
Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

# Um Método para Melhoria de Dados Estruturados de Imóveis

Lucas Nunes de Souza

Proposta de Trabalho de Graduação

Orientador: Prof. Luciano de Andrade Barbosa

Recife
Abril de 2017

# Resumo

A Internet contém uma grande quantidade de dados que podem ser usados para diversos propósitos, por exemplo, para modelagem de dados, análises estatísticas e previsões. No entanto, dados obtidos da web podem conter vários problemas como campos faltando, duplicatas ou informações erradas. Esses problemas podem surgir de algum erro nas fontes de dados ou no método de obtenção dos dados. Detectar e tratar esses problemas pode exigir muito trabalho e geralmente requer o uso de diferentes ferramentas e técnicas estatísticas. O principal objetivo desse trabalho é criar uma solução para ajudar o usuário nesses processos, de forma que ele possa utilizar, de forma integrada, diferentes métodos estatísticos e de visualização para ajudá-lo no processo de melhoria da qualidade de dados.

**Palavras-chave:** recuperação de informação, qualidade de dados, web

# Abstract

The Internet contains a large amount of data that can be used for various purposes, for example, for data modeling, statistical analysis and forecasting. However, data obtained from the web may contain various problems such as missing fields, duplicates, or wrong information. These problems may arise from some error in the data sources or in the method of obtaining the data. Detecting and treating these problems can require a lot of work and usually needs the use of different statistical tools and techniques. The main goal of this work is to create a solution to help users in these processes, so that they can use different statistical and visualization methods in an integrated way to help them in the process of data quality improvement.

**Keywords:** information retrieval, data quality, web

# Sumário

CAPÍTULO 1

# Contexto

Na era da internet em que vivemos, entretenimento e informações são predominantemente digitais e a quantidade de dados que são gerados cada dia é enorme. Podemos encontrar todo tipo de informação na Web, desde informações e preços de produtos até dados e artigos científicos. Esses dados podem ser usados, por exemplo, para tarefas de previsões e tomadas de decisão [1].

Com o aumento de automação na forma de adquirir os dados, a preocupação com a qualidade dos dados aumenta. Normalmente é necessário que os dados estejam no correto estado e representem fielmente o mundo real [2]. A presença de dados incorretos ou inconsistentes podem distorcer significativamente resultado de análises, potencialmente negando os benefícios de usar métodos e algoritmos de previsão e análise de dados [3].

No contexto da Web, estamos interessados em dados estruturados, aqueles que possuem atributos e valores de um determinado domínio. Por exemplo, um produto numa loja online pode ter atributos preço, nome, estado, etc. Esses tipos de dados são de mais facil manipulação e análise. Dados extraídos de sites com informação estruturada podem conter problemas como: valores faltando, dados duplicados, erro de formatação, entre outros. As duas principais fontes de erros de dados obtidos da web são:

- Fonte dos dados: Problemas podem sugir tanto na entrada manual de dados, campos errados e faltando, ou na geração automática de páginas, erros de formatação e valores duplicados. Por exemplo, um site de imóveis pode conter valores faltando como vagas ou suites, informações normalmente não obrigatórias. por exemplo,....

- Método extração: Métodos de extração automatizados podem ter problemas para extrair informação devido a diferença de estrutura que cada página pode ter. Até em um único domínio, não pode ser garantido que os dados vão estar no mesmo formato.

Para poder lidar com esses problemas, diversos métodos têm sido propostos na área de qualidade de dados e limpeza de dados. Como exemplo temos outliers nos dados, que podem causar efeitos negativos aos processos que forem eventualmente utilizá-los [4]. O processo de melhoria dos dados possui alguns desafios. Métodos diferentes podem ter resultados diferentes ou até mesmo um mesmo método pode ser melhor para um determinado conjunto de dados que outro. O processo de limpeza também pode criar novos erros [5]. Esse processo pode ser considerado interativo, no qual o usuário pode analisar cada etapa e decidir qual o melhor resultado [6]. Desse modo, pretende-se criar nesse trabalho uma ferramenta de suporte à melhoria na qualidade de dados estruturados.

CAPÍTULO 2

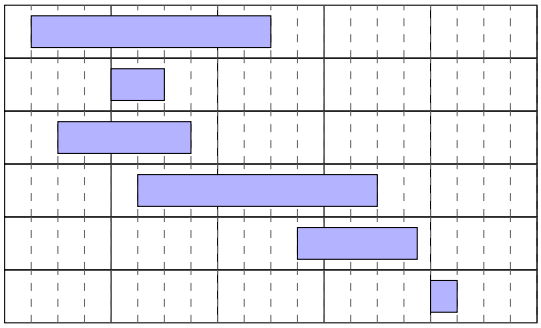
# Objetivo

O objetivo do trabalho é criar uma ferramenta que auxilie o tratamento e estudo de dados adquiridos da web. Para isso usaremos técnicas estatísticas para tratar dos problemas mais comuns encontrados nos dados. A ferramenta permitirá que o usuário escolha quais técnicas serão usadas nos dados e vizualizar o resultado de cada etapa, podendo obter os dados tratados.

CAPÍTULO 3

# Cronograma

| Atividades | Mar | Abr | Mai | Jun | Jul |
|---|---|---|---|---|---|
| Revisão Bibliográfica | | | | | |
| Elaboração da proposta | | | | | |
| Análise dos problemas nos dados | | | | | |
| Desenvolvimento da ferramenta | | | | | |
| Relatório final | | | | | |
| Apresentação | | | | | |



CAPÍTULO 4

# Possíveis Avaliadores

São possíveis avaliadores do trabalho a ser produzido conforme especificado nesta proposta:

- Bernadette Farias Lóscio

CAPÍTULO 5

# Assinaturas

---

Lucas Nunes de Souza
Aluno

---

Luciano de Andrade Barbosa
Orientador

# Referências Bibliográficas


[1] S. K. J. W. Xu Chu, Ihab F. Ilyas, "Data cleaning: Overview and emerging challenges," *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, 2016.

[2] M. v. d. L. Edwin de Jonge, "An introduction to data cleaning with r," *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, 2013.

[3] J. M. Hellerstein, "Quantitative data cleaning for large databases," *United Nations Economic Commission for Europe (UNECE)*, February 2008.

[4] S. Seo, "A review and comparison of methods for detecting outliers in univariate data sets," *Master of Science thesis. University of Pittsburg*, 2006.

[5] D. S. Laure Berti-Équille, Tamraparni Dasu, "Discovery of complex glitch patterns: A novel approach to quantitative data cleaning," *Data Engineering (ICDE), 2011 IEEE 27th International Conference on Data Engineering*, May 2011.

[6] T. Dasu, "Data glitches: Monsters in your data," *Handbook of Data Quality*, p. 163–178, 2013.