Ferramenta de monitoramento web para apoio em observatórios de tendências: um estudo de caso Lattes

Paloma Reis Lucas[1], Wellington Rangel dos Santos[2], Carlos Vinícius Vasconcelos Rodrigues[3], Marcelo Vicente de Paula[4]

Resumo

Atualmente a internet produz um grande volume de dados diariamente. Esses dados são utilizados estrategicamente para o monitoramento de assuntos diversos, de interesse de observatórios de tendências. Os observatórios utilizam a Tecnologia da Informação e Comunicação (TIC) para auxiliar nos diagnósticos realizados. Uma das técnicas de TIC utilizadas para esse fim é a raspagem web. O objetivo deste trabalho é propor um software de monitoramento de dados públicos da internet a partir de temas de pesquisa da Embrapa Agroenergia. Foi utilizada a base de currículos do Lattes no experimento para desenvolver um software especialista em duas etapas. A primeira etapa consistiu em extrair dados da internet através de raspagem web e a segunda etapa tratou e transformou os dados brutos em informação. Os resultados forneceram insights que possibilitam a identificação de áreas de interesse, padrões de colaboração e o mapeamento da produção científica no Brasil.

Termos para indexação: observatório de tendências, raspagem web, produção científica.

Introdução

Atualmente a internet produz um grande volume de dados diariamente. Explorar e analisar grandes quantidades de dados para detectar mudanças emergentes podem gerar vantagens competitivas e moldar ambientes futuros (Jamra et al., 2022). Esses dados, gerados a cada instante, precisam ser monitorados de forma contínua e imediata para gerar valor. Essas informações são úteis para estudos de tendências e fazem parte de observatórios, presentes em algumas organizações.

Um observatório de tendências permite que a organização atue como radar que antecipa sinais, tendências e indicadores para serem utilizados de forma estratégica em uma empresa de ciência, tecnologia e inovação (De La Vega, 2007; Enjunto, 2010; Parreiras; Antunes, 2013). A tomada de decisões estratégicas e de ação imediata está se tornando uma tarefa complexa para empresas e formuladores de políticas, uma vez que o ambiente está sujeito a mudanças emergentes (Jamra et al., 2022). Por isso, cada vez mais os observatórios utilizam a Tecnologia da Informação e Comunicação (TIC) para auxiliar nos diagnósticos realizados.

A TIC contribui com técnicas e ferramentas de monitoramento que podem colaborar para a análise de dados massivos em um menor espaço de tempo. Uma das técnicas utilizadas para esse fim é a raspagem web. Essa técnica é usada para obter automaticamente algumas informações de um site ou de serviços web, em vez de copiá-las manualmente (Vargiu; Urru, 2012). A raspagem web consiste na coleta sistemática de páginas da internet por meio de um robô (Mauro et al., 2018). Esses dados são utilizados para o monitoramento de assuntos diversos, de interesse dos observatórios. Sem um robô, esse monitoramento é humanamente inviável.

[1] Analista de Sistemas, mestre em Governança, Tecnologia e Inovação, Embrapa Agroenergia, paloma.lucas@embrapa.br
[2] Cientista da Computação, Embrapa Agroenergia, wellington.santos@embrapa.br
[3] Analista de Sistemas, especialista em administração de redes Linux, Embrapa Tabuleiros Costeiros, vinicius.rodrigues@embrapa.br
[4] Analista de Sistemas, mestre em Gestão do Conhecimento e da Tecnologia da Informação, Embrapa Agroenergia, marcelo.paula@embrapa.br

De forma prática, essas ferramentas podem rastrear artigos científicos, empresas privadas, canais de notícias, redes sociais, fóruns, fontes oficiais do governo, como a base de dados da plataforma Lattes e do Instituto Brasileiro de Geografia e Estatística (IBGE), ou outros dados na web. A raspagem web permite que uma ampla gama de informações de diferentes fontes seja capturada. Possibilita-se, então, uma visão mais abrangente e detalhada das mudanças comportamentais e do surgimento de necessidades e inovações tecnológicas sobre temas de interesse. É possível também que, nesse monitoramento, sejam identificados perfis de indivíduos atuantes em assuntos de relevância para os observatórios.

Assim, o objetivo deste trabalho é propor um software de monitoramento de dados públicos da internet a partir de temas de pesquisa da Embrapa Agroenergia. Esse software será capaz de extrair dados, tratá-los e transformá-los em informação para apoio à tomada de decisão futura, com o diferencial de se comportar proativamente alertando os interessados, periodicamente, sobre informações relevantes quando elas surgem. A base de currículo do Lattes será utilizada como estudo de caso.

Materiais e métodos

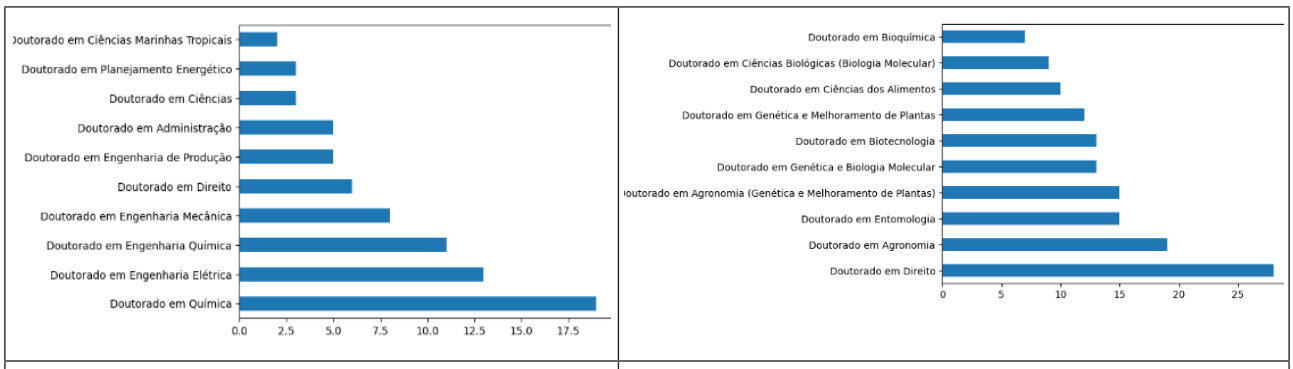
Um software especialista foi desenvolvido em duas etapas. A primeira etapa consiste em extrair informação da internet de acordo com o tema a ser monitorado. Para isso, utilizamos técnicas de consumo de dados na web com a biblioteca Selenium (Nyamathulla et al., 2021). Para a realização deste trabalho utilizamos uma amostra de dados de 698 currículos de pesquisadores da plataforma Lattes, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Selecionamos dois temas relacionados a tecnologias desenvolvidas pela Embrapa Agroenergia: “Hidrogênio verde”, com 167 currículos, e “Organismos geneticamente modificados (OGM)” com 531 currículos. Os seguintes indicadores de produção foram utilizados: a) Artigos Completos Publicados em Periódicos (ACPP); b) Trabalhos Publicados em Anais de Eventos (TPAE); c) Resumos Publicados em Anais de Eventos (RPAE); d) Livros (LV); e) Capítulos de Livros (CLV); f) Apresentações de Trabalho (AT); g) Programas de Computador sem Registro (PCR); h) Produtos (PT); i) Trabalhos Técnicos (TT); j) Orientações Concluídas de Mestrado (OCM); e l) Orientações Concluídas de Doutorado (OCD). A segunda etapa do software foi responsável por tratar e transformar os dados brutos em informação no formato de relatórios gráficos, análises estatísticas e uso de aprendizado de máquina. O software foi programado usando a linguagem Python (versão 3.10.12) com os seguintes pacotes: Selenium (versão 4.10.0), ChromeDriverManager (versão 3.8.6), Matplotlib (versão 3.7.1), Numpy (versão 1.22.4), Pandas (versão 1.5.3), Scikit-learn (versão 1.2.2) e ambientes compartilhados de programação Jupyter Notebook, RStudio e Google Colaboratory.

Resultados e discussão

A ferramenta desenvolvida mostrou eficácia na coleta e transformação dos dados, que identificam perfis de pesquisadores brasileiros e estrangeiros a partir de temas que representam algumas das linhas de pesquisa da Embrapa Agroenergia. A pesquisa avaliou quatro propostas de análises descritivas para cada um dos temas.

Distribuição por formação no doutorado

Para o tema “Hidrogênio verde”, 18% dos pesquisadores são doutores em Química (19 ocorrências) ou Engenharia Química (15 ocorrências) (Figura 1). Doutores em Direito lideram o grupo de pesquisadores “OGM”, provavelmente como especialistas na discussão sobre leis, ética, biossegurança e outros sistemas regulatórios na esfera do Direito.



“Hidrogênio verde” | “OGM”

Figura 1. Os dez cursos de doutorado mais frequentes em “Hidrogênio verde” e “OGM”.

Produção científica

Em média, os dois grupos de temas analisados apresentaram algumas diferenças. “OGM” teve maior número médio de artigos completos publicados em periódicos e em resumos publicados em anais de eventos quando comparado com “Hidrogênio verde” (Tabela 1). Este último se difere em trabalhos publicados em anais de eventos, com quase o dobro de média. Entretanto, todas as medidas apresentaram altos valores de desvio padrão. Isso indica que a produção científica não é distribuída entre os pesquisadores, ou seja, poucos têm números elevados e a maioria ainda está construindo uma progressão em publicações. Há uma oportunidade de monitoramento ao longo do tempo desses números a fim de se obter um indicativo de tendência da área avaliada. O crescimento em publicações e sua distribuição mais homogênea podem indicar se a área está ganhando ou perdendo força entres os pesquisadores.

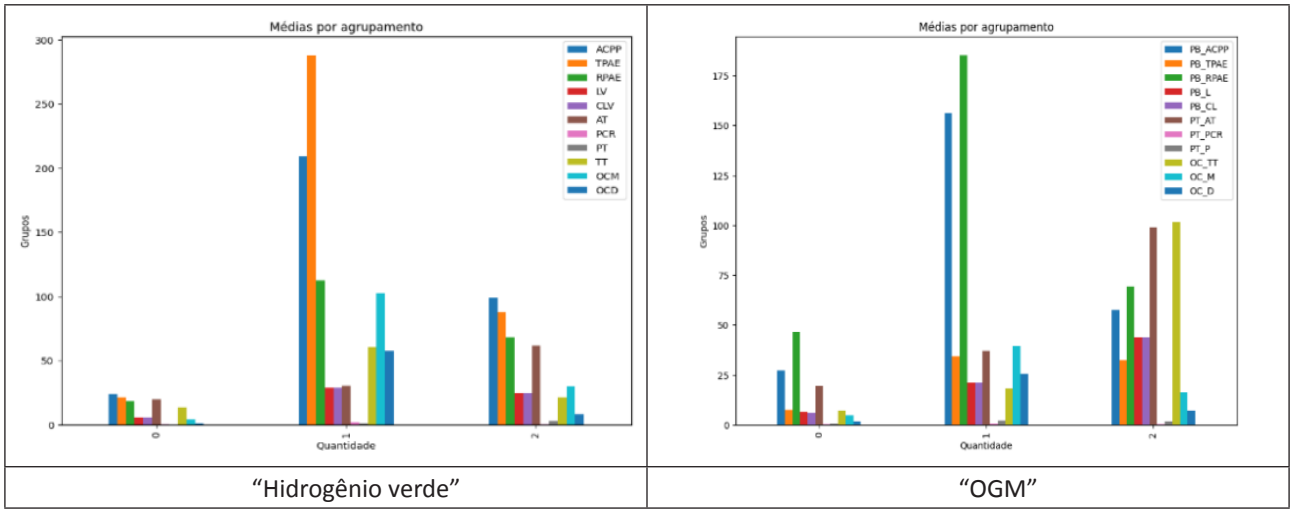
Tabela 1. Média e desvio padrão dos quatro maiores indicadores entre “Hidrogênio verde” e “OGM”.

Tema	Artigos Completos Publicados em Periódicos	Trabalhos Publicados em Anais de Evento	Resumos Publicados em Anais de Eventos	Livros ou Capítulos de Livros
Hidrogênio verde	39,7 (56,2)	37,2 (52,4)	28,1 (47,9)	18,6 (46,1)
OGM	46,9 (62,8)	12,7 (25,0)	67,0 (80,8)	20,9 (34,2)

Bolsa por produtividade

Entre todos os pesquisadores analisados, 15,3% têm bolsa de produtividade. Separamos os pesquisadores em três grupos usando o algoritmo de aprendizado de máquina não supervisionado k-Means (Bock, 2007). Para identificar o melhor número de grupos, usamos o teste Elbow (Cui, 2020). Para ambos os temas “Hidrogênio verde” e “OGM”, o resultado mostrou que o grupo ‘0’ foi composto por pesquisadores com indicadores de produtividade em menor patamar (Figura 2). O grupo ‘1’, por pesquisadores que apresentaram os maiores valores dos indicadores de produção. O último grupo apresentou números em maior equilíbrio entre esses indicadores. Quanto à bolsa de produtividade para o “Hidrogênio verde”, o grupo ‘0’ teve 10,3% de pesquisadores com bolsa, o grupo ‘1’ teve 100% de bolsistas e o grupo ‘2’, 53,6% com bolsa de produtividade. Para “OGM”, o grupo ‘0’ teve 6,1% de pesquisadores com bolsa, o grupo ‘1’ teve 57,5% de bolsistas e o grupo ‘2’ teve 21,9% de bolsistas de produtividade. Esses resultados mostraram que a bolsa de produtividade aumenta à medida que os indicadores de publicação aumentam. Apesar de o grupo de

pesquisadores em OGM apresentar três vezes mais indivíduos, este grupo tem menor percentual de bolsistas (14,1%) do que o grupo dos pesquisadores em “Hidrogênio verde” (19,2%).

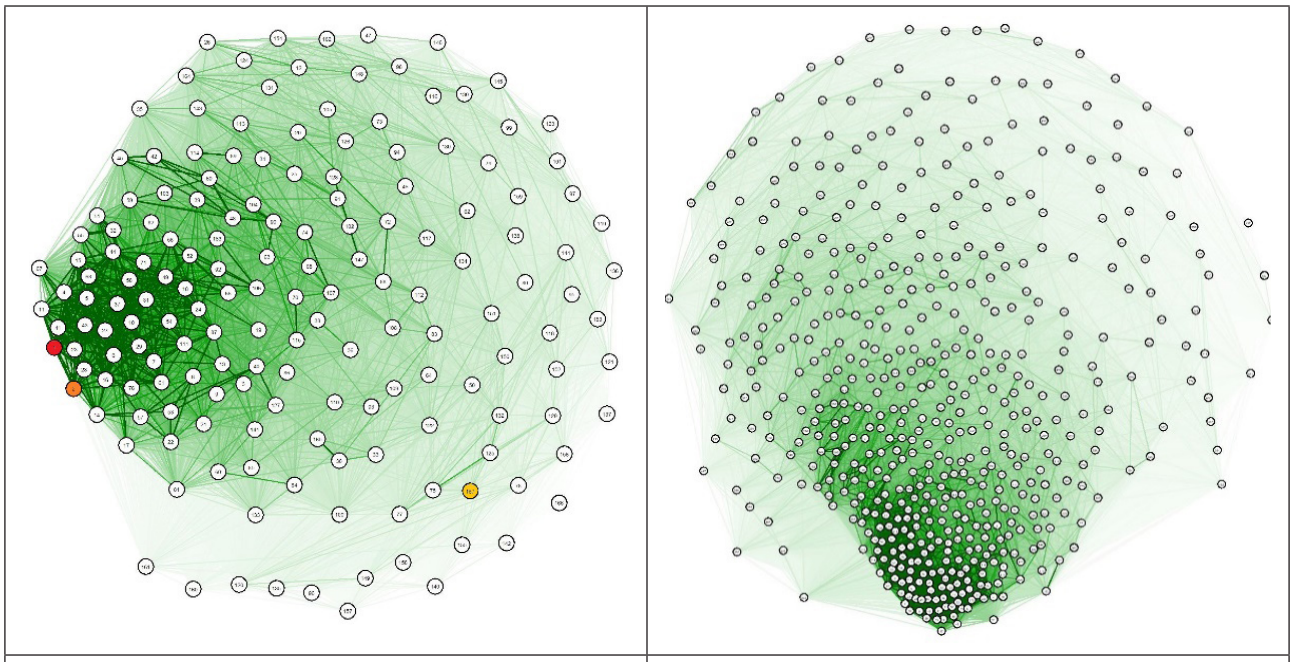


“Hidrogênio verde” | “OGM”

Figura 2. Três agrupamentos de pesquisadores por seus indicadores de produtividade.

Similaridade entre pesquisadores

A partir dos indicadores de produção científica analisados, o resultado da técnica medida de distância euclidiana apresentou-se como uma ferramenta viável como métrica de comparação entre perfis similares. A Figura 3 mostra uma rede onde cada ponto representa um pesquisador avaliado, e as arestas são as distâncias entre eles. Por exemplo, para o “Hidrogênio verde”, o pesquisador “1” (em vermelho) tem nove pontos de distância do pesquisador “2” (em laranja) e 1.141 pontos do pesquisador “167” (em amarelo). Uma rede de colaboração pode ser facilitada a partir dessa métrica. Pesquisadores experientes podem colaborar melhor com pesquisadores mais produtivos no tema de interesse.



“Hidrogênio verde” | “OGM”

Figura 3. Três agrupamentos de pesquisadores por seus indicadores de produtividade. A área mais densa representa perfis mais similares entre os indicadores de produtividade.

Conclusão

O presente estudo demonstra a importância da análise de dados obtidos por raspagem da web. O estudo de caso da plataforma Lattes contribui para a compreensão das tendências e dos perfis dos pesquisadores de acordo com a linha de pesquisa da Embrapa Agroenergia. Os resultados fornecem insights valiosos para formuladores de políticas públicas, instituições acadêmicas e pesquisadores, possibilitando a identificação de áreas de interesse, padrões de colaboração e o mapeamento da produção científica no Brasil. A abordagem de raspagem se mostrou uma ferramenta poderosa para a extração de informações relevantes de grandes volumes de dados, impulsionando a pesquisa acadêmica e a tomada de decisões em áreas do conhecimento importantes para a corporação.

Referências bibliográficas

BOCK, H. H. Clustering methods: a history of k-means algorithms. In: BRITO, P.; BERTRAND, P.; CUCUMEL, G.; CARVALHO, F. de. (Ed.). **Selected contributions in data analysis and classification**. Berlin: Springer, 2007. p. 161-172.

CUI, M. Introduction to the k-means clustering algorithm based on the elbow method. **Accounting, Auditing and Finance**, v. 1, n. 1, p. 5-8, 2020.

DE LA VEGA, I. Tipología de observatorios de ciencia y tecnologia: los casos de América Latina y Europa. **Revista Española de Documentación Científica**, v. 30, n. 4, p. 545-52, 2007.

ENJUNTO, N. Razón de ser de los observatorios. In: JORNADA OBSERVANDO OBSERVATORIOS: ¿NUEVOS AGENTES EN EL TERCER SECTOR? 2010, Madrid. **[Anais...]** Madrid: Plataforma del Voluntariado de España, 2010. Disponível em: <https://plataformavoluntariado.org/wp-content/uploads/2018/10/observando-observatorios.-nuevos-agentes-en-el-tercer-sector.pdf>. Acesso em: 25 jul. 2023.

JAMRA, H. A.; SAVONNET, M.; LECLERCQ, E. BEAM: a network topology framework to detect weak signals. **International Journal of Advanced Computer Science and Applications**, v. 13, n. 4, p. 16-27, 2022.

MAURO, A. de; GRECO, M.; GRIMALDI, M.; RITALA, P. Human resources for Big Data professions: a systematic classification of job roles and required skill sets. **Information Processing & Management**, v. 54, n. 5, p. 807-817, 2018.

NYAMATHULLA, S.; RATNABABU, P.; SHAIK, N. S.; LAKSHMI, B. A review on selenium web driver with python. **Annals of the Romanian Society for Cell Biology**, v. 25, n. 4, p. 16760-16768, 2021.

PARREIRAS, V. M. A.; ANTUNES, A. M. de S. Aplicação de foresight e inteligência competitiva em um centro de P&D empresarial por meio de um observatório de tendências: desafios e benefícios. **Revista Gestão & Conexões**, v. 1, n. 1, p. 55-73, 2013.

VARGIU, E.; URRU, M. Exploiting web scraping in a collaborative filtering-based approach to web advertising. **Artificial Intelligence Research**, v. 2, n. 1, p. 44-54, 2012.