MOVIMENTO
ISSN: 1982-8918

ENSAIOS

POTENCIALIDADES E LIMITES DO PROCESSAMENTO DE DADOS EM PESQUISAS SOBRE A PRODUÇÃO CIENTÍFICA

POTENTIAL AND LIMITS OF DATA PROCESSING IN RESEARCH ON SCIENTIFIC PRODUCTION

POTENCIAL Y LÍMITES DEL PROCESAMIENTO DE DATOS EN INVESTIGACIONES SOBRE LA PRODUCCIÓN CIENTÍFICA

https://doi.org/10.22456/1982-8918.120556

Leoncio José de Almeida Reis* <leojar_edf@yahoo.com.br>

* Universidade Federal do Paraná. Matinhos, PR, Brasil.

Resumo: Este trabalho versa sobre potencialidades e limites relacionados à utilização de processamentos de dados para auxiliar na produção e sistematização de conhecimento científico. Objetiva, através de um exercício experimental envolvendo a utilização de algoritmo, discutir a viabilidade do uso de técnicas de coleta automatizada para levantamento e produção de dados utilizáveis no âmbito das pesquisas científicas. Como demonstração, busca reproduzir de maneira automatizada processos relacionados à coleta de dados de pesquisa anteriormente publicada neste periódico, descrevendo metodologicamente como foram organizados e desenvolvidos a extração e o tratamento desses dados. Como resultado, constata que o processamento automatizado pode ser uma alternativa produtiva e eficiente para auxiliar nas sistematizações e análises sobre o acumulado crescente de publicações no campo científico, podendo abrir novos caminhos metodológicos de pesquisa na Educação Física, especialmente considerando o volume de dados passível de coleta e análise em redes sociais, fóruns e outras plataformas na web.

Palavras-chave: Processamento Eletrônico de Dados. Armazenamento e Recuperação da Informação. Bibliometria.

Recebido em: 11 jan. 2022
Aprovado em: 19 abr. 2022
Publicado em: 25 jul. 2022

1 INTRODUÇÃO

Este trabalho versa sobre potencialidades e limites no que se refere à utilização de processamento computacional para auxiliar na produção e sistematização de conhecimento científico. Objetiva, através de um exercício experimental envolvendo a utilização de algoritmo especificamente construído para esse fim, apresentar a funcionalidade e discutir a viabilidade do uso de técnicas de coleta automatizada para levantamento e produção de dados utilizáveis no âmbito das pesquisas científicas. Para tanto, busca reproduzir por meio de extração automatizada a coleta de dados manualmente realizada em pesquisa publicada na revista *Movimento* (pesquisa produzida por Dias *et al.,* 2017, conforme explicado adiante).

Em termos mais gerais, está organizado no sentido de a) problematizar a quantidade de dados e informações acumulados na internet, bem como seu crescimento vertiginoso; b) discutir a inviabilidade ou mesmo a impossibilidade de tratar do crescente volume de dados e informações sem a utilização de processamento computacional - o que é válido também no âmbito da produção científica; c) apresentar e demonstrar a utilização de técnicas para coleta automatizada de dados e suas potencialidades no que se refere especialmente à sistematização da produção científica em periódicos; d) ponderar sobre as dificuldades e limites de sua utilização.

2 DILÚVIO INFORMACIONAL

Com a invenção e o desenvolvimento das tecnologias computacionais, textos, imagens e sons, antes restritos à materialidade concreta dos meios físicos que os continham, tornaram-se passíveis de serem armazenados digitalmente (isto é, armazenados através de estados binários, representado com os dígitos 0 e 1), podendo ser muito mais facilmente produzidos e reproduzidos; organizados e manipulados; armazenados e colocados em circulação.

No atual estágio de desenvolvimento computacional e de integração global propiciado pela rede mundial de computadores, a quantidade de dados e informações produzidos chega a patamares nunca antes vistos. Não é com espanto que se ouve que a humanidade produziu mais nas últimas décadas em termos de volume de dados e de informação do que em toda sua história (PIMENTA, 2017).

Um olhar de relance sobre algumas métricas da produção de dados na internet revela a dimensão desse fluxo informacional. O Quadro 1 apresenta a quantidade aproximada de informação produzida em apenas 1 segundo:

Quadro 1: Circulação estimada de dados

Plataforma	Quantidade (em 1 segundo)	Tipo do dado
Twitter	9.755	tweets enviados
Instagram	1.118	fotos postadas
Google	97.358	buscas realizadas
YouTube	92.766	vídeos visualizados
E-mails	3.090.044	e-mails enviados (67% deles SPAM)

Fonte: Internet Live Stats (2021)

Contemporânea ao atual cenário informacional é a expressão *Big data,* que é usada para fazer referência ao imenso a) volume, b) velocidade e c) variedade de dados produzidos, e cujo armazenamento, tratamento e manipulação é tido como valioso e imprescindível para usos sociais, políticos e mercadológicos.

> Em uma sociedade da informação como a atual, marcadamente exponencial no tocante à produção de registros, dados e informação, lidamos hodiernamente com perfis e contas virtuais concernentes as nossas atividades sociais, políticas, culturais, sexuais e econômicas. Do *facebook* [sic] ao *ResearchGate*, passando pelo *Tinder*, pela *Amazon* ou até mesmo o *Avaaz*, todas plataformas com fins muito diferentes, mas que produzem ambos dados brutos (*raw data*) sobre nós e sobre nossas práticas cotidianas. […] A cada clique, a cada toque, compartilhamos informações. Por vezes estas mesmas compõem, estruturam, depoimentos, entrevistas, imagens, vídeos e documentos digitalizados. Em outros contextos apenas alimentam algoritmos com o intuito de produzir mais meta-dados [sic] e informação direcionada ora ao mercado, ora ao Estado. (PIMENTA, 2017, p. 14)

Escândalos recentes, como no caso do pleito presidencial estadunidense de 2016, que elegeu Donald Trump, marcado pelo vazamento de dados pessoais de usuários da plataforma Facebook e pelo suposto e indevido uso político desses dados para manipulação do eleitorado através do direcionamento de anúncios emocionalmente apelativos; ou ainda na eleição presidencial brasileira de 2018, que elegeu Jair Bolsonaro, marcada pela circulação exaustiva das chamadas *fake news* (DOURADO, 2020), se por um lado mostram a importância e a potencialidade dos usos dos dados, por outro chamam a atenção para os potenciais riscos do acesso irrestrito e indiscriminado à informação e à comunicação.

No contexto do *Big data*, o volume, a velocidade e a variedade de dados e informações são tão vertiginosos que o tratamento não poderia ser feito de outra forma senão automatizada, com o uso do poder de processamento dos computadores. Tanto pela quantidade, quanto pela sua intangibilidade, esses dados e informações jamais poderiam, portanto, serem efetivamente “manipulados” (no sentido original do termo, derivado do latim *manus*, mão).

No âmbito da ciência, vivemos num cenário no qual se avolumam publicações de pesquisas científica nos mais diversos formatos textuais: artigos, dissertações, teses e livros. Pesquisa bibliométrica que avaliou a produção científica brasileira de 1990 até 2017 identificou quase seis milhões de artigos[1] registrados no currículo de pesquisadores cadastrados na plataforma Lattes (NASCIMENTO *et al*., 2021), com a produção científica brasileira anual saltando de aproximadamente 40 mil artigos em 1990 para mais de 391 mil em 2016. Ao fim, os autores chegaram à conclusão de que “no período analisado, a produção científica brasileira cresceu com taxas superiores à média mundial, e as projeções da sua curva exponencial de crescimento não indicaram saturação ou a proximidade de um teto” (NASCIMENTO *et al*., p. 69, 2021.

O crescimento substantivo da produção científica nas últimas décadas, se por um lado origina avanços e desenvolvimento da ciência, por outro gera um acúmulo sem fim de informação e conhecimentos eventualmente fadados à invisibilidade ou

1 Precisamente, a pesquisa identificou 5.886.968 artigos.

mesmo ao esquecimento. Problematiza-se, nessa direção, a existência de artigos que nunca serão citados, e outros, sequer lidos (CASTIEL; SANZ-VALERO, 2007).

Mesmo com uma eventual estagnação ou desaceleração da produção científica nacional (NASCIMENTO *et al*., 2021), é possível pensar que, num cenário bastante próximo, trabalhos quantitativos de sistematização e levantamento de informações em publicações científicas somente poderão ser realizados de maneira produtiva, eficiente e abrangente com o auxílio de processamento computacional. Do contrário, a quantidade de informação à disposição exigirá ou a imposição de recortes muito específicos ou a mobilização de um corpo grande de colaboradores.

A partir de um exercício experimental, buscou-se verificar como a utilização de processamento computacional pode auxiliar no levantamento e sistematização de informações em periódicos e artigos científicos.

3 EXTRAÇÃO AUTOMATIZADA DE DADOS

Considerando pesquisas que tratam especificamente da sistematização da produção científica, há procedimentos e métodos de coleta de dados e informações que, muito embora tenham sido realizados manualmente, poderiam ter sido automatizados?

Para responder a essa questão, selecionou-se dentro da produção recente veiculada na revista *Movimento* pesquisa que abordasse especificamente a sistematização da produção científica e que tivesse envolvido, na sua execução, levantamento de volume significativo de dados.

Assim, foi selecionada e tomada como material de referência a pesquisa "Estudos do lazer no Brasil em princípios do século XXI: panorama e perspectivas", de Dias *et al*. (2017). Os autores se propuseram a realizar uma análise bibliométrica dos artigos publicados em periódico da revista *Licere* de 2000 a 2010. Os artigos originais e de revisão selecionados para a análise totalizaram 198 trabalhos, de 268 autores e coautores, distribuídos em 23 fascículos. Pela abrangência, profundidade e qualidade da investigação realizada, considerou-se que o referido trabalho ofereceria razoável parâmetro para o exercício aqui proposto.

Metodologicamente, o exercício planeado envolve: a) inventariar quais tipos de dados foram coletados na referida pesquisa; b) identificar quais foram os "passos" ou "caminhos" prováveis realizados para obtenção desses dados; c) verificar quais desses são passíveis de automatização; d) implementar a automatização parcial ou total desses passos.

Quanto ao inventário, foi possível identificar que foram levantados, quantificados ou classificados: a) número de autores por artigo; b) palavras-chave de cada artigo; c) referências bibliográficas listadas em cada artigo; d) classificação do tipo de referência bibliográficas (se livro, tese, dissertação, artigo publicado em periódico ou congresso); e) identificação de artigos listados nas referências que foram publicados na própria revista examinada; f) autores e obras mais citados nas referências; g) área de formação (na graduação, no mestrado e no doutorado)

dos autores que publicaram quatro ou mais artigos na revista e h) suas referências bibliográficas mais utilizadas.

Descreve-se no Quadro 2, de forma genérica, o algoritmo que qualquer pesquisador ou indivíduo precisaria executar para obtenção dos dados levantados, tratados e discutidos naquela pesquisa. Algoritmo nada mais é do que uma sequência lógica, finita e organizada de instruções que descrevem como um problema deve ser resolvido ou como uma tarefa deve ser executada.

Quadro 2 – Tarefas relacionadas à obtenção dos dados e seu posterior processamento

1. Acessar o site do periódico;
2. Acessar cada fascículo e, dentro de cada fascículo:
   2.1. Selecionar as publicações do tipo: “Artigos Originais” e “Artigos de Revisão”
   2.2. Acessar cada artigo selecionado;
      2.2.1. Localizar o nome de cada autor do artigo e guardá-los em uma tabela;
      2.2.2. Contar o número de autores do artigo;
      2.2.3. Localizar cada palavra-chave do artigo e guardá-las em uma tabela;
      2.2.4. Localizar cada referência bibliográfica e guardá-las em uma tabela;
3. Para cada referência bibliográfica guardada, classificá-la como livro, artigo, tese, dissertação, anais de congresso ou outro;
4. Para cada referência classificada como artigo, registrar se foi publicado na própria revista ou não.
5. Contabilizar e elencar os autores mais citados nas referências bibliográficas;
6. Contabilizar e elencar as obras mais citadas nas referências bibliográficas;
7. Contabilizar e elencar as palavras-chave mais citadas;
8. Contabilizar e listar os autores com quatro os mais artigos publicados;
9. Acessar o site da Plataforma Lattes e, para cada um dos autores com quatro ou mais artigos publicados:
   9.1. Buscar o currículo de cada autor;
   9.2. Localizar e guardar a área de formação na graduação, no mestrado e no doutorado;

Fonte: dados da pesquisa.

Em termos de complexidade, a realização manual dessas tarefas é bastante simples e as habilidades requeridas não vão muito além de certa familiaridade com artigos científicos e do domínio de planilhas eletrônicas.

O problema maior, de fato, reside na inevitável repetição. Um dado ilustrativo: as tarefas n. 2.2.4 e n. 3 tiveram que ser realizadas pelos autores da referida pesquisa não menos do que 4.038 vezes (pois esse foi o total catalogado de referências bibliográficas extraídas dos 198 artigos investigados). Lidar manualmente com esse volume de informações, além de cansativo e enfadonho, pode mais facilmente resultar em erros.

Pela natureza desses dados (todos armazenados em formato digital), pelo fato de sua obtenção poder ser descrita através de uma sequência lógica de instruções, e pela notável repetitividade requerida por esse processo, é certamente mais recomendado e produtivo (mas nem sempre mais viável) que levantamentos desse tipo sejam realizados de forma automatizada, a partir de algoritmos computacionais.

Em análise preliminar, verificou-se que as tarefas acima listadas poderiam ser realizadas completamente ou parcialmente com o auxílio de processamento computacional.

Contudo, embora tenha sido verificada a possibilidade de automatizar todos os processos (tarefas) descritos, tal empreitada demandaria esforço e investimento aquém dos propósitos deste artigo, que não é o de reproduzir exatamente a pesquisa utilizada como referência, mas demonstrar a exequibilidade e discutir potencialidades e limites da automatização, o que inclui, também, explicitar complexidades envolvidas em determinadas partes do processo.

Algumas tarefas são mais complexas de automatizar, especialmente aquelas que envolvem classificação baseada em reconhecimento de padrões, como as tarefas n. 3, 4 e 5. No caso da tarefa n.3, por exemplo, de separar as referências em livro, capítulo, tese, dissertação, artigo etc., embora seja uma tarefa bastante óbvia para qualquer iniciado no universo acadêmico, “ensinar” o computador a realizar tal tarefa requeria certo investimento.

Por exemplo, a mera identificação e separação, em uma dada referência, do título da obra e de sua autoria podem ser mais ou menos complexa dependendo de quão rigorosa e efetiva é a padronização das normas de uma dada revista. Isso porque no processo de elaboração e desenvolvimento do algoritmo é necessário que sejam previstas e tratadas todas ou parte considerável das situações possíveis.

A facilidade e a rapidez com que um leitor familiarizado com os protocolos acadêmicos faz essa identificação e separação esconde um complexo processo cognitivo que envolve o acionamento de componentes na memória associados à escrita e ao reconhecimento de certos padrões (das normas, de palavras relativas a nomes próprios, de pontuações etc.).

Assim, uma vírgula a mais ou a menos, embora não gere dificuldade alguma para o leitor familiarizado, que rápida e facilmente interpretaria a nova informação, poderia, no caso do processo automatizado, produzir dados e informações completamente errôneos.

Também pode haver dificuldades no processo de extração dos dados em virtude de dificuldades erguidas pelos próprios detentores da informação. É o caso das tarefas n. 9, 9.1 e 9.2, que demandam a busca do currículo do autor na plataforma Lattes. Por questões de segurança e sigilo de informações, o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) impõe barreiras para impedir a tentativa de obtenção de informações e dados por meio de extração automatizada (utilizando, por exemplo, serviços para certificação de que a consulta a um determinado currículo está sendo realizada por um humano e não um algoritmo computacional: como quando clicamos em caixas dizendo “eu não sou um robô”). No caso do uso acadêmico e científico desse tipo de informação, é possível obter todos os dados públicos da plataforma Lattes diretamente do CNPq, devendo, no entanto, ser solicitados formalmente à instituição, após viabilização de Protocolo de Cooperação Técnica entre universidade-CNPq.[2]

Adiante será apresentado quadro síntese (Quadro 4) indicando quais tarefas foram de fato automatizadas. A seguir, serão abordadas as tarefas relacionadas a extração dos dados, isto é, a sua obtenção e armazenamento.

2 Conforme informações do próprio *site*: http://lattes.cnpq.br/web/plataforma-lattes/extracao-de-dados/

4 *WEB SCRAPING*

Na ciência da computação, é chamada de *web scraping* a técnica de utilização de algoritmo para extrair dados disponíveis na *web* (HADDAWAY, 2015), armazenando-os em informação estruturada para posterior análise. É uma das etapas relacionados ao processo de mineração de dados.

No Quadro 3 são listados os dados a serem extraídos automaticamente, acrescidos de informações relativas à fonte, isto é, ao local digital de onde os dados serão obtidos.

Quadro 3 – Dados para extração

	Dados	Fonte	Formato
A	Número do fascículo	Página da Revista	Página HTML
B	Título do artigo	Página do Artigo	Página HTML
C	Subtítulo do artigo	Página do Artigo	Página HTML
D	Nomes dos autores	Página do Artigo	Página HTML
E	Instituições dos autores	Página do Artigo	Página HTML
F	Palavras-chaves	Página do Artigo	Página HTML
G	Resumo	Página do Artigo	Página HTML
H	Referências	Página do Artigo	Página HTML
I	Tipo do artigo	Página do fascículo	Página HTML

Fonte: dados da pesquisa.

Visando extrair e armazenar os dados desejados, desenvolveu-se uma aplicação utilizando a linguagem de programação R (R CORE TEAM, 2021). Resumidamente, tal aplicação consiste num algoritmo especificamente construído para obter dados da revista examinada e não de outra. O algoritmo indica precisamente o que o programa deve fazer e quais dados buscar (ex.: acessar o *site* da revista; “abrir” cada um dos fascículos; “clicar” em cada um dos artigos de cada fascículo; identificar e extrair os dados desejados; armazená-los em uma tabela; etc.).

Nesses processos, um dos principais desafios é traduzir para o computador exatamente o que se quer extrair, visto que o conteúdo textual da página exibida no navegador nada mais é, para o computador, do que um emaranhado de letras e palavras, não sendo distinguível, por exemplo, qual dessas letras ou palavras representa um nome de autor, uma palavra-chave ou qualquer outra informação. Também por isso é mais adequado utilizar o termo “extrair” do que “coletar”, visto que os dados não estão simplesmente à espera da coleta e o algoritmo precisa conter instruções suficientemente precisas de como encontrá-los.

Com o algoritmo desenvolvido[3], foi possível realizar sua execução e aguardar a extração automatizada, a qual foi realizada com sucesso.

O tempo total cronometrado para a extração das informações do total de 1064 publicações veiculadas entre 1998 e 2021 na revista analisada foi de aproximadamente 1h42min, com uma média aproximada de cinco segundos por artigo.

3 O desenvolvimento deste tipo de algoritmo requer conhecimentos associados ao campo das ciências da computação, como lógica de programação e domínio de linguagens de programação que disponibilizam ferramentas destinadas a esse fim (como Python ou R), além de conhecimentos próprios do funcionamento de páginas HTML.

A demora relativa na execução desse algoritmo está relacionada menos a sua complexidade de processamento (bastante simples do ponto de vista computacional) e mais ao tempo necessário para o tráfego de informações na *web* (mais especificamente, ao tempo necessário para aguardar que uma página da internet seja carregada), variável que depende tanto da velocidade de conexão de quem requisita as informações quanto do servidor que as fornece.

A título de comparação, levou-se aproximadamente 3min58 segundos para a coleta manual de um único artigo (copiando cada um dos dados listados no Quadro 3 para uma célula em uma planilha eletrônica, incluindo cada referência listada). Nesse ritmo, a coleta manual geraria uma jornada de trabalho estimada em 70h20min para obtenção dos dados na sua totalidade, conforme apresentado na Tabela 1.

Tabela 1 – Tempo gasto com a extração

Quantidade de artigos	Manual	Automatizada
1	3min58seg *	5seg **
1064	70h20min **	1h42min *

Fonte: dados da pesquisa.
* Tempo cronometrado
** Tempo estimado

Cumpre lembrar que não se está considerando o tempo gasto com o desenvolvimento do algoritmo, que não foi possível computar, mas que é significativo. Contudo, como discutiremos ao fim, mesmo que o desenvolvimento do algoritmo exigisse tempo equivalente ou até maior que a coleta manual, sua utilização ainda se justificaria pelo fato de a) garantir a integridade dos dados (ausência de erros na coleta manual) e b) permitir a reprodutibilidade da operação no futuro com um volume maior de dados.

5 TRATAMENTO E MANIPULAÇÃO

De posse dos dados extraídos automaticamente, outras tarefas que podem ser realizadas com auxílio de processamento computacional é o seu tratamento: eliminação de registros incompletos, indesejados ou cuja informação não está dentro dos recortes da pesquisa.

Manteve-se, tal como na pesquisa utilizada como referência, a opção por delimitar a pesquisa somente a artigos publicados na seção “Artigos originais” e “Artigos de Revisão” do periódico, descartando os demais (tarefa n. 2.1). O número total de publicações extraídas automaticamente e selecionados para continuidade do trabalho estão sumarizadas a seguir na Tabela 2.

Tabela 2 – Número de publicações por seção da revista *Licere* (1998-2021)

Seção	Número de publicações	Publicações selecionadas
Artigos de Revisão	189	189
Artigos Originais	652	652
Fique Por Dentro	37	0
Relatos de Experiências	33	0
Sobre nossa capa	2	0
Tome Ciência	151	0
TOTAL	1.064	841

Fonte: dados da pesquisa.

A vantagem da coleta automatizada, nesse caso, é que a decisão sobre as informações a serem descartadas pode ser feita *a posteriori*, através de comandos simples, visto que já se dispõe de todos os dados coletados.

Em comparação, é mais fácil descartar uma informação indesejada coletada automaticamente do que, no caso de um levantamento manual, ter que voltar às fontes para recuperar uma informação cuja coleta inicial não havia sido prevista.

Na sequência, para exemplificar possibilidades de produção de informação[4] a partir dos dados coletados e também para demonstrar a viabilidade de uso de processamento computacional no tratamento e manipulação dos dados, trataremos especificamente da tarefa n. 7, referente à contagem das palavras-chave mais utilizadas nos artigos “de revisão” e “originais” da revista.

Importante ressaltar que as informações apresentadas estão praticamente em estado bruto. Sua utilização para fins de análise requereria um esforço de tratamento que não foi completamente realizado, até porque o tema aqui apresentado está relacionado ao exercício técnico e à discussão do processamento automatizado de dados e não propriamente ao conteúdo extraído ou ao seu mérito. Portanto, deliberadamente não se fará qualquer reflexão nesse sentido.

Foram descartados 106 artigos (do total de 841 selecionados, restando, portanto, 753) cujas palavras-chave não foram informadas na própria página (por não existirem tais informações ou pelo fato de se tratarem de pesquisas mais antigas, publicadas originalmente em versão impressa, particularmente para as publicações anteriores ao ano de 2007).

Após tratamento preliminar identificaram-se 752 palavras-chave distintas para um total de 753 artigos analisados. A frequência das 15 palavras-chave mais utilizadas está listada na Tabela 3.

4 Opera-se aqui com a distinção bastante comum no âmbito da ciência da computação entre dado, informação e conhecimento. Dado é uma entidade matemática puramente sintática, podendo ser descritos por estruturas de representação. De acordo com Setzer (2001, n.p.), são “sequências de símbolos quantificados ou quantificáveis”. Dados também podem ser definidos como unidades básicas a partir das quais as informações poderão ser elaboradas ou obtidas. São, portanto, fatos brutos, ainda não organizados nem processados. Por meio de operação de “processamento de dados” é que tais dados serão quantificados, organizados e manipulados de maneira a produzir informação nova. Já informação é o conjunto de dados que permite a extração de algum significado, podendo, num estágio seguinte, favorecer a geração e/ou obtenção de conhecimento (LIMA; ALVARES, 2012).

Tabela 3 – Frequência absoluta e relativa de palavras-chave publicadas entre 2007-2021

Palavras-chave	Freq. absoluta.	Freq. relativa ao total de artigos
atividades de lazer	516	68,53%
esportes	116	15,41%
políticas públicas	93	12,35%
educação física e treinamento	49	6,51%
cultura	47	6,24%
lazer	45	5,98%
história	35	4,65%
jogos e brinquedos	34	4,52%
futebol	31	4,12%
educação	24	3,19%
idoso	24	3,19%
adolescente	23	3,05%
trabalho	19	2,52%
criança	18	2,39%
turismo	18	2,39%

Fonte: dados da pesquisa.

Em estado bruto, a lista discriminava palavras com ligeiras diferenças na grafia, seja por conta da flexão de número ("políticas públicas" — no plural — aparecia citada 70 vezes, enquanto "política pública" — no singular —, 21 vezes) ou de diferenças em relação a acentuação. Para correção, recorreu-se ao auxílio computacional: foi eliminada a acentuação das palavras, foram convertidos todos os caracteres em letra minúscula e aplicado algoritmo para detecção de similaridade entre palavras, gerando, por fim, novo agrupamento[5]. Procedimentos que não resolveram por completo a questão das similaridades, o que exigiria nova rodada de verificação e tratamento de cada caso individualmente (e, portanto, manualmente).

Identificaram-se também casos de palavras-chave distintas, mas com mesmo significado semântico, como, por exemplo: "jogos de vídeo", citada 13 vezes, e "jogos digitais" e "jogos eletrônicos", ambas citadas uma vez cada. Casos desse tipo até poderiam ser tratados e resolvidos com a utilização de algoritmos, os quais, por exemplo, recorreriam a buscas automatizadas *online* em dicionários de sinônimos ou banco de descritores. Contudo, a complexidade e o esforço exigidos na implementação de tal solução não se justificaria para esta pesquisa em específico, especialmente considerando o pequeno volume de dados utilizado. Já para um volume maior, envolvendo, por exemplo, palavras-chave de todos os artigos de periódicos relacionados à área da Educação Física, o esforço certamente valeria a pena.

Em todo caso, tornar plenamente úteis esses dados (referentes às palavras-chave) ainda exigiria um cuidadoso trabalho manual no seu refinamento e categorização. Tal tarefa poderia eventualmente auxiliar na construção de um banco de descritores apropriados à revista em questão ou na elaboração de parâmetros para auxiliar autores e/ou editores na definição de palavras-chave mais apropriadas.

5 Utilizou-se como limite de corte de similaridade o índice de 0,86, do método "osa", da função "stringsim", do pacote "StringDist" (VAN DER LOO, 2014). Contudo, isso não foi suficiente para tratar automaticamente todas as similaridades identificadas.

De qualquer forma, o uso computacional na automatização da extração e em parte do tratamento/manipulação auxilia consideravelmente na realização do trabalho.

Com o conjunto de dados coletados, outras operações e/ou cruzamentos poderiam ser facilmente realizados, de modo a produzir informações relativas a, por exemplo, frequência de autores e suas instituições, a distribuição das publicações ao longo dos anos, as referências e os autores mais citados, entre outras. Entende-se, contudo, que seria incoerente apresentá-los aqui sem o devido tratamento e sua subsequente (e esperada) análise.

Para concluir apresentamos no Quadro 4 a síntese dos dados brutos (*raw data*) extraídos automaticamente juntamente com uma avaliação sobre a realização de cada tarefa listada anteriormente no Quadro 2.

Quadro 4 – Automatização realizada e síntese dos dados brutos extraídos

Tarefa	Realizada automatização?	Dados brutos obtidos
1, Acessar o *site* do periódico	Sim	1 periódico (1998-2021)
2. Acessar cada fascículo	Sim	70 fascículos
2.1 Selecionar as publicações do tipo: “Artigos Originais” e “Artigos de Revisão”	Sim	841 de 1064 artigos
2.2 Acessar cada artigo selecionado	Sim	-
2.2.1 Localizar o nome de cada autor do artigo e guardá-los em uma tabela	Sim	2.107 autores
2.2.2 Contar o número de autores do artigo	Sim	Média de 2,5 autores p/ artigo
2.2.3 Localizar cada palavra-chave do artigo e guardá-las em uma tabela	Sim	2.361 palavras-chaves
2.2.4 Localizar cada referência bibliográfica e guardá-las em uma tabela	Sim	22.523 referências (de 838 artigos com referências)
3 Para cada referência bibliográfica guardada, classificá-la como livro, artigo, tese, dissertação, anais de congresso ou outro	Não (complexidade, falta de padronização)	-
4 Para cada referência classificada como artigo, registrar se foi publicado na própria revista ou não	Sim	725 ref. da própria revista
5 Contabilizar e elencar os autores mais citados nas referências bibliográficas	Não (complexidade, falta de padronização)	-
6 Contabilizar e elencar as obras mais citadas nas referências bibliográficas	Não (complexidade, falta de padronização)	-
7 Contabilizar e elencar as palavras-chave mais citadas	Sim	752 palavras-chaves únicas
8 Contabilizar e listar os autores com mais artigos publicados	Sim	1.310 autores únicos
9 Acessar o *site* da Plataforma Lattes e, para cada um dos autores com quatro ou mais artigos publicados:	Sim	-
9.1 Buscar o currículo de cada autor	Não (acesso restrito)	-
9.2 Localizar e guardar a área de formação na graduação, no mestrado e no doutorado	Não (acesso restrito)	-

Fonte: dados da pesquisa.

Para realização plena das tarefas n. 3, 4 e 5 seria necessário o desenvolvimento de algoritmo especificamente construído para identificar, em meio ao amontoado de letras que compõe uma referência, quais caracteres e palavras designam especificamente autores e quais se referem aos títulos das obras.

Ainda seria necessário prever e tratar certas exceções provocadas por erros (acidentais de digitação), equívocos (por desconhecimento) ou mesmo simples omissões (intencionais) na forma de se referir aos autores e suas obras.

Na tabela de autores que publicaram na revista, por exemplo, foram encontrados registros diferentes que, em caso de tratamento manual, poderiam ser facilmente associados a um único autor, mas que, no caso de tratamento automatizado, exigiriam buscas e confirmações através de outras plataformas (como Lattes ou Orcid)[6]. O mesmo ocorreu na tabela de referências, onde foram encontrados pelo menos 13 registros diferentes para denotar o mesmo autor[7].

6 POTENCIALIDADES E LIMITES

Tentou-se, por meio do exercício apresentado, mostrar que a extração e tratamento automatizado de dados pode ser uma alternativa produtiva e eficiente para auxiliar nas sistematizações e análises sobre o acumulado cada vez maior de publicações no campo científico.

Em trabalhos dessa natureza, a coleta e o levantamento de dados acaba sendo uma parte do processo de pesquisa que demanda envolvimento considerável e a realização de tarefas repetitivas e triviais. Agilizar esse processo é importante porque permite que tempo e esforço despendidos com o trabalho braçal da coleta sejam redirecionados ao exercício intelectual da análise e discussão, que é, afinal, a razão última desses tipos de trabalho.

Em comparação com pesquisas realizadas com a coleta manual de dados, a principal vantagem da automatização dos processos, além da evidente e óbvia velocidade, é o fato de que o volume de informações coletadas para análise independe da dedicação, envolvimento e atuação do pesquisador. Resolvido o problema da automatização, é praticamente insignificante a diferença entre coletar cem, mil ou dez mil artigos.

Outra vantagem é a reprodutibilidade. Uma vez automatizado, o processo pode ser repetido quantos vezes for necessário, o que pode ser interessante para acompanhar o lançamento de novas publicações. Além disso, o algoritmo pode ser compartilhado, permitindo a reprodução de estudos ou a realização de novas coletas por outros pesquisadores.

O principal limite é que a automatização é um processo artesanal custoso, que exige o envolvimento de especialista do campo das tecnologias computacionais,

6 Por exemplo, encontraram-se diferentes registros com nome e último sobrenome idênticos, que o algoritmo poderia supor se tratar de mesma pessoa (ex.: Cleber Augusto Dias; Cleber Augusto Goncalves Dias; Cleber Dias), mas cuja suposição não poderia ser confirmada tão facilmente, haja vista, por outro lado, a constatação de registros com mesmo nome e sobrenome, mas sem que necessariamente se referissem a mesma pessoa (ex.: Priscilla Pinto Silva poderia ser tanto as autoras Priscilla Ramos Pinto de Freitas Silva, quanto Priscilla Pinto Costa da Silva).

7 Disparadamente um dos mais citados nas referências, Nelson Carvalho Marcellino aparece referenciado sob diferentes combinações: com e sem abreviações, com e sem o nome do meio, com e sem vírgula após o sobrenome, com e sem ponto após o nome, com letras em excesso ou com a falta delas (ex.: Marcelino e Marcelinno), etc.

podendo não ser viável dependendo da dimensão da pesquisa e do volume de dados a ser coletado. Ademais, a coleta automatizada não dispensa por si só o trabalho de refinamento e de curadoria dos dados coletados, que é tão mais complexa quanto maior a quantidade de dados coletados.

Por fim, entende-se que esse tipo de solução pode auxiliar trabalhos de pesquisa sobre temas relacionados à Educação Física envolvendo grandes volumes de dados, obtidos não somente em plataformas de periódicos científicos, mas em redes sociais, fóruns, *blogs* e outras plataformas na *web*, o que abre novos caminhos metodológicos de pesquisa.

REFERÊNCIAS

CASTIEL, Luis David; SANZ-VALERO, Javier. Entre fetichismo e sobrevivência: o artigo científico é uma mercadoria acadêmica? **Cadernos de Saúde Pública**, v. 23, n. 12, 2007.

DIAS, Cleber *et al*. Estudos do lazer no Brasil em princípios do século XXI: panorama e perspectivas. **Movimento**, v. 23, n. 2., p. 601-616, abr./jun. de 2017.

DOURADO, Tatiana Maria Silva Galvão. **Fake news na eleição presidencial de 2018 no Brasil.** Tese (Doutorado em Comunicação e Cultura Contemporâneas) - Universidade Federal da Bahia, 2020.

HADDAWAY, Neal R. The use of web-scraping software in searching for grey literature. **Grey Journal**, v. 11, n. 3, p. 186-90, 2015.

INTERNET LIVE STATS. **[Elaboration of data by International Telecommunication Union (ITU), World Bank, and United Nations Population Division]**. Disponível em: https://www.internetlivestats.com/. Acesso em: 27 nov. 2021.

LIMA, José Leonardo Oliveira; ALVARES, Lilian. Organização e representação da informação e do conhecimento. *In*: ALVARES, Lilian. (org.). **Organização da informação e do conhecimento:** conceitos, subsídios interdisciplinares e aplicações. São Paulo: B4 Editores, 2012. 248 p. cap. 1, p. 21-48.

NASCIMENTO, Dandara Souza Araújo *et al*. Projeções exponenciais da ciência brasileira: modelos e análises quantitativas da produção científica nacional publicada nos últimos 30 anos. **Informação & Informação**, v. 26, n. 1, p. 53 – 73, jan./mar. 2021.

PIMENTA. Ricardo Medeiros. Nosso futuro em um post. cultura da velocidade, big data e a novo desafio dos “peixes” para os historiadores da era digital. **Revista Transversos**: Revista de História, n. 11, p. 9-22, dez. 2017. Disponível em: https://www.e-publicacoes.uerj.br/index.php/transversos/article/view/31510/22479.

R CORE TEAM. **The R project for statistical computing**. Disponível em: https://www.R-project.org/ Acesso em: 15 nov. 2021.

SETZER, Valdemar Waingort. **Os meios eletrônicos e a educação**: uma visão alternativa. São Paulo: Editora Escrituras, 2001. (Coleção Ensaios Transversais, v. 10).

VAN DER LOO, Mark. The stringdist package for approximate string matching. **R Journal,** v. 6, n. 1. p. 111-122, 2014.

Abstract: This paper deals with potentials and limits related to the use of data processing to assist in the production and systematization of scientific knowledge. It aims, through an experimental exercise involving the use of an algorithm, to discuss the feasibility of using automated collection techniques for surveying and producing data that can be used in scientific research. As a demonstration, it seeks to automatically reproduce processes related to the collection of research data previously published in this journal, describing methodologically how the extraction and treatment of these data was organized and developed. As a result, it finds that automated processing can be a productive and efficient alternative to assist in the systematization and analysis of the growing accumulation of publications in the scientific field, which may open new methodological paths for research in Physical Education, especially considering the volume of data subject to collection and analysis on social networks, forums and other web platforms.

Keywords: Electronic Data Processing. Information Storage and Retrieval. Bibliometrics.

Resumen: Este trabajo aborda las potencialidades y los límites relacionados con el uso del procesamiento de datos para ayudar en la producción y sistematización del conocimiento científico. Su objetivo, a través de un ejercicio experimental que implica el uso de un algoritmo, es discutir la viabilidad del uso de técnicas de recolección automatizada para la obtención y producción de datos que se puedan utilizar en el ámbito de las investigaciones científicas. A modo de demostración, se busca reproducir de manera automatizada procesos relacionados con la recolección de datos de una investigación previamente publicada en esta revista, describiendo metodológicamente cómo se organizó y desarrolló la extracción y el tratamiento de esos datos. Como resultado, se constata que el procesamiento automatizado puede ser una alternativa productiva y eficiente para ayudar en la sistematización y análisis de la creciente acumulación de publicaciones en el campo científico, lo que puede abrir nuevos caminos metodológicos para la investigación en Educación Física, especialmente considerando el volumen de datos que se pueden recolectar y analizar en redes sociales, foros y otras plataformas web.

Palabras clave: Procesamiento Automatizado de Datos. Almacenamiento y Recuperación de la Información. Bibliometría.

CONFLITO DE INTERESSES

O autor declarou que não existe nenhum conflito de interesses neste trabalho.

CONTRIBUIÇÕES AUTORAIS

Leoncio José de Almeida Reis: Conceituação; Curadoria de dados; Metodologia; Programas; Escrita.

FINANCIAMENTO

O presente trabalho foi realizado sem o apoio de fontes financiadoras..

ÉTICA DE PESQUISA

A pesquisa seguiu os protocolos vigentes nas Resoluções 466/12 e 510/2016 do Conselho Nacional de Saúde do Brasil.

COMO REFERENCIAR

REIS, Leoncio José de Almeida. Potencialidades e limites do processamento de dados em pesquisas sobre a produção científica. **Movimento**, v. 28, e28037, jan./dez. 2022. DOI: https://doi.org/10.22456/1982-8918.120556

RESPONSABILIDADE EDITORIAL

Alex Branco Fraga*, Elisandro Schultz Wittizorecki*, Mauro Myskiw*, Raquel da Silveira*

*Universidade Federal do Rio Grande do Sul, Escola de Educação Física, Fisioterapia e Dança, Porto Alegre, RS, Brasil.