Large-Scale Data Cleaning Using Hadoop

Chen Li
UC Irvine

Joint work with Michael Carey, Alexander Behm, Shengyue Ji, Rares Vernica
Overview

- Importance of information
- Importance of information quality
- Data cleaning
- Large scale
- Hadoop
Data Quality Example 1

<table>
<thead>
<tr>
<th>Star</th>
<th>Title</th>
<th>Year</th>
<th>Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keanu Reeves</td>
<td>The Matrix</td>
<td>1999</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>Samuel Jackson</td>
<td>Iron man</td>
<td>2008</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>Schwarrzeneger</td>
<td>The Terminator</td>
<td>1984</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>Samuel Jackson</td>
<td>The man</td>
<td>2006</td>
<td>Crime</td>
</tr>
</tbody>
</table>
Data Quality Case 2

Large-Scale Data Cleaning Using Hadoop
Outline

- Three data cleaning applications
  - Document Cleaning
  - Fuzzy string search
  - Record linkage (with preliminary results)
- Conclusion
Data Cleaning Application 1

Document Cleaning
Wikipedia: “Wikify” project
Wiki Linking Policies

- “Do not link terms that most readers are familiar with”
- Neither too many nor too few links

How to detect violations automatically?
An example wiki page

St Antony's Forane Church, Kurumpanadam

From Wikipedia, the free encyclopedia

This article is an orphan, as few or no other articles link to it. Please introduce links to this page from other articles related to it. (August 2008)

This article may need to be wikified to meet Wikipedia's quality standards. Please help by adding relevant internal links, or by improving the article's layout. (August 2008)

Kurumpanadam Forane Church is situated on a beautiful hillock about 6 km east of Changanacherry and 800 metres north of Perumbanachi Junction. This area was a part of Changanacherry Parish in ancient days and later from 1836 to 1837 under the jurisdiction of Thuruthi Parish. The Mooppannermar (Senior Fathers) who had visited Kurumpanadu to select aspirants for priesthood selected Chorikkavungal Zackarias and Muckattukunnel Thomas from this area. They requested the Catholics of this area to erect a church. Following their request, 93 Catholic families who had inhabited Kurumpanadu took the initiative and erected the first Catholic Church here on 13 June 1837. It had stood on the north of the present church. As the church was not spacious enough, the construction of a new church was planned. Under the leadership of Rev. Fr. Thuruthi, was the then vicar of Kurumpanadu. The main role in the construction of the new church was played by Rev. Fr. Abraham Marthanan, the first vicar of Kurumpanadu. The construction of the church was completed in 1844.

Local Seminary

There had been a local seminary at Kurumpanadu in 1852. The fathers of the Carmelite Order and the European Missionaries of St. Francis Xavier were ordained by the Bishop of Varapuzha in 1852.

The Vestition of Puthencham Thoran was the munificent and saintly Puthencham Thoran from Edathuva who had founded Francisian munnar sabha in kerala received the religious habit (sabha vastram) in 1868 at Kurumpanadu church from the then vicar Rev. Fr. Palakkunnel Vellikan. This incident is recorded in the church annals as. Puthencham Thoran vowed that he would wear the habit (sabha vastram) of Anichakar Prakoskose’s munnar sabha on 26 december 1868 which was the Feast of St. Esthanapone. With my unworthy hands I enobled him with the habit of Francisian munnar sabha on the steps of Mar Antony’s church at kurumpanadu. It was not the custom of Malayali males to cover the upper part of their body in those days."

St. Peter's English Middle School was established at Kurumpanadu in 1921 under the leadership of Rev. Fr. Jacob Kandankali. Rev. Fr. George Muckattukunnel (Junior) who was a seminarian in those days took much effort in the establishment of the school. H.E Archbishop Mar Joseph Powathil was a former student of this school. This Middle School was upgraded

http://en.wikipedia.org/wiki/St_Antony%27s_Forane_Church_,_Kurumpanadam

Large-Scale Data Cleaning Using Hadoop
Water quality

At the lake's south-east side it connects with Round Lake at the SR 500 bridge. The 43,000 acres (174 km²) around these lakes collect precipitation and form the Lacamas Creek which feeds into both lakes. This area is mostly large open fields and private farm or ranch land. The resulting runoff contains high levels of fertilizers which leads to nutrient abundance in both lakes. This nutrient abundance has caused plant overgrowth and algae problems. Consequently only a few fish species are able to live in either lake (brown trout, rainbow trout, largemouth bass, bluegill, crappie, yellow perch, catfish and sturgeon). The government continues to help with this issue.

Too many trivial links!

http://en.wikipedia.org/wiki/Lacamas_Lake
Challenge

- Corpus-wide analysis of links and entities
- Implicit “join” operations on page entities (even fuzzy)
- Scale:
  - millions of pages (> 3M for English)
  - # of entities/links could be even larger (even more)
- Plan: Hadoop
Data Cleaning Application 2

Fuzzy string search
Example: Web Search

Errors in queries

Errors in data

Bring query and meaningful results closer together

http://www.google.com/jobs/britney.html
Fuzzy String Search: Formulation

Find strings similar to a given string

Functions: edit distance, jaccard, cosine, etc.
Performance is important!
-10 ms: 100 queries per second (QPS)
- 5 ms: 200 QPS
“q-grams” of strings

universal

2-grams
Similar strings have many common grams
q-gram inverted lists

<table>
<thead>
<tr>
<th>id</th>
<th>strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>rich</td>
</tr>
<tr>
<td>1</td>
<td>stick</td>
</tr>
<tr>
<td>2</td>
<td>stich</td>
</tr>
<tr>
<td>3</td>
<td>stuck</td>
</tr>
<tr>
<td>4</td>
<td>static</td>
</tr>
</tbody>
</table>

2-grams

- at → 4
- ch → 0 → 2
- ck → 1 → 3
- ic → 0 → 1 → 2 → 4
- ri → 0
- st → 1 → 2 → 3 → 4
- ta → 4
- ti → 1 → 2 → 4
- tu → 3
- uc → 3
T-occurrence Problem

Find elements whose occurrences $\geq T$
Optimizing inverted index

- Compressing inverted lists
- Using variable-length grams
- All these techniques need a cost-based analysis on the effects of grams on query performance
- Computationally expensive
Challenge: Scale!

- **PubMed publications**: 18 million records
- **GeneBank**: 108 million sequences
- **Google frequent Web word tokens**: 1 trillion

Parallel computing (Hadoop) to the rescue!
Data Cleaning Application 3

Record Linkage

with preliminary results
Record Linkage

<table>
<thead>
<tr>
<th>Phone</th>
<th>Age</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>Brad Pitt</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>Arnold Schwarzenegger</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>George Bush</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>Angelina Jolie</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>Forrest Whittaker</td>
</tr>
</tbody>
</table>

No exact match!

<table>
<thead>
<tr>
<th>Name</th>
<th>Hobbies</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brad Pitt</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forest Whittacker</td>
<td></td>
<td></td>
</tr>
<tr>
<td>George Bush</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angelina Jolie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arnold Schwarzenegger</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fuzzy Join

\[ \text{dist}(T1.A, T2.A) \leq k \]

Self-Join: \( T1 = T2 \)
Two Critical Steps

Step 1: Sorting grams based on frequencies
Step 2: Computing record ID pairs with common grams
Stage 1 – Phase 1

Large-Scale Data Cleaning Using Hadoop
Stage 1 – Phase 2

Large-Scale Data Cleaning Using Hadoop
Step1 – Analysis

- Limitations:
  - Second phase is just to sort the tokens
  - Using one Reduce (not parallelizable)

- A possible optimization:
  - Eliminate the second MR phase
  - Sort in the Reduce of Phase 1

- Result: 10M records, 10 nodes,
  - 2-MR algorithm: 81 seconds
  - 1-MR: 82 seconds
Step 1: Lesson Learned

- Extra steps are needed to get right key-value pairs

- 1-MR solution might not be faster than 2-MR solution
Stage 2 – Phase 1

Large-Scale Data Cleaning Using Hadoop
Stage 2 – Phase 2

Large-Scale Data Cleaning Using Hadoop
Experiments

![Graph showing time (seconds) vs. dataset size (times DBLP) for different algorithms]

- Alg 1
- Alg 2
- Alg 3
Outline

- Three data cleaning applications
  - Document Cleaning
  - Fuzzy string search
  - Record linkage (with preliminary results)

- Conclusion
UCI ASTERIX Project

- Managing large amounts of semi-structured data using parallel computing
- $2.7M from NSF (thanks, Jim!)
- Other campuses: UCR and UCSD