazer decisões mais críticas e entender o mercado, são públicas, armazenados em dados públicos (TEO E

CHOO, 2001).

A IC, baseia-se em três princípios, que são a classificação e o armazenamento das informações, a análise e interpretação dos dados e a disseminação da informação. Sendo que a inteligência dará as empresas uma vantagem competitiva pois irá fornecer bases para que as empresas sejam capazes de conhecer melhor seu concorrente, e ser capaz assim de ter um planejamento muito mais adequado (TEO E CHOO, 2001).

2.5.1. Objetivos da Inteligência Competitiva

Conforme as empresas vão crescendo, e vão atingindo um certo grau de maturidade, é necessário que as empresas, apresentem um processo decisório regular e previsível, baseado no histórico das decisões tomadas anteriormente. E conforme as empresas tem esta maturidade, vai existindo assim uma convicção a respeito do ambiente competitivo, tendo um processo decisório bem definido, desta maneira, pode acontecer dos gestores, reduzirem a importância de certos aspectos do ambiente.

E assim, ocorre os chamados pontos cegos, pois existe uma diferença entre como os gestores estão encarando o ambiente, e como o fato está ocorrendo. Normalmente, isto é mais intenso em empresas de baixo nível de maturidade, mas pode ocorrer com empresas já consolidadas, e bastante maduras, pois os gestores podem ter um nível muito elevado de certeza, devido à convicção que eles têm a respeito do ambiente (CASTRO E ABREU, 2006).

Assim a IC, tem o papel de evitar com que as empresas criem estes pontos cegos, ou seja, evitar com que ocorra uma supremacia da convicção destes gestores, a ponto de não levar em consideração o que está ocorrendo no ambiente (CASTRO E ABREU, 2006).

2.6. Semântica no Big Data

Para conseguir extrair todo o potencial que este movimento do Big Data consegue oferecer, é necessário organizar o conteúdo dentro da Web, de uma forma que os dados ali contidos apresentem uma semântica em sua estrutura. Desta forma, ontologias se apresentem como uma solução para este problema, pois conseguem contextualizar as informações que se relacionam com esta.

3. Ontologia e Recuperação Semântica

Neste capítulo serão descritos conceitualmente ontologias e a relação entre esta e a computação. Também serão mostrados conceitos de Web Semântica.

3.1. Web Semântica

Em 1989 Tim Berners Lee propões a criação da Web, idealizando também posteriormente a criação da primeira versão do "*HyperText Markup Language*" (HTML), que é a linguagem de formatação de documentos de links de hipertexto, que se tornou o formato básico para a publicação dentro da Web (W3C, 2014) (BERNERS-LEE, 1989).

A partir de então, a Web passa a crescer e disponibilizar informações dos mais variados tipos, sendo estas informações preparadas principalmente para a leitura humana (BERNERS-LEE, 1989)

Desta forma, embora a Web tenha sido criada com o intuito de possibilitar o fácil acesso, intercâmbio e a recuperação da informação, em seu início foi gerada de uma maneira totalmente descentralizada e acabou crescendo de uma maneira exponencial. Sendo que hoje existe uma grande quantidade de informações, mas quando há a necessidade de recuperar algo, os resultados obtidos são poucos satisfatórios (SOUZA E ALVARENGA, 2004).

De forma contrária a isto, a Web foi concebida não com a intenção de apenas o ser humano entender o conteúdo ali presente, mas também para que as máquinas conseguissem fazer a leitura e conseguir trabalhar em cima daquelas informações. Na busca de resolver este problema, surge o termo Web Semântica.

Tim Berners-Lee propõem colocar um sentido nos termos da Web, para que além dos humanos, as máquinas também consigam entender os textos. E assim consigam estruturar as informações, fazendo conjuntos de regras de inferências para automatizar o raciocínio.

A partir de então, surgiram várias representações e maneiras para fazer da Web Semântica uma realidade. Uma dessas maneiras é ilustrada na Figura 1, onde nesta proposta, existem várias camadas para conseguir aplicar a Web Semântica. As camadas deste modelo são descritas abaixo.

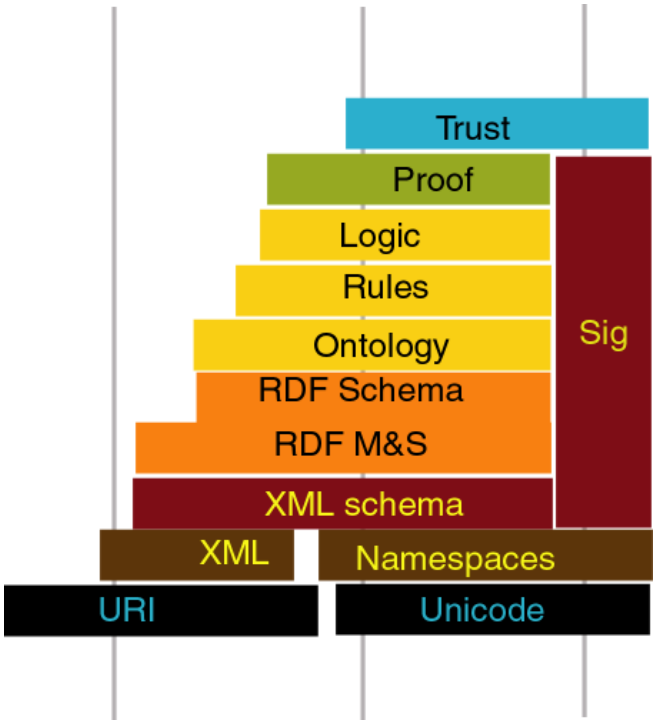

Figura 2: Estrutura da Web Semântica (W3C, 2014h)

- URI (*Uniform Resource Identifier* – Identificador de Recursos Uniforme): conjunto de caracteres para a identificação de um recurso (W3C, 2014b);
- Unicode: define um conjunto e padrão universal de codificação (UNICODE, 2008);
- XML (*Extensible Markup Language* – Linguagem de Marcação Extensível): é um sistema de representação de informação estruturada (W3C, 2014c);
- Namespace: um conjunto de nomes, identificada por uma referência URI.
- XML Schema: expressam os vocabulários compartilhados e permitem que as máquinas vejam as regras feitas pelas pessoas (W3C, 2014d);
- RDF M&S: um modelo para intercâmbio de dados na web, e tem características que facilitam a fusão de dados (W3C, 2014e);
- RDF Schema: um vocabulário para fazer a modelagem de dados de RDF (W3C, 2014f);
- *Ontology*: será tratado com mais clareza ainda neste capítulo;
- *Rules*: nela é feita a conversão das informações que estão dentro de um documento para outro, criando regras de inferência (PRADO, 2004).

- *Logic*: tem a intenção de transformar o documento em uma linguagem lógica, fazendo inferências e funções, para que duas aplicações de RDF sejam conectadas
- *Proof*: pode-se depois de passar por várias camadas, fazer uma prova deste documento, ou seja, pode-se provar hipóteses a partir das informações.
- *Sig*: assinatura, para verificar a autonomia do documento.
- *Trust*: tendo a assinatura do documento, pode-se saber a confiança nesta informação.

Dziekaniak (2004) diz que a semântica não está apenas relacionada ao conteúdo de um recurso, mas também na relação deste com os outros conteúdos da WEB. Logo é necessário que os recursos da Web sejam muito expressivos, para que os agentes e máquinas consigam processar a informação e entender seu valor.

Assim a Web Semântica, trará um significado às páginas, propiciando desta maneira com que os agentes e máquinas encontrem um ambiente que promova buscas e a recuperação da informação (BEERNERS-LEE, 2001a).

A Web Semântica não tem a intenção de criar uma nova Web, e sim de trazer um entendimento sobre a atual Web, onde a informação possa além de ser entendida por pessoas ser entendida por máquinas (PRAZERES, 2004).

Uma possibilidade de aderir semântica as páginas Web é através do uso de Ontologias, este tema será melhor explorado a seguir.

3.2. Definição de Ontologia

A palavra ontologia vem de *ontos* (ser, ente) e *logos* (saber, doutrina), e de maneira estrita significa o "estudo do ser". Surgiu do estudo de filósofos, ainda na época de Aristóteles, e era usada neste contexto para fazer uma abordagem do ser enquanto ser, ou seja do ser de uma maneira geral. Mais tarde ainda na filosofia, o termo ontologia passou a ser mais usado para saber aquilo que é fundamental ou irredutível, comum a todos os seres.

Dentro da Computação, Guarino (1998) diz que a ontologia é uma teoria lógica que representa um vocabulário pretendido, ou seja, é uma contextualização de algo particular existente no mundo. Neste sentido observa-se que com uma ontologia você consegue definir contextos e domínios particulares do mundo.

Gruber (1993) diz que em um contexto de múltiplos agentes, a ontologia poderia definir o contexto, o vocabulário daquele domínio, servindo assim de base para a comunicação entre os agentes, e para conseguir fazer suas extrações no conhecimento em que eles estão presentes. Gruber ainda diz que a ontologia é uma especificação explícita de uma conceitualização.

Posteriormente Borst (1997) complementa esta definição de Gruber dizendo que a ontologia é uma especificação formal de uma conceitualização compartilhada. Desta maneira traz que um dos principais objetivos da ontologia é o compartilhamento para o reuso destas informações.

Segundo Santarém Segundo (2010) a Ciência da Computação utilizou a ontologia quando se refere a aquisição de conhecimentos a partir de dados semiestruturados, utilizando da ontologia para aplicar técnicas e métodos, para processar as informações.

Santarém Segundo ainda diz que as ontologias vêm com o principal objetivo de ter um vocabulário compartilhado, onde essas informações possam ser trocadas, e usadas para outros usuários. Sendo que estes usuários são tanto seres humanos quanto agentes inteligentes.

Partindo disto, Guarino (1997) diferencia os tipos de ontologia, de acordo com sua utilização:

- Ontologia de topo (*top-level ontology*): tem uma função de descrever conceitos gerais, como o tempo, objeto, matéria, e que não estão dentro de um problema ou domínio particular. É aplicado na conceitualização de conceitos muito grandes e utilizados em grandes comunidades de usuários;
- Ontologia de domínio (*domain ontology*): já tem uma função de descrever conceitos de um domínio particular. São exemplos disto, áreas do conhecimento, como medicina, ciência da computação, entre outros;
- Ontologia de tarefa (*task ontology*): resolvem uma tarefa (um problema) dentro de um domínio. Ou seja dentro de um domínio, trata de algo específico, como uma doença dentro da medicina, ou compra e vendas de veículos.
- Ontologia de aplicação (*application ontology*): descrevem conceitos tanto de um domínio específico quanto de uma tarefa, que são especializações de ambas as ontologias. Estes conceitos correspondem a papéis desempenhados por entidades de domínio durante a execução de uma atividade.

Berners-Lee (2001b) diz que para uma semântica dentro da web funcione, é importante que a máquinas tenham acesso a coleções estruturadas de informações e que tenham regras de inferências que conduzam a máquina no processo de busca automatizada.

Dentro deste processo a ontologia aparece como uma solução neste sentido, pois a ontologia, conforme visto nos conceitos apresentados acima, será um conjunto estruturado de informações.

3.3. Construção da Ontologia

Segundo Clark (1999), uma ontologia é organizada em hierarquias de conceitos, por causa de não refletir de forma ideal nenhum formalismo específico, então pode-se considerar uma ontologia como a materialização do nível de conhecimento.

Também para a construção da ontologia, Gruber (1995) destaca alguns pontos que devem ser considerados na construção da ontologia:

- **Clareza**: é necessário que as escolhas dos termos tenham objetividade, e a definição deve ser independente do contexto social ou computacional. Sendo que as definições devem ser feitas com linguagem natural.
- **Coerência**: a ontologia deve ser consistente, para possibilitar que as inferências feitas a partir delas sejam consistentes.
- **Extensibilidade**: para que a ontologia, possa receber atualizações e incorporações de novos termos sem mudar os conceitos que já haviam sido definidos;
- **Codificação baixa**: para que não exista uma dependência de tecnologias ou de um tipo específico de codificação para a representação do conhecimento, pois o compartilhamento da informação, pode ser feito em ambientes diversos e diferentes do que foi feito a ontologia inicialmente.
- **Mínimo compromisso ontológico**: para permitir compartilhamento e o reuso da ontologia.

Gomez-Perez (1999) traz ainda autores que trouxeram outros pontos que devem ser considerados para a construção da ontologia:

- **Distinção da Ontologia**: onde as classes da ontologia devem ser distintas.
- **Diversificação da hierarquia para aumentar a força fornecida por múltiplos mecanismos de herança**: se existe conhecimento suficiente usado na ontologia e

existem muitos tipos diferentes de critérios para ser usados, é mais fácil inserir novos conceitos e herdar as propriedades de diversos critérios e pontos de vistas.

- **Modularidade**: para não existir dependências entre os módulos existentes.
- **Minimizar a distância semântica entre conceitos irmãos**: desta maneira conceitos similares, serão representados como subclasses de uma classe, enquanto conceitos menos semelhantes ficarão mais afastados na hierarquia.
- **Padronizar os nomes quando for possível**: para que não exista uma inconsistência nos nomes das classes.

3.4. Metodologias de Construção da Ontologia

Várias metodologias foram desenvolvidas para fazer a construção da ontologia, ou seja a engenharia da ontologia.

Falbo (1998) diz que independente do domínio, a construção de uma ontologia é uma tarefa bastante complexa, e a partir disto, alguns mecanismos de decomposição são necessários para facilitar este processo.

É interessante notar que não existe uma metodologia definida de como se deve construir uma ontologia, não existindo um consenso de qual metodologia se deva utilizar, assim, normalmente os desenvolvedores acabam fazendo sua própria metodologia (MARTIMIANO, 2006).

Para a construção da ontologia deste trabalho, foi utilizada a Metodologia definida por Noy e McGuiness (2001), que explicam uma forma de se realizar a engenharia da ontologia.

3.4.1. Metodologia da Noy e McGuiness

Neste contexto Noy (2001) definiu como deve ser o processo da construção da ontologia, para que esta ontologia não seja falha, e não apresente defeitos durante o seu funcionamento.

Noy (2001) explica os sete passos que são necessários para a construção de uma ontologia, esses passos estão descritos abaixo:

1. Determinar o Domínio e o Escopo da Ontologia: este momento é fundamental para se ter um escopo bem definido do que a ontologia irá representar. Neste passo

algumas perguntas devem ser respondidas como: “qual domínio a ontologia irá cobrir?”, “quem irá usar a ontologia?”, “quem irá usar e manter a ontologia?”.

2. Reutilizar Ontologias Existentes: como definido em alguns conceitos, a ontologia se refere ao uso compartilhado de algumas informações. Logo um dos passos para a construção de uma ontologia, é a verificação da existência de alguma ontologia semelhante construída, e reaproveita-la no projeto, apenas a melhorando ou a adaptando. Neste contexto existe algumas bibliotecas para o compartilhamento de ontologias, como “Ontolingua Ontology Library” (http://www.ksl.stanford.edu/software/ontolingua/) ou a “DAML Ontology Library” (http://www.daml.org/ontologies/)
3. Levantar termos importantes: é fundamental para a construção da ontologia fazer um levantamento de vários termos que são importantes para aquele cenário, para ter uma base de quais serão as classes, propriedades e hierarquia da ontologia.
4. Definir classes e sua hierarquia: Uschold e Gruninger (1996) definem que existe diversas maneiras para desenvolver uma hierarquia de classes da ontologia. Como a *top-down*, onde a modelagem começa dos conceitos mais gerais, e posteriormente a construção da hierarquia dos conceitos abaixo. Também existe o *bottom-up*, onde o processo de desenvolvimento começa dos conceitos mais específicos, e depois vai para os conceitos mais genéricos. E existe uma terceira forma, chamado de combinação, onde este processo é uma combinação entre o método de *bottom-up* e *top-down*, sendo definidos os conceitos mais importantes a princípio, e depois, feito uma generalização e uma especificação destes conceitos. O método para a construção deve ser escolhido segundo a necessidade e o conhecimento acerca do domínio.
5. Definir propriedades das classes: Nesta etapa deve-se observar os termos listados no passo três, e então, observar quais termos são propriedades de alguma classe, e assim encaixa-las dentro desta.
6. Restrições das Propriedades: Neste passo deve-se observar as restrições de cada propriedade, como cardinalidade e os tipos de dados das propriedades.
7. Criação de instâncias: aqui, é criado objetos (instâncias) das classes, pois muitas vezes estes valores já devem estarem definidos dentro da ontologia. Assim, deve ser criado um objeto da classe e ser preenchido as propriedades da classe, cujo a instância foi criada.

A partir destes passos, é possível então, construir uma ontologia que siga regras, e tenha uma boa consistência.

3.5. Linguagens para construção da ontologia

Existem muitas linguagens que podem ser utilizadas para a construção de uma ontologia como o RDF, OWL, Ontolingua/KIF entre outros. Todas estas linguagens conseguem representar uma ontologia, tendo que ser verificado as necessidades de cada caso, para assim poder escolher qual é a mais adequada (PRADO, 2004).

Em 2004, a W3C recomendou a linguagem OWL (W3C, 2014g) para a construção de ontologia. A seguir é descrito com mais detalhes esta linguagem.

3.5.1. OWL

A OWL (*Web Ontology Language* – Linguagem de Ontologia para Web) é uma linguagem criada para representar uma ontologia. Ela é usada em aplicações que necessitam processar o conteúdo de uma informação e deixa-la disponível para uma máquina realizar a leitura. A OWL proporciona uma maior capacidade de interpretação dos conteúdos da Web pelos computadores do que utilizando linguagens como o XML, o RDF e o RDF Schema (NAKAMURA, 2011).

Neste sentido, o uso da OWL possibilita que vocabulários mais ricos sejam adicionados para fazer a descrição das classes, para assim fazer comparações entre as classes, restringir cardinalidades e características das propriedades.

A linguagem OWL foi desenvolvida para satisfazer as necessidades da Web Semântica, num sentido de agregar o contexto nas informações.

A linguagem OWL é dividida em três sub linguagens que devem ser escolhidas conforme a necessidade de cada projeto:

- OWL Lite: é uma definição mais simples de hierarquia de classes e com restrições mais simples, onde por exemplo a cardinalidade que pode ser adotada nesta sublinguagem, seria apenas valores 0 e 1. É mais utilizada para uma migração mais simples de tesauros e outras taxonomias. Oferecendo uma formalidade muito mais baixa.

- OWL DL (*Descriptions Logics*): esta sublinguagem já apresenta uma expressividade máxima, fazendo com que esta expressividade seja computável, ou seja, permitindo com o computador consiga fazer a leitura desta expressividade. Mas apresenta algumas restrições como por exemplo uma classe não poder ser instância de outra classe.
- OWL *Full*: é a linguagem OWL completa com todas as suas funções sem restrições, sem ter a garantia que um sistema conseguirá entender o que está descrito naquela OWL.

3.5.1.1. Elementos do OWL

Todas as classes do OWL são subclasses de "*owl:Thing*", e a linguagem OWL possibilita que estas classes tenham propriedades de elementos e de restrições. Abaixo segue algumas destas propriedades (ANTONIOU E HARMELEN, 2004):

- Propriedade de elemento
    - Propriedades de Objeto: utilizado para relacionar um objeto com outro, exemplo: "supervisor de".
    - Propriedade de Tipo de Dados: utilizado para relacionar objetos com tipos de dados. Um exemplo disto seria os dados como "telefone", "idade", entre outros.
- Propriedade de restrição
    - Todos os valores de (*owl:allValuesFrom*): é utilizada para definir quais são os valores possíveis que a propriedade especificada por *owl:onProperty* pode ter.
    - Tem o valor (*owl:hasValue*): define um valor determinado para a propriedade especificada por *owl:onProperty* pode ter.
    - Algum valor de (*owl:someValuesFrom*): tem a função de determinar a classe e a ocorrência de pelo menos um valor dentre as propriedades.
    - Cardinalidade Mínima (*owl:minCardinality*): restringe o valor mínimo de um relacionamento.
    - Cardinalidade Máxima (*owl:maxCardinality*): restringe o valor máximo dentro de um relacionamento.
- Propriedades Especiais

- Propriedade Transitiva (*owl:TransitiveProperty*): define que uma propriedade é transitiva, seguindo um sentido de “tem melhor qualidade que”, “é ancestral de”.
- Propriedade Simétrica (*owl:SymmetricProperty*): define a simetria entre as classes, como “é similar a”.
- Propriedade Funcional (*owl:FunctionalProperty*): define uma propriedade que tem pelo menos um valor para cada objeto, como “idade”, “altura”.
- Propriedade Funcional Inversa (*owl:InverseFunctionalProperty*): define uma propriedade que dois objetos não podem ter o mesmo valor, por exemplo um campo id de identificação, onde cada valor deve ser único.

3.6. Ambiente de Desenvolvimento da Ontologia

Para se desenvolver ontologias utilizando como linguagem o OWL, existe o ambiente Protégé (STANFORD, 2014), que auxilia neste processo, ajudando para que a modelagem e a construção da ontologia se torne mais simplificada.

3.6.1. Protégé

O Protégé é uma ferramenta usada para o desenvolvimento de sistemas baseados em conhecimento. Esta ferramenta permite que seja construída uma ontologia de domínio ou uma base de conhecimento, permitindo, para auxiliar neste processo, a construção de diagramas e de gráficos.

O sistema é *open-source* construído em Java e pode ser instalado direto no computador desktop ou ser executado direto na Web (PRADO, 2004). A tela inicial do sistema pode ser vista na figura 2.

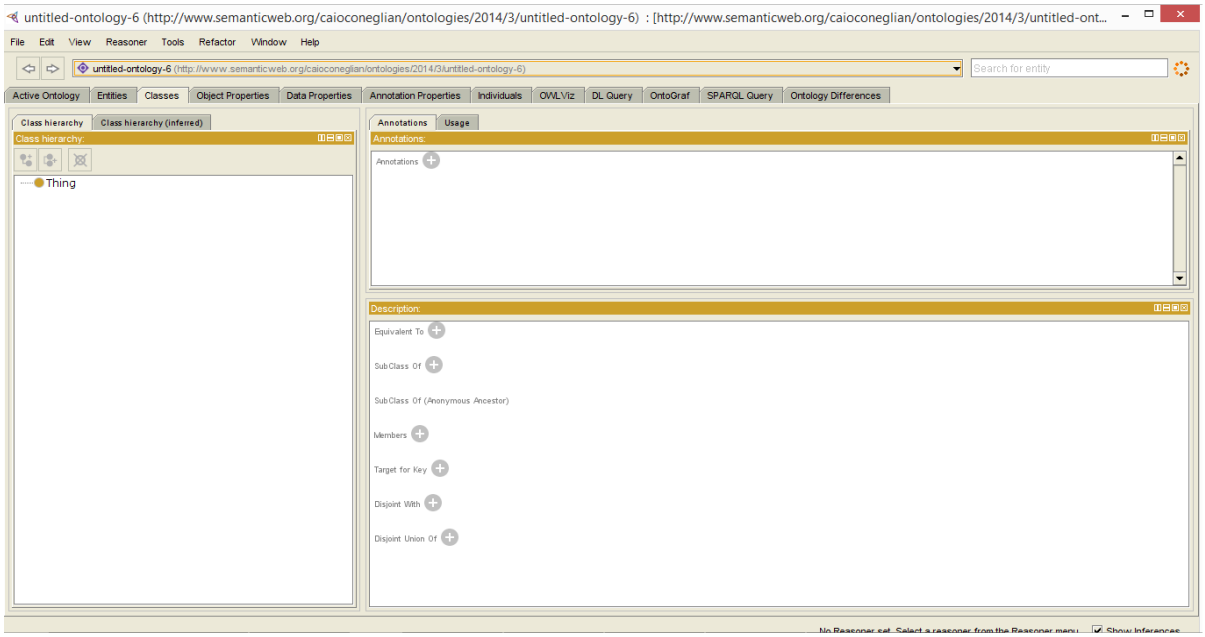

Figura 3: Tela Protégé

Nesta figura é possível verificar que dentro deste sistema tem a aba *classes*, *object properties*, *data properties*. Na aba *classes* são construídas as classes e as relações de hierarquia entre elas, já na aba *object properties* são construídas as relações entre duas classes e na aba *data properties*, fica as propriedades de dados de cada classe (por exemplo idade, data de nascimento, CPF).

4. Proposta de Recuperação da Informação

Os sistemas de informação tradicionais são incapazes de lidar de forma eficiente com todas as novas fontes de dados dinâmicas e de contextos múltiplos de informações que têm principalmente a Internet como plataforma.

São encontrados problemas em recuperar, padronizar, armazenar, processar e utilizar informações geradas por diversas fontes heterogêneas que servem de base para alimentar os sistemas de apoio à decisão das organizações.

Para resolver esta problemática foi proposta a criação de uma arquitetura de Recuperação de Informação no contexto de Big Data como pode ser visto na Figura 4.

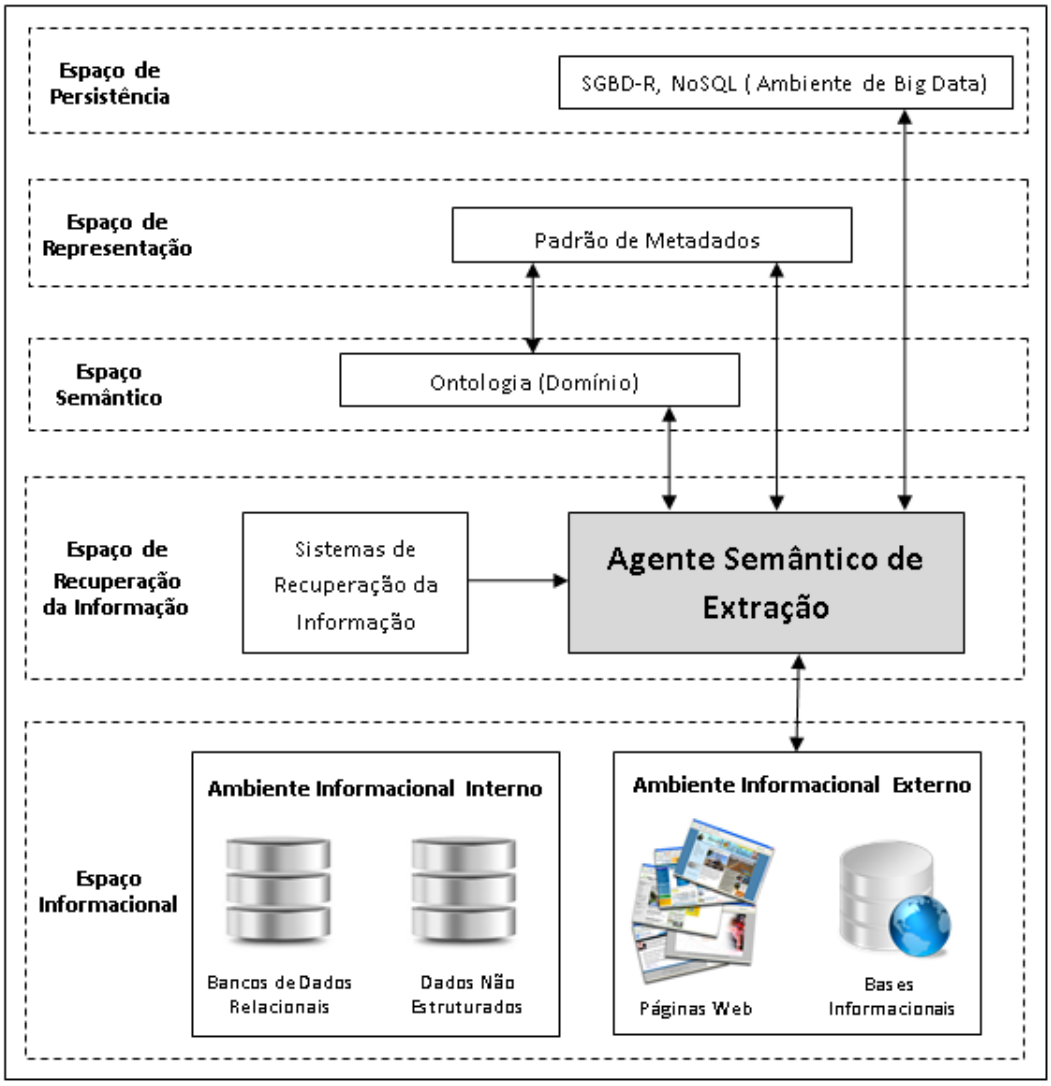

Figura 4: Arquitetura de Contextualização do Agente Semântico de Extração

A arquitetura proposta, contempla a ideia de ser realizada uma recuperação de informações tanto em ambientes internos (banco de dados) quanto externos (WEB), utilizando-se de um agente de extração, que para analisar o domínio da informação usa de ontologias.

Este trabalho tratará das camadas do espaço semântico, do espaço de Recuperação da Informação e do Espaço Informacional. Tratando da questão de recuperar, processar e utilizar informações diversas.

Esta arquitetura é dívida em cinco partes que serão exploradas a seguir:

4.1. Espaço de Persistência

No espaço de persistência ocorre o armazenamento das informações que são extraídos pelo agente de busca. Este armazenamento pode ocorrer tanto em Sistemas Gerenciadores de Banco de Dados relacionais, quanto em Banco de Dados NoSQL (*Not Only SQL* – Não apenas SQL). A persistência pode ocorrer destas duas maneiras pelo fato de que as informações extraídas podem ser, tanto dados estruturados quanto dados não estruturados.

Desta forma os dados estruturados podem ser armazenados em cima de bancos de dados relacionais, que apresentam regras bem definidas, e consegue dar uma integralidade maior aos dados que são armazenados.

Já dentro da Web, existe uma grande quantidade de dados que são semiestruturados ou não-estruturados, ou seja, não seguem regras, ou não tem uma estrutura exatamente definida. Assim, é necessário que estes dados sejam armazenados em bancos de dados NoSQL, que apresentam características de não ter um esquema totalmente definido, e que permite uma flexibilidade maior ao armazenar estas informações.

4.2. Espaço de Representação

O espaço de representação tem a função de definir os padrões de metadados para que seja realizado a busca pelo agente e que assim consiga posteriormente definir como os dados serão representados e armazenados nos bancos de dados.

4.3. Espaço Semântico

O espaço semântico tem a função de inserir uma semântica nas buscas realizadas pelo agente de busca. Sendo possível que a busca realizada pelo agente, leve em consideração o contexto na qual aquela informação está inserida.

Esta semântica pode ser alcançada através do uso de uma estrutura ontológica, que analisa o domínio do contexto que se deseja buscar as informações.

4.4. Espaço de Recuperação de Informação

O espaço de recuperação de informação é onde é representado o sistema de recuperação de informação, que tem a função de ser o gerenciador das buscas e de realizar a recuperação propriamente dita.

Neste espaço também, fica o Agente Semântico de Extração, este agente ficará responsável por extrair os dados dos ambientes informacionais. Este agente utiliza da ontologia para conseguir realizar a busca semântica.

4.5. Espaço Informacional

O espaço informacional contempla toda a Web e as bases de dados internas, que serão utilizados como fontes para a extração do agente de buscas. Portanto, todos os dados estão dentro do espaço informacional, que necessita ser extraído, para se tornar conhecimento para quem for utilizá-lo.

Neste trabalho foi construído esta arquitetura de forma parcial, sendo realizado o espaço semântico, onde foi construída uma ontologia. Também foi utilizado o Agente Semântico de Extração e o espaço informacional. Sendo também construída toda a relação entre estes espaços.

Esta arquitetura busca provar o uso de ontologias para conseguir inserir semântica, dentro de um contexto de Big Data, que faz uso de um número muito grande de informações.

Para provar isto, este projeto, funciona de maneira que, o espaço informacional são bases de dados de artigos científicos, no caso, foi utilizado a base de dados do IEEE Xplore (http://ieeexplore.ieee.org).

Na figura 5, é mostrado o processo feito pelo sistema. O usuário realiza uma busca sobre algum tema, o agente extrai das bases de dados resumos referentes a este tema. Estes resumos irão passar por um processo, onde estes serão analisados, levando em consideração se as palavras contidas neste resumo, estão presentes no domínio daquele tema procurado. Isto será possível, utilizando uma ontologia construída, que trata de um tema específico na área de pesquisa científica.

Neste trabalho, a ontologia trata-se da área de Banco de Dados, portanto, este processo funcionará por buscas realizadas neste domínio.

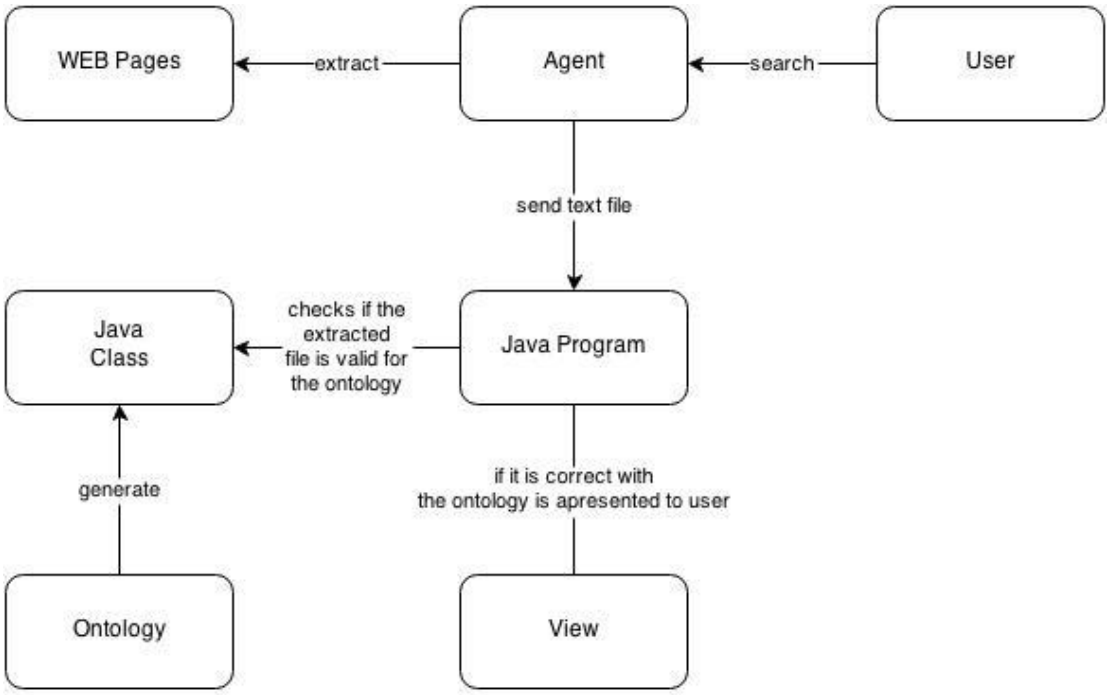

Figura 5: Processo realizado pelo sistema de extração

É possível verificar que o processo é finalizado quando é apresentado ao usuário as informações extraídas pelo agente, após passar pela ontologia. Buscando apresentar um resultado mais consistente, com uma semântica bem estruturada.

5. Modelagem e Implementação da Ontologia

A ontologia encontra-se no espaço semântico da arquitetura, ou seja, será a ontologia a responsável pela busca ser mais semântica e menos sintática.

A ontologia necessária para a utilização deste projeto, é uma ontologia que deve tratar de um domínio específico, onde uma área do conhecimento é representada em sua totalidade, com a função de ser utilizada para a determinação se algumas informações estão ou não contidas dentro daquele contexto.

Seguindo esta necessidade, foi verificado que a ontologia que foi construída é classificada, segundo Gomes-Perez (1999), como uma ontologia de domínio, pois trata de um domínio mais específico de uma área do conhecimento.

Esta ontologia tem a função de representar uma área do conhecimento, para a utilizar na verificação dos artigos, determinando se estes estão contidos nesta área do conhecimento. Devido ao fato do autor, ter um conhecimento mais amplo na área de banco de dados, foi utilizado este domínio para a implementação da ontologia.

Neste sentido, a ontologia representa a área de Banco de Dados como um todo, abrangendo, os tópicos de pesquisa relacionada à esta área.

Para a construção desta ontologia, foi utilizado o método de Noy (2001), que determina os sete passos para a construção da ontologia. Os passos desta metodologia aplicados a este projeto são demostrados abaixo:

1. Determinar o Domínio e o Escopo da Ontologia: o domínio é a área de Banco de Dados, abrangendo os tópicos de pesquisa mais comum nesta área;
2. Reutilizar Ontologias Existentes: foi pesquisado nas principais bibliotecas online de ontologias, para verificar se havia ontologias que tratavam de Banco de Dados como um todo, não sendo encontrada nenhuma ontologia que atendesse esta necessidade;
3. Levantar termos importantes: foram levantados os seguintes termos: SQL, NoSQL, Modelo, Datawarehouse, relacionamento, bancos relacionais, bancos orientados a documentos, bancos orientados a colunas, bancos orientados a grafos, restrições, normalização, segurança, esquemas, instâncias, transação, objetos, administração, esquemas, álgebra relacional, modelo entidade relacionamento, modelo entidade relacionamento estendido, projeto de banco de dados relacionais, diagrama ER,

MongoDB, CouchDB, Cassandra, Neo4J, Big Table, Oracle, MySQL, PostgreSQL, Firebird, Microsoft SQL Server;

4. Definir classes e sua hierarquia: foi definida utilizando mapas mentais, as classes e as relações de hierarquia entre elas. Na figura 6, é representado esta relação

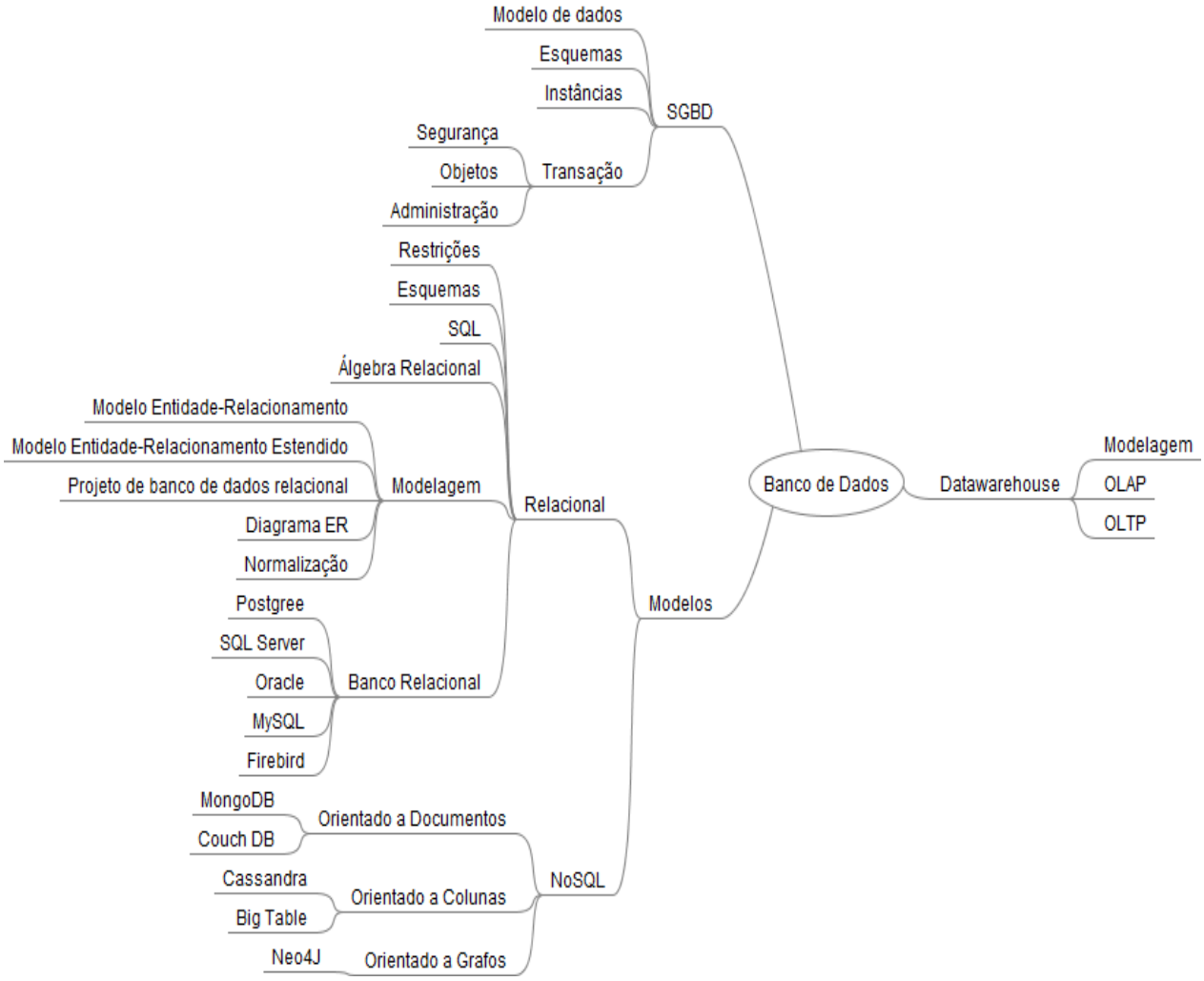

Figura 6: Mapas mentais representação a relação hierárquica da ontologia

5. Definir propriedades das classes: este passo não foi realizado devido o fato que nesta ontologia, não há a necessidade de levar em consideração as propriedades de cada nó da ontologia, pois o mais importante é a relação entre as classes propriamente dita;
6. Restrições das Propriedades: como não há propriedades, não é necessário tratar das restrições entre estas;

7. Criação de instâncias: Não há a necessidade de criar instâncias, pois as instâncias serão propriamente os termos retirados pelo agente de extração.

Posteriormente a construção da ontologia, seguindo a metodologia de Noy, foi realizada a implementação da ontologia utilizando o software Protégé (STANFORD, 2014), onde utilizando o esquema de mapas mentais mostrado na figura 6, foi realizado a construção da ontologia, onde após a realização da modelagem pelo Protégé, é gerado um arquivo OWL que representa a ontologia.

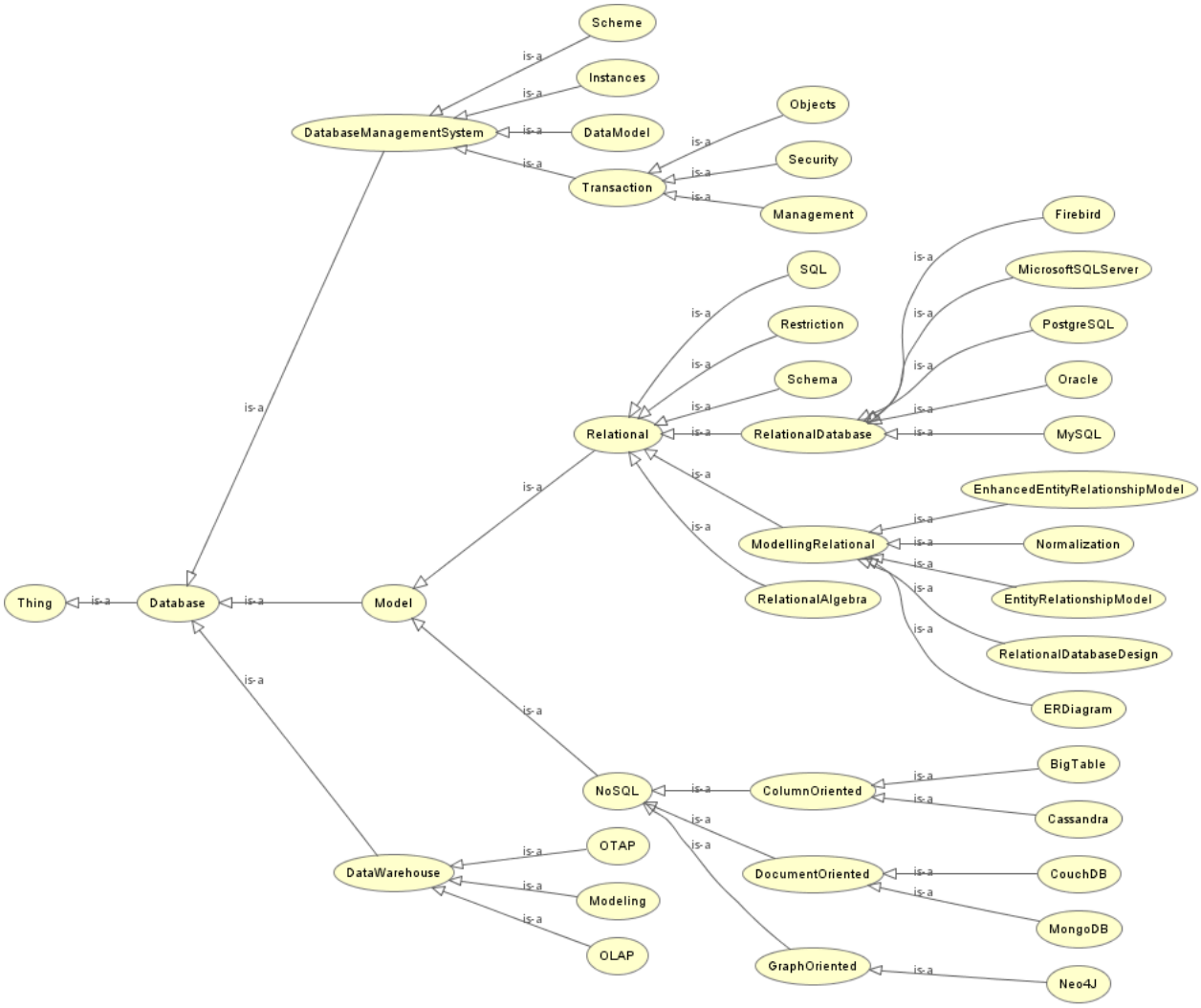

Figura 7: Relação das classes feitas no Software Protégé

A ontologia modelada pelo software Protégé, pode ser visualizado através da figura 7, que mostra as relações da ontologia. Nesta modelagem, a ontologia, já foi construída em inglês, pelo fato que as fontes de informações que serão retirados os artigos são da língua

inglesa.

Esta modelagem mostrada na figura 7, representa um arquivo OWL, que mostra as classes e as relações entre elas. Este arquivo OWL, foi utilizado para que a ontologia, fosse representada em classes Java, através do software Owl2Java (2009), que realiza esta transformação, descrevendo toda estas relações entre as classes da ontologia, mesmo nas classes Java.

6. Agente de Extração e Integração com a Ontologia

Após ser realizado a implementação da ontologia e a transformação desta em classes Java. Foi possível iniciar a integração da ontologia com o agente de buscas.

A implementação consistiu na integração do agente de buscas com a ontologia, ou seja, a comunicação das informações que são extraídas, com o intuito de dar semântica a busca. Desta maneira, o agente extrai um texto de uma página, e um algoritmo irá avaliar se aquela informação está dentro do contexto da ontologia, e se aquela informação de fato será útil para o usuário.

6.1. Extração da informação

O agente extrai da página do IEEE Xplore (http://ieeexplore.ieee.org), os resumos, baseado na pesquisa que o usuário executa. Baseado na localização dos resumos no HTML na página, o agente extrai as informações, e transforma isto numa cadeia de String. Na figura 8 é visto o processo de funcionamento do robô de busca.

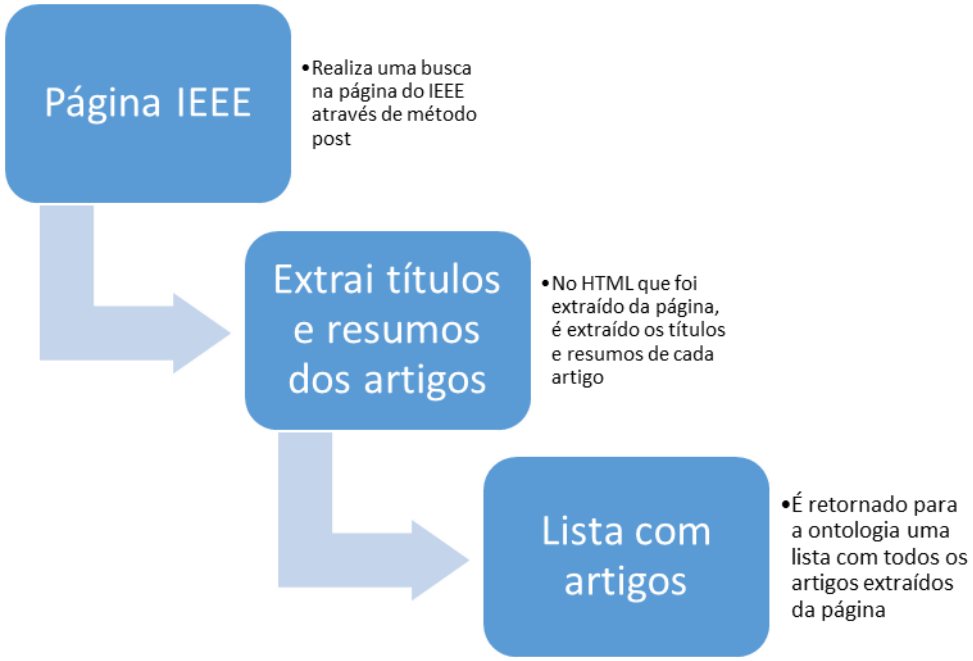

Figura 8: Diagrama com estrutura do robô de extração

Como mostrado na figura 8, é possível visualizar que o processo do agente é divido em três fases: busca na página, extração dos títulos e resumos e devolução ao programa

principal uma lista com os artigos.

- Busca na página HTML: esta primeira fase se caracteriza por realizar uma busca no sistema de busca do IEEE Xplore, de forma que a busca realizada se caracteriza por uma requisição a este sistema, sendo inserido na url, qual é o tema que o usuário deseja buscar. Por exemplo, caso o usuário deseje realizar uma busca sobre Datawarehouse, o agente irá abrir uma conexão, e buscar no seguinte endereço (http://ieeexplore.ieee.org/search/searchresult.jsp?newsearch=true&queryText=datawarehouse). A partir disto a página do IEEE, irá retornar um HTML, contendo os artigos relacionados a este tema. Na figura 9 é mostrada como é a página HTML do retorno.

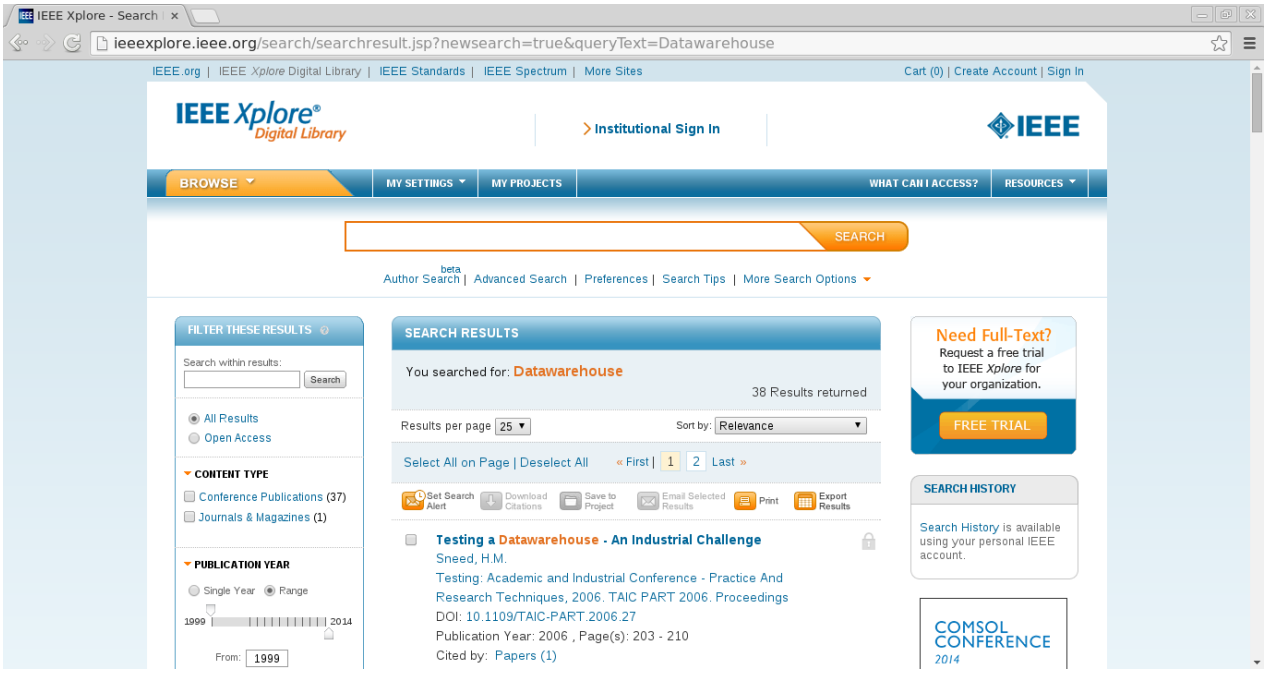

Figura 9: Página de retorno do IEEE Xplore

- Extração de títulos e resumos da página: após o retorno do HTML, o agente extrai deste, o título e o resumo de cada artigo. Isto é possível por uma análise da página HTML, verificando as *tags* cujo os dados dos resumos e dos títulos estão inseridos. Desta maneira para cada artigo é criado um objeto Java que contém os dados do título, do resumo e do link para o acesso ao artigo completo. Para realizar esta retirada de dados dentro de uma página HTML, foi utilizada a ferramenta JSOUP (2014). Esta ferramenta funciona como um *HTML Parser*, ou

seja, trabalha com a página HTML, de maneira que consiga extrair os dados das classes, tags e estruturas do HTML.

- Criação de uma lista com os artigos extraídos: por fim, o agente cria uma lista contendo todos os artigos que foram extraídos da página HTML. Esta lista será utilizada pelo programa principal que irá unir a ontologia com este agente de recuperação de informação.

Desta forma, este robô de busca, consegue realizar uma extração sintática dos artigos contidos na base de dados do IEEE Xplore, pois, o robô de busca recupera os artigos que foram indexados pela própria base de dados, criando uma lista com todos os artigos que foram apresentados, para ser utilizado na ontologia.

6.2. Integração da Ontologia com o Agente de Extração

Para que o programa tenha de fato a semântica apresentada, o programa faz o uso da ontologia, para avaliar quais dos resultados que foram extraídos da base de dados, são de fato úteis, e tem relação com o contexto daquela busca.

Esta integração acontece em cinco momentos:

- Primeiramente, é verificado onde o termo pesquisado pelo usuário se encontra dentro da ontologia. Por exemplo, se o usuário realiza uma busca de Datawarehouse, o sistema irá verificar onde este termo está dentro da ontologia.
- Depois são obtidos, quais são as classes hierarquicamente superior e inferior ao termo pesquisado. No exemplo do Datawarehouse, serão obtidos, as classes inferiores: OLAP, OTAP e modeling, e a classe superior Database. É possível visualizar este processo na figura 10, onde são visualizados apenas as classes que tem relação com o termo pesquisado, no caso Datawarehouse.

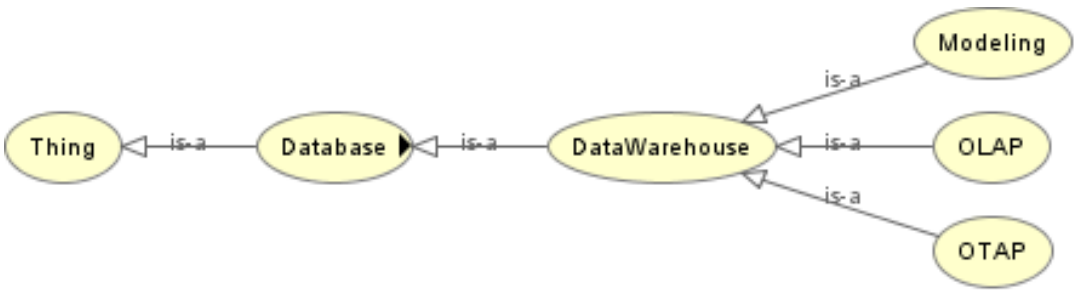

Figura 10: Relações da classe, do termo pesquisado

- Posteriormente é verificado dentro do resumo e do título dos artigos pesquisados, se contém ou não, os termos que fazem parte daquela hierarquia do termo pesquisado. No exemplo do Datawarehouse, seria verificado se os termos OLAP, OTAP, modeling, datawarehouse e database, estão contidos dentro dos resumos e dos títulos daqueles artigos extraídos.
- Após, é realizado uma comparação entre quantidade de termos que estão na hierarquia e os que estão contidos dentro do resumo e do título daquele artigo. Resultando assim uma porcentagem da quantidade de termos que estão na hierarquia, que estão dentro do resumo e do título daquele artigo. No mesmo exemplo, se conter os termos Database, OLAP, Datawarehouse e modeling, dentro de um artigo, vai conter quatro dos cinco termos da hierarquia, o que resulta numa porcentagem de 80% dos termos.
- Por fim, é apresentado ao usuário todos os artigos que alcançaram uma porcentagem acima dos 35%.

6.3. Interação do Usuário com o Programa

O usuário na primeira tela pode escrever o tema que ele deseja realizar sua busca. No caso do programa que foi implementado, o usuário necessariamente precisa realizar uma busca relacionado a banco de dados. Na figura 11, é possível visualizar a tela para o usuário realizar a busca.

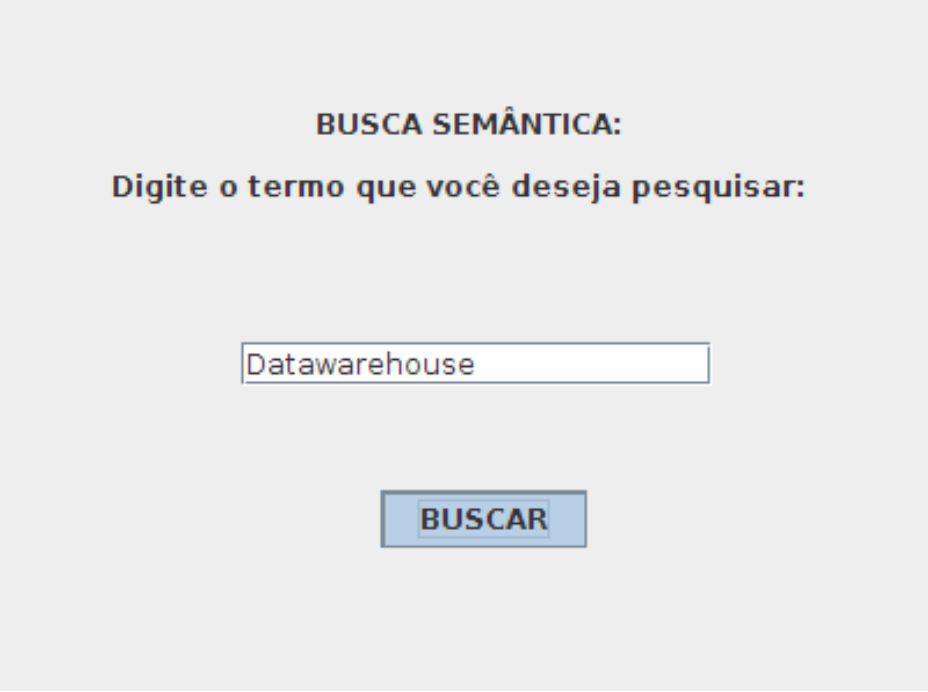

Figura 11: Tela de interação com o usuário para realizar a busca

Após o usuário escrever o que ele necessita, o sistema irá fazer os passos descritos nos capítulos 6.1. e 6.2., onde o sistema faz a integração da pesquisa do usuário, com a extração realizada no site do IEEE Xplore, com a ontologia.

Após realizar estes passos, o sistema retorna para o usuário, uma tela contendo quais são os artigos e os links destes artigos, que o sistema extraiu e verificou que tinha relação com a busca realizada pelo usuário. Este resultado é possível visualizar na figura 12, onde são apresentados os nomes e os links, para que o usuário possa acessar ao artigo completo.

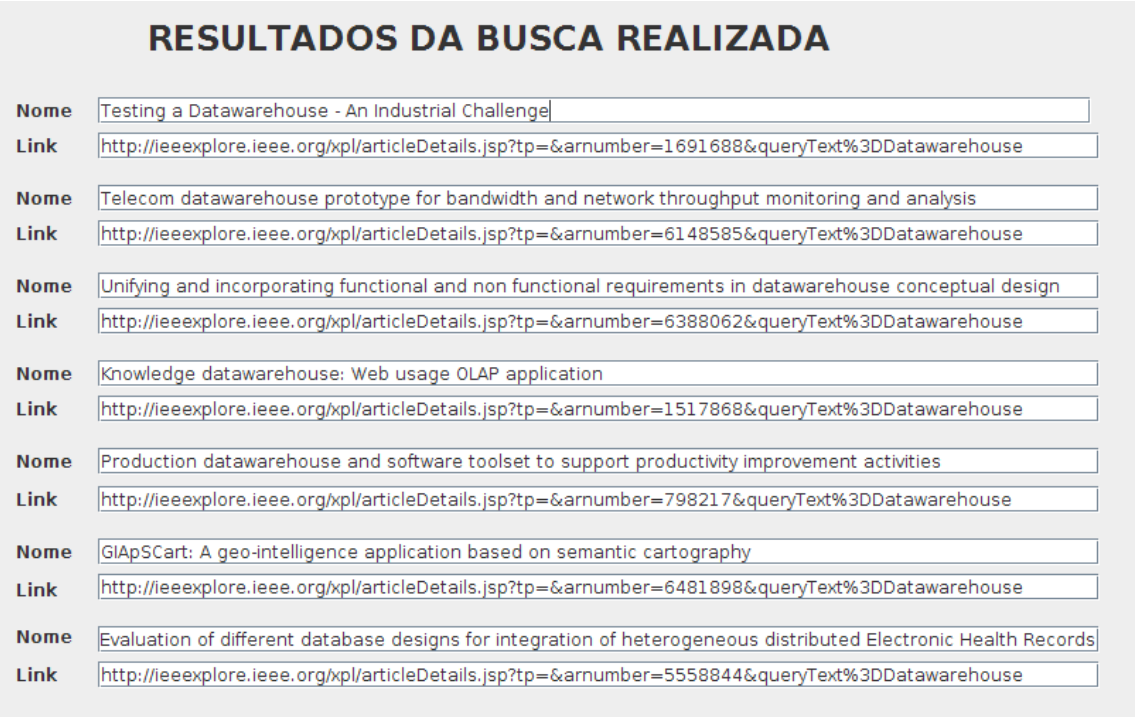

Figura 12: Tela de resultados da busca realizada

7. Resultados

Como teste para averiguar se o sistema está extraindo e verificando a semântica dos artigos extraídos, foi feita uma busca com o usuário pesquisando pelo termo "Datawarehouse", como mostrado na figura 11.

A hierarquia do termo Datawarehouse são os termos: Database, Datawarehouse, OLAP, OTAP e modeling.

Na tabela 2, é possível visualizar todos os títulos dos artigos que foram extraídos do site do IEEE, a quantidade dos termos da cadeia da ontologia que foram encontrados no resumo e no título, a relação entre os termos encontrados no artigo e os termos da cadeia da ontologia do termo "Datawarehouse" (no caso será a porcentagem resultante da divisão entre a quantidade de palavras encontradas na ontologia por 5, que são os termos contidos na hierarquia da cadeia de ontologia) e se este artigo atende ou não ao requisito mínimo de pelo menos 35% dos termos contidos no resumo e no título.

Tabela 2: Análise dos Artigos Extraídos

Título	**Qtd. de palavras encontadas**	**%**	**Atende ao requisito?**
Testing a Datawarehouse - An Industrial Challenge	2	40	**SIM**
Telecom datawarehouse prototype for bandwidth and network throughput monitoring and analysis	3	60	**SIM**
Unifying and incorporating functional and non functional requirements in datawarehouse conceptual design	3	60	**SIM**
Knowledge datawarehouse: Web usage OLAP application	2	40	**SIM**
Datawarehouse and dataspace — information base of decision support syste	1	20	NÃO
The implementation of datawarehouse in Batelco: a case study evaluation and recommendation	1	20	NÃO
E-Business Model Approach to Determine Models to Datawarehouse	1	20	NÃO
Production datawarehouse and software toolset to support productivity improvement activities	2	40	**SIM**
A genomic datawarehouse model for fast manipulation using repeat region	1	20	NÃO
A datawarehouse for managing commercial software release	1	20	NÃO
Modeling Analytical Indicators Using DataWarehouse Metamodel	1	20	NÃO
An SLA-Enabled Grid DataWarehouse	1	20	NÃO

Business Metadata for the DataWarehouse	1	20	NÃO
A partition-based approach to support streaming updates over persistent data in an active datawarehouse	1	20	NÃO
Study of localized data cleansing process for ETL performance improvement in independent datamart	1	20	NÃO
Visualizing Clouds on Different Stages of DWH - An Introduction to Data Warehouse as a Service	0	0	NÃO
GIApSCart: A geo-intelligence application based on semantic cartography	2	40	**SIM**
JISBD 2008 + TELECOM I+D 2008 = INTRODUCTIONS	0	0	NÃO
Normed principal components analysis: A new approach to data warehouse fragmentation	0	0	NÃO
Enriching hierarchies in multidimensional model of data warehouse using WORDNET	0	0	NÃO
The fragmentation of data warehouses: An approach based on principal components analysis	0	0	NÃO
Evaluation of different database designs for integration of heterogeneous distributed Electronic Health Records	2	40	**SIM**
Keynote talk data warehouses: Construction, exploitation and personnalisation	1	20	NÃO
Security Analysis of Future Enterprise Business Intelligence	0	0	NÃO
QVT transformation by modeling: From UML model to MD model	1	20	NÃO

No caso de 25 artigos, 7 foram os que atenderam aos requisitos, sendo estes apresentados aos usuários, esta apresentação pode ser visualizada da figura 12.

Para visualizar como o programa faz a análise dos resumos e dos títulos, abaixo na figura 13, é apresentado um artigo dos que atenderam aos requisitos.

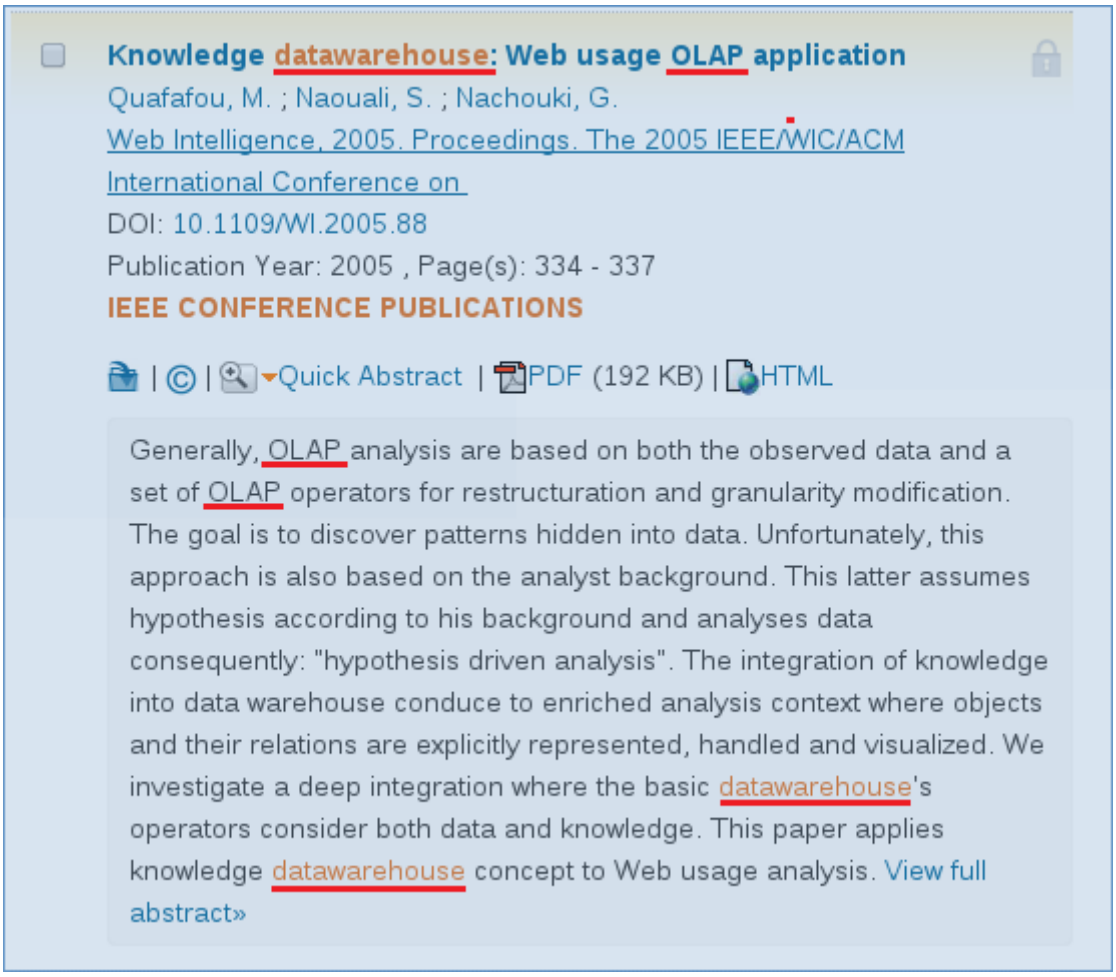

Figura 13: Exemplo de um artigo analisado.

Este artigo, como visto na tabela 2, apresentou 2 dos 5 termos da hierarquia da ontologia. Neste caso apresentou os termos OLAP e Datawarehouse. Na figura 13 está grifado em vermelho os termos que foram encontrados pelo programa.

8. Conclusões

Este trabalho apresenta o uso de ontologias na melhoria do processo de Recuperação de Informação.

O objetivo desta pesquisa é aderir semântica ao processo de Recuperação da Informação, utilizando das informações dentro do contexto do Big Data, para realizar um processo que agregue mais valor às buscas realizadas pelo usuário.

Para comprovar este objetivo, foi utilizado o domínio de pesquisas científicas, em que o usuário ao realizar uma busca em bases de dados de artigos científicos, se depara com o problema de ter uma quantidade muito grande de documentos, sendo que boa parte destes, não são de fato úteis, não atendendo às necessidades que o usuário possui.

Foi, então, criado uma ontologia e um robô de buscas e realizada a conexão entre estes para alcançar desta maneira o objetivo inicial.

Para a realização de testes, no sentido de averiguar o real funcionamento deste processo, o robô de buscas foi implementado com a capacidade de extrair artigos da base de dados do IEEE Xplore, e a ontologia foi construída utilizando o domínio da disciplina de banco de dados.

Após a realização de testes, foi observado que o uso de ontologia para o agente de pesquisa é uma maneira eficaz para se obter informações de valor e conseguir atender as necessidades informacionais do usuário.

A ontologia pode ser eficiente no presente processo, porque se torna uma forma de organizar a informação semântica, e assim, apenas a informação significativa será apresentada ao usuário.

Embora o termo Web Semântica é usado já a alguns anos, ainda existe uma limitação em seu uso, porque grande parte da Web está organizada de uma forma sintática, em que a maioria das páginas são criadas para que apenas o ser humano consiga ler o que lá está escrito, sem serem estruturadas de uma maneira que agentes computacionais consigam extrair os dados ali contidos dentro de um contexto, com um significado implícito dentro do HTML.

O agente de extração consegue retirar os documentos da Web e um programa consegue por meio do uso de ontologia, tratar as informações, conseguindo assim apresentar resultados mais relevantes aquele usuário.

Desta maneira os resultados obtidos com a utilização do protótipo desenvolvido,

consegue refinar bastante a quantidade de artigos apresentados aos usuários. Esta pesquisa, busca portanto, fazer com o que o usuário obtenha, em um processo de Recuperação de Informação, resultados mais expressivos e que apresente maior valor. Assim, o usuário conseguirá avaliar informações mais expressivas, e não perderá tempo com aqueles dados que não tem atende suas necessidades.

Portanto, para tratar a questão de como inserir uma inteligência na recuperação de páginas Web que não apresentam uma contextualização de suas informações, esta pesquisa propõe que o processo de aderir semântica a estas páginas ocorra fora da Web, ou seja, a extração das páginas ocorra de maneira sintática, e a partir do que foi extraído, ocorra uma análise das informações, inserindo desta forma semântica a este processo. Este método se mostrou muito eficiente, pois consegue de fato realizar uma busca mais inteligente, que vai além de simples fórmulas de buscas, que observam apenas a sintaxe dos textos, e consegue analisar o contexto na qual os documentos extraídos estão inseridos, e assim visualizar se aquele documento atende ao que o usuário necessita.

Referências Bibliográficas

Antoniou, G. e Van Harmelen, F. **A semantic web primer.** MIT press, 2004.

Baeza-Yates, R.; Ribeiro-Neto, B. **Modern information retrieval.** New York: ACM; Harlow: Addison-Wesley, 1999.

Bentlet, P. J. **Biologia digital: como a natureza está transformando nossa tecnologia e nossas vidas.** São Paulo: Berkeley Brasil, 2002.

Beppler, Fabiano D. et al. **Uma arquitetura para recuperação de informação aplicada ao processo de cooperação universidade-empresa**. KM Brasil, São Paulo, Brasil, 2005.

Berners-Lee, T. **Information Management: A Proposal.** 1989. Disponível em <http://www.w3.org/History/1989/proposal.html> acesso em 09 de julho de 2014.

Berners-Lee, T. **Semantic Web Road Map.** 1998. Disponível em <http://www.w3.org/DesignIssues/Semantic.html> acesso em 12 de julho de 2014.

Berners-Lee, T., Lassila, O. e Hendler, J. **The semantic web. Scientific American**, New York, v. 5, 2001a.

Berners-Lee, T., Hendler, J. e Lassila, O. **The semantic web**. Scientific american 284.5. 28-37. 2001b.

Beyer, M. A., e Laney, D. **The importance of 'big data': a definition.** Stamford, CT: Gartner. 2012.

Borst, W. N. **Construction of engineering ontologies for knowledge sharing and reuse.** 1997. 227 f. Tese (Doutorado). Centre for Telematics for Information Technology, University of Twence, Enschede. 1997.

Brasiliano, A. C. R. **A Fuga Involuntária das Informações Estratégicas nas Empresas: Fragilidades nas Redes Humanas.** 2002. Disponível em <http://www.abraic.org.br> acesso em 20 de março de 2014.

Castro, J. M., e Abreu, P. G. **Influência da inteligência competitiva em processos decisórios no ciclo de vida das organizações**. Ciência da Informação 35.3. 15-29. 2006.

Clark, D. **Mad cows, metathesauri, and meaning.** Intelligent Systems and their Applications, IEEE 14.1. 75-77. 1999.

De Diana, M., e Gerosa, M. A. **Nosql na web 2.0: Um estudo comparativo de bancos não-relacionais para armazenamento de dados na web 2.0.** 2010.

Deters, J. I., e Adaime, S. F. **Um estudo comparativo dos sistemas de busca na web.** Anais do V Encontro de Estudantes de Informática do Tocantins. Palmas, TO. 189-200. 2003.

Dziekaniak, G. V., e Kirinus, J. B. **Web semântica**. 2004.

Falbo, R. A. **Integração De Conhecimento Em Um Ambiente De Desenvolvimento De Software.** 1998. 215 f. Tese (Doutorado em Ciências em Engenharia de Sistemas e Computação) – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro. 1998.

Ferneda, E. **Aplicando algoritmos genéticos na recuperação de informação**, DataGramaZero: Revista de Ciência da Informação, Rio de Janeiro, v. 10, n. 1, fev. 2009. Disponível em: <http://www.dgz.org.br/fev09/F_I_aut.htm>. Acesso em: 13 de outubro de 2014.

Ferneda, E. **Recuperação da Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação.** 2003. 147 f. Tese (Doutorado em Ciência da Informação) – Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo. 2003.

Graham-Rowe, D., et al. **Big data: science in the petabyte era.** Nature 455. 1-50. 2008.

Gruber, T. R. **A translation approach to portable ontology specifications.** Knowledge acquisition 5.2. 199-220. 1993.

Gruber, T. R. **Toward principles for the design of ontologies used for knowledge sharing?** International journal of human-computer studies 43.5. 907-928. 1995.

Guarino, N. **Formal ontology in information systems.** Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy. Vol. 46. IOS press, 1998.

Guarino, N. **Understanding, building and using ontologies.** International Journal of Human-Computer Studies 46.2. 293-310. 1997.

Gómez-Pérez, A. **Ontological engineering A state of the art.** Expert Update: Knowledge Based Systems and Applied Artificial Intelligence 2.3. 33-43. 1999.

JSOUP. **Java HTML Parser.** Disponível em: <http://jsoup.org/> acesso em: 14 de setembro de 2014

Kaisler, S., et al. **Big data: Issues and challenges moving forward.** System Sciences (HICSS), 2013 46th Hawaii International Conference on. IEEE, 2013.

Kakhani, M. K., Kakhani, S., e Biradar, S. R. **Research Issues in Big Data Analytics.** 2013.

Katal, A., Wazid, M., e Goudar, R. H. **Big data: Issues, challenges, tools and Good practices.** Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013.

Martimiano, L. A. F. **Sobre a estruturação de informação em sistemas de segurança computacional.** 2006. 185 f. Tese (Doutorado em Ciências em Engenharia de Sistemas e Computação) – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro. 1998.

Mayer-Schönberger, V., e Cukier, K. **Big data: A revolution that will transform how we live, work, and think**. Houghton Mifflin Harcourt, 2013.

McAfee, A., et al. **Big Data.** The management revolution. Harvard Bus Rev 90.10. 61-67. 2012.

Modesto, L. R. **Representação e Persistência para acesso a Recursos Informacionais Digitais gerados dinamicamente em sítios oficiais do Governo Federal.** 2013. 103 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2013.

Mooers, C. **Zatocoding applied to mechanical organization of knowledge.** American Documentation, Washington, v. 2, n. 1, p.20-32. 1951.

Nakamura, L. H. V. **Utilização de Web Semântica para Seleção de Informações de Web Services no Registro UDDI uma abordagem com qualidade de serviço.** 2012. 148 f. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional). – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos. 2011.

Noy, N. F., e McGuinness, D. L. **Ontology development 101: A guide to creating your first ontology.** 2001.

Owl2Java. **A Java Code Generator for OWL**, 2009.

Prado, S. G. D. **Um Experimento no Uso de Ontologias para Reforço da Aprendizagem em Educação à Distância.** 2004. 177 f. Tese (Doutorado em Engenharia). – Escola Politécnica, Universidade de São Paulo, São Paulo. 2004.

Prazeres, C. V. S. **Serviços Web Semântica: da modelagem à composição.** 2009. 189 f. Tese (Doutorado em Ciência da Computação). – ICMC, Universidade de São Paulo, São Carlos. 2004.

Prescott, J. E. **The evolution of competitive intelligence.** International Review of Strategic Management 6. 71-90. 1995.

Sagiroglu, S., e Sinanc, Duygu. **Big data: A review.** Collaboration Technologies and Systems (CTS), 2013 International Conference on. IEEE, 2013.

Salton, G.; Buckley, C. **Term-weighting approaches in automatic text retrieval.** Information Processing & Management, Oxford v. 24, n. 5, p. 513 – 523, 1988.

Santarem Segundo, J. E. **Representação Iterativa: um modelo para Repositórios Digitais.** 2010. 224 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2010.

Silva, T. M. S. **Extração de informação para busca semântica na web baseada em**

ontologias. 2003.

Stanford University. **Protégé**. Disponível em <http://protege.stanford.edu/> acesso em 3 de maio de 2014.

Souza, R. R., e Alvarenga, L. **A Web Semântica e suas contribuições para a ciência da informação.** Ciência da Informação, Brasília 33.1. 132-141. 2004.

Teo, T. S. H., e Choo, W. Y. **Assessing the impact of using the Internet for competitive intelligence.** Information & management 39.1. 67-83. 2001.

UNICODE. **O que é Unicode?** 2008. Disponível em <http://www.unicode.org/standard/translations/portuguese.html> Acesso em: 25 de agosto de 2014.

Uschold, M., e Gruninger, M. **Ontologies: Principles, methods and applications.** The knowledge engineering review 11.02. 93-136. 1996.

Wiesner, Kevin et al. **Recovery mechanisms for semantic web services**. In: Distributed Applications and Interoperable Systems. Springer Berlin Heidelberg, 2008. p. 100-105.

W3C. **Conhecendo o W3C.** Disponível em <http://www.w3c.br/Sobre/ConhecendoW3C> acesso em 09 de outubro de 2014.

W3C. **The need for a universal syntax.** 2014b. Disponível em <http://www.w3.org/Addressing/URL/uri-spec.html> acesso em 09 de outubro de 2014.

W3C. **XML Essentials.** 2014c. Disponível em <http://www.w3.org/standards/xml/core> acesso em 09 de outubro de 2014.

W3C. **XML Schema.** 2014d. Disponível em <http://www.w3.org/XML/Schema.html> acesso em 09 de outubro de 2014.

W3C. **RDF.** 2014e. Disponível em <http://www.w3.org/RDF/> acesso em 09 de outubro de 2014.

W3C. **RDF Schema 1.1.** 2014f. Disponível em <http://www.w3.org/TR/rdf-schema/> acesso em 09 de outubro de 2014.

W3C. **OWL**. 2014g. Disponível em <http://www.w3.org/TR/owl-features/> acesso em 9 de outubro de 2014.

W3C. **Web Semântica.** 2014h. Disponível em < http://www.w3.org/2001/Talks/0228-tbl/slide5-0.html> acesso em 9 de outubro de 2014.

Zikopoulos, P., e Eaton, C. **Understanding big data: Analytics for enterprise class hadoop**

and streaming data. McGraw-Hill Osborne Media, 2011.