ANALYSIS OF THESIS ADVISING PATTERN IN MASTER PROGRAM ON FACULTY OF AGRICULTURE IPB USING DATA MINING APPROACH

ANALISIS POLA PEMBIMBINGAN TESIS BIDANG PERTANIAN IPB MENGGUNAKAN PENDEKATAN DATA MINING

Oleh :
Lilis Syarifah
Imas Sukaesih Sitanggang
Pudji Muljono
Program Studi Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Pertanian Bogor

Abstrak. Tesis merupakan laporan hasil belajar mahasiswa yang diselesaikan sebagai persyaratan kelulusan untuk program Magister. Memilih topik studi memengaruhi pelaksanaan penelitian. Oleh karena itu, topik penelitian mampu meningkatkan kualitas institusi akademik, namun sejumlah besar dokumen tesis yang tersedia di repositori masih kesulitan dalam mendapatkan informasi yang terkait dengan topik yang sering atau jarang dipelajari sebelumnya. *Association Rule Mining* dapat digunakan untuk menemukan informasi pada item tertentu. Penelitian ini bertujuan untuk menganalisis sistem pola pemberian saran dalam program Magister Pertanian berdasarkan pembimbing dan topik penelitian mereka pada metadata tesis Institut Pertanian Bogor (IPB) dari repositori IPB dan dokumen teks ringkasan menggunakan pendekatan *text mining.* Data dikumpulkan dari situs repositori IPB dan diproses menggunakan bahasa pemrograman R. Hasil pola dari penelitian adalah bahwa asosiasi yang paling populer pada pembimbing terjadi pada nilai dukungan 0,00793 atau setara dengan tujuh tesis dan empat topik populer adalah insektisida botani, pemanasan global, padi, dan perubahan penggunaan lahan. Hasil analisis bisa menjadi informasi yang berguna untuk menjadi referensi bagi penelitian masa depan atau menentukan pembimbing yang tepat pada bidang pertanian.

Kata Kunci: *Apriori Algorithm, Association Rule Mining,* **Institut Pertanian Bogor,** *Text Mining*

Abstract. The thesis is student study report which is accomplished as a requirement of graduation for Master program. Selecting study's topic and advisors influence implementation of the study. Therefore, study's topic is able to improve academic institution quality, however a large number of thesis documents on the repository cause difficulty to get information related to advisor's expertness and the frequent or rare topic is former studied. Association rule mining can be used to mine information on the related item. This study aims to analyze advising patterns system in Master program on Agriculture based on supervisors and their topic research on metadata thesis of IPB repository and text documents of summary using data mining approach. The datas were collected from the repository of Bogor Agricultural University website and processed using R language programming.

Pattern result of the reseach were that the most popular association on supervisor was occurred at support value of 0.00793 or equivalent to 7 theses and four popular topics were Botanical insecticide, Global warming, Upland Rice, and Land Use Change. The analysis result could be useful information to be reference or suggest future research or appropriate supervisor among agricultural.

Keywords: Apriori Algorithm, Association Rule Mining, Bogor Agricultural university, Text Mining

INTRODUCTION

Bogor Agricultural University (IPB) is known as the best university in the agriculture sector in Indonesia based on QS World University Ranking data by Subject Agriculture and Forestry in 2016 (QS, 2016). As leading institution, IPB constantly tries to always keep going and increase university quality and its academic community, especially on the agriculture sector. According to the IPB (2016) statement quality that through a high commitment to the quality, high efficiency, and accountability, IPB produces graduates who are competent agriculture, marine, and tropical biosciences sector for the improvement of the nation. The thesis is one of academic activity that has to be accomplished by a college student in various level of education as graduation requirement such as bachelor degree, master degree and doctoral to produce qualified graduates.

The quality of thesis can reflect the university's quality, thus the information of thesis can be a resource for increasing qualified university. Information of IPB thesis can be obtained from IPB scientific repository as storage media based on a web of thesis and IPB academic community study. According to the Mhunpiew (2013) The success of a thesis or a dissertation for a graduate student relies upon the roles of their supervisor. The right thesis supervisor able to support completing the thesis optimally. Student thesis history can be a reference to gain expertise of lecture which is from student study topic frequently supervised and who has been being the pair supervising often. In addition to the information be found on the master thesis also can be the reference to the university to see the frequent topic being student study object, hence can support university to develop and increase the quality by study program, especially on the agriculture sector.

Information of IPB thesis can be obtained from IPB scientific repository as storage media based on a web of thesis and IPB academic community study and from IPB library archive. The total number of thesis result on IPB scientific repository until July 2017 is 71 793 and for Magister program is 12 477, whereas 1 590 thesis is from Magister thesis in the agriculture sector (PPIP, 2015). On the IPB library data, the number of master thesis student agriculture of faculty is 883. The data from web scientific repository or IPB library achieve can be used to mine the information related to the master thesis student agriculture of faculty. However, the great number of the thesis is the constraint of mining information. Mining the information related to the master thesis student of agriculture faculty that have a huge number can be solved by using data mining method.

Data mining techniques that are widely applied in various study sector are association rule mining and text mining. A structured and unstructured data text

mining able to aid finding accurate information or knowledge and feature on text document (Bhujade *et al.*, 2011). According to Han *et al.* (2012) association rule mining is one of data mining method that can be applied for discovery of correlation relashionship between associated items.

This study analysed thesis supervising pattern in Faculty of Agriculture IPB using data mining method. Analysing was done by applying the apriori algorithm in association rule mining and text mining. The purpose of this study is mining the information related to the frequency of association supervising with the study topic occurred concurrently and the most frequent of topic study being on thesis Faculty of Agriculture. By the information discovered based on analyzing, it is expected to aid giving information related to thesis faculty of agriculture IPB in order to support developing and increasing the quality of the university, especially in the agriculture sector.

LITERATURE REVIEW

Applying association rule mining in thesis document was done by Erman *et al.* (2016) for finding term combination existed frequently and grouping summary thesis document. Khan *et al.* (2015) applied association rule mining technique on Korean text document to extract a useful pattern between hidden word and information. Kulkarni *et al.* (2016)

reviewed the use of association rule mining in text mining to discover knowledge from a set of web documents. A previous research by Setiawan (2016) aims to group summary documents to accelerate searching for IPB postgraduate student thesis document in the IPB repository.

Another related study was also proposed by Shatnawi *et al.* (2014) using association rule mining to find patterns of associations between subjects that have been registered by student in previous semesters. The system has provided a list of association rules that show certain students who choose subjects that are registered by students who have similarity. Angeline (2013) used Apriori algorithm to extract a collection of rules on education institution data and show good result on academic examinations. Angeline (2013) analyzed the data provided to classify students based on academic achievement. This result helped to identify students on average and below average of academic quality level to improve their performance to produce good results.

RESEARCH METHOD

Stage of study shows the phase structure of supervising pattern with topic study in the agriculture sector. Based on Picture 1, this study was started by collecting the magister thesis of Agriculture Faculty, IPB. Thereafter R

language programming was used to analyze the data in order to discover the hidden knowledge and information included. Stage of study is referred in Picture 1.

Data Collection

Data used thesis metadata on Faculty of Agriculture IPB. IPB thesis metadata consists of identity of particular thesis obtained from IPB library and DIDSI (Directorate of Integration and Information System) IPB. Analyzing conducted on this study used magister thesis metadata of agriculture faculty IPB from IPB library on the period of 2001 to 2017 and DIDSI IPB on the period of 1981 to 2016 which is as part of IPB institution managing information system of IPB scientific repository.
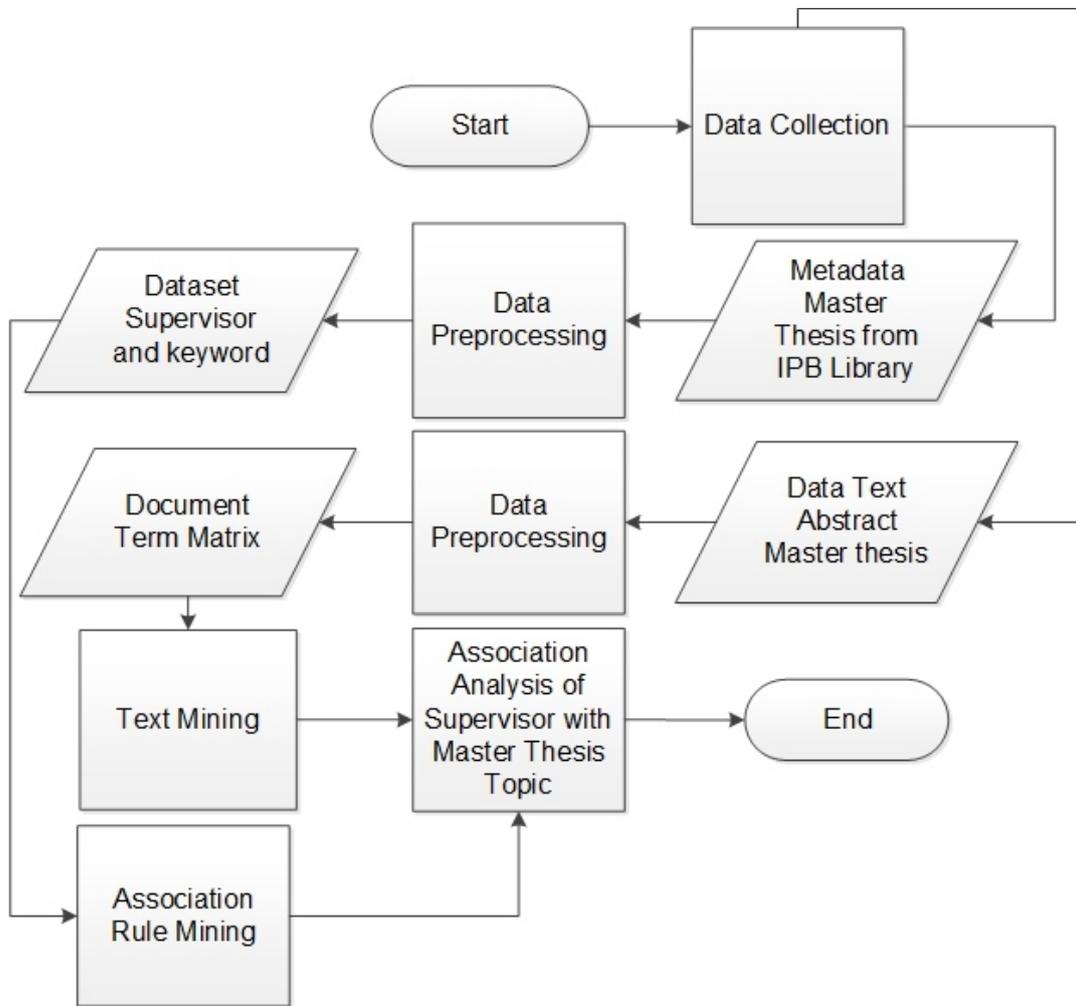
Picture 1
Stage of study preprocessing

Data Preprocessing

Preprocessing is the beginning step to prepare the data before performing data mining technique. The following is several preprocessing conducted on this study:

1) Preprocessing dataset of library metadata thesis
   Data preprocessing was carried out by reducing the data.
2) Data preprocessing of summary text

Data preprocessing for summary text consists of data reduction, data transformation by changing the format file, case folding, remove the punctuation, removing the number, remove the white space, stemming, remove stopwords, make the document-term matrix, and remove sparcity (remove term with existing frequency smaller than the sparcity value from the document-term matrix. Sparsity value is resulted from the calculation of sparcity value from the term. The bigger sparsity value, the number of the existing terms are more.

Association Rule Mining

Association rule mining is the process for discovering information or knowledge on the association between itemsets in transaction databases, relational database and data warehouses (Novitasari et al., 2015). The structure of association rule is IF (antecedence)-THEN (consequent) and for each rule counts its value of support and confident, besides that it also can predict the combination among attribute.

On this study, data as the result of preprocessing from library IPB is dataset including main supervisor, member supervisor and keyword. Association rule mining is utilized to find association rule between the main supervisor with member supervisor and association rule between main supervisor and member supervisor with the keyword. Association rule were generated used an apriori algorithm with specifying the smallest existing value (minsup /minimal support), association rule minimum length (minlen/ minimum length), confident value (conf), and association target.

Text Mining

The result of preprocessing text mining on repository data from DIDSI is documented term matrix. Document-term matrix gives the information of term existing in summary. To obtain information on study topic from the document-term matrix, text mining is conducted to mine existing frequency of term combination coming up on the document-term matrix using an apriori algorithm with association rule mining technique. Mining the frequency appearing of term combination was carried out by determining the target which is the frequency of occurrence over the set word (term), testing the minimum appearance value (minsup / minimal support), and the minimum length of term occurrence frequency (minimum length / minlen).

RESULTS AND DISCUSSION

Data Collection

Data resource of this study is IPB library and DIDSI (Directorate of Integration and Information System) IPB. Library data used are thesis metadata on Faculty of agriculture IPB. The number of data used are 883 theses on the period 2001 to 2017. Based on "The Guideline for Writing Scientific Papers IPB", the rule of writing keyword begins on 2001 which is the first edition of the guideline for writing scientific papers IPB. Therefore, data collection process including keyword information was performed with the agriculture master thesis from 2001. Field of metadata utilized are supervisor and keyword.

Metadata resources from DIDSI have 1368 of agriculture master thesis since 1981 to 2016. The data used are 483 summary on faculty of agriculture. Data summary obtained, later on is saved as the text format file and will be applied on data preprocessing for text mining step.

Library Data

1) Preprocessing

The study process utilizing the IPB library data was started with the preprocessing on master thesis faculty of agriculture IPB. Preprocessing data was done by reducing data process. The data used are main supervisor, member supervisor, and keyword. The result of data reduction was task relevant dataset

for association rule mining method using apriori algorithm.

2) Association of Supervisor

Mining the related information between main supervisor and member supervisor are executed by specifying several parameter values. For minimum frequent pattern value of association rule (minsup/ minimum support), the value tested is 0.004, the minimum confident value is 0.01, the minimum length of the rule (minlen/ minimum length) is two, and the expected result object is association rule. The result of association rule mining is 15 rules as shown in Table 1.

Table 1
Association Of Supervisor

Rule	Main Supervisor	Member Supervisor	Support	Number of Thesis
R1	NK	SWA	0.00793	7
R2	D	DP	0.0068	6
R3	T	DW	0.00566	5
R4	MAC	DG	0.00566	5
R5	SRPS	KM	0.00566	5
R6	KM	W	0.00566	5
R7	AAN	KMH	0.00453	4
R8	AJ	IL	0.00453	4
R9	DS	S	0.00453	4
R10	MM	SAA	0.00453	4
R11	S2	GS	0.00453	4
R12	BPWS	MA	0.00453	4
R13	DAS	SU	0.00453	4
R14	RP	DE	0.00453	4
R15	SRPS	HS	0.00453	4

Table 1 denotes the association between member supervisor and the main supervisor. Supervising percentage on NK as the main supervisor with SWA as member supervisor have the highest frequent pattern value and the following association rule with the high value of support is D as main supervisor and DP support 0.00793 or as much as 7 thesis and 0.0068 (6 thesis) respectively.

Association rule R3 up to R6 have the same occurrence value which is as the third order, likewise on R7 till R15 as the fourth order with each total occurrence as much as 0.00566 (5 thesis) for support value on the third biggest order of association rule and the following value is 0.00453 (4 thesis). Associations between main supervisor and member supervisor are represented on Picture 2.
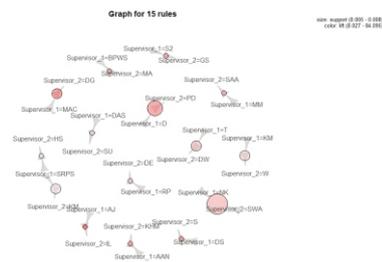
Picture 2
Visualization of 15 supervisors association rule

Appearance frequency of association between supervisors visually represented by the size of circle symbol linking the supervisor on Picture 2. The size of circle symbol declares the occurrence frequency (support) from association rule, the color of circle symbol shows the frequency of association toward another item (lift) and the arrow indicate the channel of the association term on association rule. On Picture 2 NK as the main supervisor with SWA as member supervisor which is symbolized by the biggest size of a circle corresponding to the frequent pattern value of Table 1.

Based on 15 the association rules,

visualization of an association rule is dominated by circle symbol with the smallest size which occurrence 9 times on the frequent pattern value 0.00453 (4 thesis) for each smallest value of association rule. One of the association rules has the highest value for the association toward another term (lift) that is initialized by the strongest color of circle symbol on the association rule if the main supervisor is AJ, so the member supervisor is IL.

The more terms associated with other terms the more concentrated the color on the circle symbol. Association between terms is showed by the arrow in association rules. The association channel between terms on Picture 2 can be represented by SRPS as the main supervisor have association supervising with member supervisor KM and HS. Member supervisors that associate with the same main supervisor are found in association rule R5 and R15 on Table 1.

3) **Association of Supevisor with keyboard**

Association of supervisor and the keyword is obtained by testing the smallest occurrence value (minsup/ minimum support) of 0.0012, the minimum length of rule testing is two, 23 association rule of main supervisor, member supervisor and the keyword.

Table 2
Association Rule Of Supervisor And Keyword

Rule	Supervisor	Topic	Support	Number of Thesis
R1	{Supervisor_1= D; Supervisor_2= DP}	Botanical insecticide	0.00453	4
R2	{Supervisor_1=AJ; Supervisor_2= IL }	Global warming	0.003398	3
R3	{Supervisor_1=BSP; Supervisor_2= DW }	Upland Rice	0.003398	3
R4	{Supervisor_1=SRPS; Supervisor_2= KM}	Land Use Change	0.003398	3
R5	{Supervisor_1=HS2; Supervisor_2=S3}	Nutrient balance	0.002265	2
R6	{Supervisor_1= DAS; Supervisor_2= SU }	Bioremediation	0.002265	2
R7	{Supervisor_1= AAN; Supervisor_2= KHM }	Rhizosphere bacteria	0.002265	2
R8	{Supervisor_1=MG; Supervisor_2=AK.}	Nigella sativa L.	0.002265	2
R9	{Supervisor_1= S2; Supervisor_2= GS}	M. javanica	0.002265	2
R10	{Supervisor_1=BPS; Supervisor_2= ISD}	Coleus blumei	0.002265	2

Table 2 represents the association between supervisor and the keyword. The meaning of association is that if the main supervisor supervises with member supervisor, so the topic of the student master thesis is related to the keyword. Base on Table 2, Association value is association rule R1 has the highest support of value 0.00453 (4 thesis), and the rule is if the main supervisor D and member supervisor DP, so the topic of student master thesis is related to "Botanical insecticide".

Another association rule that has the high support value is related to the the topic master thesis "Global Warming", "Upland Rice" and "Land Use Change" with support 0.003398 (3 thesis on the association rule R2, R3 and R4. Support value on association rule R5, R6, R7, R8, R9, R10 is same and so do for the association rule outside Table 2 with support value 0.002265 (or 2 thesis).

Text Mining

Text mining is performed on 483 summary data text of master thesis Faculty of Agriculture IPB which collected from DIDSI IPB. Data preprocessing in text mining uses functions in tm library which is available in R. The result of data preprocessing is 284 terms its frequency. Term occurrence frequency is represented on wordcloud visualization on Picture 3.

Picture 3
Wordcloud of Summary Term Frequency on Master Thesis Faculty of Agriculture IPB

Picture 3 shows term visualization which appearances on more than 50

summary documents. Font size is an indication of words quantities and the colour indicates the similarity of words quantities. The most frequent term is {use} which is illustrated on Picture 3 as the biggest word. The Frequent terms that were extracted on more than 200 summary documents are use, land, area, develop, plant, seed, fertility, treatment, soil, rice, region, genotype, agriculture, forest, fruit, district, potential, field, water, yield, control, and disease with the terms frequency of 1125, 809, 754, 749, 747, 461, 332, 320, 298, 272, 268, 257, 253, 244, 242, 230, 228, 222, 219, 208, 205,

and 204, respectively.

Mining Itemset Frequency with Apriori Algorithm

For further process, sparse terms were removed. Mining the frequent itemset was conducted using apriori algorithm at the support value of 0.022 (10 thesis), the minimum length of itemset (minlen) of 4 terms, the association object is frequent itemset. Next process was determining the term on agriculture sector from document-term matrix. The results are 12 frequent itemsets. The frequent itemsets are represented on Table 3.

Table 3
Frequent Itemsets Of Summary Agriculture Thesis

Frequent Itemset	Itemset	Original Word	Support	Number of Thesis
F1	land soil use farm		0.028985	14
F2	land use plantat forest	Plantation	0.026915	13
F3	soil use fertil nutrient	Fertilizer	0.024845	12
F4	land use plantat farmer	Plantation	0.022774	11
F5	land use farm farmer		0.022774	11
F6	land water area forest		0.022774	11
F7	land use water forest		0.022774	11
F8	land rice use forest		0.022774	11
F9	land use farm plantat	Plantation	0.022774	11
F10	land rice use plantat	Plantation	0.022774	11
F11	land rice use farm		0.022774	11
F12	land soil use water		0.022774	11

Table 3 show the most frequent itemsets on the summary text were land soil use farm with support value as much as 14 thesis or 0.028985. Among the frequent itemsets the frequent itemset F4 up to F12 have the lowest support value of 0.022774 or 11 thesis.

CONCLUSION

Based on association rule mining on thesis metadata of IPB library Archive, this study extracts the most frequent supervising combination between main supervisor and member supervisor on the master thesis Faculty of Agriculture. The result shows that if NK as the main

supervisor, so the member supervisor is SWA, with support of 0.00793 or 7 thesis, nevertheless the master thesis topic of their supervising student are diverse.

The supervisor combination between D as main supervisor and DP as member supervisor have a high frequency of occurrence which is 4 master thesis. This combination has thesis topic related to "Botanical insecticides" as many as 4 master thesis.

Content analysis from summary text of thesis results the frequent terms. The ten terms that have high support value are use, land, area, develop, plant, seed, fertile, treatment, soil, and rice. The combination of the terms which are frequently found on summary of thesis are land, soil, use, and farm with occurred on 14 master thesis of Faculty of Agriculture IPB.

REFERENCES

Angeline, D.M.D. (2013). Association Rule Generation for Student Performace Analysis using Apriori Algorithm. *The SIJ Transactions on Computer Science Engineering & its Applications*. 1(1): India.

Bhujade, V., & Janwe, N.J. (2011). Knowledge Discovery in Text Mining Technique Using Association Rules Extraction. *International Conference on Computational Intelligence and Communication Systems.*

Erman, L.M., & Sitanggang, I.S. (2016). Clustering Undergraduate Computer Science Student Thesis Based on Frequent Itemset. *I. J. Information Technology and Computer Science.* 8(11), 1-7. doi: 10.5815/ijitcs.2016.11.01

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining : concepts and techniques*. Third Edition. Elsevier.

Institut Pertanian Bogor. (2016). *Visi, Misi, dan Kebijakan Mutu*.

Khan, I.A., Woo, J., Seo, J.H., & Choi, J.T. (2015). Text Mining: Extraction of Interesting Association Rule with Frequent Itemsets Mining for Korean Languang from Unstructured Data. *International Journal of Multimedia and Ubiquitous Engineerig*. 11:11-20.

Kulkarni, M., & Kulkarni, S. (2016). Knowledge Discovery in Text Mining using Association Rule Extraction. *International Journal of Computer Applications*. 143(12).

Mhunpiew, N. (2013). *A Supervisor's Roles for Successful Thesis and Dissertation*. US-China Education Review A. 3(2), 119-122.

Novitasari, W., Hermawan, A., & Abdullah, Z. (2015). A Method of Discovering Association Rules from Student Admission Dataset. *International Journal of Software Engineering and Its Applications*. 9(8): 51-66. doi: 10.14257/ijaseia.2015.9.8.05

Perpustakaan Pusat Institut Pertanian Bogor. (2015). MT-Agriculture.

Symonds Q. (2016). *Bogor Agriculture University*.

Setiawan, A. (2016). *Clustering Dokumen Ringkasan Tesis Mahasiswa Pascasarjana IPB Berbasis Frequent Itemsets Menggunakan Algoritme Bisecting K-Means*. IPB (Bogor Agricultural University). (in Bahasa)

Shatnawi, R., Althebyan, Q., Ghalib, B., & Al-Maolegi, M. (2014). *Building a Smart Academic Advising System Using Association Rule Mining*. arXiv preprint arXiv:1407.1807.