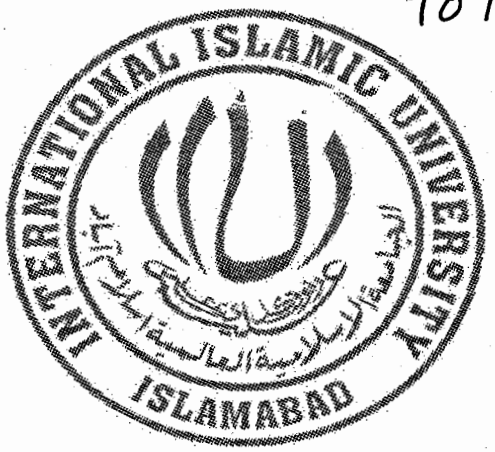
Sequential Pattern Mining for Feature Based Opinion Analyzer (SPM-FBOA)

T07485



By:

Naveed Anwer Butt
252-FAS\MSCS-F05

Supervisor:

Dr. Mian Muhammad Awais
Associate Professor
Lahore University of Management Science-LUMS

Co-Supervisor:

Muhammad Imran Saeed
Assistant Professor
International Islamic University Islamabad-IIUI

International Islamic University, Islamabad
Faculty of Basic & Applied Sciences
Department of Computer Science

Accession No TH 7485

MS
006.3
BUS

1- Data mining

Sequential Pattern Mining for Feature Based Opinion Analyzer (SPM-FBOA)

T07485

BY

Naveed Anwer Butt

Supervisor

Dr. Mian Muhammad Awais

Co-Supervisor

Muhammad Imran Saeed

THESIS SUBMITTED TO

DEPARTMENT OF COMPUTER SCIENCE,

FACULTY OF BASIC AND APPLIED SCIENCES,

INTERNATIONAL ISLAMIC UNIVERSITY, ISLAMABAD, PAKISTAN

as a partial fulfillment of the requirement for the award of the degree of

MASTER OF SCIENCE

IN

COMPUTER SCIENCES

INTERNATIONAL ISLAMIC UNIVERSITY, ISLAMABAD
FACULTY OF BASIC AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER SCIENCE

Dated: 20-10-2010

Final Approval

It is certified that we have read the thesis titled "**Sequential Pattern Mining for Feature Based Opinion Analyzer (SPM-FBOA)**" submitted by **Naveed Anwer Butt Registration No. 252-FAS\MSCS\F05.** It is our judgment that this thesis is of sufficient standard to warrant its acceptance by **International Islamic University, Islamabad** for the degree **MS in Computer Science.**

COMMITTEE

External Examiner:

Dr. Nisar Ahmed,
Assistant Professor,
Faculty of Electronic Engineering, GIKI,
Sawabi

Internal Examiner:

Mr. Asim Munir,
Assistant Professor,
DCS, FBAS, IIUI

Supervisor:

Dr. Mian Muhammad Awais,
Associate Professor.
Lahore University of Management Science, LUMS, Lahore

Co-Supervisor:

Mr. Muhammad Imran Saeed,
Assistant Professor,
DCS, FBAS, IIUI

DECLARATION

I, hereby declare that "**Sequential Pattern Mining for Feature Based Opinion Analyzer (SPM-FBOA)"** neither as a whole nor as a part thereof has been copied out from any source. I have developed this project and the accompanied report entirely on the basis of my personal efforts made under the sincere guidance of my supervisor. No portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institution of learning.

Naveed Anwer Butt

252-FAS\MSCS-F05

ACKNOWLEDGEMENT

The humble accomplishment of this thesis would not have been possible without the contribution of many individuals, to whom I express my appreciation and gratitude. Firstly, I am deeply indebted to my principal supervisor Dr. Mian Muhammad Awais (Associate Professor, Lahore University of Management Sciences, Lahore Pakistan) who guided me every step of the way and was a source of inspiration. His patience, insights, research style and the ability to draw research questions from literature have been integral to the success of this work and to my career development as a researcher. Without his professional guidance and help, this work would not have been achieved. I am also grateful to him for providing me with various supports to conduct this study and many invaluable opportunities to let me be involved in many professional activities, which are very beneficial for my future academic career. Thanks are also presented to my co-supervisor, Mr. Muhammad Imran Saeed (Assistant Professor, International Islamic University Islamabad-IIUI) for his assistance and many valuable suggestions throughout my study.

I am profoundly grateful for all the support of my (late) father and my (late) mother. This thesis is dedicated to them. I owe my loving thanks to my wife who stood by my side through the years, and who has lost a lot due to my research work.

In the last but the most I would like to thanks my brother in law Dr. Asad Imtiaz Khan (Medical Specialist, WAPDA) who motivate me to do the MS thesis.

Finally, thanks to all the people who helped and encouraged me through these years, both in working and private life, friends, colleagues and relatives.

ABSTRACT

Volume and fast crescents complexity of the modern databases make the need for technologies to summarize the information contained, increasingly important. Rules to define relations between pattems also called rule mining is one of the major task in data mining. There are basically two categories of rules, namely the association rules and sequence rules

Today, a large number of websites allow users to post reviews about products they bought. There are thousands of reviews of customers related to one product. So it is difficult for the manufacturers and also for the customers to have an idea about the product from these large reviews. In this research we aim to summarize the customer reviews in factual form as done using two data mining techniques that is Sequential pattern Mining and Association Rule Mining.

This dissertation deals with discovery of product features to which customers have opinions from a preprocessed Web usage data by extraction of Association rules and Sequential Patterns (SPs) and with very low support.

The proposed system is a supervised learning approach can model information extraction system perfectly, which extracts fine-grained features, and associated opinions, from online product reviews to identify product features with improved precision and recall. In addition, for extracting features, we simply combine natural language processing techniques with data mining. There exist a variety of web data sources which generate vast amounts of data, with inherent sequential nature. Much of product feedback from external reviews available online through websites, discussion forums, mailing lists, blogs and portals also called Consumer Generated Content accumulate in human-written unstructured documents.

As Web-log data is sequential in nature, Sequential pattern Mining (SPM) is particularly well suited for such data.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

1. Introduction

> "The desire of knowledge, like the thirst of riches, increases ever with the acquisition of it."
>
> — Laurence Sterne

This chapter starts with the background of the problem. We will discuss the brief background, overview of the problem, its impact in different domains and our purpose of the research.

1.1 Overview

" What people are saying" "What do people have opinion about?", perhaps the issue of what is important in a world where millions of people express their opinions on any topic at the available electronic forums such as blogs, news sites, product review sites or social media. Today, companies have great need for continuous assessment of reputation of services and products and how marketing campaigns will receive from target audience.

With the dramatically quick and explosive growth of information available over the Internet, blogging and blog pages are grown rapidly. This becomes the most popular way to express opinions and sentiments. According to the blog search engine of Technorati [94], by the end of 2008, there were 133 million blogs on the global internet, which are indexed by Technorati. Figure 1.1 shows the state of the blogosphere at 2008.The figure shows the amount of digital information is constantly increasing, and search engines are still the means for accessing this information. The demand for targeted information extraction is constantly growing.

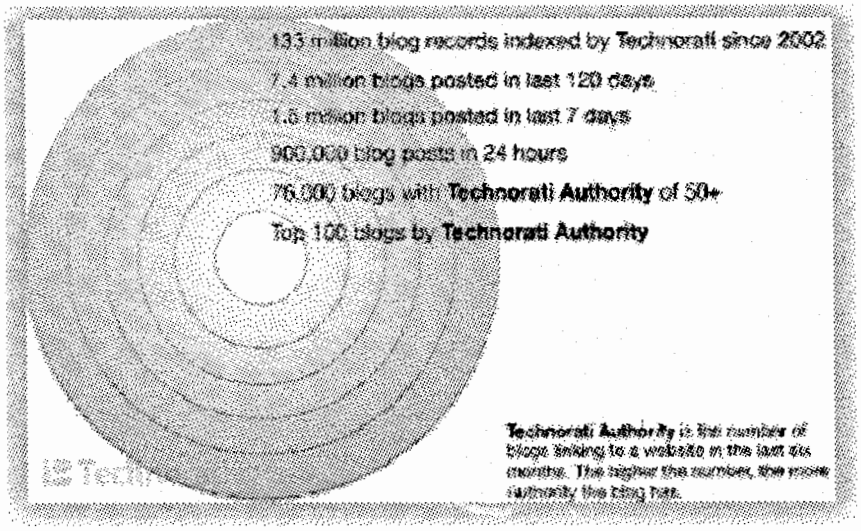


Figure 1.1 State of the Blogosphere [94]

Products or services are often discussed by customers on the web. Online communities are interesting for market analysis and research. Official company sites usually tell a certain side of the story. Valuable data relevant for market research on the web is neither easy accessible nor process able. Time expenses to collect and evaluate data needed for a better market understanding are still tremendous. On the other hand opinion mining aims at developing technology for exploiting the rich and dynamic resource of factual information and human opinions available on the internet.

Today, a large number of websites i.e. Amazon.com allow users to post reviews about products they bought. Such information can be used for marketing and product improvements. There are thousands of reviews of customers related to one product. So it is difficult for the manufacturers and also for the customers to have an idea about the product from these large reviews. However, It is time consuming job if do this manually. As an example, many manufacturers needs information in a structure format, which is easier to use and it will helpful for the automation of this process[48] .

It is highly desirable to mine such opinions and then measure the polarity and strength of the expressed opinions and produce a summary of the opinions.

However, Understanding, effectively analyzing and summarizing opinions from huge amount of unstructured information such as product reviews is an intellectually very challenging task in recent research and practically very useful and it is also a subject of this research.

This thesis focus on product reviews on the portal sites provided by Bing Liu. Mining opinions from online reviews, however, it is a complex process that requires more than just text-mining techniques. This complexity is due to some natural inherent process issues. Firstly, data to be examined has to be crawled from sites where role of spiders or Web search engines can be important. In addition, the separation of reviews from non-reviews is necessary so that opinion summarizing process can occur [79].

1.2 Background of the Problem

The Web 2.0 provides two types of textual information can be generally classified into two main categories, "*Fact and Opinions*". " *Facts are objective sentences about entities and events in the world while opinions are subjective sentences that reflect people's sentiment about entities and event.*" Search has become the default way of interacting with user generated content. The growth of data volume is rapidly shifting to user generated content in the past few years.

The motivation towards the search for hidden knowledge in text collections is due to the availability of large collections of e-documents. Consequently, there is growing research interest in the general topic of text mining.

This tendency increased interest in technologies for the automatic extraction or analysis of the personal opinions of Web documents contributions to blogs. Such technologies Test packs an alternative to traditional questionnaire-based social or customer research, Recommender systems and administrative community dying dignity even for web users seeking advice on certain consumer products in their interest. Since search engines are not looking for opinions, because opinions are difficult to express in brief. Opinion search is therefore one of the major challenges which

can provide contextually relevant information for organizations and businesses today.

The opinions are much harder to describe than facts. Notices sources are usually written informally (or worse) and very diverse. You are able to provide short to describe this context, analytical efforts. Thus, the precision sentiment extraction is generally much lower, but it can be induced by appropriate measures that are appropriate for the source and targets.

Although sentiment classifications at both level e.g., (document and sentence level) are useful, but cannot find what holders of opinion liked and disliked. We have to go the feature level. The goal is to be found in what might reviewers (opinion holders) liked and disliked, as the number of comments on an object may be significant, an opinion summary should be made desirable a structured summary that easy to visualize and compare.

1.3 Objectives and Goals

Opinion mining problem can be decomposing into the following series of subtasks:

1. Identify and extract commented features of the product that the reviewers have expressed their opinions on, called product features. For instance, in the sentence "*the picture quality of this camera is amazing*" the product feature is "*picture quality*".

2. Identify opinions regarding product features.

3. Sentiment Analysis: Determining the semantic orientation and strength i.e. whether the opinions on the features are positive, negative or neutral. In the above sentence, the opinion on the feature "*picture quality*" is positive.

4. Produce a summary of the opinion based on multiple entities comments

This dissertation deals with the first and second task: to find all the explicit and implicit product features on which reviewers have expressed their opinion words (positive or negative)

1.4 Problem Statement

Let A = {A1, A2 ... An} is the set of products from one brand or more than one brand that the users interested to express their opinion. Each product Ai has a set of reviews R = {R1, R2... Rk}. Each review is sequence of sentences Rj = {S1, S2 ... Sjm}. These reviews could be from one site from multiple sites.

Definition (Product Feature): A feature f in the series Rj is an attribute or a component of the product, where users have commented on. If f appears in a sentence Rj then it will be a frequent feature, and if f does not appear in a sentence Rj then it will be mentioned only infrequent feature

In the following example "Player" and "look" are frequent features/

"Player works and looks great."

Similarly "support" is infrequent feature in following example.

"Apex doesn't answer the phone."

Definition (opinion): The opinion of a feature f in review is a set of consecutive sentences which shows positive or negative opinion on f.

In a sentence at least one sentence expresses opinion on a feature. But in one sentence more than one feature could be present on which user has expressed his/her opinion.

"Great quality picture and features."

"The apex 2600 has been a steady performer for me."

Our task: The research problem that is to be addressed during the research work is problem involving massive, unstructured and complex datasets, solutions using

innovative data mining algorithms, novel sequential pattern approaches, the objective evaluation of analysis and solutions. Extraction of fine-grained features with a good rate of identification has been a huge challenge for researchers conducting research in the field of mining associated opinions. But each feature-opinion extraction and modeling technique is not suitable in all types of environment. Our solution is to use supervised data mining technique such as Apriori-based Algorithms e.g. sequential pattern mining for building up a representation in the form of sequential rules to extract frequent and infrequent product features and opinions words which are used on these features from tagged data by applying NLP to find the interesting patterns from the reviews, with improvement in both recall and precision to summarize all the product reviews, which will increase the accuracy of feature and opinion sentence extraction. Therefore, an approach is developed in this thesis by using different techniques for achieving high detection rate.

1.5 Motivation and Need

A large number of consumers turn to see what a significant impact on the purchasing decision, found online after a recent survey by Deloitte's Consumer Products Group that almost two thirds (62%) of consumers read reviews online consumer written.

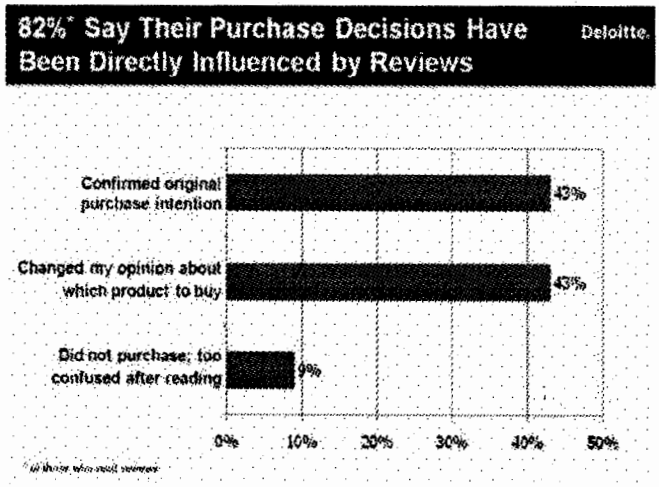


Figure 1.2 Purchase Decisions Influenced by Reviews [94]

Figure 1.3 shows that more than eight out of 10 (82%) of those who read the articles, said that their purchasing decisions were influenced directly by the ratings. People have to acknowledge the contributions of both the first purchase and to change them, the study found. In addition, said 69% of respondents said they have shared online with friends, family or colleagues, the report says .

Other motivational factors are as follows:

- Products or services are often discussed by customers on the Web
- Online communities are interesting for market analysis and research
- Official company sites usually tell a certain side of the story
- Valuable data relevant for market research on the Web is neither easy accessible nor processable
- Time expenses to collect and evaluate data needed for a better market understanding are still tremendous.
- To find out the customer opinion regarding to product's features that which features of the product customer likes and which features are disliked.
- To find out the customer opinion regarding to the overall product (positive, negative or neutral)
- There are a lot of customer reviews on the web in raw form which is not easy for potential customer to read and extract the overall opinion about product.
- Reviews are increasing day by day, it becomes difficult and time consuming for online companies to manage and extract the meaningful reviews for potential customers as well as for company itself.
- Reviews should be in summarized form for online communities as they are interesting for market analysis and research.

1.6 Contributions

In this research a new knowledge discovery model is proposed with an attempt to effectively exploit the discovered patterns in a large data collection using data

mining approaches. This model uses pattern taxonomies as features to represent knowledge based on the state-of-art data mining techniques such as sequential pattern mining.

The summarized contributions are briefed as follows.

- A knowledge discovery model based on pattern taxonomies is proposed
- The state-of-art data mining techniques are used
- Provide a comprehensive comparison of
    - Apriori vs GSP

There is need to overcome a number of challenges in the systems that can handle subjective information. We will highlight these challenges and will explain our purpose of research.

1.7 Purpose of Research

Systems which can process subjective information need to overcome a number of challenges. We will highlight these challenges and our purpose is to explain the research.

1. There are mainly two types of information on the Web, for example, facts and opinions. Current search engines to find facts and not seeking opinions. Subjective material is hard to find, because opinions are hard to express with keywords but facts the facts can be expressed with keywords subject. For example such queries are hard to handle e.g. "How do people think of Nokia 6660 Cell phones?" Current search ranking strategy is not appropriate for opinion retrieval/search. If we could be able to integrate the application into a search engine, it should be ascertained whether the user is looking for the subjective material or not. A checkbox could be provided to the user so that user could indicate what he/she desired.

2. The additional challenge in addition to the open problem of determining which documents are relevant to the query is simultaneously or subsequently, that the provision to check what document or the part of the document, as the material or the comments contain. If the text is retrieved from the review aggregation sites, write on review oriented information in a relatively stereotyped format; it would be relatively simple as we can examine the contributions from Amazon.com and Eopinion.com. But we know a lot of blogs contain subjective content and can be seen, therefore, a place still very important to review and may be more important than commercial sites for queries that policy concerns, the people, not products, etc., but yet the documents referred to in blogs can be very common in the content, style, presentation, and even the level of grammar.

3. Another important problem is to identify overall sentiment as well as specific opinions in relation to certain features or topics in question, as required. Some of the pages have this type of collection is easier to do,for example, reviews posted to Yahoo! Movies have to give marks for pre-defined characters of the films. For example, care must taken when quotes in a newspaper article, care are included to the articles in every offer the right person to express a similar attribute and these are held on some points should be noted.

   Now the system has found the general atmosphere of the documents, The system must feel the information it was produced in a manner appropriate to present summary. This, some or all of the following measures [73]:

   1. Aggregation of registered "votes" on different levels.
   2. Extraction of selective opinions
   3. Representation of points of disagreement and points of consensus.
   4. Opinion owner communities identification
   5. Identification of authority levels among opinion holders.

1.8 Applications Domains

Opinion mining and sentiment analysis has two direct audiences: Consumer and Business organization. A number of enterprise and small companies that have adopt or start to implement opinion mining modules. There are also other applications in different fields that are going to be flourished; We try to list some of the possibilities

1.8.1 Review Related Websites

- Provide consumer opinions about a range of products to companies that would result in enhancing informed decision making in marketing and product improvements.
- Provide valuable feedback direct from consumers available on multiple forums.
- To help industry in order to optimize business processes and boost quality.
- To support Product manufacturer in market intelligence, decision support tasks and product benchmarking
- Provide subjectivity analysis and Polarity Analysis, Feature Based Analysis & Summarization
- To help industry in reduction of cost, enhance research & development and increase sales.
- To find reviewers reasons of liking and disliking of product?

1.8.2 Consumer Awareness

Opinion mining is popular due to the decision support for consumers. Today, most of the popular sites like amazon.com and Epinions collecting consumer opinions against a wide range of products. Theses repositories make guide consumers to make informed decisions based on:

- Guidance received from opinions of other consumers while purchasing a product
- Knowledge is presented in a graphical format for easy comparison of product features.
- Purchasing trends thus reducing lead time to purchasing a product

About Dell Small Business, It is to be noticed opinions are submitted about delivery time, shipping charges, support and web site navigation and other various features.

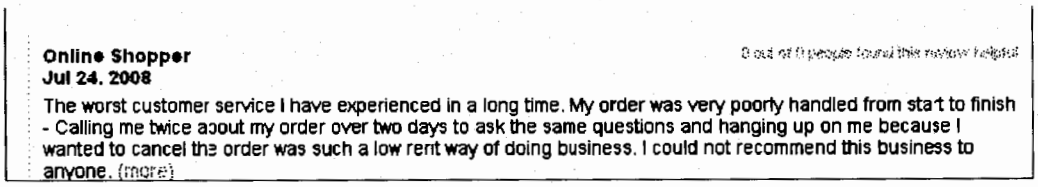


Figure 1.3 Consumer Review in free Format [6]

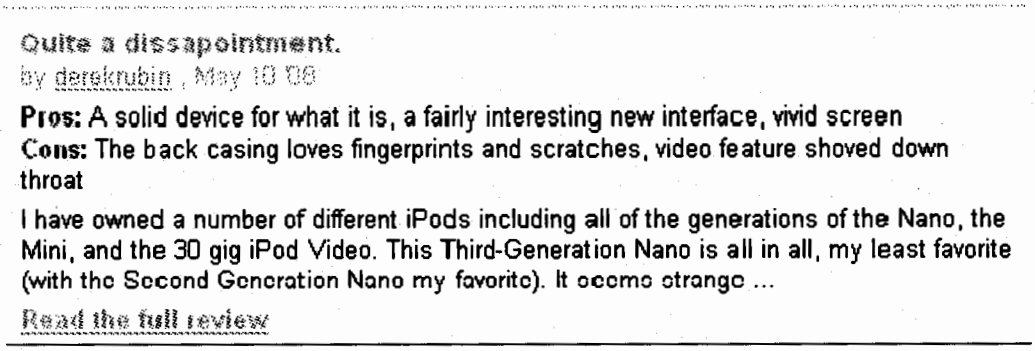


Figure 1.4 Consumer Reviews [The State of the Art] [6]

1.8.3 Marketing- Ads Placements

Opinion mining can play a vital role while someone wants to place ads. On the websites ads are placed on the sidebar to attract people. This could be done by the overall opinion of the users. If a vast class of users show their likeness about the product, it could be very helpful for the companies to place ads. Also if most of the people dislikes the product it would be easy to remove the ads from such sides. It is because; the negative opinion of most users can put a bad impact on the

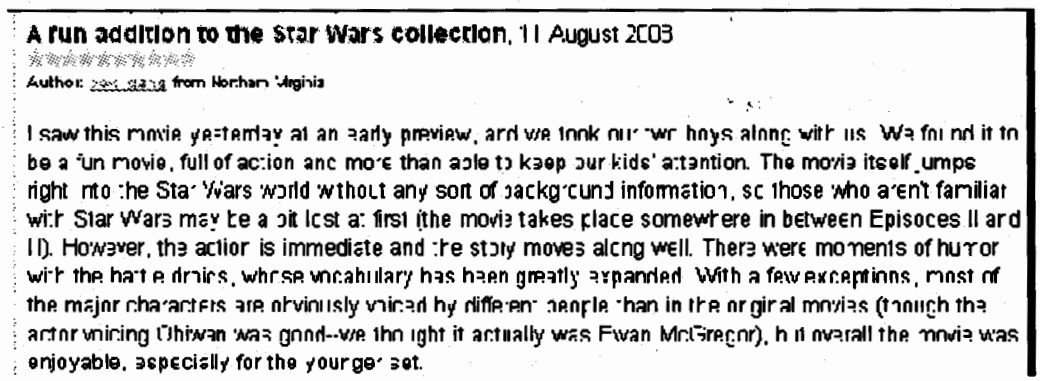
sale of the product. There is another point to be mentioned, companies can place ads when users are showing negative opinions about their competitor's product. This could be very interesting point that one can advertise their product while people are fed up from the same product or another company. Opinion mining also helps in Placing Ads in the user-generated contents where the competitor's product is being criticized.

1.8.4 Entertainment

Other popular use of opinion mining is in the field of entertainment. Today, moviegoers and television viewers at home can quickly review the recent releases and popular movies and programs. There have a number of websites, the online ratings for movies and TV programs. The following is an excerpt from a popular Internet Movie Database IMDB.

"Christopher Nolans second bundle of joy The Dark Knight EXCEEDED all of my expectations!!! I can HONESTLY tell you that: as good as Jack Nicholson was in Batman'89 he is CHILDS PLAY compared to this Joker. He is sadistic, psychotic, and downright SCARIER and PSYCHOLOGICALLY disturbing than the previous incarnation of The Clown Prince of Crime and Ledger gives it his all to do him justice. The action is great, and the plot is deeper and engrossing." [6].

In addition, most opinions are expressed on the films in this way as Figure 1.6.


A fun addition to the Star Wars collection, 11 August 2003

Author [illegible] from Northern Virginia

I saw this movie yesterday at an early preview, and we took our two boys along with us. We found it to be a fun movie, full of action and more than able to keep our kids' attention. The movie itself jumps right into the Star Wars world without any sort of background information, so those who aren't familiar with Star Wars may be a bit lost at first (the movie takes place somewhere in between Episoces II and III). However, the action is immediate and the story moves along well. There were moments of humor with the battle droids, whose vocabulary has been greatly expanded. With a few exceptions, most of the major characters are obviously voiced by different people than in the original movies (though the actor voicing Obiwan was good--we thought it actually was Ewan McGregor), but overall the movie was enjoyable, especially for the younger set.

Figure 1.5 Movie Viewer Reviews in free Format [6]

1.8.5 Government

The field of extraction and analysis of public perception is well adapted to different types of government agencies..Government is also interested in prevailing opinions on public policy. Most news websites have a facility to express their opinions upon various issues. That could be a best source for public viewpoints .Through opinion mining someone can get familiar about the public opinions upon government policies. Elections candidates can learn more about the specifics of an opinion poll. They can easily identify their strengths and weakness lie according to their electorate and also sense the satisfaction and dissatisfaction among the electorate.

1.8.6 Education

In systems of e-learning, the opinions of the user can be used to evaluate the academic institutions and universities have great concerns about the word of mouth of services offered, courses, programs and teachers. Curtin University of Technology offers their students state of the art, forums for discussion and module for submitting weekly peer reviews. This arrangement act as sources of rich content by the user that can be mined. The result of research by [33] suggests that e-learning management systems introduced by the tertiary institutions is still in its infancy.

1.8.7 BLOG Mining

- To grab people's opinions and emotions about recent survey on a subject.
- For the determination of prevailing opinion about products or technologies.
- To detect motives and arguments from the court decision or legislative debate.
- To discover the interaction in medical abstracts drug.

1.9 Thesis Outline

The rest of the work is as follows;

Chapter 2: This chapter explains different techniques used for opinion mining and a complete survey of the related work. Also we will discuss some drawbacks of most of the previously used methods and highlight our method of research. In the end of this chapter we are proposing our model for opinion mining.

Chapter 3: In this chapter we have given on overview of opinion mining. All the methods used for opinion mining and sentiment classifications are discussed. Moreover, we have briefly explained our approach.

Chapter 4: We have explained sequential pattern mining in a very brief manner. Different algorithms are discussed which are used for mining sequential patterns. It also pinpoints the current work and future trends. We have highlighted their limitations and have explained the proposed algorithms which we are using in our research.

Chapter 5: This chapter starts with the proposed methodology as explained in Chapter1. We have explained our work step by step and also have compared our results with the previous results on this area.

Chapter 6: This chapter presents the description of benchmark dataset and performance measures, along with the application of proposed model to opinion mining. A detailed Analysis of the comparison results of experiments is also presented in this chapter. In this chapter we are using sequential pattern identification for the opinion mining. We have clearly explained our work and also shown our results gathered after implementing frequent pattern mining.

Chapter 7: This chapter will conclude our work and draws the direction for future work.

CHAPTER 2

2. Review of the Literature

> "A man's feet should be planted in his country, but his eyes should survey the world."
>
> — George Santayana

As an introduction and to place this research in perspective this chapter introduces other research on relative areas. This survey leads us the foundation for the development of effective tools to improve the build function is based opinion mining and leads us to the motivation, the context for opinion mining and categorization of current literature in the way to make clear, research opportunities. This research is close related to review mining. Morinaga et al. [66] proposed a text mining techniques to mining the product reputation, but they do not focus on product features. In the next few years, there are some researches [30, 121] that focus on mining product features. This research work is based on these works to improve.

2.1 Introduction

Opinion Mining (OM), an emerging discipline, defined as an intersection of information retrieval and computational linguistic techniques deals with the opinions expressed in a document. The field is aimed at solving the problems associated with opinions about products, politics, review sites and newsgroup postings, etc. There are different techniques for summarizing customer reviews like Data Mining, Information Retrieval, Text Classification and Text Summarization.

Although OM is closely related to the areas of Data Mining, Information Retrieval, Text Classification and Text Summarization, yet significant differences exist. The OM differs from data mining in the sense that it focuses on extracting

subjective information from the data repository while data mining extracts objective information. If using information retrieval, which concerns the document subject compared, OM differs in the sense that it is concerned with the opinion of a document. Text Classification classifies a document according to the topic while OM classifies a document according to the polarity. The area of Text Summarization deals with free form, topic-related documents. On the contrary, OM deals with structured and opinion related visualized form.

We divided the literature review in three sections according to the domains in which we reviewed the literature. The research in Opinion Mining can be categorized into three distinct areas, namely:

1. Development of linguistic resources
2. Sentiment classification
3. Extracting and summarizing opinion expression

Opinion mining has some basic components; depending on these components we start our work. **Opinion holder:** *the person or organization that holds a specific opinion on a particular object.* **Object:** *on which an opinion is expressed.* Opinion: a view, attitude, or appraisal on an object from an opinion holder.

2.1.1 Development of Linguistic Resources

A linguistic resource could be used for extraction of stakeholders' opinions so that the documented user sentiments could be classified. The OM has threefold tasks related to a linguistic resource. These include:

1. Determining Subjectivity i.e., finding the nature of a given term;
2. Determining whether a particular term a positive or negative
3. Determining Strength of the linguistic term attitude.

The linguistic resources development methods can be broadly classified into two categories:

2.1.1.1 Appraisal Theory

The Appraisal Theory forms the linguistic foundation of OM It provides a framework for language resources, the authors and speakers supported to express inter-subjective and ideological position. The Appraisal defined by the Appraisal Theory helps in evaluating all possible uses of language.

2.1.2 Development Methods

The Development Methods include the following four main methods:

2.1.2.1 Conjunction Method

Hatzivassiloglou, V. et al, in 1997 [28], , based on the hypothesis that the adjectives used to have 'and' conjunctions usually similar orientations are based, however, "but" with an opposite orientation is used.The method involves extraction of all conjunctive adjectives from the corpus. The extracted conjunctive adjectives are then acted upon by the log-linear regression model for determining if ever connected two of the adjectives same or different orientation.. A clustering algorithm is then applied for clustering the words with same orientation. Higher frequency group is labeled as positive.

2.1.2.2 PMI Method

Pointwise Mutual Information (PMI) method involves association between the linguistic terms. The association is measured inline with the information theory and statistics. Turney and Littman in 2003 [71] have described that the terms tend to occur together in documents with similar orientation. Baroni, M. et al, in 2004 [10], have described that the subjective adjectives tend to occur in the near of other subjective adjectives. A modified version of PMI has also been used for predicting semantic orientation based on the seed term set.

2.1.2.3 WordNet Expansion Method

Hu et al, in 2004 [31] have proposed this method based on the hypothesis that the "adjectives usually share the same orientation as their synonyms and opposite orientation as their antonyms". The method involves assignment of orientation of all adjectives by exploring the cluster graphs of the used set of seed adjectives.

2.1.2.4 Gloss Use Method

Esuli and Sebastiani in (2005 & 2006) [17, 18], have proposed this method based on the hypothesis that "*terms with similar orientation have similar glosses*" and "*terms without orientation have non-oriented glosses*". The classification method used in the Gloss Use Method involves a binary input of Positive and Negative seeds set. The lexical relationship is discovered from a thesaurus or dictionary in an iterative process for building synonym term sets. A textual representation of each term of the set is generated by collating all the glosses of the term present in a machine-readable repository. Each representation is then acted upon by the standard text indexing techniques for conversion to a vectorial form. Thereafter, a binary classifier is trained on the developed seed set and afterwards evaluated on the test set.

Method	Intuition	Accuracy	Characteristics
Conjunction Method	Adjectives and conjunctions usually have similar orientation when used only with opposite orientation	78.08 %	The First try
			Test Data: 1336 adjectives
PMI Method	Words of similar orientation tend to co-occur in documents	87.13 %	No limitation
			Much time required
WordNet Expansion Method	Generally adjectives share the same orientation as their synonyms and opposite orientation in front of their antonyms	N / A	Limited to WordNet
Gloss Use Method	Terms with the same orientation have similar glosses	87.38 %	SentiWordNet (All word in WordNet)
	Terms without orientation have non-oriented glosses		Accuracy depends on the quality of thesaurus

Table 2.1 Characteristics of five Methods for opinion development Methods

2.2 Sentiment Classification

The process of identifying sentiments or the polarity of some of the text or the document is named as Sentiment Classification.

The classification of sentiments is usually performed at:

1) Document–level; 2) Sentence–level; 3) Phrase–level; and 4) Feature–level;

2.2.1 Document Level Opinion Mining

The research for classifying sentiments is usually done at the document-level. Document level sentiment classification of documents (e.g. articles) to the total weight of authors expressed sentiments related. Three methods namely: PMI, Machine Learning and NLP Combined; have been widely used for document-level sentiment classification.

2.2.1.1 Point Wise Information (PMI) Method

Some of the previous work on document level used this method. PMI measures the strength of semantic association between two phrases. As Turney in 2002 [96] used PMI-IR algorithm which is the combination of the PMI and information retrieval (IR) to measure the similarity of two different words or phrases. PMI-IR basically estimates the PMI by giving queries to the search engines and keeping the record of the number of hits. On the basis of this semantic association, the average sentiment orientation of the phrase is being calculated and the review is then classified as recommended or not recommended.

2.2.1.2 Machine Learning (ML) Methods

Machine learning plays an important role in opinion mining. Many researchers used different machine learning techniques for the evaluation of the document. Researchers have used three different techniques for opinion mining related to machine learning methods.

2.2.1.3 Natural Language Processing (NLP) Methods

NLP concerns with the interaction between computer and human languages. With NLP methods, information from computer databases could be converted into human readable language. Many researchers have used NLP techniques along with the machine learning techniques for the sentiment orientation of the documents. Wilson, T . Wiebe , J . et al, in 2005 [102], combines machine learning and NLP techniques for sentiment analysis at clause level. Similarly Whitelaw, C. et al, in 2005 [101], used appraisal theory along with support vector machines.

A comparison of these methods is shown in Table 2.2 below:

Method	Characteristics	Pros	Cons
PMI	Use phrase PMI	• Simple; • Need not priory polarity dictionary;	• Loss of contextual meaning; • Slow (Time to get PMI);
Machine Learning	Bag of Words	• Simple; • Need not priory polarity dictionary;	• Loss of contextual meaning; • Need learning phase;
	Unigram to bigram or n-gram		
	SVM, NB, MaxEnt		
NLP Combined	Based on ML	• Consider contextual meaning; • Easily extendible for various purpose;	• Need priory polarity dictionary; • Syntactic Analysis Overhead;

Table 2.2 Characteristics of Methods of Document Level Opinion Mining

Pang and Lee in 2004 [72], improve the sentiment polarity by applying only subjective sentences. To checks whether a sentence is subjective or not they

remove the objective sentences with the help of a subjectivity detector. Accuracy of film reviews classification increased to 6.4%.

A machine learning technique Maximum entropy analysis is used by Mehra, Khandelwal and Patel in 2002 [60], to identify sentiment of the movie reviews. Three text classification algorithms (Nearest Neighbor Classification, Naïve Bayes Classification Algorithm, and Maximum Entropy Classification) have been used. The reviews they used, to serve as the input for the text analysis, were obtained from the newsgroup and movies reviews. Their work consists of three steps. First step is User Interaction, in which they allow users to rank the movies according to their taste or choice. The second step Creation of Training and Test Tests, they split the ranking of movies by users into two steps. In step three Document Pre-processing in which the reviews are preprocess, they read all the reviews according to their categories and implemented their data sets as hash arrays to improve the efficiency of system. Three hash function tables are used for document preprocessing. From these hash arrays, they extracted the words and implemented maximum entropy classifier. Then they choose the words extracted from maximum entropy classifier as features for further processing.

Another work for automatically collecting technical terms associated with a specification given seeds have been proposed by Satoshi Sato and Yasuhiro Sasaki in 2003 [86], Then they used seed term to find the related term from the web. Some queries were generated against each term for collecting top pages. A corpus is also generated of the sentences containing these seed terms for producing candidate term. To extract terminology most researcher use statistics information (such as term frequency and n-grams), but to replace statistics Hulth [36] developed a supervised algorithm adding linguistic knowledge (such as syntactic features). The results prove that a higher precision has achieved by extracting NP-chunks than n-grams.

Zhongchao Fei, Liu Jian, and Wu Gengfeng [121], in Sentiment Classification Using Phrase Patterns application to build a sense of classification opinions phrase used to classify samples. They add special tags to some words in the text

and then get the tags to fit within a sentence with some phrase patterns and the atmosphere. In this paper, the method is basically build a set pattern and then calculate the assessment of orientation to unsupervised learning algorithm. You are adding the sentence alignment, the document classification phase. And for every sentence that they add the mood for orientation, to the rate and classification of the text according to their sums. And procedures used in this paper achieved an accuracy will be to identify reviews of several sites of 86% during exercise. [121]

A new approach AMOD is introduced to automatically extracts positive or negative adjectives in the context of relevant domain by Harb et al, in 2008 [26]. AMOD works in three phases. First step deals with the automatic extraction of documents containing positive or negative opinions for a specific domain. Two different sets containing positive and negative words are defined. From these seed words, they search on the web and collected the documents containing the seed words and generate corpora. Then from these corpora, they focused of extracting adjectives from the documents relevant to the domains. Relevant adjectives are those which are highly correlated with seed word. Then they applied a number of methods to filter out the unwanted adjectives. After filtering, classification is applied to rank each document according to negative or positive opinions obtained in previous phase. This is done by finding out the difference between positive and negative adjectives. If the difference is positive the document will classify as neutral. Experiments conducted on training sets in the domain of "cinema" shows that this approach is able to extract relevant adjectives for a specific domain.

2.2.2 Sentence Level Opinion Mining

Document-level sentiment classification is too coarse for most applications. Therefore the research trend is shifting towards Sentence level. In sentence level sentiment analysis the main focus is on to identify the subjective sentences. The classification of the award is done with objective and subjective judgments. There are two main opinion mining tasks at the sentence level:

- Task1 is to identify subjective or opinionated sentences and the classes for it are objective and subjective.
- Task 2 is to form sentiment classification of sentences and classes for it are positive, negative and neutral.

Most of the work on sentence level sentiment analysis focused on the identification of subjective sentences in the news articles. All techniques use some forms of machine learning.

A lot of work is done in sentence level opinion mining. Sentence level opinion mining Using learnt patterns Rilloff and Wiebe [83], Subjectivity and polarity (orientation) Yu and Hatzivassiloglou [111], and Finding strength of opinions at the clause level Wilson et al.[102], all these are a notable work in this regard.

To find the strength of opinions Wilson et al. [103], used a new idea of syntactic clues. For finding the opinions strength and subjectivity they use wide range of features. They also use the same method for classifying the subjectivity for even deeply nested clauses.

Ellen Riloff, Janyce Wiebe and William Phillips in 2005 [82] conducted study in 2005 by implementing automatic filtering system to ensure the accuracy of the information. that system seem to guarantee better with subjective classification

They illustrate an IE system that uses a subjective sentence classifier to filter its extractions. For This purpose they experimented with many strategies and used AutoSlog-TS extraction pattern learning algorithm (Riloff et al.).First of all they used the Rule based classifiers on unlabeled corpus of articles from the world press to train the data, after that is used to train naïve base classifier but the end result system was not effective on their IE experiments for the MUC-4 terrorism domain. Thus, they retrained it on the MUC-4 training set. Finally, they achieved 52% Recall and 42 % Precision on test data set with 397 extracted patterns and without any subjectivity classification. Their final IE system with subjectivity

filtering achieved the 48% Precision and 51% recall that is 4% greater and 1% low with baseline respectively.

Riloff, E. & Weibe, J. in 2003 [83], presents a bootstrapping approach at sentence level. In this approach, to automatically identify the subjective or objective sentences, a higher precision classifier is being used. With this method, a series of samples from identified subjective and objective sets won and then the same rules be applied to more subjective and objective sentences and the whole process could be repeated as necessary to extract. The experimental results demonstrate that extraction pattern learner gathers pattern biased towards subjective texts.

Kim & Myaeng in 2007 [47], present "Opinion Analysis based on Lexical Clues and their Expansion" . They suggest that the semi-supervised learning method in high-precision seed rules. The main purpose of this method is the better combination of rule-based algorithms and techniques of machine learning. Their work begins with the discovery of subjectivity in the record. In the second step assigns subjective sentences in a system of three categories: positive, negative or neutral. In the third and final step, they try to identify the opinion holders for the subjective judgments. The experimental results show that the system reaches 45% of the power to extract identifying opinionated sentences and 35% of the power holders opinion.

Youngho Kimand Sung Hyon Myaeng [110] provides a method to identify the holder of an opinion on an anaphora resolution technique is based. This technique exploits the use of lexical and structural information. Supposedly, the process required lexical information (eg, "said") to improve the performance of syntactic features. They developed lexical and syntactic features for anaphora resolution of noun phrases (ie, opinion holder) is used. The system to determine whether the owners are anaphoric opinion or not and find actual owners. They tested method that is available on-line news articles, from NTCIR-6 and show hypothesis. The goal is that the task of solving novel approach to working reference resolution. The system consists of two main components, the subjectivity of classification

And labeling are the owner of the opinion. The system suggests a statistical learning model based on conditional probabilities. They develop guided lexical and syntactic features for the model on the training examples. After completion of all steps of each sentence is associated with a triplet <Opinionated, Anaphoricity, Holder>.. To test their methodology on online news documents and obtained 72.22% accuracy for the task of non-anaphoric opinion holder resolution and 69.89% in accuracy for the task of anaphoric opinion holder identification. The idea of the lexical evidence based opinion mining of Kim & Myaeng in 2007 [47], they have used both rule based approach and machine learning based approach. Their work begins with the identification of the subjectivity of each sentence. The next step, they are classified the sentence into positive, negative or neutral. After finding the emotional classification of the sentence, attempted to the opinion of the holder of each sentence that is classified. To identify these opinion holders, they developed a set of rules on the basis of the acquired lexical information. For the opinion of the holder, they focused on the person or organization and see it as their opinion holders classified in each set.

Another approach for the opinion mining [Ghose, Ipeirotis & Sundararajan 2007] used is by using econometrics. They have examined the effect of opinions on the pricing power of the merchants. It is like that if a customer comments on some product then how much it could be effected on the product whether he wants to purchase it or not. In this research they have described the data into two categories. Transaction Data contains the details of the transactions and Reputation Data contains the reputation of each merchant. On the basis of the data obtained they combine the econometric techniques with natural language processing techniques to find the semantic orientation and then evaluated the feedback strength. They have used a number of methods to calculate the scoring of a product purchase by the customer on the basis of customer opinion on that product.

2.2.3 Feature Level Opinion Mining

Classify evaluative texts say not to the level of the document or sentence level, what the opinion holder and may dislikes. A positive document to an object does not mean that the opinion holder has positive opinions on all aspects or features of the object. Similarly, a negative document does not mean that the opinion holder dislikes everything about the object. In an evaluative document (for example, a product review), owner of opinion writing usually both positive and negative aspects of the object, even though the general mood on the object can positively or negatively. To such detailed aspects, goes to the feature level is required [49].

Ana-Maria Popescu & Oren Etzioni in 2005 [5], presented extracting product features and opinion from reviews. They introduced an unsupervised information extraction system OPINE for extracting the frequent features, their opinion words and polarity of opinion words from customer reviews. OPINE consists of two main components of PMI-FEATURE function to find the precise feature extraction and the use of relaxation labeling in order, the semantic orientation of potential opinion words leads. [35] and [43] both had presented systems to extract product features but OPINE achieve 22% high precision than other systems.

Bing Liu, Minqing Hu, Junsheng in 2005 [53], presented Opinion Observer: Analyzing and Comparing Opinions on the Web. They proposed Opinion Observer system which shows the strength and weakness of opinion of competing products from the customer reviews which is very helpful for manufacturer and new customers. Opinion observer works in 3 steps. First step is to extract all the explicit and implicit features of product. Second step is to create opinion segments of features which consist on sentences express positive and negative opinion. In third step positive opinion set of features created. They also proposed a supervised rule mining technique based on language pattern to automatically extract the product features from Pros and Cons format of customer reviews. To extract the features first of all tag the reviews, replace all the features by word feature to make it generalize, use n gram modeling to make shorter segments, generate the rules using association rule mining CBA and extract the frequent

features by matching the patterns. Finally creates the positive and negative opinion sets of corresponding features from the given pros and cons reviews. Their Experimental results show that system is highly effective.

Mohammed Salem Binwahlan, Naomie Salim, and Ladda Suanmali [55] presented Swarm Based Features Selection for Text Summarization. Their task is to find the weights of each feature based on selection score. They use particle swarm optimization to find the effects of feature structure on feature selection and assumed that combined features have high priority then individual features. As in text summarization already used features are the base to score the sentences so Swarm optimization selects the features and their weights to assign the score to sentences and selected top scored sentences as summary. The final results of features weights show that feature structure plays an important role in the features selection process.

Jingye Wang and Heng Ren in 2009 [39] presented Feature-based Customer Review Mining in 2009. The paper is about to find out the overall rating of product on the basis of customer reviews on features of product by using feature level classification. This system works in seven steps. First of all extract the customer feedbacks from the websites and add them to the review database; second, use POS Tagging to tag these reviews; third, by employing the subject/object review separation described in (Yeh, 2006), eliminate the objective descriptions, which are not related to the opinion of the customers, from the reviews. Then PMI-IR [96] algorithm is used to calculate the mutual information between the review and the polarized words to generate the weight of the ratings. The overall rating is the weighted average of each single rating coming along with the product review. Then find the opinion words along with the features. Finally, for each product, the program returns an overall rating and several pairs of <feature, opinion words> as the summary of the reviews. [39]

Liu, Hu & Cheng in 2005 have introduced the concept of comparing opinions on multiple competing products. Their technique was to handle the reviews where user is asked to write the Pros and Cons separately and also have to write the

detailed review. In their work, they facilitate the user with some visual effects by which user can easily examine the positive and negative features of a certain product and also able to compare it with other products of the same kind. In this research, a new techniques based on language pattern mining is introduced to deal with such kind of the reviews and extract product features where Pros and Cons are separately defined along with the detail review of the product.

Although many researches are being made for the feature level opinion mining but to group the synonyms, no much work is present. [Carenini, Ng & Zwart 2005] gave the idea to deal with such problems where different users can express their views with different words on the same feature of the product. They did not give any idea to extract feature, but using previous work for feature extraction they added user defined taxonomy of features. By finding the new user defined features, the main concept of their research was the similarity matching and grouping the features which match with each other.

Instead of using machine learning and information retrieval technologies, another approach to extract features with particle swarm optimization is introduced by [Liangtu & Xiaoming 2007]. Their technique starts with the selection of the features and then coding them and in the next step they run optimization procedure to find the optimized solution.

Ding, Liu & Yu in 2008 introduced a holistic lexicon based approach to handle opinion words that are context dependent. Also they deal with a number of words or phrases which have effect on the opinion words. They have showed that not all the opinion words extracted have the same effect as they show. For example; *"not like"* in this phrase the word like has the positive orientation but complete phrase has a negative orientation. There are other words and phrases are also present which effect the opinion word but could not be find out by traditional techniques.

2.3 Opinion Summarization

Most of the people think that opinion summarization is same as text summarization but actually it is not. Text summarization is to construct the summery of long text and that summary completely express the subject of original long text while opinion summarization is to give the overall sentiment of large amount of reviews. Some of the Systems such as Opinion Observer, WebFoutain, Kanayama's system and OPINE use the linguistic resources in opinion summarization.

Some of the opinion summarization systems are:

OPINE: in 2005 [78]. This is a Web-based, domain independent information extraction system. This system performs the four main tasks

- Extract the product features
- Extract the opinion regarding to product features
- Determination of opinion polarity
- Rank the opinion according to its strength

It extracts the product characteristics on which comments the user directly with PMI. It uses explicit feature to identify potential opinion phrases, which is assigned in the area of explicit product feature. After extracting the expression of opinion by using the technique of unsupervised classification views are marked. Accordingly, the series (feature, ranked opinion list) tuples are extracted.

Bing Liu, Minqing Hu, Junsheng presented Opinion Observer in 2005[53]. A Web-based Sentiment Analysis System is used by Hu et al. for analyzing and comparing opinions. As with other systems use compression opinion, this system also to the product properties of nouns or noun phrases extract by the Association Miner or CBA. You only use adjectives as opinion words and use the method of exploring WordNet before polarity to assign opinion words. To determine the

TH 7485

polarity of the opinioned rates, which is a set of function and position of words classified as the dominant orientation contains. All features and the extract opinion words with distinction between positive and negative opinions stored in a database. Then the result is presented in graphical form.

2.3.1 WebFoutain (2005)

This system was developed by Yi et al. [109], this system also use to extract the features from given document using bBNP (beginning definite Base Noun Phrase). bBNP extract the definite base noun phrases at the beginning of sentences followed by a verb phrases. Two linguistic resources (sentiment lexicon and sentiment pattern database) are used to assign the opinion word to the concerning feature by traversing and parsing the document. Both the linguistic resources are used to assign the polarity to features and store the sentiment assignment patterns of predicates. So a simple wed interface can develop or the end user which shows the list of sentimental sentences of the given problem.

2.3.2 Review Seer (2003)

Review Seer [15] uses statistical models (e.g. statistical substitution, linguistic substitution) depending on the extraction of terms characteristic of the given material. To improve the performance feature extraction, N-grams proximity and substrings have been applied. In addition, the authors have used Naive Bayes classifier ensembles with positive opinions and negative attribution score under feature extracted. The results were presented as mere opinion sentences.

Christopher Scaffidi, Kevin Bierhoff, Eric Chang, Herman Ng and Chun Jin in 2007 [88] presented Red Opal: product-feature scoring from reviews. This system also uses to extract feature and their corresponding opinion words from the given reviews. Features are extracted by the frequent nouns and noun phrases and then assign the sentiments based on star rating. Results are shown on web interface by descending the order of each feature and the opinion word.

2.3.3 Kanayama's et al. 2007

Kanayama et al. [28] proposes a sentiment analysis system, with a high accuracy

rate and at a low cost from a transfer-based machine translation engine. Sentiment units of assessments by the full analysis and top-down tree-matching extracted with a syntactic parser and matching patterns and lexicons polarity. Sentiments can be extracted easily summarized and visualized in different formats.

System	Sentiment Resource	Syntactic Analysis	Extracting Opinion Expression		Presentation
			Feature Extraction	Sentiment Assignment	
Review Seer(2003)	Thumbs up/down	No	Probabilistic model Naïve Bayes Classifier		List sentences contain the feature term
Red Opal (2007)	Star rating		Frequent noun and noun phrase	Average star rating	Order products by score of each feature
Opinion Observer (2004)	Linguistic Resource		CBA miner Infrequent feature selection	WordNet exploring Dominant polarity of each phrase	Bar graph
Kanayama's System (2004)		Yes	Sentiment unit Modifying the machine translation framework		N/A
WebFountain (2005)			bBNP heuristic	Sentiment lexicon Sentiment pattern database	List sentences which bear sentiment of a product
OPINE (2005)			Web PMI	Relaxation labeling	N/A

Table 2.3 Characteristics of the six systems for opinion summarizations[28]

A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining

Is different from previous approaches, relying on natural language processing techniques .Turney [96] or statistical information, Hu and Liu [35] Wei Jin Ho & [39], they propose a new framework integrates linguistic features (eg part -of-speech, sentences' internal formation patterns, and the surrounding contextual clues of words / phrases) in the automatic learning of lexicalized HMMs supported. A company based method to convert each opinion sentence in a series of product groups, institutions, organizations and opinion orientations were manually labeled with the tag sets and compared to bootstrapping process approach can extract the high confidence data by self-learning.

The survey results show that the proposed approach of machine learning much better than the reference system based on rules mining company notes, recognition set opinion and the opinion polarity classification rarely mentioned

facilities are identified effectively and efficiently, which has been ignored or under analyzed by methods previously proposed.

Mining Sentiment Classification from Political Web Logs by Kathleen T. Durant & Michael D. Smith [44] in his research address a difficult problem than classification of tradition text by applying Naïve Bayes and SVMs on political web log posts for sentiment classification. The results clearly indicate that a Naïve Bayes classifier significantly outperforms Support Vector Machines at a confident level of 99%, with a confidence interval of [1.425, 3.488].

2.4 Limitations

Although the researchers some very useful techniques for Opinion Mining have outlined, but there are a number of questions that remain. We will discuss these points step by step and see how these techniques to produce a number of disadvantages in the processing of the emotional orientation of the document.

- In Bing Liu approach opinion always classified only in two categories positive and negative but Neutral opinion also expressed sometimes. Liu considers only adjective as opinion words but opinion can also expressed as adverb, adjectives and verb. For example "like" is a verb but also an opinion word. His approach finds the implicit features because it extracts the sentences contain at least one feature word. So the features commented by customer indirectly are ignored.
- Lexicon based methods use for opinion mining has not an effective method to deal with context dependent words. For example, the word can express "small" either positive or negative opinion about the product properties. For a mobile phone if customer comments that "size of mobile phone is small" this sentence does not show either size is positively opinioned or negatively.

- Drawbacks of document level classification are :

    - It does not give details of what people likes or dislikes because writer comments only the specific aspects of product.

- Document level classification is not applicable on forums and blogs as they contains only few opinion sentences because their focus is not the evaluation of the product so that it is only a few comments from users
- The document-level opinion mining predicts the orientation of document as a whole and whether a document has a positive or negative direction.
- The document level opinion mining is basically calculate the orientation of the document as whole and predicts whether a document has a positive orientation or negative. To this end, researchers have used different techniques to all records with positive or negative orientation and on the average of these rates, they classify to extract the document. But a positive phrase dose not indicates that the holder likes everything and similarly a negative phrase does not indicate that the holder dislikes everything and Similarly, not a negative pharse indicates that the owner dislikes everything. Just imagine for a moment, if a document containing review in which holder has likes the movie but dislikes its sound and picture quality. The overall sentiment of the document is negative but the holder still like the movie. So such kind of reviews shows the wrong classification. Similarly if some user dislikes the movie but likes everything else, again the review will be classified as positive due to the average orientation of the positive phrases.

- Drawbacks of Sentence level classification are :

  - Same case at the sentence level opinion mining. A sentence could contain lots of information from the user. A user can express different views in a single sentence. If a user expresses his likeness of picture quality and dislike the sound of the movie, the review will be ranked as neutral at sentence level. As such kind of the sentences the average orientation of positive and negative phrases will be equal and one cannot find out what user wants to convey. Similarly if user expresses his opinion about the

likeness of the movie in just one sentence and the rest of the sentences are expressing the user dislike some feature then the classification of the document could be wrong and will create negative impact.

- At feature level opinion mining, such kinds of problems are being solved. As it classifies the document at feature level and extracts the entire features on which user have expressed his/her opinion. In our research work we are focusing on this type of opinion mining. In previous work on feature based opinion mining [31] have used Apriori based approach to extract the product features. We will discuss the cost of Apriori based algorithms in chapter 4 and will examine the complexity one could face while using such algorithms. Moreover Apriori is not for the sequential pattern mining. It is usually used to generate the interesting rules from the dataset. By keeping these points in mind we are proposing a new approach for the opinion mining which will prove on the basis of the results that it is more efficient and accurate than the previous used techniques for feature based opinion mining.

- Drawbacks of NLP are :

  - A drawback of the NLP approach is that it could really cut very badly if they are used grammatically incorrect text. Currently a large part of the web based sentiment data fall into this category, methods to detect and correct bad English, if any, would be necessary before using them on a larger scale.

2.5 Proposed Methodology

Based on previous topics and related works, we can now present our proposed techniques. Figure 2.1 shows the methods and stages of our research.

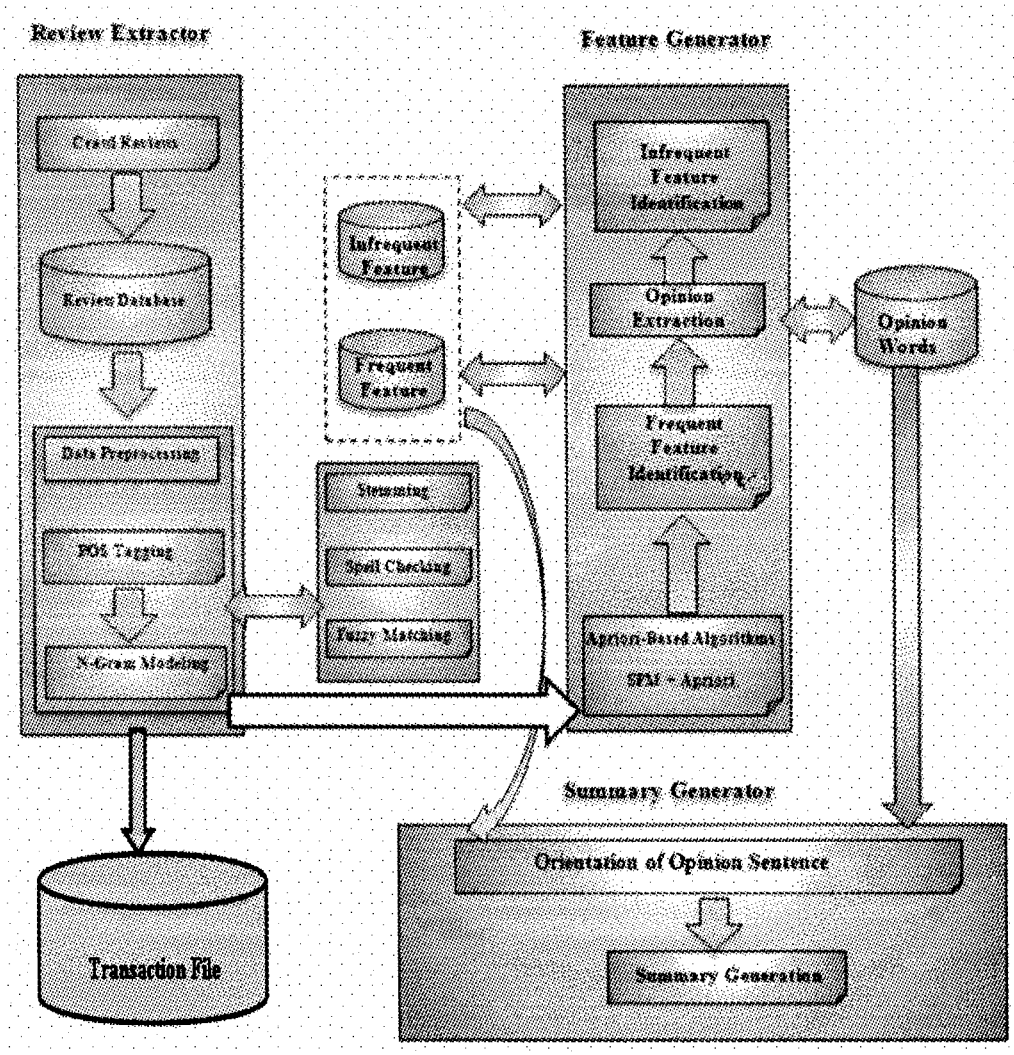


Figure 2.1 Feature and Opinion Words Extraction

In this work, we have provided the reviews are available in text form. From the review database first task is to extract frequent feature. That we have following the same methodology as in [31] explained in 2004. Our approach is for the evaluations of free format, to explain where users post their opinions in an open format. For the feature and opinion extraction following steps are needed.

- Apply Part of Speech tagging on the review database
- Prepare Tri-Model on tagged Data
- Replace the actual feature word with "$Feature" and adjectives words in a sentence with "$Opnion".

- Extract noun/noun phrases from tagged datasets
- Apply stemming on each noun/noun phrase
- Apply spell checking and fuzzy matching on each phrase
- Apply Association Rule Mining and Sequential Pattern Mining for extracting interesting and sequential patterns.
- Apply these patterns and rules to Extract noun & adjective sequences
- Finally noun/noun phrases extracted are stored in frequent feature database
- From frequent features extract opinion words.
- Stemming. Spell checking and fuzzy matching will be implemented on each opinion word
- Extracted opinion words are stored in opinion words database
- From opinion words extract infrequent features
- Again stemming, spell checking and fuzzy matching will be implemented on each infrequent features
- Extracted infrequent features will be stored in infrequent features database
- Extract and store these opinion words in a database
- Test and Find Accuracy
- Compare the results of association rule mining and sequential pattern mining

We have our proposed methodology explained step by step. The details for each step will be explained in Chapter 5 and 6.

CHAPTER 3

3. Opinion Mining

> "An information retrieval system will tend *not* to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it."
>
> – Calvin N. Mooers

This chapter gives a brief overview of the opinion mining. We will discuss different steps of opinion mining.

3.1 What is Opinion Mining?

Refers to identification of opinions about various objects on web. Opinion mining deals with the extraction of opinions from the text. Textual information consists of two things that are facts and opinions. Facts are the objective statements and Opinions are the subjective statements that show the people's sentiments about tangible events in the world. Objective mean sentiments not are involved and subjective mean sentiments are engage. Opinion can be defined as a view, attitude, or appraisal on an product. Facts are always true but opinion may or may not be true. One fact equivalent to multiple facts but one opinion is not equivalent to multiple opinions. Opinion holder is a person who has some opinion or sentiment for any object.

Opinion mining is a recent discipline at the crossroads of information retrieval and computational linguistics which tries to detect the opinions expressed in the natural language texts. Opinion mining deals with the extraction of opinions from the text. Opinion Extraction is a specified method of information extraction, delivering inputs for opinion mining. Opinion mining, also known as sentiment analysis, subjectivity analysis, reviews mining and appraisal extraction. Sentiment analysis and sentiment classification are sub-areas of opinion extraction and opinion mining. Opinion mining deals with the computational treatment of

- Positive: good, excellent, fine, etc.
- Negative: bad, worst, horrible, etc.
- Neutral: yellow, liquid, vertically, etc.

3.3 Product Reviews Mining

In this age of e-commerce products are being sold on web by the merchants. After selling their products and services merchants wants the reviews of their customers about those products to get familiar with the sentiments of customers that how much their product is admired or criticized by the customers even in all over the world. Companies will come to know that how much the product needs to be improved, and how much it is demanded in a particular region. Online reviews documents taken by the company's products sites can be in hundreds and thousands which is not possible for a reader to read at all to perceive their required information. These documents must be summarized which also makes the customer able to decide whether the product meets his requirements or not if so, than he will purchase otherwise moves on to the alternative options. This summarization totally based on particular feature (feature-based opinion summarization) of that product which is opinioned positively or negatively by the customer. There are multiple techniques to mine these features from the whole text.

3.4 Feature-Based Opinion Mining and Summarization:

The feature based opinion mining model focus on reviews. Its objective is what opinion holder liked and disliked. This model gives a complete formulation of the opinion mining problem. It identifies the values of information that should be describes how a structured opinion summary can be produced from unstructured texts.

3.4.1 Types of Features

Basically there are three main types of feature.

opinion, sentiment, and subjectivity in text. The goal of opinion mining is to enable computers to recognize expressed emotions with effective computing.

3.1.1 Fact or Opinion?

Before going further we must know what a fact or opinion. "An opinion is a state that is not open to objective observation or verification Quirk et al., in 1985 "[81]. Opinion mining deals with the extraction of opinions from the text. Textual information consists of two things that are facts and opinions. Facts are the objective statements and Opinions are the subjective statements that show the people's sentiments about tangible events in the world. Objective mean sentiments not are involved and subjective mean sentiments are engage. Opinion can be defined as a view, attitude, or appraisal on an product. Facts are always true but opinion may or may not be true. One fact equivalent to multiple facts but one opinion is not equivalent to multiple opinions. Opinion holder is a person who has some opinion or sentiment for any object.

Much of what you read in newspapers or magazines is a mix of factual information and the opinions of the author. Often the opinions are disguised as fact, to make the author's argument seem more believable.

To understand the opinion mining let us consider some examples.

- *What is the opinion of students about UOG?*
- *What is the public opinion on Iraq war?*
- *People who do not like this product and why?*

These are just few example of how one can express his/her thoughts. Opinion mining is to mine in these thoughts and identify them as positive or negative.

3.1.2 Early History

A huge burst of research activity has recently moved to opinion mining recently. However, it has a fixed interest rate for a while. Although in 2001 was the beginning of awareness of the research problems and opportunities of opinion

mining and sentiment analysis collected [11, 43,48, 85,94,8,44] . There are still land rush to the hundreds of papers are published on the subject.

Some factors include:

- Implementation of machine learning methods in natural language processing and information retrieval;
- The availability of datasets to be trained by the rapid growth of digital data have been made available in recent years, particularly the development of ratings collection web-sites.
- Emergence of Business intelligence applications
- Realization of the intellectual challenges that the area offers

3.2 Components of Opinion Mining

There are basically three components of opinion [51].

- **Opinion Holder:** Someone who holds a specific opinion on a particular product. It could a single person or an organization.
- **Object:** Object is an entity about which a person or organization express their opinion. This entity could be a person, product, organization, topic, and some event or even could be opinion. Basically, object is a hierarchy of concepts and their sub-concepts, where each concept can be associated with a set of attribute or properties. E.g.
    - *Picture quality of the player is good*

  In the above example domain is the *DVD player* and instance is *player* whereas attribute is *picture quality* which is represent an object. [31] used the word "feature" to represent the object.
- **Opinion:** Opinion is a view, attitude or appraisal which an opinion holder expressed about an object. Basically opinion could be three types

- Positive: good, excellent, fine, etc.
- Negative: bad, worst, horrible, etc.
- Neutral: yellow, liquid, vertically, etc.

3.3 Product Reviews Mining

In this age of e-commerce products are being sold on web by the merchants. After selling their products and services merchants wants the reviews of their customers about those products to get familiar with the sentiments of customers that how much their product is admired or criticized by the customers even in all over the world. Companies will come to know that how much the product needs to be improved, and how much it is demanded in a particular region. Online reviews documents taken by the company's products sites can be in hundreds and thousands which is not possible for a reader to read at all to perceive their required information. These documents must be summarized which also makes the customer able to decide whether the product meets his requirements or not if so, than he will purchase otherwise moves on to the alternative options. This summarization totally based on particular feature (feature-based opinion summarization) of that product which is opinioned positively or negatively by the customer. There are multiple techniques to mine these features from the whole text.

3.4 Feature-Based Opinion Mining and Summarization:

The feature based opinion mining model focus on reviews. Its objective is what opinion holder liked and disliked. This model gives a complete formulation of the opinion mining problem. It identifies the values of information that should be describes how a structured opinion summary can be produced from unstructured texts.

3.4.1 Types of Features

Basically there are three main types of feature.

Feature about reviews: These are those features that are about the review which are given by customer.

Feature about reviewers: These are those features that are about the person who give the review.

Feature about product reviewed: These are those features that are about the products on which customer give the review.

3.5 Opinion Formats on Web

There are three formats in which customers can give opinions. These three formats of opinions are used to carry out research experiments. Results are produced using these data formats. These formats are as follows:

- Free format
- Pros and Cons format
- Mix format

3.5.1 Free format

In this format user can write freely comments on any product. The user is not bounded to follow any format. In our research experiments we are using free format of reviews. Example from amazon.com of customer review on Canon Power Shot in a free format:

"Picture Quality very poor, or it involves too many menu settings to take decent shot .I recently purchased this camera to take pictures of the indoor Children's programs I do. Even when I changed the camera, to Program Mode, Manual, and Auto, all the images had a lot of noise in the colors. Even if you adjust the ISO to take faster pictures, the graininess becomes overwhelming....". [33]

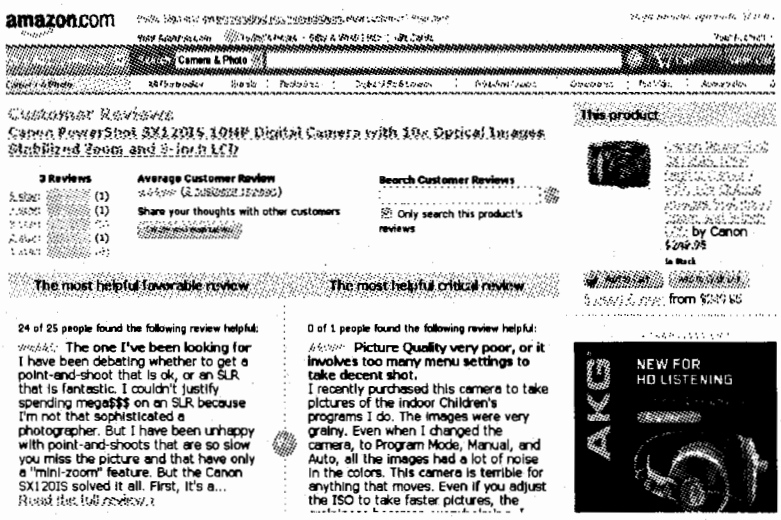


Figure 3.1 Free format

3.5.2 Pros and Cons Format

In this format the users have to write pros and cons about the product separately. Example from Epinions.com of customer review on Canon PowerShot SX10 IS 10-Megapixel Digital Camera in a Pros and Cons format:

Pros: Price, 20x zoom, 28mm wide angle, solid build, features, face detection, battery life resolution.

Cons: Uses 4 AA batteries, heavier than Panasonic

Mega-zoom cameras (cameras that have 10x optical zoom or over) are fun. **Before digital mega-zoom cameras were available, you would have to buy and, not less important, carry around a bunch of lenses and a camera body to be able** to get to 10x.....

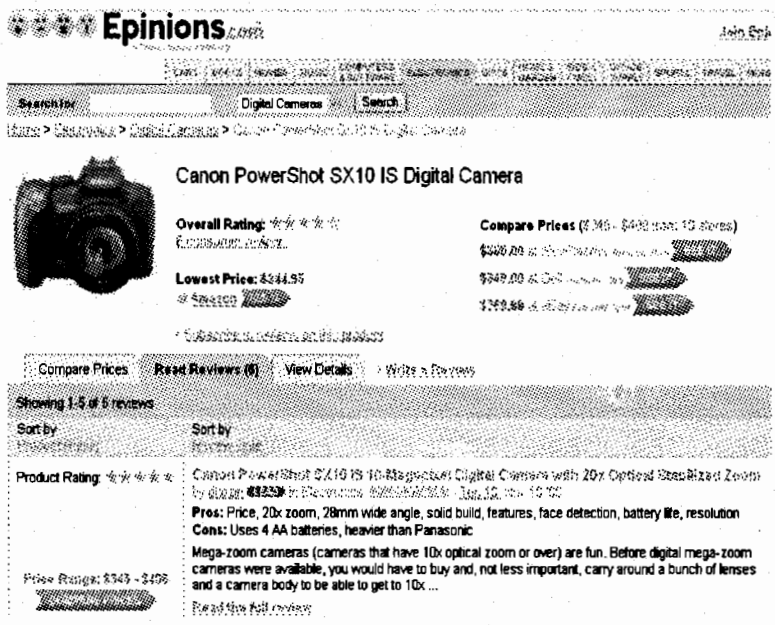


Figure 3.2 Pros and cons format

3.5.3 Mix Format

In this format users give comment in pros and cons format along with the detailed reviews, therefore it is called mix format. Example from Cnet.com of customer review on digital camera Samsung GX-1S in a mix format:

"Samsung GX-1S

Reviewed on 04/14/2006

Updated on 02/27/2007

The bottom line: Beginners will applaud the Samsung GX-1S's automated features, but advanced shooters might wince at the menu-centric controls……"

Figure 3.3 Mix format

3.5.4 Steps for Feature-Based Opinion Summarization

1. Extract the opinion features and rank these features according to their frequencies (no. of occurrences in review document).
2. Identify how many opinions are positive or negative along each feature; associate each opinion word with its feature.

Example:

Summary of the review of cell phone is:
Sound quality:
Positive 355 (potential customer review)
Negative 17 (potential customer review)
Size:
Positive 275 (potential customer review)
Negative 45 (potential customer review)
And so on…

The above like picture helps the customer more to get feel about a particular product and its particular feature in which customer is more interested.

3.5.5 Summarization Technique

Summarization system takes the product name and entry document of customer reviews as input, processes it and finally displays the summary of reviews as an output.

3.6 Related Technologies in Opinion Mining

Researchers have used various methods for opinion mining. Some well-known methods in respect with opinion mining have given the overview of some techniques regarding to opinion mining.

3.6.1 Data Mining Methods

Mining is a full-fledge domain that is used for the extraction of interesting knowledge (rules, regularities, patterns, constraints) from data. It structures the data in such a way that information can be extracted easily in the desirable form. Application areas for the data mining are marketing, investment, fraud Detection, manufacturing and telecommunications. There are two primary goals of data mining that are: prediction and description. In predictive data mining the variables are involved to know the values of the unknown data sets. Whereas descriptive data mining is used for giving the description of the pattern extracted from the data.

Mining further has different sub domains such as Data mining, Text mining and Opinion mining. Data mining deals with the extraction of patterns from data, Text mining deals with the extraction of high quality of information from the text, and this extraction is done through the statistical pattern learning and opinion mining deals with the extraction of opinions about different products from the customers.

Data mining can be applied to any size of data. Many researchers have used data mining techniques to extract knowledge from data for the sentiment classification. [35] have shown a very good approach to combine data mining with NLP techniques which shows very interesting results.

3.6.2 Natural Language Processing (NLP) Methods

Natural language processing (NLP) is a field of computer science concerned with the interactions between computers and human (natural) languages. Part of speech tagging (POS) from NLP can defined as assigned text to each word in review on the basis of characteristics of the word and the context in which it occurs. It helps

for text analyzing. POS tagging is a perfect example of the more general sequence tagging task, in which a sequence of words is mapped with a one-to-one mapping between words and tags.

Major tasks in NLP are following.

- Natural language generation
- Natural language understanding
- Information retrieval
- Automatic summarization
- Information extraction

Many researchers have used NLP techniques along with the machine learning techniques for the sentiment orientation of the documents. [102] combines machine learning and NLP techniques for sentiment analysis at clause level. Similarly [103] used appraisal theory along with support vector machines.

3.6.3 Text Mining

Text mining is an application of data mining. Text mining is used to find out innovative knowledge." A key element in the text mining is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining finds overall trends in textual data" [15]. Text mining is done for the news groups, emails and documents etc. Text mining provided a roadmap for the opinion mining.

3.6.4 Sentiment Analysis

The data about the customers' reviews is subjective. It contains sentiments of the customers regarding products and services. So for the resourceful information retrieval there is also a need to understand the customers' sentiments hidden in the reviews, for this purpose sentiment analysis is done. When the customers' express their opinions about the products, they write something positive, negative or

neutral about the product or service. Subjectivity and polarity are important measures in the sentiment analysis. For the correct prediction and analysis about the products it is important to do the proper sentiment analysis.

3.6.5 Informational Retrieval Methods

Information retrieval concerns with the searching for document which are of human interest and also the information within the document. In information retrieval methods user enters a search query into the system. This method extracts all the relative objects within the system which matches the input query. It could be a single object or a number of relevant objects. Most of the information retrieval methods compute a numeric score as how accurate each objects matches to the input query and hence it ranks the object according to the calculated score. The objects having the higher score are then shown to the user. Many researchers have used these kind of methods like [15] and [26] for the opinion extraction from the documents.

3.7 Tools for Opinion Mining

There are some tools for the opinion mining but some of the important are discussed below:

- **Sentimetrix** : Provides sentiment mining technologies to social media analysis shops.
- **Sentiment Metrics** : A competitor to Visible Technologies and others in the enterprise facing space.
- **OpinMind :** OpinMind has a new look, but its ad (audience) targeting system is differentiated by their opinion mining technology.
- **Summize** : "Summize connects you with the opinions of millions of people on zillions of products. We scour the web for the latest user reviews and blogger discussions, surmise their sentiments, and summarize them here."[123]

CHAPTER 4

4. Sequential Pattern Mining

> "In the 21st century, you have to be able to monitor where information is going and who's tapping into it, you have to be able to do massive data mining, and nobody can do that today."
>
> – Curt Weldon

In this chapter we will discuss sequential patterns and different algorithms and approaches for solving this problem. Also, the approach we are using for opinion mining will be discussed briefly.

4.1 Introduction

Sequential pattern discovery has an active research field in recent years. The application of sequential pattern mining covers a wide spectrum, suggests the analysis user access patterns of a site, the protein motif discovery from the analysis of the workload from a large computer system of the child-abuse cases, etc. The variety of applications that there may not be possible to apply a single model of sequential patterns for all these problems. Each application may require a unique model and solution. To develop meaningful sequential patterns models, some research projects in last years have been established that produce efficient algorithms for mining. Most of these models are one of the following four categories, patterns, frequent, regular pattern, statistically significant patterns, and the approximate model. We describe something from the state of the art execution within these fields in this chapter.

4.2 Types of Patterns

Before going further, we must first discus some very interesting kinds of patterns which will help us to understand the domain and our purpose of using the sequential patterns. There are basically four kinds of patters.

- **Trend Analysis:** is to find the evolution patterns of attributes over time, they can be long-term trend movements, cyclic movements or variations, seasonal movement and irregular/random movements. This method is widely used in stock market.
- **Similarity Search:** tries to find sequences that differ only slightly. Similarity searching is a blurry matching process that can tolerate some differences within a certain threshold. Based on the length of sequences we can try to match, sequence matching can be classified as: subsequence matching and whole sequence matching.
- **Sequential Patterns:** mining is trying to find the relationships between occurrences of sequential events, to find if there exists any specific order of the occurrences. We can find the sequential pattern of specific individual items; also we can find the sequential patterns cross different items.
- **Periodical Patterns:** are those recurring patterns in the time-series database, periodicity can be daily, monthly, seasonal or yearly.

Although from last several years time-related data mining has been a widely researched area. But sequential databases consist of patterns without the relation of time. Also our research has also no concern with the time related database. So we will only discuss sequential patterns in this section.

4.3 Characteristics of Sequence Data

For sequential data mining, the several distinct characteristics of sequence data lead to many opportunities and challenges. These include the following:

- Due to diversities of applications, sequences may exist in a large variation of length. For example, the length of a gene can be as large as over 100K, and as small as several hundreds.

- Someone may/may not be interested in Absolute positions in sequence. Someone may want to look for patterns which can occur anywhere in the sequences. But the positions changes when insertion/deletion performs.

- "The relative ordering/positional relationship between elements in sequences is often important. In sequences, the fact that one element occurs to the left of another is usually different from the fact that the first element occurs to the right of the second. Moreover, the distance between two elements is also often significant. The relative ordering/positional relationship between elements is unique to sequences, and is not a factor for relational data or other high dimensional data such as microarray gene expression data". [90]

- "Patterns can be substrings or subsequences. Sometimes a pattern must occur as a substring (of consecutive elements) in a sequence, without gaps between elements. At other times, the elements in a pattern can occur as a subsequence (allowing gaps between matching elements) of a sequence ". [90]

4.4 Sequential Patterns

Sequential pattern mining has been an active research topic from past several years. Sequential pattern mining was first introduced by [Agrawal & Srikant 1995].

> *"Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specific minimum support threshold, sequential pattern mining is to find all frequent subsequences, i.e., the subsequences*

whose occurrence frequency in the set of sequences is no less than minimum support."

Sequential patterns indicate the correlation between transactions while association rule represents relationship within transactions. In association rule mining, the results are about which items are brought together frequently and these items should be from the same transaction. On the other hand, sequential pattern mining shows which items are brought together in a certain order and those could be from different transactions. Sequential patterns can be used in different areas like; mining user access patterns for web sites, using the history of symptoms to predict certain disease and by using sequential patterns, retailers can control their inventory and make it more efficient.

Sequential pattern mining is the process of extract sequential patterns, whose support exceed from a predefined minimal support threshold. As the database could be very large and a huge number of patterns can be extracted which are of different interest of users, usually a minimum support is pre-defined by the user. With the help of this support, those patterns which are of no interest to the user can be pruned out.

4.5 Sequential Pattern Mining

Sequential pattern mining [75] is an essential task in sequence data mining. The sequential pattern mining problem was first addressed by Agrawal and Srikant (1995) and was defined to be:

> "Given a database of sequences, where each sequence consists of a list of transactions ordered by transaction time and each transaction is a set of items, sequential pattern mining is to discover all sequential patterns with a user-specified minimum support, where the support of a pattern is the number of data-sequences that contain the pattern."

Garofalakis, Rastogi and Shim (1999) described it as

> "Given a set of data sequences, the problem is to discover sub-sequences that are frequent, i.e. the percentage of data sequences containing them exceeds a user-specified minimum support",

while Masseglia, Poncelet and Teisseire (2000) describe it as

> ". . . the discovery of temporal relations between facts embedded in a database",

and Zaki (2001b) as

> ". . . to discover a set of attributes, shared across time among a large number of objects in a given database".

Extraction of sequential patterns exceeding a user defined minimal support threshold is called sequential pattern mining [91]. The limit of threshold depends upon the length of sequence and to different requirements. This would eliminate the sequential patterns of no interests and increase the efficiency of mining process. There is another metric named as surprise introduced by Yang et al. in 2001 [105] to measure the interestingness of those sequences that do not satisfy the support threshold.

4.6 Application Areas of Sequential Pattern Mining

Sequential pattern mining is used in a great spectrum of areas.

- In computational biology, sequential pattern mining is used to analyze the mutation patterns of different amino acids
- Analyzing medical records of patients for temporal patterns between diagnosis, symptoms, examination results, and treatment etc.
- Business organizations use sequential pattern mining to study customer behaviors.

- Discovering relationships between stock market events
- Discovering patterns among different socio-economic events
- Analyzing data from scientific experiments conducted over a period of time.
- Discovering sequential relationship between different telecommunication switches and alarms triggering and in system performance analysis and telecommunication network analysis
- studying the workload of a large computer system to the child-abuse cases
- analyzing user access patterns of a web site to the protein motif discovery

4.7 Sequential Pattern Mining Approaches

From the last several years, the approaches proposed by researchers for mining sequential patterns do not utilize the universal formulation of patterns [63]. Due to the great diversity of applications, most of the algorithms are designed to solve specific application related research problems. In the following, some general and basic algorithms suitable for variety of problems are presented and will briefly discus the algorithm we are using in this research.

4.7.1 Apriori-Based Algorithms

Most of the fundamental algorithms for sequential pattern mining are based on the Apriori property proposed in association rule mining by Agrawal and Srikant 1994. "This property states the any sub-pattern of a frequent pattern must be frequent". A series of Apriori-like algorithms were suggested on the basis of this heuristic:

- AprioriAll [Agrawal and Srikant 1995]
- AprioriSome [Agrawal and Srikant 1995]
- DynamicSome [Agrawal and Srikant 1995]
- GSP [Srikant and Agrawal 1996]
- SPADE [Zaki 2001]

- Afterward another a series of data projection based algorithms were proposed, which includes
  - FreeSpan [Han et al. 2000]
  - PrefixSpan [Pei et al. 2001].
  - MEMISP [Lin and Lee 2002]
  - SPIRIT [Garofalakis et al. 1999]

4.7.2 Horizontal Database Format

In horizontal formatting, the original data is sorted then customer sequence database is generated by transaction time. As the original data, Table 4.1(a), is sorted first by Customer Id and then by Transaction Time which results in a transformed customer sequence database, Table 4.2(b), where the timestamps from Table 4.2(a) are used to determine the order of events, which is used as the basis for mining. The mining is then carried out using a breadth-first approach [4].

Table 4.1 Horizontal Formatting Data Layout – adapted from Agrawal and Srikant (1995)

Customer Id	Transaction Time	Items Bought
1	June 25 '03	30
1	June 30 '03	90
2	June 10 '03	10, 20
2	June 15 '03	30
2	June 20 '03	40, 60, 70
3	June 25 '03	30, 50, 70
4	June 25 '03	30
4	June 30 '03	40, 70
4	July 25 '03	90
5	June 12 '03	90

Table 4. 1(a) Customer Transaction Database

Customer Id	Customer Sequence
1	⟨ (30) (90) ⟩
2	⟨ (10 20) (30) (40 60 70) ⟩
3	⟨ (30 50 70) ⟩
4	⟨ (30) (40 70) (90) ⟩
5	⟨ (90) ⟩

Table 4.1 (b) Customer Sequence version of the Database

4.7.3 Horizontal Database Format Algorithms

4.7.3.1 AprioriAll, AprioriSome and DynamicSome

In data mining Apriori is a standard algorithm for learning association rule first identified by [Agrawal & Srikant 1995].They proposed three algorithms AprioriAll, AprioriSome and AprioriDynamic. The Apriori family of algorithms has typically been used to discover intra-transaction associations and then to generate rules about the discovered associations, however the sequence mining task is defined as discovering inter-transaction associations sequential patterns across the same, or similar data. It is not surprising then that the first algorithms to deal with this change in focus were based on the Apriori algorithm (Agrawal and Srikant, 1994) using transactional databases as their data source.

The mining process in their work was decomposed in five phases.

Sort phase: This phase effectively transforms the dataset from the original transaction database to a sequence database by sorting the dataset by major key and minor key. Table 4.2(a) shows the sorted transaction data [91].

Table 4.2 Large Itemsets and a possible mapping – Agrawal and Srikant (1995).

Customer-id	Transaction-time	Purchased-items
1	Oct 23' 02	30
1	Oct 28' 02	90
2	Oct 18' 02	10, 20
2	Oct 21' 02	30
2	Oct 27' 02	40, 60, 70
3	Oct 15' 02	30, 50, 70
4	Oct 08' 02	30
4	Oct 16' 02	40, 70
4	Oct 25' 02	90
5	Oct 20' 02	90

Table 4.2 (a) Sorted Transaction Data

L-Itemset phase: The function of this phase is to find the set of all litemsets L, according to the predefined support threshold. For example, the minimal support is 40%, using the data from Table 4.1 (a), the minimal support count is 2. A possible mapping for the large itemsets is depicted in Table 4.2(b) [3][4].

Large Itemsets	Mapped To
(30)	1
(40)	2
(70)	3
(40, 70)	4
(90)	5

Table 4.2 (b) Large Itemsets MinSupp=40%

Transformation phase: To make mining more efficient there is a need to repeatedly determine which of a given set of long sequences are contained in a customer sequence, each customer sequence is transformed by replacing each transaction with the set of litemsets contained in that transaction. Sequences are not retained that do not contain any litemsets and transactions that do not contain any litemsets are dropped. But dropped sequences still contribute to the total count. This is depicted in Table 4.3.

Table 4.3 The transformed database including the mappings – Agrawal and Srikant (1995).

Customer-id	Customer Sequence	Transformed DB	After Mapping
1	⟨(30)(90)⟩	⟨{(30)}{(90)}⟩	⟨{1}{5}⟩
2	⟨(10, 20)(30)(40, 60,70)⟩	⟨{(30)}{(40)(70) (40,70)}⟩	⟨{1}{2,3,4}⟩
3	⟨(30, 50, 70)⟩	⟨{(30)(70)}⟩	⟨{1, 3}⟩
4	⟨(30)(40, 70)(90)⟩	⟨{(30)}{(40)(70)(40, 70)}{(90)}⟩	⟨{1}{2, 3, 4}{5}⟩
5	⟨(90)⟩	⟨{(90)}⟩	⟨{5}⟩

Transformed Database

Sequence phase: This phase mines the set of litemsets to discover all frequent sequential patterns generated from the transformed sequential database.

Maximal phase: since we are only interested in maximum sequential patterns, this is designed to find all maximal sequences among the set of large sequences. Pruning takes place to those sequential patterns that are contained in other super sequential pattern in this phase [91].

4.7.3.2 Difference between AprioriAll, AprioriSome and DynamicSome

Maximal phase is applicable to all of the algorithms. The AprioriAll algorithm counts all of the sequences whereas the AprioriSome and DynamicSome are designed to only produce maximal sequences. Both these algorithms first counting longer sequences and only counting shorter ones which are not contained in longer ones. This leads to save time by not counting non-maximal sequences. The method of finding all subsets of a given itemset is similar in all three algorithms. Only the implementation method of counting the sequences produced makes the difference in the three algorithms [99].

4.7.3.3 Generalized Sequential Patterns (GSP)

GSP also an Apriori based algorithm proposed by [Srikant & Agrawal 1996]. GSP integrates with time constraints and relaxes the definition of transaction. This algorithm makes multiple passes over the data. The support of each item is determined in the first pass and shows which items are frequent i.e. having minimum support. By this 1-element frequent sequence is generated. These frequent sequences will be used as seed set for the next pass. The support of these

sequences is calculated during the pass over data. At the end, the algorithm determines the actually frequent candidate sequences. If there are no more frequent candidates then the algorithm terminates. Basically for this algorithm, we need to specify two detail i.e. candidate generation and counting candidate. GSP has shown some better results than AprioriAll. In AprioriAll the counts for the candidates is higher than GSP and AprioriAll has to find which frequent item sets are present in each element of data-sequence. This is a slow approach to find the candidate sequences. A GSP-based algorithm called MFS (Mining Frequent Sequences) was proposed by [Zhang et al. 2001]. Rather than scanning the database several times, MFS proposed a two-stage algorithm. MFS produces same set of frequent sequences as GSP but reduces the cost of I/O. A GSP-based algorithm called MFS (Mining Frequent Sequences) was proposed by Zhang et al. in 2001[117]. Rather than scanning the database several times, MFS proposed a two-stage algorithm. MFS produces same set of frequent sequences as GSP but reduces the cost of I/O.

4.7.3.4 PSP

The PSP algorithm proposed by Masseglia, Cathala and Poncelet in 1998 [56] focused upon improvement to perform retrieval optimizations. For discovery of frequent sequences scan approach and candidate generation are used. For optimizing retrieval the PSP approach organizes the candidates in a prefix-tree according to their common elements which results in less memory overhead Instead of using hash tables at each internal node of the candidate tree.

Initial sub-sequences common to several candidates are stored only once by the tree structure used in this algorithm. And the terminal node of any branch stores the support of the sequence to any considered leaf inclusively. The support value of candidates is added by navigating to each leaf in the tree and then simply incrementing the value. This method is much faster than the GSP approach [101].

"Figure 4.5 illustrates a comparison of tree structures by using the set of frequent 2-sequences: L2 = {(10) (30)}, {(10) (40)}, {(30) (20)}, {(30) (40)}, {(40 10)}."

The state of the tress clearly shows the reduction of overhead of the PSP approach after generating the 3-candidates [103].

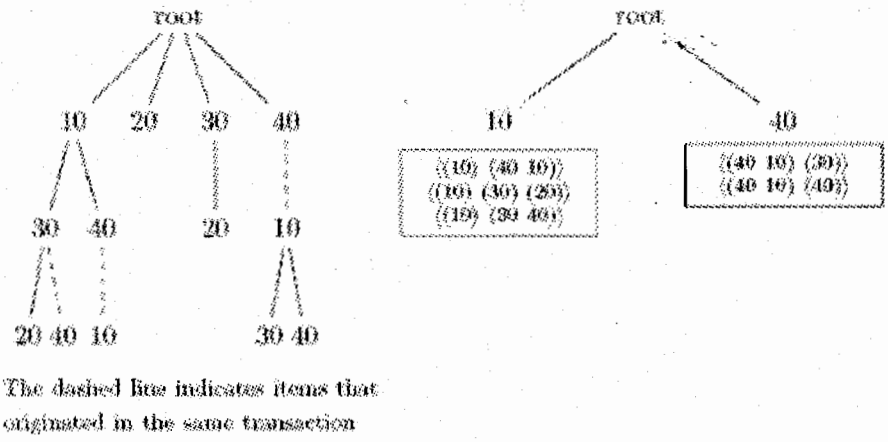


Figure 4.1 The Prefix-tree of PSP (left tree) and the hash tree of GSP (right tree) showing storage after candidate -3 generation – Masseglia et al. (1998)

4.7.3.5 RE-Hackle: Regular Expression-Highly Adaptive Constrained Local Extractor

The RE-Hackle algorithm (Regular Expression-Highly Adaptive Constrained Local Extractor) by Albert-Lorincz and Boulicaut (2003b) [3] use the same technique and has already been discussed for the SPIRIT family of algorithms, however, uses a hierarchical representation of Regular Expressions which it stores in a Hackle-tree rather than the Finite State Automaton used in the SPIRIT algorithms.

4.7.3.6 MSPS: Maximal sequential Patterns using Sampling

Luo and Chung in 2004 [54] implements super sequence frequency pruning for mining maximal frequent sequences by combining the approach taken in GSP in the MSPS (Maximal Sequential Patterns using Sampling). Because of the implementation of two pruning techniques named as subsequence infrequency and super sequence frequent pruning the proposed algorithm reduce greater space than others. The super sequence frequency pruning method considers any subsequence of a frequent sequence as a frequent from candidates item set. MSPS proves to be an efficient scalable algorithm.

4.7.4 Vertical Database Format

In vertical database format, data is organized in a fashion where the rows of the database consist of object-time stamped pairs associated with an event. His helps in generating id lists for each event that consists of the object-timestamp rows.

An example of this type of format is shown in Tables 4.4(a) and 4.4(b). Yang, Wang, Yu and Han in 2002 [106] proves through their results that the generation and counting of candidates becomes much easier and in case of long patterns it performs better and reduce processing time.

(a) Input Sequence Database

Sequence Id	Time	Items
1	10	C D
1	15	A B C
1	20	A B F
1	25	A C D F
2	15	A B F
2	20	E
3	10	A B F
4	10	D G H
4	20	B F
4	25	A G H

(b) Id-Lists for the Items

A		B		D		F	
SID	EID	SID	EID	SID	EID	SID	E
1	15	1	15	1	10	1	
1	20	1	20	1	25	1	
1	25	2	15	4	10	2	
2	15	3	10			3	
3	10	4	20			4	
4	25						

SID: Sequence id

EID: Time

Table 4.4 Vertical Formatting Data Layout – Zaki (2001b)

4.7.5 Vertical Database Format Algorithms

4.7.5.1 SPADE: Sequential Pattern Discovery using Equivalence Classes

SPADE (Sequential PAttern Discovery using Equivalence classes) is sequential pattern mining algorithm which adopts vertical data format [115] and its variant cSPADE (constrained SPADE) [113] use combinatorial properties and lattice based search techniques and allow constraints to be placed on the mined sequences. In this algorithm all the sequences are discovered with three pass on

the database or just once on some pre-processed data. It decomposes the mining problem into sub-problems which can be fitted in the main memory. In this approach a vertical id list is formed for each item. With sequence identifier of the sequence this item appeared in and their corresponding time stamps that each list contains this information.

Two search strategies are proposed for finding sequences in the lattices:

1. Breadth-first search: the lattice of equivalence classes is explored in a bottom-up manner and all child classes at each level are processed before moving to the next.

2. Depth-first search: all equivalence classes for each path are processed before moving to the next path [99].

4.7.5.2 SPAM: sequential Pattern Mining using a Bitmap Representation

SPAM (Sequential PAttern Mining using A Bitmap Representation) Ayres, Flannick, Gehrke and Yiu, in 2002 [6] uses a novel depth-first traversal of the search space with effective pruning mechanisms and a vertical bitmap database is used in which each bitmap has a bit corresponding to each element of the sequences which enables efficient support counting.[99]

The algorithm used two steps of PSP and GSP algorithms. Using a lexicographic sequence lattice or tree (the same type as used in PSP) and using an S-step process and Itemset extended using an I-step process. This process is exactly similar as adopted in GSP [92] and PSP [57].

4.7.5.3 Cache-based Constrained Sequence Miner

The CCSM (Cache-based Constrained Sequence Miner) algorithm proposed by Orlando, Perego and Silvestri in 2004 [68] uses a level-wise approach. This algorithm overcomes most of the problems attendant

with this type of algorithm. The algorithm used k-way intersections of id-lists to compute the support of candidates combined with a cache that stores intermediate id-lists for future reuse [99].

4.7.5.4 LAPIN-SPAM: Last Position Induction Sequential Pattern Mining

LAPIN-SPAM (Last Position INduction Sequential PAttern Mining) algorithm by Yang and Kitsuregawa in 2005[107] is based on the same approach as SPAM Ayres et al., in 2002[6]. The only difference exits in candidate verification and counting methods based on the observation that if the last position of item α is smaller than, or equal to, the position of the last item in a sequence s, then item α cannot be appended to s as a (k+1)-length sequence extension in the same sequence [108].

4.8 Projection-Based Algorithms

This category is more efficient one than others. The researchers recognized earlier the need for new method because of shortcomings of Apriori-like property.

- **Potentially huge sets of candidate sequences.**
  Even for moderate seed set Apriori type algorithms generates large number candidate sequences. This is due to the fact that it includes all the possible permutations of the elements.
- **Multiple scans of databases.**
  The scanning process becomes larger to check a large set of candidates by some method of pattern matching. To find a sequential pattern (abc)(abc)(abc)(abc)(abc)(abc), an Apriori-like method must scan the database at least 15 times.[90]

The recognition of these problems in the first instance in Association Mining gave rise to, in that domain, the frequent pattern growth paradigm and the FP-Growth algorithm [24]. This new methodology named as frequent pattern growth removes the need for the candidate generation and prune steps by representing squeezed

frequent sequences into a frequent pattern tree and then dividing this tree into a set of projected databases, which are mined separately [25].

4.8.1 Free Span: Frequent pattern-projected Sequential Pattern Mining

FreeSpan (Frequent pattern-projected Sequential Pattern Mining) is projection based algorithm proposed by Han, Pei, et al, 2000 [24].This approach works on divide and conquers phenomena. The aim of FreeSpan is to integrate the mining of frequent sequences with that of frequent patterns and use projected sequence databases to confine the search and growth of the subsequence fragments (Han et al., 2000).[99]

The generation of candidate sub-sequences is significantly reduced by using projected sequences.

4.8.2 PrefixSpan

To overcome the limitation of GSP and AprioriAll in case of long sequential patterns the authors of PrefixSpan (Prefix-projected Sequential Pattern Mining) [29] build on the idea of prefix projection to extract the complete set of frequent sequential patterns. The idea of PrefixSpan was given by [Pei et al. 2001]. PrefixSpan is the most efficient approach for sequential pattern mining among all other well known algorithms like GSP, SPAD, FreeSpan and Apriori. PrefixSpan is capable of dealing with very large databases with a very efficient manner. PrefixSpan mainly employs the method of database projection to make the database much smaller for the next pass and obviously makes the algorithm speedier. As there is no candidate generation in this approach. It only recursively projects the database according to their prefix. [Pei et al. 2001] has proposed a number of algorithms for PrefixSpan as level-by-level projection, bi-level projection and pseudo projection. Latter on in their next version of PrefixSpan [Pei et al. 2004] have dropped the bi-level projection, in performance improvement is only marginal in certain cases, but can barely offset its overhead in many cases.

For this approach, ordering or items within an element does not affect the sequential mining. PrefixSpan algorithm mines the sequential patterns in following steps;

- ❖ Find length-1 sequential patterns
- ❖ Divide search space
- ❖ Find subsets of sequential patterns

[Pei et al. 2004] have proved in their work that sequential patterns mined after these steps are same as generated from GSP and FreeSpan. More over them have mentioned three main benefits of using PrefixSpan.

- ❖ No candidate sequence needs to be generated by PrefixSpan
- ❖ Projected databases keep shrinking

4.8.3 Summery Pattern Growth Algorithms

Algorithm Name	Author	Year	Notes
Pattern Growth			
FREquENt pattern-projected Sequential PAtterN mining (FreeSpan)	Han, Pei, Mortazavi-Asl, Chen, Dayal and Hsu	2000	Projected sequence database
PREFIX-projected Sequential PAttern mining (PrefixSpan)	Pei, Han, Mortazavi-Asl, Pinto, Chen, Dayal and Hsu	2001	Projected prefix database
Sequential pattern mining with Length-decreasing suPport (SLPMiner)	Seno and Karypis	2002	Length-decreasing support

Table 4.5 A summary of pattern growth algorithms.

4.9 SPM in our Technique

In our technique the two algorithms we are applying Apriori and GSP which are SPAM base. In SPAM technique apply Apriori algorithm for extracting rules by giving different minimum support and metric type values. How much support is minimum then best or accurate rule will be generated. Metric type contains four values (Lift, Confidence, Leverage, Conviction). In Apriori time stamp not matter

(DNA sequencing). It is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no time stamp (Winepi and Minepi).

NN VBZ	1st level sequence
NN NNP RBS	2nd level sequence

Sequential Pattern Mining Steps:

1. Use tagger to tag the product reviews and then make the TRI MODEL in (W1, W2, W3) form.
2. Make the .arff extension file and then by using WEKA tools we find the different combinations of TRI MODEL
3. Apply APRIORI and G.S.P. (Generalized sequential pattern) algorithms in WEKA
4. Extract the combinations in the sequence of (W1, W2, W3), (W1, W2) or (W2, W3) compare these sequences to remaining reviews without applying above algorithms.

4.10 Future Trends

For efficiently discovering sequential patterns, there are several methods are available. Today, there is a significant and growing interest in the creation of algorithms to accommodate different datasets which are applicable for a large number of applications. Existing algorithms are applicable mostly on data of type binary and static [58].

Sequential pattern mining is also applicable in a much older yet related field is that of approximate string matching and searching, where the strings are viewed as sequences of tokens.

From the last several years, researchers have widely studied the sequential pattern mining problems and hence there exist a number of approaches used by different people. In this section we will discuss some general and basic algorithms and then we will outline their limitations and will briefly discus the algorithm we are using in this research.

CHAPTER 5

5. Purposed Solution

In this chapter we present the steps of our overall process we went through to the research done. The complete architecture for the extraction of frequent and infrequent feature will be discussed. Also we will see how to extract opinion words from the review database. This chapter begins a part of speech tagging, and then we discuss some problems in the data set that we encountered during our research and continue towards feature words and opinion extraction.

5.1 Introduction

With the rapid expansion in customer reviews, it is difficult to follow for retailers, manufacturers of the product, or customers and it is difficult to follow for retailer, manufacturer of the product or customer and to receive a complete view of the customer opinions in relationship to products of the interest. Therefore the development of automatic technical summary of customer's opinion is crucial. Opinion summary consists generally of two major tasks:

Product feature & opinion extraction

- Opinion orientation identification.

Because of the considerable effectiveness of opinion summary, the product feature extraction is substantial because its effectiveness is greatly influence the performance of the determination of the orientation of the opinion and the crucial effectiveness of the summary of the opinion. Consequently, we concentrate on the extraction of product features and extraction of the opinions of the reviews. This study assumes the fact that nouns or noun phrases are greatly be product features, by using sequential algorithm pattern mining exploitation of model to find all frequent itemsets which are frequent nouns and their combinations. Additionally,

adjectives are mainly used, in order to express subjective opinion of the adjacent adjective, is identified like word opinion.

Therefore, for searching infrequent features the bag of opinion words are used to look in these sentences do not find frequent features. Then Frequent and infrequent features considered as product features in combination. Following is the overview of the life cycle of an opinion mining project. The research model inside led to a number of following 3 phases

The system architecture can be divided into three sub-tasks:

(1) Review Extractor (3) Feature Generator (4) Summary Generation respectively.

The input of our system is web reviews page. Then, apply to our system, the output is the summary of user reviews .The summary is composed by product features and its' opinion sentences. In the below section, we will discuss each main task.

5.2 Reviewer Extractor

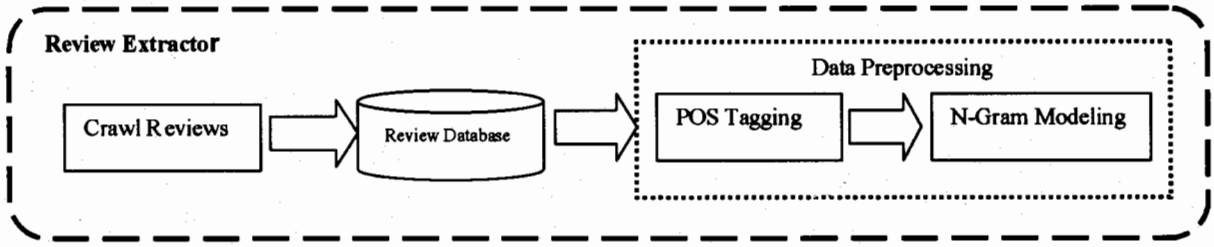


Figure 5.1 Review Extractor Architecture

First, we extract user reviews from web sites. Then, we do sentence segmentation with these reviews. Sentence segmentation step is prepared for sequential pattern mining. Each sentence is an itemset for mining. We will introduce in the next section.

5.2.1 Blog Crawler

One of the most important parts of the application is the blog crawler. The crawler has a really hard work, as is necessary to analyze as many as data that can achieve

good accuracy results. If not need to analyze is done with enough data, the results show the opinions of only limited group of people, but it is an objective of the general opinion about a product charge. So it must be crawled as many blogs as you may achieve good results, but there are some hardware limitations in this matter. The blogosphere contains very large amounts of data, but the storage capacity is low, the crawler needs fast computers with high memory to crawl all of the blogosphere, it is crawled only part of the blogosphere. It is a hypothesis that if improving the hardware specifications and crawled increased part of the blogosphere to create the application to better results.

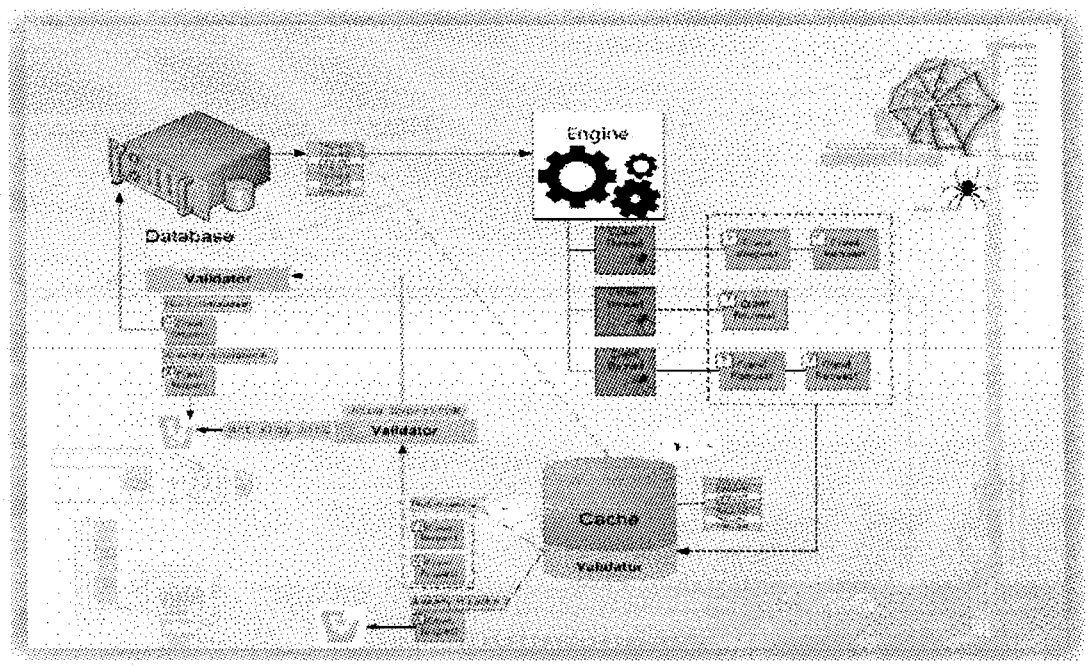


Figure 5.2 Crawler Architecture [96]

5.2.2 Data Set

As discussed in section 3.2 there are three formats of data used to extract opinions of customers about different products. We used free format data sets for opinion extraction of customers about different products. This data is also adopted by Bing Liu. This data was subjective and representing the sentiments of the customers about the products. These products are Linksys router, Canon G3, Nokia 6600, Apex AD2600 Progressive-scan DVD player, Hitachi router and Norton antivirus. The data is available in the following form.

Small [+1] ##I want to start off saying that this camera is small fo r a reason. ##Some people, in their reviews, complain about it s small size, and how it doesn't compare with larger cameras. camera[+3],size[+2]##I'm in high school, and this camera is p erfect for what I use it for, carrying it around in my pocket so I can take pictures whenever I want to, of my friends and of f unny things that happen. Memory [-2] ##the only thing I don' t like is the small size (8 MEG) memory card that comes with it. room[-2]##I have to move pictures off of it every day so I have room for more pictures the next, and I don't have enoug h money to buy the 256 MEG card that I've had my eye on fo r a while. Memory [-1][s],battery[-1]##A larger memory card and extra battery are good things to buy. Pictures [-2] ##Othe r than that pictures taken in the dark are not as nice as I'd like them,

Figure 5.3 Product Review

5.3 Data Preprocessing

In order to design a concluding data record from raw data, which is drawn into the modeling tools. The data preparation phase covers all activities.

Data preprocessing include cleaning of data by removing numbers, HTML tags, website's own information given on the web pages, symbols, spelling mistakes and the extra useless information e.g. date, name of the reviewer or reference of any third person who is not seems to be relevant. Someone said: "my brother gifted me this camera". In the described case, camera is the word for which we are collecting reviews but mentioning the name of reviewer or who gave him is useless for us. So, we had to remove such words too. Data preprocessing is also very much useful for information retrieval from the user generated contents on web, blogs, and discussion forums etc. And it promotes feature extraction precision. Tasks of data preparation are probable repeated times and not in each prescribed order to be accomplished. Tasks cover as transformation and cleaning of data for modeling the tools. Each document is preprocessed with word stemming and stops words removal and transformed into a set of transactions based on its nature of document structure.

Our Data preprocessing process includes following steps for the suitable configuration of data

5.3.1 Part-of-Speech Tagging (POS)

As discussed in section 3.2 there are three formats of data used to extract features and opinions of customers about different products. We used free format data sets for opinion extraction of customers about different products. On such kind of the data it would be very difficult to find out the features and opinions. So there would be the need for such system which will convert such text in to the noun phrases. Because our basic understanding is that usually nouns or noun phrases in review sentences represent the features of the products or the products' names, the adjectives are the opinions.

GO tagger tag whole file and also their meaning of keywords described below which show that this word is verb or noun etc. Here is the tagger named as GO tagger.

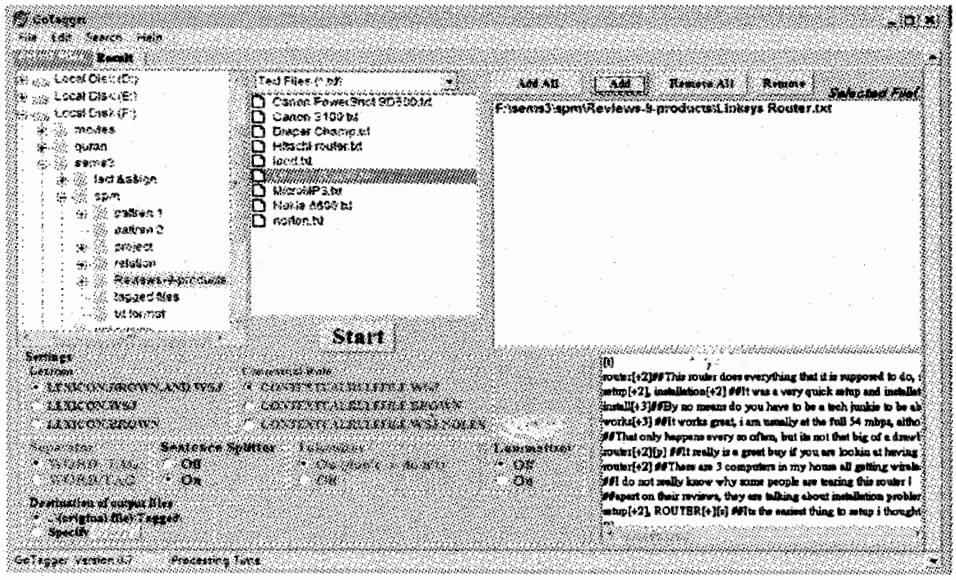


Figure 5.4 GO Tagger

To better understand the tagging let us take an example.

"This Camera Is Ideal For People Who Want More Power But Do N't Want To Spend Dollars On A Camera".

After applying GO tagger this sentence will look like the following figure.

This_DT Camera_NNP Is_VBZ Ideal_NNP For_IN People_NNS Who_WP Want_VBP More_RBR Power_NNP But_CC Do_VBP N't_NNP Want_VB To_TO Spend_VB S_NNP Dollars_NNPS On_IN A_DT Camera_NNP

Figure 5.5 Product Review with Tags

The tagger we used ,declares "VB" for verbs, "NNP" for "proper nouns", "JJ" for "Adjectives" etc.

There are other classes but our concern is only with the noun group. As all the reviews are being tagged we can easily extract noun/noun phrases from the generated tagged output. As we know that every feature is a noun/noun phrase so other tagged information is not required but could be very helpful as the work goes on.

The detailed list of all the tags we used is prescribed below:

Tag	Meaning
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition/subord. Conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund/present
VBN	Verb, past participle
VBP	Verb, non-3rd ps. sing.
VBZ	Verb, 3rd ps. sing. Present
WDT	wh-determiner
WP	wh-pronoun
WP$	Possessive wh-pronoun
WRB	wh-adverb

Table 5.1GO Tagger List of all Tags

5.3.2 Data Cleaning

Data cleaning is a process of detecting and removing inaccurate records from the data. There are number of flaws that we identify. Some of the possible flaws are mentioned below:

5.3.2.1 Defects in Dataset

In this research web crawling phase is not applied because the data sets used are available in the form of text. Therefore we are not talking about the crawling of reviews on the web. We use the same data set 1 was used by [Hu and Liu, 2004].

How do we implement the NLP Tagger to our database, we have examined a number of errors in the data set, which we downloaded from the website of the author.

Flaw 1:

- The type of dataset selected for this research is of TYPE 3 in which, customer is free to write in any case. In the given data set the word "I" is not in capital case. After tagging the dataset it was found that the GO

tagger considered small " i " as noun where ever it was used. For example following is a sentence from the dataset.

Example 1:

> "i *tried to contact apex to return it i missed the Amazon return period and It took months to get through to their customer service dept .* "

In the above example the sentence is starting with the word "i" and also there is another "i" present within the sentence. When GO tagger has applied on it following was the output as shown in Figure 5.6.

```
i_NN tried_VBD to_TO contact_VB apex_NN to_TO return_VB it_PRP
i_NN missed_VBD the_DT Amazon_NNP return_NN period_NN and_CC
It_PRP took_VBD months_NNS to_TO get_VB through_IN to_TO
their_PRP$ customer_NN service_NN dept_NN ._.
```

Figure 5.6 Output with small "i"

The above figure indicates that GO tagger considering small "i" as noun wherever it exits in the sentence. By this effect, there would be such a large number of nouns which have no effect in relation to our research and will also affect the resultant confusion matrix. But if we replace these small "i" with capital "I" then

```
I_PRP tried_VBD to_TO contact_VB apex_NN to_TO return_VB it_PRP
I_PRP missed_VBD the_DT Amazon_NNP return_NN per iod_NN
and_CC It_PRP took_VBD months_NNS to_TO get_VB through_IN
to_TO their_PRP$ customer_NN service_NN dept_NN ._.
```

Figure 5.7 Output with capital "I"

the output will be quite improved. Figure 5.7 shows the output of the same sentence with capital "I". By comparing two figures, it is obvious that in figure 5.6 there are nine nouns but in figure 5.7 it reduce to seven. This type of conversion has great impact in reducing the ratio of unwanted nouns.

Flaw 2:

- In our data set all the first letter of each sentence is in small letters. It also creates ambiguity after POS tagging.

  Let us consider an example from the same dataset.

Example 2:

> *"its fast-forward and rewind work much more smoothly and consistently than those of other models i 've had ."*

As applying GO-tagger on this sentence the output is as shown in Figure 5.8.

```
its_PRP$ fast_RB -_: forward_RB and_CC rewind_NN work_NN much_RB
more_RBR smoothly_RB and_CC consistently_RB than_IN those_DT
Of_IN other_JJ models_NNS i_NN '_POS Ve_NNP had_VBD had_
```

Figure 5.8 Output with first letter as small

This shows the following effect after converting the first letter of the sentence to capital and has found the following output.

```
its_PRP$ Fast_NNP -_: Forward_NNP and_CC rewind_NN work_NN
much_RB more_RBR smoothly_RB and_CC consistently_RB than_IN
those_DT Of_IN other_JJ models_NNS i_NN '_POS Ve_NNP had_VBD
had_VBN ._.
```

Figure 5.9 Output with first letter as capital

After exchanging the first letter in the word "fast-forward" to "Fast-Forward" the tagger consider them as separate words and tagged them to class noun from adverb class. But still in fact this conversion is not sufficient. There are a number of other words of same nature in the dataset i,g auto-focus, auto-mode. The words of this nature represent same meaning and in fact represent a single feature but POS tagger treats it differently. It tags separately at some places like. An example of that problem is.

and_CC even_RB after_IN the_DT warning_NN i_NN had_VBD enough_JJ juice_NN left_VBN to_TO finish_VB the_DT evening_NN with_IN . **autofocus_JJ**. I_PRP found_VBD best_JJS results_NNS in_IN most_RBS situation_NN with_IN manual/center_NN **auto_NN focus_NN** ,_, manual_JJ mode_NN ,_, ISO_NNP 100_CD (_(

Figure 5.10 An Example of the problem

Similarly same for appearance in review files for auto mode. They represent same feature so we consider it as same feature. Similarly there may other words exist in file like above example but we consider all of them as a single.

Flaw 3:

- The next flaw is find out in case of helping words like "couldn't", "can't", "wouldn't", etc., where the words join with the symbol "'". Across data while such words were separated by spaces. For example, the above words were written in the data as "could n't", "can 't", would n't", etc., where we can see these words separated by a space. For observing its effect after tagging, consider the following example.

  "the deal Amazon has going can 't be beat and if you're considering buying this machine, do so from this site. unfortunately it can 't play all of the dvd . we don't thought it was just the player. player can't play."

Figure 5.11 shows the output of the sentence above where these words are present.

The_DT deal_NN Amazon_NNP has_VBZ going_VBG **can_MD '_POS t_NN** be_VB beat_VBN and_CC if_IN you_PRP 're_VBP considering_VBG buying_VBG this_DT machine_NN ,_, do_VBP so_RB from_IN this_DT site_NN ._. unfortunately_RB it_PRP **can_MD '_POS t_NN** play_NN all_DT of_IN the_DT dvd_NN ._. we_PRP **do_VBP n't_RB** thought_VBD it_PRP was_VBD just_RB the_DT player_NN ._. player_NN **ca_MD n't_RB** play."_CD

Figure 5.11 Output for the negation words having spaces

From the above illustration, it being cleared that the GO tagger tagged words like these as two separate words, and especially the words that show a negative impact on the sentence, but shows no such thing by tagger used.

Consider now Figure 5.12 with output generated with GO tagger, these premises have been removed. In Figure 5.11, we can see that the two words still not treated as a word in its entirety but they are now belonging to the class of negation. Thus, these kinds of words should be treated according to their class because they can be useful for future work, treats.

The_DT deal_NN Amazon_NNP has_VBZ going_VBG ca_MD **n't_RB** be_VB beat_VBN and_CC if_IN you_PRP 're_VBP considering_VBG buying_VBG this_DT machine_NN ,_, do_VBP so_RB from_IN this_DT site_NN ._. unfortunately_RB it_PRP **ca_MD n't_RB** play_VB all_DT of_IN the_DT dvd_NN ._. we_PRP do_VBP n't_RB thought_VBD it_PRP was_VBD just_RB the_DT player_NN ._. player_NN **ca_MD n't**_RB play."_CD

Figure 5.12 Output for the negation words after removing spaces

Here we have described some significant shortcomings in the datasets, as it could be much impact of our research in the future. We have tried to overcome the marked mistakes. We all have the "I" words converted into capital, converts the first letter of the individual sentences to upper case and removes the spaces from the words, if the user provides a negative item on the Action function.

5.3.3 Stemming

Stemming is to remove suffix by automatic means in an operation. This is very useful in information retrieval field. There are many words with a common stem which have similar meaning. For example;

Connect
Connected
Connection
Connecting
Connections

All these words belong from same class connect but have different suffix. When we are extracting product features the possibility of such words is very high. If we ignore stemming then all the above words will be treated individually and hence will decrease the accuracy of the system. Just consider during the extraction of nouns/noun phrases we are encountered with such words and in the absence of stemming there would be five different kind of nouns/noun phrases will be extracted although all these have the same meaning. But due to stemming we will get only one feature which has occurred five times and hence the accuracy will be increases. We have used porter stemmer [77] in our research. It will convert all given above words into connection.

5.3.4 Fuzzy Matching and Spell Checking

A problem arises by removing all the suffixes during stemming. If there are words such as families stemmer then convert the word into famili after removing the suffix but it is not the right word. The right word is the family but not stemmer handles these kinds of problems. To overcome such problems we need to verify the spelling of each word. By check the spelling of famili be converted to family, and therefore we can overcome these problems. There are a number of spelling techniques such as use of Yahoo and Google APIs. But this type of techniques you need to be online, as these APIs check the spelling of a word from the Internet. There are other techniques that do not need to be online. We use Microsoft Office Word 2007 spell checker in our research.

Fuzzy matching is match to two strings, how much match both strings together. For example, if two strings autofocus and auto-focus then these two strings have the same meaning but are expressed to by two different ways. When we extract features, these strings can cause problems. Stemmer and spell checker are not suitable to cope with such problems then the two strings are considered individual not one . It will simply reduce the precision of our system. To do so with such problems, we implement fuzzy string matching. Instead of a fuzzy matching algorithm we have proposed our algorithm for handling these kinds of words in our system. We have extracted all the words containing '-' character (let's say in

file A). Then we have checked each word which do not contain '-' character during the extraction of nouns/noun phrases and matched them with each word in file A while removing the character '-'. If file A does contain the word then replace it with the same word. Here is our logic.

for each word z in the database
if (z contains character '-') then
save in file A as word q
for each word x in the database
if (x does not contain character '-') then
for each word q in file A
y= remove the character '-' from q
if (x = y) then
replace the word x=q

Figure 5.13 Fuzzy matching for two strings

5.4 Generalization

Data generalization is a process to get a clear view of problem/solution of data. In this process we generalized our data so that it may be testable for the other products. This research work use a supervised pattern learning approach to extract product features usually complete sentences e.g. "The pictures are very clear".

Before identifying the sequential patterns we must arrange the dataset. Usually object features are nouns/noun phrases and the nearby adjectives are usually opinion words which modify the object feature. In the original dataset all the sentences expressing opinion of the user on a specific object feature are tagged. We can find the sequential patterns from product features and opinion words. For this we replace the actual feature words with the word {$feature _ <tag>} having tags NN, NNS, NNP and NNPS, where $feature represents a feature.

Similarly replacing the actual adjectives words in a sentence with "$Opnion". Use of "$Feature" or replacement of feature word with "$Feature" is necessary because different products have different features and the replacement have

ensured that we common language patterns that can be used for each product feature to be found. Following is example after replacing noun features with "$ feature" and shown adjectives with "$ Opnion", as in Figure 5.16.

Example:

```
It_PRP 's_VBZ $Opnion_RB $Opnion _JJ $Feature_NN with_IN a_DT
$Opnion _RB $Opnion _JJ $Feature _NN $Feature _NN $Feature _NN
$Feature _NN ,_, and_CC it_PRP has_VBZ a_DT $Opnion _JJ $Feature _NN
$Feature _VBN.
```

Figure 5.14 Replacing the product features and opinion words to identify sequential patterns

Usually all JJ tags are adjective but all NN (nouns) are not features. We haven't used those sentences which don't have any orientation on the product features. As those sentences will only produce junk sequential patterns which, are not of our interest. After that we prepared the dataset by removing all the values before other tags. Then final file will in the following form as shown in Figure 5.17:

```
PRP VBZ $Opnion_RB $Opnion _JJ $Feature_NN IN DT $Opnion _RB
$Opnion _JJ $Feature _NN $Feature _NN $Feature _NN $Feature _NN ,_, CC
PRP VBZ DT $Opnion _JJ $Feature _NN $Feature _VBN.
```

Figure 5.15 Removing the values before remaining tags to identify sequential patterns

Use then a N-gramm, in order to produce shorter segments from the long.

5.5 n-Gram Modeling

Once data cleaning are complete, we convert tagged data into *n*-Grams, a standard technique in language processing.

n-gram: a sequential list of n words, used to encode the likelihood that the phrase will appear in the future

"An N-Gram model is a type of probabilistic model for predicting the next item in a sequence. N-grams are used in various areas of statistical natural language processing and genetic sequence analysis ".[33]

Then we build *n*-Gram Model. *n*-Gram modeling is used to convert unstructured data into structure data. *n*-Gram involves splitting sentence into chunks of consecutive words of length "n". *n*-Gram Modeling has many types; some of important are as follows:

- Uni-gram modeling
- Bi-gram modeling
- Tri-gram modeling
- Penta- gram modeling

The *n*-Gram with size 1 is known as unigram; with size two is bigram and with size 3 is as trigram and so on. For example,

"I don't know what to say"

1-gram (unigram): I, don't, know, what, to, say

2-gram (bigram): I don't, don't know, know what, what to, to say

3-gram (trigram): I don't know, don't know what, know what to, etc.

..... n-gram

It been used in many research works but most of the work is based on tri-gram. Min, KH, Wilson, WH, Kang, and BH compares the use of methods based on *n*-Gram in their research paper and resulted as the word trigrams method resulted in 83.8% precision, 81.2% recall ration and on the other hand the pentagrams method resulted in 86.6% precision, 82.9% recall ratio [61].

In modeling step of our research we applied Tri-Gram (3 words with their POS tags) on our cleaned data.

After that word stemming is performed. This will reduce a word to its stem.

NN,JJ_$Opnion,PRP
JJ_$Opnion,PRP,VBP
IN,DT,NN_$Feature
DT,NN_$Feature,VBZ
NN_$Feature,VBZ,JJ_$Opnion
VBZ,JJ_$Opnion,IN
JJ_$Opnion,IN,DT
IN,DT,NN_$Feature
DT,NN_$Feature,DT
NN_$Feature,DT,NNS
IN,PRPS$,JJ_$Opnion

Figure 5.16 Tri-Gram Model

As our interest is to extract features and the opinions related with them that why we seized all Noun tags (NN, NNP, NNS, etc.) along with all Adjective tags (JJ) and the Verb tags (VB, VBP, VBZ, VBN, VBD etc.) to make proper sense of our model.

5.6 Feature Generator

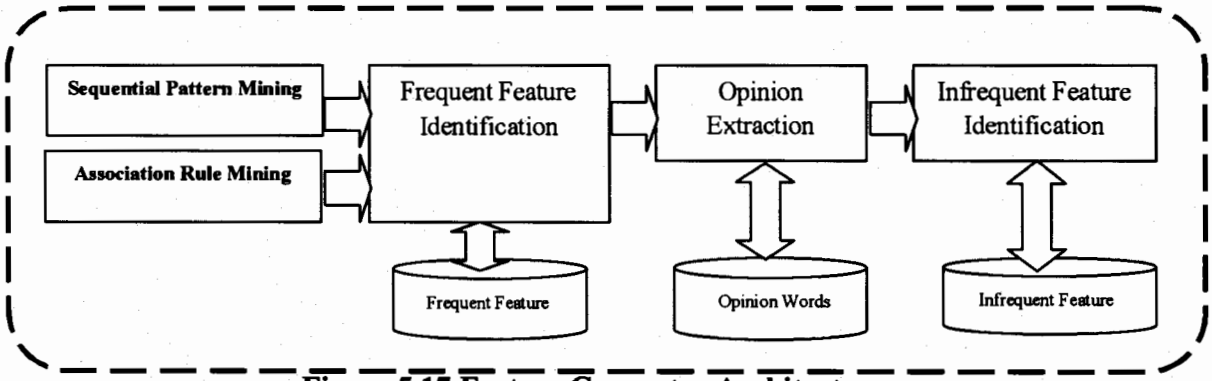


Figure 5.17 Feature Generator Architecture

In this section, we solve the problem of finding explicit feature. The method we used is based on Hu's work. The concept of his work is that he thought the frequent noun/noun phrases in the review are important features which people care about. They used association rule mining [48] technique to find these features, but he did not use associate rule; he just found the frequent itemsets. An itemset is a set of words or a phrase that occurs together in some sentence.

Consider of word sequence, we use sequential pattern mining technique to find frequent feature. The algorithm we used is GSP [29] to do sequential pattern mining. It finds all frequent itemsets in the transaction set and consider the word sequence. We set minimum support as 0.2 % (i.e. an itemset as frequent if it appears more than 1%), but not all the frequent features are real features.

In the same way Association rule mining is used to find the association rules. By applying different minimum support generate different rules and in the end gets best rules. Best rules are those that are valid sequence, usually from W1, W2, W3, or is are W1, W2, or W2, W3.

5.7 Identification of Frequent Features

In the previous section, we extracted all nouns / noun phrases of these reviews. The next step is to identify frequent features from these extracted nouns/noun phrases. Frequent features are the ones who talked most of the people. Previous work on identifying frequent features [Hu and Liu, 2004] used association rule miner, CBA Liu, Hsu & MA in 1998 [52], based on the Apriori algorithm [1]. The aim was to identify all frequent itemsets of all transactions.

Basically Apriori operates in two steps. The first step is to identify all frequent itemsets satisfying the minimum support and to generate rules in the second step. In this technique, the total numbers of frequent itemsets generated are 2n-1and from these frequent itemsets [35] has extracted frequent features with some further preprocessing (we discuss this process later). Just consider the nouns/noun phrases extracted from the Figure 5.13 which are as follows;

In this section, we solve the problem of finding explicit feature. The method we used is based on Hu's work. The concept of his work is that he thought the frequent noun/noun phrases in the review are important features which people care about. They used association rule mining [48] technique to find these features, but he did not use associate rule; he just found the frequent itemsets. An itemset is a set of words or a phrase that occurs together in some sentence.

Consider of word sequence, we use sequential pattern mining technique to find frequent feature. The algorithm we used is GSP [29] to do sequential pattern mining. It finds all frequent itemsets in the transaction set and consider the word sequence. We set minimum support as 0.2 % (i.e. an itemset as frequent if it appears more than 1%), but not all the frequent features are real features.

In the same way Association rule mining is used to find the association rules. By applying different minimum support generate different rules and in the end gets best rules. Best rules are those that are valid sequence, usually from W1, W2, W3, or is are W1, W2, or W2, W3.

5.7 Identification of Frequent Features

In the previous section, we extracted all nouns / noun phrases of these reviews. The next step is to identify frequent features from these extracted nouns/noun phrases. Frequent features are the ones who talked most of the people. Previous work on identifying frequent features [Hu and Liu, 2004] used association rule miner, CBA Liu, Hsu & MA in 1998 [52], based on the Apriori algorithm [1]. The aim was to identify all frequent itemsets of all transactions.

Basically Apriori operates in two steps. The first step is to identify all frequent itemsets satisfying the minimum support and to generate rules in the second step. In this technique, the total numbers of frequent itemsets generated are 2n-1and from these frequent itemsets [35] has extracted frequent features with some further preprocessing (we discuss this process later). Just consider the nouns/noun phrases extracted from the Figure 5.13 which are as follows;

"front panel button layout feature set"

In the above set there are six nouns which are being extracted but CBA will generate 26-1 which is 31 features with given constraints. Note that there are only two noun groups, in Figure 5.15 are in place and even within these noun groups there is only two noun phrases, the user expressed his/ her opinion. So the cost for the production of frequent features over CBA is very high. [Hu & Liu 2004], also has the support of at least 1% of the review sentences and considered that the product feature contains no more than three words. So from six words 31 frequent features are generated.

In our approach, we did not apply Apriori algorithm to generate frequent features. If there are three items *DVD, DVD player and Apex DVD player* then Apriori will generate the following itemsets with minimum support 1.

DVD	support=3
Player	support=2
DVD player	support=2
Apex	support=1
Apex DVD	support=1
Apex player	support=1
Apex DVD player	support=1

But those are all different items. Although word *DVD* is available in all products, but *DVD* separate expressing a disc, *DVD player* is a device which plays DVD and could be have any brand and *Apex DVD player* is a branded player to play a DVD. Thus, Apriori is not appropriate for such kind of items. Our approach will extract only the nouns / noun groups that users have expressed their opinion and no need to generate all possible frequent itemsets.

As we have already shown that are extracted from Figure 5.13 only two noun phrases, as follows.

"front panel button layout"

"feature set"

What features are generated, there is always the need to find support for each noun / noun phrase in the dataset. Therefore, we have used the approach similar to Apriori but we are considering the whole noun phrase as one item. Also, the minimum support for any item is set to 1% of the review sentences. Now we can define it as follows.

Definition: *If any item in the set of items has the support minimum to 1% of the total review sentences then the item would be the frequent feature.*

We do not bind the length of the noun phrase as it could be any length. We identify the frequent features with the following manner.

for all the nouns/noun phrases in the database
find the support of each noun/noun phrase
if(support of noun/noun phrase > 1% of the review sentences)
save noun/noun phrase as frequent feature
else
eliminate the noun/noun phrase

Figure 5.18 calculating the support of each noun/noun phrase

Before going further, we must keep in mind one thing. Any product feature is that on which user has expressed his/her opinion. When we extract a product feature that does not hold any opinion then this function is not a product feature. For example

"I've owned 6 or 7 DVD players since 1998."

The user talks about the DVD player, but does not express any opinion on the product. Thus, in such sentences the "DVD player" is not a feature, so we can get the list of frequent features after finding the opinion words. If we could not find any opinion words on some feature then that would not be our frequent feature.

By using this technique we have reduced the feature pruning. [31] has proposed two techniques for feature pruning. We will discuss them and will show how we have succeeded in reducing these steps.

Compactness Pruning: Compactness pruning checks only those words which have the length greater than one. This method states that if the distance between two words in any phrase is less than or equal to three and such phrases have the support greater than one in the review sentences then the phrase is said to be compact. But if the distance between any two words in a phrase is greater than three or these phrases are not compact in at least two sentences then such words are not features and pruned out from the set of features. To better understand the logic let us reconsider the example from Figure 5.13. From the nouns extracted in this figure the Apriori algorithm will generate one item as "layout feature set". If we look at the figure 5.8 then it is clear that the distance between "layout" and "feature set" is greater than three. Just assume that the phrase had the support greater than our given supports then such kind of words are handled with compactness pruning. But our purposed technique does not extract the sequences like that and hence we do not need the compactness pruning.

Redundancy Pruning: Similarly for those features which consists only a single words the redundancy pruning was proposed. It states that is any feature having length one and having support greater than the given support but appears as a subset of any feature phrase from the dataset then if its individual support is less than three which is called as *pure support* then these features will be pruned out. For example *DVD, DVD player* and *Apex DVD player* are three features. Just consider that the support of each one is 10, 3 and 3 respectively then the individual support of *DVD* would be 4. Because *DVD* is occurring three times in both *DVD player* and *Apex DVD player* and hence total occurrence of *DVD* will be ten as describe in previous examples. The individual support of *DVD* which is 4 is called *pure support*. So the redundancy pruning says that if any feature which is a member of any other feature phrase has the pure support less than three it

would not be a feature. But in the above case *DVD* has the pure support greater than three so it is a feature. In our approach we deal all the features like *DVD, DVD player* and *Apex DVD player* as separate feature and occurrence of *DVD* in other phrases does not impact on the support of *DVD* individual that's why we do not need the redundancy pruning.

The above explanation clearly indicates the benefits of our approach. Not only the cost of Apriori is eliminated but also the cost of compactness and redundancy pruning is also reduced in our approach.

5.8 Extraction of Opinion Words

Opinion words can express positive or negative opinion against some product features. Typically, these types of words are around the feature words in a sentence and can extract by using frequent words. Following are some sentences we are looking for

Sentence 1: *"It's very sleek looking with a very good front panel button layout, and it has a great feature set."*

Sentence 2:"Loved the slim design."

Feature word looking exists after the opinion word very sleek in sentence 1. Similarly, front panel button layout is right after to very good and feature set is near to great opinion word. In the same manner, in sentence 2 feature design is near to opinion word slim. We are following the same approach as [35] proposed for the extraction of opinion words.

After examining the above two sentences, it is obvious, the words expressing the opinion on some feature are objective words. So the opinion words could be extracted in the following way:

- For each sentence in the review database, if it contains any frequent feature, then the *nearby adjectives* would be the opinion words. Extract all these adjectives and store you them as opinion words.
- The opinion words modify the adjacent features. We must kept in mind one thing while extracting opinion words is that opinion words modify the adjacent feature words not the adjacent noun/noun phrase. We have extracted frequent features from nouns/noun phrases but it is not necessary that all the frequent features are noun/noun phrase. Just look at the first sentence in the above example, *look* is not a noun as cleared from Figure 5.13 but still is a frequent feature because in other sentences it appears as noun. So while extracting opinion words, extract with respect to frequent features not with the respect to nouns/noun phrases. Moreover we sue stemming, spell checking and fuzzy matching for each of the opinion words.

There is also the possibility that one frequent feature is being expressed by more than one opinion words. That's why we are using the term *nearby* so that if there are more than one opinion words which modifies a frequent feature could be extracted. We extract the opinion words in the following manner.

for each sentence in the review database
if (it contains one or more frequent features)
extract all the nearby adjectives as opinion words

Figure 5.19 Extraction of opinion words

It is clear from above diagram that we do not have to extract all the adjectives from a sentence, but rather than this we can extract only those adjectives which are nearby to the given frequent feature. We have to keep in mind one thing that nearby means all the adjectives which are near to given frequent feature but not nearest to any other feature present in that sentence.

5.9 Identification of Infrequent Features

Frequent features are the features that have most people talked about it. We may call it hot features. But there are certain hot features that are not typical opinioned by most of the reviewers called as infrequent features. However, these features may be helpful. The approach to find common features is unable to extract the infrequent features. Because they occur in very rare number of sentences and in the process of seeking support from each noun / noun phrase they could be eliminated. It must, therefore, another way to find such features. Before we continue, we examine some of the sentences that contain infrequent features.

"This is the best DVD player I've purchased."
"Have used it frequently and have had no problems."

Both sentences share the same opinion on the player. But in the first sentence, the DVD player is the common feature that the user is an expression of opinion. But in the second sentence of the user uses have no words to express product functionality, but still express his opinion. But again, this sentence is containing the opinion of the user on the drive. For the infrequent features, we can use the opinion words as one adjective used to describe different object. We follow the same approach as described by [31] for infrequent feature extraction.

- For each sentence in the review database, if it contains no frequent feature, but contains one or more opinion words, then the nearest noun / noun phrase must be found. If any nearest noun/noun phrase is found store it as infrequent feature in the set of features.

It should be clarified in the process of opinion word extraction, we used nearby term, but for the extraction of infrequent features that we have used the term nearest. These two differ from each other. The nearest noun / noun phrase is that the opinion word modifies.

This is a simple heuristic to determine the nearest noun / noun phrase needs the understanding of natural language processing is hard to understand without POS

tags. Here is the procedure for the extraction of features uncommon.

for each sentence in review database

if (it contains no frequent feature but contains one or more opinion words)

find the nearest noun/noun phrase around each opinion word which opinion word modifies and store as infrequent feature

Figure 5.20 Extraction of infrequent features

Results for the individual infrequent feature that we need to apply, stemming, spell checking and fuzzy matching. This method works well enough to find some infrequent features. There are still some problems in infrequent feature extraction. As we have used opinion words for the extraction of infrequent features but usually describe the most interesting features and the ones that are not really features people use common adjectives .As we have used opinion words for the extraction of infrequent features, but usually people use common adjectives to describe many objects, including both the most interesting features are those that are not really features. Because some of these nouns /noun phrases could be obtained at this stage, not product features.

This could increase the error rate of the system. However, this is not a very serious problem. Since we are infrequent features in comparison to the frequent features, the ratio is very low. Because, frequent feature are more important than the infrequent features, the number of infrequent features is very small and hence will not affect most of the users.

The experimental results will be that due to the extraction of rare feature of the error rate increased a little, but on the other hand, the accuracy of the system increases, so that we can neglect the error rate at this time, because we are always more accurate results.

5.10 Orientation Identification for Opinion Words

Prediction of the orientation of opinion sentence is to identify complete sentence as positive or negative. Orientation of opinion sentence is not the scope of our research but we can use the same technique as used by [Hu & Liu 2004]. They have given three different kinds of sentences.

- The opinion words are either positive or negative i.e. two positive and three negative. It means that sentence has negative orientation. The sentence will be positive in the opposite condition.
- The number of positive and negative opinion words are equal i.e. sentence has one negative opinion and one positive opinion.
- All other cases.

For the first case the orientation is simple. For the second case they have used the average orientation of the opinion words. For the third case they have used the knowledge of previous sentence and examine the orientation of previous opinion sentences.

We have extracted opinion words and now we have to find the sentiment orientation (i.e. positive or negative) of each opinion word. We are just giving the idea to identify the orientation as our work focused on the extraction of product features and not on the orientation of opinion words. The sense orientation of Word shows the direction that the word deviates from the norm for its disposition group. Words that represent the desirable state, for example, exceed, excellent, have a positive orientation. Words that represent the undesirable condition. e.g. bad, worst, have a negative orientation.

But there are some other words like play, work, which neither represents desirable neither undesirable state can be denoted as neutral words. So while predicting the sentiment orientation of opinion words we can encounter with three kinds of words i.e. positive, negative and neutral. But to find the opinion word orientation we must focus on positive and negative terms because neutral term does not indicate any thing whether the user likes or dislikes.

[Hu & Liu 2004] have used the WordNet [60]. WordNet organize adjectives in bipolar clusters. The cluster of good / bad are shown in the figure and consists of two half-clusters, one for the good and bad for others. Each half is managed by a cluster head synset. Satellite synsets are set by the head synset, the similar meanings of the adjective to the head followed.The other half cluster is headed by the reverse pair [19].

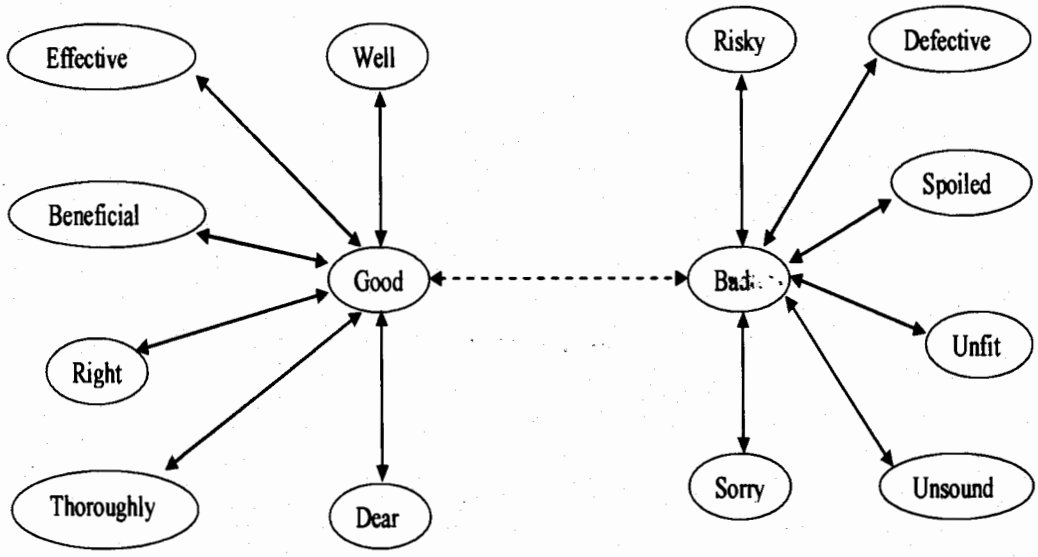
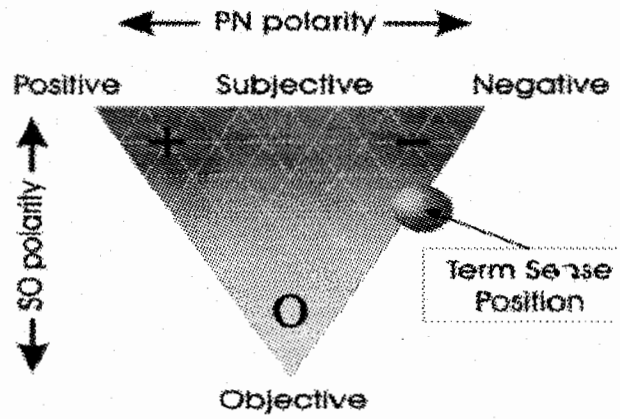


Figure 5.21 Bipolar adjective structure

The opinion means that we have extracted are adjectives that have the same orientation as their synonyms, and share as compared to their orientation as antonyms. [35] have used the same idea to the orientation of each opinion word predict. There is another way to find the orientation of words, is by aSentiWordNet [19]. SentiWordNet a lexical resource, which means each synset of WordNet three Sentiment Scores assigns positive, negative and objective. The sum of the scores is always one.

Figure 5.22 Graphical representation of SentiWordNet [Esuli & Sebastiani 2006]



The Figure 5.24 is representing the SentiWordNet. The experimental results showed by [18] of SentiWordNet express its accuracy on WordNet. We are leaving the opinion words orientation for our future work. We are recommending some techniques to find the orientation of opinion words.

5.11 Summary Generation

The final task of feature based opinion mining is to generate the summary for the whole document. In the summary generation all the features which have the same feature word are collected together and their total occurrences in the document are collect and on the bases of collect information positive and negative ranks are given to each feature. [31] have given this idea. In their approach they describe the summary generation in two steps.

- In the first step each feature and its opinion is put into a positive and negative category and for each positive or negative one a count is computed i.e. how many positive or negative times the feature has occurred.
- Feature Ranking are based on review appearances
- The features consisting of single words are placed before the feature phrases as user has more interest in single words feature.

Following is an example of showing the summary.

Feature: **Player**

Positive: 5

- Player works and looks great – if you can get the DVD's to play.
- it works and looks great - if you can play DVD's.
- .

...

Negative: 10

- This player is not worth any price and I recommend that you don't purchase it.
- The DVD player just wouldn't recognize them.

...

CHAPTER 6

6. Experiments and Results

This chapter describes the experimental evaluation of our proposed approach for opinion mining. We are using sequential pattern mining to identify interesting patterns from the review database and extract the frequent and infrequent features and opinion words. To fulfill this chapter, three aspects are discussed including experimental datasets, performance measures, and evaluation procedures. The bin liu document collection is chosen as our benchmark dataset. Most of the standard performance measures (i.e. precision, recall, FB-measure, accuracy) are used for evaluating the experimental performance. The discussion and analysis of experiments are split into two categories. In this research, we have applied Aprioi and GSP algorithms on the final data sets and prominent features separately. Text preprocessing for each document is applied before both the learning and evaluation phases. Term stemming and fuzzy matching techniques are also used in this stage. We also compare the results of frequent itemset mining (association rule mining) and sequential pattern mining methods.

6.1 Experimental Datasets

Several standard benchmark datasets are available for experimental purposes. We have used all the five datasets available of customer reviews on five electronic products i.e. 2 digital cameras, 1 DVD player, 1 mp3 player and 1 cellular phone.

Review Source: amazon.com

The corpus can also be obtained from the following web site:

[http://www.cs.uic.edu/~liub/]

The datasets were tagged and all the features were tagged manually on which users have expressed their views and those sentences which does not contain any user opinion were not tagged such information is not needed and also opinions

were identified as positive or negative (i.e. the orientation). So for each sentence which contains any product feature a manual list of product features is generated.

6.2 Preparing the Datasets

Before identifying the sequential patterns we must arrange the dataset. As we have explained before that usually object features are nouns/noun phrases and the nearby adjectives are usually opinion words which modify the object feature in chapter 5. In the original dataset all the sentences expressing opinion of the user on a specific object feature are tagged and manually features are being extracted. We can find the sequential patterns from product features and opinion words. For this we have change all the words in the datasets manually which represents a product feature with the word "$Feature" and all the opinion words which represents the opinion on the product feature with the word "$Opinion". For example following is sentence from the review dataset.

"It's very sleek looking with a very good front panel button layout, and it has a great feature set."

This sentence contains three product features *looking, front panel button layout* and *feature set* and the opinion words which modify these features are *very sleek, very good* and *great* respectively. After generating the output from GO tagger as shown in Figure 5.13 we have replace all the feature words with the word "$Feature" and all the opinion words with the word "$Opinion" as shown in figure 6.1. By this method we have replaced all the product features and opinion words of those review sentences which contains opinion on some product feature. We haven't used those sentences which don't have any orientation on the product features. As those sentences will only produce junk sequential patterns which, are

It_PRP 's_VBZ $Opnion_RB $Opnion _JJ $Feature_NN with_IN a_DT $Opnion _RB $Opnion _JJ $Feature _NN $Feature _NN button_NN $Feature _NN ,_, and_CC it_PRP has_VBZ a_DT $Opnion _JJ $Feature _NN $Feature _VBN.

Figure 6.1 Replacing the product features and opinion words to identify sequential patterns

As the dataset is being prepared we can run sequential pattern mining to find the interesting sequential patterns. For this we have used the GSP algorithm [4]. We have set minimum lower bound support 0.02.

6.3 Tri-Model

In modeling step of our research we applied Tri-Gram (3 words with their POS tags) on our cleaned data. And the main reason for using modeling rather than taking full sentences, that important features are not ignored. The use of long sentences rather than n-grams leads to generate false or fraudulent rules to make complications. A tri-model file is generated as follows.

NN,JJ_$Opnion,PRP

JJ_$Opnion,PRP,VBP

IN,DT,NN_$Feature

DT,NN_$Feature,VBZ

NN_$Feature,VBZ,JJ_$Opnion

VBZ,JJ_$Opnion,IN

JJ_$Opnion,IN,DT

IN,DT,NN_$Feature

DT,NN_$Feature,DT

NN_$Feature,DT,NNS

IN,PRPS$,JJ_$Opnion

Figure 6.2 Tri-Gram Model

6.4 Tools

6.4.1 Weka: Machine Learning Software in Java

WEKA stands for "Waikato Environment for Knowledge Analysis". For solving a number of real world problems WEKA provides a collection of algorithms of different domains which can either be applied directly to a dataset or called from users' own Java code. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. WEKA machine learning software uses ARFF2 (Attribute-Relation File Format)

ARFF documents have two clear sections: The first section is the header information, which one follows, the data information. The header of the ARFF file contains the name of the relationship, a list of attributes (columns in the data), and their types.

An example header dataset looks like this:

```
@relation breast-cancer
@attribute age {'10-19','20-29','30-39','40-49','50-59','60-
69','70-79','80-89','90-99'}
@attribute menopause {'lt40','ge40','premeno'}
@attribute node-caps {'yes','no'}
@attribute deg-malig {'1','2','3'}
@attribute breast {'left','right'}
@attribute breast-quad
{'left_up','left_low','right_up','right_low','central'}
@attribute 'irradiat' {'yes','no'}
@attribute 'Class' {'no-recurrence-events','recurrence-events'}
@data
```

6.4.2 Bing Liu's Annotated Sentences

Following is a small data set of 6 sentences taken from Bing Liu's Data Set, annotated for product features:

- Camera [+2] ##this camera is perfect for an enthusiastic amateur photographer.
- Picture [+3], macro [+3] ##the pictures are razor-sharp, even in macro.
- Size [+2][u]##it is small enough to fit easily in a coat pocket or purse .
- Weight [+1][u]##it is light enough to carry around all day without bother .
- Manual [+2] ##the manual does a fine job filling in any blanks that remain.
- camera[+2][p], use[+1][u], feature[+2]##it 's easy to use , and yet very feature rich .

6.4.3 Data Set for Weka

Following is the 3-gram data set of the above six sentences including 50+ records:

(DT This) (NN feature$) (VBZ is) (NN feature$) (VBZ is) (JJ perfect) (VBZ is) (JJ perfect) (IN for) (JJ perfect) (IN for) (DT an) (IN for) (DT an) (JJ enthusiastic) (DT an) (JJ enthusiastic) (NN amateur) (JJ enthusiastic) (NN amateur) (NN photographer) (NN amateur) (NN photographer) (. .) (DT The) (NNS feature$) (VBP are) (NNS feature$) (VBP are) (JJ razor-sharp) (JJ razor-sharp) (RB even) (IN in) (RB even) (IN in) (NN feature$) (IN in) (NN feature$) (. .) (PRP It) (VBZ is) (JJ feature$) (VBZ is) (JJ feature$) (RB enough) (JJ feature$) (RB enough) (TO to) (RB enough) (TO to) (VB fit) (TO to) (VB fit) (RB easily) (VB fit) (RB easily) (IN in) (RB easily) (IN in) (DT a) (IN in) (DT a) (NN coat) (DT a) (NN coat) (NN pocket) (NN coat) (NN pocket) (CC or) (NN pocket) (CC or) (NN purse) (CC or) (NN purse) (. .)	(PRP It) (VBZ is) (JJ feature$) (VBZ is) (JJ feature$) (RB enough) (JJ feature$) (RB enough) (TO to) (RB enough) (TO to) (VB carry) (TO to) (VB carry) (IN around) (VB carry) (IN around) (DT all) (IN around) (DT all) (NN day) (DT all) (NN day) (IN without) (NN day) (IN without) (VBP bother) (IN without) (VBP bother) (. .) (DT The) (NN feature$) (VBZ does) (NN feature$) (VBZ does) (DT a) (VBZ does) (DT a) (JJ fine) (DT a) (JJ fine) (NN job) (NN job) (VBG filling) (IN in) (VBG filling) (IN in) (DT any) (IN in) (DT any) (NNS blanks) (DT any) (NNS blanks) (IN that) (NNS blanks) (IN that) (VBP remain) (IN that) (VBP remain) (. .) (PRP It) (VBZ is) (JJ easy) (VBZ is) (JJ easy) (TO to) (JJ easy) (TO to) (VB feature$) (TO to) (VB feature$) (CC and) (VB feature$) (CC and) (RB yet) (CC and) (RB yet) (RB very) (RB yet) (RB very) (VBZ feature$) (RB very) (VBZ feature$) (JJ rich) (VBZ feature$) (JJ rich) (. .)

6.5 Sequential Patterns

GSP algorithm generates all possible sequential patterns with the given input. We studied the sequential patterns generated against the data input (as shown in Figure 6.2) and found some very interesting patterns. Some of these patterns are shown in Figure 6.3. From the figure we can find that there are sequential patterns

which are identifying that object features are usually noun/noun phrase and its opinion word is the adjective which is nearby to that noun/noun phrase. But we have found some other patterns as well.

These patterns show that not only adjectives represent the opinion of a specific product feature, but there is the possibility that opinions words are expressing in a sentence are not adjectives.

```
- 1-sequences

[1] <{JJ}> (1)
[2] <{DT}> (6)

- 2-sequences
[1] <{f_NNP}{NN}> (1)
[2] <{f_NNP}{IN}> (1)
[3] <{f_NNP}{PRP}> (1)

- 3-sequences
[1] <{f_NNP}{NN,PRP}> (1)
[2] <{f_NNP}{IN,f_NN}> (1)
[3] <{f_NNS}{IN,DT}> (1)

- 4-sequences
[1] <{f_NN,NN}{NN,DT}> (1)
[2] <{f_NN,NN}{NN,VBZ}> (1)
[3] <{f_NNP,IN}{IN,f_NN}> (1)
[4] <{f_NNS,IN}{IN,DT}> (1)
```

```
GeneralizedSequentialPatterns
=============================

Number of cycles performed: 4
Total number of frequent sequences: 119

Frequent Sequences Details (filtered):
```

```
- 1-sequences

[1] <{JJ}> (1)
[2] <{DT}> (6)
[3] <{NN}> (7)
[4] <{VBZ}> (3)
[5] <{IN}> (3)
[6] <{f_NNP}> (6)
[7] <{PRP}> (1)
[8] <{VBN}> (1)
[9] <{PRP$}> (1)
[10] <{f_NNS}> (2)
[11] <{TO}> (1)
```

```
[18] <{PRP$,NN}> (1)
[19] <{f_NN,NN}> (3)
[20] <{A_JJ,NN}> (3)
[21] <{NN}{DT}> (1)
[22] <{NN}{VBZ}> (1)
[23] <{NN,VBZ}> (1)
[24] <{f_NNP,VBZ}> (1)
[25] <{PRP,RB}> (1)
[26] <{NN,IN}> (1)
[27] <{IN}{IN}> (2)
[28] <{f_NNP,IN}> (1)
[29] <{f_NNS,IN}> (1)
[30] <{f_NN,IN}> (1)
```

[12] <{VBD}> (1)
[13] <{VB}> (1)
[14] <{f_NN}> (9)
[15] <{A_JJ}> (4)
[16] <{WRB}> (1)
[17] <{DT}> (7)
[18] <{NN}> (7)
[19] <{VBZ}> (2)
[20] <{RB}> (1)
[21] <{IN}> (4)
[22] <{f_NNP}> (6)
[23] <{PRP}> (1)
[24] <{VBN}> (1)
[25] <{PRP$}> (1)
[26] <{f_NNS}> (2)
[27] <{TO}> (1)
[28] <{VB}> (1)
[29] <{f_NN}> (7)
[30] <{WDT}> (1)
[31] <{A_JJ}> (4)
[32] <{WRB}> (1)
[33] <{A_JJS}> (1)

- 2-sequences

[1] <{f_NNP}{NN}> (1)
[2] <{f_NNP}{IN}> (1)
[3] <{f_NNP}{PRP}> (1)
[4] <{f_NNP}{f_NN}> (1)
[5] <{f_NNS}{IN}> (1)
[6] <{f_NNS}{DT}> (1)
[7] <{f_NN}{NN}> (2)
[8] <{f_NN}{f_NN}> (1)
[9] <{f_NN}{DT}> (1)
[10] <{f_NN}{VBZ}> (1)
[11] <{f_NN}{f_NN}> (1)
[12] <{NN,DT}> (1)
[13] <{VBZ,DT}> (2)
[14] <{IN,DT}> (2)
[15] <{VB,DT}> (1)
[16] <{f_NN,DT}> (1)
[17] <{NN}{NN}> (2)

[15] <{f_NNP,IN}{f_NN}> (1)
[16] <{f_NNS,IN}{IN}> (1)
[17] <{f_NNS,IN}{DT}> (1)
[18] <{f_NNP}{NN,PRP}> (1)
[19] <{f_NNP,f_NNP}{NN}> (1)
[20] <{f_NNP,f_NNP}{PRP}> (1)
[21] <{f_NN}{f_NN,f_NN}> (1)
[22] <{f_NN,f_NN}{f_NN}> (1)
[23] <{f_NN,f_NN}{f_NN}> (1)

[31] <{IN}{DT}> (1)
[32] <{IN}{f_NN}> (1)
[33] <{DT,f_NNP}> (3)
[34] <{f_NNP}{NN}> (1)
[35] <{f_NNP,f_NNP}> (3)
[36] <{f_NNP}{PRP}> (1)
[37] <{NN,PRP}> (1)
[38] <{WRB,VBN}> (1)
[39] <{f_NN,PRP$}> (1)
[40] <{f_NN,f_NNS}> (1)
[41] <{A_JJ,f_NNS}> (1)
[42] <{f_NNP,TO}> (1)
[43] <{TO,VB}> (1)
[44] <{JJ,f_NN}> (1)
[45] <{NN,f_NN}> (1)
[46] <{IN,f_NN}> (1)
[47] <{VBN,f_NN}> (1)
[48] <{VBD,f_NN}> (1)
[49] <{f_NN}{f_NN}> (1)
[50] <{f_NN,f_NN}> (2)
[51] <{f_NN}{f_NN}> (1)
[52] <{NN,WDT}> (1)
[53] <{DT,A_JJ}> (2)
[54] <{NN,A_JJ}> (1)
[55] <{VBZ,A_JJ}> (1)
[56] <{f_NNS,WRB}> (1)
[57] <{DT,A_JJS}> (1)

- 3-sequences

[1] <{f_NNP}{NN,PRP}> (1)
[2] <{f_NNP}{IN,f_NN}> (1)
[3] <{f_NNS}{IN,DT}> (1)
[4] <{f_NN}{NN,DT}> (1)
[5] <{f_NN}{NN,VBZ}> (1)
[6] <{f_NN}{f_NN,f_NN}> (1)
[7] <{NN}{NN,DT}> (1)
[8] <{NN}{NN,VBZ}> (1)
[9] <{f_NN,NN}{NN}> (2)
[10] <{f_NN,NN}{DT}> (1)
[11] <{f_NN,NN}{VBZ}> (1)
[12] <{IN}{IN,DT}> (1)
[13] <{IN}{IN,f_NN}> (1)
[14] <{f_NNP,IN}{IN}> (1)

- 4-sequences

[1] <{f_NN,NN}{NN,DT}> (1)
[2] <{f_NN,NN}{NN,VBZ}> (1)
[3] <{f_NNP,IN}{IN,f_NN}> (1)
[4] <{f_NNS,IN}{IN,DT}> (1)
[5] <{f_NNP,f_NNP}{NN,PRP}> (1)
[6] <{f_NN,f_NN}{f_NN,f_NN}> (1)

Figure 6.3 Sequential Patterns output with GSP algorithm

In GSP algorithm we only take level 2 and level 3 sequences.

6.5.1 Elimination of Extra Rules:-

In this step all the repeated sequences exits in the same file or in other files have removed. For example, sequence

<{DT,NN_F}> (3)

exist in canon SD 500 and also found in canon S100. All repeated Rules are represented as bold.

6.5.2 Best Combinations

For Feature Extraction the best combinations or rules are

- 2-sequences

```
[1] <{ VBN,f_NN }>
[2] <{ VBD,f_NN }>
[3] <{ f_NNS,WRB }>
```

- 3-sequences

```
[1] <{f_NNP}{NN,PRP}> (1)
[4] <{f_NN}{NN,DT}> (1)
[5] <{f_NN}{NN,VBZ}> (1)
[10] <{f_NN,NN}{DT}> (1)
```

These are train data rules that are getting from one product Canon SD 500. We use same procedure for remaining train files.

For Opinion Extraction the best combinations or rules are:

```
<{A_JJ,NN}>
<{A_JJ,f_NNS}>
<{DT,A_JJ}>
<{NN,A_JJ}>
<{VBZ,A_JJ}>
```

6.6 Association Rule Mining

Apirori algorithm generates all possible Association Rules with the given input. We studied the Association Rules generated against the data input (as shown in

Figure 6.2 and found some very interesting rules. Some of these patterns are shown in Figure 6.4.

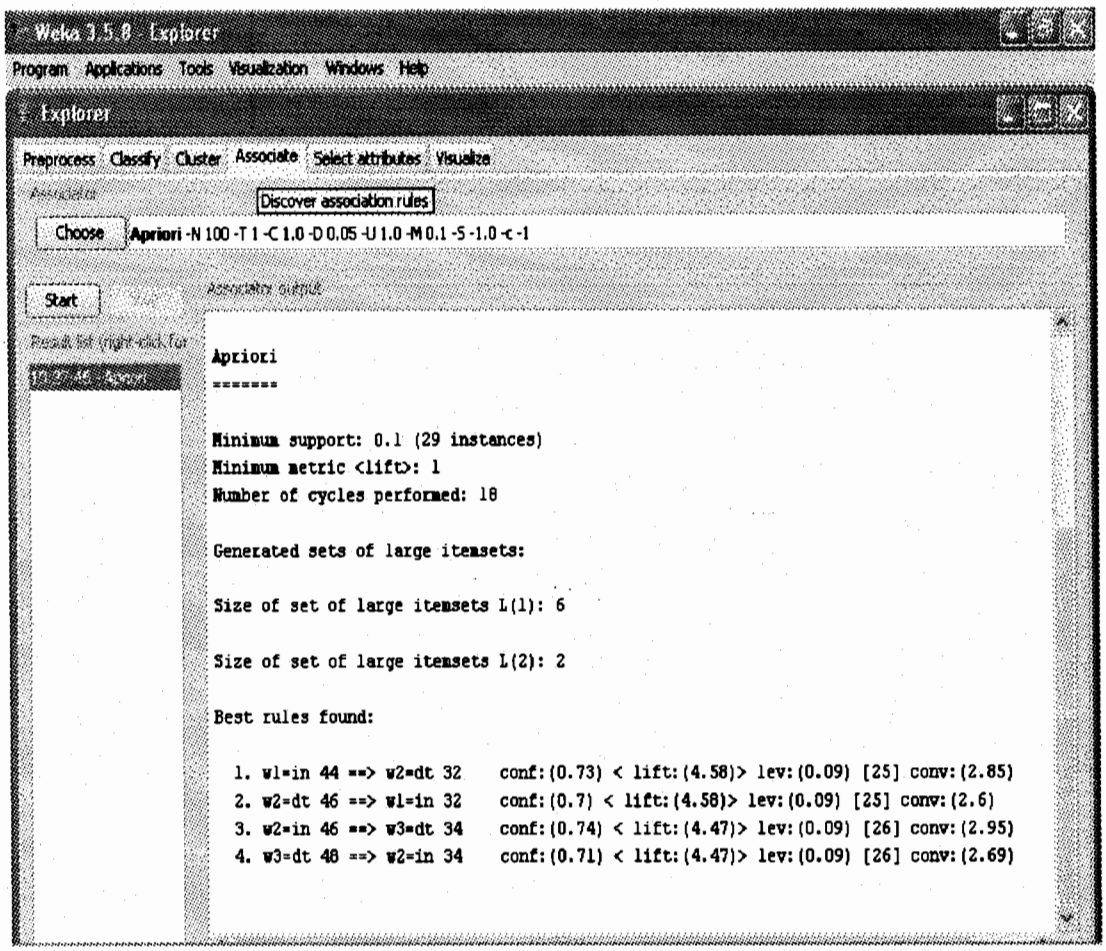


Figure 6.4 Association Rule Mining output with Apriori algorithm

In APRIORI we set the

Lower bound min support = 0.01

Number of maximum rule =100

Upper bound min support=1.0

We use different values of the above data, but these are the values most appropriate because it gives us rules with maximum noun and adjective in the right combination so our need is to find as many rules as we can because many rules that we have our results are better.

All the frequent features are not real features because some are repeated and some are indicated by tagger but in reality it's not features. There are also some unnecessary features and ones we are not interested in. We should exclude all of these features by some means. After applying APRIORI algorithm by using WEKA and applying the above mention values we get different rules.

6.6.1 Assigning Different Minimum Support

There are basically two types of Apriori and GSP algorithms. One is the Original Apriori which uses single minimum support and second is the MS-Apriori which use multiple minimum supports. Similarly for GSP the original GSP, which uses a single minimum support, and MS-GSP, which uses multiple minimum supports. When using only a single minimum support generally of the opinion that all the elements in the data of the same nature and / or have similar frequencies in the database. But it is in fact not in the real case, There are many applications in which some elements appear frequently in the data, during some other objects appears rare. If the frequencies of the individual parts differ very much, we push on two problems:

1. By setting minimum support too high, rules will not extracted involving infrequent or rare items in the data.
2. Therefore, we have to set low minimum support to find rules involving frequent and infrequent items

For example in a supermarket, people buy tea and cooking much rarer than butter and milk. The situation is the same for online stores. Thus, in the general sense, the long-term or expensive goods are purchased less frequently, but each of them generates more profit. It is therefore important to understand the rules with less frequent items. However, we must do so without letting frequent items produce too many insignificant rules with very low support.

Following are the output with different values of minimum support.

Lower bound min support = 0.1

Metric type =confidence

Number of rule =10

Upper bound min **supp**ort=1.0

Review Name: - CANON POWER SHOT SD500

OUTPUT: No large item sets and rules found!

Lower bound min support = 0.07

Metric type =lift

Number of rule =20

Upper bound min support=1.0

Review Name: - CANON POWER SHOT SD500

OUTPUT: No large item sets and rules found!

Lower bound min support = 0.03

Metric type =lift

Number of rule =10

Upper bound min support=1.0

Review Name: - CANON POWER SHOT SD500

OUTPUT:

1. w2=DT 148 ==> w3=$Feature_NN 55
2. w1=DT 167 ==> w2= $Feature _NN 57
3. w1=A_JJ 197 ==> w2=NN 46

Lower bound min support = 0.01

Metric type =lift

Number of rule =100

Upper bound min support=1.0

Review Name: - CANON POWER SHOT SD500

OUTPUT:

1. w2=TO 25 ==> w3=VB 16
2. w2=PRP 59 ==> w3=VBP 16
3. w2=PRP 59 ==> w3=VBD 14
4. w1=IN 108 ==> w2=DT w3=$Feature_NN 18
5. w2=DT 148 ==> w3=$Feature _NNP
6. w1=IN 108 ==> w2=DT 41
7. w2=IN 119 ==> w3=DT 29
8. w1=RB 56 ==> w2=A_JJ 23
9. w1=DT 167 ==> w2=$Eeature _NNP 20
10. w2=VBZ 48 ==> w3=A_JJ 20
11. w2=$Opinion_JJ 171 ==> w3=NNS 15
12. w2=JJ 40 ==> w3=$Feature _NN 18
13. w1=IN w2=DT 41 ==> w3=$Feature_NN 18
14. w1=VBZ 52 ==> w2=$Opinion _JJ 18
15. w1=JJ 39 ==> w2=$Feature _NN 16
16. w2=NN 147 ==> w3=VBZ 16
17. w2=RB 64 ==> w3=$Opinion _JJ 22
18. w1=$Feature_NN 230 ==> w2=VBZ 20
19. w1=$Opinion _JJ 197 ==> w2=NNS 16
20. w2=DT 148 ==> w3=$Featur e _NN 55
21. w2=IN 119 ==> w3=PRP
22. w1=DT 167 ==> w2=$Feature _NN 57
23. w1=$Feature_NN 230 ==> w2=CC 19
24. w1=$Opinion _JJ 197 ==> w2=NN 46
25. w2=$Opinion _JJ 171 ==> w3=NN 32
39. w1=IN 108 ==> w2=$Opinion _JJ 19
40. w1=VBZ 52 ==> w3=$Opinion _JJ 19
41. w1=NN 117 ==> w3=$Feature_NNS 14
42. w1=$Feature_NN 230 ==> w3=DT 38
43. w1=$Feature_NN 230 ==> w3=VBZ 20
44. w1=IN 108 ==> w3=$Feature_NN 34
45. w1=$Opinion _JJ 197 ==> w3=PRP 20
46. w1=NN 117 ==> w3=$Feature_NN 30
47. w1=$Opinion _JJ 197 ==> w3=NN 25
48. w1=$Feature_NN 230 ==> w3=NN 29
49. w2=IN 119 ==> w3=PRP
50. w1=DT 167 ==> w2=$Feature _NN 57
51. w1=$Feature_NN 230 ==> w2=CC 19
52. w1=$Opinion _JJ 197 ==> w2=NN 46
53. w2=$Opinion _JJ 171 ==> w3=NN 32
54. w3=NN 121 ==> w2=A_JJ 32
55. w2=$Feature_NN 203 ==> w3=CC 16
56. w2=$Feature_NN 203 ==> w3=VBZ 17
57. w1=$Feature_NN 230 ==> w2=PRP 18
58. w2=$Feature_NN 203 ==> w3=NN
59. w2=$Opinion _JJ 171 ==> w3=$Feature _NNS
60. w1=$Opinion _JJ 197 ==> w2=$Feature _NNS
61. w2=DT 148 ==> w3=A_JJ 32
62. w2=$Opinion _JJ 171 ==> w3=IN 24
63. w1=NN 117 ==> w2=IN 16
64. w1=$Opinion _JJ 197 ==> w2=IN 26

26. w3=NN 121 ==> w2=A_JJ 32	65. w1=$Feature_NN 230 ==> w2=NN 37
27. w2=$Feature_NN 203 ==> w3=CC 16	66. w1=$Feature_NN 230 ==> w2=IN 29
28. w2=$Feature_NN 203 ==> w3=VBZ 17	67. w1=IN 108 ==> w2=$Opinion _JJ 19
29. w1=$Feature_NN 230 ==> w2=PRP 18	68. w1=VBZ 52 ==> w3=$Opinion _JJ 19
30. w2=$Feature_NN 203 ==> w3=NN	69. w1=NN 117 ==> w3=$Feature_NNS 14
31. w2=$Opinion _JJ 171 ==> w3=$Feature _NNS	70. w1=$Feature_NN 230 ==> w3=DT 38
32. w1=$Opinion _JJ 197 ==> w2=$Feature _NNS	71. w1=$Feature_NN 230 ==> w3=VBZ 20
33. w2=DT 148 ==> w3=A_JJ 32	72. w1=IN 108 ==> w3=$Feature_NN 34
34. w2=$Opinion _JJ 171 ==> w3=IN 24	73. w1=$Opinion _JJ 197 ==> w3=PRP 20
35. w1=NN 117 ==> w2=IN 16	74. w1=NN 117 ==> w3=$Feature_NN 30
36. w1=$Opinion _JJ 197 ==> w2=IN 26	75. w1=$Opinion _JJ 197 ==> w3=NN 25
37. w1=$Feature_NN 230 ==> w2=NN 37	76. w1=$Feature_NN 230 ==> w3=NN 29
38. w1=$Feature_NN 230 ==> w2=IN 29	

6.6.2 Elimination of Extra Rules

Valid rules are those that are in the sequence of one of the following:

- (W1, W2, W3)
- (W1, W2)
- (W1, W3)

We extract the one to follow the order and remove the repeated rules. The final rules set are:

6.6.3 Best Combinations

1. w1=IN 108 ==> w2=DT w3=$Feature_NN 18
2. w2=JJ 40 ==> w3=$Feature_NN 18
3. w1=$Feature_NN 230 ==> w2=VBZ 20
4. w1=DT 167 ==> w2=$Feature_NN 57
5. w1=$Feature_NN 230 ==> w2=CC 19
6. w1=$Feature_NN 230 ==> w2=PRP 18
7. w2=$Feature_NN 203 ==> w3=NN 32
8. w1=$Opinion_JJ 197 ==> w2=$Feature_NNS 18
9. w1=$Feature_NN 230 ==> w2=IN 29
10. w1=$Feature_NN 230 ==> w3=DT 38

11. w1=IN 108 ==> w3=$Feature_NN 34
12. w1=NN 117 ==> w3=$Feature_NN 30

6.7 Performance Measures

It is important that how to measure the performance for an information system. In this section, some of the common measures that have been used in the literature are described. To evaluate experimental results, several standard measures such as precision and recall are used. Precision is the proportion of retrieved documents that are relevant to the topic, and the recall is the proportion of relevant documents that were retrieved. Can be defined for a binary classification problem the judgement within a contingency table, as shown in Table 6.1.

	Human Judgment		
System Judgment		*Yes*	*No*
	Yes	TP	FN
	No	FP	TN

Table 6.1 Confusion Metrics

According to the definition in this table, the precision and recall are denoted by the following formulas.

Where *TP* (True Positive) is the number the system correctly identifies as positives; *FP* (False Positive) is the number the system falsely identifies as positives; FN (False Negative) is the number the system fails to identify; TN (True Negative).

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

The above formula will calculate the recall of the system and to find the precision of the system we have to follow the following equation.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

So we will be calculating the remaining values of the confusion metrics for our calculations. On the basis of calculated recall and precision we have listed all the results while running feature extraction on each datasets in the table 6.2.

6.8 Experimental Results

We designed three experiments. The experiments are frequent feature extraction, Infrequent feature extraction and opinion sentence extraction respectively. We discuss each experiment respectively in the bellow section.

6.8.1 Feature Extraction Results

Product Name	No. of manual features	Frequent Features			Infrequent Features	
		Recall	Precision	Accuracy	Recall	Precision
Digital camera 1	45	0.9333	0.9334	90.31%	0.667	0.667
Digital camera 2	48	0.9579	0.9579	85.57%	0.75	0.75
Cellular phone	50	0.9856	0.986	85.56%	0.6	0.6
Mp3 player	23	0.9494	0.9494	89.97%	0.6667	0.6667
DVD player	44	0.9531	0.9540	84.62%	0.8	0.8
Average	**42**	**0.95866**	**0.95614**	**87.19%**	**0.69668**	**0.69674**

Table 6.2 Recall and Precision of frequent and infrequent feature generation by Sequential Pattern Mining

Table 6.2 gives all the precision and recall results which we have evaluated at each step, which uses sequential pattern Mining. Table 6.3 gives all the precision and recall results which we have evaluated at each step, which uses association mining. Column 1 lists each product. Column 2 lists the features which are calculated manually from each dataset. Column 3 and 4 is the recall and precision evaluated during the frequent feature extraction phase. Column 5 and 6 highlights the recall and precision evaluated during the infrequent feature extraction. We can examine from the above table that recall and precision decreased during the infrequent feature extraction. But that is not a major problem as number of infrequent features is very low and makes a very little effect on the results. We can examine from these two tables that precision and recall of frequent features

by using GSP is little high than Aproiri but the overall accuracy of the system with GSP is marginally higher than system with Aproiri.

Product Name	No. of manual features	Frequent features			Infrequent features	
		Recall	Precision	Accuracy	Recall	Precision
Digital camera 1	45	0.95633	0.9536	84.63%	0.667	0.667
Digital camera 2	48	0.9675	0.9676	83.13%	0.75	0.75
Cellular phone	50	0.9471	0.9471	81.03%	0.4	0.4
Mp3 player	23	0.9530	0.954	81.11%	0.5	0.5
DVD player	44	0.959	0.9590	81.15%	0.6	0.6
Average	**42**	**0.9565**	**0.9566**	**87.19%**	**0.5834**	**0.5834**

Table 6.3 Recall and Precision of frequent and infrequent feature generation by Association Rule Mining

We compared the generated features by our method to further illustrate the effectiveness of our feature extraction step[31].

Table 6.4 shows the recall and precision of FBS.

Product name	No. of manual Features	Frequent features (association mining)		Compactness pruning		P-support pruning		Infrequent feature identification	
		Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Digital camera 1	79	0.671	0.552	0.658	0.634	0.658	0.825	0.822	0.747
Digital camera 2	96	0.594	0.594	0.594	0.679	0.594	0.781	0.792	0.710
Cellular phone	67	0.731	0.563	0.716	0.676	0.716	0.828	0.761	0.718
Mp3 player	57	0.652	0.573	0.652	0.683	0.652	0.754	0.818	0.692
DVD player	49	0.754	0.531	0.754	0.634	0.754	0.765	0.797	0.743
Average	**69**	**0.68**	**0.56**	**0.67**	**0.66**	**0.67**	**0.79**	**0.80**	**0.72**

Table 6.4 Recall and Precision at each step of the system [Hu & Liu 2004]

The above table shows the results at each step including the pruning steps. As compared with our results we can easily evaluate that our system produces high recall and precision of frequent feature The overall accuracy of the system is higher than evaluated by [31]. The above results in table 6.2 clearly show that we

have extracted the features without association mining and also we have reduced the cost of both compactness and redundancy pruning. We have used the same datasets to compare our results. We have concluded the result that our system much faster and efficient than the previous system proposed for the extraction of features and produces same results with doing much less effort.

6.8.2 Opinion Extraction Results

We have implemented our approach on all the datasets we have used for the frequent and infrequent feature extraction. Table 6.5 shows the evaluation results of Opinion Words feature generation by Sequential Pattern Mining and Table 6.6 shows the evaluation results of Opinion Words feature generation by Association Rule Mining.

The average recall and precision of opinion sentence extraction is 85% with accuracy rate of 97% when applying GSP. The average recall and precision of opinion sentence extraction is 70% with accuracy rate of 95% when applying Aproiri. From these results we can easily examine that the evaluation results of the system uses GSP is higher than Aproiri. So our approach using the sequential pattern mining to identify sequential patterns does produce interesting results.

Product Name	No. of manual Opinion Words	Opinion Sentence Extraction		
		Recall	Precision	Accuracy
Digital camera 1	61	0.8529	0.8529	97.27%
Digital camera 2	53	0.8292	0.8292	97.63%
Cellular phone	51	0.776	0.7776	98.48%
Mp3 player	46	0.9025	0.9058	97.47%
DVD player	42	0.8933	0.8933	97.11%
Average		**0.8508**	**0.8518**	**97.52**

Table 6.5 Recall and precision of Opinion Words feature generation by Sequential Pattern Mining

Product Name	No. of manual Opinion Words	Opinion Sentence Extraction		
		Recall	Precision	Accuracy
Digital camera 1	61	0.7142	0.7143	96.80%
Digital camera 2	53	0.72	0.72	94.99%
Cellular phone	51	0.6363	0.6364	97.03%
Mp3 player	46	0.625	0.625	94.17%
DVD player	42	0.80	0.80	92.58%
Average		**0.6991**	**0.6994**	**95.10%**

Table 6.6 Recall and precision of Opinion Words feature generation by Association Rule Mining

To further illustrate the effectiveness of our opinion extraction step, we compared the our method with terms found by the well-known and publicly available term extraction and indexing system, FBS(Feature-Based Summarization), [31] Table 6.7 shows the recall and precision of FBS.

Product Name	FBS	
	Recall	Precision
Digital camera 1	0.719	0.643
Digital camera 2	0.634	0.554
Cellular phone	0.675	0.815
Mp3 player	0.784	0.589
DVD player	0.653	0.607
Average	**0.693**	**0.642**

Table 6.7 Results of opinion sentence extraction of FBS

Table 6.8 compares the evaluation results of FBS and GSP. The results shows average increase in the recall and precision is 0.16 % and 0.2 % with GSP.

Product Name	FBS		GSP	
	Recall	Precision	Recall	Precision
Digital camera 1	0.719	0.643	+0.1339	+0.2099
Digital camera 2	0.634	0.554	+0.188	+0.2452
Cellular phone	0.675	0.815	+0.101	-0.0374
Mp3 player	0.784	0.589	+0.1185	+0.3168
DVD player	0.653	0.607	+0.2403	+0.2863
Average	**0.693**	**0.642**	**+0.157**	**+0.2098**

Table 6.8 Comparing the results of FBS and GSP

Table 6.9 compares the evaluation results of FBS and Apriori. The results shows average increase in the recall and precision e.g. 0.61 % and 5.74 % with Apriori.

Product Name	FBS		Apriori	
	Recall	Precision	Recall	Precision
Digital camera 1	0.719	0.643	-0.048	+0.0713
Digital camera 2	0.634	0.554	-0.434	+0.166
Cellular phone	0.675	0.815	-0.0384	-0.1786
Mp3 player	0.784	0.589	-0.115	+0.036
DVD player	0.653	0.607	+0.187	+0.193
Average	**0.693**	**0.642**	**+0.0061**	**+0.0574**

Table 6.9 Comparing the results of FBS and Aproiri

In summary, we can conclude our techniques produce more accurate results and are of very promising especially for prediction of sentence orientation. . We believe they can be used in practical contexts.

6.9 Orientation of Opinion Sentence

Prediction of the orientation of opinion sentence is to identify complete sentence as positive or negative. Orientation of opinion sentence is not the scope of our research but we can use the same technique as used by [35]. They have given three different kinds of sentences.

- The opinion words are either positive or negative i.e. two positive and three negative. It means that sentence has negative orientation. The sentence will be positive in the opposite condition.
- The number of positive and negative opinion words are equal i.e. sentence has one negative opinion and one positive opinion.
- All other cases.

For the first case the orientation is simple. For the second case they have used the average orientation of the opinion words. For the third case they have used the knowledge of previous sentence and examine the orientation of previous opinion sentences.

6.10 Summary Generation

The final task of feature based opinion mining is to generate the summary for the whole document. In the summary generation all the features which have the same feature word are collected together and their total occurrences in the document are collect and on the bases of collect information positive and negative ranks are given to each feature. [35] have given this idea. In their approach they describe the summary generation in two steps.

- In the first step each feature and its opinion is put into a positive and negative category and for each positive or negative one a count is computed i.e. how many positive or negative times the feature has occurred.
- All features are ranked according to their appearances in the review. The features consisting of single words are placed before the feature phrases as user has more interest in single words feature.

Following is an example of showing the summary.

Feature: **Player**

Positive: 5

- Player works and looks great – if you can get the DVD's to play.
- This is the best DVD player I 've purchased.

...

Negative: 10

- This player is not worth any price and I recommend that you don't purchase it.
- The DVD player just wouldn't recognize them.

...

CHAPTER 7

7. Conclusion

> When I examine myself and my methods of thought, I come to the conclusion that the gift of fantasy has meant more to me than any talent for abstract, positive thinking.
>
> – Albert Einstein

7.1 Conclusion

In conclusion, Opinion Mining in Web 2.0 is very important, and trend of buying and expressing their opinions on the web is increasing dramatically day by day. Collecting meaningful information from this user generated content became an important task. Before anyone purchases a product online, they usually surf the web to find useful advice and take a final decision. However, it is difficult to read all related reviews. So, how exactly the nontrivial features from review and associated opinions is a topical issue in recent research.

The proposed system is a supervised information extraction system, extracts the fine features, and identify the associated reviews of online reviews of product characteristics with improved accuracy and compared recall with [36] Hu's work. The aim is to provide a feature-based summary of a large number of customer reviews of a product sold online.

The main purpose of this research is to compare the association rule mining and sequential pattern mining. We used the algorithm *GSP* to find the interesting patterns from the database. Based on these sequential patterns we extract the frequent features and opinion words. We also perform the same procedure with association rule mining by using algorithm *Apriori*. In our experimental results, it proves that sequential pattern mining produce better results than the association rule mining by proposed approach.

Our system consists of three components including review extractor, feature generator and sentiment identifier. Finally, we run a series of experiment to evaluate the performance between Hu's work and our work. The experiment results demonstrate our work has high precision and recall than Hu's work. Besides, the feature appraisal system can be extended to shopping web, music forum and news web. In this research, we have a number of techniques for mining features and opinion ratings based on data mining and methods of natural language processing.

Although opinion mining is to predict the orientation of the document, but the core object of feature based opinion mining is to extract the entire product feature and their opinion words which modify any product feature. While there are other steps involved to predict the document orientation like finding the polarity of each opinion word, predicting the sentiment of each sentence and producing a complete summary for the document.

Our scope of research is to extract product features and their opinion words and to extract the opinion sentences. The rest of the steps we are leaving for the future work. With the best of our knowledge of our proposed approach is new in this area for extraction of product features.

The another great valuable outcome of this research is that we have reduced the cost of extracting frequent features used by Hu[].

We have extracted almost the same number of features without the use of the following steps.

- Generating frequent features using CBA which is based on Apriori algorithm.
- No need to use compactness pruning
- No need to use redundancy pruning

By reducing these steps the cost of extracting frequent features is quite low. We have searched and investigated many works about this subject and we believe that

using supervised approach might create more accurate results for sentiment analysis.

7.2 Possible Future Work

In this dissertation, we have concentrated on the research of Web content mining for Web product reviews via opinion mining paradigms. The theoretical and experimental studies have shown the effectiveness and applicability of the proposed model and approach.

The future work can be continued along the following directions:

7.2.1 Focusing on Other Review Formats

In our research we are using the free format of customers' reviews. We processed this data and produced results. Therefore another important direction for research is to use other two formats of opinions which are pros and cons format and mix-format and analyze the result with these formats.

7.2.2 Implementing on Other Levels of Mining

Apply the proposed approach on the other levels i.e., document level and sentence level.

7.2.3 Comparison of Results with Other Techniques

Our focus will also be on the comparison of our technique's results with the other techniques' results. A complete comparison will help us out for checking that where there is a need to improve the results and how much the results are satisfactory.

7.2.4 Application to other Domains

Extending the scope of current research in other related areas. The proposed approach of opinion mining thus has a tremendous scope for practical applications.

References

" The person who can combine frames of reference and draw connections between ostensibly unrelated points of view is likely to be the one who makes the creative breakthrough."

– Denise Shekerjian

[1] Agrawal, R. and Srikant, R. 1994, Fast Algorithms for Mining Association Rules, *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94).*

[2] Agrawal, R. and Srikant, R. 1995, Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering (ICDE'95).

[3] Albert-Lorincz, H. and Boulicaut, J.-F. (2003a), A framework for frequent sequence mining under generalized regular expression constraints, in J.-F. oulicaut and S. Dzeroski, eds, 'Proceedings of the Second International Workshop on Inductive Databases KDID', RudjerBoskovic Institute, Zagreb, Croatia, Cavtat-Dubrovnik, Croatia, pp. 2–16.

[4] Albert-Lorincz, H. and Boulicaut, J.-F. (2003b), Mining frequent sequential patterns under regular expressions: A highly adaptive strategy for pushing contraints, in D. Barbar'a and C. Kamath, eds, 'Proceedings of the Third SIAM International Conference on Data Mining', SIAM, San Francisco, CA.

[5] Ana-Maria Popescu , Oren Etzioni, Extracting Product Features and Opinions from Reviews, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, p.339-34

[6] Ayres, J., Flannick, J., Gehrke, J. and Yiu, T. (2002), Sequential pattern mining using a bitmap representation, in '8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, Edmonton, Alberta, Canada, pp. 429–435.

[7] B. Liu, Web data mining; Exploring hyperlinks, contents, and usage data, Opinion Mining. Springer, 2006.

[8] B. Pang and L. Lee. 2008, Opinion Mining and Sentiment Analysis, Foundations and Trendsin Information ReterievalVol. 2, Nos. 1-2 (2008) 1-135.

References

" The person who can combine frames of reference and draw connections between ostensibly unrelated points of view is likely to be the one who makes the creative breakthrough."

– Denise Shekerjian

[1] Agrawal, R. and Srikant, R. 1994, Fast Algorithms for Mining Association Rules, *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94).*

[2] Agrawal, R. and Srikant, R. 1995, Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering (ICDE'95).

[3] Albert-Lorincz, H. and Boulicaut, J.-F. (2003a), A framework for frequent sequence mining under generalized regular expression constraints, in J.-F. oulicaut and S. Dzeroski, eds, 'Proceedings of the Second International Workshop on Inductive Databases KDID', RudjerBoskovic Institute, Zagreb, Croatia, Cavtat-Dubrovnik, Croatia, pp. 2–16.

[4] Albert-Lorincz, H. and Boulicaut, J.-F. (2003b), Mining frequent sequential patterns under regular expressions: A highly adaptive strategy for pushing contraints, in D. Barbar´a and C. Kamath, eds, 'Proceedings of the Third SIAM International Conference on Data Mining', SIAM, San Francisco, CA.

[5] Ana-Maria Popescu , Oren Etzioni, Extracting Product Features and Opinions from Reviews, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, p.339-34

[6] Ayres, J., Flannick, J., Gehrke, J. and Yiu, T. (2002), Sequential pattern mining using a bitmap representation, in '8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, Edmonton, Alberta, Canada, pp. 429–435.

[7] B. Liu, Web data mining; Exploring hyperlinks, contents, and usage data, Opinion Mining. Springer, 2006.

[8] B. Pang and L. Lee. 2008, Opinion Mining and Sentiment Analysis, Foundations and Trendsin Information RetierevalVol. 2, Nos. 1-2 (2008) 1-135.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.

[10] Baroni, M. and Vegnaduzzo, S. 2004, Identifying subjective adjectives through web-based mutual information. In E. Buchberger, editor, In Proceedings of the Conference for the Processing of Natural Language and

[11] C. Cardie, J. Wiebe, T. Wilson, and D. Litman, "Combining low-level and summary representations of opinions for multi-perspective question answering,"inProceedings of the AAAI Spring Symposium on New Directions

[12] Camelin, N., Damnati, G., Béchet, F. and De Mori, R. 2006, Opinion Mining in a Telephone Survey Corpus, International Conference on Spoken Language Processing (Interspeech 2006).

[13] D. K. Evans, L.-W. Ku, Y. Seki, H.-H. Chen, and N. Kando, "Opinion analysis across languages: An overview of and observations from the NTCIR6 opinion analysis ilot task," in Proceedings of the Workshop on Cross-Language Information Processing, vol. 4578 (Applications of Fuzzy Sets Theory) of Lecture
Notes in Computer Science, pp. 456–463, 2007.

[14] D. Pierce, E. Riloff, T. Wilson, D. Day, and M. Maybury, "Recognizing anorganizing opinions expressed in the world press," in Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, 20

[15] Dave, K.., Lawrence, S. and Pennock, D. 2003, Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews, Proceedings of International World Wide Web Conference (WWW'03).

[16] Dey, L. 2007, Web and Text Mining for opinion/trend analysis, IEEE Conference of WI and IAT, (IEEE'07).

[17] Esuli, A. and Sebastiani, F. 2005, Determining the semantic orientation of terms through gloss analysis. In Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management, Bremen, DE.

[18] Esuli, A. and Sebastiani, F. 2006, SentiWordNet: A publicly Available Lexical Resource for Opinion Mining, Proceedings of the Fifth International Conference on Language Resource and Evaluation (LREC'06).

[19] Fellnaum, C. 1998, WordNet: an Electronic Lexical Database, MIT press'98.

[20] Freespan: frequent pattern-projected sequential pattern mining, in '6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, Boston, MA, USA, pp. 355–359.

[21] Ganapathibhotla, M. and Liu, B. 2008, Mining Opinions in Comparative Sentences, Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08).

[22] Garofalakis. M.. Rastogi, R., and Shim, K. (1999). SPIRIT: sequential pattern mining with regular expression constraints. Proc. of Int'l Conference on Very Lnrge dam base.(VLDB). pp. 223-234.

[23] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences,"in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2003.

[24] Han, J. and Pei, J. (2000), 'Mining frequent patterns by pattern growth: Methodology and implications', SIGKDD Explorations Newsletter 2(2), 14–20.

[25] Han, J., Pei, J., Mortazavi-Asl, B., Chen, Q., Dayal, U. and Hsu, M.-C. (2000),Freespan: frequent pattern-projected sequential pattern mining, in '6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM Press, Boston, MA, USA, pp. 355–359.

[26] Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F. and Poncelet, P. 2008, Web Opinion Mining: How to extract opinions from blogs?,International Conference on Soft Computing as Transdisciplinary Science an

[27] Hatzivassiloglou ,V. and McKeown, R. 1997, Predicting the semantic orientation of adjectives, Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, p.174-181

[28] Hiroshi, K., Tetsuya, N., and Hideo, W. 2004. Deeper sentiment analysis using machine translation technology. InProceedings of the 20th international Conference on Computational Linguistics (Geneva, Switzerland

[29] http://www.worsleyschool.net/socialarts/factopinion/factopinion.html - Last Accessed on Feburaray, 10, 2010

[30] Hu, M. and Liu, B. 2004, Mining and Summarizing Customer Reviews, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04).

[31] Hu, M. and Liu, B. 2004, Mining Opinion Features in Customer Reviews, Proceedings of Ninth International Conference on Artificial Intelligence (AAAI'04).

[32] Hulth, A. and Megyesi,B.B. 2006, A Study on Automatically Extracted Keywords in Text Categorization, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the

[33] Hulth, A. Improved automatic keyword extraction given more linguistic knowledge. In proceedings of EMNLP'2003.in A. Jorge, L. Torgo, P. Brazdil, R. Camacho and J. a. Gama, eds, 'PKDD in Computer Science, Spring

[34] I. Elgort, E-learning adoption: Bridging the chasm, presented at ascilite 2005 conference proceedings, 2005.

[35] Identifying subjective adjectives through web-based mutual information. In E. Buchberger, editor, In Proceedings of the Conference for the Processing of Natural Language and Speech KONVENS), pages 17--24, Vienn

[36] J. M. Wiebe, Tracking point of view in narrative,Computational Linguistics,vol. 20, pp. 233–287, 1994.

[37] J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, T. Wilson, D. Day, and M. Maybury, "Recognizing and organizing opinions expressed in the world press," in Proceedings of the AAAI Spring Symposium on New Directions in Question Answering, 2003.

[38] J. Zabin and A. Jefferies, Social media monitoring and analysis: Generating consumer insights from online conversation, Aberdeen Group Benchmark Report, January 2008.

[39] Jin, W. and Ho, H.H. 2009, A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining, In Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009.

[40] Jindal, N. and Liu, B. 2006, Mining Comparative Sentences and Relations, Proceedings of National Conference on Artificial intelligence (AAAI'06).

[41] Jindal, N. and Liu, B. 2007, Review Spam Detection, Proceedings of the 16th International Conference on World Wide Web (WWW'07) (Poster Paper).

[42] Jindal, N. and Liu, B. 2008, Opinion Spam and Analysis, Proceedings of the First ACM Conference on Web Search and Data Mining (WSDM'08).

[43] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of WWW, pp. 519–528, 2003.

[44] K. Tateishi, Y. Ishiguro, and T. Fukushima, "Opinion information retrieval from the internet," Information Processing Society of Japan (IPSJ) SIG Notes, 2001, vol. 69, no. 7, pp. 75–82, 2001. (Also cited as "A reputation search engine that gathers people's opinions from the Internet", IPSJ Technical Report NL-14411. In Japanese).

[45] Kathleen T. Durant, Michael D. Smith: Predicting the Political Sentiment of Web Log Posts Using Supervised Machine Learning Techniques Coupled with Feature Selection. WEBKDD 2006: 187-206

[46] Kim, S. and Hovy, H. 2004, Determining the Sentiment of Opinions, Proceedings of the 20th International Conference on Computational Linguistics (COLING'04).

[47] Kim, Y. and Myaeng, S. 2007, Opinion Analysis based on Lexical Clues and their Expansion, Proceedings of NII Test Collection for Information Retrieval (NTCIR-6'07).

[48] L. B. Hu M, Mining and summarizing customer reviews., presented at Proceedings of the tenth ACM SIGKDD international conference onKnowledge discovery and data mining 22:2525, 2004.

[49] L. Dini and G. Mazzini, "Opinion classification through information extraction," in Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining),

[50] Lin,W.H., Wilson,T., Wiebe,J. and Hauptmann, A. 2006, Which Side are You on? Identifying Perspectives at the Document and Sentence Levels, Proceedings of the 10th Conference on Computational Natural Language Le

[51] Liu, B. 2007, Web Data Mining, Springer.

[52] Liu, B., Hsu, W. and Ma, Y. 1998, Integrating Classification and Association Rule Mining, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98).

[53] Liu, B., Hu, M. and Cheng, J. 2005, Opinion Observer: Analyzing and Comparing Opinions on the Web, Proceedings of the 14th International Conference on World Wide Web (WWW'05).

[54] Luo, C. and Chung, S. M. (2004), A scalable algorithm for mining maximal frequent sequences using sampling, in '16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004)', IEEE Compu

[55] M.S. BinWahlan, Salim,N and L. Suanmali, (2009), Swarm Based Features Selection for Text Summarization, International Journal of Computer Science and Network Security", 9(1): pp. 175-179.

[56] Mário J. Silva, Paula Carvalho, Luís Sarmento, Eugénio Oliveira, Pedro Magalhães, The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics.New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence p. 565-576, October, 2009. Universidade de Aveiro.

[57] Masseglia, F., Cathala, F. and Poncelet, P. (1998), The PSP approach for mining sequential patterns, in '2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98)', Vol. 1510 of LNAI, Springer Verlag, antes, France, pp. 176–184.

[58] Masseglia, F., Poncelet, P. and Teisseire, M. (2000), Incremental mining of sequential patterns in large databases, VLDB'99', Edinburgh, Scotland, pp. 223–234.

[59] Masseglia, F., Teisseire, M. and Poncelet, P. (2005), Sequential pattern mining: A survey on issues and approaches, in 'Encyclopedia of Data Warehousing and Mining', Information Science Publishing.

[60] Mehra, N., Khandelwal, S. and Patel, P. 2002, Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews, Stanford University, USA.

[61] Miller, S., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. 1990, Introduction to WordNet: An on-line Lexical Database, International Journal of Lexicography (special issue), 3(4):235-312.

[62] Min, KH, Wilson, WH, Kang, BH :Effectiveness of methods for syntactic and semantic recognition of numeral strings, tradeoffs between number of features and length of word N-grams, University of Tasmania 2007.

[63] Mining Sequential Patterns, Research report TTE1-2001-10.

[64] Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews Kushal Dave, Steve Lawrence, David M. Pennock]Table 5.

[65] Morinaga, S., Yamanishi, K., Tateishi, K., and Fukushima, T. Mining product reputations on the web.

[66] Narayan, R., Liu, B. and Choudhary, A .2009, Sentiment Analysis of Conditional Sentences, Proceedings of Conference on Empirical Methods in Natural language processing (EMNLP'09).

[67] Nasraoui O., Spiliopoulou M., Srivastava J., Mobasher B., Masand B.,WebKDD 2006: web mining and web usage analysis post-workshop reportACM SIGKDD Explorations Newsletter 8(2): 84-89, 2006.

[68] Orlando, S., Perego, R. and Silvestri, C. (2004), A new algorithm for gap constrained sequence mining, in 'SAC Proceedings of the 2004 ACM Symposium on plied Computing (SAC)', ACM Press, Nicosia, Cyprus, pp. 540–547

[69] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association", ACM Trans. Inf. Syst., vol. 21(4), pp. 315–346, 2003.

[70] P. Mika. Microsearch: An Interface for Semantic Search. In Proc. of the Workshop on Semantic Search (SemSearch 2008) at the 5th European Semantic Web Conference (ESWC 2008) , Tenerife, Spain, volume 334 of CEUR

[71] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424, 2002.

[72] Pang, B. and Lee, L. 2004, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts,Proceedings of the Association for Computational Linguistics (ACL'04).

[73] Pang, B. and Lee, L. 2008, Opinion Mining and Sentiment Analysis,Foundation and Trends in Information Retrieval, vol. 2, Nos. 1-2 (2008) 1-135.

[74] Pang, B., Lee, L. and Vaithyanathan, S. 2002, Thumbs up? Sentiment Classification Using Machine Learning Techniques,Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP'02)

[75] Pinto, H., Han, J., Pei, J., Wang, K., Chen, Q., &Dayal, U. (2001). Multi-Dimensional Sequential Pattern Mining.Proceedings of the 10th International Conference on Information and Knowledge Management, Atlanta, USA, 81-88.

[76] Popescu, A.-M. andEtzioni, O. 2005, Extracting Product Features and Opinions from Reviews, Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP'05).

[77] Porter, M.F. 1980, An Algorithm for Suffix Stripping, Originally Published Program, 14 no. 3, pp 130-137.

[78] Qi Sul, Xinying Xul, Honglei Guo2, Zhili Guo2, Xian Wu2, Xiaoxun Zhang2, Bin Swen1and Zhong Su2:Hidden Sentiment Association in Chinese Web

[79] Qiang Ye, Ziqiong Zhang, and Rob Law. 2008, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches..Expert Systems with Applications (2008) doi:10.1016/j.esw

[80] R. M. Tong, "An operational system for detecting and tracking opinions in on-line discussion," in Proceedings of the Workshop on Operational Text Classification (OTC), 2001.

[81] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, A comprehensive grammar of the English language. Longman, 1985.

[82] Rillof,E., Wiebe, J., and Philips, W. 2005, Exploiting Subjectivity Classification to Improve Information Extraction, Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05).

[83] Riloff, E. and Wiebe, J. Learning Extraction Patterns for Subjective Expressions ,Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing(EMNLP'03).

[84] S. Das and M. Chen, "Yahoo! for Amazon: Extracting market sentiment from stock message boards," in Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 2001.

[85] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the Web," in Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pp. 341–349, 2002.

[86] Savary, L. and Zeitouni, K. (2005), Indexed bit map (ibm) for mining frequent sequences, in A. Jorge, L. Torgo, P. Brazdil, R. Camacho and J. a. Gama, eds, 'PKDD Knowledge Discovery in Databases: PKDD 2005, 9th European Conference on Principles and Practice of Knowledge Discovery in Databases', Vol. 3721 of Lecture Notes in Computer Science, Springer, Porto, Portugal, pp. 659–666.

[87] Scaffidi, C., Bierhoff, K., Chang, E., Felker, M., Ng, H., and Jin, C. 2007. Red Opal: product-feature scoring from reviews. In Proceedings of the 8th ACM Conference on Electronic Commerce (San Diego, Californi

[88] Seno, M. and Karypis, G. (2002), Slpminer. An algorithm for finding frequent sequential patterns using length-decreasing support, Technical Report #02-023, niversity of Minnesota.

[89] Sequence Data Mining (Springer, 2007)

[90] Sequential pattern mining : A survey by QiankunZho, Sourav S. Bhowmick, Nanyang Technological University Singapore.2nd SIAM International Conference on Data Mining (SDM'02)', SIAM, Arlington,2001).2002.

[91] Srikant, R. and Agrawal, R. (1996), Mining sequential patterns: Generalizations and performance improvements, in P. M. G. Apers, M. Bouzeghoub and G. Gardarin, eds, '5th International Conference on Extending Database Technology, (EDBT'96)', Vol. 1057 of LNCS, Springer, Avignon, France, pp. 3–17.

[92] Su, Q., XU, X., Guo, H., Guo, Z., WU, X., Zhang, X., Swen, B., and Su, Z. 2008, Hidden Sentiment Association in Chinese Web Opinion Mining, International World Wide Web Conference Committee (IW3C2'08).

[93] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the Conference on Knowledge Capture (K-CAP), 2003.

[94] Technorati, Inc. http://technorati.com; Last Accessed at 20.05.2009

[95] Turney, P. 2002, Thumbs Up or Thumbs Down? Sentiment Orientation Applied to Unsupervised Classification of Reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'0

[96] under regular expressions: A highly adaptive strategy for pushing contraints, in update of frequent sequences. In PAKDD. 186{197.VA, USA, pp. 457–473.

[97] Very Large Databases, VLDB'99', Edinburgh, Scotland, pp. 223–234.

[98] Viktor Pekar and ShiyanOuHLSS: Discovery of Subjective Evaluations of Product Features in Hotel Reviews, University of Wolverhampton, United Kingdom [58-Thesis].

[99] Wang W., Yang J. Mining Sequential Patterns from Large Data Sets.

[100] Web Log Miner

[101] Whitelaw, C., Grag, N. and Argamon, S. 2005, Using Appraisal Groups for Sentiment Analysis, Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM'05).

[102] Wilson, T., Wiebe, J. and Hoffmann, P. 2005, Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis, Proceeding of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP'05).

[103] Wilson, T., Wiebe, J. and Hwa, R. 2004, Just how mad are you? Finding strong and weak opinion clauses, Proceeding of National Conference on Artificial Intelligence (AAAI'04).

[104] Xiaowen Ding , Bing Liu , Philip S. Yu, A holistic lexicon-based approach to opinion mining, Proceedings of the international conference on Web search and web data mining, February 11-12, 2008, Palo Alto, Calif

[105] Yang, J., Wang, W., and Yu, P. S. 2001. Infominer: mining surprising periodic patterns. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discoveryand data mining. ACM Press, 395.

[106] Yang, J., Wang, W., Yu, P. S. and Han, J. (2002), Mining long sequential patterns in a noisy environment, in 'ACM SIGMOD International Conference on Management of Data (SIGMOD'02)'

[107] Yang, Z. and Kitsuregawa, M. (2005), Lapin-spam: An improved algorithm for mining sequential pattern, in 'ICDEW '05: Proceedings of the 21st International Conference on Data Engineering Workshops', IEEE Computer Society, Tokyo, Japan, p. 1222.

[108] Yang, Z., Wang, Y. and Kitsuregawa, M. (2005), Lapin: Effective sequential pattern mining algorithms by last position induction, in 'The 21st International onference on Data Engineering (ICDE 2005)', Tokyo, Japan.

[109] Yi, J. and Niblack, W. 2005. Sentiment Mining in WebFountain. In Proceedings of the 21st internationalConference on Data Engineering (Icde'05) - Volume 00 (April 05 - 08, 2005). ICDE. IEEE Computer Society, Washington, DC, 1073-1083

[110] Youngho Kim, Sung-HyonMyaeng,Opinion Analysis based on Lexical Clues and their ExpansionInformation and Communications University 119, Moonji-ro, Yuseong-gu, Daejeon, 305-714, South Korea

[111] Yu,H. and Hatzivassiloglou ,V. 2003,Towards answering opinion questions:Separating facts from opinions and identifying the polarity of opinionsentences, Proceedings of EMNLP-03, 8th Conference on Empirical Methodsin Natural Language Processing(EMNLP'03), pages 129–136.

[112] Zaki, M. J. (1998), Efficient enumeration of frequent sequences, in '7th InternationalConference on Information and Knowledge Management', ACM Press, Bethesda,Maryland, United States, pp. 68–75.

[113] Zaki, M. J. (2000), Sequence mining in categorical domains: Incorporating constraints,inA. Agah, J. Callan and E. Rundensteiner, eds, '9th International Conference on Information and Knowledge Management (CIKM2000)', ACM Press, McLean, VA,USA, pp. 422–429.

[114] Zaki, M. J. (2001a), 'Parallel sequence mining on shared-memory machines', Journal of Parallel and Distributed Computing 61(3), 401–426.

[115] Zaki, M. J. (2001b), 'SPADE: An efficient algorithm for mining frequent sequences',Machine Learning 42(1/2), 31–60.

[116] Zaki, M. J. and Hsiao, C.-J. (2002), CHARM: An efficient algorithm for closed itemset mining, in R. L. Grossman, J. Han, V. Kumar, H. Mannila and R. Motwani, eds,

[117] Zhang, M., Kao, B., Cheung, D. W., and Yip, C. L. 2002. E±cient algorithms for incremental update of frequent sequences. In PAKDD.186{197.

[118] Zhang, M., Kao, B., Yip, C., and Cheung, D. 2001. A gsp-based efficient algorithm for mining frequent sequences. In Proc. 2001International Conference on Artificial Intelligence (IC-AI 2001).

[119] Zheng, Q., Xu, K., and Ma, S. 2002. When to update the sequential patterns of stream data.

[120] Zheng, Q., Xu, K., Ma, S., and Lv, W. The algorithms of updating sequential patterns.

[121] ZhongchaoFei, Jian Liu, and Gengfeng Wu: Sentiment Classification Using Phrase Patterns, Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04).

[122] Zhuang, L., Jing, F. and Zhu, X.Y. Movie Review Mining and Summarization. In proceedings of CIKM'06,pp.43-50.

[123] http://datamining.typepad.com/data_mining/2008/04/opinion-ining.html -Last Accessed on Feburaray, 10, 2010