A

Project Report

on

EFFECTIVE PATTERN DISCOVERY FOR TEXT MINING

Submitted in Partial Fulfillment of
the Requirements for the Degree
of

Bachelor of Engineering

in

Computer Engineering

to

North Maharashtra University, Jalgaon

Submitted by

Dipali Sonawane
Tejal Shirole
Kajal Patil
Priyanka Patil
Amol Patil

Under the Guidance of

Mrs. Nilima Patil

DEPARTMENT OF COMPUTER ENGINEERING

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY,
BAMBHORI, JALGAON - 425 001 (MS)
2016 - 2017

SSBT's COLLEGE OF ENGINEERING AND TECHNOLOGY, BAMBHORI, JALGAON - 425 001 (MS)

DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the project entitled *Effective Pattern Discovery For Text Mining*, submitted by

Dipali Sonawane
Tejal Shirole
Kajal Patil
Priyanka Patil
Amol Patil

in partial fulfillment of the degree of *Bachelor of Engineering* in *Computer Engineering* has been satisfactorily carried out under my guidance as per the requirement of North Maharashtra University, Jalgaon.

Date: September 30, 2016
Place: Jalgaon

Mrs. Nilima Patil
Guide

Prof . Dr. Girish K. Patnaik
Head

Prof. Dr. K. S. Wani
Principal

Acknowledgements

First of all we would like to extend our deep gratitude to almighty God, who has enlightened me with power of knowledge. We would like to express our sincere gratitude towards Principal Prof. Dr. K. S. Wani (SSBT, COET Jalgaon) for his encouragement during the work. We wish to express our sincere and deep gratitude to Prof. Dr Girish K. Patnaik (Head of Computer Department) for giving us such a great opportunity to develop this project. Inspiration and Guidance are invaluable in all aspects of life especially on the fields of gratitude and obligation and sympathetic attitude which We received from our respected project guide, Mrs.Nilima Patil whose guidance and encouragement contributed greatly to the completion of this Project. We would like to thanks to all faculty members of Computer Engineering Department and all friends for their co-operation and supports in making this project successful. We would also like to thanks my parents for supporting us and helping us. We acknowledge our sincere gratitude to all who have directly or indirectly helped me in completing this project successfully.

Dipali Sonawane
Tejal Shirole
Kajal Patil
Priyanka Patil
Amol Patil

Contents

List of Figures

Abstract

Data mining techniques have been proposed for mining useful patterns in text documents.In domain of text mining discovered pattern is still an open issue that how it is effectively use and update . Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Text Mining presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.

Chapter 1

Introduction

Many data mining techniques have been proposed for mining useful patterns in text documents. Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge in text documents to help users to find what they want. data mining techniques have been used for text analysis by extracting occurring terms as descriptive phrases from document collections.

In this chapter, Section 1.1 describes background of the project. Section 1.2 describes motivation of the project. Section 1.3 describes problem definition of project. Section 1.4 describes the objective of the project. Section 1.5 describes organization of the report.Finally, section 1.6 contains Summary.

1.1 Background

This project aim to proposed approach is used to improve the accuracy of evaluating term weights. Because, the discovered patterns are more specific than whole documents. To avoiding the issues of phrase-based approach to using the pattern-based approach. Pattern mining techniques can be used to find various text patterns.

1.2 Motivation

In our work an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents.

1.3 Problem Definition

Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge. The advantages of term based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

Many data mining techniques have been proposed in the last decade. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. However, using these discovered knowledge (or patterns) in the field of text mining is difficult and ineffective. The reason is that some useful long patterns with high specificity lack in support . All frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. An effective pattern discovery technique has been proposed to overcome the low-frequent and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents.

1.4 Objective

- Objective are:

1. **D**ata Preparation:
    - To load the list of all documents.
    - The user to retrieve one of the documents.
    - This document is given to next process.
    - That process is preprocessing.
2. **P**reprocessing:
    - The retrieved document preprocessing is done in module.
    - There are two types of process is done.

- Stop Words: Stop words are words which are filtered out prior to, or after, processing of natural language data.
- Text Steaming: Stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms.

3. Pattern Matching.

1.5 Organisation of the Report

Chapter 1 describes the introduction, background , motivation , problem defination of project with its scope and objective . Chapter 2 describes system analysis. Chapter 3 describes system requirements specification which includes software, hardware, functional and non functional requirements. Chapter 4 describes system design with the help of various unified modeling language diagram.

1.6 Summary

This chapter covers the introduction of the project and it's background, problem definition, motivation and objective of the project. In next chapter, the system analysis is described.

Chapter 2

System Analysis

System analysis is the process of gathering and interpreting facts, diagnosing problems and using the facts to improve the system.

In this chapter,Section 2.1 describes the literature Survey. Proposed System is discussed in section 2.2. Section 2.3 describes feasibility study. Project scheduling is described in section 2.4. Initially, section 2.6 contains Summary.

2.1 Literature Survey

Mining Closed Sequential Patterns in Large Sequence Database" V. Purushothama Raju1 and G.P. Saradhi Varma ,2015. Information Retrieval using Pattern Deploying and Pattern Evolving Method for Text Mining" ishakha D. Bhope, Sachin N. Deshmukh,2015.

2.2 Proposed System

- An effective pattern discovery technique, is discovered
- Evaluates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns
- Solves Misinterpretation Problem
- Considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and tries to reduce their influence for the low-frequency problem.
- The process of updating ambiguous patterns can be referred as pattern evolution.
- The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents.
- In General there are two phases

- Training and Testing
- In training phase the d-patterns in positive documents (D) based on a min sup are found, and evaluates term supports by deploying dpatterns to terms
- In Testing Phase to revise term supports using noise negative documents in D based on an experimental coefficient
- The incoming documents then can be sorted based on these weights.

2.3 Feasibility Study

Feasibility study conducted once the problem is clearly understood. Feasibility study is a high level capsule version of the entire system-analysis and design process. The objective is to determine quickly and at the minimum expense how to solve the problem and to determine the problem is solved. The system has been tested for feasibility in the following ways.

- Technical feasibility
- Operational feasibility
- Economical feasibility

2.3.1 Technical feasibility

A study of function, performance and constraints may effect the ability to achieve an acceptable system so ,that necessary function and performance are achieved with in the constraints uncovered during system analysis. The software developed for the automation text mining is MYSQL as backend and JAVA as front end. Since the software is platform independent and has predefined functions and constraints such as to locate the charges, validating functions etc.,so the project is technically feasible.

2.3.2 Operational feasibility

The purpose of this project is to develop a computerized system which facilitates text mining in precvious technique there were biggest drawback of polysemy and synonymy. All the operators of this project are trained in this area. So this project is operational feasible.

2.3.3 Economical feasibility

Economic analysis includes a broad range of concerns that include cost benefit analysis ,strategies, cost of resources needed for development. For text mining many techniques,

they performed text mining very well but there were biggest disadvantage polysemy and synonymy for in our . Since the cost of resources for development of system satisfies the organization, the software is economically feasible.

2.4 Risk Analysis

There are various types of risk are present so different categories of risk are as follows:

Technical risks Technical risk is simply the risk associated directly with the knowledge base being employed and it's technical aspects including such things as understanding, reproducibility and the like. Exposure to loss arising from activities such as design and engineering, manufacturing, technological processes and test procedures.

Business risks The term business risk refers to the possibility of inadequate profitt or even loss due to uncertainties e.g., changes in tastes, preferences of consumers, strikes, increased competition, change in government policy, obsolesce etc. Every business organization contains various risk elements while doing the business. Business risks implies uncertainty in profits or danger of loss and the events that could pose a risk due to some unforeseen events in future, which causes business to fail.

Project risks A project risk is an uncertain event that, if it occurs, has a positive or negative effect on the prospects of achieving project objectives. Project risk can be defined as an unforeseen event or activity that can impact the project's progress, result or outcome in a positive or negative way. Efective risk management strategies allow to identify your project's strengths, weaknesses, opportunities and threats. By planning for unexpected events, user can be ready to respond if they arise. To ensure project's success, define how user will handle potential risks so user can identify, mitigate or avoid problems.

2.5 Project Scheduling

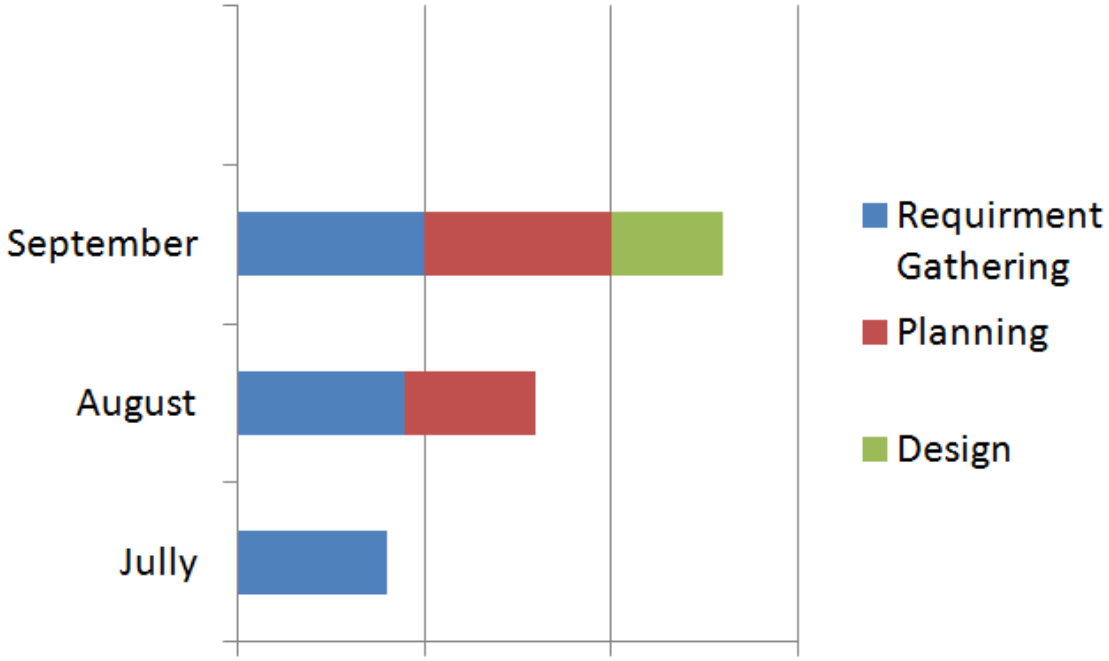

Figure 2.1: Project Scheduling

2.6 summary

In this chapter, the system analysis is described, literature survey and proposed system to over come the problem of the existing system is described. In addition also described feasibility study, risk analysis and project scheduling. In the next chapter, system requirement specifications are described.

Chapter 3

System Requirements Specification

Software Requirement Specification is the oficial statement of what is required to the system developers.It should include both user requirements and a detailed specification of the system requirements.Requirement analysis is done in order to understand the problem the software system is to solve.

In this chapter section 3.1 describes the hardware requirements. Software requirements are explained in section 3.2. Finally, section 3.3 contains Summary.

3.1 Hardware Requirements

- Processor : intel(R)Core(TM)i3-3227U CPU @ 1.90GHz
- Harddisk : 40GB
- RAM : 256MB

3.2 Software Requirements

- Operating System : ubantu 14.04
- Programing Language : JAVA
- Database : MYSQL
- Tool : jdk7

3.3 Summary

In this chapter, hardware requirements, software requirements are explained. In next chapter, the System Design is described through various UML diagram.

Chapter 4

System Design

System Design chapter will provides graphical structure of the project by using various UML diagrams. System design provides the understanding and procedural details necessary for implementing the system recommended in the system study. Design is a meaningful engineer- ing representation of something that is to be built. In the software engineering context, design focuses on four major areas of concern are data, architecture, interfaces and components.

In this chapter, section 4.1 describes system architecture of the project. E-R Diagram is describe in section 4.2. Data flow diagram is described in section 4.3. Various UML diagrams are describe in section 4.4. Section 4.5 gives summary.

4.1 System architecture

The Figure.1 depicts the process flow of system that consists of loading document for pre-processing by users preference. The text preprocessing, in which the retrieved document is passed through two processes such as stop word removal and text stemming. In first process words which are filtered out prior to, or after, processing of natural language data are called as stop words. The second process for reducing inflected (or sometimes derived) words to their stem base or root form called as stemming. In pattern taxonomy process, the documents are split into paragraphs and are considered as a separate document from which set of terms are extracted are called the patterns. In pattern deploying the discovered patterns are summarized using d-pattern algorithm. The pattern evolving process is used to identify the noisy patterns in documents. In which sometimes, the system falsely identified negative document as a positive. So, noise is occurred in positive documents, these noised patterns named as offender and if partial conflict offender contains in positive documents, the reshuffle process is applied.

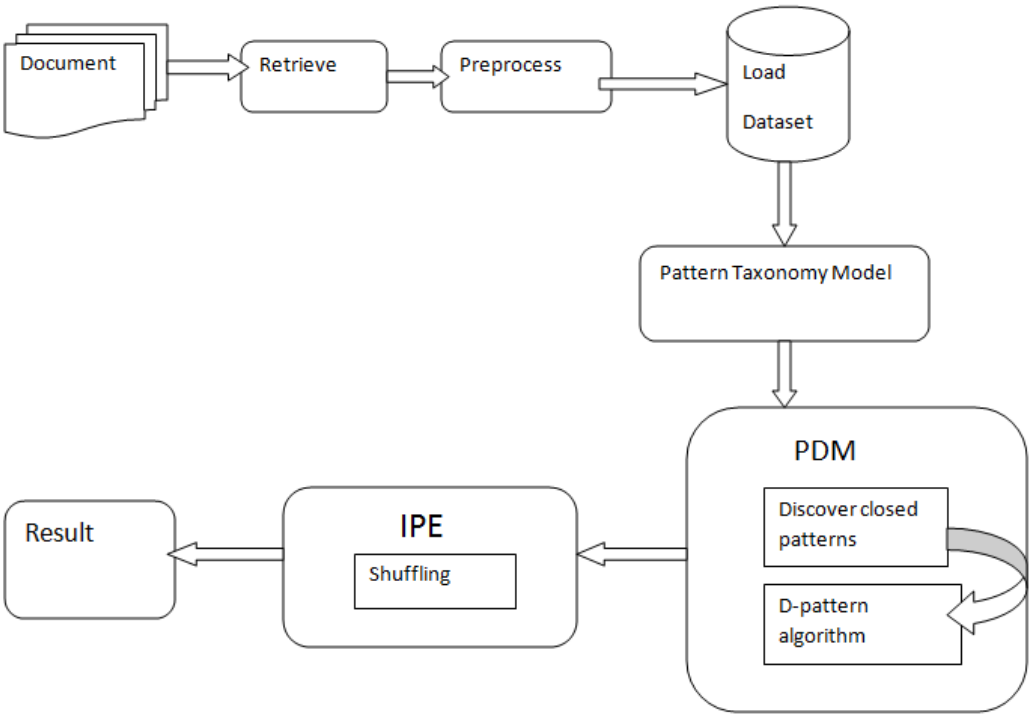

Figure 4.1: System Architeture

4.2 UML Diagrams

4.2.1 Usecase Diagram

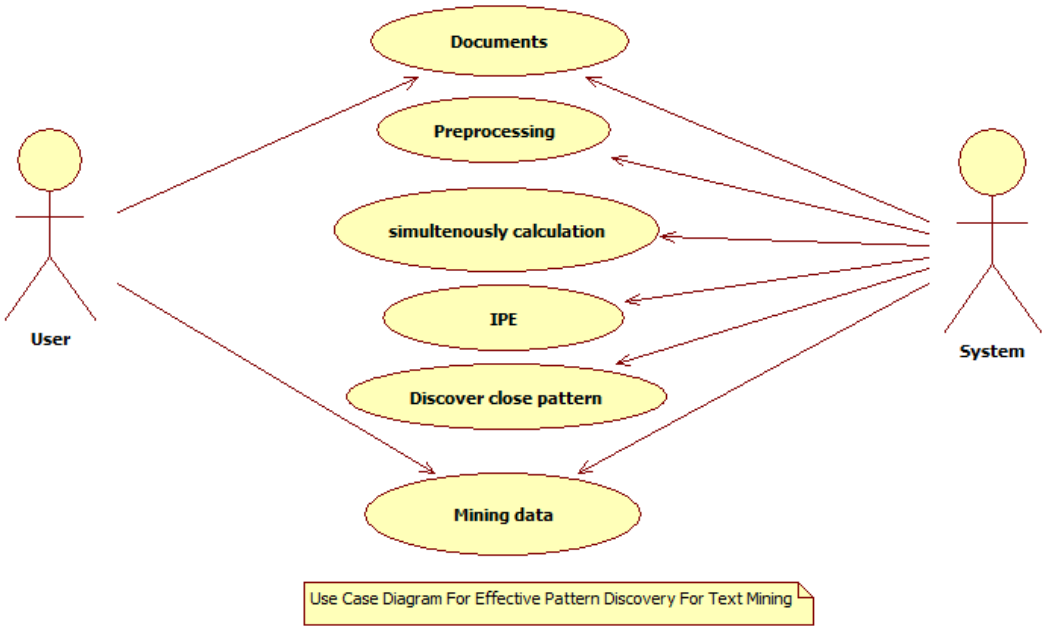

Figure 4.2: Usecase Diagram

4.2.2 Sequence Diagram

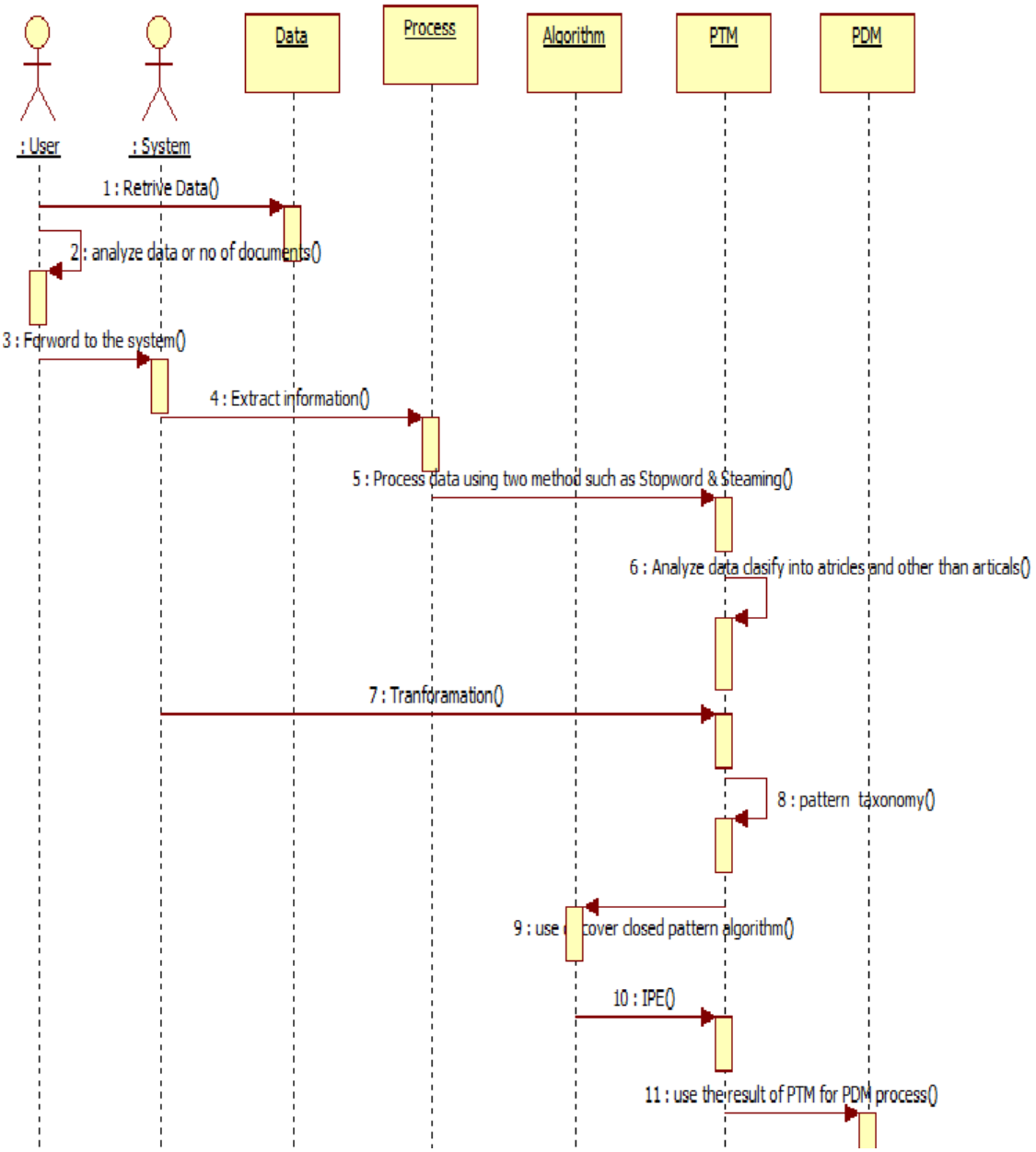

Figure 4.3: Sequence Diagram

4.2.3 Class Diagram

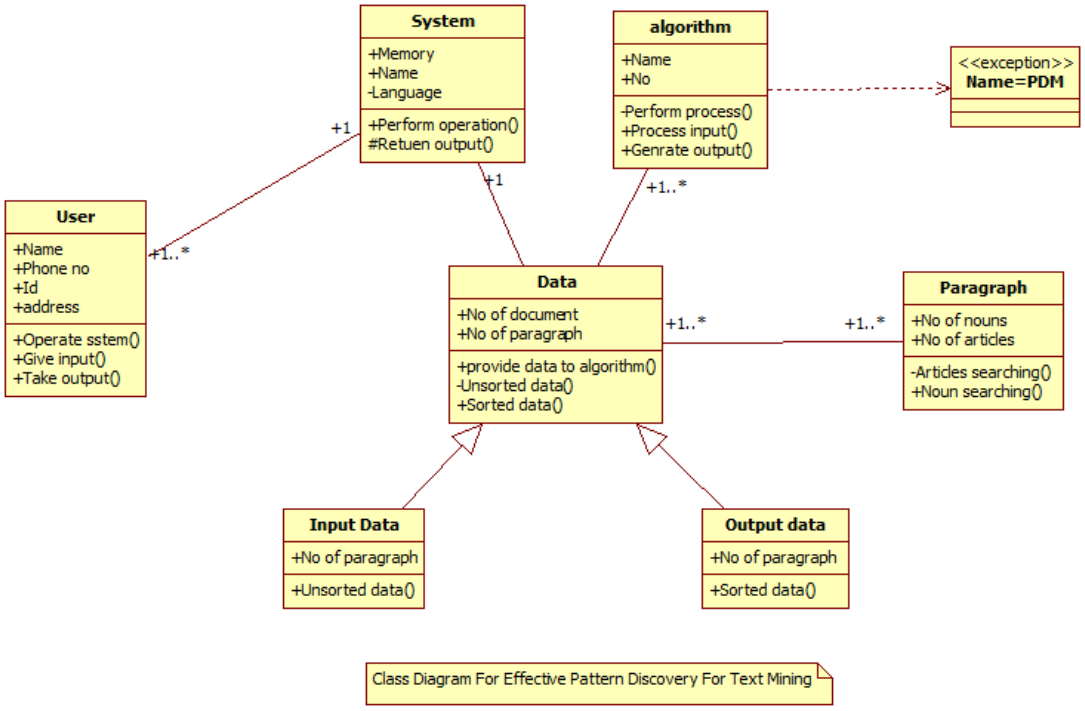

Figure 4.4: Class Diagram

4.2.4 Statechart Diagram

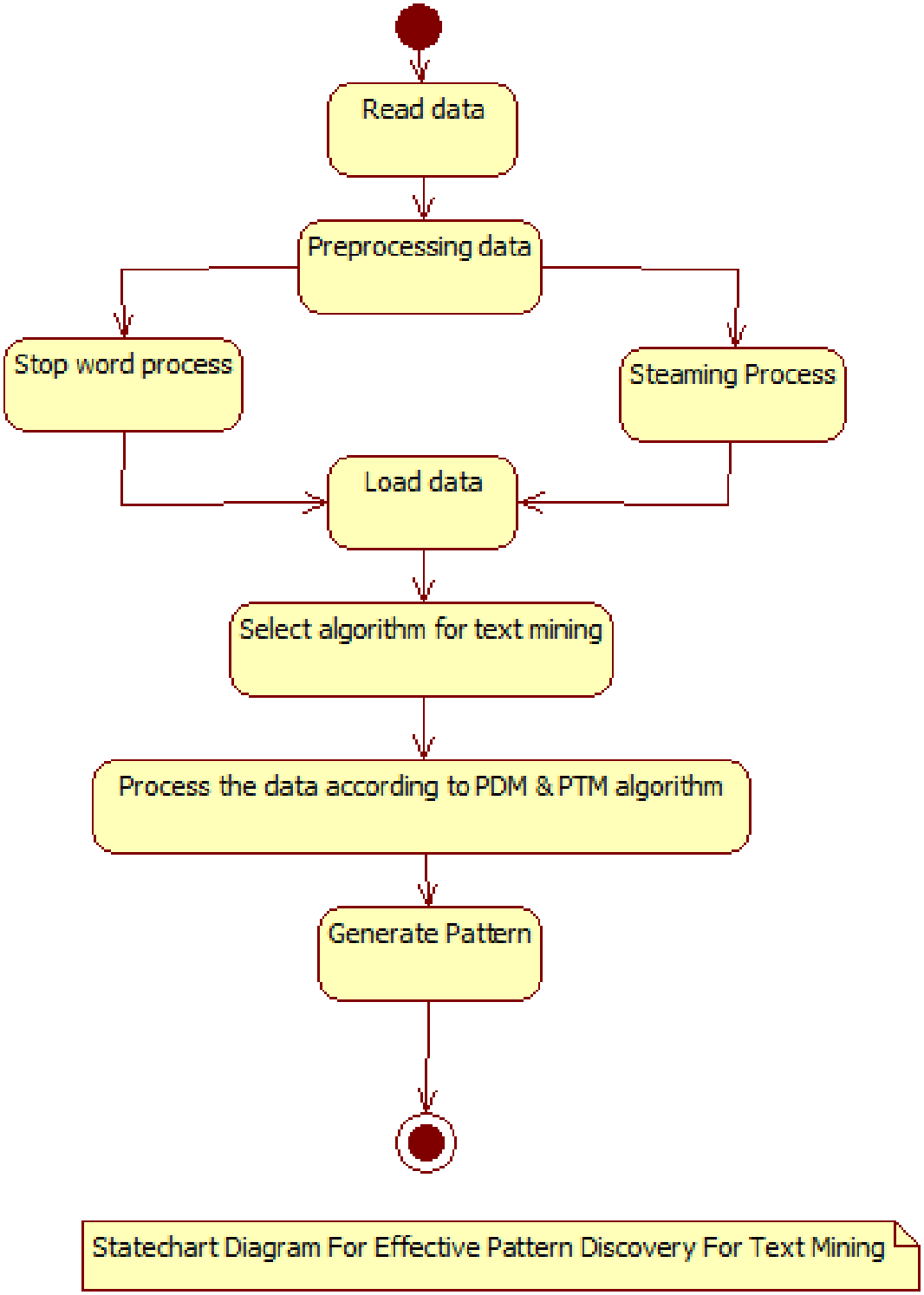

Figure 4.5: Statechart Diagram

4.2.5 Activity Diagram

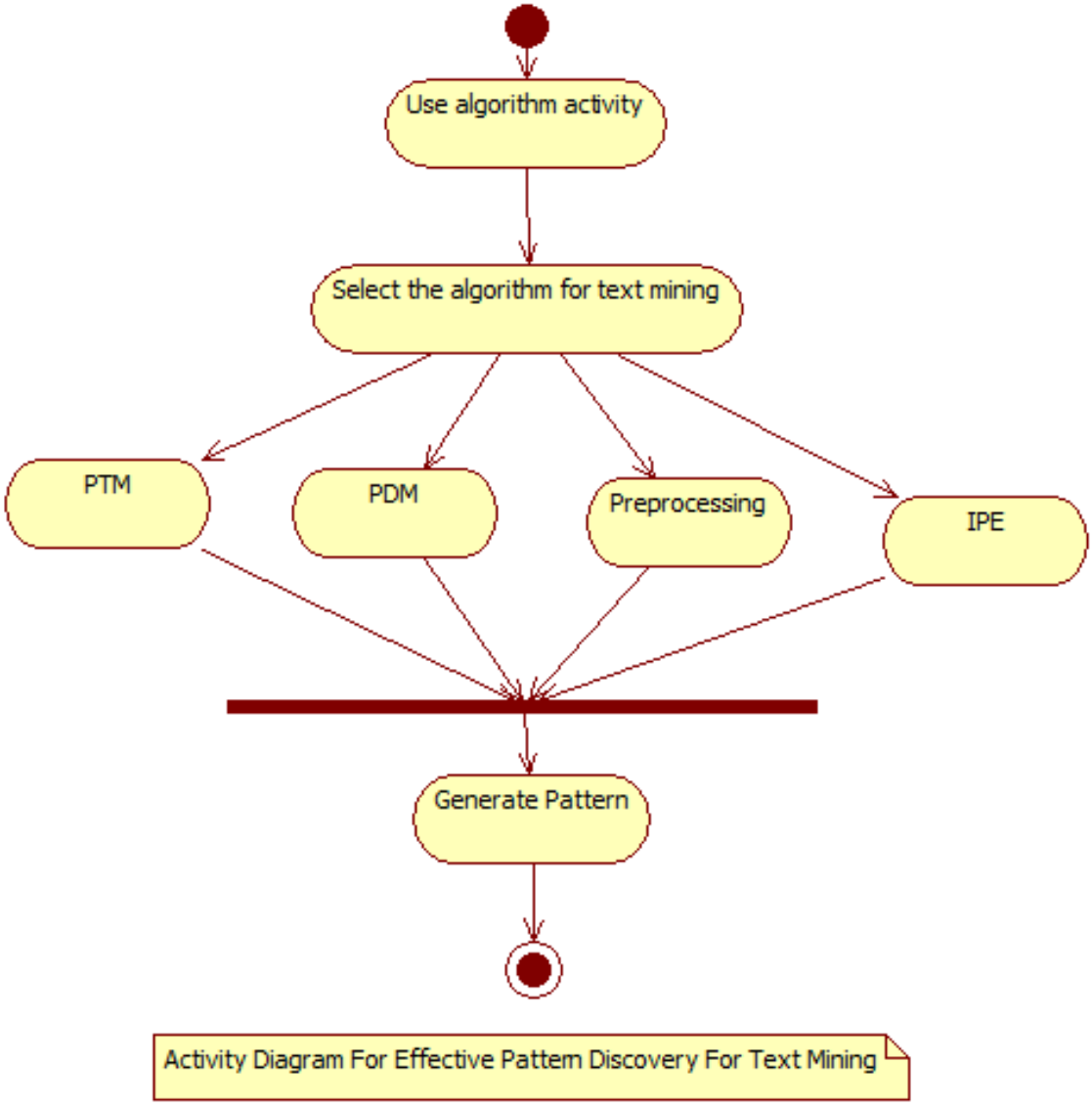

Figure 4.6: Activity Diagram

4.2.6 Component Diagram

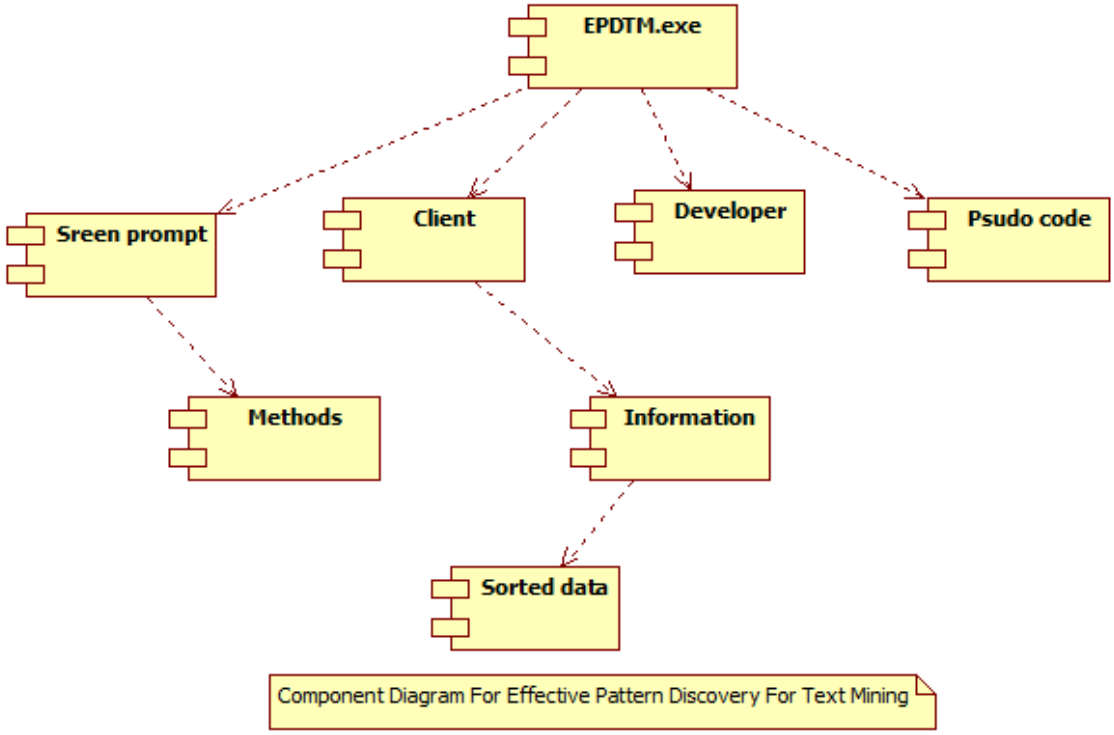

Figure 4.7: Component Diagram

4.2.7 Deployment Diagram

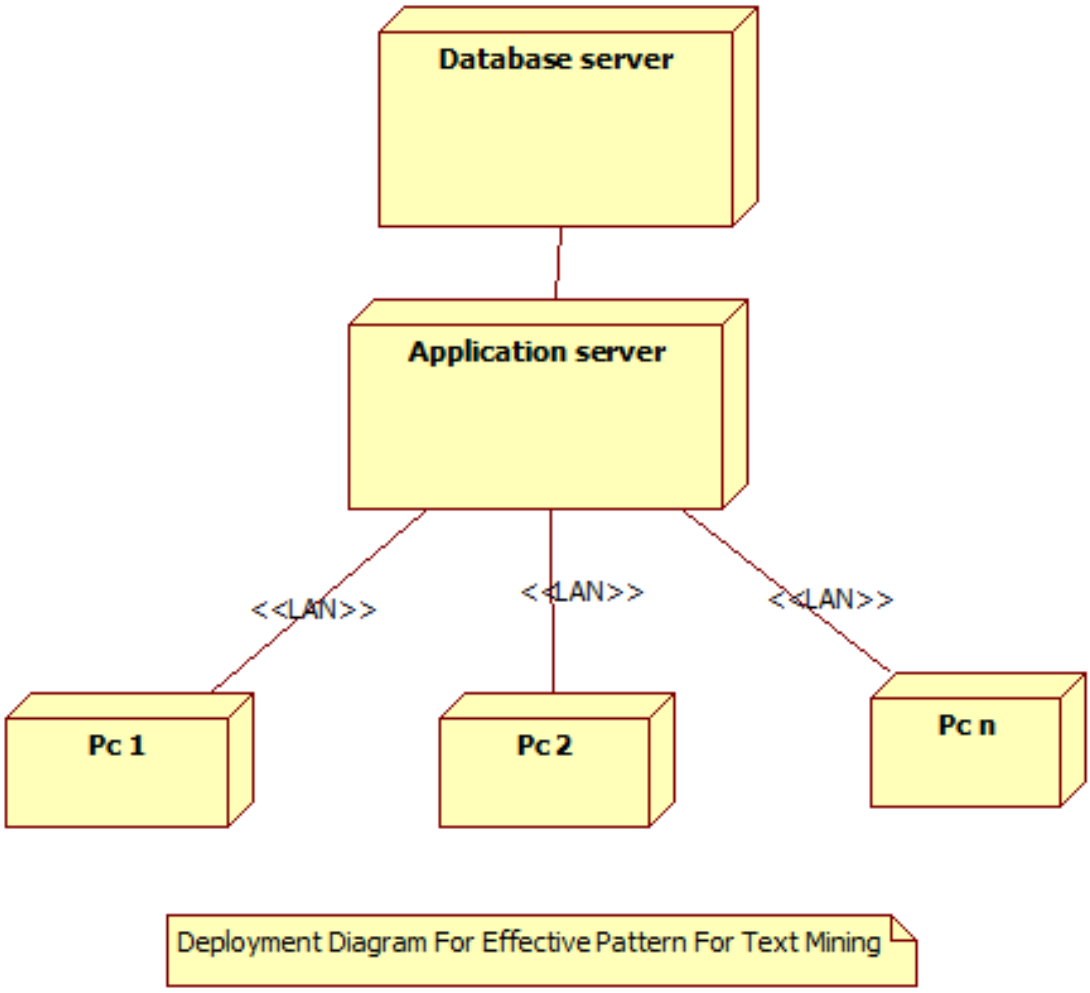

Figure 4.8: Deployment Diagram

4.3 Summary

In this chapter, architecture of system, E-R diagram, Data flow diagram and UML diagrams are described. In the next chapter, implementation are described.