Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Programa de Pós-Graduação em Ciência da Computação

Aprendizagem de Ontologias a partir de Textos

Lucelene Lopes, Renata Vieira

Relatório Técnico N^0 056

Porto Alegre, Outubro de 2009

Resumo

Neste relatório são apresentados de forma detalhada o conceito e as etapas de Aprendizagem de Ontologias (*Ontology Learning*), em particular, no caso desta aprendizagem ser feita a partir de textos. Como todo o processo de construção de ontologias, a Aprendizagem de Ontologia também visa construir uma representação de conhecimento conceitual de um domínio específico que pode ser utilizada em diversas áreas para diferentes aplicações. Porém, nesta abordagem, a construção de ontologia é feita através de métodos automáticos e semi-automáticos de extração de conhecimento originários da área de Aprendizagem de Máquina. Esses métodos buscam reduzir o custo na construção de ontologias, bem como na sua representação estrutural. Cabe salientar que o objetivo deste relatório é uma revisão bibliográfica sem a ambição de propor novas técnicas ou abordagens.

Capítulo 1

Introdução

Ontologias são representações formais de um modelo de domínio. A uma ontologia podemos associar uma base de conhecimento que instancia conceitos e relações desta ontologia. O termo Aprendizagem de Ontologias (*Ontology Learning*) foi introduzido originalmente por Alexander Madche e Steffen Staab [26], e pode ser descrito como processo de aquisição de um modelo de domínio a partir de dados.

Sendo assim, a Aprendizagem de Ontologia pode ser vista como um caso particular de construção de ontologia onde, ao invés de se utilizar o conhecimento de um especialista para construir uma ontologia de maneira manual, utiliza-se um processo de inferência a partir de um volume considerável de dados de forma semi-automatica [11].

Segundo Madche e Staab, a Aprendizagem de Ontologias visa a integração de várias áreas do conhecimento para facilitar a construção de ontologias, em particular a área de Aprendizagem de Máquina. A automatização de todo o processo de construção de ontologias não é possível com as atuais tecnologias, neste sentido o que se busca é um processo semi-automático que minimize a intervenção humana [26].

Neste caso, a área de Aprendizagem de Máquina tem um potencial de contribuição bastante grande, por se tratar de uma área onde existem métodos, técnicas e ferramentas consolidadas [27]. O processo usual de Aprendizagem de Máquina é menos ambicioso que a Aprendizagem de Ontologias, ainda que seja em muitos casos bastante complexo. No entanto, diversas técnicas de extração de conhecimento podem ser adaptadas a certas fases da Aprendizagem de Ontologias, principalmente quando a aprendizagem é feita sobre textos.

Quando a Aprendizagem de Ontologias é feita sobre fontes textuais não estruturadas, é denominada Aprendizagem de Ontologias a partir de Textos [11]. Este processo é bastante complexo sendo necessário estruturá-lo em etapas onde apenas algumas delas poderão ser automatizadas.

O objetivo deste relatório é uma revisão de literatura sobre Aprendizagem de Ontologia a partir de Textos exposta de maneira didática. Sendo assim, esse relatório está organizado, além desta introdução, em duas seções e uma breve conclusão. A seção 2 descreve formalmente Ontologias, utilizando-se de um exemplo apresentado em detalhe. A seção 3 apresenta Aprendizagem de Ontologias a partir de Textos, através de suas etapas. Finalmente a conclusão sumariza a contribuição e sugere trabalhos futuros.

Capítulo 2

Ontologias

Ontologia é uma especificação formal de uma conceitualização [16]. De um ponto de vista formal uma ontologia é uma estrutura [11]:

$$\mathcal{O} := (C, \leq_C, R, \sigma_R, \leq_R, A, \sigma_A, T)$$

Composta de:

- Quatro conjuntos disjuntos:
  - C - identificadores de conceitos;
  - R - identificadores de relação;
  - A - identificadores de atributos; e
  - T - tipos de dados (inteiros, *strings*, etc);
- Um semireticulado superior $\leq_C$ definido sobre os elementos de C (conceitos) chamado de hierarquia de conceitos ou taxonomia, que possui:
  - um supremo $raiz_C$;
  - uma relação de subconceito e superconceito entre dois conceitos c_1 e c_2 pertencentes a C que diz que c_1 é um subconceito de c_2, caso $c_1 \leq_C c_2$, e que c_2 é um superconceito de c_1;
  - adcionalmente caso não exista um conceito c_3 tal que $c_1 \leq_C c_3 \leq_C c_2$, diz-se que c_1 é um subconceito direto de c_2 e, analogamente, c_2 é um superconceito direto de c_1, estas relações denota-se como $c_1 \prec c_2$;
- Uma função σ_R: $R \to C^+$ que estabelece relações entre conceitos, chamada assinatura de relação, estas funções definem uma relação do conjunto R e dois conjuntos de conceitos de C, respectivamente:

  - domínio (*domain*) que diz quais conceitos podem originar a relação; e
  - intervalo (*range*) que diz que conceitos podem ser destino da relação;

- Uma ordem parcial $\leq_R$ sobre R que estabelece uma ordem de precedência de certas relações sobre outras, chamada hierarquia de relação, que de forma análoga a hierarquia de conceitos define:

  - os conceitos de subrelação e superelação que diz que duas relações r_1 e r_2 pertencentes a R onde $r_1 \leq_R r_2$ são: r_1 uma subrelação de r_2 e, analogamente, r_2 uma superelação de r_1; e
  - os conceitos de subrelação e superelação diretas quando não existe uma relação r_3 tal que $r_1 \leq_R r_3 \leq_R r_2$, que denota-se $r_1 \prec r_2$;

- Uma função σ_A: $A \rightarrow C \times T$, similar a função σ_R, mas que relaciona atributos ao invés de conceitos, chamada assinatura de atributos.

Para exemplificar as definições apresentadas, considera-se o exemplo da Figura 2.1. Nesta ontologia os conjuntos C, R, A e T são:

$$\begin{aligned} C = \; & \{ser\_vivo, pessoa, cachorro, homem, \\ & neto, mulher, macho, femea\} \end{aligned}$$

$$\begin{aligned} R = \; & \{pai\_de, inv\_pai\_de, \\ & mae\_de, inv\_mae\_de, \\ & parente\_de, \\ & marido\_de, esposa\_de, \\ & dono\_de, pertence\_a\} \end{aligned}$$

$$A = \; \{idade\}$$

$$T = \; \{inteiros\}$$

O semireticulado superior $\leq_C$ possui o supremo $raiz_C = ser\_vivo$ e sua hierarquia é:

$$\begin{aligned} pessoa &\prec ser\_vivo \\ cachorro &\prec ser\_vivo \\ homem &\prec pessoa \\ neto &\prec homem \\ mulher &\prec pessoa \\ macho &\prec cachorro \\ femea &\prec cachorro \end{aligned}$$

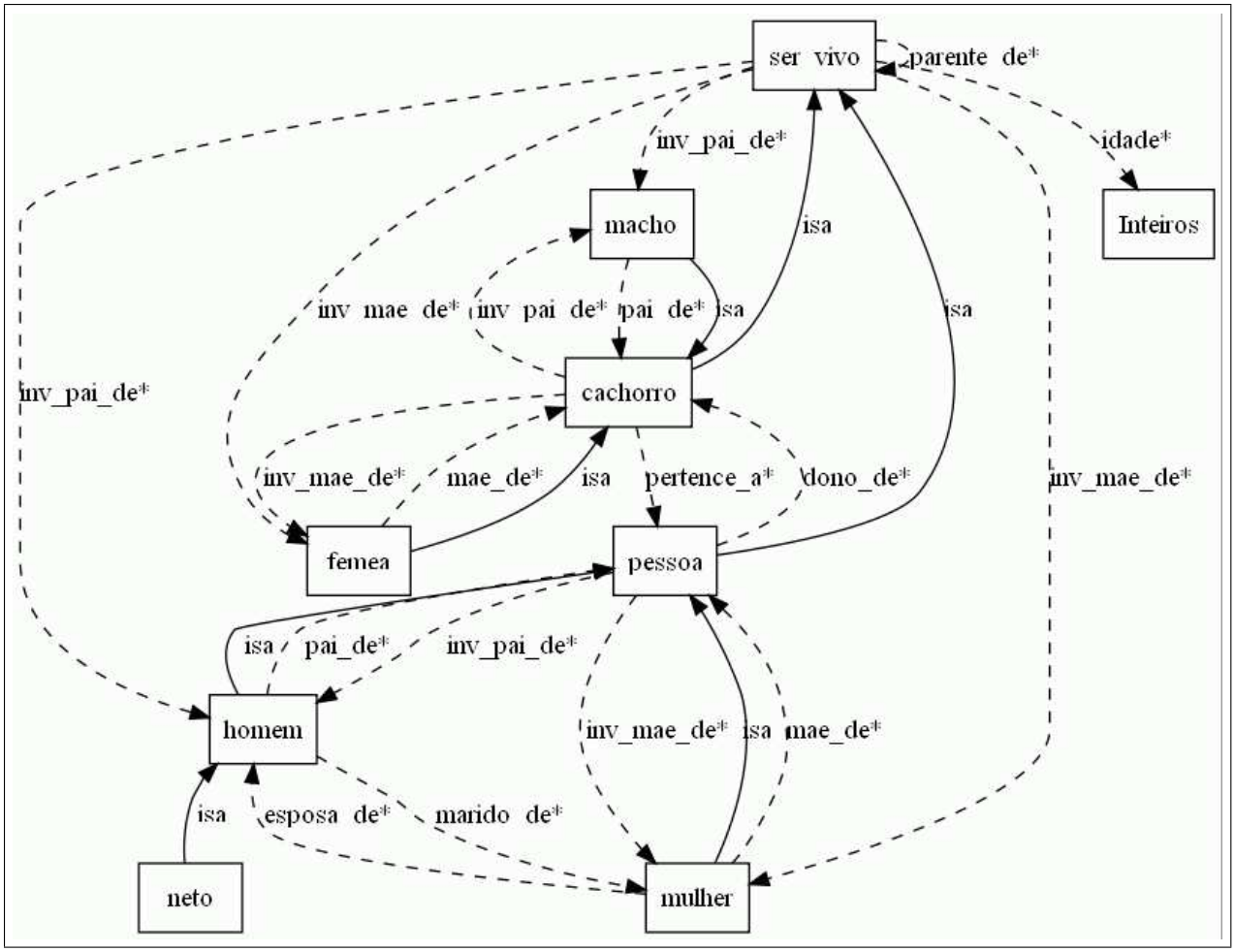

Figura 2.1: Ontologia Exemplo

Utilizando a notação σ_R(relação) = {domain,range}, a função σ_R define:

$$
\begin{aligned}
\sigma_R(pai\_de) &= (\{homem, macho\}, ser\_vivo) \\
\sigma_R(mae\_de) &= (\{mulher, femea\}, ser\_vivo) \\
\sigma_R(inv\_pai\_de) &= (ser\_vivo, \{homem, macho\}) \\
\sigma_R(inv\_mae\_de) &= (ser\_vivo, \{mulher, femea\}) \\
\sigma_R(parente\_de) &= (ser\_vivo, ser\_vivo) \\
\sigma_R(marido\_de) &= (homem, mulher) \\
\sigma_R(esposa\_de) &= (mulher, homem) \\
\sigma_R(dono\_de) &= (pessoa, cachorro) \\
\sigma_R(pertence\_a) &= (cachorro, pessoa)
\end{aligned}
$$

A ordem parcial $\leq_R$ define:

$$\begin{aligned} pai\_de &\prec parente\_de \\ inv\_pai\_de &\prec parente\_de \\ mae\_de &\prec parente\_de \\ inv\_mae\_de &\prec parente\_de \end{aligned}$$

Utilizando uma notação análoga a utilizada para a função σ_R, a função σ_A define apenas:

$$\sigma_A(idade) = (ser\_vivo, inteiros)$$

2.1 Sistema de Axiomas

Usualmente, define-se junto com uma ontologia um conjunto de axiomas que permite estabelecer propriedades necessárias entre conceitos, relações e atributos desta ontologia. Formalmente, um sistema de axiomas $\mathcal{S}$ de uma ontologia $\mathcal{O}$ é definido pela tripla:

$$\mathcal{S} := (AS, \alpha, \mathcal{L})$$

Composta de:

- uma linguagem lógica $\mathcal{L}$;
- um conjunto de axiomas AS que pode fazer referência a conceitos, relações e atributos;
- um mapeamento $\alpha : AS \rightarrow AS_{\mathcal{L}}$

Logo, este conjunto de axiomas pode ser utilizado para definir restrições entre conceitos, tipicamente conceitos que são disjuntos, como por exemplo definições de conceitos disjuntos na ontologia exemplo da Figura 2.1:

- $\forall x(pessoa(x) \rightarrow \neg cachorro(x))$
- $\forall x(homem(x) \rightarrow \neg mulher(x))$
- $\forall x(macho(x) \rightarrow \neg femea(x))$

Igualmente usual na área é utilizar axiomas para definir relações simétricas, ou seja, definir que duas relações tem um comportamento análogo. Por exemplo, na ontologia utilizada nesta seção é possível definir as seguintes simetrias de relações:

- Se um homem é marido de uma mulher, esta será sua esposa:
  $\forall x(marido\_de(x, y) \leftrightarrow esposa\_de(y, x))$;

- Se uma pessoa é dona de um cachorro, este cachorro pertence a esta pessoa: $\forall x(dono\_de(x, y) \leftrightarrow pertence\_a(y, x))$;
- Se um ser vivo é parente de outro, este também é seu parente: $\forall x(parente\_de(x, y) \leftrightarrow parente\_de(y, x))$;
- Se um ser vivo é pai de outro, este será seu filho (inverso de pai): $\forall x(pai\_de(x, y) \leftrightarrow inv\_pai\_de(y, x))$;
- Se um ser vivo é mãe de outro, este será seu filho (inverso de mãe): $\forall x(mae\_de(x, y) \leftrightarrow inv\_mae\_de(y, x))$;

Pode-se ainda utilizar axiomas para definir outros tipos de restrições, como por exemplo definir que as relações de paternidade só existem entre duas pessoas, ou entre dois animais. Isto somado ao fato de que somente indivíduos *mulher* ou $femea$ podem ser mães e, analogamente, somente indivíduos *homem* e *macho* podem ser pai, resulta nas seguintes restrições:

- $\forall x(pessoa(x) \wedge pai\_de(y, x) \rightarrow homem(y))$
- $\forall x(cachorro(x) \wedge pai\_de(y, x) \rightarrow macho(y))$
- $\forall x(pessoa(x) \wedge mae\_de(y, x) \rightarrow mulher(y))$
- $\forall x(cachorro(x) \wedge mae\_de(y, x) \rightarrow femea(y))$
- $\forall x(pessoa(x) \wedge inv\_pai\_de(x, y) \rightarrow homem(y))$
- $\forall x(cachorro(x) \wedge inv\_pai\_de(x, y) \rightarrow macho(y))$
- $\forall x(pessoa(x) \wedge inv\_mae\_de(x, y) \rightarrow mulher(y))$
- $\forall x(cachorro(x) \wedge inv\_mae\_de(x, y) \rightarrow femea(y))$

Outro exemplo de restrição permite definir que um ser vivo possa ter apenas um pai e uma mãe, ou ainda que um cachorro possa ter apenas um dono:

- $\forall x(\exists y\ pai\_de(y, x) \wedge \exists z pai\_de(z, x) \rightarrow z = y)$
- $\forall x(\exists y\ inv\_pai\_de(x, y) \wedge \exists z inv\_pai\_de(x, z) \rightarrow z = y)$
- $\forall x(\exists y\ mae\_de(y, x) \wedge \exists z mae\_de(z, x) \rightarrow z = y)$
- $\forall x(\exists y\ inv\_mae\_de(x, y) \wedge \exists z inv\_mae\_de(x, z) \rightarrow z = y)$
- $\forall x(\exists y\ dono\_de(y, x) \wedge \exists z dono\_de(z, x) \rightarrow z = y)$
- $\forall x(\exists y\ pertence\_a(x, y) \wedge \exists z pertence\_a(x, z) \rightarrow z = y)$

Uma forma mais sofisticada de utilizar os axiomas também pode ser utilizado para definir subconceitos a partir de um axioma, por exemplo, podemos dizer que um homem é neto quando ele é filho de alguém que também é filho de alguém, ou seja:

$$\begin{aligned} \forall x(neto(x) \quad \rightarrow \quad & \exists y(pai\_de(y,x) \vee mae\_de(y,x)) \ \wedge \\ & \exists z(pai\_de(z,y) \vee mae\_de(z,y))) \end{aligned}$$

2.2 Base de Conhecimento

Uma vez definida a ontologia e o sistema de axiomas, a ontologia é populada através da definição de instâncias para conceitos, relações e atributos.

De um ponto de vista formal, isto é feito através da definição de uma base de conhecimento:

$$\mathcal{KB} := (I, \iota_C, \iota_R, \iota_A)$$

Composta de:

- um conjunto I de identificadores de instâncias, ou simplesmente instâncias;
- uma função $\iota_C : C \rightarrow \mathfrak{P}(I)$, chamada instanciação de conceitos, que define para cada conceito $c \in C$ qualquer subconjunto[1] de I;
- uma função $\iota_R : R \rightarrow \mathfrak{P}(I^+)$, chamada instanciação de relações, que define para cada relação $r \in R$ qualquer tupla[2] contendo elementos de I;
- uma função $\iota_A : A \rightarrow (I \cup_{t \in T} [\![t]\!])^+$, chamada instanciação de atributos, que define para cada atributo $a \in A$ um par com uma instância de I e um elemento do seu tipo de dados t.

A título de exemplo, para a ontologia apresentada na Figura 2.1, acrescenta-se as instâncias conforme a Figura 2.2.

Assim sendo, considera-se o seguinte conjunto I:

$$\begin{aligned} \text{I} \quad := \{ \quad & \text{Mateus, Lucas,} \\ & \text{Joao, Maria,} \\ & \text{Rita, Ines,} \\ & \text{Lobo, Sultao,} \\ & \text{Pitucha, Huli } \} \end{aligned}$$

[1] A notação $\mathfrak{P}(I)$ denota o conjunto com todos os subconjuntos possíveis do conjunto I.
[2] A notação I^+ denota todos os conjuntos possíveis de tuplas formadas por elementos de I.

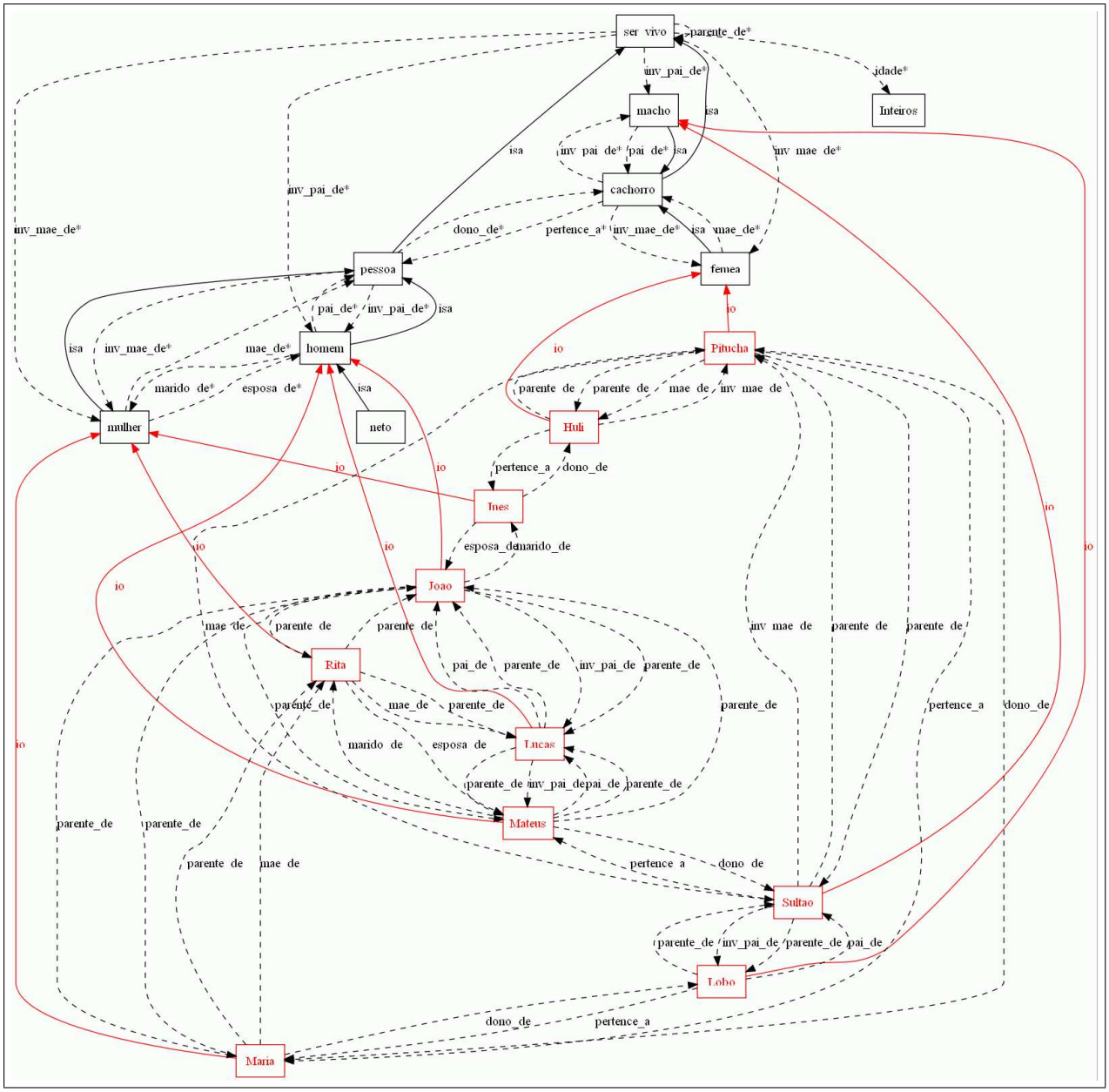

Figura 2.2: Ontologia Exemplo e Todas Instâncias

As instanciações de conceitos são:

$$\begin{aligned} \iota_C(homem) &:= \{ \text{Mateus, Lucas, Joao} \} \\ \iota_C(mulher) &:= \{ \text{Maria, Rita, Ines} \} \\ \iota_C(macho) &:= \{ \text{Lobo, Sultao} \} \\ \iota_C(femea) &:= \{ \text{Pitucha, Huli} \} \end{aligned}$$

As instanciações de relações são:

$$
\begin{array}{lll}
\iota_R(pai\_de) & := \{ & \text{(Mateus,Lucas), (Lucas,Joao),} \\
 & & \text{(Lobo,Sultao) } \} \\
\iota_R(mae\_de) & := \{ & \text{(Maria,Rita), (Rita,Ines),} \\
 & & \text{(Pitucha,Huli) } \} \\
\iota_R(parente\_de) & := \{ & \text{(Mateus,Joao), (Joao,Maria),} \\
 & & \text{(Rita,Joao) } \} \\
\iota_R(marido\_de) & := \{ & \text{(Mateus,Rita) } \} \\
\iota_R(esposa\_de) & := \{ & \text{(Ines,Joao) } \} \\
\iota_R(dono\_de) & := \{ & \text{(Mateus,Sultao), (Maria,Pitucha),} \\
 & & \text{(Maria,Lobo), (Ines,Huli) } \}
\end{array}
$$

As intanciações de atributos são:

$$
\begin{array}{lll}
\iota_A(idade) & := \{ & (Mateus, 65), \\
 & & (Lucas, 43), \\
 & & (Joao, 22), \\
 & & (Maria, 56), \\
 & & (Rita, 30), \\
 & & (Ines, 21), \\
 & & (Lobo, 6), \\
 & & (Sultao, 3), \\
 & & (Pitucha, 8), \\
 & & (Huli, 5) \}
\end{array}
$$

2.3 Extensões

Aplicando-se a uma ontologia $\mathcal{O}$ instânciada por uma base de conhecimentos $\mathcal{KB}$ e levando-se em consideração um sistema de axiomas $\mathcal{S}$ é possível popular esta definição com instanciações adicionais decorrentes do semireticulado $\leq_C$, da ordem parcial $\leq_R$ e da aplicação dos axiomas. Estas extensões são definidas como $[\![c]\!]$, para conceitos $c \in C$, $[\![r]\!]$, para relações $r \in R$ e $[\![a]\!]$, para atributos $a \in A$.

Considerando o exemplo apresentado, decorre do semireticulado $\leq_C$:

$$
\begin{array}{lll}
[\![pessoa]\!] & := \{ & \text{Mateus, Lucas, Joao,} \\
 & & \text{Maria, Rita, Ines } \} \\
[\![cachorro]\!] & := \{ & \text{Lobo, Sultao,} \\
 & & \text{Pitucha, Huli } \} \\
[\![ser\_vivo]\!] & := \{ & \text{Mateus, Lucas, Joao,} \\
 & & \text{Maria, Rita, Ines,} \\
 & & \text{Lobo, Sultao,} \\
 & & \text{Pitucha, Huli } \}
\end{array}
$$

Decorre dos axiomas de simetria:

$$
\begin{array}{lll}
[\![parente\_de]\!] & := \{ & (\text{Joao,Mateus}), \\
 & & (\text{Maria,Joao}), \\
 & & (\text{Joao,Rita}) \} \\
[\![inv\_pai\_de]\!] & := \{ & (\text{Lucas,Mateus}), \\
 & & (\text{Joao,Lucas}), \\
 & & (\text{Sultao,Lobo}) \} \\
[\![inv\_mae\_de]\!] & := \{ & (\text{Rita,Maria}), \\
 & & (\text{Ines,Rita}), \\
 & & (\text{Huli,Pitucha}) \} \\
[\![marido\_de]\!] & := \{ & (\text{Joao,Ines}) \} \\
[\![esposa\_de]\!] & := \{ & (\text{Rita,Mateus}) \} \\
[\![pertence\_a]\!] & := \{ & (\text{Sultao,Mateus}), \\
 & & (\text{Pitucha,Maria}), \\
 & & (\text{Lobo,Maria}), \\
 & & (\text{Huli,Ines}) \}
\end{array}
$$

Decorre do axioma que define o subconceito *neto*:

$$[\![neto]\!] := \{ \text{Joao} \}$$

Decorre da ordem parcial $\leq_R$:

$$
\begin{array}{lll}
[\![parente\_de]\!] & := \{ & (\text{Mateus,Lucas}), \\
 & & (\text{Lucas,Joao}), \\
 & & (\text{Lobo,Sultao}), \\
 & & (\text{Maria,Rita}), \\
 & & (\text{Rita,Ines}), \\
 & & (\text{Pitucha,Huli}), \\
 & & (\text{Lucas,Mateus}), \\
 & & (\text{Joao,Lucas}), \\
 & & (\text{Sultao,Lobo}), \\
 & & (\text{Rita,Maria}), \\
 & & (\text{Ines,Rita}), \\
 & & (\text{Huli,Pitucha}) \}
\end{array}
$$

Consideradas todas as instanciações e as extensões apresentadas ficam definidos os indivíduos e seus respectivos conceitos, relações e atributos conforme descrito na Tabela 2.1.

Indivíduo	Conceitos	Relações	Atributos
Mateus	*homem*	*pai_de*(Lucas)	*idade*(65)
	pessoa	*parente_de*(Joao)	
	ser_vivo	*parente_de*(Lucas)	
		marido_de(Rita)	
		dono_de(Sultao)	
Lucas	*homem*	*pai_de*(Joao)	*idade*(43)
	pessoa	*inv_pai_de*(Mateus)	
	ser_vivo	*parente_de*(Joao)	
		parente_de(Mateus)	
Joao	*homem*	*inv_pai_de*(Lucas)	*idade*(22)
	pessoa	*marido_de*(Ines)	
	ser_vivo	*parente_de*(Mateus)	
	neto	*parente_de*(Maria)	
		parente_de(Rita)	
		parente_de(Lucas)	
Maria	*mulher*	*mae_de*(Rita)	*idade*(56)
	pessoa	*parente_de*(Joao)	
	ser_vivo	*parente_de*(Rita)	
		dono_de(Pitucha)	
		dono_de(Lobo)	
Rita	*mulher*	*mae_de*(Lucas)	*idade*(30)
	pessoa	*esposa_de*(Mateus)	
	ser_vivo	*parente_de*(Joao)	
		parente_de(Lucas)	
Ines	*mulher*	*esposa_de*(Joao)	*idade*(21)
	pessoa	*dono_de*(Huli)	
	ser_vivo		
Lobo	*macho*	*pai_de*(Sultao)	*idade*(6)
	cachorro	*parente_de*(Sultao)	
	ser_vivo	*pertence_a*(Maria)	
Sultao	*macho*	*inv_pai_de*(Lobo)	*idade*(3)
	cachorro	*inv_mae_de*(Pitucha)	
	ser_vivo	*parente_de*(Lobo)	
		parente_de(Pitucha)	
		pertence_a(Mateus)	
Pitucha	*femea*	*mae_de*(Sultao)	*idade*(8)
	cachorro	*mae_de*(Huli)	
	ser_vivo	*parente_de*(Sultao)	
		parente_de(Huli)	
		pertence_a(Maria)	
Huli	*femea*	*inv_mae_de*(Pitucha)	*idade*(5)
	cachorro	*parente_de*(Pitucha)	
	ser_vivo	*pertence_a*(Ines)	

Tabela 2.1: Instanciação e Extensões da Ontologia Exemplo

Capítulo 3

Aprendizagem de Ontologias a partir de Textos

O processo de aprendizagem de ontologia a partir de textos segundo [11] é composto de oito etapas seguidas da população da ontologia. No entanto para facilitar o entendimento, nesse relatório optou-se por aglutinar estas tarefas em 5 grupos:

- extração de termos: este grupo trata das tarefas de extração léxica de termos e definição de sinônimos definida em [11];
- definição de conceitos: neste grupo estão as tarefas responsáveis pela definição de que elementos vão compor o conjunto C e o semireticulado superior $\leq_C$ conforme a definição formal da seção anterior;
- definição de relações e atributos: neste grupo encontra-se as tarefas que definem os elementos de R A e T, bem como as funções σ_R e σ_A e a ordem parcial $\leq_R$;
- instanciação de axiomas: este grupo é composto pelas tarefas de instanciação de axiomas (conjunto AS) e sua representação em uma linguagem lógica $\mathcal{L}$ (α);
- população de ontologias: este grupo é composto pelas tarefas de instanciação de conceitos e relações a partir de textos, ou seja, criar a base de conhecimentos $\mathcal{KB}$ conforme a definição formal da seção anterior.

3.1 Extração de Termos

A tarefa de extração de termos é o ponto de partida para a Aprendizagem de Ontologias a partir de Textos. Portanto, há necessidade de cuidado redobrado nesta etapa para não comprometer a qualidade das etapas seguintes. Está tarefa consiste em estabelecer um conjunto de termos relevantes com significado para um determinado domínio. Um termo é uma palavra ou conjunto de palavras que possui uma semântica associada ao domínio de interesse. Usualmente, a extração de termos está baseada em métodos de recuperação

de informação através de indexação de termos [29], ou em métodos de Processamento de Linguagem Natural [4].

Esses métodos, em sua maioria, utilizam técnicas estatísticas para o processamento de extração de termos, além de estabelecer os termos estes devem ser associados segundo sua semântica em sinônimos. Sinônimos são termos que podem ser empregados com pelo menos um significado equivalente, sem necessariamente serem sinônimos perfeitos. Por exemplo, apesar de "cachorro" e "melhor amigo do homem" serem termos diferentes, em determinados contextos podem ser empregados como sinônimos.

A entrada desta etapa é um conjunto de textos, o resultado intermediário é uma lista de termos e o resultado final é uma lista de conjuntos de termos que possuem para o domínio escolhido uma semântica equivalente entre si.

3.2 Definição de Conceitos e Hierarquia

A tarefa de identificação de conceitos está baseada na busca de similaridade semântica entre os termos de um contexto. Nesse sentido, a busca de similaridade semântica se torna semelhante à indetificação de sinônimos.

No caso de identificação de sinônimos, procura-se termos diferentes que podem ser substituídos sem alteração de significado. Por exemplo, em um determinado contexto o termo "cachorro" pode ser substituído pelo termo "cão". Para identificação de conceitos busca-se termos que também são utilizados de maneira similar, porém sua substituição muda o significado no contexto. Por exemplo, os termos "cachorro" e "gato" fazem parte de um mesmo conceito (animais) mas não são sinônimos.

Entre as técnicas de extração de conceitos pode-se observar três diferentes abordagens:

- abordagem baseadas em agrupamento que consideram grupos (cluster) de termos relacionados como conceitos [19, 24, 28].

- técnicas de redução de dimensões que revelam conexões inerentes entre palavras que levam a formação de grupos [32, 21].

- abordagem de um ponto de vista extensional, ou seja, apartir de alguns conceitos dados se estende suas definições para novos conceitos através de interpretação composicional [13, 12, 35].

A indução de uma hierarquia entre os conceitos detectados segue um de três paradigmas:

- busca de padrões léxico-sintáticos que explora a extrutura interna de frases nominais para extração de relações taxonômicas. Ainda que bastante precisa, essa técnica tem pouca aplicabilidade, pois, tais padrões não são encontrados com frequência [5].

- algoritmos de agrupamento hierárquico baseado na hipótese distribuicional de Harris. Neste caso, a busca de hierarquia é frequentemente concatenada com a detecção de conceitos também feita por agrupamento [8, 3, 14].

- análise de co-ocorrência de termos que busca a hierarquia de acordo com a co-ocorrência de termos na mesma sentença, parágrafo ou documento. Por exemplo, um termo t_1 é mais específico que um termo t_2 se t_2 aparece em todos documentos nos quais aparece t_1 e o contrário não é verdade [30].

3.3 Definição de Relações e Atributos

A definição de relações e atributos tem como objetivo encontrar conceitos em C que possuem uma relação ontológica não taxonomica entre eles. No caso do conjunto de relações (R) busca-se uma relação não taxonomica entre dois ou mais conceitos, por exemplo, entre dois seres vivos existe uma relação de parentesco. No caso de atributos (A), busca-se uma relação entre um conceito e um valor de um tipo de dado definido (T), por exemplo, um ser vivo tem uma idade que é um valor numérico pertencente a $\mathbb{N}$.

Uma vez definidas as relações e atibutos, é necessário definir seus nomes correspondes de acordo com as ocorrências do corpus. Por exemplo, a relação de parentesco pode ser chamada *parente_de*, enquanto o atributo de idade será simplesmente *idade*. Em seguida é necessário determinar o nível correto de abstração de acordo com a hierarquia $\leq_C$ para estabelecer o domínio e intervalo de cada relação $r \in R$, bem como as informações correspondentes para os atributos $a \in A$. Finalmente, é preciso identificar as possíveis hierarquias entre as relações, ou seja, a ordem parcial $\leq_R$.

Determinar as relações é uma das mais complexas dentre as tarefas de construção de ontologias a partir de textos. Poucas abordagens foram empregadas com este propósito e seu sucessso é discutível. Dentre elas cita-se o trabalho de Madche e Staab [34] que baseia-se numa variante do algoritmo de extração de regras de associação que procura a co-ocorrência de termos em setenças [11]. O trabalho de Ciaramita *et alli* [7] segue a mesma linha respeitando a hierarquia de conceitos e baseando-se em dependências sintáticas encontradas no texto. No entanto, esses trabalhos, segundo Cimiano, são apenas abordagens superficiais que estão distantes de prover uma solução satisfatória para a definição de relações e atributos a partir de texto.

3.4 Instanciação de Axiomas

Para a definir os axiomas de uma ontologia, se assume a existência de um conjunto de axiomas AS que possui definições usuais como por exemplo, disjunção de conceitos e simetria de relações. Desta forma, a tarefa a ser feita consiste em instanciar esses axiomas de AS descobrindo a partir do contexto, por exemplo:

- que conceitos são disjuntos: Haase e Volker [17] propõem uma abordagem que procura termos coordenados e experessões como "homens e mulheres" que indicam uma provável disjunção destes conceitos. Nota-se que disjunções não ocorrem necessariamente com apenas um par de conceitos, por exemplo, uma expressão "peixes, cães e gatos" pode indicar que estes três conceitos são disjuntos.

- que relações são simétricas: Lin e Pantel [23] propõem uma abordagem que analisa similaridade em caminhos de árvores de depêndencias que possam sugerir simetrias de relações, ou seja, um certo número de inversões em domínios e intervalos entre duas relações pode indicar uma simetria como, por exemplo, nas funções *dono_de* e *pertence_a* da ontologia da seção anterior.

No que diz respeito a outros axiomas além dos usuais, ou seja, restrições específicas a cada ontologia individualmente, pouco pode ser feito com as tecnologias atuais, pois esta é a área menos pesquisada no que diz respeito à aprendizagem de ontologias [11]. A busca deste tipo de axioma é ao mesmo tempo complexa e relativamente de rara aplicação, pois não é frequente encontrar este tipo de restrições em ontologias. Os poucos trabalhos aproximativos nesta área foram feitos por Shamsfard e Barforoush [33] tentando derivar axiomas de expressões condicionais quantificadas e Lavrac e Dzeroski [22] buscando aplicar programação lógica indutiva a grandes conjuntos de dados de treino.

3.5 População de Ontologias

A tarefa de popular ontologias a partir de textos consiste em instanciar conceitos, relações e atributos por meio de tarefas de reconhecimento de entidades nomeadas [11], ou seja, construir a base de conhecimentos $\mathcal{KB}$ conforme a definição formal da seção anterior.

Enquanto instanciar relações e atributos em um corpus é uma tarefa muito difícil que requer conhecimento completo da linguagem natural, e portanto está além da fronteira tecnológica atual, a instaciação de conceitos tem sido proposta com relativo sucesso por diversas pesquisas na área. Intuitivamente, esta maior dificuldade de detectar relações ao invés de conceitos faz sentido. Por exemplo na ontologia proposta na Seção 2, é mais fácil depreender de um texto quem são os homens, quem são as mulheres e quem são os cachorros, do que as relações que existem entre eles.

Ainda que somente a instaciação de conceitos seja viável no momento, seus objetivos tem sido modestos. A maioria dos trabalhos busca classificar entidades nomeadas sobre um conjunto finito, conhecido e, frequentemente, pequeno de conceitos. São exemplos os trabalhos de Hirshman e Schinchor [20] que instanciam apenas três conceitos: *pessoas*, *localidades* e *organizações*; e de Fleischman e Hovy [15] que instanciam oito classes: *atletas*, *políticos*, *religiosos*, *empresários*, *artistas*, *cientistas* e *policiais*.

Porém, alguns trabalhos são mais ambiciosos ao tentarem classificar números maiores de conceitos, como é o caso de Hahn e Schnattinger [18] que classificam entidades em 325 conceitos e Alfonseca e Manandhar [1] que classificam sobre 1200 conjuntos de sinônimos. Mais ambicioso ainda é o trabalho de Evans [13] que faz ao mesmo tempo a detecção dos conceitos e sua instanciação ao mesmo tempo, ou seja, este trabalho não parte de um conjunto conhecido de conceitos.

Finalmente, Cimiano [11] propõe duas metodologias de instanciação de conceitos:

- População baseada em corpus: uma metodologia que parte de uma hierarquia de conceitos pré-definida que contém um grande número de conceitos (da ordem de centenas). Esta aboradgem calcula medidas de similaridade e, segundo o autor, funciona

de forma independente do corpus utilizado. Esta abordagem mais próxima de técnicas tradicionais foi proposta por Cimiano e Volker [9].

- Aprendizado por *Googling*: uma metodologia moderna que a partir de um conjunto de padrões independente de domínio busca classificar entidades contidas em um texto a partir de resultados obtidos no Google para estas entidades. Esta abordagem baseia-se na idéia de que o conhecimento global sobre um termo (expresso pela busca no Google) supera o conhecimento individual que possa ser ter sobre este termo. Esta forma de classificação está disponível no sistema C-PANKOW [10].

Capítulo 4

Conclusão

O objetivo deste relatório foi traçar um panorama das técnicas disponíveis para aprendizado de ontologias a partir de texto.

Uma das conclusões naturais ao final desta revisão bibliográfica é o fato indiscutível que esta área ainda apresenta muitos desafios e uma quantidade enorme de questões em aberto. Apesar disso, muitas pesquisas tem sido feitas e a compreensão de cada uma delas é um trabalho futuro bastante grande devido a complexidade das técnicas envolvidas. Notável também é a grande variedade das abordagens na área que vai desde trabalhos baseados em estudos sociais, como é o caso da abordagem de aprendizagem por Googling [10], até trabalhos completamente baseados em lógica induditiva, como é o caso de detecção de axiomas gerais [22].

Um trabalho futuro igualmente necessário consiste em observar outras técnicas práticas aplicadas na extração de termos como o trabalho de Bourigault e Lame [4] sobre textos jurídicos em francês e métodos estatísticos sofistificados baseados em amostragem como o trabalho de Baroni e Bernardini [2] que propõe um método sofisticado chamado BootCat. Além destes, outros trabalhos semelhantes podem ser incluídos, pois, como foi dito, esta área ainda carece de muitas pesquisas e muito precisa ser pesquisado.

Referências Bibliográficas

[1] ALFONSECA , E.; MANANDHAR, S. Extending a lexical ontology by a combination of distributional semantics signatures. In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW), pp. 1-7, 2002.

[2] BARONI, M.; BERNADINI, S. BootCaT: Bootstrapping Corpora and Terms from the Web. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), pp.1313-1316, 2004.

[3] BISSON, G.; NEDELLEC, C; CANAMERO, L. Designing clustering methods for ontology building- The Mo'K workbench. In Proceedings of the ECAI Ontology Learning Workshop, 2000.

[4] BOURIGAULT, D.; LAME, G. Analyse distributionnelle et structuration de terminologie - Application à la construction d'une ontologie documentaire du Droit, TAL, 43-1, pp. 1-22, 2002.

[5] BUITELLAAR, P.; OLEJNIK, D.; SINTEK, M. A Protégé plug-in for ontology extraction from text based on linguistic analysis. In Proceedings of the 1st European Semantic Web Simposium (ESWS), 2004.

[6] BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. In: Buitelaar, P.; Cimiano, P.; and Magnini, B. (Ed.). Ontology Learning from Text: Methods, Evaluation and Applications, v. 123 of Frontiers in Artificial Intelligence and Apllications. IOS Press, 2005.

[7] CIARAMITA, M.; GANGEMI, A.; RATSCH,E.; SARIC, J.; ROJAS, I. Unsupervised learning of semantic relations between concepts of molecular biology ontology. In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI), 2005.

[8] CIMIANO, P.; HOTHO,A.; STAAB,S. Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In Proceedings of the European Conferenc on Artificial Intelligence (ECAI), 2004.

[9] CIMIANO, P.; HOTHO,A.; STAAB,S. Learning concept hierarchies from text corpora using formal concept analysis. Journal of Artificial Intelligence Research (JAIR), v. 24, pp. 305-339, 2005.

[10] CIMIANO, P.; LADWIG,G.; STAAB,S. Gimme the contxt: Context-driven automatic semanti annotation with C-PANKOW. In Porceedings of the 14th Word Wide Web Conference (WWW), pp. 332-341, 2005.

[11] CIMIANO, P. Ontology Learning and Population from Text - Algorithms, Evaluation and Applications. Springer, 2006.

[12] ETZIONI, O.; CAFARELLA, M.; DOWNEY, D.; POPESCU, A.-M.; SHAKED, T.; SODERLAND, S.; WELD, D.; YATES,A. Methodsfor domain-independent information extraction from the web: An experimental comaprison. In Proceedings of the 19th National Conference on Artificial Intelligence (AAAI), 2004.

[13] EVANS,R. A framework for named entity recognition in the open domain. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP), 2003.

[14] FAURE, D.; NEDELLEC, C. A corpus-based conceptual clustering method for verb frames and ontology. In Velardi, P., editor, Proceeding of the LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications, 1998.

[15] FLEISCHMAN, M.; HOVY, E.Fine grained classification of named entities. In Proceedings of the 19th International Conference on Computational Linguistics (COLING), pp. 1-7, 2002.

[16] GRUBER, T. Toward principles for the design of ontologies used knowledge sharing. In Formal Analysis in Conceptual Analysis and Knowledge Representation. Kluwer, 1993.

[17] HAASE, P.; VOLKER, J. Ontology learning and reasoning- dealing with uncertainty and inconsistency. In Proceedings of the Workshop on Uncertainty Reasoning of the Semantic Web (URSW), pp 45-55, 2005

[18] HAHN, U.; SCHNATTINGER, K. Ontology engineering via text understanding. In Proceedings of the 15th IFIP World Computer Congress, pp. 429-442, 1998.

[19] HINDLE, D. Noun Classification from predicate-argument structures. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 1990.

[20] HIRSCHMAN, L.; CHINCHOR, N. Muc-7 named entity task definition. In Proceedings of the 7th Message Understanding Conference (MUC-7), 1997.

[21] LANDAUER, T.; DUMAIS, S. Asolution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychological Rewiew, v. 104, pp. 211-240, 1997.

[22] LAVRAC, N.; DZEROSKI, S. Inductive Logic Programming: Techiniques and Applications. Ellis Horwood, 1994.

[23] LIN, D.; PANTEL, P. DIRT - discovery of inference rules from text. In Proceedings of the ACM SIGKDD Conference on Knowlwdge Discorevy and Data Mining, pp. 323-328, 2001.

[24] LIN, D.; PANTEL, P. Concept discovery from text. In Proceedings of the International Conference on Computacional Linguistics (COLLING), 2002.

[25] MADCHE, A.; STAAB, S. Semi-automatic Engineering of Ontologies from Text. In: *Proceedings of the 12th Intenational Conference on Software Engineering and Knowledge Engineering*, 2000.

[26] MADCHE, A.; STAAB, S. Ontology learning for the semantic web. IEEE Intelligent Systems, v. 16, nr. 2, pp. 72-79, 2001.

[27] MITCHELL, T. Machine Learning. McGraw-Hill, 1997.

[28] REINBERGER, M.; SPYNS,P. Unsupervised text mining for the learning of dogma-inspired ontologies. In Ontology Learninf from Text: Methods, Applications and Evaluation, 123 in Frontier in Artificial Intelligence and Applications, IOS Press, 2005.

[29] SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. Automatic text structuring and summarization. Information Processing and Management, v. 33, nr. 2, pp. 193-207, Elsevier, March 1997.

[30] SANDERSON, M.; CROFT,B. Deriving concept hierarchies from text. In Proceedings of the SIGIR Conference on Research and Development in Information Retrieval, pp. 206-213, 1999.

[31] SHUTZE, H.; BUITELAAR, P. RealExt: A tool for relation extraction from text in ontology extension. In Proceedings of the International Semantic Web Conference, pp. 593-606, 2005.

[32] SHUTZE, H. Word space. In Advances in Neural Information Processings Systems 5, pp. 895-902, 1993.

[33] SHAMSFARD, M.; BARFOROUSH, A. Learning ontologies from natural language texts. Human-Computer Studies, v. 60(1), PP. 17-63, 2004

[34] STAAB,S.; ERDMANN, E.; MADCHE, A. Engineering ontologies using semantic patterns. In Proceedings of the IJCAI Workshop on E-Business and Intelligent Web, 2001.

[35] VELARDI, P.; NAVIGLI, R.; CUCHIARELLI, A.; NERI,F. Evaluation onf OntoLearn, a methodology for automatic population of domain ontologies. In Ontology Learning from Text: Methods, Aplications and Evaluation, nr. 123 in Frontiers in Artificial Intelligence an Applications, pp. 92-106, IOS Press, 2005.