University Library Data Analysis to Help Book Collection Procurements using C4.5 Algorithm

(Case Study: Udayana University Postgraduate Library)

Anak Agung Gede Oka Kessawa Adnyana[1*], I Nyoman Darma Kotama[2], Yanu Prapto Sudarmojo[3]

[1,2,3] Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University
[*]agungkez@ unud.ac.id

Abstract Library is one of many requirements for a university to be labeled as qualified university. Udayana University's Library collections is one major concern. To keep a good collections that can fulfill the student and other academic purposes fulfilled. To create an efficient procurement, we need to keep tracks on every transactions that happening, to determine whether it is important to procure certain books from a certain study field by going through number of lending, number of novelty and book counts. Using C4.5 to learn about past procurement we are going to create an effective procurement.

***Index Terms*—C4.5, Graduation, Data Mining, Library.**

I. INTRODUCTION

World Class University in 2025 is a long term vision that Udayana University want to achieve. Within that vision exist some parts of University department that should be organized and optimized. As example the University Postgraduade Program Library that exist from 2011 [1]surving fast growing enviroment and technology until now.

Postgraduade Program Library of Udayana University members is consists of ±9000 students and ±500 lecturers [2]. With that much of demands, the library should be up to date with the book collection to support its member research and studies. The problems in updating book collection is occurs in book procurements process. Book procurements process has 4 main purposes; (1) to update collection as soon as possible, (2) collection accuracy, (3) simplify workload and keep the cost low, (4) keep a good relationship with the collection vendors [3]. In other words, a procurement process in Library involves many process to make a decision so the procurement can be simple and low cost.

Data Mining concept and algorithm have a long history in supporting the business decision. C45 Algorithm is a data mining concept that can be used to data classification. C45 had several success in past research such as algorithm to predict the student graduation and study results [4][5], best lecturer classification [6] and telco customer classification [7]. With this success stories and vast application of the algorithm, we try to propose a data analysis using C45 to help the book procurement decision so it can fulfill the third main purpose of the procurement process; simple and low cost.

II. PURPOSE OF PAPER

We hope this research will lay out a clear blueprint about Library book procurement data that will help the upper management decide which book to prioritize and ignored. Also we hope this blueprint can help future research about University Library, especially Decision Support System that involves data classification.

(p-issn: 2579-5988, e-issn: 2579-597X)

III. LITERATURE REVIEW

A. University Library Book Procurement Process

Book procurements process is library is a part of library function that procures and develop its own books collections in purpose to gather new information[8].This procurements process is usually done in getting a new source of collection, or add more copies into existing collection[9].

Badan Pengawasan Keuangan dan Pembangunan (BPKP) is Indonesia government agency that regulates and supervise about how a government agent should spend and do financing. Udayana University is bound to government, and BPKP stated that there are 2 steps involved in Library procurement process[10].

1. Selection
   In this part selection process is done by the selected expert that already knows what the library needs and should update. What kind of collection that library should update.
2. Ordering
   After all the selected collection updates is decided. Ordering a collection can be done through several kind such as; (1) buying, (2) trading and (3) gifting

B. C4.5 Algorithm Principle

C4.5 algorithm works by generating the initial decision tree throught the training sample set [11]. Decision tree in a form of classifier is made by the algorithm. The decision tree is a structure that consists of two types of node, which are leaf nodes and the decision node. Leaf nodes represents a class, and decision node has a branch and sub tree for each output according to the difference property values [12]. The steps of C4.5 algorithm is represents as follow

- **Step 1:** Set T is a tuple training set for the class tag assuming that *n* output test is chosen, then the training sample set *T* should pe partitioned into subsets $\{T_1, T_2, \ldots T_n\}$. So we can calculate the entropy of the set T (in bits):

$$Info(T) = -\sum_{i=1}^{k} ((freq(C_i, T)/|T|) \times \log_2(freq(C_i, T)/|T|)) \quad (1)$$

- **Step 2:** The training sample set is divided according to a particular property value, and then the information entropy of property *T* is

$$Infox(T) = -\sum_{i=1}^{n} ((|T_i|/|T|) * Info(T_i)) \quad (2)$$

- **Step 3:** The information gain refers to the difference between the original information requirement and the new one. Through Eq. (1) and Eq. (2), we may get a gain standard, that is

$$Gain(X) = Info(T) - Infox(T) \quad (3)$$

- **Step 4:** Although the gain standard is beneficial to construct the compact decision tree, it has a serious flaw that test has great deviation for many outputs situation. So it must be addressed by standardization, that is

$$Split - Info(X) = -\sum_{i=1}^{n} ((|T_i|/|T|) \log_2(|T_i|/|T|)) \quad (4)$$

The new gain standard is:

$$Gain - ratio(X) = gain(X) / Split - Info(X) \quad (5)$$

IV. RESEARCH METHOD

A. Research Methods

To get the best result , we separated our research methods into 5 tasks:

1. Literature Researches
   In this part we gather all the required information about Data Mining Algorithm, tools and technologies that can help us to research this blueprints.
2. Interview with the Library Management
   We interviewed the library management and issuing a database request with range time 3 years.
3. Design and Implementation
   After we gather enough data about C4.5, we calculate and try to figure out how the data will layed out
4. Result Analysis
   After we draw finish with the equation and results, then we analyze it to match the core problems of the Library
5. Conclusion
   In this part we write the conclusion of the research into a research paper.

V. ANALYSIS AND RESULTS

A. C 4.5 Algorithm on Library Data

1 Defining Training Data

We use procurement data combined with subject's current books situations that are described as the table I.

TABLE I
PROCURMENT TRAINING TABLE

	Subject	Count	Novelty	Lend	Proq
1	Excel	505	0	1	Iya
2	Ekonomi	493	7	25	Iya

3	Management	207	23	0	Tidak
4	Hukum	176	15	14	Iya
5	Pendidikan	123	1	0	Iya
6	Chemistry	120	1	14	Iya
7	Accounting	117	55	1	Tidak
8	Bisnis	103	9	0	Tidak
9	Marketing	97	16	2	Tidak
10	Perpajakan	86	21	15	Iya

2 Transforming numerical value into enum value

Transforming each cell into group of enum value, on this research we use STDEV to divide every cells into a group of value. For example, on column *count*, every cell higher than STDEV(*count*) will be considered as many and represented with 1, the other will be considered as few and represented with 0.

TABLE II
PROCUREMENT TRAINING TABLE WITH ENUM VALUE

	Subject	Count	Novelty	Lend	Proq
1	Excel	1	0	0	Iya
2	Ekonomi	1	0	1	Iya
3	Management	1	1	0	Tidak
4	Hukum	1	0	1	Iya
5	Pendidikan	0	0	0	Iya
6	Chemistry	0	0	1	Iya
7	Accounting	0	1	0	Tidak
8	Bisnis	0	0	0	Tidak
9	Marketing	0	0	0	Tidak
10	Perpajakan	0	1	1	Iya

3 first gain ratio

To retrieve gain ratio, we need to retrieve gain from each attribute and find which attribute has the highest gain ratio.

4 Creating the tree

The tree is created by taking the highest gain ratio. The selected attributes will chosen as the tree node and eliminated from the looping process. The other attributes will follow the looping process until the tree well created

5 Continuing the looping

We repeat Step 1 to 4 until the last attribute used, so a full grow decision tree formed.

TABLE II
PROCUREMENT TRAINING TABLE WITH ENUM VALUE

Attr	Val	count	ent	gain	split	ratio
Count	1	4	1.0000			
	0	6	0.9183			
				0.0200	0.9710	0.0206
Novelty	1	3	0.9183			
	0	7	0.8631			
				0.0913	0.8813	0.1036
Lend	1	4	0.0000			
	0	6	0.9183			
				0.4200	0.9710	0.4325

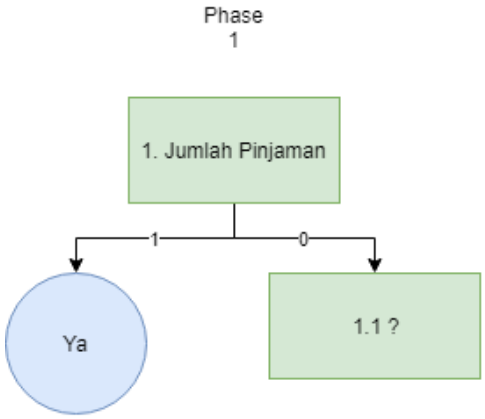

Fig. 1. Phase 1 of C4.5 tree creation

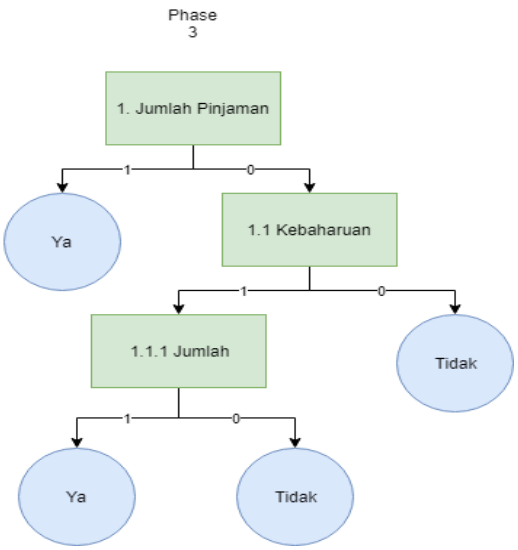

Fig. 2. Phase Final of C4.5 tree creation

VI. Conclusion

After we follow through the rules of C4.5 methods and collected the data from the Library Collections, we successfully created a way to help the book procuring process. From the data we found that three data can be used for the suggestion, (1) Amount of lender / Jumlah Pinjaman, (2) Novelty (Kebaharuan), (3) Amount of books (Jumlah Buku).

VII. Sugestion

We suggest, future study around this topic is about implementing the Data design into a real case, as example Decision Support System.

References

[1] "UNUD | Pascasarjana - Sejarah Program S2 Magister Ilmu Ekonomi." [Online]. Available: https://fe.unud.ac.id/pascafeb/pages/view/sejarah-program-s2-magister-ilmu-ekonomi. [Accessed: 21-May-2019].

[2] "SRV4 PDDIKTI : Pangkalan Data Pendidikan Tinggi." [Online]. Available: https://forlap.ristekdikti.go.id/perguruantinggi/detail/NEFGNzM0OTgtMzgwMy00RjQ5LUFBNjQtQ0FGODJDNTU0Q0NF. [Accessed: 22-May-2019].

[3] G. E. Evans, *Developing Library and Information Center Collections. Library Science Text Series*. ERIC, 1995.

[4] J. H. Jaman, "Prediksi Kelulusan Mahasiswa Dengan Metode Algoritma c4. 5," *Syntax J. Inform.*, vol. 2, no. 02, 2016.

[5] M. A. Sembiring, "Penerapan metode decission tree algoritma c45 untuk memprediksi hasil belajar mahasiswa berdasarkan riwayat akademik," *JURTEKSI R. Vol 3 No 1*, vol. 3, 2016.

[6] S. Lestari and A. S. Karim, "Model Klasifikasi Kinerja Dan Seleksidosen Berprestasi Dengan Algoritma C. 45," *Pros. Sembistek 2014*, vol. 1, no. 02, pp. 340–350, 2015.

[7] A. Zulkifli, "Metode C45 Untuk Mengklarifikasi Pelanggan Perusahaan Telekomunikasi Seluler," *Riau J. Comput. Sci.*, vol. 2, no. 1, pp. 65–76, 2016.

[8] P. Soeatimah, "Kepustakawanan dan Perpustakaan," *Yogyak. Kasinius*, 1992.

[9] S. Basuki, *Pengantar Ilmu Perpustakaan dan Informasi*. Yogyakarta: PT Gramedia Pustaka Utama, 1991.

[10] "Pengadaan Bahan Pustaka." [Online]. Available: http://www.bpkp.go.id/pustakabpkp/index.php?p=pengadaanbahanperpus. [Accessed: 22-May-2019].

[11] L. Dongming, L. Yan, Y. Chao, L. Chaoran, L. Huan, and Z. Lijuan, "The application of decision tree C4.5 algorithm to soil quality grade forecasting model," in *2016 First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, Wuhan, China, 2016, pp. 552–555.

[12] Y. Zhang and L. Gong, *Principle and Technology of Data Mining*. Beijing: Publishing House of Electronics Industry, 2004.