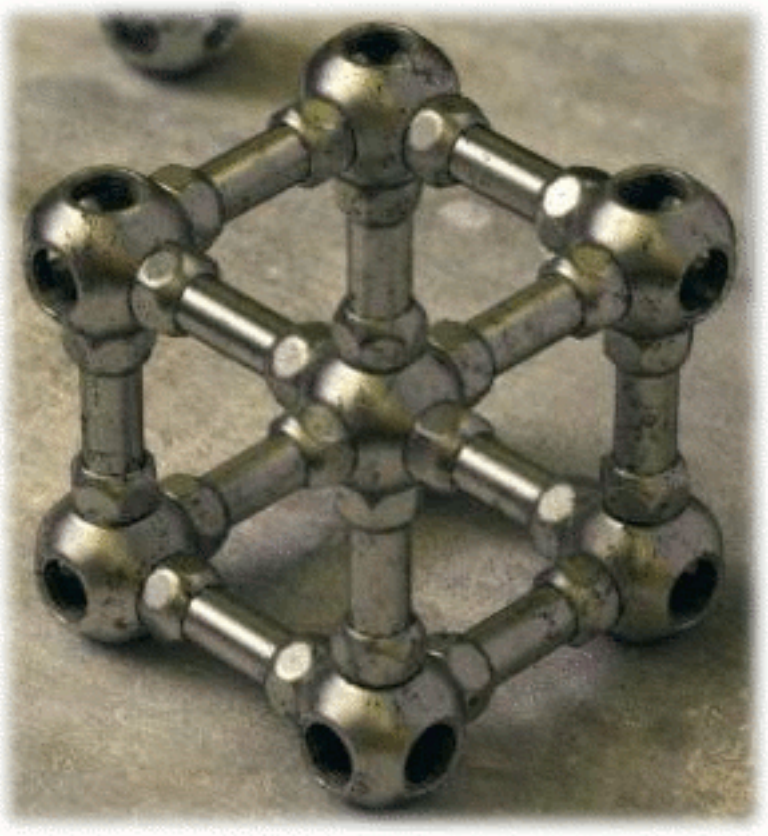


Uma Análise de Técnicas Utilizadas no Aprendizado de Ontologias [1]

Author(s):
Carlos E. Atencio-Torres
Renata Wassermann

[1]This work was supported by Fapesp Project LogProb, grant 2008/03995-5, São Paulo, Brazil.

SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL
3º ONTOBRAS
30 e 31 de agosto de 2010 - Florianópolis/SC

Uma Análise de Técnicas Utilizadas no Aprendizado de Ontologias

Carlos E. Atencio-Torres (USP) carlost@ime.usp.br

Renata Wassermann (USP) renata@ime.usp.br

Resumo: Este artigo apresenta uma análise das técnicas de aprendizado de ontologias. Revisaremos os trabalhos sobre a extração de termos, a construção de taxonomias e apresentaremos os pontos positivos e deficiências dos mesmos.

Palavras-chave: Aprendizado de Ontologias; Extração de Termos; Taxonomia de Conceitos; Descoberta de Relações

1. Introdução

O aprendizado de ontologias (AO) é uma necessidade para os sistemas baseados em ontologias, sendo um desafio para os pesquisadores desta área. No entanto, embora o interesse por este tema tenha crescido consideravelmente, os sistemas atuais ainda necessitam de muitas características exigidas pelos usuários.

Como exemplos de problemas no AO temos: o descobrimento de relações, e a precisão na extração de termos ou na construção de taxonomias. Devido a estes problemas, os sistemas atuais ainda necessitam da supervisão de um especialista. O objetivo deste trabalho é analisar tais problemas para entender melhor a natureza do AO e como poderíamos propor possíveis soluções.

O artigo está dividido de forma a apresentar: na Seção 2, os métodos e ferramentas existentes em AO; nas Seções 3 e 4, uma análise das fases extração de termos e a construção de taxonomias, respectivamente; na Seção 5, a fase de descoberta de relações; e, finalmente, na Seção 6, nossas conclusões e algumas discussões.

2. Aprendizado de Ontologias

O termo Aprendizado de Ontologias, definido por Alexandre Maedche e Steffen Stab (2001), consiste na aquisição de um modelo de domínio a partir de uma fonte de dados. Segundo Drumond e Girardi (2008), existem três tipos de dados (estruturado, semi-estruturado e não estruturado) e um tipo de aprendizado para cada um deles. Nosso

interesse é pelo aprendizado em dados não estruturados, entendido como texto sem formatação.

A metodologia de Cimiano (2006) é usada como padrão no AO e consiste em um conjunto de camadas (ver Figura 1). Primeiramente, os termos de um texto são extraídos. Em seguida, há a fase de aprendizado de conceitos e aprendizado de relações e, finalmente, ocorre a fase de aprendizado de axiomas.

FIGURA 1 – Conjunto de Camadas para AO de acordo com a metodologia de Philipp Cimiano.

Fonte: Cimiano (2006).

O trabalho de Drumond e Girardi (2008) analisa diversas ferramentas para AO como DL Learner [1], Text-2-Onto [2], WebKB [3] e Hasti (2002), as quais possuem características similares como, por exemplo, a procura de termos na etapa inicial e a construção de uma taxonomia de conceitos.

Ribeiro-Júnior (2008) desenvolveu um *plugin* no Protègè para realizar AO a partir de textos escritos em Português. Este *plugin* foi inspirado no trabalho de OntoLT de Buitelaar P. et al. (2003), que considera textos em inglês. No trabalho de Ribeiro-Junior, temos que:

– A entrada consiste em documentos pré-formatados em XCES, que é um formato XML com informação morfosintática.

[1] DLLearner: http://aksw.org/Projects/DLLearner
[2] Text-2-Onto: http://code.google.com/p/text2onto/
[3] WebKB: http://www.cs.cmu.edu/~webkb/

- A extração de termos é feita usando uma seleção de grupos semânticos de cada palavra e, em seguida, aplicando padrões léxico-sintáticos, um *ranking* de termos usando as métricas TF-IDF, C-Values e NC-Values.
- Em cada etapa, o usuário interage com o *plugin* para descartar os grupos semânticos que ele ache desnecessários, assim como os termos que não sejam relevantes para a ontologia.
- A construção da taxonomia usa os padrões léxico-sintáticos de Baségio (2006) que adaptou os padrões no inglês e francês de Hearst e Morim/Jackquemm respectivamente.
- Finalmente exporta-se a hierarquia de conceitos para o Protègè e obtém-se a ontologia.

O trabalho de Ribeiro-Júnior foi importante porque propôs uma ferramenta de AO, especificamente para o português, e mostrou as vantagens de uma técnica híbrida entre padrões léxico-sintáticos e técnicas estatísticas.

3. Análise da Extração de Termos

Um termo, como definido por Pantel P. e Ling D. (2001), é a unidade semântica de uma linguagem. A extração de termo é a base da maioria dos trabalhos da área do AO. Este processo tem como objetivo, reconhecer aquelas palavras que são representativas em um texto. Por exemplo, em um texto sobre religião podemos encontrar palavras simples como *Deus*, *Fé;* e palavras compostas como *Perdão dos Pecados*, *Amor ao Próximo;* que consideraremos como termos.

A maioria de trabalhos usam técnicas estatísticas para obter os termos, por exemplo, o trabalho de Pantel P. e Ling D. (2001) propõe usar uma métrica híbrida baseada no valor de informação mútua junto à verossimilhança logarítmica. Esse trabalho alcançou uma precisão de 74,4% e abrangência de 62,3%.

O fundamento de usar esses dois métodos juntos se deve a que a verossimilhança devolve valores altos para termos que acontecem como erros, como por exemplo “the the” no idioma inglês. Porém a informação mútua diminui a chance deste tipo de expressão, ser escolhidas como termo.

Para melhorar os resultados do Pantel, o trabalho de Tomokiyo e Hurst (2003), propôs o uso de uma fonte externa maior (*background*) para calcular as probabilidades com maior precisão que usando apenas o texto a ser analisado (*foreground*). Para cada fonte foi criado um *modelo de linguagem*, que é uma estrutura de dados para armazenar probabilidades de ocorrências de uma palavra (unigram) ou várias palavras (N-gram) apareceram no texto.

Tomokiyo e Hurst também usaram o ponto de divergência KL (Kullback-Leibler) como a medida de ineficácia por assumir uma certa distribuição de probabilidade *p* quando a verdadeira é *q*. A divergência KL é conhecida também como *entropia relativa*.

Finalmente, o trabalho de Deane (2005) compara diferentes métricas, a MutualRank entre elas, e explica que os métodos estatísticos não tem bons resultados devido à incorreta suposição da distribuição de probabilidade (como por exemplo distribuição normal ou de Poisson), mas que a distribuição Zipf é a mais apropriada porque ela se amolda melhor em distribuições assimétricas (*skewed distribution* em inglês), as quais são mais comuns em ocorrências de palavras.

A conclusão de Deane com respeito à incorreta suposição da distribuição de probabilidade também é defendida pelo trabalho de Duan et al. (2008), que apresentou uma nova proposta para extração de termos baseada em alinhamento de sequências, a mesma que era usada em alinhamento de cadeias de DNA.

As principais vantagens do método de alinhamento são: (i) a capacidade de identificar frases em palavras não continuas e (ii) o pouco rigor estatístico para dar ao sistema flexibilidade no caso de mudar de domínio. Os resultados mostraram maior abrangência, e menor precisão que os métodos estatísticos que usavam *modelos de linguagens*.

4. Análise de Construção de Taxonomias

De acordo com Cimiano (2006), existem três métodos para adquirir uma hierarquia de conceitos:

1. Baseado em Similaridade,
2. Teoría de conjuntos (set-theoretical em inglês), e

3. Clustering.

4.1 Baseados em Similaridade

O trabalho de Hearst (1992) propôs seis padrões léxico-sintáticos para ser aplicados a textos em inglês, no entanto Baségio (2006) fez a adaptação desses seis padrões para a língua portuguesa (ver Tabela 1).

Tabela 1: Padrões de Hearst adaptados para o português. Em que NP : Noun Phrase (Sintagma Nominal) e SUB : Substantivo

	Padrão Original	**Adaptação**
h1	NP such as {(NP,)*(and – or)} NP	• SUB como {(SUB,)*(ou—e)} SUB • SUB tal(is) como {(SUB,)*(ou—e)} SUB
h2	such NP as {(NP,)*{(and – or)} NP	• tal(is) SUB como {(SUB,)*(ou—e)} SUB
h3	NP {,NP}* {,} or other NP	• SUB {,SUB}* {,} ou outro(s) SUB
h4	NP, {NP}* {,} and other NP	• SUB {,SUB}* {,} e outros SUB
h5	NP {,} including {NP.}*{and – or} NP	• SUB {,} incluindo {SUB,}*{ou—e} SUB
h6	NP {,}especially {NP,}*{and – or} NP	• SUB {,} especialmente {SUB,}*{ou—e} SUB • SUB {,} principalmente {SUB,}*{ou—e} SUB • SUB {,} particularmente {SUB,}*{ou—e} SUB • SUB {,} em especial {SUB,}*{ou—e} SUB • SUB {,} em particular {SUB,}*{ou—e} SUB • SUB {,} de maneira especial {SUB,}*{ou—e} SUB • SUB {,} sobretudo {SUB,}*{ou—e} SUB

Fonte: Ribeiro-Júnior (2008).

Por exemplo na sentença “Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use”, depois de aplicar o padrão h1

extraímos a relação: hiponíımia(“Gelidium”, “red algae”), que indica que “red algae” tem como subconceito “Gelidium”.

Para a sentença “... works by such authors as Herrick, Goldsmith and Shakespare”, depois de aplicar o padrão h2, extraímos as relações:

- hiponímia(“author”,“Herrick”);
- hiponímia(“author”,“Goldsmith”); e
- hiponímia(“author”,“Shakespeare”).

que indica que “Herrick”, “Goldsmith” e “Shakespeare” são subconceitos de “author”.

4.2 Teoria de Conjuntos

A *análise de conceitos formais* (ACF) é uma teoria de análise de dados que identifica estruturas conceituais entre conjunto de dados em que se define *objetos formais* e *atributos formais*. A informação de quais atributos pertencem a que objetos é representada por uma função binária chamada *relação de incidência*.

Cimiano descreve um exemplo do domínio das viagens turísticas da seguinte forma. A partir de um texto, uma lista de verbos e os objetos relacionados a esses verbos são extraídos (ver Tabela 2).

Tabela 2: Exemplo de verbos e objetos extraídos de um corpus textual

Verbo	Objetos
reservar	Hotel, apartamento, carro, bicicleta, excursão, viagem.
alugar	Apartamento, carro, bicicleta.
dirigir	Carro, bicicleta.
montar	Bicicleta.
acompanhar	Excursão, viagem.

Fonte: Cimiano (2006).

O seguinte passo consiste em adjetivizar os verbos para representar as características dos objetos, como veremos na Tabela 3.

Tabela 3: Domínio de Viagens Turísticas como um contexto formal

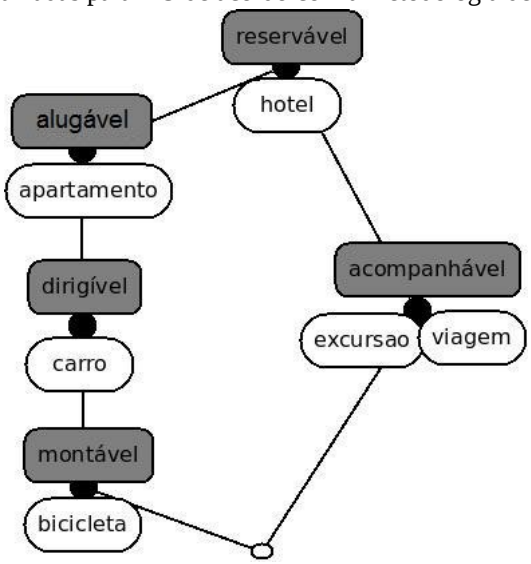
	Reservável	Alugável	Dirigível	Montável	Acompanhável
hotel	X				
apartamento	X	X			

carro	X	X	X		
bicicleta	X	X	X	X	X
viagem	X				X

Fonte: Cimiano (2006).

Após da construção da tabela objeto-atributos, aplica-se o algoritmo Ganter para obter o reticulado na Figura 2.

FIGURA 2 – Conjunto de Camadas para AO de acordo com a metodologia de Philipp Cimiano.



Fonte: Adaptado de Phillip Cimiano (2006).

4.3 Clustering

A suposição básica tomada por esses métodos se baseia na hipótese da distribuição de Harris (1968), que simplesmente indica que palavras similares aparecem em contextos similares.

A maioria das abordagens sobre aprendizado de hierarquias é baseada em paradigmas não-supervisionados, e portanto não são recomendáveis pois existem casos em que esses métodos não conseguem agrupamentos corretos.

5. Descoberta de Relações

A descoberta de relações conceituais é uma etapa pouco desenvolvida, uma vez que o maior interesse da comunidade era melhorar as etapas básicas no AO.

De acordo com Shamsfard e Barforoush (2003), existem dois tipos de relações conceituais:

1. **Taxonômicas**, representam relações de generalização ou especialização. Pertencem a este tipo as relações de hiponímia. De acordo com Sánchez (2007) existem dois métodos para a descoberta de relações deste tipo: (i) usando os padrões de Hearst, os quais vimos na Tabela 1; e (ii) achando padrões do tipo [[ADJETIVO*]SUBSTANTIVO], onde a expressão completa forma a ser apenas uma sub-classe do substantivo, esse enfoque foi abordado por Grefenstette (1997).

2. **Não-Taxonomicas**, são o resto de relações, por exemplo sinonímia (mesmo significado), meronímia (parte de), antonímia, possessão, e outras relações que são aprendidas por algum sistema de aprendizado. Sánchez (2007) indica que a descoberta de relações deste tipo deve se ocupar de duas tarefas:

   - Achar que conceitos estão relacionados e estabelecer uma relação entre eles.
   - Etiquetar tal relação.

O trabalho de Botero (2008) propôs um algoritmo de *clustering* para achar relações semânticas. Um problema deste trabalho é o nomeamento das relações, as quais podiam ter nomes compostos como "*algorithm_genetic*" entre os conceitos *algorithm* e *genetic*. Entre as conclusões, Botero deixa aberto como trabalho futuro a otimização do algoritmo com ajuste de certas variáveis.

6. Considerações Finais

Na Seção 3, vimos que os resultados de Pantel sobre precisão e abrangência, aparentemente são bons, porém só foram aplicados a um só domínio. A melhora de Tomokiyo no trabalho de Pantel não apresentou resultados quantitativos, pois a decisão sobre o que era relevante ou não tornou-se um pouco subjetiva.. A melhora de Deane mostra que a distribuição assimétricas Zipf, junto à métrica MutualRank, obtem melhores resultados. Finalmente, o trabalho de Duan apresenta um método inovador que melhora o trabalho de Deane.

Também vimos na Seção 4, que de acordo com Cimiano, a construção de hierarquias é feita de uma melhor forma com o método de análise de conceitos formais. Porém, para aplicar este método devemos ter uma grande quantidade de relações objeto-atributos. O

trabalho de Cimiano P. et al. (2003) então, descreve como adquirir essa tabela a partir de um grande corpus, usando padrões léxico-sintáticos.

A área de descoberta de relações encontra-se pouco estudada devido ao fato de ter uma base forte de conceitos e relações de taxonomias entre eles. Para tratar esta área, poderíamos supor um domínio fechado e construir padrões léxico sintáticos para descobrir regras específicas neste domínio, tal como foi feito pelo no trabalho de Santos (2002) que definiu regras para um domínio fechado.

Agradecimentos: Este trabalho recebeu apoio da FAPESP, da CAPES e do CNPq.

Referências

BASEGIO, Túlio Lima, **Uma Abordagem Semi-automática para Identificação de Estruturas Ontológicas a partir de Textos na Língua Portuguesa do Brasil.** Dissertação de Mestrado, Pontificia Universidade Católica do Rio Grande do Sul, Rio Grande do Sul, 2006.

BOTERO, Sergio William. **Extração de relações semânticas via análise de correlações de termos em documentos**. Dissertação de mestrado, Universidade Estadual de Campinas, 2008.

BUITELAAR, P.; OLEJNIK, D.; SINTEK, M. **Ontolt: A protégé plug-in for ontology extraction from himnotext**. In: Proceedings of the Demo Session of the International Semantic Web Conference (ISWC). Sanibel Island, Florida, 2003.

CIMIANO, Philipp; STAAB, Steffen; TANE, Julien. **Automatic Acquisition of Taxonomies from Text: FCA meets NLP**. Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia, 2003.

CIMIANO, Philipp. **Ontology Learning and population from text: Algorithms, evaluation and applications**, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

DEANE, Paul. **A nonparametric method for extraction of candidate phrasal terms**. ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (Morristown, NJ, USA), Association for Computational Linguistics, 2005.

DRUMOND, Lucas; GIRARDI, Rosario. **A survey of ontology learning procedures.** WONTO, 2008.

DUAN, Jianyong; LI, Ru; HU, Yi. **A bio-inspired application of natural language processing: A case study in extracting multiword expression**. Expert Syst. Appl., 2009.

GREFENSTETTE, Gregory. **SQLET: Short Query Linguistic Expansion Techniques: Palliating One or Two-word queries by providing intermediate structure to WWW pages**. In Proceedings of RIAO '97, 1997.

HARRIS, Z.S. **Mathematical structures of language, Wiley**, 1968.

HEARST, M. A. **Automatic acquisition of hyponyms from large text corpora**. In Proceedings of the 14th International Conference on Computational Linguistics, 1992.

RIBERO-JUNIOR, Luiz Carlos. **OntoLP: Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa**. Dissertação de Mestrado, Universidade do Vale do Rio dos Sinos, São Leopoldo, Brasil, 2008.

MAEDCHE, Alexandre; STAAB, Steffen. **Ontology Learning for the semantic web**. IEEE Intelligent Systems, 2001.

PANTEL, Patrick; LING, Dekang. **A statistical corpus-based term extractor**. AI '01: Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence (London, UK), Springer-Verlag, 2001.

SANCHEZ, David. **Domain Ontology Learning from the Web**. Tesi Doctoral, Universitat Politécnica de Cataluya, Tarragona, Espanha, 2007.

SANTOS, Maria A. M. R dos. **Extraindo Regras de Associação a Partir de Textos**. Dissertação de Mestrado, Pontifícia Universidade Católica do Paraná, 2002.

SHAMSFARD, M.; BARFOROUSH, A. **An introduction to hasti: An ontology learning system**. In; Proceedings of 6th Conference on Artificial Intelligence and Soft Computing (ASC), 2002.

SHAMSFARD, Mehrnoush; BARFOROUSH, Ahmad Abdollahzadeh. **The state of the art in ontology learning: a framework for comparison**, Knowl. Eng. Rev. 18 (2003).

TOMOKIYO, Takashi; HURST, Matthew. **A language model approach to keyphrase extraction**. Proceedings of the ACL 2003 workshop on Multiword expressions (Morristown, NJ, USA),Association for Computational Linguistics, 2003.