Engenharia de Dados e Biblioteconomia: A Modernização do Catálogo Coletivo Nacional de Publicações Seriadas (CCN)

Bruno Costa[1], João Gabriel Viana[3], Gabriélle Santos[4], Greicy Santos[2], Tainá Batista de Assis[2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro (IFRJ)
[2]Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)
[3]Instituto Federal de Educação, Ciência e Tecnologia de Goiás
[4]Universidade Federal do Mato Grosso do Sul (UFMS)

bruno.costa@ifrj.edu.br, jggviana202@gmail.com, gabrielle.helpis@gmail.com
greicysantos@ibict.br, taina@ibict.br

Abstract. *The National Collective Catalog of Serial Publications (CCN) is a public access catalog that brings together information about serial, technical and scientific publications, as well as monographic series and collection indices available in the collections of Brazilian libraries. Under the administration and coordination of the Brazilian Institute of Information in Science and Technology (Ibict), the CCN aims to promote the dissemination, identification, and localization of serial publications available throughout the country. This presentation aims to report the challenges and strategies applied in the modernization of the CCN system from the perspective of Data Engineering.*

Resumo*. O Catálogo Coletivo Nacional de Publicações Seriadas (CCN) é um catálogo de acesso público que reúne informações sobre as publicações seriadas, técnicas e científicas , bem como de séries monográficas e índices de coleção disponíveis nos acervos das bibliotecas brasileiras. Sob a administração e coordenação do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), o CCN visa promover a difusão, a identificação e a localização de publicações seriadas disponíveis em todo o País. Esta apresentação tem por objetivo relatar os desafios e as estratégias aplicadas na modernização do sistema CCN sob a perspectiva da Engenharia de Dados.*

1. Introdução

O Catálogo Coletivo Nacional de Publicações Seriadas (CCN) é um catálogo de acesso público que reúne informações sobre periódicos, monografias seriadas, obras de referência e outras publicações semelhantes disponíveis nas bibliotecas brasileiras. Essas bibliotecas disponibilizam o seu acervo compondo, assim, a chamada "Rede CCN". Sob a administração e coordenação do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), o CCN visa promover a difusão, a identificação e a localização de publicações seriadas disponíveis no país, padronizando a descrição bibliográfica e de coleções conforme critérios internacionais. Atualmente, a Rede CCN conta com centenas de bibliotecas e o sistema permite consultas pela Internet.

Esta apresentação tem por objetivo relatar os desafios e as estratégias aplicadas na modernização do sistema CCN sob a perspectiva da Engenharia de Dados. Serão demonstradas as técnicas utilizadas na identificação de inconsistências nos dados, além de protótipos desenvolvidos para a correção automática de informações catalográficas.

2. Catálogo Coletivo Nacional de Publicações Seriadas (CCN)

Criado em 1954, o Catálogo Coletivo Nacional de Publicações Seriadas (CCN) é um serviço do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Unidade de Pesquisa vinculada ao Ministério da Ciência, Tecnologia e Inovação (MCTI), que tem por objetivo reunir as informações sobre os títulos e as coleções de publicações seriadas técnicas e científicas, nacionais e estrangeiras, presentes nos acervos bibliográficos das instituições brasileiras.

Uma consulta ao CCN equivale a pesquisar em centenas de catálogos bibliográficos simultaneamente, aumentando assim a abrangência da pesquisa, a otimização da busca, a economia tempo e recurso e sem custos. Isso porque o CCN mantém em seu banco de dados as informações de publicações seriadas técnicas e científicas de mais de 600 bibliotecas brasileiras, com mais de 62 mil títulos de periódicos e mais de 3 milhões de transcrições de coleções. As bibliotecas que integram o CCN são, em sua maioria, de instituições brasileiras de ensino e pesquisa, distribuídas por todas as regiões do Brasil.

3. Inconsistências nas Transcrições do CCN

A norma que define os elementos de dados que compõem a descrição da coleção de uma publicação seriada do CCN e as regras para a sua transcrição baseou-se, inicialmente, na forma de apresentação de dados do padrão ISO 690/1975 Bibliographical Reference - Essential and Supplementary Elements vem sendo aperfeiçoada ao longo dos anos de existência do catálogo**.** Com base nessa norma incorporada de características do padrão NISO Z39.44, consideradas relevantes para catálogos coletivos, uma transcrição deve descrever o ano, volume e fascículo de uma publicação seriada [2, 3]. A transcrição **1987 20(1),** por exemplo**,** descreve que a publicação é do ano de **1987** com o volume **20** com o fascículo **1.**

Em uma exploração prévia dos dados da base de coleções do CCN foi possível identificar manualmente as inconsistências mais comuns nas descrições das

transcrições. A inconsistência mais frequente identificada foi em relação a abreviação do ano e dos volumes, por outro lado, a de menor ocorrência foi em relação à repetição de volumes em uma mesma data de publicação. Na transcrição *1996-2003 25-32*, por exemplo, foi identificado um erro no agrupamento de anos e na *1977 6(1), 6(2), 6(3)* um erro de repetição de volumes de um mesmo ano. Também foi identificado erro de números de fascículos escritos com zero à esquerda como em *2001 30(01,02,03)*, texto especial escrito incorretamente como em *1972-1983 2-12 IN* e ausência de ponto e vírgula entre cada período como da transcrição como *1977 6(1)_1977 7(1)*, por exemplo.

Outras inconsistências identificadas foram em relação ao espaçamento. Espaço entre o volume e o fascículo como na transcrição *1977 6 (1-2)* e dentro do fascículo como em *1978 7(1)*. Ausência de espaço entre ano e volume como a transcrição *199625(1,3)* e após o ponto e vírgula que separa o períodos das transcrições como em *1995/1999 24-28;2000 29 (1,3)*. Já a transcrição *1989 18(2) ; 1990 19(1) ; 1993 22(3) ; 1997 26(2) ; 1998 27(1-2)* contém erro de espaço antes do ponto e vírgula de cada período. As inconsistências apresentadas foram identificadas na exploração inicial dos dados do CCN. No entanto, diversos outros tipos de erros foram relatados pelos usuários do sistema.

3. Identificação Automática de Transcrições Incorretas

Para a identificação de inconsistências nas mais de 3 milhões de transcrições do CCN seguiu-se duas linhas de trabalho paralelas, desempenhadas pela mesma equipe de trabalho. A primeira consistiu em uma abordagem determinística, na qual o padrão definido na Norma ISO 690/1975 foi especificado em expressão regular (regex).

O desenvolvimento expressão regular foi feito de forma interativa e incremental, com validação contínua dos profissionais do Ibict até sua finalização. A estratégia utilizando a abordagem determinística foi de criar um regex capaz de identificar transcrições corretas de acordo com as normas, ou seja, transcrições que não descem match com o regex eram consideradas erradas.

Na segunda abordagem, utilizou-se algoritmos probabilísticos com aprendizagem de máquina. A justificativa de utilizar uma abordagem probabilística era didática, considerando que após a identificação automática das transcrições, seguia-se a necessidade de classificação dos diversos tipos de erros, além daqueles citados

anteriormente. Assim, foi desenvolvido um modelo utilizando aprendizagem supervisionada e o modelo foi aplicado às transcrições. O modelo consiste em uma rede neural utilizando o algoritmo *Sequential* da biblioteca TensorFlow. Este modelo recebe apenas uma entrada e espera uma saída. Ele passa os dados e flui em ordem sequencial da abordagem de cima para baixo até que os dados cheguem ao final do modelo. Um modelo sequencial é relevante quando há uma pilha simples de camadas. Nesta pilha, cada camada tem um tensor de entrada e um tensor de saída.

Considerando que a expressão regular era capaz de identificar as transcrições corretas, a comparação do número de transcrições identificadas pela abordagem probabilística, indicava a acurácia do modelo e, também, a maturidade da equipe para avançar na criação de modelos de classificação para os diferentes tipos de erros nas transcrições.

4. Resultados Iniciais e Desafios Futuros

Utilizando a expressão regular, das 3.291.203 transcrições de coleções do CCN, identificaram-se 2.079.732 transcrições corretas (~63%). Utilizando o modelo criado com aprendizagem de máquina, após diversas configurações e calibragens, foi possível identificar 2.435.478 transcrições corretas, o que resultou em 83% de acurácia do modelo.

Apesar de a expressão regular ter eficácia na identificação de transcrições corretas, o exercício de configuração e calibragem do modelo supervisionado permitiu maior aprofundamento da equipe de desenvolvimento na utilização de abordagens probabilísticas para a correção das transcrições. Neste sentido, novos estudos estão sendo desenvolvidos para o desenvolvimento de modelos capazes de identificar outros tipos de erros nas transcrições. Em seguida, pretende-se analisar a viabilidade de correções automáticas nas transcrições. Neste sentido, busca-se a participação no CBSoft como espaço de diálogo com especialistas da área no sentido de vislumbrar-se possíveis caminhos para o desenvolvimento das soluções.

Referências

BRASIL. 2018 Ministério da Ciência, Tecnologia e Inovação. Anexo IV- ORIENTAÇÕES PARA TRANSCRIÇÃO DE DADOS DE COLEÇÃO NO CATÁLOGO COLETIVO NACIONAL DE PUBLICAÇÕES SERIADAS - CCN. Brasília, DF: IBICT, 2018.

FITZGERALD, Michael. 2106 Introdução às Expressões Regulares. 2. ed. São Paulo: Novatec Editora