A EVOLUÇÃO DA PESQUISA EM CIÊNCIA DA INFORMAÇÃO: UMA ANÁLISE DE TRÊS REVISTAS INTERNACIONAIS POR *TOPIC MODELING* USANDO *LDA*

Luciano Rossi
Doutor em Ciência da Computação
Universidade Cruzeiro do Sul (UNICSul)
E-mail: luciano.rossi@cruzeirodosul.edu.br

Fabio Castro Gouveia
Doutor em Química Biológica
Fundação Oswaldo Cruz (Fiocruz)
E-mail: fgouveia@gmail.com

Jesús P. Mena-Chalco
Doutor em Ciência da Computação
Universidade Federal do ABC (UFABC)
E-mail: jesus.mena@ufabc.edu.br

INTRODUÇÃO

A Ciência da Informação (CI) é uma área interdisciplinar cujo objeto de estudo é a informação. Dentre as diferentes vertentes de atuação, há uma linha referente à produção e comunicação científica. A interação, a comunicação e a difusão científica evoluem ao longo do tempo, sendo essa evolução observável a partir dos interesses de pesquisa de seus atores.

O objetivo deste trabalho é mapear a evolução dos interesses de pesquisa, na CI, ao longo da última década. Para tal, os interesses de pesquisa foram representados por meio da designação de tópicos usando a Alocação Latente de Dirichlet (Latent Dirichlet Allocation, LDA) (BLEI; NG; JORDAN, 2003), um método de aprendizado não supervisionado. Esse método foi aplicado ao conjunto de resumos de artigos publicados na última década em três importantes veículos de comunicação da área: *Journal of Informetrics* (JOI), *Journal of the Association for Information Science and Tech*[1] (JASIST) e *Scientometrics*.

Os resultados mostram as proporções de publicações dos periódicos nos tópicos e a sua associação com as palavras que os compõem. Esses mapas foram construídos de

1 Este periódico era denominado *Journal of the American Society for Information Science and Technology* até o ano de 2013.

modo a representar cinco intervalos bienais. Além disso, evidenciaram-se os interesses de pesquisa dos autores atuantes no Brasil. Para esse caso, consideraram-se como publicações brasileiras aquelas que apresentaram, ao menos, um coautor associado a uma instituição brasileira.

TRABALHOS CORRELATOS

Este trabalho está alinhado com o mapeamento dos interesses de pesquisa dos acadêmicos/cientistas atuantes em uma área interdisciplinar como é a CI. Nesse sentido, há diferentes trabalhos com objetivos similares. A comunidade dos físicos foi objeto de estudo no trabalho de Aleta e demais autores (2019), no qual foram explorados artigos provenientes da base da American Physical Society (período 1980-2006). Considerando os diferentes estágios de carreira dos acadêmicos, foram evidenciadas as modificações no interesse de pesquisa desses atores, por meio de tópicos obtidos a partir de um esquema de classificação já estabelecido.

Por outro lado, Lamba e Madhusudhan (2019) consideraram um conjunto de artigos, provenientes do DESIDOC *Journal of Library and Information Technology*, com o objetivo de mapear os interesses de pesquisa dos acadêmicos na Índia. Os autores utilizaram o LDA para a inferência de tópicos, baseada nos artigos completos. Além disso, há no trabalho uma análise comparativa entre os interesses de pesquisa na Índia contrastados com outros países. A abordagem de Lamba e Madhusudhan (2019) é similar a este estudo, exceto por considerar, como dados de entrada, o texto completo de publicações em um contexto local, contrastando com a utilização de resumos em um contexto global, como é o caso aqui.

Consideramos que este trabalho está em sintonia com uma tendência, contemporânea e crescente, de se mapear a evolução do conhecimento científico. Além disso, a utilização de dados obtidos junto a importantes periódicos da área da CI e a utilização do método LDA, resultam em uma representação temporal inédita na área.

MATERIAIS E MÉTODO

Foram considerados três importantes periódicos da área da CI, mais especificamente: JOI , JASIST, e *Scientometrics*. A opção por esses periódicos foi feita considerando (i) o fato de serem revistas internacionais, (ii) a reconhecida qualidade evidenciada por critérios objetivos de indexação em bases internacionais (WoS e Scopus) e (iii) a maturidade editorial evidenciada por mais de uma década de atuação na área. Os dados foram obtidos por meio da Web of Science,[2] na forma de um conjunto de registros referentes às publicações

2 Coleta realizada por intermédio do acesso ao site Periodicos.Capes. Ver: www.periodicos.capes.gov.br

realizadas na última década (2010-2019[3]). Dentre os diversos atributos disponíveis sobre a publicação, foram considerados: (i) o resumo, (ii) a lista de autores com a identificação dos respectivos países de origem e (iii) ano da publicação. A coleta dos dados resultou em 5.691 registros. Os registros considerados para este estudo foram aqueles que apresentavam valores válidos para os três atributos descritos anteriormente. Assim, registros que não apresentavam resumo ou autores ou ano de publicação foram desconsiderados, o total resultante foi de 5.193 (91,2%) registros completos.

A Tabela 1 sumariza as quantidades de registros completos coletados, estratificados por periódicos e biênios. Adicionalmente, os registros referentes às publicações com, ao menos, um(a) autor(a) brasileiro(a) foram destacados.

Tabela 1 - Descrição dos registros considerados, estratificados por biênio e periódicos

Período	JOI (867)		JASIST (2.085)		*Scientometrics* (2.739)		Total (5.691)	
	Mundo	Brasil	Mundo	Brasil	Mundo	Brasil	Mundo	Brasil
2010/11	125 (18,6%)	2 (1,6%)	364 (54,2%)	5 (1,4%)	183 (27,2%)	2 (1,1%)	672 (100,0%)	9 (1,3%)
2012/13	165 (19,2%)	3 (1,8%)	371 (43,1%)	4 (1,1%)	324 (37,7%)	12 (3,7%)	860 (100,0%)	19 (2,2%)
2014/15	164 (13,3%)	1 (0,6%)	375 (30,4%)	6 (1,6%)	694 (56,3%)	24 (3,5%)	1.233 (100,0%)	31 (2,5%)
2016/17	167 (12,7%)	6 (3,6%)	420 (31,8%)	12 (2,9%)	733 (55,5%)	37 (5,1%)	1.320 (100,0%)	55 (4,2%)
2018/19	147 (13,3%)	1 (0,7%)	259 (23,4%)	7 (2,7%)	702 (63,4%)	29 (4,1%)	1.108 (100,0%)	37 (3,3%)
Total	**768 (14,8%)**	**13 (1,7%)**	**1.789 (34,5%)**	**34 (1,9%)**	**2.636 (50,8%)**	**104 (4,0%)**	**5.193 (100,0%)**	**151 (2,9%)**

Fonte: elaborado pelos autores.

Os resumos das publicações, referenciados daqui em diante como documentos, foram utilizados para a identificação dos tópicos de interesse dos(as) seus(suas) respectivos(as) autores(as). O objetivo desse processo é obter classes que organizem as publicações por similaridade. Nesse sentido, as classes identificadas são definidas como tópicos (arranjo de palavras), os quais são ordenados em função de sua frequência relativa nos documentos.

Os documentos foram tratados de modo que os elementos textuais irrelevantes (*stop words*[4]) fossem retirados. No contexto da computação, as *stop words* são as palavras irrelevantes para o objetivo pretendido, as quais são removidas antes ou depois do processamento do documento. As palavras flexionadas resultantes foram agrupadas pelo

3 Dados coletados em dezembro de 2019 (os registros referentes a esse mês não foram considerados).

4 Adicionalmente consideramos os seguintes cinco termos como *stop-words* dada sua pouca relevância para análise de tópicos (termos genéricos): *article, paper, research, result*, e *study*.

seu respectivo lema, isto é, na forma de dicionário (*lemmatization*). A lematização é a técnica de se representar os verbos no infinitivo e os substantivos e adjetivos no masculino singular. Já a *Topic Modeling* é um tipo de modelagem estatística para a identificação de tópicos abstratos em um conjunto de documentos e tem crescido ao longo dos últimos anos o interesse por ela. (LEYDESDORFF; NERGHES, 2017) Assim, o método LDA é um tipo de modelagem que classifica os textos de um conjunto de documentos em tópicos específicos. (BLEI; NG; JORDAN, 2003)

O método LDA admite, como requisito, um número arbitrário de *k* tópicos, para os quais cada palavra nos documentos é associada. Em um processo iterativo, são calculadas as proporções de palavras associadas aos tópicos, assim como de tópicos associados aos documentos. Dessa forma, em cada iteração, o tópico de cada palavra é atualizado com o produto das proporções, até convergir (estabilizar). Ao final do processo, os documentos têm um *score* para cada tópico, que indica a probabilidade do documento ser relativo ao respectivo tópico, assim, o tópico com maior *score* é aquele que representa o documento. Similarmente, os tópicos tem um *score* para cada palavra que os compõem, indicando a sua proporção naquele tópico. Nesse sentido, cada tópico é um arranjo de todas as palavras que compõem o conjunto de documentos, o que diferencia um tópico de outro é a proporção das palavras, a qual ordena o arranjo.

Há duas decisões importantes a serem tomadas no contexto do método LDA: a escolha do número de tópicos que são considerados e a definição do número de palavras que representam esses tópicos. Nesse contexto, foram realizados experimentos de classificação, variando de um a dez tópicos, observou-se que o aumento do número de tópicos resulta em uma maior interseção entre eles, i.e., houve muitas palavras comuns a mais de um tópico. O intervalo de número de tópicos, considerado nos experimentos, teve por objetivo verificar a existência de uma correlação entre número de tópicos e interseção das palavras que compõem os respectivos tópicos. Assim, foi possível observar a correlação e definir um número de tópicos que resultasse na menor interseção possível, notando que tópicos que apresentam grande interseção são difíceis de serem caracterizados.

A partir da observação dos diferentes níveis de sobreposição, e por intermédio de uma checagem, verificando quantos agrupamentos seriam gerados pelo software VOSviewer (VAN ECK; WALTMAN, 2010), mesmo levando em conta que seu algoritmo tende ligeiramente a formação de mais agrupamentos (LEYDESDORFF; NERGHES, 2017), foi considerado que o número de quatro tópicos para o método LDA era o ideal para este estudo. Essa configuração apresenta interseção entre os tópicos, porém, o número de palavras comuns a dois ou mais tópicos é inferior ao número de palavras exclusivas em

cada tópico. Esse critério visou facilitar a caracterização dos tópicos. Quanto ao número de palavras em cada tópico, consideraram-se as dez palavras com maior *score*. Os experimentos, também, mostraram que o número de palavras e a interseção entre os tópicos são diretamente proporcionais.

RESULTADOS

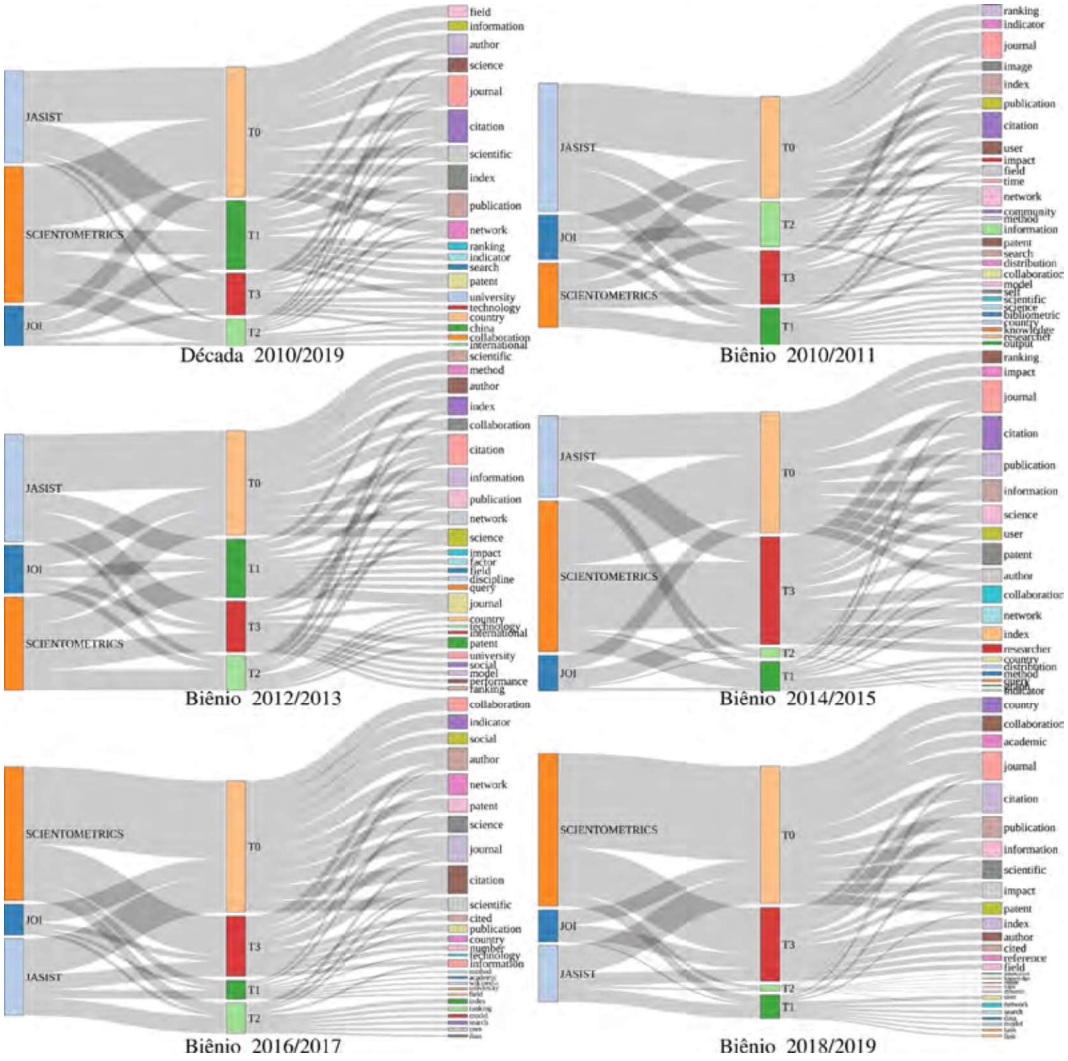
A aplicação do método LDA foi realizada em contextos diferentes. Primeiro todo o conjunto de documentos foi considerado, enquanto as demais aplicações foram feitas com os documentos particionados em biênios. Essa divisão de documentos permite: (i) obter uma visão geral sobre o interesse de pesquisa na área da CI e (ii) observar as mudanças nos interesses de pesquisa ao longo do tempo.

Os tópicos identificados, considerando todo o período, estão associados a 20 termos, sendo os principais: (i) *citation*, (ii) *journal*, (iii) *index*, (iv) *publication*, (v) *author*, (vi) *network*, (vii) *scientific*, (viii) *patent*, (ix) *science* e (x) *field*. Na Figura 1, os diagramas de associação descrevem os resultados das classificações, indicando as proporções de palavras nos tópicos e as relações entre periódicos e seus tópicos, quanto ao volume de publicações.

A caracterização dos tópicos não é determinística, pois eles representam arranjos sobre o mesmo conjunto de palavras, cuja diferenciação é dada pela ordem das mesmas. Assim, tentativas de análise são interpretativas, inclusive as descritas aqui.

O conjunto completo dos documentos abrange a última década (2010/19). Entre os tópicos identificados, o T0 aparece com maior volume de publicações (2.527). O periódico *Scientometrics* foi o que mais publicou artigos associados a esse tópico (1.181) e, também, foi líder em número de publicações (2.636) nesse período, seguido pelos periódicos JASIST (1.789) e JOI (768), respectivamente. A análise por biênios tem por objetivo verificar a evolução dos interesses dos pesquisadores na área. Assim, o biênio inicial (2010/11) revela uma predominância, em volume de publicações, do periódico JASIST (364). Além disso, o tópico T0 se destaca como o mais considerado pelos periódicos nesse período (288) e, em sua composição, observa-se que os termos *ranking* e *indicator* aparecem de forma exclusiva.

Figura 1 - Diagramas de associação entre periódico-tópico e tópico-palavra. As alturas dos blocos, em cada coluna, indicam as proporções de participação nos relacionamentos[5]

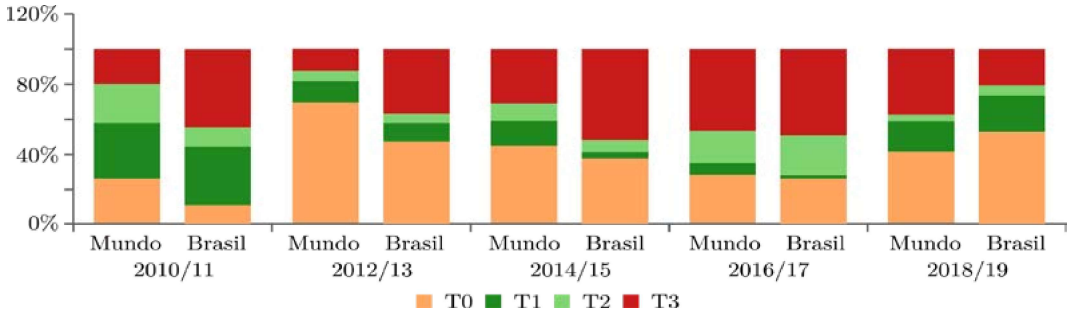


Fonte: elaborado pelos autores.

O biênio seguinte (2012/13) teve como característica principal o equilíbrio, quanto ao volume de publicações, entre os periódicos JASIST (371) e *Scientometrics* (324), sendo que as publicações deste último periódico foram realizadas de forma equilibrada entre os quatro tópicos. O tópico To se manteve como o mais considerado com 363 publicações, porém, os termos exclusivos associados a ele foram *scientific* e *method*.

5 Versões interativas dos diagramas estão disponíveis em: https://rossi-luciano.github.io/EBBC2020/

Figura 2 - Percentuais de publicações por tópico em cada biênio. Os valores referem-se ao conjunto completo (Mundo) e ao caso específico do Brasil



Fonte: elaborado pelos autores.

O periódico *Scientometrics* se consolidou como o líder, em número de publicações (694), no biênio 2014/15, o que se mantém até o momento. Os tópicos T0 e T3 reuniram, de maneira equilibrada, o maior volume de publicações neste período, com 558 e 496 publicações, respectivamente. Assim, em T0 observa-se os termos exclusivos *ranking* e *impact*, por outro lado, o termo *researcher* é observado somente em T3.

O biênio 2016/17 foi marcado pela ampliação da liderança do periódico *Scientometrics*, em relação aos demais, com 733 publicações, das quais 426 referiram-se a temas ligados ao tópico T0 que foi o preferido das publicações na área neste período, com 720 publicações. Assim, os termos exclusivos, associados ao tópico T0 são *collaboration*, *indicator* e *social*.

Finalmente, no biênio 2018/19, observa-se a manutenção do padrão iniciado em 2014/15, com a *Scientometrics* e o tópico T0 concentrando os maiores volumes de publicações, sendo ao todo 702 e 633 publicações, respectivamente. Os termos exclusivos em T0 são *collaboration* e *academic*, sendo esse tópico formado majoritariamente por artigos da *Scientometrics* (435).

Por outro lado, é importante destacar que o Brasil está representado, no escopo analisado, por um pequeno grupo de artigos que corresponde a 2,9% do total. Nesse sentido, a Figura 2 apresenta os comparativos entre os tópicos mais representativos por autores(as) no mundo e no Brasil. Nos três primeiros biênios os(as) autores(as) brasileiros(as) publicaram, majoritariamente, artigos associados ao tópico T3, contrastando com os(as) autores(as) no mundo. Para o biênio 2016/17 há, praticamente, uma equidade entre os dois grupos. No biênio 2018/19, o tópico de preferência dos(as) brasileiros(as) foi o T0.

CONCLUSÕES

O mapeamento da pesquisa, por meio da detecção de seus respectivos tópicos, pode resultar em novas descobertas sobre a evolução do conhecimento científico. Neste trabalho

buscou-se evidenciar a evolução da pesquisa em CI por meio da análise de publicações em seus principais periódicos. Os resultados mostram que os tópicos de atuação apresentam uma composição dinâmica, que se altera ao longo do tempo.

A abordagem, aqui considerada, caracteriza-se como um passo inicial de uma exploração mais aprofundada para a área da CI. Assim, como trabalho futuro, pretende-se considerar o estudo da similaridade entre os veículos de publicação a partir dos tópicos. Nesse contexto, uma maior especificidade na composição dos tópicos pode ser útil para investigar o surgimento, a divisão, a união e a transformação da pesquisa em CI.

REFERÊNCIAS

ALETA, A.; MELONI, S.; PERRA, N. *et al.* Explore with caution: mapping the evolution of scientific interest in physics. *EPJ Data Science*, Heidelberg, v. 8, n. 1, p. 1-15, 2019.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, Cambridge, v. 3, p. 993-1022, 2003.

LAMBA, M.; MADHUSUDHAN, M. Mapping of topics in DESIDOC journal of library and information technology, India: a study. *Scientometrics*, Dordrecht, v. 120, n. 2, p. 477-505, 2019.

LEYDESDORFF, L.; NERGHES, A. Co-word maps and topic modeling: a comparison using small and medium-sized corpora (N<1,000). *Journal of the Association for Information Science and Technology*, Hoboken, v. 68, n. 4, p. 1024-1035, 2017.

VAN ECK, N. J.; WALTMAN, L. Software survey: vosviewer, a computer program for bibliometric mapping. *Scientometrics*, Dordrecht, v. 84, n. 2, p. 523-538, 2010.