INTER FACES CIENTÍFICAS

EXATAS E TECNOLÓGICAS

ISSN IMPRESSO - 2359-4934
ISSN ELETRÔNICO - 2359-4942
http://dx.doi.org/10.17564/2359-4942.2018v3n2

BHEXTRACT – EXTRAÇÃO DE DADOS DE SITES DE REVISTAS CIENTÍFICAS NACIONAIS SOBRE EDUCAÇÃO

BHEXTRACT - DATA EXTRACTION FROM WEBSITES OF NATIONAL EDUCATION SCIENTIFIC JOURNALS

BHEXTRACT - EXTRACCIÓN DE DATOS DE SITIOS DE REVISTAS CIENTÍFICAS NACIONALES SOBRE EDUCACIÓN

Gabriel Menezes da Silva[1]
Antônio Cleverson dos Santos[2]
Layse Santos[3]

RESUMO

Bases dispersas inviabilizam ou no mínimo dificultam a descoberta de informações relevantes ao trabalho do pesquisador no campo da educação. Assim, o objetivo deste estudo é utilizar a web scraping para extrair dados de revistas científicas que contenham produções que abordam sobre educação, com o intuito de centralizá-los numa base de dados visando auxiliar na busca de informações. Foi realizada uma revisão de literatura sobre inteligência artificial e mineração de dados, seguido de um estudo de caso por meio de um protótipo de extração. Como resultado, foram obtidos 1237 artigos de 3 periódicos e uma base de dados unificadas com essas produções científicas foi construída.

PALAVRAS-CHAVES

Mineração de Dados. Artigos. Inteligência Artificial. Estudo de Caso. Educação.

ABSTRACT

Scattered bases makes harder the discovery of relevant information for the scientific papers of education researchers. Therefore, the purpose of this study is to extract data from scientific journals containing educational papers, using web scraping, and create a centralized database to assist further information searches. It was made a literature revision about artificial intelligence and data mining, followed by a case study using a extraction prototype. As a result, 1237 articles were obtained from 3 scientific journals and a unified database was built with the data from these papers.

KEYWORDS

Data mining. Papers. Artificial Intelligence. Case Study. Education

RESUMEN

Las bases dispersas impiden o al menos dificultan el descubrimiento de informaciones relevantes al trabajo del investigador en el campo de la educación. Así, el objetivo de este estudio es utilizar la **web scraping** para extraer datos de revistas científicas que contengan producciones que aborden sobre educación, con el propósito de centrarlos en una base de datos buscando auxiliar en la búsqueda de informaciones. Se realizó una revisión de literatura sobre inteligencia artificial y minería de datos, seguida de un estudio de caso por medio de un prototipo de extracción. Como resultado, se obtuvieron 1237 artículos de 3 periódicos y una base de datos unificada con esas producciones científicas fue construida.

PALABRAS CLAVE

minería de datos. artículos. inteligencia artificial. estudio de caso. Educacíon

1 INTRODUÇÃO

A produção intelectual brasileira é prejudicada pela falta de organização do órgão dos documentos (AMORIM; 2000). Além disso, bases dispersas inviabilizam ou, no mínimo dificultam a descoberta de informações relevantes ao trabalho do pesquisador no campo da educação. A inteligência artificial juntamente com a mineração de dados proporciona o poder de extrair informações úteis a partir de dados brutos, determinando vantagens em um mundo onde valores como produtividade e competitividade estão em alta (ROCHA; CORTEZ; NEVES, 2008).

Diante desse contexto, este projeto de pesquisa tem como principal objetivo utilizar da computação, mais especificamente das técnicas de inteligência artificial e mineração de dados, para extrair dados de revistas científicas que contenham produções de qualificação qualis, que sejam de extrato A1 do quadriênio 2013/2016, que abordam sobre a educação, com o intuito de centralizá-las em um único banco de dados, utilizando algoritmos computacionais para auxiliar na busca de informações sobre esse assunto, para que no futuro seja possível a criação de um portal similar ao **IEEE** e **ACM**, mas que contenha apenas produções da área da educação.

Segundo Gondra (2000), o emprego de novos recursos para suporte à pesquisa educacional, é uma imposição da atualidade aos investigadores da área. Segundo Galvão e Lopes (2010) é necessário que surjam novas ferramentas conceituais que possibilitem um melhor refinamento das análises dos dados com o intuito de criar técnicas que permitam organizar o inventário de fontes bibliográficas por meio da informatização de banco de dados.

Embora exista um conjunto de sites com revistas científicas sobre a educação, estas estão dispersas, dificultando ao pesquisador na busca de informação, assim, neste contexto, torna-se necessário a criação de uma base unificada de informações sobre as publicações da área da educação.

O artigo está organizado da seguinte forma: a fundamentação teórica, na seção 2, a metodologia utilizada, na seção 3, o desenvolvimento, na seção 4, os resultados atingidos na seção 5, a conclusão na seção.

2 ESTUDO TEÓRICO

Nesta seção, são abordados alguns conceitos que embasam este projeto científico.

O processo de descoberta de conhecimento ou **Knowledge Discovery in Databases** (KDD), cujo termo foi cunhado em 1989, no âmbito das discussões do **International Joint Conference on Artificial Intelligence**, em **Detroit** (Michigan/EUA). A ocasião reuniu os principais pesquisadores em aprendizagem de máquina, banco de dados, lógica difusa, aquisição de conhecimentos, entre outras áreas (PIATETSKY-SHAPIRO, 1990). O processo KDD visa transformar dados brutos em conhecimento de alto-nível (ROCHA; CORTEZ; NEVES, 2008), o que possibilita aplicá-lo na mineração de dados sobre temas e contextos específicos no âmbito educacional.

A extração, organização e descoberta de padrões, de forma sistêmica, pode apoiar análises de dados históricos, assim a mineração de dados surge da necessidade de extrair informações potencialmente úteis de uma ou várias bases de dados (ROCHA; CORTEZ; NEVES, 2008), com o intuito de utilizá-las para algum propósito, assim é possível, por meio de algoritmos, empregar técnicas de extração e análise de dados em diversas linguagens ou ferramentas.

Web scraping significa extrair dados da internet de forma que não seja por intermédio de uma pessoa, usando um navegador web e sim por meio de algoritmos computacionais (MITCHEL, 2015). Assim, após a extração dos dados, eles podem ser salvos em um banco de dados que pode ser relacional ou não-relacional.

Os bancos de dados relacionais possuem um alto fator de crescimento, porém, quanto maior o tamanho mais custoso se torna essa escalabilidade, seja pelo custo de novas máquinas, seja pelo aumento de especialistas que irão manipular esse banco de dados (IANNI, 2018). Diante desse contexto, para este projeto foi escolhido o MongoDB, banco de dados não relacional que permite uma escalabilidade menos custosa e trabalhosa, reunindo características que torna possível trabalhar com dados semiestruturados ou crus, vindo

de diversas origens (arquivos de multimídia, **web sites** etc). Os dados serão disponibilizados para acesso em um **web service** que utiliza a arquitetura REST.

A arquitetura **Representational State Transfer** (REST) representa uma nova possibilidade na criação de **web services**, possibilitando leveza e simplicidade nos pacotes de dados que serão trafegados na rede, fazendo desnecessária a criação de camadas intermediárias para encapsular os dados (CÉSAR, 2018). Assim, esse modelo foi escolhido para agilizar a entrega de resultados.

3 MÉTODO

A pesquisa é de caráter descritivo e exploratório e emprega um estudo de caso como método. O uso de estudos de caso permite um estudo específico sobre um determinado caso, ou seja, busca conhecer em profundidade "como" e/ou "por que" um determinado fenômeno ocorre (RUNESON *et al.*, 2012). No caso da presente pesquisa, pretende-se responder "como?" criar uma base centralizada, unificando dados das revistas científicas sobre educação. Para a realização do estudo de caso, foram utilizadas algumas das etapas propostas por Fayyad, descritas a seguir (FAYYAD; PIATESKY-SHAPIRO; SMYTH, 1996):

Etapa 1: Entender o domínio onde será aplicado o processo de KDD, identificando os objetivos do ponto de vista do cliente;
Etapa 2: Selecionar uma série de dados onde será executado o processo;
Etapa 3: Limpeza e pré-processamento dos dados;
Etapa 4: Geração de uma base de dados centralizada.

4 DESENVOLVIMENTO

Os algoritmos e a **Application Programming Interface** (API) produzidos neste projeto realizam **web scraping**. Assim, nesta seção será descrito o cenário e as características dos mesmos.

4.1 CENÁRIO

Atualmente no portal de periódico da CAPES, a área de educação possui classificado 4203 periódicos (PLATAFORMA..., 2018), tornando complexo a busca individual de trabalhos neste conjunto. Assim, para facilitar a busca de artigos científicos que abordam a educação, estão sendo desenvolvidos algoritmos e uma API de consulta para extrair informações de revistas que abordam esse tema com o intuito unificá-los em uma única base de dados.

4.2 CARACTERÍSTICAS

Para execução e análise deste projeto foram desenvolvidos algoritmos em Python, a 4° linguagem mais utilizada mundialmente como demonstrado na Figura 1 (TIOBE, 2018).

Figura 1 – A tabela apresenta a classificação das linguagens de programação mais utilizadas mundialmente

Apr 2018	Apr 2017	Change	Programming Language	Ratings	Change
1	1		Java	15.777%	+0.21%
2	2		C	13.589%	+6.62%
3	3		C++	7.218%	+2.66%
4	5	^	Python	5.803%	+2.35%
5	4	v	C#	5.265%	+1.69%
6	7	^	Visual Basic .NET	4.947%	+1.70%
7	6	v	PHP	4.218%	+0.84%
8	8		JavaScript	3.492%	+0.64%

Fonte: TIOBE, 2018

Além disso, essa linguagem possui uma grande comunidade, facilitando na resolução de dúvidas durante a construção dos algoritmos e uma ampla biblioteca de códigos que aborda inteligência artificial e mineração de dados. A biblioteca escolhida foi a **Beatiful Soap**.

A API de consulta foi construída em Javascript, a 8° linguagem mais utilizada mundialmente de acordo com a Figura 1, que oferece as mesmas vantagens que a linguagem Python. O banco de dados utilizado para armazenar as informações foi o MongoDB.

Os algoritmos desenvolvidos em Python, por meio do uso dos métodos disponibilizados pela **Beatiful Soap**, analisam e obtéêm um padrão de como cada página dos portais como: **Open Journal System e Scielo** se comportam, com o intuito de obter dados como: título, resumo e **abstract**.

Os dados extraídos das revistas foram armazenados no banco NoSQL e disponibilizados para consulta por meio de API, desenvolvida em Nodejs. A qual fornece uma interface de consulta com filtros de busca por nome de autor, palavras chaves, título, data de publicação, entre outros dados relacionados aos periódicos e seus artigos.

O algoritmo analisa a url da requisição e de acordo com as cláusulas passadas como parâmetro, constrói a melhor consulta de forma transparente e performática.

5 RESULTADOS

Nas seções abaixo estão descritos os resultados obtidos durante a execução deste trabalho científico que foram obtidas a partir das etapas baseadas na proposta de Fayyad, descritas na seção 3.

5.1 ETAPA 1: ENTENDER O DOMÍNIO

O domínio desta produção científica consiste nas revistas, que abordam pesquisas na área da educação, disponíveis no portal de periódico capes e com classificação qualis de extrato A1 do quadriênio 2013/2016.

A classificação qualis-periódicos se refere ao sistema usado para classificar e estratificar os artigos publicados em periódicos científicos dos programas de pós-graduação. A classificação é feita anualmente por cada área de avaliação. As produções científicas são qualificadas de acordo com os seguintes indicativos de qualidade: A1, o mais elevado; A2; B1; B2; B3; B4; B5 e C. Esta classificação não define de forma absoluta a qualidade dos periódicos, pois o mesmo pode ter avaliações diferentes a depender da área a qual ele é avaliado e da pertinência do seu conteúdo para a mesma (CAPES, 2018).

O Hirsch-Index (ou h-index), introduzido por Jorge Hirsch em 2005, é um dos parâmetros mais utilizados e aplicados no âmbito científico e acadêmico. Por meio dele é possível classificar os cientistas baseado em suas produções (EGGHE, 2010). Diante desse contexto, para este trabalho científico, foram escolhidos periódicos com h-index maior ou igual a cinco.

5.2 ETAPA 2: SELECIONAR UMA SÉRIE DE DADOS ONDE SERÁ EXECUTADO O PROCESSO

Primeiramente, uma pesquisa resultou na seleção dos periódicos Qualis, conforme determinado no domínio, para que a partir delas sejam executados os algoritmos criados para que se possa extrair os dados requisitados. Os periódicos obtidos estão na Figura 2.

Figura 2 – Revistas científicas com grande impacto no âmbito acadêmico nacional

ISSN	Título	Estrato	Hindex
0120-0534	REVISTA LATINOAMERICANA DE PSICOLOGIA	A1	19
1678-7153	PSICOLOGIA: REFLEXÃO E CRÍTICA	A1	19
1678-4634	EDUCAÇÃO E PESQUISA	A1	18
1806-9584	REVISTA ESTUDOS FEMINISTAS	A1	17
1413-2478	REVISTA BRASILEIRA DE EDUCAÇÃO	A1	17
1809-449X	REVISTA BRASILEIRA DE EDUCAÇÃO	A1	17
1806-9584	ESTUDOS FEMINISTAS	A1	17
1984-0411	EDUCAR EM REVISTA	A1	15
0104-4060	EDUCAR EM REVISTA	A1	15
0104-4060	EDUCAR EM REVISTA (IMPRESSO)	A1	15
1982-5765	AVALIAÇÃO: REVISTA DA AVALIAÇÃO DA EDUCAÇÃO SUPERIOR	A1	15
0103-863X	PAIDEIA (RIBEIRAO PRETO)	A1	13
1980-850X	CIÊNCIA & EDUCAÇÃO (ONLINE)	A1	12
1677-941X	ACTA BOTANICA BRASILICA	A1	12
1646-401X	REVISTA LUSOFONA DE EDUCACAO	A1	11
1645-7250	REVISTA LUSÓFONA DE EDUCAÇÃO	A1	11
0871-9187	REVISTA PORTUGUESA DE EDUCAÇÃO	A1	10
2238-0094	REVISTA BRASILEIRA DE HISTÓRIA DA EDUCAÇÃO	A1	8
1519-5902	REVISTA BRASILEIRA DE HISTÓRIA DA EDUCAÇÃO	A1	8
2237-2660	REVISTA BRASILEIRA DE ESTUDOS DA PRESENÇA	A1	5
2237-2660	REVISTA BRASILEIRA DE ESTUDOS DA PRESENÇA [EPERIODICO]	A1	5
0102-0188	REVISTA BRASILEIRA DE HISTÓRIA	A1	5

Fonte: Elaborado pelos autores

5.3 ETAPA 3: LIMPEZA E PRÉ-PROCESSAMENTO DE DADOS

Nesta pesquisa, a limpeza e o pré-processamento de dados consiste no fato dos algoritmos coletarem toda a página do periódico e dela extrair apenas o título, link e resumo dos artigos científicos. Foram coletados 179 Journals com a execução do algoritmo feito em Python. Como exemplificação de um dos resultados obtidos, na Figura 3 estão as informações de uma produção científica que está na Revista de Estudos Feministas:

Figura 3 – Título, link e resumo de um artigo contido na Revista de Estudos Feministas

```
Um fim à negligência em relação aos problemas da mulher negra!
https://periodicos.ufsc.br/index.php/ref/article/view/53336
O artigo é a tradução de um ensaio publicado originalmente em 1949, pela intelectual e ativista negra Claudia Jones na revista Political Af
fairs. No ensaio, Jones demonstra as origens e as múltipmas dimensões da dinâmica do sistema de opressão a que eram submetidas mulheres neg
ras e critica a inabilidade dos comunistas estadunidenses em mobilizá-las. Ao argumentar que as mulheres negras compunham a fração superexp
lorada da classe trabalhadora, a autora as posiciona como parecela central da militância internacional contra o fascismo e o imperialismo.
```

Fonte: Elaborado pelos autores

Foram consultados, pela API, 3 revistas e 1058 artigos. Os resultados obtidos pela mesma podem ser vistos por meio dos respectivos links que permitem consultar: Título, **abstract**, palavras-chave e autores, quantidade de revistas consultadas e quantidade de artigos coletados: http://bvhep-com-br.umbler.net/articles/all,http://bvhep-com-br.umbler.net/magazines/count, http://bvhep-com-br.umbler.net/articles/count:

Figura 4 – Título, abstract, palavras-chave e autores de algumas revistas coletadas pela API

Fonte: Elaborado pelos autores

5.4 ETAPA 4: GERAÇÃO DE UMA BASE DE DADOS CENTRALIZADA

Os dados obtidos pela API, como podem ser vistos na Figura 2, são salvos em um banco de dados de forma centralizada para posterior consulta.

6 CONCLUSÃO

Este trabalho teve como objetivo a criação de uma base unificada de informação que aborda sobre as publicações da área da educação, tendo como resultado um banco de dados que centraliza essas produções relativa aos periódicos com classificação Qualis. Este resultado é de grande importância para o crescimento do âmbito acadêmico da educação no Brasil por unificar dados de diversos periódicos, facilitando a busca por parte dos pesquisadores.

Este trabalho é o passo inicial para a criação de um portal onde será possível procurar produções científicas que tratem sobre o tema, permitindo que pesquisadores tenham acesso facilitado a base de conhecimento sobre educação, incentivos para produzir mais devido a maior facilidade para divulgar os seus trabalhos científicos e maior facilidade em obter referências bibliográficas necessárias para a escrita do seu artigo, o que mostra que esta pesquisa tem um alto potencial de continuidade e aprofundamento. Além disso, como trabalhos futuros, enseja implementar um modelo de classificação da produção, facilitando a busca por novos conhecimentos.

REFERÊNCIAS

AMORIM, Eliane Dutra. Arquivos, pesquisa e as novas tecnologias. In: FARIA FILHO, Luciano Mendes (Org.). **Arquivos, fontes e nova tecnologia:** questões para a história da educação. Campinas: Autores Associados/ Bragança Paulista: Universidade São Francisco, 2000. p.89-99.

CÉSAR, Paulo. **Primeiros passos com o serviço REST**. Disponível em: <https://www.devmedia.com.br/primeiros-passos-com-os-servicos-rest/28912>. Acesso em: 3 mar. 2018.

CLASSIFICAÇÃO da produção atual. Disponível em: <http://www.capes.gov.br/avaliacao/instrumentos-de-apoio/classificacao-da-producao-intelectual>. Acesso em: 7 maio 2018.

EGGHE, Leo. The Hirsch index and related impact measures. **Annual review of information science and technology**, v.44, n.1, p.65-114, 2010.

FAYYAD, U.M.; PIATESKY-SHAPIRO; SMYTH, P. "From Data Mining to Knowledge Discovery: An Overview", In: **Advances in Knowledge Discovery and Data Mining, AAAI Press**, 1996.

GALVÃO, Ana Maria de Oliveira; LOPES, Eliane Marta Teixeira. **Território plural:** a pesquisa em história da educação. São Paulo: Ática, 2010.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. Data mining: conceitos, técnicas, algoritmos, orientações e aplicações. 2.ed. Rio de Janeiro: Elsevier, 2015.

GONDRA, José G. A leveza dos bits. In: FARIAS FILHO, Luciano Mendes. **Arquivos, fontes e novas tecnologias:** questões para a História da Educação. Campinas: Editores Associados, 2000. p.3-17.

IANNI, Vinicius. **Introdução ao Banco de Dados NoSQL**. Disponível em: <https://www.devmedia.com.br/introducao-aos-bancos-de-dados-nosql/26044>. Acesso em: 3 mar. 2018.

MITCHELL, Ryan. **Web Scraping com Python:** Coletando dados na web moderna. São Paulo, Brasil: Novatec.

PIATETSKY-SHAPIRO, Gregory. Knowledge Discovery in Real Databases: a report on the IJCAI-89 Workshop. **Artificial Intelligence Magazine**, Palo Alto, v.11, n., p.68-70, 1990.

PLATAFORMA Sucupira. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf>. Acesso em: 8 maio 2018.

ROCHA, Miguel; CORTEZ, Paulo; NEVES, José Maia. **Análise inteligente de dados, algoritmos e implementação em Java**. Lisboa: FCA, 2008.

RUNESON, Per *et al*. Case study research in software engineering: Guidelines and examples. **John Wiley & Sons**, 2012.

TIOBE Index for February 2018. Disponível em: <https://www.tiobe.com/tiobe-index/>. Acesso em: 3 mar. 2018.

Recebido em: 10 de Fevereiro de 2018
Avaliado em: 20 de Março de 2018
Aceito em: 5 de Abril de 2018

1 Graduando em Ciência da Computação; Membro do GPTIC.
Email: gabriel.menezes96@souunit.com.br

2 Graduando em Ciência da Computação; Membro do GPTIC.
Email: antonio.cleverson@souunit.com.br

3 Bacharel em Ciências da Computação; membro do GPITIC.
Email: santoslay3@gmail.com