Chapter 1 : INTRODUCTION

This chapter presents an overview of the thesis. Section 1.1 discusses motivation, section 1.2 discusses problem statement and research questions, section 1.3 discusses objectives, section 1.4 explains the hypotheses, section 1.5 explains the scope and limitation, section 1.6 explains about research method, section 1.7 discusses contribution, and section 1.8 discusses thesis overview.

1.1 MOTIVATION

Earthquakes are natural disasters caused by convection activity in the earth's core resulting in the movement of the earth's plates colliding, moving away from each other, or just shifting between plates. These movements cause cracks in the land called faults at the location of earthquakes because the land is not strong enough to withstand the activities of the two plates when they are colliding [1]. Countries in the world that are prone to earthquake activity are Indonesia, Japan, Chile, and many others.

Indonesia is a country with a high potential for earthquakes because Indonesia is a country traversed by the Ring of Fire which causes frequent earthquakes and volcanic eruptions [1], [2]. According to BMKG, Indonesia is prone to a small or large scale earthquakes , based on earthquake data recorded, many studies have been carried out on earthquakes, but no one has been able to predict the possibility of earthquakes well. One of the earthquake research projects uses Complex Network Analysis (CNA) for earthquake network analysis.

Complex Network Analysis (CNA) is a method used to analyze network patterns consisting of nodes and edges [3]–[7] This technique can be used to represent an earthquake network where the node is the location of the earthquake and the edge shows the relationship between nodes as seen from the sequence of events in time. The purpose of CNA is to find out which nodes have a major influence in the series of earthquake activities so that these nodes can give the appropriate weight to their influence on the network.

There are several techniques for predicting earthquake patterns, Joel et al [8] observed the pattern of earthquakes using the correlation of each earthquake hit using the Pearson Correlation and Time Series. The results of this study are pairs of areas where the earthquake occurs that have a strong relationship, but the results of the study do not predict the location of the next earthquake. Another study used several Machine Learning techniques to predict Earthquakes [9], [10], but this study could not predict where the next earthquake would be, the focus of both studies was on the accuracy of the model. Meanwhile, I Made Murwantara et al [11] used several machine learning methods to represent the location of the next earthquake based on longitude, latitude, magnitude, and depth.

Earthquake Prediction pays attention to repeating patterns, therefore there is potential for using the Sequential Pattern Mining method to predict Earthquakes. Sequential Pattern Mining is used to get data on the occurrence of certain items together, but by paying attention to the order of occurrence of items based on the time of the earthquake [12], [13]. It is called a pattern because the occurrence of these events can occur many times. This potential is strengthened by previous studies using the MSS method. Asri Inna Khoirun Nissa et al [14] conducted a study using Sequential Pattern Mining with the SPADE algorithm to find the display pattern of web activities accessed by students, lecturers and employees. Using the Prefixspan algorithm, Jian Pe et al [15] proved that prefixspan is better than the Apriori based GSP and Freespan algorithms for large sequence lengths. A study conducted by Manika Verma et al [16], analyzed the performance of three algorithms, namely GSP, Spade, and Prefixspan, and the results showed that prefixspan was superior in all test parameters, namely execution time, memory required, and number of frequent sequences. The table summary of previous method can be seen in table 1-1

Table 1-1 Summary of Previous Method

Category	Researcher	Year	Methods	Description	Result
Machine Learning	I.M. Murwantara et al.[11]	2020	Multinomial LR, SVM and Naive Bayes	Indonesia earthquake prediction over the next 30 years using historical data	Root Mean Square Error evaluation is 0.751 using support vector machine
	Bilal Adlam et al [17]	2021	SVR-HNN	Using seismic indicators, earthquake prediction along the Chaman fault in Baluchistan	The accuracy is 81.2% using supervised learning with category “Yes” and “No” means that earthquake happened or not,.
Deep Learning	Q.Wang et al. [18]	2017	LSTM	Earthquake prediction by studying the temporal-spatial association of earthquakes in various regions.	LSTM with decomposition with accuracy value is 87.59%
	R. Li et al. [19]	2020	CNN	Combining explicit and implicit earthquake features for earthquake prediction.	Accuracy is 92.42% divide of 5 class magnitude
	Bilal Aslam1 et al [20]	2021	ANN	Numerous studies have been conducted to identify the region's seismic characteristics for improved disaster management.	Evaluation using Accuracy with value is 69%

From table 1.1 the is one of potential improvements with Sequential Rule Mining (SRM) method because this method suitable for earthquake dataset. SRM has been used for another domain for example tourist destination, and customer data of a telecommunication service provider. For evaluation of this method using confidence value each rule that created. Vu, H. Q. et al [18] using SRM and get one of confidence value from the rule (Monaco => Paris) is 93.2%. Another research using SRM, Husák, M. et al [19] success to implemented on the customer data of a telecommunication service provider. The highest confidence value of a rule is 97.50%.

1.2 Problem Statement and Research Question

The use of machine learning to predict earthquakes has been carried out in research [9]–[11]. This study resulted in an accuracy of 90.6% using SVR-HNN, 65% accuracy using Linear Programming Boost ensemble classifier , and 0.751 Root Mean Square Error (RMSE) using SVM. However, this study cannot predict future earthquakes because this study relies on input parameters for the classification process, so the system's prediction of earthquakes will come out when an earthquake has occurred. There are two approaches for earthquake prediction, namely using Sequential Pattern Mining[21] and using Sequential Rule Mining[22]. Comparison between proposed method and others can be seen in figure 1-2.

Table 1-2. Strength and weakness of type of methods

Method	Strength	Weakness
ML/DL Classification	- Powerful with predefined class (Supervised Learning) and can have multiple input parameters to predict class	- Not relate to predict the future earthquake, because most of research using this method to know where class of the new input of earthquake. Classification means that categorized the input.
ML/DL Regression	- Can predict next earthquake location with evaluation of RMSE - Can predict next earthquake the next few days	- RMSE value contain continuous value. It means that the evaluation error of method has range value. - Not paying attention to the relationship between the location of the earthquake and the plates of the earth, even though it is important.

Pattern Mining	- Can advance to analysis of each rule, not only get the evaluation of method but also researcher can analyse each rule or each location interconnected, - Can predict next earthquake with time range that have been predefined.	- Limited input of parameter, input method only sequence of items because the main purpose of this method to know the pattern relation each node - have to add evaluation scheme to get 1 value metrics from the method

The Sequential Pattern Mining (SPM) method gives a potential solution to predict future Earthquake activity. The characteristics of SPM to find data that has a sequence, it is believed that it can be used to predict earthquakes that will occur in the future. This means that it depends on data testing, if the data contain patterns from the model, the location of the next earthquake can be predicted. SPM has been used to observe patterns of web activity accessed by students, lecturers and staff[14], novel buying patterns [15], and to observe the performance of GSP (Generalized Sequential Pattern), SPADE (Sequential Pattern Discovery using Equivalence Class) and Prefix Span (Projected Sequential Pattern Mining Prefix) [16], The best evaluation result is the Prefix Span method in terms of execution time, memory, and frequent sequence count. The problem of using SPM is that it cannot find the evaluation of method. Researcher using SRM to obtain confidence value as a evaluation method.

The Sequential Rule Mining (SRM) helps researchers evaluate each rule so that they can build accurate model. They need to analyse complicated data and make these data into simple pattern built by SPM, it is necessary to do a Social Network Analysis (SNA) to represent the network of related nodes. The basis of using CNA refers to research [15] which uses CNA to determine which nodes have the most influence in the network using Centrality Measurement (CM).

As mentioned in table 1-2, related to the limitation of sequential pattern mining algorithm researchers did not covers longitude, latitude and the magnitude of earthquakes occurrence. therefore in this study researchers only cover the earthquake prediction in a form of pattern mining. Pattern discovery from historical datasets is done by calculating the number of occurrences of an item in the sequence database and projected database so as to produce the output of several sequence patterns. To calculate the accuracy of these patterns, it is necessary to have an accuracy calculation scheme that can be seen in table 3-9 . The test data was tested as a starting location earthquake to review the pattern of the model being built.

This research proves that the most important node has the highest weight compared to other nodes. The main research questions of this research is:

Can the centrality method in CNA improve the performance of the Sequential Rule Mining model for earthquake prediction?

1.3 Objectives

The objectives of this research include:

a. Obtaining the next earthquake prediction from the rule found in the SRM method.
b. Recognizing the effect of the centrality technique in CNA for weighting and confidence value of Sequential Rule Mining in predicting Earthquakes.

1.4 Hypothesis

Research [13], [15], [16], [21] applies SPM to observe patterns in each transaction, in a series of earthquakes that can be analyzed using SPM to determine the subsequent earthquake. To facilitate the analysis of SPM patterns, SNA can be applied to represent earthquake networks. Research [7], [8] describe networks that are influential in disaster management. The most influential node is calculated using the centrality measurement, the largest centrality value indicates that the node is important. Up to this moment, SPM is only on frequent itemset. Thus, it is important to pay attention to the pattern of both the frequent itemset and the weight of a node. For these reasons, it is expected that the combination of centrality measurement techniques in CNA and SPM can build models that represent more precise predictions and improve the evaluation of the models being developed. The hypothesis of this study is that using the Centrality Measurement technique as a weighting on CNA can improve the performance of Sequential Pattern Mining when it is compared to the one without using weighting.

1.5 Problem Scope and Limitation

To ensure that the scope of this problem does not extend to unrelated aspects, it is necessary to define the scope of the problem. The scope of the problem in this study is as follows:

a. The data is taken from BMKG from 2008 to 2022.

b. The data set is filtered by Magnitude > 4 . The purpose of filtering is to observe destructive earthquakes or felt by people. Meteorology and Geophysics Agency (BMKG) [1] uses the unit to measure the strength of an earthquake, namely the MMI scale. The MMI scale category can be seen in the table 1-3.

Table 1-3 Characteristic of Earthquake

Magnitude Scale	MMI	Category	Description
1-1.9	I	Micro	Vibrations are not felt except in exceptional circumstances by a few people
2-2.9	I to II	Minor	The vibrations are felt by several people, the light objects that are hung swayed.
3-3.9	II to III	Minor	Vibrations are felt in the house. There is a vibration as if a truck is passing.
4-4.9	IV to VI	Light	Vibrations are felt and the windows/doors and walls are creaking and making noisy sound. As if a heavy truck is passing.
5-5.9	VI to VII	Strong	The Vibrations are felt by all residents. Most people are shocked and running outside, the plaster of the walls fell off and make minor damage.
6-6.9	VII to X	Strong	The vibrations damage strong buildings, the frames of the houses become misaligned, and experience many cracks. The houses look a bit moved from its foundation. The pipes in the houses are broken.
7-7.9	X to XII	Major	The vibrations destroy buildings but few buildings remain standing. The bridge is broken, and there is a valley. The underground pipe cannot be used at all, the ground is split, the rails are very curved.
8+	X to XII	Great	The vibrations destroy buildings completely. The ground is waving. The scene is darkened. Objects are thrown into the air.

c. Earthquake location data are in Indonesia.

d. The series of earthquakes is reviewed based on the time and distance of occurrence.

1.6 Research Methodology

The research methodology applied in this thesis consists of several steps.

a. Problem Identification

This step aims to identify problems in the recommendation of learning materials, define the problems discussed in this thesis, and make improvements to the selected problems.

b. Model Design

The design model defines the framework of the system that is developed as a description of how the system works. The model design consists of the SPM process and Centrality measurement which can be combined into the system.

c. Data Collection and Processing

Data are source of information that can be used to gain knowledge. The data are taken from the BMKG scrapping website to obtain earthquake events according to the specified time span.

d. Implementation

At this stage, the model that has been designed in the previous stage is implemented into a program. The dataset is processed using the SPM method to obtain patterns and then processed using the SRM method to obtain the rules and evaluation results.

e. Experiment

The experimental scheme is carried out for the needs of working analysis of a method and as a proof of research hypotheses. Parameters of time, location, and confidence value are observed to obtain the best accuracy in the experiment

f. Analysis of experiment results

This stage is to analyze the results of the experimental schemes carried out in the previous stage. The results of the accuracy values of each experiment are further analyzed to get the conclusions of the experiment

1.7 Contribution

The main contribution of this research is an improved method for earthquake prediction using SRM. Recent studies have focused on hybrid methods that incorporate system constraints for rule generation efficiency. This study proposes a combination of the SRM method and the calculation of Centrality Measurement for the formation of earthquake as predictions. This research uses accuracy for method evaluation. Using accuracy enables people to compare methods.

1.8 Thesis Overview

This thesis is organized into five chapters consisting of introduction, literature review, algorithm design, and implementation, experiment and data analysis, and conclusions. The explanation of each chapter is as follows:

a. Introduction

   This chapter discusses the background of the problems raised in this thesis, the definition of problems and research questions, research questions, objectives to be achieved, an overview of the research methodology and an overview of the thesis.

b. Literature Review

   This chapter discusses the theoretical basis of the concepts and theories that support and are applied in this final project. Literature Review refers to the theoretical basis used as a basic reference for the guidelines for the final project.

c. Research Method

   This chapter describes the general description of the system and the proposed model for solving the problems in this final project. In this design, an overview of the system is described, both in terms of detailed system structure, proposed model, data flow, and system usage scheme.

d. Experiment and Analysis

   This chapter explains the purpose of testing and test scenarios for solving the problems that exist in this final project. At the analysis stage of the test results, it is explained about the data derived from the test results, the graph of the test results and the analysis.

e. Conclusion

   This chapter describes the conclusions of this study. The conclusion includes the results of the final analysis that has been carried out and explains the answers to the problem formulation formulated in this thesis. At the recommendation points, suggestions are outlined regarding the possibilities that can be developed from the results of this research.