

Modelagem e análises dos dados do Portal de Revistas da USP

Marcela Megumi Terakado, Marcos Kazuya Yamazaki

Orientadora: Profa. Dra. Kelly Rosa Braghetto

Instituto de Matemática e Estatística da USP



IME-USP

Introdução

Atualmente, o Portal de Revistas da USP, que é gerenciado pelo Sistema Integrado de Bibliotecas da USP (SIBi-USP), utiliza um modelo de dados relacional para armazenar as informações dos artigos publicados nas revistas da USP. Este banco contém os dados dos artigos, revistas, autores, métricas sobre estes artigos, entre outras informações. Como o relacionamento entre os dados é mais relevante que os atributos deles, este trabalho propôs uma nova modelagem do banco de dados para um melhor aproveitamento de possíveis análises. Sendo o foco, de uma das análises, a influência e impacto de artigos e autores tanto no meio acadêmico como regionais.

Objetivos

Os objetivos deste trabalho foram:

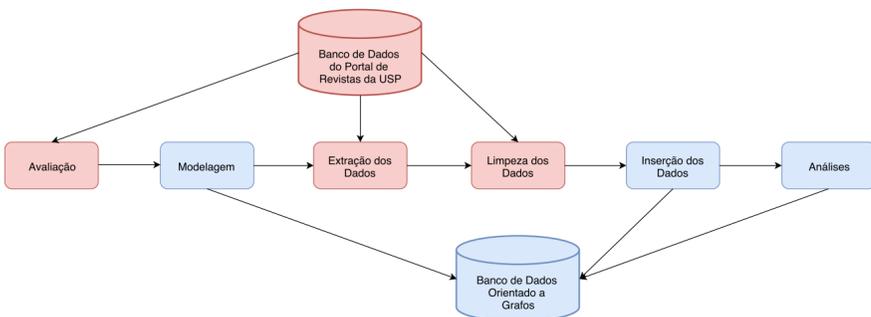
- Analisar o banco de dados do Portal de Revistas da USP
- Propor uma modelagem alternativa utilizando um modelo de dados orientado a grafos
- Realizar análises sobre os dados inseridos no banco de dados de grafos
- Gerar visualizações destas análises

Conceitos

Os conceitos utilizados neste trabalho foram:

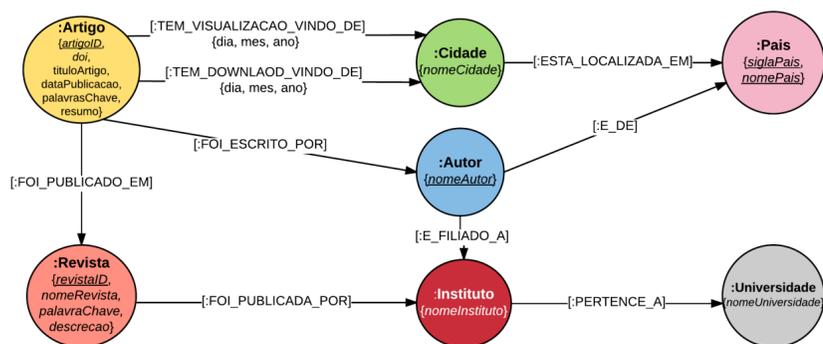
- Banco de dados orientado a grafos, Neo4j
- Métodos para detecção de comunidade: Girvan-Newman [3], Louvain [1], Autovetores [2] e Caminhada aleatória [4]

Fluxograma do Trabalho

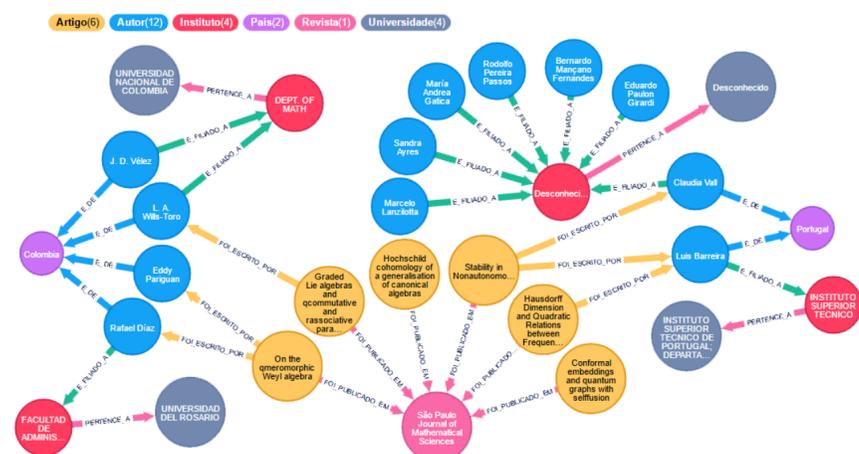


Modelagem do banco de dados de grafos

Diagrama contendo os nós e relacionamentos do banco modelado.



Consulta no banco de dados Neo4j após a criação dos nós, propriedades e relacionamentos e inserção dos dados extraídos do banco de dados do Portal de Revistas da USP.



Análises

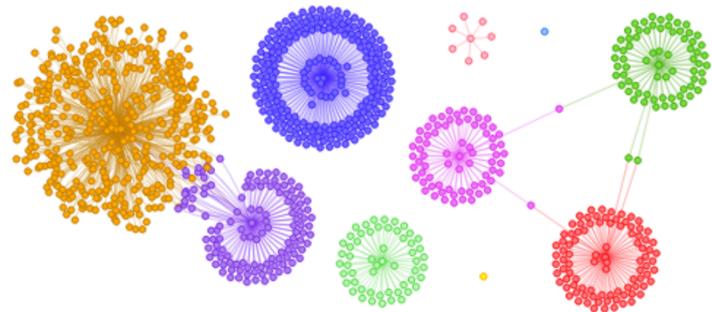
Das análises de detecção de comunidade, foram realizados 2 experimentos:

- grafo completo, contendo todos autores e artigos
- subgrafo contendo 1100 autores, 2137 artigos e 1118 arestas entre os autores

No primeiro experimento, obtivemos os seguintes resultados:

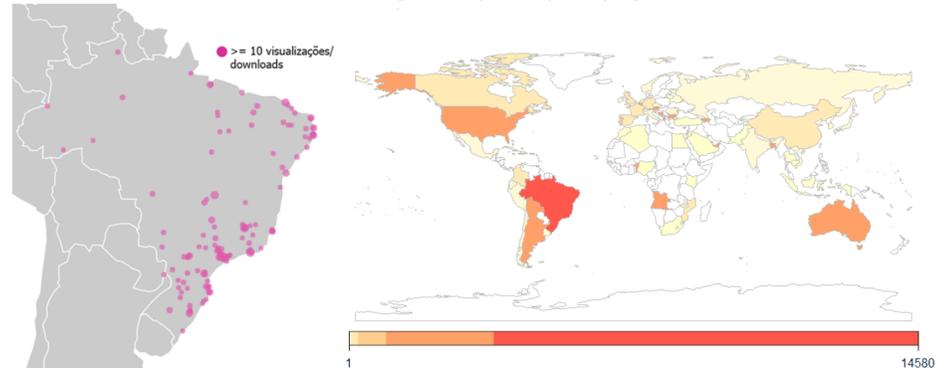
- Louvain: 30600 comunidades detectadas
- Autovetores de Matrizes: 33495 comunidades detectadas
- Random-walk: 30484 comunidades detectadas

No segundo experimento, todos os métodos encontraram 10 comunidades.



Outras análises realizadas sobre os dados inseridos no banco de dados de grafos foram em relação às métricas. A métrica mais explorada neste trabalho foi a localização geográfica dos acessos aos artigos. Para medir o impacto de cada artigo em cada região foram gerados: gráficos dos acessos a um artigo realizados no Brasil e mundialmente.

Métricas do artigo: On the qmeromorphic Weyl algebra



Considerações Finais

Com este trabalho foi possível verificar o quão importante é um estudo dos dados que serão guardados e uma modelagem que esteja de acordo com as especificações realizadas antes da criação e utilização de um banco de dados. Além disso, foi constatado que o modelo de dados de grafos é mais eficiente do que um relacional quando os relacionamentos são mais significativos do que os valores em si. E, por fim, as análises visuais se mostraram úteis em diversos aspectos, como por exemplo a visualização das comunidades dos autores contidos no banco de dados.

Trabalhos Futuros

Neste trabalho, só foram abordados os métodos de detecção de comunidade que levam em consideração a estrutura do grafo, ou seja, só consideram a conexão dos autores entre si. Um trabalho futuro seria rodar outros métodos de detecção de comunidade que consideram o conteúdo também. Os conteúdos seriam as palavras-chave ou as palavras contidas no título ou no resumo dos artigos.

References

- [1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- [2] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.
- [3] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- [4] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. In *International Symposium on Computer and Information Sciences*, pages 284–293. Springer, 2005.

Agradecimentos

Agradecemos às pessoas do SIBi-USP pela disponibilização dos dados utilizados neste trabalho.