Data Science – A Software Perspective

Swedhana Viswanathan
Faculty Advisor: Dr. Jean Muhammad
Hampton University
100 E.Queen Street
Hampton Virginia 23668
swesairi@gmail.com

ABSTRACT
Information science is a powerful tool to study the personality trait of a person as individual or a group or as an entity of a society. Data science helps in storing the information, sieving the required information from the rest, analyzing the data to produce qualitative and quantitative products and securing the data. This paper is an attempt to briefly study data science in a software perspective with data mining as the topic of interest. Various methods and techniques used in data mining and analysis are considered. Few examples of data mining in various fields are discussed. Commercially available data mining/analysis tools are looked into. The author has taken Rapid Miner -a commercially available data mining and predictive analysis tool for this paper and briefly reviewed its text analysis tool. A use case analysis to explore the performance of the tool for various document types was conducted using the tool. Rapid miner is open sourced written in JAVA. This is an initial study of the tool and cannot be considered as the ultimate critique analysis.

CCS Concepts
• **Information systems~Data mining**

Keywords
Data Mining; Rapid Miner; Text Processing

1. INTRODUCTION
Human beings are always fascinated by anything and everything remotely interesting. We have come a long way from wild dwellers, food gatherers to space exploring and accessing our needs with just a click. The last 50 years has shown a great growth technology wise and people are more connected than ever. This has led to huge accumulation of information or data of all kinds. Over the last couple of decades, study of the data generated through various sources has become a field of its own and has created new technical innovations to surface. Data available can be exploited for both productivity and destruction. It becomes necessary for every governmental and commercial agency to collect and filter information from various sources which are of importance for their particular field of interest, analyze and envisage the outcome. Researchers study the distributed data of varied types from their experiments and use data analysis techniques to do a qualitative observation in the data pattern. It is always not necessary that the data available is sufficient or direct to extract the information. It is the science and techniques which are applied to the data make it usable.

ADMI 2016, March 31-April 3, 2016, Winston-Salem, NC, USA.

2. OBJECTIVE AND METHODOLOGY

This paper attempts to understand the process involved in data science to transform data to products. Data mining- a part of the process is studied in depth with importance given to the software aspects. Various techniques and elements in data-mining are explored and discussed. The need for software tools and the requirements for a data predictive analysis tool are studied. A software tool is tested with a use-case. Its ease of access, adaptability and performance are reviewed.

3. DATA SCIENCE
3.1 Knowledge Discovery from Database
One definition of data science is the study of the generalizable extraction of knowledge from data [15]. However, data science does not pertain just to data however it's a wider arena creating data products, creation of data applications, creating products which directly or indirectly use data. A single application can utilize, analyze, and produce outcomes from multiple non related data and data types [7].

Though data science was initially presumed to be an extension of statistics, it is an area of study which encompasses multiple fields, integrate information from various sources, analyze the information according to the necessity and create usable product of the data. Data Science is a growing field and is expected to be a job generating area. More awareness on data Science as a career is being explored; with Universities now providing courses which caters to this field. The exorbitant amount of data, the necessity to manage and analyze it creates a requirement for automation. Applications and methodologies are devised and updated regularly for precision handling. Data Science as an entity is still in the early stages. Hence it would not be surprising if the current process documentation and knowledge extraction methodology are changed completely in few years. Transformation of raw data to knowledge and data products is also known as Knowledge Discovery from Database (KDD). This term was first coined in the early 90s. The process of conversion involves the below basic steps which can be achieved through various technologies and methods (see Figure 1).

3.1.1 Selection
This is the first step in KDD process which involves choosing the right data sets for evaluation by identifying the type of knowledge which needs to be extracted and understanding the customer requirements.

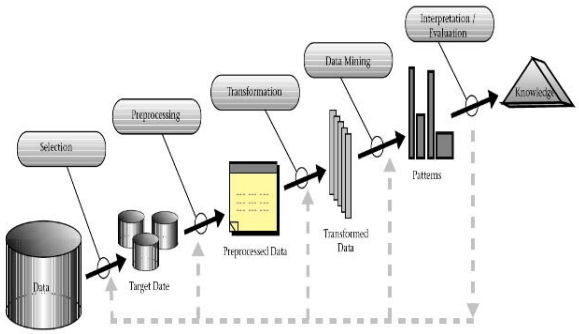

Figure 1. Knowledge Discovery from Database [8]

3.1.2 Preprocessing

Raw data can be anything from just numbers, random audio tapes to huge amount of ASCII and binary values. It is mandatory to remove the noises in the data, align it in a DMBS, identifying the entities, schema and mapping the inter-related values.

3.1.3 Transformation

Once the data is aligned, it can be transformed into a format which can assist in better processing. Normalization of values, converting numbers into functions, adding fields to correlate the data are few examples of data transformation.

3.1.4 Data-Mining

Once the data is aligned multiple techniques can be used to analyze and study the data. Algorithms and data modeling should be designed to read the data Depending on the requirement data patterns and analysis can be formed. These patterns can then be visualized.

3.1.5 Evaluation

The resultant pattern from the previous step can be evaluated and unwanted information can be removed. The performance of the devised algorithms is tested and checks the precisions.

3.1.6 Knowledge

With the relevant and available analyzed information, result can be utilized as per needs. Extracted knowledge can be used to create data products, pave way to future analysis, further analysis can occur etc...

3.2 Data Scientist

Data Scientist is a recent term which has been coined over the last few years for the people who analyze and work with data. However, a data scientist not just analyze the data, but also evaluate the necessary information from the data by designing algorithms, writing automation scripts to process the data and create valuable and creative knowledge out of it. A Data Scientist is not necessarily a statistician or a computer engineer, but can belong to any other domain as long as he/she is capable of visualizing relevant information from raw data [14].

3.2.1 Data Scientist's responsibilities [18]

Data Seeking: Data is never available in a ready-made form. It is necessary to seek out the right data. If required, data should also be created from other data sources by various methods.

Data Visualization: Before proceeding with analyzing the data, a data scientist should try to visualize the data present and the product outcome of the analysis. This helps in focusing on the choosing the right type of tool/methodology to work on the data.

Data Integration: Data can be present in the various types, software, stored in various hardware and be of different bandwidth. A data scientist should be able to recognize and extract the data and bring out a relationship between different data sets.

Data Consistency: Data extracted could be of different types. A data scientist is responsible for ensuring the integrity of data across modules.

3.2.2 Data to knowledge conversion

The curiosity and the cleverness of a data scientist can change the outlook of the data derived from various sources into creative data products. This can be achieved by recognizing the patterns of the data, identifying the link and the lead provided by the data. Knowledge to project the future of the product can help in maintaining provisions to upgrade it as required.

3.2.3 Software engineers as Data Scientists

Though a Data Scientist need not be a software engineer essentially, it is necessary for him/her to know the software tools to study the data, build automation scripts to read and analyze the data, process, develop architecture for the product. A software engineer essentially turns the amalgamated information into a working prototype [4].

3.3 Data Mining

The vital step in KDD process is Data mining, which refers to application of algorithms and other data analysis tools to study the patterns within the gathered data. This step widely uses multiple software tools to complete the process. Data mining aims mainly at predictive analysis. This step particular by itself has direct business applications. This process has multiple stages.

- Analytical Process: Also known as Online Analytical Processing is the selective extraction of data and viewing it in different points of view [20]
- Data Exploration: This is the stage where the data is prepared, extracted and large data sets are transformed. There are multiple models which are utilized in this stage.
- Validation and Verification: In this stage, multiple models used in predictive analysis are compared and the best model for the scenario is chosen. This stage is very important in data mining as it helps in stabilizing the data
- Deployment: Final stage where the chosen predictive analysis model from the previously discussed stages is applied to the data and the outcome is expected.

3.3.1 *Data Mining Methods*

Data mining has the following tasks [8]

- Classification: Defining the data into one or more predetermined classes.
- Regression: Finding a right function which could be a right predictive model for the data.
- Clustering: Identifying a set of categories to fit the data in. These categories can be pre-available set.
- Summarization: This process helps in providing a compact representation of the data sets.
- Association rule learning: Also known as dependency modelling, this task works towards finding a relation between the variables in the data. This happens in two levels; 1) Structural level and 2) Quantitative level.

3.3.2 *Data Mining Techniques*

Some of the data mining techniques generally used are [9]: Statistics, Machine learning, Database Systems, Neural Networks, Visualizations etc.

3.3.3 *Data Mining Tools*

Data mining utilizes few many algorithms and software tools to do the predictive analysis. Some of these algorithms commonly used are Hierarchical Methods, Partitioning Methods, Grid-Based Methods, Constraint-Based Clustering, Scalable clustering, High Dimensional Data Algorithms [3]. Some of the commercial available Data-Mining software tools are EXCEL, ORACLE SQL, MATLAB, SAS, SAP, RAPID MINER, KNIME, TERADATA, TIBCO etc... With the data mining becoming a very sought out profession for Software Engineers, many software tools are being developed as open sources which cater to multiple research fields.

4. GARTNER REPORT

The Gartner Magic Quadrant for advanced analytics platforms was released in February 2015. The major criteria for the evaluation were the product/service provided by the platform, the customer experience, market understanding and innovation (see Figure 2**Figure**). The report has a detailed description of the requirements to be a part of the quadrant and the vendors chosen where completely analyzed for their strengths and weakness [12]. The report has placed Rapid Miner – an open sourced predictive analysis tool as a leader. This paper will continue to explore Rapid miner as a Data Mining tool.

Figure 2: Gartner Report [12]

5. TEXT MINING

A well-conceived fact about knowledge in business is 80% of the relevant information is found from unstructured texts [10]. Text mining is an extension of data mining to find interesting pattern in textual content. It uses natural language processing to study through the preprocessing, transformation and extraction [11]. In text mining, sentences are looked like set of words and the words are individually analyzed. There is an increased and growing interest in text mining in varied disciplines. In the field of life sciences, researches use text mining algorithm and software tools to predict the outcome of experiments, dig into the varied data available to find the pattern, compare and contradict data from various sources and find a common link, help the labs to determine and identify the latest trend in research.

In linguistics, text mining is used in determining the word pattern, spotting trend in social media, detecting spam messages, identifying critical information from set of insignificant data [2]. In business, market research uses text analysis to identify the customer requirements and determine the next expected product change. Job markets use text analysis to compare the keywords of the job requirement and the submitted resumes. Generally, in text mining the following pre-processing techniques are generally undertaken before the data is analyzed [1] (see Figure 3).

- Create Corpus: Collecting the data from documents.
- Pre-processing and cleaning the text
- Tokenization: Fragmenting the text to countable words.
- Stemming: Identifying the common words.
- Stop words: Eliminating the prepositions, adjectives, articles etc.
- Transformation: Converting the case of the words
- Reducing the dimensionality
- Knowledge Extraction: Using the data mining methods

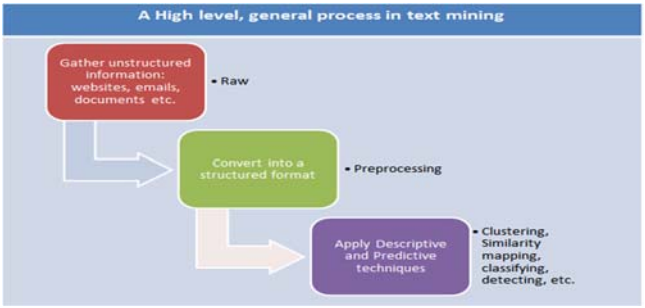

Figure 3 : Text mining process [5]

6. RAPIDMINER

6.1 Background and Overview

RapidMiner is a predictive analysis software platform. Initially known as YALE (Yet another Learning environment), it was developed in 2001 at University of Dortmund, Germany by Ralf Klinkenberg, Ingo Mierswa & Simon Fischer. It was later founded as a separate company in 2007, with its head-quarters at Boston, USA. This is a complete KDD software suite with modern Graphical user Interface, enabling users to perform various types of data analysis without actually programming [19] [6]. The software is open sourced and is written in JAVA. The following links have the rapid miner documentation and code-source. It incorporates WEKA and R scripts for its big data analysis [17].

The RapidMiner Suite consists of 4 products: -

- RapidMiner Studio: GUI for Data mining and Predictive analysis (see Figure 5)
- RapidMiner Server: A dedicated server to back up the predictive analytics
- RapidMiner Radoop: A RapidMiner extension to work on Hadoop Big Data environment
- RapidMiner Cloud: 25 GB cloud repository and accessibility to analyze data on cloud.

This platform is adaptable across domains such as Banking, Insurance, Education, Life Sciences, and Manufacturing etc… RapidMiner suite is used by PayPal, EBay, Telenor, Lufthansa, Cisco, and George Washington University etc...

Rapid Miner studio has multiple user interactive pages. It provides hands-on tutorials for first time users. There are multiple process control algorithms and operators for predictive analysis which are custom provided by Rapid Miner (see Figure 4 and Figure 5). This being an open source tool, users can also download the R and WEKA script plugins, develop their own operators and processes. There are also interactive tutorials available for first time users. The data can be uploaded in multiple formats such as a CSV file, XML file, database table and others. It also provides a local repository to store the data. A defined process list by the user is stored in the design page. The same design can be utilized to analyze multiple data sets.

Rapid Miner studio 6.1 is available to download for a trial period of 14 days. For academicians (students, professors, researchers), free licensing is provided for 1-year period to learn and use.

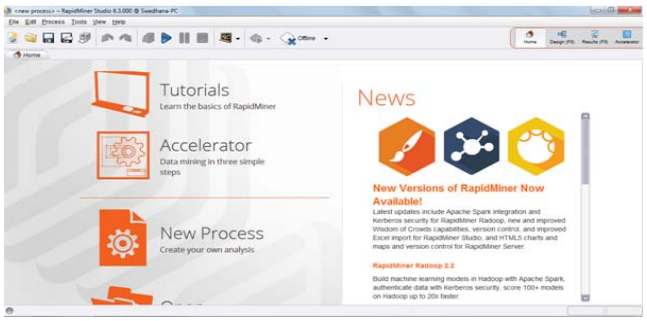

Figure 4 : RapidMiner Welcome Page

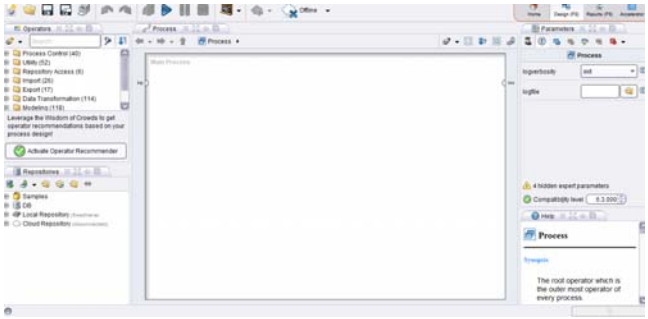

Figure 5 : RapidMiner Design Interface

6.2 Text processing

In RapidMiner text mining is available as a plug-in which can be downloaded from the website. The user can go to the help tab in GUI, click on Manage Extensions and choose the text processing plug-in [13]. The GUI will restart and the text processing operator will be available on the design interface (see Figure 6)**Error! Reference source not found.**. This operator has in-built sub operators for reading the data, pre-processing, extraction of information. The RapidMiner text processing operator utilizes the Naïve-Bayes algorithm [13], a clustering technique for tokenization and extraction techniques. It can read document in .txt format, html, pdf and xml. The tokenization process removes the tags which are present in html and xml files.

6.3 User Interface Example

A basic text processing test case is taken into consideration [16]. A text document is read from the file, it is tokenized to words, the stop words are filtered, words are then normalized for their cases, the redundant words with lengths less than 3 and greater than 25 are removed. The document is then processed to find the words, its count and it is listed in alphabetical order. The preprocessing is conducted using the operators (see Figure 7) and on executing the operation, the analyzed document is produced (see Figure 8).

6.4 Comparative Analysis

A same document in multiple formats is taken in this analysis for consideration. The process discussed in the previous section is used to analyze this document. The PDF format takes the maximum time to analyze which is about 2 seconds while the text and HTML documents takes less than 1 second. The data which is once analyzed from a document is stored in cache and it can be retrieved on recursion.

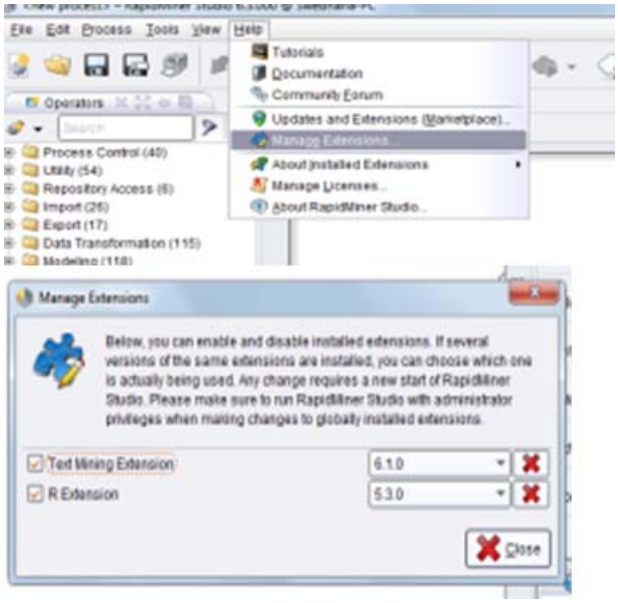

Figure 1: RapidMiner Text processing extension

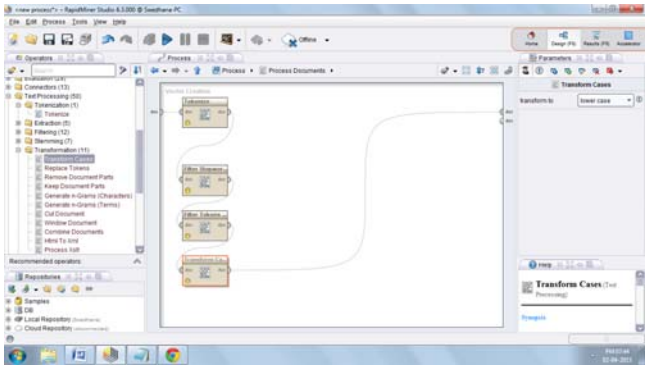

Figure 7: Word preprocessing operators

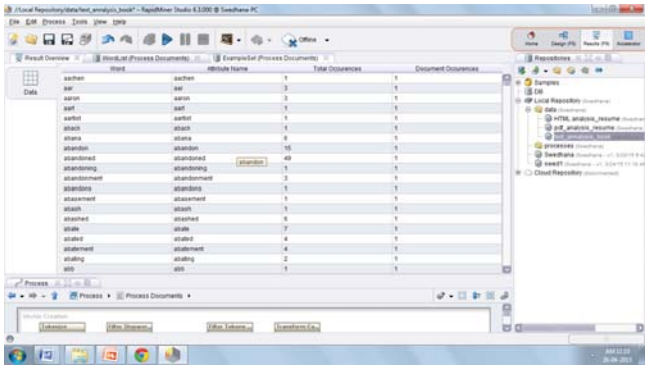

Figure 8: Analyzed Document

7. CONCLUSION

This paper is a detailed study on Data Science with special emphasis on to Data Mining. An open sourced predictive analysis tool - RapidMiner was chosen after analyzing the Gartner report. The tool was explored briefly for text mining procedures. A test case was chosen to analyze a document. The same use case was considered to analyze the same document in multiple formats and its performance with respect to time was evaluated. This paper can be used by beginners to understand about Data science and it shows a career path for software engineers to explore in this area.

8. REFERENCES

[1] Text Mining: The Next Data Frontier:2014. *http://www.scientificcomputing.com/blogs/2014/01/text-mining-next-data-frontier*.

[2] Why Text Mining May Be the Next Big Thing | TIME.com: 2012. *http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/*

[3] Berkhin, P. 2006. A survey of clustering data mining techniques. *Grouping multidimensional data*. J. Kogan et al., ed. Springer. 25-71.

[4] Data Scientist: The Sexiest Job of the 21st Century: 2012. *https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/*.

[5] 3 ways to use text mining with RapidMiner to juice up your job search: 2014.*http://www.simafore.com/blog/bid/111839/3-ways-to-use-text-mining-with-RapidMiner-to-juice-up-your-job-search*.

[6] RapidMiner at CeBIT 2010: the Enterprise Edition, Rapid-I and Cloud Mining - Data Mining - Blog.com: 2010. *http://www.data-mining-blog.com/cloud-mining/rapidminer-cebit-2010/*.

[7] Dhar, V. 2013. Data science and prediction. *Communications of the ACM*.

[8] Fayyad, U. et al. 1996. From Data Mining to knowledge discovery in databases. *AI Magazine*.

[9] Friedman, J. 1998. Data Mining and Statistics: What's the connection? *Computing Science and Statistics*. 29, 1 (1998), 3-9.

[10] Unstructured Data and the 80 Percent Rule: 2008.*http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/*.

[11] What Is Text Mining? 2003. *http://people.ischool.berkeley.edu/~hearst/text-mining.html*.

[12] Magic Quadrant for Advanced Analytics Platforms: 2015.*http://www.gartner.com/technology/reprints.do?id=1-2AHPOU0&ct=150225&st=sb*. Accessed: 2015- 03- 04.

[13] Hofmann, M. and Klinkenberg, R. *RapidMiner*.

[14] 8 Skills You Need to Be a Data Scientist | Udacity: 2014.*http://blog.udacity.com/2014/11/data-science-job-skills.html*.

[15] What is data science? 2010. *https://www.oreilly.com/ideas/what-is-data-science*.

[16] Analytics and Visualization of Big Data: Text Processing Tutorial with RapidMiner: 2013. *http://auburnbigdata.blogspot.com/2013/03/text-processing-tutorial-with-rapidminer.html*.

[17] Mierswa, I. 2012. *RapidMiner Studio*. RapidMiner.

[18] Building data science teams - O'Reilly Radar: 2011.*http://radar.oreilly.com/2011/09/building-data-science-teams.html*.

[19] RapidMiner Review - Butler Analytics: 2015. *http://butleranalytics.com/rapidminer-review/*.

[20] What is OLAP (online analytical processing)? – Definition from WhatIs.com: *http://searchdatamanagement.techtarget.com/definition/OLAP*.