

Uma Abordagem Baseada em Ontologias para a Predição de Ligações em Redes de Colaboração Científica

Thiago Henrique Dias Araujo

TEXTO APRESENTADO
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
O EXAME DE QUALIFICAÇÃO
DE
MESTRE EM CIÊNCIAS

Programa: Mestrado em Ciência da Computação

Orientadora: Prof^a. Dr^a. Renata Wassermann

São Paulo, março de 2016

Uma Abordagem Baseada em Ontologias para a Predição de Ligações em Redes de Colaboração Científica

Esta é a versão original do texto elaborado pelo candidato Thiago Henrique Dias Araujo para o exame de qualificação apresentado ao Instituto de Matemática e Estatística da Universidade de São Paulo como requisito para obtenção de título de Mestre em Ciências.

Resumo

ARAUJO, T. H. D. **Uma Abordagem Baseada em Ontologias para a Predição de Ligações em Redes de Colaboração Científica**. 2016. 30 f. Exame de qualificação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

A comunidade científica pode ser enxergada como uma rede em que cada pesquisador se relaciona com outros através de colaborações diversas como, por exemplo, na coautoria de um artigo científico. Alguns trabalhos aplicam técnicas de aprendizado de máquina para prever ligações entre os participantes de uma rede, tratando do problema conhecido como Predição de Ligações. Entretanto, algumas dessas metodologias apresentam certas limitações, ou por utilizarem formas pouco expressivas de representação do domínio, ou por analisarem apenas a estrutura da rede, sem levar em consideração as características intrínsecas dos participantes dessa rede. A proposta do presente trabalho é modelar uma ontologia capaz de indicar características próprias dos pesquisadores da Plataforma Lattes e suas relações com outros pesquisadores, extraíndo conhecimento prévio sobre o domínio e, posteriormente, utilizá-lo no enriquecimento de um modelo de aprendizado que faça predição de novas colaborações. Esperamos que esta abordagem aqui apresentada seja mais eficaz, e que contribua para uma melhor colaboração entre pesquisadores.

Palavras-chave: redes de colaboração científica, ontologia, aprendizado de máquina, predição de ligações, plataforma lattés.

Abstract

ARAUJO, T. H. D. **Uma Abordagem Baseada em Ontologias para a Predição de Ligações em Redes de Colaboração Científica**. 2016. 30 f. Exame de qualificação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

The scientific community can be seen as a network of people whose relationships are built by means of various types of collaborations, like the co-authoring of a scientific paper. Some applications of machine learning to the problem of link prediction have been done before, but these methods have some limitations, because they employ representations of the domain that are not very expressive. And some of these works only analyse the structure of these networks, without taking into consideration the characteristics and attributes of the people that integrate these networks. Our proposal is to create an ontology able to show the characteristics of researchers and their relationships with others, using information from Plataforma Lattes, and extract background-knowledge about this domain that will be used to enrich a machine learning model capable of predicting new collaborations and co-authorships. We hope that our methodology will be more efficient and that it will help researchers form better relationships and collaborations.

Keywords: scientific collaboration networks, ontology, machine learning, link prediction, plataforma lattes.

Sumário

Lista de Abreviaturas	ix
Lista de Figuras	xi
Lista de Tabelas	xiii
1 Introdução	1
1.1 Motivação	2
1.2 Objetivos	2
1.3 Metodologia	3
1.4 Contribuições	3
1.5 Organização do Trabalho	4
2 Conceitos	5
2.1 Aprendizado Supervisionado	5
2.2 Conhecimento prévio (<i>background knowledge</i>)	5
2.3 Hierarquia de Representações em Problemas de Aprendizado	6
2.4 Ontologia	6
3 Trabalhos Correlatos	9
3.1 scriptLattes	9
3.2 Predição de Ligações	9
3.2.1 Importância de um pesquisador, financiamento, e quantidade de colaboradores em um projeto	9
3.2.2 Predição utilizando Aprendizado Supervisionado	10
3.2.3 Grafos Relacionais com Atributos	10
3.2.4 Predição de Citações	10
3.3 Ontologias	11
3.3.1 Plataforma Lattes	11
3.4 Justificativa	11
4 Desenvolvimento	13
5 Conclusões	19
6 Cronograma	21

Referências Bibliográficas

23

Lista de Abreviaturas

ILP	(<i>Inductive Logic Programming</i>)
OWL	(<i>Web Ontology Language</i>)
SPARQL	(<i>SPARQL Protocol and RDF Query Language</i>)
SQL	(<i>Structured Query Language</i>)
SRL	(<i>Statistical Relational Learning</i>)
SVM	(<i>Support Vector Machine</i>)
SWRL	(<i>Semantic Web Rule Language</i>)

Lista de Figuras

Lista de Tabelas

4.1	Matriz de relacionamentos entre pesquisadores	16
6.1	Cronograma de atividades	21

Capítulo 1

Introdução

A internet não é apenas um repositório quase infinito de informações aleatórias em páginas da web. Em uma análise mais apurada, ela passa a ilustrar imensas cadeias de significados. Com o surgimento da área de Mineração de Dados esse imenso conjunto de dados passou a ser analisado de forma mais eficiente e utilizado como base para a construção de novos conhecimentos.

Por essa razão, quando falamos da quantidade maciça de informação disponível, conhecida como *Big Data*, devemos pensar na grande quantidade de **significados** que podemos extrair desses dados. Para [Kay \(2014\)](#), o que chamamos de *Big Data* não é algo relevante pela sua quantidade, mas sim pelo que ele chama de *Big Meaning*, isto é, toda essa imensa riqueza de significados presente na web que podemos estudar.

Nesse panorama, surgem novas possibilidades de extração de informação e de construção do conhecimento de forma automatizada. Através de modelos de classificação e de análise, tornou-se possível, segundo [Halevy et al. \(2009\)](#), construir sistemas que entendem uma frase em determinada língua e a traduzem para outra, ou que interpretam e classificam textos, tudo isso usando como exemplo apenas os inúmeros textos contidos na web.

Ainda como desdobramento dessas novas técnicas, tornou-se possível analisar as interações entre seres humanos em redes sociais, como nas redes de colaboração científica, que contêm informações de autores, publicações e seus temas, em diversas áreas do saber. É possível identificar comunidades interessadas em temas similares, grupos de pesquisadores e mudanças nos interesses dessas comunidades. Através disto, é possível saber mais a respeito do dia-a-dia da produção científica e do desenvolvimento da Ciência em geral ao longo do tempo.

Com especial interesse, podemos analisar essas redes de colaboração, que formam um subconjunto da comunidade científica e avaliar o desenvolvimento da Ciência em todo o mundo, descobrindo quais são as principais contribuições e tendências, como se dá o relacionamento entre pesquisadores e qual a importância e influência de grupos de pesquisa em suas comunidades. Esse tipo de análise é interessante em levantamentos sobre produtividade e impacto de uma pesquisa, relevância e popularidade de um projeto ou uma área, como formas de se mensurar e quantificar o progresso científico, entre outros aspectos.

Várias técnicas de análise podem ser empregadas nesse problema, desde as mais simples, como relatórios de produtividade, até as mais complexas, como o agrupamento de pesquisadores em categorias específicas ou a classificação de trabalhos de forma automática. Na literatura, encontramos algumas ferramentas muito úteis para essa análise. Uma delas trata do problema da predição de ligações entre pesquisadores, que permite prever, com boa segurança, novas colaborações, através da verificação das características de trabalhos publicados, interesses de pesquisa, e outras particularidades.

Como é usual no método científico, encontramos algumas limitações nessas técnicas, geralmente derivadas de métodos da área de Aprendizado de Máquina. Algumas também não levam em consideração características dos pesquisadores, mas apenas da estrutura da rede. Existem representações mais expressivas que poderiam ser aplicadas ao domínio com facilidade, e várias técnicas relacionadas à área de Sistemas Baseados em Conhecimento que possibilitariam simplificar essa tarefa e

aumentar sua eficácia.

Com o presente trabalho, esperamos contribuir para esse tipo de análise, aplicando ideias vindas da área de Sistemas Baseados em Conhecimento, e comparando o resultado com outros trabalhos, afim de analisar a possível eficácia das técnicas propostas.

1.1 Motivação

Uma rede de colaboração científica, como outras redes sociais, é uma estrutura relacional que podemos facilmente representar por um grafo com atributos, como fez [Cervantes \(2014\)](#), ou por uma ontologia, como a desenvolvida por [Anauê e Yamate \(2009\)](#), ou até mesmo por um banco de dados relacional. Também possui uma natureza multi-relacional, pois uma publicação pode ter vários tipos de relacionamento diferentes, como citações, autoria, veículo de publicação, dentre outros. E um autor pode se relacionar com outro via coautoria de um artigo, participação na mesma banca ou conferência, fazendo parte do mesmo grupo de pesquisa, ou tendo interesses similares, dentre inúmeras outras relações possíveis. Todos esses atributos podem contribuir na análise da estrutura desse tipo de rede.

Entretanto, se desejamos aplicar um modelo de aprendizado para analisar alguma característica da rede, geralmente utilizamos modelos que recebem como entrada tabelas de atributos e valores. Partindo dessa abordagem, faz-se necessário transformar uma representação relacional mais rica (como um grafo) em uma representação mais simples (como uma tabela), que nos leva a dizer que uma ferramenta de aprendizado supervisionado desse tipo trabalha apenas com representações pouco expressivas. Podemos chamar isso de um "problema de expressividade."

Segundo [Raedt \(2008\)](#), existem várias representações possíveis na modelagem de um problema de aprendizado, algumas mais expressivas, outras menos. Ele mostra em seu trabalho que existe uma hierarquia dessas representações em uma ordem crescente de expressividade. Discutimos essa ideia no capítulo sobre trabalhos correlatos.

Além da análise da estrutura da rede, podemos investigar outras características, como o perfil de um pesquisador, seu currículo, sua experiência e quem são seus pares, ou a área de pesquisa em que estão inseridas as suas publicações mais importantes. Porém, não é nada trivial transformar esse conhecimento em um conjunto de tabelas de atributos e valores.

A principal motivação deste trabalho é a possibilidade de se utilizar conhecimento prévio do domínio, também chamado de *background-knowledge*. A proposta é construir uma ontologia que representa o conhecimento acerca de uma rede de colaboração, sendo capaz de inferir novas relações, regras e derivar novas características das entidades pertencentes ao domínio. Assim, é possível extrair esse conhecimento (*background-knowledge*) e transformá-lo em novos atributos das entidades presentes na rede, enriquecendo os dados a serem explorados pelo modelo de predição.

A principal vantagem do uso de uma ontologia é que o conhecimento extraído é declarativo, compacto e de alto-nível, o que simplifica a sua validação e apreciação. Sem contar que esse conhecimento a respeito de uma entidade do domínio pode se espalhar para outras entidades, gerando outros novos atributos. Esperamos, com isso, melhorar a eficácia do modelo.

1.2 Objetivos

O objetivo do presente trabalho é demonstrar que um modelo de predição de ligações em uma rede de colaboração científica, quando enriquecido com conhecimento prévio (*background knowledge*) extraído de uma ontologia, torna-se mais eficaz em sua tarefa de predição. Para isso, espera-se cumprir os seguintes passos:

- Modelar uma Ontologia para representar o conhecimento a respeito de uma rede de colaboração científica, contendo informações sobre pesquisadores, publicações, colaborações e instituições.

- Construir consultas em SPARQL na ontologia, as quais serão utilizadas como conhecimento prévio. Algumas dessas consultas são:
 - Qual a área de pesquisa de um pesquisador?
 - Quem são os pesquisadores que ele orientou?
 - Há quantos anos um dado pesquisador trabalha em uma instituição?
- Construir um modelo de predição de ligações baseado no trabalho de [Cervantes \(2014\)](#).
- Enriquecer o modelo de predição com o conhecimento prévio extraído da ontologia, executá-lo, e comparar sua eficácia com o modelo básico.
- Propor novas consultas à ontologia que possam ser utilizadas como conhecimento prévio aplicado ao modelo e fazer novos testes.

1.3 Metodologia

Os objetivos traçados serão alcançados através da execução dos seguintes itens:

1. **Estudo dos resultados de [Cervantes \(2014\)](#):**
O objetivo desse item é entender o modelo de representação das redes de colaboração como grafos com atributos e o modelo de predição utilizando aprendizado supervisionado para que possa ser implementado e enriquecido.
2. **Estudo sobre modelagem de Ontologias e Consultas via SPARQL:**
Pretendemos encontrar uma forma de representação do conhecimento a respeito da rede de colaboração que seja útil e possa ser extraído e reutilizado.
3. **Propor uma forma de extração de conhecimento prévio da ontologia:**
Almejamos propor um método de extração das informações obtidas via consultas SPARQL na ontologia, construção de uma base de conhecimento prévio, e utilização desse conhecimento no enriquecimento do modelo, via adição de novos atributos às entidades da rede de colaboração científica.
4. **Executar a ideia proposta:**
Desenvolver um algoritmo que recebe como entrada um conjunto de dados sobre pesquisadores, suas publicações e seus coautores, que seja capaz de popular a ontologia e extrair dela o conhecimento prévio, enriquecer o modelo de predição, e executá-lo.
5. **Realizar teste da ideia com informações da Plataforma Lattes:**
Preparar um ambiente de testes, que contenha um conjunto de dados extraídos da Plataforma Lattes através do *scriptLattes* e a ontologia já estruturada. Depois disso, devemos popular a ontologia e extrair daí conhecimento prévio, rodar o modelo de predição com e sem o enriquecimento e comparar os resultados.

1.4 Contribuições

A principal contribuição que esperamos alcançar com este trabalho é a construção de um modelo que utilize aprendizado supervisionado e seja capaz de receber informação (*background-knowledge*) extraída de uma ontologia e de tratar o problema de predição de ligações em uma rede de colaboração científica de forma eficiente.

O interesse desta contribuição é a proposta de uma forma de representação mais compacta e expressiva do conhecimento prévio, algo que pode permitir uma fácil validação, expansão, generalização e reutilização desse conhecimento, além de uma consequente melhoria da capacidade de predição desse modelo.

1.5 Organização do Trabalho

No Capítulo 2, apresentamos alguns conceitos teóricos que servem para dar um melhor entendimento do trabalho. Já no 3, exploramos alguns trabalhos relacionados, algumas questões que ainda não foram exploradas por outros pesquisadores e justificamos a importância deste projeto de pesquisa. Finalmente, no Capítulo 5, delimitamos o escopo do projeto e discutimos algumas conclusões. O 6 trata do cronograma de atividades necessárias para a execução do projeto.

Capítulo 2

Conceitos

Alguns conceitos teóricos são importantes para o entendimento do presente projeto. Aqui, discutimos os fundamentos mais importantes.

2.1 Aprendizado Supervisionado

O aprendizado supervisionado é um tipo de aprendizado de máquina em que o agente, durante a fase de treinamento, observa alguns exemplos de entradas (*inputs*) e de saídas (*outputs*) esperadas e aprende a classificá-las, através de uma função de transformação de entradas em saídas. Depois disso, esse agente observa uma base de testes com inúmeros novos exemplos e deve classificá-los de acordo com o seu modelo interno. Dizemos que um modelo é eficaz se maximiza o conjunto de observações classificadas corretamente e minimiza as observações classificadas incorretamente.

2.2 Conhecimento prévio (*background knowledge*)

Conhecimento prévio, segundo [Russell e Norvig \(2009, capítulo 19\)](#), geralmente é representado como um conjunto geral de teorias em uma lógica de primeira ordem. Essas teorias são compostas por hipóteses que devem explicar ou classificar corretamente um conjunto de observações (ou atributos). Essas observações são sentenças lógicas que descrevem algo sobre o mundo. As hipóteses devem poder ser generalizadas e devem ser aplicáveis a novos exemplos. Outra propriedade importante é que a teoria seja consistente, ou seja, uma hipótese não pode gerar falsos positivos ou falsos negativos.

Esse conhecimento prévio pode ser cumulativo: novas observações podem gerar novas hipóteses, que enriquecem o conhecimento prévio e o modelo como um todo, tornando mais eficaz a sua capacidade de predição e aumentando a generalidade das hipóteses.

Uma das áreas que se concentra nesse tipo de problema de aprendizado é chamada de *Inductive Logic Programming* (ILP), que é uma intersecção entre Aprendizado de Máquina e Programação Lógica. Os programas lógicos são compostos por fatos ou predicados que descrevem exemplos (observações do mundo), conhecimento prévio, e hipóteses. Partindo disto, o programa deriva de forma indutiva um conjunto de novas hipóteses que tenham como consequência lógica todos os exemplos positivos e nenhum dos negativos. Esse conjunto de hipóteses é também chamado de teoria. As novas hipóteses geradas devem ser consistentes com as teorias já presentes no programa. O interessante dessa técnica é o uso de conhecimento prévio definido de uma forma declarativa e compacta, o reaproveitamento de hipóteses e a garantia da consistência.

Com o uso de conhecimento prévio, é reduzida a complexidade do aprendizado, pois as novas hipóteses geradas devem ser consistentes com as hipóteses anteriores, e isso reduz o conjunto de possibilidades que o algoritmo precisaria considerar, além de conseguir explicar parte das novas observações, através do reaproveitamento das hipóteses encontradas ou declaradas anteriormente (idem). Por isso, podemos dizer que problemas expressos de forma relacional podem ser tratados muito bem com algoritmos de ILP e com o uso de conhecimento prévio.

2.3 Hierarquia de Representações em Problemas de Aprendizado

Segundo Raedt (2008), existem vários tipos de representações que podem ser utilizadas para modelar um problema de aprendizado. Ele mostra em seu trabalho que também existe uma hierarquia dessas representações em uma ordem crescente de expressividade, exibida a seguir:

- **Representações Booleanas** (*Boolean Learning*): cada percepção contém itens ou proposições verdadeiras (ou *presentes*) e falsas (ou *não-presentes*), e esperamos encontrar uma regra que defina essas observações. Esta é a representação com o menor grau de expressividade.
- **Aprendizado por tabelas de valores e atributos** (*Attribute-Value Learning*): o conjunto de percepções ou experiências é apresentado como uma tabela única onde cada linha é um exemplo e cada coluna é um atributo, e desejamos que o modelo aprenda a classificar uma experiência nova ou a prever o valor de um atributo de acordo com os exemplos vistos anteriormente durante o treinamento. Esse tipo de representação é a mais comumente aplicada na literatura.
- **Representações Multi-Instância** (*Multi-Instance Representations*): muito parecida com a representação por valores e atributos, só que, neste caso, classes podem conter múltiplos exemplos, ou seja, um exemplo pode ter um atributo que depende do valor contido em outro exemplo.
- **Aprendizado Relacional** (*Relational Learning*): múltiplos exemplos ou relações entre exemplos e hipóteses podem aparecer e é possível construir essa representação com múltiplas tabelas em um banco de dados. É bastante útil quando o domínio possui uma natureza relacional.
- **Programas Lógicos** (*Logical Programs*): é um modelo que recebe como entrada as diversas observações (exemplos) e aprende a sintetizar um programa lógico com regras gerais (hipóteses) capazes de derivar estes exemplos e classificar outros conjuntos de observações. Existem alguns trabalhos na literatura cuja linguagem de representação utilizada foi o PROLOG e que usavam ferramentas de ILP. Tal representação é a mais expressiva.

Apesar dessa hierarquia de expressividade, Raedt nos mostra que as representações são equivalentes em sua capacidade de representação: o que pode ser modelado por uma pode ser adaptado e transformado em outra representação de mais baixo nível e vice-versa. Entretanto, um modelo de mais alto nível é declarativo e compacto. Quando transformado em um modelo de nível mais baixo, necessita de um número muito maior de exemplos, atributos, tabelas, colunas e outras entidades, o que dificulta a análise e a apreciação dessas informações. Portanto, a escolha correta da representação é fundamental para o pesquisador que deseja explorar um problema, pois isto simplifica a análise do domínio.

Raedt também conclui que todas essas representações podem ser modeladas naturalmente com lógica de predicados, que é relacional por natureza. Também dá exemplos de vários algoritmos de aprendizado utilizando programação com lógica, como *SVMs*, *k-nearest neighbors* e *redes bayesianas*, que podem ser utilizados na modelagem do problema.

2.4 Ontologia

Diferentemente da noção filosófica de Ontologia como o estudo da natureza do Ser, para a Ciência da Computação uma ontologia é uma estrutura relacional, um modelo de especificação formal explícita de conceitos de um determinado domínio. Esses conceitos podem dizer respeito a entidades, ou a relações entre entidades desse mesmo domínio.

Segundo Guarino *et al.* (2009), uma ontologia é um artefato computacional que possui um modelo formal de representação da estrutura de um domínio, contendo as entidades relevantes, bem como as relações que emergem da observação desse mesmo domínio, porém utilizando apenas as

que são úteis para algum determinado fim. A modelagem de classes e relações entre entidades é feita através de definições formais em uma lógica de descrição, de forma compacta e de alto nível de abstração.

Uma ontologia tem por finalidade responder a algumas perguntas, também chamadas de Questões de Competência. No fundo, são informações sobre algum domínio que desejamos que ela nos responda.

Um exemplo de domínio possível é a comunidade científica, com seus diversos pesquisadores e suas relações com os demais. Uma ontologia que representa esse domínio deve ser capaz de capturar as entidades relevantes, organizando as informações em conceitos e relações, para que seja capaz de responder a algumas perguntas, como por exemplo: *Quem é coautor de um artigo A?*. Nesse caso, ela pode descrever classes de entidades como *Pesquisador* e *Artigo* e relações binárias entre dois indivíduos, como a relação *colaborou_com* e *publicou_artigo*.

O mais importante é que essa definição formal possa ser entendida por um computador em um formato padronizado, para que a resposta seja encontrada de forma automática. Para isso, uma ontologia pode utilizar a inferência lógica para responder a esse tipo de pergunta, como também descobrir novas relações entre entidades, através do uso de *reasoners*, que são programas capazes de inferir consequências lógicas a partir de uma base de fatos.

Por tudo isso, podemos dizer que uma ontologia é uma ferramenta bastante poderosa, capaz de descrever um domínio de forma compacta, permitir consultas complexas e derivar novos conhecimentos de forma automática.

Capítulo 3

Trabalhos Correlatos

Neste capítulo, apresentamos algumas contribuições importantes que tratam do problema de análise de redes de colaboração científica e predição de ligações, e outros trabalhos relacionados que serviram de inspiração para este projeto de pesquisa.

3.1 scriptLattes

A Plataforma Lattes é uma plataforma vinculada ao CNPq, e a mais importante base integrada de currículos, grupos de pesquisa e instituições de ensino do Brasil, registrando informações valiosas sobre as atividades de pesquisa e o perfil de pesquisadores de diversas áreas do Saber em todo o país.

O projeto *scriptLattes*, proposto por [Mena-Chalco e Junior \(2009\)](#), é um sistema capaz de fazer mineração dos dados de currículos presentes na Plataforma Lattes e de gerar vários relatórios acadêmicos, além de disponibilizar informações sobre as publicações dos pesquisadores brasileiros, fazendo desambiguação dos autores e artigos e exportando dados sobre coautoria e outros tipos de colaboração.

O nosso trabalho vai utilizar uma base de mais de 4 milhões de currículos extraídos da Plataforma Lattes, gentilmente cedida pelo professor Jesús P. Mena-Chalco, como base de testes para o modelo aqui proposto.

3.2 Predição de Ligações

Vários trabalhos presentes na literatura exploram o problema da predição de ligações (*link prediction*) em redes sociais. Esse problema possui diversas aplicações, como na análise e reconstrução de redes e em sistemas que utilizam informações pessoais para sugerir novos contatos ou novos amigos. Outros pretendem detectar membros de redes terroristas com o intuito de prevenir ataques.

Uma outra aplicação interessante, discutida no item [3.2.4](#), conseguia encontrar artigos ou documentos relacionados e sugerir citações.

Discutimos a seguir alguns trabalhos relevantes.

3.2.1 Importância de um pesquisador, financiamento, e quantidade de colaboradores em um projeto

Em [Ebadi e Schiffauerova \(2015\)](#), encontramos uma análise muito apurada sobre uma rede de colaboração científica que serviu para levantar os fatores que determinam a importância de um pesquisador nessa rede, sua centralidade e sua produtividade. Foram feitas várias constatações interessantes. Descobriram, por exemplo, que grupos ligados a organizações que produziam muitos artigos importantes acabavam tendo uma performance melhor do que outros grupos.

Outro fato interessante é que a experiência e os anos de trabalho de um indivíduo fazem com que ele seja mais conhecido em sua comunidade. Tendo acesso a dados sobre financiamento e fomento à

pesquisa, constataram que pesquisas fomentadas por indústrias e empresas acabam atraindo mais colaboradores.

Essa análise também foi baseada em medidas feitas na estrutura da rede, como a centralidade dos vértices, *eigenvectors* e coeficientes de agrupamento, número médio de coautores (ligações) de um dado pesquisador, dentre outros indicadores. Com essas medidas, foi possível encontrar os líderes de diversas comunidades, que são pessoas com grande influência local.

Descobriram, ainda, que os pesquisadores mais produtivos e com o trabalho de melhor qualidade também eram os mais colaborativos. Também mostraram a influência que o financiamento causa nas redes locais de colaboração ao longo do tempo, fazendo com que os cientistas que recebem financiamento e ocupam posições de liderança busquem coautores em outras comunidades mais distantes, ampliando o seu canal de conexões.

3.2.2 Predição utilizando Aprendizado Supervisionado

Hasan *et al.* (2006) aplicou e comparou a eficácia de diversas técnicas de aprendizado supervisionado, tais como *SVMs*, *Árvores de Decisão*, *Multilayer Perceptrons*, modelos de classificação utilizando *kNN* (*k-nearest neighbors*), *Naive Bayes*, e *RBF Networks*. Sua aplicação foi na detecção de membros de redes terroristas.

O uso de *SVMs* mostrou-se melhor do que as outras técnicas, por obter uma taxa de acerto um pouco maior e por possibilitar a escolha das características importantes a serem utilizadas pelo modelo através de um método automatizado de filtragem.

O aspecto mais importante desse projeto é que a escolha do algoritmo não parece fazer tanta diferença assim na eficácia do modelo. Podemos, com isso, concluir que uma boa escolha das características é um dos fatores mais importantes.

3.2.3 Grafos Relacionais com Atributos

Um experimento muito interessante e inspirador foi explorado por Cervantes (2014), que modelou a rede de colaboração científica através de grafos relacionais com atributos para mostrar as relações de coautoria entre pesquisadores. Foi também criado um modelo baseado em *SVMs* capaz de prever novas colaborações (ligações) entre pesquisadores a partir de dados de treinamento extraídas da Plataforma Lattes via *scriptLattes*.

O modelo proposto permitiu uma série de outras análises relevantes da estrutura dessa rede, como a identificação de pesquisadores mais importantes, ou mais colaborativos, que correspondem aos vértices com mais conexões. Também foi possível identificar comunidades dentro de diferentes áreas formadas por componentes conexos desse grafo.

Por sua flexibilidade e alto nível de expressividade, esse modelo em grafo com atributos pode ser facilmente expandido. Entretanto, o modelo de aprendizado utilizado recebe como entrada tabelas de atributos e valores. Por causa disso, é preciso transformar uma representação relacional mais expressiva (um grafo) em uma representação mais simples (uma tabela) para que o algoritmo funcione.

3.2.4 Predição de Citações

Uma aplicação de Aprendizado Estatístico Relacional (SRL) foi feita por Alexandrin Popescul e Ungar (2003) em um sistema de predição de citações, que nada mais é do que um modelo de predição de ligações entre documentos. Analisando informações a respeito de duas publicações, o sistema calculava a probabilidade de uma publicação citar a outra. Com isso, o sistema poderia sugerir trabalhos correlatos que pudessem ser citados por um autor durante a escrita de algum artigo ou outro documento.

O grande desafio desse modelo de aprendizado é a seleção de atributos (*features*) que auxiliem na tarefa de classificação. Como o problema está modelado de forma relacional, a quantidade de

atributos pode crescer indefinidamente. É preciso, portanto, encontrar uma forma inteligente de adicionar um atributo e avaliar sua relevância, através de algum tipo de verificação.

Em um trabalho posterior, [Popescul e Ungar \(2007\)](#) sugerem um modelo interessante de seleção de atributos: a geração de novos atributos se dá via busca no conjunto de atributos possíveis, e sua seleção é feita através de um teste que avalia a significância estatística da inclusão desse atributo no modelo de predição.

Aplicado ao mesmo problema de predição de citações, o conjunto de atributos possíveis surgiam de consultas SQL relacionadas às entidades pertencentes a um banco de dados relacional, e o algoritmo explorava esse conjunto e selecionava os atributos mais relevantes segundo sua significância estatística. O interessante desse trabalho é que o algoritmo, ao gerar novos atributos, adicionava essas novas informações no mesmo banco de dados como novas relações entre entidades, enriquecendo ainda mais o modelo.

Nossa proposta é parecida, pois também desejamos gerar novos atributos para essas entidades, explorando algumas consultas possíveis, extraíndo conhecimento e formalizando esse resultado como *background-knowledge*, que será usado para enriquecer o modelo de aprendizado. A principal diferença está no uso de uma ontologia em vez de um banco de dados relacional.

3.3 Ontologias

Uma rede social é uma estrutura relacional que podemos facilmente representar por um grafo ou por uma ontologia. Alguns trabalhos modelaram ontologias para tratar de problemas relacionados à Plataforma Lattes e outras redes de colaboração, que discutimos a seguir.

3.3.1 Plataforma Lattes

No caso da Plataforma Lattes, [Anauê e Yamate \(2009\)](#) construíram uma ontologia capaz de responder a diversas questões de competência. O usuário consulta a base usando linguagem natural, e a ontologia consegue encontrar a informação desejada, através de inferência. A vantagem dessa metodologia é a facilidade na modelagem, nas ferramentas de inferência, e na expressividade da representação. [Galego \(2013\)](#) também desenvolveu uma ontologia para esse domínio, mas seu interesse era trabalhar com Web Semântica, gerar relatórios e detectar inconsistências.

3.4 Justificativa

A principal hipótese a ser testada é se o enriquecimento de um modelo de aprendizado supervisionado com conhecimento prévio extraído de uma ontologia pode aumentar a eficácia desse modelo quando aplicado ao problema de predição de ligações, aqui definido como a colaboração entre pesquisadores em uma rede de colaboração. Essa proposta será comparada com os trabalhos de [Hasan et al. \(2006\)](#) e [Cervantes \(2014\)](#).

O uso de uma ontologia é justificado por sua expressividade, algo que simplifica a criação de consultas geradoras de novos atributos para as entidades da rede, e a possibilidade da descoberta de novas características, através do uso de inferência lógica. Também é simples encontrar e representar novos tipos de relações entre entidades, bem como organizá-las em diferentes classes. Outra vantagem secundária é a uniformização do vocabulário a respeito do domínio, e a fácil reutilização desses conceitos e conhecimentos em outras aplicações de domínios semelhantes.

Além disso, os resultados do modelo de predição podem ser aproveitados posteriormente para expandir a própria ontologia inicial. As predições resultantes do modelo podem ser adicionadas à ontologia como novas relações. E isto pode gerar novos atributos úteis e novos tipos de classificação que podem ser extraídos novamente, derivando um conhecimento prévio ainda mais rico.

Escolhemos aplicar esse modelo à Plataforma Lattes por ser um domínio bem utilizado e por termos disponível uma base de currículos bastante extensa, da ordem de 4 milhões de arquivos.

Quanto ao modelo de aprendizado supervisionado, serão utilizados os mesmos que estão presentes na literatura para que se faça uma justa comparação.

Capítulo 4

Desenvolvimento

A proposta inicial do projeto consiste em rodar um experimento em pequena escala. Extraímos informações de um número pequeno de currículos Lattes, com interesse nas informações básicas dos pesquisadores do Departamento de Computação do Instituto de Matemática e Estatística da USP. Dados como nome, área de atuação, publicações, coautores, local de trabalho ou residência, e data em que começou a trabalhar no instituto.

Com base nessas informações, modelamos uma ontologia utilizando OWL através do software Protégé. Esse software facilita a construção, edição, e análise de ontologias.

Essa ontologia tem por finalidade organizar o conhecimento a respeito de pesquisadores e sua origem acadêmica, suas publicações, e os relacionamentos entre essas entidades. Com isso, é possível acessar informações estruturadas acerca da produção bibliográfica e científica dos docentes e alunos do instituto, servindo de base para esse experimento.

A ontologia possui a seguinte estrutura (TBOX):

Class: Grupo \equiv foaf:Group

SubClassOf: Agent

Descrição: *Classe que é pai de diferentes outras classes que definem grupos de pes.*

Relações:

membros \equiv member

Class: Grupo_de_Pesquisa

SubClassOf: Grupo

Descrição: *Classe dos grupos de pesquisa científica.*

Class: Organização \equiv foaf:Organization

SubClassOf: Agent

Descrição: *Define uma organização, que pode ser uma Universidade, por exemplo.*

Relações:

faz_parte \equiv sub_organizacao: Uma **Organização** pode fazer parte de outra **Organização**, por exemplo: o Instituto de Física faz parte da USP.

possui \equiv super_organizacao: Uma **Organização** pode possuir outra **Organização**, por exemplo: A USP possui o Instituto de Matemática e Estatística.

Class: Universidade

SubClassOf: Organização

Descrição: *Classe das universidades e instituições de ensino.*

Class: Instituto
 SubClassOf: Organização
 Descrição: *Classe dos institutos de ensino e pesquisa e faculdades.*

Class: Departamento
 SubClassOf: Organização
 Descrição: *Classe dos departamentos ligados a institutos de pesquisa ou faculdades.*

Class: Revista
 SubClassOf: Organização
 Descrição: *Classe das revistas científicas que publicam artigos.*
 Relações:
 publicou: *Uma **Revista** pode publicar um **Artigo**.*
*É o inverso da relação **publicado_em**.*

Class: Pessoa \equiv foaf:Person
 SubClassOf: Agent
 Descrição: *Classe-pai dos tipos de pessoa.*
 Relações:
 cursou: *Uma **Pessoa** pode cursar algum **Curso**.*
 estudou_em: *Uma **Pessoa** pode ter estudado em alguma **Universidade, Instituto, Faculdade**.*
 membro_de: *uma **Pessoa** pode ser membro de um **Grupo**.*
*É o inverso da relação **membro**.*
 autor: *Uma **Pessoa** pode ser autora de um **Documento**.*
*É o inverso da relação **tem_autores**.*
 trabalhou_em: *uma **Pessoa** pode trabalhar ou ter trabalhado em uma **Organização**.*

Class: Aluno
 SubClassOf: Pessoa
 Descrição: *Classe dos indivíduos que são alunos de algum curso.*
 Relações:
 tem_orientador: *Um **Aluno** pode ser orientado por um **Professor** ou outro pesquisador.*

Class: Professor
 SubClassOf: Pessoa
 Descrição: *Classe dos professores.*
 Relações:
 orienta: *Um **Professor** pode orientar um **Aluno**. É o inverso da relação **tem_orientador**.*
 tem_orientador: *Um **Professor** pode ser orientado por um outro **Professor**, mesmo que no passado.*

Class: Curso
 SubClassOf: Thing
 Descrição: *Classe dos cursos de graduação e pós-graduação.*
 Relações:
 cursado_por: *Um **Curso** pode ser cursado por **Pessoas**.*
 tipo_de_curso: *Um **Curso** pode ser de Pós-Graduação ou*

Graduação.

oferecido_por: Pode ser oferecido por uma instituição.

Class: Graduação

SubClassOf: Tipo de Curso

Descrição: *Classe dos cursos de graduação.*

Class: Pós Graduação

SubClassOf: Tipo de Curso

Descrição: *Classe dos cursos de pós-graduação.*

Class: Documento \equiv foaf:Document

SubClassOf: Thing

Descrição: *Classe dos documentos produzidos por alunos, professores e pesquisadores.*

Relações:

tem_autores: Um documento possui um ou mais autores do tipo

Pessoa. *Essa propriedade é importante pois indica uma relação de **coautoria**.*

Class: Publicacao

SubClassOf: Documento

Descrição: *Classe dos artigos científicos e outras publicações de revistas, simpósios e outros eventos.*

Relações:

*publicado_em: Um **Publicacao** pode ser publicado em uma **Revista** ou um **Evento**.*

Class: Tese

SubClassOf: Documento

Descrição: *Classe-pai das Teses científicas de doutorado e dissertações de mestrado.*

Class: Eventos

SubClassOf: Thing

Descrição: *Classe-pai dos eventos científicos, conferências, e simpósios.*

Relações:

*publicou: Um **Evento** pode ter publicações de **Artigos**.*

Class: foaf:Spatial Thing

Descrição: *Classe dos indivíduos que possuem alguma localização geográfica. **Evento** é uma delas.*

Classe: Local

SubClassOf: Thing

Descrição: *Classe dos locais geográficos; são localizações.*

Relações:

*localizado \equiv foaf:based near Relaciona **Organização** ou **Evento** com um local geográfico.*

Class: País

SubClassOf: Local

Class: Continente

SubClassOf: Local

A ontologia foi populada com as informações dos currículos lattes, e algumas consultas também farão parte da definição da ontologia. Essas consultas são escritas na linguagem SPARQL. Duas consultas que nos interessam agora respondem às seguintes perguntas:

- Qual a área de pesquisa de um pesquisador?
- Quem são os pesquisadores que ele orientou?
- Há quantos anos um dado pesquisador trabalha em uma instituição?

A primeira questão de competência é extraída diretamente do currículo lattes. A segunda questão pode ser respondida através da relação *orienta* das instâncias da classe *Professor*. Um professor terá uma lista de orientandos, que também são pesquisadores.

Já a questão a respeito dos anos de trabalho em uma dada instituição

O conhecimento prévio, que chamaremos de BK , será extraído da ontologia através das consultas SPARQL, e estruturado em uma matriz de relações, da seguinte forma: seja P o conjunto de pesquisadores presentes na ontologia, e R o conjunto de relações possíveis entre dois pesquisadores. Cada linha dessa tabela resultante corresponde a um elemento do conjunto $P \times P$, e cada coluna indica as diversas possíveis relações entre um par de pesquisadores, assumindo o valor 1 se a relação existir, e 0 se não existir.

Como exemplo, seja $P = \{p_1, p_2, p_3\}$, o conjunto de relações $R = \{\textit{orienta}, \textit{coautor}\}$, e sabemos que p_1 orienta p_2 , pois p_1 foi orientador de p_2 em seu doutorado. E também sabemos que p_1 coautor p_2 e p_1 coautor p_3 e p_2 coautor p_3 , pois os três publicaram um artigo juntos. A matriz resultante será:

	orienta	coautor
p_1p_2	1	1
p_1p_3	0	1
p_2p_3	0	1

Tabela 4.1: Matriz de relacionamentos entre pesquisadores

Durante o treinamento do modelo de predição, o conjunto de treinamento conterá a coluna coautor, que define se esse exemplo pertence ou não à categoria "coautor". Durante os testes, o modelo deverá prever se existe ou não uma relação de coautoria a partir das outras propriedades contidas nas demais colunas, classificando cada um dos exemplos em uma classe 1 (existe relação de coautoria), ou -1 (não existe relação de coautoria). A classificação será validada através da coluna coautor, e será possível medir a eficácia do método de predição.

O conjunto de informações geradas através do grafo com atributos será enriquecido com as informações extraídas do conjunto BK . Chamamos aqui de enriquecimento a adição de novos atributos aos elementos do conjunto P . Cada pesquisador possui um conjunto de atributos relevantes, e esse conjunto será expandido com as informações do conjunto BK .

Vamos testar inicialmente dois modelos de predição: Máquinas de Vetores de Suporte e Árvores de Decisão. SMVs são importantes porque... Árvores de decisão são...

O modelo receberá como estrada um conjunto de treinamento onde cada linha corresponde a um elemento do conjunto $P \times P$, sendo cada coluna uma propriedade de um dos pesquisadores, ou uma relação entre eles extraída de BK .

Um dos modelos será treinado com dados apenas dos pesquisadores (conjunto inicial sem o enriquecimento), e será testado, tendo sua eficácia será analisada.

Outro modelo será treinado apenas com o conjunto enriquecido com as informações de BRK , será testado, e sua eficácia será analisada.

Por fim, faremos a comparação da eficácia entre ambos os modelos, verificando se há melhoria na eficácia quando utilizamos conhecimento prévio.

Também é de interesse avaliar se os atributos adicionais extraídos de *BK* são relevantes. O modelo SVM é

O ambiente de testes recebe um conjunto de currículos e um arquivo OWL correspondente à estrutura da ontologia, porém sem nenhuma instância.

(1) O programa será responsável por ler os currículos e popular a ontologia com esses dados.

(2) Um dos módulos será responsável por construir o grafo com atributos e extrair o conjunto de dados sobre pesquisadores e o vetor de características.

(3) Após isso, deve rodar consultas SPARQL e extrair o conhecimento prévio *BK*.

(4) Com esses arquivos gerados, o programa vai ler os dados extraídos por (2) e enriquecer o conjunto *BK* com o vetor de características.

(5) O programa deve instanciar um modelo de predição e treiná-lo com as informações apenas dos vetores de características (6) O programa deve instanciar outro modelo de predição e treiná-lo com as informações enriquecidas em *BK*

(7) O programa deve rodar os testes de ambos os modelos, calculando sua eficácia, e deve gerar um relatório comparando ambas as metodologias.

(8) O programa deve ainda avaliar se as características contidas nos atributos de *BK* são relevantes.

Capítulo 5

Conclusões

Uma ontologia é uma forma mais expressiva de representação de conhecimento, e com ela, podemos gerar várias informações úteis sobre o domínio estudado. O uso dessa ferramenta no estudo das redes de colaboração trará uma melhor utilização do conhecimento e uma melhor exploração do domínio, auxiliando na descoberta de novos significados.

Esperamos, com o modelo proposto no presente trabalho, obter um aumento da eficácia de modelos de predição de ligações, através do uso desse conhecimento extraído da ontologia. Com isso, desejamos avançar o estado-da-arte no tratamento do problema de predição de ligações, e construir algo que possa ser utilizado por outros que se interessem por esse tipo de problema.

Capítulo 6

Cronograma

O cronograma de atividades está descrito na tabela a seguir.

	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov
1	X								
2	X	X							
3		X	X						
4			X	X					
5					X	X	X		
6					X	X	X	X	
7									X

Tabela 6.1: *Cronograma de atividades*

Item 1. Estudar o domínio e modelar uma Ontologia.

Item 2. Modelar as consultas à ontologia que servirão para enriquecer o modelo de predição.

Item 3. Construir um modelo de predição de ligações baseado nos trabalhos da literatura, fazendo experimentos com diferentes técnicas de aprendizado supervisionado.

Item 4. Propor e implementar um algoritmo de extração de conhecimento da ontologia e enriquecimento do modelo de predição.

Item 5. Montar ambiente de testes e elaborar alguns testes de comparação.

Item 6. Analisar os resultados e escrever a dissertação.

Item 7. Defender a dissertação.

Referências Bibliográficas

- Alexandrin Popescul e Ungar(2003)** Rin Popescul Alexandrin Popescul e Lyle H. Ungar. Statistical relational learning for link prediction. Em *In Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003*. Citado na pág. 10
- Anauê e Yamate(2009)** Costa Anauê e Fabio Yamate. Semantic Lattes: uma ferramenta de consulta de informações acadêmicas da base Lattes baseada em ontologias, 2009. Citado na pág. 2, 11
- Cervantes(2014)** Everlyn Perez Cervantes. Análise de Redes de Colaboração Científica: Uma Abordagem Baseada em Grafos Relacionais com Atributos, 2014. Citado na pág. 2, 3, 10, 11
- Ebadi e Schiffauerova(2015)** Ashkan Ebadi e Andrea Schiffauerova. How to become an important player in scientific collaboration networks? *Journal of Informetrics*, 9(4):809–825. ISSN 17511577. doi: 10.1016/j.joi.2015.08.002. URL <http://dx.doi.org/10.1016/j.joi.2015.08.002><http://linkinghub.elsevier.com/retrieve/pii/S1751157715000565>. Citado na pág. 9
- Galego(2013)** Eduardo Ferreira Galego. Extração e Consulta de Informações do Currículo Lattes baseada em Ontologias Eduardo Ferreira Galego Dissertação apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo para obtenção do título de Mestre em Ciências da Computação. página 70. Citado na pág. 11
- Guarino et al.(2009)** Nicola Guarino, Daniel Oberle e Steffen Staab. *Handbook on Ontologies*, chapter What Is an Ontology?, páginas 1–17. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-92673-3. doi: 10.1007/978-3-540-92673-3_0. URL http://dx.doi.org/10.1007/978-3-540-92673-3_0. Citado na pág. 6
- Halevy et al.(2009)** Alon Halevy, Peter Norvig e Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12. ISSN 1541-1672. doi: 10.1109/MIS.2009.36. URL <http://dx.doi.org/10.1109/MIS.2009.36>. Citado na pág. 1
- Hasan et al.(2006)** Mohammad Al Hasan, Vineet Chaoji, Saeed Salem e Mohammed Zaki. Link prediction using supervised learning. Em *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*. Citado na pág. 10, 11
- Kay(2014)** Alan Kay. The future doesn't have to be incremental, 2014. URL <https://www.youtube.com/watch?v=gTAghAJcOIo>. DEMO Enterprise 2014. Citado na pág. 1
- Mena-Chalco e Junior(2009)** Jesús Pascual Mena-Chalco e Roberto Marcondes Cesar Junior. scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39. ISSN 0104-6500. doi: 10.1007/BF03194511. URL <http://dx.doi.org/10.1007/BF03194511>. Citado na pág. 9
- Popescul e Ungar(2007)** Alexandrin Popescul e Lyle H. Ungar. Feature generation and selection in multi-relational statistical learning. Em Lise Getoor e Ben Taskar, editors, *Introduction to Statistical Relational Learning*, chapter 16, páginas 453–475. The MIT Press. Citado na pág. 11

Raedt(2008) Luc De Raedt. *Logical and Relational Learning*. Cognitive Technologies. Springer Berlin Heidelberg, Berlin, Heidelberg. ISBN 978-3-540-20040-6. doi: 10.1007/978-3-540-68856-3. URL <http://www.springer.com/us/book/9783540200406><http://link.springer.com/10.1007/978-3-540-68856-3>. Citado na pág. 2, 6

Russell e Norvig(2009) Stuart Russell e Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edição. ISBN 0136042597, 9780136042594. Citado na pág. 5