Data Mining and Sharing to Create Usable Knowledge, Implementation in Small Business in Indonesia

Haikal Achmad, Vanessa F. Sabur, Adita Pritasari*, Hendy Reinaldo

School of Business & Management, Bandung Institute of Technology, Bandung 40132, Indonesia

*Corresponding author: adita.pritasari@sbm-itb.ac.id

Abstract

Data mining is one of the current methods for a company to create new knowledge from the mass amount of information source and organized it into usable data for the company's specific target. Conventionally, only large corporation uses data mining as it needs to collect massive amounts of data and process to sort it into usable forms of information; also because it has necessary customer base and target value that justified the cost and effort. However, with the growth of internet usage and web based technology, with its low barrier entry and almost cost free nature, a small company at current time should also be able to take advantage of data mining to reach its target. The paper will show some possible proposed model of data mining suitable for use by small companies in Indonesia and some of its perceived advantages.

Keywords: Data mining, knowledge management, small business

■1.0 INTRODUCTION

Knowledge management in its current form has evolved to the point where it has converted into various methods and theory that has surpassed its previous limitation. For some, knowledge management is all about people and their action; to another it is all about the system and the asset of the knowledge itself. Nevertheless, one of the current branches of knowledge management itself is about data mining (Stankosky, 2005). The branch of knowledge management discipline can be seen in Figure 1 below
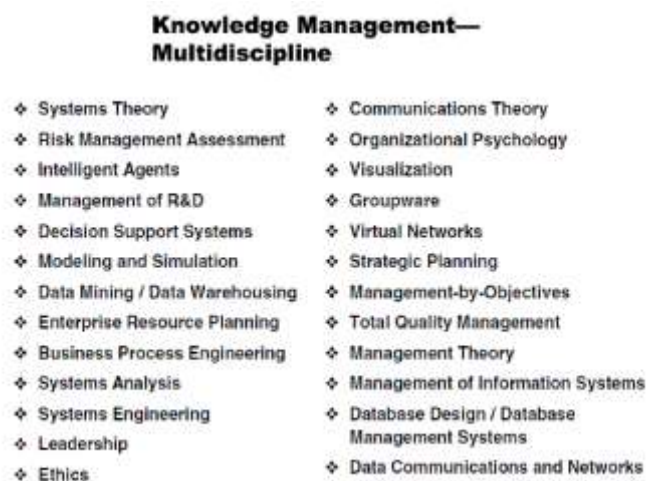
**Knowledge Management—
Multidiscipline**

- ❖ Systems Theory
- ❖ Risk Management Assessment
- ❖ Intelligent Agents
- ❖ Management of R&D
- ❖ Decision Support Systems
- ❖ Modeling and Simulation
- ❖ Data Mining / Data Warehousing
- ❖ Enterprise Resource Planning
- ❖ Business Process Engineering
- ❖ Systems Analysis
- ❖ Systems Engineering
- ❖ Leadership
- ❖ Ethics
- ❖ Communications Theory
- ❖ Organizational Psychology
- ❖ Visualization
- ❖ Groupware
- ❖ Virtual Networks
- ❖ Strategic Planning
- ❖ Management-by-Objectives
- ❖ Total Quality Management
- ❖ Management Theory
- ❖ Management of Information Systems
- ❖ Database Design / Database Management Systems
- ❖ Data Communications and Networks

Figure 1 Branch of KM (Stankosky, 2005)

Data mining is a process of collecting the data needed and summarizing into useful information. The information given by data mining should be processed further into knowledge for making the next step. With even further process, innovation can be achieved as application of using Data Mining as can be seen in Figure 2.
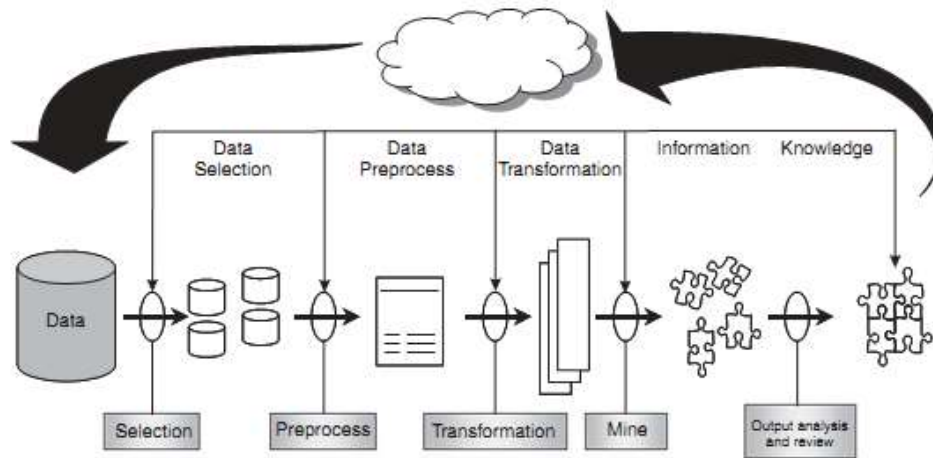

Figure 2　Data Mining Process (Cabena et al., 1997)

Small and Medium Enterprise (SME) can be defined, based from 2008 Law No. 20 Ministry of SMEs and Cooperatives, using the number of net assets the enterprise has and total annual sales gained. Small-sized enterprises are enterprises which have net assets with the value from Rp. 50 Million to Rp. 500 Million or with total annual sales from Rp. 300 Million to Rp. 2.5 Billion. Medium-sized enterprises are enterprises which have net assets with the value from Rp. 500 Million to Rp. 10 Billion or with total annual sales from Rp. 2.5 Billion to Rp. 50 Billion. From the World Bank Enterprise Survey, size of the enterprise is defined by the number of employee: from 5-19, the firm is categorized as small and from 20 – 99 it is a medium-sized firm. SMEs have important roles in developing countries due to their unique characterization which are (Tambunan, 2006): (1) their number is huge, and they are scattered widely throughout the rural areas and therefore they may have a special "local" significance for the rural economy; (2) they are labour intensive; (3) SMEs in developing countries are mainly agriculturally based activities; (4) these enterprises finance their operations overwhelmingly by personal savings of the owners, supplemented by gifts or loans from relatives or from local informal moneylenders, traders input suppliers, and payment in advance from consumers; and (5) SMEs provide an avenue for the development entrepreneurship. These SMEs are engaged in agriculture industry, non-oil mining industry and non-oil processing industry.

Recently, the Indonesian government revitalized the role of SMEs for export growing in manufacturing goods with large-scale exporting companies. With the total of 678,415 SMEs in 2012 and 12.77% contributed to non-oil export (Ministry of SMEs and Co-operative, 2012), Indonesia has the potential of developing SMEs for future economic development. The total non-oil goods being exported from SMEs is 14.93% of the total non-oil export. In the near future, the competition is not only from local enterprises, but also from international enterprises; hence SMEs need to develop their businesses. For sustaining, they need information or knowledge for improvement and innovation. These needs can be gained by using data mining in their business process.

■2.0　DATA MINING MODEL

Data mining is known for its usage in technology firm in the world such as Google, Facebook, Microsoft, etc. After all it is in their nature to be easier to process data that is already in their possession. However, with recent change in technology and the web, even small entity and a single person could already collect necessary data needed for a data mining of certain usage. Data mining itself in its base is a simple principle of creation of knowledge, without the social aspect of a company hierarchy and staffing/training. It means it is enabled an entity to get the benefit of knowledge in a more objective state. How knowledge discovery process happens can be seen in Figure 3.
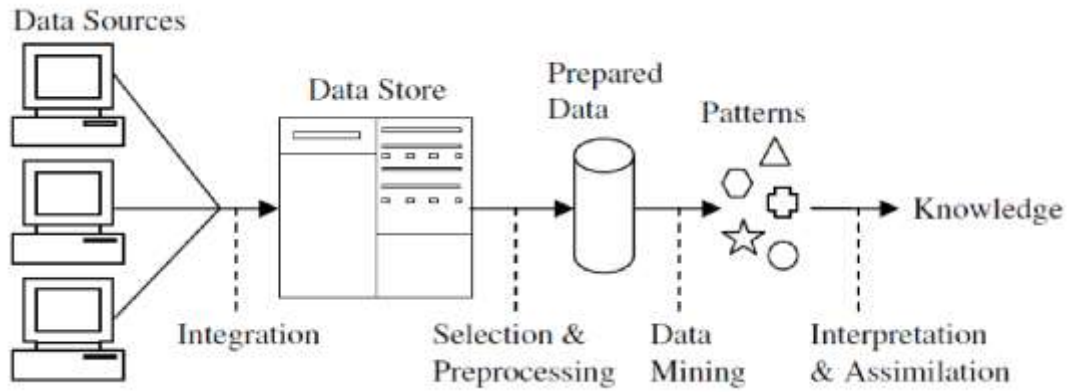
Figure 3 Knowledge Discovery Process (Bramer, 2007)

Most of the current data mining process involved web based info collecting through various account and plug-in which organized data of a person (potential/returning customer) to a single repository and building its oncology profile. Data mining process mapping using web based info collection and plug-in using can be done by using Google Desktop as interface as described in Figure 4 below. Figure 5 shows an example of info collecting result on Standard Web Data Mining.

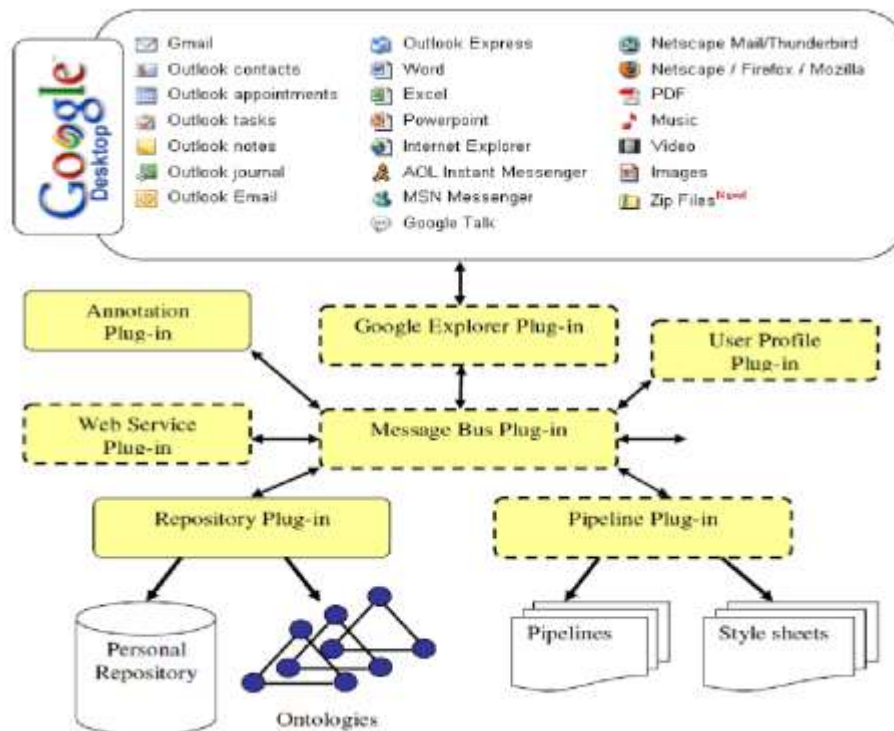**Figure 4** LIFE based Semantic Data Mining Modelling (Reimer & Karagiannis, 2007)

Each plug-in in the framework itself is collecting various data of the person and categorized it on a simple true or false algorithm and deposit it to a database. Even in this stage, there is a lot of redundancy data which is not useful for the company, but is used only to categorize the data itself into a subset of smaller information.

Figure 5   Info Collected on Standard Web Data Mining (Reimer & Karagiannis, 2007)

The problem is, of course, that the model needs numerous amount of plug-in collector and big amount of data to be processed. This means it applied mainly to big corporation with specific skilled staff and big budgeted. Even the big corporation can buy the standardized program for data mining itself. But for the small business, a simpler and specific step system would help them to achieve their intended knowledge.

There are two main approaches in data mining, which are KDD (Knowledge Discovery in Databases) related approach and CRISP-DM (Cross Industry Standard Process for Data Mining) related approach. Almost all the other approaches were made based on these two main approaches. There are three approaches that seem feasible for SMEs which are:

1. KDD Process

KDD process (see Figure 6 for the model) is interactive and iterative (with many decisions made by the user), involving nine steps, which are (Mariscal, Marban, and Fernandez, 2010):

a. Learning the application domain

It includes developing an understanding of the relevant prior knowledge and the goals of the application.

b. Creating a target data set

It includes selecting a data set or focusing on a subset of variables or data samples on which discovery is to be performed.

c. Data cleaning and preprocessing

It includes basic operations, such as removing noise or outliers if appropriate, collecting the necessary information to model or account for the noise, deciding on strategies for handling missing data fields, and accounting for time sequence information and known changes, as well as deciding database management system issues; such as data types, schema and mapping of missing and unknown values.

d. Data reduction and projection

It includes finding useful features to represent the data, depending on the goal of the task, and using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations of the data.

e. Choosing the function of data mining

It includes deciding the purpose of the model derived from the data mining algorithm (e.g. summarization, classification, regression and clustering).

f. Choosing the data mining algorithm

It includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate and matching a particular data mining method with the overall criteria of the KDD process.

g. Data mining

It includes searching for patterns of interest in a particular representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modelling, dependency, association rules and line analysis.

h. Interpretation

It includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns and translating the useful ones into terms understandable by users.

i. Using discovered knowledge

It includes incorporating this knowledge into the performance system, taking actions based on the knowledge or simply documenting it and reporting it to interested parties, as well as checking for and resolving potential conflicts with previously believed (or extracted) knowledge.
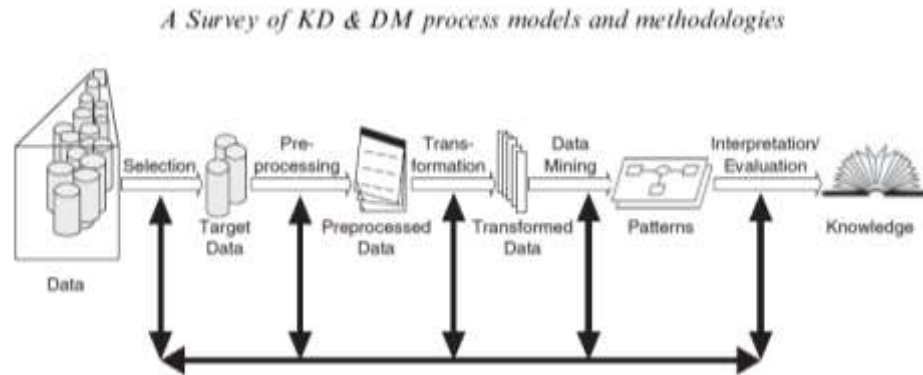
Figure 6   KDDM Data Model

2. SEMMA

SEMMA was developed by the SAS institute for a logical organization of the functional tool set of SAS enterprise miner for carrying out the core tasks of data mining (Mariscal, Marban and Fernandez, 2010). SEMMA is similar to KDD, with the differences that the 1st phase of data selection and the last loop phase of knowledge are omitted (see Figure 7). Hence, SEMMA has no phase to choose whether a resulting data is worthy as knowledge or not.
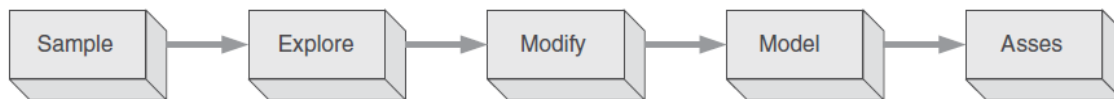
Figure 7 SEMMA Data Model

3. CRIPS – DM

CRIPS-DM model is a modified KDD model. CRIPS-DM is considered the factor standard for developing data mining and knowledge discovery projects. CRIPS-DM forms a hierarchal process model consisting of set of task described at four levels of abstraction (from general to specific). This model can be described as in Figure 8
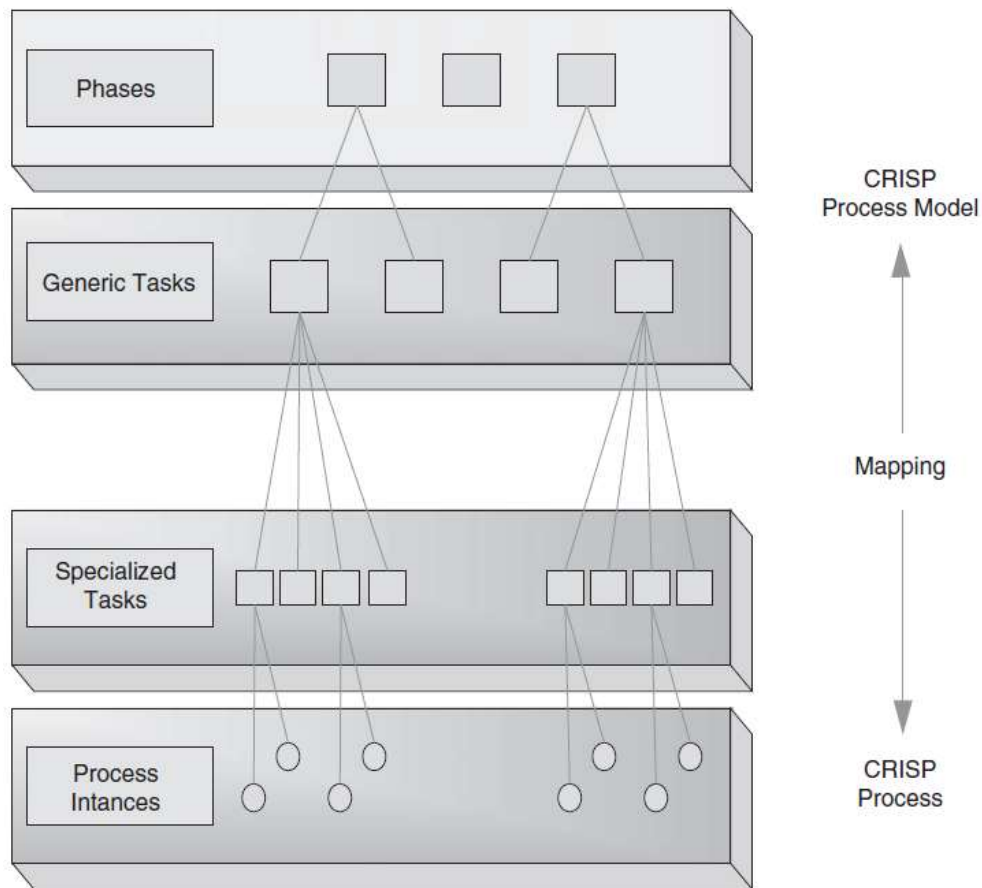
Figure 8 CRIPS-DM Data Model

Each of these three approaches has its own advantages and disadvantages. Some of each element will be combined into a feasible approach as a solution for developing SMEs.

■3.0 PROPOSED APPROACH AND ANALYSIS

As recently technology price has been affordable for SMEs in developing their business, they can use it for data mining for knowledge creation. The proposed approach for SME data mining can be used offline from computer or gadgets, or online as a plug-in using social media. The proposed model for SME data mining is made for giving a sustainable business and initiated innovation in the business.

The proposed approach is a combination of CRIPS-DM and SEMMA. This approach results in a bit wider SEMMA system with specialized process instances as the result in Figure 9.
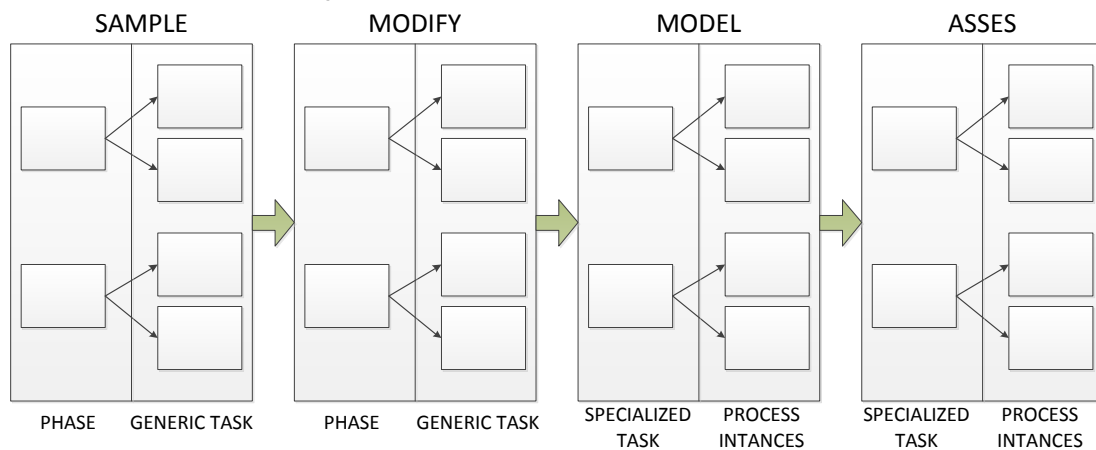

Figure 9 Proposed Approach Combination of SEMMA and CRIPS-DM

Using this approach, SMEs doesn't need a high budget technology to be applied in their own system. The data or information gathering needed can be done by offline system (data have been owned before) or online using social media. The system would consist of linear process operated by one operator, where the sampling phase would be done automatically after the 1st setup, gathering data from historic consumer data and the company social account. No feedback loop included in this system to simplify the process and maintain consistency of the SOP.

The modifying and modelling process is where the operator (most probably also one of the owners) has the most influence. The default model of the software probably did not match everyone needs, and thus some creative and analytic thinking from the operator (as the analyst) is needed. In modifying phase, redundant and unimportant data should be purged to build a clean and easy to analyze the model.

The assess process would also be the transformation phase where the knowledge actually used by the company into useful decision. For small company most of decision regularly involved marketing, production and product design. The effect of data mining on financial modelling for small business is usually not that big.

The cost estimation of this system takes around Rp. 50 Million for the initial cost, which still less than 50% of most small business capital by definition (Dep. Koperasi, 2014). The details of the cost for implementing data mining can be seen in Table 1. The SAS cost taken from www.sas.com, hardware cost from bhinneka.com and connection cost from speedy.com (14 Nov 2014).

Table 1 Cost of Implementing Data Mining

	Initial Setup Cost (IDR)	Monthly Regular Cost (IDR)
SAS program	38,000,000	1,000,000
Officer training	5,000,000	4,000,000
Hardware cost	15,000,000	
Connection cost	300,000	300,000
TOTAL Cost (initial & / month)	**58,300,000**	**5,300,000**

The IRR (Internal Rate of Return) would be varied depending on the profit of the business, thus modelling the return rate would be pointless. Based on the implemented model, it is hoped that the result would be an affordable model for small business in Indonesia to be used as a tool for gathering information to develop their business.

■4.0 CONCLUSION

The proposed concept for SMEs in Indonesia will be the combination of CRISP-DM and SEMMA, and gives wider SEMMA system with specialized process instances. Using this approach will give lower cost, so that it will be affordable for most SMEs, which can be useful for them. Finally, by applying this approach on the data mining, it is hoped that more SMEs will implement the data mining to improve their performance.

References

Bramer, M. (2007). *Principles of Data Mining*. London: Springer-Verlag London Limited.
Cabena, P., et al. (1997). *Discovery Data Mining From Concept to Implementation*. Prentice Hall.
Kementrian Koperasi dan Usaha Kecil Menengah Indonesia (2014). *Perkembangan Data Usaha Mikro Kecil Dan Menengah Indonesia (UMKM) Dan Usaha Besar (UB) Tahun 2011 – 2012.*<Assessed from www.depkop.go.id on 10 Nov 2014>.
Mariscal, G., Marba, O., & Fernandez, C. (2010). *A Survey Of Data Mining And Knowledge Discovery Process Models And Methodologies*. Cambridge: Cambridge University Press.
Reimer, U. & Karagiannis, D. (2006). *Practical Aspects of Knowledge Management*. Springer Publishing.
Stankosky, M. (2005). *Creating the Discipline of Knowledge Management The Latest in University Research*. Elsevier Books.
Tambunan, T.T.H. (2006*). Development of Small & Medium Enterprises in Indonesia from the Asia-Pacific Perspective*. Jakarta: LPFE-Usakti.