UNIVERSIDADE FEDERAL DE SANTA CATARINA

TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO

RONALDO TADEU MURGUERO JUNIOR

UM SISTEMA DE MANUTENÇÃO SEMIAUTOMÁTICA DE ONTOLOGIAS A PARTIR DO RECONHECIMENTO DE ENTIDADES

Araranguá, 22 de fevereiro de 2013

RONALDO TADEU MURGUERO JUNIOR

UM SISTEMA DE MANUTENÇÃO SEMIAUTOMÁTICA DE ONTOLOGIAS A PARTIR DO RECONHECIMENTO DE ENTIDADES

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

Ronaldo Tadeu Murguero Junior

UM SISTEMA DE MANUTENÇÃO SEMIAUTOMÁTICA DE ONTOLOGIAS A PARTIR DO RECONHECIMENTO DE ENTIDADES

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 este trabalho a todos que contribuíram direta ou indiretamente em minha formação acadêmica.

AGRADECIMENTOS

Agradeço a todos que direta ou indiretamente contribuíram no decorrer desta caminhada, especialmente:

A Deus, a quem devo minha vida.

A minha família que sempre me apoiou nas escolhas tomadas. Em especial meu Pai, Ronaldo minha mãe, Sandra minha irmã, Juliana e meu Cunhado, Ricardo Cavalheiro.

A Flávia Maria por sempre me incentivar, compreender e apoiar nos momentos difíceis.

Ao orientador Prof. Alexandre Leopoldo Gonçalves que teve papel fundamental na elaboração deste trabalho.

Ao Prof. Flávio Ceci por ter colaborado neste trabalho com a cessão da biblioteca de reconhecimento de entidades.

Aos meus colegas pelo companheirismo e disponibilidade para me auxiliar em vários momentos.

RESUMO

Uma quantidade cada vez maior de informações está disponível em formato textual e eletrônico. Essas informações contêm padrões textuais, tais como, conceitos, relacionamentos, regras, entre outros, podendo ser de grande auxílio na integração com outros sistemas ou mesmo, para auxiliar processos de tomada de decisão. Contudo, existe uma grande preocupação em como recuperar, organizar, armazenar e compartilhar estes padrões considerando uma formalização adequada. Neste sentido, a área de Extração de Informação promove suporte através de técnicas que analisam o texto e extraem padrões tidos como relevantes. Após a fase de extração, torna-se necessária a correta atribuição dos padrões para classes de um domínio em particular, em que estes passam a se chamar entidades. Tal processo é realizado através da subárea chamada de Reconhecimento de Entidades. Além disso, visando o compartilhamento e a manutenção de determinado domínio de conhecimento, as entidades devem ser armazenadas em um meio que possibilite atingir tais objetivos. Neste contexto a área de Ontologia se insere. Para demonstrar a viabilidade da proposição deste trabalho foi desenvolvido um protótipo voltado às fases de extração e reconhecimento de entidades, bem como, a adição dessas entidades em uma ontologia para posterior manutenção. O processo de manutenção envolve a participação de um especialista de domínio responsável por validar os conceitos e modificar estes para as suas devidas classes quando necessário. Sendo assim, a manutenção pode ser entendida como semiautomática. De modo geral, a aplicação do protótipo em alguns cenários permitiu demonstrar que o sistema proposto é capaz de obter resultados satisfatórios, ainda que iniciais, mesmo que não exista conhecimento prévio de determinado domínio.

Palavras-chave: Extração de Informação; Reconhecimento de Entidades Nomeadas; Ontologia; Manutenção de Ontologia.

ABSTRACT

An increasing amount of information is available in textual and electronic format. This information has textual patterns, such as concepts, relationships, rules, among others. It can be valuable whether integrated with other systems or even to support decision making processes. However, there is great concern about how to retrieve, organize, store and share these patterns considering a suitable formalization. In this sense, the Information Extraction area promotes support through techniques that analyze the text and extract patterns regarded as relevant. After extraction phase it becomes necessary the correct assignment of patterns to classes in a particular domain. Thus, these patterns are called entities. This process is accomplished through the Named Entity Recognition area. Additionally, aiming sharing and maintenance of a specific knowledge domain, entities should be stored in a way that allows achieve these goals. In this context the Ontology area stands. To demonstrate the feasibility of the proposed work we have developed a prototype toward pattern extraction and entity recognition phases, as well as the addition of these entities into ontology for subsequent analyses. The maintenance process involves the participation of a domain expert which is responsible for the concepts validation, as well as by moving these entities to the properly classes when needed. Thus, maintenance can be understood as semiautomatic. In general, the application of the prototype in some scenarios demonstrated that the proposed system, although in an initial stage, is able to obtain satisfactory results even without prior knowledge of a particular domain.

Keywords: Information Extraction; Named Entity Recognition; Ontology; Ontology Maintenance.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACE – *Automatic Content Extraction*

DAML - *DARPA agent markup language*

EI – Extração de Informação

HTML – *HyperText Markup Language*

MUC – *Message Understanding Conference*

NER – *Named Entity Recognition*

NLP – *Natural Language Processing*

OIL - *Ontology Inference Layer* ou *Ontology Interchange Language*

OWL – *Web Ontology Language*

OWL – *Web Ontology Language*

RDF – *Resource Description Framework*

SWRL – *Semantic Web Rule Language*

UFSC – Universidade Federal de Santa Catarina

URI – *Uniform Resource Identifiers*

WEB – *World Wide Web*

W3C – *World Wide Web Consortium*

XML – *eXtensible Markup Language*

SUMÁRIO

1. INTRODUÇÃO

Uma quantidade cada vez maior dos dados existentes estão disponíveis na forma textual e eletrônica. Esses arquivos textuais contêm informações e, possivelmente padrões, que se analisados corretamente podem auxiliar no processo de tomada de decisão.

Estes arquivos que estão disponíveis em meio eletrônico encontra-se principalmente na World Wide Web, ou simplesmente WEB. A WEB muitas vezes confunde-se com a Internet, esta criada durante a Guerra Fria para possibilitar a troca de informação. Porém, a WEB pode ser vista como um ambiente com interfaces mais amigáveis e intuitivas para o acesso ao crescente repositório de documentos, possuindo um diferencial em relação à Internet que são os hipertextos (SOUZA; ALVARENGA, 2004). Segundo Branski (2004), hipertextos funcionam como pontes que ligam dois pontos, na informática isto é chamado de link. Links tem a capacidade de ligar palavras ou frases de uma pagina web a outros recursos da internet, fazendo com que o usuário explore a WEB.

Desde a sua criação a WEB tem crescido, segundo alguns autores, a taxas exponenciais (LYMAN; VARIAN, 2003; HIMMA, 2007). Segundo o estudo de Lyman e Varian (2003) a WEB era constituída por cerca de 167 *terabytes* de páginas estáticas. Contudo considerando, páginas geradas em tempo real a partir dos dados em base de dados, e-mails e mensagens instantâneas este valor atingia mais de 532 *terabytes*.

No periodo de 1986 a 2007 segundo Hilbert (2011) a taxa de computação cresceu a uma porcentagem anual de 58%. Estes dados começaram a mudar não sendo somente dos modos citados anteriormente estando agora também no novo paradigma da WEB, chamado de WEB 2.0. Este novo paradigma foi citado pela primeira vez por O'Reilly e MediaLive International onde foi citado que a Web 2.0 era composta agora por ferramentas que possibilitavam aos próprios usuário produzirem conteúdo na WEB (O'REILLY, 2007). Nesse sentido, redes sociais, wikis e blogs se tornaram meios amplamente utilizados para popular a WEB.

De modo geral a informação divide-se em três tipos, sendo, Estruturada, quando o texto segue determinados padrões, possui uma estrutura definida, estes padrões são formatos de escrita que a EI conhece facilitando a sua consulta, Semiestruturada, quando os textos possuem certo padrão, mas não é todo escrito desta maneira e Não estruturado, também chamado de texto livre, são os textos lidos diariamente, seja ele em um site de noticias em blogs ou ate mesmo na Wikipédia®, ou seja, não seguem um padrão de escrita pré-definido (ÁLVAREZ, 2007).

Considerando o volume de informação e a sua natureza, tarefas voltadas ao armazenamento, gerenciamento e disponibilização se tornam desafios principalmente para organizações que desejem obter alguma vantagem competitiva. Como afirma Ceci (2010), as informações em sua maioria encontram-se dispersas sem uma adequada formalização e estrutura, necessitando de processamento adequado para que seja possível extrair padrões, regras, e tendências que possam prover alguma utilidade no processo de tomada de decisão.

Uma das ferramentas utilizadas para enfrentar estes desafios é a Extração de Informação (EI) que se preocupa com a extração de elementos relevantes em uma coleção de documentos (GRISHMAN, 1997; MUSLEA, 1999; ZAMBENEDETTI, 2002). Uns dos componentes chaves da extração de informação são os seus padrões de extração ou regras de extração.

Dentro da Extração de Informação, encontra-se o Reconhecimento de Entidades Nomeadas, do inglês, *Named Entity Recognition* (NER). NER pode ser definida como um mecanismo que faz a associação entre termos, por exemplo, nomes próprios, a determinadas classes, sendo então chamadas de entidades (CUNNINGHAM, 2002; GROVER et al., 2002). Uma entidade pode representar tanto objetos do mundo fisico quanto abstrato, tais como, ”Albert Eisnten” sendo classificado como pessoa ou ”Araranguá” que seria classificada como cidade. Como afirma Gonçalves (2006), uma entidade quando extraída do texto pode ser vista como um vetor *E* com uma descrição, uma classe e informações adicionais, ou seja, E={descrição,classe,<informações adicionais>}. As informações adicionais podem ser as posições onde a entidade ocorre no texto.

Após o processo de extração de informação e sua categorização, torna-se necessário algum formalismo visando explicitar o conhecimento de um determinado domínio de problema. Para Schreiber (2002), conhecimento representa informações e dados que auxiliam a pessoa a tomar uma decisão, a realizar tarefas e a criação de novas informações e conhecimento, sempre sofrendo modificações ao longo do tempo. Está além deste trabalho a

discussão epistemológica sobre o conceito conhecimento no que tange a possibilidade deste ser explicitado em algum meio ou não. Assume-se assim, como afirma Krogh e Roos (1995), que o "conhecimento é uma entidade universalmente representável que pode ser armazenada em computadores, bases de dados, arquivos e manuais". Contudo, torna-se necessário a representação utilizando alguma estrutura formal.

Existem vários mecanismos que permitem representar formalmente determinado domínio do conhecimento, entre eles, redes semântica, frames e ontologias (RUSSELL; NORVIG, 1995). A partir da representação torna-se possível sua posterior recuperação e utilização de tal modo que promova o desenvolvimento de sistemas computacionais. Neste trabalho tem-se como foco a representação do conhecimento através de Ontologias.

Segundo Gruber (1995), ontologia se constitui em uma "especificação explicita de uma conceitualização". Em 1998 Studer et al. (1998), complementa a definição de Gruber afirmando que uma ontologia é uma "especificação formal e explícita de uma conceitualização compartilhada". Para representar uma ontologia torna-se necessário algum formalismo visando à concepção de sistemas computacionais. Neste sentido, Maedche et al. (2003) propôs o formalismo de 5-tuplas, sendo composto por conceitos, relacionamentos, hierarquia de relacionamentos e um conjunto de axiomas.

Segundo Almeida (2003) as ontologias podem funcionar sobre fonte de dados proporcionando uma melhor organização e uma recuperação de informações mais eficiente, ainda desempenhando um papel importante da comunicação entre os agentes, usuários, e os dados envolvidos.

A combinação das áreas de Extração de Informação e Ontologia permite facilitar o processo de manter uma estrutura formal de conhecimento atualizada. Este processo é visto como semiautomático, pois possui uma etapa automática em que entidades são extraídas de textos e armazenadas na ontologia através de algoritmos, mas que, na etapa seguinte depende de um especialista de domínio para refinar e manter a ontologia atualizada, por exemplo, através da correção de classificações inadequadas, criação de novas classes, propriedades e axiomas.

1.1 PROBLEMÁTICA

Com o aumento do uso da WEB cada vez mais as informações estão disponíveis através de paginas estáticas ou dinâmicas, wiki e blogs. Além disso, muita informação encontra-se nas redes corporativas. Tais informações estão em sua grande maioria em modo

textual, ou seja, sem estrutura, gerando desafios em como realizar a extração de conteúdo relevante para posterior utilização, de modo que se possa auxiliar em processos de tomada de decisão.

Para lidar com tal demanda tem-se utilizado técnicas de extração de informação, mas especificamente a NER, que promove o reconhecimento e nomeação de padrões encontrados em textos. Por reconhecimento entende-se a extração em si do padrão e por nomeação a atribuição deste padrão para uma determinada classe. Vale mencionar que de modo geral, para que a extração de informação obtenha sucesso é necessário um conjunto de textos para que determinado padrão tenha relevância estatística, caso contrário, este não será identificado.

Após a fase de reconhecimento é comum o armazenamento destas entidades em estruturas formais para a representação do conhecimento de determinado domínio. Entre os meios de se formalizar determinado domínio encontram-se as ontologias que se constituem em uma conceitualização explicita e formal de um determinado domínio que se deseja representar.

Esses elementos criam suporte que promovem a manutenção de ontologias que pode ser considerada como semiautomática, pois a intervenção de um especialista sobre o domínio em determinado ponto do processo é fundamental para garantir a qualidade e a evolução do conhecimento que uma ontologia representa.

Outro ponto importante a considerar é a dinamicidade do conhecimento, ou seja, o conhecimento de determinado domínio tende a evoluir mais rapidamente do que a capacidade de representar este em uma estrutura formal. Neste sentido, o desenvolvimento de sistemas que auxiliem na etapa de manutenção de ontologia extraindo as principais entidades se mostra relevante. Por outro lado, constitui-se em desafio uma vez que identificar tais entidades não é tida como uma tarefa trivial considerando os aspectos de cada língua e a necessidade de constituição de alguma base de conhecimento inicial que promova suporte ao processo de extração de entidades.

Desse modo tem-se como pergunta de pesquisa “Como criar um sistema que possibilite a extração de entidades de determinado domínio a partir de fontes de informação textual, bem como, a representação destas entidades em uma estrutura formal de conhecimento?”.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo geral do trabalho consiste na proposição de um sistema que possibilite a extração de entidades e manutenção semiautomática do conhecimento de determinado domínio de problema.

1.2.2 Objetivos Específicos

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

- Analisar as técnicas de extração de informação e nomeação de entidades, bem como, as formas atuais de representação do conhecimento obtido pelo processo de extração;
- Propor os modelos lógicos e físicos, que permitam suportar o desenvolvimento de um sistema de extração de entidade e manutenção de ontologias;
- Desenvolver um protótipo voltado à manutenção semiautomática de ontologia que permita demonstrar a viabilidade do sistema proposto neste trabalho;
- Realizar uma discussão dos resultados obtidos através da utilização do protótipo de manutenção semiautomática de ontologia.

1.3 METODOLOGIA

O trabalho será desenvolvido com base em uma pesquisa aplicada materializada através da implementação de um protótipo de extração de entidades e manutenção semiautomática de ontologia. A metodologia de desenvolvimento deste trabalho é dividida como segue:

- Revisão da literatura científica sobre Extração de Informação e Reconhecimento de Entidades, bem como, sobre Ontologias;
- Proposição de uma visão lógica e física do sistema que guia a implementação do mesmo;
- Desenvolvimento de um sistema (protótipo) voltado à extração de entidades e manutenção semiautomática de ontologias;
- Detalhamento das funcionalidades e discussão de um cenário de aplicação;

- Análise dos resultados obtidos através da utilização do protótipo;
- Apresentação das considerações finais do trabalho, assim como potenciais pontos de aprimoramento futuro.

1.4 ORGANIZAÇÃO DO TEXTO

O documento esta dividido em 6 capítulos. No primeiro capitulo apresenta-se o projeto, expondo uma breve contextualização e apresentando a problemática vislumbrada, assim como, os objetivos geral e específicos.

No segundo capítulo é realizada uma revisão sobre a área de Extração de Informação promovendo um maior detalhamento desta, abordando conceitos reconhecimento de entidades nomeadas, identificação de padrões, analisadores de discurso e integração e preenchimento de valores.

O terceiro capítulo é realizado uma contextualização sobre área de Ontologia, dando um detalhamento maior do processo de construção de ontologia, utilizando como exemplo o método 101. Também apresenta as principais regras utilizadas para a construção de ontologia.

O quarto capítulo detalha o sistema visando promover suporte para o seu desenvolvimento. Divide-se assim em duas etapas em que a primeira apresenta a visão lógica, que promove um entendimento geral do sistema e a segunda apresenta a visão física, que apresenta e detalha a interconexão dos componentes tecnológicos do sistema.

O quinto capítulo apresenta um cenário de aplicação do sistema e discute os resultados obtidos através da utilização do protótipo neste cenário. Por fim, o sexto capítulo apresenta as considerações finais e detalha propostas de trabalhos futuros.

2. EXTRAÇÃO DE INFORMAÇÃO

2.1 INTRODUÇÃO

Como citado anteriormente o aumento de informação pelos mais diversos meios de comunicação e formatos, principalmente na forma de texto traz desafios em como extrair conteúdo relevante a partir de informações em linguagem natural. Entre o ferramental disponível para lidar com tais desafios cita-se a área de Extração de Informação.

A Extração de Informação (EI) faz parte do Processamento da Linguagem Natural (*Natural Language Processing* - NLP), e tem como foco identificar padrões em bases textuais. Liddy (2001), NLP pode ser entendida como um conjunto de técnicas computacionais voltadas à representação e análise de textos que ocorrem em um ou mais níveis de análise linguística, com o propósito de auxiliar aos usuários em uma série de tarefas e aplicações.

De maneira geral, pode-se afirmar que a função primária da EI consiste em isolar os fragmentos dos textos importantes e agrupá-los em alguma estrutura permitindo futuras recuperações (ZAMBENEDETTI, 2002). Por exemplo, em um texto que trata do assunto "extração de informação", seria adequado encontrar outros assuntos relacionados, tais como, reconhecimento de entidade e manutenção de ontologia, descartando aquilo que não é relevante a determinado domínio.

2.2 ARQUITETURA DE EXTRAÇÃO DE INFORMAÇÃO

A arquitetura de Extração de Informação esta dividida em duas partes. Na primeira etapa são extraídos fatos individuais, por exemplo, o nome de uma proteína ou de uma pessoa, Em seguida é realizada a integração dos fatos visando produzir fatos novos. Por fim, os fatos são passados para um formato de saída requerido (ZAMBENEDETTI, 2002). Por integração, pode-se entender a análise do contexto através de coocorrência dos fatos, da extração de relacionamentos que possibilitam a inferência de novos fatos.

Em 1997 foi apresentada uma estrutura de EI simplificada a partir do modelo proposto na sexta edição da MUC (*Message Understanding Conference*) (GRISHMAN, 1997). A estrutura proposta possui dois estágios. O primeiro estágio é responsável por algumas etapas, entre elas, a análise léxica do texto (cada sentença é analisada visando identificar suas partes essenciais, por exemplo, nome, verbo, artigo, preposição) e o reconhecimento de nomes (cada termo é analisado e é adicionado a uma classe, por exemplo, nome, organização, etc.), Além disso, esta estrutura utiliza um analisador sintático/semântico para verificar o contexto em que o termo está inserido visando identificar relacionamentos entre os nomes, por exemplo, "trabalha com". De posse de nomes e relacionamentos é realizada a análise de padrões com o objetivo de identificar eventos relevantes para determinado domínio de análise, por exemplo, o padrão "Pessoa_1 trabalha com Pessoa_2" poderia ser analisado e preenchido visando um melhor entendimento de determinado domínio.

O segundo estágio objetiva realizar a análise dos termos no contexto do texto (análise de discurso) passando então para o preenchimento de *templates* que podem ser integrados a sistemas computacionais, por exemplo. A Figura 1 apresenta a arquitetura.

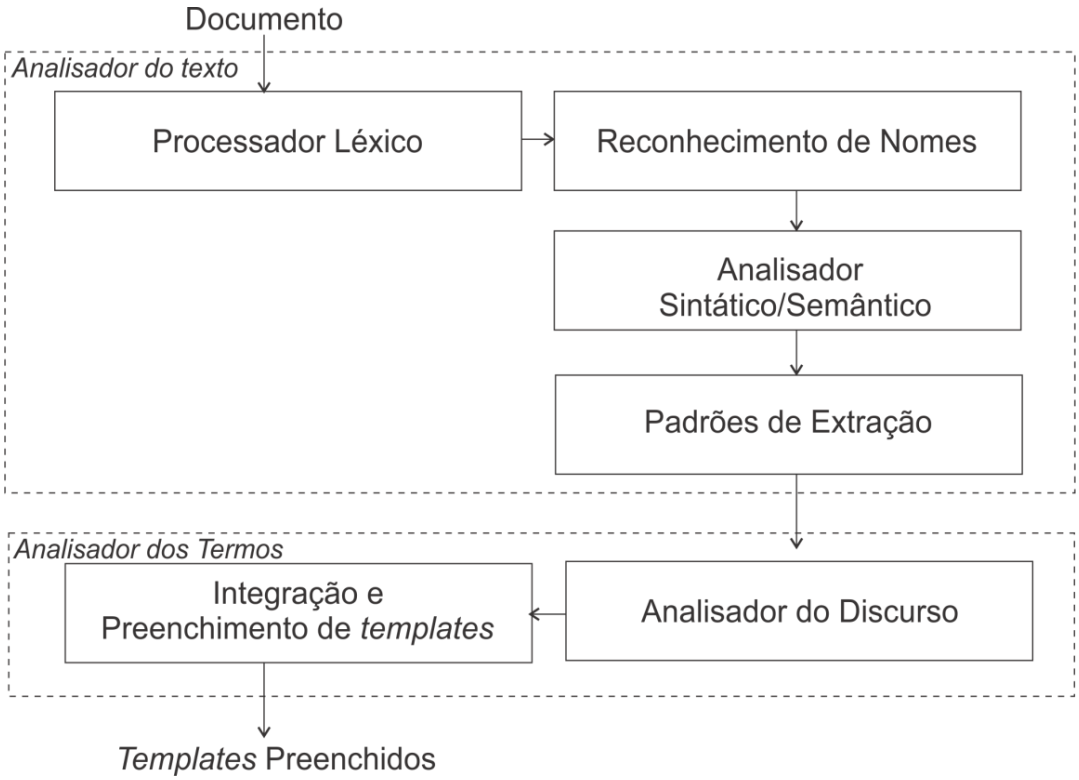


Figura 1. Estrutura de um sistema de extração informação

Fonte: Adaptada de Grishman (1997)

Cada um dos seis módulos de extração de informação será detalhado nos tópicos subsequentes, pois são à base do modelo proposto neste trabalho.

2.2.1 Processador Léxico

De modo geral o processador léxico separa o texto em termos visando ter um melhor controle sobre o texto, pois todos os termos serão analisados e procurados em um dicionário a fim de determinar suas características (ZAMBENEDETTI, 2002).

No contexto da Extração de Informação esses dicionários se referem a um conjunto de termos que auxiliam na identificação ou rótulo de determinado termo, também chamados de tabelas léxicas, *lexicons* ou *gazetters*. A Tabela 1 apresenta exemplos de valores para a formação de três tabelas léxicas, sendo uma para designar **Lugares**, outra para designar **Instituições de Ensino** e outra para designar **Pessoas**.

Lugar
Araranguá
Florianópolis

Instituição de Ensino
UFSC
Unisul

Pessoas
Alexandre
Ronaldo

Tabela 1 – Exemplos de tabelas léxicas

2.2.2 Reconhecimento de Entidades

Após toda a estruturação das ligações e referências é necessário extrair termos relevantes e atribuir estes para suas classes. A técnica mais comum utilizada para isto é o reconhecimento de entidades nomeadas (NER).

A NER é fundamental na área de EI, pois é responsável pela identificação e classificação de entidades em textos. Esse tópico foi inicialmente tratado na sexta edição do MUC como uma subtarefa dentro da conferência (GRISHMAN; SUNDHEIM, 1996). Outras ideias foram sendo incorporadas a MUC tendo crescido em importância e abrangência. Mais recentemente vem sendo incorporado o conceito de extração automática de conteúdo (ACE - *Automatic Content Extraction*), que tem como foco, além da extração de entidades, a extração de relacionamentos e eventos (NIST, 2008). Isto pode ser encontrado em trabalhos de autores como Santos (2011) e Ceci et al. (2012).

Após a categorização de padrões textuais para uma determinada classe, por exemplo, Pessoa, Organização, Cidade, Tempo, Data e Expressões numéricas, entre outras, os padrões agora nomeados para entidades são armazenados, de maneira geral, em uma base de dados.

2.2.2.1 Identificação de Padrões (*Collocation*)

Collocation segundo Manning e Schütze (1999) "é uma expressão que consiste em duas ou mais palavras que correspondem a uma forma comum de dizer as coisas", ou seja, formam uma maneira natural de concatenação de palavras. Segundo Firth (1957), *collocation* pode ser definida como "declarações dos lugares habituais ou costumeiros da palavra". São as frases costumeiras da população, padrões sutis, derivando também segundo Stubbs (1996) de obras de historia, literária e culturais sendo uns dos meios do formalismo do vocabulário local. Para Choueka (1988) *collocation* pode ser definida como "uma sequência de duas ou mais palavras consevutivas, que tem características de uma unidade sintática e semântica".

De modo geral, através do uso de funções estatísticas, determinada sequência de palavras que formam uma c*ollocation* pode ser identificada. Manning e Schütze (1999) ressalta algumas formas de reconhecimento de *collocation*, entre elas:

a. Frequência

Pode ser considerada a forma mais simples de se estabelecer uma *collocation*, pois conta a quantidade de vezes que duas palavras quaisquer coocorrem ao longo do texto. A utilização da frequência conjunta de duas palavras, ainda que imprecisa, demonstra algum tipo de evidência de relacionamento.

b. Média e Variância

A frequência conjunta revela indícios para a constituição de estruturas frasais sejam estas formadas por palavras ocorrendo na sequência, ou a certa distância. Neste caso, quando duas palavras se encontram distantes o uso da média e variância pode promover uma forma mais acurada de se estabelecer uma *collocation*, pois se utiliza do conceito de janela entre as palavras.

Primeiramente, é necessário calcular a média dos deslocamentos entre as palavras, sendo dada pela equação $\overline{d} = sn$, onde ***s*** representa a soma das distancias e ***n*** o numero de vezes que a palavra aparece no texto. Por outro lado, a variância calcula o grau de desvio das distância conforme a seguinte equação $S^2 = \frac{\sum_{i=1}^{n}(d_i - \overline{d})^2}{n-1}$ onde, ***n*** é o numero de vezes que as duas palavras ocorrem (coocorrência), ***di*** é a distancia da *i*th coocorrência, e $\overline{\boldsymbol{d}}$ é a

média das distâncias. Quando as distâncias forem sempre iguais a variância é zero, do contrário, quando não apresentam um padrão de relacionamento a variância será alta.

c. Teste de Hipótese

Apesar de muitas palavras ocorrerem com elevada frequência e baixa variância estas não podem ser utilizadas como *collocation*, pois criam ocorrências ao acaso e não um padrão relevante. No teste de hipótese é atribuída uma hipótese nula H_0, ou seja, não existe associação entre duas palavras *A* e *B*. A partir disto deve ser calculado um valor de significância de modo que se possa rejeitar ou aceitar a hipótese H_0. Manning e Schütze (1999) apresentam alguns exemplos de cálculos baseados em teste de hipótese, entre eles: a) teste *t*: analisa a média e variância de um termo de onde é retirada a sua hipótese nula. b) teste de hipótese das diferenças: quando se tem duas palavras distintas, mas que denotam o mesmo significado. c) teste qui-quadrado ($\boldsymbol{x^2}$): analisa a independência do termo através de uma comparação entre as frequências observadas e as esperadas; e d) razão de probabilidade: similar ao teste qui-quadrado, contudo, ao invés de se basear em uma hipótese nula é atribuída uma hipótese alternativa, de modo que não sejam analisadas somente as palavras que coocorrem frequentemente.

d. Informação Mútua

Esta estratégia possui sua motivação na teoria da informação e objetiva medir o nível de associação entre palavras utilizando como base a coocorrência. A Informação Mútua analisa a probabilidade de um par de palavras aparecerem frequentemente de maneira conjunta ao invés de aparecer isoladamente. De modo geral, quando existir um relacionamento entre as palavras este terá um valor maior do que zero.

2.2.2.2 Nomeação/Classificação de Entidades

A nomeação de entidades possui varias definições e formatos diferentes. Segundo Marrero et al. (2012), pode-se analisar e classificar termos segundos alguns critérios como categoria gramatical, designação rígida, identificação única e domínio de aplicação.

- Categoria Gramatical (Nomes próprios): Esta categoria se utiliza dos critérios de nomeação e estrutura gramatical para reconhecer uma entidade como relativa à sua categorização. Podendo ser designada para seres e realidades únicas, como nomes comuns e substantivos próprios. Para realizar esta categorização a NER

analisa alguns pontos como a falta de inflexibilidade nos substantivos e em caso de nomes próprios deve-se iniciar com letra maiúscula.

- Designação Rígida: Define unicamente determinado conceito em um determinado domínio de análise, por exemplo, **Petrobrás®**. Neste caso, **Petrobrás®** refere-se a uma empresa (classe organização) e considerando a sua especificidade dificilmente existirá outra entidade com o mesmo nome. Por outro lado, o termo **Jaguar** não possui uma designação rígida, uma vez que este por se referir a um carro/empresa, avião ou animal. Mesmo o nome de uma pessoa pode não possuir uma designação rígida visto que podem existir homônimos.

- Identificação Única: Diferentemente da designação rígida o termo aqui deve possuir uma exemplificação única e neste caso o tratamento da ambiguidade torna-se necessário. Como mencionado anteriormente, o termo jaguar pode ser atribuído para mais de uma classe.

- Domínio de Aplicação: Esta classificação está ligada a termos que não apresentem alguma especificação clara ou que se muito abrangente, como é o caso do conceito de tempo. Marrero et al. (2012) afirma que para uma classificação mais adequada pode se utilizar as cinco perguntas do jornalismo: o que, onde, quando, quem e por que.

Para exemplificar o resultado do reconhecimento de entidades a Figura 2 apresenta uma matéria do site de noticias Contato.net[1].

UFSC Araranguá

Jovens aprendem noções de programação em projeto

Fonte: Contato Internet / Cristina Possamai (SC04023JP)

Araranguá – O projeto da Universidade Federal de Santa Catarina (UFSC), em parceira com a Escola de Educação Básica de Araranguá surgiu da necessidade de envolver instituições públicas com a comunidade acadêmica. O intuito é mostrar aos adolescentes como se pode utilizar o computador como uma ferramenta no processo aprendizagem e incentivar a interação entre os estudantes do ensino médio e os acadêmicos da Universidade Federal.

A equipe de instrutores do projeto foi composta pelos acadêmicos Iury Melo e Mauricio Justo Izé do curso de Engenharia e Computação e os universitários Samuel Ghisleri Minatto e Sabrina Pitz Lima, de Tecnologias da Informação e Comunicação. Além das professoras da UFSC, Eliane Pozzebon e Silvia Helena Mangili.

Figura 2 - Exemplo de texto com a identificação de entidades reconhecidas

[1] http://noticias.contato.net/?acao=noticia&noticia=079502

Na Figura 2 acima os termos destacados em azul claro correspondem a lugares, os em verde correspondem a instituições de ensino, e os em vermelho a pessoas. Kozareva (2006) afirma que a NER, delimita o inicio e o fim de palavras sendo de grande auxílio na identificação de entidades compostas por várias palavras como "Universidade Federal de Santa Catarina". Em alguns casos determinado termo não é reconhecido e este pode então ser atribuído para uma classe geral para posterior avaliação ou descarte por um especialista de domínio.

Outra situação comum que acontece com um termo ao ser classificado é a que ele pode possuir mais de um significado, gerando desse modo ambiguidade. Por exemplo, a palavra **Java** pode ser tanto uma ilha quanto uma linguagem de programação de computadores. Para estes casos Kozareva (2006) sugere a utilização de listas de termos (*lexicons*). Listas de termos podem ser identificadas como um dicionário dividido em classes (cada classe possui um *lexicon* associado) onde são encontradas palavras relevantes em um texto, diferente de repositórios de preposições e determinações (como a, ante, segundo) que são fixas (MANNING; SCHÜTZE, 1999). A utilização dessas listas promove um desafio adicional, uma vez que um termo em particular pode fazer parte de mais de uma classe. Nesse sentido, o processo de reconhecimento de entidade requer algoritmos específicos para lidar com a ambiguidade. Para Stevenson e Wilks (2003) existem três algoritmos básicos para a eliminação de ambiguidade:

- Preferências Semânticas: Utiliza-se de uma seleção hierárquica organizada, acerca do texto que esta sendo analisado. Esta hierarquia possui a palavra e seu significado, contendo também adjetivos e características acerca do substantivo a fim de facilitar a interpretação para a eliminação da ambiguidade.

- Análise de Especialista: Os termos são agrupados em uma estrutura conforme alguma regra considerando o pressuposto que o ser humano é condicionado a classificar este termo por conhecimento sobre a palavra e não sobre a regra.

- Contexto da Palavra: Utiliza-se de um mecanismo de análise do texto mais amplo para verificar o contexto em que o termo esta inserido, ou seja, analisa os demais termos que aparecem antes e depois para chegar a uma conclusão. Por exemplo, se o termo **Jaguar** estivesse circundado por outros termos formando a sentença "Jaguar XF® esportivo" o sistema poderia atribuir o termo a classe **Carro**. Para tal, é necessário que cada classe possua um

conjunto de palavras que permitam estabelecer o contexto que vão além das constantes na tabela léxica.

2.2.3 Analisador Sintático/Semântico

O analisador sintático/semântico tem como função organizar o texto para as fases subsequentes. Grishman (1997) ressalta que os analisadores são de grande importância, pois muitas vezes os textos correspondem a frases nominais e possuem estruturas de relações gramaticais. O autor ressalta ainda que os resultados obtidos com esse analisador não são sempre satisfatórios, pois a uma grande variação na quantidade de estrutura sintática.

2.2.4 Padrões de Extração

Uma das características mais importantes em um processo de extração de informação são os padrões de extração, ou seja, as regras que serão utilizadas para realizar a extração que vão além da identificação de *collocations*. Neste caso, um sistema de EI tem que aprender com os exemplos fornecidos pelos usuários e tentar extrair o conteúdo. Conforme apresentado por Muslea (1999) a formalização dos padrões ou regras de extração se baseiam em restrições sintáticas e semânticas que ajudam a identificar as informações relevantes dentro de um texto. Várias abordagens são descritas pelo autor, entre elas, AutoSlog, LIEP, PALKA, CRYSTAL, entre outras.

2.2.5 Analisador de Discurso

Para Álvarez (2007) nesta etapa é realizada o relacionamento entre os diferentes termos do texto, considerando o relacionamento entre as sentenças e incluindo o tratamento de algumas tarefas como:

- Análises de frases que se referem a apostos e outros grupos nominais;
- Resolução de correferência, quando um pronome já citado se refere a outro já citado anteriormente;
- Descoberta de relacionamento entre as partes dos textos que traz suporte para estrutura de extração.

A Tabela 2 exemplifica de maneira simplificada o analisador de discurso. A entidade A1 possui um tipo “Local” e uma descrição “Araranguá”, assim como a entidade A2 com tipo “Instituição de Ensino” e descrição “UFSC”. O termo que se encontra entre estas duas

entidades tem o evento A3, que neste exemplo é do tipo “Possui” e conecta uma entidade de tipo “Local” com outra de tipo “Instituição de Pesquisa”. Sendo assim será feita a associação de que “Araranguá possui UFSC”.

Entidade A1	Tipo: Local, Descrição: Araranguá
Entidade A2	Tipo: Instituição de Ensino, Descrição: UFSC
Evento A3	Tipo: Possui, Local: A1, Instituição de Ensino: A2

Tabela 2 - Exemplo de Analisador de Discurso
Fonte: Adaptada de Grishman (1997)

2.2.6 Integração e Preenchimento

Na ultima etapa após os dados estarem formatados, os *templates* são preenchidos com as informações relevantes e apresentados ao usuário para que ele possa tornar estas informações válidas ou não. O preenchimento destes *templates* é realizado através das descobertas de termos e conexões entre estes que são realizadas durante todo o processo de Extração de Informação (GRISHMAN, 1997).

2.3 TIPOS DE DADOS DA EXTRAÇÃO DE INFORMAÇÃO

Como foi citado anteriormente o texto ao qual é realizada a extração de informação nunca possui uma estrutura bem definida. Segundo Álvarez (2007), para se tentar amenizar este fator existem técnicas que, conforme a estrutura do texto deve ser selecionada para se obter melhores resultados. Os tipos de textos são: Estruturado, Semiestruturado e Não estruturado. A seguir é dada uma exemplificação de cada tipo de texto:

Estruturado: Para um texto ser considerado estruturado este deve possuir alguma regularidade no seu formato de apresentação. Esta estrutura regular permite que facilmente um sistema de EI identifique cada elemento de interesse com base em regras uniformes (ÁLVAREZ, 2007).

Semiestruturado: Como no Estruturado ele também possui alguma regularidade no seu formato de apresentação, porém possui algumas irregularidades como, campos com nulos ou até mesmo ausentes e variações na ordem dos dados (SILVA; BARROS, 2005).

Não estruturado: Este tipo de texto não possui formatação envolvendo basicamente sentenças de linguagem natural (SILVA; BARROS, 2005). Álvarez (2007) complementa afirmando que estes são os textos encontrados na Web nos mais diferentes tipos de formato.

3. ONTOLOGIA

3.1 INTRODUÇÃO

A palavra Ontologia vem da junção de dois termos *ontos* (ser) e *logos* (palavra), tendo seu inicio na Grécia antiga com o objetivo de lidar com as questões filosóficas do mundo, ou seja, o estudo da natureza, do ser, da realidade e de questões de metafísica, sendo introduzida nesta área por Aristóteles, grande filósofo grego. Enquanto disciplina da filosofia, ontologia tem como objetivo o fornecimento de sistemas de categorização para organizar a realidade (BREITMAN, 2006).

No início dos anos 90 pesquisadores da área de Inteligência Artificial começaram a introduzir a ontologia na informática (LIMA, 2005). Segundo Gruber (1995), uma ontologia pode ser vista como uma "especificação explicita de uma conceitualização". Conceitualização pode ser entendida como uma visão abstrata simplificada do mundo que se deseja representar, enquanto Explícita representa a definição clara dos elementos que compõem este mundo. A definição de Gruber era considerada por muitos pesquisadores ampla, pois não dava suporte ao desenvolvimento de aplicações, e também por possuir uma alta carga filosófica, não sendo completamente definida para a computação. Borst (1997) exemplifica esta representação como a hierarquia da classificação animal onde, animais que possuem hábitos semelhantes podem pertencer à mesma classe. Neste sentido, toda base de conhecimento compromete-se com alguma conceitualização, explicita ou não. Com este pensamento Borst (1997) complementa a definição de Gruber dizendo que uma ontologia é uma "especificação formal de uma conceitualização compartilhada". O conceito de compartilhada reflete a necessidade de a ontologia representar o conhecimento consensual.

Em 1998, Studer et al. (1998) afirma que toda ontologia também deve ser explícita, indicando que os conceitos e as restrições de um domínio são explicitamente definidos. Sendo assim, a definição passou a ser "uma especificação formal e explícita de uma conceitualização compartilhada".

Neste sentido, Maedche (2003), propôs o formalismo de 5-tuplas. A estrutura conceitual da 5-tupla é O:= {C, R, H^C, rel, A^o} composta de:

- Dois disjuntos C e R cujos elementos são conceitos e relacionamentos, respectivamente;
- Uma hierarquia de conceitos H^c x H^c, em um relacionamento direto, chamado de hierarquia de conceitos ou taxonomia, onde A1 é um subconceito de A2, H^c (A1,A2);
- Uma função, rel: R = C x C, que relaciona os conceitos de modo não taxonômico;
- Os axiomas, A^o que ligam estes conceitos expressos em sua linguagem própria.

A proposta de Maedche é amplamente utilizada, pois ela permite a representação na maioria das linguagens atuais voltadas à representação de ontologias. Independente da estruturação escolhida é interessante analisar que a ontologia tem sido utilizada para descrever artefatos com variados graus de estruturação e diferentes propósitos (BREITMAN, 2006).

Uma ontologia é composta de alguns conceitos básicos, sendo:

a) Classes: conjuntos abstratos de coisas ou coleções (por exemplo, uma classe que representa pessoas);

b) Indivíduos: objetos em uma ontologia que possuem pelo menos um nome e um valor;

c) Relacionamento (Propriedade): ligação entre os objetos na ontologia (HORRIDGE, 2011).

A Figura 3 apresenta uma estrutura de ontologia sobre a composição de uma pizza com cobertura e base e o relacionamento entre tais componentes. Neste exemplo a classe Pizza possui relacionamentos com as classes Base e Cobertura.

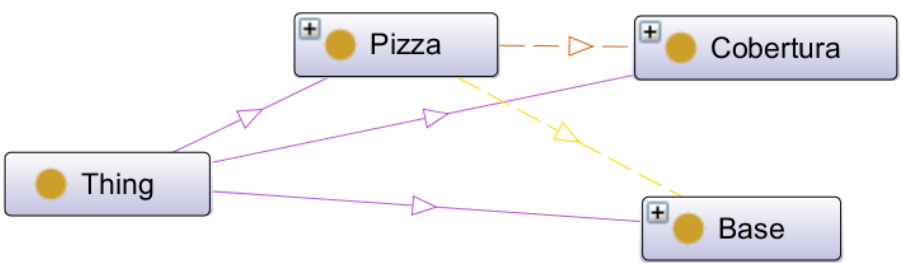


Figura 3 - Exemplo de Estrutura Básica de uma Ontologia
Fonte: Adaptada de Horridge (2011)

O projeto de modelagem de uma ontologia está fortemente relacionado a determinado domínio, bem como, ao conhecimento que o especialista possui deste domínio. Neste sentido, especialistas de um mesmo domínio podem produzir ontologias diferentes. Levando isto em consideração Guarino (1998) propôs uma estrutura (Figura 4) que possibilita a definição de diferentes níveis de uma ontologia facilitando a representação de determinado domínio.

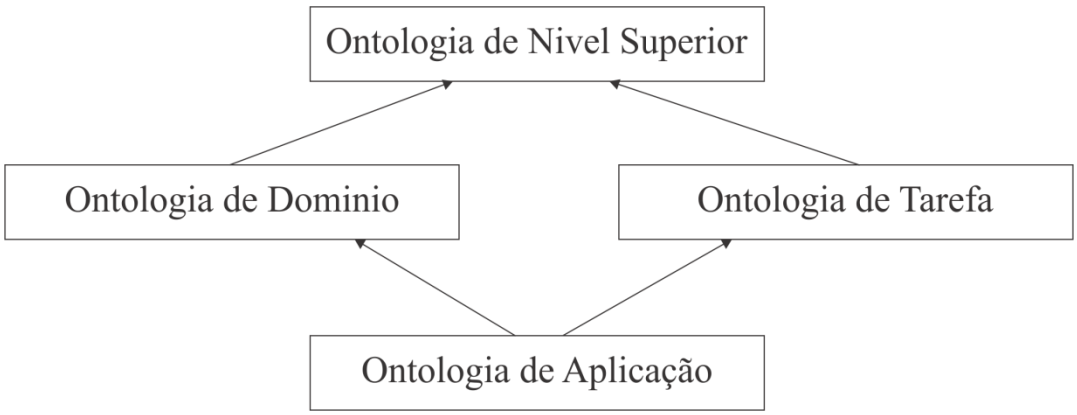


Figura 4 - Tipos de ontologias de acordo com o seu nível de dependência
Fonte: Adaptada de Guarino (1998)

- Ontologias de nível superior: descrevem conceitos genéricos, tais como, espaço, tempo e eventos. A princípio estas são independentes de domínio e podem ser utilizadas no desenvolvimento de novas ontologias;
- Ontologias de domínio: descrevem o vocabulário relativo a um domínio específico através da especialização de conceitos presentes na ontologia de nível superior;
- Ontologias de tarefas: descrevem o vocabulário relacionado a uma tarefa ou atividade através da especialização de conceitos presentes na ontologia de nível superior;
- Ontologias de aplicação: é o nível mais específico, pois correspondem a papéis desempenhados por entidades do domínio no desenrolar de alguma tarefa.

Uma ontologia pode definir um vocabulário comum para pesquisadores que desejam compartilhar informações acerca de um domínio, incluindo máquinas que interpretem conceitos acerca de um determinado domínio e as relações entre eles (NOY;McGUINNESS,

2001). Segundo Noy e McGuinness (2001), existem algumas razões pela qual uma ontologia é necessária, entre elas:

- Para compartilhar entendimento comum da estrutura de informação entre pessoas e agentes de s*oftware*;
- Para permitir a reutilização de um domínio do conhecimento;
- Para tornar explícitas as suposições de um determinado domínio;
- Para separar o conhecimento de domínio do conhecimento operacional;
- Para analisar o conhecimento do domínio.

Visando promover um entendimento maior sobre o assunto torna-se necessário o aprofundamento das metodologias de desenvolvimento de ontologia, bem como, as linguagens necessárias para representar computacionalmente tal estrutura.

3.1.1 Metodologia

Uma metodologia constitui-se dos passos que devem ser tomados para um determinado processo, sendo também uma forma de conduzir a pesquisa ou um conjunto de regras, ou seja, é a explicação minuciosa, detalhada, rigorosa e exata de toda ação desenvolvida no método da pesquisa.

Na área de Ontologia existem algumas metodologias, tais como, Methontology (FERNANDEZ; GOMEZ-PEREZ; JURISTO, 1997), NeOn (GÓMEZ-PEREZ; SUÁREZ-FIGUEROA, 2008), ontoKEM (RAUTENBERG; GOMES FILHO, et at., 2010) e 101 (NOY; MCGUINNESS, 2001).

Como saliente Breitman (2006) não existe uma única maneira de se construir o domínio de uma ontologia, sempre existem várias alternativas, pois é um processo não linear, envolve muitas iterações e refinamentos são necessários para se chegar ao modelo adequado.

Para exemplificar o processo de desenvolvimento uma ontologia neste trabalho será detalhada a metodologia 101 proposta por Noy e McGuinness (2001). Esta metodologia utiliza-se de sete passos, que são:

a) Determinar o domínio e escopo da ontologia;
b) Considerar o reuso de outras ontologias;
c) Enumerar os termos importantes da ontologia;
d) Definir classes e hierarquia de classes;

e) Definir propriedades de classes;

f) Definir os valores de propriedades;

g) Criar instâncias.

3.1.1.1 Determinar o domínio e o escopo da ontologia

Para iniciar o domínio e o escopo da ontologia Noy e Mcguinness (2001) sugerem que sejam respondidas algumas questões, entre elas:

- Qual domínio a ontologia irá cobrir?
- Para qual propósito se utilizará a ontologia?
- Para quais informações a ontologia deve fornecer respostas?
- Quem vai utilizar e manter a ontologia?

As respostas a estas perguntas podem ser modificadas a todo o momento durante o processo de construção da ontologia, mas são de grande importância, pois em determinado momento elas auxiliam a limitar o alcance do modelo (NOY; MCGUINNESS 2001).

Uma maneira também utilizada para se determinar o escopo da ontologia é através de perguntas, ditas como questões de competência. Estas perguntas elaboradas servirão como teste de eficiência da ontologia. Para realizar este teste são levados em conta alguns aspectos, tais como: a) A ontologia contêm informações suficientes para responder a esses tipos de perguntas?; e b) Será que as respostas exigem um determinado nível de detalhe ou de representação de uma determinada área? Estas questoes não precisam ser exaustivas são apenas um esboço de competencia.

Considerando uma ontologia de pizza que representa a elaboração de uma pizza com cobertura e base, teriamos as seguintes questões de competencia:

- Quais temperos combinados ficam melhor para o paladar?
- Pizza quatro queijo harmoniza com vinho tinto?
- Combinação entre ingredientes como coração e calabresa resulta em uma boa pizza?

3.1.1.2 Considerar o reuso de outras ontologias

Existem muitas ontologias disponíveis em formato eletrônico que podem ser importadas para o ambiente de desenvolvimento. De modo geral, o formalismo em que estas ontologias são expressas não é algo tão relevante, uma vez que muitos sistemas de representação de conhecimento podem importar e exportar (NOY; MCGUINNESS 2001).

Reutilizar uma ontologia quase sempre vale a pena sendo que se pode aperfeiçoar e expandir as fontes existentes para um domínio em particular. Atualmente existem algumas bibliotecas que disponibilizam modelos de ontologia, por exemplo, Ontolingua[2], DAML[3] e o Dublin Core[4].

3.1.1.3 Enumerar os termos importantes da ontologia

Quando existe a necessidade de definir ou explicar para o usuário é interessante criar uma lista de termos. Neste sentido se justifica responder algumas questões que auxiliem na formulação da lista, tais como: Quais são os termos que se gostaria de incluir? e Quais são as propriedades desses termos?

Nesta etapa da construção da lista de termos não deve existir a preocupação com redundâncias ou detalhamento de seus relacionamentos. Este tipo de relacionamento será tratado nos próximos passos.

Analisando o domínio de um sistema capaz de manter registros de currículos os seguintes termos poderiam ser considerados: área, produção, formação, titulação, ano, título, periódicos, congressos, endereço, autoria, competidor, santa catariana.

3.1.1.4 Definir classes e a hierarquia de classes

Classes podem ser definidas como um conjunto que contem indivíduos e estes indivíduos possuem requisitos que o classificam em determinada classe. Por exemplo, em uma classe **Felino** todos os animais que fazem parte da família dos felinos possuem requisitos para fazer parte da classe felino, como gato, leão e tigre (HORRIDGE, 2011).

Para se definir uma hierarquia de classes existem algumas abordagens, como exemplo tem-se:

- Topo-para-base (*top-down*): Esta classificação começa com a definição dos termos mais abrangentes (pai ou superclasse) e posteriormente os termos mais específicos (filhos ou subclasse).

[2] http://www.ksl.stanford.edu/software/ontolingua/

[3] http://www.daml.org/ontologies/

[4] http://dublincore.org/

- Base-para-topo (*bottom up*): Nesta definição é realizado o inverso da *top-down*, ou seja, primeiramente são criados os termos específicos e depois estes são agrupados em termos mais abrangentes.
- Combinação: Como o nome já sugere esta é a combinação entre as duas abordagens citadas acima a *top-down* e *bottom up*. Nesta abordagem, podem-se definir algumas classes abrangentes e outras mais especificas fazendo a relação entre elas.

A utilização destas abordagens depende muito das características e do ponto de vista dos especialistas de domínio. Se este tem uma visão mais sistemática do domínio a abordagem *top-down* provavelmente será utilizada.

Para organizar as classes em uma hierarquia taxonômica utilizam-se questionamentos sobre a organização dessa estrutura, por exemplo, "*Se uma classe A é superclasse de uma B, então toda instancia de B também é instancia de A".* Isto tende a guiar a formulação correta de uma estrutura taxonômica.

Como exemplo apresenta-se abaixo (Figura 5) uma hierarquia de classes para o domínio que mapeia a elaboração de uma pizza:

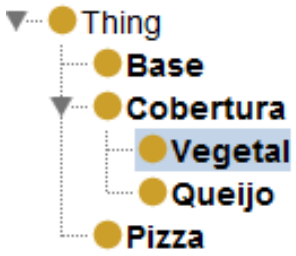


Figura 5 - Hierarquia de classes de um domínio

3.1.1.5 Definir as propriedades das classes

A classe em si não provê informações suficientes para responder a determinada pergunta de competência. Neste sentido, logo após a definição das classes deve-se definir a estrutura interna destas.

A partir da lista de termos definida no passo anterior esta deve ser analisada para verificar quais desses termos representam propriedades de classes. Pensando um uma ontologia de pizza um exemplo de propriedade para a classe Base seria o "tipo de massa". Em OWL este tipo de propriedade é chamada de propriedade de dado.

Na linguagem OWL uma das características mais importantes é a definição das propriedades de objetos que conectam duas classes através dos conceitos de domínio (domain) e escopo (range). Neste sentido, o domínio indica que a propriedade se aplica a determinada classe, enquanto o escopo indica que os valores de uma propriedade em particular são instâncias de uma determinada classe.

3.1.1.6 Definir os valores das propriedades

Esta etapa depende muito da linguagem que esta sendo utilizada, pois cada uma estabelece um valor diferente para a propriedade, como exemplo tem-se cardinalidade que em alguns sistemas elas assumem somente um valor enquanto outras permitem cardinalidade múltiplas.

Na linguagem OWL é permitido utilizar tipos de dados nos preenchimentos de valores de propriedades, como cadeia de caracteres, número, valores booleanos e listas enumeradas de elementos (BREITMAN, 2006).

3.1.1.7 Criar Instâncias

O último passo requer a criação de instâncias e o preenchimento de seus valores, para definir esta instância individual é necessário:

a) escolher uma classe;

b) a criação de uma instância individual desta classe;

c) o preenchimento dos valores.

A Figura 6 mostra a criação da instancia na classe pizza.

Figura 6 – Criação de instancia.

3.1.2 Linguagens

Para representar uma ontologia pode-se utilizar de vários métodos como textos, tabelas e gráficos (RABELO, 2012). Porém, a utilização destes métodos não é aconselhável,

pois cada um poderia vir em um modelo diferente e isto dificultaria a interpretação. Para tal, é aconselhável quando se trata de comunicação entre sistemas a utilização de linguagens que sejam formais.

No contexto de Ontologia algumas linguagens foram desenvolvidas baseadas em XML (*eXtensible Markup Language)* entre elas tem-se o RDF (*Resource Description Framework*), a OWL (*Web Ontology Language) e* a SWRL (*Semantic Web Rule Language*) normatizadas pela W3C (*World Wide Web Consortium*). Conforme afirmavam Souza e Alvarenza (2004), RDF e OWL seriam utilizadas pela maioria das ontologias no futuro. Nas seções a seguir estas três linguagens serão descritas considerando suas características e funcionalidades.

3.1.2.1 RDF

Segundo o W3C (2012b) o RDF é uma estrutura para representar informações na Web. A descrição dos dados e dos metadados em RDF é realizada através de um esquema de triplas composto pelo recurso, pela propriedade e pelo valor (SOUZA; ALVARENGA, 2004). A Figura 7 apresenta uma tripla RDF composta por sujeito, predicado e objeto, O sujeito representa a referência para determinado recurso através de um identificador URI (*Uniform Resource Identifiers)*, o predicado denota o relacionamento entre o sujeito e o objeto, e o sujeito pode ser tanto um recurso (identificado por um URI) quanto um literal (que descreve o valor do predicado).

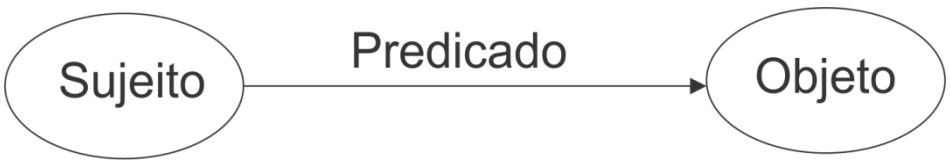


Figura 7 - Tripla RDF (*Resource Description Framework*)
Fonte: Adaptada de W3C

Os identificadores URI (*Uniform Resource Identifiers*) são sequência de caracteres que não contem qualquer caractere de controle como exemplos têm *http://www.example.org/staffid*. (W3C, 2012b).

Este padrão RDF, permite que qualquer indivíduo e ou organização publique na Web de forma que agentes ou softwares posam interpretar e agir sobre esta informação de forma mais inteligente (SOUZA; ALVARENGA, 2004). Souza e Alvarenga ainda afirmam que RDF fornece alguns benefícios como:

- Disponibiliza um ambiente consistente para publicação e utilizar dados da Web.
- Disponibiliza uma sintaxe padronizada.
- Permite que aplicações possam tomar decisões inteligentes e automatizadas sobre suas informações

A estrutura básica de um RDF para representar determinado recurso é composta por sujeito, predicado e objeto como dito anteriormente e é apresentada na Figura 8.

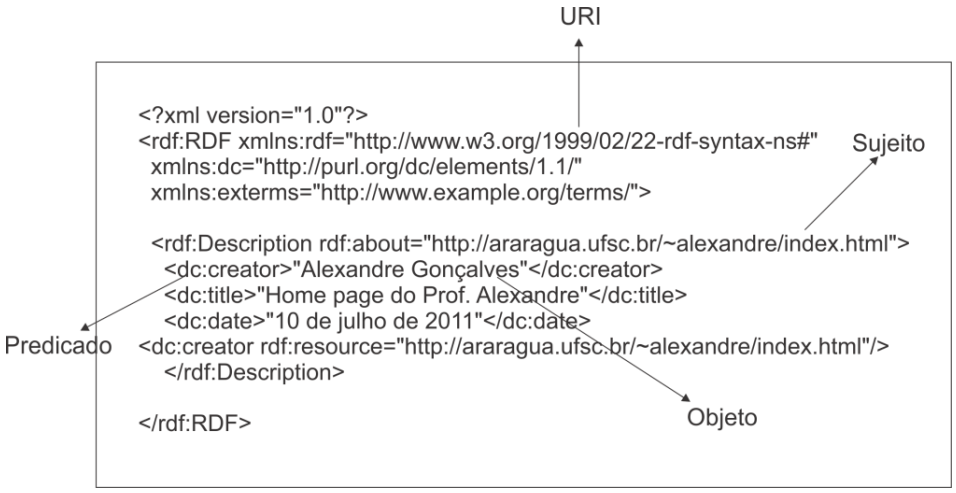


Figura 8 - Estrutura de um recurso em RDF

3.1.2.2 OWL

A OWL (*Web Ontology Language*) é uma derivação da DAML+OIL (*DARPA agent markup language + Ontology Inference Layer* ou *Ontology Interchange Language*), que eram duas linguagens de formalismo de ontologias. De modo geral, é projetada com o objetivo de facilitar a representação semântica de recursos na Web (W3C, 2012a). Lima (2005) e Chen et al. (2012), citam as sub-linguagens de OWL, sendo:

- OWL-Lite: Possui restrições e é indicada para usuários que utilizam uma hierarquia simples.
- OWL-DL: Indicada para usuários que queiram o máximo de expressividade, completude (todas as conclusões são computáveis) e decidibilidade (todas as conclusões terminaram em um tempo computacional finito).
- OWL-Full: É usada por usuários que queiram o máximo de independência de RDF e expressividade, sem garantia formal.

A OWL possui alguns elementos básicos sendo eles:

a) Classe: utilizadas para descrever o conceito mais básico do domínio;

b) Indivíduos: membro de determinada classe;

c) Propriedades: relacionamentos binários que servem para descrever fatos em geral a respeito de determinado domínio;

d) Restrições: delimitam os indivíduos que pertencem àquela classe (BREITMAN, 2006).

3.1.2.3 SWRL

A proposta da utilização da linguagem SWRL (*Semantic Web Rule Language*) é a de utilizar regras para acessar os dados desejados. Possui como base a combinação da linguagem DL e OWL-Lite e para a execução das regras tem como base a sublinguagem RuleML (*Rule Markup Language*) (W3C, 2012c; CHEN, 212). Como afirma Chi (2009) SWRL é baseado em regras semânticas.

Uma regra em SWRL possui a forma de uma implicação entre um antecedente (corpo) e o consequente (cabeça). Isto significa que sempre que as condições especificadas no antecedente forem verdadeiras, então as condições especificadas no consequente também devem ser mantidas verdadeiras (W3C, 2012c).

Para se escrever a regra de SWRL utilizam-se átomos, que são conjunções estabelecidas. Estes átomos podem ter as seguintes formas: *C*(*x*), *P*(*x,y*), *sameAs*(*x,y*), *differentFrom*(*x,y*), em que *C* descreve uma classe, *P* uma propriedade, *x* e *y* são indivíduos ou valores de dados em OWL que podem ser ou não atribuídos a variáveis para futuras conclusões, *sameAs* indica que duas instâncias quaisquer são as mesmas e *differentFrom* indica que duas instâncias quaisquer são diferentes (W3C, 2012c).

Através uma regra SWRL é possível, por exemplo, determinar o IMC (Índice de Massa Corporal) de uma pessoa atribuindo-a para uma subclasse (IMC_I, IMC_III ou IMC_III) dependendo de suas propriedades de dado (peso e altura). A Figura 9 apresenta uma regra em SWRL que possibilita o cálculo necessário para determinar o IMC de uma pessoa.

```
Pessoa(?x), Altura(?x, ?altura), Massa(?x, ?massa),
divide(?imc, ?massa, ?dob),
greaterThanOrEqual(?imc, 25),
lessThan(?imc, 30),
multiply(?dob, ?altura, ?altura) -> IMC_I(?x)
```

Figura 9 - Exemplo de uma regra SWRL para cálculo do IMC

3.1.3 Manutenção de Ontologias

Uma característica importante das ontologias é que estas evoluem com o tempo e isto se deve principalmente a natureza dinâmica de determinado domínio do conhecimento. Conforme afirma Yu (2006), ontologias evoluem inevitavelmente, seja por causa de melhorias necessárias ou porque o domínio que a ontologia representa mudou e a representação deste não reflete mais a realidade. Ainda neste sentido, nota-se que a manutenção de uma ontologia é um processo contínuo, pois o conhecimento não é estático motivo que leva as ontologias a sofrerem constantes atualizações (CECI, 2010).

Segundo Rafi et al. (2009), a manutenção de ontologia é bastante ampla, pois engloba vários aspectos, tais como: a combinação, fusão, integração, alinhamento, mapeamento, articulação e a tradução de diferentes conceitos a fim de promover adaptações ou evoluções no domínio de interesse. A manutenção de ontologia considerando o aspecto estrutural, inclusão ou alteração de classes/subclasses e propriedades é geralmente realizada manualmente. Por outro lado, a extração de instância que possibilitem a população da ontologia e em estudos mais recentes, a extração de relacionamentos, pode ser realizada por algoritmos especializados.

Além do exposto acima, para que manutenções ocorram tornam-se necessárias ferramentas que auxiliem e facilitem o processo de manutenção ontologias. Entre as ferramentas destacam-se o OntoLancs (GACITUA et al. 2009; GACITUA;SAWYER, 2008), OntoLearn (NAVIGLI et al. 2003; MISSIKOFF et al. 2002) e o Protégé® (PARK; KIM; BAE, 2011) que foi o utilizado neste trabalho. Desta forma, o processo de manutenção por ser entendido como semiautomático, em que o conhecimento de especialistas de domínio, algoritmos especializados e ferramentas computacionais promovem suporte à manutenção de ontologias.

No caso do presente trabalho a manutenção é realizada de maneira semi automaticamente, pois se utiliza de um algoritmo que realiza o preenchimento de classes com com as entidades extraídas de texto livro visando facilitar a etapa de manutenção de modo que o usuário se preocupe com a validação destas entidades.

4. SISTEMA PROPOSTO

Este capítulo exemplifica o sistema proposto dividindo-o em duas visões, sendo a primeira a visão lógica que descreve o sistema de maneira mais geral, e a segunda a visão física, que detalha os componentes tecnológicos utilizados no sistema.

4.1 VISÃO LÓGICA

A visão lógica (Figura 10) representa um conjunto de passos que possibilitam a conexão entre o conteúdo textual, possivelmente vindo de várias fontes de informação, até a extração e armazenamento de entidades em uma base de conhecimento.

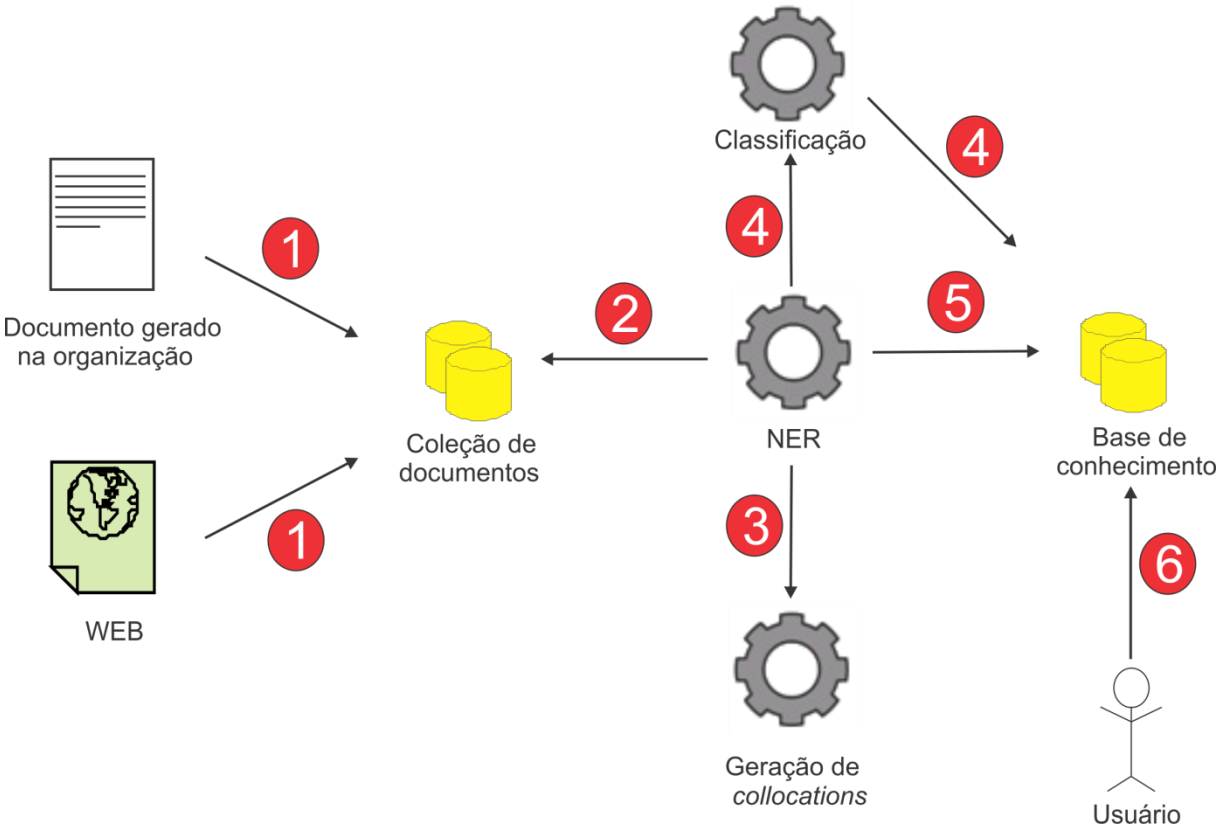


Figura 10 - Visão lógica do sistema proposto

A seguir são detalhados os passos que envolvem a visão lógica da arquitetura proposta:

1) Os textos vindos da Web ou de uma base de dados organizacional são armazenados em um repositório para que se possa iniciar a etapa seguinte responsável pela tarefa de população da ontologia;

2) Como base essencial do processo os documentos contidos no repositório são coletados de modo que os algoritmos de extração e nomeação de padrões possam ser aplicados;

3) Neste passo em geral utiliza-se a estratégia de geração de *collocations*, em que são analisadas as frequências individuais e conjuntas entre dois termos visando determinar se esse par possui relevância estatística. Caso tal relevância seja atingida considera-se como um termo válido passível de ser analisado no próximo passo. O resultado desse processo é sempre um vetor de determinado documento, em que, este vetor é composto por padrões que representam termos.

4) Em seguida ocorre o processo de classificação em que o vetor de determinado documento é recebido para análise individual de cada termo. Neste sentido, o termo é comparado com a lista de palavras de cada classe constante na base de conhecimento para determinar a que classe este pertence, por exemplo, uma Pessoa ou Organização. Com a atribuição dos termos para classes, estes são adicionados ao vetor de retorno que além dos dados vindos da extração de informação possui agora a classe a qual eles pertencem.

5) Após as etapas de extração e classifica o processo de NER realiza a atualização da base de conhecimento. Através da utilização do vetor de retorno este é aplicado na base de conhecimento. A base de conhecimento é uma estrutura que representa determinado domínio de conhecimento através de suas classes, relações e propriedade e instâncias.

6) Como etapa final existe a necessidade de interação de um usuário (especialista de domínio) com o objetivo de verificar e validar o processo automático de extração de entidades e população da ontologia. Por exemplo, o especialista pode determinar se determinada instância está ou não classificada corretamente. Também é possível descartar determinada entidade caso esta tenha sido erroneamente classificada. Estas informações podem então serem utilizadas posteriormente visando melhorar o desempenho do processo de NER.

4.2 VISÂO FÍSICA

A visão física apresenta o detalhamento dos processos e componentes tecnológicos e como eles se interconectam oferecendo uma visão mais tecnológica do sistema proposto. Este detalhamento é apresentado na Figura 11.

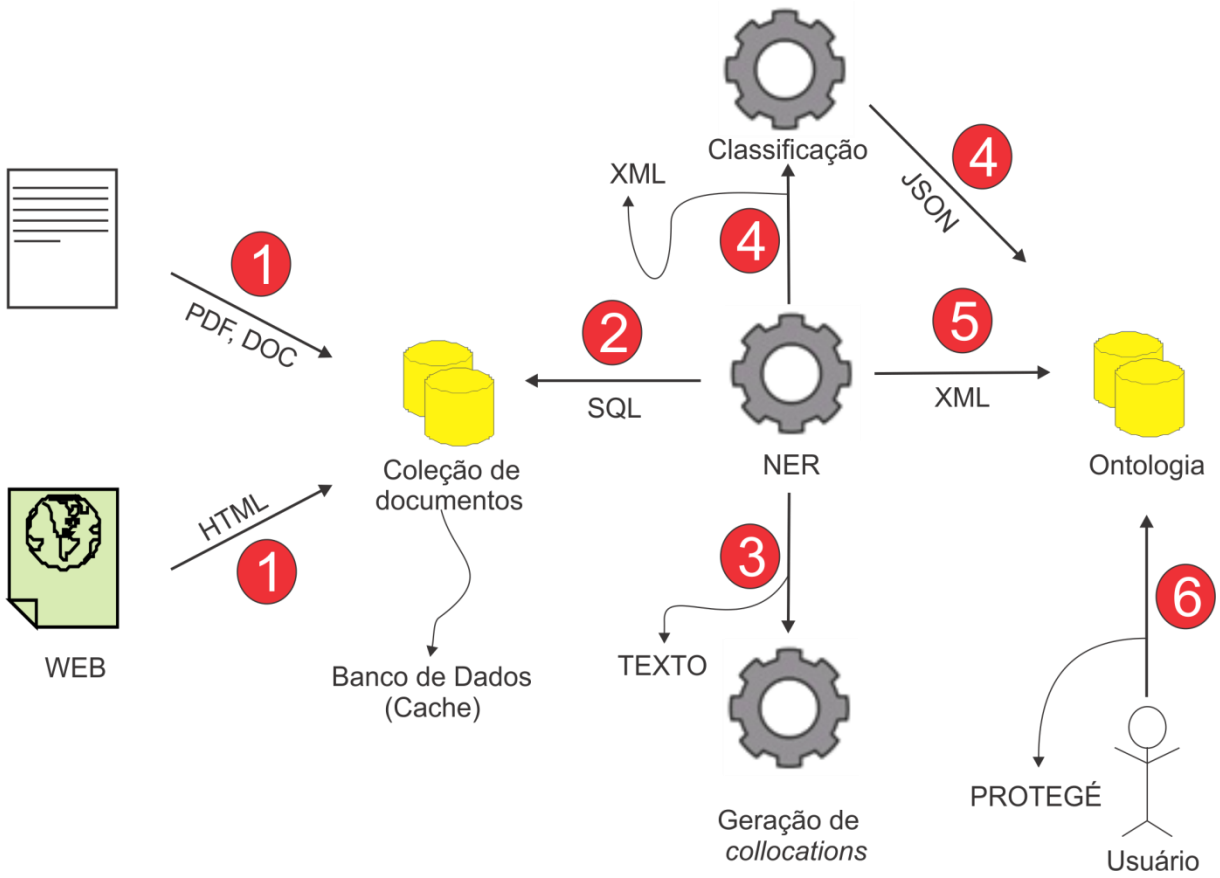


Figura 11 - Visão física do sistema proposto

Em seguida são detalhados os passos que envolvem a visão física do sistema proposto:

1) Os textos tanto de uma organização quanto da Web possuem um formato. Os de organização em sua maioria são de formato PDF e DOC, já os da Web estão no formato HTML. Estes textos podem ser armazenados em um banco de dados ou mesmo um sistema de arquivo qualquer e formam assim um repositório. Na implementação do sistema utilizou-se um banco de dados relacional conforme a estrutura apresentada na Figura 12.

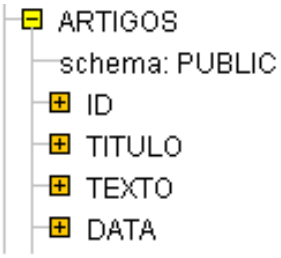


Figura 12 - Estrutura do Banco de Dados da coleção de documento

A estrutura é uma tabela composta por um identificador do artigo (ID), um campo chamado TITULO que armazena o título do artigo, um campo TEXTO que contém o documento completo, e um campo DATA que representa a data de publicação do documento.

2) Nesta etapa o sistema realiza uma consulta no banco de dados utilizando SQL. Esta consulta é realizada de maneira temporal, ou seja, é realizada em ciclos de tempo. Para tal, utiliza-se uma data armazenada no sistema representando a última consulta e que irá ser o valor atribuído à variável na consulta (configurado na cláusula WHERE) que possibilita recuperar somente os documentos que ainda não foram processados. Cada linha resultante da consulta possui um conjunto de campos conforme apresentado na Figura 12.

3) De posse dos documentos o processo de NER inicia visando extrair os padrões mais relevantes constantes em um texto. Em geral utilizam-se estratégias para a identificação de *collocation* conforme descrito no Capítulo 2. Tal abordagem pode ser custosa uma vez que é necessário analisar a coleção para que se consiga estabelecer as frequências conjuntas e desse modo obter relevância estatística.

Visando resolver isto se utilizou um algoritmo mais simples em que são extraídos termos e suas combinações (também chamados de *n*-gramas). A partir disto os termos são adicionados a uma lista que contém ainda a frequência em que o padrão ocorre no texto. Ao final os itens da lista são ordenados de forma decrescente sendo possível estabelecer alguma estratégia de corte para evitar termos pouco relevantes. A simplicidade do método é adequada, pois não exige que se analise todo o conjunto de documentos. Contudo, a precisão do processo pode ser reduzida. Contudo, à medida que o processo de NER é executado, termos que constam na ontologia, pois foram considerados importantes em algum momento, podem ser recuperados de um documento mesmo que este não tenha relevância neste documento. Deste modo, o processo pode ser visto como incremental.

Por fim, após a extração dos termos é gerado um vetor do documento em formato XML que será enviado para o passo seguinte, a classificação.

4) O vetor gerado na etapa anterior em formato XML é a base para o processo de classificação. Para tal, são utilizadas tabelas léxicas em que cada uma possui um conjunto de termos que descrevem determinada classe. Por exemplo, a classe que representa Área do Conhecimento possui termos como “reconhecimento”, “entidade”, “nomeada”, “base”, “dados”, “inteligência” e “artificial”. Além da classe mencionada a implementação do protótipo considera mais três, sendo Pessoa, Organização e Projeto.

As informações de cada tabela constam na ontologia associadas a classe correspondente, ou seja, existe um campo na ontologia vinculado a classe que possui todos os termos que a define. Como apresentada na visão física a comunicação com a ontologia é realizada através do padrão JSON. Contudo, vale mencionar que a obtenção do conteúdo das tabelas léxicas a partir da ontologia não foram implementadas no protótipo, elas foram adicionadas diretamente no código do protótipo.

Após o reconhecimento das entidades estas são armazenados em um vetor que contém os seus itens mais relevantes de um documento. Este vetor, representado em formato XML, possui a estrutura conforme a Figura 13. Cada elemento do XML representa uma entidade composto pelo identificador desta e pela sua classe.

```
<vetor>
        <entidade id=”Ronaldo” classe=”Pessoa”/>
        <entidade id=”UFSC” classe=”Organizacao”>
</vetor>
```

Figura 13 - Estrutura do vetor de entidades em XML

5) O vetor de retorno em formato XML obtido através dos passos 3 e 4 é então adicionado à ontologia.

6): O último passo envolve o refinamento da ontologia a medida que novos documentos chegam a base de dados e novas entidades são recuperadas. Para tal, é necessário que o responsável pela manutenção possua determinado conhecimento sobre o domínio, sendo este chamando de especialista. A manutenção da ontologia pode ser realizada por ferramentas como o Protégé®.

Um exemplo de uma ontologia em que o especialista pode realizar a manutenção é apresentado na Figura 14. Esta ontologia é composta por algumas classes. A classe “Pessoa” representa o(s) autor(es) dos documentos ou outras pessoas encontradas no texto. A classe “Projeto” contém nomes de projetos que foram identificados nos textos. A classe “Area” armazena termos que representam áreas ou especializações de um determinado domínio. A

última classe, “Organizacao”, se refere a nomes de organizações extraídas no processo de NER. As classes na ontologia estão conectadas por propriedades e permitem definir os axiomas que declaram o domínio.

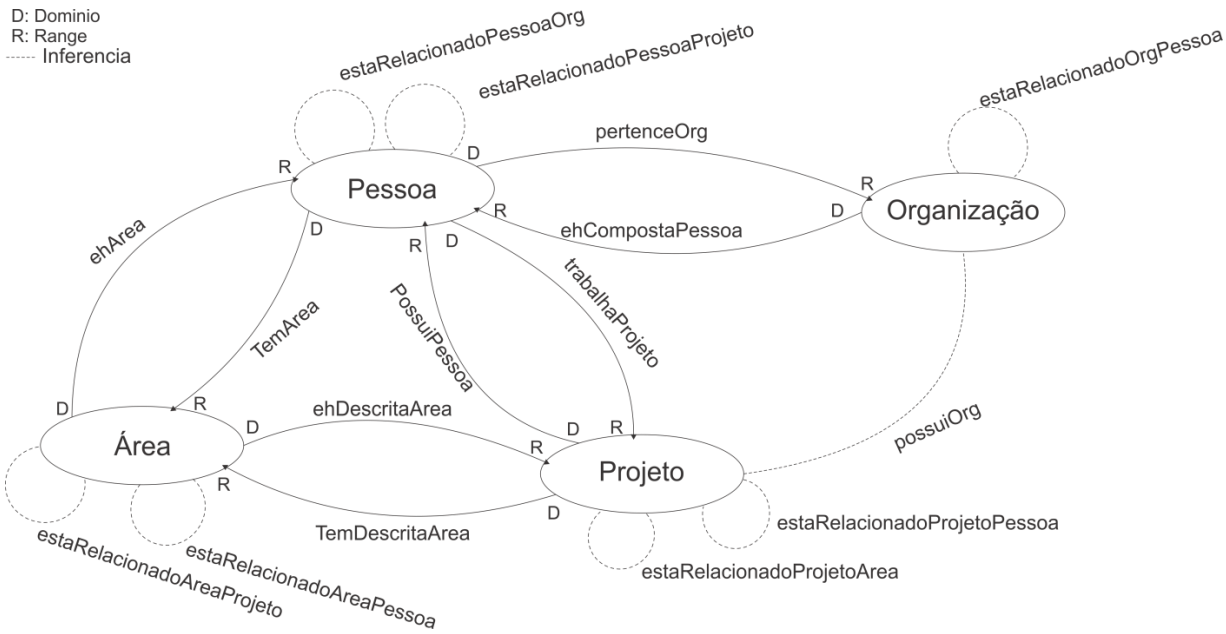


Figura 14 – Modelo da ontologia proposta

De modo geral, o processo de NER e a adição das entidades na ontologia permite apenas a definição de instâncias (indivíduos) para uma determinada classe. A atribuição das propriedades (dados e objetos) deve ser realizada pelo especialista do domínio. Sendo assim, o processo detalhado através das visões lógica e física é considerado semiautomático.

4.2.1 Detalhamento do Protótipo

Para um melhor entendimento do trabalho proposto esta seção detalha o desenvolvimento do protótipo de extração de entidades e manutenção de ontologia. Para descrever de maneira geral o processo de reconhecimento de entidades (NER) é utilizado um diagrama de sequência (Figura 15) com o intuito de apresentar os objetos e a interação entre eles.

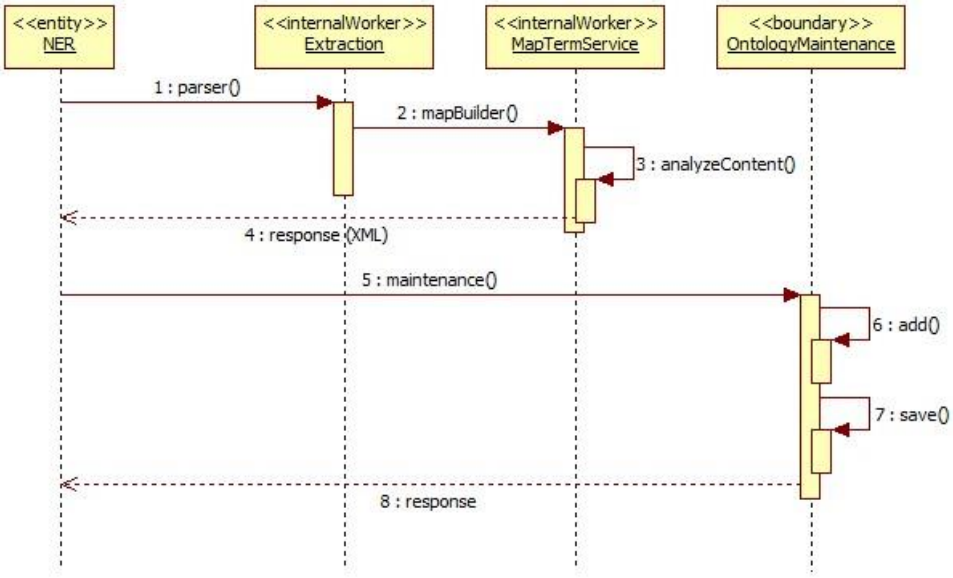


Figura 15 - Diagrama de sequência do processo de NER

O processo de NER inicia através da execução do método *parser*() da Classe *Extraction* que irá varrer um conjunto de documento, sejam eles documentos da internet ou de bases de documento previamente armazenados (*cache*), extraindo destes documentos o conteúdo textual. No caso desse trabalho os documentos estão armazenados em um Banco de Dados Relacional.

Para cada documento é invocado o método *mapBuilder*() também da classe *Extraction* que irá extrair os termos mais relevantes do texto através do algoritmo descrito no capítulo anterior. Após isso, para cada termo extraído é invocado o método *analyzeContent*() que, com base nas tabelas léxicas, tenta atribuir o termo à uma classe. Uma tabela léxica é entendida como um conjunto de termos que geram um contexto para uma classe em particular. Exemplos disto podem ser verificados na Tabela 3. Como mencionado anteriormente, as tabelas léxicas estão diretamente codificados no código fonte do protótipo.

Classe	**Termos**
Pessoa	Alexandre, Leopoldo, Gonçalves, Aloizio, Mercadante, Barack, Obama, Camargo, Engler, Paulo, Speller, Daniel, Bell, Dilma, Rousseff, Eduardo, José, Arruda.
Organização	Agência, Espacial, Brasileira, AEB, FAPESP, Universidade, CEITEC, CNE, Folha, São, Paulo, USP, MEC, Embaixada, Brasil, Unicamp, Hospital, Universitário.
Local	Amazonas, América, Sul, Brasil, Calota polar, Groenlândia, Mato Grosso, Mato Grosso do Sul, Distrito Federal, São Paulo, Estados Unidos, Golfo do México, Itália.

Tabela 3 - Relação de classes e termos em que cada classe é representada por uma tabela léxica

Vale ressaltar que determinado termo pode ser designado para mais de uma classe podendo durante o processo de NER gerar ambiguidade. No contexto do trabalho esta questão não foi tratada. Ao término do método *analyzeContent*() é devolvido como resultado um XML conforme apresentado no Capítulo anterior em que cada termo agora possui uma classe associada.

De posse do XML é invocado o método *maintenance*() da classe *OntologyMaintenance*. Este método obtém cada elemento extraindo a descrição e a classe da entidade a partir do XML de modo que estas informações possam ser adicionadas na ontologia. O processo de inclusão das informações na ontologia, classes e indivíduos, é realizado pelo método *add*(). Ao término o método *add*() é invocado o método *save*() que persiste a ontologia em um arquivo no formato OWL.

O usuário acessa um sistema utilizando uma ferramenta de edição de Ontologia. Para o presente trabalho utilizou-se o *Protégé*®. Uma ontologia editada por esta ferramenta é descrita através de um arquivo OWL que contêm as classes e os indivíduos. A Figura 16 representa através de um caso de uso a interação do usuário (especialista do domínio) com a ontologia.

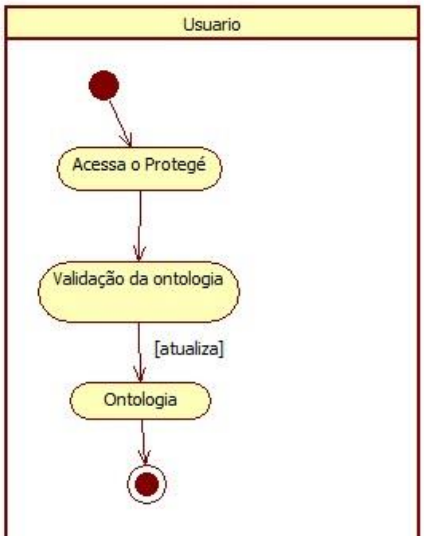


Figura 16 - Diagrama de caso de uso do processo de manutenção da ontologia

Como durante o processo de NER nem todos os termos são nomeados para determinada classe, muito desses indivíduos são atribuídos para uma classe “Geral”, como representado na Figura 17.

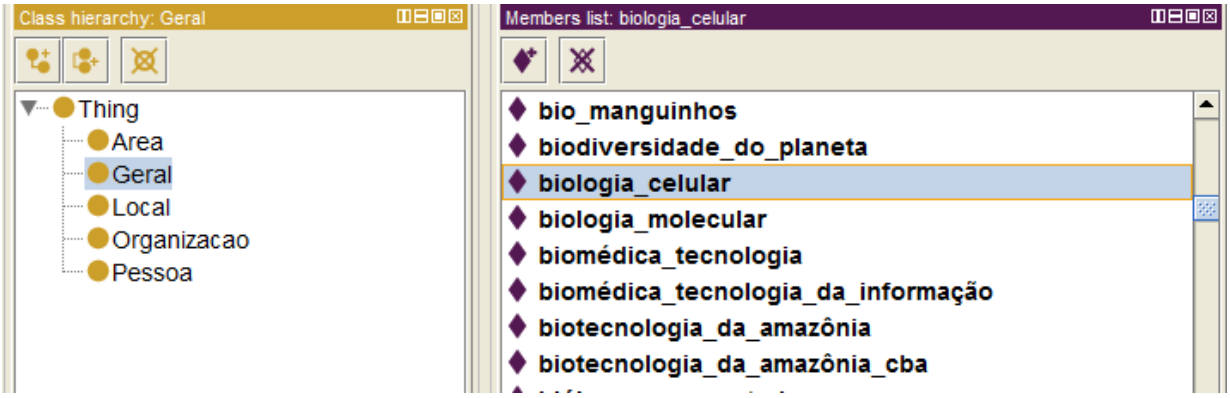


Figura 17 – Classificação dos termos e suas classes

O passo seguinte consiste na validação dos indivíduos da ontologia por intermédio do usuário, em que este se certifica que determinado termo (indivíduo) foi atribuído corretamente para uma determinada classe. Caso este não esteja classificado adequadamente, o usuário pode realizar a alteração da classe. Como exemplo pode-se ter o individuo "biologia_celular" atribuído pelo processo de NER à classe "Geral". Contudo, por se tratar de uma área de pesquisa o usuário poderia melhor classificá-la como pertencente à classe "Área" (Figura 18).

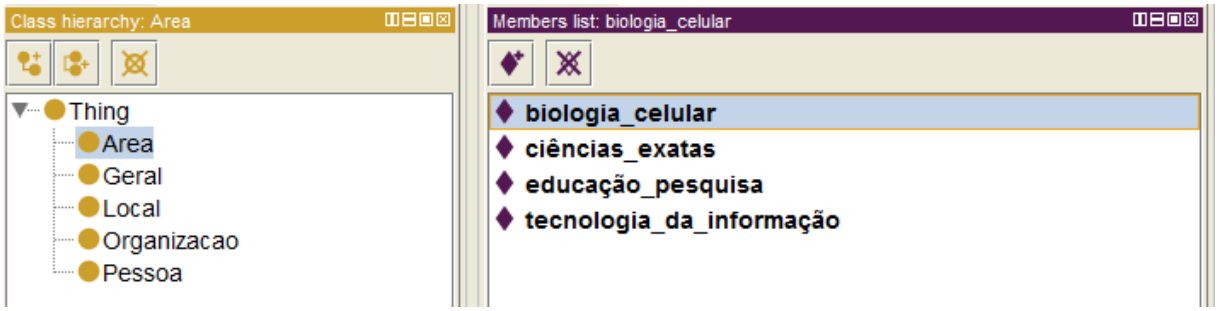


Figura 18 – Classificação realizada pelo usuário

5. APRESENTAÇÃO DOS RESULTADOS

5.1 INTRODUÇÃO

A apresentação dos resultados demonstrada neste capítulo visa permitir uma visão das fases de extração de entidades e manutenção de uma ontologia. Este capítulo está dividido em duas partes, sendo:

- Cenário de Aplicação: Apresenta de maneira geral o cenário informando características de onde os textos foram coletados visando a realização da extração de entidades e manutenção da ontologia;
- Exemplos de Extração de Entidades: Analisa os resultados obtidos através da execução do protótipo desenvolvido, apresentando exemplos de textos, o vetor formado a partir destes textos após o NER e a ontologia carregada com as entidades.

5.2 CENÁRIO DE APLICAÇÃO

Como cenário de aplicação foi utilizado como fonte de informação um conjunto de artigos a partir do Jornal da Ciência[5] disponibilizados publicamente em seu site. Ao todo foram coletados, manualmente, 100 artigos sendo estes armazenados em um Banco de Dados Relacional.

A partir disso desenvolveu-se uma classe que conecta ao banco de dados, coleta cada registro e envia as classes que realizam extração de entidades e atualização da ontologia. Cada registro possui informações sobre o título, a data de publicação e o texto do artigo como demonstrado na Figura 12.

[5] http://www.jornaldaciencia.org.br/index2.jsp

Todos os cenários que serão discutidos iniciam pela recuperação de um documento qualquer que se encontra na base de dados. A partir desse documento o processo de NER é aplicado visando a extração de entidades e subsequentemente o preenchimento da Ontologia.

Considerando a base de 100 documentos que possuem em média 2,8 KBytes ou 430 palavras por documentos, sendo considerados documentos pequenos. O tempo total de processamento, incluindo a extração e nomeação das entidades e a atualização da ontologia, é de 33 segundos. Para se ter um referencial, considerando um documento com 2,5 KBytes representando uma dissertação de mestrado de 130 páginas, o tempo de processamento total (extração de entidades e atualização da ontologia) foi de 6 segundos. Pode-se notar que grande parte do tempo envolve a conexão com o banco de dados e atualização da ontologia, pois considerando que os 100 documentos e o documento da dissertação possuem praticamente o mesmo tamanho em Kbytes, os tempos de processamento na fase de extração são similares para ambos.

5.3 EXEMPLOS DE EXTRAÇÃO DE ENTIDADES

As discussões apresentadas a seguir são baseadas em três documentos obtidos a partir da base de dados que representa a amostra coletada do Jornal da Ciência. Para cada documento será apresentado o texto e as entidades que deveriam ser extraída e as entidades que realmente foram extraídas. Isto objetiva demonstrar o atual estágio de desenvolvimento do protótipo do sistema de reconhecimento de entidades. Além disso, será apresentado o resultado final após a inclusão pelo protótipo das entidades diretamente na ontologia.

Como primeiro cenário tem-se um pequeno trecho do texto “Deputados apontam prioridades para área de ciência e tecnologia” [6] (Figura 19), em que são destacados em amarelo a expectativa de identificação de entidades e em vermelho o que o processo de NER conseguiu extrair deste recorte do texto.

[6] http://www.jornaldaciencia.org.br/Detalhe.jsp?id=76118

[...] As prioridades coincidem com as propostas defendidas pelo ministro da Ciência e Tecnologia, Aloizio Mercadante, em seu discurso de posse. Ele também destacou como prioridade a transição para uma economia "verde e criativa", com sustentabilidade ambiental. [...]

[...] Orçamento outro aspecto defendido por Piau é a ampliação dos recursos destinados à ciência e à tecnologia. Hoje, o investimento no setor é de 1,25% do Produto Interno Bruto (PIB). "Temos de chegar, a médio prazo, a pelo menos a 2,5%, que é o índice americano de investimento", afirmou. Segundo Rollemberg, recursos orçamentários significativos devem ser destinados a iniciativas estratégicas para o país, como o programa espacial e programas voltados ao desenvolvimento da bioenergia, da biotecnologia e da nanotecnologia, além de redução dos danos causados ao meio ambiente. [...]

Figura 19 – Textos destacados os termos esperados e os extraídos pelo processo de NER (cenário um)

Como resultado da extração de termos é criado um vetor de entidades Figura 20, em formato XML, como foi citado no capítulo anterior em que são apresentados os itens encontrados e as classes a que eles pertencem. Vale mencionar que existem mais termos no vetor em relação à marcação em vermelho da figura anterior, pois aqui são apresentadas todas as entidades extraídas considerando o texto completo.

```
<temos>
        <item id= "ciência_e_tecnologia" tipo= "te" classe= "Area" />
        <item id= "espacial_brasileira" tipo= "te" classe= "Organizacao" >
        <item id= "banda_larga" tipo= "te" classe= "Geral" >
        <item id= "lei_do_bem" tipo= "te" classe= "Geral" >
        <item id= "recursos_orçamentarios" tipo= "te" classe= "Geral" >
        <item id= "programa_espacial" tipo= "te" classe= "Geral" >
        <item id= "plano_nacional" tipo= "te" classe= "Geral" >
</termos>
```

Figura 20 – Vetor de entidades (cenário um)

De posse do vetor as informações são salvas na ontologia. A Figura 21 apresenta através da ferramenta *Protégé*®, o indivíduo "ciência_e_tecnologia" pertencente à classe "Area".

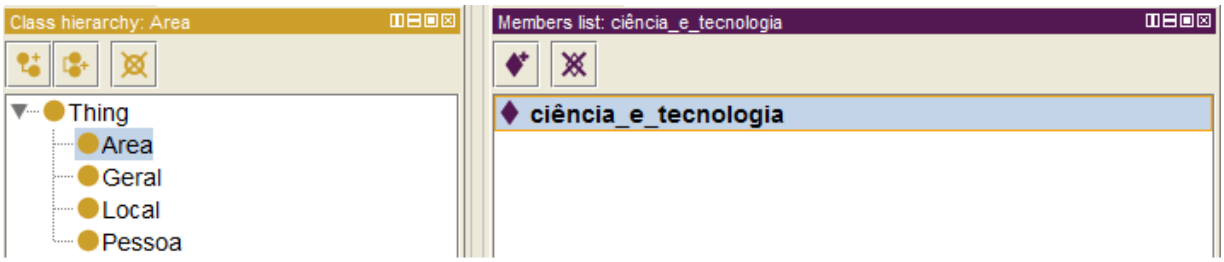


Figura 21 – Apresentação da ontologia extraída do *Protégé*

Para o segundo cenário foi utilizado o texto “Morre o sociólogo Daniel Bell” [7], (Figura 22), utilizando como referência para destaque das palavras os mesmos critérios do exemplo anterior, sendo amarelo para a expectativa e vermelho para o que realmente foi extraído. O texto, assim como o anterior, também é um fragmento.

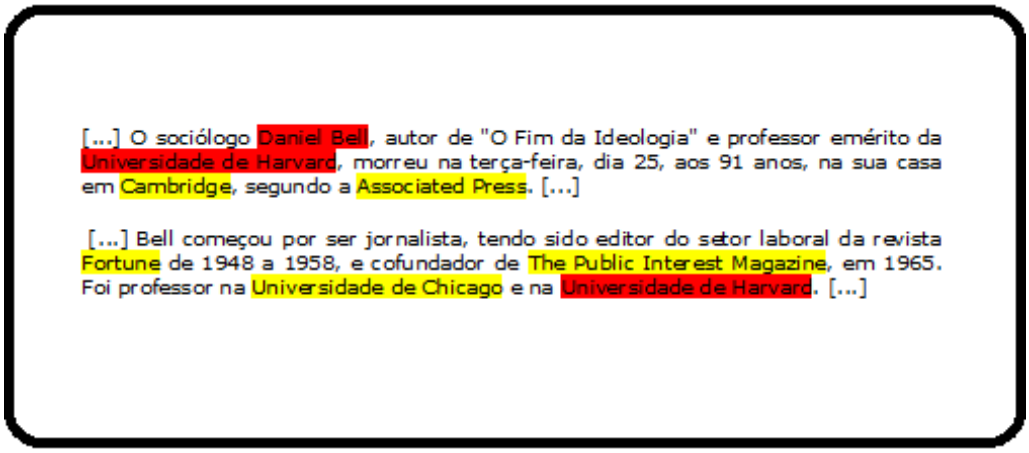


Figura 22 – Texto destacando os termos esperados e os extraídos pelo processo de NER (cenário dois)

O vetor de entidades extraídas no processo é apresentado na Figura 23 em formato XML em que cada elemento é composto pela descrição da entidade e por sua classe.

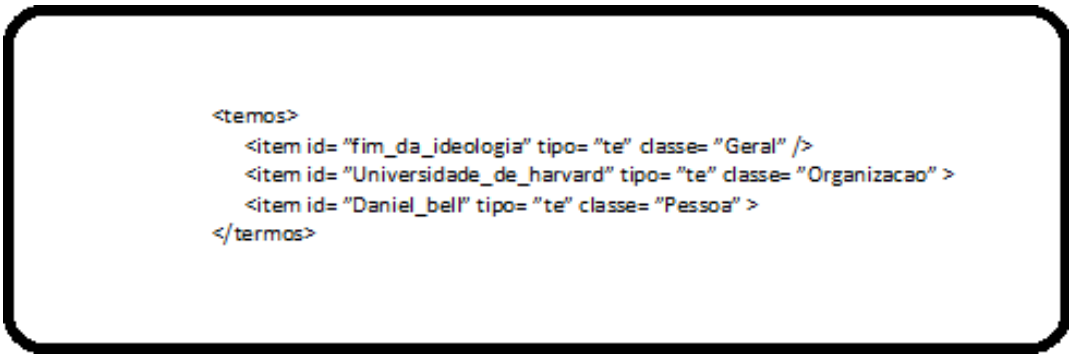


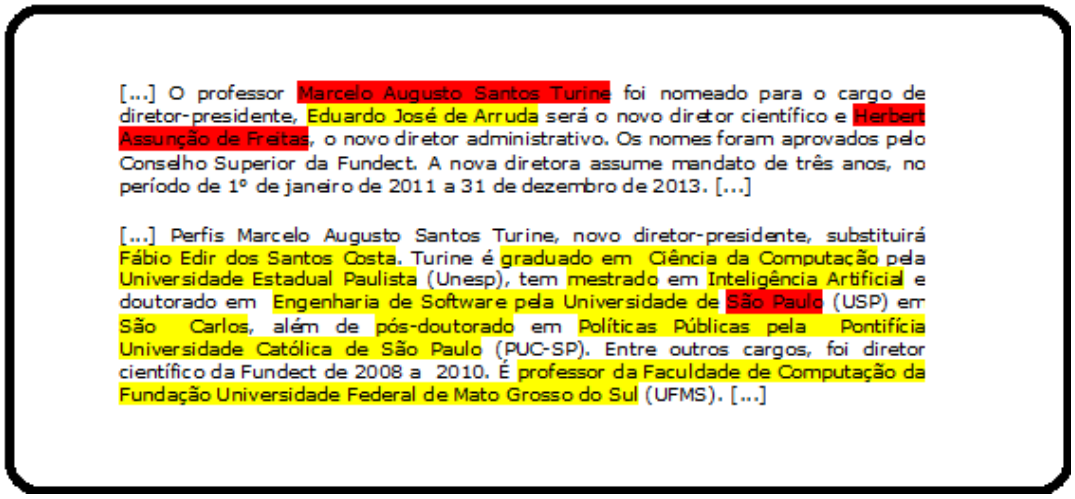
Figura 23 – Vetor de termos (cenário dois)

A Figura 24 apresenta o conteúdo da classe Pessoa após a atualização a partir do vetor de entidades. Vale mencionar que o caractere “_” é utilizado um vez que OWL um indivíduo precisa de um URI e o caractere espaço não é adequado caso seja necessário algum tipo de interoperabilidade entre ontologias. Uma forma de resolver isto é adicionar ao indivíduo uma propriedade de dados, por exemplo, chamada “nome”, para conter a descrição original da entidade.

[7] http://www.jornaldaciencia.org.br/Detalhe.jsp?id=76127

Figura 24 - Apresentação da ontologia extraída do *Protégé*® (cenário dois)

No cenário três foi utilizado o texto "Nova diretoria da FAP do Mato Grosso do Sul é nomeada" [8] (Figura 25), utilizando como destaque dos termos o mesmo utilizado nos cenários um e dois, onde os termos esperados que sejam extraídos pelo processo de NER estão destacados em amarelo e os extraídos estão em vermelho.



[...] O professor Marcelo Augusto Santos Turine foi nomeado para o cargo de diretor-presidente, Eduardo José de Arruda será o novo diretor científico e Herbert Assunção de Freitas, o novo diretor administrativo. Os nomes foram aprovados pelo Conselho Superior da Fundect. A nova diretora assume mandato de três anos, no período de 1º de janeiro de 2011 a 31 de dezembro de 2013. [...]

[...] Perfis Marcelo Augusto Santos Turine, novo diretor-presidente, substituirá Fábio Edir dos Santos Costa. Turine é graduado em Ciência da Computação pela Universidade Estadual Paulista (Unesp), tem mestrado em Inteligência Artificial e doutorado em Engenharia de Software pela Universidade de São Paulo (USP) em São Carlos, além de pós-doutorado em Políticas Públicas pela Pontifícia Universidade Católica de São Paulo (PUC-SP). Entre outros cargos, foi diretor científico da Fundect de 2008 a 2010. É professor da Faculdade de Computação da Fundação Universidade Federal de Mato Grosso do Sul (UFMS). [...]

Figura 25 - Texto destacando os termos esperados e os extraídos pelo processo de NER (cenário três)

O vetor de termos extraído do texto é apresentado na Figura 26 utilizando o mesmo formato dos anteriores em XML e com as entidades e suas classes.

```
<termos>
        <item id="Herbert_assunção_de_freitas" tipo="te" classe= "Pessoa"/>
        <item id="Marcelo_augusto_santos_turine" tipo="te" classe= "Pessoa"/>
        <item id="Mato_grosso" tipo="te" classe= "Local"/>
        <item id="Mato_grosso_do_Sul" tipo="te" classe= "Local"/>
        <item id="São_Paulo" tipo="te" classe= "Local"/>
        <item id="Ciencia_e_tecnologia" tipo="te" classe= "Area"/>
        <item id="José_de_arruda" tipo="te" classe= "Geral"/>
        <item id="Universidade_estadual" tipo="te" classe= "Geral"/>
        .
        .
</termos>
```

Figura 26 - Vetor de termos (cenário três)

[8] http://www.jornaldaciencia.org.br/Detalhe.jsp?id=76119

A Figura 27 apresenta o conteúdo da classe Local após a atualização a partir do vetor de entidades. Neste caso, todas as entidades classificadas como “Local” foram adicionadas a ontologia.

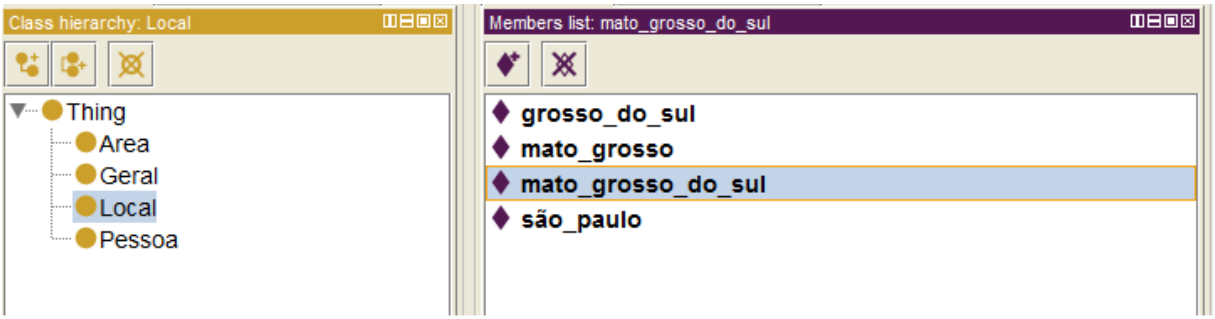


Figura 27 - Apresentação da ontologia extraída do *Protégé*® (cenário três)

De modo geral, o sistema proposto e desenvolvido neste trabalho ainda encontra-se em seus estágios iniciais. Isto pode ser evidenciado pelo número reduzido de entidades reconhecidas nos três cenários apresentados. Várias estratégias poderiam melhorar a precisão do reconhecimento, contudo, muitas dessas impactam no aumento do tempo de processamento, o que se forem consideradas grandes coleções de documentos podem ser proibitivas.

Outro fator a mencionar é a falta de precisão do processo atual, ou seja, o reconhecimento de muitos termos que são atribuídos à classe Geral. A Tabela 4 apresenta um resumo do processo considerando os três textos dos cenários acima quanto ao total de entidades esperadas, ou seja, obtidas através de anotação manual, o número de entidades reconhecidas em comparação com as entidades esperadas e por último, o total de entidades extraídas onde a maior parte foi atribuída à classe Geral. Este resultado pode ser visto principalmente no cenário 1 e 2, enquanto que o 3 cenário permitiu um resultado melhor.

Cenário	Esperado	Esperado-Reconhecido	Extraído
1	16	2	7
2	9	2	3
3	29	14	37

Tabela 4 – Números obtidos no processo de NER considerando os 3 cenários

Tal fato pode ser visto como uma deficiência, mas ao longo de várias interações e considerando algumas melhorias no algoritmo de NER, resultados quanto à precisão podem ser melhorados substancialmente. Isso se justifica uma vez que o usuário tem a opção de interagir com a Ontologia e com o seu conhecimento realizar a classificação correta ou o

descarte do termo. Pensando em um processo incremental que considere novamente a ontologia na fase de extração e nomeação dos padrões (entidades) os erros de classificação podem ser minimizados.

6. CONSIDERAÇÕES FINAIS

O objetivo geral desse trabalho foi desenvolver um sistema que permitisse a extração de entidades e a construção e a manutenção semiautomática de uma ontologia. Neste sentido, foi realizada uma revisão das áreas de extração de informações e ontologia visando dar suporte ao desenvolvimento do trabalho.

A Extração de Informação constitui-se em uma subárea da área de Processamento de Linguagem Natural, que é amplamente utilizada para reconhecimento de padrões, aplicações de estatística e métodos de aprendizagem. A partir do estudo da área este trabalho propôs uma abordagem estatística simplificada conforme detalhamento apresentado no Capítulo 4. A extração de termos relevantes constitui-se no primeiro passo para o processo. O passo seguinte envolve a nomeação dos termos atribuindo a cada um deles para determinada classe. Nesse sentido, o Reconhecimento de Entidades Nomeadas (NER) é responsável pela identificação e classificação de entidades em textos, ou seja, o preenchimento de classes como, Pessoa e Organização.

A partir do resultado prévio de processo de NER torna-se necessário o armazenamento em algum meio físico, por exemplo, um banco de dados. Neste trabalho utilizou-se uma Ontologia, pois esta permite representar vários conceitos de representação de conhecimento que não estão disponíveis estruturas de dados tradicionais. Uma Ontologia permite representar conceitos abstratos de determinado domínio tais como Classes, Indivíduos e os relacionamentos entre eles que podem ser afirmados ou inferidos. Além disso, a ontologia permite o seu reuso, ou seja, a utilização de uma mesma ontologia em diferentes contextos de aplicação visando à disseminação do conhecimento de um domínio.

Para validar os conceitos envoltos nesse trabalho foi desenvolvido um protótipo voltado à extração de entidades e manutenção semiautomática de Ontologias. Ainda que a precisão do processo de extração seja baixa, quando comparado às entidades esperadas e as que foram corretamente identificadas, pode-se afirmar que o protótipo atendeu as

expectativas, pois possibilita a extração de entidades de maneira rápida sem qualquer conhecimento prévio, ou seja, não existe a necessidade de uma fase de treinamento. Além disso, pode ser aplicado a textos pequenos uma vez que a relevância estatística não é obtida considerando-se a palavra como um todo, mas sim pequenos *tokens* chamados de *n*-gramas.

Outro ponto importante a mencionar é a diferença entre a proposição de visão lógica e física em relação à implementação do protótipo. Por questões de tempo alguns passos importantes não foram realizados, por exemplo, a obtenção das tabelas léxicas (conjunto de palavras que definem determinada classe) não vem da ontologia. Estas tabelas foram configuradas e definidas manualmente no código fonte do protótipo. Menciona-se ainda a não implementação de um analisador de expressões regulares, o que, em muitos casos, aumenta a precisão do reconhecimento de entidades.

Ao longo do projeto foram vislumbradas novas possibilidades de melhorias para o projeto que não puderam ser desenvolvidas. Entre as possibilidades destacam-se a implementação de um analisador de expressões regulares e a criação de uma estrutura de processamento distribuída de modo que fosse possível a separação entre os processos de extração e classificação de termos. Outra possibilidade envolve a melhoria no algoritmo de reconhecimento de entidades, por exemplo, a implementação de um algoritmo de duas fases, em que a primeira fase analisa e computa os *n*-gramas relevantes para, em um segundo momento, realizar o reconhecimento. Deste modo, *n*-gramas que não são relevantes em um primeiro momento teriam sua importância alterada ao fim do processo considerando toda a coleção, e deste modo, poderiam melhorar o reconhecimento de termos. Finalmente, mas sem exaurir as possibilidades, o processo de NER poderia ser aperfeiçoado se, de maneira incremental, o conhecimento armazenado na ontologia fosse utilizado cada vez que este fosse executado.

REFERÊNCIAS

ALMEIDA, Mauricio B. Roteiro para a construção de uma ontologia bibliográfica através de ferramentas automatizadas. **Perspectivas em Ciência da Informação**, v. 8, n. 2, p. 164-179, 2003.

ÁLVAREZ, Alberto Cáceres. **Extração de Artigos Científicos:** uma abordagem baseada em indução de regras de etiquetagem. 2007. 131 f. Dissertação (Mestrado) - Usp, São Carlos, 2007.

BORST, Willem Nico. **Construction of Engineering Ontologies for Knowledge Sharing and Reuse**. 1997. 38 f. Tese (Doutorado) - University Of Twente, Enschede, 1997.

BRANSKI, Regina Meyer. **Recuperação de informações na Web**. **Perspectivas em Ciência da Informação**, v. 9, n. 1, p. 70-87, 2004.

BREITMAN, Karin Koogan. **Web semântica:** a internet do futuro. Rio de Janeiro: LTC, 2006. 190 p.

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.

CECI, Flávio; PIETROBON, Ricardo; GONÇALVES, Alexandre Leopoldo. Turning Text into Research Networks: Information Retrieval and Computational Ontologies in the Creation of Scientific Databases. **Plos One**, v. 7, n.1, e27499, 2012.

CHEN, Rung-Ching; HUANG, Yun-Hou; BAU, Cho-Tsan; CHEN, Shyi-Ming. A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. **Expert Systems with Applications**, v. 39, p.3995-4006, 2012.

CHI, Y. L. Ontology-based curriculum content sequencing system with semantic rules. **Expert Systems with Applications**, n. 36, v. 4, p. 7838–7847, 2009.

CHOUEKA, Yaacov. **Looking for needles in a haystack or locating interesting collocational expressions in large textual databases**. Proceedings of the RIAO, 1988. p. 609-624.

CUNNINGHAM, H. **GATE**: a General Architecture for Text Engineering. **Computers and the Humanities**, v. 36, p. 223-254, 2002.

FERNANDEZ, M.; GOMEZ-PEREZ, A; JURISTO, N. METHONTOLOGY: **From Ontological Art Towards Ontological Engineering**, AAAI Technical Report SS-97-06, 1997.

FIRTH, J. R. A synopsis of linguistic theory 1930-1955. **Studies in Linguistic Analysis,** p. l-32, 1957.

GACITUA, R.; ARGUELLO CASTELEIRO, M.; SAWYER, P.; Des, J.; Perez, R.; Fernandez-Prieto, M.J.; Paniagua, H. A collaborative workflow for building ontologies: A case study in the biomedical field. **Proceedings of the Third International Conference on Research Challenges in Information Science (**RCIS), 2009. p.121-128.

GACITUA, R.; SAWYER, P. Ensemble Methods for Ontology Learning - An Empirical Experiment to Evaluate Combinations of Concept Acquisition Techniques. **Proceedings of the Seventh IEEE/ACIS International Conference on Computer and Information Science (ICIS 2008)**, 2008. p. 328-333.

GÓMEZ-PEREZ, A.; SUÁREZ-FIGUEROA, M. C.; NeOn Methodology: Scenarios for Building Networks Ontologies. **Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management Patterns (EKAW)**, 2008.

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.

GRISHMAN, Ralph: Information Extraction: Techniques and Challenges. **Proceedings of the International Summer School on Information Extraction (SCIE)**, 1997. p. 10-27.

GRISHMAN, Ralph; SUNDHEIM, Beth: Message Understanding Conference- 6: A Brief History. **Proceedings of the 16th conference on Computational linguistics (COLING)**, p. 466-471, 1996.

GROVER, C.; GEARAILT, D. N.; KARKALETSIS, V.; FARMAKIOTOU, D.; PAZIENZA, M. T.; VINDIGNI, M. Multilingual XML-Based Named Entity Recognition for E-Retail Domains. **Proceedings of the International Conference on Language Resources and Evaluation (LREC)**, p. 1060-1067, 2002.

GRUBER, Thomas R.. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. **International Journal Human-computer Studies**, v. 43, n. 5-6, p.907-928, 1995.

GUARINO, Nicola. Formal Ontology and Information Systems. Proceedings of the **Formal Ontology in Information Systems (FIOS)**, Trento, Italy, Amsterdam: IOS Press, 1998.

HILBERT, Martin; LÓPEZ, Priscila. The World's Technological Capacity to Store, Communicate, and Compute Information. **Science**, v. 332, n. 6025, p. 60-65, 2011.

HIMMA, K. The concept of information overload: A preliminary step in understanding the nature of a harmful information-related condition. **Ethics and Information Technology**, v. 9, n. 4, p. 259-272, 2007.

HORRIDGE, Matthew. **A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools Edition 1.3.** Manchester: The University Of Manchester, 2011. Disponível em: < http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_1.pdf >. Acesso em: 17 set. 2012.

KOZAREVA, Zornitsa. Bootstrapping named entity recognition with automatically generated gazetteer lists. **Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, 2006. p. 15-21.

KROGH, G. V.; ROOS, J. **Organizational Epistemology**. New York, NY: St. Martin's Press, 1995.

LIDDY, E. D. Natural Language Processing. **Encyclopedia of Library and Information Science**, 2nd Ed., New York: Marcel Decker Inc, 2001.

LIMA, Júnio César de; CARVALHO, Cedric Luiz de. **Ontologias - OWL (Web Ontology Language)**. 2005. Disponível em: < http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-05.pdf>. Acesso em: 17 set. 2012.

LYMAN, Peter; VARIAN, Hal R. **How much information**? Executive summary. 2003.

MAEDCHE, A.; MOTIK, B.; STOJANOVIC, L.; STUDER, R.; VOLZ, R. Ontologies for enterprise knowledge management. **IEEE Intelligent Systems**, v. 18, n. 2, p. 26-33, 2003.

MANNING, Christopher D.; SCHÜTZE, Hinrich. **Foundations of Statistical Natural Language Processing**. Cambridge: Mit Press, 1999.

MARRERO, Mónica; URBANO, Julián; SÁNCHEZ-CUADRADO, Sonia; MORATO, Jorge; GÓMEZ-BERBÍS, Juan Miguel; Named Entity Recognition: Fallacies, challenges and opportunities. **Computer Standards & Interfaces**, In press, 2012.

MISSIKOFF, M.; NAVIGLI, R.; VELARDI, Paola. Integrated approach to Web ontology learning and engineering. **Computer**, v. 35, n. 11, p. 60-63, 2002.

MUSLEA, Ion. **Extraction Patterns for Information Extraction Tasks:** A Survey. Marina Del Rey: American Association for Artificial Intelligence, 1999.

NAVIGLI, R.; VELARDI, Paola; GANGEMI, A. Ontology learning and its application to automated terminology translation. **Intelligent Systems**, v. 18, n. 1, p. 22-31, 2003.

NIST, ACE08-EVALPLAN.V1.2D. **Assessment of Detection and Recognition of Entities and Relations Within and Across Documents.** The Ace 2008 Evaluation Plan, 2008. Disponível em: <http://www.itl.nist.gov/iad/mig/tests/ace/2008/doc/ace08-evalplan.v1.2d.pdf>. Acesso em: 17 set. 2012.

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, 2001.

O'REILLY, Tim. What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. **Communications & Strategies**, n. 1, p.17-37, 2007.

PARK, Ji-Hyun; KIM, Kyung-Hoon; BAE, Jae-Hak J. Analysis of shipbuilding fabrication process with enterprise ontology. **Computers in Human Behavior**, v. 27, p. 1519-1526, 2011.

RABELO, Ricardo J. **Ontologia.** Disponível em: <http://www.das.ufsc.br/~rabelo/Ensino/DAS5316/MaterialDAS5316/Ontologia.pdf>. Acesso em: 28 ago. 2012.

RAFI, Muhammad; QURESHI, H.; KHATOON, H. **Ontology Maintenance via Multi-agents.** Proceedings of the Fifth International Joint Conference on INC, 2009, p. 955-959.

RAUTENBERG, S.; GOMES FILHO, A. C.; TODESCO, J. L.; GAUTHIER, F. Á. O. Ferramenta ontoKEM: uma contribuição à Ciência da Informação para o desenvolvimento de ontologias. **Perspectivas em Ciência da Informação**, v. 15, p. 239-258, 2010.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**: A Modern Approach. New Jersey: Prentice Hall, 1995.

SANTOS, J. E. Bastos dos. Automatic Content Extraction on Semi-structured Documents. **Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)**, 2011. p. 1235–1239.

SCHREIBER, G. et al. **Knowledge Engineering and Management: The CommonKADS Methodology**. Cambridge, Massachusetts: The MIT Press, 2002.

SILVA, E.; BARROS, F. A.; PRUDENCIO, R. B. C. Uma Abordagem de Aprendizagem Híbrida para Extração de Informação em Textos Semi-Estruturados. **Anais do V Encontro Nacional de Inteligência Artificial (ENIA)**, 2005. p. 504-513.

SOUZA, Renato Rocha; ALVARENGA, Lídia. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, v. 33, n. 1, p. 132-141, 2004.

STEVENSON, Mark; WILKS, Yorick. Word-Sense Disambiguation. In: MITKOV, Ruslan (eds). **The Oxford Handbook of Computational Linguistics.** New York : Oxford University Press, 2003. p. 249-254.

STUBBS, Michael. **Text and corpus analysis: computer-assisted studies of language and culture.** Oxford: Blackwell, 1996. 288 p.

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

W3C. **OWL 2 Web Ontology Language Document Overview (Second Edition).** 2012a. Disponível em: <http://www.w3.org/TR/2012/REC-owl2-overview-20121211/>. Acesso em: 21 dez. 2012.

W3C. **Resource Description Framework (RDF): Concepts and Abstract Syntax**. 2012b. Disponível em: <http://www.w3.org/TR/rdf-concepts/>. Acesso em: 21 dez. 2012.

W3C. **SWRL: A Semantic Web Rule Language Combining OWL and RuleML.** 2012c. Disponível em: <http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/>. Acesso em: 21 dez. 2012.

YU, Alexander C. Methods in biomedical ontology. **Journal of Biomedical Informatics**, v. 39, p. 252-266, 2006.

ZAMBENEDETTI, Christian. **Extração de Informações sobre Bases de Dados Textuais.** 2002. 144 f. Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002.