Compreensão e Preparação dos Dados do Censo da Educação Superior para Mineração de Dados

Data Preparation of Higher Education Census for Data Mining

Gabrieli Demenjon Mathias*, André Luis Schwerz†

RESUMO

Os dados do Censo da Educação Superior anualmente são disponibilizados de forma aberta pelo governo brasileiro. Entretanto, esses dados ainda são pouco explorados pela comunidade acadêmica. Por outro lado, acredita-se que a aplicação de uma metodologia de mineração de dados a fim de descobrir padrões e regras possam produzir conhecimentos relevantes sobre a educação superior brasileira para a sociedade. Neste sentido, este artigo apresenta as etapas de compreensão e preparação dos dados do Censo da Educação Superior aplicados aos anos de 2010 a 2019 seguindo a metodologia CRISP-DM. Como resultado, apresenta-se uma base de dados relacional e um código usado para converter os dados de um formato semiestruturado para uma estrutura normalizada, livre de redundâncias e inconsistências.

Palavras-chave: Mineração de Dados, Censo da Educação Superior, CRISP, Preparação dos Dados.

ABSTRACT

The Brazilian government makes available annually the Census of Higher Education as open data, including several data on Brazilian higher education. However, these data are still little explored by the academic community. The main goal is to apply a data mining methods in order to find patterns and rules that produce relevant knowledge about Brazilian Higher Education for society. More specifically, this paper presents the stage of understanding and preparing data following the CRISP-DM methodology on data from the Higher Education Census applied to the years 2010 to 2019. As a result, we present a normalized database and a code used to convert the data from a semi-structured format to a normalized structure without redundancies and inconsistencies.

Keywords: Data Mining, Higher Education, CRISP, Data Preparation.

1 INTRODUÇÃO

Recentemente, estão se popularizando iniciativas para a disponibilização de dados abertos em várias áreas do governo. Em especial, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) disponibiliza, de forma aberta, diversos dados por meio do Censo Educação Superior, do ENADE e do ENEM. Tema deste trabalho, o Censo da Educação Superior é realizado anualmente pelo Inep. Seu objetivo é reunir informações sobre as instituições de ensino superior, seus cursos de graduação presencial ou a distância, cursos sequenciais, vagas oferecidas, inscrições, matrículas, ingressantes e concluintes, e informações sobre docentes nas diferentes formas de organização acadêmica e categoria administrativa. Entretanto, essa grande quantidade de dados ainda permanece pouco explorada pela comunidade acadêmica e pelo próprio Inep em suas pesquisas internas.

* Técnico Integrado em Informática, Universidade Tecnológica Federal do Paraná, Campo Mourão, PR, Brasil; gabidemenjon@gmail.com
† Universidade Tecnológica Federal do Paraná, Campus Campo Mourão (Campo Mourão); andreluis@utfpr.edu.br

Por esse motivo, acredita-se que a aplicação do processo de mineração para analisar grande quantidade de dados abertos, principalmente do Censo de Educação Superior, a fim de descobrir padrões e regras significativas possa gerar conhecimentos úteis e compreensíveis para subsidiar a tomada de decisão aos diversos níveis de gestores públicos.

A Mineração de Dados (ZAKI & MEIRA Jr, 2014) consiste na aplicação de técnicas e de algoritmos com o intuito de extrair padrões em bases de dados, cujo objetivo é descobrir conhecimento útil nestes dados (BRITO et al., 2015). Na Mineração de Dados, paradigmas computacionais convergem e se aglutinam em torno de conceitos como: árvores de decisão, regras de associação, regressão linear, regras de indução, redes neurais artificiais, inferência bayesiana, lógica de programação, entre outros.

Para todo processo de mineração de dados, inicialmente, os dados devem ser compreendidos e preparados para que hipóteses e modelos inteligentes possam ser propostos. Essa primeira etapa certamente é uma das atividades mais trabalhosa de todo o processo. O *dataset* do Censo da Educação Superior consiste de 41 GB de dados em texto puro, semiestruturados e organizados em arquivos csv (do inglês, *comma separeted values*). Esses arquivos apresentam dados com inúmeros problemas de padronização e inconsistências aos longos dos anos. Desta forma, a questão que norteia esse trabalho é: pode-se realizar um estudo que permita eliminar consistências e redundância no *dataset* do Censo da Educação Superior deixando-o apto para ser usado em experimentos de mineração de dados? A fim de resolver esse problema, este artigo apresenta o estudo realizado para a compreensão e preparação dos dados do Censo da Educação Superior aplicados aos anos de 2010 a 2019. Neste trabalho, produziu-se um banco de dados relacional devidamente normalizado e implementou-se a conversão dos dados dos arquivos semiestruturados para essa base de dados normalizada. Com isso, pesquisadores e alunos poderão mais facilmente identificar e selecionar variáveis de interesse para formular suas hipóteses de pesquisa.

Este trabalho esta organizado da seguinte forma. Na Seção 2, descreve-se brevemente o método de mineração de dados adotado e a organização dos dados disponibilizados pelo Inep. A Seção 3 apresenta o modelo de dados de forma simplificada e seus principais elementos. Por fim, na Seção 4, a conclusão e os trabalhos futuros são apresentados.

2 MÉTODO

O CRISP-DM (do inglês, *Cross-Industry Standard Process for Data Mining*) (SHEARER, 2000) é um modelo de processo padrão que descreve abordagens comuns usadas por especialistas em mineração de dados. O CRISP-DM divide o ciclo de vida de um projeto de mineração de dados em seis fases cíclicas e interativas:

- Compreensão do negócio – fase inicial do processo na qual os objetivos e requisitos do projeto a partir de sua perspectiva de negócio são convertidos em um problema de mineração de dados.
- Compreensão dos dados – análise preliminar dos dados em que podem ser identificados problemas na qualidade dos dados e subconjuntos com anomalias que podem ser utilizados para formular hipóteses sobre informações escondidas.
- Preparação de dados – consiste nas ações necessárias para se construir o conjunto final de dados, chamado de entrada.
- Modelagem – os modelos inteligentes são selecionados e aplicados para a visão de dados final.
- Avaliação – os modelos construídos devem ser revisados e avaliados quanto ao atendimento dos objetivos do negócio.
- Implantação – a criação do modelo não é o fim do projeto, o conhecimento obtido precisa ser organizado e apresentado de uma forma que o usuário final possa usufruí-lo.

Este trabalho dedica-se as fases de compreensão e preparação dos dados do Censo da Educação Superior. Esses dados são disponibilizados anualmente em arquivos formatos csv. Há várias dificuldades em manipular esses arquivos: (i) diferentes variáveis coletadas ao longo dos anos com o mesmo significado; (ii) falta de padrão dos valores armazenados; (iii) grande volume de linhas para serem processadas; (iv) informação redundante entre os arquivos e ao longo dos anos.

Junto aos dados, todo ano o Inep publica um Dicionário de Dados que contém as informações coletadas em cada uma das tabelas. Embora o dicionário traga informação histórica de todos os campos, ela é incompleta e, em alguns dados, inconsistente. Os principais problemas são os casos de campos com nomes trocados e os campos faltantes. O dicionário de dado contém também informações do tipo do dado. Entretanto, essa informação é inconsistente ao longo dos anos. Um exemplo é o campo TP_SEXO da tabelas de aluno e docente. Nos anos mais recentes, eles são representados pelos inteiros 1 para feminino e 2 para masculino. Nos primeiros anos analisados, M e F para masculino e feminino, respectivamente. Masculino também é representado com zero e feminino com 1 em outros anos. No processamento desses dados, foi adotado uma padronização, normalmente, seguindo a nomenclatura mais recente.

Há também um grande volume de dados para ser processado. A Tabela 1 apresenta o volume de dados em cada um dos anos analisados. São cerca de 41 GB de dados em 113,8 milhões de linhas.

Tabela 1 – Número de linhas e tamanho em Gigabytes dos arquivos tratados

Ano	Tamanho (GB)	Linhas (milhões)	Ano	Tamanho (GB)	Linhas (milhões)
2010	2,6	8,7	2015	5,8	11,8
2011	3,1	9,4	2016	5,9	12,1
2012	3,9	10,1	2017	2,8	12,3
2013	5,1	10,5	2018	3,2	13,2
2014	5,5	11,4	2019	3,3	14,0

Fonte: Autoria própria (2021).

Neste artigo, optou-se por trabalhar com dados das últimas dez edições publicadas (2010 a 2019). Essa escolha deu-se por dois motivos. O primeiro é que o tempo para processar cada um dos arquivos é relativamente alto. Um exemplo é o processamento da tabela de alunos de 2019 que levou mais do 7 horas de execução. O segundo motivo é a falta de padronização das coletas mais antigas. Fazer esse tratamento tornaria o trabalho ainda mais custoso.

Uma das principais contribuições deste trabalho é a eliminação dos dados redundantes. Todo ano informações estáticas são novamente coletadas e informação redundantes são registradas. Por exemplo, no arquivo que armazena informações sobre os cursos, em cada linha, além da referência a IES que oferta aquele curso, dados adicionais característicos à IES, como categoria administrativa e organização acadêmica, são adicionados de forma redundante, uma vez que essa informação já esta disposta no arquivo de IES.

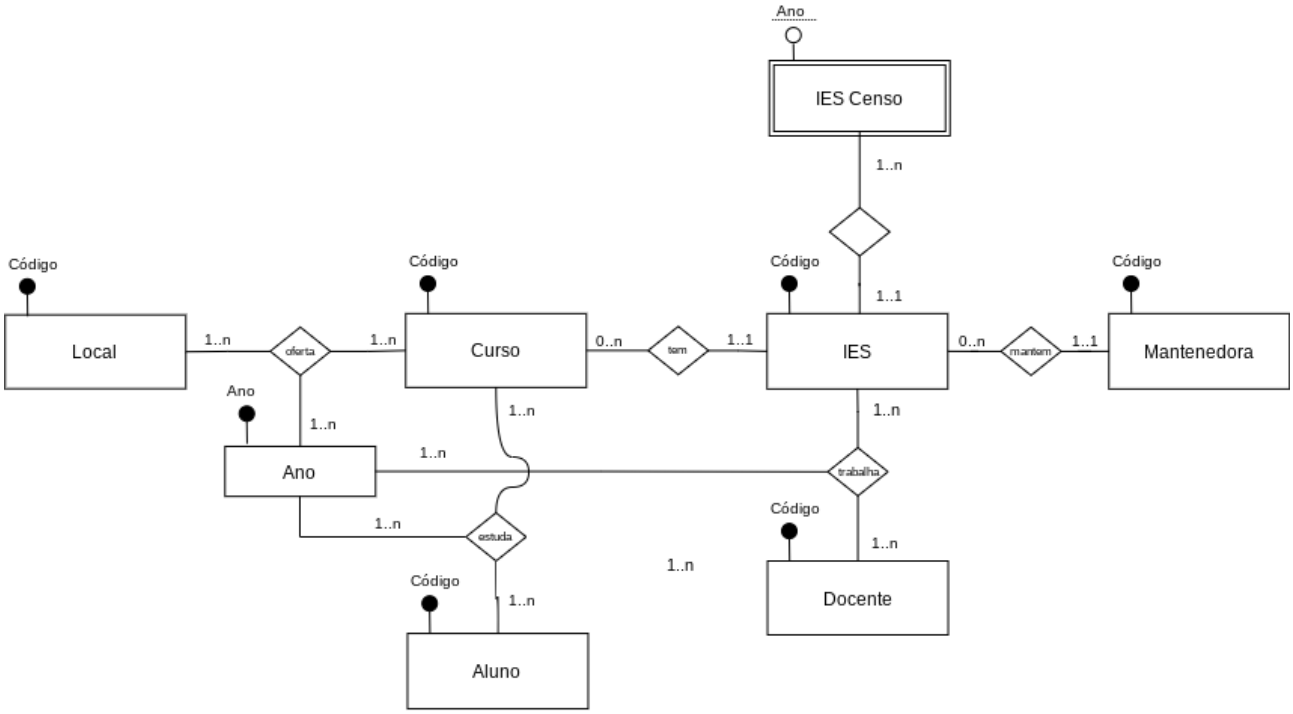
3 RESULTADOS

Nesta seção apresenta-se os artefatos produzidos neste trabalho: (i) o esquema de banco de dados relacional e (ii) o código para conversão dos dados semiestruturados para a base de dados relacional.

O esquema de banco de dados relacional foi proposto para eliminar inconsistências e redundâncias apresentadas nos arquivos semiestruturados dispostos pelo Inep. O Modelo de Entidade-Relacionamento ilustrado na Figura 1 contém de forma resumida as entidades e seus devidos relacionamentos. Vale destacar

que entidades periféricas (por exemplo, municípios, estados, países, etc.), atributos das entidades, e entidades que representam o histórico de evolução de cada uma das entidades foram omitidos do diagrama por uma questão de clareza.

Figura 1 – Modelo de Entidade-Relacionamento para o Censo da Educação Superior



Fonte: Autoria Própria (2021).

A seguir, apresenta-se brevemente cada uma das entidades e os seus relacionamentos.

- Mantenedora – representa as mantedoras das instituições. Mantenedora é a pessoa jurídica de direito público ou privado ou pessoa física que provê os recursos necessários para o funcionamento de outras entidades ou incubadoras.
- IES – representa-se as Instituições de Ensino Superior (IES). Para cada IES tem-se uma mantenedora representada pela entidade Mantenedora.
- IES Censo – representa os dados da IES que são dependentes ao ano que em o censo foi realizado. Por exemplo, quantidade de alunos ativos e gastos com pessoal.
- Local – representa os locais em que os cursos são oferecidos.
- Curso – representa os cursos oferecidos pela IES. Cursos são ofertados em Locais. Para cada ano do censo, novos locais podem ser incorporados a um curso, principalmente, nos cursos de Ensino á Distância.
- Docente – representa as informações coletadas dos docentes nas IES. A cada ano os dados sobre os docentes são atualizados no censo.
- Aluno – representa todos os alunos registrados nos cursos. A anonimização faz com que cada aluno receba um identificador. Infelizmente, não é mantido referência aos identificadores entre os anos de coleta. Isso impede, por exemplo, descobrir o tempo exato que um aluno demorou para conclusão do curso.

Por fim, foi projetado, mas omitido na Figura 1 por uma questão de clareza, uma entidade de histórico para cada uma das entidades IES, Mantenedora, Curso, Docente e Local. Acontece que essas entidades

possuem informações que eventualmente mudam no decorrer dos anos. Por exemplo, uma IES pode mudar de organização acadêmica ou, até mesmo, de nome, mas mantém seu identificar único ao longo dos anos. Para armazenar essa informação, projetou-se então entidades fracas chamadas de histórico em que anota-se os valores anteriores ao valor atual de um determinado campo.

Após projetar o modelo de dados, as tabelas foram implementadas no sistema de banco de dados MySQL (MYSQL, 2021). Para isso, utilizou-se a ferramenta MySQL Workbench (WORKBENCK, 2021) para facilitar a interação com o MySQL. A conversão dos dados foi implementada em scripts na linguagem de programação Python usando as bibliotecas: **pymysql**, para conexão com o banco de dados; e **Pandas**, para manipulação e tratamento dos dados oriundos dos arquivos em formato csv.

A expressiva quantidade de dados dificulta a execução dos scripts em computador convencionais (usando HD convencionais) por causa do tempo de execução. O arquivo completo de alunos de 2019 levou cerca de sete horas para execução. Por causa disso, no momento da escrita deste trabalho não temos a estatística completa dos dados convertidos no banco de dados, pois para os testes, executou-se o código de conversão apenas com amostras aleatórias dos arquivos com maior volume de linhas.

4 CONCLUSÃO

Este artigo apresentou as etapas de compreensão e preparação dos dados do Censo da Educação Superior dos anos de 2010 a 2019 para mineração de dados. Neste trabalho, produziu-se um banco de dados relacional devidamente normalizado e implementou-se a conversão dos dados dos arquivos semiestruturados para essa base de dados normalizada. Como contribuição, pesquisadores e alunos poderão mais facilmente identificar e selecionar variáveis de interesse para formular suas hipóteses de pesquisa e executar algoritmos de mineração de dados. Como trabalho futuro, pretende-se disponibilizar a base de dados com os dados convertidos para acesso remoto em um servidor da UTFPR.

AGRADECIMENTOS

O presente trabalho foi realizado com o apoio da Conselho Nacional de Desenvolvimento Científico e Tecnológico CNPq - Brasil e da Universidade Tecnológica Federal do Paraná/Brasil. A autora principal foi bolsista PIBIC-EM do CNPq.

REFERÊNCIAS

BRITO, Daniel Miranda de; PASCOAL, Túlio Albuquerque; ARAÚJO, Jairo Gustavo G. de O.; LEMOS, Marcílio O.; RÊGO, Thaís Gaudêncio do. Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de data mining. In: XX Congreso Internacional de Informática Educativa. [S.l.: s.n.], 2015. (TISE'15), p. 459–463.

MySQL Community Server. Versão 8.0.26. [S. l.]: Oracle Corporation, 2021. Disponível em: https://dev.mysql.com/downloads/. Acesso em: 10 set. 2021.

SHEARER, Colin. The crisp-dm model: The new blueprint for data mining. Journal of Data Warehousing, v. 5, n. 4, 2000.

WORKBENCK MySQL Workbench. Versão 8.0.26. [S. l.]: Oracle Corporation, 2021. Disponível em: https://dev.mysql.com/downloads/workbench/. Acesso em: 10 set. 2021.

ZAKI, Mohammed J.; MEIRA Jr, Wagner. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.