Big data cleansing based on qualitative attributes

Mika Huttunen
Helsinki University

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings. It is an *alternate* style which produces a *tighter-looking* paper and was designed in response to concerns expressed, by authors, over page-budgets. It complements the document *Author's (Alternate) Guide to Preparing ACM SIG Proceedings Using LaTeX2$_\epsilon$ and BibTeX*. This source file has been written with the intention of being compiled under LaTeX2$_\epsilon$ and BibTeX.

Keywords

big data, data cleansing, data analysis, data transformation

1. INTRODUCTION

Nowadays companies are gathering large amounts of data, and its becoming easier to use data to help their businesses in all kinds of decision making, analytics, or for example automating human-involved tasks. Data can be collected in different ways like via user input and, all kinds of sensors. Humans often make data input errors via misspelling, and sensors may grab unwanted noise along the data they're designed to catch. Not to mention that today, the variety of data is also large which leads into collecting data of different formats together. Using *dirty*, or invalid data can lead into making incorrect decisions. These can sometimes cause serious issues - especially in health care and financing sides [2].

Data cleansing is the solution to the above-mentioned problem. It's a process that contains steps for *detecting* possible *errors* in data, and generating data *transformations* to repair them. Data cleansing can be done either based on *qualitative* or *quantitative* techniques.

Quantitative techniques focus on error detection and correction based on numerical attributes of data [1]. If for example some data is visualised in two-dimensional space, it's fairly easy to detect *outliers*, and thus possible errors in the data. Outlier detection is an example of quantitative data *analysis*. My seminar report focuses on the latter, qualitative techniques of data cleansing, and their scalability to big data.

Qualitative techniques are used to detect and correct data errors based on *integrity constraints*, or *patterns* in the data [2]. In an *SQL* database, an example of an integrity constraint could be such that each person should have unique social security number. A pattern on the other hand could be such that for each person, their ZIP code would define the city they live in. If we have several people with ZIP code being 00100, and city as Helsinki, and yet another person with same ZIP code, but city being Turku, the latter mentioned person should be considered erroneous. Some data transformation should be applied for this latter mentioned person like changing its ZIP code, or city or possibly even removing the whole person from the dataset.

The rest of the report is will be divided as follows. In section 2, I will summarise, where we are with big data cleansing at the moment. I'll introduce several big data cleansing techniques in more detail. The techniques can practically be divided into *de-duplication* methods, *sampling*, *incremental cleansing*, and *distributed cleansing*. I'll also discuss what kinds of problems are yet to be solved.

In section 3, I will go through actual implementations for big data cleansing techniques. Section 4 discusses weak points in the papers, and in section 5, I show my own ideas to the research problem. Finally section 6 summarises the whole report.

2. LITERATURE REVIEW

- Määrittele ensin dirty / low quality data
- Seuraavaksi hyvä selittää, mistä qualitative data analysisissä ja sen eri tekniikoista on kyse

3. GENERAL DATA CLEANSING PROCESS

This section discusses background related to data cleansing. It introduces a general process for data cleansing containing data analysis and data transformation steps.

It also introduces specific data quality related problems and different kinds of methods and frameworks for handling them.

Quantitative Data Cleaning for Large Databases

- Quanttitative data (outlier detection) - Categorical data (same thing mentioned with a different name, misspellings)

An Efficient Data Cleaning Algorithm Based on Attributes Selection

- SNM & MPN algorithms for duplicate records detection
- Improved algorithm for duplicate error detection - O(n log

n) - not suitable for big data?

Qualitative Data Cleaning

- What kind of errors, how and where to detect them - Data repairing – Trusting integrity constraints / rules, data or both? – Automatic / human guided (training ML model, suggesting fixes etc.) – Repairing on place / generating a model for repairing

From the actual paper / book:

- "Minimal cost for repair"

3.1 - WHAT TO REPAIR

- Trusting integrity constraints – NADEEF is a holistic repairing algorithm based on user-defined rules (p. 337)

- Trusting data – No examples necessary to be given rather than mentioning this kind of rules cleansing approach

- Trusting (= "changing") both data and constraints:

3.2 - HOW TO REPAIR

3.2.1 - Automatic Repairing

- Cardinality- / Cost-minimal repair – Algorithm 8 for automatic repair procedure - Unverified fixes, may introduce new errors during the process

3.2.2 - Human guided repair

- Guided Data Repair uses ML classifier - KATARA - Data Tamer (very tied to ML)

3.3 - WHERE TO REPAIR

One-shot cleaning

- "Most of the proposed data repairing techniques (*all discussed so far*) identify errors in the data, and find a unique fix of the data either by minimally modifying the data according to a cost function or by using human guidance (Figure 3.17(a))."

Probabilistic Cleaning - Probabilistic deduplication (probability based duplication "removal")

4. BIG DATA CLEANING - Pyrkimys vähentää tarvittavaa ihmiskommunikaatiota - Mahd. vain datan pienen osajoukon käsittely ja tämän perusteella todennäköisyyksiin perustuvia vastauksia / jatkokäsittelyä

- Deduplicating – Blocking, windowing, canopy clustering

- Sampling (SampleClean)

- Incremental data cleaning – Entity resolution (ER) algorithm

- Distributed data cleaning – MapReduce, Spark, Dedoop (with ER) – BigDansing (runs on top of a data processing platform like DBMS or MapReduce)

- Problems with data partitioned across multiple sites (p. 379) – Objective to minimize data shipment cost

5. CONCLUSION

- "Scalability. Large volumes of data render most current techniques unusable in real settings. "

3.1 Type Changes and *Special* Characters

We have already seen several typeface changes in this sample. You can indicate italicized words or phrases in your text with the command `\textit`.

4. ISSUES WITH BIG DATA

This section discusses problems that arise when data cleansing is done with big data.

5. BIG DATA CLEANSING FRAMEWORKS

This section discusses tools and frameworks designed particularly for handing big data. There are BigDansing and KATARA, but what else is available?

Table 1: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
π	1 in 5	Common in math
$	4 in 5	Used in business
Ψ_1^2	1 in 40,000	Unexplained usage

6. WEAK POINTS ON PAPERS

This section discusses weak points of the papers related to the papers current ongoing situation with big data cleansing.

7. OWN IDEAS

Give new ideas/algorithms/experiments on this research problem. This part is very important, because it shows the potential of the author to be an independent innovative researcher. This section can be as long as possible.

Come up with a new idea / algorithm that could be used for dealing with big data and does something better than the existing tools available...

8. CONCLUSION

Summarize the research problem and the main contributions of previous papers. The main weakness of previous works could be also mentioned here. Some future works can be described as well.

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

9. REFERENCES

[1] J. M. Hellerstein. Quantitative data cleaning for large databases, 2008.

[2] I. F. Ilyas and X. Chu. Trends in cleaning relational data: Consistency and deduplication. *Found. Trends databases*, 5(4):281–393, Oct. 2015.

10. EXAMPLES

end the environment with table*, NOTE not table!

As was the case with tables, you may want a figure that spans two columns. To do this, and still to ensure proper "floating" placement of tables, use the environment **figure*** to enclose the figure and its caption. and don't forget to end the environment with figure*, not figure!

THEOREM 1. *Let f be continuous on $[a, b]$. If G is an antiderivative for f on $[a, b]$, then*

$$\int_a^b f(t)dt = G(b) - G(a).$$

Definition 1. If z is irrational, then by e^z we mean the unique number which has logarithm z:

$$\log e^z = z$$

Table 2: Some Typical Commands

Command	A Number	Comments
`\alignauthor`	100	Author alignment
`\numberofauthors`	200	Author enumeration
`\table`	300	For tables
`\table*`	400	For wider tables

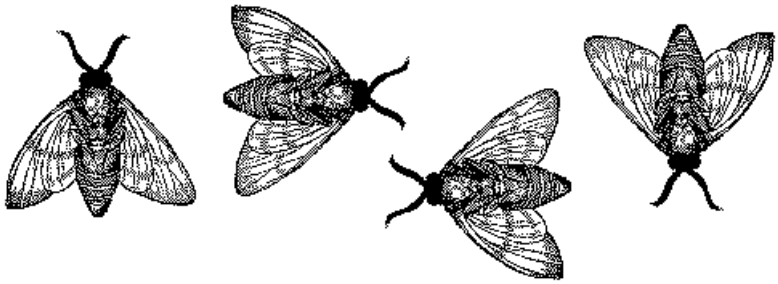

Figure 1: A sample black and white graphic that needs to span two columns of text.

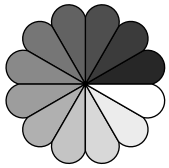

Figure 2: A sample black and white graphic that has been resized with the `includegraphics` command.

PROOF. Suppose on the contrary there exists a real number L such that

$$\lim_{x\to\infty} \frac{f(x)}{g(x)} = L.$$

Then

$$l = \lim_{x\to c} f(x) = \lim_{x\to c} \left[gx \cdot \frac{f(x)}{g(x)} \right] = \lim_{x\to c} g(x) \cdot \lim_{x\to c} \frac{f(x)}{g(x)} = 0 \cdot L = 0,$$

which contradicts our assumption that $l \neq 0$. □