**NSF/ASME Student Design Essay Competition 2019**
Anaheim, California

Challenges in the Design of Complex Systems

# Stay Competitive in Big Data: A Text Mining Platform for Knowledge Discovery in Manufacturing Enterprises

**Shan Peng**
University of Oklahoma

## Abstract

Manufacturing enterprises will be immersed in a huge amount of data, especially textual data, in the year of 2035. A successful company should be able to overcome the problem of "rich data but poor knowledge." In this paper, I define five characteristics for a company to stay competitive in the big data environment, and propose the concept of TMPKDME - A Text Mining Platform for Knowledge Discovery in Manufacturing Enterprises. The role of TMPKDME is to maximize the utility of accumulated data and help a manufacturing company discover knowledge. Six research questions and the associated hypotheses are also identified for the interests of both industry and academia.

**Key Words**

Big data, Text Mining, Knowledge Discovery, Machine Learning, Platform

## 1. Introduction

With the advances in information and communication technology, we are living in a "big data" era. As estimated by Tx Zhuo, a Managing Partner from Karlin Ventures, by 2020, every person online will create roughly 1.7 megabytes of new data every second of every day, and that's on top of the 44 zettabytes (or 44 trillion gigabytes) of data that will exist in the digital universe by that time [1]. The same thing is happening in the design and manufacturing industry, where the data is not only produced by humans, but also by machines. Because of the globalization of supply chain and technologies such as 5G, Internet of Things (IoT), Digital Twins, Augmented Reality (AR), and Cyber-Physical Systems (CPS), the volume of data grows at an unprecedented rate in digital manufacturing environments. According to Wang and Lai [2], about 80% of the available large-scale electric and digital data are texts in nature. These data may be related to design, products, machines, processes, materials, inventories, maintenance, planning and control, assembly, logistics, performances etc. [3]. The textual data exists in the form of descriptive formats which include project management documents, design manuals, operation handbooks, service reports about repair information, manufacturing quality documentation and customer help desk notes [4]. However, having large-scale of textual data doesn't mean that enterprises gain the corresponding amount of knowledge from the data, the "rich data but poor knowledge" problem [5] can still happen. Data are just the description of raw facts. Most of the human-and-machine-generated textual data are often highly unstructured and heterogeneous. The capability to discover knowledge from the accumulated, unstructured, and heterogeneous textual data is critical for manufacturing enterprises to be successful in a highly competitive global environment.

I foresee that textual data in manufacturing enterprises will keep growing exponentially and reach a mind-boggling amount in 2035 that may plunge these enterprises into the "rich data but poor knowledge" problem. To address this issue, in this paper I propose a Text Mining Platform for Knowledge Discovery in Manufacturing Enterprises (TMPKDME). This paper is organized as follows. In Section 2, I define the characteristics of a successful manufacturing company in 2035, and introduce the framework and key functionalities of TMPKDME to facilitate the company staying competitive. In Section 3 I propose six research questions for the design and realization of TMPKDME, and propose the hypotheses related to the research questions. In Section 4, I summarize the paper, speculate the future, and emphasize the importance of TMPKDME.

## 2. The Characteristics of a Successful Company in 2035

### 2.1 What is a successful company?

A global manufacturing company in the year 2035 will have to face a lot of challenges, especially the challenge of a huge amount of textual data accumulated from the projects, products, processes, and customers, etc. Given this challenge, I define the characteristics of a successful company as shown in Figure 1 and describe the details of the characteristics.

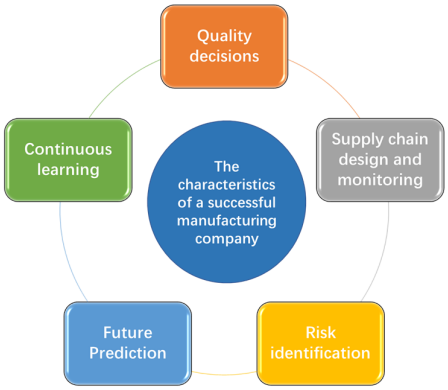


Figure 1. The Characteristics of a Successful Company in 2035

- **The capability to rapidly make quality decisions**

Manufacturing companies make many decisions, such as the decision about the design of product families to satisfy various customer needs, the decision about making the products by themselves or outsourcing, the decision about whether or not to adopt a new technology, the selection of suppliers or partners, the decision about the internal organization structure, etc. As the environment changes fast, in order to stay competitive a company should be able to rapidly make quality decisions in response to changes. These decisions can be supported by analyzing textual documents such as market reports and technical papers that are internal or external of the company.

- **The capability to design and monitor the entire supply chain**

Products are becoming more and more complex. The realization of these product requires a company to have an extensive supply chain that is distributed globally. To ensure the quality of the products and secure the production capacity, a company should be firstly able to design an appropriate global supply chain by considering the properties of different suppliers, then able to monitor the operating conditions of the entire supply chain for well managing it once it is established. The design of a supply chain depends on the analysis of the advantages/disadvantages of suppliers. The monitoring of a supply chain requires a company to acquire the status reports from its suppliers and integrate these reports into a big picture so as to have an overview of the operating conditions of the supply chain.

- **The capability to identify risks in the dynamic market**

Risks are ubiquitous in a dynamic market. The risks may include, for example, variation of customer demands, shortage or increasing prices of materials/components, crisis in public relationship, and adjustments in partners' strategies etc. A successful company should be able to identify the potential risks from huge amount of textual data in the related domains, and make responding plans before the risks cause adverse consequences or losses. Even though the risks become crisis, the company can stay resilient and recover quickly.

- **The capability to predict customer demands and technology developing trends**

Being able to predict the future is critical for a manufacturing company to gain competencies in the

future. Customers and technologies are two important aspects based on which a company could win the competition in the market. Customers are sources where a company gets orders and makes profits. Technology is the core element for a company to reduce cost, improve quality and productivity. The company should be able to predict customer demands by analyzing textual resources such as online customer reviews, and to predict technology developing trends by analyzing academic literature and patent databases. Base on the prediction, the company can make the right investment in product development, manufacturing, and sale so as to quickly win future market shares.

- **The capability to learn continuously**

Being able to learn from the past can help a manufacturing company develop itself in a healthier way. Past experience (represented by documented cases) can be classified into successful and failure cases. Identifying the characteristics and reasons of successful cases facilitates a company understanding what principles, policies, or actions it should insist on. Analyzing the root causes of past failure cases helps a company avoid adverse consequences in similar situations in the future. The capability to continuously learn from both successful and failure cases will benefit the company in making better decisions.

### 2.2 A Text Mining Platform for Knowledge Discovery in the Context of Big Data

In order to facilitate a manufacturing company to gain the competencies identified in Section 2.1, I propose a Text Mining Platform for Knowledge Discovery in Manufacturing Enterprises (TMPKDME). The overview of TMPKDME is shown in Figure 2. TMPKDME consists of three parts: text sources, program, and decision makers. Text sources are classified as external and internal documents. External documents include customer related resources such as online product reviews, supplier related resources such as financial reports and capability reports etc., public resources such as ASME database, AIAA database, and US patent database etc. Internal documents include technical reports, product handbooks, maintenance reports and failure reports etc. These textual documents are input to a program for analysis and generating visual results. Key functionalities of the program include data cleaning, data storage and various text mining algorithms. Data cleaning is to extract key components from the raw textual contents. Data storage is to save the cleaned data using structured formats such as tables and ontologies. Various algorithms such as word frequency, text similarity, text network, machine learning algorithms are used to discover the hidden patterns in the data. Finally, the discovered patterns are visualized using different graphs or charts to facilitate decision makers such as designers, engineers, and managers gaining knowledge from the data. This knowledge can help decision makers: 1) make quality decisions, 2) design and monitor the supply chain, 3) identify potential risks, 4) predict the future, 5) learn continuously.

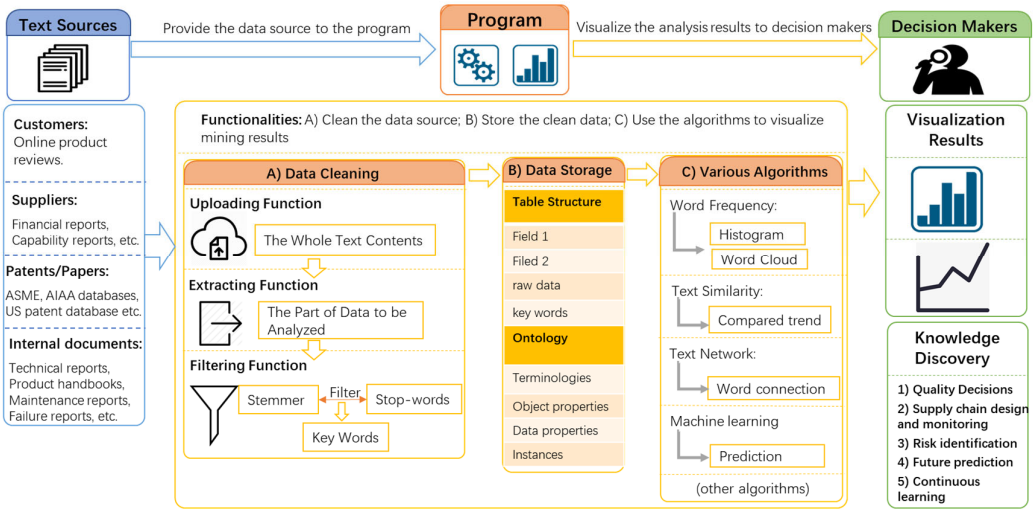


Figure 2. A Text Mining Platform for Knowledge Discovery in Manufacturing Enterprises

In the year of 2035 when manufacturing industry is immersed in big data, the role of TMPKDME is to maximize the utility of accumulated data and help a manufacturing company discover knowledge and get rid of the "rich data but poor knowledge" problem. To realize TMPKDME, in Section 3 I propose six research questions and the associated hypothesis.

## 3. Research Questions for Designing and Developing the Text Mining Platform

### 3.1 Textual Document Formalization

In a manufacturing company, the textual documents may come from different domains such as material, design, manufacturing, maintenance etc. Due to the difference of standards in different domains, the associated documents are heterogenous and multi-structured. There is a need to formalize these documents in TMPKDME. To address this, I propose a research question and an associated hypothesis as follows.

**Research Question 1**: How can the heterogenous and multi-structured textual documents be formalized to facilitate computing in TMPKDME?
**Hypothesis 1**: Ontologies can be used to formalize the heterogenous and multi-structured documents.

An ontology is an explicit specification of a conceptualization [6]. Domain terminologies are formally defined in an ontology using explicit specifications, which can be used to annotate a document from a particular domain. Figure 3 is an ontology example. Annotating documents using the terms in this ontology will facilitate the interpretation of the documents in TMPKDME.

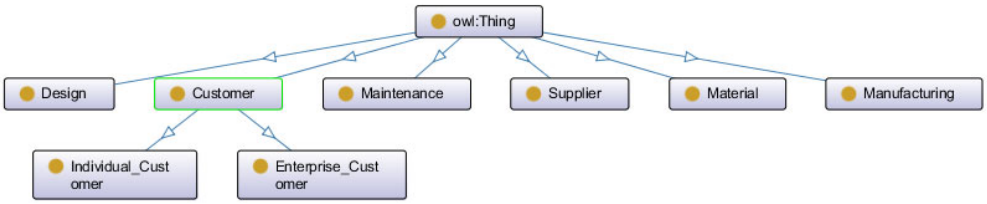


Figure 3. Ontology-Based Data Formalization

### 3.2 Data Cleaning

Textual documents include both topic-related words and topic-unrelated words. The realizability of the text mining results depends on the statistics and analysis of topic-related words. Topic-unrelated words are the noise in the text mining process. There is a need to filter the topic-unrelated words in TMPKDME so as to ensure the computing efficiency (i.e., to avoid too much computing resource are used in analyzing topic-unrelated words) and result reliability. To address this, I propose a research question and an associated hypothesis as follows.

**Research Question 2**: What is a approach to filter topic-unrelated words in textual documents in TMPKDME?
**Hypothesis 2**: Natural Language Processing (NLP) methods such as stemming and stop words filtering can be used to filter topic-unrelated words in textual documents.

Figure 4 is a sample text from an essay submitted by engineering students. The stemming method is to transform a given word to its root word, e.g., "forming" is transformed to "form" and "learned" is transformed to "learn". The stop words filtering method is to eliminate the preposition and conjunction words from a given text, e.g., Words "a", "the", "by", "of" are removed from the sample text. The clean text will be analyzed in subsequent steps in TMPKDME.

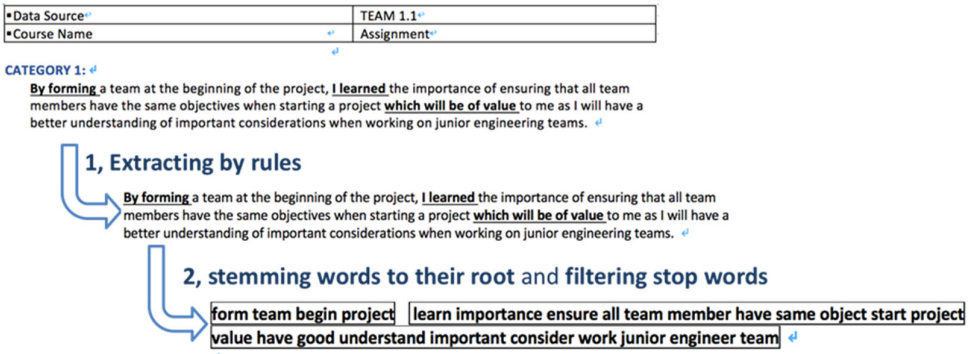


Figure 4. Data Extraction and Filtering

### 3.3 Scalability in Data Storage

Textual documents in manufacturing companies are usually distributed in storage mediums such as company-owned servers, employees' laptops, mobile devices, and portable hard disk drives. The problems of distributed storage include the difficulty in collecting the scattered data, and the difficulty to scale the storage capacity up when the data increase rapidly. There is a need to enable

scalability in data storage in TMPKDME, and to enable uniform management of the scattered textual resources. To address this, I propose a research question and an associated hypothesis as follows.

**Research Question 3**: What is an appropriate data storage architecture that is scalable and facilitate uniform management of the scattered textual resources?
**Hypothesis 3**: A cloud-based architecture can be used for scalable data storage and uniform data management.

Cloud storage is a model of data storage in which the digital data is stored in logical pools [7]. The physical storage spans multiple servers (sometimes in multiple locations), and the physical environment is typically owned and managed by a hosting company. A cloud-base data storage architecture, as shown in Figure 5, is a scalable and elastic architecture which is able to rapidly expand the storage capacity by some simple configurations. It also provides a uniform, accessible environment for managing the scattered textual data, which can furthered be processed using cloud computing.

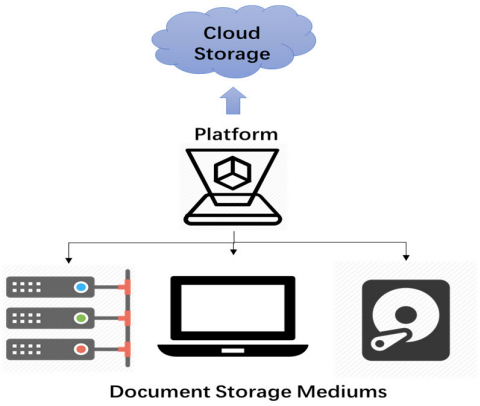


Figure 5. A Cloud-Based Data Storage Architecture

### 3.4 Pattern discovery using Various Algorithms

There are patterns hidden within an individual document and across a set of documents. These patterns may include the relation between the number of occurrence of some particular words and the topic of an individual document, the relation between the co-relation strength of a set of phrases and the topic of an individual document, the relation between the similarity of a set documents and the closeness of their associated topics, and the existence of some known or unknown topic groups. There is need to discover the hidden patterns in a huge amount of textual documents in TMPKDME. To address this, I propose a research question and an associated hypothesis as follows.

**Research Question 4**: How to discover different patterns hidden within individual documents and across a set of documents?
**Hypothesis 4**: Algorithms such as word frequency algorithm, text network algorithm, text similarity algorithm, and text clustering algorithm can be used to discover different patterns hidden within individual documents and across a set of documents.

Word frequency is a measurement of the occurrence numbers of words in a specific document. Higher word frequency means higher closeness between the word and the topic of the document. Text network is an word net identified from a specific document with a level of co-relation strength among a set of phrases. Higher co-relation strength among a set of phrases means higher closeness between these phrases and the topic of the document. Text similarity is a measurement of similarity between two textual documents, it can be to determine if a specific document belongs to a given topic. Text clustering is about the identification of topic groups given a set of documents.

### 3.5 Visualization for Gaining Insights

The results generated by text mining algorithms are a set of numbers and words. For example, the result generated by word frequency algorithm is formatted as "{'word1':200,'word2':150,…}". Results in this representation is not intuitive enough for decision makers to understand the meaning of the numbers and words in a big picture. There is need to visualize the text mining results in TMPKDME so as to facilitate decision makers gaining insights, knowledge, and making decisions. To address this, I propose a research question and an associated hypothesis as follows.

**Research Question 5**: How to visualize the text mining results so as to facilitate decision makers gaining insights, knowledge, and making decisions?
**Hypothesis 5**: Visualization tools such as word cloud, word histogram, and word net etc. can be used to visualize the text mining results.

Figure 6 is a visualized result of the word frequency algorithm using word cloud. The word with the biggest size means that it has the highest frequency in the document, and the topic of the document can be defined by this word. Word histogram is the ranking of words by their frequencies in the document. Word net is used to show the co-relation strength among different words.


Figure 6. A Visualized Result Using Word Cloud

### 3.6 Prediction Based on Machine Learning

The value of the accumulated textual documents in a manufacturing company is anchored in that these historical documents can be used to predict the future. The prediction of the future must be relied on a mathematical model that can be used to infer the future. There is a need to enable the building of prediction models in TMPKDME so as to facilitate decision makers to identify future directions of the company. To address this, I propose a research question and an associated hypothesis as follows.

**Research Question 6**: What is the method to build prediction models based on the accumulated textual documents?
**Hypothesis 6**: Machine learning methods can be used to build mathematical models for the prediction of the future.

The huge amount of historical textual documents form a rich training set for building mathematical models using machine learning methods. Machine learning include both supervised learning and unsupervised learning. The former is used for processing the documents with tags, such as “success” and “failure”. The latter is used for processing the documents without tags.

## 4. Closure

The world is already and will be continuously living in an era of “big data”. In the year of 2035, a manufacturing company will face an incredible amount of data, especially textual data. A company’s competitiveness in the big data environment is heavily relied on the capability to discover the rich knowledge from the accumulated textual data. Given the challenges of big data, in this paper I define five characteristics of a successful company in 2035. These characteristics include the capability to rapidly make quality decisions, the capability to design and monitor the entire supply chain, the capability to identify risks in the dynamic market, the capability to predict customer demands and technology developing trends, and the capability to learn continuously. To facilitate the achievement of these capabilities, I propose a Text Mining Platform for Knowledge Discovery in Manufacturing Enterprises (TMPKDME). The role of TMPKDME is to maximize the utility of accumulated data and help a manufacturing company discover knowledge and get rid of the “rich data but poor knowledge” problem. For the realization of TMPKDME, I propose six research questions and the associated hypotheses. These research questions are related to textual document formalization, data cleaning, scalability in data storage, pattern discovery, visualization, and machine-learning-based prediction. The research questions and hypotheses should be interesting to both industry and academia. I believe TMPKDME can well support manufacturing companies in the era of “big data”.

## References


[1]. Zhuo, T., 2016, "'Big Data' Is No Longer Enough: It's Now All About 'Fast Data'," Entrepreneur. https://www.entrepreneur.com/article/273561.

[2]. Yu, L., Wang, S., and Lai, K., 2005, "A rough-set-refined text mining approach for crude oil market tendency forecasting," International Journal of Knowledge and Systems Sciences, 2(1), pp. 33-46.

[3]. Choudhary, A. K., Harding, J. A., and Tiwari, M. K., 2008, "Data mining in manufacturing: a review based on the kind of knowledge," Journal of Intelligent Manufacturing, 20(5), p. 501.

[4]. Ur-Rahman, N., and Harding, J. A., 2012, "Textual data mining for industrial knowledge management and text classification: A business oriented approach," Expert Systems with Applications, 39(5), pp. 4729-4739.

[5]. Wang, X., and McGreavy, C., 1998, "Automatic classification for mining process operational data," Industrial & Engineering Chemistry Research, 37(6), pp. 2215-2222.

[6]. Gruber, T. R., 1993, "A Translation Approach to Portable Ontology Specifications," Knowledge Acquisition, 5(2), pp. 199-220.

[7]. Wikipedia, 2019, "Cloud Storage," https://en.wikipedia.org/wiki/Cloud_storage.