Uma Metodologia para a Extração de Ontologias a partir de Páginas *Web*

Lucas Pupulin Nanni e Sérgio Roberto Pereira da Silva
UEM - Universidade Estadual de Maringá
Grupo de Sistemas Interativos Inteligentes
Av. Colombo, 5790. Jardim Universitário.
CEP 87020-900 Maringá, PR.

{lucasnanni, sergio.r.dasilva}@gmail.com

Resumo. *A dificuldade de recuperação de informação relevante e de qualidade na Web vem se agravando muito nos últimos anos com o aumento vertiginoso da disponibilidade de conteúdos. Grande parte desta dificuldade advém da atual impossibilidade de identificar o contexto da pesquisa que o usuário está realizando. Este trabalho aborda uma metodologia para extração de ontologias de páginas Web que possam ser aplicadas na representação do modelo do usuário e do contexto de busca para apoiar um sistema de recomendação para busca na Web.*

1. Introdução

O crescimento descontrolado da informação disponível na *Web* causa um fenômeno denominado de "sobrecarga de informação" [1, 2]. Um dos mais visíveis problemas desta sobrecarga é o esforço necessário para se encontrar uma informação desejada na *Web*. Esta sobrecarga de informação, aliada a ausência de mecanismos para garantir a qualidade da informação recuperada, torna o processo de recuperação de informação na *Web* complexo e de baixa qualidade. Nos últimos anos, a atenção tem sido voltada para sistemas de recuperação de informação adaptativos [3, 4, 5] que consideram modelos dos interesses do usuário e do contexto de pesquisa, rearranjando os resultados da pesquisa em função da relevância de cada documento aos interesses do usuário.

Este trabalho está imerso no projeto SARI*Web (http://www.din.uem.br/gsii)*, o qual visa a especificação e desenvolvimento de um sistema adaptativo de recuperação de informação para a *Web* que considere o contexto de interesse do usuário, a relevância e a qualidade da informação resultante do processo de recuperação. A intenção de considerar o contexto de interesse do usuário implica em determiná-lo, e para tal sugerimos que ontologias que o modelem sejam extraídas a partir das páginas *Web* referenciadas pelo usuário. Além de estabelecer uma metodologia de extração e emersão de ontologias a partir de páginas *Web* escritas em língua inglesa, o objetivo deste trabalho também considera a criação de um extrator automático de ontologias capaz de ser integrado ao projeto SARI*Web*.

2. Ferramentas, metodologias e técnicas

Para a realização deste trabalho foram utilizadas como ferramentas de desenvolvimento a linguagem de programação *Python (http://www.python.org)*, a ferramenta de processamento de linguagem natural *NLTK (http://www.nltk.org)*, o processador de *HTML Beautiful Soup (http://www.crummy.com/software/BeautifulSoup)*, a ferramenta de detecção e extração de texto estruturado *Tika (http://tika.apache.org)* e a base de dados léxicos *WordNet (http://wordnet.princeton.edu)*.

O processo foi dividido entre as etapas de extração de texto e de emersão da ontologia. Cada etapa resultou em um módulo específico implementado em *Python*.

2.1. Módulo de Extração de Texto

Para a extração de texto das páginas *Web* foram consideradas duas situações. A primeira, quando o conteúdo a ser extraído é detectado como *HTML*, o módulo responsável pela extração decide pela utilização da ferramenta *Beautiful Soup* para o processamento e extração do texto estruturado da página *Web*, valendo-se das marcações `<p>`, `<h1>`, `<h2>`, `<h3>`, `<h4>`, `<h5>`, `<h6>` e `<title>`. A segunda, quando o conteúdo é detectado como uma formatação diferente à *HTML*, a extração é realizada pela ferramenta *Tika*, a qual detecta o formato correspondente e extrai o texto estruturado de forma adequada.

Após a extração, o texto é fragmentado em sentenças e *tokens* com o auxílio de expressões regulares. Os *tokens* são normalizados, filtrados e classificados a fim de definir um conjunto formado apenas de substantivos ordenados por sua frequência. Para fins de melhora de desempenho computacional, o conjunto é dividido segundo a proporção 20/80. A escolha desta proporção é resultante do uso da combinação da lei de Zipf com o princípio de Pareto [6], conforme detalhado em [7]. Finalmente, *synsets* [8] são atribuídos aos *tokens* pertencentes à fração dos 20% mais frequentes, tornando-os *tags* semânticas prontas para a emersão da ontologia.

No processo de atribuição de significado aos *tokens*, foi observado que grande parte dos termos pesquisados na *WordNet* eram ambíguos, *i.e.*, possuíam um conjunto de *synsets* relacionados a eles. Isto exigiu que um processo de desambiguação fosse aplicado, permitindo a cada *token* ser representado por um único *synset*. Deste modo, no processo de atribuição de sentidos foi empregada a métrica *lesk* adaptada para a *Web* [9], considerando a sentença de ocorrência de cada *token*. A Figura 1 ilustra essa situação.

A desambiguação restringida às sentenças em que o *token* ocorre retorna resultados mais precisos, uma vez que em uma sentença os *tokens* possuem uma relação natural, a qual proporciona *synsets* mais contextualizados.

2.2. Módulo de Emersão de Ontologias

A emersão das ontologias, isto é, o processo de criação de uma estrutura ontológica para o conjunto de *tags*, se baseou no processo de estruturação semântica proposta por [10],

porém mais simplificado e considerando como base um conjunto de *tags* distintas e propriamente conceituadas.

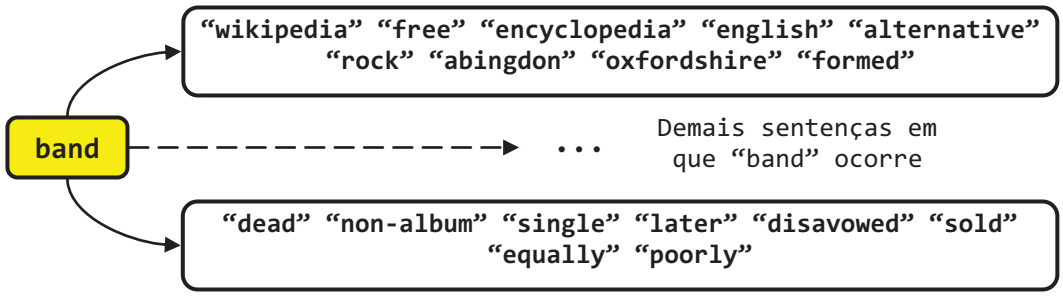

Figura 1. *Token* "band" e suas respectivas sentenças de ocorrência.

Para cada *tag* foi construída uma estrutura que a conectava à raiz da *WordNet*, como ilustrada pela Figura 2. As relações consideradas foram a de hiperônimo *(is-a)* e a de hipônimo *(kind-of)*, presentes na rede *WordNet*. Entretanto, a metodologia proposta pode ser facilmente estendida com as relações de homônimos e merônimos, ampliando a representatividade da ontologia emersa. A esta *tag*, associada à estrutura que a liga as demais *tags* do sistema, é dado o nome de *tag* semântica.

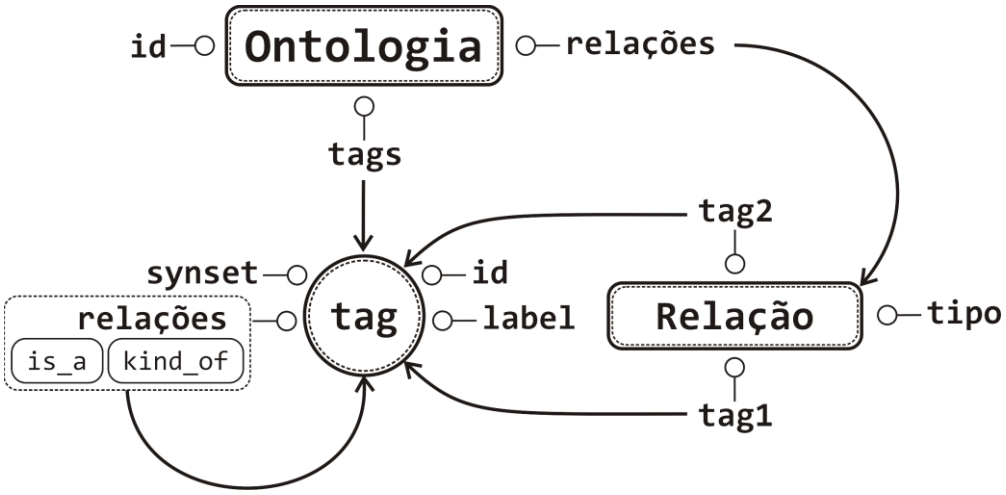

Figura 2. Estrutura básica das ontologias emersas.

As relações entre conceitos (*tags*) foram criadas de forma sequencial, "*tag* a *tag*", exigindo que a unicidade entre os elementos internos da estrutura ontológica fosse garantida antes das *tags* serem adicionadas à ontologia. Para a construção do "caminho" entre a *tag* e a raiz `entity` foi utilizado, além da interface à *WordNet* fornecida pela *NLTK*, o método `hypernym_paths` implementado pela mesma ferramenta. A Figura 3 exemplifica a estrutura das relações criadas para a *tag "performance"*.

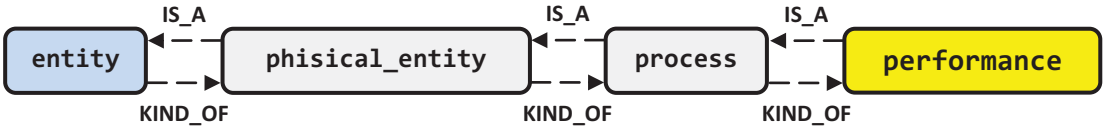

Figura 3. Estrutura de relações para a *tag "performance".*

3. Avaliação da Ontologia Emersa

A metodologia de emersão proposta neste trabalho produz uma ontologia "leve" [11] e simplificada o suficiente para que seja suportada pela *"Web Ontology Language Lite" (OWL-Lite)*, permitindo sua persistência e compartilhamento em sistemas similares ao que integra este trabalho. A Figura 4 ilustra um fragmento do grafo gerado para a ontologia extraída do recurso *Web http://en.wikipedia.org/wiki/California*.

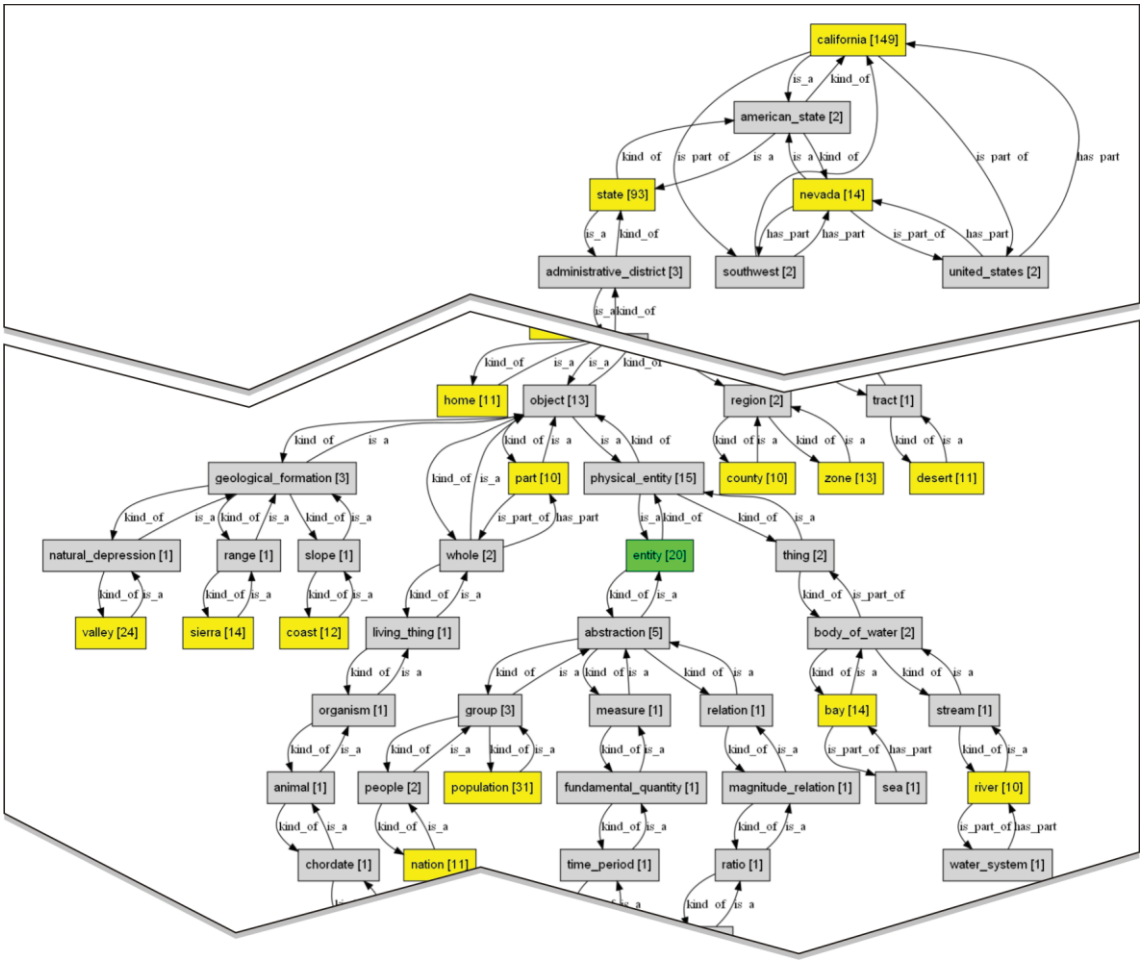

Figura 4. Fragmento da representação da ontologia emersa para o recurso *http://en.wikipedia.org/wiki/California*.

Além disso, o desempenho computacional do sistema proposto ficou em níveis satisfatórios com a simplificação empregada, reduzindo o conjunto de termos extraídos em 80% e utilizando as relações sugeridas. A Tabela 1 relaciona as médias ponderadas

de tempo experimentadas na aplicação do processo em 500 amostras de artigos aleatórios da *Wikipedia (http://en.wikipedia.org/wiki/Special:Random)*. Os tempos foram ponderados em relação à quantia de *tags* consideradas na extração de cada artigo.

É importante frisar que os tempos foram experimentados considerando a execução do protótipo pelo interpretador *Python 2.6.6* de 32 bits, sobre o sistema operacional *Windows 7*, também de 32 bits. A máquina utilizada possuía ainda um núcleo de processamento *Intel Core 2 Duo*, de 64 bits e 2,40 GHz de frequência, além de 3,0 GB de memória *RAM*, padrão *DDR3*. Essa descrição corrobora com o fato de que os testes foram realizados sobre uma máquina pessoal de pequeno porte, justificando a satisfabilidade do desempenho.

Etapa do processo	**Média ponderada do tempo (*s*)**	**Intervalo (*s*)**
Extração de Termos	1,699	[1,310; 6,661]
Filtragem dos Termos	0,468	[0,016; 19,250]
Atribuição de Sentidos	0,787	[0,062; 14,196]
Emersão da Ontologia	0,012	[0,000*; 0,702]
Tempo Total	2,965	[1,419; 25,678]

*valor experimentado por falta de precisão numérica computacional.

Tabela 1. Tempos experimentados nas etapas do processo para 500 amostras aleatórias da Wikipédia.

No entanto, não foi possível avaliar a qualidade da ontologia resultante, já que esta característica está diretamente relacionada ao seu fim. Os aspectos que influenciam a emersão de uma ontologia estão presentes na seleção do texto do qual ela será extraída, na escolha dos termos a serem inseridos na ontologia, no processo de desambiguação empregado e na forma de criar as relações ontológicas. Para o fim aqui proposto, é suficiente o uso de ontologias leves, pois deste modo o desempenho computacional de seu processamento pode ser mantido em tempos aceitáveis, apoiando o processamento *online* das páginas *Web*.

Na abordagem proposta, antes de se tornarem parte da ontologia, os termos mais promissores passam por um processo de seleção que requer a sua existência léxica na *WordNet*. Essa seleção é essencial, visto que a estrutura final da ontologia será baseada nas relações semânticas existentes na *WordNet*, o que em parte garante sua qualidade. Por outro lado, devido a esta filtragem, não há garantia alguma de que termos fortemente representativos estejam presentes na ontologia, podendo causar a perda de representatividade da ontologia emersa.

Um exemplo prático que elucida essa situação é dado pelo conteúdo presente na página *Web "http://en.wikipedia.org/wiki/Battlecruiser"*, na qual o extrator classifica o termo *"Battlecruiser"* em segundo lugar, com ocorrência de 37 vezes. Mesmo sendo

um termo de alta frequência e de grande representatividade, ele não será eleito para participar da ontologia, visto que não pôde ser encontrado na *WordNet*.

Outro ponto a ser avaliado, é a redução da representatividade da ontologia pelo fato de verbos e adjetivos não serem considerados na emersão. A decisão de não incluí-los na ontologia é devida a *WordNet* implementar uma árvore independente para cada uma dessas classes gramaticais, tornando difícil considerá-las em uma única representação.

4. Conclusões

Este trabalho procurou possibilitar a emersão de ontologias a partir das páginas *Web* visando apoiar o processo de modelagem de usuário e de contexto necessários ao processo de busca adaptativa do projeto SARI*Web*. Para tal foi apresentado um processo de extração de termos e sua conversão em conceitos usando a *WordNet*. Ao conjunto de conceitos resultantes foram aplicadas simplificações para que o desempenho computacional da emersão das ontologias ficasse dentro de limites aceitáveis para uso online.

As primeiras avaliações deste processo se mostraram promissoras. No entanto, ainda existem alguns problemas em relação à qualidade final das ontologias emersas e sua aplicação à modelagem de interesses dos usuários e de contexto. Estes problemas estão sendo estudados no projeto SARI*Web* com a realização de teste de modelagem de contexto de interesse do usuário e sua utilização na reordenação de resultados de busca.

Referências

[1] Lyman, P., "How Much Information?," University of California. USA. http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/, March 2008, 28.

[2] Himma, K., "The Concept of Information Overload: A Preliminary Step in Understanding the Nature of a Harmful Information-Related Condition," Ethics and Information Technology, Vol. 4, No. 4, December 2007, pp. 259–272.

[3] Micarelli, A, et. al., The Adaptive Web, Personalized Search on the World Wide Web, Vol. 4321 of LNCS, Springer-Verlag, Berlin, Heidelberg, p. 195–230.

[4] Ardissono, L, et al., "A Multi-Agent Infrastructure for Developing Personalized Web-Based Systems," ACM Transactions on Internet Technology, Vol. 5, No. 1, February 2005, pp. 47–69.

[5] Bunt A., Carenini G., Conati, C., The Adaptive Web, Adaptive Content Presentation for the Web, Vol. 4321 of LNCS, Springer–Verlag, Berlin, Heidelberg, p. 409–432.

[6] Newman, M., "Power Laws, Pareto Distributions and Zipf's Law," Contemporary Physics, Vol. 46, No. 5, May 2006, pp. 323-351.

[7] Borth, M., Uma Abordagem de Recomendação de *Tags* Semânticas para Sistemas Baseados em *Tagging*, dissertação de mestrado, Universidade Estadual de Maringá, Departamento de Informática, 2011.

[8] Fellbaum, C., WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, US, 1st edition, May 1998.

[9] Banerjee, S.; Pedersen, T., Computational Linguistics and Intelligent Text Processing, An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet, Vol. 2276 of LNCS, Springer-Verlag Berlin, Heidelberg, February 2002, p. 117–171.

[10] Nanni, L., "Prototipação de um Sistema de Recomendação de *Tags* para o *TagManager*," XIX EAIC – Encontro Anual de Iniciação Científica, Unicentro, Guarapuava, 2010.

[11] Giunchiglia, F., Marchese, M., Zaihrayeu, I., Journal on Data Semantics VIII, Encoding Classifications into Lightweight Ontologies, Vol. 4380 of LNCS. Springer-Verlag, Berlin, Heidelberg, 2007, p. 57–81.