Disciplina Ciência de Dados Aplicada e Ciência de Dados para Todos

Relatório 3

Autor: João Paulo Tavares Cavalcante Matrícula: 13/0029335

1. Introdução e contextualização

Atualmente a ciência dos dados está presente em tudo o que fazemos, análises de dados são feitas a todo momento sobre qualquer coisa que podemos imaginar. Essas análises revelam muitas coisas sobre os donos desses dados, padrões, gostos, preferências, além da possibilidade de se relacionarem entre si. A responsabilidade gerada em torno desses dados obtidos é grande, conhecer pessoas e seus hábitos podem tanto ajuda-las quanto prejudica-las, dependendo do objetivo.

Quantidades gigantescas de dados são produzidas a todo momento por nós, analisar esses dados geram informações valiosas para diversos interessados, empresas por exemplo podem analisa-los para melhor ofertar determinados produtos e assim ganhar com anúncios, um governo pode analisar os dados de sua população para melhor oferecer melhores serviços, as aplicações são inúmeras.

Nesse relatório o foco será nas aplicações acadêmicas de um determinado tipo de dados referentes a programas de ensino, em especifico ciências do comportamento, uma área da psicologia responsável por análises comportamentais de uma determinada população.

O programa de ensino sobre ciências do comportamento analisado possui cursos de mestrado e doutorado que tem como finalidade promover a base científica e gerar docentes e pesquisadores nesse campo da psicologia. Os cursos possuem duas grandes áreas de concentração, a análise do comportamento e a cognição e neurociências do comportamento.

A área da análise do comportamento tem como objetivo conhecer melhor os processos básicos do comportamento não só humano, mas também animal e investigar os fenômenos associados a esses comportamentos, além de interpretá-los e saber como são influenciados por relações diversas estabelecidas por inúmeras razões naturais.

A área de concentração referente à cognição e neurociências do comportamento tem o seu foco voltado para produzir conhecimento sobre os processos que envolvem a parte cognitiva e da psicobiologia do comportamento humano e animal, mas diferente da análise do comportamento, deseja-se obter dados a respeito dos processos relacionados a memória, percepção, emoção e como o ambiente em que vive afeta esses processos.

Dentro dessas áreas de concentração ainda possuímos duas linhas de pesquisas distintas para cada uma. Em análise do comportamento temos as linhas processos comportamentais básicos e análise comportamental aplicada. Já na área da cognição e neurociência do comportamento temos pesquisas sobre cognição, percepção e sensação e também cérebro, drogas e comportamento.

A linha de pesquisas sobre processos comportamentais básicos busca compreender como as relações funcionais estabelecidas por contingências ambientais diversas afetam os processos básicos do comportamento, isso aplicado tanto a humanos quanto animais.

A segunda linha de pesquisa dentro da mesma área de concentração refere-se a análise comportamental aplicada que tem como objetivo principal investigar e interpretar fenômenos comportamentais que ocorrem em seus ambientes naturais usando metodologias obtidas de pesquisas experimentais no campo de aprendizagem, na mesma linha também procura-se saber como desenvolver e avaliar procedimentos que promovam o estabelecimento e manutenção de repertórios comportamentais eficazes na resolução de problemas em diferentes contextos.

Na outra área de concentração do programa de ensino da pós-graduação temos a primeira linha de pesquisa sendo a cognição, percepção e sensação. “Os objetivos dessa linha de pesquisa consistem em investigar, por meio de métodos experimentais ou observacionais, sob quais condições ou fatores cognitivos, sensoriais e perceptuais influenciam o desempenho humano em sua interação com o ambiente, visando desenvolver modelos teóricos e estratégias de investigação em psicologia e neurociências do comportamento”.

A outra linha de pesquisa dessa área de concentração tem como tema cérebro drogas e comportamento, seu objetivo é avaliar o comportamento humano e como este afeta e é afetado pelo cérebro. Também tem como foco avaliar a dependência de drogas psicotrópicas no comportamento e gerar pesquisas relacionadas a diversos transtornos de memória, ansiedade e estresse.

Os assuntos abortados nesse programa de ensino são de interesse mundial, linhas de pesquisas sobre comportamento podem resolver problemas diversos que há muito estão presentes no cotidiano de uma sociedade como a nossa. Dada a importância do tema, outras universidades do Brasil e também em outros países provavelmente possuem linhas de pesquisa semelhantes em seus programas de ensino contribuindo para a descoberta de melhores soluções para problemas em todo o mundo.

Nesse relatório será abordado justamente a questão da internacionalização do conhecimento já produzido pelos pesquisadores que estão no programa de ensino da pós-graduação em ciências do comportamento oferecida pela UnB. Veremos o número de publicações científicas que foram produzidos em outros países, assim como o Brasil para que seja possível ser feita a comparação.

2. Referencial

A mineração de dados em larga escala vem passando por grandes mudanças nos últimos anos, os profissionais estão melhores capacitados e o conhecimento de análise de dados melhor difundido na sociedade. Quanto maior o entendimento sobre os dados melhor é para o seu aproveitamento prático, mas nem sempre se pode confiar nos dados ou nos resultados que são ditos produzidos por um conjunto de dados analisados.

A validade da análise de dados é primordial para se chegar à uma conclusão plausível onde esses dados podem ser aplicados de forma útil, uma análise que não seja clara pode levar a resultados ilusórios e não efetivos na hora de serem aplicados. Questões éticas sempre envolveram a análise de dados, visto que a divulgação de resultados sobre um determinado assunto pode mudar os dados analisados ou mesmo até a fabricação de resultados enganosos sem fidelidade ao verdadeiro resultado obtido.

Métricas dificilmente são verificadas quando resultados são mostrados em canais de televisão e jornais, mas a demanda por credibilidade vem aumentando nos últimos tempos e o debate sobre métodos mais apropriados e questões de ética sobre o assunto é discutido em vários encontros que abordam o tema.

A cienciometria sofre muito com a manipulação de dados de forma errada, muitos autores defendem a regulamentação de como dados nesse tema devem ser analisados e divulgados, mas outros acreditam que somente isso não resolverá o problema de credibilidade e veracidade dos dados obtidos.

3. Metodologia

A metodologia utilizada nesse relatório leva em consideração o apresentado em aula pelo professor e o que foi passado em relatórios anteriores com o adicional de um novo método de análise de dados, o CRISP-DM. Ciência de dados é um assunto relativamente novo no contexto em que se encontra hoje, por isso não é fácil achar metodologias feitas tendo em mente somente a análise de dados, inúmeras metodologias existem para fabricação de software, para fazer um automóvel, mas para mineração de dados não existia algo parecido antes do CRISP-DM, uma metodologia criada inteiramente com foco em dados.

Para que haja um maior entendimento dos métodos utilizados na análise de um certo conjunto de dados a metodologia utilizada precisa ser replicada e é justamente onde entra o CRISP-DM, afim de padronizar a metodologia utilizada entre diferentes profissionais da área de análise de dados para que assim se obtenha os mesmos resultados com o mesmo conjunto de dados. O CRISP-DM consiste de seis passos simples mostrados a seguir:

Entendimento do negócio: Esta é a etapa em que o profissional deve buscar uma compreensão adequada do problema que necessita ser resolvido. É preciso buscar detalhes sobre como a questão afeta a organização e quais são os principais objetivos e expectativas em relação ao trabalho como um todo. Uma faculdade quer entender por qual motivo seu número de alunos diminuiu 40% em um determinado mês. Muitos cancelaram suas matrículas no final do segundo semestre e a instituição precisa tirar proveito desses dados afim de evitar a repetição dessa evasão.

Compreensão dos dados: Após a primeira etapa o objetivo é inspecionar, organizar e descrever todos os dados disponíveis. É fundamental a avaliação do profissional em busca de quais dados podem ser relevantes para decifrar o problema. Por exemplo, dados de vendas, do desempenho escolar desses alunos, das redes sociais relacionadas à instituição de ensino, de pagamento e faturas atrasadas, dentre outros.

Preparação dos dados: Definidos, organizados e bem inspecionados, nesta etapa o profissional deverá conduzir os dados tecnicamente. É preciso preparar as bases de dados, definir o formato que será necessário para a análise e ajustar demais questões técnicas.

Modelagem: Neste quarto momento, são selecionadas e aplicadas as técnicas de mineração de dados mais apropriadas, dependendo dos objetivos identificados na primeira fase.

Avaliação: Considerada uma etapa de pós trabalho, mas ainda assim extremamente importante para a vitalidade do ciclo, a quinta fase pede o acompanhamento dos resultados objetivos e a avaliação da aplicabilidade confiável dos conhecimentos obtidos.

Desenvolvimento: Todo o conhecimento que for obtido por meio do trabalho de mineração e modelagem agora poderá ser aplicado de forma prática. O ideal aqui é dar uma entrega mais palpável e aplicável ao cliente a partir das análises dos dados feitas pela equipe.

4. Resultados

A contextualização de resultados encontrados pela mineração de dados é um importante procedimento a ser feito para poder dar sentido aos dados, números soltos e sem significado aparente podem apenas passar despercebido ou mesmo dar impressões erradas para quem os vê.

Na busca pela contextualização dos dados, esse relatório teve como foco analisar diferentes conjuntos de dados, mas focar somente nos essenciais na hora de mostrar os resultados, assim não embaralhando números e resultados diferentes do proposto pela análise inicial.

Os resultados obtidos condizem com o esperado após se aplicar a metodologia CRISP-DM com foco na internacionalização das publicações produzidas pelo programa de ensino da pós-graduação em ciências do comportamento da UnB. Esses resultados foram obtidos utilizando-se a linguagem R na plataforma de desenvolvimento RStudio e com o auxílio de diversos pacotes para manipular os dados.

O data-set escolhido foi o de perfis disponibilizado na base da ciência do comportamento e foi analisado utilizando diversas ferramentas disponibilizadas pela linguagem R. Outros data-sets não são o foco, mas também foram analisados na busca de um melhor contexto para ser mostrado na parte dos resultados obtidos.

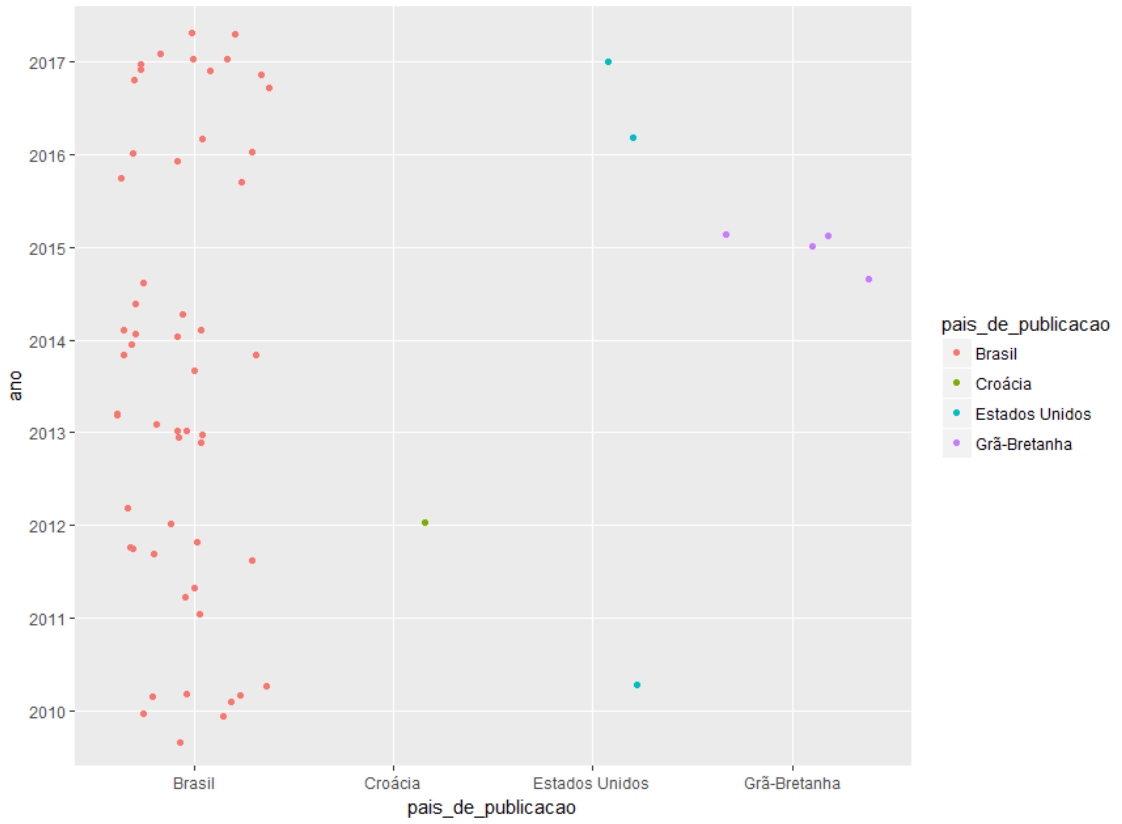
Inicialmente a análise foi feita utilizando as bibliotecas e ferramentas descritas abaixo:

```
library(ggplot2)
library(jsonlite)
library(dplyr)
library(tidyr)
```

Com a finalidade de melhorar a interpretação e manipulação dos dados as bibliotecas acima foram utilizadas.

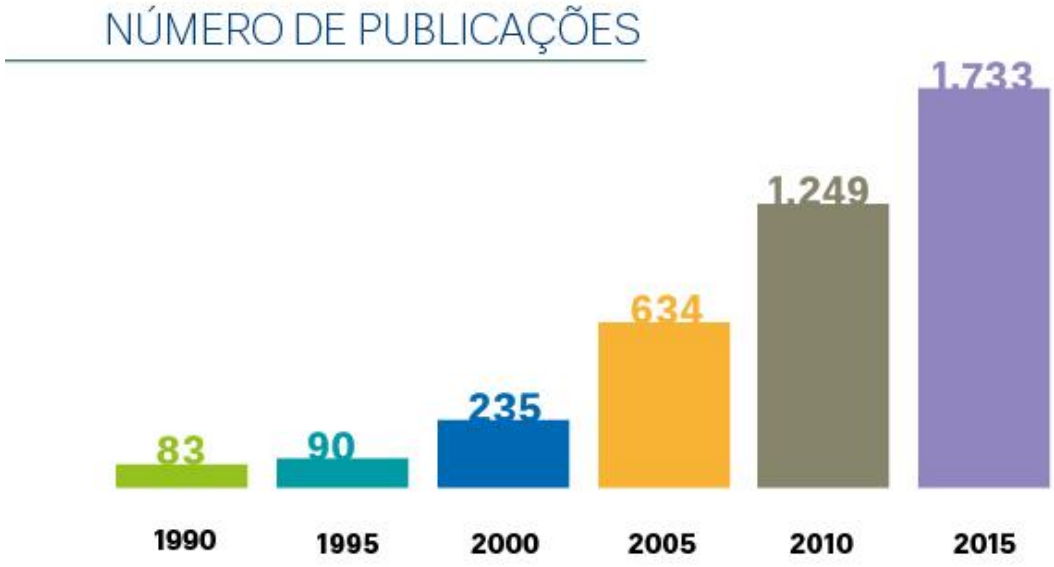
```
comportprof <- fromJSON ("CiComport.profile.json")
PRODBIBLIO <- lapply(comportprof, function(x) x$producao_bibiografica)
PAIS <- lapply(PRODBIBLIO, function(x) x$CAPITULO_DE_LIVRO)
PAIS_bind <- bind_rows(PAIS[[1]], PAIS[[2]], PAIS [[3]], PAIS[[4]], PAIS [[5]], PAIS [[6]], PAIS[[7]], PAIS[[8]], PAIS[[9]
ggplot(PAIS_bind, aes(y = ano, x = pais_de_publicacao, col = pais_de_publicacao)) + geom_jitter()
```

A partir da análise acima obtemos um gráfico que nos mostra as publicações realizadas nos diferentes países, como queríamos saber de início.



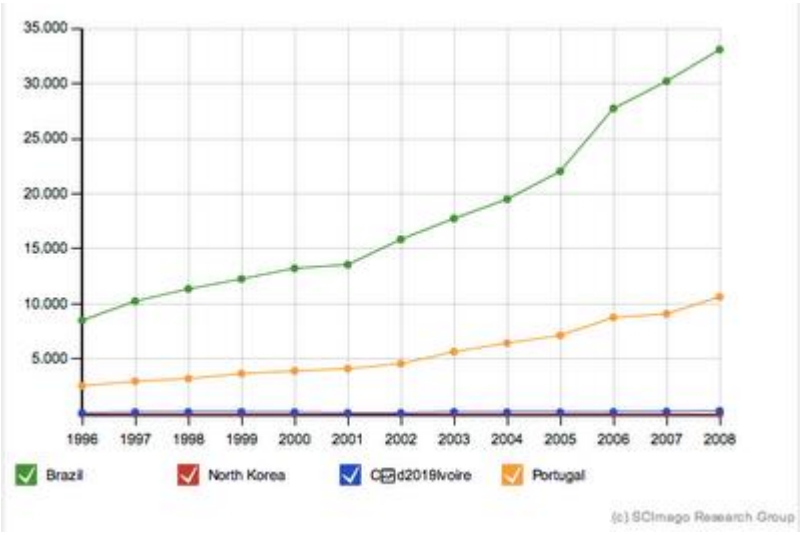
Como esperado, a maioria das publicações são brasileiras, mas temos outras internacionais em três diferentes países. O número de publicações pode parecer pequeno, mas precisamos levar em conta o número de pesquisadores do programa, que são apenas 19 segundo o arquivo de profiles. Temos também um aumento nas publicações desde 2010, o que mostra um maior engajamento nas áreas da ciência do comportamento.

Para obtermos uma melhor comparação do número de publicações, podemos pegar os dados da UnB como um todo:



Temos aqui um aumento que corrobora as informações do primeiro gráfico, também afirmando o maior número de publicações acadêmicas e o seu aumento com o passar dos anos, melhorando a comparação com a base de dados da ciência do comportamento.

Para melhorar a comparação podemos também pegar o histórico de publicações do Brasil e também a comparação com outros países



Nessa imagem temos que o crescimento no número de publicações brasileiras foi muito alto nos últimos anos, dando sustentação para o que já havia sido falado acima a

respeito do crescimento de produções acadêmicas. Mesmo quando comparamos a outros países temos o maior índice de crescimento voltado favoravelmente aos números brasileiros, colocando o Brasil como grande produtor de publicações acadêmicas e também com um grande crescimento.

Referência Bibliográfica

CRISP-DM - http://www.bigdatabusiness.com.br/se-voce-se-interessa-por-big-data-precisa-entender-o-crisp-dm/

De Rijcke, S., & Rushforth, A.D. (2015). To intervene, or not to intervene, is that the question? On the role of scientometrics in research evaluation. Journal of the Association for Information Science and Technology, 66(9), 1954-1958. doi:10.1002/asi.23382.

Linhas de pesquisa - http://www.ppg-cdc.unb.br/linhas-de-pesquisas

Pesquisas UnB - https://www.noticias.unb.br/publicacoes/117-pesquisa/747-publicacoes-cientificas-em-periodicos-internacionais-crescem-38-em-cinco-anos

Produção cientifica brasileira - http://sbi.org.br/copa-do-mundo-de-producao-cientifica/