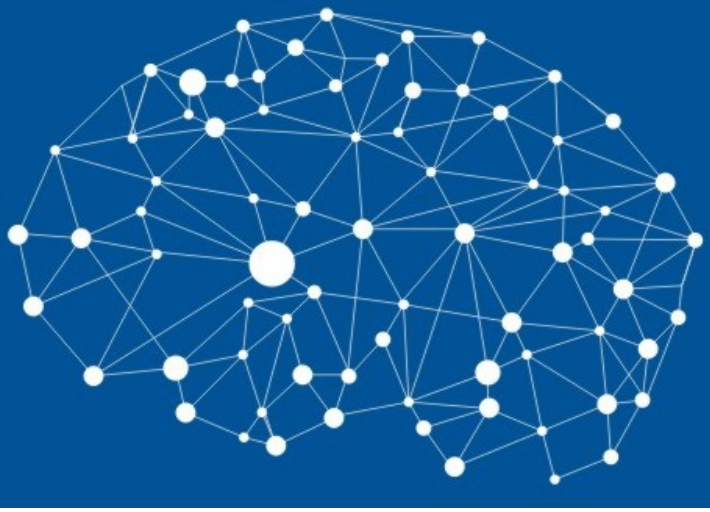


Técnicas para desambiguação de citações: o uso de Regressão a partir do ano e volume

Luciano Antonio Digiampietri[1]; Rogerio Mugnaini[2]

DIGIAMPIETRI, L. A.; MUGNAINI, R.. Técnicas para desambiguação de citações: o uso de Regressão a partir do ano e volume In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais**... São Paulo: USP, 2016. p. A101

[1]PPGSI, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo; [2]PPGCI, Escola de Comunicações e Artes, Universidade de São Paulo

Técnicas para desambiguação de citações:
o uso de Regressão a partir do ano e volume

Eixo temático: Métodos, Técnicas e Ferramentas para Estudos Bibliométricos e Cientométricos

Modalidade: Apresentação oral

1 INTRODUÇÃO

A análise de citações ganhou um papel central nos processos avaliativos em diversos níveis do amplo sistema de ciência e tecnologia. Observa-se o uso deste tipo de análise em variados níveis de agregação, fenômeno este que se orininou na análise de literatura, consolidando-se pricipalmente na avaliação de revistas científicas, até chegar às avaliações em nível de um país (VESSURI; GUÉDON; CETTO, 2013). No nível macro pode-se mencionar a avaliação nacional de programas de pós-graduação, como se observa por exemplo no Brasil, porém que se baseia na classificação de revistas (usado principalmente o Fator de Impacto do Journal Citation Reports, e crescentemente, o índice h) para composição de indicadores de programas de pós-graduação (MUGNAINI, 2015). Na execução do *Excellence in Research for Australia* (ERA) a classificação de revistas é realizada de maneira muito similar ao que se realiza no Qualis brasileiro, com a diferença de que desde a avaliação de 2009 decidiu-se que as áreas de artes e humanidades não utilizariam indicadores de impacto, fazendo com que estudos fossem realizados questionando a utlização de análise de citação para avaliação nas ciências sociais (HADDOW; GENONI, 2010).

Por outro lado existe entre os dois países uma preocupação comum com a melhor avaliação das revistas nacionais não indexadas na Web of Science, e para se poder realizar análise de citação, um esforço de padronização de informações das referências bibliográficas é essencial. O desafio está em associar a referência bibliográfica de um artigo (citante) ao documento fonte (citado). Para isso informações da referência, como sobrenome do(s) autor(es), título da revista, volume e ano, precisarão haver sido escritos corretamente pelo autor citante, para que coincida com as informações do registro bibliográfico do artigo citado. Um estudo realizado sobre as citações recebidas na Web of Science por revistas croatas revelou que a informação com mais incidência de erro (37%) foi o título da revista (ANDREIS; JOKIC,

2008). Este tipo de problema afeta diretamente o cálculo do Fator de Impacto do Journal Citation Reports, prejudicando a avaliação da revista, como ocorreu com duas revistas, que tinham suas citações confundidas – *Educational Research* e *Educational Researcher* –, cujas citações foram computadas em favor de uma delas (*Educational Researcher*) por quase 20 anos.

No Brasil um caso típico ocorre com a abreviação "J PEDIAT", que é utilizada em citações ao Jornal de Pediatria (brasileiro) e ao Journal of Pediatrics (norte-americano), confundindo-os. Este tipo de problema pode ser resolvido com certa simplicidade, caso duas revistas que são citadas de maneira homônima tenham iniciado em anos diferentes. Por exemplo, no ano 2000 a revista brasileira publicou o volume 76, ao passo que a norte-americana publicou os volumes 136 e 137, por essa razão a ambiguidade pode ser desfeita ao se considerar ano e volume. Ferramentas computacionais podem ser utilizadas para distinção de casos como este, denominando esta operação como “resolução de entidades”.

Para avaliação deste tipo de solução, foi proposta uma metodologia que compara três estratégias diferentes para, a partir da consideração do ano e volume de todos os registros da base Web of Science, determinar a revista correta, utilizando para tanto a técnica estatística de regressão linear.

2 METODOLOGIA

Os dados foram cedidos pelo *Centre for Science and Technology Studies* (CWTS), que mantém contrato com a Thomson Reuters, dispondo assim de dados completos (em meados do ano de 2015), e contabilizando cerca de 890 milhões de registros, correspondentes às referências bibliográficas dos artigos indexados na base. Pelo fato dos registros de citação estarem no nível de artigos, um processamento de tabulação foi realizado a fim de agregá-los ao nível de volume da revista (ou seja, desconsiderando nome de autor, título do artigo, número do fascículo, páginas, entre outras informações), resultando em cerca de 69 milhões de registros. Então dois conjuntos de dados foram obtidos da Web of Science: o primeiro é composto pelos registros de título, ISSN, ano e volume de cada uma das revistas indexadas pela Web of Science que somavam à época 15.532 revistas diferentes, e totalizando 407.874 registros diferentes de ano e volume; já o segundo conjunto é composto por 7.049.289 registros de referências, isto é, referências a artigos citados por publicações cadastradas na

própria base. Destes registros, foram utilizados 5.283.496 correspondendo àqueles cujo ISSN, ano e volume já foram resolvidos/identificados pelo CWTS quando da identificação do par de artigos (citante e citado) permitindo uma análise de desempenho da estratégia proposta (a porção de referências descartadas são as citações a artigos de outras revistas, não indexadas na base, e, portanto, não estando livre de ambiguidade).

Os dados obtidos foram organizados em arquivos texto no formato CSV (*Comma-Separated Values*). Um arquivo contém os dados referentes aos registros das revistas, cada linha contendo ISSN, título, ano e volume e este arquivo foi ordenado por ISSN, ano e volume. O segundo arquivo contém dados das referências presentes em cada publicação, além de conter o título, ano e volume citados este arquivo também contém o ISSN, ano e volume considerados corretos por um processo de resolução semiautomático previamente executado.

CÁLCULO DO VOLUME ESTIMADO

Com base no provável ISSN de uma revista e o ano de publicação encontrado nos dados de citação, é estimado um volume para essa citação. Três estratégias foram desenvolvidas para a realização deste cálculo. A primeira considera todas as informações dos registros de ano e volume de uma revista e gera uma regressão linear para o cálculo do volume com base no ano de publicação. Por exemplo, se há 20 informações diferentes de ano e volume para uma mesma revista, todas estas informações serão utilizadas para o cálculo da regressão linear. A segunda estratégia utiliza apenas dois valores de ano e volume para gerar a regressão. Neste caso, dado um ano e uma revista a estratégia procura nos registros de ano e volume da respectiva revista os dois anos mais próximos ao ano passado como entrada e realiza a estimativa do volume com base nesses valores. Por fim, a última estratégia é semelhante à segunda, porém, no processo de busca por ano e volume mais próximos ao ano passado como entrada (isto é, o ano encontrado na referência bibliográfica) a estratégia verifica se o ano da citação atual já consta no cadastro de anos e volumes para a respectiva revista. Se sim, esta estratégia retorna o volume correspondente, caso contrário o volume é estimado da mesma forma que na segunda estratégia.

COMPARAÇÃO COM O VOLUME REAL

O volume presente nas citações é comparado com o volume estimado pelas três diferentes estratégias e são verificados os acertos, bem como a diferença entre a estimativa e o valor registrado, conforme será apresentado na próxima seção.

3 RESULTADOS

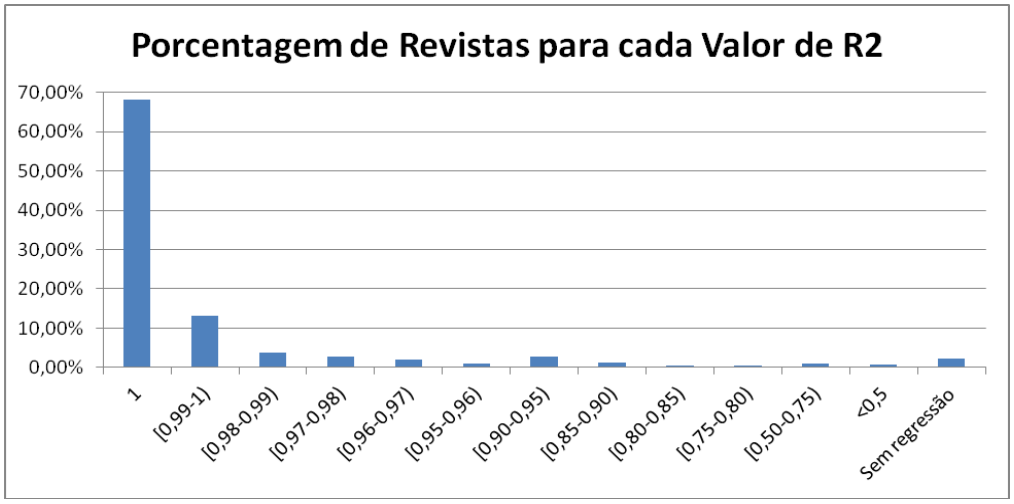
Nesta seção são detalhados os resultados do uso das três estratégias utilizadas para a estimativa do valor do volume com base em um cadastro de ISSN, ano e volume.

A estratégia que utiliza todos os dados cadastrados de ano e volume para calcular uma equação que corresponde à regressão linear desses dados (aqui chamada de regressão geral) produz, para cada revista que possua ao menos dois registros de ano e volume, uma função no formato: *volume = a + b*ano*, sendo *a* e *b* constantes e ano a variável. Dada uma citação para a respectiva revista e o ano encontrado na citação é possível calcular o volume. Destaca-se que se uma revista publicar mais de um volume por ano é possível que o valor estimado esteja errado em uma ou mais unidades. Conforme será apresentado, para a maioria das revistas este problema não ocorre, porém, sistemas que utilizem a estimativa do volume para auxiliar no processo de resolução de revistas devem levar isto em consideração.

Das 15.532 revistas diferentes que possuíam dados de ano e volume, 15.190 (97,8%) possuíam mais de um registro de ano e volume e, consequentemente, foi possível gerar uma função de regressão linear para cada uma delas. Para 10.148 destas revistas (65%) a função apresenta o seguinte formato: *volume = -AnoInicial + 1*ano*, sendo *AnoInicial* o ano em que a revista começou a ser publicada.

A Figura 1 apresenta a distribuição da porcentagem de revistas em relação à medida R2 da função de regressão linear calculada sobre os dados de ano e volume. Observa-se que a função de regressão possui um casamento perfeito (isto é, R2=1) para todos os anos e volumes de mais de 68% das revistas cadastradas. Adicionalmente, 91% das funções de regressão das revistas possuem R2 maior ou igual a 0,95.

Figura 1 – Porcentagem das revistas para cada valor de R2



Fonte: os autores

Uma diferença fundamental entre a primeira estratégia e as demais é que ela necessita apenas das funções de regressão linear (uma para cada revista) para realizar a estimativa do volume de um registro de citação. Já as demais estratégias precisam de todas as informações de ano e volume de cada revista para calcular a regressão considerando apenas os dois pontos mais próximos do ano citado. Apesar do custo computacional e de armazenamento adicionais de se utilizar todos os anos e volumes, uma vantagem é que é possível verificar o casamento exato entre o ano e volume citado e os dados cadastrados. Esta característica é utilizada pela terceira estratégia.

Conforme apresentado, foram utilizados dados de 5.283.496 de registros de citações para se verificar a estratégia proposta. Deste total, apenas 70,15% possuíam o volume cadastrado na base de anos e volumes para a respectiva revista. A Tabela 1 apresenta a distribuição dos erros da estimativa de volume com base nos dados da revista e ano presentes nos dados de citação. Observa-se que a regressão utilizando todos os dados de ano e volume para cada revista (regressão geral) foi capaz de estimar corretamente (sem nenhum erro) o volume de apenas 34,45% dos registros de citação. Por outro lado, as estimativas que tiveram um erro máximo de uma unidade totalizam 60,77% dos dados registros de citação para esta estratégia. A estratégia que calcula a regressão baseada apenas nos dados dos dois anos mais próximos ao ano citado (Regressão dois pontos) obteve resultados superiores à regressão geral, estimando volumes sem erro para quase metade dos registros de citações (49,41%) e com estimativas com erro menor ou igual a 3 para mais de 80% dos registros.

Tabela 1 – Distribuição dos erros de estimativa de volume para cada proposta

Erro	Regressão geral	Regressão dois pontos	Regressão dois pontos e casamento exato
0	34,45%	49,41%	70,18%
Até 1	60,77%	71,98%	78,21%
Até 2	67,43%	77,15%	80,76%
Até 3	71,49%	80,04%	82,48%
Até 4	74,13%	81,58%	83,41%
Até 5	76,21%	82,80%	84,23%
Até 6	77,86%	83,76%	84,89%
Até 7	79,12%	84,41%	85,36%
Até 8	80,21%	84,96%	85,77%
Até 9	81,35%	85,70%	86,40%
Até 10	82,94%	87,09%	87,73%
Maior que 10	100,00%	100,00%	100,00%

Fonte: os autores

Dentre as três estratégias utilizadas, a que obteve os melhores desempenhos foi a que combina a regressão utilizando dois pontos com o casamento exato dos dados da citação em relação ao cadastro de anos e volumes. Esta estratégia identificou sem erros o volume de mais de 70% dos registros de citações e com erro de no máximo uma unidade 78,21% dos registros. Observa-se pelos dados da Tabela 1 que para todos os intervalos de erro a combinação da regressão utilizando dois pontos e o casamento exato foi superior às demais estratégias. Por outro lado, a regressão utilizando todos os dados disponíveis da respectiva revista apresentou os piores resultados.

4 CONSIDERAÇÕES FINAIS

Este artigo apresentou o uso de três técnicas baseadas na regressão linear para auxiliar no processo de resolução de entidades em dados de citações a revistas.

As estratégias foram comparadas utilizando-se dados de citações de mais de cinco milhões de registros de citações da Web of Science. Observou-se que a combinação entre o casamento exato entre os dados de citações e os registros de ano e volume das revistas, combinado com o uso da regressão linear utilizando-se apenas os dois registros de ano e volume mais próximos ao ano citado, foi a estratégia que obteve melhores resultados, tanto

considerando-se apenas o acerto exato do ano (acurácia acima de 70%) como também ao se permitir uma margem de erro de uma ou mais unidades.

Desta forma, o uso de tal estratégia apresenta elevado potencial para ser utilizada no processo de resolução de dados de citações, em especial nos casos em que informações tradicionalmente utilizadas, como o título das revistas, não são suficientes para se realizar a resolução de entidades (por exemplo, quando se identifica ambiguidade entre o título de duas ou mais revistas).

AGRADECIMENTOS

O trabalho apresentado neste artigo foi parcialmente financiado pela FAPESP (projetos nº 2012/00255-6 e 2015/07891-3, CAPES e CNPq (projeto nº 2036046/2013-0). Agradecemos a Rodrigo Costas, pesquisador do CWTS que extraiu os dados da base, e cujos comentários permitiram o aprimoramento deste estudo, assim como ao CWTS, pela concessão dos dados.

REFERÊNCIAS

ANDREIS, M.; JOKIC, M. An impact of Croatian journals measured by citation analysis from SCI-expanded database in time span 1975–2001. **Scientometrics**, v. 75, n. 2, p. 263-288, 2008.

HADDOW, G.; GENONI, P. Citation analysis and peer ranking of Australian social science journals. **Scientometrics**, v. 85, n. 2, p. 471-487, 2010.

LANGE, L. L. The impact factor as a phantom: Is there a self-fulfilling prophecy effect of impact? **Journal of Documentation**, v. 58, n. 2, p. 175-184, 2002.

MUGNAINI, R. Ciclo avaliativo de periódicos no Brasil: caminho virtuoso ou colcha de retalhos? In: ENANCIB - Encontro Nacional de Pesquisa em Ciência da Informação, 16., 2015, João Pessoa. **Anais...** João Pessoa: UFPB, 2011. v. 16.

VESSURI, H.; GUÉDON, J. C.; CETTO, A. M. Excellence or quality? Impact of the current competition regime on science and scientific publishing in Latin America and its implications for development. **Current Sociology**, p. 0011392113512839, 2013.