

UNIVERSIDADE FEDERAL DE SANTA CATARINA

TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO

**LEONARDO DAITX DE BITENCOURT**

**UM SISTEMA VOLTADO À INDEXAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO  
INTEGRADO À ONTOLOGIA**

**Araranguá, 22 de fevereiro de 2013**

LEONARDO DAITX DE BITENCOURT

UM SISTEMA VOLTADO À INDEXAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO INTEGRADO À  
ONTOLOGIA

Trabalho de Conclusão de Curso submetido à Universidade Federal de Santa Catarina como parte dos requisitos necessários para a obtenção do Grau de Bacharel em Tecnologias da Informação e Comunicação. Sob a orientação do Professor Alexandre Leopoldo Gonçalves.

**Araranguá, 2013**


**Leonardo Daitx de Bitencourt**

**UM SISTEMA VOLTADO À INDEXAÇÃO E RECUPERAÇÃO DE INFORMAÇÃO  
INTEGRADO À ONTOLOGIA**

Trabalho de Conclusão de Curso submetido à  
Universidade Federal de Santa Catarina, como  
parte dos requisitos necessários para a  
obtenção do Grau de Bacharel em Tecnologias  
da Informação e Comunicação.



Professor Alexandre Leopoldo Gonçalves, Dr.  
Presidente da Banca - Orientador



Professora Olga Yevseyeva, Dra.  
Membro



Professor Flávio Ceci, M. Eng.  
Membro

**Araranguá, 22 de fevereiro de 2013**

*Dedico esse trabalho a todas as  
pessoas que acreditam no meu potencial, em  
especial a minha família, pessoas dignas e  
honradas que sempre a meu lado, próximos ou  
distantes, nutriram minhas esperanças de dias  
melhores, senão pelo apoio em minha  
formação acadêmica também pelos exemplos.*

## AGRADECIMENTOS

*Agradeço ao bom Deus que me ilumina e me  
guia.*

*Aos meus pais Darci e Vilma que sempre me  
apoiaram nos estudos.*

*Aos meus irmãos Leandro e Evandro que sempre  
serviram de exemplo para meu empenho, esforço e  
dedicação.*

*Ao professor e orientador Alexandre Leopoldo  
Gonçalves que mesmo com seu tempo escasso  
sempre esteve disposto a auxiliar nesse trabalho.*

*Aos colegas que conviveram comigo o período da  
faculdade dividindo experiências.*

*À família Ghisleri Minatto que muitas vezes me  
deram apoio durante o curso.*

*À Jaini Cândido pelo incentivo.*

*A todos que contribuíram direta ou indiretamente  
para que eu pudesse chegar até aqui, colaborando  
para meu crescimento pessoal e intelectual.*

*Tenha em mente que tudo que você aprende na escola é trabalho de muitas gerações. Receba essa herança, honre-a, acrescente a ela e, um dia, fielmente, deposite-a nas mãos de seus filhos.*

*Albert Einstein*

## RESUMO

O aumento da quantidade de informação disponibilizada tanto na internet quanto nas organizações geram desafios, principalmente se considerada a questão de como recuperar conteúdo relevante. Muitas instituições necessitam de métodos de recuperação de informação aprimorados tendo em vista que a informação tornou-se um recurso essencial e o uso adequado desta é de suma importância em cenários competitivos. Além da questão da utilidade da informação, menciona-se como desafio a própria evolução dos motores de busca uma vez que, para satisfazerem requisitos cada vez mais complexos torna-se necessário a utilização de semântica. Para viabilizar essa evolução, cada documento pertencente a determinado *corpus* necessita ter seus principais conceitos e os seus relacionamentos identificados e armazenados em estruturas adequadas. Entre essas estruturas encontram-se os índices invertidos, para a realização de buscas textuais, e as ontologias, visando a capacidade de realização de inferências. Neste sentido, o presente trabalho apresenta uma proposição de integração das áreas de Recuperação de Informação e Ontologia. Para a avaliação da proposição realizada neste trabalho desenvolveu-se um protótipo e em que este foi aplicado sobre uma base exemplo contendo artigos da área de Ontologia. Visando garantir que o usuário tenha uma visão integrada da informação, o protótipo realiza consultas em duas bases (índice textual e ontologia) de forma coordenada e demonstra de forma mais abrangente as informações que compõem o contexto de consulta por ele informado. Considerando os objetivos do trabalho e analisando os resultados da integração das informações pode-se concluir que o trabalho os cumpre, pois se acredita que o mesmo seja capaz de fornecer informações que contribuem para que o usuário obtenha um entendimento mais completo de determinado domínio de interesse de maneira interativa e iterativa.

Palavras-chave: Ontologia; Indexação; Recuperação de Informação; Busca Semântica.

## **ABSTRACT**

The fast information growing on the Internet as well as on the organizations show challenges, especially when considering the question of how to retrieve relevant content. Many organizations require improved methods of information retrieval taking into account that the information has become essential and its appropriate use is extremely important in competitive scenarios. Beyond that, we can mention the need for evolution of the search engines, whereas in order to meet increasingly complex requirements becomes necessary to use semantics. To make this evolution feasible each document belonging to a particular corpus needs to have its main concepts and its relationships identified and stored in appropriate structures. Among these structures are inverted indexes and ontologies in order to perform textual searches and to make inferences, respectively. In this sense, this work presents a proposition for the integration of Information Retrieval and Ontology areas. For the evaluation of the proposed work it was developed a prototype in which was applied on a sample base containing articles from Ontology area. Aiming to ensure for the user an integrated view, the prototype search for information on two bases (textual index and ontology) in a coordinated way and thus demonstrating more fully the details that compose the context of the search. Analyzing the results obtained from the integration of information it can be concluded that the work has achieved its objectives, since we believe that it is able to provide content that helps users to get a more complete understanding of a particular area of interest.

**Keywords:** Ontology; Indexing; Information Retrieval; Semantic Search.



## LISTA DE ILUSTRAÇÕES

Figura 1 – Representação do conjunto de termos nos documentos .....	33
Figura 2 - Gráfico do cosseno.....	39
Figura 3 - Estrutura de um índice invertido por nome de autores cadastrados na plataforma Science Direct ( <a href="http://www.sciencedirect.com">http://www.sciencedirect.com</a> ) .....	41
Figura 4 - Similaridade entre o modelo relacional e o modelo ontológico.....	45
Figura 5 - Exemplo de uma hierarquia de ontologia .....	47
Figura 6 - Tipos de ontologia de acordo com seu nível de dependência .....	48
Figura 7 - Passos para a construção de uma ontologia segundo a metodologia 101 .....	49
Figura 8 - Organização das metodologias por tipo .....	51
Figura 9 - Representação em Grafo do arquivo RDF .....	53
Figura 10 - Modelo de arquitetura da Web Semântica “ <i>Layer Cake</i> ” proposto por Berners-Lee (2000).....	61
Figura 11 - Exemplo de anotação semântica .....	64
Figura 12 – Visão lógica do sistema de busca semântica .....	67
Figura 13 - Diagramas de atividades do passo 1 ao 3.....	69

Figura 14 - Diagramas de atividades do passo 4 ao 7.....	70
Figura 15 – Visão física do sistema de busca semântica.....	72
Figura 16 - Grafo da ontologia de documentos .....	76
Figura 17 - Exemplo de apresentação dos resultados de uma consulta obtidas da ontologia...	77
Figura 18 - Exemplo de apresentação dos resultados de uma consulta obtidas do índice textual .....	78
Figura 19 - Diagrama de sequência do protótipo.....	79
Figura 20 – Resultado da pesquisa pelo descritor "Ontologies" .....	83
Figura 21 – Resultado da pesquisa pelo descritor "OWL" .....	84
Figura 22 – Resultado da pesquisa pelo documento " <i>Translating the Foundational Model of Anatomy into OWL</i> " .....	85
Figura 23 – Resultado da pesquisa pela autora “Natalya F. Noy” .....	86
Figura 24 – Propriedades do indivíduo Dan Suciú.....	87
Figura 25 - Propriedades do indivíduo Daniel L. Rubin .....	88
Figura 26 - Propriedades do indivíduo Mark A. Musen.....	88
Figura 27 - Rede de relações dos autores que circundam a autora “Natalya F. Noy” .....	89

## **LISTA DE TABELAS**

Tabela 1 - Matriz de termos existentes nas obras de Machado de Assis.....	32
---	----

## LISTA DE ABREVIATURAS E SIGLAS

**API** – *Application Programming Interface.*

**CKML** – *Conceptual Knowledge Markup Language.*

**DAML** – *DARPA Agent Markup Language.*

**FaCT** – *Fast Classification of Terminologies.*

**HTML** – *HyperText Markup Language.*

**IA** – *Inteligência Artificial.*

**IDEF5** – *Integrated DEFinition for Ontology Description Capture Method.*

**KAON** – *Karlsruhe Ontology.*

**NeOn** – *Network Ontologies.*

**OIL** – *Ontology Interchange Language.*

**OML** – *Ontology Markup Language.*

**OWL** – *Web Ontology Language.*

**RDF** – *Resource Description Framework.*

**RDFS** – *Resource Description Framework Schema.*

**RI** – *Recuperação de Informação.*

**SHOE** – *Simple HTML Ontology Extensions.*

**SPARQL** – *SPARQL Protocol and RDF Query Language.*

**SRI** – *Sistema de Recuperação de Informação.*

**SWRL** – *Semantic Web Rule Language.*

**tf-idf** – *Term frequency - Inverse document frequency.*

**TI** – *Tecnologias da Informação.*

**TOVE** – *Toronto Virtual Enterprise.*

**TREC** – *Text REtrieval Conference.*

**URI** – *Uniform Resource Identifier.*

**W3C** – *World Wide Web Consortium.*

**WWW** – *World Wide Web.*

**XML** – *eXtensible Markup Language.*

**XOL** – *Ontology Exchange Language.*

## SUMÁRIO

<b>1. INTRODUÇÃO</b> .....	<b>16</b>
1.1 <i>PROBLEMÁTICA</i> .....	18
1.2 <i>OBJETIVOS</i> .....	20
1.2.1 Objetivo Geral.....	20
1.2.2 Objetivos Específicos.....	20
1.3 <i>METODOLOGIA</i> .....	20
1.4 <i>ORGANIZAÇÃO DO TEXTO</i> .....	21
<b>2. RECUPERAÇÃO DE INFORMAÇÃO</b> .....	<b>23</b>
2.1 <i>HISTÓRICO</i> .....	23
2.2 <i>MODELOS</i> .....	30
2.2.1 Modelo Booleano.....	31
2.2.2 Modelo Vetorial.....	34
2.3 <i>MODELO DE ARMAZENAMENTO</i> .....	39
<b>3. ONTOLOGIA</b> .....	<b>43</b>
3.1 <i>METODOLOGIAS</i> .....	48
3.2 <i>LINGUAGENS</i> .....	51
3.2.1 RDF.....	52
3.2.2 OWL.....	54
3.2.3 SWRL.....	58
3.3 <i>WEB SEMÂNTICA</i> .....	59
3.4 <i>ANOTAÇÃO SEMÂNTICA</i> .....	63
3.5 <i>BUSCA SEMÂNTICA</i> .....	65
<b>4. SISTEMA PROPOSTO</b> .....	<b>67</b>
4.1 <i>VISÃO LÓGICA</i> .....	67
4.2 <i>VISÃO FÍSICA</i> .....	70
4.2.1 Detalhamento do Protótipo.....	78

<b>5. ANÁLISE DOS RESULTADOS.....</b>	<b>81</b>
5.1 <i>CENÁRIO DE APLICAÇÃO</i> .....	81
5.2 <i>DISCUSSÃO</i> .....	82
<b>6. CONSIDERAÇÕES FINAIS.....</b>	<b>90</b>

## 1. INTRODUÇÃO

A popularização das tecnologias da informação e comunicação vem contribuindo diretamente para o aumento expressivo de conteúdo informacional de forma estruturada e não estruturada. O fator determinante para o armazenamento de grande quantidade de informação se deve aos avanços do poder computacional de armazenamento e processamento de dados.

Por outro lado, o crescimento progressivo e intenso de informação desencadeia a necessidade de novas técnicas para armazenamento organizado dos dados visando permitir acesso rápido e coerente a um determinado conteúdo. Bovo (2011) afirma que uma das vantagens de informações em demasia são as oportunidades que o seu uso adequado pode propiciar às pessoas para a tomada de decisões. Isto gera desafios em como armazenar, recuperar, e transformar essa informação em conhecimento. Neste sentido, possuir capacidade de armazenamento e meios eficientes de lidar com a informação torna-se crucial para facilitar processos de tomada de decisão nas organizações.

Segundo Hilbert (2011) em 2007, a humanidade já contava com recursos computacionais capazes de armazenar  $2,9 \times 10^{20}$  bytes, comunicar quase  $2 \times 10^{21}$  bytes e realizar  $6,4 \times 10^{18}$  instruções por segundo em computadores de uso geral e estimativas revelam que no mesmo ano 94% da informação armazenada era digital. Em 2002 mais de 500 milhões de pessoas já acessavam cerca de 3 bilhões de documentos na Web diariamente (FENSEL, 2002).

Conforme Ceci (2010), grande parte dos dados que as organizações possuem estão disponíveis na forma textual e eletrônica, cujo conteúdo se refere a informações e ativos de conhecimento como redes de relacionamento, competências e dados que auxiliam na tomada de decisões e permitem a criação de bases de conhecimento. Contudo, para uma organização



somente o conhecimento não gera lucro, ele só se concretiza com a tomada de ações efetivas (HEBELER; VAN DOREN, 1997). O conhecimento é visto como parte tão fundamental que segundo Nurmi (1998), é aceito como o quarto fator de produção, juntamente com terra, trabalho e capital, tamanha é a sua importância e apesar da produção em massa superar o conhecimento em volume de negócios, este tem um potencial maior em termos de desenvolvimento.

Dessa forma, segundo Cao, Li e Gao (2009), a agregação de recursos de tecnologia da informação no dia a dia das pessoas promove cada vez mais facilidades para se produzir conteúdo. Por outro lado, gera desafios, pois se tornam necessários métodos computacionais apropriados para que o ser humano seja capaz de extrair conhecimentos relevantes das bases de dados, ainda mais porque a informação pode estar fragmentada em diferentes bases de dados, e a busca manual de conteúdo se constituiria em perdas de recursos para a organização.

Informações textuais na Web estão basicamente na forma de artigos científicos e páginas Web, mas trabalhos acadêmicos estão em menor número, enquanto que as páginas Web se proliferam sem controle de qualidade ou custos de publicação (PAGE et al., 1998). Por essas razões, os sistemas de recuperação de informação devem ser capazes de possibilitar buscas em grandes quantidades de informações de maneira adequada, visto que a falta de uniformidade nos dados fornecidos por documentos é um problema para aplicações que precisem navegar no seu conteúdo.

Para Hogan et al. (2011), na perspectiva do usuário os mecanismos de busca atuais estão longe de ser a solução consumada para pesquisas na Web. O autor se justifica ao afirmar que o motivo disso é a falta de respostas diretas a perguntas, em que o resultado é sempre uma seleção de documentos a partir da Web. Afirma ainda que recentemente o Google™ vem oferecendo respostas diretas a determinados modelos comuns de perguntas como “Quanto é 100 dólares em reais?” ou “Quanto é 200 vezes 3700?”, mas essa funcionalidade é restrita a um pequeno subconjunto de consultas.

Existem meios de permitir maior capacidade de consulta com maior relevância às respostas a partir da aplicação de semântica em sistemas de recuperação de informação. Uma dessas maneiras é por meio da representação do conhecimento utilizando-se de ontologias. Ontologia é uma tecnologia chave para permitir processamento de informação orientada a

semântica, e é estimado que a próxima geração de sistemas de gerenciamento da informação vá contar com modelos conceituais na forma de ontologia (MAEDCHE et al., 2003). Neste sentido, representam a organização do contexto de um domínio do conhecimento codificado computacionalmente. Sua definição e utilização serão detalhadas no Capítulo 3.

Contudo, preencher uma ontologia e mantê-la atualizada pode representar um desafio. De certo modo, torna-se necessário a utilização de estratégias que permitam extrair a semântica dos documentos visando manter a ontologia e, conseqüentemente, os sistemas que dela dependem atualizados. A extração significa em si possibilitar a anotação de conceitos e relacionamentos entre estes conceitos diretamente no texto, ou seja, possibilitar a anotação semântica. A anotação semântica de documentos se faz uma técnica crucial para permitir a partilha de informação, troca de conhecimento e maior fidelidade em resultados obtidos da consulta a documentos. Essa técnica define formalmente o conteúdo de textos, o que torna possível a maior precisão e relevância da informação retornada (ZHANG; SHEN, 2009) e adiciona informações para que agentes de software possam atuar sobre seus campos. Entretanto, é ainda um processo muito manual, pois a anotação semântica semiautomática não é capaz de realizar classificação com grande precisão (PIPITONE; PIRRONE, 2012).

## 1.1 PROBLEMÁTICA

Assumindo que o conhecimento é um recurso essencial para uma organização, métodos que objetivam desenvolver o conhecimento e aumentar a capacidade de resolução de problemas e habilidades organizacionais configuram-se em ferramentas que permitem o aumento da competitividade organizacional (STEIL, 2002). Hansen, Bigruam e Tierney (1999) destacam a importância das empresas de gestão do conhecimento para a ascensão do gerenciamento de informações dentro das empresas, criando métodos eficientes de administração do conhecimento e transformando-o em estratégias de gestão.

O autor Goulart (2007) considera a busca coordenada de informações que estão sob o domínio de uma empresa de base tecnológica um desafio que se torna uma oportunidade se realizada de forma coerente, pois aprimora a habilidade decisória dos empreendedores e assegura a capacidade de inovação empresarial por meio de programas estrategicamente relevantes. Essa afirmação estimula a união de esforços para que sejam construídos sistemas

que contem com o auxílio de mecanismos de inferência e portanto consigam realizar buscas por informações de maneira otimizada em bases de dados.

Um problema central em sistemas de recuperação de informação está relacionado à previsão de quais documentos são relevantes e quais não são em uma consulta. Essa decisão geralmente fica por conta de um algoritmo, que por esse motivo se torna parte fundamental no SRI (Sistema de Recuperação de Informação). O algoritmo necessita de uma política de premissas para a determinação da relevância de um documento, e o modelo de recuperação adotado determina as previsões do que é mais ou menos relevante (RIBEIRO-NETO; BAEZA-YATES, 1999; KORFHAGE, 1997; MANNING; RAGHAVAN; SCHÜTZE, 2009).

Os modelos de recuperação de informação utilizados atualmente se baseiam na busca por palavras chaves, dessa forma a importância dos documentos retornados é altamente dependente do quanto o usuário conhece sobre o assunto para realizar as consultas. Essa limitação é inerente ao método sintático de consulta na Web, de forma que o usuário que pouco conhece sobre o domínio que deseja pesquisar deverá inicialmente tomar nota sobre alguns conceitos do domínio para então poder pesquisar o que deseja. Por outro lado, se o usuário tem conhecimento dos termos corretos a serem pesquisados a Web Sintática pode corresponder bem à sua necessidade.

A maioria dos modelos tradicionais de recuperação de informação se baseia na exatidão do que é pesquisado sintaticamente (FANG; ZHAI, 2006), portanto, algumas vezes é necessário um conjunto de termos para que o sistema retorne aquilo que o usuário almeja realmente pesquisar. Uma busca por “jaguar”, por exemplo, nos mecanismos atuais promoverá sempre um questionamento referente a qual jaguar se deseja pesquisar, uma vez, que jaguar é um termo que se refere a mais de um conceito. Isso demonstra um problema decorrente da falta de semântica nos sistemas de busca e indexação de informação.

Desse modo tem-se como pergunta de pesquisa “Como conceber um sistema de recuperação de informação que permita aos seus usuários obterem um melhor entendimento de determinado domínio de análise?”.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo Geral

Propor um sistema de recuperação de informação textual integrado a ontologias que, por meio de inferências, proporcione um melhor entendimento de determinado domínio de análise.

### 1.2.2 Objetivos Específicos

Visando atingir o objetivo principal, alguns objetivos específicos são requeridos, entre eles:

- Pesquisar alguns dos modelos clássicos de recuperação de informação, bem como, estudar modelos que admitem o uso de semântica no processo de recuperação de informação;
- Propor um arcabouço (visão lógica e física) que guie o desenvolvimento de um Sistema de Recuperação Semântica de Informação;
- Desenvolver um protótipo de Sistema de Recuperação Semântica de Informação para demonstrar a viabilidade da integração com ontologia.
- Discutir a implementação das funcionalidades do protótipo, bem como, analisar os resultados obtidos por meio da utilização do mesmo.

## 1.3 METODOLOGIA

O desenvolvimento deste trabalho possui como base uma pesquisa aplicada em que conceitos relacionados às áreas de Recuperação de Informação e Ontologia são utilizados na construção de um sistema que objetiva integrar estas áreas. A metodologia ocorrerá nas seguintes etapas.

- Levantamento do referencial teórico sobre as áreas de Recuperação de Informação e Ontologia;

- Modelagem de uma ontologia de objetivo mais geral capaz de representar o conhecimento contido em documentos;
- Proposição de uma visão lógica e física que possibilite guiar o desenvolvimento do trabalho;
- Construção de um protótipo que permita a integração de sistemas de recuperação de informação tradicionais e ontologias visando possibilitar um ferramental que facilite o entendimento do domínio a que se refere determinada coleção de documentos;
- Avaliação dos resultados obtidos a partir da utilização do protótipo.

#### 1.4 ORGANIZAÇÃO DO TEXTO

O documento foi organizado em 6 capítulos. O primeiro capítulo se constitui em uma introdução do conteúdo, apresentando a importância dos conceitos que são abordados nesse trabalho, assim como a problemática que envolve os sistemas de recuperação de informação atuais, o objetivo geral e os objetivos específicos que compõem esse documento.

É feita uma explanação acerca dos conceitos de recuperação de informação no segundo capítulo, abordando história da evolução das técnicas de armazenamento e recuperação de informação, e modelos que serão utilizados no protótipo desenvolvido para realizar a busca textual no índice.

No terceiro capítulo se discute conceitos de ontologia que abrange as metodologias utilizadas e seus passos sistemáticos para a criação de uma ontologia de forma correta, as linguagens existentes que por meio das quais a ontologia pode ser codificada, conceitos de Web Semântica que foi o que impulsionou e popularizou o uso de ontologias e uma breve explicação sobre a necessidade de anotação semântica para o uso coordenado da ontologia.

A apresentação do sistema proposto foi objeto do quarto capítulo, com visão detalhada da arquitetura no modelo lógico, demonstrando os passos a serem executados pelo protótipo até que a informação chegue ao usuário, e no modelo físico, que apresenta a interação das camadas em um nível descritivo mais técnico para esclarecer quais as ferramentas utilizadas e com que recursos o protótipo é constituído.

A explicação de como se deu o desenvolvimento do protótipo é realizada no quinto capítulo com a análise dos resultados obtidos a partir de testes simulados que puseram à prova a capacidade do protótipo de cumprir aquilo que é proposto, qual seja, apresentar a informação integral sobre um determinado termo pesquisado e o que envolve o termo objeto da busca de acordo com o que é conhecido sobre ele e armazenado na ontologia.

As considerações finais, que é apresentada no sexto capítulo, expõem os pontos do sistema que podem ser melhorados e as possibilidades para a confecção de trabalhos futuros tendo como base o protótipo construído.

## 2. RECUPERAÇÃO DE INFORMAÇÃO

A RI (Recuperação de Informação) pode ser definida como a busca por conteúdo de natureza não estruturada que satisfaz uma necessidade de informação. Não estruturado se refere a recursos (dados) que não possuem claramente uma estrutura semântica para auxiliar no processo computacional. Tais recursos geralmente são documentos de texto presentes em grandes coleções armazenados em computador (MANNING; RAGHAVAN; SCHÜTZE, 2009).

### 2.1 HISTÓRICO

Segundo Frakes e Baeza-Yates (1992), a RI de hoje é suportada por computadores e não se limita a realizar pesquisas por palavras chaves. Numerosas técnicas vêm sendo desenvolvidas nos últimos 50 anos, de forma que já existem estruturas capazes de armazenar índices de grande porte, algoritmos de consulta mais sofisticados para realização mais rápida de pesquisas, métodos de compressão de dados, hardwares específicos, entre outros atributos.

Os recursos de sistemas de recuperação se desenvolveram com o aumento da velocidade dos processadores e capacidade de armazenamento. Um SRI se faz necessário quando uma coleção de dados atinge um tamanho tal que técnicas tradicionais e não automatizadas para organizar conteúdo informacional já não são mais capazes de suportar tamanha informação, e assim como a Lei de Moore, que apresenta uma projeção do aumento contínuo da velocidade dos processadores, há também duplicação da capacidade de armazenamento digital a cada dois anos (SANDERSON; CROFT, 2012). Esse panorama demonstra que são necessários estudos para o aprimoramento maior das técnicas de recuperação de informação para atender demandas cada vez mais específicas.

A RI enquanto área de pesquisa pode ser considerada consolidada e com influência em diversas outras áreas. Singhal (2001) afirma que um grande número de trabalhos sobre RI surgiram em meados da década de 1950, porém registros de estudos relacionados a essa área ocorrem desde meados da década de 1940.

Um exemplo é citado por Vannevar Bush (BUSH, 1945). Ele notou que a ciência proporciona mais rápida comunicação entre os indivíduos, e a documentação do que é descoberto perdura por muitas gerações. O reuso dessas descobertas gera informação em demasia de forma que as publicações se estenderam muito além da capacidade humana de fazer uso real delas. Segundo o autor, a mente humana opera por associação e por isso propôs uma máquina pessoal que realizava busca de informação mecanizada sem métodos de indexação seguindo a lógica do pensamento humano. A máquina recebeu o nome de “Memex” e nela o usuário poderia guardar seus livros e registros e a busca dos documentos era realizada com certa velocidade e flexibilidade tornando-se uma extensão da memória do usuário.

Nota-se que estudos relacionados à recuperação de informação automatizada foram inicialmente realizadas em dispositivos mecânicos, mas já nos anos de 1950 se falava em recuperação de informação utilizando computadores. Segundo Singhal (2001), muitos trabalhos emergiram nos anos de 1950 para a realização de busca de informação em arquivos de texto automaticamente com um computador.

Um exemplo da popularização do uso de técnicas para a recuperação de informação com computadores na década de 1950 é Luhn (1957), que propôs armazenamento virtual de informação, processamento de dados, indexação de formas variadas, dicionários de índices e técnicas abstratas de codificação. O autor concluiu que a comunicação de ideias por meio das palavras é efetuada com base em estatísticas, por isso afirma que ao realizar uma pesquisa pode-se utilizar diferentes técnicas de consulta que varia entre o levantamento de palavras-chave de um texto por edição manual até a análise interpretativa por fórmulas lógicas de conceitos bem definidos. Evidencia-se então que nessa época já havia sido formulada técnicas de utilização de palavras como entidade base para a recuperação de informação em computadores.

Segundo Sanderson e Croft (2012), em 1960 os sistemas de recuperação de informação já se encontravam em aplicações comerciais. Um dos pesquisadores que se



destacou na época foi Gerard Salton que formou e liderou um grupo importante no âmbito da recuperação de informação na Universidade de Harvard (Cambridge, MA), e depois na Universidade de Cornell (Ithaca, NY). O grupo produziu muitos relatórios técnicos que estabeleceram conceitos que ainda hoje são foco de estudo e de pesquisa. Uma dessas áreas é a formalização de algoritmos para calcular o grau de relevância de documentos referentes a uma consulta.

Essa preocupação na avaliação e aprimoramento de uma variedade de métodos automáticos para análise de informação e de pesquisa resultou num projeto de um sistema de recuperação de documentos experimental chamado SMART, cujo diretor era Gerard Salton. No sistema SMART documentos e consultas eram tratadas em uma linguagem automática de análise de conteúdo, que ocorria por diversos procedimentos incorporados no sistema, entre eles o método estatístico e de análise sintática de linguagem (SALTON, 1966).

Outros avanços que ocorreram na RI na década de 1960 foram o agrupamento de documentos com conteúdo semelhante e a associação estatística de termos com significado semântico similar que, em síntese, possibilitou o aumento do número de documentos encontrados com uma requisição por meio da expansão da consulta com variações lexicais, ou com palavras semanticamente associadas (SANDERSON; CROFT, 2012).

A década de 1970 foi marcada pela utilização de técnicas que levavam em consideração as ocorrências de uma palavra num conjunto de documentos. Nessa década Jones (1972) havia determinado que um termo geralmente possui grande variação de significado. Em suas pesquisas o autor constatou que um termo frequentemente usado em um conjunto de documentos funciona como um termo inespecífico. Para haver consultas razoáveis seria necessário utilizar um vocabulário eficaz para uma coleção de documentos com assuntos previamente conhecidos, de forma que fosse suficiente para representar o conteúdo dos documentos individualmente, distinguindo cada um dos demais. Dessa forma um vocabulário auxiliaria na determinação da relevância de cada documento, pois as decisões são influenciadas pelas relações entre os termos e por como o grupo de termos escolhidos caracterizam coletivamente o conjunto de documentos.

Técnicas mais sofisticadas no que concerne premissas probabilísticas surgiram nessa década. A abordagem supracitada ilustra um exemplo disso, quanto maior a frequência de uma palavra em um conjunto de documentos menos importante ela é por ser mais generalista,

enquanto que uma palavra menos comum se refere a informações mais específicas presentes num menor número de documentos, que permite um melhor ajuste de relevância no retorno de informação.

Nos anos 1970 e 1980 muitas tecnologias se desenvolveram e muitos modelos de recuperação de informação foram criados. Neste período as tecnologias de recuperação de informação foram melhoradas com base nos avanços da década de 1960. Os avanços realizados na área ocorreram sobre todos os aspectos a que se referem os fatores decisivos para uma pesquisa eficiente de informações. As novas técnicas que surgiram nesse período mostraram-se eficazes em grandes coleções de pequenos artigos disponíveis aos pesquisadores da época, mas como não se tinha disponibilidade de artigos grandes não se tinha noção de sua escalabilidade das técnicas de recuperação de informação (SINGHAL, 2001).

Em 1975 Salton, Wong e Yang (1975) trabalharam na criação do modelo espaço vetorial. Segundo os autores a recuperação de documentos poderia ocorrer por meio de comparações entre documentos e os dados de entrada fornecidos em uma pesquisa, e tiveram a suposição de que a indexação pode ocorrer por meio da distância entre as entidades, desta forma o valor de um sistema de indexação pode ser expresso como uma função da densidade do espaço objeto, a correlação entre as entidades ocorre inversamente à densidade de espaço.

O modelo de espaço vetorial descreve o processo de recuperação utilizado nos sistemas de investigação, mas atualmente o processo de classificação proposto por Salton são raramente utilizados, todavia, a visualização de documentos e consultas como vetores em um espaço de grande dimensão é ainda comum (SANDERSON; CROFT, 2012).

Outros pesquisadores influentes da década de 1970 foram Robertson e Jones (1976) e Rijsbergen (1979) que também realizaram pesquisas na área de recuperação de informação utilizando modelos probabilísticos. Robertson e Jones examinaram técnicas estatísticas para a exploração de informação relevante pelo peso dos termos pesquisados. Rijsbergen ao escrever seu livro afirmou que a maior parte dos trabalhos em recuperação de informação era não probabilística, somente na década de 1970 é que houve algum processo de implementação significativa com métodos probabilísticos, entretanto, segundo ele métodos probabilísticos já eram mencionadas nos anos sessenta, mas por algum motivo desconhecido as ideias não obtiveram grande aceitação.

A década de 1980 teve desenvolvimentos baseados na década que a precedeu. Nessa década Porter (1980) com base em relatos na literatura da década de 1970 desenvolveu um programa que com base em regras retirava o sufixo das palavras. Em seus estudos afirmou que, partindo do pressuposto de que um documento é representado por um vetor de palavras ou termos, numa coleção de documentos cada documento é descrito basicamente pelo título e eventualmente pelo seu resumo. A remoção de sufixos de palavras por métodos automáticos para realizar buscas é, segundo ele, uma estratégia que é útil na área de recuperação de informação, ignorando-se a questão da origem das palavras.

Conforme Sanderson e Croft (2012), outro evento que destacou a importância da década de 1970 para os desenvolvimentos da década de 1980 são os algoritmos que atribuíam peso aos termos de uma consulta. Foi uma época em que os modelos formais de recuperação foram estendidos e houve a produção de variações de algoritmos que calculavam o peso da frequência de termos e a frequência inversa dos documentos (*tf-idf*), técnica proposta por Luhn (1957). Salton e Buckley (1988) ilustram isso ao resgatar metodologias que realizam ponderações sobre o número de vezes que um determinado termo aparece num documento. Nessa técnica são utilizados vetores que realizam comparações globais entre consulta e vetores de documentos classificados. Um sistema que utiliza essa premissa recupera primeiramente os itens considerados de maior relevância para o usuário.

Mais tarde Deerwester et al. (1990) estabeleceram uma abordagem para indexação automática e recuperação baseada no modelo de indexação semântica latente (Latent Semantic Indexing – LSI) que tentava superar as deficiências da recuperação por tratar a falta de informação que uma palavra oferece sobre seu contexto. Utilizavam-se técnicas estatísticas para estimar esta estrutura latente, e impedir a carência de conteúdo nas palavras além de obter a estrutura semântica dos documentos a fim de melhorar a detecção de informação com base nos termos presentes na consulta. Esse modelo inibiu o problema da pura combinação de palavras de consulta com as palavras de um documento existente nos métodos que existiam até então. Com base em pesquisas os autores perceberam que palavras individuais não fornecem evidências sobre o tema central ou significado de um documento, então utilizaram a descrição dos termos e documentos com base na sua estrutura semântica latente para a realização de indexação e recuperação.

Nos anos 1990, não obstante os avanços sucessivos na recuperação de informação automatizada, estudos mostraram que a maioria das pessoas preferia obter informação de outras pessoas a sistemas de recuperação, porém na última década, o perseverante trabalho de otimização das técnicas de recuperação de informação levou os motores de busca da Web para níveis mais elevados de qualidade, promovendo maior satisfação às pessoas de forma que ao longo do tempo a pesquisa na Web tornou-se o método de busca preferido pelas pessoas (MANNING; RAGHAVAN; SCHÜTZE, 2009).

O princípio da utilização de grandes coleções de documentos aconteceu com a Text REtrieval Conference (TREC) realizada em novembro de 1992. A conferência tinha por objetivo reunir grupos de pesquisa para discutir a respeito da utilização de grandes coleções de documentos de testes em seus trabalhos. No âmbito de recuperação de informação se obteve uma variedade de técnicas de recuperação relatados incluindo métodos usando dicionários automáticos, classificação do peso de um termo, técnicas de linguagem natural, feedback de relevância. Os resultados foram executados por critérios de avaliação uniforme permitindo a comparação da eficácia das diferentes técnicas e análise de como as diferenças entre os sistemas afetaram o desempenho. O TREC foi composto para incentivar a pesquisa em recuperação de informação usando grandes coleções de dados (HARMAN, 1993).

Segundo Singhal (2001) o TREC contribuiu de forma determinante para a criação e modificação de muitas técnicas antigas bem para recuperação de informação sobre grandes coleções de dados. TREC também ramifica a recuperação de informação a domínios relacionados, como recuperação de informação falada, filtragem de informações, interações do usuário com um sistema de recuperação, entre outros.

A década de 1990 foi então marcada com desenvolvimentos que em geral foram essenciais para a construção do que se tem atualmente. Parte desse desenvolvimento aconteceu com a união de esforços para construir ferramentas que viriam a ser apresentadas em conferências como a já mencionada TREC, ou por parte de pessoas visionárias que criaram ferramentas que permitiram grandes desenvolvimentos, como é o caso de Tim Berners-Lee que criou um sistema de hipertexto interligado, a WWW (*World Wide Web*).

Tim Berners-Lee havia criado a WWW no final de 1990, de lá para cá o número de sites disponíveis na Web cresceu muito e para lidar com esse crescimento motores de busca da Web começaram a surgir no final de 1993. O advento da Web iniciou estudos de novos

problemas que abrange a recuperação de informação. Esse período foi marcado pela maior impulsionamento da interação entre comunidades comerciais e pesquisas referentes à recuperação de informação. As ideias criadas até esse período foram sendo mais utilizadas no setor de pesquisas comerciais (SANDERSON; CROFT, 2012).

Com a difusão da WWW surgiu nessa década pesquisas referentes à nova realidade que a recuperação de informação necessitava suprir. O acesso à informação ficou facilitado fazendo com que abordagens já estabelecidas tivessem que ser repensadas para se adaptar a novos problemas levando em consideração a utilização de um usuário leigo.

Segundo Ponte e Croft (1998) apesar de intensos estudos acerca dos conceitos de indexação e recuperação de informação, existia um problema que era a falta de um modelo adequado de indexação. A concepção de um melhor modelo poderia resolver o problema, mas propostas indevidas não trariam resultados satisfatórios, por isso o autor sugeriu recuperação com base em modelagem probabilística de linguagem, avaliando cada documento individualmente. Da mesma forma Hiemstra (1998) apresentou um novo modelo probabilístico de recuperação de informações baseado em modelagens que definem que os documentos e consultas são definidos por uma sequência ordenada de termos individuais, modelo essencial para o processamento da linguagem estatística natural. A autora usou a técnica de ponderação do peso da frequência de termos e a frequência inversa do documento (*tf-idf*) de documentos com uma nova interpretação probabilística de ponderação para se chegar uma melhor compreensão de mecanismos estatísticos de classificação.

Os algoritmos desenvolvidos em recuperação de informação foram os primeiros métodos a serem empregados para a busca na rede mundial de computadores, em 1996 a 1998, entretanto a pesquisa na Web foi aprimorada mesmo em sistemas que se utilizam da ligação cruzada disponível na Web (SINGHAL, 2001). Uma das principais referências dessa abordagem é o PageRank™ que se utiliza dos links que interligam as páginas Web para determina importância a elas.

Segundo Page et al. (1998), quando o PageRank (sistema de classificação do Google) foi proposto, a utilização da contagem de links para eleger a importância de um site já era estudada, mas o PageRank solucionou alguns problemas de precisão e importância que permeava o método de ranqueamento de contagem de links, entretanto o PageRank no seu

estágio inicial ainda assim mantinha alguns gargalos nos cálculos de relevância, que são citados por (GUPTA; JINDAL, 2008).

O número de referências que aponta a um determinado site pode não corresponder à noção de importância das pessoas, pois um único link apontando para um site importante pode fazer um site mais relevante a outro que possui vários links apontando para sites desconhecidos, tornando a técnica inviável (QIAO et al., 2010). O problema de precisão da contagem de links tangia o fato de um documento recém publicado e com poucas referências não poderia ter alta relevância, é o caso das notícias por exemplo (SATO; UEHARA; SAKAI, 2004).

Já nos últimos anos muitas pesquisas passaram a ser feitas para a construção de sistemas de consulta curta, que se limita a algumas palavras, por outro lado, há também pesquisas relacionadas a consultas por meio de perguntas mais próximas da linguagem falada. Nesse período pesquisadores foram desenvolvendo técnicas que fornecem respostas mais focadas para questões mais detalhadas que resultou em aplicações como Siri® da Apple™ e Watson® da IBM™ (SANDERSON; CROFT, 2012).

## 2.2 MODELOS

Segundo Chiaramella e Chevallet (1992) a noção do modelo de recuperação tem para a recuperação de informação a mesma importância fundamental como a noção de modelo de dados num domínio de um banco de dados. Assim, primeiramente um modelo de recuperação preocupa-se com os aspectos de implementação que definem as capacidades e limitações inerentes a qualquer sistema derivado de recuperação de informação. Para o autor um modelo de recuperação pode ser definido como um conjunto de três elementos principais:

- Um modelo para os documentos, que inclui o conteúdo semântico e os atributos contextuais do documento, tais como, autor, editor, título, entre outros;
- Um modelo para as consultas que inclua o conteúdo semântico e os atributos contextuais dentro de uma linguagem de consulta que serão usados para expressar as necessidades de informação dos usuários;
- Uma função correspondente que define a forma com que uma consulta é comparada com qualquer documento modelado, ou seja, a função correspondente que implementa a noção de relevância do sistema.

Dependendo de como os problemas são abordados diferentes técnicas de recuperação de informação podem ser utilizadas. Alguns autores como Amazonas et al. (2008), Osuna-Ontiveros, Lopez-Arevalo e Sosa-Sosa (2011), Poltronieri (2006), Cardoso (2000) e Souza (2006) afirmam que os modelos clássicos para a recuperação de informação são os modelos Booleano, Espaço Vetorial e Probabilístico. Alguns desses autores citam ainda a existência de outros modelos menos utilizados no domínio da recuperação de informação, entre eles, Redes Bayesianas, Indexação Semântica Latente e Redes Neurais Artificiais.

A seguir serão detalhados os modelos booleano e vetorial, pois constituem a base para o desenvolvimento do sistema de recuperação de informação proposto neste trabalho.

### 2.2.1 Modelo Booleano

A lógica como ciência que determina a distinção do verdadeiro e do falso data da época aristotélica e prevaleceu até o século XIX (FERNEDA, 2003). George Boole em 1847 permitiu o que foi considerado o renascimento dos estudos acerca da lógica nos tempos modernos, a partir da publicação de um documento que apresentava uma linguagem simbólica que permitiu mais tarde a criação de uma álgebra não numérica nomeada "Álgebra Booleana" cujos resultados eram baseados em processos dedutivos embasados nos princípios lógicos tradicionais (NEWMAN, 1956). A maioria dos sistemas de recuperação de informação dependem fortemente da capacidade de executar operações booleanas, afinal em sistemas sofisticados o usuário utiliza linguagem natural para realizar consultas e o sistema a converte em expressões booleanas (FRAKES; BAEZA-YATES, 1992).

O modelo de recuperação booleana utiliza consultas de texto livre, ou seja, é utilizada uma linguagem com operadores para construir expressões de consulta. Apesar de décadas de pesquisa acadêmica sobre as vantagens da recuperação com classificação de relevância, o modelo booleano foi o principal modelo utilizado durante três décadas, até 1990 com o advento da *World Wide Web* (MANNING; RAGHAVAN; SCHÜTZE, 2009).

A consulta Booleana é baseada em conceitos de lógica e álgebra Booleana, com termos agrupados por conectivos lógicos, tipicamente AND, OR e NOT, considerados suficientes para expressar qualquer combinação lógica de termos. A pesquisa pode ser expandida com valores decorrentes, isto é, pode se utilizar técnicas para reduzir uma palavra

na sua raiz, bem como fazer uso de um dicionário de sinônimos ou uma lista de termos relacionados (KORFHAGE, 1997).

Segundo Beppler (2008) em virtude do resultado de uma consulta ser binário, o modelo não tem a capacidade de determinar qual documento satisfaz melhor aos termos inseridos numa consulta. Devido a sua condição prática de funcionamento, uma grande vantagem que define esse modelo é o formalismo que oferece e simplicidade com que opera.

O modelo booleano de recuperação de informação possui alguns problemas e limitações. Cooper (1988) afirma que consultas devem ser realizadas de forma não amigável e geralmente as primeiras consultas retornam resultado nulo ou muitos resultados, exigindo alterações na consulta. Destaca ainda que o modelo não possui a determinação de relevância de documentos de acordo com os termos pesquisados e não é possível atribuir peso aos termos inseridos numa consulta.

Abaixo são apresentadas algumas obras do escritor Machado de Assis para ilustrar o funcionamento do modelo booleano. A matriz é representada por colunas representando as obras e linhas representando os termos. Os campos preenchidos com o valor “1” indicam que o termo consta na obra, caso contrário o valor é “0”.

	Quincas Borba	Dom Casmurro	Esaú e Jacó	Memorial de Aires	Memórias Póstumas de Brás Cubas	Ressurreição	A Mão e a Luva
Vagaroso	1	1	0	0	1	0	1
Ciência	1	1	1	0	1	1	1
Enseada	1	0	1	1	0	0	1
Alcunha	1	1	0	0	1	0	0
Fidalgo	1	1	0	0	1	0	1
Generosidade	1	1	1	0	0	1	0
Brasil	1	1	1	1	1	0	0
Talento	1	1	1	1	1	0	1
Valsar	1	0	0	1	1	1	0
Influência	1	0	1	1	1	1	1
Obstáculo	1	1	1	0	0	1	1

**Tabela 1** - Matriz de termos existentes nas obras de Machado de Assis

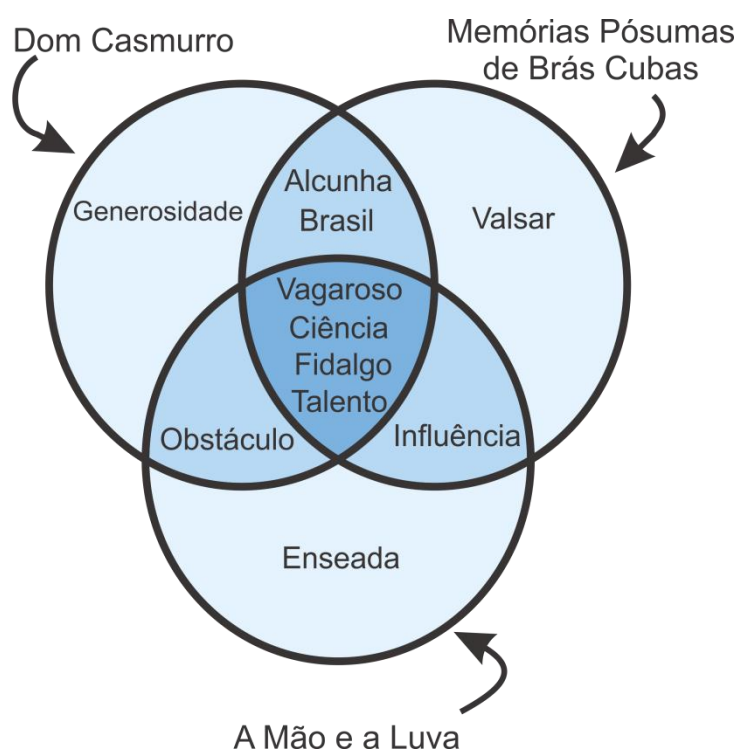
Supondo que o usuário realize uma consulta “Ciência AND Brasil AND Talento AND Alcunha AND NOT Enseada” o algoritmo se apossa dos vetores de cada um dos termos



consultados para determinar quais documentos serão retornados. Então é realizada uma operação AND bit a bit semelhante ao que se segue:

$$1110111 \text{ AND } 1111100 \text{ AND } 1111101 \text{ AND } 1100100 \text{ AND } \text{NOT } 1011001 = 0100100$$

Os documentos que satisfazem a consulta são “Dom Casmurro” e “Memórias Póstumas de Brás Cubas”. Para uma representação gráfica do que ocorre numa consulta booleana usaremos as obras “Dom Casmurro”, “Memórias Póstumas de Brás Cubas” e “A mão e a Luva”. A Figura 1 representa o exemplo do modelo booleano:



**Figura 1** – Representação do conjunto de termos nos documentos

Para o modelo booleano de consulta o documento se constitui numa sequência de termos. Na maioria dos documentos existe uma estrutura adicional denominado metadados responsável por reunir informações sobre o documento como autor, título, data de publicação, formato do documento. Os documentos possuem um conjunto de índices associados a cada campo do metadados e por meio desse mecanismo é possível realizar a seleção dos documentos para satisfazer uma consulta (MANNING; RAGHAVAN; SCHÜTZE, 2009).

Beppler (2008) apresenta uma visão de mais baixo nível do funcionamento do modelo booleano ao afirmar que os documentos são representados por um conjunto de termos indexados, e a recuperação de um documento ocorre somente se seu conteúdo corresponde verdadeiramente a uma consulta com expressão booleana.

### 2.2.2 Modelo Vetorial

O modelo de espaço vetorial é um dos modelos mais utilizados em tarefas de recuperação de informação. Têm grande aceitação principalmente pela sua simplicidade conceitual e o uso da proximidade espacial de linhas retas para determinar a proximidade semântica entre documentos (MANNING; SCHÜTZE, 1999). A autora Cardoso (2000) cita ainda como vantagens desse modelo a facilidade que provê em computar similaridade com eficiência e o bom comportamento em coleções genéricas de documentos.

Segundo Radovanović, Nanopoulos e Ivanović (2010) o modelo de espaço vetorial é popular e amplamente aplicado que representa cada documento como um vetor que armazena pesos dos termos. Com base nos pesos são realizados cálculos para se obter a medida de similaridade que permite realizar a listagem de documentos pesquisados em ordem de relevância. Segundo Manning, Raghavan e Schütze (2009) o modelo vetorial é a representação de um conjunto de documentos como vetores que armazenam a importância relativa de um termo no documento em um espaço vetorial. Este modelo permite operações de recuperação de informações que variam quanto à relevância de um determinado documento a partir de uma consulta, assim como a possibilidade de utilização em tarefas de classificação e agrupamento de documentos.

Segundo Gomes (2009) classificação e agrupamento de documentos são técnicas de mineração de textos. A classificação é o processo de aprendizagem que mapeia dados de entrada em classes de saída e é considerada uma abordagem de aprendizado supervisionado por causa do uso de informações pré-classificadas. De acordo com Maia e Souza (2010) agrupamento é a classificação de documentos baseado em análises comparativas em um determinado número de classes, realizando o agrupamento de documentos, também denominado *clusterização*, sem se basear em algum agrupamento prévio e é isso que lhe confere a condição de considerá-lo um algoritmo de mineração de dados não supervisionado. Gomes (2009) afirma que agrupamento é a organização de documentos agrupados conforme o grau de associação de similaridade entre os membros do grupo.

O modelo de espaço vetorial não é um modelo rígido e por isso permite variações na definição de qual método será utilizado na atribuição de peso aos termos, com representantes proeminentes que são a técnica de ponderação *tf-idf* e de similaridade de cosseno (RADOVANOVIĆ; NANOPOULOS; IVANOVIĆ, 2010).

Segundo Manning e Schütze (1999) esse modelo representa documentos e consultas num espaço  $n$ -dimensional, em que o número de dimensões é igual ao número de termos presentes em um documento ou consulta. Os documentos mais relevantes nessa abordagem são aqueles que têm o menor ângulo de distância do vetor de consulta, ou seja, os documentos cujo vetor, considerando uma série de variáveis que serão apresentadas, apresenta maior grau de similaridade com o vetor de consulta.

Além disso, documentos e consultas são representados por vetores de termos. Os vetores de termos armazenam ocorrências únicas dos termos dos documentos ou consultas e a partir do vetor é realizado um conjunto de operações para determinar a similaridade de um documento com uma consulta (CARDOSO, 2000).

Na literatura este é um modelo conhecido como *saco de palavras* (bag of words) porque a ordem dos termos inseridos na consulta é ignorada, mas o número de ocorrências de cada termo é válido, ao contrário do modelo Booleano que não se importa com o número de palavras num documento, mas somente se o termo existe ou não. Sob este ponto de vista numa consulta *Leandro é mais alto do que Evandro*<sup>1</sup> o documento com a expressão *Evandro é mais alto do que Leandro* também seria retornado (MANNING; RAGHAVAN; SCHÜTZE, 2009).

Segundo Alzghool e Inkpen (2008) e Cardoso (2000) o cálculo de frequência de documento de um termo acontece com o cálculo que define a frequência inversa do documento para dimensionar o peso de um termo. O cálculo é feito da seguinte forma:

$$idf_t = \log \frac{N}{DF_t}$$

Onde  $N$  é o número total de documentos numa coleção,  $DF$  é a frequência do documento e  $t$  representa o termo, de forma que o termo que é menos frequente possui um  $idf$  alto, um termo mais frequente terá um  $idf$  mais baixo.

Existem variadas formas para a obtenção do  $tf$ . O  $tf$  é a frequência do termo, ou seja, o número de vezes que um termo aparece num documento. Alguns cálculos realizados para

---

<sup>1</sup> As consultas propostas estão destacadas porque a utilização de aspas pode ser entendida como parte da consulta. A utilização das aspas alteraria na consulta, pois os motores de busca atuais entenderiam os termos como uma frase e só retornariam os documentos que tivessem as palavras na mesma ordem.

conseguir esse valor são abordados por Alzghool e Inkpen (2008); Wang e Merialdo (2010); Czauderna et al. (2011) e Maiti, Mandal e Mitra (2011).

A obtenção do peso individual dos termos em cada documento se faz com a combinação entre a frequência do termo e a frequência inversa do documento. Segundo Amati e Rijsbergen (2002), Paltoglou e Thelwall (2010) e Czauderna et al. (2011) a ponderação *tf-idf* atribui a um termo  $t$  um peso em relação à um documento  $d$  dado pela fórmula:

$$w_{t,d} = tf_{t,d} * idf_t$$

onde  $w$  representa o peso do termo no documento. Se o termo  $t$  possuir muitas ocorrências em poucos documentos  $d$ , apresentará um peso maior, caso possua poucas ocorrências em um único documento, ocorrer em muitos documentos ou em todos terá um peso menor.

Após a ponderação têm-se cada documento como um vetor com cada posição correspondendo ao peso de cada termo fornecido pela fórmula do *tf-idf* no documento. Após todos os vetores formulados obtêm-se uma matriz que tem dimensões  $d \times t$  (documento x termo), essencial para classificação dos documentos (MANNING; RAGHAVAN; SCHÜTZE, 2009).

Os pesos dos termos nos documentos são necessários para realizar a operação do cosseno de similaridade nos termos. O *tf-idf* é somente uma normalização para apresentar uma matriz resultante de pesos cujo qual algum método matemático os utilizará para determinar a similaridade entre os vetores que compõem as linhas dessa matriz. Os autores Egghe e Michel (2002), Jones e Furnas (1987), Salton e Buckley (1988) e Korfhage (1997) discutem sobre alguns modelos mais conhecidos de cálculo de similaridade ordenada, ou seja, que determinam relevância por meio da proximidade entre vetores, entre eles existem a medida de Jaccard, Dice, cosseno, medidas de sobreposição (*overlap measure*), produto dos pontos, modelo difuso, modelo probabilístico. Nesse documento a técnica de similaridade que será apresentada é a do cosseno.

Segundo Gonçalves (2006) o cosseno calcula o ângulo de distância de similaridade entre dois vetores, com resultados que variam entre 1.0 ( $\cos(0^\circ) = 1.0$ ) para vetores apontando na mesma direção, 0.0 ( $\cos(90^\circ) = 0.0$ ) para vetores ortogonais e -1.0 ( $\cos(180^\circ) = -1.0$ ) para

vetores apontando em direções opostas, mas neste trabalho a abordagem utilizada será somente de variância entre 0.0 e 1.0. O cálculo do cosseno é dado pela seguinte fórmula:

$$\cos \theta(vc, d) = \frac{\sum_{i=1}^t (w_{i,vc} * w_{i,d})}{\sqrt{\sum_{k=1}^t (w_{k,vc})^2} * \sqrt{\sum_{j=1}^t (w_{j,d})^2}}$$

onde,  $w_{i,vc}$  e  $w_{k,vc}$  representam o vetor de frequências normalizadas dos  $i_{th}$  e  $k_{th}$  termos do vetor que se refere à consulta, e  $w_{i,d}$  e  $w_{j,d}$  representa o vetor de frequências normalizadas dos  $i_{th}$  e  $j_{th}$  termos do vetor que se refere ao documento.

Para esclarecer o funcionamento do cálculo do cosseno, assume-se que se têm uma coleção com 2 documentos denominados doc1 e doc2 com conteúdos “*O relógio pode despertar incessantemente a qualquer momento.*” e “*O despertar de uma pessoa esgotada é um ato penoso.*” respectivamente. A partir de uma técnica de restrição de um domínio de conteúdo utilizando-se de axiomas ou taxonomias presentes numa ontologia pode-se extrair palavras mais relevantes que definem um documento em uma coleção segundo um conjunto de regras.

Vamos assumir que foi extraído de cada documento um vetor que se restringe a dois termos e são capazes de representar fielmente os documentos dentro de um domínio do conhecimento existente em uma coleção. Uma consulta realizada com os termos “relógio” e “despertar” formaria um vetor de consulta dessa forma:  $cons = \{(\text{relógio}, 0.7), (\text{despertar}, 0.6)\}$ . O vetor é representado da seguinte forma:  $vetor = \{(\text{termo } 1, \text{ peso } 1), \dots, (\text{termo } n, \text{ peso } n)\}$ , com pesos de qualquer valor entre 0 e 1.

Após a realização da ponderação do *tf-idf* se obtém os vetores dos documentos da coleção. Iremos supor que os vetores resultantes sejam  $doc1 = \{(\text{relógio}, 0.6), (\text{despertar}, 0.4)\}$  e  $doc2 = \{(\text{relógio}, 0.0), (\text{despertar}, 0.7)\}$ .

A partir desses dados é possível realizar o cálculo de similaridade a partir da fórmula do cosseno. A seguir é apresentado o cálculo realizado entre o vetor de consulta e o doc1:

$$\cos \theta(cons, doc1) = \frac{(0,7 * 0,6) + (0,6 * 0,4)}{\sqrt{(0,7)^2 + (0,6)^2} * \sqrt{(0,6)^2 + (0,4)^2}}$$

$$\cos \theta(\text{cons}, \text{doc1}) = \frac{0,42 + 0,24}{\sqrt{0,49 + 0,36} * \sqrt{0,36 + 0,16}}$$

$$\cos \theta(\text{cons}, \text{doc1}) = \frac{0,66}{\sqrt{0,85} * \sqrt{0,52}}$$

$$\cos \theta(\text{cons}, \text{doc1}) = \frac{0,66}{0,6648}$$

$$\cos \theta(\text{cons}, \text{doc1}) = 0,99$$

$$\cos \theta(\text{cons}, \text{doc1}) = 99\% \text{ de similaridade}$$

A mesma operação deve ser realizada com o vetor de consulta e o doc2:

$$\cos \theta(\text{cons}, \text{doc2}) = \frac{(0,7 * 0,0) + (0,6 * 0,7)}{\sqrt{(0,7)^2 + (0,6)^2} * \sqrt{(0,0)^2 + (0,7)^2}}$$

$$\cos \theta(\text{cons}, \text{doc2}) = \frac{0,0 + 0,42}{\sqrt{0,49 + 0,36} * \sqrt{0,0 + 0,49}}$$

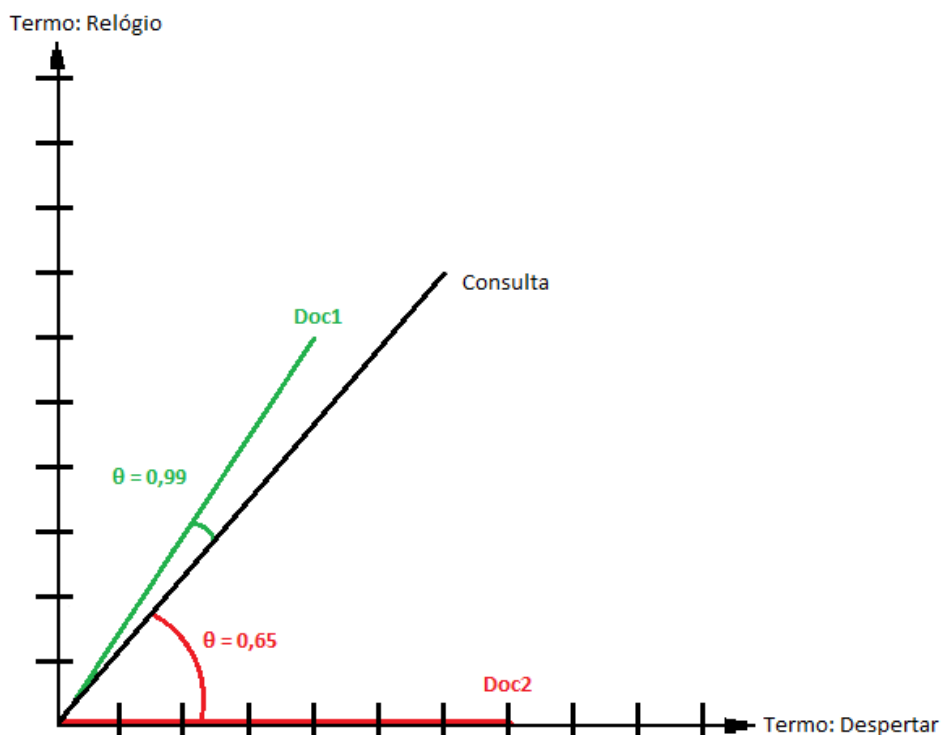
$$\cos \theta(\text{cons}, \text{doc2}) = \frac{0,42}{\sqrt{0,85} * \sqrt{0,49}}$$

$$\cos \theta(\text{cons}, \text{doc2}) = \frac{0,42}{0,6453}$$

$$\cos \theta(\text{cons}, \text{doc2}) = 0,65$$

$$\cos \theta(\text{cons}, \text{doc2}) = 65\% \text{ de similaridade}$$

Na Figura 2 pode ser observada uma representação gráfica do resultado obtido no cálculo para determinar a porcentagem de similaridade dos vetores. A apresentação do gráfico teve como base o gráfico apresentado no livro de Manning, Raghavan e Schütze (2009).



**Figura 2** - Gráfico do cosseno

O gráfico possui duas dimensões por ter sido utilizado dois termos no vetor de consulta. Aplicando os vetores ao cálculo de similaridade é possível verificar a similaridade dos vetores em porcentagem, visto que o resultado sempre ficará entre 0 e 1 (incluindo esses valores). No exemplo apresentado chega-se a um resultado de 99% no grau de similaridade entre doc1 e o vetor de consulta ao passo que doc2 possui similaridade de 65% com o vetor de consulta. Dessa forma é possível estabelecer que o doc1 é mais relevante que o doc2

### 2.3 MODELO DE ARMAZENAMENTO

Segundo Lima (2003) indexação envolve atividades cognitivas para a compreensão do texto e a composição da representação do documento. A indexação pode ser realizada sobre texto, imagens e formatos de áudio e vídeo (ZOBEL; MOFFAT, 2006) e gera um índice. O índice é criado a partir de uma série de passos em qualquer que seja o tipo de dado indexado. O primeiro é o *Tokenize* (Análise Léxica) que consiste separação e armazenamento de *tokens*, nesse processo pode ser retirado acentos das palavras e também fazer com que fiquem em caixa baixa, o segundo é o *Analysis* que consiste em retirar *tokens* pouco relevantes em pesquisas como artigos, preposições, pontuações, espaços em branco, entre

outros e em terceiro e último o *Stemming* (radicalização) que consiste em reduzir os *tokens* remanescentes em sua palavra base (AMAZONAS et al., 2008).

Um *token* é uma sequência de caracteres que podem ser tratados como uma unidade na gramática estabelecida na linguagem de programação (APPEL, 2004). Nenhum mecanismo de busca indexa texto diretamente, para isso acontecer o texto precisa ser dividido em elementos atômicos individuais que são os *tokens*. Cada *token* corresponde aproximadamente a uma palavra. O processo de indexação pode ser subdividido em quatro passos: (a) aquisição do conteúdo, (b) construção do documento, (c) análise do documento e (d) indexação do documento. É durante a fase de análise do documento que é definido como que os campos textuais no documento serão divididos dentro de uma série de *tokens*, para só então os *tokens* serem indexados (escritos dentro de arquivos no índice) numa arquitetura segmentada (HATCHER; GOSPODNETIĆ; MCCANDLESS, 2009).

Christen (2012) cita cinco técnicas de indexação mais recentemente desenvolvidas, entre elas o “(1) *Traditional Blocking*”, “(2) *Sorted Neighbourhood Indexing*” que possui três abordagens, a “(2.1) *Sorted Array Based Approach*”, “(2.2) *Inverted Index Based Approach*” e “(2.3) *Adaptive Sorted Neighbourhood Approach*”, “(3) *Q-gram Based Indexing*”, “(4) *Suffix Array Based Indexing*” que possui como variação a “*Robust Suffix Array Based Indexing*”, “(5) *Canopy Clustering*” que tem como variação duas abordagens que são “(5.1) *Threshold Based Approach*” e “(5.2) *Nearest Neighbour Based Approach*” e finalmente “(6) *String-Map Based Indexing*”.

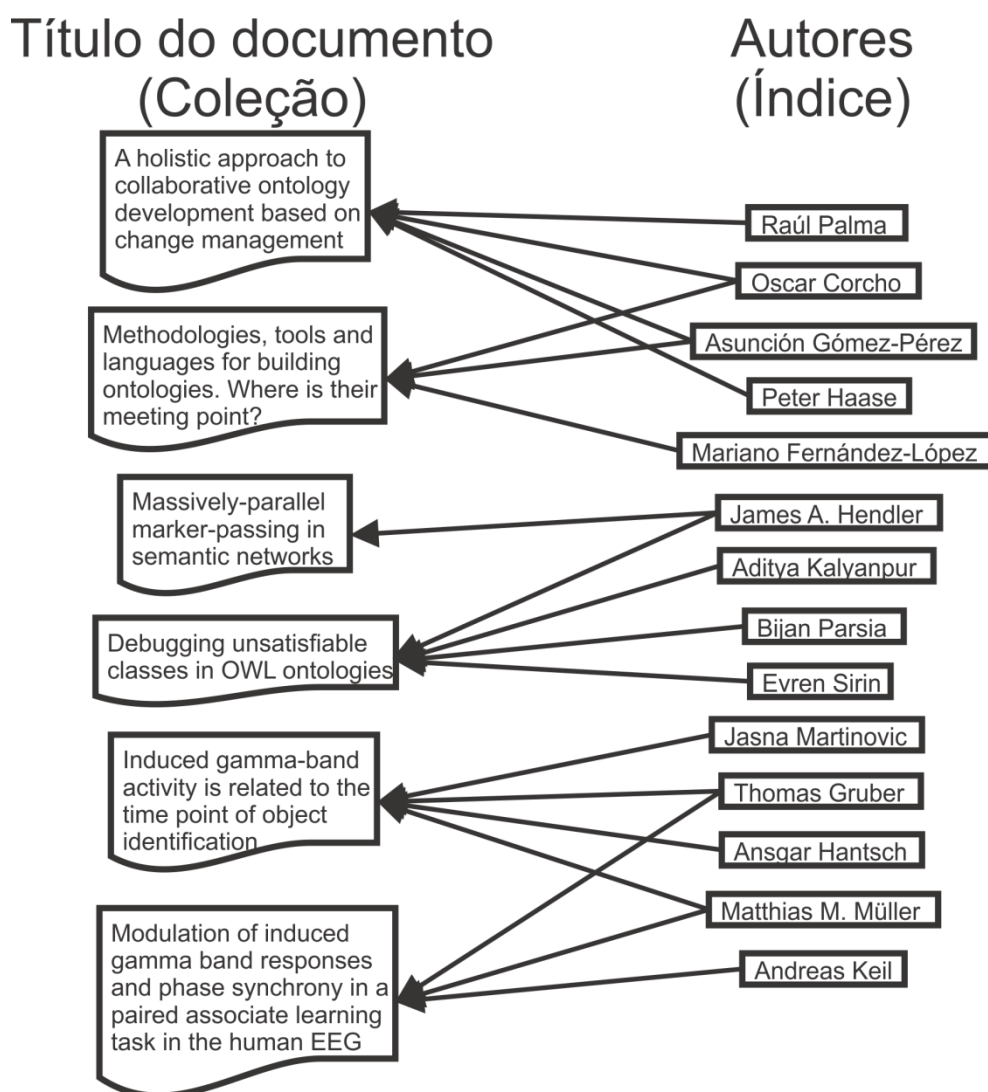
Amazonas et al. (2008) afirma que entre os existentes os tipos de índices mais comuns são o “índice invertido” que é baseado no mapeamento de cada *token* relacionando ao conjunto de documentos a qual ele pertence e o “índice sequencial” que consiste em uma lista de pares de documentos e lista de *tokens*, ordenados pelos documentos.

Amazonas et al. (2008) explica ainda que a construção e atualização de um índice invertido é uma tarefa dispendiosa uma vez que isso demanda a pesquisa de todos os *tokens* a serem inseridos ou alterados, um trabalho que pode ocasionar um aumento de tempo significativo (*overhead*). Em contrapartida a abordagem do índice sequencial reduz o *overhead* que ocorre no índice invertido, pois as inserções sempre ocorrem no final e as atualizações ocorrem diretamente no par do documento a ser atualizado. Em tempo de



execução o índice sequencial é modificado para índice invertido para reduzir o tempo de pesquisa.

Na Figura 3 é possível visualizar o funcionamento da indexação utilizando a técnica de índice invertido, com uma coleção de documentos que possuem autores associados. É possível notar que uma mesma instância de autor poder estar relacionado a mais de um documento.



**Figura 3** - Estrutura de um índice invertido por nome de autores cadastrados na plataforma Science Direct (<http://www.sciencedirect.com>)

No modelo apresentado acima foi exposto uma simplificação de como ocorre a relação de instâncias de pessoas e documentos por um relacionamento de autoria. Essa relação não é restrita a duas entidades como é apresentado no modelo anterior entre documentos e

autores, é possível que haja também índices que interliguem os autores com instituições de pesquisa, bem como, quais instituições os documentos pertencem.

O universo de informação pode gerar um grafo que se expande em escalas cada vez maiores, por exemplo, se assumirmos nesse domínio os cursos que são oferecidos nas instituições de pesquisa, os departamentos elas possuem, as instituições e departamentos que estão associados aos autores dos documentos, instituições concorrentes, entre outros. A estrutura de indexação é um artifício para representar um volume considerável de informações e com navegabilidade facilitada para realizar consultas.

O autor Zobel e Moffat (2006) afirma que em motores de busca os documentos são armazenados num repositório com um índice que é utilizado para realizar análise de correspondência, isso permite indicar quais documentos devem ser retornados numa consulta. O autor cita que um sistema de banco de dados lida com consultas complexas arbitrariamente por uma chave de registro única retornando todos os registros correspondentes, enquanto que consultas submetidas aos motores de busca consistem em uma lista de termos e frases que retorna um número fixo de documentos com base em estatísticas de similaridade, técnica que permite determinar relevância dos resultados. Considerações relacionadas ao cálculo de relevância da arquitetura proposta são descritas na seção 2.2 .

### 3. ONTOLOGIA

Ontologia é um termo que vem do grego *ontos* (ser) + *logos* (palavra) e foi introduzida na filosofia no século XIX pelo filósofo alemão Rudolf Göckel para distinguir o estudo do “ser” do estudo de diversos tipos de seres em ciências naturais. Na filosofia ontologia preocupa-se em com o fornecimento de sistemas de categoria que representam uma visão do mundo. O primeiro sistemas de categorias conhecido foi proposto por Aristóteles. No século III a.C. Porfírio organizou a estrutura de categorias proposta por Aristóteles em um diagrama constituído em uma árvore conhecido como “Árvore de Porfírio” (BREITMAN; CASANOVA; TRUSZKOWSKI, 2007).

Ontologia é um ramo da filosofia, mas o papel da ontologia em tecnologias da informação é diferente da função que exerce na filosofia. Existem discordâncias entre autores na literatura sobre o que é ontologia na área de TI, mas em resumo pode-se dizer que ontologia em sistemas de informação é uma linguagem formal concebida para representar um determinado domínio do conhecimento (ZUÑIGA, 2001).

No contexto de IA, ainda segundo Gruber (1995), ontologia pode ser descrita como a definição de um conjunto de termos de representação, pois há a associação de nomes de entidades no universo do discurso, por exemplo, classes relações, funções entre outros objetos. Conforme o autor entende-se por universo do discurso o conjunto de objetos que podem ser representados num domínio do conhecimento formal. As relações entre os objetos são retratadas no vocabulário de representação do conhecimento. Segundo Meersman (1999) universo do discurso em ontologia é a aplicação, domínio ou conjunto de conceitos linguísticos que constituem o domínio de conhecimento abordado.

Nota-se que ontologia é usada com diferentes sentidos. A diferença mais expressiva reside principalmente entre o sentido filosófico e o sentido computacional que teve ascensão nos últimos anos na comunidade de engenharia do conhecimento (STAAB; STUDER, 2009). Ontologia também é utilizada em diferentes áreas da ciência da computação, entre elas, destacam-se a inteligência artificial, a representação do conhecimento, o processamento de linguagem natural, a Web Semântica, entre outras. Supõe-se que é por essa razão que há algumas divergências entre suas múltiplas definições (BREITMAN; CASANOVA; TRUSZKOWSKI, 2007).

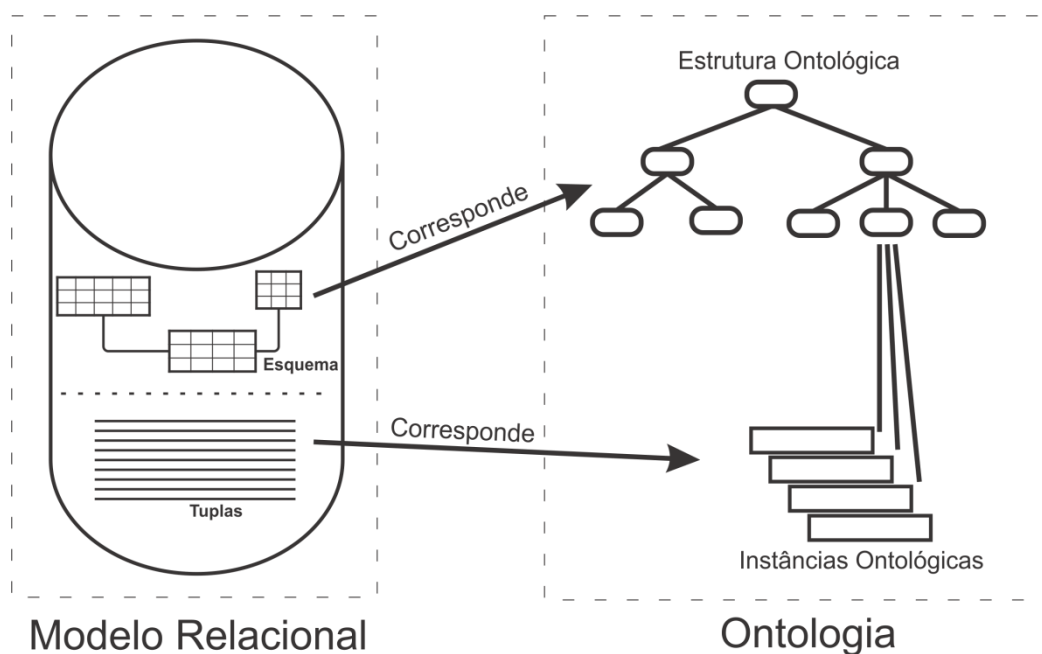
Para Sridharan; Tretiakov e Kinshuk (2004) ontologia pode ser considerada como um vocabulário de termos e relações entre esses termos em um dado domínio. Meersman (1999) afirma que ontologia denota uma coleção de objetos linguísticos organizados de diversos modos diferentes, e que de uma forma limitada e bem definida podem representar a interpretação semântica de um domínio do mundo real. Gruber (1995, p. 908) oferece uma definição sucinta sobre o que é ontologia ao afirmar que “é uma especificação explícita de uma conceitualização”. Na mesma linha de raciocínio Borst (1997, p. 12) afirma que “uma ontologia é uma especificação formal de uma conceitualização compartilhada”.

Os autores Studer, Benjamins e Fensel (1998, p. 184) baseados nas proposições de Gruber e Borst estabeleceram uma definição própria de ontologia que é “uma especificação formal e explícita de uma conceitualização compartilhada” em que, “conceitualização” corresponde a um modelo abstrato de algum fenômeno no mundo, “explícita” significa que os tipos de conceitos usados e as restrições ao seu uso são declarados explicitamente, “formal” quer dizer que a ontologia deve ser legível por máquinas e “compartilhado” diz respeito a um conhecimento que é assumido como consensual.

Considerando a Web o modelo mais comum de ontologia tem na sua base uma taxonomia, que define classes de objetos e relações entre eles, e um conjunto de regras de inferência, que oferece a um sistema a capacidade de expressar conclusões com base em um universo de regras que representam uma verdade (BERNERS-LEE; HENDLER; LASSILA, 2001). Essa definição corresponde à proposição de funcionamento do 5-tupla apresentada por Maedche (2002), que segundo o autor é um modelo simples, amplamente aceito e é facilmente mapeado nas linguagens de representação de ontologias existentes. Essa estrutura permite a

representação da maioria das linguagens de representação do conhecimento atuais (BREITMAN; LEITE, 2004).

O autor Maedche apresentou um modelo de ontologia que se assemelha ao modelo relacional de um banco de dados. A estrutura ontológica compreende um conjunto de instâncias que são como extensão dos conceitos. Tanto o modelo relacional quanto a ontologia são uma espécie de padrão para organizar o conhecimento, e possuem algumas semelhanças semânticas. No modelo relacional, entidades e relacionamento são expressos por relações, portanto uma relação pode ser correspondente a um conceito ontológico ou relação que é organizada hierarquicamente na ontologia. Além disso, as restrições de atributo na base de dados podem ser convertidos em axiomas ontológicos, e as tuplas podem constituir instâncias ontológicas. Há casos em que a informação espalhada por diversas relações podem precisar ser integrados num único conceito ontológico (LI; DU; WANG, 2005). A representação de equivalência entre uma ontologia e um banco de dados relacional é descrita na Figura 4:



**Figura 4** - Similaridade entre o modelo relacional e o modelo ontológico

Fonte: Adaptada de Li, Du e Wang (2005)

Para construir uma base ontológica representando um domínio do conhecimento deve-se primeiramente ter visão completa do domínio para poder ser modelado as regras que compõem o âmbito desse contexto.

Sridharan, Tretiakov e Kinshuk (2004) corroboram que ontologia auxilia na construção de um domínio do conhecimento e permite a representação da estrutura do conhecimento para facilitar a integração de bases de conhecimento, independentemente da heterogeneidade das fontes de informação.

Conforme Almeida e Bax (2003) ontologias não apresentam necessariamente a mesma estrutura, entretanto existem características e componentes básicos que são comuns e estão presentes em muitas delas. Conforme Noy e McGuinness (2001) e Gruber (1995) ontologias em geral contêm classes (domínio do discurso), relações, funções, objetos e axiomas.

Para se chegar à especificação de determinado domínio torna-se necessário a utilização de metodologias. Uschold e King (1995) apresentam sua proposta de metodologia para o desenvolvimento de ontologias em quatro fases: (a) identificar o propósito, (b) construir a ontologia, que abrange captura da ontologia (identificação dos conceitos chaves e relacionamentos no domínio e produção de textos precisos para as relações), codificação da ontologia e integração de ontologias existentes, (c) avaliação e (d) documentação.

Existem pesquisas acerca de metodologias que abordem especificamente o desenvolvimento e manutenção de ontologias. Entre alguns exemplos podem ser citados o TOVE, Enterprise Model Approach, Methontology e o IDEF5 (JONES; BENCH-CAPON; VISSER, 1998). Uma discussão mais detalhada sobre as metodologias de construção de ontologias será realizada na seção 3.1 .

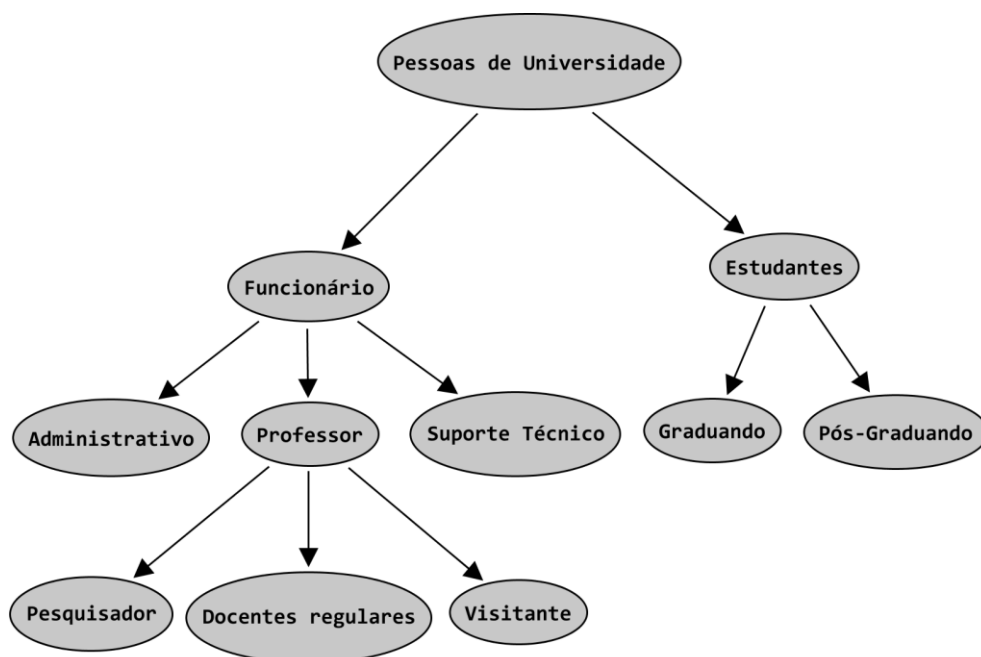
Além de metodologias são necessários no processo de desenvolvimento de ontologia ferramentas e linguagens para especificação das mesmas. Na literatura existe um vasto número de documentos que citam as ferramentas que auxiliam na criação de ontologias como o Protégé (KNUBLAUCH et al, 2005), WebODE (ARPÍREZ et al., 2001), NeOn (HAASE et al., 2008), Ontolingua (FARQUHAR; FIKES; RICE, 1997), WebOnto (DOMINGUE, 1998), OntoEdit (SURE et al., 2002), KAON (BOZSAK et al., 2002), OIL (FENSEL et al., 2001), OilEd (BECHHOFER et al., 2001), entre outros.

No que tange a questão das linguagens já em 2004 se mencionava a elaboração de variados padrões para construção e compartilhamento de ontologia baseados em XML (*eXtensible Markup Language*), com sucintas modificações nas marcações, entre elas, SHOE

(*Simple HTML Ontology Extensions*), XOL (*Ontology Exchange Language*), OML (*Ontology Markup Language*) e CKML (*Conceptual Knowledge Markup Language*), RDFS (*Resource Description Framework Schema*) denominada OIL (*Ontology Interchange Language*) e seu sucessor DAML+OIL (SOUZA; ALVARENGA, 2004).

Recentemente, se têm realizado pesquisas acerca de tecnologias referentes às linguagens de ontologias como RDF, RDFS, OWL, SWRL e demais linguagens prescritas pelo padrão W3C (*World Wide Web Consortium*), em especial a linguagem OWL, que a W3C designa como a linguagem padrão e que deverá ser utilizada na maioria das descrições dos dados de ontologias no futuro (HEO; KIM, 2008).

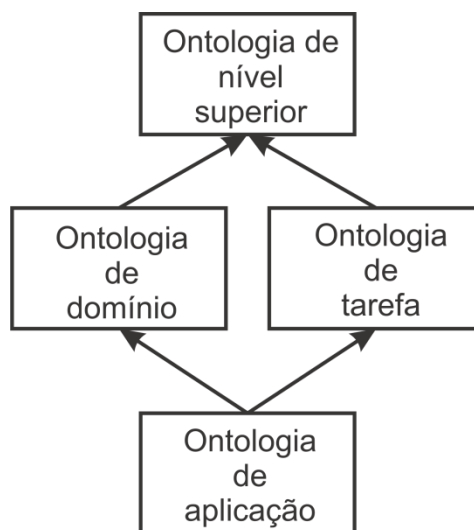
De modo geral, uma ontologia deve descrever formalmente um domínio e tipicamente é formada por uma lista finita de termos e relacionamentos entre esses termos. Os termos representam conceitos (classes de objetos do domínio) importantes, em um ambiente universitário, por exemplo, os conceitos “funcionário”, “estudantes”, “palestras” e “disciplinas” são alguns conceitos importantes (ANTONIOU; HARMELEN, 2008). No exemplo a seguir (Figura 5) é apresentado um modelo simplificado do que seria uma ontologia de domínio de universidade:



**Figura 5** - Exemplo de uma hierarquia de ontologia

Fonte: Adaptada de Antoniou e Harmelen (2008)

Ontologias possuem diferenças que residem no seu nível de generalidade, estabelecendo tipos de ontologia (GUARINO, 1997). As relações entre esses tipos de ontologias são apresentadas na Figura 6.



**Figura 6** - Tipos de ontologia de acordo com seu nível de dependência

Fonte: Adaptada de Guarino (1997)

Ontologia de nível superior se refere a conceitos mais gerais como espaço, tempo, matéria, objeto, evento, ação, entre outros, que independem de um problema particular ou domínio. Ontologias de domínio e de tarefa representam respectivamente o vocabulário relacionado a um domínio genérico (medicina, engenharia) ou a uma tarefa genérica ou a atividade, especializando o que foi descrito na ontologia de nível superior. Ontologias de aplicação descrevem conceitos em função tanto de um determinado domínio quanto de uma tarefa, o qual muitas vezes são especializações de ambas as ontologias relacionadas (domínio e tarefa) (GUARINO, 1997).

### 3.1 METODOLOGIAS

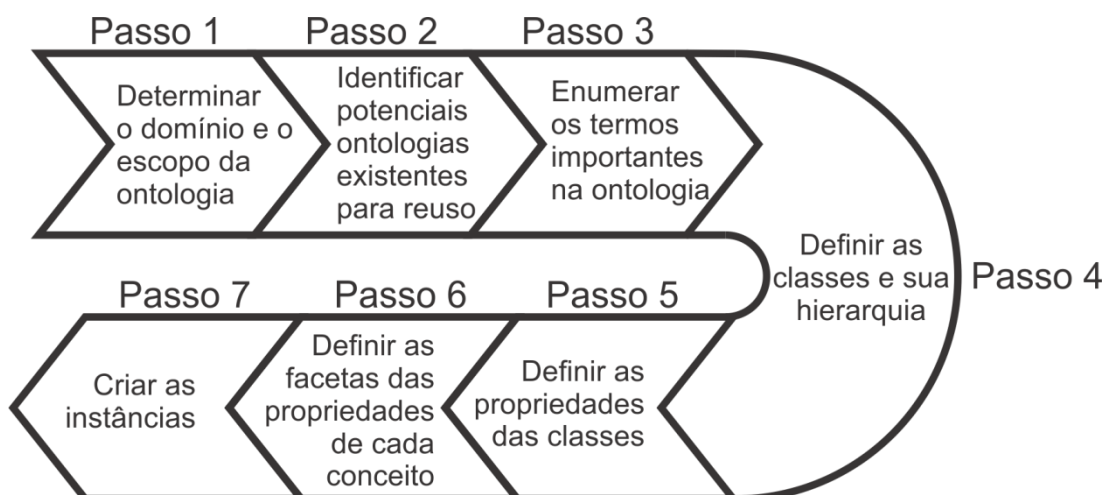
É importante ter o conhecimento do porquê se deseja construir uma ontologia e qual será o seu uso. Segundo Mattos, Simões e Farias (2007) uma metodologia na área de linguagens de representação do conhecimento é uma técnica que por meio de estudos dos métodos visa a criação de um padrão para construção de uma ontologia. Contudo, como afirmam Noy e McGuinness (2001), não existe uma metodologia que seja considerada a mais correta para projetar uma ontologia.



Existem metodologias voltadas para a construção de ontologias, para a construção de ontologias em grupo, para o aprendizado sobre a estrutura de ontologias e para a integração de ontologias. As diversas metodologias existentes possuem aspectos diferentes e não é provável a existência de uma unificação das propostas em uma só metodologia (ALMEIDA; BAX, 2003).

Os autores Noy e McGuinness (2001) declaram três regras como sendo fundamentais para o desenvolvimento de ontologias e que em determinados casos ajudam na tomada de decisões de projeto, são elas: (a) não há uma maneira correta de modelar um domínio, mas sim alternativas viáveis, sendo que a melhor opção muitas vezes depende da aplicação que se deseja construir; (b) o desenvolvimento de ontologias é necessariamente um processo iterativo; e (c) conceitos em ontologia devem ser próximos a objetos (físicos ou lógicos) e relacionamentos no seu domínio de interesse. Estes são mais propensos a serem substantivos (objetos) ou verbos (relacionamentos) em sentenças que descrevem seu domínio. O que vai guiar o desenvolvedor da ontologia sobre quais decisões tomar sobre modelagem, organização e desempenho é o escopo da tarefa a qual a ontologia será submetida.

Os mesmos autores citados no parágrafo anterior apresentam a metodologia 101 que oferece uma perspectiva facilitada de desenvolvimento de ontologias. A metodologia utiliza como apoio uma sequência lógica e simples de ações e define como devem acontecer as atividades para construir uma ontologia coerente e apropriada. A sequência é apresentada na Figura 7 em um modelo sintetizado:



**Figura 7** - Passos para a construção de uma ontologia segundo a metodologia 101

Fonte: Adaptada de Noy e McGuinness (2001)

No passo 1 são analisados questionamentos tais como: Qual é o domínio abordado? Pelo que está se utilizando ontologia? A ontologia deve responder por quais questões? Quem utilizará ou manterá a ontologia? Se a ontologia for utilizada no processamento de linguagem natural pode ser necessário incluir sinônimos. Reutilização é útil para possibilitar comunicação com outras aplicações, aperfeiçoamento e manter um relativo nível de formalismo. Listar termos auxilia a fazer declarações sobre eles e relacioná-los (NOY; MCGUINNESS, 2001).

Definir as classes ajuda a definir de que tipo são as instâncias, e conseqüentemente suas superclasses. As propriedades das classes precisam ser definidas porque somente as classes não conseguem definir o domínio da ontologia. É necessário definir as propriedades porque elas podem ter diferentes facetas que descrevem o tipo de valor, valores permitidos, cardinalidade dos valores, e outras características dos valores que a propriedade pode tomar. Por último instanciar as classes e preencher com valor as propriedades (NOY; MCGUINNESS, 2001). Portanto até o passo 3 as atividades são voltadas ao levantamento e análise dos requisitos e a partir do passo 4 as tarefas se referem à construção da ontologia em si.

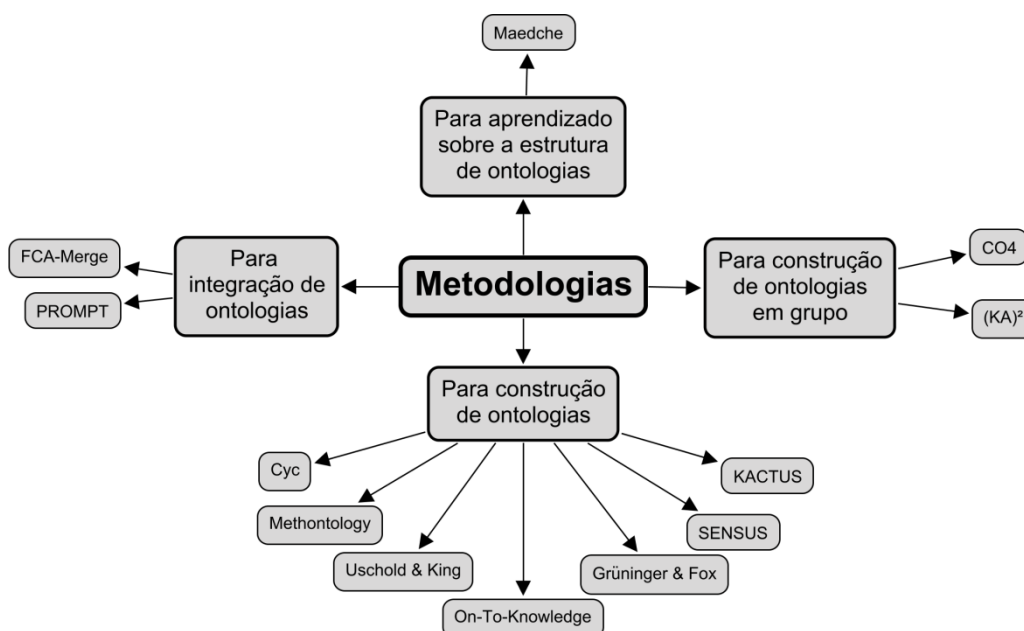
Almeida e Bax (2003) afirmam que a construção de uma ontologia é uma tarefa dispendiosa, por isso além das metodologias, existem algumas ferramentas que facilitam o processo de sua construção. Para manter uma relativa comparabilidade entre as ferramentas existem critérios definidos para a sua construção, geralmente utilizam linguagem de representação para a construção das ontologias.

Segundo Noy e McGuinness (2001) e Corcho, Fernández-López e Gómez-Pérez (2003), a construção de uma ontologia parte da utilização conjunta de uma metodologia, uma ferramenta que auxilie na construção de uma ontologia e uma linguagem de ontologias. Os diferentes *frameworks* necessários para a construção de uma ontologia apropriada possuem diferentes abordagens e finalidades para a construção de ontologias específicas.

Muito embora as metodologias possuam abordagens diferentes sobre como atender as demandas, a criação de uma ontologia em geral possui alguns passos fundamentais inerentes à criação de qualquer ontologia, não importando qual tarefa ela irá resolver. Esses passos constituem-se na seleção de uma metodologia em si, escolha de qual ferramenta auxiliará no desenvolvimento da ontologia e uma linguagem de construção. As escolhas

podem ou não determinar qual será a escolha seguinte, em virtude de nem todas as ferramentas darem suporte a todas as linguagens de ontologia (SILVA; SOUZA; ALMEIDA, 2008).

Existe um conjunto de nomenclaturas que identificam as ferramentas que compõem o grupo de recursos que auxiliam na criação de uma ontologia. Na Figura 8 são apresentadas algumas dessas ferramentas e suas respectivas características para tornar mais explícito pelo que cada ferramenta é responsável conforme é citado por (ALMEIDA; BAX, 2003).



**Figura 8** - Organização das metodologias por tipo

Fonte: Adaptada de Almeida e Bax (2003)

Algumas das metodologias citadas neste trabalho são discutidas mais detalhadamente em documentos como a metodologia de Grüninger e Fox em (GRÜNINGER; FOX, 1995), a de Uschold e King em (USCHOLD; KING, 1995); Methontology em (FERNÁNDEZ-LÓPEZ; GÓMEZ-PÉREZ; JURISTO, 1997), a metodologia Cyc em (REED; LENAT, 2002), metodologia Sensus em (SWARTOUT et al., 1997) e método 101 (NOY; MCGUINNESS, 2001).

### 3.2 LINGUAGENS

Entre as linguagens disponíveis atualmente para a criação e manipulação de ontologias encontram-se o RDF, RDF Schema (RDFS), o OWL e o SWRL.

### 3.2.1 RDF

RDF (*Resource Description Framework*) é um modelo padrão para permitir troca de dados na Web. Tem como função estender a estrutura de ligação da Web utilizando as URIs para nomear a relação existente entre os recursos presentes na Web. Essa forma abstrata de um recurso estar ligado a outro por uma relação que os conecta de forma lógica é denominado "tripla" (RDF WORKING GROUP, 2004).

Arquivos RDF fornecem maneiras de expressar declarações sobre recursos, usando propriedades nomeadas e valores, mas em si o RDF não fornece meios para definição de classes e propriedades de aplicações específicas. Em vez disso, tais classes e propriedades são descritas como um vocabulário RDF, usando extensões no RDF providos pela linguagem de descrição de vocabulário, denominada RDF Schema. O RDF Schema é, em alguns aspectos, similar a uma linguagem de programação orientada a objetos, pois proporciona os mecanismos necessários para descrição de classes e propriedades assim como as suas utilidades, mas diferem no sentido de que as classes RDF apenas descrevem informações adicionais sobre os recursos RDF, à medida que uma classe na linguagem de programação possui métodos e atributos (MANOLA; MILLER, 2004).

O RDF, se adotado em larga escala, permite a adoção de mecanismos mais avançados de busca, mas também é necessário a adaptação dos motores de busca para que indexem RDF, que permite navegação de dados estruturados com resultados agregados de vários documentos. Alguns motores de busca que vem sendo implementado RDF são o Swoogle, Falcons, WATSON e Sindice (HOGAN et al., 2011).

Bizer, Berners-Lee e Heath (2009) dizem que sites não devem ser ligados somente por URLs, mas também vinculá-los por meio de RDF, pois possui declarações tipadas ligando recursos arbitrários, diferente do HTML (*HyperText Markup Language*) que segundo Klyne e Carroll (2004) não possuem tipagem, portanto não têm mapeamento léxico-valor. O projeto "Linking Open Data" contribui para a regulamentação de emprego de Web Semântica nos motores de busca, em grande parte por meio de arquivos RDF com conteúdos exportados da Wikipedia, BBC, New York Times, publicações científicas, entre outros (HOGAN et al., 2011).

O RDF possui recursos que geralmente representam um objeto do mundo real que possui atributos. Cada recurso possui um URI (*Uniform Resource Identifier*), ou seja, um identificador único de um recurso da Web. A partir dos recursos é possível estabelecer declarações (*statements*) que definem a relação de um recurso com um valor correspondente (denominado literal) por meio de uma propriedade. Propriedades são responsáveis por descrever de forma explícita as relações entre os recursos, possibilitando uma visão completa do contexto sem ambiguidades. A declaração é uma tripla atributo-valor consistindo em um recurso, uma propriedade e um valor, algumas vezes mencionados como sujeito, predicado e objeto por possuírem estrutura léxica semelhante (ANTONIOU; HARMELEN, 2008).

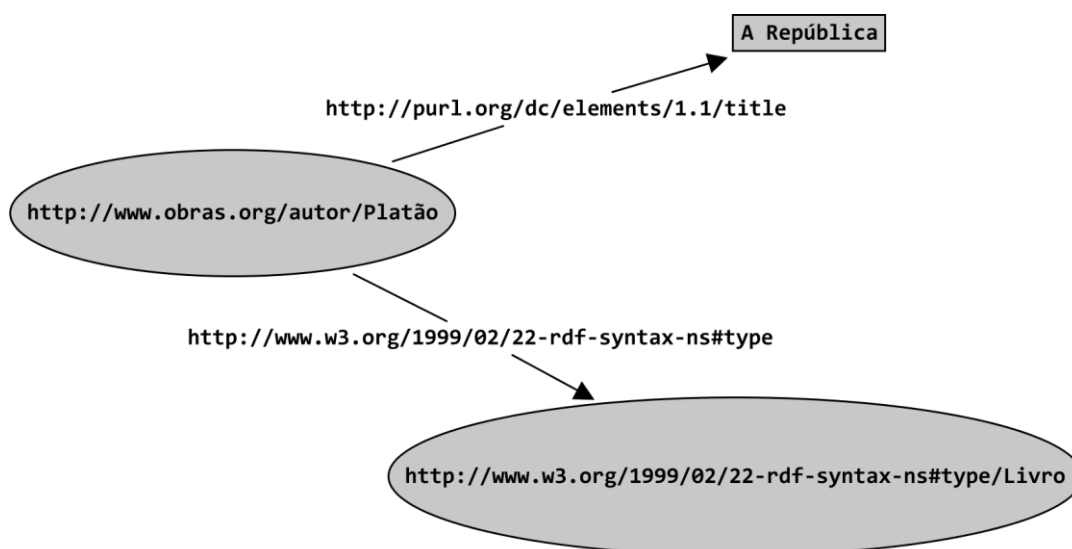
A seguir é apresentada uma exemplificação de uma tripla RDF com sujeito, predicado e objeto:

*Platão é autor do livro “A República”*

O arquivo RDF que representa essa afirmação é representado da seguinte forma:

```
<?xml version="1.0"?>
  <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description rdf:about="http://www.obras.org/autor/Platão">
      <dc:title>A República</dc:title>
      <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-
ns#type/Livro"/>
    </rdf:Description>
  </rdf:RDF>
```

Em grafo o RDF é representado na Figura 9:



**Figura 9** - Representação em Grafo do arquivo RDF

### 3.2.2 OWL

Em 10 de fevereiro de 2004 o W3C introduziu como recomendação formal de ontologia a OWL (*Web Ontology Language*), que foi desenvolvida para a Web e fundamentada sobre lógica descritiva. Sua criação tinha como objetivo superar limitações e problemas de outras linguagens que a precederam, por exemplo, o XML (MATOS, 2008).

Segundo Matos (2008) arquivos XML proporcionam sintaxe para documentos semi-estruturados, mas não agrega semântica aos marcadores e por isso não apresentam nenhum significado no processamento de um documento. O XML Schema estende a XML com um aglomerado de tipos de dados fornecendo um esquema para os documentos XML.

De modo geral, a OWL é uma linguagem construída sobre o RDF e RDF Schema com sintaxe RDF baseada em XML. As instâncias são definidas por meio de descrições RDF (MATOS, 2008).

A OWL é uma linguagem baseada em lógica, de tal maneira que o conhecimento expresso em OWL possa ser fundamentado por programas de computador para conferir a consistência de um conhecimento ou tornar explícito um conhecimento implícito. É uma linguagem planejada para representar o conhecimento rico e complexo sobre as coisas, grupos e suas relações (OWL WORKING GROUP, 2012).

As ontologias são documentos OWL que podem ser publicadas na Web e fazem referências a outras ontologias ou são referenciadas a partir de outras ontologias. OWL é parte do conjunto de tecnologias desenvolvidas pela W3C assim como RDF, RDFS, SPARQL (*SPARQL Protocol and RDF Query Language*), entre outros (OWL WORKING GROUP, 2012).

OWL é uma linguagem de especificação para a Web Semântica influente, fazendo com que se torne um veículo natural para fornecer controle de acesso num determinado contexto (FININ et al, 2008). Segundo Breitman (2006), a *Web Ontology Language* (OWL) é uma revisão da linguagem DAML+OIL, criada para atender necessidades como:

- Construção de ontologias
  - Criar uma ontologia;

- Explicitar conceitos fornecendo informações sobre os mesmos;
- Explicitar propriedades fornecendo informações sobre as mesmas.
- Explicitar fatos sobre um determinado domínio
  - Fornecer informações sobre indivíduos que fazem parte do domínio em questão.
- Racionalizar sobre ontologias e fatos
  - Determinar as consequências do que foi construído e explicitado.

Conforme Horridge et al. (2004) a linguagem OWL pode se dividir em três sub-linguagens que são o OWL-Lite, OWL-DL e OWL-Full sendo o primeiro o menos expressivo, o segundo com expressividade intermediária e o último a mais expressiva das linguagens.

Os autores McGuinness e Harmelen (2004) relatam com mais detalhes cada tipo. O OWL-Lite permite gerar hierarquia de classificação e restrições simples, possui cardinalidade binária, oferece um pacote de migração rápida para um tesouro ou outras taxonomias e possui uma complexidade formal reduzida. OWL-DL oferece expressividade intermediária, completude computacional e decidibilidade. Inclui todos os construtores da linguagem OWL, mas podem ser utilizados somente sob certas restrições. O OWL-Full oferece expressividade máxima e liberdade sintática do RDF sem garantias computacionais, permite utilizar ontologias para aumentar o significado do vocabulário pré-definido e é pouco provável que um software de raciocínio suporte completamente todos os recursos do OWL-Full (MCGUINNESS; HARMELEN, 2004).

Em muitos domínios existe grande quantidade de informação armazenada em formatos XML, que tem impulsionado o desenvolvimento de ferramentas para a substituição de XML para OWL (O'CONNOR; DAS, 2010), tais ações são importantes para a aquisição do conhecimento.

Além disso, proposições mais formais como a de Maedche (2002) vem sendo sugeridas. Nesse modelo uma ontologia é composta de 5 tuplas visando prover um meio de representar determinado domínio de conhecimento por meio de uma linguagem. O modelo de 5-tuple possui a seguinte estrutura:

$$O := \{C, R, H^c, rel, A^0\}$$

Consistindo em:

- Dois conjuntos disjuntos de  $C$  e  $R$  cujos elementos são chamados de conceitos e relações, respectivamente.
- Uma hierarquia de conceitos  $H^c$ :  $H^c$  é uma relação dirigida  $H^c \subseteq C * C$  que é denominado hierarquia de conceitos ou taxonomia.  $H^c(C_1, C_2)$  significa que  $C_1$  é um subconceito de  $C_2$ .
- Uma função  $rel: R \rightarrow C * C$  que relaciona conceitos de forma não taxonômica. A função  $dom: R \rightarrow C$  com  $dom(R) := \Pi_1(rel(R))$  fornece o *domain* (domínio) de  $R$ , e  $range$  (alcance)  $R \rightarrow C$  com  $range(R) := \Pi_2(rel(R))$  fornece seu *range*.
- Um conjunto de axiomas da ontologia  $A^0$ , expressada numa linguagem apropriada.

Traçando um paralelo com ontologia, o modelo 5-tuple proposto por Maedche possui o  $C$  que representa o conjunto de conceitos declarados na ontologia e  $R$  as relações que fazem a conexão lógica dos conceitos num domínio do conhecimento.  $H^c$  faz referência à hierarquia de conceitos modelada para representar uma taxonomia de uma temática. A função  $rel$  faz alusão à relação não taxonômica dos conceitos, ou seja, uma relação cruzada entre os conceitos ( $C$ ) que a partir de um processo de raciocínio lógico e inferências pode-se construir as propriedades dos objetos somente por um processo dedutivo, por exemplo, *todo cetáceo é um mamífero e nada, baleia é um cetáceo, logo baleia é mamífero e nada*, e por fim o  $A^0$  que é um conjunto de verdades assumidas sem ambiguidades para admitir o processo de raciocínio lógico sobre o universo limitado de conhecimento.

OWL possui primitivas de restrição denominadas *domain* e *range* que permitem modelar e organizar objetos inseridos no contexto da ontologia. *Domain* são restrições de domínio sobre o primeiro argumento de uma relação binária, e *range* são restrições de alcance referente ao segundo argumento numa relação binária (GRUBER, 1993). Os exemplos apresentados a seguir demonstram o conceito de domínio e alcance:



```

<ObjectPropertyDomain>
  <ObjectProperty IRI="#temSintoma"/>
  <Class IRI="#Paciente"/>
</ObjectPropertyDomain>

<ObjectPropertyRange>
  <ObjectProperty IRI="#temSintoma"/>
  <Class IRI="#Sintoma"/>
</ObjectPropertyRange>

```

O código define uma propriedade de objeto “temSintoma” que tem como domínio a classe “Paciente” e como alcance a classe “Sintoma”, restringindo nesse contexto que o paciente tem uma propriedade que diz que ele possui um sintoma e uma classe sintoma que instancia as doenças.

Existem dois tipos de propriedades em OWL que permitem modelar a relação entre as classes, são a propriedade de objeto, que liga indivíduo a indivíduo por uma relação declarada (*ObjectProperty*), e os dados da propriedade que liga indivíduos aos seus atributos e valores (*DataProperty*). Abaixo são apresentadas em exemplos as definições de propriedades:

```

<DataPropertyDomain>
  <DataProperty IRI="#Nome"/>
  <Class IRI="#Paciente"/>
</DataPropertyDomain>

<DataPropertyRange>
  <DataProperty IRI="#Nome"/>
  <Datatype abbreviatedIRI="xsd:string"/>
</DataPropertyRange>

<ObjectPropertyDomain>
  <ObjectProperty IRI="#temDoenca"/>
  <Class IRI="#Paciente"/>
</ObjectPropertyDomain>

<ObjectPropertyRange>
  <ObjectProperty IRI="#temDoenca"/>
  <Class IRI="#Doenca"/>
</ObjectPropertyRange>

```

No exemplo acima “Nome” é uma propriedade de dado (um atributo) da classe “Paciente”, desta forma possui como domínio (*domain*) global a classe “Paciente” e alcance (*range*) para a classe “string”, indicando o tipo de dado que esse atributo armazena. Na propriedade de objeto “temDoenca” é uma relação entre as classes “Paciente” e “Doenca” indicando que uma instancia de paciente possui alguma doença.

### 3.2.3 SWRL

A SWRL (*Semantic Web Language Rule*) é a linguagem padrão para regras de Web Semântica (ORLANDO, 2012). Um número expressivo de áreas no âmbito da computação estão utilizando a linguagem SWRL, como no gerenciamento de consumo de energia em residências (ROSSELLÓ-BUSQUET et al., 2011), no aprimoramento de navegação em sistemas de tutoria (VESIN et al., 2011), mecanismo de cálculo utilizando regras para a construção de redes de coautoria de publicações (AHMEDI; ABAZI-BEXHETI; KADRIU, 2011), representação de um domínio do conhecimento para criar ambiente inteligentes com auxílio de sensores atuadores (SADOUN et al., 2011), gerenciamento inteligente de fotos digitais (CHAI et al., 2010), gestão de batimentos cardíacos com base em regras para prever anomalias nos sinais (TANANTONG; NANTAJEEWARAWAT; THIEMJARUS, 2011), entre outros.

SWRL é uma linguagem que possui uma sintaxe abstrata de alto nível que se baseia em OWL-DL e OWL-Lite que são linguagens derivadas da OWL bem como a *Rule Markup Language* (RuleML) (HORROCKS et al., 2004). Na pilha da Web Semântica as regras estão no topo da ontologia. A linguagem de regras é útil para expressar diferentes formas de regras como o padrão para o encadeamento de ontologias, regras de ponte para raciocínio através de domínio, regras de mapeamento para integração de dados entre ontologias, regras de consulta para expressar consultas complexas sobre a Web e meta regras para a engenharia de ontologias (GOLBREICH; IMAI, 2004).

De acordo com Horrocks et al. (2004) SWRL é uma extensão direta da semântica utilizada em OWL que possui definição de *binding* (ligação) e extensões de interpretação OWL que permite mapeamento de variáveis para os elementos de um domínio. Segundo Wagner, Giurca e Lukichev (2006) sistemas baseados em regras estão ficando mais comuns na modelagem de sistemas. O que se espera disso é que sob uma mesma abordagem o sistema seja capaz de realizar trocas de informação com outro sistema mesmo que utilize outra linguagem baseada em regras. Tendo em vista que a interoperabilidade entre sistemas baseados em regras era até então limitada, houve destaque de interesses acerca dessa propriedade na padronização de regras dos sistemas que deu início a um determinado número de linguagens incluindo o SWRL (O'CONNOR et al., 2005).

Conforme O'Connor et al. (2005) assim como muitas outras linguagens baseada em regras, as regras SWRL são escritos em pares antecedente-consequente, em que em sua terminologia o antecedente se refere à regra de corpo e o consequente se refere à cabeça. A cabeça e o corpo consistem de uma conjunção de um ou mais átomos.

Os átomos e são formados por um predicado e um ou mais argumentos, de forma que a quantidade e o tipo de argumentos são determinados pelo tipo do átomo, que por sua vez, é definido pelo tipo de predicado utilizado (SILVA, 2012).

Mesmo com essa formalização o SWRL possui alguns problemas. Segundo Orlando (2012) conforme o número de regras cresce, os desenvolvedores passam a enfrentar problemas no seu gerenciamento, pois com o tempo as regras tomam uma forma tal que é difícil o entendimento e erros se tornam mais suscetíveis. Outro agravante é se o conjunto de regras for construído por mais de uma pessoa e o fato de essa linguagem carecer de ferramentas que operacionalizem o trabalho de criação dessas regras. A criação dessas ferramentas seria uma possível solução por facilitar a visualização e edição colaborativa das regras.

### **3.3 WEB SEMÂNTICA**

Em empresas o uso de ontologia para motores de busca de informação auxiliar no processo de decisão estratégica é recente, o intenso uso da ontologia se deu com o advento da Web Semântica e foi nesse segmento que se desenvolveu. Em virtude disso a finalidade dessa seção será apresentar um histórico sintetizado da ontologia apresentando juntamente o início da Web Semântica visto que a história da ontologia é delineada em conjunto com a evolução da Web Semântica.

Guha, McCool e Miller (2003) definem a Web Semântica como uma extensão da Web atual, na qual a informação possui significado bem definido, permitindo melhor cooperação entre computadores e pessoas. É o conceito de se ter dados na Web definidos e ligados de um modo que possam ser usados para descoberta, automatização, integração e reuso entre diversas aplicações. A Web Semântica contém recursos correspondentes a objetos de mídia comuns utilizados na Web atual como páginas Web, imagens e vídeos e, além disso, corresponde a objetos como pessoas, lugares, organizações e eventos. Esses objetos se

relacionam com muitos tipos distintos de ligação, diferente do que ocorre atualmente na Web, em que as ligações ocorrem somente por *hiperlinks*.

A ontologia na área da computação foi introduzida pela Web Semântica mencionada no documento (BERNERS-LEE; HENDLER; LASSILA, 2001). Nesse documento os autores afirmam que a Web Semântica não é separada da Web atual, mas sim uma extensão em que a informação recebe significados bem definidos permitindo que computadores e pessoas trabalhem em cooperação.

A Web atual não provê funcionalidades muito além do que o armazenamento de informação em que cuja coleta precisa ser interpretada e filtrada, utilizando técnicas de pesquisa sobre informação correlacionada por meio de ligações entre os documentos (HEO; KIM, 2008). Existe uma grande variedade de recursos disponíveis na Web, semanticamente descrito ou não, abrangendo páginas da Web, diferentes tipos de documentos, aplicações, formulários, listas e até mesmo esquemas que representam estruturas de dados (BLANCO; VILA; MARTINEZ-CRUZ, 2008).

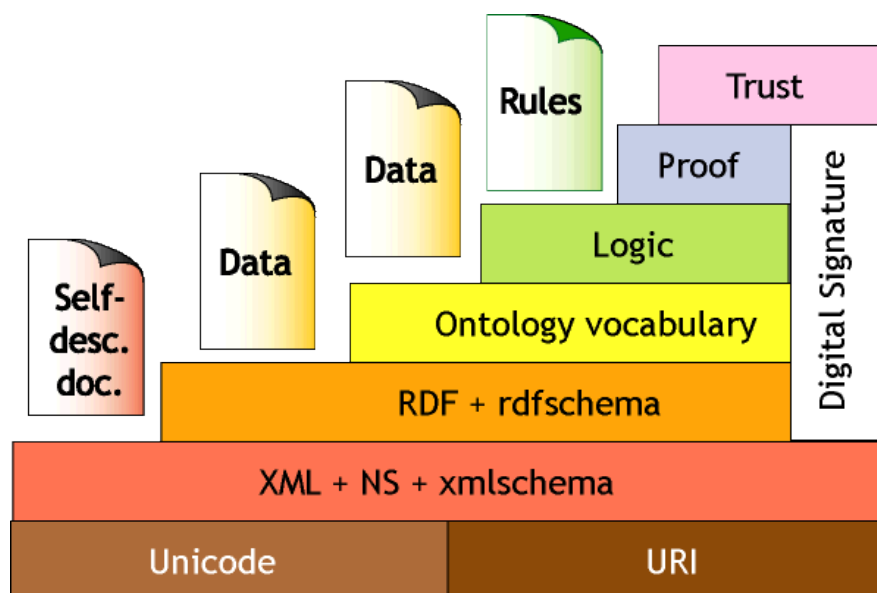
Para solucionar a limitação de relevância das consultas utilizando semântica, pesquisas para linguagens de ontologia (RDF, RDFS, OWL e assim por diante) prescritas como padrão da W3C e para as tecnologias relacionadas foram realizadas ativamente (HEO; KIM, 2008).

Segundo Breitman (2006) na década de 1990 houve muitas propostas de linguagem para criação de ontologias e foram criadas uma série de linguagens baseadas em princípios de inteligência artificial, a maior parte delas com base em lógica de primeira ordem. De acordo com Matos (2008) lógica de primeira ordem se caracteriza como um sistema de raciocínio cujo cada sentença ou declaração é dividida em sujeito e predicado em que o predicado define as propriedades de um único sujeito.

Acompanhando a evolução surgiram novas tecnologias conhecidos como “linguagens leves” (*lightweight*), entre eles o HTML, XML e RDF. O XML foi promissor porque permitiu estruturar documentos e validar a informação recebida (BREITMAN, 2006). O RDF e XML são linguagens que podem realizar anotação semântica no conteúdo de um documento.

Anotação semântica cria um link entre um objeto no texto ao seu descritor semântico, por meio de um mecanismo de marcação visual e a ontologia auxilia na criação dessa anotação na Web Semântica. Para uma anotação semântica ser criada é necessário que uma ontologia esteja associada ao documento e os termos presentes no documento reconhecidos pela ontologia são transformados em *links* para outros recursos semanticamente conectados (OLIVEIRA, 2006).

Segundo Breitman (2006) em 2000 Tim Berners Lee propôs um modelo em camadas para a arquitetura futura da Web com o intuito de sugerir uma arquitetura que reestruturasse a Internet em cima do que já existe em tecnologia. O modelo tem como primeira camada o HTML em conjunto com o XML, na segunda camada o RDF. Esse modelo conceitual e minimalista que apresenta a arquitetura da Web Semântica em camadas com interoperabilidade é representado na Figura 10.



**Figura 10** - Modelo de arquitetura da Web Semântica “*Layer Cake*” proposto por Berners-Lee (2000)

Antoniou e Harmelen (2008) e Passin (2004) discorrem sobre algumas das camadas apresentadas no modelo:

- XML (*eXtensive Markup Language*) é caracterizado por ser uma linguagem que possibilita a construção e envio de documentos pela Web. O documento é estruturado com um vocabulário definido pelo usuário;

- RDF (*Resource Description Framework*) é definido como um tipo de dados básico que descreve informações sobre recursos da Web. O RDF não depende do XML. Esquema RDF é a camada que fornece os mecanismos para modelar e organizar objetos da Web, entre eles as classes e subclasses, propriedades e subpropriedades, relacionamentos e restrições de domínio (*domain*) e alcance (*range*);
- Camada de vocabulário de ontologia são linguagens utilizadas para definir vocabulários e estabelecer o uso de palavras e termos no contexto de um vocabulário específico;
- Camada de Lógica, usada para melhorar a ontologia incrementalmente, expandindo esquema RDF e permitindo representações de relação mais complexas entre objetos da Web. Permite escrever declarações específicas de um conhecimento;
- A camada Proof envolve o processo dedutivo e a representação de verificações em linguagens Web e verificação de validação;
- A camada Trust é responsável por fazer verificações de assinaturas digitais para mensurar a confiança dos dados operacionais e informações prestadas.

Breitman (2006) afirma que o RDF surgiu para fornecer um modelo formal de dados de sintaxe para codificar metadados, permitindo interoperabilidade entre aplicativos e primitivas básicas para a criação de ontologias simples. Mais tarde foi desenvolvido uma extensão do RDF denominada RDFS que ofereceu primitivas de modelagem para a construção de hierarquias, classes, propriedades, subclasses e subpropriedades.

Segundo Heo e Kim (2008) existe recentemente o interesse por Jena2, que é um *framework* Java™, que inclui um ambiente de programação e motor de inferências baseada em regras para as linguagens como RDF, OWL, SPARQL. Nos últimos anos foram propostas algumas linguagens de extensões ao RDFS para ontologias entre elas SHOE, OIL, DAML (*DARPA Agent Markup Language*), DAML+OIL e OWL (BREITMAN, 2006).

### 3.4 ANOTAÇÃO SEMÂNTICA

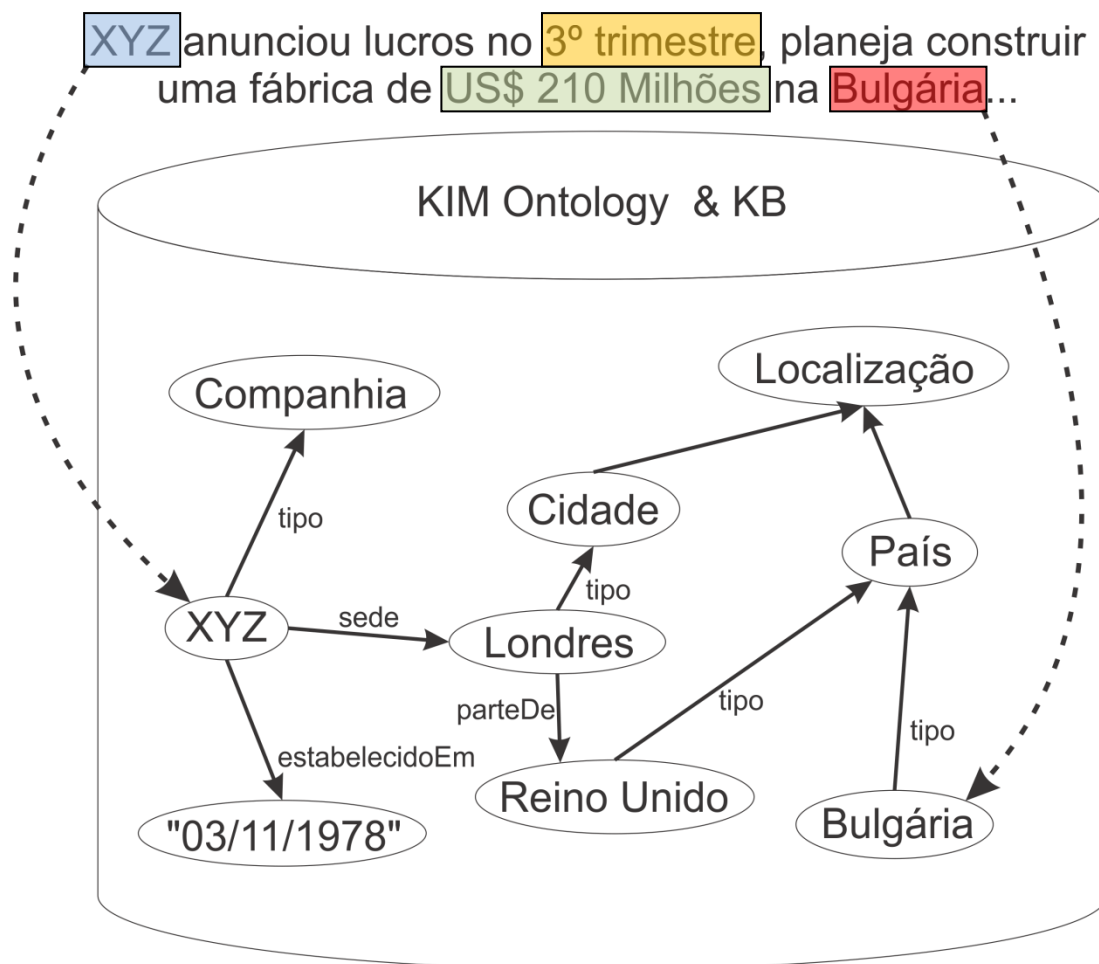
Anotação Semântica é a atribuição de significados formais a fragmentos de textos, permitindo que softwares possam navegar na informação para realizar consultas e retornar a resposta ao usuário com mais precisão. Essa técnica faz com que dados sejam encapsulados juntamente com a sua semântica, provido por um modelo semântico que é uma ontologia ou um esquema conceitual (KIYAVITSKAYA et al., 2009). Um documento anotado semanticamente tem informações que permite que motores de busca possam inferir sobre o seu conteúdo.

Segundo Nunes e Fileto (2007) buscas sobre documentos sem estruturação semântica é sintática e não permitem consultas com relacionamento semântico entre os termos. A ontologia representa a associação semântica entre os conceitos de um domínio e a anotação semântica de documentos explicita seu conteúdo pela associação dos conceitos expressado na ontologia. Toda essa arquitetura aliado aos padrões estabelecidos na Web permitem que esse metadados seja amplamente processável e as consultas passam a se favorecer de relações semânticas presentes na ontologia para exprimir as ambiguidades e recuperar termos semanticamente relacionados.

Anotação semântica é admitir valores definidos em certas estruturas do conhecimento, como em ontologias, cujo uso adequado evita que informações sejam mal descritos em virtude de fenômenos que ocorrem na linguagem natural (como as ambiguidades). As anotações podem ser descritos em triplas RDF, permitindo que objetos não explicitamente informados na consulta sejam apresentados se necessário (RIGO, 2011).

A anotação semântica das informações deve ser feita seguindo os padrões estabelecidos na modelagem da ontologia. A anotação deve ser bem definida, não ambígua e fácil de entender, pois é a ferramenta que fornecerá o elo entre a informação armazenada no documento e o modelo formal de domínio. Se o formalismo utilizado for a ontologia, a anotação semântica deve seguir a hierarquia ontológica (NUNES; FILETO, 2007). O processo de anotação pode ocorrer sobre qualquer tipo de texto digital (mesmo domínios de banco de dados), e sua estrutura fornece padrões para agentes de software recuperar informações autonomamente.

Kiryakov et al. (2004) explica que a tarefa de anotação semântica se dá em dois passos: (a) anotar e ligar (referenciar a) as entidades nos documentos e (b) indexar e recuperar documentos que tenham relação com as entidades referidas. O processo pode ser visualizado na Figura 11.



**Figura 11** - Exemplo de anotação semântica

Fonte: Adaptada de Kiryakov et al. (2004)

O processo de anotação se dá inicialmente identificando as entidades nomeadas no documento (nome de uma pessoa, objeto, entre outros) e posteriormente ligar as entidades com suas descrições semânticas na ontologia por meio de um URI (*Uniform Resource Identifier*). O resultado do processo é associado a uma ontologia e gravado em um repositório. A anotação pode ter uma representação intrusiva, que é a anotação gravada no próprio documento, e não intrusiva, cuja anotação é armazenada separadamente sem alterar o arquivo, para tanto utiliza ponteiros para referenciar os termos (NUNES; FILETO, 2007).



### 3.5 BUSCA SEMÂNTICA

De acordo com Zou et al. (2008) busca semântica refere-se ao processo envolvendo um algoritmo de anotação semântica visando extrair conceitos ou instâncias constantes em uma ontologia de domínio, anotar o conjunto de documentos para o repositório de recursos de domínio e gerar o repositório de índice semântico. De acordo com o termo de consulta do usuário o mecanismo de busca realiza a tarefa de consulta a partir do repositório de índices semânticos e os resultados de pesquisa com recurso semântico são devolvidos ao usuário.

Fazzinga e Lukasiwicz (2010) afirmam que não existe uma definição única sobre a noção de busca semântica na Web. O que é tido como um consenso é que se trata de uma forma melhorada de busca na Web, onde é extraído o significado e estrutura das consultas de usuário, assim como dos demais conteúdos da Web para servir de ferramental de análise durante o processo de consulta na Web.

Segundo Nagarajan e Thyagarajan (2012), a busca semântica tem como intuito a partilha de informação cujo objetivo central é aproximar o homem e a máquina para buscar sobre um conteúdo de forma colaborativa. Esse mecanismo de busca aproxima-se dessa possibilidade a partir do momento que consegue superar limitações referentes à incapacidade de identificar as relações entre os termos de busca, que seria o caso numa eventual pesquisa entre *sistemas de recomendação de livros* ou *livros de sistemas de recomendação* (FERNÁNDEZ et al., 2011).

De acordo com Hendler (2010), existem duas funcionalidades essenciais em uma busca semântica. A primeira objetiva fornecer resultados mais informativos ao usuário em vez de simples páginas como acontece nos mecanismos de busca atuais e a segunda procura ajudar o usuário na identificação de pesquisas adicionais que podem ser úteis.

O processo de busca semântica melhora os resultados de pesquisa de duas formas, sendo: (a) os resultados assumem a forma de uma lista de documentos independentes que se estendem pelo uso de informações adicionais implícitas; e (b) se a pesquisa denotar um ou mais conceitos do mundo real, a busca semântica será útil para a compreensão destes conceitos e seu respectivo contexto (GUHA; MCCOOL; MILLER, 2003).

No contexto da Web, para tornar o resultado não ambíguo, as consultas e textos se utilizam de dados de redes semânticas. Os fundamentos de lógica da Web semântica permitem a recuperação inteligente de dados e ajuda a lidar com a heterogeneidade semântica (LUPIANI-RUIZ et al., 2011). A consulta passa por uma fase de reformulação para torná-las estruturadas. Seja o método de busca por palavras-chave ou linguagem natural, cada termo é mapeado para um conceito ontológico que pode ser uma propriedade ou entidade de classe, para então se realizar as combinações de correspondência semântica dos termos (FAZZINGA et al., 2011).

De acordo com Fazzinga et al. (2011) o uso de semântica na Web está intrinsecamente ligado ao contexto de organização de dados, que depende da adoção de ontologias e envolvimento de raciocínio sobre as conexões existentes entre os recursos na Web. Ainda segundo o autor, devido a essa técnica, a busca semântica é capaz de analisar a consulta Web bem como as páginas Web de acordo com seu significado, podendo retornar exatamente as páginas semanticamente relevantes à consulta.

O desempenho de pesquisa semântica em relação aos métodos baseados em palavras-chave é um motivador, pois a busca semântica é aplicada visando estender buscas gerais permitindo mudanças na experiência de pesquisa por parte do usuário. Em vez de simplesmente identificar uma página útil, os sistemas retornam as informações de páginas que o usuário está procurando de forma imediata (HENDLER, 2010). Os autores Tümer, Shah e Bitirim (2009), afirmam que o motor de pesquisa considerado ideal seria o que fosse capaz de retornar precisamente a informação que o usuário deseja encontrar.

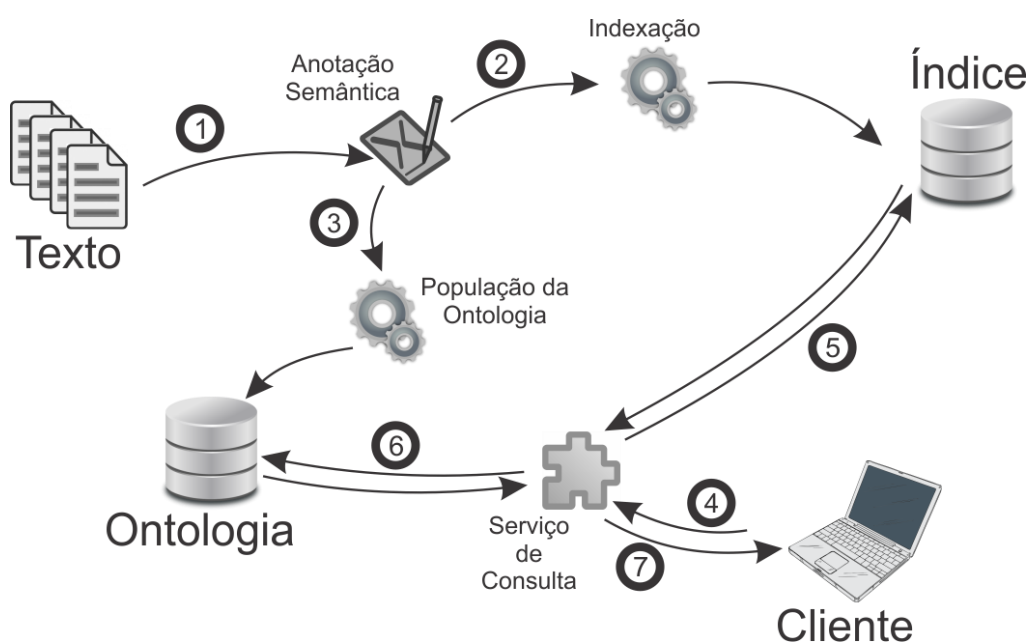
O uso de semântica em buscas é um tema tão proeminente que o desenvolvimento de uma nova tecnologia de busca para a Web Semântica é muito estudada tanto por empresas privadas quanto em pesquisas acadêmicas, com ênfase no desenvolvimento de formas de pesquisa e coleta de dados codificados numa representação formal e adição de semântica em pesquisas web (FAZZINGA et al., 2011). Tais empresas, principalmente as empresas de Web, estão trabalhando autonomamente em algoritmos especializados para atender suas próprias necessidades na área de busca de informação sem esperar pelos avanços de pesquisas científicas, e grande parte desse desenvolvimento contribuirá significativamente para melhorar a infraestrutura de busca semântica (HENDLER, 2010).

## 4. SISTEMA PROPOSTO

Nesse capítulo será apresentada a visão lógica e física do sistema proposto neste trabalho. A seção promove informação sobre a arquitetura, técnicas e ferramentas utilizadas visando clarificar as etapas de concepção e desenvolvimento do trabalho.

### 4.1 VISÃO LÓGICA

A proposta deste trabalho incute a união de um conjunto de partes logicamente dispostas para realizar a operacionalização do processamento do sistema. A visão lógica objetiva apresentar de forma clara como é constituída a arquitetura do sistema e como se dará a dinâmica do processo como um todo. Os passos são descritos de forma genérica para oferecer o entendimento geral do seu funcionamento. A visão lógica do sistema é representada na Figura 12:



**Figura 12** – Visão lógica do sistema de busca semântica

Esta é uma proposição em camadas, que é constituída no nível mais alto pela camada do cliente, que recebe a informação pré-processada, seguida pela camada de serviço que é a estrutura que realiza o acesso integrado às informações de ontologia e índice textual e pela camada capaz de realizar as operações de indexação e manutenção dos dados, e no nível mais baixo, a camada de base de dados que é composta por uma coleção de documentos textuais.

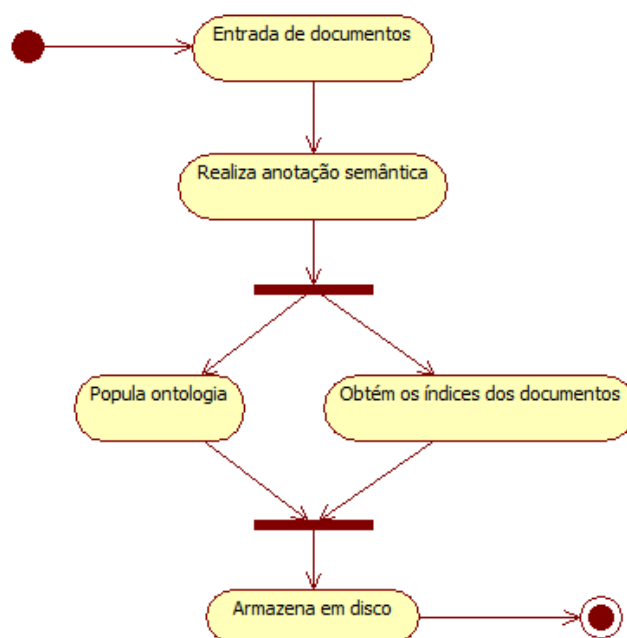
Uma discussão sobre os passos realizados pela arquitetura para seu completo funcionamento é explanado abaixo. Os passos citados a seguir são enumerados na Figura 12 para permitir uma melhor visualização do processo:

**PASSO 1:** Documentos não estruturados são anotados semanticamente para atribuir uma formatação padronizada formal para que o processo de população de ontologias possa realizar as operações adequadamente. Esse processo consiste em utilizar documentos não estruturados e gerar a partir deles, uma representação padronizada explicitando o que significam os diversos elementos presentes nestes documentos.

**PASSO 2:** Neste passo é realizada a indexação dos documentos anotados em uma estrutura de índice invertido com o objetivo de possibilitar a realização de consultas sobre documentos completos.

**PASSO 3:** Ainda com base nos documentos anotados é realizada a população da ontologia, ou seja, são adicionadas as classes, indivíduos e propriedades de objetos e de dados com o objetivo de proporcionar conteúdos adicionais obtidos por meio de inferências. Desse modo, espera-se respostas que possibilitem ao usuário, a partir de uma consulta inicial, explorar novos caminhos visando um melhor entendimento do domínio de interesse.

A Figura 13 apresenta a sequência de atividades que ocorrem do passo 1 ao passo 3:



**Figura 13** - Diagramas de atividades do passo 1 ao 3

**PASSO 4:** O usuário, tendo uma necessidade de busca, envia uma solicitação ao serviço de consulta que realiza as operações de localização dos documentos relevantes e das informações adicionais advindas da ontologia, sejam estas, declaradas explicitamente ou inferidas.

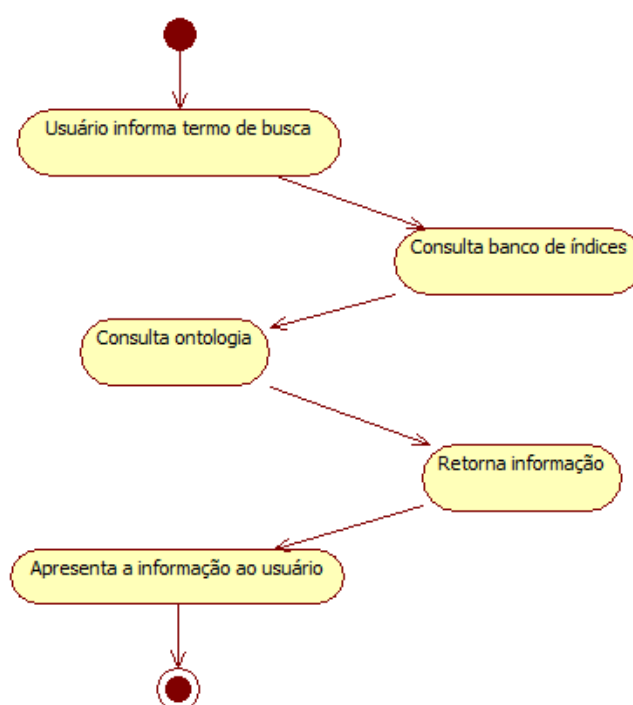
**PASSO 5:** Neste passo o serviço de consulta realiza uma busca por todos os documentos que satisfazem os termos informados pelo usuário, ou seja, a expressão de busca. Basicamente, o serviço irá retornar alguns campos utilizados para apresentação, por exemplo, título, URL, autores de documentos, e um campo adicional que informa a relevância do documento frente a expressão de busca. Essa relevância é obtida através do cálculo do cosseno permitindo que a lista esteja ordenada de maneira decrescente pela relevância do documento em que os mais relevantes são apresentados primeiro.

**PASSO 6:** O serviço de consulta além de localizar os documentos que satisfazem determinada expressão de busca também utiliza essa informação para realizar a pesquisa na base de conhecimento (representada pela ontologia) para que sejam agregadas novas informações referentes aos documentos obtidas por meio de um processo de inferência. Como retorno, são devolvidas informações que representam as possíveis classes onde o termo de

busca foi encontrado, bem como, as propriedades que relacionam o termo (entendido como um indivíduo na ontologia) com outros indivíduos de diferentes classes.

**PASSO 7:** Esse último passo apresenta as informações de forma simples ao usuário contendo, tanto os documentos quanto as informações adicionais da ontologia obtidas por meio de um processo de inferência. Esse resultado objetiva propiciar ao usuário diferentes maneiras de analisar determinado domínio e satisfazer de forma mais adequada uma necessidade de consulta.

A Figura 14 apresenta a sequência de atividades que ocorrem do passo 4 ao passo 7:



**Figura 14** - Diagramas de atividades do passo 4 ao 7

## 4.2 VISÃO FÍSICA

A visão física do sistema apresenta os componentes tecnológicos que o constitui e é detalhada na Figura 15. Os diversos componentes do sistema trabalham em conjunto para proporcionar uma experiência aprimorada na navegação de busca de informações, a partir de uma aplicação Java (cliente) que integra as funcionalidades dos mecanismos de indexação, anotação semântica e ontologias. O sistema como um todo foi desenvolvido utilizando o

*framework* Apache Lucene™ (versão 3.0.3) como meio de construção e consultas de índices textuais e a API OWL™ (versão 3.4.1) para armazenar as informações complementares sobre os termos incorporados ao domínio do conhecimento por meio de uma ontologia.

O processo como um todo necessita inicialmente que os documentos pertencentes ao *corpus* sejam anotados semanticamente. Com os documentos já anotados realiza-se a indexação destes por meio do Lucene que terão seus conteúdos armazenados em uma estrutura de índice invertido.

Além disso, é realizada a população de ontologia utilizando-se para tal a API OWL, responsável por armazenar as informações em uma ontologia no formato OWL. Desse modo, o índice e a ontologia formam a base para a realização de consultas sobre documentos e informações adicionais que auxiliem no entendimento de determinado domínio de interesse do usuário.

O serviço de consulta foi desenvolvido de maneira simplificada através de uma aplicação Java que recebe do usuário determinado conteúdo de interesse (termos de busca) e, a partir disto, realiza a consulta no índice textual e localiza na ontologia todas as informações adicionais referentes ao conteúdo de interesse, afirmadas ou inferidas. Ainda é de responsabilidade do serviço de consulta apresentar a informação obtida ao usuário composta pelo conjunto de documentos e das informações da ontologia que satisfazem a consulta.





configurado é então enviado a base de índices através da classe de indexação do Lucene (*IndexWriter*) que cria (ser for a primeira indexação) ou atualiza o índice.

Quando o documento é indexado o índice fica estruturado em um conjunto de campos que formam o conteúdo do documento. Neste trabalho o índice é composto pelos campos “key” que representa a chave primária do documento, ou seja, seu número identificador para futuras atualizações, “title” que é o título do documento, “year” sendo o ano em que o documento foi publicado, “author” que consiste no conjunto de autores que escreveram o documento, “institution” indicando em qual instituição os autores trabalham, “abstract” é o resumo do documento e “keyword” são os descritores ou palavras-chave.

De maneira geral, a indexação consiste na criação de três arquivos que são: *\_ $N$ >.cfs*, *segments.gen*, *segments\_<math>N</math>*, em que o arquivo a) *\_ $N$ >.cfs* é a condensação dos arquivos de índice em um único onde o  $N$  representa a evolução do índice uma vez que para a realização da atualização uma nova cópia do arquivo é gerada e ao término está é integrada (compactada) novamente no arquivo com um novo valor; b) “*segments.gen*” contém o sufixo dos segmentos atuais como uma forma redundante de se determinar a persistência (*commit*) dos dados em índices mais recentes; e c) “*segments\_<math>N</math>*” é o arquivo que faz referência a todos os segmentos ativos uma vez que durante o processo de atualização novos segmentos são criados mas somente se tornam ativos após a persistência (*commit*) (HATCHER; GOSPODNETIĆ; MCCANDLESS, 2009).

Entretanto, antes da indexação, é necessário submeter os arquivos a uma política de uso dos termos que estão presentes nos documentos para que haja uma organização formal dos termos adicionados ao índice. Essa política de organização é determinada pelo “analisador” escolhido para a situação. O analisador é responsável por selecionar pedaços de textos, denominados *tokens*, que julgar importante segundo os padrões de um determinado idioma.

**PASSO 3:** Para a execução de uma busca semântica é necessário que os documentos que agora estão armazenados em um índice textual sejam adicionados também a uma ontologia levando em consideração as marcações semânticas. Os documentos anotados explicitamente por um usuário ou por auxílio de algum sistema fazem então parte da ontologia sendo estas informações ditas como afirmadas.

**PASSO 4:** Além das informações incluídas na ontologia no passo anterior a partir dos documentos anotados (informações afirmadas) torna-se necessária a inclusão manual de informações adicionais. Nesse sentido, a adição de regras permite aumentar a relevância das informações que podem ser obtidas de uma ontologia através de um processo de inferência.

Analisando a visão física proposta é requerida a utilização de uma ferramenta que possibilite a população da ontologia, ou seja, a inclusão de regras. No contexto do trabalho utilizou-se o Protégé para a inclusão de regras no formato SWRL.

**PASSO 5:** Neste passo o usuário submete uma consulta, ou seja, o argumento de busca desejado à aplicação Java e recebe a informação processada para que esta possa ser disponibilizada para avaliação.

Assim como na indexação, os termos de entrada de uma consulta passam pelo mesmo analisador para a extração de *tokens* e posteriormente a efetiva consulta dos termos no índice. A geração dos *tokens* depende do analisador utilizado, entre eles existem o *WhitespaceAnalyzer*, *SimpleAnalyzer*, *StopAnalyzer*, *BrazilianAnalyzer*, *ChineseAnalyzer*, *DutchAnalyzer*, *RussianAnalyzer* e o *StandardAnalyzer* que é considerado o analisador mais sofisticado (HATCHER; GOSPODNETIĆ; MCCANDLESS, 2009). O detalhamento do retorno da consulta será apresentado no passo 8.

**PASSO 6 e 7:** De forma a integrar a busca textual com a busca sobre informações complementares presentes na ontologia foram utilizadas funções capazes de retornar a que classe na ontologia pertence o termo (indivíduo) pesquisado, quais são as suas relações com outros indivíduos e quais são seus atributos.

Na concepção do modelo de ontologia foi construído um diagrama que demonstra a comunicação entre as classes, as propriedades de dado que os conectam, e o que é inferido ou declarado com afirmações. A ontologia foi concebida para armazenar informações de um pequeno contexto acadêmico, com dados sobre documentos publicados e informações a respeito do documento como autor, descritores, instituição, resumo, título e ano de publicação.

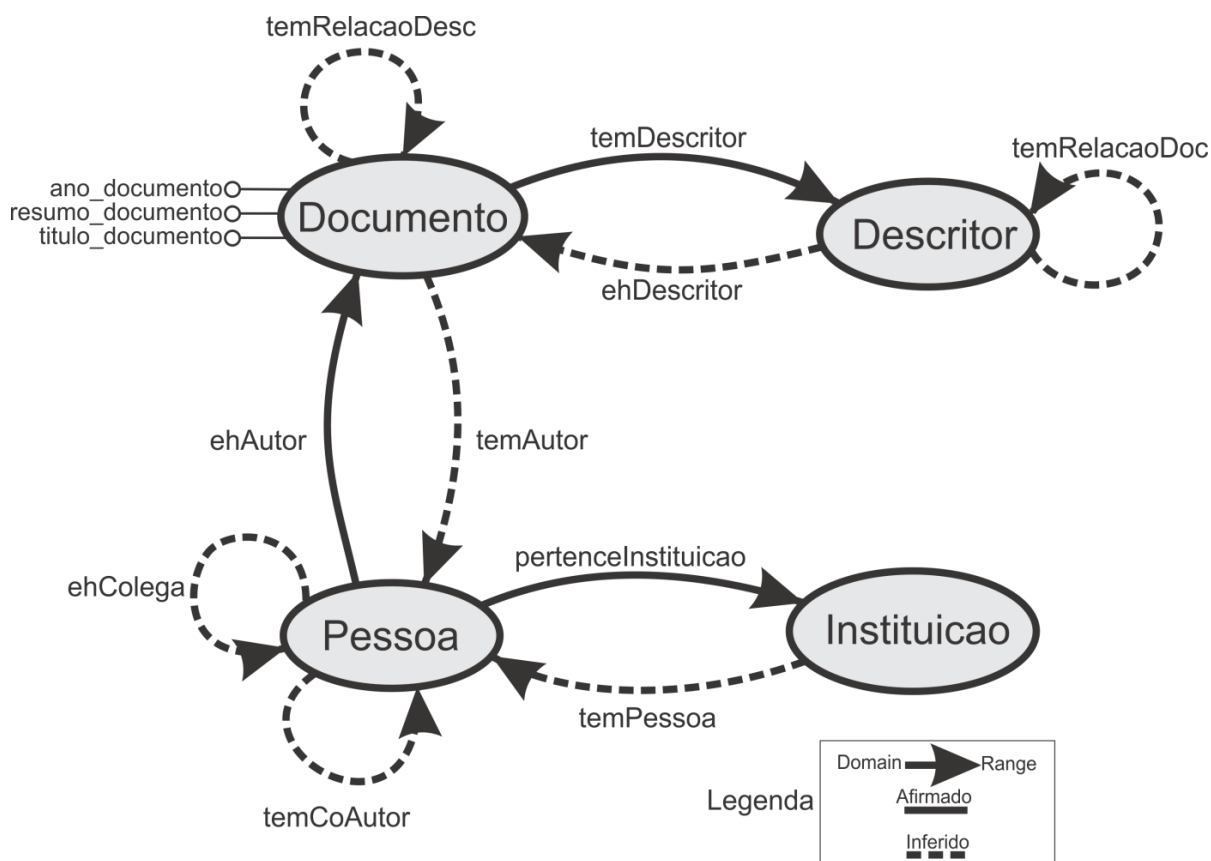
Um ponto principal a ser considerado no sistema se refere à representação dos documentos na ontologia. Estes possuem campos tais como título, resumo e ano que não são

compartilhados, ou seja, pertencem especificamente ao documento e, portanto, é adequado que sejam propriedades de dado. A identificação de um documento acadêmico consiste ainda na presença de um número de autores. Os autores são indivíduos que potencialmente escreverão diferentes documentos e terão assim diversos coautores. Dessa forma, motiva-se a utilização de uma classe que possibilite a definição de relações de coautoria.

Uma pessoa que publica um documento acadêmico necessariamente estará vinculada a uma organização, por exemplo, uma universidade ou um instituto de pesquisa. Uma mesma organização pode ter mais de uma pessoa associada e essa possibilidade torna pertinente o uso de uma classe para representar instâncias de organização.

Os documentos possuem ainda descritores que definem seu conteúdo a partir de alguns poucos termos. O uso de descritores como classe motiva a análise de relação de conteúdos entre documentos a partir da existência de descritores iguais em ambos, permitindo navegabilidade nos documentos por conceitos.

Essa configuração resultou, como demonstra a Figura 16, em uma ontologia que contém 4 classes e 10 propriedades de objeto. Das propriedades de objeto 7 delas são inferidas, das quais 4 são por meio de regras SWRL e 3 a partir da propriedade inversa. As outras 3 propriedades de objeto são afirmadas, ou seja, declaradas explicitamente. As classes foram dispostas para atender a necessidade de devolver as informações em conformidade com os campos presentes nos documentos indexados.



**Figura 16** - Grafo da ontologia de documentos

A ontologia possui três autorreferências e o processo de atribuição dos indivíduos no *domain* e no *range* são realizados por meio de regras SWRL que estão descritas abaixo:

- 1) `temDescritor(?doc, ?desc), temDescritor(?doc2, ?desc), DifferentFrom (?doc, ?doc2) -> temRelacaoDoc(?doc, ?doc2)`
- 2) `temAutor(?doc, ?aut), temAutor(?doc, ?aut2), DifferentFrom (?aut, ?aut2) -> temCoAutor(?aut, ?aut2)`
- 3) `ehDescritor(?desc, ?doc), ehDescritor(?desc2, ?doc), DifferentFrom (?desc, ?desc2) -> temRelacaoDesc(?desc, ?desc2)`
- 4) `pertenceInstituicao(?pes, ?inst), pertenceInstituicao(?pes2, ?inst), DifferentFrom (?pes, ?pes2) -> ehColega(?pes, ?pes2)`

A primeira regra mencionada estabelece que se dois documentos diferentes possuírem alguma palavra-chave em comum, será atribuído a ambos uma propriedade de objeto informando que esses documentos têm conteúdo relacionado. A segunda regra determina que se em um mesmo documento houver mais de um autor, será atribuído uma propriedade de objeto a cada indivíduo estabelecendo relação de coautoria. Por sua vez, a terceira regra retrata a relação entre duas palavras-chave que ocorrem no mesmo documento,

objetivando permitir uma maior navegabilidade através de conceitos. Por último, a quarta regra permite ao usuário saber quais são os colegas de trabalho de uma determinada instância da classe pessoa por meio de uma relação de trabalho na mesma instituição.

**PASSO 8:** Numa situação em que se consulte o nome de um autor, o resultado obtido é uma lista de informações sobre o autor presente na ontologia e a lista de documentos consultados no índice e que tenham como autor o nome pesquisado. Considerando a ontologia, é possível ter-se informações sobre quais pessoas são coautores do indivíduo, quais documentos ele é autor, seu nome completo e e-mail. Considerando o índice de documentos é possível visualizar o título, o ano e os demais autores dos documentos que tem como autor o termo pesquisado. O exemplo a seguir (Figura 17 e Figura 18) apresenta as informações obtidas a partir da ontologia e a partir do índice textual considerando o indivíduo (termo utilizado na busca) “Natalya F. Noy”.

```

=====
||                               Dados da ontologia                               ||
=====
Termo:
  Natalya F. Noy

Classe:
  (Processo afirmado) Pessoa

Propriedades de Objeto:
  (Processo afirmado) ehAutor >> "Translating the Foundational Model of Anatomy into OWL"
  (Processo afirmado) ehAutor >> "Where to publish and find ontologies? A survey of ontology libraries"
  (Processo afirmado) ehAutor >> "vSPARQL: A view definition language for the semantic web"
  (Processo inferido) temCoAutor >> "Dan Suciu"
  (Processo inferido) temCoAutor >> "Daniel L. Rubin"
  (Processo inferido) temCoAutor >> "Landon T. Detwiler"
  (Processo inferido) temCoAutor >> "Mathieu d'Aquin"
  (Processo inferido) temCoAutor >> "Marianne Shaw"
  (Processo inferido) temCoAutor >> "James Brinkley"
  (Processo inferido) ehColega >> "Paea LePendu"
  (Processo inferido) ehColega >> "Daniel L. Rubin"
  (Processo inferido) ehColega >> "Nigam H. Shah"
  (Processo inferido) ehColega >> "Mark A. Musen"
  (Processo afirmado) pertenceInstituicao >> "Stanford University"

Propriedades de Dado:
  (Processo afirmado) email_pessoa >> "noy@stanford.edu"
  (Processo afirmado) nome_pessoa >> "Natalya F. Noy"

```

**Figura 17** - Exemplo de apresentação dos resultados de uma consulta obtidas da ontologia

```

=====
||                               Dados da indexação                               ||
=====
Número de documentos retornados ao consultar o índice: 3

1º documento
Título: Translating the Foundational Model of Anatomy into OWL
Ano: 2008
Autor(es): Natalya F. Noy; Daniel L. Rubin

2º documento
Título: Where to publish and find ontologies? A survey of ontology libraries
Ano: 2012
Autor(es): Mathieu d'Aquin; Natalya F. Noy

3º documento
Título: vSPARQL: A view definition language for the semantic web
Ano: 2011
Autor(es): Marianne Shaw; Landon T. Detwiler; Natalya F. Noy; James Brinkley; Dan Suciu

```

**Figura 18** - Exemplo de apresentação dos resultados de uma consulta obtidas do índice textual

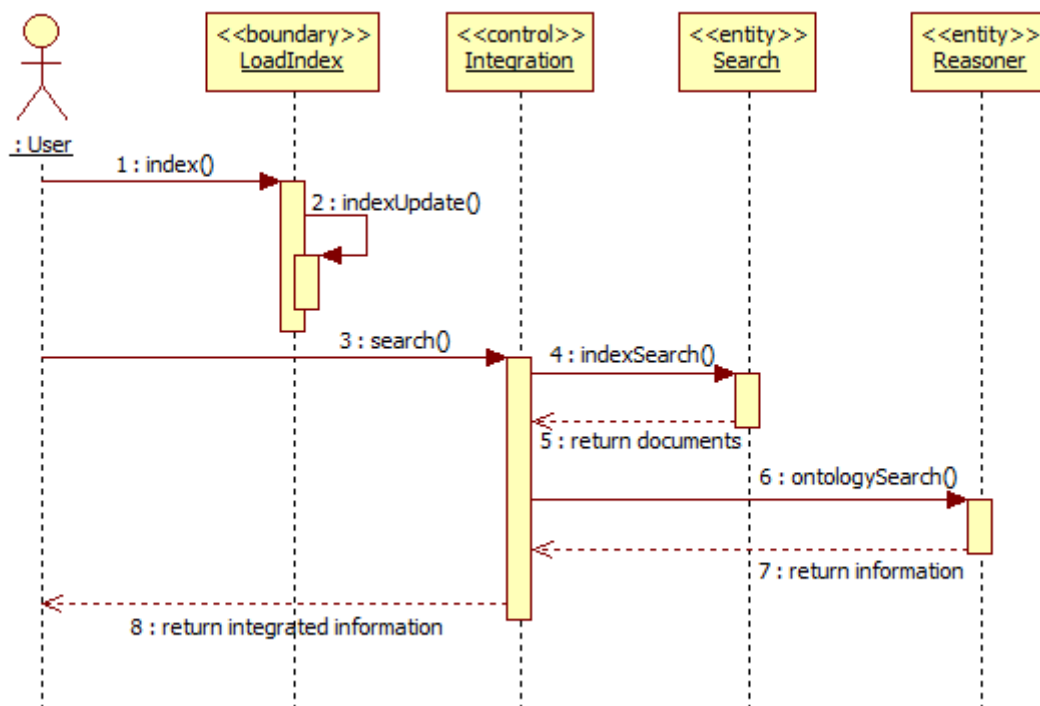
#### 4.2.1 Detalhamento do Protótipo

O protótipo foi construído com funcionalidades básicas de recuperação de informação baseados em um *framework* já consolidado, Lucene. Entretanto, o objetivo principal foi promover a integração desse sistema com uma ontologia de modo que fosse possível a obtenção de resultados que permitissem o entendimento de determinado domínio de aplicação.

Sua funcionalidade principal é fornecer ao usuário navegabilidade a partir de complementos informativos retornados como resultados de uma consulta. A partir das propriedades retornadas o usuário pode realizar de maneira interativa e iterativa novas pesquisas que se referem ao termo pesquisado.

A pesquisa de informações sobre os documentos exige um processo sistemático de consultas envolvendo a busca coordenada de informações em um índice textual e em uma ontologia. Para realizar essa operação é necessária a indicação de um termo inicial de consulta.

O conjunto de documentos que fazem parte do *corpus* compõe uma pequena lista de documentos existente numa base de dados científica, apenas para fins de testes e avaliação de resultados. A seguir é apresentado o diagrama de sequência do protótipo (Figura 19) que possibilita uma visão geral do fluxo e inter-relação das classes.



**Figura 19** - Diagrama de sequência do protótipo

O protótipo foi organizado em quatro classes com a finalidade de encapsular o comportamento conforme suas funcionalidades, bem como, consolidar o processo de consulta informacional.

Para que o procedimento de realização de pesquisa possa acontecer é necessário que o conjunto de documentos de interesse seja indexado. Isso ocorre utilizando-se o método *index()* da classe *LoadIndex*. Também é possível indexar novamente um documento que já exista no índice caso este tenha sofrido alguma modificação. Vale mencionar que a ontologia foi preenchida manualmente.

Logo após ser realizada a indexação, o usuário submete uma consulta através do método *search()* à classe *Integration* que é responsável por realizar a integração completa do sistema, ou seja a realização de uma consulta ao índice textual através do método *indexSearch()* da classe *Search* e de uma consulta na ontologia através do método *ontologySearch()* da classe *Reasoner*. No caso do método *indexSearch()* é indicado em qual campo do índice o termo deve ser aplicado no processo de consulta.

A classe *Search* através do método *indexSearch ()* se encarrega de realizar a tradução do conteúdo de pesquisa, ou seja, analisar o(s) termo(s) de busca transformando-o(s) para o modelo de processamento do Lucene e então executar a pesquisa. Com o resultado da pesquisa sob seu domínio é delimitado quantos resultados vão aparecer na tela e se eles aparecerão em ordem decrescente de relevância. Neste trabalho os documentos recuperados em uma consulta são apresentados em ordem de relevância limitados no máximo a dez documentos por consulta. Como resultado, um vetor é preenchido com os documentos retornados de forma que apenas percorrendo os campos do vetor é possível apresentar o que foi obtido através da consulta.

O segundo passo envolve a invocação do método *ontologySearch()* da classe *Reasoner*. Nesta classe a ontologia, armazenada em arquivo com formato OWL, é carregada para a memória e também é criado um objeto que recebe as referências de dados que estão presentes na ontologia para permitir a obtenção dos indivíduos e suas informações relacionadas, sejam afirmadas ou inferidas.

Para possibilitar a obtenção das informações inferidas a partir da ontologia é requerida a definição de qual será o raciocinador (*reasoner*). Entre os raciocinadores mais conhecidos existem o FaCT++, RacerPro, HermiT e o Pellet. Por ser o mais popular e melhor conhecido este último foi escolhido para ser utilizado no sistema.

Como passo final a classe *Integration*, que recebeu tanto os documentos que satisfazem a consulta quanto às informações vindas da ontologia, integra esses dados visando a apresentação ao usuário. A partir disso, a informação pode ser utilizada para a realização de novas consultas com o intuito de entender mais detalhadamente o conteúdo que se encontra disponível no *corpus*.



## 5. ANÁLISE DOS RESULTADOS

No que tange a análise dos resultados cabe frisar que o foco do trabalho foi projetar e desenvolver um sistema, não havendo portanto diagnóstico comparativo com outras propostas ou ferramentas computacionais. A avaliação será realizada com base na capacidade do sistema em retornar corretamente o que foi pesquisado, ou seja, recuperar a partir dos termos de uma consulta, todos os documentos e todas as informações da ontologia, sejam estas afirmadas ou inferidas.

Além disso, tem como escopo analisar os resultados obtidos através da utilização do protótipo considerando uma base de testes contendo artigos na área de Ontologia. A partir desta análise deseja-se demonstrar a capacidade do sistema em contribuir para que o usuário entenda determinado conteúdo a partir da navegação entre os conceitos do(s) domínio(s) que a coleção de documentos representa. O capítulo foi dividido em:

- Cenário de Aplicação: Com a discussão acerca do cenário utilizado para a avaliação do protótipo, sendo justificado como se ocorreu a seleção dos documentos que compuseram o *corpus* (conjunto de documentos);
- Discussão: Discorre sobre os resultados obtidos a partir da análise de diferentes contextos aplicados, com avaliação da utilidade do protótipo em pesquisas reais a partir da informação retornada na tela;

### 5.1 CENÁRIO DE APLICAÇÃO

Sabe-se que o processo de pesquisa textual e recuperação de informação é uma prática comum executada por pessoas. Sistemas que possibilitam a realização de buscas proveem em geral funcionalidades que facilitam o processo. Contudo, os sistemas de

recuperação de informação em sua maioria, não possuem a capacidade de prover informações implícitas por meio da utilização de inferências. Neste sentido, o protótipo tenta explorar esta lacuna.

A aplicação do protótipo foi realizada em âmbito acadêmico visando pesquisar métodos ou estratégias para a melhoria do processo de entendimento do conteúdo coberto por determinada coleção de documentos. Nos exemplos apresentados na próxima seção será realizada uma apresentação de consultas simples realizadas detalhando as possibilidades de análise feitas sobre o que é retornado e como um usuário poderia navegar na informação e obter resultados adicionais aos obtidos em SRIs tradicionais a partir de algumas interações.

Para a elaboração do cenário houve uma preocupação em selecionar documentos que possuíssem autores em comum, que os autores possuíssem uma relação de colegas de trabalho, ou seja, que atuassem na mesma instituição, que os descritores (palavras-chave) ocorressem em mais de um documento visando demonstrar as intersecções do domínio (área de Ontologia) ao usuário que realiza a consulta.

O protótipo se utiliza de um conjunto de 12 documentos que são reproduções resumidas dos documentos originais contidos na plataforma da *Science Direct*. Os documentos escolhidos tratam basicamente sobre o assunto de ontologias e são compostos no índice textual pelos campos de chave (um número sequencial que identifica os documentos durante o processo de atualização), título, ano, autor, instituição, resumo e palavras-chave.

## 5.2 DISCUSSÃO

No primeiro cenário supõe-se que um usuário que esteja iniciando na área de Ontologia, tenha interesse em buscar documentos que abordem o conteúdo de compartilhamento e reuso de ontologias e também aprender quais conceitos envolvem esta temática, buscando descobrir qual a abrangência da área. Desta forma, o usuário faz uma consulta no sistema com intenção de selecionar os documentos que possuam o termo “*Ontologies*” nas palavras-chave, campo “*keyword*”.

Até chegar ao documento que considere útil, o usuário pode navegar nas outras informações de autores como colegas de instituição ou então coautores. A seguir são apresentadas algumas das propriedades de objeto que são retornados pelo sistema a partir de

uma consulta com o termo “*Ontologies*” (Figura 20). A navegação pelos demais conceitos da ontologia deve ser realizada manualmente, ou seja, o usuário deve realizar uma nova consulta com o novo termo. Isso é requerido uma vez que está fora do escopo do trabalho o desenvolvimento de uma interface amigável.

Para demonstrar o potencial de investigação do conteúdo em torno do termo “*Ontologies*” pode-se supor que o usuário tenha interesse em outro descritor que esteja relacionado, neste caso “*OWL*”. Desta forma, refazendo a consulta com o termo “*OWL*” destacado na Figura 20 é obtido o resultado apresentado na Figura 21.

```

=====
||                               Dados da ontologia                               ||
=====
Termo:
  Ontologies

Classe:
  (Processo afirmado) Descritor

Propriedades de Objeto:
  (Processo inferido) temRelacaoDesc >> "ITSM"
  (Processo inferido) temRelacaoDesc >> "Translational medicine applications"
  (Processo inferido) temRelacaoDesc >> "RDF views"
  (Processo inferido) temRelacaoDesc >> "Knowledge representation"
  (Processo inferido) temRelacaoDesc >> "ITIL"
  (Processo inferido) temRelacaoDesc >> "OWL"
  (Processo inferido) temRelacaoDesc >> "Process modeling"
  (Processo inferido) temRelacaoDesc >> "Vocabularies"
  (Processo inferido) temRelacaoDesc >> "SWRL"
  (Processo inferido) temRelacaoDesc >> "SQWRL"
  (Processo inferido) ehDescritor >> "Applying an ontology approach to IT service management ..."
  (Processo inferido) ehDescritor >> "vSPARQL: A view definition language for the semantic web"

Propriedades de Dado:

=====
||                               Dados da indexação                               ||
=====
Número de documentos retornados ao consultar o índice: 3

1º documento
Título: vSPARQL: A view definition language for the semantic web
Ano: 2011
Autor(es): Marianne Shaw; Landon T. Detwiler; Natalya F. Noy; James Brinkley; Dan Suciú

2º documento
Título: Applying an ontology approach to IT service management for business-IT integration
Ano: 2012
Autor(es): Maria-Cruz Valiente; Elena Garcia-Barriocanal; Miguel-Angel Sicilia

3º documento
Título: Combining OWL ontologies using -Connections
Ano: 2006
Autor(es): Bernardo Cuenca Grau; Bijan Parsia; Evren Sirin

```

**Figura 20** – Resultado da pesquisa pelo descritor "Ontologies"

Dentre as propriedades de dado retornadas o usuário resolve então selecionar o documento “*Translating the Foundational Model of Anatomy into OWL*” (em destaque na Figura 21) cujo um dos descritores é “OWL”. O resultado é apresentado na Figura 22.

```

=====
||                               Dados da ontologia                               ||
=====
Termo:
OWL
Classe:
(Processo afirmado) Descritor
Propriedades de Objeto:
.
.
(Processo inferido) ehDescritor >> "Translating the Foundational Model of Anatomy into OWL"
(Processo inferido) ehDescritor >> "Applying an ontology approach to IT service management..."
(Processo inferido) ehDescritor >> "Swoop: A Web Ontology Editing Browser"

=====
||                               Dados da indexação                               ||
=====
Número de documentos retornados ao consultar o índice: 3

1º documento
Título: Swoop: A Web Ontology Editing Browser
Ano: 2006
Autor(es): Aditya Kalyanpur; Bijan Parsia; Evren Sirin; Bernardo Cuenca Grau; James Hendler

2º documento
Título: Translating the Foundational Model of Anatomy into OWL
Ano: 2008
Autor(es): Natalya F. Noy; Daniel L. Rubin

3º documento
Título: Applying an ontology approach to IT service management for business-IT integration
Ano: 2012
Autor(es): Maria-Cruz Valiente; Elena Garcia-Barriocanal; Miguel-Angel Sicilia

```

**Figura 21** – Resultado da pesquisa pelo descritor "OWL"

A partir das informações referentes ao documento “*Translating the Foundational Model of Anatomy into OWL*” se o usuário decidir realizar uma consulta com a autora Natalya F. Noy (em destaque na Figura 22) ele obterá o resultado apresentado na Figura 23.

```

=====
||                               Dados da ontologia                               ||
=====
Termo:
  Translating the Foundational Model of Anatomy into OWL

Classe:
  (Processo afirmado) Documento

Propriedades de Objeto:
  (Processo afirmado) temDescritor >> "Foundational Model of Anatomy"
  (Processo afirmado) temDescritor >> "OWL"
  (Processo afirmado) temDescritor >> "Semantic Web"
  (Processo afirmado) temDescritor >> "Ontology"
  (Processo inferido) temAutor >> "Daniel L. Rubin"
  (Processo inferido) temAutor >> "Natalya F. Noy"
  .
  .
  .
  (Processo inferido) temAutor >> "Natalya F. Noy"
  (Processo inferido) temAutor >> "Natalya F. Noy"

Propriedades de Dado:
  (Processo afirmado) ano_documento >> "2008"
  (Processo afirmado) resumo_documento >> "The Foundational Model of Anatomy (FMA) represents...
  (Processo afirmado) titulo_documento >> "Translating the Foundational Model of Anatomy into OWL"

=====
||                               Dados da indexação                               ||
=====
Número de documentos retornados ao consultar o índice: 1

1º documento
Título: Translating the Foundational Model of Anatomy into OWL
Ano: 2008
Autor(es): Natalya F. Noy; Daniel L. Rubin

```

**Figura 22** – Resultado da pesquisa pelo documento "*Translating the Foundational Model of Anatomy into OWL*"

Na tela de apresentação dos dados da “Natalya F. Noy” (Figura 23), o usuário pode notar que ela é autora de um documento com título “*Where to publish and find ontologies? A survey of ontology libraries*” em que ele se interessa, pois se trata de uma revisão e ele está começando suas pesquisas na área.

```

=====
||                               Dados da ontologia                               ||
=====
Termo:
  Natalya F. Noy

Classe:
  (Processo afirmado) Pessoa

Propriedades de Objeto:
  (Processo afirmado) ehAutor >> "Translating the Foundational Model of Anatomy into OWL"
  (Processo afirmado) ehAutor >> "Where to publish and find ontologies? A survey of ontology libraries"
  (Processo afirmado) ehAutor >> "vSPARQL: A view definition language for the semantic web"
  .
  .

Propriedades de Dado:
  (Processo afirmado) email_pessoa >> "noy@stanford.edu"
  (Processo afirmado) nome_pessoa >> "Natalya F. Noy"

=====
||                               Dados da indexação                               ||
=====
Número de documentos retornados ao consultar o índice: 3

1º documento
Título: Translating the Foundational Model of Anatomy into OWL
Ano: 2008
Autor(es): Natalya F. Noy; Daniel L. Rubin

2º documento
Título: Where to publish and find ontologies? A survey of ontology libraries
Ano: 2012
Autor(es): Mathieu d'Aquin; Natalya F. Noy

3º documento
Título: vSPARQL: A view definition language for the semantic web
Ano: 2011
Autor(es): Marianne Shaw; Landon T. Detwiler; Natalya F. Noy; James Brinkley; Dan Suciu

```

**Figura 23** – Resultado da pesquisa pela autora “Natalya F. Noy”

Ao selecionar o documento “*Where to publish and find ontologies? A survey of ontology libraries*” o usuário pode verificar o ano em que o documento foi publicado, o resumo, o título, os autores do documento, documentos relacionados e os descritores.

É importante salientar que o sistema fornece informações que podem ajudar o usuário durante suas pesquisas de modo que este possa ir além daquilo que almejava encontrar inicialmente. Visão similar é apresentada no trabalho de Beppler (2008), em que é proposto um modelo de recuperação de informação onde um ambiente interativo possibilita a experiência de buscas mais naturais.

No segundo cenário, o usuário percebe que a autora Natalya F. Noy escreve documentos relacionados à área de Ontologia e resolve pesquisar sobre sua rede de relações profissionais e publicações que circundam sua atividade profissional.

A autora Natalya F. Noy continuará sendo utilizada como exemplo porque de um total de 30 autoras cadastradas na ontologia, 9 estão relacionadas a ela, deste modo, é a pessoa que tem o maior número de relacionamentos considerando colegas e coautores.

Ao consultar o nome da autora o usuário pode notar que ela possui uma rede de relacionamentos abrangente. O usuário então pode passar a pesquisar sobre as pessoas relacionadas para identificar a representatividade que eles têm no âmbito acadêmico. Foram selecionado 2 coautores (Figura 24 e Figura 25) e 1 colega de trabalho (Figura 26) de “Natalya F. Noy” que possuem contextos semelhantes, ou seja, poucos coautores, poucos colegas de trabalho e com pouca produção entre eles e os demais autores para poder exemplificar como o usuário pode identificar os autores.

```

=====
||                               Dados da ontologia                               ||
=====
Termo:
  Dan Suciú

Classe:
  (Processo afirmado) Pessoa

Propriedades de Objeto:
  (Processo afirmado) ehAutor >> "vSPARQL: A view definition language for the semantic web"
  (Processo inferido) temCoAutor >> "Landon T. Detwiler"
  (Processo inferido) temCoAutor >> "Marianne Shaw"
  (Processo inferido) temCoAutor >> "James Brinkley"
  (Processo inferido) temCoAutor >> "Natalya F. Noy"
  (Processo inferido) ehColega >> "Landon T. Detwiler"
  (Processo inferido) ehColega >> "Marianne Shaw"
  (Processo inferido) ehColega >> "James Brinkley"
  (Processo afirmado) pertenceInstituicao >> "University of Washington"
  .
  .
  .

```

**Figura 24** – Propriedades do indivíduo Dan Suciú

Dados da ontologia	
Termo:	Daniel L. Rubin
Classe:	(Processo afirmado) Pessoa
Propriedades de Objeto:	
(Processo afirmado) ehAutor >>	"Translating the Foundational Model of Anatomy into OWL"
(Processo inferido) temCoAutor >>	"Natalya F. Noy"
(Processo inferido) ehColega >>	"Paea LePendu"
(Processo inferido) ehColega >>	"Nigam H. Shah"
(Processo inferido) ehColega >>	"Mark A. Musen"
(Processo inferido) ehColega >>	"Natalya F. Noy"
(Processo afirmado) pertenceInstituicao >>	"Stanford University"
.	
.	
.	

**Figura 25** - Propriedades do indivíduo Daniel L. Rubin

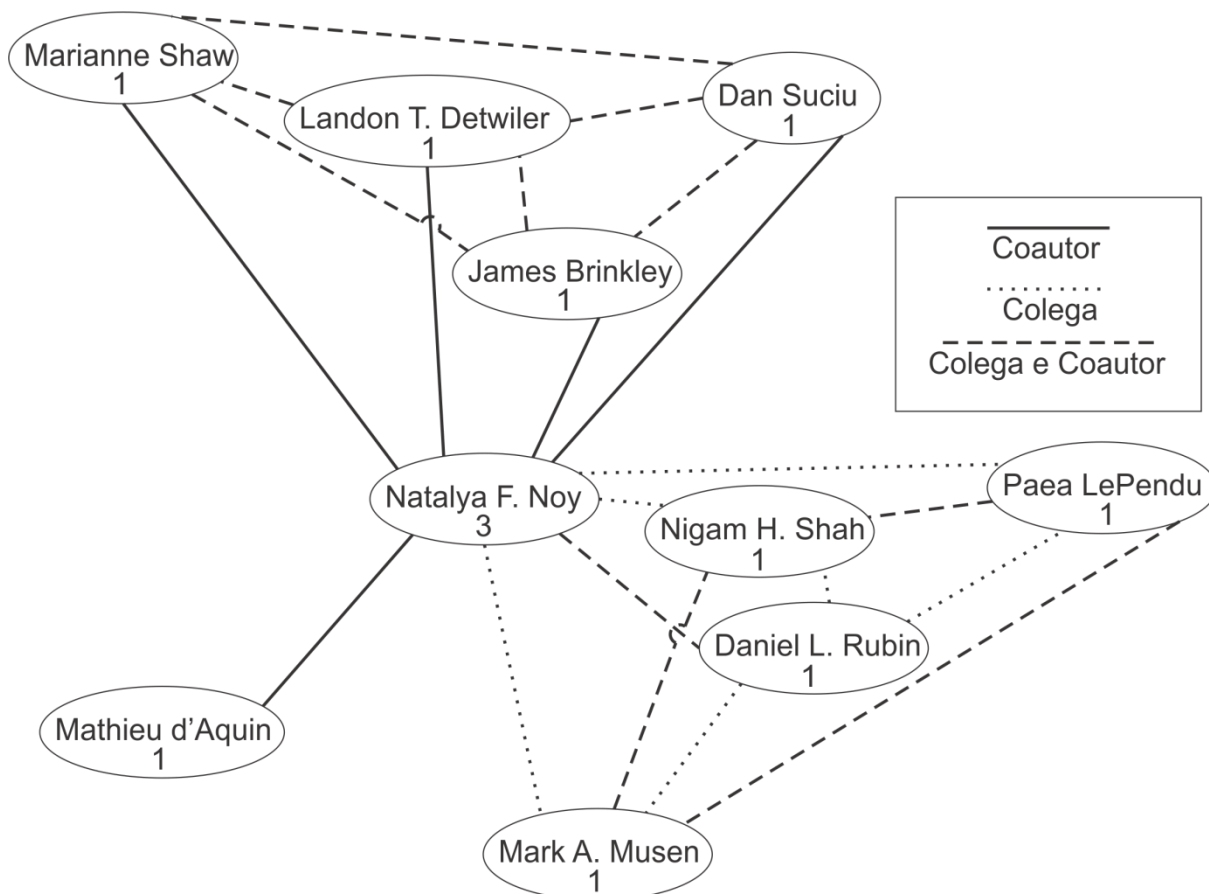
Dados da ontologia	
Termo:	Mark A. Musen
Classe:	(Processo afirmado) Pessoa
Propriedades de Objeto:	
(Processo afirmado) ehAutor >>	"Enabling enrichment analysis with the Human Disease Ontology"
(Processo inferido) temCoAutor >>	"Paea LePendu"
(Processo inferido) temCoAutor >>	"Nigam H. Shah"
(Processo inferido) ehColega >>	"Paea LePendu"
(Processo inferido) ehColega >>	"Daniel L. Rubin"
(Processo inferido) ehColega >>	"Nigam H. Shah"
(Processo inferido) ehColega >>	"Natalya F. Noy"
(Processo afirmado) pertenceInstituicao >>	"Stanford University"
.	
.	
.	

**Figura 26** - Propriedades do indivíduo Mark A. Musen

O usuário ao navegar entre as pessoas cadastradas no sistema irá notar que o autor possui mais ou menos publicações assim como autores relacionados. Com essa análise pode-se identificar qual autor é mais influente, assim como, identificar quais são os autores de referência na área pesquisada.

Esse processo permite que naturalmente uma pessoa possa descobrir quais são os autores mais influentes de uma determinada área, pelo volume de publicações e coautores. Na Figura 27 é possível observar na forma de grafo qual é o autor mais representativo na área de Ontologia. O número presente dentro de cada elipse é a quantidade de publicações que o autor realizou.





**Figura 27** - Rede de relações dos autores que circundam a autora "Natalya F. Noy"

Com base nos dados retornados pela aplicação é possível perceber que a maioria dos autores se relacionam com outros 4 autores. Outro fator que deve ser levado em consideração é que autores com maior número de coautores são mais importantes em termos de referência do que os autores que tem mais relações de colegas de trabalho.

A autora "Natalya F. Noy" é a mais representativa em relação a coautorias, totalizando 6 autores diferentes, e em número de publicações, totalizando 3 publicações. A funcionalidade de fornecer informações sobre autores, suas relações diretas com outros autores e o número de documentos publicados constitui-se no foco do protótipo. Com algumas iterações nas pesquisas é possível encontrar facilmente qual é o autor que é a referência na área.

Por fim, é importante ressaltar que essa configuração não necessariamente condiz com a realidade. Os cenários foram modelados para demonstrar uma possível utilização do sistema e seu funcionamento no contexto acadêmico.

## 6. CONSIDERAÇÕES FINAIS

O objetivo geral desse trabalho foi apresentar um sistema que possibilitasse aplicar semântica aos motores de busca. A partir da análise dos sistemas de recuperação mais utilizados foram buscadas referências para entender o funcionamento destes e sua evolução. Com base nos dados obtidos foi possível realizar estudos e avaliações para definir as possibilidades de construção de um protótipo que permita a adição de conteúdo complementar ao disponível no *corpus*, ou seja, a semântica.

Atualmente, se um usuário deseja conhecer melhor autores de uma determinada área, este deve realizar um processo de pesquisa que poderá ser demorado quando considerado o modelo atual dos mecanismos de busca. Necessitará assim, de um período de reconhecimento para saber quais autores são mais mencionados em textos da área. Neste sentido, o protótipo desenvolvido pode ser visto como um facilitador, uma vez que o usuário pode-se utilizar de caminhos alternativos para analisar e, possivelmente, entender o contexto de determinados autores. O processo de busca por documento com assuntos relevantes também ficou facilitado. Os meios de determinar relacionamento entre documentos através de assuntos ou autores contribuem para que o usuário tenha um melhor proveito de suas pesquisas.

Quanto ao funcionamento geral, o protótipo demonstra potencial no que tange o entendimento de determinado domínio, sendo este o foco principal do trabalho. A análise do desempenho em relação ao tempo de consulta tanto no índice quanto na ontologia não é relevante neste trabalho, pois o índice textual e a ontologia são pequenos. Neste sentido, as APIs OWL e Lucene, corresponderam de maneira satisfatória no provimento de um mecanismo de busca e um mecanismo de representação de conhecimento e inferência.

Apesar disto, escolhas foram tomadas durante o processo de construção do protótipo quando sua composição lógica e física ainda não estava efetivamente consolidada. A primeira delas refere-se ao processo de anotação semântica automatizada. Em função da complexidade desse processo, seria requerida a utilização de técnicas de Processamento de Linguagem Natural, o que vai além do escopo do trabalho. Dessa forma a anotação semântica e a ontologia foram preenchidas manualmente considerando os documentos selecionados na Plataforma *Science Direct*.

Existem ainda possibilidades de se aprimorar a forma de indexação textual, entretanto, durante a construção do protótipo foi utilizada uma versão do Lucene (3.0.3) que não é capaz de indexar subcampos, o que permite a composição de estrutura hierárquica em cada campo. Atualmente, toda a informação, ainda que tenha hierarquia, deve ser concatenada em um mesmo campo o que em muitos casos pode misturar os conteúdos promovendo um resultado errôneo em uma busca. Para o protótipo isto seria de grande auxílio, pois todas as informações afirmadas ou inferidas que constam na ontologia poderiam residir em subcampos do índice, aumentando o desempenho das consultas se considerados grandes índices.

É importante mencionar que outra limitação existente é que as consultas quando submetidas à ontologia devem ser realizadas com termos na mesma forma em que foram armazenados, além disso, as buscas realizadas na ontologia são processadas de maneira sequencial, em que todos os indivíduos são inspecionados. Uma ontologia em si não possui uma estrutura de índice, contudo, isso poderia ser resolvido por meio da integração da ontologia em um índice textual como discutido anteriormente, ou mesmo, em bancos de dados específicos para este fim.

Durante o desenvolvimento do trabalho outras possibilidades foram vislumbradas como trabalhos futuros. Entre estas possibilidades menciona-se o desenvolvimento de uma interface que permita a integração, tanto da sistemática de navegação discutida neste trabalho quanto à utilização de grafos para a apresentação das informações. Além disso, a adição de técnicas de análise de rede pode facilitar em muito a compreensão em como indivíduos de determinado domínio do conhecimento se interconectam. Deste modo, o usuário poderia encontrar autores de referência em alguma área muito mais facilmente do que navegar manualmente nas informações para identificar quais destes possuem maior influência em determinada área.

Pode-se vislumbrar ainda o desenvolvimento de um sistema de anotação semântica capaz de extrair padrões relevantes de determinado documento (chamados de entidades) e os relacionamentos entre estes. Estas informações poderiam ser atribuídas diretamente à ontologia reduzindo o esforço operacional em mantê-la atualizada. Entretanto, vale mencionar que, apesar da evolução dos métodos e técnicas na área de extração de informação, a participação de um especialista em determinado domínio do conhecimento para realizar possíveis correções e evoluções em uma ontologia é fundamental.

## REFERÊNCIAS

AHMEDI, Lule; ABAZI-BEXHETI, Lejla; KADRIU, Arbana. A uniform semantic web framework for co-authorship networks. In: Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC), 9. 2011, Sydney. **Proceedings...** Sydney: DASC, 2011. p. 958 - 965.

ALMEIDA, Mauricio B.; BAX, Marcello P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ci. Inf.**, Brasília, v. 32, n. 3, p. 7-20, Dec. 2003.

ALZGHOOL, Muath; INKPEN, Diana. Clustering the topics using TF-IDF for model fusion. In: PIKM '08 Proceedings of the 2nd PHD Workshop on Information and Knowledge Management, 2. 2008, Napa Valley. **Proceedings...** Napa Valley: ACM, 2008. p. 97 - 100.

AMATI, Gianni; RIJSBERGEN, Cornelis Joost Van. **Probabilistic models of information retrieval based on measuring the divergence from randomness**. ACM Transactions on Information Systems, Pittsburgh, v. 20, n. 4, p.357-389, out. 2002.

AMAZONAS, André Mano et al. Integrando motores de indexação de dados para a construção de sistemas de recuperação de informação em ambientes heterogêneos. **JISTEM J.Inf.Syst. Technol. Manag. (Online)**, São Paulo, v. 5, n. 2, 2008. Disponível em: <[http://www.scielo.br/scielo.php?script=sci\\_arttext&pid=S1807-17752008000200002&lng=en&nrm=iso](http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-17752008000200002&lng=en&nrm=iso)>. Acesso em: 27 Ago. 2012. <http://dx.doi.org/10.4301/S1807-17752008000200001>.

ANTONIOU, Grigoris; HARMELEN, Frank Van. **A semantic web primer**. 2nd Cambridge, MA: MIT Press, 2008.

APPEL, Andrew W. **Modern compiler implementation in Java**. 2nd Cambridge: The Press Syndicated Of The University Of Cambridge, 2004.

ARPÍREZ, Julio C. et al. WebODE: a scalable workbench for ontological engineering. In: K-CAP '01 Proceedings of the 1st International Conference on Knowledge Capture, 1. 2001, Victoria. **Proceedings...** Victoria: ACM, 2001. p. 6 - 13.

BECHHOFER, Sean et al. OilEd: a reason-able ontology editor for the semantic web. In: KI 2001: Advances in Artificial Intelligence, 24. 2001, Vienna. **Proceedings...** Vienna: Springer, 2001. p. 396 - 408.

BEPPLER, Fabiano Duarte. **Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico**. 2008. 135 f. Tese (Doutorado) - Universidade Federal de Santa Catarina, Florianópolis, 2008.

BERNERS-LEE, Tim. **Semantic web on XML**. Washington: World Wide Web Consortium, 2000. 17 slides, color. Acompanha texto. Documento apresentado na XML 2000 Conference. Disponível em: <<http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>>. Acesso em: 25 Set. 2012.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. **Scientific American**, [S. L.], v. 284, n. 5, p.34-43, maio 2001.

BIZER, Christian; BERNERS-LEE, Tim; HEATH, Tom. Linked data: the story so far. **International Journal On Semantic Web And Information Systems**, v. 5, n. 3, p.1-22, 20 Out. 2009.

BLANCO, Ignacio J.; VILA, M. Amparo; MARTINEZ-CRUZ, Carmen. The use of ontologies for representing database schemas of fuzzy information. **International Journal Of Intelligent Systems**, [ S. L.], v. 23, n. 4, p.419-445, 2008.

BORST, Willem Nico. **Construction of engineering ontologies for knowledge sharing and reuse**. 1997. 38 f. Tese (Phd) - University Of Twente, Enschede, 1997.

BOVO, Alessandro Botelho. **Um modelo de descoberta de conhecimento inerente à evolução temporal dos relacionamentos entre elementos textuais**. 2011. 155 f. Tese (Doutorado) - Curso de Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2011.

BOZSAK, Erol et al. KAON: Towards a large scale semantic web. In: (EC-WEB'02) E-Commerce and Web Technologies Third International Conference, 3. 2002, Aix-en-provence. **Proceedings...** Aix-en-provence: Springer, 2002. p. 231 - 248.

BREITMAN, K. K.; CASANOVA, M. A.; TRUSZKOWSKI, W. **Semantic web: concepts, technologies and applications**. London: Springer, 2007. 327 p. ISBN 9781846285813.

BREITMAN, Karin Koogan. **Web semântica: a internet do futuro**. Rio de Janeiro: LTC, 2006. xvii, 190 p. ISBN 978-85-216-1466-1.

BREITMAN, Karin Koogan; LEITE, Julio Cesar Sampaio do Prado. Lexicon Based Ontology Construction. In: LUCENA, Carlos et al. **Software engineering for multi-agent systems II: research issues and practical applications**. Heidelberg: Springer-Verlag, 2004. p. 19-34.

BUSH, Vannevar. As we may think. **The Atlantic Monthly**, Boston, v. 176, n. 1, p.101-108, Jul. 1945.

CAO, Luhui; LI, Qingzhong; GAO, Xiang. A framework for personal information integration in organizations. In: Web Information Systems and Applications Conference, 6. 2009, Busan. **Proceedings...** Busan: IEEE, 2009. p. 206 - 209.

CARDOSO, Olinda Nogueira Paes. Recuperação de informação. **Infocomp: Journal of Computer Science**, Lavras, v. 2, n. 1, p.33-38, 2000.

CECI, Flávio. **Um modelo semiautomático para a construção e manutenção de ontologias a partir de bases de documentos não estruturados**. 2010. 131 f. Dissertação (Mestrado) - Universidade Federal de Santa Catarina, Florianópolis, 2010.

CHAI, Yanmei et al. Intelligent digital photo management system using ontology and SWRL. In: The 7th International Conference on Computational Intelligence and Security (CIS), 7. 2010, Nanning. **Proceedings...** Nanning: CIS, 2010. p. 18 - 22.

CHIARAMELLA, Y.; CHEVALLET, J. P. About retrieval models and logic. **The Computer Journal**, Grenoble, v. 35, n. 3, p.233-242, 1992.

CHRISTEN, Peter. A survey of indexing techniques for scalable record linkage and deduplication. **IEEE Transactions On Knowledge And Data Engineering**, [S. L.], v. 24, n. 9, p.1537-1555, Set. 2012.

COOPER, William S. Getting beyond boole. **Information Processing & Management**, [S. L.], v. 24, n. 3, p.243-248, 1988.

CORCHO, Oscar; FERNÁNDEZ-LÓPEZ, Mariano; GÓMEZ-PÉREZ, Asunción. Methodologies, tools and languages for building ontologies. Where is their meeting point? **Data & Knowledge Engineering**, [S. L.], v. 46, n. 1, p.41-64, 2003.

CZAUDERNA, Adam et al. Traceability challenge 2011: using TraceLab to evaluate the impact of local versus global IDF on trace retrieval. In: TEFSE '11 Proceedings of the 6th International Workshop on Traceability in Emerging Forms of Software Engineering, 6. 2011, Waikiki. **Proceedings...** Waikiki: ACM, 2011. p. 75 - 78.

DEERWESTER, Scott et al. Indexing by latent semantic analysis. **Journal Of The American Society For Information Science**, [Silver Spring], v. 41, n. 6, p.391-407, 1990.

DOMINGUE, John. Tadzebao and WebOnto: discussing, browsing, and editing ontologies on the web. In: Eleventh Workshop on Knowledge Acquisition, Modeling and Management, 11. 1998, Banff. **Proceedings...** Banff: EKAW, 1998.

EGGHE, L.; MICHEL, C. Strong similarity measures for ordered sets of documents in information retrieval. **Information Processing and Management: an International Journal**, v. 38, n. 6, p. 823-848, 2002.

FANG, Hui; ZHAI, Chengxiang. Semantic term matching in axiomatic approaches to information retrieval. In: PROCEEDINGS OF THE 29TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 29, 2006, Seattle. **Proceedings...** New York: Acm, 2006. p. 115 - 122.

FARQUHAR, Adam; FIKES, Richard; RICE, James. The Ontolingua server: a tool for collaborative ontology construction. **International Journal Of Human-computer Studies**, [S. L.], v. 46, n. 6, p.707-727, 1997.

FAZZINGA, Bettina et al. Semantic Web search based on ontological conjunctive queries. **Web Semantics: Science, Services and Agents on the World Wide Web**, [S. L.], v. 9, n. 4, p.453-473, 2011.

FAZZINGA, Bettina; LUKASIEWICZ, Thomas. Semantic search on the Web. **Semantic Web: Interoperability, Usability, Applicability**, [S. L.], v. 1, n. 1-2, p.89-96, 2010.

FENSEL, Dieter et al. OIL: An ontology infrastructure for the semantic web. **IEEE Intelligent Systems**, [New York], v. 16, n. 2, p.38-45, 2001.

FENSEL, Dieter. Ontology-based knowledge management. **Computer**, [S. L.], v. 35, n. 11, p.56-59, 2002.

FERNÁNDEZ, Miriam et al. Semantically enhanced Information Retrieval: An ontology-based approach. **Web Semantics: Science, Services and Agents on the World Wide Web**, [S. L.], v. 9, n. 4, p.434-452, 2011.

FERNÁNDEZ-LÓPEZ, Mariano; GÓMEZ-PÉREZ, Asunción; JURISTO, Natalia. Methontology: from ontological art towards ontological engineering. In: Proceedings of the AAAI-97 Spring Symposium Series on Ontological Engineering, 14, 1997, Stanford. **Proceedings...** Stanford: American Association for Artificial Intelligence, 1997. p. 33 - 40.

FERNEDA, Edberto. **Recuperação de informação: análise sobre a contribuição da ciência da computação para a ciência da informação**. 2003. 147 f. Tese (Doutorado) - Universidade de São Paulo, São Paulo, 2003.

FININ, T. et al. ROWLBAC: representing role based access control in OWL. In: SACMAT '08 Proceedings of the 13th ACM Symposium on Access Control Models and Technologies, 13. 2008, Estes Park. **Proceedings...** Estes Park: ACM, 2008. p. 73 - 82.

FRAKES, William B.; BAEZA-YATES, Ricardo. **Information retrieval: data structures and algorithms**. Upper Saddle River: Prentice Hall, 1992. 464 p.

GOLBREICH, Christine; IMAI, Atsutoshi. Combining SWRL rules and OWL ontologies with Protégé OWL Plugin, Jess, and Racer. In: International Protégé Conference, 7. 2004, Bethesda. **Proceedings...** Bethesda: 2004. p. 1 - 4.

GOMES, Roberto Miranda. **Desambiguação de sentido de palavra dirigida por técnicas de agrupamento sob o enfoque da mineração de textos**. Rio de Janeiro, 2009, 118p. Dissertação de mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

GONÇALVES, Alexandre Leopoldo. **Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à engenharia e gestão do conhecimento**. 2006. 196 f. Tese (Doutorado em Engenharia de Produção) ênfase em Inteligência Aplicada - Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis.

GOULART, Mauro Sérgio Boppré. Uso da informação empresarial no processo de decisão estratégica em empresas de base tecnológica - EBTS: o caso do centro empresarial para laboração de tecnologias avançadas - CELTA. **Perspect. ciênc. inf.**, Belo Horizonte, v. 12, n. 1, Abr. 2007.



GRUBER, Thomas R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, Stanford, v. 5, n. 2, p.199-220, 1993.

GRUBER, Thomas R. Toward principles for the design of ontologies used for knowledge sharing. **International Journal Human-computer Studies**, [S. L.], v. 43, n. 5-6, p.907-928, 1995.

GRÜNINGER, Michael; FOX, Mark S. Methodology for the design and evaluation of ontologies. In: Proceedings of International Joint Conference on Artificial Intelligence's (IJCAI-95) Workshop on Basic Ontological Issues in Knowledge Sharing, 14. 1995, Montreal. **Proceedings...** Toronto: Ijcai, 1995. p. 1 - 10.

GUARINO, Nicola. Semantic matching: formal ontological distinctions for information organization, extraction, and integration. In: Information Technology, International Summer School, SCIE-97, 1. 1997, Frascati. **Proceedings...** Padova: Springer Verlag, 1997. p. 139 - 170.

GUHA, R.; MCCOOL, Rob; MILLER, Eric. Semantic search. In: Proceedings of the 12th International Conference on World Wide Web (WWW '03), 12. 2003, New York. **Proceedings...** New York: ACM, 2003. p. 700 - 709.

GUPTA, Samriti; JINDAL, Alka. Contrast of link based web ranking techniques. In: International Symposium on Biometrics and Security Technologies, 2. 2008, Islamabad. **Proceedings...** Islamabad: IEEE, 2008. p. 1 - 6.

HAASE, Peter et al. The NeOn Ontology engineering toolkit. In: WWW 2008: DEVELOPERS TRACK, 17. 2008, Beijing. **Proceedings...** Beijing: WWW2008, 2008. p. 1 - 2.

HANSEN, Morten T.; BIGRUAM, Nitin; TIERNEY, Thomas. What's your strategy for managing knowledge? **Harvard Business Review**, v. 77, n. 2, p. 106-116, Mar./Apr. 1999.

HARMAN, Donna. Overview of the first TREC conference. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development In Information Retrieval, 16. 1993, Pittsburgh. **Proceedings...** Gaithersburg,: ACM, 1993. p. 36 - 47.

HATCHER, Erik; GOSPODNETIĆ, Otis; MCCANDLESS, Michael. **Lucene in action:** covers Apache Lucene v.3.0. 2nd Greenwich: Manning Publications, 2009.

HEBELER, John W.; VAN DOREN, Doris C. Unfettered leverage: the ascendancy of knowledge-rich products and processes. **Business Horizons**, v. 40, n. 4, p. 2-10, 1997.

HENDLER, James. Web 3.0: The Dawn of Semantic Search. **Computer**, [S. L.], v. 43, n. 1, p.77-80, 2010.

HEO, Sun-young; KIM, Eun-gyung. A study on the improvement of query processing performance of OWL data based on Jena2. In: International Conference on Convergence and Hybrid Information Technology 2008 (ICHIT '08), 3. 2008, Busan. **Proceedings...** Busan: IEEE, 2008. p. 678 - 681.

HIEMSTRA, Djoerd. A linguistically motivated probabilistic model of information retrieval. In: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, 2. 1998, University Of Twente,. **Proceedings...** London: Springer-Verlag, 1998. p. 569 - 584.

HILBERT, Martin; LÓPEZ, Priscila. The world's technological capacity to store, communicate, and compute information. **Science**, United States, v. 332, n. 6065, p.60-65, 01 Abr. 2011.

HOGAN, Aidan et al. Searching and browsing Linked Data with SWSE: the semantic web search engine. **Web Semantics: Science, Services and Agents on the World Wide Web**, v. 9, n. 4, p.365-401, Dez. 2011.

HORRIDGE, Matthew et al. **A practical guide to building OWL ontologies using the Protégé-OWL plugin and CO-ODE tools edition 1.0**. Manchester: The University Of Manchester, 2004.

HORROCKS, Ian et al. **SWRL: a semantic web rule language combining OWL and RuleML**. W3C, 2004. Disponível em: <<http://www.w3.org/Submission/SWRL/>>. Acesso em: 21 Nov. 2012.

JONES, Dean; BENCH-CAPON, Trevor; VISSER, Pepijn. Methodologies for ontology development. In: PROC. IT&Knows Conference of the 15th IFIP World Computer Congress, 15. 1998, Budapeste. **Proceedings...** Budapeste: Chapman-hall, 1998. p. 62 - 75.

JONES, Karen Spärck. A statistical interpretation of term specificity and its application in retrieval. **Journal Of Documentation**, Cambridge, v. 28, n. 1, p.11-21, 1972.

JONES, William P.; FURNAS, George W. Pictures of relevance: a geometric analysis of similarity measures. **Journal of the American Society for Information Science**, [New York], v. 38, n. 6, p.420-442, 1987.

KIRYAKOV, Atanas et al. Semantic annotation, indexing, and retrieval. **Journal Of Web Semantics**, [S. L.], v. 2, p.49-79, 2004.

KIYAVITSKAYA, Nadzeya et al. Cerno: Light-weight tool support for semantic annotation of textual documents. **Data & Knowledge Engineering**, [S. L.], v. 68, n. 12, p.1470-1492, 2009.

KLYNE, Graham; CARROLL, Jeremy J. **Resource Description Framework (RDF): concepts and abstract syntax**. W3C, 2004. Disponível em: <<http://www.w3.org/TR/rdf-concepts/>>. Acesso em: 26 Jul. 2012.

KNUBLAUCH, Holger et al. The Protégé OWL experience. In: Fourth International Semantic Web Conference (ISWC2005), 4. 2005, Galway. **Proceedings...** Galway: ISWC, 2005. p. 1 - 11.

KORFHAGE, Robert R. **Information storage and retrieval**. New York: John Wiley & Sons, Inc., 1997.

LI, Man; DU, Xiaoyong; WANG, Shan. A semi-automatic ontology acquisition method for the semantic web. In: WAIM'05 Proceedings of the 6th International Conference On Advances in Web-Age Information Management, 6. 2005, Hangzhou. **Proceedings...** Heidelberg: Springer-Verlag, 2005. p. 209 - 220.

LIMA, Gercina Ângela Borém. Interfaces entre a ciência da informação e a ciência cognitiva. **Ci. Inf.**, Brasília, v. 32, n. 1, Apr. 2003.

LUHN, Hans Peter. A statistical approach to mechanized encoding and searching of literary information. **Ibm Journal of Research and Development**, Riverton, v. 1, n. 4, p.309-317, Out. 1957.

LUPIANI-RUIZ, Eduardo et al. Financial news semantic search engine. **Expert Systems With Applications**, [S. L.], v. 38, n. 12, p.15565-15572, 2011.

MAEDCHE, Alexander et al. Ontologies for enterprise knowledge management. **IEEE Intelligent Systems**, [New York], v. 18, n. 2, p.26-33, 2003.

MAEDCHE, Alexander. **Ontology learning for the semantic web**. Norwell: Kluwer Academic Publishers, 2002.

MAIA, Luiz Cláudio; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspect. ciênc. inf.**, Belo Horizonte, v. 15, n. 1, Abr. 2010.

MAITI, Saptaditya; MANDAL, Deba P.; MITRA, Pabitra. Tackling content spamming with a term weighting scheme. In: Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, 2011, Beijing. **Proceedings...** Beijing: ACM, 2011. p. 1 - 5.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **An introduction to information retrieval**. Cambridge: Cambridge University Press, 2009.

MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. Cambridge: Mit Press, 1999.

MANOLA, Frank; MILLER, Eric. **RDF primer: W3C**, 2004. Disponível em: <<http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#rdfschema>>. Acesso em: 07 Out. 2012.

MATOS, Ely Edison da Silva. **CelOWS: um framework baseado em ontologias com serviços web para modelagem conceitual em biologia sistêmica**. 2008. 135 f. Dissertação (Mestrado) - Curso de Modelagem Computacional, Universidade Federal de Juiz de Fora, Juiz De Fora, 2008.

MATTOS, Merisandra Côrtes de; SIMÕES, Priscyla Waleska Targino de Azevedo; FARIAS, Renan Figueredo. A metodologia Methontology na construção de ontologias. **Revista de Iniciação Científica**, Criciúma, v. 5, n. 1, p.1-12, 2007.

MCGUINNESS, Deborah L.; HARMELEN, Frank Van. **OWL Web Ontology Language: overview**. W3C, 2004. Disponível em: <<http://www.w3.org/TR/owl-features/>>. Acesso em: 23 Out. 2012.

MEERSMAN, Robert. The use of lexicons and other computer-linguistic tools in semantics, design and cooperation of database systems. In: The Proceedings of the Second International Symposium on Cooperative Database Systems for Advanced Applications (CODAS'99), 2. 1999, Heidelberg. **Proceedings...** Heidelberg: Springer-Verlag, 1999. p. 1 - 14.

NAGARAJAN, G.; THYAGHARAJAN, K. K. A machine learning technique for semantic search engine. **Procedia Engineering**, [S. L.], v. 38, p.2164-2171, 2012.

NEWMAN, James R. **Volume three of the world of mathematics**. New York: Simon And Schuster, 1956. 4 v.

NOY, Natalya F.; MCGUINNESS, Deborah L. **Ontology development 101: A Guide to Creating Your First Ontology**. Stanford: 2001. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

NUNES, Anselmo Maciel; FILETO, Renato. Uma arquitetura para recuperação de informação baseada em semântica e sua aplicação no apoio a jurisprudência. In: Iii Escola Regional de Banco De Dados (ERBD), 3, 2007, Caxias do Sul. **Anais...** Caxias do Sul: UCS, 2007. p. 1 - 10.

NURMI, Raimo. Knowledge-intensive firms. **Business Horizons**, p. 26-31, May/June 1998.

O'CONNOR, Martin et al. Supporting rule system interoperability on the semantic web with SWRL. In: Fourth International Semantic Web Conference (ISWC2005), 4. 2005, Galway. **Proceedings...** Heidelberg: Springer-Verlag, 2005. p. 974 - 986.

O'CONNOR, Martin J.; DAS, Amar K. Acquiring OWL ontologies from XML documents. In: K-CAP '11 Proceedings of the Sixth International Conference on Knowledge Capture, 6. 2010, Banff. **Proceedings...** Banff: ACM, 2011. p. 17 - 24.

OLIVEIRA, Edgard Costa. **Autoria de documentos para a Web Semântica: um ambiente de produção de conhecimento baseado em ontologias**. 2006. 207 f. Tese (Doutorado) - Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2006.

ORLANDO, João Paulo. **Usando aplicações ricas para internet na criação de um ambiente para visualização e edição de regras SWRL**. 2012. 112 f. Dissertação (Mestrado) - Curso de Ciências da Computação e Matemática Computacional, Departamento de Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012.

OSUNA-ONTIVEROS, Daniel; LOPEZ-AREVALO, Ivan; SOSA-SOSA, Victor. A topic based indexing approach for searching in documents. In: International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE 2011), 8. 2011, Merida City. **Proceedings...** Merida City: IEEE, 2011. p. 1 - 6.

OWL WORKING GROUP. **OWL**. W3C, 2012. Disponível em: <<http://www.w3.org/2001/sw/wiki/OWL>>. Acesso em: 06 Out. 2012.

PAGE, Lawrence et al. **The PageRank citation ranking: bringing order to the web**. Stanford: Universidade de Stanford, 1998.

PALTOGLOU, Georgios; THELWALL, Mike. A study of information retrieval weighting schemes for sentiment analysis. In: ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 48. 2010, Uppsala. **Proceedings...** Uppsala: ACM, 2010. p. 1386 - 1395.

PASSIN, Thomas B. **Explorer's guide to the semantic web**. Greenwich: Manning Publications Co., 2004.

PIPITONE, Arianna; PIRRONE, Roberto. Cognitive linguistics as the underlying framework for semantic annotation. In: International Conference on Semantic Computing, 6. 2012, Palermo. **Proceedings...** Palermo: IEEE, 2012. p. 52 - 59.

POLTRONIERI, Anderson. **Modelo gráfico de recuperação de informação semântica**. 2006. 110 f. Dissertação (Mestrado) - Universidade Federal do Espírito Santo Centro Tecnológico, Vitória, 2006.

PONTE, Jay M.; CROFT, W. Bruce. A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 21. 1998, Melbourne. **Proceedings...** Amherst: ACM, 1998. p. 275 - 281.

PORTER, M. F. An algorithm for suffix stripping. **Program: electronic library and information systems**, Cambridge, v. 14, n. 3, p.130-137, Jul. 1980.

QIAO, Shaojie et al. SimRank: A Page Rank approach based on similarity measure. In: International Conference on Intelligent Systems and Knowledge Engineering, 5. 2010, Hangzhou. **Proceedings...** Hangzhou: IEEE, 2010. p. 390 - 395.

RADOVANOVIĆ, Miloš; NANOPOULOS, Alexandros; IVANOVIĆ, Mirjana. On the existence of obstinate results in vector space models. In: Proceedings Of The 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval SIGIR, 33. 2010, Geneva. **Proceedings...** New York: ACM, 2010. p. 186 - 193.

RDF WORKING GROUP. **RDF**. W3C, 2004. Disponível em: <<http://www.w3.org/RDF/>>. Acesso em: 05 Out. 2012.

REED, Stephen L.; LENAT, Douglas B. Mapping ontologies into Cyc. In: Proceedings of the 18th NAT'L Conf. Artificial Intelligence Workshop Ontologies for the Semantic Web (AAAI '02), 18. 2002, Edmonton. **Proceedings...** Austin: IEEE, 2002. p. 1 - 6.

RIBEIRO-NETO, Berthier; BAEZA-YATES, Ricardo. **Modern information retrieval**. 1. ed. Harlow: Addison Wesley, 1999. 513 p.

- RIGO, Wanderson. **Semântica e visualização para anotação e recuperação de informação**. 2011. 184 f. Dissertação (Mestrado) - Curso de Ciências da Computação, Universidade Federal de Santa Catarina, Florianópolis, 2011.
- RIJSBERGEN, C. J. Van. **Information retrieval**. 2nd ed. Newton: Butterworth-Heinemann, 1979.
- ROBERTSON, S. E.; JONES, K. Spärck. Relevance weighting of search terms. **Journal Of The American Society For Information Science**, Silver Spring, v. 27, n. 3, p.129-146, 1976.
- ROSSELLÓ-BUSQUET, Ana et al. OWL ontologies and SWRL rules applied to energy management. In: UKSIM 13TH International Conference on Modeling and Simulation, 13. 2011, Cambridge. **Proceedings...** Lyngby: UKSIM, 2011. p. 446 - 450.
- SADOUN, Driss et al. An ontology for the conceptualization of an intelligent environment and its operation. In: 10TH Mexican International Conference on Artificial Intelligence (MICAI), 10. 2011, Puebla. **Proceedings...** Puebla: MICAI, 2011. p. 16 - 22.
- SALTON, G.; WONG, A.; YANG, S. A vector space model for automatic indexing. **Communications Of The ACM**, New York, v. 18, n. 11, p.613-620, Nov. 1975.
- SALTON, Gerard. The SMART System: Retrieval results and future plans. In: **SCIENTIFIC REPORT**, 11, 1966, Ithaca. **Information storage and retrieval**. Ithaca: Cornell University, 1966. p. 1 - 9.
- SALTON, Gerard; BUCKLEY, Christopher. Term-weighting approaches in automatic text retrieval. **Information Processing & Management**, Ithaca, v. 24, n. 5, p.513-523, Jan. 1988.
- SANDERSON, Mark; CROFT, W. Bruce. The history of information retrieval research. **Proceedings Of The IEEE**, [S. L.], v. 100, n. 13, p.1444-1451, 13 maio 2012.
- SATO, Nobuyoshi; UEHARA, Minoru; SAKAI, Yoshifumi. FTF-IDF scoring for fresh information retrieval. In: International Conference On Advanced Information Networking And Applications, 18. 2004, Fukuoka. **Proceedings...** Fukuoka: IEEE, 2004. p. 165 - 170.
- SILVA, Adriano Rívollí da. **Aprimorando a visualização e composição de regras SWRL na Web**. 2012. 129 f. Dissertação (Mestrado) - Curso de Ciências de Computação e Matemática Computacional, Universidade de São Paulo, São Carlos, 2012.
- SILVA, Daniela Lucas da; SOUZA, Renato Rocha; ALMEIDA, Maurício Barcellos. Ontologias e vocabulários controlados: comparação de metodologias para construção. **Ci. Inf.**, Brasília, v. 37, n. 3, Dec 2008.
- SINGHAL, Amit. Modern information retrieval: a brief overview. **Bulletin Of The IEEE Computer Society Technical Committee On Data Engineering**, Madison, v. 24, n. 4, p.35-42, 2001.
- SOUZA, Renato Rocha. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspect. ciênc. inf.**, Belo Horizonte, v. 11, n. 2, Ago. 2006.

SOUZA, Renato Rocha; ALVARENGA, Lídia. A Web Semântica e suas contribuições para a ciência da informação. **Ci. Inf.**, Brasília, v. 33, n. 1, Apr. 2004.

SRIDHARAN, Bhavani; TRETIAKOV, Alexei; KINSHUK. Application of ontology to knowledge management in web based learning. In: IEEE International Conference on Advanced Learning Technologies (ICALT'04), 4. 2004, Joensuu. **Proceedings...** Joensuu: IEEE, 2004. p. 663 - 665.

STAAB, Steffen; STUDER, Rudi. **Handbook on ontologies**. 2nd ed. Berlin; Springer, c2009. xix, 811 p. ISBN 9783540709992.

STEIL, Andrea Valéria. **Um modelo de aprendizagem organizacional baseado na ampliação de competências desenvolvidas em programas de capacitação**. 2002. 218p. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2002.

STUDER, Rudi; BENJAMINS, V. Richard; FENSEL, Dieter. Knowledge engineering: principles and methods. **Data & Knowledge Engineering**, [S. L.], v. 25, n. 1-2, p.161-197, 1998.

SURE, York et al. OntoEdit: collaborative ontology development for the semantic web. In: (ISWC'02) First International Semantic Web Conference, 1. 2002, Sardinia, Italy. **Proceedings...** Sardinia, Italy: Springer, 2002. p. 221 - 235.

SWARTOUT, Bill et al. Toward distributed use of large-scale ontologies. In: Proceedings of The 10th Banff Knowledge Acquisition For Knowledge-Based Systems Workshop, 10. 1997, Providence. **Proceedings...** Marina Del Rey: IEEE, 1997. p. 138 - 148.

TANANTONG, Tanatorn; NANTAJEEWARAWAT, Ekawit; THIEMJARUS, Surapa. Towards continuous electrocardiogram monitoring based on rules and ontologies. In: 11th IEEE International Conference On Bioinformatics And Bioengineering (BIBE), 11. 2011, Taichung. **Proceedings...** Pathumthani: BIBE, 2011. p. 327 - 330.

TÜMER, Duygu; SHAH, Mohammad Ahmed; BITIRIM, Yıltan. An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, Yahoo, Msn and Hakkia. In: ICIMP'09 Fourth International Conference on Internet Monitoring and Protection, 4, 2009, Venice. **Proceedings...** Venice: IEEE, 2009. p. 51 - 55.

USCHOLD, Mike; KING, Martin. Towards a methodology for building ontologies. In: Workshop On Basic Ontological Issues in Knowledge Sharing, Held in Conjunction with IJCAI-95, 14. 1995, Montreal. **Proceedings...** Edinburgh: University Of Edinburgh, 1995. p. 1 - 13.

VESIN, Boban et al. Rule-based reasoning for altering pattern navigation in programming tutoring system. In: 15th International Conference On System Theory, Control, And Computing (ICSTCC), 15. 2011, Sinaia. **Proceedings...** Novi Sad: ICSTCC, 2011. p. 1 - 6.

WAGNER, Gerd; GIURCA, Adrian; LUKICHEV, Sergey. A usable interchange format for rich syntax rules integrating OCL, RuleML and SWRL. In: Reasoning On The Web, 2006, Edinburgh. **Proceedings...** Edinburgh: ROW2006, 2006. p. 1 - 8.

WANG, Feng; Merialdo, Bernard. Weighting informativeness of bag-of-visual-words by kernel optimization for video concept detection. In: VLS-MCMR '10 Proceedings of the International Workshop on Very-Large-Scale Multimedia Corpus, Mining And Retrieval, 9. 2010, Firenze. **Proceedings...** Firenze: ACM, 2010. p. 55 - 58.

ZHANG, Tong-Zhen; SHEN, Rui-Ming. Learning objects automatic semantic annotation by learner relevance feedback. In: International Conference On Biomedical Engineering And Informatics, 2. 2009, Tianjin. **Proceedings...** Tianjin: IEEE, 2009. p. 1 - 4.

ZOBEL, Justin; MOFFAT, Alistair. Inverted files for text search engines. **ACM Computing Surveys**, New York, v. 38, n. 2, p.1-56, 2006. Article No.: 6.

ZOU, Guobing et al. An ontology-based methodology for semantic expansion search. In: Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 5. 2008, Shandong. **Proceedings...** Shandong: IEEE, 2008. p. 453 - 457.

ZÚÑIGA, Gloria L. Ontology: its transformation from philosophy to information systems. In: Proceedings of the International Conference on Formal Ontology in Information Systems, 2. 2001, Maine. **Proceedings...** Maine: FOIS, 2001. p. 187 - 197.