Métricas para ontologias no formato OBO: Um estudo utilizando o Cytoscape

Ana Carolina Gama e Silva Assaife[1], Fernanda Bruno dos Santos[1], Jonice de Oliveira Sampaio[1], Maria Luiza Machado Campos[1]

[1]Programa de Pós Graduação em Informática (PPGI)
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro– RJ – Brazil

{assaife,fernanda.bruno}@ufrj.br

Abstract. *The importance of the role of ontologies is growing and also its size and complexity. Therefore, the evaluation of ontologies becomes essential because it makes possible to determine some of their fundamental characteristics that are important for designing ontologies, improving quality of existing ones, preventing and reducing the need for future maintenance. This paper aims, through literature review, some methods for evaluating ontologies and through them, suggests a set of metrics in order to relate their theoretical meaning to the meaning in the ontology through an experiment.*

Keywords: *ontology, metrics, bioinformatics, Cytoscape*

Resumo. *A importância do papel das ontologias vem crescendo e, consequentemente também, seu tamanho e complexidade. Portanto, a avaliação de ontologias se torna essencial pois torna possível a determinação de algumas de suas características fundamentais que são importantes para projetar ontologias, melhorar qualidade das já existentes, prevenir e reduzir a necessidade de manutenção futura. Este trabalho aponta, através de revisão de literatura, alguns métodos de avaliação de ontologias e através deles, sugere um conjunto de métricas com o objetivo de relacionar o seu significado teórico ao significado na ontologia através de um experimento.*

Palavras chave: *ontologia, métricas, bioinformática, Cytoscape*

1. Introdução

O rápido crescimento da Internet impulsiona o compartilhamento de informação entre os usuários. É neste contexto que a Web Semântica se expandiu tendo como principal objetivo criar uma estrutura de representação do conteúdo de páginas da Web, possibilitando atingir esse compartilhamento de forma que os computadores consigam interpretar dados e, portanto, facilitando a utilização da informação pelos usuários. Nesse contexto, as ontologias são utilizadas no compartilhamento de conhecimento e visam a apoiar a representação explícita e formal de uma conceituação sendo um mecanismo importante no suporte à interoperabilidade de sistemas.

Com o aumento da popularidade das ontologias, o desenvolvimento de atividades que utilizam essa representação de domínio cresceu e se expandiu para diversas áreas. Em especial, as comunidades internacionais da área da genômica proporcionaram esse crescimento de informações biológicas descentralizadas. A disponibilidade crescente e necessidade de gerência e compartilhamento de ontologias

favoreceram a criação do consórcio de terminologias conhecido como Open Biological and Biomedical Ontologies, OBO, (Smith et al. 2007) possibilitando assim, a troca e reuso do conhecimento pelas comunidades acadêmicas e científicas. Diante de uma grande massa de dados e, consequentemente, de um volume muito grande de informações disponíveis para os usuários, torna-se fundamental que estas informações possam estar descritas ou contextualizadas de forma a facilitar suas localizações, usos e interoperabilidades.

Com o crescente número de ontologias e expansão de seu reuso se torna importante conhecer, analisar e comparar elementos destas ontologias. Além disso, é necessário também avaliar sua estrutura para que seu entendimento seja mais completo e, portanto, melhor. Em especial, no processo de integração e mapeamento entre ontologias, uma avaliação mais acurada das características estruturais, dentre outras, constitui valioso instrumento de apoio.

Outra questão que deve ser destacada é o pouco conhecimento de ferramentas que atendam a necessidade de analisar a ontologia e seus elementos. Algumas vezes, recorre-se a uma análise manual de termos. Tendo em conta que essas tarefas demandam tempo e muito esforço, corre-se um risco de o resultado obtido não ser preciso o suficiente quando comparado com um método automatizado. Este problema pode ser causado pela falta de profissionais de Ciência da Computação capacitados em Bioinformática que buscariam automatização do processo necessário para análise.

O objetivo deste trabalho é levantar, experimentar e analisar um conjunto de ferramentas livres e de fácil manuseio que realizam análises de ontologias utilizando-se de métricas variadas que possam ser úteis aos interesses dos pesquisadores. Além disso, visa apontar uma série de métodos para avaliação de ontologias disponíveis, no domínio da biomedicina, bem como uma discussão de métricas que permitem avaliar e conhecer essas estruturas. Serão propostos, também, experimentos que buscam evidenciar o significado obtido através das métricas no contexto da ontologia selecionada, a GO Slim, utilizando a ferramenta Cytoscape e seus plugins.

2. Revisão de Literatura

Esta seção irá tratar de métodos já existentes para avaliar ontologias disponíveis na literatura com uma visão voltada para integração deles em um único *framework* com foco teórico.

De acordo com Gangemi et al. (2005), é possível identificar três tipos principais de métricas para avaliação de ontologias: métricas estruturais, que estão relacionadas com a representação da ontologia como um grafo, métricas funcionais que estão relacionadas com a utilização da ontologia e seus componentes, e por fim as métricas relacionadas a usabilidade que dependem do nível de anotação da ontologia considerada.

O trabalho de Yao et al. (2005) traz um conjunto de medidas que avaliam o grau de relacionamento entre elementos em uma ontologia, chamadas métricas de coesão. Uma ontologia tem alto grau de coesão se suas entidades são fortemente relacionadas. A utilidade dessas métricas vem da idéia de que conceitos agrupados em uma ontologia devem estar relacionados a um determinado domínio de modo a atingir objetivos semelhantes.

Os estudos de Gómez-Pérez (2003) discutem, dentre outros assuntos, avaliação de conteúdo de ontologias, que tem como principal objetivo detectar inconsistências ou redundâncias antes de utilizá-las em aplicações. Ela está relacionada com o paradigma da representação de conhecimento que possui ligação com a linguagem em que a ontologia é implementada.

A pesquisa de Noy (2004) é focada nas necessidades dos consumidores de ontologias que necessitam saber quais delas são adequadas às suas necessidades. É mencionado, também, que é importante não só um sistema para avaliar ontologias de um ponto de vista genérico mas também modos práticos para que esses consumidores possam descobrir e avaliar as ontologias em questão.

A aplicação OntoClean, que foi proposta por Welty et al. (2003), tem como objetivo detectar inconsistências formais e semânticas nas propriedades definidas por uma ontologia. Sua função principal é a avaliação formal de propriedades definidas na ontologia por meio de uma estrutura taxonômica pré definida de meta propriedades.

Através da revisão dos trabalhos citados foi selecionado um conjunto de métricas candidatas para a realização de um experimento apresentado na seção 4. Neste experimento, será possível visualizar não só o significado teórico das métricas mas o que elas significam no contexto da ontologia em questão.

3. Ferramentas e métricas para ontologias

Esta seção irá tratar das ferramentas utilizadas para realização do experimento bem como as métricas que cada uma delas trata.

De acordo com Sure e Vrandecic (2007), métricas são necessárias para se avaliar ontologias no decorrer das fases de construção e de aplicação, possibilitando uma compreensão rápida e simples a respeito do que está sendo modelado por essas estruturas, desse modo, facilitando o controle de sua futura evolução. Com o objetivo de calcular essas métricas e, consequentemente, seus significados será utilizado o Cytoscape, um software de bioinformática de código aberto e gratuito usado para visualizar interações de redes moleculares (Shannon et al. 2003).

A ontologia GO Slim será utilizada para realizar o estudo. Ela é uma versão editada da ontologia GO (Ashburner et al. 2000), também chamada de *GO Full*., e contém um subconjunto de termos pertencentes a GO completa. Sendo assim, ela contuma visão de alto nível dos processos biológicos, funções moleculares e/ou localizações celulares sem todo o detalhamento contido na GO completa. Vale lembrar que ela pertence a Gene Ontology que faz parte do Consórcio OBO.

Para está análise, serão utilizados três plugins do Cytoscape para realizar o cálculo das métricas e a visualização da ontologia como um grafo. O primeiro deles é o ClusterViz que é um *plugin* desenvolvido para visualização e análise de *clusters* de uma rede.Três importantes algoritmos de criação de grafos de *clusters* que são FAG-EC, EAGLE e MCODE foram implementados no *plugin*. Para esse estudo, foi utilizado o algoritmo MCODE, visando encontrar regiões fortemente conectadas em ontologias. A partir do resultado da análise de *clusters* nas ontologias propostas, é possível esperar que os termos das ontologias encontrados no grupo tenham similaridade entre eles, explicando, dessa forma, a relação forte entre esses conceitos e apontando a presença de possíveis redundâncias.

Outro *plugin* útil que fornece métricas de uma ontologia é o CytoHubba. Com ele, é possível fazer uma classificação de nós, escolhendo o algoritmo dentre os disponíveis. Ou seja, em uma ontologia importada, a ferramenta é capaz de disponibilizar o *ranking* dos termos, usando algoritmos topológicos.

Dentre os critérios de ordenação disponíveis, para o experimento, foi selecionado o *degree*, ou grau de um nó em um grafo. Essa medida é simples de ser calculada. Para grafos não direcionados, a medida é correspondente ao número de ligações do vértice. Já para grafos direcionados, o grau é correspondente a soma das arestas que entram com as arestas que saem do vértice. No contexto das ontologias, *degree* é a soma das relações que envolvem aquele termo analisado. O grau de um nó em uma ontologia pode trazer o sentido de ambigüidade. Um termo que possui somente uma relação para outro termo, possui um significado único que pode ser descrito por esse relacionamento. Portanto, baixos valores de *degree* facilitam o entendimento das relações entre os conceitos analisados. Já um termo com alto valor dessa medida, possui um alto valor de relações associadas. Dessa forma, há a dificuldade de estabelecer um significado único para um termo em estudo, dependendo da qualidade das muitas relações associadas ao termo, trazendo uma possível ambigüidade.

O terceiro e último plugin utilizado neste estudo é o Network Analyzer (Assenov 2008). Ele será útil para analisar e visualizar ontologias computando parâmetros que descreverão a topologia para redes direcionadas e não direcionadas com o objetivo de coletar estatísticas sobre uma única ontologia selecionada.

Uma qualidade deste *plugin* é que ele pode ser utilizado em dois momentos no experimento. O primeiro envolve o cálculo de estatísticas da ontologia como um todo, permitindo assim que o usuário visualize um resumo do conteúdo da ontologia. Esse sumário torna-se importante na medida em que as ontologias são atualizadas, modificadas e reutilizadas para o uso em conjunto ao longo do tempo. Com o grande número de versões que passam a estar disponíveis, é necessário comparar as mudanças entre versões e, se possível, acessar rapidamente o tipo de mudança efetuada. Dessa forma, os valores obtidos nos cálculos facilitam a visão do usuário e o entendimento de uma determinada ontologia que está em constante desenvolvimento e manutenção. Em um segundo momento, o *plugin* é utilizado com uma abordagem diferente. Isto é, o cálculo passa a ser feito para um determinado nó isolado e não mais para a ontologia inteira. Uma qualidade da ferramenta que deve ser ressaltada é a possibilidade do tratamento de centralidades como atributos, permitindo o enriquecimento da análise do Cytoscape. De acordo com Scardoni et al. (2009), centralidades são parâmetros de nós que permitem identificar nós que possuem posição relevante em comparação ao universo da rede analisada.

Medidas relacionadas a caminhos mais curtos são úteis nos experimentos propostos, pois fornecem um alto grau de conhecimento sobre o conceito e caracterizam a posição do termo na ontologia, disponibilizando mais dados que o *degree*. Nesse estudo, duas dessas medidas foram selecionadas: *closeness centrality* e *betweenness centrality*.

Deve-se destacar que a *closeness centrality* pode ser obtida através do inverso da média dos caminhos mais curtos entre um nó *n* e todos os outros nós do grafo. Ela representa a influência que o nó em questão exerce sobre as interações dos outros nós na rede como um todo. No contexto da ontologia, os termos que possuem altos valores

dessa medida possuem os menores caminhos para os outros da ontologia. Ou seja, estão mais pertos dos outros termos e são partes essenciais do fluxo da rede.

No que se refere à *betweenness centrality*, considerando três nós distintos *a, b, c*, a *betweenness centrality* do nó *a* é calculada a partir da razão entre o número de caminhos mais curtos entre *b* e *c* e o número de caminhos mais curtos entre *b* e *c* que passam por *a*. É possível observar que quanto maiores os valores encontrados, maior será a quantidade de caminhos mais curtos entre outros termos que passam por esse nó. A *betweenness centrality* é uma forma de obter um alto grau de informação sobre o nó, ja que é capaz de ilustrar a sua posição em um sentido global. Essa medida mostra que o nó com alto valor exerce uma função de controle e influência sobre a rede, funcionando como uma espécie de ponte entre complexos da ontologia.

4. Experimento

Nesse trabalho, experimentos foram propostos a fim de ilustrar a importância da métrica tirada de uma ontologia. Sendo assim, a GO Slim foi usada como entrada dos experimentos por ser uma ontologia com poucos conceitos, de alto nível e que possibilita a visão do nó em comparação com o todo. Os três ramos da GO Slim, *cellular component*, *biological process* e *molecular function*, são tratados como ontologias distintas nesse experimento. Para apoiar essa análise, um conjunto de ferramentas foi selecionado. O Cytoscape é a ferramenta de edição e visualização de grafos selecionada, já para calcular estatísticas e análises os plugins Clusterviz, Network Analyzer e CytoHubba foram utilizados.

4.1 – Análise de *Clusters*

Essa etapa do experimento foi apoiada pelo *plugin* ClusterViz e foram mantidos os parâmetros sugeridos pela documentação da ferramenta, objetivando encontrar regiões fortemente conectadas.

Após concluir essa parte do experimento, foi possível observar que somente no ramo *cellular component* foi encontrado um *cluster*, utilizando os parâmetros propostos. Ao permitir a criação de *clusters* maiores, a partir da alteração dos parâmetros, foram encontradas regiões conectadas também nos outros dois ramos da GO Slim. Porém, esses não foram analisados, pois, ao aumentar o tamanho do *cluster,* a análise perdeu o propósito. Sendo assim, o *cluster* do ramo *cellular component* possui três termos: *organelle*, *nucleolus* e *nucleus*. Pode-se concluir que ou esses elementos trazem uma redundância por estarem tão fortemente ligados ou representam uma inconsistência de representação ao compará-los com os outros termos da ontologia.

4.2 – Análise de *Degree*

Essa etapa do experimento foi feita usando o *plugin* CytoHubba visando obter uma classificação dos termos da ontologia que possuem mais ligações com outros termos.

A partir dos resultados obtidos, pode-se concluir que quanto mais ligações, mais relações envolvem o conceito analisado e mais significado está agregado ao conceito.

No ramo *biological process,* o termo com maior valor de grau é a própria raiz da ontologia, com valor 32. Observa-se que aproximadamente 50% das relações dessa ontologia são com a raiz, o que mostra uma falta de detalhamento na descrição dos

termos dessa ontologia. O mesmo acontece com o ramo *molecular function,* evidenciando termos que são folha diretamente ligados por uma relação *part-of* ou *is_a* com a raiz.

Já no ramo *cellular component,*a raiz também possui o maior valor de grau, mas os termos *cytoplasm* e *organelle* também têm um alto valor, podendo ser interpretados assim como subdomínios da ontologia *cellular component,* já que os outros termos da ontologia possuem muitas ligações com os dois destacados.

Vale ressaltar que ao analisar somente o *out-degree,* deve-se perceber se o excesso de arestas de saída dá significado da descrição dos relacionamentos entre os conceitos de difícil compreensão.

4.3 – Análise da Ontologia

Esta etapa do experimento foi realizada utilizando o *Network Analyzer* no intuito de obter uma visão geral da estrutura da ontologia, usando a abordagem de enxergar a ontologia como um grafo.

O resumo da ontologia permite que o usuário tenha noção, principalmente, do seu tamanho e suas características como número de nós, quantidade de ligações entre termos, a média de termos a que cada nó está ligado. A cada nova versão dessas ontologias é possível perceber, comparar os valores obtidos e concluir seu diâmetro, que significa a maior distância entre dois nós da ontologia, seu raio, correspondente à menor distância entre dois nós da ontologia. Além disso, é possível verificar algumas possíveis inconsistências na sua construção como um nó com a ligação para ele mesmo e termos isolados, isto é, sem ligações.

O resumo de cada ontologia analisada pode ser visto na Tabela 1. É possível perceber, a cada nova versão dessas ontologias, pode-se comparar os valores obtidos e concluir se as modificações feitas foram a nível estrutural ou não.

Tabela 1. Resumo das ontologias usando o Network Analyzer

	Molecular Function	*Biological process*	*Cellular Component*
Número de Componentes Conexos	1	1	1
Diâmetro	4	4	4
Raio	1	1	1
Número Médio de Vizinhos	2,049	2,36	3,611
Número de nós	41	50	36
Número de arestas	42	63	65
Densidade	0,051	0,048	0,103
Número de nós isolados	0	0	0
Número de Loops	0	0	0

4.4 – Análise das Centralidades

Esta etapa do experimento foi realizada utilizando o *Network Analyzer* objetivando a obtenção da relevância de um nó em relação a ontologia em questão.

Após a geração dos resultados das métricas *betweenness centrality* e *closeness centrality,* foi percebido que os termos que possuiam maiores valores dessas métricas eram termos mais complexos da ontologia, ou seja, os mais gerais. Esses termos são importantes uma vez que possuem caminhos mais curtos passando por eles o que indica que esses termos possuem relação com grande parte dos outros termos da ontologia. Logo, são termos centrais em questão de estrutura.

Foi notado, também, que para os três ramos da GO Slim esta idéia se repetiu, ou seja, termos com maiores valores de centralidade eram termos mais gerais da ontologia representados pelas raízes *cellular component, molecular function* e *biological process.*

5. Conclusão

Este artigo tratou da importância de se avaliar ontologias no cenário atual de crescimento do uso e relevância das ontologias nas mais diversas áreas do conhecimento. O estudo apontou um conjunto de métodos de avaliação disponíveis na literatura e selecionou algumas métricas a fim de realizar um experimento que apontou ferramental apropriado para os cálculos.

Pode-se concluir que o ramo de avaliação de ontologias está em crescente expansão uma vez que as ontologias estão se tornando cada vez mais importantes e complexas demandando métricas para que seja possível conhecê-las melhor. Desse modo, pode-se trabalhar na qualidade e estimar custos de sua utilização, por exemplo.

As métricas selecionadas se mostraram úteis para determinar características da ontologia e dos termos pertencentes à ela. Foi possível, também, identificar como estão dispostos os relacionamentos entre termos e, consequentemente, perceber seus papéis nos ramos da ontologia selecionada.

Segundo Cross e Pal (2005), o resumo da ontologia é necessário, por exemplo, no resultado de buscas de programas como o Swoogle (Ding et al. 2004). As estatísticas sobre as ontologias buscadas orientariam os usuários na escolha de uma terminologia apropriada ao uso. O trabalho sugere mecanismos e mostra aplicações de como obter essas informações.

Como trabalho futuro, seria interessante ampliar o domínio do experimento, sem fixar a análise à área da bioinformática a fim de analisar se os resultados teriam comportamento diferenciado.

References

Assenov, Y., Ramírez, F., Schelhorn, S., Lengauer, T., Albrecht, M., (2008), "Computing topological parameters of biological networks", Bioinformatics (Oxford, England), v. 24, n. 2 (Jan.), p. 282-284.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., et al., (2000), "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium", Nature Genetics, v. 25, n. 1 (Maio.), p. 25-29.

Cross, V., Pal, A., (2005), "Metrics for Ontologies". In: NAFIPS 2005 - 2005 Annual Meeting of the North American Fuzzy Information Processing SocietyNAFIPS 2005 - 2005 Annual Meeting of the North American Fuzzy Information Processing Society, p. 448-453, Detroit, MI, USA.

Gangemi, A., Cantenacci, C., Ciaramita, M. And Lehmann, J. (2005) "A theoretical framework for ontology evaluation and validation", Semantic Web Applications and Perspectives SWAP2nd

Gómez-Pérez, A. (2003) "Ontology Evaluation". In Handbook on Ontologies, pages 251-254

Ding,L , Finin, T., Joshi, A. , Pan, R. , Scott Cost, R. , Peng, Y., Reddivari, P., Doshi ,V., and Sachs,J. (2004). Swoogle: a search and metadata engine for the semantic web. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04). ACM, New York, NY, USA, 652-659.

Noy, N. (2004) "Evaluation by Ontology Consumers". IEEE Intelligent Systems 1541-1672

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504.

Sure, Y. e Vrandecic, D. (2007) "How to Design Better Ontology Metrics". In: 4th European Conference on The Semantic Web, pages 311–325

Welty, C., Kalra, R. and Chu-Carrol, J.(2003) "Supporting ontological analysis of taxonomic relashionships". In *Proceedings of the ISWC-03* Workshop on Semantic Integration

Yao, H., Orme, A. M., and Etzkorn, L. (2005). Cohesion metrics for ontology design and application. *Journal of Computer Science*, 1(1):107–113.