DOI: 10.5748/9788599693131-14CONTECSI/COMM-4986

EXTRACTION AND IDENTIFICATION AGENT OF SEMANTIC STRUCTURES IN DIGITAL INFORMATIONAL ENVIRONMENTS

Thiago Aparecido Gonçalves da Costa (Centro Universitário Eurípides de Marília, São Paulo, Brasil) – thiago.gcosta13@gmail.com
Elvis Fusco (Centro Universitário Eurípides de Marília, São Paulo, Brasil) - fusco@univem.edu.br
Marcos Luiz Mucheroni (Universidade de São Paulo, São Paulo, Brasil) - mmucheroni@eca.usp.br
Fábio Dacêncio Pereira (Centro Universitário Eurípides de Marília, São Paulo, Brasil) - prof.fabiopereira@gmail.com
Colaboradores:
Caio Saraiva Coneglian (Universidade Estadual Paulista, São Paulo, Brasil) - caio.coneglian@gmail.com
Edward David Moreno Ordonez (Universidade Federal de Sergipe, Sergipe, Brasil) - edwdavid@gmail.com

In the current scenario of the Internet with the massive production of information, the Information Extraction and Treatment areas are being challenged by the volume, variety and speed of semi-structured and unstructured complex data that must be found and judged as to their value and truthfulness. In this context, the innovation process has become the focus of many companies and is being explored as a means to improve the competitiveness and positioning of companies in new markets. This work aims to establish a computational mechanism for extracting and identifying semantic structures from specific informational sources, where the informational space will be the news site of FAPESP (Foundation for Research Support of the State of São Paulo). In addition, we obtain results ranging from the creation of specific and general semantic extractor to the RDF model for persistence of metadata and data in digital informational environments.

Keywords: innovation, extraction of information, web semantic.

AGENTE DE EXTRAÇÃO E IDENTIFCAÇÃO DE ESTRUTURAS SEMÂNTICAS EM AMBIENTES INFORMACIONAIS DIGITAIS

No cenário atual da Internet com a produção massiva de informações, as áreas de Extração e Tratamento da Informação estão sendo desafiadas pelo volume, variedade e velocidade de dados semiestruturados e não estruturados de natureza complexa que devem ser encontrados e julgados quanto ao seu valor e veracidade. Neste contexto, ao processo de inovação tornou-se o foco de muitas empresas e está sendo explorado como meio de melhorar a competitividade e posicionamento de empresas em novos mercados. Este trabalho tem como objetivo estabelecer um mecanismo computacional de extração e identificação de estruturas semânticas de fontes informacionais específicas, onde o espaço informacional será o site de notícias da FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo). Além disso, obtêm-se resultados que vão desde a criação de extrator semântico específico e geral até o modelo de RDF para persistência de metadados e dados em ambientes informacionais digitais.

Palavras-Chave: inovação, extração de informação, web semântica.

Agradecimentos: Os autores agradecem ao CNPQ pelo apoio financeiro.

1. INTRODUÇÃO

A crescente geração massiva de dados está testando a capacidade das mais avançadas tecnologias de recuperação, armazenamento, tratamento, transformação e análise de informações. As áreas de Gestão e Recuperação da Informação e Apoio à Decisão estão sendo desafiadas pelo volume, variedade e velocidade de uma imensidão de dados semiestruturados e não estruturados de natureza complexa que devem ser encontrados e julgados quanto ao seu valor e veracidade.

O cenário exposto anteriormente evidencia-se pelo fenômeno chamado de Big Data. Desse modo, segundo Manyika et al. (2011) Big Data é definido como um conjunto de dados onde o tamanho está além das capacidades dos bancos de dados típicos, portanto, necessita-se de ferramentas de software específicas para captura, armazenamento, gerenciamento e análise. Além disso, Schroeck et al. (2012) descreve Big Data como sendo uma combinação de volume, variedade, velocidade e veracidade, no qual proporciona para o mercado global atual vantagens competitivas.

Nos ambientes que são caracterizados por Big Data, há a necessidade de empregar novas tecnologias que vão desde a captação dos dados até a persistência, logo, para a obtenção de informação líquida é imprescindível a utilização de mecanismos de extração de dados adaptados a esse novo contexto, já para a persistência os bancos de dados tradicionais não são pertinentes para o processamento, recuperação e armazenamento, portanto, neste cenário é aconselhável a utilização de bancos não relacionais, como MongoDB, Cassandra, DynamoDB, SimpleDB etc.

Segundo Malik et. al. (2011) caracteriza-se os mecanismos de extração de dados ou mecanismos de obtenção de informação líquida como um processo de captação de informações úteis de páginas HTML, onde essas técnicas são semelhantes as utilizadas por motores de busca, mas possuem outro viés que é a estruturação de dados não estruturados e posteriormente sua análise e armazenamento numa base de dados.

Bancos de dados não relacionais mais conhecidos como NoSQL ("*Not Only SQL*") possuem vantagens e desvantagens na utilização em comparação aos bancos relacionais. De acordo com Han et. al. (2016) é vantajoso a sua utilização, pois há o apoio ao armazenamento de grande volume de dados, são fáceis de expandir, tem baixo custo e dispõe de uma leitura e escrita veloz. No entanto, esse tipo de armazenamento falta a possibilidade de transações e relatórios, além de não suportar SQL que é a linguagem de *query* mais comumente utilizada pela indústria. Dessa forma, é evidente que uma das maneiras benéficas de persistência em cenários de Big Data é a utilização NoSQL.

Com a finalidade de adicionar significado ao que foi extraído pelo mecanismo de captação de informação líquida é necessário combinar o robô de extração com conceitos de Web Semântica, no qual fundamenta-se em aderir significado a páginas Web para a manipulação e processamento de conteúdo por computadores. Portanto, ao realizarmos a extração de um conteúdo desejado num domínio específico obteremos dados semiestruturados e uma busca textual nessas informações em vez de ser sintática será semântica, logo o extraído terá significado e valor que será indispensável para a tomada de decisão. A partir disto, a modelagem *Resource Description Framework* (RDF) aparece como solução na busca de inserir semântica neste processo.

O RDF é o modelo de descrição de informação recomendado pela W3C (*World Wide Web Consortium*), utilizado para representar informações disponíveis na Web. Dessa forma, utilizando o SPARQL (SPARQL Protocol and RDF Query Language) linguagem de consulta e protocolo de acesso de dados em RDF há a possibilidade de obtenção de busca

semântica, ou seja, possibilita-se 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 mais relevante.

É inegável que atualmente tem crescido substancialmente o volume dos dados, portanto torna-se um grande desafio para as áreas de Ciência da Computação e Ciência Informação buscar formas de gerir, acessar e controlar as informações contidas em cenários de Big Data.

Desse modo, esta pesquisa tem como objetivo o desenvolvimento de um mecanismo computacional de extração e identificação de estruturas semânticas em ambientes informacionais digitais, onde futuramente com o apoio de uma estrutura de classificação de informação seja possível o desenvolvimento de soluções computacionais que possam recuperar informações que poderão ser utilizadas para apoiar a tomada de decisão nos processos de inovação nas organizações. Os objetivos específicos são:

- Identificar técnicas adequadas para identificação e extração de informações (estruturadas e não estruturadas) que compõem o espaço informacional delimitado;
- Implementar de agentes inteligentes de extração automática/semiautomática de informações (robôs de extração de informações);
- Aplicar semântica nas informações extraídas através de RDF.

Além disso, a metodologia desta pesquisa foi dividida em seis etapas iniciais, logo segue:

- Levantamento bibliográfico, pesquisa de trabalhos correlatos e tecnologias: revisão sistemática sobre temas como: recuperação de informação, robôs de extração;
- Desenvolvimento de robôs de extração para o cenário delimitado;
- Nesta etapa foi utilizada inicialmente a API JSoup, indicada para extração e manipulação de dados a partir de uma URL. Os robôs de extração são capazes de analisar o conteúdo de uma página WEB e a API Jsoup oferece recursos para analisar as TAGs e conteúdo HTML a partir de uma URL definida. Cada mecanismo de extração proposto nesta etapa analisou um elemento do espaço informacional delimitado e obter metadados e dados do mesmo;
- Definir métricas de extração, essas que vão desde a definição do limite da busca até o domínio da extração;
- Mapear e classificar as informações extraídas;
- Nesta etapa os resultados da extração de cada robô foram mapeados em RDF, logo foi utilizada a API RDF disponível na plataforma APACHE Jena.

O domínio de aplicação ou espaço informacional do mecanismo de extração de informação proposto nesta pesquisa será o site de notícias da FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), pois o mesmo possui diversas informações relevantes sobre ambientes de inovação, parques tecnológicos, incubadoras de empresas, centros e núcleos de inovação tecnológica que são assuntos pertinentes que orientam o estudo.

2. REFERENCIAL TEÓRICO

Neste capítulo é descrito o referencial teórico utilizado nesta pesquisa.

2.1 Web Semântica

O cenário da Web em seus primórdios possuía como característica o

desenvolvimento de páginas para internet, onde eram criadas por programadores ou engenheiros com a finalidade de compartilhar informações. No entanto, com o passar do tempo surgiram técnicas e ferramentas que proporcionaram a familiarizados ou não da arte da programação a possibilidade de criar seu próprio site, Breitman K. K. (2005).

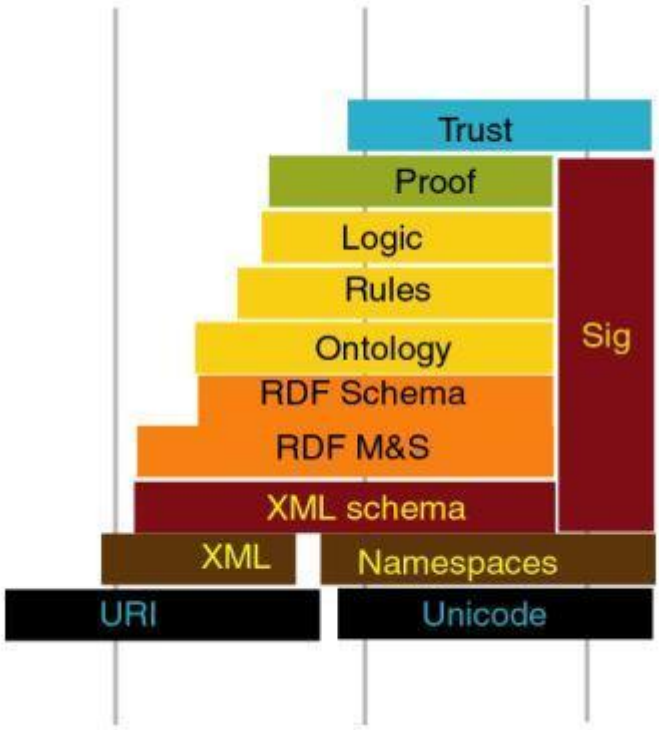
Com o passar dos anos o cenário da Web tende-se a crescer cada vez mais, no entanto atualmente mantém ainda características de seus primórdios em que páginas na maioria das vezes são orientadas para pessoas e não para serem manipuladas ou processadas por outros *softwares*.

A solução para os *softwares* conseguirem interpretarem as páginas e as mesmas serem desenvolvidas direcionadas para os mecanismos informacionais digitais é chamada de Web Semântica, onde tende a ser uma estruturação semântica dos dados na Web para serem processados e interpretados por máquina.

Desse modo, segundo Berners-Lee (2001) em seu artigo denominado "*The Semantic Web*" o autor define que a Web Semântica é uma nova forma de Web que expõe significado para os computadores e consequentemente será uma revolução de oportunidades para o meio. Além disso, o autor exemplifica didaticamente as implicações da Web Semântica no cotidiano das pessoas. Neste exemplo, um usuário necessita marcar uma consulta com um médico de uma determinada área da medicina, portanto ele notifica o computador e lhe informa algumas restrições. A máquina navega pela internet procurando médicos que estejam perto da residência do usuário, que sejam conveniados ao seu plano de saúde e possuam uma boa reputação. Logo, de uma maneira inteligente o computador compara as agendas de consulta do médico e seu horário de atendimento com a agenda do cliente, dessa forma oferece opções de atendimento. O usuário deve se preocupar somente com o horário que melhor lhe convém.

A partir disso, surgem várias formas de se fazer a Web Semântica tornar-se realidade, portanto na imagem a seguir segue o modelo camadas estipulado pelo W3C:

Figura 1 - Camadas da Web Semântica



Fonte: Estrutura da Web Semântica (W3C, 2014h)

Cada camada a seguir é descrita segundo Coneglian (2014, p. 35-36):

- URI (Uniform Resource Identifier – Identificador de Recursos Uniforme): conjunto de caracteres para a identificação de um recurso (W3C, 2014b, apud Coneglian, 2014, p. 35-36);
- Unicode: define um conjunto e padrão universal de codificação (UNICODE, 2008 apud Coneglian, 2014, p.35-36);
- XML (Extensible Markup Language – Linguagem de Marcação Extensível): é um sistema de representação de informação estruturada (W3C, 2014c, apud Coneglian, 2014, p.35-36);
- 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, apud Coneglian, 2014, p.35-36);
- RDF M&S: um modelo para intercâmbio de dados na web, e tem características que facilitam a fusão de dados (W3C, 2014e, apud Coneglian, 2014, p.35-36);
- RDF Schema: um vocabulário para fazer a modelagem de dados de RDF (W3C, 2014f, apud Coneglian, 2014, p.35-36);
- Ontology: é um modelo de dado que representa um conjunto de conhecimento e o relacionamento entre eles dentro de uma base informacional;
- Rules: nela é feita a conversão das informações que estão dentro de um documento para outro, criando regras de inferência (PRADO, 2004,

apud Coneglian, 2014, p.35-36);

- 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.

2.2 JSOUP

Segundo Hedley (2016), Jsoup é uma biblioteca para a linguagem de programação Java que possibilita trabalhar com HTML, onde há a possibilidade de extrair e manipular dados, analise da estrutura HTML de uma URL ou sequência de caracteres, manipulação de elementos HTML e texto, além disso, possibilita-se encontrar elementos estruturais através de passagem de DOM (*Document Object Model*) e CSS.

Ademais, Jsoup implementa a especificação WHATWG HTML5 que proporciona trabalhar com DOM da mesma forma que os navegadores contemporâneos operam, ou seja, proporciona uma árvore de objetos mais sensata possível ao usuário da ferramenta.

2.3 Apache Jena

Apache Jena é um *framework* de código aberto criado pela *Apache Software Fundation* para a criação de aplicações contendo conceitos de Web Semântica e Dados Conectados. O Jena possui uma série de ferramentas, portanto segue algumas ferramentas:

- API de RDF: ferramenta para a criação e leitura de RDF;
- ARQ (SPARQL): mecanismo que proporciona a consulta em RDF utilizando uma arquitetura denominada ARQ;
- API de OWL: ferramenta que possibilita o trabalho com modelos ontológicos descritos em OWL (Web Ontology Language).

3. DESENVOLVIMENTO

Neste capítulo é descrito como foi desenvolvido o agente de extração de estruturas semânticas em espaços informacionais digitais.

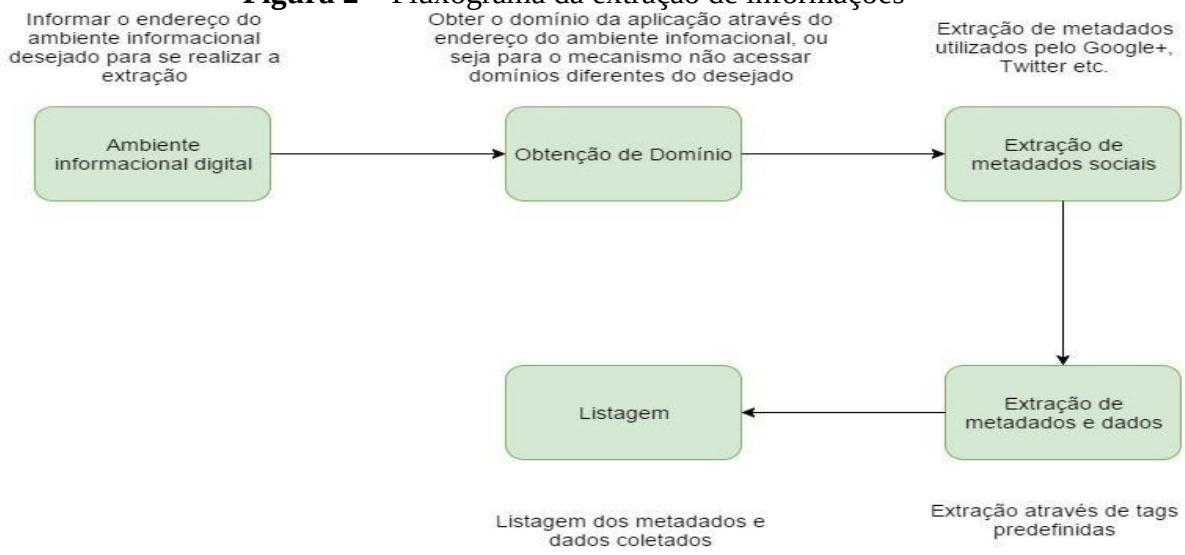
3.1 Espaço informacional

Primeiramente, esta pesquisa utilizou como espaço informacional a seção de notícias do site da FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo), pois o mesmo retém um acervo considerável de informações referentes a inovação, pesquisa científica, parques tecnológicos, centros de inovação, etc.

3.2 Extração da informação

O mecanismo elaborado nesta pesquisa é um *web crawler* que extrai metadados e dados de ambientes informacionais digitais. O agente faz essa busca no HTML da página por *tags* predefinidas, logo caso ache um *link* que esteja dentro do domínio do ambiente informado e não ultrapasse a profundidade de pesquisa definida as informações mineradas são transformadas em cadeia de *String*. Desse modo, podemos observar o processo do mecanismo de extração no fluxograma a seguir:

Figura 2 – Fluxograma da extração de informações



Fonte: Elaborado pelos autores

Como mostrado na figura 2 o mecanismo de extração é dividido em 5 etapas: definição do ambiente informacional, obtenção do domínio, extração de metadados sociais, extração de metadados e dados e listagem contendo listas referentes a metadados e dados de links específicos.

- **Ambiente informacional**: nesta etapa caracteriza-se pela definição de qual será o espaço informacional digital que será extraído e seu endereço Web.
- **Obtenção do domínio da aplicação**: a aplicação desenvolvida trata-se de *web crawler*, logo o mecanismo de obtenção de informação líquida possui a característica de ao encontrar um *hiperlink* guardá-lo em uma lista para posteriormente acessá-lo e realizar essa operação diversas vezes, no entanto, isso pode acarretar um *loop* infinito caso não for tratado quais links o mesmo pode acessar.

  Portanto, o agente pegará o link informado e extrairá o domínio do espaço informacional, por exemplo para o link da seção de notícias do site Inova Marília (http://www.inovamarilia.com.br/category/noticias/) o domínio é (www.inovamarilia.com.br).

  Ademais, o agente ao encontrar um link válido, logo o mesmo adiciona em uma lista de links para serem acessados futuramente, outrora após um endereço ser acessado é inserido em uma outra lista de endereços eletrônicos que já foram conectados. Desse modo, toda vez que for acessar um endereço Web é retirado um da lista de acessados e verificado se o mesmo já foi conectado.

- **Extração de metadados sociais**: A conexão e obtenção de dados intrínsecos dentro da estrutura HTML é feita a partir de uma ferramenta chamada Jsoup no qual é uma biblioteca em Java para manipular e extrair dados, classes, tags, estruturas do HTML advindos de uma página da Internet.
  Portanto, após realizar a conexão com uma determinada página é necessário extrair de sua estrutura as informações que se deseja e atualmente é comum alguns sites possuírem incorporado no *head* algumas *tags* para indexação em redes sociais, mecanismos de busca, etc., desse modo, as tags que estes dados estão inseridos são:
  - "og:title": título da página;
  - "og:description": breve descrição da página;
  - "og:url": endereço eletrônico;
  - "og:site_name": nome do ambiente informacional digital.
- **Extração de metadados e dados**: primeiramente, o mecanismo de obtenção de informação líquida tende a procurar por hiperlinks para guardar em uma lista e acessá-los futuramente, logo foi definida uma política de acesso de endereços Web, onde só é possível acessar aqueles que estiverem no mesmo domínio e não estiverem na expressão regular de restrição (".*\\.(png|jpg|gif|bmp|pdf|ppt|pptx|jpeg|xml|csv)"), nesta expressão é aceito todos os links que não tiverem o término descriminado entre barras.
  Além disso, há outro critério para o acesso de endereços eletrônicos o de profundidade, ou seja, só é permitido o acesso de links que tiverem a profundidade percorrida menor que o informado.
  As *tags* do tipo *head* (h1, h2, h3, h4, h5, h6) segundo o W3C (*World Wide Web Consortium*) são utilizadas em cabeçalhos de páginas Web, dessa forma, levando em consideração que metadados são dados que descrevem dados e no contexto desta pesquisa essas *tags* servem para descrever o conteúdo que será explicitado posteriormente a sua aparição, logo conclui-se que essas *tags* são metadados e seus subsequentes os dados.
  Inclusive, nesta pesquisa adota-se como estrutura HTML específica provedora de conteúdo caracterizado como dado, sendo os "*p*", "*pre*", "*span*", "*i*", "*strong*", "*a*", em que, tais estruturas representam respectivamente: parágrafo, texto pré-formatado, elemento estilizado, itálico, negrito, endereço eletrônico.
  Na figura a seguir demonstra-se numa página tal relação:

Figura 3 - Exemplo da página com metadados e dados

Fonte: Adaptado do site da FAPESP (http://www.fapesp.br/9480)

Dessa forma, foi criado uma classe Java chamada “Dado” que encapsula todas as informações necessárias para prover metadados e dados ao usuário do mecanismo de extração e seus atributos são:

- O tipo da informação extraída (metadado ou dado);
- A *tag* que o conteúdo se encontra;
- O conteúdo;
- O endereço eletrônico do conteúdo extraído do espaço informacional;

As informações coletadas são convertidas para o formato String e é criado um objeto Java da classe explicitada anteriormente para conter todo esse conhecimento adquirido.

- **Listagem**: No final o agente de obtenção de informação líquida cria um objeto Java de uma classe denominada “MDado” que possui como atributos dois objetos da classe “Dado”, ou seja um relacionamento um para um de metadados e dados. Por exemplo: na figura 1 o metadado é “Centros virtuais de pesquisa em agricultura serão financiados”, logo o dado é “A Fapesp anuncia [...]”. Logo, após essa instanciação é adicionado esse objeto no final de uma lista de “MDado”. Inclusive, essa criação de objetos com metadados e dados é realizada para cada dado seguinte. Desse modo, essa lista será utilizada no programa principal para estruturar os resultados no formato RDF através da biblioteca do Jena, onde possui uma API para criar e ler grafos RDF.

Conclui-se que o agente de extração e identificação de estruturas semânticas em ambiente informacionais digitais, extrai e identifica metadados e dados como explicitado anteriormente usufruindo das peculiaridades da estrutura HTML do site, logo o mesmo produz uma lista com todo o conhecimento agregado da extração.

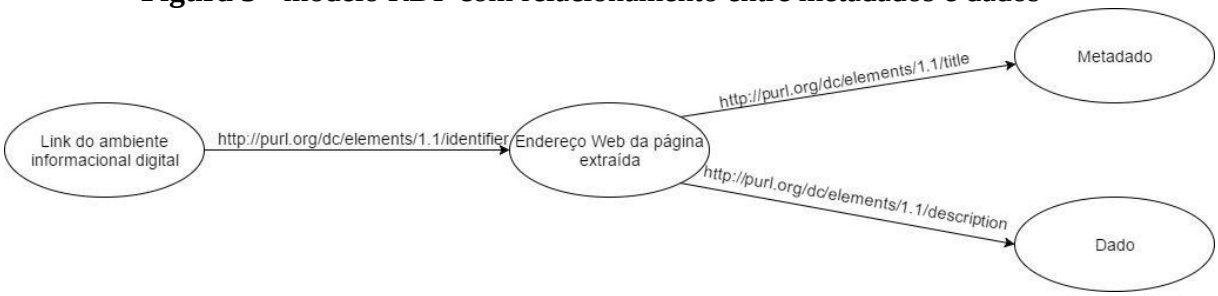
3.3 Modelagem RDF

Para este trabalho foi estipulado um modelo de RDF com base nas informações advindas do agente de extração de informação e no padrão Dublin Core.

Primeiramente, para a criação do modelo RDF é necessário a utilização da lista de resultados obtida pelo robô de extração e seguir este passo-a-passo:

1. Crie um modelo padrão RDF com utilização da biblioteca do Jena;
2. Logo após, elabore um *resource*;
3. Feito isso, basta você adicionar propriedades a este *resource*.

Figura 3 - modelo RDF com relacionamento entre metadados e dados



Fonte: Elaborado pelos autores

Desse modo, na figura anterior representa-se em forma de grafos a estrutura RDF proposta por esta pesquisa, no qual suas propriedades foram elaboradas com auxílio do padrão de metadados Dublin Core, onde possui o objetivo de descrever dados advindos de páginas da internet.

3.4 Integração do agente com o RDF

Para que de fato o mecanismo de extração tenha a semântica proposta pela pesquisa é necessário integrar o robô com a estrutura RDF.

Portanto, precisa-se criar um laço capaz de percorrer a lista com os resultados obtidos anteriormente, onde em cada propriedade do RDF deve atribuir uma informação contida nos objetos. Por exemplo:

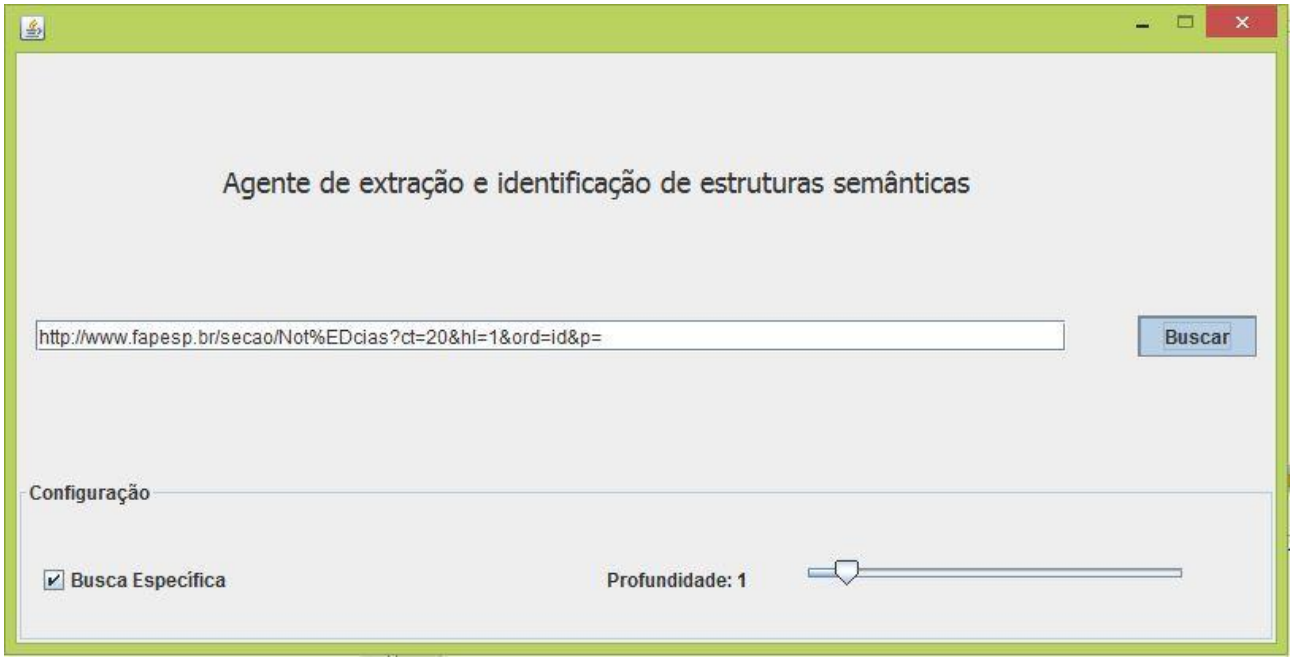
a. *Identifier*, para o endereço web de cada URL extraída;
b. *Title*, apropriado para o conteúdo do metadado;
c. *Description*, apropriado ao conteúdo proveniente do dado.

Feito isso, bastar utilizar um método da própria biblioteca do Jena capaz de exportar para um arquivo de texto ou no próprio console o RDF gerado.

3.5 Interação do usuário com o programa

O usuário ao executar a aplicação depara-se com a primeira tela do programa, onde ela exige que se digite o endereço do ambiente informacional. Além disso, há a possibilidade de se configurar qual será a profundidade da consulta para buscas específicas. Na figura a seguir é possível observar essa tela.

Figura 4 - Tela da interação do usuário com o Agente de Extração e identificação de estruturas semânticas



Fonte: Elaborado pelos autores

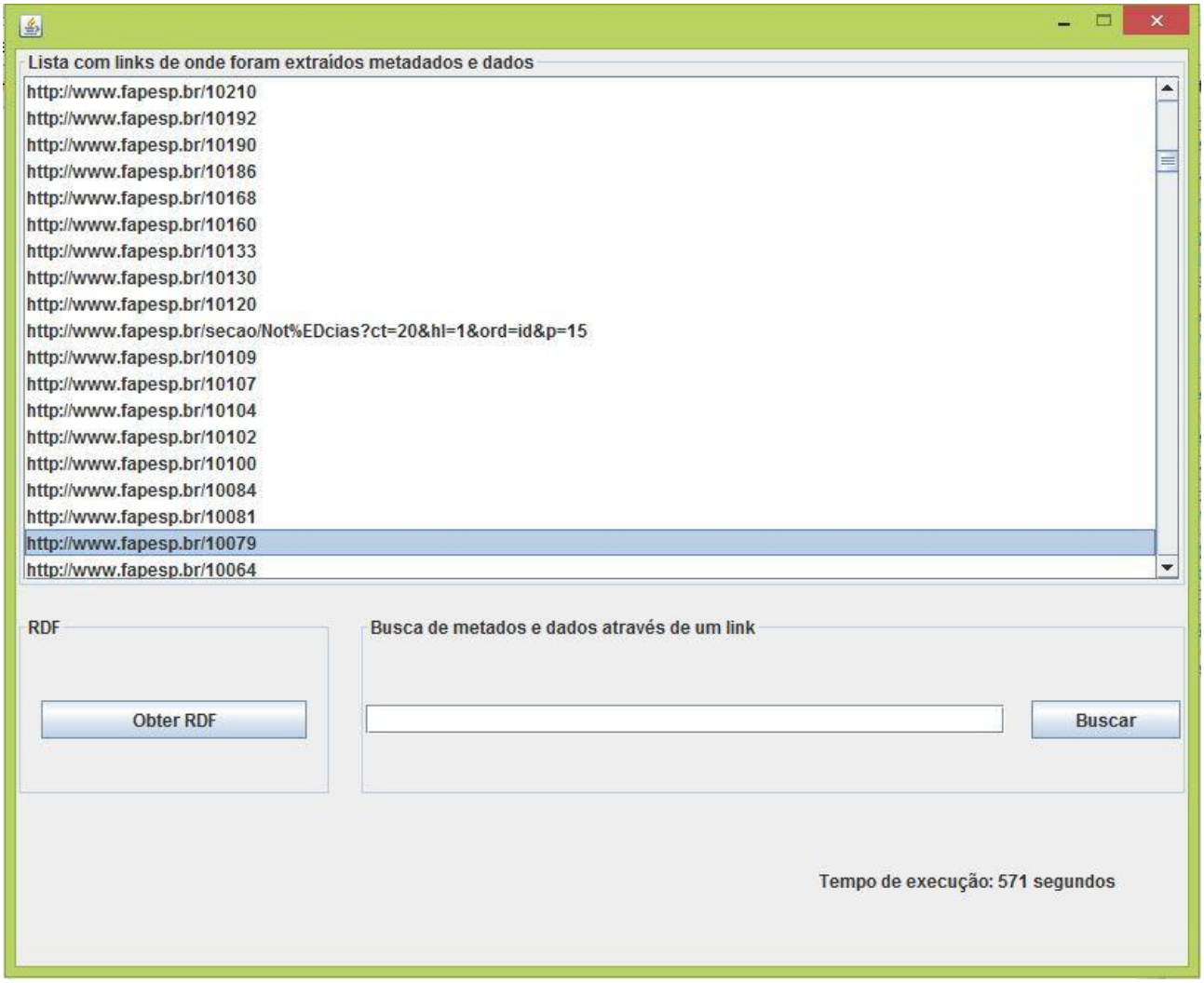
Após o usuário informar qual o endereço do ambiente informacional digital ou caso como no mostrado na figura anterior o caminho de uma determinada seção, por exemplo a seção de notícias da FAPESP, houve a necessidade de demarcar que é uma busca específica e informar ao programa qual será a profundidade.

Desse modo, o agente pegará o link informado e extrairá o domínio do espaço informacional, por exemplo para o link da seção de notícias da Fundação de Amparo à Pesquisa do Estado de São Paulo (http://www.fapesp.br/secao/Not%EDcias?ct=20&hl=1&ord=id&p=) o domínio é (www.fapesp.br).

O mecanismo de extração buscará por *tags* predefinidas de indexação de redes sociais e tags HTML que provém conteúdo para o ambiente informacional, como <h></h> responsável pelo *header* da aplicação, de parágrafo etc. Logo, levando-se em consideração que é mais comumente utilizado para designar o conteúdo que descreve a página a partir de *tags* de cabeçalho é feito uma classificação daquilo que são metadados e dados.

Posteriormente, com uma lista de metadados e dados há a possibilidade de colocá-los em uma estrutura semântica chamada RDF. Após realizar estes passos será aberta uma nova tela contendo o tempo que durou o processo de extração, estruturação e disponibilização da informação via RDF, além de uma lista contendo o endereço Web das páginas que foram extraídas.

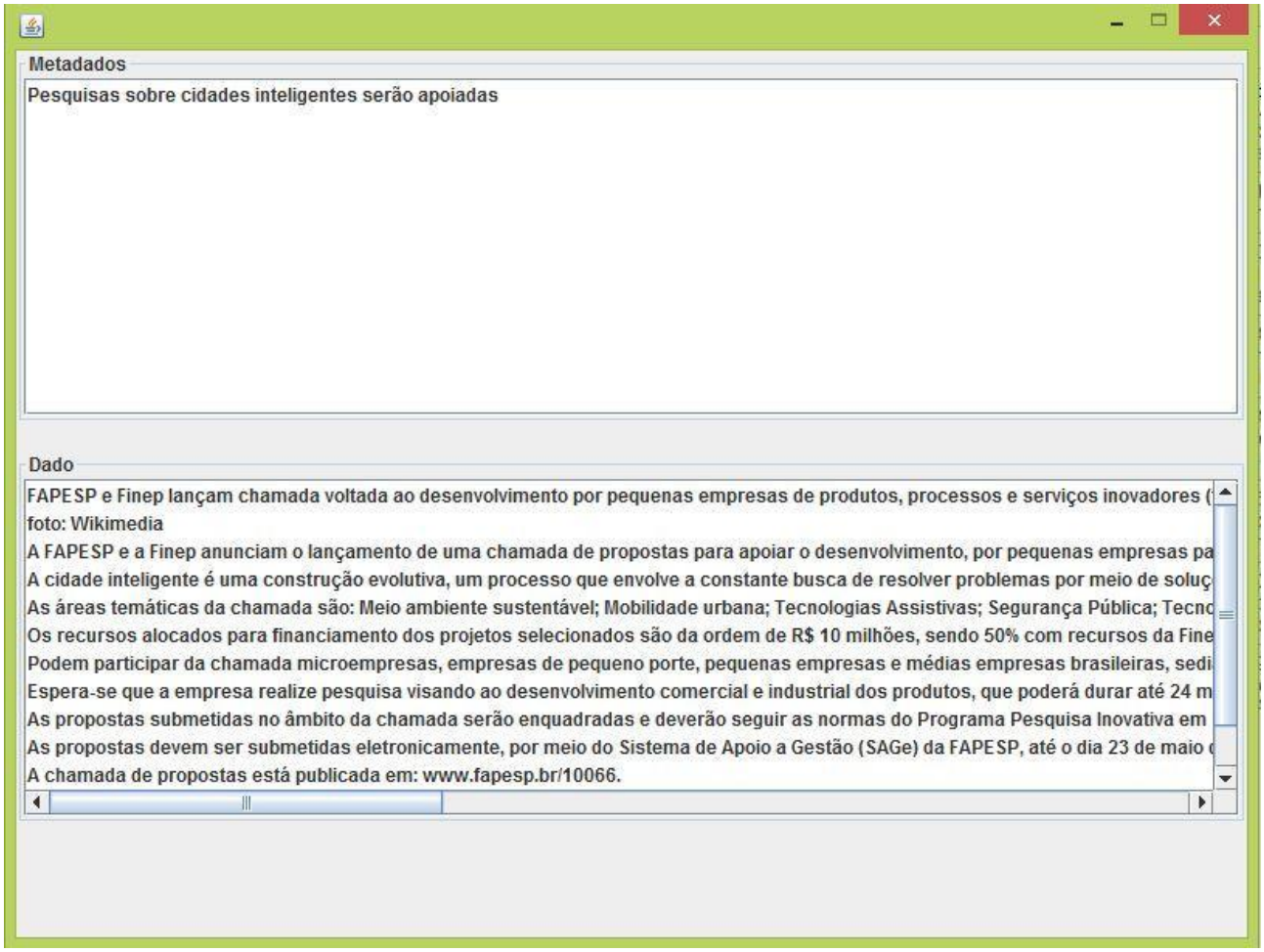
Figura 5 - Tela de resultados do agente



Fonte: Elaborado pelos autores

Portanto, para a visualização dos metadados e dados na íntegra basta colocar um dos links extraídos informados no campo de pesquisa da tela de resultado do agente e clicar no botão "buscar". Portanto, podemos observar tal feito com o link: http://www.fapesp.br/10079.

Figura 6 - Tela de metadados e dados do agente



Fonte: Elaborado pelos autores

3.6 Validação

Para a validação do agente de extração de estruturas semânticas em ambientes informacionais digitais foi realizado uma pesquisa com seis usuários especialistas da área de Ciência da Computação sobre a congruência entre os resultados obtidos com o mecanismo de extração e os metadados e dados que os mesmos encontraram em páginas que foram extraídas.

3.6.1 Metodologia de validação do agente

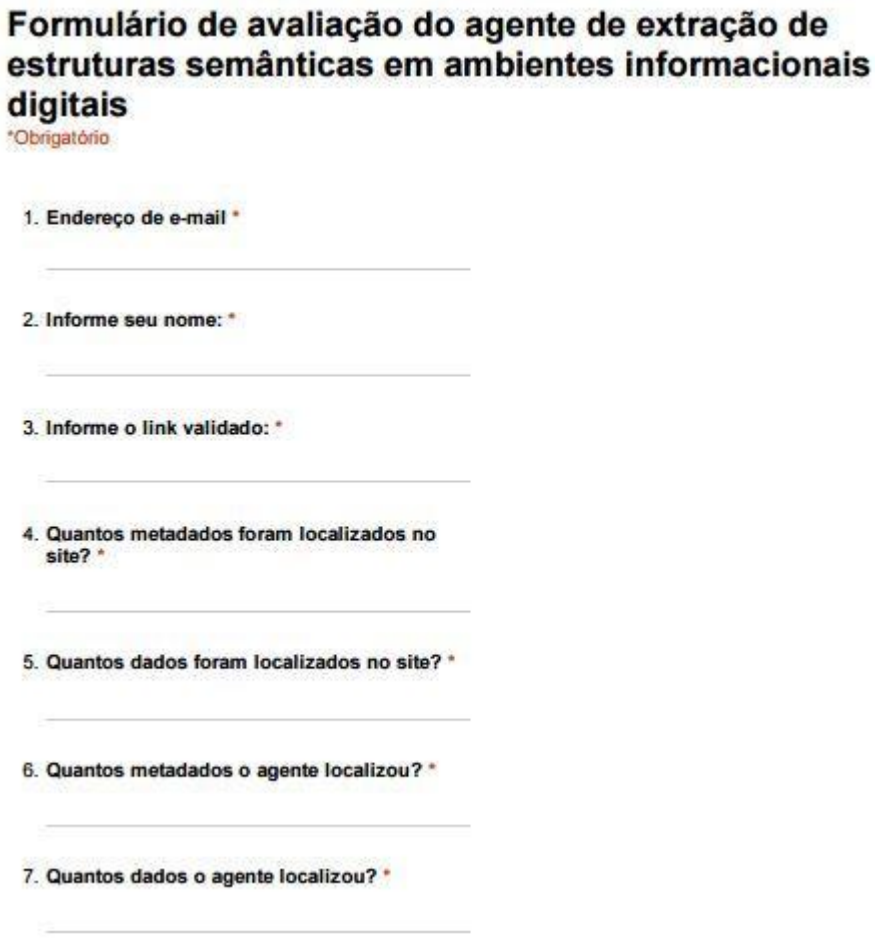
- Criar formulário de avaliação da ferramenta;
- Definir os espaços informacionais que serão utilizados para a extração;
- Definir os usuários especialistas que serão os avaliadores;
- Extrair os metadados e dados das fontes informacionais selecionadas;
- Explicar ao usuário especialista a forma de avaliação da ferramenta e demonstrá-lo manualmente metadados e dados em uma página para reforçar o seu conhecimento;
- Pedir ao usuário para selecionar uma URL aleatória;
- Pedir ao usuário para contar quantos metadados e dados ele acha ao acessar o endereço Web selecionado;
- Analisar a quantidade de metadados e dados que a ferramenta retornou;

- Comparar quantitativamente os resultados obtidos pelo usuário com o da ferramenta;
- Obter gráficos com a comparação dos resultados obtidos.

3.6.2 Formulário de validação da ferramenta

O formulário de validação do agente de extração de estruturas semânticas foi criado com auxílio da ferramenta Google Forms, no qual segue a representação do guia na imagem a seguir:

Figura 7 - Formulário de validação do agente



Fonte: Elaborado pelos autores

Conforme mostrado anteriormente, o formulário de avaliação do mecanismo de extração foi elaborado para que houvesse uma congruência entre os resultados informados pelos usuários especialistas e os que o agente entregou como resposta.

4. RESULTADOS

Como teste para verificar se o agente está extraindo os metadados e dados dos ambientes informacionais digitais corretamente houvesse a necessidade de escolher dois ambientes, logo foram selecionados o site da FAPESP (www.fapesp.br) e o Inova Marília

(www.inovamarilia.com.br).

Dessa forma, como explicitado anteriormente na seção 3.6.2 os entrevistados escolheram links aleatoriamente e responderam o questionário com o auxílio de um avaliador juntamente com o sistema em execução.

Figura 8 - Gráfico da relação entres os resultados extraídos no site da FAPESP

Fonte: Elaborado pelos autores

Figura 9 - Gráfico da relação entres os resultados extraídos e os observados no site do Inova Marília

Fonte: Elaborado pelos autores

Portanto, com base nos gráficos obtidos do primeiro cenário o agente encontrou 13 metadados e os entrevistados 16, já em relação aos dados o mecanismo encontrou 113 e os usuários especialistas 128, desse modo, realize-se dois cálculos de porcentagem sobre os

dois montantes de metadados e dados e uma média aritmética entre os resultantes é possível obter que a acurácia do agente para esse ambiente informacional é de 84,766% com base nos endereços eletrônicos validados.

Além disso, conforme as representações gráficas dos resultados obtidos no segundo espaço informacional digital o agente encontrou 10 metadados e os usuários também 10, já em relação aos dados o robô achou 165 e os usuários especialistas 148, portanto, ao realizar uma conta de porcentagem simples entre os montantes de metadados e dados e com os resultados aplicar uma média aritmética é possível obter o valor de 108,94%.

Conclui-se que o agente de extração e identificação de estruturas semânticas de ambientes informacionais digitais obteve uma exatidão média de 96,853%. No entanto, ao analisarmos o percentual de precisão do segundo ambiente observa-se que o valor ultrapassa 100% e isso ocorre, pois, ao realizar a entrevista com os usuários especialistas considera-se que há erro humano de análise de metadados e dados o que alavanca esse resultado em prol do agente.

5. CONSIDERAÇÕES FINAIS

Este trabalho possui como característica demonstrar a utilização de agentes informacionais e RDF em ambientes digitais.

O objetivo desta pesquisa é o desenvolvimento de um agente de extração semântico para ambientes informacionais digitais, onde a informação capturada pelo mecanismo de extração são os metadados e dados de páginas da Web.

Para o desenvolvimento do mecanismo de extração foi utilizado a linguagem de programação Java e ferramentas, como a biblioteca de manipulação de HTML Jsoup e a API de Web Semântica para a criação de grafos RDF chamada Jena.

Para comprovar este objetivo, foram utilizadas duas bases informacionais digitais que são o site do Inova Marília (http://www.inovamarilia.com.br/) e o da FAPESP (http://www.fapesp.br/), inclusive para restringir o domínio e validar a extração específica do web crawler focou-se somente na seção de notícias desses sites com o grau de profundidade sendo um.

Ao extrairmos ambos os sites deparamos com alguns tipos de dificuldades essas que vão desde a definição de tags para atribuir os metadados e dados, a ausência de metatags sociais o que dificulta a indecisão por redes sociais e a obtenção de uma informação mais concisa referente ao domínio extraído até a carga temporal de processamento por meio do sistema desenvolvido.

Para a validação da ferramenta foram selecionadas seis pessoas da área de Ciência da Computação, onde alguns já possuíam conhecimento referente ao conceito de metadados, dados e Web Semântica, outrora aqueles que não apresentavam tal ideia foram lhes mostradas. Além disso, os questionamentos feitos a eles foram pensados para que seja possível a validação da ferramenta por meios estatísticos e já se esperava algumas incongruências por parte dos avaliados quanto do sistema.

O resultado percentual de validação da ferramenta é considerado satisfatório, já que ultrapassa 95% em quantidade de acertos em relação as informações respondidas pelos entrevistados quanto nas capturadas no ambiente informacional.

Desde o surgimento do conceito de Web Semântica por Tim Berners Lee em 2001 faz-se alguns anos, porém o cenário da Web atualmente tende a não ter se modificado substancialmente, pois grande parte da rede mundial de computadores está organizada de forma sintaticamente e as páginas são criadas com o intuito do ser humano ler determinada

informação e não para outros softwares extraírem o conteúdo.

Portanto, como trabalho futuro para essa pesquisa define-se a utilização de mecanismos ontológicos para aumentar a semântica no processo de extração.

6. REFERÊNCIAS

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

Breitman, K. K., **Web Semântica: A Internet do Futuro**, Rio de Janeiro: LTC, 2005.

Coneglian, C. S. **Agente Semântico de Extração Informacional no Contexto de Big Data**. Marília, 2014.

Han J., Haihong E., Le. G., Du, J. **Survey on NoSQL Database**, 6th International Conference on Pervasive Computing and Applications (ICPCA), 2011.

HEDLEY, J. **Jsoup: Java HTML Parser**. 2009-2016. ed. [S.l.], 2016.

Malik, S. K., Rizvi R., **Information Extraction using Web Usage Mining, Web Scrapping and Semantic Annotation**, International Conference on Computational Intelligence and Communication Systems, 2011.

Manyika, J. et al., **Big data: The next frontier for innovation, competition, and productivity**. McKinsey Global Institute, 2011.

Schroeck, M. et al., **Analytics: The real-world use of big data**, 2012.