Recuperação da Informação em Ambientes Semânticos: uma ferramenta aplicada à publicações científicas

Caio Saraiva Coneglian, *Universidade Estadual Paulista - UNESP*; Elvis Fusco, *Centro Universitário Eurípides de Marília - UNIVEM*

***Resumo*— Os dados disponíveis na Web estão crescendo exponencialmente, oferecendo informações de alto valor agregado às organizações. Tais informações podem estar dispostas em diversas bases e em formatos variados, como vídeos e fotos em mídias sociais. Contudo, dados não estruturados apresentam grande dificuldade para a recuperação da informação não atendendo eficientemente as necessidades informacionais dos usuários, pois ocorre problemas em compreender o sentido dos documentos armazenados na Web. 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, tratamento e recuperação de informações no contexto do Big Data nas mais variadas fontes informacionais 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.**

***Palavras-Chaves*—web semântica, ontologia, agente de extração, recuperação da informação**

***Abstract* --The data available on the web are growing exponentially, offering information of high value to organizations. Such information may be arranged in different bases and in different formats like videos and photos in social media. However, unstructured data present great difficulty for the retrieval of information not meet efficiently the information needs of users, because it occurs trouble understanding the meaning of Web documents. In the context of an Information Recovery architecture, this research aims to implement a semantic extraction agent in the Web context to allow the retrieval, processing and retrieval of information in the context of Big Data in various informational sources to serve as a basis for the implementation of information environments to assist the process of information retrieval using ontology to add semantics to the recovery process and presentation of results to users, thus being able to meet their needs.**

***Index Terms* -- Semantic web, ontology, extracting agent, information retrieval**

I. Introdução

As instituições e organizações estão diante de um universo onde ocorre um aumento exponencial das informações geradas tanto internamente quanto externamente. O final do século XX iniciou uma era onde praticamente tudo o que é gerado fica disponível digitalmente, se tornando assim um grande desafio para as mais avançadas tecnologias de armazenamento, tratamento, transformação e análise de informações. Testando, assim, as áreas do tratamento e recuperação da informação, no que diz respeito ao volume, variedade e velocidade de uma inundação de dados semiestruturados de natureza complexa.

Dentro desse contexto, surgiu um conceito de Big Data que classifica os dados gerados em ambientes informacionais digitais, principalmente aqueles que tem a Internet como plataforma. Big Data é definido como 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 [22]. Beyer e Laney [2] define também tal conceito, dizendo que o Big Data contém três características essências, o alto volume, alta velocidade e alta variedade das informações, que requerem novos meios para processamento e análise dos dados, permitindo uma melhor tomada de decisão, nova descoberta do conhecimento e otimização de processos.

Coneglian, Fusco e Santarem Segundo [20] exploram e exemplificam as características expostas por Beyer e Laney, dizendo que o alto volume dos dados dentro de Big Data representa conjuntos de dados de grande magnitude; a alta variedade remete a heterogeneidade, complexidade e variabilidade dos dados gerados, podendo ser em formatos de vídeos, figuras, textos entre outros e; a alta velocidade diz respeito ao fluxo constante de consultas geradas em tempo real com informações para a tomada de decisão.

Na publicação do Journal of Science [22] 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 [2] 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 [4] 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 [5] 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 [18], 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 [18] 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.

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.

II. Recuperação da 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 [21].

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 [21].

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 [17].

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 [14].

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.

O termo Recuperação da Informação foi trazido pela primeira vez em 1951, por Mooers [15], 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 [7].

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 [7].

A. 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 [17] 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 se verifica 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 [17].

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.

III. 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 [9] 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 [8] 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 [3] 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 [17] 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 [10] 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 [1] 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.

A. Metodologia para Construção de Ontologias

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

Falbo [6] 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 [13].

Para a construção da ontologia deste trabalho, foi utilizada a Metodologia definida por Noy e McGuiness [16], que explicam uma forma de se realizar a engenharia da ontologia.

Neste contexto Noy e McGuiness [16] 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 e McGuiness [16] 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; 2) Reutilizar Ontologias Existentes; 3) Levantar termos importantes; 4) Definir classes e sua hierarquia; 5) Definir propriedades das classes; 6) Restrições das Propriedades; 7) Criação de instâncias.

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

IV. Arquitetura

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 1.

Figura 1: Arquitetura de Contextualização do Agente Semântico

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 do Agente Semântico e do Espaço Informacional. Tratando da questão de recuperar, processar e utilizar informações diversas.

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 2, 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.

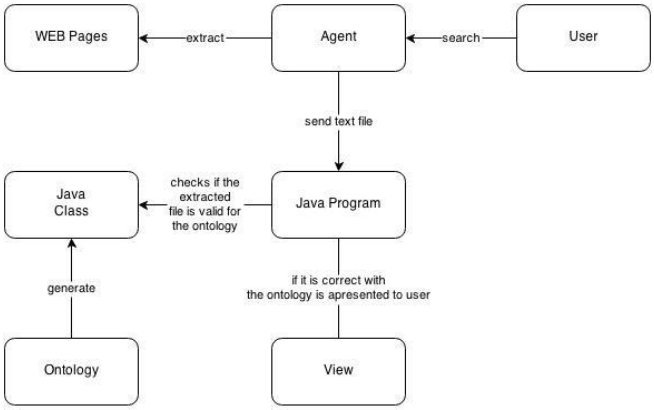

Figura 2: Processo realizado pelo sistema de extração

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.

V. Resultados

A. Construção 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 Gómez-Perez [11], 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 [16], 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.
5. Definir propriedades das classes: este passo não foi realizado devido ao 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 [19], foi realizado a construção da ontologia, onde após a realização da modelagem pelo Protégé, é gerado um arquivo OWL que representa a ontologia.

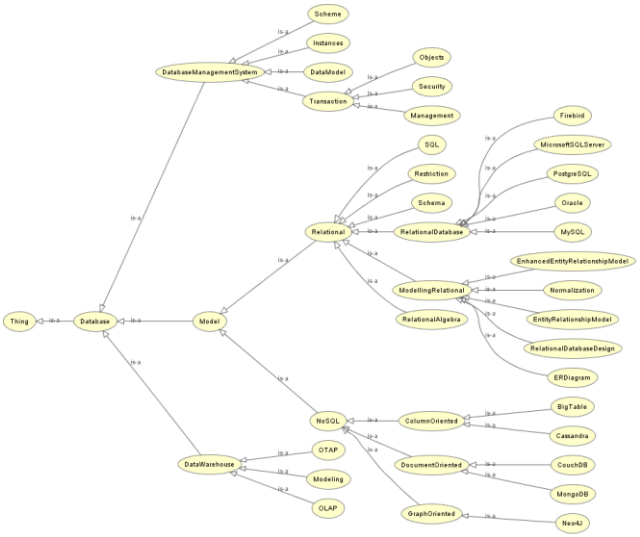

Figura 3: Relação de Classes da Ontologia

A ontologia modelada pelo software Protégé, pode ser visualizado através da figura 3, 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.

B. 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.

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.

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 4 é mostrada como é a página HTML do retorno.

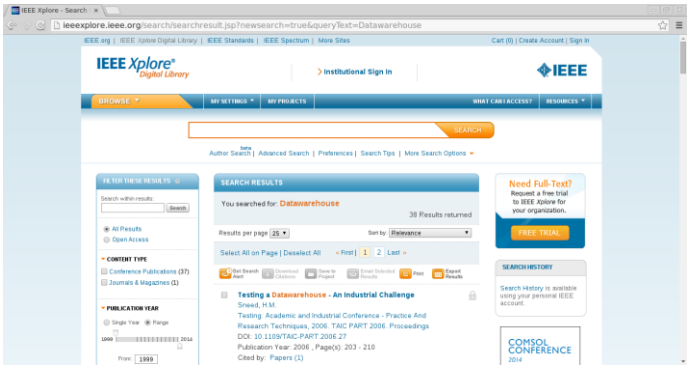

Figura 4: HTML extraído pelo agente

• 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 [12]. 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.

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 5, onde são visualizados apenas as classes que tem relação com o termo pesquisado, no caso Datawarehouse.

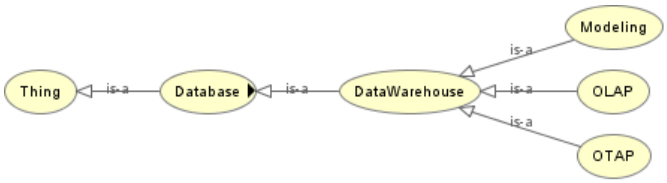

Figura 5: Fragmento da relação de classes da ontologia do termo Datawarehouse

• 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%.

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.

Após o usuário escrever o que ele necessita, 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 6, onde são apresentados os nomes e os links, para que o usuário possa acessar ao artigo completo.

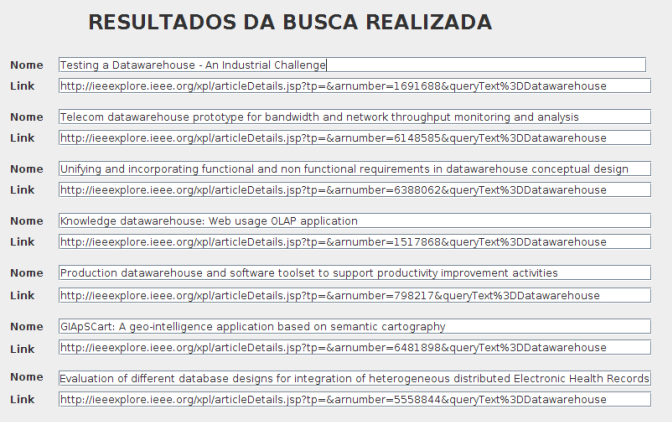

Figura 6: Tela de Resultados Apresentados ao Usuário

C. Testes

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 5.

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

Na tabela 1, 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.

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

Tabela 1: Análise dos artigos extraídos.

No caso de 25 artigos, 7 foram os que atenderam aos requisitos, sendo estes apresentados aos usuários, esta apresentação pode ser visualizada da figura 6.

Para visualizar como o programa faz a análise dos resumos e dos títulos, abaixo na figura 7, é apresentado um artigo dos que atenderam aos requisitos.

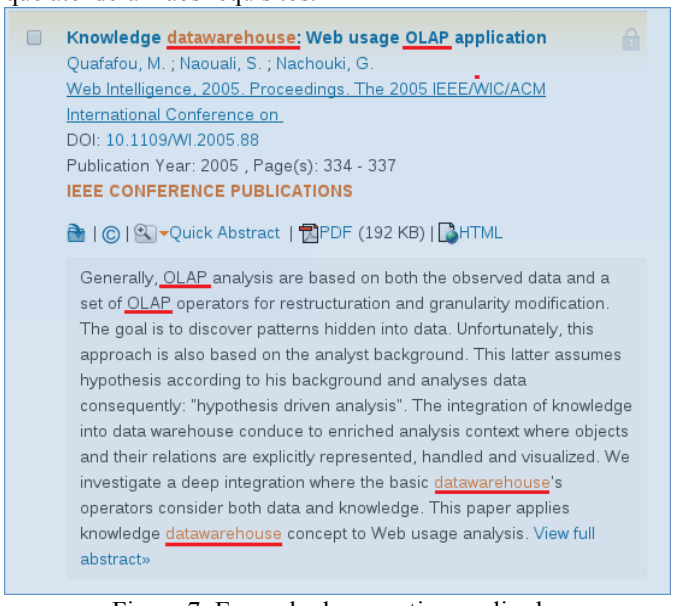

Figura 7: Exemplo de um artigo analisado.

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 7 está grifado em vermelho os termos que foram encontrados pelo programa.

VI. Considerações Finais

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.

VII. AGRADECIMENTOS

O trabalho apresentado neste artigo foi fomentado pela FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo), processo 2013/16369-3.

REFERÊNCIAS

[1] Berners-Lee, T., Hendler, J. e Lassila, O. (2001) *The semantic web.* Scientific american 284.5. 28-37.

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

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

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

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

[6] Falbo, R. A. (1998) *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.

[7] Ferneda, E. (2003) *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.

[8] Gruber, T. R. (1993) A translation approach to portable ontology specifications. Knowledge acquisition 5.2. 199-220.

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

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

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

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

[13] Martimiano, L. A. F. (2006) *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.

[14] Modesto, L. R. (2013) 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.

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

[16] Noy, N. F., e McGuinness, D. L. (2001) Ontology development 101: A guide to creating your first ontology.

[17] Santarem Segundo, J. E. (2010) *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.

[18] Silva, T. M. S. (2003) Extração de informação para busca semântica na web baseada em ontologias.

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

[20] Coneglian, C. S.; Fusco, E.; Santarem Segundo, J. E. (2015) Semantic Information Retrieval Platform Applied To Scientific Papers Extraction. In: CONTECSI 12ª Conferência Internacional sobre Sistemas de Informação e Gestão de Tecnologia.

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

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