Em Questão
ISSN: 1807-8893
ISSN: 1808-5245
emquestao@ufrgs.br
Universidade Federal do Rio Grande do Sul
Brasil

Mapeamento do conhecimento científico: uma proposta de método baseado em Genealogia Acadêmica

Rossi, Luciano; Mena-Chalco, Jesús Pascual
Mapeamento do conhecimento científico: uma proposta de método baseado em Genealogia Acadêmica
Em Questão, vol. 24, 2018
Universidade Federal do Rio Grande do Sul, Brasil
Disponível em: https://www.redalyc.org/articulo.oa?id=465658737010
DOI: https://doi.org/10.19132/1808-5245240.172-192

Artigos

Mapeamento do conhecimento científico: uma proposta de método baseado em Genealogia Acadêmica

Luciano Rossi
Universidade Federal do ABC, Brasil
luciano.rossi@ufabc.edu.br

Jesús Pascual Mena-Chalco
Universidade Federal do ABC , Santo André,, Brasil
jesus.mena@ufabc.edu.br

DOI: https://doi.org/10.19132/1808-5245240.172-192
Redalyc: https://www.redalyc.org/articulo.oa?id=465658737010

Recepção: 03 Dezembro 2016
Aprovação: 24 Setembro 2018

Resumo:

A estruturação do conhecimento fornece os meios para organizar o saber formal em categorias que facilitem sua gestão e contribuam para sua disseminação. Entretanto os modelos formais de classificação do conhecimento não permitem a observação das relações existentes entre as diferentes categorias ou do fluxo desse conhecimento entre membros da comunidade científica. Neste trabalho, buscamos estruturar o saber científico, representado por mapas da ciência, considerando a estrutura hierárquica fornecida pela Genealogia Acadêmica. Utilizamos, como estudo de caso, a descendência acadêmica de Johann Bernoulli representada em forma de grafo de genealogia e as informações biográficas disponibilizadas pelo Wikipedia em concordância com o padrão desenvolvido pelo Mathematics Subject Classification, além de um glossário de áreas da matemática. Para inferir os tópicos do conhecimento aos matemáticos que não possuíam informações, utilizamos um procedimento de propagação de tópicos que considera a topologia do grafo de genealogia. Consideramos que identificar e estudar o fluxo do conhecimento científico entre gerações de pesquisadores é uma tarefa importante para entender como o conhecimento do estado da arte foi difundido. Contudo, no nosso entendimento, essa tarefa foi pouco explorada pela comunidade científica devido à inexistência de conjuntos de dados. Acreditamos que este trabalho permitirá a definição de um novo método computacional para o estudo de fluxo de conhecimento. A relevância do trabalho recai na possibilidade de descoberta de novas informações que auxiliem a identificar, analisar e estruturar os mapas da ciência.

Palavras-chave: Fluxo de conhecimento, Mapas da ciência, Tópicos de conhecimento, Genealogia acadêmica.

Abstract:

The structuring of knowledge provides the means to organize formal knowledge into categories that facilitate its management and contribute to its dissemination. However, the formal models of knowledge classification do not allow the observation of the relationships between the different categories or the flow of this knowledge among members of the scientific community. In this work, we seek to structure scientific knowledge, represented by maps of science, considering the hierarchical structure provided by academic genealogy. We used, as a case study, all the academic descendants of Johann Bernoulli represented in the form of genealogy graph and the biographical information provided by Wikipedia following the standard developed by Mathematics Subject Classification as well as a glossary of areas of mathematics. To infer knowledge topics from mathematicians who did not have information, we used a topic propagation procedure that considers the topology of the genealogy graph. We believe that identifying and studying the flow of scientific knowledge among generations of researchers is an essential task in understanding how state-of-the-art knowledge has spread. However, in our understanding, this task was little explored by the scientific community due to the lack of data sets. We believe that this work will allow the definition of a new computational method for the study of knowledge flow. The relevance of the work lies in the possibility of discovering new information that helps to identify, analyze and structure the maps of science.

Keywords: Flow of knowledge, Maps of science, Topics of knowledge, Academic genealogy.

1 Introdução

Grande parte do conhecimento acerca da origem e do desenvolvimento do conhecimento científico, representado na forma de disciplinas ou tópicos, é resultado predominantemente da análise de publicações (CRONIN; SUGIMOTO, 2014). Neste contexto, a observação de possíveis influências entre esses tópicos do conhecimento envolve a identificação das áreas de atuação de acadêmicos, que são derivadas de suas respectivas informações biográficas, e das relações entre essas áreas, estruturadas por meio de relacionamentos de orientação acadêmica. O problema de identificar o fluxo do conhecimento passa pela obtenção de uma rede de tópicos, aqui representada por um mapa de influência entre tópicos/disciplinas científicas, que evidencie a estrutura topológica[1] derivada dos seus relacionamentos. Neste trabalho, a expressão mapa da ciência está relacionada com uma rede de tópicos do conhecimento, o qual representa o objeto de estudo que viabiliza as análises referentes ao fluxo do conhecimento científico, os quais constituem, respectivamente, o objetivo e o problema de pesquisa deste trabalho.

A criação de uma rede de tópicos do conhecimento possibilita a realização de análises sobre as influências existentes entre esses tópicos e, consequentemente, sobre a interdisciplinaridade no desenvolvimento da ciência (DAMACENO; ROSSI; MENA-CHALCO, 2017). Comumente, a representação de um mapeamento deste tipo é feita por meio de uma estrutura matemática denominada grafo direcionado. Um grafo é uma estrutura abstrata de dados que permite representar diferentes elementos e seus relacionamentos como vértices e arestas, respectivamente (ROSSI; FREIRE; MENA-CHALCO, 2017). O presente trabalho tem como objetivo propor um método de estruturação de tópicos do conhecimento que resulte na criação de uma rede (ou um grafo) baseada em dados de Genealogia Acadêmica (GA), a qual é definida como o estudo da herança intelectual que é perpetuada por meio de relacionamentos de orientação acadêmica (CRONIN; SUGIMOTO, 2014). O método contempla (i) a prospecção de dados de GA, (ii) a estruturação hierárquica dos dados em grafo de GA, (iii) a identificação das áreas de atuação dos acadêmicos, (iv) a inferência das áreas de atuação por meio da propagação de tópicos no grafo de GA e (v) a criação do grafo de tópicos do conhecimento, no qual os vértices e as arestas representam, respectivamente, os tópicos e suas influências que são herdadas dos relacionamentos de orientação acadêmica. Como estudo de caso, este estudo utiliza a GA do proeminente matemático suíço Johann Bernoulli (1667-1748). A escolha deste conjunto de dados se justifica pela notoriedade de J. Bernoulli no campo da matemática. Johann Bernoulli foi o décimo filho de um casal atuante no ramo do comércio de especiarias, área na qual não demonstrou adequação necessária. Na Universidade de Basileia estudou inicialmente Medicina, porém, devido à influência e tutoria do irmão mais velho Jacob, Johann inicia sua trajetória na Matemática. Além de suas contribuições para matemática, destacando-se na área de cálculo infinitesimal, há diversas referências a tutoria de alunos e/ou colegas ilustres como l'Hôpital, Leonhard Euler e Daniel Bernoulli. A identificação das áreas de atuação dos matemáticos que formam a descendência de Bernoulli considerou as informações biográficas disponíveis pela enciclopédia *Wikipedia* (WIKIPEDIA, 2019) e em dois dicionários de tópicos baseados no sistema *Mathematics* Subject Classification (MSC), além de um glossário de áreas da matemática. Como demonstração do método de estruturação de tópicos do conhecimento, destacamos a criação do grafo de tópicos de pesquisa e sua análise descritiva, a qual considera os tópicos do conhecimento, as relações hierárquicas existentes entre eles e o processo de transição entre os tópicos.

2 Trabalhos correlatos

A estruturação e análise de redes de conhecimento são objetos de estudo em diferentes áreas. O trabalho de Mohammadi e Thelwall (2014) utilizou redes de publicações científicas que são relacionadas por critérios específicos como, por exemplo, os leitores dessas publicações, com o objetivo de analisar a estrutura resultante em função da preferência desses leitores. Por outro lado, as redes podem descrever elementos distintos,

como publicações científicas e patentes, de modo a evidenciar a transferência do conhecimento científico teórico para aplicações tecnológicas, como descrito no trabalho de Ding et al. (2017). Segundo Sorenson et al. (2006), as redes de patentes podem ser úteis, também, para análises que envolvam a identificação do impacto que a localização geográfica das fontes de conhecimento exerce em diferentes atores. A análise do fluxo da informação em redes que representam diversos atores envolvidos em processos de negócios e/ou desenvolvimento tecnológico pode contribuir para a gestão da inovação, conforme descrito por Inomata e Rados (2011). Por fim, Bufrem et al. (2017) demonstra que as redes de citações podem ser importantes como objeto de estudo para a propagação do conhecimento.

Os processos de identificação de tópicos, que representem conhecimento específico, também são amplamente documentados na literatura. Um diferencial importante entre as metodologias consideradas é a fonte de informações utilizada. A identificação de tópicos pode utilizar como fonte de informações fóruns on-line de discussão sobre tópicos de aprendizagem (TOBARRA et al., 2014) ou ainda a análise de expressões em títulos de publicações científicas (BOSCHMA; HEIMERIKS; BALLAND, 2014).

É relevante considerar o ineditismo da abordagem deste artigo, visto que as pesquisas similares, descritas anteriormente, sobre o fluxo do conhecimento são comumente desenvolvidas considerando áreas ligadas à gestão ou sobre a mobilidade geográfica de acadêmicos. Essas pesquisas buscam a identificação dos impactos socioeconômicos que a produção e a disseminação do conhecimento exercem sobre a comunidade de forma geral, baseando-se em redes estruturadas por citações de patentes, relações de coautorias em artigos científicos e citações bibliográficas.

Por fim, destaca-se que este trabalho corresponde a uma melhora substancial do trabalho apresentado por Rossi e Mena-Chalco (2018), dado que foram consideradas duas novas características: (i) a identificação do fluxo do conhecimento sob uma perspectiva histórica, no qual são observadas as dinâmicas das atuações em tópicos do conhecimento, e (ii) a utilização de redes sociais estruturadas por meio de relacionamentos de orientação acadêmica (grafos de GA).

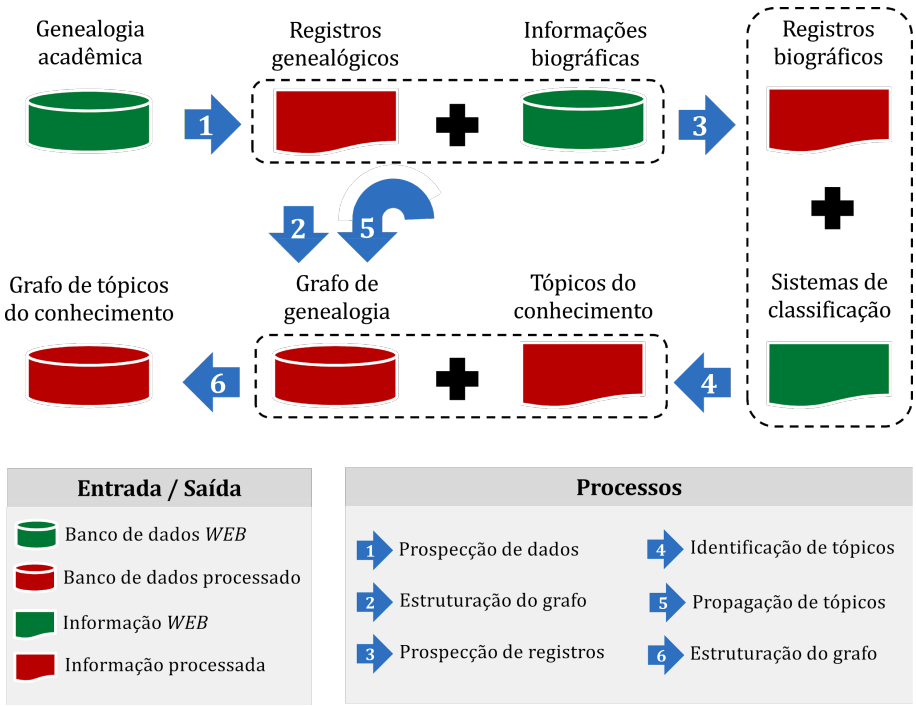

FIGURA 1
Fluxograma do método de estruturação de tópicos do conhecimento
Fonte: Elaborado pelos autores (2019).

3 A proposta do método de estruturação de tópicos do conhecimento

O método proposto está estruturado por meio de seis etapas que consistem da (1) prospecção dos dados genealógicos, (2) estruturação do grafo de genealogia, (3) prospecção de registros biográficos, (4) identificação dos tópicos associados aos acadêmicos, (5) propagação dos tópicos e (6) estruturação do grafo de tópicos. A Figura 1 sumariza o método considerado neste trabalho na forma de fluxograma, no qual são representados os elementos envolvidos (entradas, processos e saídas). Os elementos na cor verde foram obtidos por meio de busca em repositórios Web e os elementos na cor vermelha são resultado de processamento. Nesta seção são detalhados cada um dos elementos representados no fluxograma.

3.1 Estruturação do grafo de genealogia acadêmica

Um grafo direcionado G(V,E) é uma representação matemática na qual G(V) e G(E) são conjuntos de vértices e arestas, respectivamente. Um vértice v $\in$ (V) pode representar diferentes elementos do mundo real, como por exemplo um acadêmico. Uma aresta (v,u) $\in$ G(E) é um par ordenado de vértices e representa uma relação entre os vértice v e u.

O processo de estruturação de um grafo de genealogia acadêmica considera a organização hierárquica de acadêmicos em função dos seus relacionamentos de orientação. Os vértices no grafo representam acadêmicos e os relacionamentos utilizados neste processo são, por exemplo, orientações de mestrado e doutorado (ROSSI; MENA-CHALCO, 2014).

Os dados utilizados para a estruturação do grafo podem ser obtidos por coleta direta junto a bancos de dados genealógicos ou por métodos computacionais que possam inferir os relacionamentos entre os acadêmicos de bancos de dados não vocacionados para a genealogia.

Na Figura 1, o processo 2 identifica a estrutura do grafo de genealogia acadêmica. A entrada do processo são os registros genealógicos, resultantes do processo de prospecção de dados genealógicos (processo 1). A saída do processo é o grafo de genealogia. Neste trabalho consideramos um banco de dados não relacional, orientado a grafos, para o armazenamento dos dados.

3.2 Prospecção de registros biográficos

Os bancos de dados genealógicos podem disponibilizar informações adicionais sobre o acadêmico. Essas informações são, comumente, atributos de identificação como: nome do acadêmico, nome da instituição na qual ocorreu a titulação, o ano de titulação, a área de atuação no momento da titulação ou em um tempo específico, dentre outros. Porém, esses atributos não são suficientes para a inferência do campo de atuação ao qual o acadêmico se dedicou durante sua trajetória.

O processo de prospecção de registros biográficos – identificado na Figura 1 como processo 3 –, tem por objetivo coletar informações biográficas do acadêmico que sejam úteis para caracterizar sua atuação. Esse processo utiliza como entradas o conjunto de registros genealógicos e os repositórios Web de informações biográficas, como por exemplo o *Wikipedia*. Assim, com a utilização de algum atributo do acadêmico, normalmente o nome, realiza-se uma busca pela sua biografia na plataforma considerada coletando todo o registro (página *html*) identificado. Esses registros são processados de modo a retirar elementos textuais irrelevantes (*stop words*) e marcadores de linguagem (*tags html*).

Este processo tem como saída uma lista dos termos (tópicos) mais frequentes, observados na biografia do acadêmico. Assim, cada acadêmico é associado a uma lista de termos que são considerados para a identificação da área de atuação dos acadêmicos pertencentes ao grafo de genealogia. Note que é improvável que a busca

resulte em registros biográficos de todos os acadêmicos. Comumente, há registros deste tipo somente para os acadêmicos mais relevantes em sua área de atuação (no escopo acadêmico ou histórico).

3.3 Identificação de tópicos do conhecimento

Os sistemas de classificação do conhecimento são esquemas hierárquicos que organizam o conhecimento formal em classes. Essas classes podem ser mais ou menos abrangentes, de acordo com o sistema utilizado, e podem ser uma área do conhecimento, uma disciplina, um termo específico, dentre outras possibilidades. Neste trabalho denominamos de tópico do conhecimento toda classe formal documentada por um sistema de classificação. Assim, a depender do sistema considerado, um tópico pode representar uma disciplina ou um termo mais específico.

A Figura 1 evidencia o processo de identificação de tópicos do conhecimento como processo 4. Podem ser consideradas diferentes abordagens para a realização desse processo. Por exemplo, uma abordagem na qual se realiza uma busca direta pelo tópico no registro biográfico do acadêmico. Por outro lado, em uma outra possibilidade, um dicionário de termos derivados dos sistemas de classificação é utilizado para a busca.

Dependendo da abordagem considerada, o resultado do processo de identificação de tópicos pode ser: (i) um conjunto de tópicos, (ii) um vetor de frequências de tópicos ou (iii) um vetor de frequências de termos que compõem os tópicos, sempre associados aos vértices do grafo de genealogia como novos atributos. Conforme descrito anteriormente, o processo de identificação dos tópicos e sua posterior associação aos vértices (rotulação) é possível somente para parte dos vértices. Isto ocorre devido à inexistência de registros biográficos para todos os acadêmicos e/ou à restrição do processo de identificação dos tópicos a partir dos registros.

3.4 Propagação de tópicos do conhecimento

Considerando que as etapas descritas anteriormente não resultam na rotulação (identificação e associação de tópicos aos vértices) de todos os acadêmicos do grafo de genealogia, o processo de propagação de tópicos (identificado como processo 5 na Figura 1) assume este objetivo. O processo utiliza como entradas o grafo de genealogia e a lista de tópicos, sendo que a saída é a rotulação dos vértices no grafo. A seguir, descrevemos as abordagens consideradas neste trabalho.

3.4.1 Propagação de frequências de termos

Para esta abordagem, os vértices do grafo de genealogia estão parcialmente rotulados com um vetor de frequências de termos que compõem os tópicos. A rotulação de um vértice não rotulado v^N considera seus vértices adjacentes rotulados u^R e w^R como dois conjuntos distintos, ambos formados pela ascendência e descendência rotuladas de v^N, respectivamente. O conjunto de ascendentes rotulados de v^N é:

$$A(v^N) = \{u^R : (v^N, u^R) \in G(E)\}$$

e o conjunto de descendentes rotulados de v^N é:

$$D(v^N) = \{w^R : (w^R, v^N) \in G(E)\}.$$

Tratando-se de transmissão de conhecimento, consideramos que há uma perda de conhecimento sobre um determinado termo na relação de orientação acadêmica, ou seja, um orientador tem mais domínio[2] sobre um determinado tópico em relação ao seu orientado. Consideramos ainda que o domínio sobre um tópico é representado pela frequência dos termos no vetor, frequências maiores indicam maior domínio, por outro lado, pouco domínio sobre um termo é representado por menores frequências. Neste contexto, justifica-se a separação da adjacência rotulada de um vértice não rotulado em ascendentes e descendentes.

Um rótulo é atribuído a um vértice quando esse tem pelo menos um vértice adjacente rotulado. Considerou-se, assim, a média ponderada das frequências dos adjacentes para a atribuição dos rótulos. Para os adjacentes ascendentes, há um decremento na frequência média para cada vértice adjacente observado, objetivamos assim representar a redução do domínio do tópico no processo de orientação. Por outro lado, os adjacentes descendentes têm um incremento na mesma proporção. O incremento do valor, que descreve a frequência média do rótulo, para a vizinhança descendente assim como o correspondente decremento para a vizinhança ascendente, permitem simular/emular uma forma de transmissão de conhecimento.

3.4.2 Propagação iterativa de tópicos

Nesta abordagem, os vértices do grafo de genealogia estão parcialmente rotulados com um conjunto de tópicos do conhecimento. A propagação neste caso é um processo iterativo de um vértice rotulado para os adjacentes não rotulados. A primeira iteração considera somente a descendência rotulada do vértice não rotulado, a atribuição de tópicos é feita por meio da interseção dos conjuntos de tópicos de seus descendentes rotulados. Caso haja apenas um descendente rotulado, seus tópicos serão atribuídos ao vértice não rotulado. Nessa iteração são submetidos ao processo todos os vértices não rotulados do grafo. Se um vértice não apresenta descendência rotulada, ou se a interseção for um conjunto vazio, ele mantém a condição de não rotulado. Na iteração seguinte, o processo de propagação é repetido com a interseção da ascendência rotulada. A descendência e ascendência rotuladas são utilizadas de maneira alternada até um determinado número de iterações que é definido por um limiar de rotulação de vértices. A definição de um limiar é importante porque a estrutura do grafo não garante a rotulação total de seus vértices.

3.5 Estruturação do grafo de tópicos do conhecimento

Neste último processo, o grafo de genealogia rotulado é utilizado como entrada para a estruturação do grafo de tópicos do conhecimento (na Figura 1 esse é o processo 6). Assim como no processo anterior, este processo também constitui um dos principais objetivos de pesquisa. Considerou-se, então, uma abordagem, baseada no produto cartesiano entre os tópicos associados ao orientador e ao orientado, para realizar a estruturação do grafo de tópicos. A estruturação do grafo de tópicos do conhecimento considera como vértices no grafo os tópicos identificados e suas arestas são definidas pelo produto cartesiano entre os conjuntos de tópicos de vértices adjacentes.

4 Um estudo com o método de estruturação de tópicos do conhecimento

Os dados considerados para demonstrar o método de estruturação de tópicos do conhecimento foram: (i) um conjunto de dados genealógicos, (ii) um conjunto de registros biográficos e (iii) dois sistemas de classificação. Esses dados são descritos a seguir.

4.1 Banco de dados e grafo de genealogia acadêmica

Considerou-se, como estudo de caso, um conjunto de dados composto pelos descendentes acadêmicos do matemático suíço Johann Bernoulli (1667-1748). Esses dados estão disponíveis no *Mathematics Genealogy Project* (MGP), uma plataforma *web*, desenvolvida por iniciativa da Universidade da Dakota do Norte, nos EUA, que tem por objetivo reunir dados sobre todos os doutores em matemática do mundo. Dentre diversos atributos, disponíveis pela plataforma, destaca-se as relações de orientação acadêmica de cada matemático.

A descendência acadêmica de Johann Bernoulli foi prospectada junto à plataforma MGP por meio de consultas recursivas a partir de um identificador único atribuído ao matemático. Foram obtidos, em 23 de setembro de 2017, 100.221 registros de matemáticos que pertencem à descendência de Bernoulli, incluindo o próprio. Esses registros estão conectados por 110.061 relacionamentos de orientação acadêmica. Foi utilizado um banco de dados não relacional orientado a grafos para a estruturação desses registros, no qual cada matemático é representado por um vértice no grafo e os relacionamentos de orientação são representados por arestas direcionadas que interligam orientador a orientado.

O grafo de genealogia resultante consiste de uma única componente conexa cuja origem é o vértice que representa J. Bernoulli, o qual pertence à geração inicial. Os vértices adjacentes ao vértice de J. Bernoulli formam a geração um, e assim sucessivamente. Os vértices do grafo estão distribuídos em 18 gerações. O maior número de vértices está concentrado na geração 12 (25.337), consequentemente, essa geração apresenta o maior número de orientações recebidas, e a geração anterior (11) o maior número de orientações realizadas, as quais são representadas pelo grau de entrada (28.327) e pelo grau de saída (28.254), respectivamente. Com relação ao grau de entrada, há uma linearidade entre as gerações, visto que o grau de entrada médio é próximo de um em todas as gerações. Por outro lado, a geração com maior grau de saída médio é a geração seis, na qual cada matemático dessa geração orientou, em média, 4,38 alunos.

O grafo, anteriormente descrito, apresenta uma estrutura típica observada em grafos de genealogia acadêmica. Notamos que, a partir da geração 12 há uma diminuição no número de matemáticos em cada uma das gerações posteriores. O número de matemáticos em cada geração depende da idade acadêmica da geração anterior. Matemáticos com menos tempo de titulação orientaram menos alunos, proporcionalmente. Por outro lado, jovens matemáticos apresentam uma probabilidade maior de ainda serem atuantes na formação da geração posterior.

4.2 Informações biográficas

A identificação dos tópicos do conhecimento foi feita com a utilização das informações disponíveis pela *Wikipedia*. A opção por essa fonte de informações considerou que, apesar da falta de referências formais, que validem a informação disponível, e do dinamismo característico dessa fonte de dados biográficos, a iniciativa de um anônimo em escrever sobre um determinado matemático pode evidenciar sua importância para a área. Assim, partiu-se da premissa que a existência de um registro no *Wikipedia*, de algum modo, evidencia a contribuição do matemático para o desenvolvimento da área. Outra limitação, proveniente da utilização da *Wikipedia*, está ligada à falta de estrutura dos dados considerados e da possibilidade de existência de homônimos que não sejam necessariamente o indivíduo de interesse. Nesses casos, o dicionário de tópicos

tem papel importante, visto que ele restringe a inferência de um tópico a um indivíduo, considerando um sistema de classificação predefinido. Assim, as biografias que não expressam as características de atuação do acadêmico aderentes ao escopo considerado são descartadas.

A busca pelas páginas *web* dos 100.221 matemáticos foi feita por meio do nome completo de cada matemático, considerando o casamento exato de nomes, na versão em inglês da plataforma *Wikipedia* e resultou na identificação de 13.477 páginas (13,48%) com informações biográficas dos indivíduos.

4.3 Sistemas de classificação

Para a composição dos dicionários de tópicos do conhecimento foram considerados dois sistemas de classificação específicos para a área da matemática. O MSC[3] é um esquema de classificação alfanumérico criado pela revista *Mathematical Reviews*, publicada pela *American Mathematical Society* e pela *Zentralblatt Math*, o maior serviço internacional de revisão de artigos em Matemática Pura e Aplicada. O MSC disponibiliza uma estrutura hierárquica de três níveis para a classificação das áreas de atuação. Para este trabalho, foi considerado apenas o primeiro nível que abrange 62 tópicos. O segundo sistema considerado é um glossário de áreas (disciplinas) da matemática[4] composto por 446 tópicos.

As descrições dos tópicos do MSC foram utilizadas para a formação de um conjunto de 100 termos de interesse, os quais constituem a base para a classificação dos registros biográficos dos matemáticos. Por meio de uma busca nas páginas, foi identificada a frequência dos termos nas informações de cada matemático. Como resultado, é atribuído um vetor com as frequências de cada um dos termos aos matemáticos cujas páginas foram identificadas.

Os sistemas de classificação MSC e o glossário de áreas foram utilizados para a composição de dois dicionários que refletem seus tópicos. Nesse caso, a busca nos registros biográficos é feita diretamente pelo tópico e não pelos termos que o compõem. Note que, para descrições mais complexas como, por exemplo: "*Harmonic analysis including Fourier analysis Fourier transforms trigonometric approximation trigonometric interpolation orthogonal functions*", há a necessidade de dividi-las em outras mais específicas, como: "*harmonic analysis*", "*Fourier analysis*", "*Fourier transforms trigonometric*", "*approximation trigonometric*" e "*interpolation orthogonal functions*". Nesse caso, a ocorrência de qualquer uma das descrições implica em atribuir o tópico ao matemático em questão.

4.4 Testando o método

Os grafos de tópicos do conhecimento que são apresentados a seguir, e suas respectivas análises, referem-se à descendência de Johann Bernoulli[5].

A estruturação do primeiro grafo de tópicos considerou um dicionário de termos que compõem os tópicos disponibilizados pelo sistema MSC. Esses termos foram atribuídos como rótulos aos respectivos vértices no grafo na forma de um vetor de frequências e, após o processo de propagação dos rótulos, foi realizada a conversão de termos para tópicos. A propagação dos vetores de frequências de termos seguiu em acordo com a descrição da Seção 3.4.1 e a estruturação do grafo de tópicos seguiu em acordo com a descrição da Seção 3.5. O processo de propagação do vetor de frequências de termos é capaz de inferir informações a todos os vértices do grafo de genealogia. Porém, a atribuição de um tópico por meio dos termos não é possível em todos os casos, visto que os valores de frequências de termos podem ser nulos ou os termos com valores não nulos são insuficientes para a associação de um tópico. Dos 13.477 vértices que possuem vetores de frequências de termos, a inferência de tópicos foi possível em apenas 2.502 (18,57%) casos.

O MGP atribuiu a J. Bernoulli a classificação 92 (*Biology and other natural sciences*) disponibilizada pelo MSC. Acreditamos que essa classificação não reflete a atual influência acadêmica de J. Bernoulli e foi

atribuída, provavelmente, devido a seu trabalho de doutorado estar relacionado com um tópico da área da Medicina. Considerando a biografia de Bernoulli, nota-se que sua influência na comunidade acadêmica está relacionada com suas contribuições na área da Matemática, o que diverge da classe inferida pelo MGP. O método, anteriormente descrito, aplicado ao registro de Bernoulli no *Wikipedia* resultou no seguinte conjunto de termos e suas respectivas frequências: *biography* (1), *calculus* (5), *education* (1), *history* (3), *information* (2), *mathematics* (8), *mechanics* (2) e *theory* (2). Ainda como resultado da aplicação do método, as classificações MSC atribuídas para Bernoulli, de acordo com seu vetor de frequências, foram as seguintes: (i) *History biography,* (ii) *Mechanics* e (iii) *Mathematics education*.

A Figura 2 apresenta o ranking dos tópicos identificados em cada uma das gerações às quais os matemáticos pertencem. A ordem de representação dos tópicos nas gerações evidencia a respectiva influência exercida, ou seja, os tópicos posicionados na parte superior do eixo vertical que representa a geração são mais influentes do que aqueles posicionados na parte inferior. A influência neste caso é capturada por meio da frequência daquele tipo de relacionamento. O ranking concentra 31 tópicos do conhecimento relacionados, o que representa 50% dos tópicos disponíveis pelo MSC, considerando somente os relacionamentos exclusivos, sem repetição de tópicos e/ou gerações. Por exemplo, se há dois relacionamentos conectando os mesmos tópicos pertencentes às mesmas gerações, então esses relacionamentos são unificados sendo considerados uma única vez. Porém, considerando repetições dos atributos tópico e geração, o total observado é de 388.595 relacionamentos.

A primeira geração reúne os três tópicos de atuação de Bernoulli de acordo com a respectiva influência exercida: *History biography*, *Mathematics education* e *Mechanics*. O primeiro tópico se mantém como o mais influente nas gerações subsequentes, perdendo força a partir da 15ª geração. Este padrão pode ser influenciado pelo método. Os termos que compõem este tópico específico (*history* e *biography*) são frequentes em textos biográficos e podem ocasionar um artefato que influencia na inferência da área de atuação do acadêmico. Isto mostra a importância da escolha adequada do dicionário considerado para a completa identificação dos tópicos e sua influência nas gerações.

A estruturação do grafo de tópicos, representado na Figura 3, considerou um dicionário elaborado a partir do sistema MSC, esses tópicos foram atribuídos como rótulos aos respectivos vértices no grafo. A propagação dos tópicos no grafo de GA seguiu em acordo com a descrição da Seção 3.4.2 e a estruturação do grafo de tópicos seguiu em acordo com a descrição da Seção 3.5.

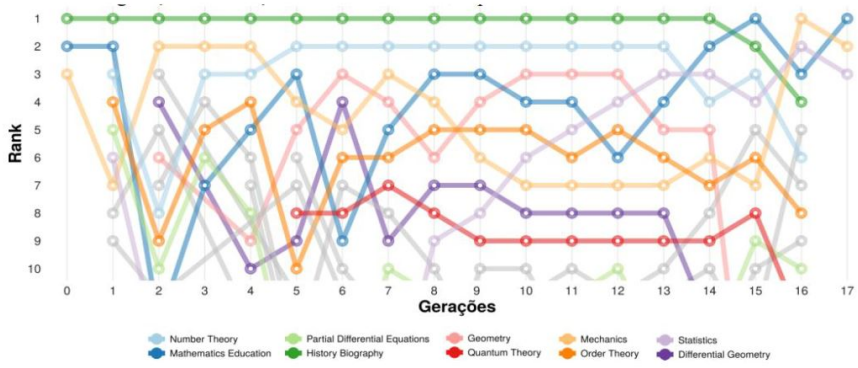

FIGURA 2
Ranking dos tópicos do MSC, representado no eixo vertical, identificados em cada geração de atuação dos matemáticos, representadas no eixo horizontal
Fonte: Elaborado pelos autores (2019).

O processo de identificação de tópicos considerou uma frequência igual a 1. O processo de propagação dos tópicos considerou um limiar de 95% de vértices rotulados para finalização, ou seja, as iterações são interrompidas quando pelo menos 95% dos vértices possuam rótulos obtidos a partir dos rótulos de sua

vizinhança. O grafo resultante apresenta 46 vértices correspondentes as tópicos identificados e 860.097 arestas que representam os relacionamentos entre os tópicos do conhecimento.

A Figura 3 apresenta o grafo resultante do método, cujos vértices representam os tópicos identificados e as arestas as principais relações de influência entre estes tópicos. Nessa representação foi mantida somente a aresta de maior frequência para cada tópico influente (aresta emergente do tópico) e influenciado (aresta incidente no tópico). O tópico *Geometry* (51) é o mais influente e influenciado, considerando o número de tópicos, e, também, é o mais influente segundo o total de arestas emergentes (99.784). *Geometry* é um *hub* (i.e, um vértice com muitas ligações) nesta rede, visto que ele concentra as relações de influência de diversos tópicos, entre os quais se destaca *Differential geometry* (7.199). Outros importantes concentradores de influência são: *Computer science*, *Statistics* e *Mechanics*.

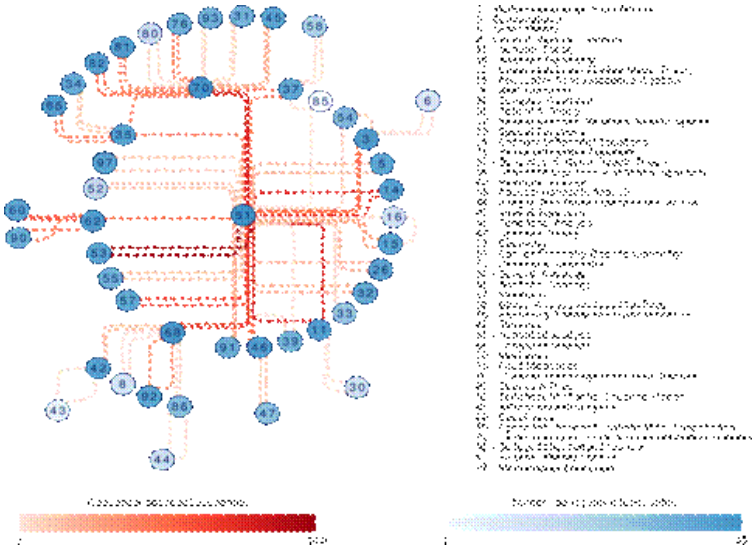

FIGURA 3
Grafo de tópicos do conhecimento onde os vértices representam os tópicos identificados e as arestas representam a relação de influência mais frequente
Fonte: Elaborado pelos autores (2019).

A estruturação deste tipo de grafo de tópicos do conhecimento pode ser importante para a observação da multidisciplinaridade na formação dos acadêmicos, em função do respectivo tópico de atuação. Tópicos que sofrem influência de diversos outros tópicos, podem ser considerados como multidisciplinares em termos de orientação, como, por exemplo, o tópico *Geometry*. Por outro lado, há tópicos com características de especialistas, no sentido de não haver muitas relações com outros tópicos. Um exemplo deste último caso é o tópico *Differential geometry*, que se relaciona apenas com *Geometry*. Outra observação importante é a presença majoritária de relações autoinfligidas, nas quais as orientações ocorrem entre os mesmos tópicos. Apesar deste comportamento parecer natural, o grande número de relações deste tipo pode ser resultado do processo de propagação.

A estruturação do segundo grafo de tópicos do conhecimento considerou o mesmo método aplicado ao grafo anterior com a utilização de um dicionário de tópicos formado a partir do glossário de áreas. O grafo resultante, representado na Figura 4, possui 26 vértices correspondentes aos tópicos identificados. Nesse último exemplo, consideramos arbitrariamente somente as arestas (relacionamentos) cujas frequências observadas foram superiores a 1.000, assim, as áreas com maior destaque, em termos de influência acadêmica, são *Analysis*, *Algebra* e *Geometry*.

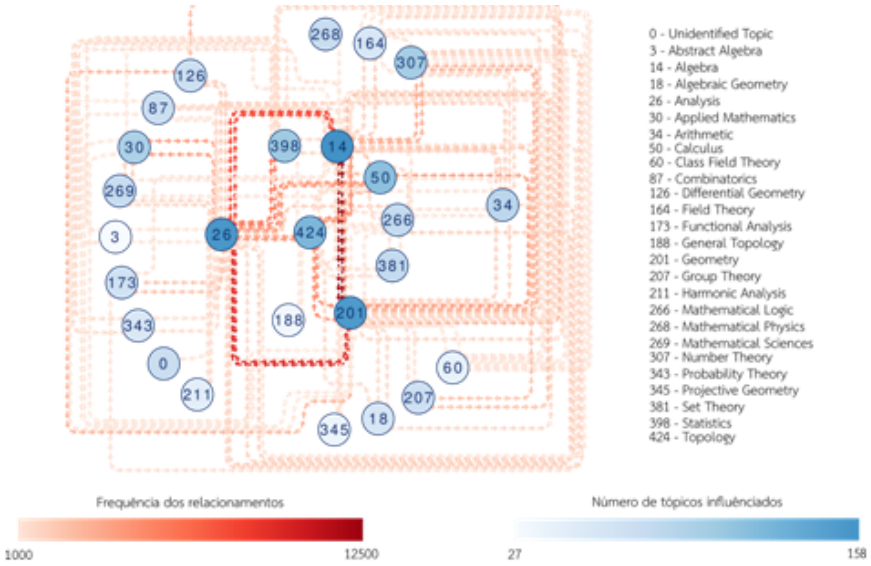

FIGURA 4
Grafo de tópicos do conhecimento onde os vértices representam os tópicos identificados e as arestas representam a relação de influência
Fonte: Elaborado pelos autores (2019).

Os grafos de tópicos do conhecimento que foram estruturados, bem como suas respectivas análises, constituem um resultado preliminar com o objetivo de demonstrar as possibilidades da abordagem e verificar a melhor parametrização dos processos envolvidos. Para todos os casos, consideramos os mesmos conjuntos de dados, registros biográficos e processo de estruturação do grafo de tópicos. A escolha do sistema de classificação é um ponto importante do método. Sistemas mais abrangentes (com maior número de tópicos) e cujos tópicos sejam mais específicos (com menor número de termos) permitem a obtenção de grafos mais assertivos. Há ainda indicadores que apontam para a utilização de um dicionário de tópicos aliado ao processo de propagação iterativa como forma de obter grafos com maior densidade. Assim como descrito anteriormente, valores maiores considerados nas frequências diminuem o número de vértices no grafo, porém, essa escolha aumenta a precisão de sua representatividade.

6 Considerações finais

O mapeamento de tópicos do conhecimento de acadêmicos e sua posterior organização por meio da estrutura fornecida pela GA, resulta em uma estrutura rica em possibilidades de exploração. Dentre essas possibilidades destaca-se a identificação do processo de transição entre os tópicos. A identificação desse processo é importante para o entendimento do padrão de desenvolvimento do conhecimento científico, estratificado por áreas do conhecimento, escalas temporais, geolocalização, dentre outros extratos possíveis.

Neste trabalho apresentamos uma proposta de método que permite construir um grafo de tópicos do conhecimento, a partir de dados abertos, que pode servir de base para descrever o processo de desenvolvimento do conhecimento científico. Acreditamos que o aprofundamento na pesquisa de modelos baseados em GA pode contribuir de forma importante para a obtenção de novos conhecimentos sobre a formação da ciência, que certamente contribuirá para o desenvolvimento da sociedade de forma geral.

A GA, como uma especialização da disciplina de Análise de Redes Complexas, constitui-se como uma rede social rica em conteúdo devido às características de seus relacionamentos. Os indivíduos que se relacionam apresentam atributos convergentes e fortemente baseados no conhecimento científico, na inovação e na evolução de forma individual e coletiva.

Como próximos passos deste trabalho pretendemos utilizar (i) o conjunto completo e atualizado dos matemáticos cadastrados no MGP, (ii) o conjunto completo e atualizado de doutores de mais de 50 áreas diferentes, prospectado da plataforma *Academic Family Tree* e (iii) o conjunto completo e atualizado de acadêmicos cadastrados na Plataforma Lattes. Pretendemos, ainda, considerar outras fontes de informação sobre os acadêmicos que possam servir de base para a identificação dos tópicos do conhecimento, como por exemplo o banco de teses e dissertações e as bases de dados de periódicos. Um desafio importante, nesta abordagem, é a ausência destas informações para os acadêmicos mais antigos, visto que estes registros não são contemplados pelas bases de dados digitais contemporâneas.

O MSC é um sistema de classificação oficial importante para a área da matemática. Porém, o primeiro nível hierárquico, que foi utilizado neste trabalho, resulta em uma classificação pouco específica. A utilização de todos os níveis hierárquicos do MSC resultará na identificação de tópicos mais específicos. Outra forma de aumentar a especificidade da classificação é a utilização dos termos identificados para os acadêmicos no processo de propagação. Isto é, considerar os termos ao invés dos tópicos para o processo de propagação e, posteriormente, identificar os tópicos. Assim, entendemos que a classificação dos acadêmicos se tornará mais específica. Realizar a rotulação sem uma classificação específica pode ser uma abordagem adicional.

Agradecimentos

Os autores agradecem a Rafael Jeferson Pezzuto Damaceno e aos avaliadores anônimos pela leitura atenta e observações para o aprimoramento deste trabalho.

Referências

BOSCHMA, Ron; HEIMERIKS, Gaston; BALLAND, Pierre-Alexandre. Scientific knowledge dynamics and relatedness in biotech cities. Research Policy, v. 43, n. 1, p. 107-114, 2014.

BUFREM, Leilah Santiago; SILVA, Fábio Mascarenhas; SOBRAL, Natanael Vitor Análise das influências intelectuais na produção científica da área de Ciência da Informação: um estudo sobre os bolsistas de produtividade em pesquisa (PQ-CNPq). Em Questão, v. 23, p. 115-141, 2017.

CRONIN, Blaise; SUGIMOTO, Cassidy R. Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact. Cambridge, Massachusetts: MIT Press, 2014.

DAMACENO, Rafael Jeferson Pezzuto; ROSSI, Luciano; MENA-CHALCO, Jesús. P. Identificação do grafo de genealogia acadêmica de pesquisadores: uma abordagem baseada na Plataforma Lattes. In: BRAZILIAN SYMPOSIUM ON DATABASES, 32., 2017, Minas Gerais. Anais... Minas Gerais: UFU, 2017.

DING, Cherng G.; HUNG, Wen-Chi; LEE, Meng-Che; WANG, Hung-Jui. Exploring paper characteristics that facilitate the knowledge flow from science to technology. Journal of Informetrics, v. 11, n. 1, p. 244-256, 2017.

INOMATA, Danielly Oliveira; RADOS, Gregório Jean Varvakis O fluxo da informação tecnológica no processo de desenvolvimento de produtos biotecnológicos: uma construção teórico-conceitual. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12., 2011, Brasília. Anais... Brasília: UnB, 2011.

MOHAMMADI, Ehsan; THELWALL, Mike Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. Journal of the Association for Information Science and Technology, v. 65, n. 8, p. 1627-1638, 2014.

ROSSI, Luciano; MENA-CHALCO, Jesús P. Caracterização de árvores de genealogia acadêmica por meio de métricas em grafos. In: BRASNAM – BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 3., 2014, Brasília. Anais... Brasília: UnB, 2014.

ROSSI, Luciano; FREIRE, Igor Leite; MENA-CHALCO, Jesús P. Genealogical index: A metric to analyze advisor-advisee relationships. Journal of Informetrics, v. 11, n. 2, p. 564-582, 2017.

ROSSI, Luciano; MENA-CHALCO, Jesús P. Criação de grafos de tópicos do conhecimento baseada em genealogia acadêmica. In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 6., 2018, Rio de Janeiro. Anais... Rio de Janeiro: UFRJ, 2018.

SORENSON, Olav; RIVKIN, Jan W.; FLEMING, Lee Complexity, networks and knowledge flow. Research Policy, v. 35, n. 7, p. 994-1017, 2006.

TOBARRA, Llanos; ROBLES-GÓMEZ, Antonio; ROS, Salvador; HERNÁNDEZ, Roberto; CAMINERO, A. Caminero Analyzing the students' behavior and relevant topics in virtual learning communities. Computers in Human Behavior, v. 31, n. 1, p. 659-669, 2014.

WIKIPEDIA, The Free Encyclopedia. Johann Bernoulli. 2019. Disponível em: https://en.wikipedia.org/w/index.php?title=Johann_Bernoulli&oldid=876077724. Acesso em: 22 jan. 2019.

Mapping of scientific knowledge: A method based on Academic Genealogy

Notas

[1] A expressão "estrutura topológica" refere-se ao estudo dos espaços topológicos os quais permitem a formalização de conceitos tais como convergência, conexidade e continuidade.
[2] Por domínio consideramos o conhecimento relativo ao termo, que é expressado pela sua frequência.
[3] Disponível em: https://mathscinet.ams.org/msc/msc2010.html. Acesso em: 31 jan. 2018.
[4] Disponível em: https://en.wikipedia.org/wiki/Glossary_of_areas_of_mathematics. Acesso em: 14 set. 2018.
[5] Disponível em: https://www.mathematik.ch/mathematiker/johann_bernoulli.php. Acesso em: 14 set. 2018.

Financiamento

Fonte: Este trabalho recebeu auxílio financeiro, na forma de bolsa de doutorado, da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).
Beneficiário: Luciano Rossi