Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Extração de novas parcerias entre pesquisadores utilizando Ontologia

Welton de Abreu Henriques

JUIZ DE FORA
NOVEMBRO, 2018

Extração de novas parcerias entre pesquisadores utilizando Ontologia

Welton de Abreu Henriques

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Regina Maria Maciel Braga

JUIZ DE FORA

NOVEMBRO, 2018

Extração de novas parcerias entre pesquisadores utilizando Ontologia

Welton de Abreu Henriques

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Regina Maria Maciel Braga
Doutora em Engenharia de Sistemas e Computação

Victor Ströele de Andrade Menezes
Doutora em Engenharia de Sistemas e Computação

Jose Maria Nazar David
Doutora em Engenharia de Sistemas e Computação

JUIZ DE FORA

28 DE NOVEMBRO, 2018

Aos meus amigos e irmãos.

Aos pais, pelo apoio e sustento.

Resumo

A cada avanço da humanidade, o conhecimento se torna mais aprofundado, sendo necessário maior número de esforços no avanço da tecnologia. Uma possibilidade é a construção de parcerias entre cientistas, pois gera intercâmbio de conhecimento e também geração de novas pesquisas. O Brasil não se diferencia nesse aspecto em relação ao mundo, sendo cada vez mais necessário parcerias para a geração de inovação. Porém, um diferencial da ciência Brasileira é a presença de uma plataforma onde a maioria das produções científicas estão cadastradas e possuem acesso público, conhecida como Plataforma Lattes. Nesse sentido, a Plataforma Lattes se comporta como um repositório de dados, onde é possível, a partir de técnicas específicas, gerar novas informações. Observando a demanda de novas parcerias e o potencial atrelado a plataforma, esse trabalho busca utilizar os dados advindos da Plataforma Lattes, que sendo introduzidos em uma ontologia possibilite a execução dos fundamentos da Web Semântica. A partir da modelagem de inferências, os dados da ontologia geram informações sobre a similaridade entre diferentes pesquisadores auxiliando assim a criação de novas parcerias. Por fim, o trabalho exibe um grafo que pode ser explorado e estudado para identificar as similaridades gerais e individuais.

Palavras-chave: Web Semântica, Ontologia, Plataforma Lattes.

Abstract

With humankind development, knowledge has become deeper and demanding effort to keep technology up to date. A possibility is making partnerships between scientists to generate knowledge and new researches. Brazil is not different form the rest of the world in this context, making arranging more partnerships necessary to bring innovation. Nevertheless, Brazilian science has a differential which is a platform where most of scientific productions are registered e have public access, known as Plataforma Lattes. In this sense, Plataforma Lattes behaves as a repository of information where it is possible to generate information from specific techniques. Observing the demand for new partnerships and the potential linked to the platform, this work seeks to use the data coming from the Plataforma Lattes, which being introduced in an ontology allows the execution of the basics of the Semantic Web. From the inference modeling, the ontology data generate information about the similarity between different researchers, thus helping to create new partnerships. Finally, the paper displays a graph that can be explored and studied to identify general and individual similarities.

Keywords: Semantic Web, Ontology, Plataforma Lattes.

Agradecimentos

Primeiramente agradeço ao meu pai e minha mãe por sempre me incentivarem a estudar e apoiar minhas escolhas respeitando minha individualidade. Também agradeço por sacrificarem parte dos esforços para que eu pudesse gastar boa parte do meu tempo estudando e aprendendo.

Agradeço a meu irmão por sempre me apoiar e também a todas as pessoas próximas que me auxiliaram de alguma forma durante toda minha graduação.

Agradeço a Regina Maria Maciel Braga Villela pela orientação, paciência, atenção e por ter me ensinado conhecimentos que vou levar para vida toda.

Agradeço a todos os docentes e funcionários que participaram da minha formação e contribuíram para meu crescimento profissional e pessoal.

Por último agradeço a Universidade Federal de Juiz de Fora por propiciar uma vivência acadêmica que meu deu liberdade de participar de tantos projetos diferentes, por aprender, por conviver com a diversidade e por conhecer tantas pessoas incríveis.

O que me preocupa não é o grito dos maus, é o silêncio dos bons.

Martin Luther King

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

WWW	World Wide Web
W3C	World Wide Web Consortium
URI	Identificador de Recurso Uniforme
HTTP	Protocolo de Transferência de Hipertexto
RDF	Resource Description Framework
XML	Extensible markup language
DTD	Definição de Tipo de Documento
OWL	Web Ontology Language
SWRL	Semantic Web Rule Language
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
HTML	Linguagem de Marcação de Hipertexto
UFJF	Universidade Federal de Juiz de Fora

1 Introdução

Atualmente com os benefícios advindos da Web, qualquer pessoa pode ter acesso a uma gama enorme de informações e também a possibilidade de se relacionar com pessoas de todos os cantos do planeta. Além de ter o acesso, essas pessoas também podem contribuir produzindo ainda mais conteúdo e enriquecendo essa grande base de dados chamada Web.

A Web que conhecemos hoje possui uma vasta quantidade de informação armazena, e é caracterizada por possuir um crescimento exponencial de dados a cada dia de acordo com Lopes (2004). No entanto, essa quantidade de informação não consegue ser processada totalmente, pois o crescimento é maior do que a capacidade de análise humana.

Nesse sentido é necessário que a Web atual evolua para outro patamar de forma que também as máquinas possam analisar essa informação, pois são mais eficientes e rápidas que os seres humanos. Essa proposta da extensão da Web foi concebida por Berners-Lee et al. (2001) com o conceito de Web Semântica.

Com a introdução da Web Semântica e a representação dos dados através de ontologia no cenário atual, os sistemas de buscas serão mais eficientes, nesse caso em vez de só indexar sites, os sistemas vão poder percorrer as informações e extrair o necessário de forma melhor que os métodos utilizados atualmente. Visualizando os benefícios, este trabalho propõe a utilização do potencial da Web Semântica aplicado no cenário da ciência brasileira de forma a gerar novas parcerias entre pesquisadores

Uma das formas que os pesquisadores possuem para produzir trabalhos significativos é a construção de parcerias. Entretanto, esse construção ocorre de forma manual, seja conhecendo novos trabalhos em eventos ou também por indicações de nome através de outros pesquisadores.

Nessa busca de novas parcerias, um local que é consultado para avaliar o perfil profissional e pessoal é a Plataforma Lattes[1]. A Plataforma Lattes possui um currículo detalhado e, geralmente, atualizado de toda a experiência de um determinado pesquisador. O currículo proposto pela Plataforma Lattes é um padrão utilizado no Brasil,

[1] http://lattes.cnpq.br/

consequentemente todo pesquisador possui um currículo Lattes.

Apesar disso, a Plataforma Lattes somente oferece uma busca direcionada por pessoa, por instituição ou por assunto. Na prática a busca direcionada por assunto não é satisfatória, pois é necessário que o usuário abra cada currículo listado e visualmente analise a informação. Outro ponto também é que a busca por assuntos multidisciplinares não traz resultados consistentes.

Vale ressaltar que cada currículo cadastrado possui várias informações relevantes. Essas informações, se analisadas de forma correta, podem auxiliar na criação de um perfil que representa um determinado pesquisador. Pensando nesse ponto, com os vários perfis existentes na Plataforma Lattes, é possível auxiliar em diminuir a distância entre pesquisadores que não se conhecem.

1.1 Justificativa

A troca de conhecimento e experiência entre pesquisadores é um dos grandes fatores que podem influenciar o crescimento e fortalecimento da ciência brasileira. Por isso, esse trabalho propõe uma solução acessível que auxilie na sugestão de criação de novas redes de relacionamento entre pesquisadores.

1.2 Objetivos gerais e específicos

Este trabalho tem por objetivo principal a extração de novas propostas de parcerias a partir da análise de similaridade entre perfis de pesquisadores. Para esse objetivo serão utilizados conceitos e técnicas da Web Semântica aplicados a base de dados da Plataforma Lattes.

Outro objetivo é a divulgação do potencial atrelado na utilização dos conceitos de Web Semântica e Ontologia em determinados cenários reais.

1.3 Metodologia

Para organizar a proposta do projeto de monografia, a sua execução foi dividida em três etapas elencadas a seguir:

Revisão bibliográfica: Nesta primeira etapa foi feita uma revisão da bibliografia buscando trabalhos relacionados e explicando conceitos que possam engrandecer e auxiliar nas atividades subsequentes.

Desenvolvimento do projeto: Durante essa etapa serão modelados os dados e desenvolvido o sistema que suportará a proposta deste trabalho.

Análise dos resultados: Na última etapa a proposta desenvolvida foi submetida a uma base de dados reais onde será avaliado os seus resultados.

2 Fundamentação teórica

2.1 Introdução

Este capítulo tem como objetivo fornecer o embasamento teórico relevante para esta monografia. Ele está organizado da seguinte forma: na seção 2.1 são abordados a definição e alguns conceitos de Web Semântica; Na seção 2.2 são apresentados os fundamentos de ontologia além de ferramentas de auxílio, implementações e descrição de algumas linguagens utilizadas. Na seção 2.3 é mostrado um histórico da Plataforma Lattes e sua importância para os dias atuais. Por último, na seção 2.4 são realizadas as considerações finais do capítulo.

2.2 Web Semântica

Proposta por Berners-Lee (1989) a World Wide Web (WWW), ou simplesmente Web, inicialmente concebida para ser um sistema interno de gerenciamento de documentos em hipertexto, armazenados em poucos servidores na rede, acabou se tornando o que chamamos de rede mundial de computadores na qual as informações disponibilizadas crescem em ordem exponencial a cada dia.

Como forma de desenvolver protocolos e diretrizes para organizar e auxiliar o crescimento da Web, foi criada uma comunidade internacional denominada World Wide Web Consortium (W3C[2]), liderada por Tim Berners-Lee e por Jeffrey Jaffe. A comunidade tem como missão a busca do potencial máximo que o WWW pode oferecer.

Atualmente, a quantidade de documentos acessíveis na Web abrange várias áreas distintas e têm objetivos diferentes dependendo do contexto, seja o entretenimento, divulgação científicas, reportagens, opiniões pessoais e etc. No entanto, boa parte do conteúdo disponibilizado é voltado para a leitura humana, e não para que computadores possam manipular a informação, fazendo com que boa parte desse conteúdo disponibi-

[2]https://www.w3.org/

lizado não possa ser aproveitado de forma eficiente de acordo com Berners-Lee et al. (2001).

Como solução para esse contexto surge uma alternativa proposta por Berners-Lee et al. (2001): "A Web Semântica não é uma Web separada, mas uma extensão da atual, na qual a informação é fornecida com seu significado bem definido, permitindo que computadores e máquinas possam trabalhar melhor em cooperação". O termo Web Semântica é também relacionado a um conjuntos de tecnologias que aplicadas criam a extensão da Web. Como forma de ilustrar, Tim Berners-Lee criou a Figura 2.1 que mostra em camadas as tecnologias utilizadas. Cada camada superior utiliza a capacidade das inferiores.

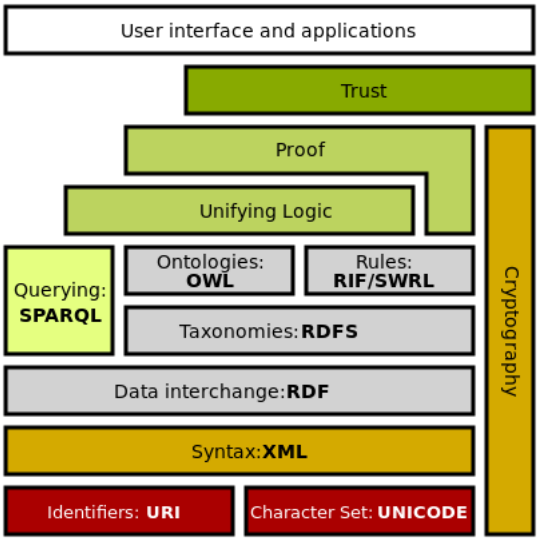

Figura 2.1: Semantic Web Stack

Outra forma de nomenclatura proposta é a divisão em 3 partes históricas da Web sendo: *Web 1.0*, *Web 2.0* e *Web 3.0*. A *Web 1.0* retoma ao início da história da Web, onde boa parte do conteúdo era disponibilizado de forma estática, produzidos majoritariamente por instituições e empresa, com pouca interação entre usuários. O termo *Web 2.0*, que é direcionado ao estado atual da Web, proposto por O'reilly (2005), possui como principal características que o conteúdo é criado, produzido e compartilhado por usuários. Um dos pontos característicos é o nascimento das redes sociais mais robustas e com maior comunidade. Por último vem a *Web 3.0*, que é um movimento que tem como principal característica a inclusão da semântica na Web e que, consequentemente, tem a ver com

adesão real dos conceitos da Web Semântica.

Um fator importante para entender o potencial da Web Semântica é o conceito de Linked Data proposto por Berners-Lee (2017). Resumidamente, a proposta é que a Web se torne uma grande base de dados conectados de forma que as pessoas e as máquinas possam explorar os dados. Para que isso ocorra de forma padronizada é necessário que existam algumas regras, por isso foram propostas 4 regras principais:

1. Use identificador de recurso uniforme (URI) para nomear as coisas;
2. Use Protocolo de Transferência de Hipertexto(HTTP) URIs para que pessoas possam procurar esses dados;
3. Quando alguém procurar uma URI, disponibilize informações úteis, usando padrões como RDF;
4. Inclua links para outras URIs de forma que exija relacionamento entre as fontes de informação

Para ilustrar melhor os resultados no cotidiano da utilização da Web Semântica e Linked Data podemos pensar em um cenário fictício onde um casal está planejando uma viagem de férias para fora do Brasil. Atualmente, existem algumas formas de realizar esse planejamento. Uma opção é entrar em contato com uma agência de viagem e comprar um pacote completo, porém boa parte desses pacotes não podem ser totalmente customizados seguindo seus gostos ou suas necessidades. Logo uma opção alternativa é pesquisar em várias agências até achar uma que melhor lhe agrade. No entanto, essa escolha pode não ser a mais adequada a suas características. Outra forma é fazer manualmente o planejamento, porém, vai ser necessário entrar em contato com empresas de venda passagens aéreas, em locais de hospedagens, em recomendações turísticas e de alimentação. Além da escolha também é necessário se atentar a questão de horário e custo da viagem. Dependendo do local, esse trabalho se torna improdutivo e estressante.

Com a proposta de Linked Data e Web Semântica basta o usuário descrever quais são suas exigências e expectativas para a viagem. Através dessa descrição o motor de busca entra em ação cruzando informações dos relacionamentos semânticos em toda a

Web de forma a mostrar o máximo possível de resultados que se adequem às exigências do usuário. Por essa razão é importante que o máximo de informações presente na Web estejam incluídas no conceito de Linked Data.

Para adoção dos conceitos citados, é necessário que exista uma forma de representação dos dados de forma a facilitar a sua interpretação. Nesse sentido são utilizados os conceitos de ontologia para estabelecimento da conceitualização do domínios do conhecimento.

2.3 Ontologia

Ontologias são os elementos centrais da Web Semântica, pois têm o trabalho de modelar o domínio de conhecimento de forma a possibilitar uma representação semântica. A ontologia possui vários significados dependendo da área estudada. Sua primeira aparição foi com Aristóteles o qual afirmava que a ontologia é a ciência que estuda o ser enquanto ser. No entanto para a área de Ciência da Computação a ontologia possui outro significado que é : "uma explícita especificação de uma conceitualização " (GRUBER, 1993). Como forma de interpretação Breitman (2005) propõe que: "conceitualização representa um modelo abstrato de algum fenômeno que identifica os conceitos relevantes para o mesmo; ser explícita significa que os elementos e suas restrições estão claramente definidos; e ser formal (especificação) significa que a ontologia deve ser passível de processamento automático".

2.3.1 Representação de conhecimento

Na construção de uma ontologia é necessária a escolha de uma linguagem de descrição que possui o trabalho de descrever o conhecimento, e também a escolha de um motor de inferência que vai entender e extrair novas informações. De acordo com Morais e Ambrósio (2007), nem todas as ontologias têm a mesma estrutura, mas a maioria delas possui alguns elementos básicos, como:

- Classes: Normalmente organizadas em taxonomias, as classes representam algum tipo de interação da ontologia com um determinado domínio podendo possuir ca-

racterísticas de herança entre elas.

- Relações: Representam o tipo de interação entre os elementos do domínio (classes).
- Axiomas: São utilizados para modelar sentenças consideradas sempre verdadeiras.
- Instâncias: São utilizadas para representar elementos específicos, isto é, os próprios dados da ontologia.

A base da estrutura de dados na Web Semântica foi proposta pela instituição W3C através do conceito de triplas, também conhecida como vocabulário Resource Description Framework (RDF). As triplas são o formato mais básico que temos para relacionar e estruturar os dados na Web Semântica. Cada tripla possui 3 elementos, sendo eles: sujeito, predicado e objeto. Na Figura 2.2 está representado visualmente o formato de um tripla:

Figura 2.2: Estrutura de tripla (EIS, 2017)

Essa declaração expressa a relação entre duas fontes. O sujeito e o objeto representam as duas fontes que serão relacionadas pelo predicado, que representa a natureza dessa relação. Ou seja, as linguagens de descrição expressam declarações acerca de dados de forma e descreve-lo e de relaciona-lo a outros dados. É possível visualizar um grafo sendo a representação do aglomerado de triplas descritas por uma determina linguagem de descrição. Na Figura 2.3 está representada uma ontologia através de um grafo.

2.3.2 Linguagens de descrição

A seguir são listadas algumas linguagens utilizadas para descrição de informações.

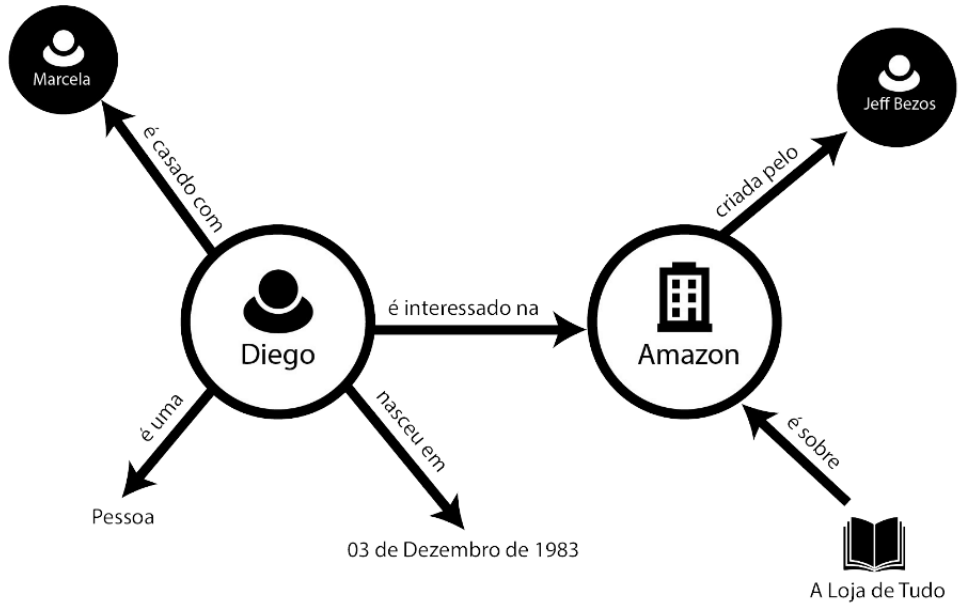

Figura 2.3: Grafo de relações de uma ontologia (EIS, 2017)

XML

Extensible markup language (XML[3]) é um formato de texto baseado em marcações (tags), criado pela W3C. Os dados são representados como uma árvore e possuem estruturas que identificam os elementos. Atualmente o XML é adotado em inúmeras aplicações que envolvem a troca de dados entre sistemas, por ser de fácil manuseio.

Para definir as estruturas de documentos XML, são utilizadas duas outras linguagens: DTD e XML Schema. A primeira é mais antiga e mais restrita, como por exemplo tipo de dados delimitados apenas como textos. Já XML Schema oferece uma linguagem mais rica porque ela mesma é baseada em XML, assim provendo um importante ganho de legibilidade. Na consulta dos dados é normalmente utilizada a linguagem XPath pois possui resposta rápida e é de fácil compreensão.

RDF e RDF Schema

De acordo com W3C[4] o Resource Description Framework (RDF) é uma estrutura para representar informações na Web. A forma de representação de dados se dá através de triplas como exemplificado na seção anterior. O design do RDF destina-se a atingir os seguintes objetivos:

[3]https://www.w3.org/XML/
[4]https://www.w3.org/TR/rdf-concepts/

- Ter um modelo de dados simples
- Ter semântica formal e inferência provável
- Usar um vocabulário extensível baseado em URI
- Usar como uma das possibilidades uma sintaxe baseada em XML
- Suportar o uso de tipos de dados de esquema XML
- Permitir que alguém faça declarações sobre qualquer recurso

A Resource Description Framework Schema (RDF Schema) é uma extensão da linguagem original RDF. Com RDF Schema é possível criar um vocabulário para a declaração de taxonomias de classes e propriedades, além da possibilidade de descrição simplificada de domínio e alcance das propriedades das classes. Costa (2009) em seu trabalho apresenta a Figura 2.4 que exemplifica a diferença entre RDF e RDF Schemas.

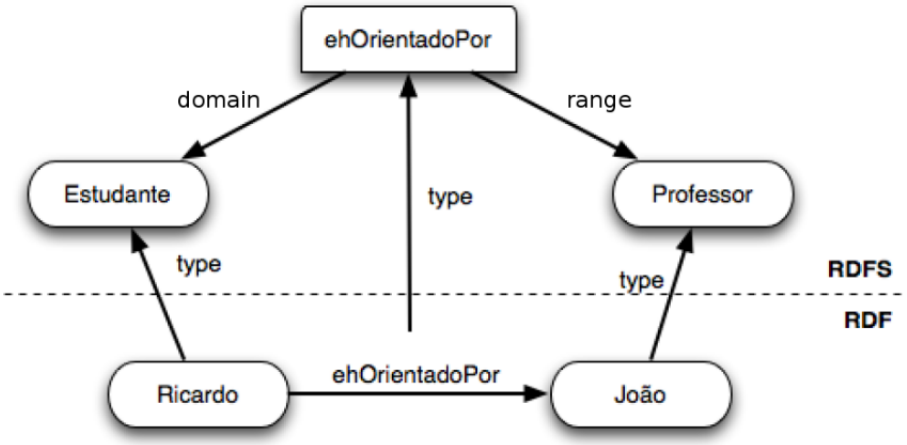

Figura 2.4: Camada RDF e RDF Schema (COSTA, 2009)

DAML+OIL e OWL

Web Ontology Language (OWL[5]) de acordo com W3C é estabelecida como linguagem padrão para descrição de ontologias para a Web Semântica. A OWL tem mais facilidades para expressar significado e semântica do que as linguagens XML, RDF e RDF Schemas.

[5]https://www.w3.org/TR/owl-features/

Ela é fruto da revisão da linguagem DAML+OIL que foi proposta a partir das fusão da linguagens de descrição DAML-ONT e OIL.

A grande diferença entre OWL e RDF Schemas é que a linguagem OWL além de oferecer as funcionalidades da antecessora também acrescenta mais vocabulários para descrever propriedades e classes. Com OWL é possível descrever que classes são disjuntas, descrever cardinalidade, igualdade, classes enumeradas e também características de propriedades como simetria. No entanto, pela sua complexidade foi necessário dividir a linguagem OWL em três sub-linguagens em que cada uma foi arquitetada segundo diferentes perspectivas:

OWL-Full: É a considera como a linguagem OWL completa com compatibilidade total ao RDF Schemas. Porém , é indecidível, ou seja, não existe atualmente um motor de inferência que seja completo e eficiente para ela.

OWL-DL: É voltada para quem busca a máxima expressividade, mantendo a completude computacional. OWL DL inclui todas as construções da linguagem OWL, mas elas podem ser usadas somente sob certas restrições. Infelizmente não possui compatibilidade total com RDF Schemas

OWL lite: Um conjunto mais restrito que o OWL-DL, que possui menos expressividade porém é mais fácil de ser compreendida pelo usuário e pelos motores de inferência. É um linguaguem utilizada também para migração de outras linguagem para a OWL.

OWL 2

O W3C evoluiu a linguagem OWL, para a linguagem denominada OWL 2[6] em 2009. A proposta foi acrescentar novas funcionalidades mantendo compatibilidade com a versão anterior agora denominada OWL 1. Algumas das funcionalidades adicionadas foram a seleção de chave para identificar um indivíduo, cadeias de propriedades, tipos de dados mais ricos e adição de propriedades assimétricas, reflexivas e disjuntas.

Foram definidos também três novos perfis (ou sub-linguagens):

OWL 2 EL: É capaz de executar algoritmos em tempo polinomial para todas

[6]https://www.w3.org/TR/owl2-overview/

as tarefas padrões de raciocínio. É particularmente adequada para aplicações com grandes ontologias, nas quais a capacidade de expressão pode ser trocada por garantia de desempenho.

OWL 2 QL: Permite que as consultas conjuntivas sejam respondidas no LogSpace usando a tecnologia de banco de dados relacionais. É adequada para aplicações em que as ontologias relativamente leves são usadas para organizar um grande número de indivíduos e onde é útil ou necessário acessar os dados diretamente através de consultas relacionais como, por exemplo, via SQL.

OWL 2 RL: Permite a implementação de algoritmos de raciocínio de tempo polinomial utilizando tecnologias de banco de dados estendidas diretamente em triplos de RDF; é particularmente adequado para aplicações em que ontologias relativamente leves são usadas para organizar um grande número de indivíduos.

SPARQL

Segundo a especificação[7] ,"SPARQL (atualmente na versão 1.1) é um conjunto de especificações que provê linguagens e protocolos para consultar e manipular grafos RDF disponíveis na Web ou em uma base de dados RDF. Assumindo que os dados estejam carregados em um serviço SPARQL, a linguagem de consultas SPARQL 1.1 pode ser usada para formular consultas que vão desde um simples esquema de reconhecimento de padrões até consultas complexas".

Vale ressaltar que SPARQL consegue trabalhar também com consultas em arquivos OWL .É capaz de compreender não apenas a sintaxe, mas também o modelo de dados do RDF e a semântica dos vocabulários do RDF Schemas e OWL.

A linguagem de busca SPARQL é apoiada no reconhecimento de padrões de grafos, logo, o mais simples deles é uma única tripla. A sintaxe se assemelha com a linguagem SQL que é utilizada para manipular informações em banco de dados. Para ilustrar esse conceito, considere o exemplo de uma query SPARQL a seguir.

```
1 PREFIX foaf: <http://xmlns.com/foaf/0.1/>
2 SELECT ?nomeCompleto
3 WHERE {
```

[7] https://www.w3.org/TR/sparql11-query/

```
4 ?pessoa foaf:nome ?nome .
5 ?pessoa foaf:conhece ?amigo .
6 } GROUP BY ?pessoa ?nome
```

2.3.3 Ferramentas e implementações

Editores de ontologia

Como forma de facilitar a criação e visualização de modelos de domínio com ontologias, existem software gráficos que buscam auxiliar nesse trabalho. Um exemplo é o Protégé.

O Protégé[8] é um software *open-source* desenvolvido pelo Centro Stanford de Pesquisa em Informática Biomédica (BMIR), e foi concebida para auxiliar no processo de criação de modelos de domínio através de ontologias. Com uma interface visual é possível editar, visualizar grafos, fazer consulta SPARQL, criar instâncias entre outras funcionalidade. Os arquivos podem ser salvos em formato RDF/XML, Turtle, OWL/XML, OWL functional, Manchester OWL, OBO, Latex e JSON-DL além de possuir compatibilidade com OWL 2. Sobre os motores de inferência, o software possui suporte para ELK, FaCT++, HermiT, Ontop, Pellet e jcel. Esse software possui vasta documentação além de suportar plugin criado por terceiros. Na Figura 2.5 é mostrado a interface do software Protégé.

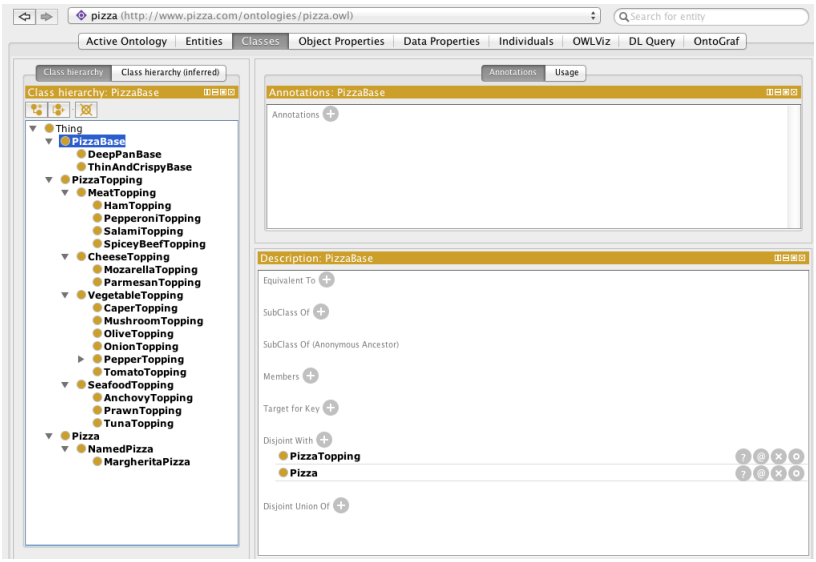

Figura 2.5: Interface Protégé (HORRIDGE et al., 20011)

[8]https://protege.stanford.edu/

Frameworks

Como forma de auxiliar a manipulação de ontologia em termos de código, existem alguns Frameworks que auxiliam nessa atividade.

Apache Jena[9] é um Framework Java para construção e manipulação de aplicações utilizando Web semântica. Sobre a responsabilidade da Apache Software Foundation, o Apache Jena busca auxiliar na criação de aplicações ferramentas e servidores para Web Semântica e *Linked-data*. Inicialmente, foi voltado para arquivos RDF, então o suporte para OWL 2 é limitado.

OWL API[10] é uma API Java *open source* para criação, edição e manipulação de ontologias em OWL. Possui compatibilidade com OWL 2 e possui suporte para 8 tipos de motores de inferência em sua versão 4.

2.3.4 Motores de inferência

Os motores de inferência exercem um trabalho muito importante na extração de novas informações na Web Semântica. Segundo Dentler et al. (2011) um motor de inferência (ou *reasoner*) é um programa que infere consequências lógicas a partir de um conjunto de afirmações ou axiomas explicitamente declarados. Uma parte dos motores de inferência já possuem suporte a OWL 2. Vale destacar alguns nomes de motores utilizados como FaCT++[11], HermiT[12], Pellet[13], JFact[14].

2.4 Plataforma Lattes

A Plataforma Lattes [15] é uma base de dados pública que propõe a integração de bases de dados de currículos, de grupos de pesquisa e de instituições em um único sistema de informações. Essa plataforma foi desenvolvida e está sob a responsabilidade do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)[16], agência vinculada ao

[9] https://jena.apache.org/
[10] http://owlcs.github.io/owlapi/
[11] http://owl.man.ac.uk/factplusplus/
[12] https://www.cs.ox.ac.uk/isg/projects/HermiT/publications
[13] https://github.com/stardog-union/pellet
[14] http://jfact.sourceforge.net/
[15] http://lattes.cnpq.br/
[16] http://www.cnpq.br/Web/guest/pagina-inicial

governo federal do Brasil, que busca o fomento da pesquisa científica e tecnológica.

A base de dados Lattes contém mais de três milhões de currículos em seu sistema. Ao buscar pelo o currículo de alguém, o usuário tem acesso a uma página HTML onde estão listados os dados cadastrados. Existe a opção de baixar o currículo em formato XML que possui dados a mais que a versão visualizada em HTML.

A Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior (CONSCIENTIAS) atualmente é responsável pela manutenção das gramáticas XML da Plataforma Lattes. A CONSCIENTIAS é uma extensão da Comunidade Virtual Linguagem de Marcação da Plataforma Lattes (LMPL) criada no ano 2000 para ser responsável pela criação do XML da Plataforma Lattes.

Processos seletivos, concursos públicos, editais de instituições de incentivo a pesquisa, seleção para pós-graduações entre outras ocasiões utilizam o Currículo Lattes do candidato como forma de análise da experiência. Nesse sentido a Plataforma Lattes tem grande importância principalmente se tratando do cenário brasileiro de ensino, pesquisa e extensão.

No ano de 2017 a Plataforma Lattes iniciou o processo de migração gradativamente para uma nova aparência e com um sistema de busca reformulado que facilita a busca e visualização de resultados, porém até o momento desse trabalho ainda não foi finalizado o processo.

2.5 Considerações finais do Capítulo

Este capítulo apresentou o referencial teórico em que se baseia este trabalho. Mostrou-se a definição de Web Semântica e discorreu sobre a proposta de Linked Data. Também apresentou um breve resumo sobre ontologia e alguns artefatos que são utilizados em sua construção. Por fim, foram apresentadas informações relevantes sobre a Plataforma Lattes. No capítulo seguinte , serão apresentados trabalhos relacionados que serviram de inspiração para esse trabalho.

3 Trabalhos relacionados

Este capítulo tem como objetivo descrever e analisar trabalhos relacionados e que serviram de base para a proposta dessa monografia.

O capítulo está organizado da seguinte forma: Na seção 2.1 é abordada sobre uma proposta de ontologia para o currículo lattes chamada OntoLattes(BONIFACIO, 2002), na seção 2.2 é detalhada a ferramenta *Semantic Lattes*(COSTA, 2009). Dentro da seção 2.3 é analisado o software scripLattes(MENA-CHALCO; JUNIOR; MARCONDES, 2009) e sua proposta de gerar informações a partir dos dados contidos na Plataforma Lattes. Na seção 2.4 a ferramenta denominada SOS Lattes(GALEGO, 2013) é descrita e, na seção 2.5, é apresentado o trabalho de Análise de Redes Sociais Científicas(STROELE, 2012). Por último, na seção 2.6 são realizadas as considerações finais do capítulo.

3.1 OntoLattes

O OntoLattes é a ontologia especificada na linguagem DAM+OIL, proposta por Bonifacio (2002), que usa como estudo de caso o Currículo Lattes.

De acordo com o autor a proposta inicial do trabalho é apresentar, avaliar e permitir uma melhor compreensão de um conjunto de conceitos, linguagens e ferramentas que são usadas na Web Semântica. Para atingir esse propósito o trabalho é aplicado a um caso de dimensão e complexidade real que é o Currículo Lattes. Para melhor entendimento do OntoLattes a Figura 3.1 ilustra o fluxo de processo global.

O primeiro passo dado em direção a construção da ontologia para o Currículo Lattes, foi a análise do documento DTD da Comunidade Virtual LMPL disponibilizado online pela Plataforma Lattes. Através deste documento, foi possível retirar os dados para a construção do esboço inicial do modelo que serve como base para a construção da ontologia OntoLattes.

Durante esse processo de análise do DTD, verificou-se quais elementos e atributos do documento poderiam ser modelados como classes e quais poderiam ser definidos como

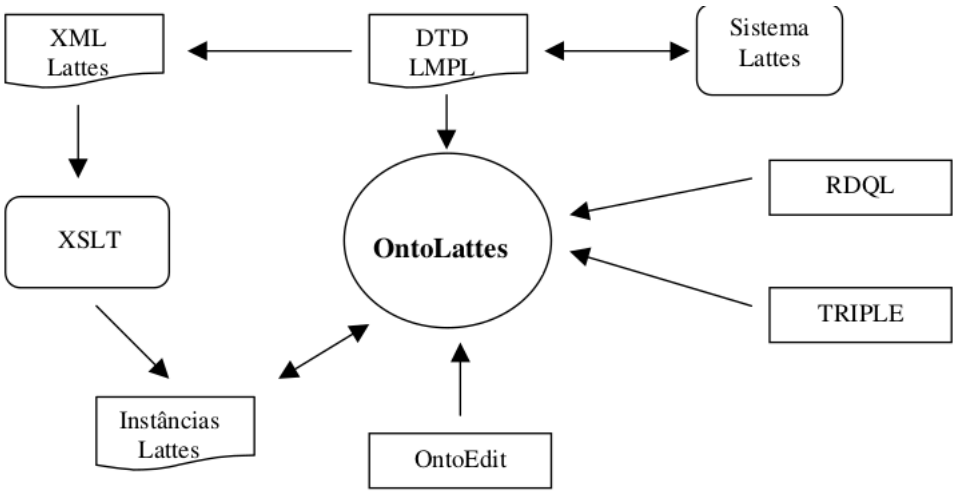

Figura 3.1: Fluxo de processo global (BONIFACIO, 2002)

propriedades no desenvolvimento da ontologia. Outro ponto de destaque é a verificação da cardinalidade dos elementos e dos tipos dos atributos prezando que a ontologia siga com fidelidade as terminologias propostas no DTD.

Assim, foram escolhidos alguns atributos pertencentes ao DTD para que se transformassem em classes na ontologia. Também durante o desenvolvimento nem todos os *slots*, relações e axiomas foram definidos na ontologia. Outro ponto adotado é que as classes, subclasses e propriedades definidas seguiram nomes idênticos aos atribuídos no DTD.

O processo de uso dos dados do documento para a especificação da ontologia ocorreu manualmente pois não existia na época uma ferramenta que faria uma migração satisfatória segundo Bonifacio (2002) . No entanto, na parte de edição e formulação da ontologia no formato DAML+OIL foi utilizada a ferramenta visual OntoEdit para auxiliar na construção.

Para o processo de instanciação foram utilizadas pequenas massas de dados de alguns currículos advindos do Plataforma Lattes. Na parte de instanciação o processo ocorreu de duas formas. Na primeira, o processo foi totalmente feito de forma manual e, na segunda forma, o processo foi semi-automatizado utilizando recursos XSLT. O XSLT é uma linguagem de transformação de documento XML para outro documento XML, como o DAML+OIL é uma extensão do modelo RDF Schemas que utiliza a sintaxe XML então é perfeitamente possível utilizar o XSLT para transformar o XML do Currículo Lattes

para o DAML+OIL.

Na etapa final de consulta da ontologia foram utilizadas as linguagem de consultas TRIPLE e a RDQL. A escolha deve-se ao fato destas duas linguagens possuírem características específicas para modelo RDF Schemas. Por extensão, também se aplicam a modelos DAML+OIL.

A abordagem de uma forma diferente de apresentação de dados do Currículo Lattes se mostrou importante e Nakashima (2004) aperfeiçou a abordando além de converter para o formato OWL. Vale ressaltar que o trabalho de Bonifácio foi talvez um dos primeiros trabalhos que aplicaram Web Semântica no contexto do Currículo Lattes em uma época que a área de Web Semântica dava seus primeiro passos.

3.2 Semantic Lattes

Costa (2009) em seu trabalho desenvolveu uma ferramenta Web chamada *Semantic Lattes*, que tem como função propor uma forma mais eficiente de consultas sobre o Currículo Lattes em comparação com o sistema utilizado atualmente na Plataforma Lattes.

O sistema atual de busca da Plataforma Lattes, consiste em o usuário escolher filtros pré-definidos pela plataforma para a direcionar os resultados. Existe também um campo textual onde o usuário pode colocar parte do nome da pessoa que está sendo buscada para restringir ainda mais os resultados. Vale ressaltar que a busca é sempre voltada a pessoas e não a produções pessoais.

Segundo Costa (2009) esta abordagem é imprecisa e pouco intuitiva. Em seu trabalho ele afirma que seria mais conveniente se o usuário pudesse realizar buscas através de perguntas escritas, como por exemplo: “Quais são os professores do Departamento de Engenharia de Computação da EPUSP?”.

Infelizmente adquirir esse tipo de resultado se torna muito difícil quando é utilizado banco de dados relacional. Um dos maiores dificultadores é pela inexistência de um relacionamento semântico de forma clara entre as tabelas de professores e departamento como exemplo. Partindo dessa observação ele propõe a utilização de conceitos e técnicas da Web Semântica para que as buscas e inferências de informações curriculares se tornem mais eficientes e precisas se adequando a essa nova perspectiva de busca.

Segundo o autor, o sistema pode ser dividido em dois fluxos: o fluxo de carga de currículos e o fluxo de consulta. Na Figura 3.2 são mostradas as atividades dos dois fluxos.

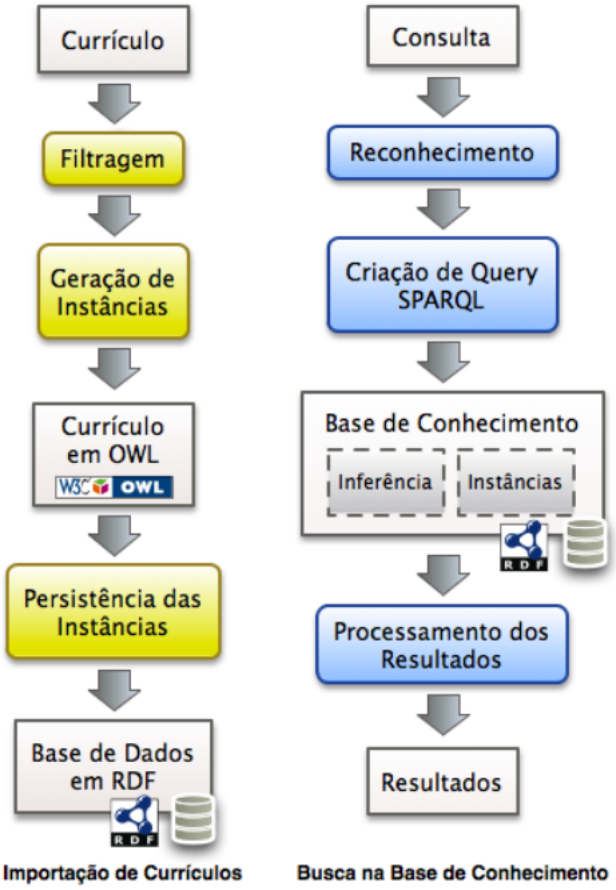

Figura 3.2: Fluxo de cada módulo da aplicação(COSTA, 2009)

No fluxo de carga de currículos o administrador faz o *upload* dos currículos em formatos XML disponibilizados na Plataforma Lattes. O sistema busca os atributos necessários nos currículos que serão inseridos como instâncias na ontologia. Para a representação das informações curriculares é utilizado um documento base formulado pelo o autor em formato OWL. Com as instâncias do currículo já em formato OWL são armazenados as triplas RDF, em uma ferramenta de persistência de dados chamada TDB. Com essa última atividade é finalizado assim o fluxo de carga.

O fluxo de consulta ocorre a cada vez que o usuário faz uma requisição de busca. Após ser digitada a pergunta e submetido o formulário, o sistema identifica as palavras chaves utilizadas e usa como base na criação das consultas em formato SPARQL. Com as

consultas SPARQL prontas o sistema busca entre as instâncias armazenadas utilizando o *framework* Jena, que provê interfaces de comunicações com o motor de inferência Pellet e a ferramenta de armazenamento de dados TDB. Com os resultados prontos o sistema organiza e exibe para os usuários a resposta da busca. Na Figura 3.3 é exibida a interface com os resultados da busca.

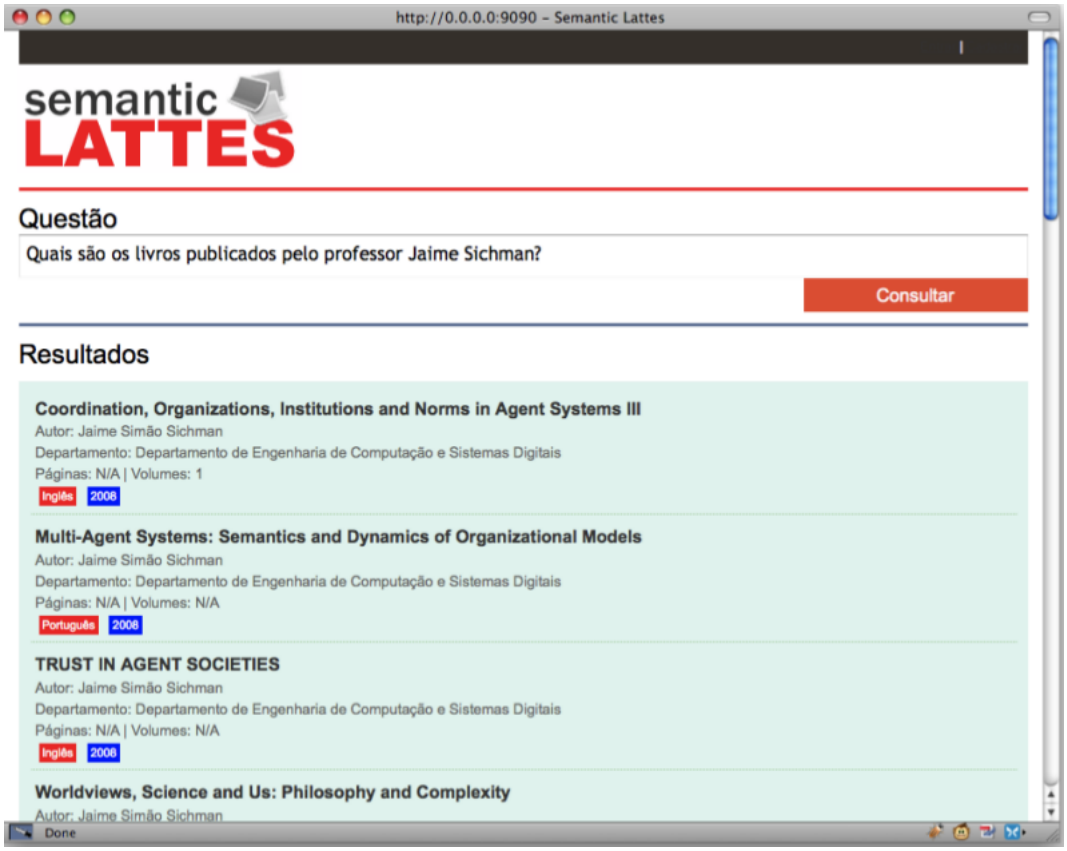

Figura 3.3: Interface com os resultados obtidos em consulta(COSTA, 2009)

Como forma de engrandecer o trabalho foi proposta também uma ontologia para relacionar as produções dos currículos com a base de dados Qualis[17] que classifica os periódicos por ordem de importância. Dessa forma o usuário também pode utilizar a classificação do Qualis como critério de busca. Um exemplo de busca seria: "Quais são os artigos Qualis B já publicados pelo professor Jaime?".

Foi obtido como resultado final do projeto uma ferramenta de consulta da base de dados da Plataforma Lattes baseada em ontologia. Essa ferramenta consegue extrair resultados mais direcionados e claros em comparação com o próprio motor de busca da Plataforma Lattes em certos contextos. Como ponto a ser melhorado em trabalhos

[17]https://qualis.capes.gov.br/

posteriores, o autor cita o tratamento de inconsistências entre os mesmos trabalhos que aparecem com títulos diferentes entre currículos. Seja isso por falta de padronização ou por erro de digitação no momento do preenchimento do currículo.

O *Semantic Lattes* tem uma relevância para mostrar uma parte do potencial que pode ocorrer quando se adota os conceitos e técnicas de Web Semântica para extração de informação.

3.3 ScriptLattes

Proposto por Mena-Chalco e Júnior (2013) o scriptLattes é um software livre que permite a criação de relatórios acadêmicos de forma automática utilizando a base de dados dos currículos extraídos da Plataforma Lattes. A abordagem do software é direcionada a análise de um grupo de pesquisadores. Dessa forma os resultados não são focados nas produções individuais e sim do grupo como um todo. Na Figura 3.4 são apresentados os tipos de produções acadêmicas que podem ser extraídos via a utilização do scriptLattes.

Na proposta inicial, o scriptLattes consultava as informações automaticamente extraindo os dados das páginas HTML contidos na Plataforma Lattes. Essa consulta ocorria quando o usuário do software especificava na entrada do sistema os IDs dos pesquisadores que deveriam ser analisados. O scriptLattes não possui suporte a análise dos Currículos Lattes em formato XML.

Vale ressaltar que, para ocorrer a extração de dados das páginas HTML, foi necessário desenvolver *parser*, baseado em análise textual permitindo identificar e extrair regiões ou trechos específicos de texto (TOMITA, 2012). Além desse procedimento, foi necessário o tratamento para separar os dados contidos, considerando as marcações do HTML, pois em muitos casos o nome do orientador, o ano e título do trabalho são incluídos na mesma frase só sendo separados por vírgulas e pontos.

Durante a fase de análise alguns cuidados foram tomados para a interpretação dos dados corretamentes. Para o casamento de produções acadêmicas idênticas foram realizadas comparações dois a dois entre todas as produções separadas previamente por ano e por tipo. Como forma de amenizar as inconsistências seja por erro de digitação ou por falta de padronização nos títulos e nomes, o software considera que se ambos os

A. Produção bibliográfica
Artigos completos publicados em periódicos Livros publicados/organizados ou edições Capítulos de livros publicados Textos em jornais de notícias/revistas Trabalhos completos publicados em anais de congressos Resumos expandidos publicados em anais de congressos Resumos publicados em anais de congressos Artigos aceitos para publicação Apresentações de trabalho Demais tipos de produção bibliográfica
B. Produção técnica
Softwares com registro de patente Softwares sem registro de patente Produtos tecnológicos Processos ou técnicas Trabalhos técnicos Demais tipos de produção técnica Total de produção técnica
C. Produção artística
D. Supervisões e orientações em andamento ou concluídas
Supervisão de pós-doutorado Tese de doutorado Dissertação de mestrado Trabalho de conclusão de curso de graduação Iniciação científica Orientações de outra natureza
E. Projetos de pesquisa
F. Prêmios e títulos
G. Eventos (participação e organização)

Figura 3.4: Tipos de produções extraidas do Currículo Lattes (MENA-CHALCO; JÚNIOR, 2013)

textos possuem 80% de equivalência, então podem ser considerados como idênticos.

A saída do software é um conjunto de relatórios no formato HTML. Eles mostram informações quantitativas das produções oriundas dos arquivos de entrada. A proposta da análise é voltada a um grupo de pesquisadores, consequentemente são produzidos gráficos que mostram o número de produções de determinado tipo durante cada ano. Outro ponto importante da saída é o grafo de colaboração entre membros de um grupo baseados exclusivamente na sua produção bibliográfica, técnica ou artística. A Figura 3.5 ilustra um exemplo de grafo de colaboração.

Inicialmente, o scriptLattes foi desenvolvido em 2005 na linguagem Perl, porém em 2011 foi reprogramado em Python incluindo funcionalidades novas e melhorando o desempenho. Ambos os códigos fontes são distribuídos na modalidade de software livre sob a licença GNU-GPL[18].

[18]https://www.gnu.org/licenses/gpl-3.0.en.html

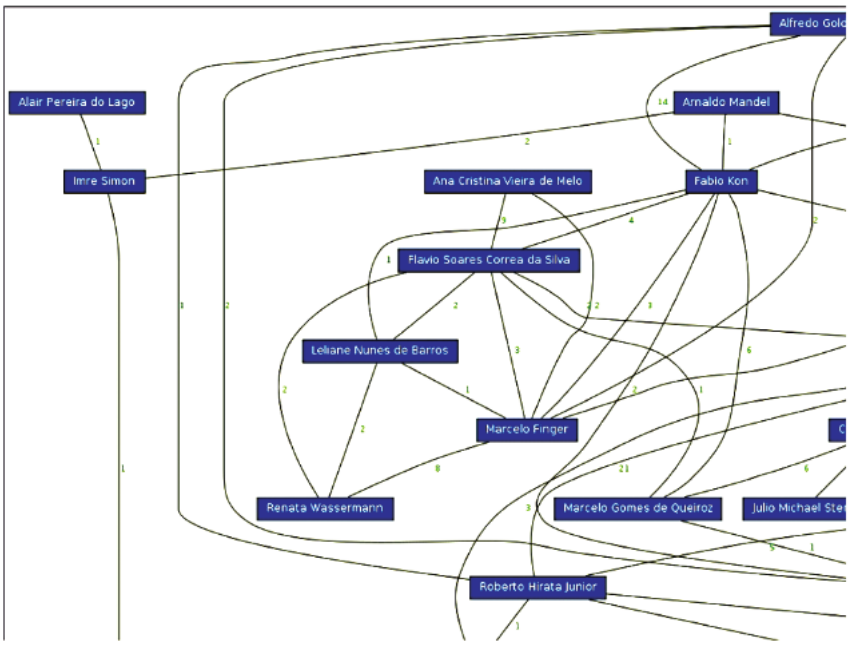

Figura 3.5: Grafo de colaboração (MENA-CHALCO; JUNIOR; MARCONDES, 2009)

Infelizmente com a adoção do captcha pela Plataforma Lattes a parte de entrada automatiza ficou inviável. Assim sendo é necessário que o usuário, de forma manual, popule o software com as páginas HTML. Além disso, a plataforma Lattes está migrando para outro *layout* de exibição sendo assim necessário uma reformulação no componente de extração de informações via documento HTML.

De acordo com Mena-Chalco, Junior e Marcondes (2009), essa ferramenta de software livre é a pioneira na prospecção de extensos conjuntos de dados acadêmicos provenientes de Currículos Lattes em formato HTML, e atualmente está sendo útil para extrair e representar conhecimento de grupos de pessoas cadastradas na plataforma Lattes, de forma simples.

3.4 SOS Lattes

Proposto por Galego (2013) a ferramenta batizada com o nome SOS Lattes tem como objetivo auxiliar na tarefa de extração e consulta de informações da Plataforma Lattes, considerando um grupo específico. O acrônimo do nome SOS Lattes significa *Semantic Ontology-based Script Lattes* que representa a união dos nomes das 3 ferramentas que serviram de base para o projeto, sendo elas: *Semantic Lattes*, OntoLattes e scriptLattes.

Observando pelo lado de sua arquitetura, a ferramenta pode ser dividida em três camadas, sendo elas denominadas pelo autor de camada de apresentação, camada de dados e camada de extração de dados da plataforma Lattes. Na Figura 3.6 é apresentada uma ilustração da sua arquitetura.

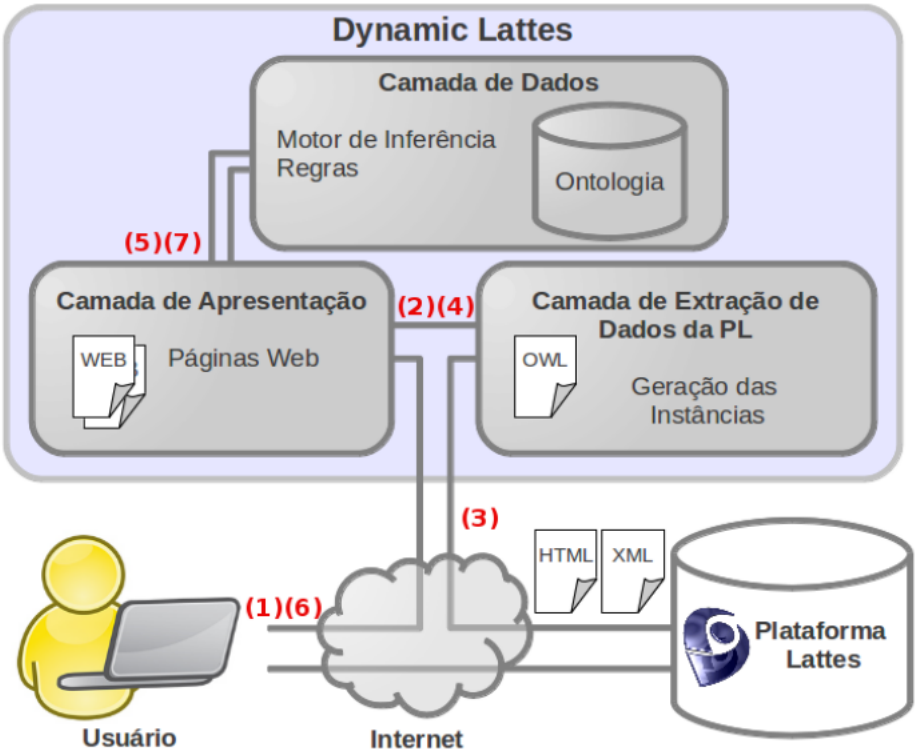

Figura 3.6: Arquitetura SOS Lattes (GALEGO, 2013)

A Camada de extração de dados da Plataforma Lattes tem por objetivo extrair os dados presentes na Plataforma Lattes. Essa extração pode ocorrer em dois formatos: HTML e XML. Além da extração essa camada é responsável por exportar os dados em formato OWL para que sejam manipulados na camada de dados. As atividades descritas foram propostas utilizando as funcionalidades já existentes no software scriptLattes e na abordagem OntoLattes para a extração e exportação de dados respectivamente. Algumas alterações nesses trabalhos se mostraram necessárias para maior adequação à arquitetura proposta no SOS Lattes.

A camada de dados tem como responsabilidade armazenar e consultar os dados. O armazenamento e cuidado com os dados ocorrer a partir da importação dos arquivos OWL cedidos pela camada de extração de dados da Plataforma Lattes. Já as buscas dos

dados ocorrem por meio de consultas por linguagem SPARQL e dos motores de inferência que ficam armazenados na camada. As funcionalidade e conhecimento adquiridos via os trabalhos *Semantic Lattes* e OntoLattes foram essenciais para sua construção dessa camada.

A camada de apresentação busca fazer a interação entre o sistema e o usuário. A interação ocorre por meio de páginas Web dinâmicas de acordo com os dados armazenados na camada de dados. As aparências das páginas são muito próximas as propostas pelo scripLattes tendo assim semelhança na interação dos usuários com a informação. Existe também um campo de texto onde o usuário pode escrever perguntas para que a ferramenta possa responder, sendo essa funcionalidade inspirada na abordagem proposta pela ferramenta *Semantic Lattes*.

O SOS Lattes apresenta grande potencial, principalmente por ser estruturado como uma aplicação Web e usar a base de conhecimento desenvolvida em trabalhos anteriores que enriquecem bastante o software.

3.5 Análise de Redes Sociais Científicas

A partir do banco de dados do Lattes, Stroele (2012), buscou em seu trabalho construir e analisar uma rede social científica multi-relacional.

Na construção desse modelo foram considerados vários fatores que exercem o papel de influenciadores na análise da rede social, como o peso do relacionamento, idade do relacionamento entre outros fatores extraídos da base de dados.

Dentre todos os objetivos deste trabalho vale destacar a criação da rede social científica multi-relacional, a busca por comunidades de pesquisas, o agrupamento dos indivíduos com relacionamento em comum além da sugestão de novos relacionamentos para melhorar a comunicação na rede social.

Ao final do trabalho, como forma de auxiliar na visualização dos resultados, foi desenvolvida uma ferramenta que permite o usuário analisar visualmente e temporariamente a rede social científica. A Figura 3.7 mostra uma forma de visualização dos dados na ferramenta.

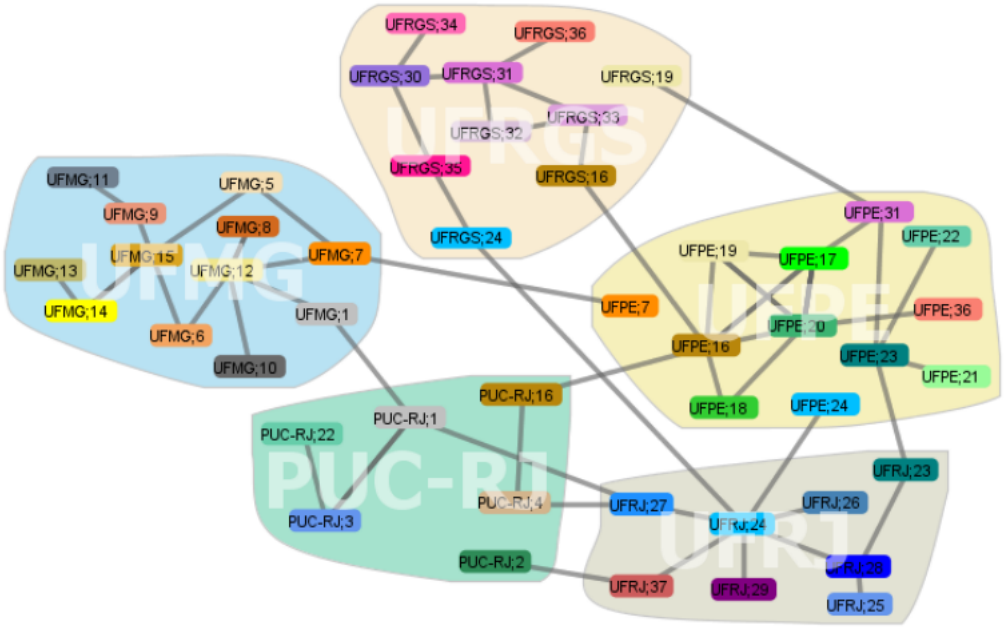

Figura 3.7: Visualização de resultados (STROELE, 2012)

3.6 Considerações finais do capítulo

Durante esse capítulo foram apresentados trabalhos relacionados, que serviram de base para a proposta desta monografia. Pode-se observar que a Plataforma Lattes serviu de inspiração para múltiplos trabalhos que buscam extrair informações de sua base dados.

Os trabalhos citados contabilizam informações em sua grande maioria dos atributos pertencentes a cada currículo, porém eles não utilizam o potencial da extração e contabilização de informações novas através da análise dos relacionamento entre entidades. Em contrapartida, o trabalho proposto por Stroele (2012) se aproveita desse potencial para extrair várias informações novas dessa relação em níveis profundos de análise.A proposta do trabalho presente busca se diferenciar dos demais trabalhos pois propõe a partir dos conceitos da Web Semântica a inferência de novas informações através dos seus relacionamentos e utiliza os dados gerados como base para extrair similaridade entre currículos.

Vale ressaltar que, infelizmente, o scriptLattes e o SOS Lattes tiveram parte das suas funcionalidades de extração de dados desativadas com a introdução do captcha pela Plataforma Lattes e também pela mudança de layout das suas páginas.

No capítulo seguinte, será apresentado a propostas do trabalho.

4 Solução Proposta

O objetivo desse trabalho é inferir novas propostas de parcerias entre pesquisadores utilizando conceitos e técnicas da Web Semântica de dados advindos da Plataforma Lattes[19]. Nesse sentido esse capítulo busca descrever as abordagens escolhidas no processo para alcançar o objetivo proposto.

4.1 Planejamento

O objetivo de inferir novas propostas de parcerias entre pesquisadores, tem como principal foco encontrar similaridades, direta ou indiretamente, entre os currículos da base dado da Plataforma Lattes.

Assim sendo, para atingir esse foco, quatro pontos chaves foram levados em conta:

1. O objetivo é inferir novas relações, então a inferência deve ser feita indiretamente, ou seja, quando um pesquisador "A" cita outro pesquisador "B" em seu próprio currículo essa relação deve ser ignorada, pois necessariamente o pesquisador "A" já conhece o "B".

2. Uma abordagem interessante para inferir indiretamente uma relação é usar intermediários, por exemplo, o pesquisador "A" cita o "B" e "C" em itens cadastrados no seu currículo, logo é possível avaliar um semelhança entre os pesquisadores "B" e "C".

3. O Currículo Lattes de um pesquisador possui diversos itens, porém nem todos apresentam dados significativos ou consistentes para inferir novas relações, por isso é necessário selecioná-los e filtrá-los.

4. As relações inferidas são simétricas, ou seja, uma relação de "A" para "B" implica também numa relação de "B" para "A".

[19] http://lattes.cnpq.br/

Levando em conta esses quatro pontos abordados e com uma análise dos itens presentes no currículo Lattes, foram criadas 6 categorias de inferências que podem contribuir para encontrar uma similaridade. A Figura 4.1 mostra uma legenda para melhor entendimento dos diagramas que se seguem.

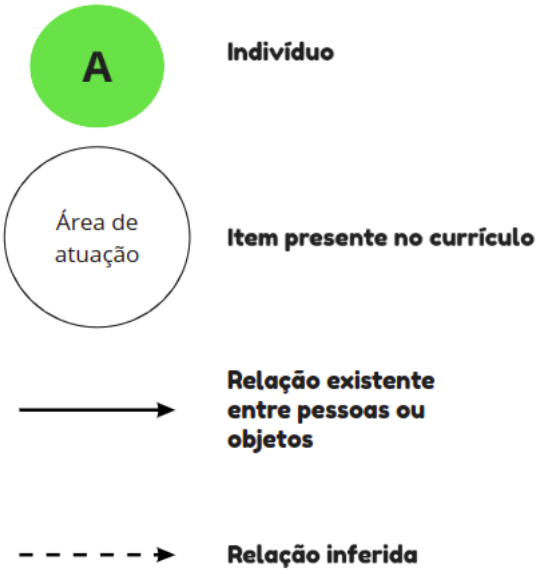

Figura 4.1: Legenda dos diagramas

Dentre as categorias de inferência citadas a seguir, não foram levados em consideração o item de artigo publicado. Essa escolha foi feita como forma de priorizar outros itens presentes no currículo e também para analisar o desempenho da ferramenta, pois a quantidade de artigos presentes poderia inviabilizar sua execução. A utilização de artigos como critério para inferência é uma das indicações para trabalhos futuros pelo seu potencial.

Área de atuação

A área de atuação é um campo geralmente preenchido em todos os currículos e que apresenta grande valor na inferência, pois especifica em diferentes níveis a área que o pesquisador pertence. No currículo lattes o campo "área de atuação" é dividido em quatro níveis de detalhamento, sendo eles "Grande Área do conhecimento", "Área do conhecimento", "Subárea do conhecimento" e "Especialidade". A cada nível seguinte a área se torna mais restrita e específica e, por esse motivo, currículos que possui o campo

"Especialidade" iguais possui mais similaridade do que currículos que só possui o campo "Grande Área" idêntico.

A inferência proposta busca gerar relações específicas para cada nível de semelhança entre área de atuação. A Figura 4.2 a seguir mostra uma representação gráfica da inferência.

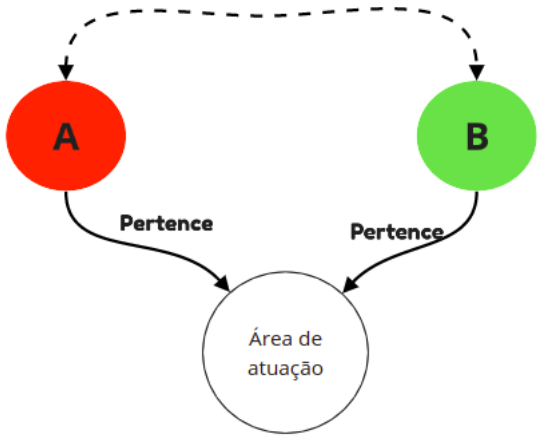

Figura 4.2: Inferência do item área de atuação

Orientador em comum

Na trajetória de um pesquisador existem várias formações profissionais, geralmente essas formações exigem um orientador. Nesse sentido, é comum que este orientador tenha de alguma forma similaridade considerando a área de conhecimento do pesquisador. Logo, dois pesquisadores que possuíam um orientador em comum também tendem a ter algum nível de similaridade. A Figura 4.3 representa essa inferência.

Participação de Banca

Na atividade de docência e pesquisa surgem muitos convites para participações como avaliador em bancas, seja nas modalidade de Lato Sensu ou Stricto Sensu. Geralmente esses convites são feitos para bancas de trabalhos que são relacionados com a área de atuação do pesquisador. Dessa forma os pesquisadores convidados possuem algum tipo de semelhança entre seus conhecimentos. Logo, nesse sentido, é possível inferir relação entre pesquisadores que participam de uma mesma banca. Na Figura 4.4 está mostrada

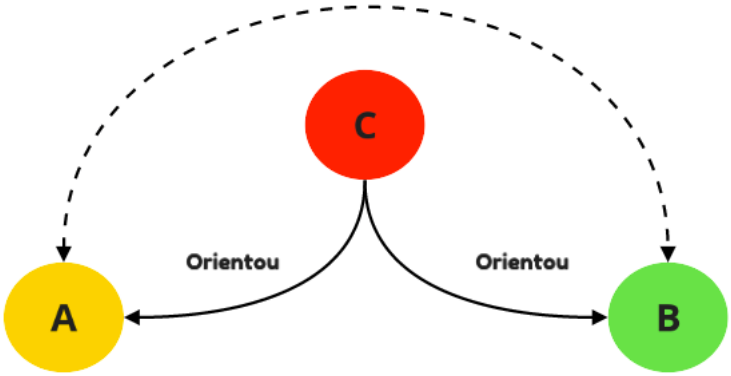

Figura 4.3: Inferência sobre a orientação

a relação de inferência sobre Banca.

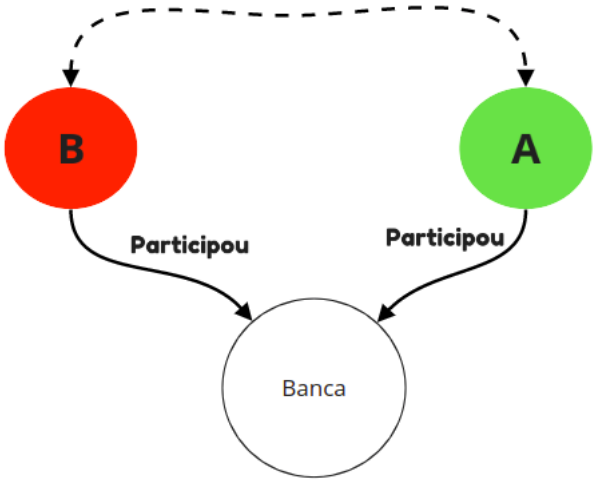

Figura 4.4: Inferência sobre a banca

Participação em projetos de pesquisa

Outra atividade que pode mostrar similaridade entre currículos é a participação de projetos de pesquisa. Porém no currículo é necessário que o pesquisador cite o trabalho e a equipe de trabalho, uma inferência diretamente entre o pesquisador e os membros da equipe estaria contrariando um dos pontos chave já relatados. Observando esse cenário foi proposta uma inferência que leva em conta um vínculo entre duas pesquisas por uma pessoa, ou seja, um pesquisador "A" tem um projeto cadastrado onde cita um usuário "B" como integrante da pesquisa, já um pesquisador "C" cita também o mesmo usuário "B" em outro projeto de pesquisa. Então existe uma ligação entre "A" e "C" que tem

como intermediário o usuário “B”. A Figura 4.5 ilustra essa inferência para facilitar o entendimento.

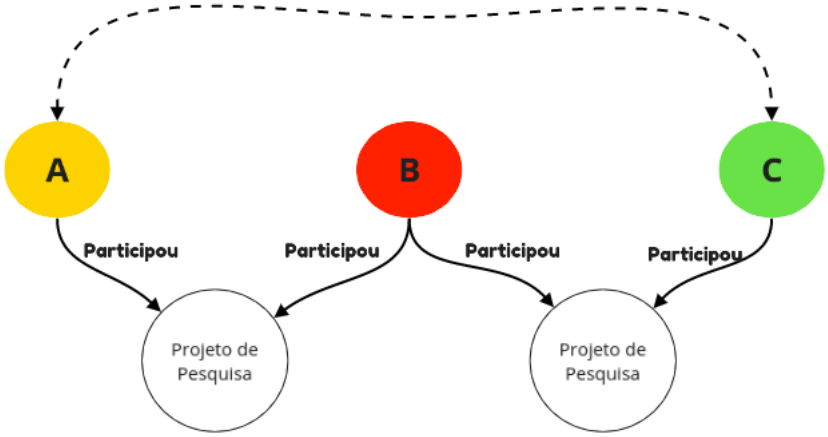

Figura 4.5: Inferência sobre projeto de pesquisa

Participação de Eventos

É muito comum existir eventos diversos que são de áreas específicas e que tendem a concentrar pesquisadores semelhantes para divulgação de conhecimento. Observando este cenário é possível propor que o item “evento” pode auxiliar bastante em comparar currículos. Para buscar esta inferência foi proposta uma abordagem onde a nova relação é criada toda vez que dois pesquisadores participarem de um mesmo evento independente se forem de anos diferentes. Na Figura 4.6 é mostrada a representação.

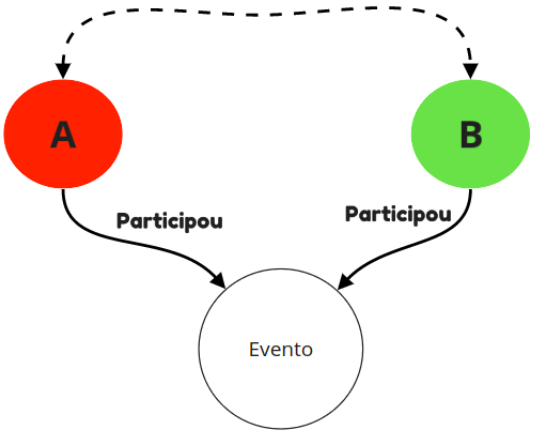

Figura 4.6: Inferência sobre eventos

Apresentação de trabalho em eventos

Ainda sobre o item "evento" no currículo, também é possível fazer inferência a partir de trabalhos apresentados em eventos, ou seja, se o usuário apresentou um trabalho em um determinado evento que outro pesquisador frequentou então é possível refletir que pode existir alguma similaridade entre esses pesquisadores. A Figura 4.7 descreve visualmente essa inferência.

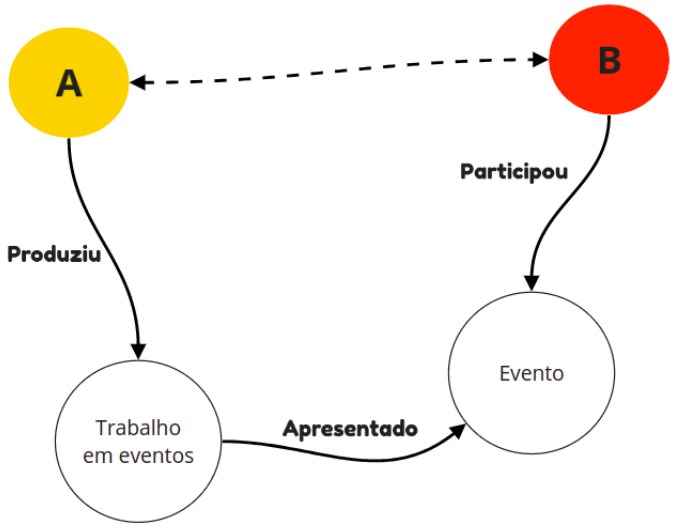

Figura 4.7: Inferência sobre trabalhos em eventos

4.1.1 Fluxo do projeto

Como forma de implementar essa proposta foi desenvolvido um sistema que ao colocar como entrada um conjunto de documentos em formato XML do currículo Lattes, seja possível gerar como saída um conjunto de pares de nomes com a pontuação atribuída a sua semelhança. Vale ressaltar que os arquivos XML aceitos pela ferramenta devem seguir o padrão proposto pela Plataforma Lattes, o DTD é disponibilizado no site[20]. Se for necessário utilizar uma base de dados externa é necessário que ocorra uma conversão anteriormente para o padrão da Plataforma Lattes.

O processo de análise geral pode ser dividido em quatro etapas como mostra a Figura 4.8. Cada etapa desse processo está descrita nas seções seguintes.

[20]http://lattes.cnpq.br/web/plataforma-lattes/extracao-de-dados

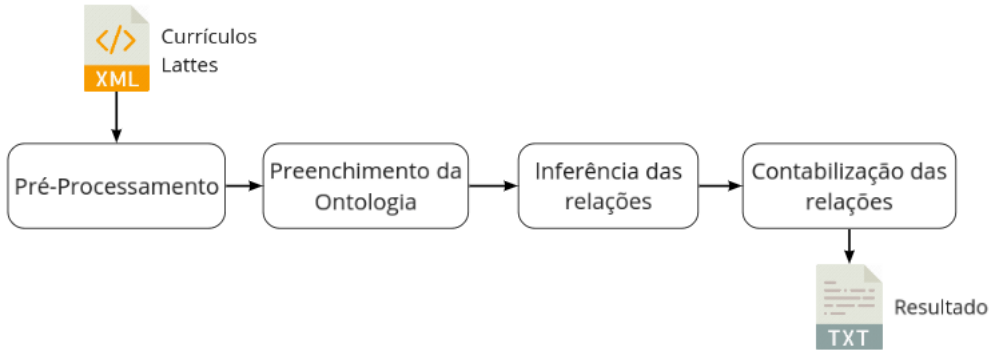

Figura 4.8: Etapas do processo de análise

4.2 Pré-Processamento

Conforme já dito a base de dados utilizada para o trabalho advém da Plataforma Lattes. Porém, os dados apresentam algumas inconsistências, duplicações e erros que podem comprometer os resultados finais. Por esse motivo é necessário que os dados passem por uma etapa de pré-processamento. Como forma de exemplificar essas inconsistências, a seguir são relatados alguns erros frequentes encontrados na base de dados.

O usuário do sistema pode cadastrar uma série de itens em seu currículo. Durante o processo de preenchimento não ocorre nenhum tipo de validação do item que está sendo cadastrado, ou seja, é possível o usuário preencher um campo de forma incorreta, como por exemplo, o usuário pode preencher o título do trabalho de forma diferente da real ou com erro de ortografia. Geralmente produções são feitas em parcerias, nesse sentido é possível que um segundo usuário cadastre o mesmo item porém com o título correto fazendo com que existam 2 produções que a princípio são iguais, porém possuem títulos diferentes.

Outro ponto é a falta de padronização dos títulos de itens, pois é possível que 3 usuários participem do mesmo evento e, no momento de cadastro, coloquem títulos equivalentes mas não sintaticamente iguais, um exemplo seria o seguinte título do evento: “4 Seminário de Iniciação Científica”, ou “ IV seminário de iniciação científica” ou “ VI SEMIC”. Todos representam o mesmo evento.

Um dos problemas é que a estrutura de dados da Plataforma Lattes não é totalmente relacional. Exemplificando, quando um usuário cadastra um artigo e adiciona

como contribuição o nome de um segundo usuário, somente o primeiro usuário tem os artigos adicionados ao seu currículo. O segundo é citado porém não é adicionado nada ao seu currículo. Dessa forma, a multiplicidade de dados iguais separados em diversos currículos é extensa.

Voltando ao processo de marcação de contribuição em artigos, também podem ocorrer alguns erros. Se o primeiro usuário digitar corretamente o nome do contribuinte da produção a Plataforma Lattes consegue preencher o campo CITAÇÃO e o campo ID-CNPQ que é um número de série único que identifica o usuário, nesse cenário não é necessário ter qualquer tipo de pré-processamento de dados, pois o ID-CNPQ facilita a identificação. A dificuldade surge quando ocorrem algumas inconsistências que o sistema não consegue identificar ou tratar. Alguns exemplos são preencher o campos NOME-COMPLETO somente com a citação do nome, preencher o nome com erro de ortografia, preencher o nome com alguma pessoa que ainda não possui currículo Lattes e mesmo que no futuro seja criado o vínculo não ocorre. Outro caso encontrado é o usuário que possui o nome de solteiro no currículo, porém naturalmente nas novas produções a pessoa é citada com o nome de casado.

Por essas e outras características é necessário que o dados passem por um processo de pré-processamento para garantir algumas correções e descartes de dados imprecisos. Dessa forma os resultados inferidos na execução do trabalho tendem possuir maior credibilidade.

Essa etapa é pré-processamento é divida em quatro partes: Normalização de Strings, Expansão dos dados, Comparação e União de dados, Redução de dados.

4.2.1 Normalização de Strings

Os dados utilizados são retirados em formato de Strings, e como não existe uma padronização de escrita é necessário um processo de normalização para diminuir algumas diferenças concretas encontradas. Dessa forma, os procedimentos seguintes são beneficiados por essa correção.

O processo de normalização ocorre logo após o dado ser retirado do arquivo XML. Como forma de guiar no processo de normalização dos dados, foi utilizada a abordagem

proposta por Euzenat, Shvaiko et al. (2007) que divide em 6 procedimentos.

Case normalisation: Consiste em converter todo o texto de letras maiúsculas para minúsculas. Por exemplo, “João” se torna “joão” e “SEMIC” se torna “semic”.

Diacritics suppression: Consiste em retirar acentuação ou sinais gráficos dos textos. Por exemplo, “maçã” se torna “maca” , “gráfico” se torna “grafico” e “it’s” se torna “its”.

Blank normalisation: Consiste em retirar excessos de espaços em brancos, tabulações e outros códigos de espaços ocultos como ”\b”,”\t”, ”\n”, ”\a”, ”\r”. Por exemplo, “João Pedro” com 3 espaços brancos para “João Pedro” com 1 espaço branco.

Link stripping: Consiste em normalizar ligação entre palavras em espaços branco. Por exemplo “Guarda-chuva” vira “Guarda chuva”.

Digit suppression: Consiste em retirar números dos textos. Por exemplo “5 Seminário” para “Seminário”.

Punctuation elimination: Consiste em retirar pontuação dos textos. “Por exemplo, “Fábio.” para “Fábio” e “Rodrigues, s. g. p.” para “Rodrigues s g p”.

Esse procedimento pode acarretar em perda de informação em alguns casos ou até em criação de falsos sinônimos, porém no caso deste projeto os danos são mínimos em comparação aos ganhos que esse procedimento pode gerar no resultado final.

4.2.2 Expansão dos dados

Como dito anteriormente, é possível que um pesquisador “A” possua um item cadastrado que cite um pesquisador “B”, porém não é possível afirmar que o usuário “B” esteja com o nome escrito corretamente, ou que o usuário “B” possua também esse item cadastrado corretamente no seu currículo. Então é necessário ocorrer um tratamento para corrigir de forma que o usuário B receba atribuição correta do item ao seu currículo.

Observando essa peculiaridade vem o processo denominado de expansão de dados onde são criadas árvores nas quais cada pessoa é um nó raiz. Toda citação de nome, seja do dono do currículo ou de alguém que ele citou se cria uma árvore independente relativa a essa pessoa com as produções que são atreladas a essa. Ou seja, se em um currículo de um usuário “A” cita 10 pesquisadores diferentes então são criadas 11 árvores independentes

onde cada folhas são as produções científicas que cada um participou.

Para ilustrar o processo de expansão foi utilizado um exemplo onde um usuário fictício “João” tem seus dados passados pelo processo de expansão de dados na Figura 4.9. Vale notar que os dados já passaram pelo pré-processamento, por isso não existe mais acentuação.

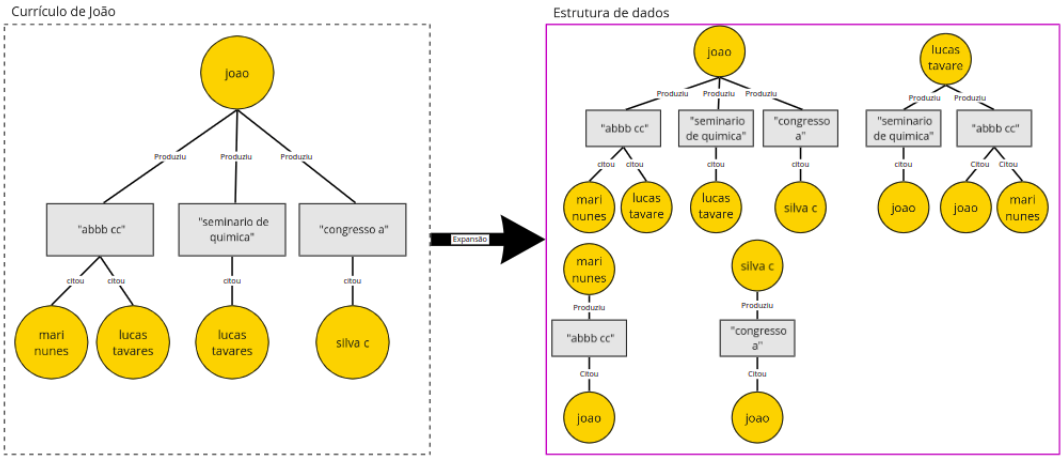

Figura 4.9: Processo de expansão dos dados

4.2.3 Comparação e União de dados

Após o processo de expansão, os dados são distribuídos em vários grafos desconectados relativos a cada expansão executada em um currículo. Nesse cenário vão existir múltiplos grafos para um mesmo pesquisador, pois ele pode ter sido citado em diferentes currículos. A proposta dessa etapa é unir esses grafos que pertencem a mesma pessoa de forma que ela também receba os nós folhas(produções científicas) sem duplicação. A melhor forma de unir é utilizar o nome do nó raiz, ou a citação ou também o id-cnpq como chave para diferenciar e comparar os usuários. O procedimento de comparação é descrito a seguir mostrando os testes feitos um a um onde a cada vez que for verdadeiro os dois grafos são unificados.

1. Comparar Id-cnpq A com id-cnpq B
2. Comparar nome A com nome B
3. Comparar citação A com nome B

4. Comparar se um nome A é subString de outro nome B

5. Comparar citação A com citação B

6. Comparar nome A com nome B utilizando algoritmo NGram.

Algoritmo NGram: A última parte do processo de comparação busca minimizar erros de ortografia entre dois tipos de String, um exemplo é o nome "Pedro Vilella" e o nome "Pedro Vilela" ou dois são o mesmo usuário, porém, por algum motivo no momento do preenchimento foi esquecido de adicionar a letra L no sobrenome.

Observando essas diferenças foi necessário utilizar o Algoritmo NGram proposto por Kondrak (2005) que busca calcular a distância entre duas Strings. Distância entre duas Strings ou também conhecida como distância Levenshtein é dada pelo número mínimo de operações necessárias para transformar uma string na outra. No algoritmo de NGram a distância varia entre 0 até 1, sendo 0 quando duas String são iguais e 1 quando a distância entre elas é "infinita". Utilizando na base de dados o algoritmo se percebeu eficiente e confiável quando se aceita a união de dados para valores abaixo de 0.25 e 0.20 em alguns casos. A Implementação do algoritmo foi extraída da biblioteca Java String Similarity[21].

O processo de comparação de nomes também ocorre com os itens de bancas e eventos cadastrados no currículo, pois possuem várias quantidades e grande variação de nomes de um mesmo item. Para amenizar essa diferença ocorrem comparações entre itens atribuídos a um mesmo indivíduo e também comparação entre itens de indivíduos diferentes.

Após terminar o processo de união é possível garantir com maior segurança que um pesquisador receba todas as atribuições de trabalhos pertencentes ao seu nome e que também não existam tantos dados duplicados.

4.2.4 Redução de dados

O último processo antes de incluir os dados na estrutura OWL é o descarte de informações desnecessárias. Nesse momento é verificado no grafo itens que só possuem vínculo a um usuário, ou seja, somente ao próprio dono. Também busca usuários que são citados so-

[21] https://github.com/tdebatty/java-string-similarity

mente uma vez, pois no fim esses usuários não apresentaram resultados relevantes além de ocupar maior espaço de processamento. Somente com a retirada desses itens desnecessários a base de dados reduz bastante o tamanho e tende a ter somente informações relevantes.

4.3 Modelagem dos dados

Um dos detalhes do objetivo do trabalho é utilizar a Web Semântica como meio para inferir novas parcerias. Por esse motivo é necessária a modelagem dos dados de forma a incluir os dados de saída da etapa de pré-processamento em uma ontologia. Como forma de mostrar o potencial da Web Semântica foi escolhida a linguagem OWL para modelar a ontologia, pois possui suporte maior para descrever relações e é a linguagem oficial recomendada pela W3C para a Web Semântica.

A modelagem dos dados é relacionada às categorias de inferências pois é através dessa modelagem que é possível extrair novas relações. Nesse sentido, a ontologia em OWL foi modelada utilizando a ferramenta visual Protégé de forma a facilitar o processo de criação.

Para representar cada indivíduo foi utilizada a criação de uma classe chamada "Pessoa" com atributos "idLattes" que corresponde ao ID único de cada usuário e o atributo "NomeCompleto" sendo o nome do pesquisador.

Para as produções cadastradas no currículo foram criadas classes específicas na ontologia que podem ser ligadas a uma classe "Pessoa" através de uma relação. As classes e os atributos criados estão representados na Figura 4.10.

Com a modelagem dos dados pronta é necessário iniciar o processo de população da ontologia, umas das formas mais fáceis e automatizadas é utilizar a biblioteca OWLAPI[22] em linguagem Java que auxilia na manipulação de ontologias. Com essa biblioteca o preenchimento da ontologia OWL com os dados gerados da etapa de pré-processamento se tornam rápida e fácil. Com a ontologia preenchida se inicia o processo de inferência que será explicado na seção seguinte.

[22]http://owlcs.github.io/owlapi/

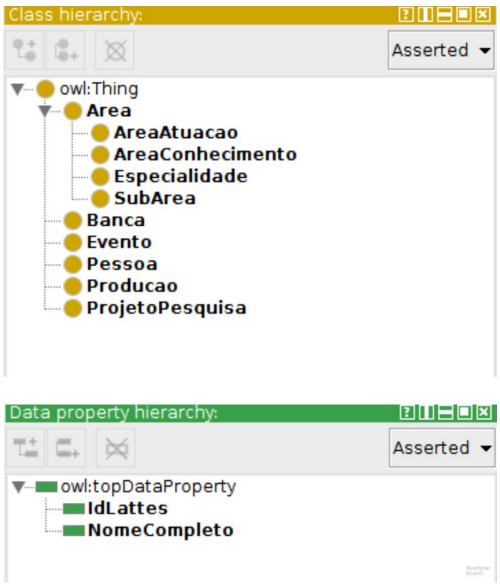

Figura 4.10: Classes e atributos da ontologia

4.4 Inferência das Relações

Inicialmente a proposta era utilizar as regras de SWRL[23] em conjunto com o motor de inferência Hermit[24] que possui suporte na biblioteca OWLAPI. Entretanto, foram encontrados dois grandes problemas que inviabilizaram a utilização dessa abordagem

O primeiro problema encontrado é que pela natureza da ontologia OWL e pelo motor de inferência, não é possível contabilizar o número de inferências criadas iguais. Exemplificando, se um pesquisador "A" for em 5 eventos diferentes entre si em que um pesquisador "B" também participou, o motor de inferência criará somente uma relação entre esses dois pesquisadores não contabilizando as outras 4 similaridades. Nesse cenário só é possível criar uma relação única entre dois indivíduos distintos na ontologia. Logo, os dados finais não iriam apresentar resultados satisfatórios pois só possuiriam 6 relações inferidas no máximo.

O segundo grande problema é o tempo de execução do motor de inferência. Mesmo em bases de dados que passaram pela etapa de pré-processamento não conseguiam apresentar resultado em tempo satisfatório. Uma possível explicação é que o motor de inferência busca ser de carácter geral, então ele tenta inferir outras relações que já vêm estabelecidas na ontologia OWL, como por exemplo inferência de classe, inferência de axi-

[23]https://www.w3.org/Submission/SWRL/
[24]http://www.hermit-reasoner.com/

oma inverso entre outras. Nessa perspectiva o motor de inferência realiza mais trabalho do que é necessário para o cenário deste trabalho. O Hermit[25] é uma dos mais famosos e completos motores de inferência, porém na perspectiva desse trabalho ele não consegue apresentar resultados apesar de largo tempo de execução.

Observando esses dois obstáculos, foi necessário utilizar uma abordagem diferente para solucionar o problema e, consequentemente, gerar resultados. Como não é possível usar um motor de inferência de caráter geral, então a proposta é criar o próprio processo de inferência utilizando os dados da ontologia.

Nessa proposta os dados necessários são retirados da ontologia e armazenados na memória de execução do programa respeitando as relações já existentes. Os dados são representados a partir de uma grafo conexo que facilita o processo para percorrer os dados. Inicialmente é escolhido um indivíduo onde ocorre buscas através das suas relações para cada tipo de inferência proposto. A Figura 4.11 ilustra o processo. É selecionado o indivíduo "A"e o critério de inferência de área de atuação. A partir dessa escolha são selecionados os itens "área de atuação"vinculados ao indivíduo (PASSO 1), para cada item selecionado são encontrado todos os usuários que também são vinculados a esse item (PASSO 2) e consequentemente são inferidas as relações entre eles (PASSO 3). O processo se repete para todos os critérios e todos os indivíduos presentes no grafo.

4.5 Contabilização das relações

As relações inferidas através do processo descrito anteriormente buscam contabilizar o número de relações entre dois usuários, pois com esse número é possível descobrir a similaridade dos currículos dos pesquisadores. No total são 6 critérios de inferência distribuídas em 9 tipos. Para trazer um resultado mais próximo da realidade é necessário atribuir pesos diferentes a categorias de inferências distintas. Os critérios de Banca e Eventos possuem grande quantidade cadastrado nos currículos, por essa perspectiva foram atribuídos o valor de 10 pontos de forma que sua contribuição seja menos predominante no resultado final. Os critérios derivados da área de atuação possuem maior valor a medida que descem no nível de hierarquia, como só é possível cadastrar no máximo 4 item por tipo,

[25]http://www.hermit-reasoner.com/

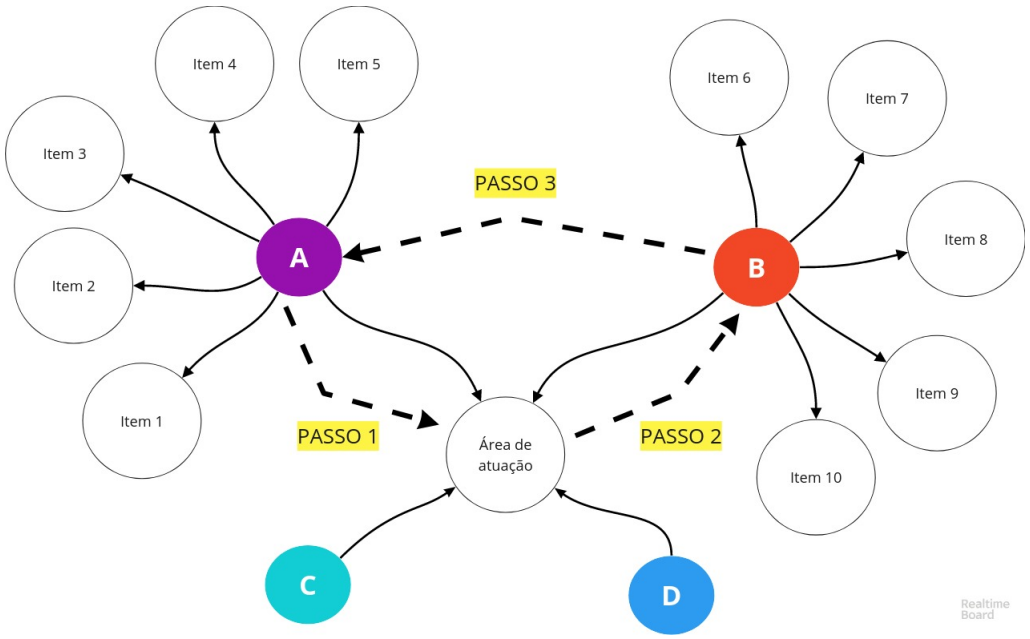

Figura 4.11: Ilustração do processo de inferência no grafo

então eles podem possuir pontuação maior que outros itens de tamanho ilimitado. Os critérios restantes receberam médio valor por ter uma presença razoável e também por apresentar poder de significância mediano. Dessa forma, na Tabela 4.1 é possível conferir as pontuações atribuídas a cada inferência. A formula 4.1 mostra como é o calculo final da similaridade, sendo um somatório de valores de critério (X) vezes a quantidade de presença dessa critério (Y).

$$\sum_{n=0}^{i} Xi * Yi \tag{4.1}$$

Tabela 4.1: Pontuação por tipo

Tipo de inferência	Pontuação por item
Área de atuação	30
Área de conhecimento	60
SubÁrea	80
Especialidade	100
Banca	10
Evento	10
Orientação	10
Projeto de evento	30
Projeto de pesquisa	20

5 Estudo Piloto

Este capítulo apresenta uma avaliação da execução e os resultados da proposta. Estamos interessados em avaliar como os conceitos de Web Semântica podem auxiliar na facilitação de colaboração entre pesquisadores. O método de avaliação escolhido foi baseado em um estudo piloto num contexto de dados reais. Um estudo piloto é normalmente aplicado na área de Engenharia de Software nas etapas iniciais do processo de avaliação. Embora não seja formal como um estudo experimental, a utilização do estudo piloto oferece uma avaliação técnica da infraestrutura, além de permitir possíveis correções e ajustes do projeto de pesquisa. Como resultado, um estudo piloto pode garantir a qualidade dos resultados auxiliando os objetivos do projeto presente.

5.1 Escolha da base de dados

Os dados escolhidos para testar o projeto foram os currículos pertencentes aos docentes vinculados ao departamento de Ciência da Computação na Universidade Federal de Juiz de Fora[26], acessados e retirados no dia 1 de Abril de 2018. Em sua totalidade são 44 currículos. Cada indivíduo apresenta currículo de tamanho diferente em decorrência dos dados cadastrados da experiências profissionais. A base de dados foi adquirida manualmente e individualmente através de *download* dos dados públicos contidos na Plataforma Lattes em documento XML.

Essa base de dados foi selecionada pois é um cenário possível de ser validado ao mesmo tempo que apresenta uma quantia de dados significativos para derivar inferências.

5.2 Resultados Obtidos

Como método para demonstrar os resultados obtidos durante todo o processo, um currículo foi selecionado da base de dados para ser acompanhado durante toda a execução do pro-

[26]http://www.ufjf.br/deptocomputacao/institucional/corpo-docente/docentes/

grama. O currículo selecionado foi da docente e pesquisadora Regina Maria Maciel Braga. O acompanhamento e os dados gerados estão descritos nos parágrafos seguintes.

Após o carregamento da base de dados na memória, o primeiro procedimento executado é a comparação de similaridade sintática entre nomes de bancas que cada currículo possui. Nesse caso, se for encontrada similaridade entre os nomes, então um dos nomes é selecionado para ser o principal e os dois itens recebem os mesmo. Neste procedimento ocorrem 2 unificações para o currículo acompanhado, e para a base de dados completa ocorrem 56 unificações.

O segundo procedimento ocorre da mesma forma que o primeiro, no entanto, o item comparado é o "evento". No currículo acompanhado tivemos 34 unificações e na base o total de 917.

O procedimento seguinte é a expansão dos membros, já descrita no capítulo anterior. Resumidamente cada nome citado no currículo se torna um indivíduo para o programa. A base total foi expandida de 44 para 6513 nomes, o currículo da Regina contribuiu para 96 nomes citados nesse procedimento.

O procedimento de união dos nomes similares é o seguinte, onde nome a nome produzido no procedimento anterior é comparado e unificado se for enquadrado como similar. Dos 6513 nomes citados anteriormente, após o final do procedimento de união sobraram 1216. Em relação ao currículo que foi acompanhado ocorreram 130 uniões de nomes, no entanto em decorrência de divergências na sintática os resultados foram comparados e unidos em 2 nomes distintos, sendo eles "regina_maria_maciel_braga" e "regina_m_m_braga_villlela". O primeiro nome é derivado do próprio currículo que está sendo acompanhando, já o segundo decorre de citações presentes em outros currículos. Confirmando a informação diretamente com a autora do currículo, foi possível confirmar que o segundo nome citado possui um sobrenome a mais pois se trata do sobrenome acrescentado após o casamento, logo se torna muito difícil conseguir informação suficiente para que no processamento o sistema identifique que os dois nomes se trata da mesma pessoa. Como dito anteriormente ocorrem 130 convergências em 2 nomes diferentes, as variações presentes citadas e sua quantidade de repetição estão presentes na Tabela 5.1. Os nomes "regina_m_m_braga_villlela" uniram entre si e todos os outros convergiram para

se unirem com o "regina_maria_maciel_braga". Vale ressaltar que o nome "braga_regina" consegue ser identificado pois é um das citações cadastradas no currículo original.

Tabela 5.1: Quantidade de nomes repetidos

Nome	Quantidade
regina m m braga villlela	2
regina maciel braga	5
regina maria maciel braga	97
regina maria maciel braga vilela	2
regina maria maciel braga villela	10
braga regina	16

Dando prosseguimento, a etapa seguinte é utilizada para remover indivíduos que apresentam valor inexpressivo. São selecionados indivíduos que possuem menos de 2 itens cadastrados para serem removidos da base de dados. Dos 1215 indivíduos restam 569 após esse procedimento. Posteriormente, existe um outro momento que ocorre uma filtragem de dados também.

Os dois próximos procedimentos se assemelham ao tratamento de banca e evento já descritos anteriormente, porém nesse caso são feitos após a expansão e união dos dados. Os nomes atribuídos a eventos são comparados interno e externamente, nesse caso temos 57 ocorrências no currículo que foi acompanhado e 886 na base de dados. Sobre a similaridade de bancas, são 2049 no total e 42 no currículo acompanhado. Após esse processo temos todos os itens tratados de forma que possuam o máximo de nomes iguais.

Finalizado esse último procedimento os dados são incluídos na ontologia em OWL. Os itens de bancas e eventos que só possuem um indivíduo vinculado são apagados na ontologia por não apresentarem importância na inferência além dos indivíduos que só possuem um item vinculado a eles. A Tabela 5.2 mostra o resultado da exclusão dos dados:

Tabela 5.2: Quantidade de itens e pessoas retirados da ontologia

Itens e pessoas	Antes	Depois	Saldo
Pessoas	570	538	-32
Eventos	1092	319	-710
Banca	1564	1420	-144

O arquivo OWL preenchido com dados é submetido a inferência das relações. O algoritmo percorre o grafo gerado a partir do arquivo OWL. Das 538 pessoas são

gerados 3024 pares de indivíduos que possuem algum tipo de relação. Os valores variam de pequenas até altas pontuações. O currículo acompanhado possui 94 relações com indivíduos diferentes que variam de valores de 10 pontos de similaridade até 2030 pontos. O gráfico da Figura 5.1 mostra um histograma com a distribuição de ponto do currículo acompanhado. Na Figura 5.2 é mostrado o gráfico da distribuição global de valores de relações em escala logarítmica.

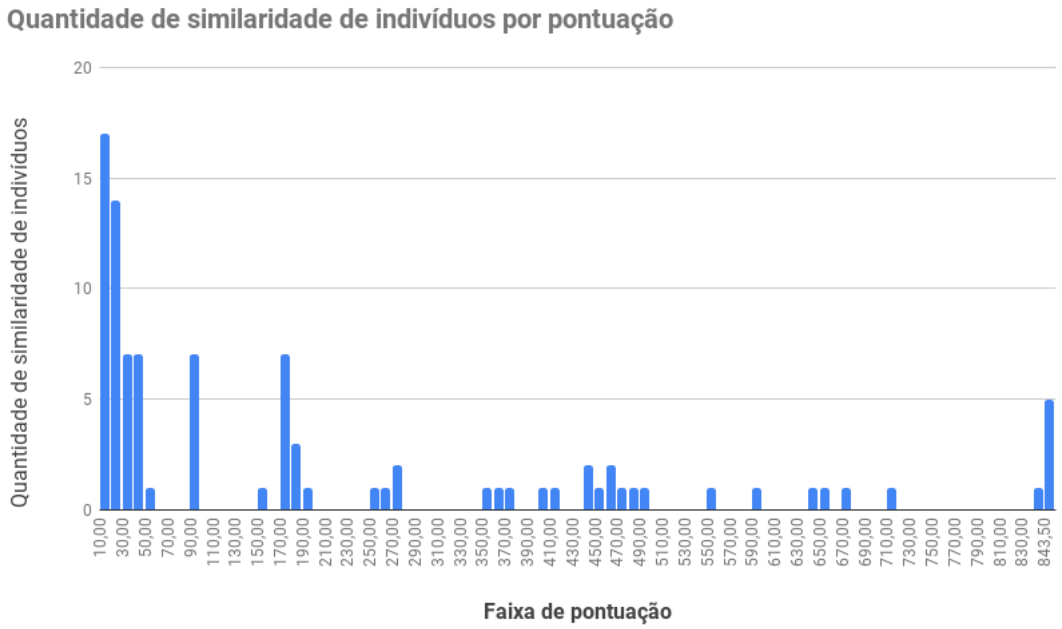

Figura 5.1: Histograma das inferências do currículo da Regina

Observando a distribuição de valores no gráfico da Figura 5.2 é possível notar uma concentração de pequenos valores em comparação a relações com valores elevados. A ultima classe a direita possui um numero significativo de indivíduos, porém essa classe engloba o intervalo de valores de 580 até 2030, enquanto as demais classes possuem intervalos de 10 pontos. A escala logarítmica foi escolhida para amenizar visualmente essa concentração e facilita visualização das pequenas quantias.

Como dito no capítulo anterior a inferência é calculada a partir de 9 tipos, onde cada tipo recebe uma pontuação para cada vez que é contabilizado. Para melhor entendimento da contribuição de cada critério para o resultado final, as duas tabelas 5.3 e 5.4 mostram o total de contribuição para o currículo acompanhado e para todos as dados respectivamente.

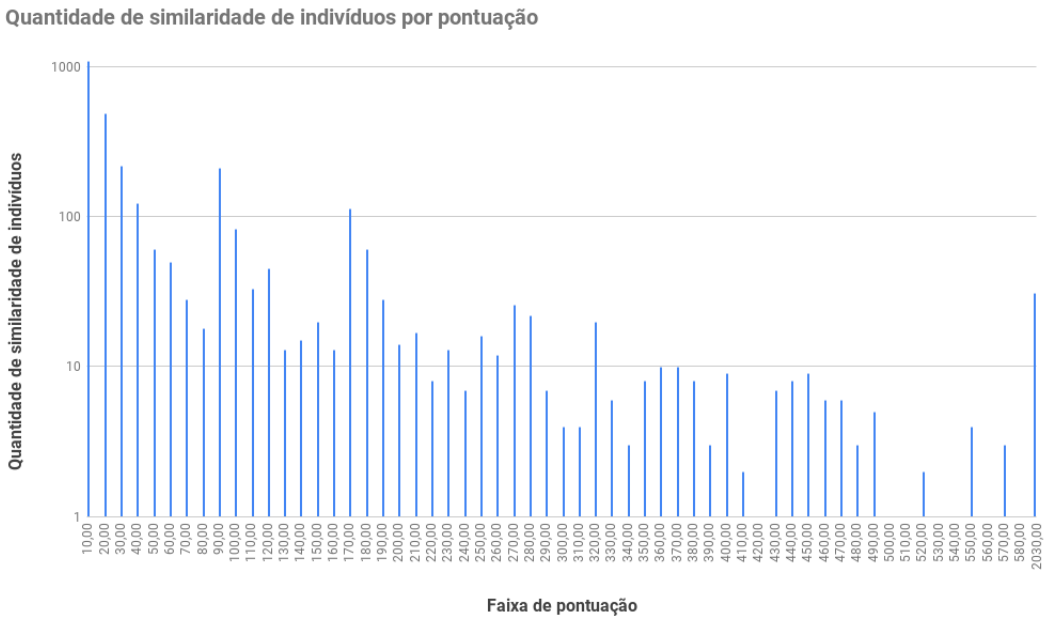

Figura 5.2: Histograma de todos os indivíduos em escala logarítmica

Tabela 5.3: Tabela de frequência e pontuação do currículo da Regina

Critério	Valor p/unidade	Quantidade	Pontuação	Frequência Absoluta
Área de atuação	30	42	1260	5,81%
Área de conhecimento	60	39	2340	10,78%
Subárea	80	39	3120	14,38%
Especialidade	100	34	3400	15,67%
Banca	10	203	2030	9,35%
Evento	10	10	100	0,46%
Orientação	10	1	10	0,05%
Projeto em eventos	30	92	2760	12,72%
Projeto de pesquisa	20	334	6680	30,78%
Total	———	794	21700	100%

Analisando a Tabela 5.4 é possível observar os critérios que mais contribuem com os resultados. Os quatro primeiros critérios correspondem a aproximadamente 55% da pontuação na base de dados. Esses critérios são vinculados a especificação de área de atuação dos indivíduos. No contexto da base de dados utilizada é esperado esse resultado pois todos os currículos são de um determinado departamento que consequentemente possuem áreas iguais ou próximas.

Tirando os critérios já citados, aquele que mais contribui é a participação de banca seguido de projeto de pesquisa. Vale notar que o valor por unidade atribuído a participação de banca é o menor, porém esse critério apresenta o maior número de

Tabela 5.4: Tabela de frequência e pontuação geral

Critério	Valor p/unidade	Quantidade	Pontuação	Frequência Absoluta
Área de atuação	30	948	28440	11,69%
Área de conhecimento	60	820	49200	20,23%
Subárea	80	465	37200	15,29%
Especialidade	100	205	20500	8,43%
Banca	10	5586	55860	22,96%
Evento	10	142	1420	0,58%
Orientação	10	206	2060	0,85%
Projeto em eventos	30	657	19710	8,10%
Projeto de pesquisa	20	1443	28860	11,86%
Total	———	10472	243250	100%

ocorrência entre todos os outros critérios.

De acordo com a Tabela 5.3, é possível notar que a frequência absoluta se mostra diferente da tendência mostrado pelos dados gerais. O item que mais contribui nesse caso é o projeto de pesquisa. No currículo Lattes estão presentes 18 projetos de pesquisas cadastrados, essa quantidade se mostra atípica se comparada com outros currículos Lattes utilizados na base de dados. Geralmente os currículos apresentam números bem menores do que 10 itens cadastrados como projeto de pesquisa.

5.3 Apresentação dos resultados

Como forma de auxiliar no entendimento dos resultado das inferências, os dados foram organizados em uma representação de grafo. Essa representação busca mostrar de forma visual a representação das relações entre pesquisadores. Como forma de dar acesso remoto foi construída uma página web que pode ser acessada no link [27]. Para a construção da página web foram utilizados o framework javascript Vue.js[28], o framework CSS Bootstrap [29] e o framework de exibição de grafo Vis.js [30]. Como forma de auxiliar na observação dos grafos, é possível movimentar o scroll do mouse para alterar o zoom no gráfico.

O grafo geral que contém todas as inferências entre pares de pessoas pode ser exibido clicando no botão “Gerar Grafo Completo”, após clicar e esperar alguns segundos

[27] https://weltonah.github.io/OntoDTLattesResult/
[28] https://vuejs.org/
[29] https://www.getbootstrap.com/
[30] http://www.visjs.org/

é gerado um grafo como na Figura 5.3. Nesse grafo cada nó representa uma pessoa e cada aresta uma relação entre indivíduos. Nesse grafo não é possível quantificar o grau de semelhança e de recomendação entre pesquisadores, somente é possível ver os indivíduos que estão relacionados entre si.

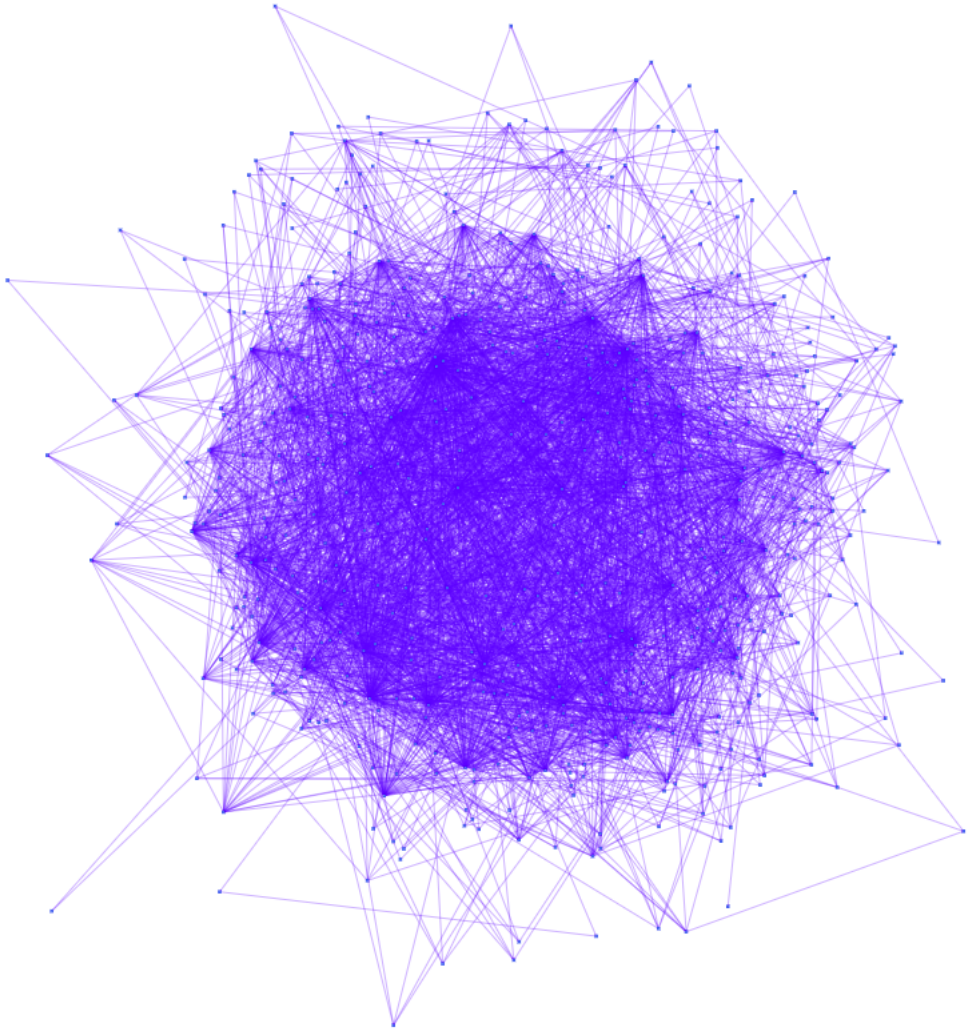

Figura 5.3: Exibição do grafo completo dos resultados

Ao encontrar e clicar no nó correspondente ao currículo que está sendo acompanhado é possível ter uma noção das relações existentes entre eles e todos os outros indivíduos. O nó relativo à pesquisadora Regina Maria Maciel Braga ele pode ser clicado para visualizar suas relações com outros pesquisadores. Na Figura 5.4 é mostrado o resultado dessa interação, as ligações em cor avermelhada são aquelas que partem do nó selecionado.

Outra forma de exibir os resultados é selecionando cada pesquisador para analisar

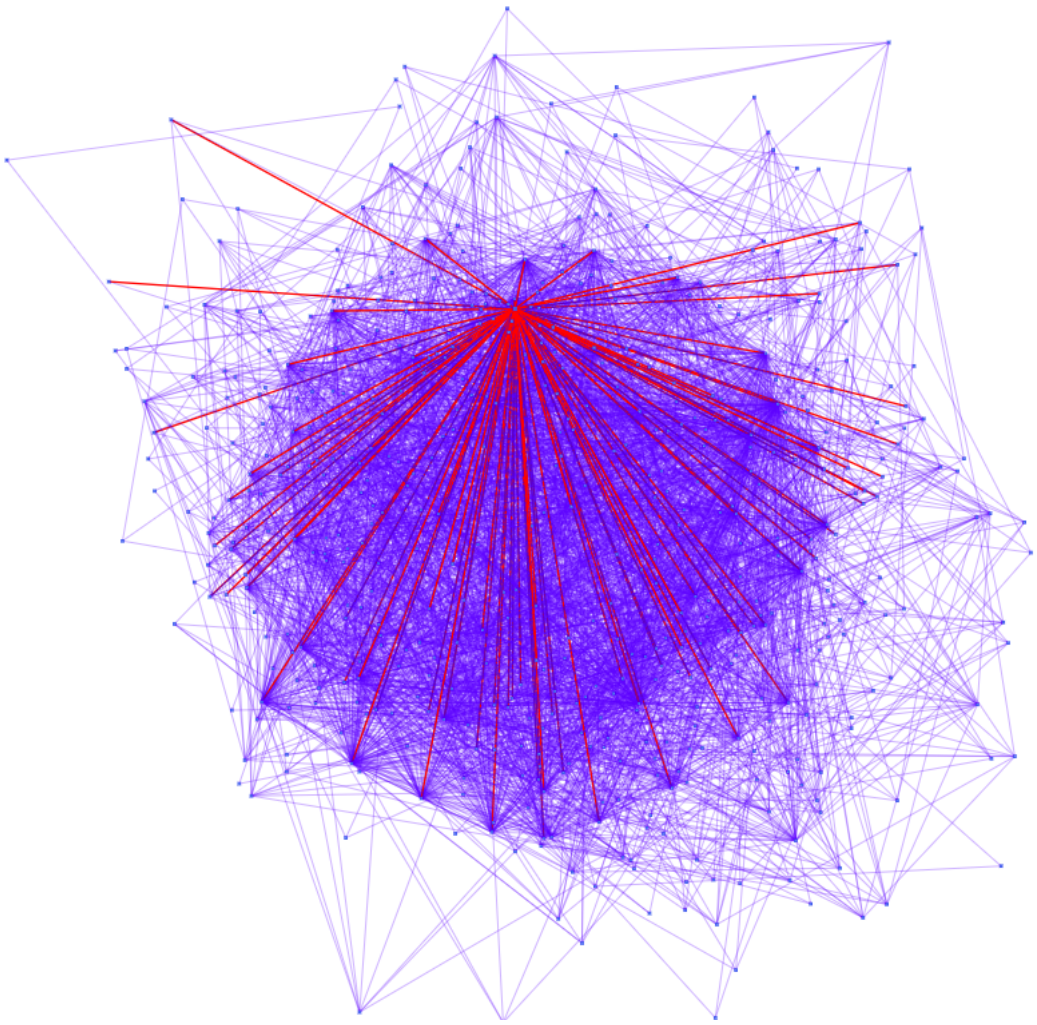

Figura 5.4: Exibição do grafo completo selecionando o nó da Regina Maria Maciel Braga

o grafo de similaridade. Nesse grafo são apresentados todos os indivíduos que apresentam algum tipo de similaridade com o pesquisador selecionado. O nó que representa o pesquisador permanece no centro e o tamanho das aresta correspondente inversamente ao grau de similaridade, ou seja, quanto mais próximo os outros nós do centro, maior é o número de inferências entre os indivíduos e, consequentemente, maior a similaridade. Na Figura 5.5 é exibido o resultado para o pesquisador Regina Maria Maciel Braga. Na Figura 5.6 é mostrada uma tabela com a pontuação relativa à similaridade de alguns indivíduos relativos a Figura 5.5.

Observando esse grafo é possível notar um grande aglomerado de indivíduos nas bordas e um número pequeno de indivíduos próximos ao centro, esses comportamentos podem ser validados pelo gráfico da Figura 5.1. Essa tendência tende a se repetir em quase todos os resultados, geralmente esses indivíduos que permanecem no centro são indivíduos

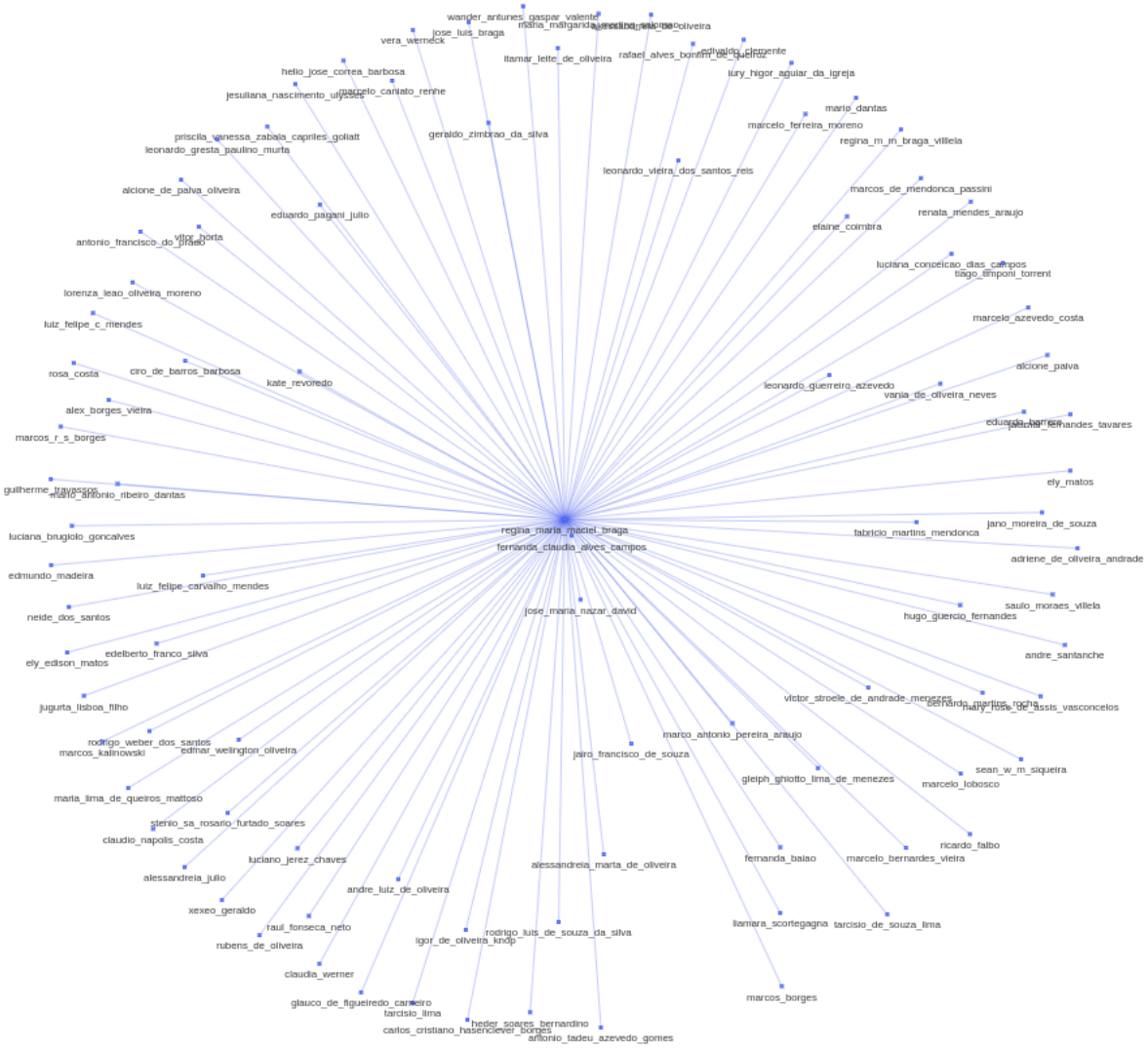

Figura 5.5: Grafo resultado para Regina Maria Maciel Braga

que tiveram seus currículos selecionados para compor a base de dados, pois possuem maior número de dados relacionados e consequentemente oferecem mais informação para gerar as inferências.

Em contrapartida, ao selecionar indivíduos que não pertençam a base de dados inicial é possível notar grafos que apresentam pouco ou quase nenhum valor expressivo. Um exemplo é o grafo da Figura 5.7.

5.4 Discussão dos resultados

Analisando os resultados finais do estudo piloto é possível observar que os valores de similaridades de pares de pesquisadores são dependentes do número de currículos incluídos

Nome	Semelhança
fernanda_claudia_alves_campos	2030
jose_maria_nazar_david	1730
jairo_francisco_de_souza	1120
marco_antonio_pereira_araujo	1000
leonardo_guerreiro_azevedo	850
kate_revoredo	840
alessandreia_marta_de_oliveira	710
victor_stroele_de_andrade_menezes	670
fabricio_martins_mendonca	650
gleiph_ghiotto_lima_de_menezes	640
luiz_felipe_carvalho_mendes	590
leonardo_vieira_dos_santos_reis	550
fernanda_baiao	490
edmar_welington_oliveira	480
andre_luiz_de_oliveira	470
eduardo_pagani_julio	460

Figura 5.6: Tabela com valores de similaridade Regina Maria Maciel Braga

na base de dados. Pois como a inferência ocorre de forma indireta, quanto maior o número de indivíduos maior a quantidade de dados disponíveis para cálculo das inferências.

Uma segunda observação é que indivíduos que somente foram citados por terceiros, e que não possuem currículos na base de dados, tendem a apresentam resultados inexpressivos. Em alguns casos, esses mesmos currículos apresentam valor numérico expressivo em relação a um pesquisador que possui seu currículo incluído na base de dados, porém ao observar o grafo gerado do primeiro indivíduos ele normalmente possui menos de 6 relações com pesquisadores diferentes.

Uma terceira observação é que indivíduos que possuem maior similaridade são indivíduos que já possuem histórico de parceria. Nesse caso, para o objetivo do trabalho seria desnecessário considerar esses resultados, pois não apresentam carácter de novidade para criar parceria entre pesquisadores. A melhor opção é analisar as relações que possuem valor no meio do espectro, porque é possível encontrar novas parcerias entre pesquisadores que nunca tiveram contato mais próximo.

Como finalização, é possível concluir que os dados encontrados no estudo piloto apresentam caráter satisfatório, pois vão de encontro com os objetivos inicialmente propostos do trabalho. Os dados quantificam corretamente a similaridade entre pesquisadores que é um reflexo da realidade.

Sendo assim, é possível utilizar o projeto em estudos experimentais para validar a sua aplicabilidade no dia a dia. A sua aplicação pode contribuir diretamente para o

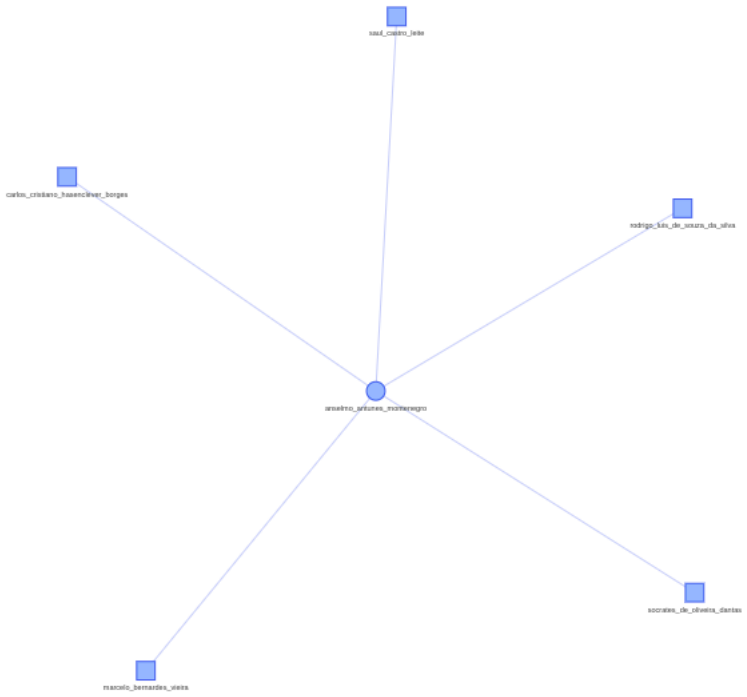

Figura 5.7: Grafo resultado para um indivíduo citado

desenvolvimento da ciência no Brasil a partir do momento que pode ser utilizada para auxiliar na criação de novas parcerias e fortalecer os laços existentes.

6 Conclusão e Trabalhos Futuros

Neste trabalho foi proposta a extração de novas parcerias entre pesquisadores através da quantificação da similaridade entre os perfis. Como forma de alcançar esse objetivo, foram utilizados conceitos de Web Semântica como forma de mostrar as possibilidades presentes da tecnologia.

A conclusão do trabalhos se mostra satisfatória e consegue alcançar os objetivos propostos inicialmente. Os dados resultantes do processo conseguem quantificar a similaridade entres pesquisadores, também é possível gerar novas informações de similaridade entre pesquisadores distantes através das inferências.

Os dados gerados servem como a analise de um cenário reais e pode ser aplicados em universidade e centros de pesquisas de forma a auxiliar a criação de novas parcerias fortalecendo a ciência brasileira.

Durante o processo de desenvolvimento surgiram alguns obstáculos, o principal foi a inviabilização de motores de inferência para o objetivo proposto. Nesse cenário, foi necessário o desenvolvimento de um algoritmo que buscava simular o trabalho e que em sua analise conseguiu suprir a necessidade. Vale destacar também a etapa de pré-processamento que conseguiu minimizar os erros advindos da base de dados tornando os resultados finais mais consistentes.

Assim como em outros trabalhos citados, é possível observar a importância da base de dados presente no Plataforma Lattes. Com possíveis modificações da plataforma, é possível criar um espaço aberto para que mais trabalhos possam explorar o potencial dos dados já existentes. Pela perspectiva desse trabalho, uma possibilidade seria a disponibilização livre dos dados em formatos RDFS ou OWL, pois a partir dessa modificação é possível que motores de inferências e aplicações possam explorar de forma mais eficiente os dados, além de possibilitar a navegação entre currículos distintos.

O trabalho presente possibilitou a demonstração de parte dos benefícios que a Web Semântica pode oferecer quando aplicada a bases de dados. Como escolha o estudo piloto auxiliou na avaliação dos resultados se mostrando consistentes.

Surge como desafios futuros o testes da ferramenta em contextos reais, podendo ser utilizados processos formais como em estudos experimentais. Outro ponto é o monitoramento dos resultados a medida que os dados de entradas vão sendo escalonados.

Para trabalhos futuros, é sugerido a proposta de criar novos critérios de inferência explorando mais a riqueza de informação contida no Lattes, uma possível abordagem é a utilização de artigos como item para derivar novas inferências. Nessa abordagem é possível também gerar resultados mais aprofundados a medida que mais itens são utilizados.

Outros fatores que possam ser utilizados são restrições que levam em considerações o tempo, podendo restringir as inferências ou fazendo uma analise de similaridade em diferentes momentos da vida profissional do pesquisador assim sendo uma analise cronológica.

A conversão direta dos arquivos XML para OWL ou RDF de forma que fiquem disponíveis online para acesso também é uma possibilidade de trabalhos futuros, dessa forma o motor de inferências criado poderia atuar em um cenário real onde os conceitos de Linked Data estivessem aplicados.

Como ultima proposta para trabalhos futuros, é a exploração dos resultados gerados utilizando redes complexas(ALBERT; BARABÁSI, 2002). O resultado obtido oferece a possibilidade de criação de um grafo, que utilizando técnicas adequadas pode auxiliar no estudo de comportamentos, agrupamentos e outros fatores que no primeiro momento não são visíveis.

Bibliografia

ALBERT, R.; BARABÁSI, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics*, APS, v. 74, n. 1, p. 47, 2002.

BERNERS-LEE, T. Linked data-design issues (2006). *URL https://www.w3.org/DesignIssues/LinkedData.html*, v. 10, n. 11, 2017.

BERNERS-LEE, T. et al. The semantic web. *Scientific american*, New York, NY, USA:, v. 284, n. 5, p. 28–37, 2001.

BERNERS-LEE, T. J. *Information management: A proposal*. [S.l.], 1989.

BONIFACIO, A. S. Ontologias e consulta semântica: Uma aplicaçao ao caso lattes. 2002.

BREITMAN, K. *Web Semântica: a internet do futuro*. LTC, 2005. ISBN 9788521614661. Disponível em: ⟨https://books.google.com.br/books?id=yJ92BAAACAAJ⟩.

COSTA, F. S. Y. A. P. D. Semantic lattes: Uma ferramenta de consulta de informações acadêmicas da base lattes baseada em ontologias. 2009.

DENTLER, K. et al. Comparison of reasoners for large ontologies in the owl 2 el profile. *Semantic Web*, IOS Press, v. 2, n. 2, p. 71–87, 2011.

EIS, D. *Introdução à Web Semântica: A inteligência da informação*. [S.l.]: Casa do Código, 2017. ISBN 9788594188076.

EUZENAT, J.; SHVAIKO, P. et al. *Ontology matching*. [S.l.]: Springer, 2007. v. 18.

GALEGO, E. F. *Extração e consulta de informações do Currículo Lattes baseada em ontologias*. Tese (Doutorado) — Universidade de São Paulo, 2013.

GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge acquisition*, Elsevier, v. 5, n. 2, p. 199–220, 1993.

HORRIDGE, M. et al. A practical guide to building owl ontologies using protégé 4 and co-ode tools edition1. 3. *The university of Manchester*, v. 107, 20011.

KONDRAK, G. N-gram similarity and distance. In: SPRINGER. *International symposium on string processing and information retrieval*. [S.l.], 2005. p. 115–126.

LOPES, I. L. Novos paradigmas para avaliação da qualidade da informação em saúde recuperada na web. SciELO Brasil, 2004.

MENA-CHALCO, J. P.; JÚNIOR, C. Prospecção de dados acadêmicos de currículos lattes através de scriptlattes. *Bibliometria e Cientometria: reflexões teóricas e interfaces. São Carlos: Pedro & João*, 2013.

MENA-CHALCO, J. P.; JUNIOR, C.; MARCONDES, R. Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, SciELO Brasil, v. 15, n. 4, p. 31–39, 2009.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Ontologias: conceitos, usos, tipos, metodologias, ferramentas e linguagens. *Relatório Técnico–RT-INF-001/07, dez*, 2007.

NAKASHIMA, M. Y. Currículo lattes e web semântica. 2004.

O'REILLY, T. *What is web 2.0*. 2005.

STROELE, V. Análise de redes sócias científicas. Universidade Federal do Rio de Janeiro, 2012.

TOMITA, M. *Current issues in parsing technology*. [S.l.]: Springer Science & Business Media, 2012. v. 126.