UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

VICTOR VIEIRA BARROS LEAL DA SILVEIRA

PYTOLOGY : UMA FERRAMENTA PARA CÁLCULO DE RELEVÂNCIA DE PREDICADOS EM BASES RDF

FORTALEZA

2018

VICTOR VIEIRA BARROS LEAL DA SILVEIRA

PYTOLOGY : UMA FERRAMENTA PARA CÁLCULO DE RELEVÂNCIA DE PREDICADOS EM BASES RDF

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. José Antônio Fernandes de Macêdo

FORTALEZA

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S591p Silveira, Victor Vieira Barros Leal da.
Pytology: Uma ferramenta para cálculo de relevância de predicados em bases RDF : estudo sobre predicados em bases RDF / Victor Vieira Barros Leal da Silveira. – 2018.
40 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Ciências, Curso de Computação, Fortaleza, 2018.
Orientação: Prof. Dr. José Antônio Fernandes de Macedo.

1. Busca Exploratória. 2. Recuperação de Informações. 3. Relevância de Predicados. 4. Grafos RDF. I. Título.

CDD 005

VICTOR VIEIRA BARROS LEAL DA SILVEIRA

PYTOLOGY : UMA FERRAMENTA PARA CÁLCULO DE RELEVÂNCIA DE PREDICADOS EM BASES RDF

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. José Antônio Fernandes de Macêdo (Orientador)
Universidade Federal do Ceará (UFC)

Prof[a]. Dr[a]. Emanuele Marques dos Santos
Universidade Federal do Ceará(UFC)

Prof. Dr. João Paulo Pordeus Gomes
Universidade Federal do Ceará(UFC)

Prof[a]. Dr[a]. Rossana Maria de Castro Andrade
Universidade Federal do Ceará(UFC)

À minha família, por me apoiar e dar forças nessa caminhada. À minha mãe por me auxiliar no dia a dia. À meu pai por me proporcionar qualidade de vida. À minha tia Geovanna por me acolher em momentos de necessidade.

AGRADECIMENTOS

Ao Prof. Dr. Antônio José Macedo por me orientar em meu Trabalho.

Ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

Ao mestrando em ciência da computação e amigo, Lucas Peres por acompanhar e auxiliar meus estudos.

Ao Pedro Henrique e Darley Barreto por serem irmãos nessa jornada.

Aos amigos de laboratório, David Araújo, Regis Pires Magalhães, Ticiane Linhares, Guilherme Estevão, Emanuel Oliveira e Erick Lima pelas discussões que acrescentaram conhecimentos a meus estudos.

Ao meu irmão por ser um grande amigo e um exemplo de determinação na minha vida.

Ao meu pai por me inspirar e motivar em minhas decisões.

À minha mãe por me mostrar que toda luta tem sua glória.

Ao meu falecido avô por ser um exemplo de superação.

Agradeço a todos os meus professores que me proporcionaram a oportunidade de crescer, não somente em conhecimento, mas como pessoa.

“O prazer de vencer é saber que todo o esforço valeu a pena.”

(Victor Vieira)

RESUMO

O crescimento de dados ligados ainda se faz presente por volta de 2018 e com isso a sua utilização e importância em nossas vidas. É comum termos esses tipos de dados armazenados em uma tecnologia que chamamos de RDF. Essa tecnologia, além dos dados, guarda suas relações, e ainda permite a representação dos dados em grafos. Os grafos gerados possuem arestas de semânticas totalmente distintas e bem variadas. Já existem ferramentas que tentam explorar as diferentes semânticas dos relacionamentos dos dados, mas ainda há falhas nos processos utilizados e dificuldade de gerar uma solução acessível. Este trabalho foca justamente em uma maneira de solucionar essas falhas através da geração de relevância para as arestas do grafo.

Palavras-chave: Busca exploratoria. Recuperacao de Informacoes. Relevância de predicados

ABSTRACT

The growth of linked data still occurs in 2018 and associated with that its use is important in our lives. It is common to have data stored in RDF technology. This kind of technology does not only store the data but also their relations. The RDF data can be represented as a graph and this kind of graph has edges with varied semantics. There already exists tools that try to explore the semantic of the edges but their solutions have some problems that this work could help. This work focus on the development of a relevance value for the predicates in RDF graphs by usign topological analysis. **Keywords:** Information Retrieve, Exploratory Search

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

EKP	Encyclopedic Knowledge Patterns
IRI	Internationalized Resource Identifiers
JSON	JavaScript Object Notation
LED	Lookup Discover Explore
NFC	Normal Form C
OWL	Ontology Web Language
RDF	Resource Description Framework
RFC	Request for Comments
SKOS	Simple Knowledge Organization System
SPARQL	Simple Protocol and RDF Query Language
SQL	Structured Query Language
URI	Uniform Resource Identifiers
W3C	World Wide Web Consortium
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

No início do seculo XXI presenciamos um aumento estarrecedor da quantidade de dados e de tecnologias para lidar com suas variedades. Em meio a esse turbilhão de tecnologia surgiu o conceito de Web Semântica, que consiste em um conjunto tecnologias e práticas com a função de estruturar e publicar os dados na Web. Tais práticas foram sendo utilizadas e assim nasceu o que chamamos de Web de Dados, cuja as grandes vantagens são fácil compartilhamento, extensibilidade e reusabilidades dos dados.

Os dados normalmente são armazenados utilizando uma tecnologia que chamamos de Resource Description Framework (RDF), o qual consiste em um modelo ou esquema que facilita a representação de informações, e consultados pela linguagem de consulta Simple Protocol and RDF Query Language (SPARQL). A tecnologia RDF auxilia o compartilhamento de informações bem como sua junção com outras fontes. Os dados em RDF normalmente estão representados em triplas no qual temos sujeito, predicado e objeto como representação dos dados e suas relações.

As triplas são facilmente representadas como grafos, porém o grafo gerado possui uma caraterística bastante peculiar. As arestas presentes nesse grafo possuem uma semântica associada a elas. Essa semântica é a representação das relações entre os dados, além disso, no mesmo grafo é possível termos uma variação de tipos de arestas muito grande. O uso de triplas como representação permite uma modelagem bastante flexível e acaba atrelando muita variabilidade e complexidade aos dados.

O armazenamento de dados é bastante importante, mas mais importante que guarda-lós é poder acessa-lós. Há vários trabalhos lidando com a recuperação de informações desses dados como mostrados em (ROA-VALVERDE; SICILIA, 2014), mas são poucos os que exploram a semântica das relações e, mesmo os que se atreveram, possuem pontos de melhoria em suas soluções.

Diante deste contexto, o objetivo desse trabalho é desenvolver uma ferramenta que permita mensurar a importância das relações de uma base RDF, utilizando medidas de centralidade, fornecendo uma métrica de relevância dessas relações.

Este trabalho estará dividido da seguinte forma, na seção 2 faremos uma fundamentação teórica para que o leitor possa entender os principais pontos aqui abordados, na seção 3 felaremos um pouco sobre os trabalhos que inspiraram a criação da ferramenta, na seção 4 mostraremos e explicaremos a ferramenta Pytology, na seção 5 os resultados e na seção 6 a conclusão do trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Esse capitulo conterá alguns conhecimentos básicos necessários para um melhor entendimento do trabalho. Apresentaremos o conceito de grafo e algumas de suas peculiaridades. Mostraremos algumas medidas de centralidades baseados na estrutura de um grafo. Introduziremos o conceito de Web Semântica e dados ligados, além de algumas de suas tecnologias inerentes. Por fim, apresentaremos também o conceito RDF. Para um maior aprofundamento nos conhecimentos aqui apresentados leia (FREEMAN, 1978) , (WEST *et al.*, 2001) e (BERNERS-LEE *et al.*, 2001).

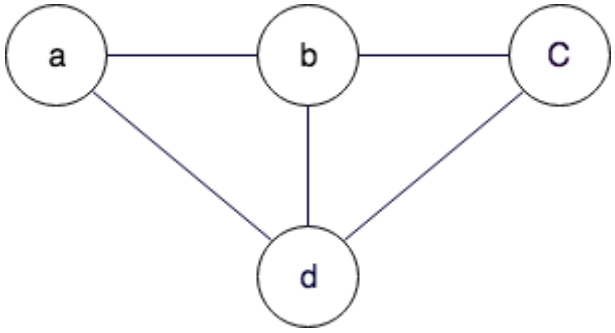
2.1 Grafos

Uma definição mais formal de grafos pode ser encontrada em (BONDY *et al.*, 1976) e ajuda a entender melhor as propriedades atreladas a essa estrutura. Aqui introduziremos alguns conceitos e propriedades básicas para ilustrar o que é um grafo. Simplificando, um grafo pode ser definido como um par ordenado G=(V,E) tal que V representam o conjunto dos vértices e E o conjunto de arestas. Todo elemento do conjunto E é um par de elementos de V e representa a aresta entre eles. Um grafo pode ser dividido em duas categorias básicas, direcionado e não direcionado, que serão explicados a seguir.

2.1.1 Grafos Não Direcionados

A Figura 1 ilustra um grafo não direcionado no qual os vértices são V = $\{$ a,b,c,d $\}$ e arestas E = $\{$ (a,b),(b,c),(c,d),(a,d),(b,d)$\}$. Em um grafo não direcionado as arestas (a,b) e (b,a) significam a mesma coisa, não há distinção entre destino e origem, o que importa é somente se há ou não ligação entre os vértices.

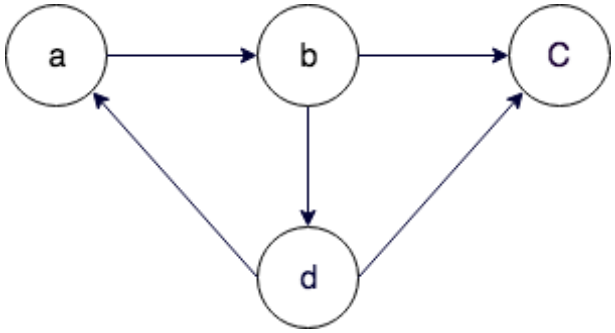
Figura 1 – Exemplo de Grafo Não Direcionado



2.1.2 Grafos Direcionados

Em grafos direcionados as arestas (a,b) e (b,a) são distintas, na Figura 2 podemos observar que a aresta (a,b) existe, enquanto a aresta (b,a) não. Resumidamente, grafos direcionados levam em consideração origem e destino das arestas, é a mesma estrutura porém é acrescentado nas arestas a informação de origem e destino.

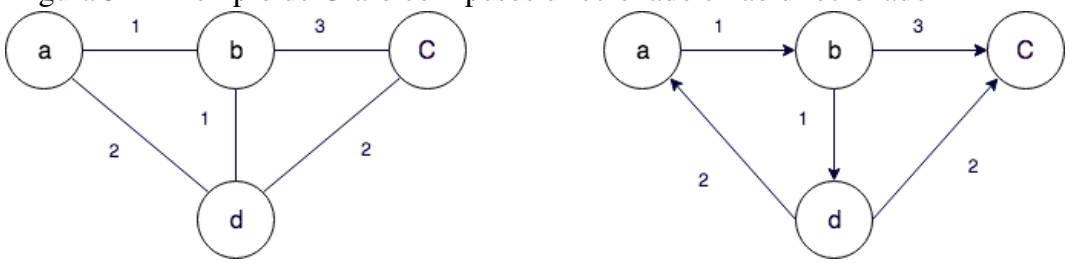
Figura 2 – Exemplo de Grafo Direcionado



2.1.3 Grafos com Pesos

Os grafos ainda podem ser acrescidos de mais informações, um outro exemplo de grafos são os grafos com pesos exemplificados na Figura 3 e que podem ou não ser direcionados. A utilização de cada representação depende da natureza do problema, grafos não direcionados podem ser utilizado em problemas como rede de relacionamentos em redes sociais, pois a relação de amizade e simétrica e se a pessoa A é amiga da pessoa B, B também é amigo de A, ou seja (A,B) = (B,A). Já grafos direcionados pode ser usado para modelar ruas de uma cidade, já que nem toda rua é de mão dupla. Grafos com pesos podem ser utilizados em qualquer um dos dois modelos, basta querermos distinguir a relação, no exemplo da rede de relacionamentos poderíamos usar para mediar o nível da amizade, e no exemplo das ruas poderíamos usar para a importância daquela rua.

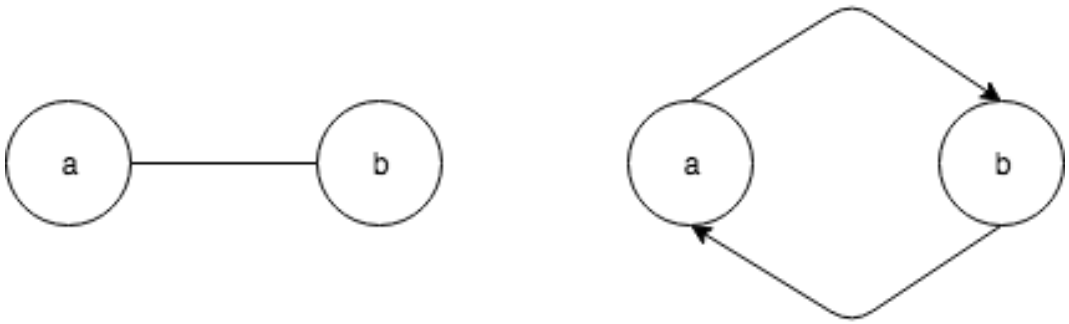
Figura 3 – Exemplo de Grafo com pesos direcionado e não direcionado



2.1.4 Grafos Fortemente Conexos

Basicamente um grafo é fortemente conexo se para qualquer par (x,y) de vértices existe um caminho entre x e y e também entre y e x. Em grafos não direcionados a informação é redundante, porém em grafos direcionados, existir caminho entre x e y não garante reciprocidade.

Figura 4 – Exemplo de Grafo Fortemente Conexo direcionado e não direcionado



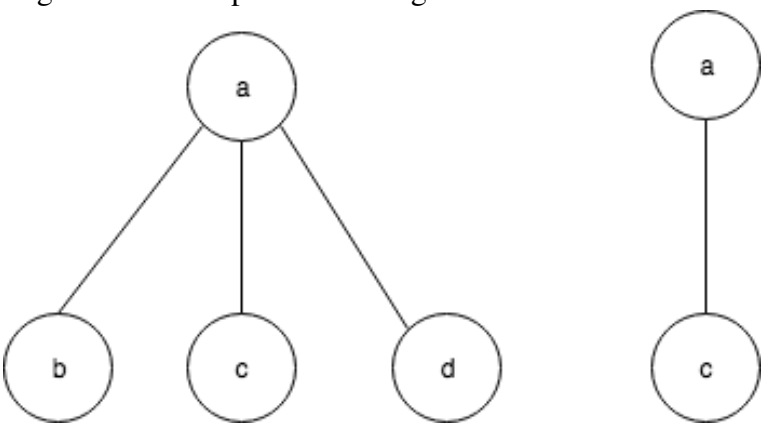
2.1.5 Subgrafos

Um grafo G' é subgrafo de um grafo G se:

- Para todo vértice v pertencente ao conjunto de vértices V' de G', v também pertence ao conjunto de vértices V de G;
- Para toda aresta a pertencente ao conjunto de arestas A' de G', v também pertence ao conjunto de vértices V de G,

Intuitivamente esta definição informal de subgrafo nos diz que G' deve estar contido em G. A Figura 5 exemplifica um grafo G à esquerda e um de seus subgrafos à direita.

Figura 5 – Exemplo de um subgrafo



2.2 Medidas de Centralidade sobre Grafos

As técnicas de centralidade são medidas que buscam identificar quais nós dos grafos são mais relevantes. Anteriormente mencionamos exemplos nos quais a estrutura de grafos pode ser utilizada. O significado da importância também varia de acordo com o que o grafo representa. Utilizando o exemplo do grafo de relacionamentos no contexto de centralidade, o nó mais importante seria a pessoa mais influente, isso é, melhor relacionada. Na explicação das técnicas a seguir suponha que temos um grafo G.

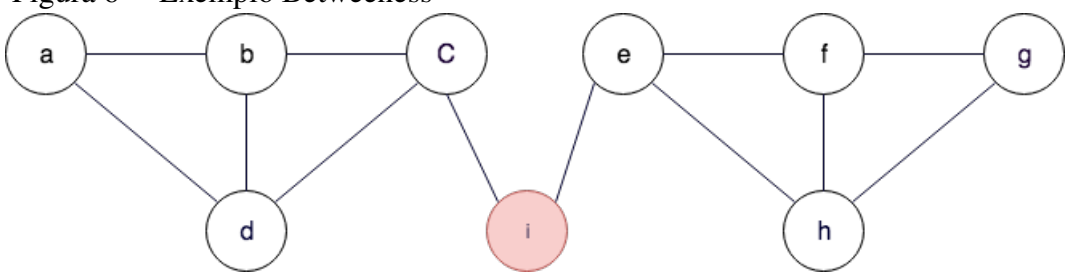
2.2.1 Betweenness

A técnica do Betweenness em particular é a única dentre as abordadas que pode ser aplicada sobre arestas. Essa medida de centralidade é calculada baseadas em caminhos mínimos. Ela calcula para cada par de vértice o número de caminhos mínimos em que um vértice aparece.

$$B_n = \sum_{s \neq n \neq t} \frac{\sigma_{s,t}(n)}{\sigma_{s,t}}$$

O $\sigma_{s,t}(v)$ é o número de caminhos mínimos que contém o vértice v entre os vértices s e t e $\sigma_{s,t}$ o número de caminhos mínimos entre eles. Para que a centralidade seja calculada sobre as arestas basta trocar o vértice v por uma aresta. Normalmente essa técnica acaba dando maior importância a nós que se encontram limiares a grupos. Isso ocorre, porque esse nó acaba sendo a ponte entre os grupos, estando presentes nos caminhos mínimos de indivíduos em grupos opostos. O exemplo da Figura 6 ilustra isso perfeitamente, em termos de vértice teríamos o vértice i com maior betweeness, caso estivéssemos a analizar arestas, seriam as (i,c),(i,e).

Figura 6 – Exemplo Betweeness



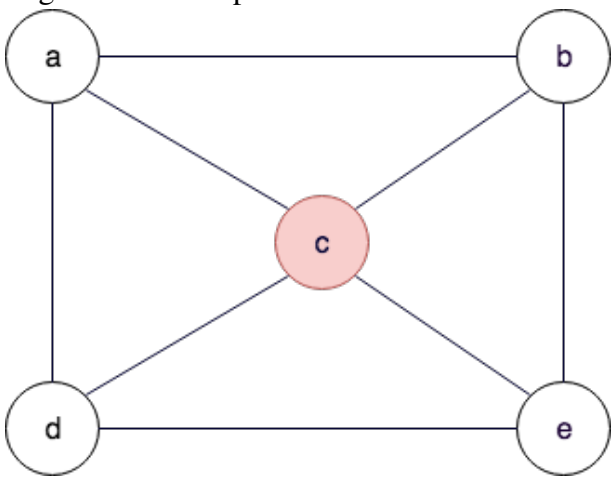
2.2.2 Closeness

O Closeness é uma medida baseada na distância que um nó tem para todos os outros da rede. A consequência direta da utilização desse método é que os nós mais centrais da topologia são os que recebem maior valor de centralidade, ou seja, quanto maior a sua centralidade, mais próximo será de todos os outros nós.

$$C_x = \frac{N}{\sum_y d_{x,y}}$$

O $d(y,x)$ representa a distância entre os vértices y e x, essa distância é igual ao número de arestas percorridas de y a x. O N representa o numero de nós no grafo. A Figura 7 exemplifica o nó com maior closeness, como podemos perceber que ele se encontra mais no centro do grafo.

Figura 7 – Exemplo Closeness



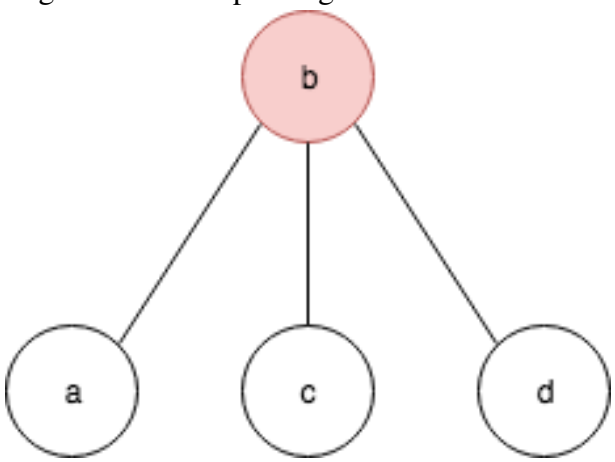
2.2.3 Degree

Essa medida de centralidade é simples, sendo basicamente uma contagem de quantos vizinhos um dado nó n possui, a intuição por traz desse cálculo é que um nó é importante se ele possui muitos vizinhos. Em casos de grafos direcionados, ele será importante se possui muitos nós que o referenciam ou se ele se liga a muitos outros, ou seja, se ele possui muitas arestas que chegam ou que saem dele.

$$x_n = \sum_k a_{k,n}$$

O x_n é a centralidade do nó n, a função $a_{k,n}$ é uma função binária e retorna 1 caso exista aresta entre k e n. Na Figura 8 podemos ver um exemplo do que seria um nó c com um alto valor de centralidade dessa métrica.

Figura 8 – Exemplo Degree



2.2.4 Eigenvectors

A métrica de Eingenvectors pode ser considerada um extensão da Degree. A medida de Degree não leva em consideração a importância dos nós que um dado nó n é ligado. Resumidamente, um nó com alta centralidade nessa métrica é um nó ligado a outros nós importantes.

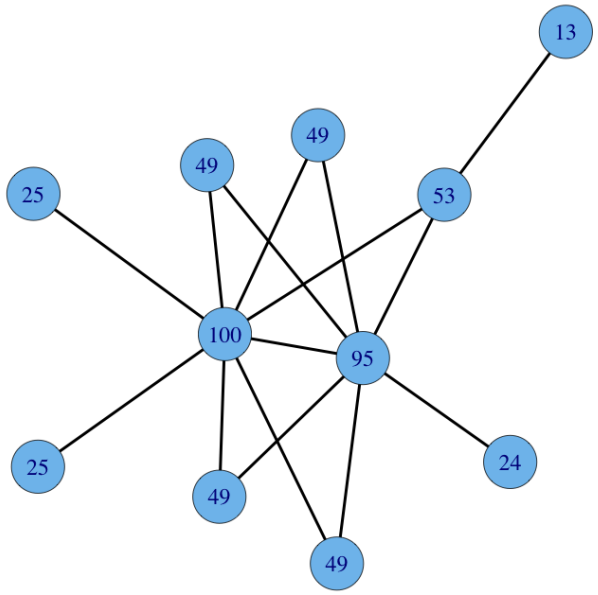
$$x_n = \frac{1}{\lambda}\sum_k a_{k,n}x_k$$

O x_n é a centralidade do nó n, a função $a_{k,n}$ é uma função binária que retorna um caso haja aresta entre os nós k e n e $\lambda \neq 0$ é uma constante. Em grafos direcionados, a medida é baseada nas arestas que chegam ao nó. A Figura 9 exemplifica um grafo com valores dessa métrica, porém Os resultados explicitados na figura não são tão intuitivos como as técnicas anteriores.

2.2.5 Katz Eigenvectors

Essa medida de centralidade resolve um problema da métrica de Eigenvectors que só funciona bem em grafos fortemente conexos. O principal problema é que em grafos direcionados que não são fortemente conexos, somente vértices em componentes fortemente conexas ou vértices que recebem arestas desses componentes que terão algum valor na métrica. Para

Figura 9 – Exemplo Eigenvector



Fonte: https://www.sci.unich.it/francesc/teaching/network/eigenvector.html

solucionar esse problema, a medida Katz também leva em consideração se o nó em questão possui muitas ligações, sejam arestas saindo ou chegando.

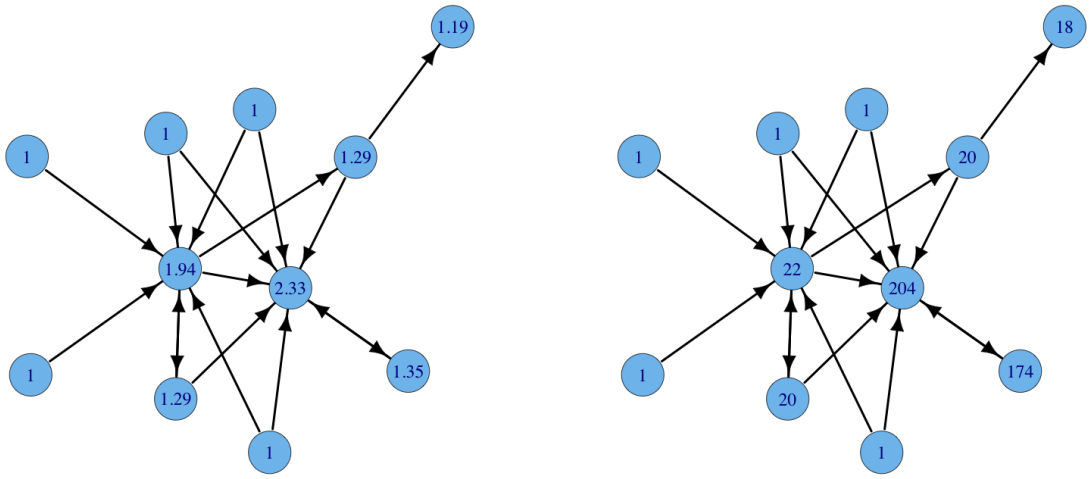
$$x_n = \alpha \sum_k a_{k,n} x_k + \beta$$

O x_n é a importância do nó n, $a_{k,n}$ é uma função que retorna 1 caso haja aresta de k para n, β e α são constantes. O parâmetro α é conhecido como fator de amortecimento. Para baixos valores de α a contribuição dada por caminhos mais longos decai rapidamente, ou seja, os valores são mais influenciados por caminhos pequenos. Para valores altos de α os caminhos mais longos são penalizamos mais suavemente. O parâmetro β é um valor de diferenciação de estado para cada nó. A Figura 10 e 11 exemplificam a diferença no uso dos parâmetros.

2.2.6 PageRank

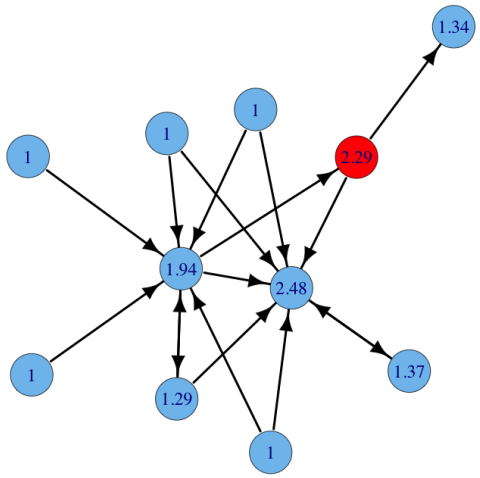
O Page Rank é um ajuste da medida de centralidade Katz. O problema que o Page Rank soluciona e que na medida Katz se um nó possui alta centralidade e se conectar com outros nós, esse outros nós possuirão alta centralidade também. Caso algum desses nós adicionados seja um nó sem tanta importância a centralidade ganha deveria ser diminuída.

Figura 10 – Exemplo Katz com parâmetros $\alpha = 0.85$ e $\beta = 1$ e $\alpha = 0.15$ e $\beta = 1$ respectivamente



Fonte: https://www.sci.unich.it/francesc/teaching/network/katz.html

Figura 11 – Exemplo Katz com parâmetros $\alpha = 0.15$ e $\beta = 2$

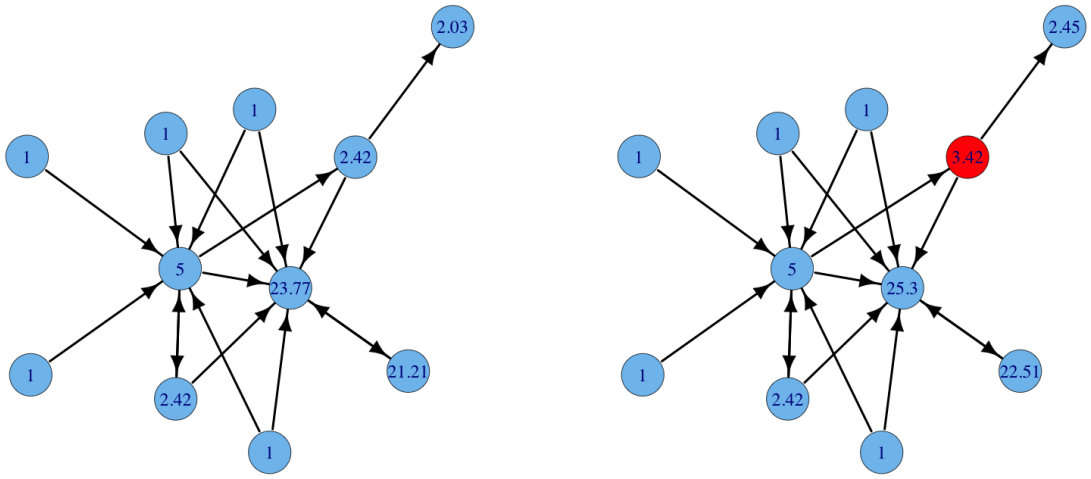


Fonte: https://www.sci.unich.it/francesc/teaching/network/katz.html

$$x_n = \alpha \sum_k \frac{a_{k,n}}{d_k} x_k + \beta$$

O x_n é a centralidade do nó n, α e β são constantes e d_k é o numero de arestas que saem do no k. As constantes representam o mesmo já visto na métrica de Katz. A Figura 12 ilustra a mudança de parâmetros no PageRank.

Figura 12 – Exemplo PageRank com parâmetros $\alpha = 0.85$ and $\beta = 1$ e $\alpha = 0.85$ and $\beta = 2$ respectivamente



Fonte: https://www.sci.unich.it/francesc/teaching/network/pagerank

2.3 Web Semântica e dados ligados

A Web Semântica é uma Web de Dados, ela é extensão da World Wide Webe e é a visão que World Wide Web Consortium (W3C) tem da web de dados ligados. As tecnologias da Web Semântica fornecem um ambiente no qual aplicações podem consultar e integrar dados e extrair inferências utilizando vocabulários. Para criar a Web de Dados é importante um grande volume de dados disponíveis em seus formatos originais, acessíveis e manuseáveis pelas ferramentas da Web Semântica. Uma outra caraterística da Web Semântica e que não somente os dados são acessíveis, mas também os seus relacionamentos. Os conjuntos de dados ligados na Web caracterizam a Web de dados. Os dados ligados são resultado dessa integração e são suportados por tecnologias como RDF, SPARQL, Ontology Web Language (OWL) e Simple Knowledge Organization System (SKOS).

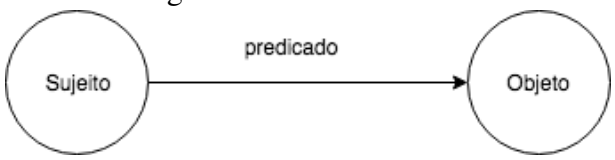
2.4 Resource Description Framework (RDF)

A semântica é o estudo do significados de símbolos e tenta explicar como extraímos significado deles, já dados com semântica são dados organizados e estruturados para que máquinas consigam entender, mais do que isso, essa organização e estruturação ajudam na identificação e no descobrimento de relações. A estruturação desses dados é definida pela RDF Schema, que provê o vocabulário para dados em RDF, é basicamente uma extensão do vocabulário básico do

RDF.

A tecnologia RDF é um framework com um conjunto de especificações para modelos de metadados mantidos pela W3C. Ela é distribuída e descentralizada e permite um modelo de dados flexível utilizado para representação de fatos e relações, possui uma estrutura simples de sujeito, predicado e objeto e, por isso, é organizado em formato de triplas. Essa organização é facilmente representável em um grafo, como exemplificamos na Figura 13.

Figura 13 – Tripla representada em formato de grafo



Existem três tipos de nós em um grafo RDF, Internationalized Resource Identifiers (IRI), literais, e nós vazios. Eles são coletivamente conhecidos como termos RDF. Os três tipos são distintos e distinguíveis.

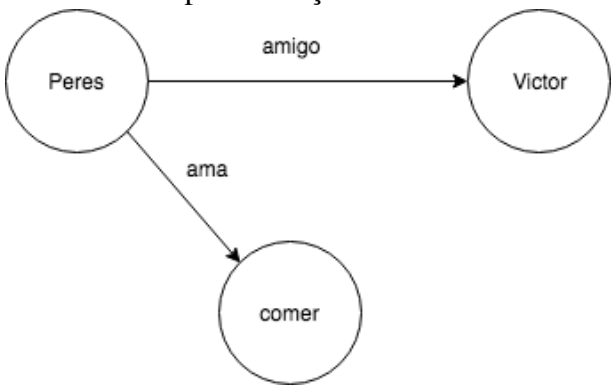
As IRI são uma generalização das Uniform Resource Identifiers (URI). Em um grafo RDF elas representam um código único de *string* que está de acordo com as especificações definidas em Request for Comments (RFC) 3987. Os literais em um grafo RDF seguem dois ou três elementos, a forma léxica, que deve seguir um codigo único e deve estar na forma Normal Form C (NFC), tipo de dado IRI, sendo um identificador de tipo e determina como a forma léxica é mapeada para valor do literal, a *tag* de linguagem, que deve ser vazio a menos que tipo de dados seja estritamente *langString*. Nós vazios são separados de IRI e literais e o RDF não faz nenhuma referência da estrutura interna desses nós.

A estrutura simples de representação em triplas nos permite modelar vários tipos de relações em um único grafo. A Figura 14 exemplifica um grafo no qual temos mais de um tipo de relação. O grafo da figura é composto por 2 conjuntos de triplas, (Peres, amigo, Victor) e (Peres, ama, comer).

Utilizar esses modelos de representação de dados facilita a extensibilidade, compartilhamento e reutilização. Uma das grandes vantagens é que, por ser facilmente extensível, podemos utilizar diversas fontes para eliminar ambiguidades. Outro ponto interessante de utilizar esses modelos é que sujeitos e predicados possuem identificadores IRI e são utilizados para especificação. Esses identificadores ajudam a especificar os dados na Web de Dados.

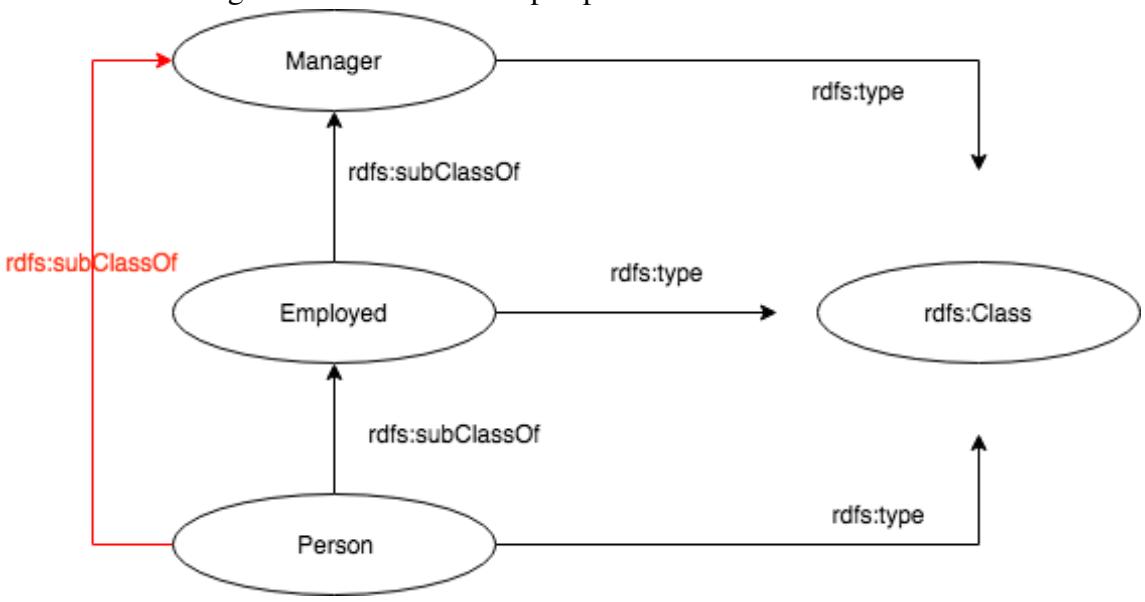
Um outro grande ponto de utilizar dados com semântica é que podemos inferir

Figura 14 – Grafo de triplas com mais de um tipo de relação



informações a partir das já existentes. O exemplo da Figura 15 que mostra um simples grafo RDF que após o processo de inferência descobrimos a relação de subclasse existente entre a classe Manager e Person.

Figura 15 – Grafo RDF após processo de inferência



O formato diferenciado do RDF requer uma linguagem de consulta específica. Essa linguagem é chamada de SPARQL e segue um formato semelhante ao Structured Query Language (SQL). A principal diferença é que os dados estão organizados na forma de triplas e possuem IRI que os identificam. Apesar da estrutura de consulta ser semelhante ao SQL, ainda possui várias nuâncias. A Figura 16 exemplifica uma consulta nessa linguagem. As triplas que formam o RDF também podem ser armazenadas em outros diversos formatos, como JSON e XML exemplificados na Figura 17.

Figura 16 – Consulta SPARQL

```
PREFIX  dc: <http://purl.org/dc/elements/1.1/>
PREFIX  : <http://example.org/book/>
SELECT  $title
WHERE   { :book1  dc:title  $title }
```

Fonte: https://www.w3.org/2001/sw/DataAccess/rq23/examples.html.

Figura 17 – Representação de uma tripla em JSON e XML

JSON

```
{
        "example" :"data",

        "triple" : {
                "subject": "http://www.example.com/person",
                "predicate": "http://www.w3.org/2000/01/rdf-schema#subClassOf",
                "object":{
                        "value":"http://www.example.com/employed",
                        "datatype":"xs:string"
                }
        }

}
```

XML

```
<sem:triple>
        <sem:subject>http://www.example.com/pearson</sem:subject>
        <sem:predicate>http://www.w3.org/2000/01/rdf-schema#subClassOf</sem:predicate>
        <sem:object datatype="http://www.w3.org/2001/XMLSchema#string">
                http://www.example.com/employed
        </sem:object>
</sem:triple>
```

3 TRABALHOS RELACIONADOS

Nessa seção abordaremos trabalhos que estão no contexto de Web semântica e que apresentam pontos de melhoria pela ferramenta desenvolvida. Estabeleceremos o conceito de busca exploratória e recuperação de informações para introduzir o contexto desses trabalhos.

Quando se trata da utilização de bases de conhecimento com semântica e dados ligados, um grande problema que surge é a acessibilidade. Principalmente quando o usuário que utiliza a base não dispõe de conhecimento nem do domínio, nem da tecnologia. Mostramos na Figura 16 um exemplo de consulta SPARQL e é evidente que serão poucos usuários capazes de efetuar algo semelhante. A complexidade da linguagem dificulta, para usuários leigos, o acesso aos dados. Esse problema é característico da busca exploratória. Para facilitar o acesso de usuário às bases de conhecimento é necessário uma interface que facilite essa interação. Mais do que o acesso, essas ferramentas também devem permitir a navegação sobre os dados. A navegabilidade é importante pois permite a aquisição de conhecimento pelo usuário. Os trabalhos (MIRIZZI; NOIA, 2010) de título From Exploratory Search to Web Search and back e (MUSETTI *et al.*, 2012) de título Aemmo: Exploratory Search bases on Knowledge Patterns over the Semantic Web tratarão desse aspecto.

Um outro problema que se tem na acessibilidade é a recuperação de informações. Como a base de conhecimentos é grande, há muitos termos de nomenclatura semelhante ou até mesmo iguais. Essa semelhança gera muita ambiguidade e torna mais difícil sabermos precisamente a que tipo de informação se faz referência. Um pequeno exemplo é, suponha que um usuário digite *manga* no buscador. Como saberemos se ele se refere a fruta manga ou a manga da camisa? Para solucionar esse problema de ambiguidade é preciso restringir o espaço de busca e filtrar as opções. Outra questão envolvendo esse processo é como recuperaremos as informações do grafo baseados nessa busca do usuário. O trabalho (ELBASSUONI; BLANCO, 2011) de título Keyword Sarch over RDF Graphs abordará esse processo.

3.1 Keyword Search over RDF Graphs

O modelo de recuperação de informações discutido em (ELBASSUONI; BLANCO, 2011) gera um conjunto de palavras chaves a partir de uma instância de dados. Essas palavras chaves são utilizadas em um índice invertido e usadas para recuperação de triplas. A representação das triplas associadas a cada palavra chave são subgrafos da instância dos dados em

RDF.

A recuperação de dados se dá pela retirada de palavras chaves da busca do usuário. Após a identificação dessas palavras, recupera-se as triplas pelos índices invertidos correspondestes. Os conjuntos de triplas recuperados gerarão um conjunto de subgrafos. Tendo os subgrafos é necessário uma etapa de ranqueamento para decidir qual subgrafo é o melhor. Essa etapa gera um valor para cada subgrafo, quanto maior esse valor, mais próximo o subgrafo é da busca entrada.

A etapa de ranqueamento é feita utilizando um modelo estatístico que consegue calcular as probabilidades dos termos e entidades. O mesmo não ocorre com os predicados, que possuem suas relevâncias setadas informalmente, baseadas somente em um palpite. Mesmo que o palpite seja feito por um especialista, pode-se perder informações. O especialista pode ter conhecimento do domínio dos dados, mas desconhece a estrutura em que as informações estão dispostas. O problema de se utilizar um palpite para mensurar a importância dos predicados é que ele pode ser vago e impreciso. A imprecisão nessa importância distorce os valores de relevância dos subgrafos. A nível de usuário, receber resultados errados é insatisfatório e prejudicial à experiência de busca.

3.2 From Exploratory Search to Web Search and back

(MIRIZZI; NOIA, 2010) apresenta uma ferramente chamada Lookup Discover Explore (LED) para busca exploratória sobre conjuntos de dados com semântica. A ferramenta tem como objetivo ajudar o usuário no processo de aprendizagem, descobrimento e entendimento de conhecimentos novos e complexos. O LED conta com DBpedia para explorar a semântica de consultas.O DBpedia é em resumo um conjunto de bases de dados ligados e estruturados disponíveis na web. Usando essas bases o LED consegue sugerir tópicos e palavras chaves para o usuário. As palavras chaves geradas pelo LED são relacionadas semanticamente. Cada palavra chave é associada a um recurso RDF vindo da Web de dados. A semântica das palavras chaves são relacionadas à consulta e não ao conjunto de resultados.

A exploração do grafo é suportadas por URI do DBpedia. Apartir de uma URI o explorador é conectado a um conjunto de propriedades predefinidas que também são do DBpedia. Nas configurações iniciais da ferramenta, são escolhidas algumas propriedades que serão utilizadas para auxiliar a exploração. No estudo ele comenta que utilizou $skos:subject$ e $skos:broader$. Mencionamos anteriormente que SKOS faz parte do conjunto de tecnologias que compõe a Web Semântica. O SKOS $subject$ e $broader$ são definições que padronizam

o predicado que identifica o conteúdo e termos semelhantes respectivamente. A justificativa da escolha se da pelo fato de que essas propriedades são independentes de domínio e muito populares em dados do DBpedia. No entanto não apresenta resultados mostrando que de fato são as melhores opções, só comenta que foi observado quer eles aparecem bastante no conjunto de propriedades selecionadas.

O LED então não mede qual tipo de predicado seria o melhor para auxiliar a navegação. A ferramenta só utiliza as propriedades predefinidas e espera um bom resultado. LED carece de uma métrica para definir as melhores propriedades de exploração.

3.3 Aemmo: Exploratory Search bases on Knowledge Patterns over the Semantic Web

(MUSETTI *et al.*, 2012) fala sobre Aemoo que é uma aplicação Web para dar suporte a busca exploratória sobre a Web Semântica. Ela utiliza uma interface simples de palavras chaves. Os dados são coletados de diversas fontes como Wikipédia, Twitter e Google News. O artigo ainda introduz o conceito de Encyclopedic Knowledge Patterns (EKP) explicado em (NUZZOLESE *et al.*, 2011). As EKP são padrões de conhecimento gerados sobre uma base de dados. Eles ajudam na exploração da Web Semântica e no enriquecimento de consulta usando conhecimentos relacionados vindos de diversas fontes da Web. Aemmo ainda aplica filtragem sobre o conhecimento recuperado para selecionar a informação a ser mostrada.

A resolução de identidade do Aemmo é feita em dois passos. Primeiro ele identifica a identidade da propriedade pela busca do usuário. No segundo passo ele procura fontes as quais mencionam esta identidade. O Aemmo ainda reconhece propriedades de tipos de acordo com a taxonomia de tipos do DBpedia. O tipo provê informações adicionais sobre a entidade e afeta como o conhecimento será mostrado ao usuário.

Em resumo podemos dizer que o Aemmo aplica uma exploração baseada em EKP. O uso delas permite filtragem, enriquecimento e extração de significados importantes nas redondezas do dado. Como a exploração de dados é baseada em EKP, acaba sendo muito dependente do domínio de geração. Podemos então dizer que a exploração está restrita aos domínios dessas EKP. Essa restrição não leva em consideração a estrutura de novos tipos de dados, a menos que sejam geradas novas EKP. O problema da geração de EKP não é trivial e mesmo que gerados não temos uma métrica para dizer qual seria o melhor predicado a se utilizar para determinada ação, por exemplo, para navegação entre grupos distintos, ou que predicado melhor permite navegar entre todos os termos.

3.4 Conclusão

Mostramos nas seções de cada trabalho relacionado os processos referentes as técnicas de cada um. Em particular mostramos que há pontos de melhoria em cada técnica. Na ferramenta LED pode-se melhorar as configurações iniciais de exploração. No Aemmo argumentamos que as EKP restringem as sugestões ao domínio de geração. No ranqueamento de subgrafos a escolha da importância dos predicados pode ser substituída por um valor que melhor represente a importância do predicado. Sabendo de todas essas falhas, podemos dizer que faltam métricas para calcular relevância de predicados em grafos com semântica.

As soluções apresentadas descartam qualquer medida de centralidade do grafo. Descartar os resultados de centralidade é descartar o significado de relevância das métricas. Isso é um problema, principalmente considerando que algumas das métricas tem relações com fluxo de informações.

4 PYTOLOGY

Mostramos que em trabalhos anteriores que a maneira que se calcula a relevância de predicados possui brechas. Não é fácil mensurar essa relevância, ainda mais quando os algoritmos de centralidade existentes não levam em consideração as diferentes semânticas presentes em diversos tipos de arestas no grafo. Desconhecemos algoritmo de centralidade baseado na semântica das ligações, só conhecemos os que analisam a topologia. Com intuito de prover essa relevância, criamos a ferramenta chamada Pytology. Essa ferramenta foi desenvolvida utilizando a linguagem de programação *Python*, por isso o nome Pytology, junção de *python* mais *ontologia*. Também foram utilizadas algumas bibliotecas para facilitar as análises, dentre elas temos : *NetworkX* e *Numpy*. A ferramenta traz consigo o conjunto de métricas de centralidade já citadas e nos permite calcular a relevância dos predicados em uma base RDF.

O Pytology é dividido em duas principais partes, uma parte de análise topológica a outra de métricas de comparação. Os módulos estão exemplificados na Figura 18 e serão explicados nas subseções seguintes. A ferramenta recebe um conjunto de dados RDF e que passam por uma etapa de objetificação. Após isso, utilizam-se as métricas de centralidade para calcular relevância. A relevância é então passada para as arestas através de um algoritmo de distribuição de importância. No final os resultados de relevância são comparados.

Figura 18 – Pytology e seus componentes

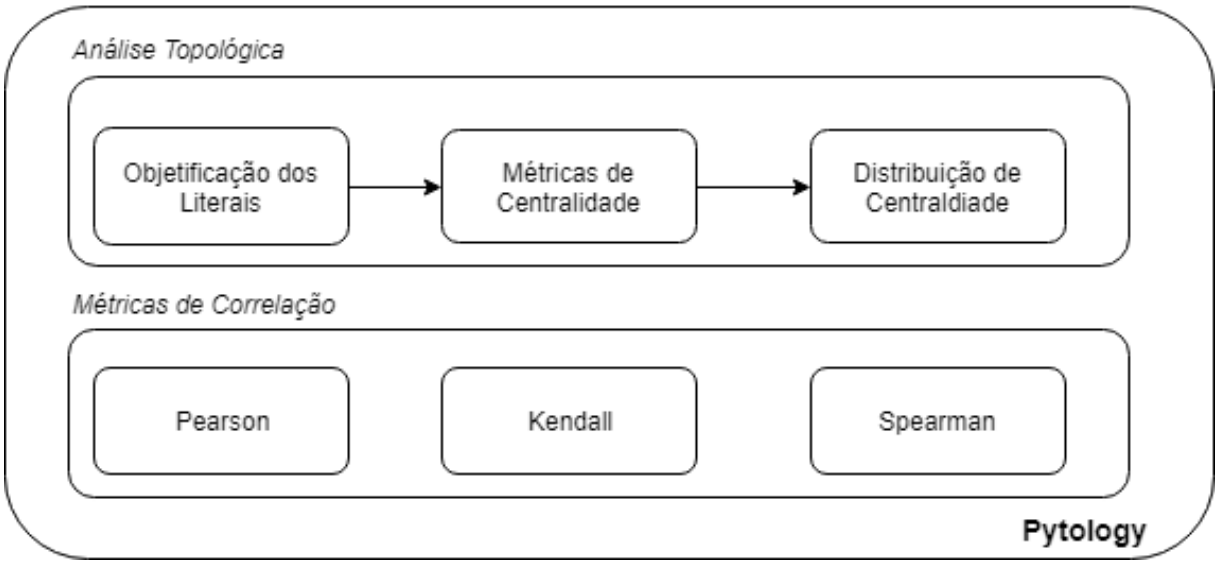
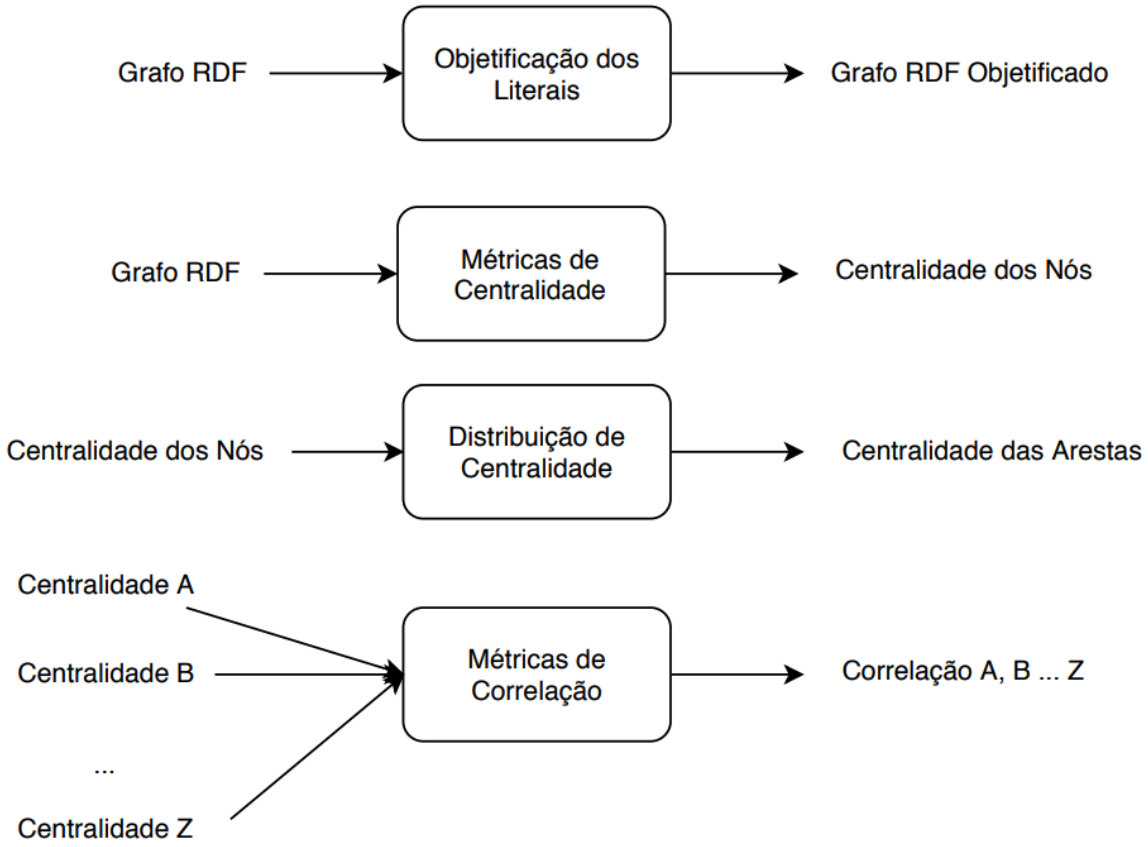


Figura 19 – Pytology e seus componentes



4.1 Análise Topológica

O modulo de Análise Topológica se refere às etapas de Objetificação de Literais, Métricas de Centralidade e Distribuição de Centralidade. Nessa seção explicaremos os detalhes de cada componente desse módulo e os problemas referentes a cada um. Essa parte da ferramenta é responsável, principalmente, por permitir o cálculo de relevância para os predicados.

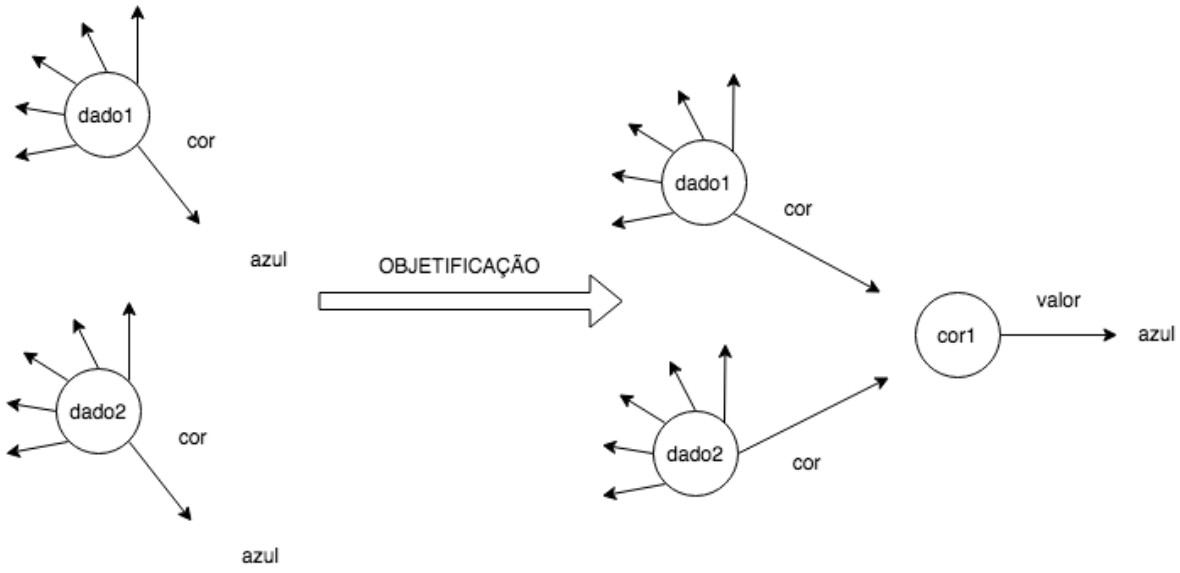
4.1.1 Objetificação de Literais

Nos grafos em RDF há três tipos de nós, IRI, literais e nós brancos. Os literais podem ser valores, como exemplo (dado1,cor,"azul"). O problema e que os valores desses literais não possuem um identificador e que caso haja uma outra ocorrência como (dado2,categoria,"azul"), elas seriam interpretadas como dois nós distintos. A ocorrência do literal acaba ficando dispersa no grafo. A dispersão dessas ocorrências altera a ordem das importâncias dos nós e é preciso corrigi-las.

O processo de objetificação transforma esses literais em instância de dados. Isso

permite que as ocorrência do literal *azul* sejam centralizadas. Assim qualquer ocorrência referente a este literal será uma referência a essa entidade. A Figura 20 exemplifica esse processo.

Figura 20 – Processo de objetificação



O tratamento é feito utilizando um novo dado para intermediar a relação, nesse exemplo criamos uma classe *cor*1 para representar a instância. Então as triplas são modificadas e se tornam as seguintes:(dado1,cor,cor1), (dado2,cor,cor1), (cor,valor,"azul"). Ainda há mais uma tripla que cria a classe *cor*1 de acordo com as especificações do RDF Schema.

4.1.2 Métricas de Centralidade

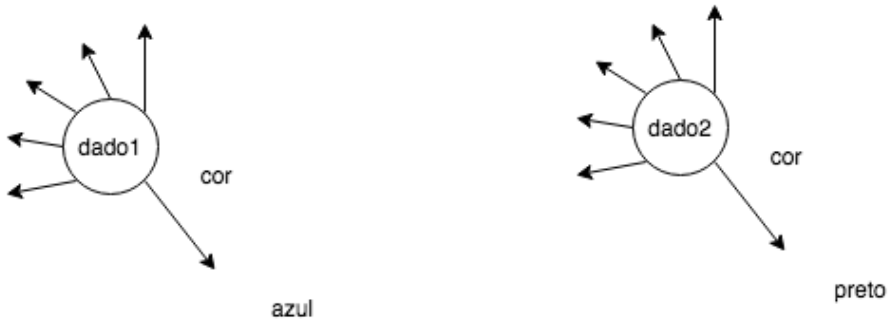
É difícil sabermos que tipo de predicado é o mais relevante em uma ontologia e como os dados estão representados em RDF e essa representação pode ser convertida em um grafo. Utilizamos as técnicas de (FREEMAN, 1978) sobre esses dados para realizar cálculos de relevância dos predicados. O Pytology utiliza as técnicas do Betweenness, Closeness, Eigenvectors, Katz-Eigenvector e Pagerank para fazer isso.

É importante mencionar que cada técnica possui uma semântica associada ao seu valor. O Closeness, por exemplo, dará maior importância a nós mais centrais do grafo, enquanto o Betweenness dará maior relevância a entidades limiares entre grupos. Vale ressaltar que das técnicas utilizadas, somente o Betweenness consegue calcular diretamente a importância para as arestas. Para resolver esse problema é que existe a etapa de Distribuição de Centralidade.

4.1.3 Distribuição de Centralidade

Os grafos em RDF podem modelar vários tipos de relações em um único grafo. Um ponto importante disso é que uma mesma relação pode ocorrer diversas vezes no mesmo grafo, a Figura 21 exemplifica esse tipo de ocorrência.

Figura 21 – Dupla ocorrência de predicado



Como existem diversas ocorrências de um mesmo predicado, não se pode atribuir diretamente o valor de centralidade do nó de origem. Caso contrário teríamos os mesmo predicado com valores de relevância distintos e isso não resolveria o problema. Para calcular o valor de centralidade é preciso levar em conta todas as ocorrências.

O Algoritmo 1 desenvolvido parte da intuição de que se um nó é muito importante, suas relações também devem ser importantes. Analogamente se um predicado aparece muito entre nós relevantes, ele deve ser muito relevante. O mesmo vale se ele aparecer muito entre nós de baixa relevância. Agora se um predicado aparece entre nós relevantes e não relevantes qual seria sua importância ?

Para decidir sua importância, distribuem-se os valores de centralidades para as arestas através de uma média ponderada. No cálculo são utilizados a frequência de ocorrência e o valor de centralidade do nó.

$$C_r = \frac{\sum_n^{n \in G} C_n * F_r^n}{F_r}$$

A relevância C_r de uma certa relação r é calculada somando, para cada nó n do grafo G, o produto entre a relevância C_n do nó n pelo número F_r^n de vezes em que a relação r aparece relacionada a n. Por fim, dividimos esse valor pelo número de ocorrências F_r da relação r em todo o grafo. A fórmula descrita representa a intuição do Algoritmo 1 exemplificado.

Algoritmo 1: Determina a relevância de cada predicado

DisTImport (N, P)
inputs : Valor de Centralidade para cada nó, Predicados entre dois vértices
output : Um conjunto de predicados com suas importâncias
Calcula a frequência para cada predicado.
Calcula a importância de cada predicado utilizando uma média ponderada pela soma das importâncias de cada nó relacionado.
Retorna o conjunto de predicados com suas importâncias.

4.2 Métricas de Correlação

São utilizadas diversas técnicas para cálculo de relevância e cada técnica gera valores de importância distintos, por consequência são gerados vários ranqueamentos diferentes. Não deseja-se descartar o resultado de uma métrica. Visto que descartar o resultado é descartar e semântica associada à técnica e essa não é a intenção. Precisa-se, de alguma maneira, decidir qual ou quais métricas utilizar.

Para isso, o Pytology possui o módulo de métricas de correlação, esse módulo permite correlacionar os resultados gerados. As técnicas utilizadas de Spearman (ZAR, 1998), Kendall (ABDI, 2007) e Pearson (SEDGWICK, 2012) buscam mensurar as relações de ordenação. Spearman avalia relações lineares e não lineares, Pearson avalia relações lineares e Kendall utiliza avaliação ordinal.

As métricas não dizem qual técnica utilizar e tão pouco quais predicados são os mais relevantes. Elas apenas geram valores de correlações. A decisão final sobre quais os predicados mais relevantes e qual ou quais técnicas serão a utilizadas cabe ao utilizador da ferramenta.

5 RESULTADOS

Os experimentos realizados partiram da execução do Pytology sobre base de dados disponíveis na internet, em particular no Datahub.io. Executamos os cálculos de centralidades Betweenness(**B**), Closeness(**C**), Eigenvectors(**E**), Katz-Eigenvector(**K**) e Pagerank(**PR**). Foi mencionado que o Betweenness pode ser executado diretamente sobre as arestas, por isso utilizamos duas abordagens do método. Na primeira, aplicamos a métrica sobre os nós e distribuímos a relevância, assim como feito nas outras medidas. Chamamos essa abordagem de Betweenness Distribuído(**BD**). Na segunda, aplicamos o cálculo diretamente para as arestas. Os resultados foram ordenados de acordo com suas relevâncias e dispostos em tabelas.

5.1 Resultados do Experimento sobre base de dados Nobeis

Tabela 1 – Top 10 relações 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
6	type	share	category	type	share	share
7	category	prizeFile	sameAs	field	category	category
8	laureate	awardFile	laureate	prizeFile	field	seeAlso
9	field	sameAs	birthPlace	awardFile	seeAlso	fileType
10	awardFile	laureate	motivation	sameAs	fileType	subClassOf

Podemos perceber 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* 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* são bem importantes, pois elas definem a representação textual de um elemento e o tipo de dado que ele é. Os resultados apresentam tanto predicados mais específicos do domínio quanto apresenta os padronizados, é interessante sabermos que as métricas

conseguem capturar os dois e compara-lós.s

As Tabelas 2, 3 e 4 referem-se às correlações entre os valores das centralidades obtidos de acordo com, respectivamente, Spearman, Pearson e Kendall. As matrizes apresentadas tendem a ser simétricas uma vez que comparamos duas vezes dois ranqueamentos distintos, é aceitável pequenas variações nos resultados, uma vez que grandes discrepâncias evidenciam possíveis erros de implementação.

Tabela 2 – Correlação entre as relevâncias de acordo com Spearman

	B	BD	C	E	K	PR
B	1.000,0	914,0	751,0	890,0	806,0	790,0
BD	914,0	1.000,0	770,0	946,0	811,0	811,0
C	751,0	770,0	1.000,0	729,0	823,0	793,0
E	890,0	946,0	729,0	1.000,0	843,0	847,0
K	806,0	811,0	823,0	843,0	1.000,0	992,0
PR	790,0	811,0	793,0	847,0	992,0	1.000,0

Valores de medida
729,0 1.000,0

Tabela 3 – Correlação entre as relevâncias de acordo com Pearson

	B	BD	C	E	K	PR
B	1.000,0	991,0	257,0	971,0	985,0	990,0
BD	991,0	1.000,0	354,0	981,0	993,0	985,0
C	257,0	354,0	1.000,0	294,0	334,0	307,0
E	971,0	981,0	294,0	1.000,0	989,0	957,0
K	985,0	993,0	334,0	989,0	1.000,0	986,0
PR	990,0	985,0	307,0	957,0	986,0	1.000,0

Valores de medida
257,0 1.000,0

Tabela 4 – Correlação entre as relevâncias de acordo com Kendall

	B	BD	C	E	K	PR
B	1.000,0	769,0	586,0	739,0	673,0	642,0
BD	769,0	1.000,0	630,0	838,0	693,0	706,0
C	586,0	630,0	1.000,0	575,0	646,0	624,0
E	739,0	838,0	575,0	1.000,0	733,0	740,0
K	673,0	693,0	646,0	733,0	1.000,0	963,0
PR	642,0	706,0	624,0	740,0	963,0	1.000,0

Valores de medida
575,0 1.000,0

Como Kendall utiliza os valores ordinais para calcular as correlações, 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, alguns ranks estão bem correlacionados, como Katz-Eigenvector, Pagerank e Eigenvectors, o que já era esperado, uma vez que suas medidas de centralidade são baseadas em vetores de pesos. Outro detalhe é que o Betweeness e o Betweeness Distribuído estão bem correlacionados. Esses resultados ajudam a fortalecer a ideia de que a distribuição das centralidades para as arestas mantém consigo a semântica da métrica utilizada. Vale evidenciar que a técnica do Closeness foi a que se apresentou nos piores valores de correlações com as técnicas, fato também comprovado pelo ranqueamento gerado pela técnica, que foi o mais distinto dos todos.

6 CONCLUSÕES E TRABALHOS FUTUROS

Os resultados apresentados se mostraram bem satisfatórios. Uma vez que conseguimos boas correlações entre medidas semelhantes. Além disso podemos perceber que predicados muito importantes se apresentam bem posicionados em quase todas as métricas. A métrica do Closeness é a que se apresenta mais distinta de todas as outras, mas é esperado, uma vez que ela é uma das mais distintas em temos de definição. Percebemos também que possuímos predicados com diferentes semânticas contidos no mesmo ranqueamento, não há qualquer distinção semântica entre eles. Os resultados podem ser utilizados para comparar predicados com semântica semelhantes e percebermos tendencia de importância entre os predicados.

Podemos ressaltar pontos de melhoria do Pytology. Atualmente Pytology executa sobre uma instância de dados, podemos utilizar as métricas de centralidade sobre esquema da base de dados também. A vantagem é gerar um pré-valor de relevância ou uma nova medida para pesar no cálculo de distribuição, além de ser mais rápido a execução, visto que o esquemas da base se apresentam bem menores que as instâncias. O Pytology também não gera um ranqueamento final, poderíamos utilizar os resultados das métricas de comparação para gerar um resultado final utilizando esquema de votação com pesos. As métricas do Pytology ainda não exploram a semântica das relações, apenas analisam a topologia da instância de dados para tentar gerar uma métrica de relevância.

REFERÊNCIAS

ABDI, H. The kendall rank correlation coefficient. **Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA**, Citeseer, p. 508–510, 2007.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific american**, JSTOR, v. 284, n. 5, p. 34–43, 2001.

BONDY, J. A.; MURTY, U. S. R. *et al.* **Graph theory with applications**. [S.l.]: Citeseer, 1976. v. 290.

ELBASSUONI, S.; BLANCO, R. Keyword search over rdf graphs. In: ACM. **Proceedings of the 20th ACM international conference on Information and knowledge management**. [S.l.], 2011. p. 237–242.

FREEMAN, L. C. Centrality in social networks conceptual clarification. **Social networks**, North-Holland, v. 1, n. 3, p. 215–239, 1978.

MIRIZZI, R.; NOIA, T. D. From exploratory search to web search and back. In: ACM. **Proceedings of the 3rd workshop on Ph. D. students in information and knowledge management**. [S.l.], 2010. p. 39–46.

MUSETTI, A.; NUZZOLESE, A. G.; DRAICCHIO, F.; PRESUTTI, V.; BLOMQVIST, E.; GANGEMI, A.; CIANCARINI, P. Aemoo: Exploratory search based on knowledge patterns over the semantic web. **Semantic Web Challenge**, v. 136, 2012.

NUZZOLESE, A. G.; GANGEMI, A.; PRESUTTI, V.; CIANCARINI, P. Encyclopedic knowledge patterns from wikipedia links. In: SPRINGER. **International Semantic Web Conference**. [S.l.], 2011. p. 520–536.

ROA-VALVERDE, A. J.; SICILIA, M.-A. A survey of approaches for ranking on the web of data. **Information Retrieval**, Springer, v. 17, n. 4, p. 295–325, 2014.

SEDGWICK, P. Pearson's correlation coefficient. **BMJ: British Medical Journal (Online)**, BMJ Publishing Group LTD, v. 345, 2012.

WEST, D. B. *et al.* **Introduction to graph theory**. [S.l.]: Prentice hall Upper Saddle River, 2001. v. 2.

ZAR, J. H. Spearman rank correlation. **Encyclopedia of Biostatistics**, Wiley Online Library, 1998.