Construção de uma Estrutura Ontológica de Domínio a partir da Wikipédia

Clarissa Castellã Xavier[1*], Vera Lúcia Strube de Lima[1]

[1] Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS) – Porto Alegre – RS – Brasil

{clarissa.xavier,vera.strube}@pucrs.br

Abstract. *Data extraction from Wikipedia for ontologies construction, enrichment and population is an emerging research field. This paper describes a study on automatic extraction of an ontological structure containing hyponymy and location relations from Wikipedia's Tourism category in Portuguese, illustrated with an experiment, and evaluation of its results.*

Resumo. *A extração de dados da Wikipédia para construção, enriquecimento e população de ontologias se mostra um campo de pesquisa emergente. Este artigo relata um estudo sobre a extração de uma estrutura ontológica contendo relações de hiponímia e localização, da categoria Turismo da Wikipédia em português, ilustrado por um experimento, e avalia os resultados obtidos.*

1. Introdução

Embora tenham origem na Filosofia, as ontologias são empregadas na Ciência da Computação com frequência crescente, sujeitas às próprias adaptações da área. Encontram-se para ontologia desde as definições mais simples – "uma ontologia pode ser um documento ou arquivo que define formalmente as relações entre termos mais gerais e termos mais específicos" [Souza e Alvarenga 2004] - até as mais complexas como aquelas de Gruber, Guarino ou Smith e Welty [Gruber 1993][Guarino 1998][Smith e Welty 2001], que relacionam classes, propriedades, instâncias, axiomas e lógicas para construir as estruturas ontológicas[1] mas que, mesmo assim, também abrigam em seu bojo, enquanto ontologias, as estruturas terminológicas mais simples.

A construção de ontologias [Maedche 2002] é um processo oneroso, tedioso e propenso a erros, e o número de ontologias de domínio disponíveis na atualidade ainda é extremamente pequeno [Hepp et al. 2006], cenário que recrudesce ainda mais no campo das ontologias em língua portuguesa [Lima et al. 2007]. Ao se buscar fontes de dados alternativas para construção ou enriquecimento de ontologias, uma opção é a

---

* Bolsista CNPq.

[1] Utilizaremos os termos "ontologia", "estrutura taxonômica" e "estrutura ontológica" intercambiadamente, e adotaremos, para ontologia, a abordagem mais aberta, que pode remeter a uma terminologia, dotada de relações semânticas simples.

Wikipédia[2], que em maio de 2009 contava com mais de 473.000 artigos em língua portuguesa. Os documentos da Wikipédia estão organizados em uma hierarquia de categorias e esta, com suas subcategorias, pode ser entendida como uma estrutura de termos, embora não seja estritamente uma estrutura arbórea, e sim uma representação mais rica construída colaborativamente. Essa estrutura permite múltipla categorização simultânea de tópicos, ou seja, algumas categorias podem ter mais de uma supercategoria [Syed et al. 2008], constituindo-se em um grafo que representa uma rede conceitual com relações semânticas não especificadas [Strube e Ponzetto 2006].

Concentrando nosso estudo no domínio Turismo da Wikipédia, analisamos os títulos e a organização de suas subcategorias. A escolha pelo domínio se deu em vista de um possível acoplamento futuro aos resultados relatados em [Baségio 2007], que realiza a extração semi-automática de uma estrutura ontológica a partir de textos jornalísticos desse domínio.

Verificamos que o emprego exclusivo da relação de hiponímia (*is-a*) é muito limitado para a construção da estrutura que pode ser extraída, e incluímos o uso da relação *located-in,* essencial para descrever com mais exatidão as ligações semânticas existentes, especialmente no domínio estudado. Por exemplo, algumas categorias apresentam em seu título um relacionamento explícito, tal como "Jardins zoológicos da Alemanha", o que pode ser melhor representado pela relação "Jardins zoológicos *located-in* Alemanha".

Para estudar a aplicabilidade dessa estratégia, realizamos um experimento cujo objetivo é extrair uma estrutura ontológica das subcategorias da categoria Turismo da Wikipédia em língua portuguesa, contendo relações do tipo hiponímia (*is-a*) e também localização (*located-in*) (Seção 3). Analisamos nossos resultados através de métricas que comparam a estrutura obtida com uma estrutura de referência mapeada manualmente (Seção 4), e discutimos esses resultados em vista de acertos e erros, bem como melhorias a serem inseridas (Seção 5). Na Seção 6 apresentamos nossas conclusões. Os resultados obtidos mostram que a alternativa de uso das categorias da Wikipédia, junto com relações do tipo *located-in*, é promissora.

Na seção que segue, reunimos comentários sobre trabalhos relacionados ao nosso.

2. Trabalhos Relacionados

Destacamos a seguir alguns dos trabalhos que ofereceram subsídios para o presente estudo, com as respectivas contribuições.

[Ponzetto e Strube 2007a] e [Ponzetto e Strube 2007b] descrevem experimentos em que é realizada a extração de taxonomias da Wikipédia em língua inglesa. [Mika et al. 2008] propõe uma solução baseada em anotação semântica da Wikipédia, apresentando um método que cria um mapa semântico com vocabulário de duas fontes: a Wikipédia e um corpus previamente anotado.

---

2 Enciclopédia disponível livremente pela internet, criada por Jimbo Wales e Larry Sanger em janeiro/2001, desenvolvida por uma comunidade de usuários que cresce exponencialmente com a adição constante de conteúdo por seus colaboradores em todo o planeta [Syed et al. 2008]. Acessível em http://www.wikipedia.org/

[Suchanek et al. 2008] apresenta uma metodologia baseada em heurísticas que, auxiliada por técnicas de Processamento da Língua Natural (PLN), extrai uma ontologia baseada no modelo *infobox*[3] da Wikipédia e conceitos da WordNet, cujas regras para extração de dados podem ser adaptadas para a extração de relações e conceitos da Wikipédia em língua portuguesa.

[Wu e Weld 2007] e [Wu e Weld 2008] propõem um sistema que gera e complementa os modelos *infoboxes* e cria uma ontologia com dados obtidos da WordNet e dos *infoboxes*. A técnica automática para geração deste modelo pode ser útil, pois ele é pouco utilizado na versão em língua portuguesa da enciclopédia.

De posse dos fundamentos e contribuições encontradas na literatura, organizamos nosso estudo e experimento.

3. Estudo e Experimento

Com o objetivo de futuramente enriquecer uma ontologia do domínio Turismo, nosso estudo visa criar, a partir da estrutura de categorias e subcategorias da Wikipédia em língua portuguesa, uma estrutura ontológica contendo relações do tipo hiponímia (*is-a)* e localização (*located-in*). Para avaliar a viabilidade da criação dessa estrutura, foi desenhado um experimento. Para apoiar o experimento, foi elaborado um protótipo implementado em PHP[4], acessando banco de dados MySql[5] e gerando uma estrutura ontológica descrita em OWL[6].

O experimento foi realizado em três etapas, vide Figura 1. Na primeira etapa, a estrutura taxonômica da categoria Turismo do banco de dados da Wikipédia é extraída. Em seguida, são obtidas as relações "*located-in*" encontradas na taxonomia. Na terceira etapa executamos a remoção dos caracteres especiais dos títulos das categorias selecionadas, a conversão de todos os caracteres para letras minúsculas, a criação da descrição OWL da estrutura taxonômica obtida e a geração do arquivo com a estrutura criada.

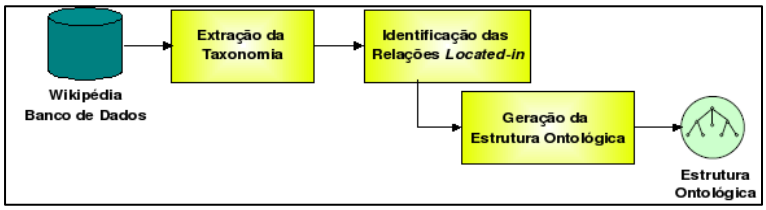

Figura 1 – Estrutura do Protótipo

Nas subseções a seguir, oferecemos a descrição mais detalhada das duas primeiras etapas.

---

[3] *Infobox* é um modelo padrão da Wikipédia que contém uma tabela que apresenta informações básicas sobre a entidade descrita no artigo em que está inserido. Por exemplo, *infoboxes* de artigos que descrevem países, costumam conter informações como o nome do país na língua nativa, sua capital e área [Suchanek et al. 2008].

[4] www.php.net

[5] Para a execução deste experimento foi utilizada uma imagem do banco de dados da Wikipédia em português de 05 de janeiro de 2009.

[6] Web Ontology Language. Disponível em http://www.w3.org/TR/owl-features/

3.1. Extração da estrutura taxonômica

Na primeira fase do experimento realizamos a seleção no banco de dados das subcategorias da categoria Turismo, em 3 níveis de profundidade (Figura 2).

O corpus da Wikipédia abrange diferentes campos do conhecimento e a organização do seu grafo de categorias viabiliza a ligação de conceitos que pertencem a domínios distintos. Realizamos uma análise do grafo das subcategorias de turismo e decidimos fixar a profundidade de nossa busca em três níveis, buscando obter o maior número de conceitos possível, sem extrapolar o domínio escolhido.

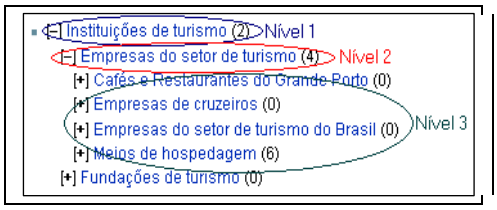

Figura 2 – Três níveis de subcategorização (categoria Turismo)

A versão em língua portuguesa da Wikipédia não usa uma versão específica da língua comum[7], motivo pelo qual efetuamos a unificação da ortografia dos títulos de categorias, definindo a ortografia brasileira como padrão. Por exemplo, em títulos contendo as palavras "atracções" e "atrações", substituímos as ocorrências de "atracções" por "atrações".

Ao final desta etapa do experimento, obtivemos uma taxonomia onde os conceitos são os títulos das categorias selecionadas e a relação hierárquica é estabelecida pela maneira com a qual a estrutura de categorias foi organizada no banco de dados da Wikipédia.

3.2. Extração de relações de localização (*located-in*)

Para obter as relações *located-in* da taxonomia extraída, implementamos duas heurísticas para inferir o relacionamento de localização a partir dos títulos das subcategorias de Turismo:

- *Heurística 1* - inferindo relacionamentos de localização em subcategorias das categorias cujo título contém "por país", "por cidade" ou "por estado", como por exemplo, "Atrações turísticas por cidade".

Para cada categoria da taxonomia, verificamos se seu título contém as expressões "por país", "por cidade" ou "por estado". Em caso positivo, inferimos que todas as suas subcategorias possuem relação *located-in* com a primeira. Neste caso, buscamos o local ao qual as subcategorias se referem, renomeamos a subcategoria para o nome deste local, e implantamos a relação *located-in.*

Exemplificando: "Atrações turísticas de Curitiba" é subcategoria de "Atrações turísticas por cidade". Para criar a relação "Atrações turísticas por cidade" *located-in* "Curitiba", selecionamos todas as categorias relacionadas com "Atrações Turísticas de Curitiba" e aplicamos a regra: *o nome da localidade (Curitiba) é a categoria*

---

7 http://pt.wikipedia.org/wiki/Wikipedia:Livro_de_estilo

selecionada cujo título possui o menor tamanho e esteja contido dentro de "Atrações turísticas de Curitiba". Ilustramos o exemplo na Figura 3.

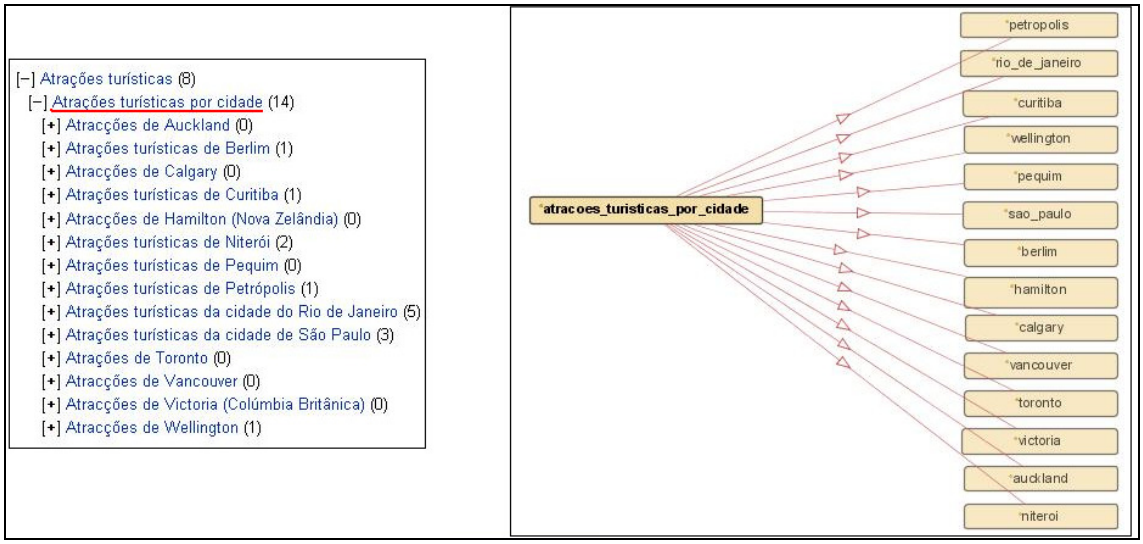

Figura 3 – Recorte da categoria "Atrações turísticas por cidade" na Wikipédia e a representação das relações extraídas através da Heurística 1

- *Heurística 2* - inferindo relacionamentos de localização em categorias contendo preposições ou contrações "de/do/da" e "em/no/na" em seu título, como por exemplo, "Aeroportos da Argentina".

Para cada categoria da taxonomia, testamos se seu título contém as preposições ou contrações "em/no/na" ou "de/do/da". Caso positivo, iremos buscar se a palavra ao lado da preposição ou contração refere-se a um local e, neste caso, inferimos que existe o relacionamento *located-in*.

Exemplificando: a categoria "Aeroportos da Argentina" possui a contração "da" em seu título e por isso inferimos ser candidata à inserção da relação *located-in* na estrutura ontológica. Para criar a relação "Aeroportos *located-in* Argentina" seguimos os seguintes passos: selecionamos todas as categorias relacionadas com "Aeroportos da Argentina". Verificamos se alguma das categorias ligadas a "Aeroportos da Argentina" possui ligação com cidade, estado ou país. Se possuir, concluímos que contém uma localidade em seu título e aplicamos as seguintes regras: *o nome da localidade é a categoria cujo título possui o menor tamanho dentre todas as selecionadas e está contido dentro de "Aeroportos da Argentina"* (neste caso, a categoria selecionada é "Argentina"); e a regra: *a classe que terá a relação de localização com a localidade é a parte inicial do título, anterior ao nome da localidade e à preposição ou contração* (neste caso "Aeroportos"). Criamos a relação "Aeroportos *located-in* Argentina". Ilustramos o exemplo na Figura 4.

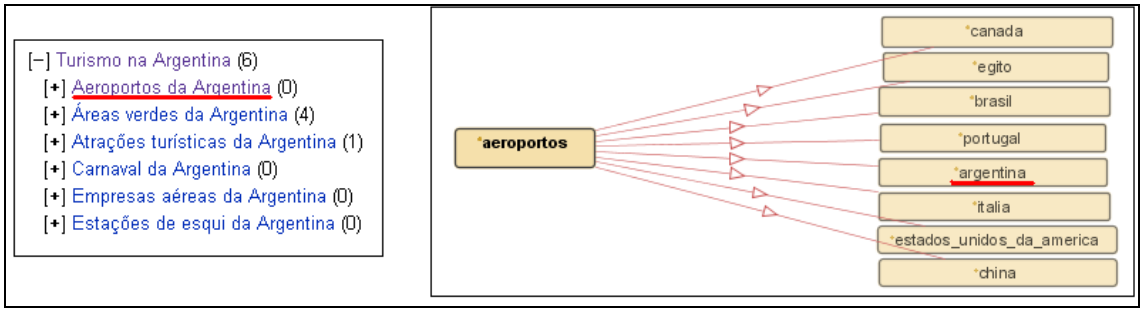

Figura 4 – Recorte da categoria "Aeroportos da Argentina" na Wikipédia e a representação da relação extraída através da Heurística 2

4. Análise dos Resultados

Para analisar o experimento, procuramos aplicar métricas para avaliar a taxonomia resultante, e as relações *located-in.*

4.1. Taxonomia

Para melhor analisar a taxonomia gerada, apresentamos a seguir as 6 classes com o maior número de subclasses (Tabela 1) e com base na bibliografia recente na área de avaliação de ontologias, as seguintes métricas:

$$\mathbf{iCnt}(\boldsymbol{C})^{8}: 509 \qquad \boldsymbol{SD}^{9}: 1{,}25$$

Tabela 1. Classes com maior número de subclasses na estrutura extraída

Classe	*NroSubclasses*	*Classe*	*NroSubclasses*
Transportes por país	79	Atrações turísticas por país	45
Turismo por país	63	Meios de hospedagem	24
Parques nacionais	48	Transporte	24

4.2. Relações *Located-in*

Para avaliar os resultados obtidos no que se refere a estas relações, utilizamos Precisão e Abrangência, medidas baseadas na comparação entre resultados esperados e resultados obtidos usuais em extração da informação, e adaptadas para ontologias [Euzenat 2007]. No caso, os conjuntos de documentos são substituídos por conjuntos de relações correspondentes. O alinhamento (*A*) retornado pelo sistema que está sendo avaliado é comparado ao alinhamento de referência (*R*). A Precisão mede a taxa de correspondência correta em relação ao total de correspondências obtidas, enquanto que a Abrangência mede o número de correspondências obtidas em relação ao total de correspondências esperadas. A Figura 5 representa estes conceitos tal como a fonte citada. Também é empregada a medida-F (média harmônica entre Precisão e Abrangência).

[8] Número de classes definidas (*number of classes defined*) [Cross e Pal 2005]

[9] Profundidade do esquema (*schema deepness*). É dada pela média do número de subclasses por classe [Tartir e Arpinar 2007]

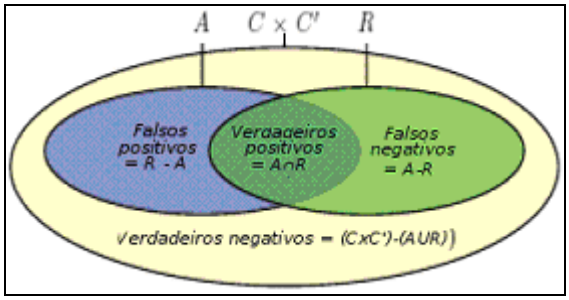

Figura 5 – Representação das medidas de Precisão e Abrangência [Euzenat 2007]

Para avaliar a qualidade da estrutura ontológica gerada pelo experimento, realizamos o mapeamento manual das relações de localização na taxonomia da categoria Turismo. A estrutura ontológica resultante deste mapeamento foi utilizada como referência e a Tabela 2 apresenta a aferição das métricas apresentadas, na estrutura ontológica extraída.

Tabela 2. Avaliação do mapeamento das relações *located-in*

Métrica	***Resultado***
Precisão	0.95
Abrangência	0.607305936073
Medida-F	0.740947075209

5. Considerações sobre os resultados

A partir da medida de Precisão obtida (Tabela 2), procuramos identificar quais os relacionamentos de localização mapeados na estrutura ontológica de referência e não mapeados pelo protótipo, e investigar os motivos das falhas.

Observa-se que apenas 3 relações mapeadas na estrutura ontológica não foram encontradas na referência (falsos positivos), todas relacionadas à categoria "Transportes por País". As categorias "Transporte ferroviário por país", "Transporte hidroviário por país" e "Transporte rodoviário por país" são subcategorias de "Transportes por País", mas não contém lugares em seu título, por isso não cabendo a inserção do relacionamento *located-in*. Esta falha deve-se a um problema na heurística descrita na Seção 3.2.1. A regra infere que todas as subcategorias de categorias que contenham "por país" em seu título apontam para lugares, cabendo a inserção do relacionamento "*located-in*", o que não se aplica neste caso descrito.

Ao analisar os resultados da Abrangência verificamos que, no total, 138 relacionamentos de localização mapeados na estrutura de referência não foram mapeados na estrutura gerada pelo protótipo. A principal falha deve-se à ausência de ligações entre as categorias cujo relacionamento de localização deveria ser mapeado na regra descrita na Seção 3.2.2 (títulos contendo "em/no/na", "de/do/da") com outras categorias ligadas à localização no banco de dados. Por exemplo, as categorias "Parques Nacionais da África do Sul" e "Estações de esqui da Argentina" deveriam ser mapeadas como "Parques Nacionais *located-in* África do Sul" e "Estações de Esqui *located-in* Argentina", conforme a heurística descrita. Todavia, no cadastro das categorias na Wikipédia, elas não estão ligadas a nenhuma categoria de sua localização, e por isso a falha na regra proposta.

6. Considerações Finais

Ao estudarmos a categoria Turismo da Wikipédia, com o objetivo de extrair uma estrutura ontológica de domínio, verificamos que a relação de hiponímia (*is-a*) é muito limitada para descrever o domínio Turismo tal como lá contido. Acreditamos que o uso da relação *located-in* é essencial para descrever com mais exatidão as ligações semânticas existentes.

Este artigo relata um estudo realizando a extração de uma estrutura ontológica contendo relações de hiponímia e localização, da categoria Turismo da Wikipédia em português, ilustrado por um experimento, e avalia os resultados obtidos. A partir das considerações acerca dos resultados obtidos, podemos propor um refinamento das heurísticas de extração da relação *located-in*, através das seguintes ações:

- Alteração da Heurística 1, inserindo a análise dos títulos que a regra infere (em vista de possuírem um local, assim como já é feito na heurística descrita na Seção 3.2.2).

- Alteração na Heurística 2, de modo a procurar se a palavra ao lado da preposição é o título de uma categoria que possui marcador adequado (município, cidade, estado ou país), ao invés de apenas buscar se as categorias ligadas à categoria em análise.

Ao que conhecemos, mesmo que ainda contenha pontos a serem aprimorados, trata-se aqui de uma iniciativa pioneira envolvendo a construção de uma estrutura ontológica com as categorias da Wikipédia em língua portuguesa, recurso que pode ser útil à construção e enriquecimento de ontologias, ainda bastante escassas, em especial em português. A avaliação de ontologias, tema recente e em consolidação, também permitirá comparações com outras ontologias extraídas da Wikipédia, nesta ou em outras línguas, bem como com outras ontologias do mesmo domínio (o que ainda e mostra difícil de ser realizado). Como sequência a este estudo, temos por objeto a análise quanto à possibilidade de extração de outras relações, bem como o uso da estrutura obtida no enriquecimento de ontologias do mesmo domínio.

Referências

[Baségio 2007] Túlio Lima Baségio (2007). Uma abordagem semi-automática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do brasil. Tese de mestrado, PUCRS, Fac. de Informática.

[Cross e Pal 2005] V. Cross e A. Pal (2005). Metrics for ontologies. Em *Fuzzy Information Processing Society, 2005*. NAFIPS 2005. Annual Meeting of the North American, páginas 448- 453.

[Gruber 1993] Thomas R. Gruber (1993). A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199-220.

[Guarino 1998] N. Guarino (1998). *Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*. IOS Press, Amsterdam, The Netherlands, The Netherlands.

[Euzenat 2007] Jérôme Euzenat (2007). Semantic precision and recall for ontology alignment evaluation. Em *Proc.-20th International Joint Conference on Artificial Intelligence (IJCAI)*, páginas 348-353, Hyderabad (IN).

[Hepp et al. 2006] M. Hepp, D. Bachlechner e K. Siorpaes (2006). Harvesting wiki consensus - using wikipedia entries as ontology elements.

[Lima et al. 2007] Vera L. Lima, Marias Nunes e Renata Vieira (2007). Desafios do processamento de línguas naturais. Em *Anais do XXVII Congresso da SBC*, páginas 2202-2216. SBC, SBC.

[Maedche 2002] Alexander Maedche (2002). *Ontology Learning for the Semantic Web* (The Kluwer International Series in Engineering and Computer Science, Volume 665). Springer.

[Mika et al. 2008] Peter Mika, Massimiliano Ciaramita, Hugo Zaragoza e Jordi Atserias (2008). Learning to tag and tagging to learn: A case study on wikipedia. *IEEE Intelligent Systems*, 23(5):26-33.

[Ponzetto e Strube 2007a] Simone P. Ponzetto e Michael Strube (2007a). Deriving a large scale taxonomy from wikipedia.

[Ponzetto e Strube 2007b] Simone P. Ponzetto e Michael Strube (2007b). Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artifcial Intelligence Research*, 30:181-212.

[Smith e Welty 2001] Barry Smith e Christopher Welty (2001). Fois introduction: Ontology-towards a new synthesis. Em *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems*, New York, NY, USA. ACM Press.

[Souza e Alvarenga 2004] Renato Souza e Lídia Alvarenga (2004). A web semântica e suas contribuições para a ciência da informação. *Ciência da Informação*, 33(1).

[Spinellis e Louridas 2008] Diomidis Spinellis e Panagiotis Louridas (2008). The collaborative organization of knowledge. *Commun. ACM*, 51(8):68-73.

[Strube e Ponzetto 2006] Michael Strube e Simone P. Ponzetto (2006) "WikiRelate! Computing semantic relatedness using Wikipedia". In *Proc. of AAAI-06,* p.1419-1424.

[Suchanek et al. 2008] Fabian M. Suchanek, Gjergji Kasneci e Gerhard Weikum (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semant.*, 6(3):203-217.

[Syed et al. 2008] Zareen Syed, Tim Finin e Anupam Joshi (2008). Wikipedia as an Ontology for Describing Documents. Em *Proceedings of the Second International Conference on Weblogs and Social Media*. AAAI Press.

[Tartir e Arpinar 2007] Samir Tartir e I. Budak Arpinar (2007). Ontology evaluation and ranking using ontoqa. Em *ICSC*, páginas 185-192. IEEE Computer Society.

[Völkel et al. 2006] Max Völkel, Markus Krötzsch, Denny Vrandecic, Heiko Haller e Rudi Studer (2006). Semantic wikipedia. Em *WWW '06: Proceedings of the 15th international conference on World Wide Web*, páginas 585-594, New York, NY, USA. ACM.

[Wu e Weld 2007] Fei Wu e Daniel S. Weld (2007). Autonomously semantifying wikipedia. Em *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, páginas 41-50, New York, NY, USA. ACM.

[Wu e Weld 2008] Fei Wu e Daniel S. Weld (2008). Automatically refining the wikipedia infobox ontology. Em *WWW '08: Proceeding of the 17th international conference on World Wide Web*, páginas 635-644, New York, NY, USA. ACM.