Pytology: Rumo ao Cálculo de Relevância sobre dados RDF

Victor V. Barros Leal[1], **José Antônio F de Macedo**[1], **Lucas Peres Gaspar**[1] **e David Araújo Abreu**[1]

[1]Insight Data Science Lab – Universidade Federal do Ceará (UFC)
Fortaleza – CE – Brazil

{victorbl,lucasperes,araujodavid}@lia.ufc.br, jose.macedo@dc.ufc.br

Abstract. *With the wide availability of RDF datasets on the Web, it becomes increasingly complex the manual analysis to understand the domains of ontologies and their levels of links. Therefore, a challenge is the semi-automatic identification of the relevant relations at the ontology, which are important to define the semantics of the data. This work presents a method to calculate relevance values to the predicates in an ontology by using topological analysis. We show the consolidation of this work with a tool named Pytology and the experimental results generated by using available datasets on the web.*

Resumo. *Com a ampla disponibilidade de bases RDF na Web, torna-se cada vez mais complexa a análise manual para entendimento dos domínios das ontologias e seus níveis de ligações. Diante disso, um desafio é a identificação semi-automática das relações relevantes na ontologia, as quais sejam importantes para definir a semântica dos dados. Este trabalho apresenta um método para calcular valores de relevância para os predicados de uma ontologia através de métricas de análise topológica. Apresentamos a consolidação do trabalho na ferramenta Pytology e nos resultados de experimentos em bases disponíveis na web.*

1. Introdução

Fontes de dados RDF têm ganhado grande importância, aumentando o número de ferramentas para manipulá-los. [Crubézy and Musen 2004] demonstra como o uso dessas fontes é importante para a solução de problemas em diversos cenários de integração de dados. No entanto, a grande quantidade de fontes RDF cria um desafio para os usuários que têm que realizar busca ou *surfing* sobre esses dados, visto que muitos usuários não possuem conhecimento prévio sobre o conteúdo de tais fontes. Esses desafios são trabalhados em áreas como Busca Exploratória [Marchionini 2006] e *Information Retrieval* [Auer et al. 2007].

O problema de recuperar informações de bases em RDF a partir de uma busca é abordado em vários artigos como mostra [Roa-Valverde and Sicilia 2014]. [Elbassuoni and Blanco 2011] apresenta a utilização de cálculos estatísticos pra identificar dados relevantes a partir de palavras-chaves, entretanto definir a relevância aos predicados é um problema nessa abordagem. No contexto de busca exploratória esse problema se apresenta de outra maneira e com poucas soluções, conforme destacado em [Mirizzi and Di Noia 2010] e [Musetti et al. 2012], onde, a partir de um termo do grafo, navega-se sobre os dados a partir dos predicados. Nesses trabalhos é possível perceber a necessidade de calcular a relevância dos predicados.

A eficiência das abordagens citadas depende de um valor de *relevância* para os predicados, os quais não são definidos automaticamente. Para contornar esse problema, os trabalhos definem, manualmente, um conjunto de valores ou regras para cada domínio de dados ou utilizam glossários com predicados mais relevantes. A relevância nesse contexto é uma medida, baseada em alguma métrica, que permite comparar os diferentes tipos de predicados.

Neste trabalho, apresentamos um método que permite calcular valores de relevância dos predicados de qualquer base RDF utilizando-se da análise topológica dos dados. A consolidação desse trabalho dá-se pelo Pytology, uma ferramenta que implementa esse método e gera valores de relevância a partir de medidas de centralidades de grafos. Destarte, como relevância é um conceito subjetivo, os valores da ferramenta servem como uma orientação baseadas em medidas já conhecidas. O resto desse trabalho está dividido da seguinte forma: na seção 2 explicaremos a ferramenta, expondo uma visão geral e como o método funciona. Na seção 3 apresentaremos alguns experimentos e, por fim, traremos nossas conclusões e trabalhos futuros na seção 4.

2. Pytology

2.1. Visão Geral

Dentre as técnicas de cálculo de centralidade sobre grafos, é comum o cálculo de valores apenas para os nós. Isso ocorre pois, geralmente, as arestas não carregam uma informação própria que as caracterize (além, claro, dos nós que elas conectam). Diante disso, é proposto um método que, a partir do cálculo de centralidade dos nós de um grafo RDF, calcula-se um valor que representa a relevância das relações. A Figura 1 demonstra a organização do Pytology.

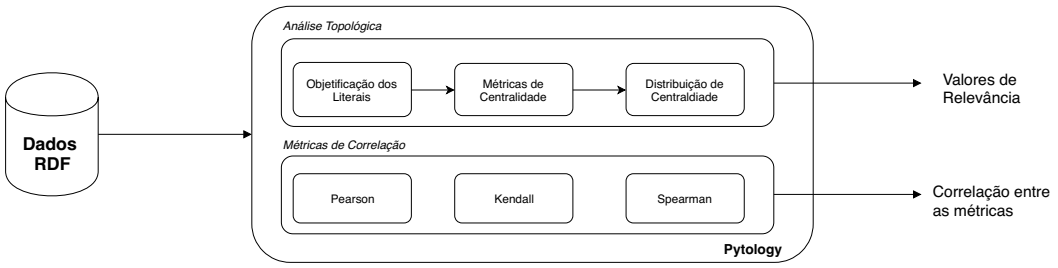

Figura 1. Visão geral das componentes do Pytology

O processo inicia a partir das instâncias de dados os quais são representadas como um grafo RDF que em particular pode conter uma ontologia. Em um grafo RDF os predicados compõem as arestas e os termos e literais os vértices. Em seguida, aplica-se um pré-processamento definido como **objetificação de literais**. Utiliza-se essa etapa para melhorar a representação das relações que apontam para literais. Em seguida, calcula-se a relevância dos nós do grafo a partir de algoritmos de centralidade em grafos. Por fim, aplica-se um cálculo para distribuir esse valores sobre as arestas.

2.2. Análise Topológica

A análise topológica é a parte principal do Pytology. Ela ocorre em três etapas: **objetificação de literais**, **cálculo de centralidade** e **distribuição de centralidade**.

2.2.1. Objetificação dos Literais

Em bases RDF, existem dois tipos de propriedades: relacionamentos, que relacionam duas instâncias de dados (classes) e atributos, que relacionam instâncias à literais (números, datas, texto, etc.). Em um grafo RDF, caso duas instâncias de dados apontem para literais iguais, como uma mesma data, por exemplo, o grafo reconheceria essas datas como dois nós diferentes. Por exemplo: suponha duas entidades com mesmo idioma (dado1, idioma, ”pt”), (dado2, idioma, ”pt”). O relacionamento ”idioma”estaria mais disperso nesse grafo, além de que haveria dois nós distintos para representar o literal ”pt”.

O processo de objetificação transforma esses literais em instâncias de dados. Isso permite que, caso dois ou mais literais sejam iguais em seu valor, eles sejam representados como um único nó no grafo, relacionando-se a todos os dados que apontam para aquele literal. Partindo do exemplo do idioma acima: cria-se um novo dado para intermediar essa relação, gerando as seguintes relações: (dado1, idioma, idioma1), (dado2, idioma, idioma1), (idioma1, valor, ”pt”). Observa-se que a informação do idioma de dado1 e dado2 não é perdida, agora apontam para um mesmo nó, o qual representa o idioma ”pt”.

2.2.2. Métricas de Centralidade

Como os dados estão representados em um grafo RDF, podemos utilizar técnicas de *network-analysis* [Freeman 1978] sobre os dados para realizar cálculos de centralidade dos nós.

Atualmente, o Pytology trabalha com as métricas $Betweenness$, $Closeness$, $Eigenvectors$, $Katz-Eigenvector$ e $Pagerank$. Com exceção do $Betweenness$, essas métricas não calculam valores para as arestas, apenas para os nós. Logo, faz-se necessário a etapa de distribuição de centralidade para podermos mensurar a valor das arestas.

É importante mencionar que cada técnica tem uma semântica associada ao seu resultado. O $Closeness$, por exemplo, dará maior importância a termos mais centrais do grafo, enquanto o $Betweenness$ dará mais importância a entidades limiares entre grupos. Logo, a semântica associada ao valor de relevância dos predicados dependerá da métrica de centralidade utilizada.

2.2.3. Distribuição da centralidade para as arestas

Em um grafo RDF, uma mesma relação pode ocorrer entre diversos pares de dados. Como existem múltiplas arestas no grafo que representam o mesmo predicado, não se pode apenas atribuir o valor de centralidade dos nós que elas conectam como seu próprio valor de relevância. Para calcular esse valor, é preciso levar em conta todas as suas ocorrências.

Se um certo predicado R ocorre entre nós de alta relevância, é provável que essa aresta deva ser de grande relevância. Analogamente, se ela ocorre entre nós de baixa relevância, ela deva ser de baixa relevância. Porém, esse predicado pode ocorrer entre nós de altas e baixas relevâncias, além de poder aparecer mais de uma vez a partir de um mesmo nó. Portanto, para distribuir as relevâncias para as arestas, realiza-se uma média ponderada com a ocorrência da relação e a relevância do nó a quem ela se refere, através

da seguinte fórmula:

$$C_p = \frac{\sum_n^{n \in G} C_n * F_p^n}{F_p}$$

A relevância C_p de um certo predicado r é calculada somando, para cada nó n do grafo G, o produto entre a relevância C_n do nó n pelo número F_p^n de vezes em que o predicado p aparece relacionada a n. Por fim, dividi-se esse valor pelo número de ocorrências F_p do predicado p em todo o grafo.

2.3. Métricas de Correlação

Como foram utilizadas muitas métricas de centralidade, existem várias possibilidades para calcular as relevâncias das relações. Precisa-se, de alguma maneira, decidir qual ou quais métricas utilizar. Descartar o resultado de uma métrica é descartar a semântica associada a técnica, e não é isso que buscamos.

Para isso, o Pytology permite correlacionar os valores gerados por mais de uma centralidade. Utilizado as métricas de Spearman[Zar 1998], Kendall[Abdi 2007] e Pearson[Sedgwick 2012] para correlacionamento de ranks. Spearman avalia relações monotônicas, lineares ou não; Pearson avalia relações lineares e Kendall que utiliza-se de avaliação ordinal.

3. Avaliação dos Experimentos

Os experimentos foram realizados executando o Pytology sobre um dataset de prêmios nobéis, disponível no Datahub.io, e executaram-se os cálculos de centralidade Betweenness(**B**), Closeness(**C**), Eigenvectors(**E**), Katz-Eigenvector(**K**) e Pagerank(**PR**). Como o Betwenness pode ser aplicado sobre arestas, foram feitos dois experimentos: o primeiro aplicando ele diretamente sobre as arestas e, em seguida, aplicando sobre os nós e calculando a distribuição. Chama-se essa segunda abordagem de Betwenness Distribuído(**BD**). Ordenaram-se as relações de acordo com suas relevâncias em forma de rank. A Tabela 1 apresenta as 5 relações mais relevantes de acordo com cada centralidade.

Tabela 1. Top 5 predicados mais relevantes de acordo com cada métrica

Posição	B	BD	C	E	K	PR
1	label	label	type	label	label	label
2	motivation	gender	year	motivation	year	year
3	gender	motivation	awardFile	gender	motivation	motivation
4	year	year	label	year	gender	gender
5	share	type	prizeFile	share	type	type

É evidente que as relações *label*, *type*, *year* e *motivation* estão bem colocadas nos ranks apresentados. Tais relacionamentos apresentam informações bem importantes sobre os dados: o rótulo do termo($label$), que pode ser um nome de país, pesquisador, prêmio, etc; o tipo do dado($type$), representando a classe que ele instancia; o ano($year$) e a motivação($motivation$) de um prêmio.

Dado o contexto dos prêmios nobéis, as informações *motivation*, que é a justificativa pelo recebimento do premio, e *year* são, de fato, bem relevantes. De um ponto de

vista mais voltado para o RDF, temos também que as informações *label* e *type* também são importantes, pois elas definem a representação textual de um elemento e o tipo de dado que ele é. As Tabelas 2, 3 e 4 referem-se às correlações entre os valores das centralidades obtidos de acordo com, respectivamente, Spearman, Pearson e Kendall.

Tabela 2. Correlação entre as relevâncias de acordo com Spearman

	C	**E**	**B**	**BD**	**K**	**PR**
C	1	0.729	0.751	0.770	0.823	0.793
E	0.729	1	0.890	0.946	0.843	0.847
B	0.751	0.890	1	0.914	0.806	0.790
BD	0.770	0.946	0.914	1	0.811	0.811
K	0.823	0.843	0.806	0.811	1	0.992
PR	0.793	0.847	0.790	0.811	0.992	1

Tabela 3. Correlação entre as relevâncias de acordo com Pearson

	C	**E**	**B**	**BD**	**K**	**PR**
C	1	0.294	0.257	0.354	0.334	0.307
E	0.294	1	0.971	0.981	0.989	0.957
B	0.257	0.971	1	0.991	0.985	0.990
BD	0.354	0.981	0.991	1	0.993	0.985
K	0.334	0.989	0.985	0.993	1	0.986
PR	0.307	0.957	0.990	0.985	0.986	1

Tabela 4. Correlação entre as relevâncias de acordo com Kendall

	C	**E**	**B**	**BD**	**K**	**PR**
C	1	0.575	0.586	0.630	0.646	0.624
E	0.575	1	0.739	0.838	0.733	0.740
B	0.586	0.739	1	0.769	0.673	0.642
BD	0.630	0.838	0.769	1	0.693	0.706
K	0.646	0.733	0.673	0.693	1	0.963
PR	0.624	0.740	0.642	0.706	0.963	1

Como Kendall basea-se em uma correlação ordinal, é esperado que os valores de relevância não sejam muito similares, uma vez que uma simples mudança na ordem do rank é suficiente para afetar a correlação. Quanto a Pearson e Spearman, ranks de técnicas semelhantes estão bem correlacionados, como Katz-Eigenvector, Pagerank e Eigenvectors e Betweeness e o Betweeness Distribuído. Esses resultados ajudam a fortalecer a ideia de que a distribuição das centralidades para as arestas mantém consigo a semântica da métrica utilizada. Percebe-se que há um certo direcionamento quando se trata de identificar os predicados mais importantes. Vale ressaltar que há casos de boas correlações com outras técnicas de semântica distinta, como evidência a tabela 3 na correlação de B com K.

4. Considerações Finais

Este trabalho apresenta uma proposta para calcular relevância de predicados em dados RDF a partir de métricas de centralidade. Aplicou-se essa técnica sobre uma grafo RDF de uma ontologia e obtiveram-se resultados demonstrando que utilizar as relevâncias de medidas de centralidade dos termos para calcular relevância das relações é algo que vale a pena ser explorado. Utilizar as métricas de correlação possibilita comparar os resultados dos ranqueamentos e permite analisar o direcionamento de relevância. Além de ser utilizado para validar o algoritmo de distribuição.

Pode-se também mencionar que não há técnica melhor ou pior, mas o que realmente é importante é saber o que se procura. Afinal cada métrica de centralidade possui uma semântica associada e ao utilizar as métricas de correlação pode-se entender quais predicados melhor atendem as técnicas correlacionadas. Permitindo então decidir qual predicado seria o mais relevante para um contexto especifico. Como trabalho futuro, pretende-se: utilizar a semântica de cada métrica e identificar o que seus valores representam no contexto de dados RDF, utilizar a semântica das relações para influenciar o calculo de relevância, usar o $schema$ do grafo RDF, se este o possuir, para efetuarmos um pré calculo sobre a relevância dos predicados.

Referências

Abdi, H. (2007). The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Crubézy, M. and Musen, M. A. (2004). Ontologies in support of problem solving. In *Handbook on ontologies*, pages 321–341. Springer.

Elbassuoni, S. and Blanco, R. (2011). Keyword search over rdf graphs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 237–242. ACM.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46.

Mirizzi, R. and Di Noia, T. (2010). From exploratory search to web search and back. In *Proceedings of the 3rd workshop on Ph. D. students in information and knowledge management*, pages 39–46. ACM.

Musetti, A., Nuzzolese, A. G., Draicchio, F., Presutti, V., Blomqvist, E., Gangemi, A., and Ciancarini, P. (2012). Aemoo: Exploratory search based on knowledge patterns over the semantic web. *Semantic Web Challenge*, 136.

Roa-Valverde, A. J. and Sicilia, M.-A. (2014). A survey of approaches for ranking on the web of data. *Information Retrieval*, 17(4):295–325.

Sedgwick, P. (2012). Pearson's correlation coefficient. *BMJ: British Medical Journal (Online)*, 345.

Zar, J. H. (1998). Spearman rank correlation. *Encyclopedia of Biostatistics*.