Big Data and Advanced Analytics

Improving Data Quality For Big Data Using Advanced Analytics

Aytac Ozkan

Supervisor: Prof.Rachid Chelouah

Department of Engineering
Ecole internationale des sciences du traitement de l'information

This dissertation is submitted for the degree of
Master of Big Data

EISTI July 2019

Abstract

Digital data play a crucial role in the information and communication technology (ICT) society: they are managed by business and govermental applications, by all kind of applications on the Web, and are fundamental in all relationships between goverments, business, and citizens.

Furthermore, quality of data is also a significant ussue for operational process of business and organizations. Some disasters are due to the presence of data quality problems, among them the use of inaccurate, incomplete, out-of-data.

As a consequence, the overall quality of the information that flows between information systems may rapidly degrade over time if both process and their inputs are not themselves subject to quality control. On the other hand, the same networked information system offers new opportunites for data quality management, including possibility of selectingsources with better quality data, and of comparing sources for the purpose of error localization and correction, thus facilitating the control and improvement of data quality in the system.

Due to the described above motivations, researchers and organizations more and more need to understand and solve data quality problems, and thus answering the following questions: What is in essence, data quality? Which teqniques, methodologies, and data quality issues are at a consolidated stage?

In this paper, we first review relevant works and discuss machine learning techniques, tools and statistical models. Second, we offer a creative data profiling frework based deep learning and statistical model algortihms for improving data quality.

Table of contents

List of figures

List of tables

Chapter 1

Introduction to Data Quality

A Web search of terms "data quality" through the search engine Google, returns about three millions of pages and indicator that data quality issues are real and increasingly important (the term data quality will be shortened to the acronym DQ)

1.1 Why Data Quality is Relevant

The consequences of poor quality of data are often experienced in everyday life, but often, without making the necessary connections to their causes.

For example, the late or mistaken delivery of a letter is often blamed on a postal service, although a closer look often reveals data-related causes, typically an error in the address, orginating in the address database.

Data quality has serious consequences of far-reaching significance, for the efficiency and effectiveness of organizations and business.

1.2 Introduction to the Concept of Data Quality

From a research perspective, data quality has been addressed in different areas, including statistics, management, and computer science. Statisticians were the first to investigate some of the problems related to data quality, by proposing a mathematical theory for considering duplicates in statistical data sets, in the late 1960's. They were followed by researchers in management, who at the beginning of the 1980's focused on how to control data manufacturing systems in order to detect and eliminate data quality problems. Only at the beginning of the 1990's computer scientists begin considering the problem of defining, measuring, and improving the quality of electronic data stored in databases, data warehouses, and legacy systems. [**?**]

Dr. Genichi Taguchi [**?**], who was a world-renowned quality engineering expert from Japan, emphasized and established the relationship between poor quality and overall loss. Dr. Taguchi (1987) used a quality loss function (QLF) to measure the loss associated with quality characteristics or parameters. The QLF describes the losses that a system suffers from an adjustable characteristic. According to the QLF, the loss increases as the characteristic y (such as thickness or strength) gets further from the target value (m). In other words, there is a loss associated if the quality characteristic diverges from the target. Taguchi regards this loss as a loss to society, and somebody must pay for this loss. The results of such losses include system breakdowns, company failures, company bankruptcies, and so forth.

Figure 1.1 shows how the loss arising from varying (on either side) from the target by Δ_0 increases and is given by $L(y)$ when y is equal to m,

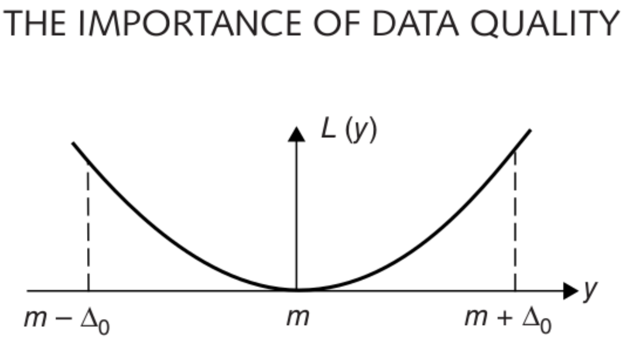

Fig. 1.1 Quality Loss Function (QLF)

the loss is zero, or at the minimum. The equation for the loss function can be expressed as follows:

$$L(y) = k(y - m)^2$$

where k is a factor that is expressed in dollars, based on direct costs, indirect costs, warranty costs, reputational costs, loss due to lost customers, and costs associated with rework and rejection. There are prescribed ways to determine the value of k. The loss function is usually not symmetrical-sometimes it is steep on one side or on both sides. Deming [**?**] says that the loss function need not be exact and that it is difficult to obtain the exact function. As most cost calculations are based on estimations or predictions, an approximate function is sufficient-that is, close approximation is good enough.

The concept of the loss function aptly applies in the DQ context, especially when we are measuring data quality associated with various data elements such as customer IDs, social security numbers, and account balances. Usually, the data elements are prioritized based on certain criteria, and the quality levels for data elements are measured in terms of percentages (of accuracy, completeness, etc.). The prioritized data elements are referred to as critical data elements (CDEs).

1.3 Data Quality and Types of Information Systems

Data are collected, stored, elaborated, retrieved, and exchanged in information systems used in organizations to provide services to business processes. Different criteria can be adopted for classifying the different types of information systems, and their corresponding architectures; they are usually related to the overall organizational model adopted by the organization or the set of the organizations that make use of the information system.

The three classifications are represented together in the classification space of Figure 1.2. Among all possible combinations, five main types of information systems are highlighted in the figure: Monolithic, Distributed, Data Warehouses,Cooperative, and Peer-to-Peer.

- In a *monolithic information system* presentation, application logic, and data management are merged into a single computational node. Many monolithic information systems are still in use. While being extremely rigid, they provide advantages to organizations, such as reduced costs due to homogenetiy of solutions and centralization of management. In monolithic systems data flows have a common format, and data quality control is facilitated by the homogenetiy and centralization of procedures and management rules.

- A *data warehouse* (DW) is a centralized set of data collected from different sources, designed to support management decision making. THe most critical problem in DW design concerns the cleaning an integration of the different data sources that are loaded into the DW, in that much of the implementation budget is spent on data cleaning activities.

- A *distributed information system* relaxes the rigid centralization of monolithic systems, in that it allows the distribution of resources and applications across network of geographically distributed systems. The network can be organized in terms of several tiers, each made of one or more computational nodes. Presentation, application logic, and data management are distributed across tiers. Usually, the different tiers and nodes have a limited degree of autonomy, data design is usually performed centrally, but to a certain extent some degree of heterogenetiy can occur, due to the impossibility of establishing unified procedures. Problems of data management are more complex than in monolithic systems, due to the reduced level of centralization.

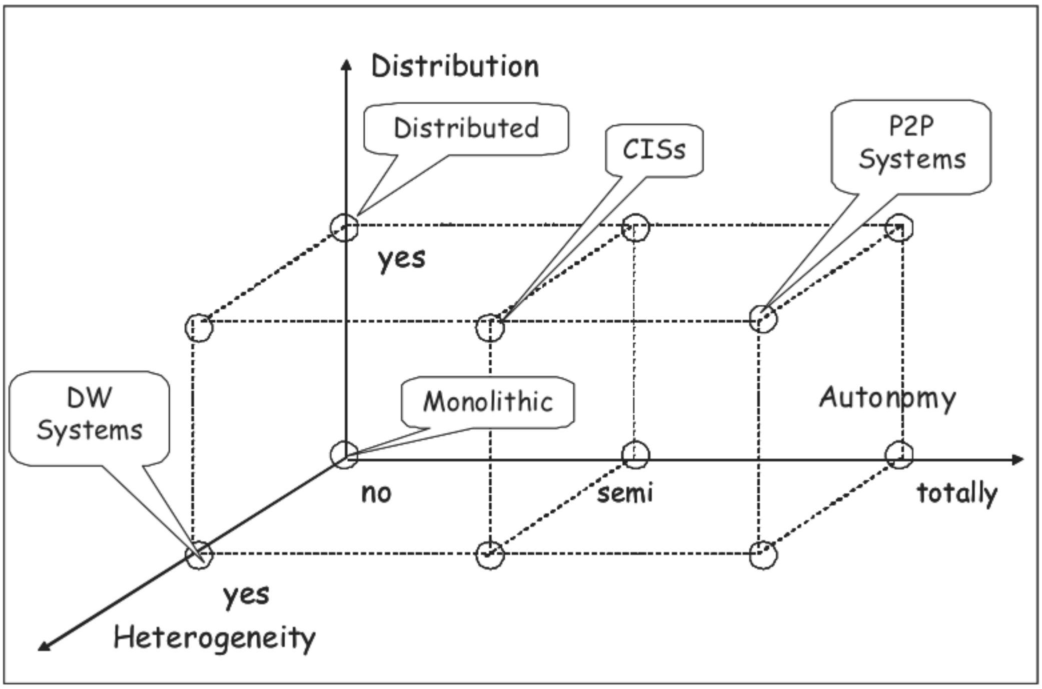

Fig. 1.2 Types of information systems

- A *coorparative information system* (CIS) can be defined as a large-scale information system that interconnects various systems of different and autonomous organizations, while sharing common objectives.

- In a *peer to peer information system* (usually abbreviated P2P), the traditional distinction between clients and servers typical of distributed systems is disappearing. A P2P system can be characterized by a number of properties: peers are highly autonomous and higly heterogeneous, they have no obligation for the quality of their services and data, no central coordination and no central database exist, no peer has a global view of the system, global behavior emerges from local intreactions.

1.4 Main Research Issues and Application Domains in Data Quality

Due to the relevance if data quality, its nature, and the variety of data types and information systems, achieving data quality is a complex, multidisciplinary area of investigation. I involves several research topics and real-life application areas

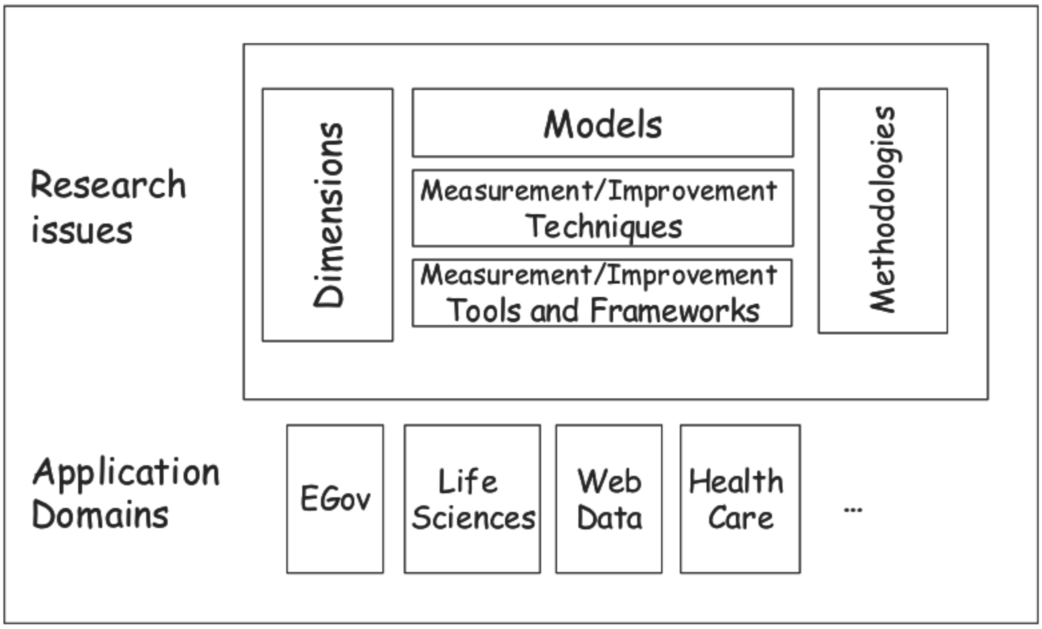

Fig. 1.3 Main issues in data quality

Chapter 2

Data Quality Dimensions

More specifically, quality dimensions can refer either to the extension of data, i.e., to data values, or to heir intension, i.e., to their schema. Both data dimensions are usually defined in qualitative way, referring to general properties of data and schemas, and and related definitions do not provide any facility for assigning values to dimensions themselves.

Specifically, definitions do not provide quantitative measures, and one or more metrics are to be associated with dimensions as separate, distinct properties. For each metric, one or more measurement methods are to be provided regarding (i) where the measurement is taken, (ii) what data are included, (iii) the measurement device, and (iv) the scale on which results are reported.

According to the literature, at times we will distinguish between dimensions and metrics, while other times we will directly metrics.

Table 2.1 Data Quality Dimensions I

Dimension Name	**Description**
Data Governance	Do organization-wide data standards exist and are they enforced? Do clearly defined roles and responsibilities exist for data quality related activities? Does data governance strive to acquire and maintain high-quality data through proactive management?
Data Specifications	Are data standards documented in the form of a data dictionary, data models, meta data, and integrity constraints?
Data Integrity	How is data integrity maintained? How are data integrity violations detected and resolved ?
Data Consistency	If data redundancy exists, how is data consistency achieved? What methods are used to bring consistency to data that has become inconsistent? If data is geographically replicated, how is the consistency and latency managed?
Data Currency	Is the data current? Do procedures exists to keep the data current and purge stale data?
Data Duplication	Are there effective procedures in place to detect and remove duplicate data?
Data Completeness	Is the data about entities complete? How is missing data managed?
Data Provenance	Is a historical record of data and its origination maintained? If the data is acquired through multiple sources and has undergone cleaning and transformations, does the organization maintain a history of all changes to the data?
Data Heterogeneity	If multi-modality data about an entity is available, is that data captured and used?
Streaming Data	How is streaming data sampled, filtered, stored, and managed for both real-time and batch processing?
Outliers	How are outliers detected and addressed? Are there versions of datasets that are outlier-free? Does each version correspond to a different method for outlier detection and treatment?
Dimensionality	Reduction Do the datasets feature dimensionality reduced versions? How many versions are available?
Feature Selection	Do datasets have versions that exclude features that are either redundant, highly correlated, or irrelevant? How many versions are available?

Table 2.2 Data Quality Dimensions II

Dimension Name	Description
Feature Extraction	Do the datasets provide a set of derived features that are informative and non- redundant, in addition to the original set of variables/features? How many such derived feature sets are available?
Business Rules	Does a process exist to identify, refine, consolidate, and maintain business rules that pertain to data quality? Do rules exist to govern data cleaning and transformations, and integrating related data of an entity from multiple sources? What business rules govern substitutions for missing data, deleting duplicate data, and archiving historical data? Are there rules for internal data audit and regulatory compliance?
Data Accuracy	Data can be syntactically accurate and yet semantically inaccurate. For example, a customer's mailing address may meet all the syntactic patterns specified by the postal service, yet it can be inaccurate. How does the organization establish the accuracy of data?
Gender Bias	Is the data free from factors that lead to gender bias in machine learning algorithms?
Confidentiality and Privacy	Are procedures and controls implemented for data encryption, data de- identification and re-identification, and differential privacy?
Availability and Access Controls	How is high data availability achieved? What security controls are implemented to protect data from unauthorized access? How are user entitlements to data access and modifications defined and implemented?

2.1 Accuracy

Accuracy [**?**] is defined as the closeness between a value v and a value v' , considered as the correct representation of the real-life phenomenon that v aims to represent. As an example if the name of a person is John, the value v' = Aytaç is correct, while the value v = Ayt is incorrect. Two kinds of accuracy can be identified, namely a syntactic accuracy and a semantic accuracy.

Let us consider a relation schema **R** consisting of **K** attributes and a relational table **r** consisting of N tuples.

Let $q_{ij}(i = 1..N, j = 1..K)$ be a boolean variable defined to correspond to the cell values y_{ij}, is syntactically accurate, while otherwise it is equal to 1.

In order to identify whether or not accuracy errors affect a matching of relational table **r** with a reference table **r'** containing correct values, we introduce a further boolean variable s_i equal to 0 if the tuple t_i matches a tuple in **r'**, and otherwise equal to 1. We can introduce three metrics to distinguish the relative importance of value accuracy in context of the tuple. The first two metrics have the purpose of giving a different importance to errors on attributes that have a higher identification power, in line with the above discussion.

The first metric called *weak accuracy error,* and is defined:

$$\sum_{i=1}^{N} \frac{\beta(q_i > 0) \wedge (s_i = 0)}{N}$$

where $\beta(.)$ is a boolean variable equal to 1 if the condition in parentheses is *true*, 0 otherwise, and $q_i = \sum_{j=1}^{K} q_{ij}$. Such metric considers the case in which for a tuple t_i accuracy errors occur $(q_i > 0)$ but do not affect identification $(s_i = 0)$.

The second metric is called *strong accuracy error*, and is defined assigning

$$\sum_{i=1}^{N} \frac{\beta(q_i > 0) \wedge (s_i = 1)}{N}$$

where $\beta(.)$ and q_i have the same meaning as above. Such a metric considers the case which accuracy errors occur $(q_i > 0)$ for a tuple t_i and actually do affect identification $(s_i = 1)$.

The third metric gives the percentage of accurate tuples matched with the reference table. It is expressed by the degree of syntactic accuracy of the relational instance **r**

$$\sum_{i=1}^{N} \frac{\beta(q_i = 0) \wedge (s_i = 0)}{N}$$

by actually considering the fraction of accurate $(q_i = 0)$ matched $(s_i = 0)$ tuples.

2.2 Completeness

Completeness can be generically defined as the extent to which data are of sufficient breadth, depth, and scope for the task at hand [**?**] [**?**]

Chapter 3

My third chapter

3.1 First section of the third chapter

And now I begin my third chapter here . . .

And now to cite some more people **? ?**]

3.1.1 First subsection in the first section

. . . and some more

3.1.2 Second subsection in the first section

. . . and some more . . .

First subsub section in the second subsection

. . . and some more in the first subsub section otherwise it all looks the same doesn't it? well we can add some text to it . . .

3.1.3 Third subsection in the first section

. . . and some more . . .

First subsub section in the third subsection

. . . and some more in the first subsub section otherwise it all looks the same doesn't it? well we can add some text to it and some more and some more and some more and some more and some more and some more and some more . . .

Second subsub section in the third subsection

… and some more in the first subsub section otherwise it all looks the same doesn't it? well we can add some text to it …

3.2 Second section of the third chapter

and here I write more …

3.3 The layout of formal tables

This section has been modified from "Publication quality tables in LaTeX*" by Simon Fear.

The layout of a table has been established over centuries of experience and should only be altered in extraordinary circumstances.

When formatting a table, remember two simple guidelines at all times:

1. Never, ever use vertical rules (lines).

2. Never use double rules.

These guidelines may seem extreme but I have never found a good argument in favour of breaking them. For example, if you feel that the information in the left half of a table is so different from that on the right that it needs to be separated by a vertical line, then you should use two tables instead. Not everyone follows the second guideline:

There are three further guidelines worth mentioning here as they are generally not known outside the circle of professional typesetters and subeditors:

3. Put the units in the column heading (not in the body of the table).

4. Always precede a decimal point by a digit; thus 0.1 *not* just .1.

5. Do not use 'ditto' signs or any other such convention to repeat a previous value. In many circumstances a blank will serve just as well. If it won't, then repeat the value.

A frequently seen mistake is to use '\begin{center}' … '\end{center}' inside a figure or table environment. This center environment can cause additional vertical space. If you want to avoid that just use '\centering'

Table 3.1 A badly formatted table

	Species I		Species II	
Dental measurement	mean	SD	mean	SD
I1MD	6.23	0.91	5.2	0.7
I1LL	7.48	0.56	8.7	0.71
I2MD	3.99	0.63	4.22	0.54
I2LL	6.81	0.02	6.66	0.01
CMD	13.47	0.09	10.55	0.05
CBL	11.88	0.05	13.11	0.04

Table 3.2 A nice looking table

Dental measurement	Species I		Species II	
	mean	SD	mean	SD
I1MD	6.23	0.91	5.2	0.7
I1LL	7.48	0.56	8.7	0.71
I2MD	3.99	0.63	4.22	0.54
I2LL	6.81	0.02	6.66	0.01
CMD	13.47	0.09	10.55	0.05
CBL	11.88	0.05	13.11	0.04

Table 3.3 Even better looking table using booktabs

Dental measurement	Species I		Species II	
	mean	SD	mean	SD
I1MD	6.23	0.91	5.2	0.7
I1LL	7.48	0.56	8.7	0.71
I2MD	3.99	0.63	4.22	0.54
I2LL	6.81	0.02	6.66	0.01
CMD	13.47	0.09	10.55	0.05
CBL	11.88	0.05	13.11	0.04

References

[] Ancey, C., Coussot, P., and Evesque, P. (2005). Examination of the possibility of a fluid-mechanics treatment of dense granular flows. *nformation Quality, Advances in Management Information Systems. M.E. Sharpe*, (4):385–403.

[] C. Batini, M. S. (2006). *Data Quality Concepts, Methodologies and Techniques*. Springer-Verlag, Berlin Heidelberg.

[] Deming, W. E. (1960). *Sample Design in Business Research*. ohn Wiley and Sons, Inc., New York.

[] Falorsi and Scannapieco (2006). *Principi Guida per la Qualità dei Dati Toponomastici nella Pubblica Amministrazione*, 12.

[] Jugulum, R. (2014). *Computing with High Quality Data*, page 1. John Wiley and Sons, Inc., New Jersey.

[] M., J. (1985). *Advances in Record Linkage Methodologies as Applied to Matching the 1985 Cencus of Tampa, Florida.*, pages 414–420.

[] Read, C. J. (1985). A solution to the invariant subspace problem on the space l_1. *Bull. London Math. Soc.*, 17:305–317.

[] Wang and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. journal of management information systems. (4):12.

Appendix A

How to install LaTeX

Windows OS

TeXLive package - full version

1. Download the TeXLive ISO (2.2GB) from https://www.tug.org/texlive/
2. Download WinCDEmu (if you don't have a virtual drive) from http://wincdemu.sysprogs.org/download/
3. To install Windows CD Emulator follow the instructions at http://wincdemu.sysprogs.org/tutorials/install/
4. Right click the iso and mount it using the WinCDEmu as shown in http://wincdemu.sysprogs.org/tutorials/mount/
5. Open your virtual drive and run setup.pl

or

Basic MikTeX - TeX distribution

1. Download Basic-MiKTeX(32bit or 64bit) from http://miktex.org/download
2. Run the installer
3. To add a new package go to Start » All Programs » MikTex » Maintenance (Admin) and choose Package Manager

4. Select or search for packages to install

TexStudio - TEX editor

1. Download TexStudio from http://texstudio.sourceforge.net/#downloads

2. Run the installer

Mac OS X

MacTeX - TEX distribution

1. Download the file from https://www.tug.org/mactex/

2. Extract and double click to run the installer. It does the entire configuration, sit back and relax.

TexStudio - TEX editor

1. Download TexStudio from http://texstudio.sourceforge.net/#downloads

2. Extract and Start

Unix/Linux

TeXLive - TEX distribution

Getting the distribution:

1. TexLive can be downloaded from http://www.tug.org/texlive/acquire-netinstall.html.

2. TexLive is provided by most operating system you can use (rpm,apt-get or yum) to get TexLive distributions

Installation

1. Mount the ISO file in the mnt directory

   ```
   mount -t iso9660 -o ro,loop,noauto /your/texlive####.iso /mnt
   ```

2. Install wget on your OS (use rpm, apt-get or yum install)

3. Run the installer script install-tl.

   ```
   cd /your/download/directory
   ./install-tl
   ```

4. Enter command ‘i’ for installation

5. Post-Installation configuration:
   http://www.tug.org/texlive/doc/texlive-en/texlive-en.html#x1-320003.4.1

6. Set the path for the directory of TexLive binaries in your .bashrc file

For 32bit OS

For Bourne-compatible shells such as bash, and using Intel x86 GNU/Linux and a default directory setup as an example, the file to edit might be

```
edit $~/.bashrc file and add following lines
PATH=/usr/local/texlive/2011/bin/i386-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

For 64bit OS

```
edit $~/.bashrc file and add following lines
PATH=/usr/local/texlive/2011/bin/x86_64-linux:$PATH;
export PATH
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;
export MANPATH
```

```
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

Fedora/RedHat/CentOS:

```
sudo yum install texlive
sudo yum install psutils
```

SUSE:

```
sudo zypper install texlive
```

Debian/Ubuntu:

```
sudo apt-get install texlive texlive-latex-extra
sudo apt-get install psutils
```

Appendix B

Installing the CUED class file

LATEX.cls files can be accessed system-wide when they are placed in the <texmf>/tex/latex directory, where <texmf> is the root directory of the user's TEXinstallation. On systems that have a local texmf tree (<texmflocal>), which may be named "texmf-local" or "localtexmf", it may be advisable to install packages in <texmflocal>, rather than <texmf> as the contents of the former, unlike that of the latter, are preserved after the LATEXsystem is reinstalled and/or upgraded.

It is recommended that the user create a subdirectory <texmf>/tex/latex/CUED for all CUED related LATEXclass and package files. On some LATEXsystems, the directory look-up tables will need to be refreshed after making additions or deletions to the system files. For TEXLive systems this is accomplished via executing "texhash" as root. MIKTEXusers can run "initexmf -u" to accomplish the same thing.

Users not willing or able to install the files system-wide can install them in their personal directories, but will then have to provide the path (full or relative) in addition to the filename when referring to them in LATEX.