Uma Abordagem Baseada em Ontologias para a Predição de Ligações entre Pesquisadores 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: Profª. Drª. Renata Wassermann

São Paulo, março de 2016

Uma Abordagem Baseada em Ontologias para a Predição de Ligações entre Pesquisadores 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 entre Pesquisadores em Redes de Colaboração Científica**. 2016. 20 f. Exame de qualificação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

A comunidade científica pode ser vista como uma rede onde cada pesquisador se relaciona com outros pesquisadores através de colaborações, como na co-autoria de um artigo científico. Alguns trabalhos utilizam ontologias para modelar o domínio das redes de colaboração, e outros aplicam técnicas de aprendizado de máquina para prever ligações entre pessoas dentro dessas redes. Entretanto, algumas limitações existem nessas metodologias por utilizarem formas pouco expressivas de representação dessas relações, ou por não aproveitarem características específicas das entidades na análise desse domínio. A proposta do presente trabalho é criar uma Ontologia capaz de indicar características próprias dessas pessoas, e descrever relações entre elas, aplicando esse conhecimento em um modelo de aprendizado capaz de descobrir e prever novas relações entre esses pesquisadores.

Palavras-chave: redes de colaboração científica, ontologia, aprendizado de máquina, predição de ligaçoes.

Abstract

ARAUJO, T. H. D. **Uma Abordagem Baseada em Ontologias para a Predição de Ligações entre Pesquisadores em Redes de Colaboração Científica**. 2016. 20 f. Exame de qualificação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2016.

Elemento obrigatório, elaborado com as mesmas características do resumo em língua portuguesa. De acordo com o Regimento da Pós- Graduação da USP (Artigo 99), deve ser redigido em inglês para fins de divulgação. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.

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

Sumário

Lista de Abreviaturas

SPARQL	(*SPARQL Protocol and RDF Query Language*)
SQL	(*Structured Query Language*)
SRL	(*Statistical Relational Learning*)
SVM	(*Support Vector Machine*)
SWRL	(*Semantic Web Rule Language*)

Lista de Figuras

Lista de Tabelas

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 começou a ser analisado e utilizado como base para a contrução de novos conhecimentos.

Por essa razão, quando falamos da quantidade maciça de informação disponível, conhecida como *Big Data*, devemos pensar também que é possível extrair muito **significado** 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 a riqueza de significados que pode ser extraída desses dados.

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 também em levantamentos sobre produtividade, impacto de uma pesquisa, relevância e popularidade de um projeto ou uma área, como formas de se mensurar e quantificar o progresso científico.

Várias técnicas de análise podem ser empregadas nesse problema, desde as mais simples, como relatórios de produtividade, até as mais complexas, como os agrupamentos 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 pessoas, que permite prever, com boa segurança, novas colaborações entre pesquisadores, através da verificação de suas características e dos seus trabalhos publicados.

Como é usual no método científico, encontramos algumas limitações nessas ferramentas, 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 simplificariam essa tarefa, e aumentariam sua eficácia.

Com o presente trabalho, esperamos contribuir para esse tipo de análise, aplicando técnicas

de Sistemas Baseados em Conhecimento, e comparando o resultado com outros trabalhos, para demonstrar a eficácia das técnicas propostas.

1.1 Motivação

Uma rede social é uma estrutura relacional que podemos facilmente representar por um grafo com atributos (Cervantes (2014)), ou por uma ontologia (Anauê e Yamate (2009)), ou até mesmo como um banco de dados. Uma rede de colaboração científica 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 em 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 dessa 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. Por causa disso, é preciso transformar uma representação relacional mais rica (como um grafo) em uma representação mais simples (como uma tabela). Portanto, podemos dizer que uma ferramenta de aprendizado supervisionado desse tipo utiliza representações pouco expressivas. Segundo Raedt (2008), existem várias representações que podem ser utilizadas para modelar 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.

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 onde estão inseridas as suas publicações mais importantes. E não é nada fácil representar esse tipo de informação em tabelas de atributos e valores.

A principal motivação do trabalho é a possibilidade de se utilizar conhecimento prévio do domínio, também chamado de *background-knowledge*. A proposta é construir uma ontologia que representa a rede de colaboração, 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 da rede de colaboração, enriquecendo os dados a serem explorados pelo modelo de predição. Esperamos, com isso, aumentar a eficiência do modelo.

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 análise e validação. Sem contar que esse conhecimento a respeito de uma entidade do domínio pode se espalhar para outras, e gerar novos atributos.

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, se torna 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 e colaborações, e instituições.
- Construir consultas em SPARQL, que utilizam regras de inferência na Ontologia, que 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, e comparar a eficácia dos dois modelos.
- 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 ítens:

1. **Estudo dos resultados de Cervantes (2014)**:
   O objetivo desse ítem 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**:
   O objetivo desse ítem é ...
3. **Propor uma forma de extração de conhecimento prévio da ontologia**:
   O objetivo desse ítem é propor um método de extração das informações obtidas via consulta da ontologia, construção de uma base de conhecimento prévio, e utilização desse conhecimento no enriquecimento do modelo, via adição de novos atributos aos vértices da rede de colaboraçào 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. Popular a ontologia, e extrair o conhecimento prévio. Rodar o modelo de predição com e sem o enriquecimento com conhecimento, e comparar os resultados.

1.4 Contribuições

A principal contribuição esperada deste 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 de tal 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, e 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 Capítulo 3, exploramos alguns trabalhos relacionados, algumas questões que ainda não foram exploradas em outros trabalhos, e justificamos a importância deste projeto de pesquisa. Finalmente, no Capítulo 4, delimitamos o escopo do projeto e discutimos algumas conclusões. O Capítulo 5 trata do cronograma de atividades necessárias para a execução do projeto.

Capítulo 2

Conceitos

Texto texto.

2.1 Fundamentos

Texto texto.

Capítulo 3

Trabalhos Correlatos

Texto texto.

3.1 Fundamentos

Texto texto.

Capítulo 4

Conclusões

Texto texto.

Capítulo 5

Cronograma

O cronograma de atividades está descrito na tabela a seguir.

	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov
1	2	3	4	5	6	7	8	9	10
2	2	3	4	5	6	7	8	9	10
3	2	3	4	5	6	7	8	9	10
4	2	3	4	5	6	7	8	9	10

Tabela 5.1: *Cronograma de atividades*

Referências Bibliográficas

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. URL http://www.pcs.usp.br/{~}pcspf/2009/Trabalhos/Cooperativo/G4/monografia.pdf. Citado na pág. 2

Cervantes(2014) Everlyn Perez Cervantes. Análise de Redes de Colabração Científica: Uma Abordagem Baseada em Grafos Relacionais com Atributos, 2014. Citado na pág. 2, 3

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

Kay(2014) Alan Kay. The future doesn't have to be incremental, 2014. URL https://www.youtube.com/watch?v=gTAghAJcO1o. DEMO Enterprise 2014. Citado na pág. 1

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/9783540200406http://link.springer.com/10.1007/978-3-540-68856-3. Citado na pág. 2