UFPE

Big Data Integration
Tópicos Emergentes

Kellyton Brito
kellyton.brito@gmail.com
IN1137 - Integração de Dados
Centro de informática - UFPE
Novembro / 2016

UFPE

Agenda

- Introdução a Big Data
- Big Data Integration: Principais Tópicos
- Big Data Integration: Tópicos Emergentes
  - CrowdSourcing
  - Seleção de Fontes
  - Profiling de Fontes
- Tópico de Discussão:
  - Big data e governos

Introdução a Big Data

"Big Data é o termo utilizado para descrever grandes volumes de dados e que ganha cada vez mais relevância à medida que a sociedade se depara com um aumento sem precedentes no número de informações geradas a cada dia".

IBM (2014)

Introdução a Big Data

- **Volume:** Número (quantidade) de dados manipulados e posteriormente analisados.
- **Velocidade:** Velocidade em que os dados são gerados e armazenados, e idealmente processados.
- **Variedade:** Diversos tipos de dados provenientes de varias fontes, sendo eles estruturados ou não estruturados.
- **Veracidade:** Confiabilidade e precisão dos dados obtidos.
- **Valor:** Vantagem que a análise dos dados podem oferecer

Introdução a Big Data

https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know

Big Data Integration – Principais Tópicos

1. Alinhamento de Esquemas
   – Esquemas distintos, mesmo significado?

2. Resolução de Entidades
   – Quais registros são referentes à mesma entidade (e quais são referentes a entidades distintas)

3. Fusão de Dados
   – Qual valor realmente reflete a realidade?

UFPE

Tópicos Emergentes:

Crowdsourcing
Source Selection
Source Profiling

UFPE

Tópicos Emergentes:

Crowdsourcing

Source Selection

Source Profiling

CrowdSourcing

Definition 5.1 (Crowdsourcing Systems) A system is a *crowdsourcing system* if it enlists a crowd of humans to help solve a problem defined by the system owners, and if in doing so, it addresses the following four fundamental challenges: How to recruit and retain users? What contributions can users make? How to combine user contributions to solve the target problem? How to evaluate users and their contributions?

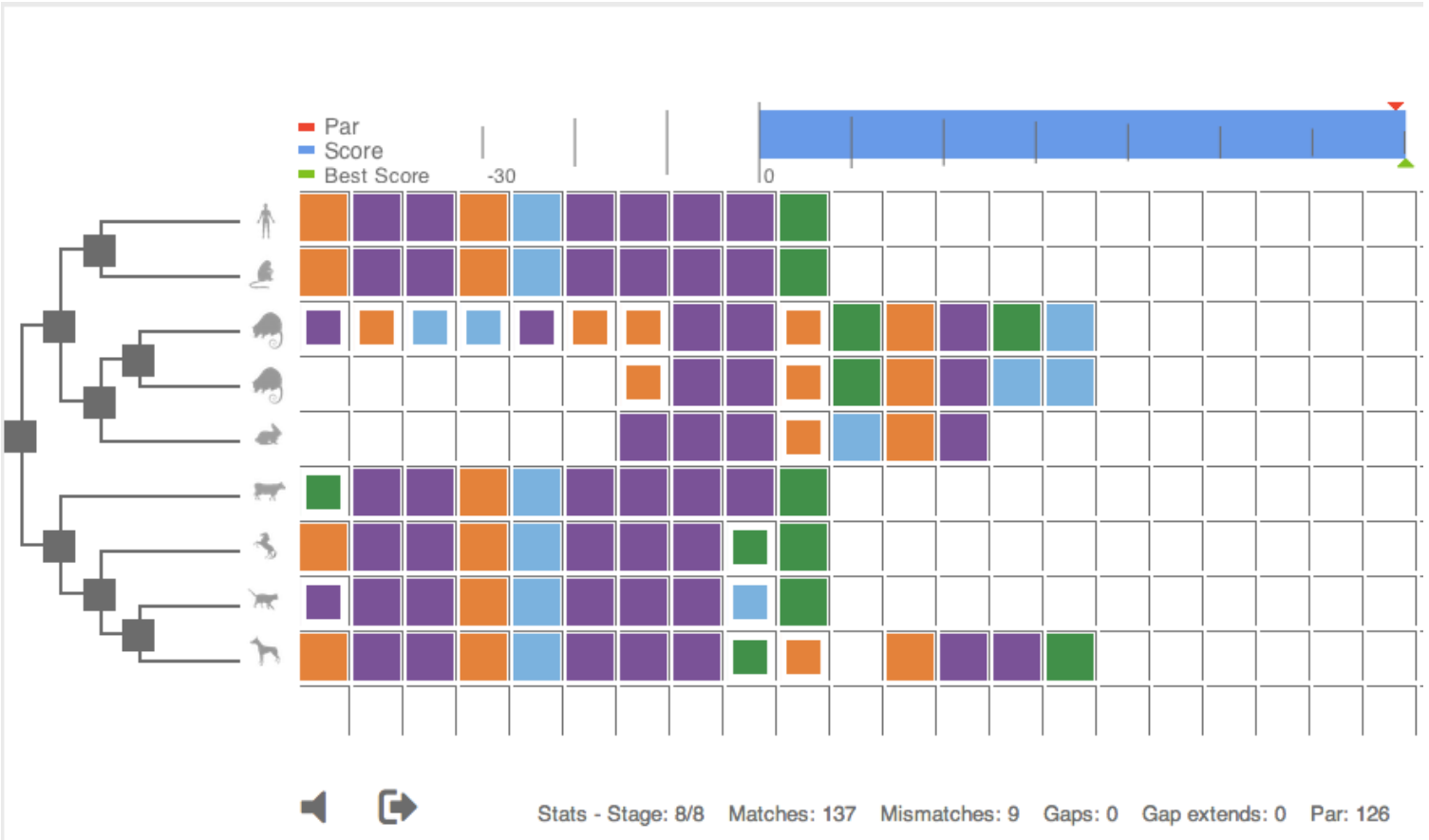
CrowdSourcing

- O que há por trás do reCaptcha?

reCaptcha: Tough on bots Easy on humans

Crowdsourcing

- Sequenciamento de DNA “jogando”
  - http://phylo.cs.mcgill.ca/



Crowdsourcing

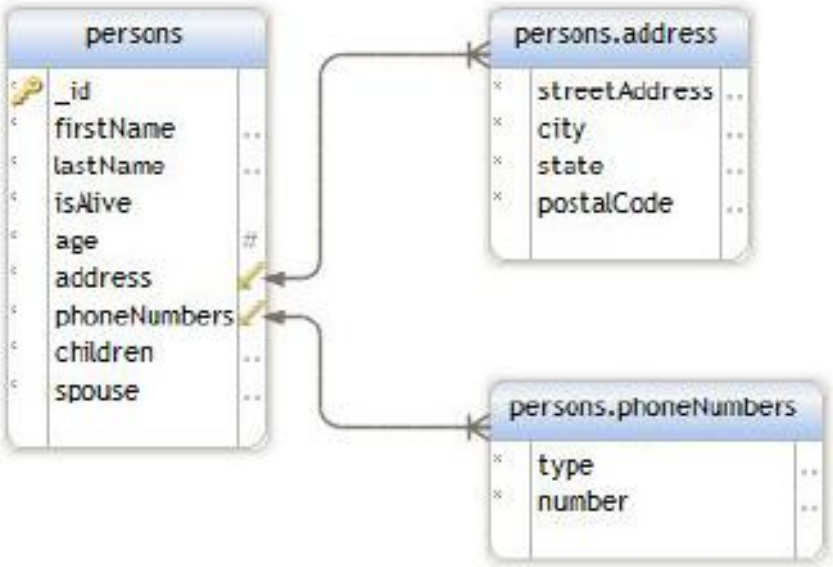
- Sequenciamento de DNA “jogando”
- Pessoas melhores do que o melhor algoritmo:

“The gamers produced roughly 350,000 solutions to various MSA problems, beating the accuracy of alignments from MULTIZ in roughly 70 per cent of the sequences they manipulated”

*http://www.nature.com/news/gamers-outdo-computers-at-matching-up-disease-genes-1.10203

Alinhamento de Esquemas

- **Alinhamento de Esquemas (by Helton)**



- Aborda desafio da ambiguidade semântica

- Visa compreender quais atributos têm o mesmo significado entre os esquemas

Alinhamento de Esquemas

- **Alinhamento de Esquemas (by Helton)**
  - Embora existam abordagens e plataformas que tratam da integração de dados em Big Data, isto ainda não é algo tão trivial de se fazer
  - Ainda existe um amplo espaço para o desenvolvimento e pesquisa de novas tecnologias que facilitem a integração destas fontes de dados de forma mais automatizada

Alinhamento de Esquemas

E se fizermos alinhamento de esquemas baseado em crowdsourcing?

Schema Mapping Suggestions

Threshold

Crowd

Crowdsourcing + Alinhamento de Esquemas

- Questões Importantes
  - Problema:
    - Muitas interações humanas para validar todas as possibilidades
  - Relações transitivas reduzem a quantidade de confirmações
    - Se A e B se referem à mesma entidade **and** B e C se referem à mesma entidade, **then** A e C se referem à mesma entidade
    - Desenvolvimento de estratégias sequenciais e paralelas

Crowdsourcing + Alinhamento de Esquemas

- Desafios do Crowdsourcing ainda existem
  - Como recrutar e manter usuários?
  - Quais contribuições os usuários podem e querem fazer?
  - Como avaliar os usuários e suas contribuições?

UFPE

Tópicos Emergentes:

Crowdsourcing

Source Selection

Source Profiling

Seleção de Fontes

- Vantagens de muitas fontes de dados:
  - Aumento da cobertura
  - Aumento da acurácia
  - Aumento da redundância
  - …

- Essas fontes tem um ***custo***

Balenceamento de custo x benefício é necessário

Seleção de Fontes

- Incluir mais fontes não necessariamente melhora o resultado

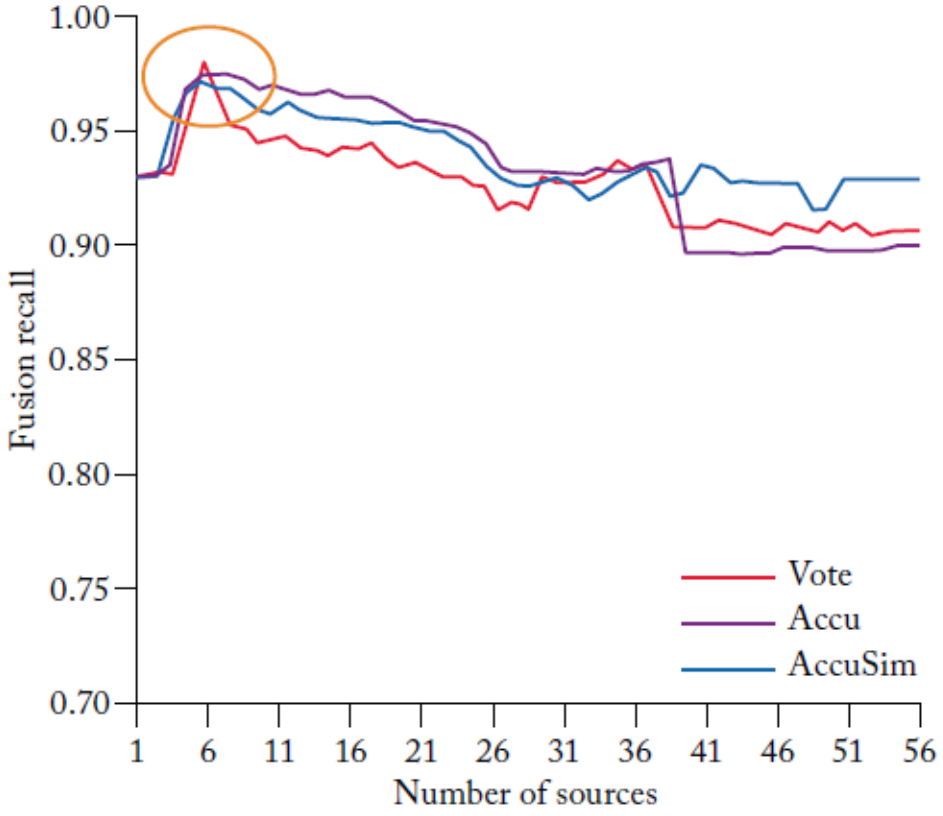


FIGURE 5.2: Fusion result recall for the Stock domain [Li et al. 2012].

Seleção de Fontes

- Baseado em métricas de custo x benefício
- Problema (e solução) de otimização:
  - Dado um custo máximo, qual conjunto de fontes maximiza o resultado
  - Desejando um benefício mínimo de X, qual conjunto de fontes maximiza o resultado?

Seleção de Fontes

- Baseado em métricas de custo x benefício

- Nova abordagem:
  - Continue adicionando fontes até o benefício marginal ser menor que o custo marginal.

ou seja

  - Dado um custo máximo, qual conjunto de fontes provê o maior lucro?

Seleção de Fontes

- Questões importantes:
  - Algoritmo guloso pode não ter bons resultados
    - Fontes marginais podem dar máximos retornos
    - Problema NP-Completo
  - Seleção das fontes ocorre antes da integração real.
    - Como estimar o custo e benefício de integrar cada fonte?

Seleção de Fontes

The work on source selection is still in its infancy

- Desafios
  - Propostas atuais consideram fontes independentes
  - Definição e estudos de métricas eficientes de custo x benefício

UFPE

Tópicos Emergentes:

Crowdsourcing

Source Selection

Source Profiling

Source Profiling

Dada a grande quantidade de fontes, como descobrir as fontes que contém dados realmente relevantes e tem a qualidade suficiente para uma dada tarefa?

Source Profiling

- Estrutura, semântica e conteúdo bem entendidos -> Sem problemas
  - Mas não é o mundo real

Objetivo:

Ajudar os usuários a entender o conteúdo das fontes, antes mesmo da decisão sobre o que será integrado

Source Profiling

- Propostas em 2 etapas:
  1. Relacionar a fonte de dados com fontes de conhecimento (Freebase, Google knowledge graph, Probase, Yago, etc)

  2. Definir a qualidade dos dados baseado em critérios específicos

Source Profiling

- Abordagens atuais tratam basicamente
  - Sumários estatísticos sobre estrutura e conteúdo das fontes
  - Tentativa de dar uma visão geral do conteúdo dos dados
- Basicamente abordagens de reengenharia
- Desafios:
  - Abordagens atuais tratam basicamente fontes relacionais e estáticas

UFPE

Big Data e Governos

Big Data and Government

- Como os governos podem tirar proveito de Big Data
- E o que eles não deveriam fazer?

- Agenda:
  - National Priorities
  - Real-time analysis
  - ICT Big Brothers
    - Privacy?

Relatório e Referências

- Relatório e referências disponível em

http://bit.ly/2gVC81o

UFPE

Big Data Integration
Tópicos Emergentes

Kellyton Brito
kellyton.brito@gmail.com
IN1137 - Integração de Dados
Centro de informática - UFPE
Novembro / 2016