Advanced Data Mining of SSD Quality Based on FP-Growth Data Analysis

Jieh-Ren Chang 1, You-Shyang Chen 2,*, Chien-Ku Lin 3,* and Ming-Fu Cheng 1

1 Department of Electronic Engineering, National Ilan University, Yilan City 260, Yilan County, Taiwan; jrchang@niu.edu.tw (J.-R.C.); richard.cheng@mphasis.com (M.-F.C.)
2 Department of Information Management, Hwa Hsia University of Technology; New Taipei City 235, Taiwan
3 Department of Business Management, Hsiuping University of Science and Technology; Taichung City 412, Taiwan
* Correspondence: ys_chen@go.hwh.edu.tw (Y.-S.C.); g9923808@yuntech.edu.tw (C.-K.L.)

Abstract: Storage devices in the computer industry have gradually transformed from the hard disk drive (HDD) to the solid-state drive (SSD), of which the key component is error correction in not-and (NAND) flash memory. While NAND flash memory is under development, it is still limited by the “program and erase” cycle (PE cycle). Therefore, the improvement of quality and the formulation of customer service strategy are topics worthy of discussion at this stage. This study is based on computer company A as the research object and collects more than 8000 items of SSD error data of its customers, which are then calculated with data mining and frequent pattern growth (FP-Growth) of the association rule algorithm to identify the association rule of errors by setting the minimum support degree of 90 and the minimum trust degree of 10 as the threshold. According to the rules, three improvement strategies of production control are suggested: (1) use of the association rule to speed up the judgment of the SSD error condition by customer service personnel, (2) a quality strategy, and a (3) customer service strategy.

Keywords: data mining; association rule; solid-state drive; quality

1. Introduction

At present, the hard disk drive (HDD) [1] and solid-state drive (SSD) [2] are the main storage devices in the computer industry. The structure of the HDD allows one to access data through the read-write mode between the magnetic read-write head and the disk and, then, store the data on the magnetic material on the disk. The surface of such disks is easily scratched by shaking and can result in functional damage. The SSD, also known as the electronic hard disk, can reduce the damage of components caused by shaking due to its structure of electronic components, but there are still cases of data loss caused by error correction in not-and (NAND) [3] flash memory damage. Unlike the HDD, the SSD mainly stores data by the flow of electrons in NAND flash memory, which is also the key to the SSD being able to read and write faster than the HDD. Furthermore, the SSD also appears to be gradually replacing the HDD, so how to manage the quality of the SSD and how to serve customers are becoming important management issues.

In order to improve the quality and storage capacity, SSD manufacturers continue to develop new types of memory, such as resistive random-access memory (RRAM) [4], parallel random-access memory (PRAM), and static random-access memory (SRAM) [5]. In addition, currently, the SSD in the industry still uses NAND flash memory. The development of R&D capability has enabled the change from the original single-level cell (SLC) and multi-level cell (MLC) to the mainstream trinary-level cell (TLC) and the next-generation quail-level cell (QLC) [6] so as to meet the market price requirements. The goals regarding the process improvement of NAND flash memory are to reduce the overall cost of the SSD and increase the storage capacity, but the limitation of the “program and erase” cycle (PE...
cycle) [7] on NAND flash memory means that no effective improvement can be achieved; thus, the durability of the SSD remains questionable.

Apriori is an algorithm for frequent itemset mining (FIM) and association rule mining of itemsets in mass transactional data [8]. The frequent pattern growth (FP-Growth) is an efficient, scalable, convenient, and improved algorithm of the Apriori algorithms for the mining of a complete set of frequent patterns (FP) by using an extended prefix-tree structure that allows compressed and crucial information about frequent patterns to be stored. The FP-Growth algorithm [9–12] presents a unique way to identify frequent itemsets without using candidate generations. Some current data mining applications, such as web usage mining and promoting education, are implemented by the FP-Growth algorithm, thus improving performance [13,14].

This study collects and analyzes the data of SSD maintenance around the world through computer company A and uses data mining technology and algorithms to identify the data rules of SSD error so as to improve the quality and service strategy of customer service personnel. The goals of this study are as follows: (1) identification of the data rules of SSD error; (2) quality improvement; (3) determination of the direction of customer service strategy; (4) optimization of “parts management” and “supply chain distribution”.

2. Literature Review

This section describes the relevant literature, namely, studies on the industry background, SSD, and the association rules mining for errors detections.

2.1. Industry Background

Advances in information technology and user requirements for the efficiency of computer products continue to increase, such as the central processing unit (CPU), power supply unit, hard disk, SSD, network card, or sound card, among which the demand for storage devices [15] is the highest, regardless of the speed of reading and writing or the capacity of storage. Major computer manufacturers, such as Hewlett Packard, Lenovo, Dell, Apple, Acer, and Asus, have upgraded their storage devices from the HDD to the new SSD to meet the needs of customers. The SSD adopts the read-write technology of semiconductor components. This is different from the HDD technology, which works with the magnetic principle and has limitations in regard to read-write technology. Not only is it unrestricted by the mechanical structure and disk speed limit of the HDD, but it also uses the read-write mode of semiconductor components to provide the SSD with better random read-write performance when reading or writing data. Although SSD manufacturers have been working on the development of NAND flash memory, the limitation of the PE cycle has not been improved.

2.2. SSD

The SSD, also called the electronic hard disk, uses semiconductor technology to store data in NAND flash memory. This is different from the HDD, which works with many mechanical structures to save data by using magnetic technology. Technically, the SSD is more suitable for notebook computers or various portable electronic products, as it allows them to take advantage of the low power consumption, small size, fall resistance, and shock resistance. The main components of the SSD are as follows:

- **SSD controller:**
  As the main core of the SSD, the SSD controller plays a role in maintaining the functional operation and adjustment of all chips on the SSD and acts as the communication bridge between the SSD and computer. Different SSD controllers [16] have different computing capabilities; the mainstream chips at present are Marvell, Silicon Motion, Phison, and JMicron.

- **NAND Flash Memory:**
NAND flash memory [17], the heart of the SSD, can retain stored data and information in the event of power failure, since it is a kind of nonvolatile memory. The HDD works with the magnetic principle to store data on the disk, while the SSD uses the current to store electrons in NAND flash memory to save data.

- Types of NAND flash memory:

NAND flash memory can be divided into four types according to the electronic structure: SLC [18], 1 bit/cell; MLC [19], 2 bit/cell; TLC [20], 3 bit/cell; QLC [21], 4 bit/cell. The sequence is SLC > MLC > TLC > QLC in terms of storage speed and SLC > MLC > TLC > QLC in terms of cost. At present, the mainstream SSD in the consumer market is the TLC chip, and the QLC chip has also been launched, which can provide more capacity and accelerate the replacement of the HDD.

- Advantage and disadvantage of the SSD:

Advantage: The SSD has the advantages of good shock resistance, low power consumption, low noise, and low heat, thereby allowing the data to be stored in it more safely, and the read-write performance is better than that of the HDD. Disadvantage: The disadvantages of the SSD are its high cost and limited PE cycle; that is, the speed of the SSD gradually slows down with the increase in PE cycle, and it cannot be recovered after being damaged. The data stored inside are almost lost when failure occurs. The bottleneck of the current SSD is represented by the NAND flash memory and SSD controller technology.

Kim et al. [22] explain that the input–output (I/O) bus on the host terminal of the computer system is limited; thus, the real read-write speed of the SSD cannot be obtained. Therefore, in-storage processing is proposed, and a method of operating the host CPU directly in flash memory of the SSD is designed to improve and accelerate the read-write capability of the main database to full efficiency. Schroeder [23] analyzed more than 130 million units of hard disk failures in the data center during actual operation and made the following observations: (1) the increase in the hard disk environment temperature is more likely to cause breakdown of the hard disk; (2) the main faults are caused by shaking; (3) the failure rate observed on the field side of the hard disk is higher than that provided by the spec at actual delivery. The SSD has gradually become the mainstream storage device, and most of the related papers discuss how SSD quality can be improved and how read-write performance can be enhanced. In this study, data mining is used to solve the problem and propose strategies to improve the quality management of the SSD by computer manufacturers.

2.3. The Association Rules Mining for Errors Detections

Apriori algorithm and FP-Growth algorithm are often used in batch transaction data for frequent transaction set mining and association rule mining. The FP-Growth algorithm implements some current data mining applications, such as detection of intrusion, web usage mining, promotion of education, and so on, thereby improving performance. More related applications that use association rules to find error detection are as follows:

Detection of intrusion: Intrusion detection is an essential professional business area. The main requirement is to develop a system that can process a large amount of network data to detect attacks more accurately and actively. Without reducing the number of features, it will be more difficult to detect attack patterns in the data and generate association rules for prediction or classification. Data mining is used to clean, classify, and inspect large amounts of network data. Due to the need to handle a large amount of network traffic, different data mining techniques (such as clustering, classification, and association rules) are very useful for analyzing network traffic. The application of association rule technology to design an effective intrusion detection system and effectively identify known and unknown attack patterns is an important research topic [24,25].

Web usage mining: Web usage mining is a method of mining user browsing and access patterns. Use website click data to capture the identity or source of web users and their behavior on the website. The FP-Growth algorithm and the Apriori algorithm...
are used to classify user behaviors to identify patterns of browsing and navigation data of internet users. Through this technology, to discover frequent user patterns and user behavior in web logs. Both algorithms are helpful in analyzing website usage patterns and the characteristics of user behavior. This can be used to enhance web design, introduce personalized services, and promote more effective browsing [13]. By applying the web usage mining method to the clickstream data recorded in the server-side log file, the online consumer behavior on the e-retailer’s website can be found. Through the combination of step graph visualization and sequential association rule mining innovative methods, through interactive measures and discovery of general navigation pattern sequence to characterize online consumer behavior. After the use of association rule technology, the results proved the cross-platform online browsing behavior in the analysis of electronic retail environment. Certain order rules are related to the increased likelihood of purchases in a conversation between mobile devices and PC devices [26].

Application of education: The campus is growing faster and more diverse. Every university hopes to obtain a larger number of outstanding students in promoting the development of the university. By using the existing techniques in data mining, a variety of methods can be used to determine the promotion strategy. Use FP-Growth algorithm for research, because it has the concept of tree construction when performing itemsets search. A software system for determining educational promotion strategies using this method was developed [13]. A collaborative educational data mining tool based on association rule mining has been developed. It is used to continuously improve e-learning courses and allows teachers with similar course profiles to share and rank the discovered information. The mining tool is intended for non-expert lecturers to use in data mining, so its internal operations must be transparent to users, and lecturers can focus on the result analysis and make decisions on how to improve e-learning courses. The data mining tools are described in the form of tutorials, and association rules can be found in web-based adaptive courses [27].

3. Related Work

This section describes the relevant analysis techniques on the data mining and association rule and its algorithm.

3.1. Data Mining

Data mining [28] can be divided into three stages: data collection, data processing, and data presentation.

- Data collection: the problem is defined, relevant data are collected, and analysis objectives are set.
- Data processing: data are organized and collected; useful information is identified and analyzed with appropriate tools and models.
- Data presentation: the program of the next step for decision making with the results of data analysis is developed.

Data need to be processed and given meaning in order to obtain valuable information, and information can be transformed into knowledge and then into decision making after being digested. Data mining can be divided into supervised and unsupervised methods. Supervised data mining mainly works with the top-down analysis method that uses models to establish and describe the relationship between a specific target variable and other variables, which can be then determined to be classification or prediction by problem type: classification is represented by the observation of a large amount of data, establishment of a class model from the rules or obtained characteristics, and the determination of the attributes in the data by category; prediction is represented by the prediction of possible future behaviors with historical data. The classification and prediction of the supervised data mining can be addressed by machine learning and deep learning. Unsupervised data mining mainly adopts the bottom-up analysis method—where there is no specific target variable marked—to find the relationship between each variable in the whole data cluster.
The user judges the importance of the results after the model is explored and establishes data in the data cluster to describe the relationship between a specific target variable and other variables. Then, unsupervised data mining can be further classified into the clustering and association rules by problem type: clustering is represented by the division of data into different clusters according to the degree of data similarity, meaning that individuals in the same cluster have a closer distance or smaller variation, and different clusters have more individual distance and larger variation. The association rule is represented by the exploration of the association of fields with a large volume of data.

The key step in the data mining process is data preprocessing, which includes data collection, cleaning, feature selection, and processing. The problems, causes, and data preparation steps of data preprocessing are shown in Table 1.

| Data Preparation Steps | Cause                                      | Problem               |
|------------------------|--------------------------------------------|-----------------------|
| Data integration       | Differences after the integration of different source data | Inconsistent data     |
|                        | Data or fields with the same meaning appear | Redundant data        |
|                        | The value of data is out of the reasonable range | Incorrect data        |
|                        | Field or value of a duplicate record       | Duplicate data        |
| Data cleaning          | Characteristics of data, improper measurement, or data input error | Outliers              |
|                        | Data omission caused by measuring equipment or human factors | Missing value         |
|                        | Error of data or deviation of data input   | Noise                 |
| Data conversion        | The data format does not meet the assumptions of mining tool | Inappropriate data scale |
|                        | Data or dimension too high                 | Too much data         |
| Data reduction         |                                            |                       |

### 3.2. Association Rule and Its Algorithm

Searching for hidden relationships between items from large-scale datasets is called association analysis. The key research approaches of association analysis include FIM and association rules. On one hand, FIM is a collection of items that often appear together. Given the past superiority, FIM has an extensive application field from the latest researches, such as extracting useful knowledge from event logs [29], using sliding windows capable of extracting tendencies from continuous data flows [30], directing membrane separator development for microbial fuel cells [31], assisting healthcare researchers in identifying medications and medication combinations that associate with a higher risk of acute kidney injury (AKI) using electronic medical records (EMRs) [32], protecting the cloud servers’ privacy of datasets from frequency analysis attack [33], and applying the FP-Growth algorithm in determining frequent itemsets in association with data mining to find customer spending habits in buying goods simultaneously [34]. On the other hand, association rules analyze the possible strong relationship between two items. In particular, both Apriori and FP-Growth are algorithms used in itemset of association rule mining in FIM and a large amount of transaction data.

The market basket analysis is a typical task for the association rules of mining, and it is the most widely used method, developed by Agrawal [8] and others at the IBM Almaden research center to identify relationships in a large amount of data. The advantage is that it can easily explain the generated rules and completely show the cross-effect of variables. However, the selected rules and conditions or set are very important; that is, they may produce cumbersome and messy results if the criteria is set slightly loose, or a relatively small number of variables may be ignored with unparsed analysis result being produced. The most common analysis data are represented by the transaction database, as it attempts to identify the potential association rules from the seemingly relevant, but not identical, customer transaction records, among which the most well-known example is the association rule, which was used to analyze the relationship between beer and diapers in American hypermarkets.
The algorithm of association rule is composed of the search data method, calculated items, and support degree, and the features of the association rule algorithm are described in Table 2 below.

Table 2. Introduction of association rule algorithm.

| Algorithm  | Main Features                                                                 | Disadvantage                                                                 | Search Method | Data Configuration |
|------------|-------------------------------------------------------------------------------|------------------------------------------------------------------------------|---------------|--------------------|
| Apriori    | The candidate itemsets are generated repeatedly to identify all of the high-frequency ones, and the rules are then deduced. FP-Growth is the deductive basis of the algorithm, and it can deal with a large amount of data effectively, thereby surpassing the limitations of Apriori. | Database needs to be searched repeatedly, and I/O time needs to be spent. | Breadth first  | Horizontal data configuration |
| FP-Growth  |                                                                               | More processing time and storage space are needed to store FP trees.          | Depth first   | Horizontal data configuration |

- **Apriori**: The Apriori algorithm is one of the classic algorithms of the association rule in the field of data mining, and it was put forward by Agrawal and Srikant [9] in 1994. The Apriori algorithm is used to deal with transaction information databases, such as lists of goods purchased by customers or lists of frequently visited web pages. The algorithms of other types of relational rules are more inclined to identify the relationship between data without transaction information or without a time stamp. The Apriori algorithm is used to search the dataset in the principle of “breadth first”, then calculate and analyze the candidate subsets with a tree structure.

- **FP-Growth**: FP-Growth was mentioned in the introduction of the prior algorithm, which needs more time and space complexity, since it searches all of the data many times before producing frequent subsets and generates a large number of candidate subsets at the same time; this is why the efficiency of the prior algorithm is low when it faces huge datasets.

- **Important parameters of the association rule**: There are three important parameters in the association rule, namely, support, confidence, and lift. The first parameter, support, represents the frequency of an item appearing in all shopping carts. If the relationship degree of two commodities A and B is analyzed, it represents the ratio of these two items in all transaction records. The greater the relationship, the more important the association rule index is.

The formula is as follows: support (A⇒B) = P(A∩B)

The second parameter is confidence. Taking the shopping basket as an example, the conditional probability of purchasing commodity A and commodity B can be used to measure the accuracy of the association rule.

The formula is as follows: confidence (A⇒B) = P(A∩B)/P(A)

The third parameter is lift, which is used to measure the impact of the association rule on commodity frequency. Only association rules with a gain greater than 1 make sense.

The formula is as follows: lift (A⇒B) = P(A|B)/P(B)

It can be seen from the calculation method that when the lift value is greater than 1, there is a positive relationship between A and B; however, some association rules are useless that can be erased when its lift value is insufficient. How the minimum degree of support and confidence for the data to be analyzed are set is the key to identifying the association rules. In this study, FP-Growth is used to calculate association rules. With this method, a set of tree structures is constructed, and it only needs to scan the database twice in the mining process so as to reduce a lot of I/O time. The minimum degrees of confidence and support are set as 90 and 10, respectively, to ensure that the association rules generated have a positive relationship. The main steps are divided into two stages: (1) establishing the FP tree; (2) mining the FP tree.
In the FP tree establishment step, [35] the database is scanned to identify the high-frequency itemsets that meet the minimum degree of support, and they are arranged according to the support degree, as shown in Table 3.

### Table 3. FP tree establishment steps.

| Transaction ID | Commodity Transaction Items | High-Frequency Items Sorted According to Support Degree |
|----------------|-----------------------------|--------------------------------------------------------|
| 1              | A, B, C, D, E, F, G         | {B}, {A}, {F}                                          |
| 2              | B, C, D, F                  | {B}, {C}, {F}                                          |
| 3              | A, B, C, F                  | {B}, {A}, {C}, {F}                                     |
| 4              | A, B, C, G                  | {B}, {A}, {C}                                          |

After the database is scanned, the support degree of each itemset is compared with the minimum support degree, and the itemset with an insufficient threshold value is be deleted, as shown in Table 4.

### Table 4. FP tree establishment steps.

| Itemsets | {A} | {B} | {C} | {D} | {E} | {F} | {G} |
|----------|-----|-----|-----|-----|-----|-----|-----|
| Support degree | 0.75 | 1   | 0.75 | 0.5 | 0.25 | 0.75 | 0.25 |

Firstly, the root node of the FP tree is established and marked as the spatial point, and then, the database is scanned again to add the transaction records of high-frequency itemsets to the FP tree in the order of Tables 3 and 4.

### 4. Results and Discussions

#### 4.1. Implementation Process of the Experimental Method

The information of this study was collected from a well-known multinational technology company, company A, in Taiwan, with more than 8000 items of SSD repair data from 2016 to 2018. These data were checked and preprocessed; then, the association rule of the SSD error status was calculated by a data analysis tool; finally, the quality improvement direction and customer service strategy were obtained. The experimental research process is shown in Figure 1.

#### 4.2. Data Preprocessing

There are 28 kinds of data attributes in the customer return SSD data attribute table of this study. Too many data attributes are not conducive to the induction of error conditions and rules; thus, data preprocessing, data integration, and cleaning are needed. The steps and results of data preprocessing are shown in Table 5.

### Table 5. Original data attribute table of the customer-returned SSD.

| Item | Details | Item | Details |
|------|---------|------|---------|
| 1    | Ship to region | 15   | Defect part number |
| 2    | Ship to country code | 16   | Key component (KC) vendor |
| 3    | Ship to country name | 17   | Big class |
| 4    | Case ID | 18   | Project |
| 5    | Repair national state (NS) | 19   | Platform |
| 6    | Repair regional office (RO) | 20   | System symptom code |
| 7    | Customer call help date | 21   | Description |
| 8    | Customer call help week code | 22   | Action code |
| 9    | Reaired date | 23   | Action code description |
| 10   | Repaired week code | 24   | Field replacement unit (FRU) error code |
| 11   | Serial No | 25   | FRU error code description |
| 12   | FG (finish good) vendor | 26   | Manufacturing (MFG) code |
| 13   | MFG week code | 27   | Category |
| 14   | MFG date | 28   | Failure period |
Figure 1. Flow chart of the experimental research. (Note: ODM is abbreviation of original design manufacturer).

- **Step 1**: The data are classified by attributes to reduce data duplication. Since the fields of ship-to region and repair RO are both the continents of SSD delivery, only the former is reserved because the contents are the same. Moreover, the ship-to country name is equal to the repair NS, both of which are country names with the same contents, so only the ship-to country name field is reserved. The ship-to country code is the code without analysis significance.

- **Step 2**: The data format is converted so as to reduce duplicate data information. The customer call help date is the repair date of the customer, and maintenance activities are currently carried out on the same day, so it is equal to the actual repair date. Moreover, the customer call help week code, repaired week code, MFG week code, and MFG code, which are the customer repair week, actual repair week, manufacturing week, and manufacturing code, respectively, are all internal codes. These are too divergent, and only the field attribute “item 28 failure period” is reserved. The failure period is the time interval of the SSD failure, and the formula of the failure period is the customer call help date minus the MFG date with the unit of days.

- **Step 3**: Data reduction can solve the problems of too much data and too high a dimension. The KC vendor field is the seller and field of components, and big class is the code of the CPU. The difference between these two is included in the project field, and the platform field is the computer platform. Therefore, the attribute of the project and platform are reserved for analysis.

- **Step 4**: Data cleaning and data conversion are reserved for the case in which the data are noisy or the data format is not suitable for analysis. Case ID and serial no are both serial numbers and have no analytical significance. The system symptom code is a code, while the system symptom code description is used to collect customer complaints. Due to language differences and bias of various narrative problems, the record of this field is too divergent and will temporarily not be include in the...
analysis. The system symptom code field is a code without analytical significance. The action code is an internal service code, and the action code description is the action initially performed by service personnel, most of which being directly exchanged for spare parts. The FRU error code is the code of a broken SSD, and the FRU error code description is the situation of SSD failure, which is helpful for the classification of SSD status in this study and is, therefore, reserved. Category is the record of the category part, but as all of the data are SSD type, it is not considered. Only item 25 is reserved.

- Results of preprocessing steps 1 to 4: after data processing, 28 data attributes are simplified to six; descriptions of the six data attributes after preprocessing are shown in Table 6.

**Table 6. Description of six data attributes.**

| Item | Details | Explain |
|------|---------|---------|
| 1    | Ship-to region | Shipping continents, five in total |
| 2    | Ship-to country | Shipping countries, 60 in total |
| 3    | Project | Computer cases, 60 cases in all |
| 4    | Platform | Motherboard differences, a total of 110 kinds |
| 5    | FRU error code description | Broken products, total five kinds |
| 6    | Failure period | Period of product failure, 14 days as a quality cycle, a total of 10 sections |

The proposed algorithm is described with the correlation analysis for the data preprocessing systematically and step-by-step. Assuming data collection can be viewed as an information system IS = (Y, A, T, n), where Y = {y_t | y_t = (a_{1t}, a_{2t}, ..., a_{nt}), t = 1, 2, ..., T} is the original dataset, which has T data objects from the business market, A = {A_i | i = 1, 2, ..., n} is the attribute set, which includes n attributes, a_{it} belongs to attribute A_i value domain. The whole procedure is described as the following pseudo codes:

```plaintext
pre-processing function = pre-proc (IS)
IS = cancel the duplicate data (IS);
IS = extract the data in the normal range (IS);
IS = data format is converted (IS);
IS = cancel the noise data (IS);
IS = cancel the duplicate attributes (IS);
return (New_IS = IS);
```

By this pre-processing function, IS information system is changed to New_IS information with different parameters for extracting association rules.

**4.3. Data Analysis**

The data analysis tool classifies the data after preprocessing into five kinds of error conditions and organizes the association rule accordingly. It then converts the numeric attributes into category attributes in the preprocessing step. The customer return date after shipment is divided into a 14-day period by time interval, which is for the purpose of matching the quality analysis process of the company. There is a quality review every 14 days before the data analysis is converted into a clear company policy model.

**4.4. Results of the Association Rule**

According to the analysis results, SSD error conditions can be divided into five types: OTHER (to be further clarified), NOT RECOGNIZED, OTHER STORAGE DEVICE NOT RECOGNIZED, DEFECTIVE SECTORS, and MEMORY CARD FUNCTION FAILURE. The association rule obtained from the experiment is based on the minimum support degree of 10, the minimum confidence degree of 90, and the lift value greater than one. Since the original association rules are generated through different frequent itemsets, considering the most interesting rules to be discussed, only the lists with highest frequent itemset association rules are presented in the following sections for five different error conditions.
4.4.1. Association Rule with the OTHER Error Condition

The results of the quadrinomial association rule by using the data analysis tool to explore the OTHER SSD error are shown in Table 7.

Table 7. Quadrinomial association rule with the OTHER error condition.

| Association Rule by Item 4 (OTHER) | Support | Confidence | Lift |
|-----------------------------------|---------|------------|------|
| Chile, Project = Drift SK, PA     | 24.3    | 1          | 1    |
| Project = Drift SK, Platform = R5371T, Chile | 17.4    | 1          | 1    |
| Project = Drift SK, Platform = R5371T, PA | 17.4    | 1          | 1    |
| Platform = E5-575, Ironman, AAP   | 13.7    | 1          | 1    |
| Chile, Failure period = 29–42, PA  | 10.3    | 1          | 1    |

4.4.2. Association Rule with the NOT RECOGNIZED Error Condition

The trinomial association rule with the NOT RECOGNIZED SSD error condition is shown in Table 8.

Table 8. Trinomial association rule with the NOT RECOGNIZED error condition.

| Association Rule by Item 3 (NOT RECOGNIZED) | Support | Confidence | Lift |
|---------------------------------------------|---------|------------|------|
| Project = Captain_SK, Platform = F5-573G    | 14.3    | 1          | 1    |
| Project = Ironman, Platform = E5-575G       | 13.8    | 1          | 1    |
| Country = China, Failure period = 15–29     | 13.0    | 1          | 1    |
| Country = China, Failure period = 42–55     | 10.6    | 1          | 1    |

4.4.3. Association Rule with the OTHER STORAGE DEVICE NOT RECOGNIZED Error Condition

The trinomial association rule with the OTHER STORAGE DEVICE NOT RECOGNIZED SSD error condition is shown in Table 9.

Table 9. Trinomial association rule with the OTHER STORAGE DEVICE NOT RECOGNIZED error condition.

| Association Rule by Item 3 (OTHER STORAGE DEVICE NOT RECOGNIZED) | Support | Confidence | Lift |
|------------------------------------------------------------------|---------|------------|------|
| EMEA (Europe, Middle East, and Africa), Germany                   | 21.4    | 1          | 1    |
| EMEA, Failure period = 42–55                                      | 19.3    | 1          | 1    |
| EMEA, Failure period = 29–42                                      | 11.3    | 1          | 1    |
| EMEA, Failure period = 15–29                                      | 18.1    | 1          | 1    |
| EMEA, Netherlands                                                 | 17.1    | 1          | 1    |
| EMEA, Failure period = 55–68                                      | 16.9    | 1          | 1    |
| EMEA, Project = Ironman                                           | 16.4    | 1          | 1    |
| EMEA, Country = Spain                                            | 14.8    | 1          | 1    |
| Platform=E5-573G, Project = Ironman                              | 12.2    | 1          | 1    |
| EMEA, Country = France                                           | 11.0    | 1          | 1    |
| EMEA, Failure period = 68–82                                     | 10.8    | 1          | 1    |

4.4.4. Association Rule With the DEFECTIVE SECTORS Error Condition

The trinomial association rule with the DEFECTIVE SECTORS SSD error condition is shown in Table 10.
4.4.5. Association Rule with the MEMORY CARD FUNCTION FAILURE Error Condition

The quadrinomial association rule with the MEMORY CARD FUNCTION FAILURE SSD error condition is shown in Table 11.

Table 11. Quadrinomial association rule with the MEMORY CARD FUNCTION FAILURE error condition.

| Association Rule by Item 4 (MEMORY CARD FUNCTION FAILURE) | Support | Confidence | Lift |
|-----------------------------------------------------------|---------|------------|------|
| Platform = E5-575G, EMEA, Ironman                        | 25.5    | 1          | 1    |
| Platform = E5-774G, EMEA, Hulk_SK                        | 14.6    | 1          | 1    |

4.5. Improvement to Quality and Customer Service Strategy

From the rules in the previous section, the customer service department can get the reasons of five SSD error conditions, which can improve the service efficiency. Taking into account the limited time and resources, to achieve executable quality improvement, the rules must be further merged or simplified. Since the rules of the more polynomials are derived from the rules of the less polynomials, the various error conditions and the variable with the highest support can be taken as the main quality improvement direction of each error. The quality improvement of five error conditions and the rules for judging the errors are provided for the reference of decision makers as follows:

- Describe the rule antecedent with the consequent NOT RECOGNIZED:
  - If the customer return case is Country = China (AAP);
  - If the customer return time is Failure period = 15–29 days;
  - If the customer return case is Project = Ironman;
  - If the customer return platform = F5-573G.

- Describe the rule antecedent with the consequent DEFECTIVE SECTORS:
  - If the customer return case is Country = China (AAP);
  - If the customer return time is Failure period = 15–29 days;
  - If the customer return case is Project = Captain_SK;
  - If the customer return Platform = F5-573G;
  - If the customer return case is Project = wish_KLS, Platform VX5 591G.

- Describe the rule antecedent with the consequent OTHERS:
  - If the customer return case is Country = Chile (PA);
  - If the customer return time is Failure period = 15–29 days;
  - If the customer return case is Project = Drift_SK;
  - If the customer return Platform = R5-371T;
  - If the customer return case is Project = Captain_SK, Platform=F5-573G;
  - If the customer return country is Chile, Failure period = 29–42, Region = PA.

- Describe the rule antecedent with the consequent MEMORY CARD FUNCTION FAILURE:
  - If the customer return case is Country = Germany;
If the customer return case is Failure period = 15–29 days;  
If the customer return case is Project = Ironman;  
If the customer return case is Platform = E5-575G;  
If the customer return case is Platform = E5-774G, Region = EME, Project = Hulk_SK.

- Describe the rule antecedent with the consequent OTHER STORAGE DEVICE NOT RECOGNIZED:
  If the customer return case is Country = Germany;  
  If the customer return case is Failure period = 42–55 days;  
  If the customer return case is Project = Ironman;  
  If the customer return Platform = E5-575G.

The analysis of the association rule can help customer service personnel to identify the types of error conditions and related quality impact and to track the problems of poor functional verification at the manufacturing end or before shipment. The analytical results of this study provide three improvement strategies (personal service strategy, product quality strategy, and customer service strategy) with key description for company A, with the aim of improving the overall SSD error status and the different SSD error conditions as shown in Table 12.

### Table 12. Summary of three improvement strategies for company A.

| Strategies             | Key description of strategies                                                                 |
|------------------------|-----------------------------------------------------------------------------------------------|
| Personal service strategy | The association rule can be used to accelerate the judgment of the SSD error condition by customer service personnel. |
|                        | (1) From relationship analysis, Ironman and Captain_SK represent the cases that need to be improved. |
|                        | (2) Quality can be improved from the corresponding rules of the case and error condition. |
| Product quality strategy | (3) The platforms that need to be improved are Platform=E5-575G and Platform = F5-573G. |
|                        | (4) Improvements to the corresponding rules of the platform and error condition are addressed. |
|                        | (1) The allocation of customer service personnel and the arrival time of maintenance logistics should be strengthened, and the delivery time should be adjusted to 19–25 days. |
|                        | (2) For the SSD with damaged parts of other storage devices, the period of sufficient logistics supply should be adjusted to 42–55 days. |
| Customer service strategy | (3) From the relationship analysis, it is necessary to strengthen the allocation of customer service personnel and the maintenance logistics in China and Germany. |
|                        | (4) Logistics distribution should be improved as per the corresponding rules of the country and error conditions. |

### 4.6. Research Discussions

This study sets the threshold of lift value 1; the rules extracted are deleted because they are useless when the lift value is less than 1. Conversely, they are useful and are kept
when the lift value is more than or equal to 1. Thus, all of the existing rules have a lift value of 1, which means they are useful. In particular, the value of 1 denotes statistically significant independence between the antecedent and consequent; thus, it can be valuaby discussed in the following statistical utilities and values. First, the lift value of most of the existing rules is 1, and only two of them are far greater than 1 in this study. The lift value of 1 is set for the rule extraction standard. Second, while there are more than 8000 data records for these experiments, there remains a limited amount of data. As researchers in the field of pattern mining, we aim to carefully record all possible rules that may be useful for discussion and subsequent research.

5. Conclusions

This study is based on the SSD error status of company A. There are five SSD error conditions, one of which is OTHERS (to be clarified). If this study can cross different enterprises in the future, or even cooperate with the original upstream SSD factory to analyze more SSD error conditions of OTHERS, more data assistance and accuracy for the improvement of quality and the formulation of error rules can be obtained. Conclusively, two main observations were made in this study: (1) The reason for selecting FP-Growth: FP-Growth is a popular, convenient, and efficient enough algorithm for many applications and has replaced LCM, which is faster and uses more memory. It is important to promote SSD quality with an advanced mining algorithm. (2) Significant contribution with novel results: The support methodology is an efficient and convenient way to promote SSD quality. This represents an important contribution for many researchers and companies in big data analysis. This study is also the first trial with novel application in terms of using the association rule for SSD quality from an industrial perspective, thereby providing the rationale for its novelty. The study results justify the capability of the association rule for data mining application areas.

Although the empirical results have significant industrial application values in this study, there is room for improvements in this research field: (1) data of different industries can be used to test this research method; (2) if lift value is set to 2, it may have different outcomes; (3) the missing values or possible hidden relations between the input variables using typical correlation analysis or suitable statistical tests can be addressed to mine different results to the final set of variables.

Author Contributions: Conceptualization, J.-R.C. and M.-F.C.; methodology, J.-R.C.; software, M.-F.C.; validation, M.-F.C.; writing—original draft preparation, M.-F.C.; writing—review and editing, Y.-S.C. and C.-K.L.; All authors have read and agreed to the published version of the manuscript.

Funding: The authors wish to sincerely thank the Ministry of Science and Technology of the Republic of China, Taiwan, grant numbers MOST 109-2221-E-146-003 for the partial financial support.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Su, C.-J.; Huang, S.-F. Real-time big data analytics for hard disk drive predictive maintenance. Comput. Electr. Eng. 2018, 71, 93–101. [CrossRef]
2. Micheloni, R. Solid-State Drive (SSD): A Nonvolatile Storage System; IEEE Xplore: Piscataway, NJ, USA, 2017; Volume 105, pp. 583–588.
3. Spinelli, A.S.; Compagno, C.M.; Lacaita, A.L. Reliability of NAND Flash Memories: Planar Cells and Emerging Issues in 3D Devices. Computers 2017, 6, 16. [CrossRef]
4. Ielmini, D. Brain-inspired computing with resistive switching memory (RRAM): Devices, synapses and neural networks. Microelectron. Eng. 2018, 190, 44–53. [CrossRef]
5. Yin, S.; Jiang, Z.; Seo, J.S.; Seok, M. XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks. *IEEE J. Solid State Circuits* 2020, 55, 1733–1743. [CrossRef]

6. Kim, J.; Park, J.K. Building Reliable Massive Capacity SSDs through a Flash Aware RAID-Like Protection. *App. Sci.* 2020, 10, 9149. [CrossRef]

7. Yang, B.-J.; Wu, Y.-T.; Chiu, Y.-Y.; Kuo, T.-M.; Chang, J.-H.; Wang, P.-Y.; Shirota, R. Evaluation of the Role of Deep Trap State Using Analytical Model in the Program/Erase Cycling of NAND Flash Memory and Its Process Dependence. *IEEE Trans. Electron Devices* 2018, 65, 499–506. [CrossRef]

8. Agrawal, R.; Imieliński, T.; Swami, A. Mining association rules between sets of items in large databases. *ACM SIGMOD Rec.* 1993, 22, 207–216. [CrossRef]

9. Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago de Chile, Chile, 12–15 September 1994; Morgan Kaufmann. pp. 487–499.

10. Han, J.; Pei, J.; Yin, Y.; Mao, R. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Min. Knowl. Discov.* 2004, 8, 53–87. [CrossRef]

11. Chang, H.-Y.; Lin, J.-C.; Cheng, M.-L.; Huang, S.-C. A Novel Incremental Data Mining Algorithm Based on FP-Growth for Big Data. In Proceedings of the 2016 International Conference on Networking and Network Applications (NaNA), Hokkaido, Japan, 23–25 July 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 375–378.

12. Zeng, Y.; Yin, S.; Liu, J.; Zhang, M. Research of Improved FP-Growth Algorithm in Association Rules Mining. *Sci. Program.* 2015, 2015, 1–6. [CrossRef]

13. Dharmaraajan, K.; Dorairangaswamy, M.A. Analysis of FP-Growth and Apriori Algorithms on Pattern Discovery from Weblog Data. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24 October 2016; 2017; pp. 170–174.

14. Ikhwan, A.; Yetri, M.; Halim, J.; Siahaan, A.P.U.; Aryza, S.; Yacob, Y.M. A novelty of data mining for promoting education based on FP-growth algorithm. *Int. J. Civil Eng. Technol. (IJICTET)* 2018, 9, 1660–1669.

15. Feygas, A.; Akritidis, L.; Bozanis, P.; Manolopoulos, Y. Indexing in flash storage devices: A survey on challenges, current approaches, and future trends. *VLDJ* 2020, 29, 273–311. [CrossRef]

16. Han, L.; Shen, Z.; Shao, Z.; Li, T. Optimizing RAID/SSD Controllers with Lifetime Extension for Flash-Based SSD Array. In Proceedings of the 19th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems, Philadelphia, PA, USA, 18–22 June 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 44–54. [CrossRef]

17. Kang, M.; Kim, Y. Natural Local Self-Boosting Effect in 3D NAND Flash Memory. *IEEE Electron Device Lett.* 2017, 38, 1236–1239. [CrossRef]

18. Ali, K.; Li, F.; Lua, S.Y.H.; Heng, C.-H. Area Efficient High Through-put Dual Heavy Metal Multi-Level Cell SOT-MRAM. *IEEE Trans. Nanotechnol.* 2020, 19, 613–619. [CrossRef]

19. Mittal, S.; Wang, R.; Vetter, J. DESTINY: A Comprehensive Tool with 3D and Multi-Level Cell Memory Modeling Capability. *J. Low Power Electron. Appl.* 2017, 7, 23. [CrossRef]

20. Shao, W.; Sha, J.; Zhang, C. Dispersed Array LDPC Codes and Decoder Architecture for NAND Flash Memory. *IEEE Trans. Circuits Syst. II Express Briefs* 2016, 65, 1014–1018. [CrossRef]

21. Liu, S.; Zou, X.; Wang, B. Quad-level cell NAND design and soft-bit generation for low-density parity-check decoding in sys-tem-level application. *Wuhan Univ. J. Nat. Sci.* 2018, 23, 70–78. [CrossRef]

22. Kim, S.; Oh, H.; Park, C.; Cho, S.; Lee, S.W. Fast, energy efficient scan inside flash memory SSDs. In Proceedings of the 37th International Workshop on Accelerating Data Management Systems (ADMS), Seattle, WA, USA, 29 August–3 September 2011.

23. Schroeder, B.; Merchant, A.; Lagissetty, R. Reliability of NAND-Based SSDs: What Field Studies Tell Us; IEEE Xplore: Piscataway, NJ, USA, 2017; Volume 105, pp. 1751–1769.

24. Lee, W.; Stolfo, S. Data Mining Approaches for Intrusion Detection. In Proceedings of the 7th USENIX Security Symposium, San Antonio, TX, USA, 26–29 January 1998; pp. 26–29. [CrossRef]

25. Agrawal, D.; Agrawal, C. A Review on Various Methods of Intrusion Detection System. *Comput. Eng. Intell. Syst.* 2020, 11, 7–15. [CrossRef]

26. Raphaeli, O.; Goldstein, A.; Fink, L. Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach. *Electron. Commer. Res. Appl.* 2020, 26, 1–12. [CrossRef]

27. Garcia, E.; Romero, C.; Ventura, S.; De Castro, C. A collaborative educational association rule mining tool. *Internet High. Educ.* 2011, 14, 77–88. [CrossRef]

28. Atluri, G.; Karpatne, A.; Kumar, V. Spatio-temporal data mining: A survey of problems and methods. *ACM Comput. Surv. (CSUR)* 2018, 51, 1–41. [CrossRef]

29. Djenouri, Y.; Belhadi, A.; Fournier-Viger, P. Extracting useful knowledge from event logs: A frequent itemset mining approach. *Knowl. Based Syst.* 2018, 139, 132–148. [CrossRef]

30. Fernandez-Basso, C.; Francisco-Agra, A.J.; Martin-Bautista, M.J.; Ruiz, M.D. Finding tendencies in streaming data using Big Data frequent itemset mining. *Knowl. Based Syst.* 2019, 163, 666–674. [CrossRef]
31. Kókók, L.; Dörgő, G.; Bakonyi, P.; Rózsénberszki, T.; Nemestóthy, N.; Bélafi-Bakó, K.; Abonyi, J. Directions of membrane separator development for microbial fuel cells: A retrospective analysis using frequent itemset mining and descriptive statistical approach. *J. Power Sources* 2020, 478, 229014. [CrossRef]

32. Abdullah, S.S.; Rostamzadeh, N.; Sedig, K.; Garg, A.X.; McArthur, E. Multiple Regression Analysis and Frequent Itemset Mining of Electronic Medical Records: A Visual Analytics Approach Using VISA_M3R3. *Data* 2020, 5, 33. [CrossRef]

33. Rong, H.; Liu, J.; Wu, W.; Hao, J.; Wang, H.; Xian, M. Toward fault-tolerant and secure frequent itemset mining outsourcing in hybrid cloud environment. *Comput. Secur.* 2020, 98, 101969. [CrossRef]

34. Asana, I.M.D.P.; Wiguna, I.K.A.G.; Atmaja, K.J.; Sanjaya, I.P.A. FP-Growth Implementation in Frequent Itemset Mining for Consumer Shopping Pattern Analysis Application. *J. Mantik* 2020, 4, 2063–2070.

35. Gatuha, G.; Jiang, T. Smart frequent itemsets mining algorithm based on FP-tree and DIFFset data structures. *Turk. J. Electr. Eng. Comput. Sci.* 2017, 25, 2096–2107. [CrossRef]