ระบบการแนะนำภาพยนตร์ที่ใช้ข้อมูลหลายมิติและข้อมูลหลายเกณฑ์แบบถ่วงน้ำหนัก

นางสาวเกียรติมา ฉัพพัณณรังษี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ ภาควิชาคณิตศาสตร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2551

MOVIE RECOMMENDER SYSTEM USING MULTIDIMENSIONAL AND WEIGHTED MULTIPLE CRITERIA DATA

Miss Keittima Chapphannarungsri

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science and Information
Department of Mathematics
Faculty of Science
Chulalongkorn University
Academic Year 2008

Thesis Title	MOVIE RECOMMENDER SYSTEM USING MULTIDIMENSIONAL AND WEIGHTED MULTIPLE CRITERIA DATA
By	Miss Keittima Chapphannarungsri
Field of Study	Computer Science and Information
Advisor	Assistant Professor Saranya Maneeroj, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

.... Dean of the Faculty of Science

(Professor Supot Hannongbua, Ph.D.)

THESIS COMMITTEE

.... Chairman

(Suphakant Phimoltares, Ph.D.)

.... Advisor

(Assistant Professor Saranya Maneeroj, Ph.D.)

.... Examiner

(Professor Chidchanok Lursinsap, Ph.D.)

.... External Examiner

(Jakkarin Suksawatchon, Ph.D.)

เกียรติมา ฉัพพัณณรังษี : ระบบการแนะนำภาพยนตร์ที่ใช้ข้อมูลหลายมิติและข้อมูลหลายเกณฑ์แบบถ่วงน้ำหนัก. (MOVIE RECOMMENDER SYSTEM USING MULTIDIMENSIONAL AND WEIGHTED MULTIPLE CRITERIA DATA) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. ศรันญา มณีโรจน์, 79 หน้า.

ในปัจจุบันนี้ ระบบการแนะนำข้อมูลส่วนใหญ่มีการนำ Content-Based Filtering, Collaborative Filtering, Demographic Filtering และ Hybrid Filtering ซึ่งให้ความสำคัญกับผู้ใช้และไอเท็ม มีงานวิจัยมากมายที่ถูกพัฒนาในส่วนของข้อมูลหลายมิติหรือไม่ก็ข้อมูลหลายเกณฑ์ งานวิจัยนี้ได้สร้างระบบการแนะนำข้อมูลที่เพิ่มคุณภาพของการแนะนำข้อมูลโดยผสมผสานทั้งการนำข้อมูลหลายมิติและการนำก็ข้อมูลหลายเกณฑ์มาใช้ โดยใช้วิธีการดังต่อไปนี้ เปลี่ยนวิธีการถ่วงน้ำหนักของข้อมูลหลายเกณฑ์ให้มีความเหมาะสมมากขึ้น, เพิ่มการถ่วงน้ำหนักโดยใช้ความถี่ในการเลือกหนังของผู้ใช้ และ มีการนำข้อมูลหลายมิติเข้ามาใช้ โดยนำการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ มาใช้ในการวิเคราะห์ข้อมูลทางด้านพฤติกรรมของผู้ใช้ระบบ และ จากการทดลองและวัดผล พบว่า ระบบการแนะนำข้อมูลที่ใช้ข้อมูลหลายมิติและข้อมูลหลายเกณฑ์แบบถ่วงน้ำหนักมีความถูกต้องมากกว่าระบบแนะนำข้อมูลแบบผสมที่ใช้กันอยู่ในปัจจุบัน

ภาควิชา คณิตศาสตร์ ลายมือชื่อนิสิต เกียรติมา ฉัพพัณณรังษี

สาขาวิชาวิทยาการคอมพิวเตอร์และสารสนเทศ ลายมือชื่ออ.ที่ปรึกษาวิทยานิพนธ์หลัก

ปีการศึกษา 2551

507 36016 23 : MAJOR COMPUTER SCIENCE AND INFORMATION

KEYWORDS : RECOMMENDER SYSTEM / CONTENT-BASED FILTERING / COLLABORATIVE FILTERING / MULTIPLE CRITERIA / MULTIDIMENSIONAL

KEITTIMA CHAPPHANNARUNGSRI: MOVIE RECOMMENDER SYSTEM USING MULTIDIMENSIONAL AND WEIGHTED MULTIPLE CRITERIA DATA. ADVISOR: ASST. PROF. SARANYA MANEEROJ, Ph.D., 79 pp.

Most of current Recommender Systems based on Content-Based Filtering, Collaborative Filtering, Demographic Filtering and Hybrid Filtering which are concentrated on user and item entities. Many research papers are improved by pointing out either Multiple Criteria Rating approach or Multidimensional approach for Recommender System. This paper proposes an advanced Recommender System to provide higher quality of recommendations by combining the Multiple Criteria rating and the Multidimensional approaches. For the Multiple Criteria approach, this paper proposed a method that changes the way of weighting to be more suitable and also concern about the frequency of the selection movie features. To do Multidimensional approach, the Multiple Linear Regression is applied to analyze the contextual information of user characteristics. According to the experimental evaluation, the combining of Multiple Criteria Rating and Multidimensional approaches provide more accurate recommendation results than the current Hybrid Recommender Systems.

Department : Mathematics	Student's Signature
Field of Study : Computer Science and Information	Advisor's Signature
Academic Year : 2008	

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor, Assistant Professor Saranya Maneeroj, at The Advanced Virtual and Intelligent Computing (AVIC) Research Center, for all her great support and patience that tremendously helped me accomplish this thesis. She also suggests the solution for solving many problems occurred while doing an experiment. I would also like to thank all my friends, and most importantly, my family for everything that they have supported me.

I dedicate this work to all the people as I mentioned above. Without them, this work will never be done. Throughout this thesis, I had encountered many problems, but they were trivial, compared to all the supports given by these people. The encouragement from them was so great and those impressions will always be remembered.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Information Overload

As the world moves into the globalization, the Internet became widely used as a source of information and very growth over the past few years. Many web sites are presenting varies of contents such as items to sell or purchase, knowledge, stories, etc [1]. Excess amounts of information are being provided, making processing and absorbing tasks very difficult for the individual. The amount of data in this online information space is increasing rapidly than the ability of the individual or online customer to process it.

As the result of the amount of information being produce form the several of people on the internet, the problem of Information Overload is arise. For example, considering on searching movie catalog for buying, looking for one or a few movies but there are large number of movies to choose. The users might spend a lot of time to read the reviews of movies in order to choose which one to watch.

One solution to this information overload problem is the use of recommender systems. Recommender systems are developed with a mission to help users find items of interest among those served by the Web application [2]. Taking the example of adding the new movies' CD in your collection but you has no idea which one is the preferred one. The user can tap on the movie recommender application. You might think of some movies in the Action genre and look over the list of Action movies that you own and decide that you would like something along the line of "The Matrix" or "Bad Boys" which are already in your "like most" of movie lists. Then the recommender system provides you with a list of five new movies to choose. After that, you decide to purchase the CD of movie "2 Fast 2 Furious".

Typically, the example of recommender system are GroupLens [3], MovieLens [4], a movie recommender system created by GroupLens research team, and Ringo[5], a music recommender system. The recommender systems help user to get the interesting information easily and filter out those which are less interesting.

1.2 The world of E-commerce

Recommender systems can be utilized to efficiently provide personalized services in most e-business domain, benefiting both the customer and merchant. Recommender systems help customers by narrowing the selected items to meet their needed. Recommender systems are changing from novelties used by a few E-commerce sites, to be a serious business tools that are re-shaping the world of E-commerce.

Many companies try to provide customers with more options. However, the businesses increase the amount of information that customers must process before they are able to select which items meet their needs. Recommender systems are used by E-commerce sites to suggest products to their customers and provide the useful information to help them decide which products to purchase. At the same time, the hope of shopkeeper is that the recommender systems will increase amount of demand. People might read more, listen to music more, or watch movies more because of the availability of accuracy, dependable and reliable method for them to get more information of things that they may like. All the business will get benefit from the increasing of sales which will normally occur when the customer is presented with more items he would likely find appealing.

1.3 Example of recommender systems

This first example focuses on recommender systems in the book section of Amazon.com. Amazon.com™ (www.amazon.com) is structured with an information page for each book, giving details of the text and purchase information. The Book Matcher feature allows customers to give direct feedback about books they have read. Customers rate books they have read on a 5-point scale from “hated it” to “loved it.” After rating a sample of books, customers get recommendations based on the opinions of other customers. Located on the information page for each book is a list of 1-5 star ratings and written comments provided by customers who have read the book in question and submitted a review. Customers have the option of incorporating these recommendations into their purchase decision.

Second, the Album Advisor feature of CDNOW™ (www.cdnow.com) works in two different modes. In the single album mode customers locate the information page for a given album. The system recommends 10 other albums related to the album in question. In the multiple artist mode customers enter up to three artists. In turn, the system recommends 10 albums related to the artists in question. CDNOW enables customers to set up their own music store, based on albums and artists they like. Customers indicate which albums they own, and which artists are their favorites. Purchases from CDNOW are entered automatically into the “own it” list. Although “own it” ratings are initially treated as an indication of positive likes, customers can go back and distinguish between “own it and like it” and “own it but dislike it.” When customers request recommendations the system will predict 6 albums the customer might like based on what is already owned. A feedback option is available by customers providing a “own it,” “move to wish list” or “not for me” comment for any of the albums in this prediction list. The albums recommended change based on the feedback.

Third, Moviefinder.com’s Match Maker (www.moviefinder.com) allows customers to locate movies with a similar “mood, theme, genre or cast” to a given movie. From the information page of the movie in question, customers click on the Match Maker icon and are provided with the list of recommended movies, as well as links to other films by the original film's director and key actors. They predict recommends movies to customers based on their previously indicated interests. Customers enter a rating on a 5- point scale -- from A to F – for movies they have viewed. These ratings are used in two different ways. Most simply, as they continue, the information page for non-rated movies contains a personalized textual prediction (go see it – forget it). In a variation of this, customers can use Powerfind to search for top picks based on syntactic criteria such as Genre, directors, or actors and choose to have these sorted by their personalized prediction or by the all customer average.

Fourth, similar to Amazon.com's Customers, Reel.com's Movie Matches (www.reel.com) provides recommendations on the information page for each movie. These recommendations consist of "close matches" and/or "creative matches." Each set consists of up to a dozen hyperlinks to the information pages for each of these "matched" films. The hyperlinks are annotated with one sentence descriptions of how the new movie is similar to the original movie in question ("Darker thriller raises similarly disturbing questions…"). The Movie Map feature of Reel.com recommends movies to customers based on syntactic features. Customers enter queries based on Genre, movie types, viewing format and/or prices, and request results be constrained to "sleepers" or "best of this genre." The recommendations are editor's recommendations for movies that fit the specified criteria.

1.4 Recommendation Strategies

There are lots of strategies to build the recommendation models including the use of Bayesians Networks [6], adaptive decision tree [7] and rule-based systems [8]. Furthermore, there are many filtering technique that support for recommending items, including “Information Retrieval”, “Content-Based Filtering (CBF)”, “Collaborative Filtering (CF)” and “Hybrid Filtering”

1.4.1 Information Retrieval (IR)

Traditionally Information Retrieval develops storage, indexing and retrieval technology for textual document. The requirement is usually a short-term interest.

1.4.2 Content-Based Filtering (CBF)

Content-Based Recommender Systems recommend items to a user based upon a description of the items and a profile of the user's interests [9].

1.4.3 Collaborative Filtering (CF)

Collaborative Filtering (CF) is one of the most successful filtering techniques which is widely use in recommendation technologies in nowadays. Collaborative Recommender Systems predict the rating of items for target user by recognizing commonalities between users on the basis of their ratings.

1.4.4 Hybrid Filtering

Hybrid Recommender Systems combine two or more recommendation techniques to gain better performance with fewer of the drawbacks of any individual one [4] such as combine Content-Based into Collaborative Filtering.

Unfortunately, the combination of Content-Based and Collaborative filtering still remain some problems. For example, lots of current Hybrid system uses the rating value towards the items for evaluating user’s preference opinions. The system does not have the capacity of recognizing the two distinct interests that represented in the same rating value. For more accurate recommendations, the interpretation of the user's interest must involve various features, related to the focus items; this is called without distinction of interest problem. It is also missing the weight of features that affect to the user's preference; this is called without weight feature problem.

Moreover, most of current Hybrid systems do not concern about the contextual information of user's characteristics such as where he saw the movie, when the movie was seen, and with whom. This contextual information might affect to the user's preference on selected movies.

1.5 Fundamental Research Points for Recommender System

In the recent research environment, there are two fundamental research points for recommender systems.

The first research point is to improve the quality of recommendations. The customers/users need recommendations they can trust to help them find products/items that they would like. If a customer/user trusts a recommender system, purchases products or selects items, and then finds out that they do not like such recommended products/items, they will be unlikely to use recommender system anymore.

For example, the Thai-Music system [4] is a hybrid recommender system. It shows that the Content-Based user profile does not cover necessary features of user's interest. In that work, they show how to cope with the forming poor neighbor problem in recommending music. To cover features of user's interest, Thai-Music system based on weighted Multiple Criteria. Unfortunately, the weighted of weighted Multiple Criteria in Thai-Music is not proper and Thai-Music do not concern about the contextual information of user's characteristic which affect to the user's preference on selected movies.

Another research point is to improve the scalability of the system (or computation time). The tremendous growth of users (customers in E-commerce sites) and items (products in E-commerce sites) poses many approaches separate the computation into two parts. In the first part, which can be done offline, they build a model, that captures the relationships between users and items, relationships between users and users, or relationships between items and items. The second part of the computation uses the model to compute the recommendations in an online process. Therefore, they can compute recommendations very rapidly.

However, the techniques used to reduce computation time can also diminish the quality of the recommendations. Conversely, this thesis only improves the quality of recommendation unlikely to support the large amount of users.

1.6 The Research Objectives

Many effective recommendation strategies have been proposed. However, it is required that an accurate usable filtering techniques should be developed for individual user in recommendation case.

This research focuses to create an advanced movie recommender system to enhance the quality and accuracy of recommender systems by:

1. To solve the *unsuitable weight feature problem* by weighting all the component of each feature.
2. To concerns about the frequency of the selection movie features.
3. To incorporate contextual information as Multidimensional that affect to the user preference on each item.

1.7 Scope of Work

The major proposed of this method is movie recommender system which based on Multidimensional and Weighted Multiple Criteria.

1. To do Multiple Criteria

- Weighting all components of each feature instead of weighting only the biggest component of each feature.

- Add more weight by using the frequency of selection.

2. To do Multidimensional user profile

- Using the contextual information of user's characteristics which are place, companion, day and time.

1.8 Domain of Recommender System

Recommender system is generally used in various domains such as movie, music, book, restaurant, etc. It is up to the researcher to choose which domain to be recommended. This thesis selected the movie domain to do the recommendation technique. This thesis also changes the domain of Thai-Music System from music to movie for evaluating the proposed method. The Thai-Music System is simulated to evaluate on the same dataset as prototype system which is MoviePlanet System.

For the Content-Based System, the selection of features is significant because Content-Based Recommender Systems recommend items to a user based upon a description of the items (or features) so the appropriate selected features will give more accurate recommendation results.

Moreover, Collaborative System utilizes the collaborative user's opinion in recommending items and do not use any information regarding to the actual content.

Furthermore, the Hybrid System that combines the Content-Based and Collaborative filtering technique which concentrates on both users and items to gain better performance of recommendation results. In order to improve the accuracy of recommendation results, this thesis focus on the Multiple Criteria of user's interest on movie features to find the similar users whose ratings strongly correlate with the target user. This thesis does not keep an eye on what the best feature is. Conversely, this thesis would like to show the way of weight the Multiple Criteria so Thai-Music is simulated to compare with the MoviePlanet on the same dataset. For example, if target user loves to watch the movies that get award so his neighbor would be the person who likes to watch the award's movie. Equally, the persons who don't like to watch awards' movies would be neighbor for each other either.

Unfortunately, the Thai-Music System concentrate on the Multiple Criteria and include weight of features in selecting items but the way of weight is not proper because they weight only the biggest component of each feature. Furthermore, Thai-Music System do not concern about the contextual information of user's characteristic that might affect to the user's preference on selected movies.

1.9 The Research Methodology

In order to achieve the above objectives, the following tasks will be carried out by means of appropriate theoretical work described below:

1. Study concepts of Recommender Systems
2. Define and state the related problems
3. Derive an algorithm to create the proposed method
4. Implement the prototype system to evaluate the proposed method
5. Write the thesis

Below is a time table covered all of the above tasks.

Table 1.1: Research methodology time table

No	Tasks	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	Study concepts of Recommender Systems																		
2	Define and state the related problems																		
3	Derive an algorithm to create the proposed method																		
4	Implement the prototype system to evaluate the proposed method																		
5	Write the thesis																		

1.10 Benefit

This proposed method will provide higher quality of Recommendation results which help user to get the interesting information more accurately.

CHAPTER 2

RECOMMENDER SYSTEM REVIEW

In this chapter, the related work of this research is shown. As mentioned, this research is attempted to increase the quality of the recommendation results. Thus, the review about background of recommender system is shown in Section 2.1 and architecture of recommender system in Section 2.2. The evolution of filtering techniques related to recommender system is provided in Section 2.3. The Thai-Music system is described in Chapter 3. A review about related works of neighborhood algorithms is shown in Chapter 4.

2.1 Background of Recommender System

In daily life, there are many books, research papers, television programs, Internet discussion postings, and web pages are published in the internet which any individual human cannot hope to review, let alone read and understand. To cope with information overload, researchers proposed different approaches to separate the interesting and valuable information from the rest.

Historically, this process was placed in the hand of editors and publishers (people given the responsibility for reviewing many documents and selecting the ones worthy of publication). Newspaper editors select what the articles that their readers want to read. Bookstores decide what books would be carrying. Television only offers a limited number of options. Nowadays, people also often read the opinions of movie, restaurant, and television critics to decide how to spend our finite time and money.

However, professional human reviewers do not solve all problems. Often, individuals' information needs and tastes differ enough to make a small number of editors ineffective. Also, the number of documents in the web, in research libraries, and in archives of discussions has grown as large as to defy systematic editing by individual editors. Accordingly, researchers have developed a wide range of systems that bring the power of computation to the problem of selecting, for each individual, the information he or she considers valuable from the enormous amount of available information.

A Recommender System is any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to get interesting or useful information in a large space of possible options.

Recommender systems have been applied to many domains and different techniques have been proposed. The systems GroupLens [3], and Ringo [5], apply collaborative filtering in the domains of Usenet News and music respectively. The Morse system [10] and Recommendz system [11], apply collaborative filtering in the movie domain. The system Krakatao[12], applies content-based filtering in the domains of online newspapers. Fab[13], and P-Tango[14], apply a hybrid approach of collaborative filtering and content-based filtering to recommend web page and online newspapers respectively.

2.2 Architecture of Recommender System

Regarding to the general architecture, a recommender system usually has: (i) background data, which is the information that the system has before starting the recommendation process, such as movie information in the movie recommender systems; (ii) input data, the information that the users have to enter in order to get recommendations; (iii) an algorithm, that combines background and input data to produce recommendations; (iv) output, the recommendations generated by the system. This process is shown in figure 2.1.

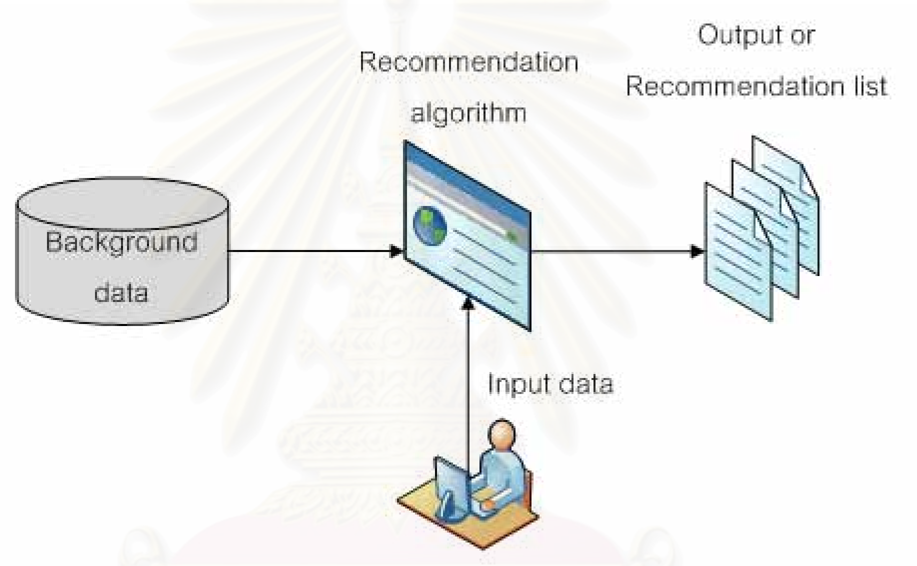


Figure 2.1 Recommendation Process

2.2.1 Input

The input of recommender system depends on the type of the employed filtering algorithm such as user's opinion on movies, contextual information of users, etc.

Rating (also called votes), which expresses the users' opinion on items. Ratings can be explicit or implicit. Explicit ratings are generally a single numeric summary rating for each item [3]. Ratings are normally provided by the user and follow a specified numerical scale (example: 1=unlike, 2=neutral and 3=like), where the higher the number represents the higher the interest. Implicit ratings have the advantage of reducing the user's burden to enter ratings, and are generally extracted from purchase records or browsing behavior, such as clicking on items. Other sources of implicit ratings being explored are the time spent reading [15] and URL references in Usenet postings [16]. Other browsing behavior indicators like mouse, keyboard and scrollbar activities have also been investigated as implicit interest indicators by Claypool [15].

Content Data, which are based on a textual analysis of documents related to the items rated by the user. The features extracted by this analysis are used as input to the filtering algorithm in order to infer a user profile.

2.2.2 Output

The output of a recommender system is a Prediction or Recommendation. A prediction is expressed as a numerical value which represents the anticipated opinion of target user towards the item. The target user refers to a user who is interacting in the system. This predicted value is needed be the same numerical scale (example: 1=unlike, 2=neutral and 3=like), as the input which refer to the opinions, is initially provided by target user. This form of recommender system output is also known as “Individual Scoring”.

A recommendation is expressed as a list of N items, which the target user is expected to like the most. This output is also known as “Top-N Recommendation” or “Ranking Scoring”.

2.3 Information Retrieval (IR)

Information Retrieval techniques allow users to express queries to select documents that match to a topic of interest. A user describes his information requirement in the form of a query to the system and the system attempts to find items that match the query within a database of items. The information requirement is usually a short-term interest; a user issues a query to an IR system to describe an immediate need. That is the IR system provides users with a search interface through with they can query a database of items. For example, when a user asks a movie website for the movie type “Action”, the system returns a list of action movie that may be helpful and may indeed lead the user to a movie that he would like to see.

IR systems may index a database of documents using the full text of the document or only document abstracts. Sophisticated systems rank query results using a variety of heuristics including the relative frequency with which the query terms occur in each document, the adjacency of query terms, and the position of query terms. IR systems also may employ techniques such as term stemming to match words such as “retrieve”, “retrieval” and “retrieving”. IR systems are generally optimized for ephemeral interest queries, such as looking up a topic in the library [17].

Internet search engines and searchable bibliographic databases are results of IR research. In the Internet domain [18], popular IR systems include Alta Vista (www.altavista.digital.com) for web pages, DejaNews (www.dejanews.com) for discussion lists postings, and the Internet Movie Database (www.imdb.com) provides extensive support for IR queries on movies.

2.3.1 The Advantage of Information Retrieval

An IR front-end is useful in a recommender system both mechanism and narrowing the scope of recommender system. For the mechanism, it is for users to identify specific items about which they would like to express an opinion. As the narrowing the scope of recommendation, a movie recommender system allows users to specifically request recommendations for newer movies, for movies released in particular time periods, for particular movie genres such as comedy, and for various combinations of movies.

2.3.2 The Limitations of Information Retrieval

The limitations of IR can be classified into three categories.

The first limitation, IR techniques are less valuable in the actual recommendation process, because they capture only the specific query, but there is no information about user preferences.

The second is that writing such a query requires the user to have a firm sense of what type of document he/she wants. That is it is difficult to find the relevant document when the user does not know appropriate query.

Finally, the IR techniques select items by comparing the content of the items and the query from the user. Therefore, the items must be of some machine parsable formats, or attributes must have been assigned to the items by hand. With current technology, media such as sound, video and some multimedia cannot be analyzed automatically for relevant attribute information, in the manner that text can be analyzed. In addition, it is not practical or possible to parse other items due to limitations of resources. For example, the contents of the Library of Congress may take decades to digitize.

2.4 Content-Based Filtering (CBF) or Information Filtering (IF)

Content-Based Filtering (CBF) or Information Filtering (IF) make recommendations by analyzing the description of the items that have been rated by the user and the description of items to be recommended [19].

A variety of algorithms have been proposed for analyzing the content of text documents and finding regularities in this content that can be serve as the basis for making recommendations. For example, text recommendation system like the news-group filtering system NewsWeeder [20] uses the words of their texts as features. A content-based recommender system learns a profile of the user's interests based on the features present in items, which the user has rated. According to features of items and users preferences, the content-based approach automatically learns and adaptively updates the profile of each user.

The type of user profile derived by a content-based recommender system depends on the learning method employed. Decision tree, neural nets, and vector-based representations have all been used.

Given a user profile, items are recommended for the user based on similarity comparisons between feature weights and those of the user profile. That is, this approach recommends items similar to those given user has liked in the past based on the contents of items [9]. The intuition behind is that if the user liked an item in the past, he tends to like other items with similar content in the future.

For example, if a user buys the "Titanic" DVD collection, the content-based system might recommend other romance drama movies, other movies actor "Leonardo DiCaprio", or other movies directed by "James Cameron".

2.4.1 Content-Based Filtering Systems

Many research projects have been using only content-based filtering to recommend items, including Maes' agents for e-mail and Usenet news filtering [21], Syskill and Webert for recommending web pages[53], NewsWeder for recommending news-group messages[20], InformationFinder for recommending textual document [22], and Lieberman's Letizia [23] employs learning techniques to classify, or recommend documents based on the user's prior actions. Moreover, Cohen's Ripper system has been used to classify e-mail [24]. Boon [25] proposed alternative approaches using other learning techniques and term frequency. The following describes three examples of content-based systems. The basic procedure is described below.

1. User enters profile of ratings (Input Data)

   The system constructs user's profile, a record of the user's tastes based on past history.
2. Learning a profile of user

   The system learning a profile of user from the movies that user has been rate. It makes recommendations by comparing a user profile with the content of each document in the collection and classifies that user's likes/dislikes.
3. Recommendation generation

   The system selects items based on the correlation between the content of the items and the user's preferences.

Pandora system [26] is a popular content-based recommender that uses trained musicians to build a model that is used to recommend music based on content. Pandora system does not scale very well, with an estimated cost of about $10 to analyze each song. As the amount of music generated every day continues to grow, Pandora will find it more and more difficult to keep up. The solution was to automate the process, to use machines instead of humans to analyze the music.

2.4.2 The Advantages of Content-Based Filtering

Content-Based Filtering can be successfully applied to recommend items. The CBF system recommends items based on correlations between the content of the items and the user's preferences. It does not require users to know the appropriate query. Thus, it can reduce the first two limitations of IR technique mentioned above. Moreover, it provides three key advantages that are not provided by Collaborative Filtering: (i) no first-rater problem, (ii) no sparisty rating problem, and (iii) no synonymy problem. The meaning of these three problems is described in the section about "Limitations of Collaborative Filtering".

The CBF technique provide the first advantage (i), because CBF recommends an item to a user if the user profile and the item share the features in common. It does not use opinions of other users.

The second (ii) and third (iii) advantages are provided by the CBF, due to the fact that, recommendations on items are generated by calculating similarity between item features and user feature. It does not use rating values on co-rated items (same rated items).

2.4.3 The Limitations of Content-Based Filtering

While Content-Based filtering techniques have been success, but they suffer certain drawbacks: (i) in some domains, such as movies or music, it cannot successfully analyze the content; (ii) no ability to provide serendipitous recommendations; and analyze the content; (ii) no ability to provide serendipitous recommendations; and (iii) no ability to filter items based on quality and taste.

First, current technology is not able to successfully analyze the content in some domains, movies or audio streams. The CBF selects items for the user's consumption based on correlations between the content of the items and the user's profile of preferences. Therefore, the items must be of some machine parsable formats, or attributes must have been assigned to the items by hand. With current technology, media such as sound, video and some multimedia cannot be analyzed automatically for relevant attribute information, in the manner that text can be analyzed. In addition, it is not practical or possible to parse other items due to limitations of resources. For example, the contents of the Library of Congress may take decades to digitize. Furthermore, reviews of items (such as movies) have been used, but it has the problem of bias of the reviewers and the reviews are not always available in digital format.

Second, it does *not provide much in the way of serendipitous discovery*. Serendipitous discovery means that system will give satisfactory recommendation results which users never think before that they will be interested in. People rely on exploration and have luck to find new items that they did not know they wanted. A person may not know they like watching day time talk shows until they accidentally turn to it. However, if the individual's previous tastes provide no indication of this new penchant, the CBF technique will never select such an item for consumption. Without the capability for exploration, the range of items provided to the user could never expand. This problem is called the "serendipitous discovery" problem.

Third, for the case of using the similarity between Content-Based user's profiles to form neighbors, it has difficultly in updating preference features in database. Unfortunately, the users' preferences are successive changed and these simulate the updating process of the preference feature database. Therefore, implementers choose to implement a small amount of features in the preference features database to avoid difficulty. Thus, this system would select poor neighbors according to the *uncovered preference features*.

For another drawback, the CBF is not able to filter items based on quality and taste. For example, the text analysis techniques are based on word analysis. Thus, they do not consider author's style of writing. In addition, many of the techniques do not consider the structures of the text, such as paragraphs and sections.

2.5 Collaborative Filtering (CF)

Collaborative Filtering system recommends items based on the opinions of other users who have the similar tastes or making recommendation by finding correlations among users in the system. It relies on the fact that people's tastes are not randomly distributed: there are general trends and patterns within the taste of a person and between groups of people, for instance, a person "Mark" loves Sci-Fi books. Therefore, it would be likely that he would be interested in seeing the new "Star Wars" movie.

If people's preferences were random, no such prediction could be made. But in reality, after getting some ideas about a person's likes and dislikes, we can often predict what he/she would like based upon intuition that we have about patterns in people's tastes.

From a real-life example, Jane might also have asked two friends, Helen and Barry, for their recommendations. Helen suggests "Pretty Woman" while Barry suggests "Face Off". From past experience, Jane knows that Helen and she have similar tastes, while Barry and she doesn't always agree with together. She therefore, accepts Helen's suggestions and decides to watch "Pretty Woman". This decision was made through Collaborative Filtering, independent of the content of the movies. Collaborative Filtering essentially automates the process of "word-of-mouth" recommendations. Except that instead of having to ask a couple friends about a few items, a CF system can ask hundreds of other people, and consider hundreds of different items, all happening autonomously automatically.

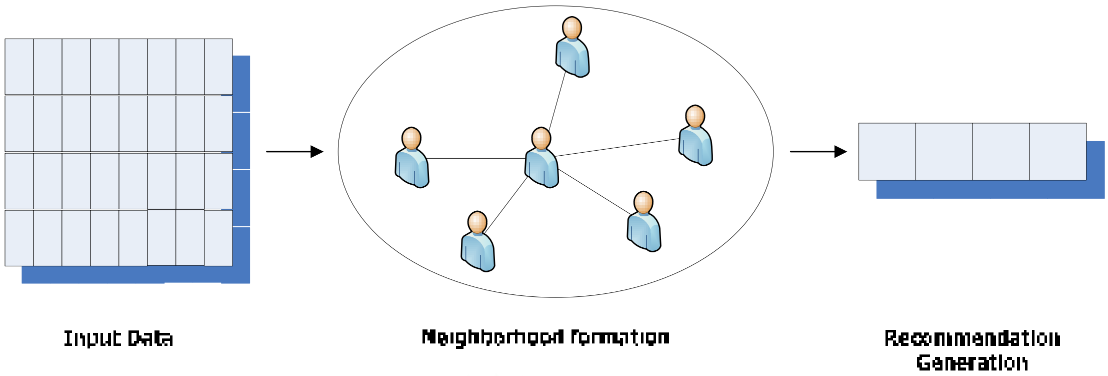


Figure 2.2: Three main parts of CF based system

Collaborative Filtering systems take advantage of this phenomenon in order to select items for their users. The basic procedure is presented in Figure 2.2 and described as follows:

1. User enters profile of ratings (Input Data)

   Over time, the system constructs user's profile, a record of the user's tastes based on past history, including explicit and implicit ratings.

2. Neighborhood formation

   The system selects the right subset of users who are most similar to the active user (a user who is interacting with the system). These similar users are called "neighbors".

   - Compare profile of an active user to the profile of other people, and weight each profile for their degree of similarity with user profile of the active user.
   - Take a group of the most similar profiles, and use them to construct an answer to some query for the active user.

3. Recommendation generation

   Neighbors' opinions or ratings are combined to form the recommendations. Then, the systems relay this information to the target user in an appropriate form.

Table 2.1: An example of user's rating matrix and co-rated items between user and user.

	Joe	Pan	Mark	Bush
Titanic	5	1	5	2
Finding Nemo	$\boxed{?}$	2	5	?
Nothing Hill	4	3	?	1
Babe	2	1	2	5
Bad Boys	5	1	4	?

Co-rated items between user "Joe" and user "Pan"

Internally, the objective of collaborative filtering can be represented as predicting missing values for cells in a user's rating matrix like the one show above.

Suppose we are trying to predict how much Joe will like "Finding Nemo". We start by finding a neighborhood for Joe. Just by looking at the matrix you can see that Joe and Mark tend to agree strongly on the past movies. There are several ways to formalize this idea of agreement, for example as Pearson correlation. Once the system selected a set of neighbors for Joe, in the most cases, the prediction value for "Finding Nemo" is simply a weighted average of the neighbor's ratings given to the "Finding Nemo".

Such a technique can be applied to a variety of problem domains. Of course, it can be used to filter items ranging from music to movies, technical journals, office equipment, restaurants, financial information, and more.

A CF system becomes more competent, while the number of users in the system increases. The more people, the greater the chance of finding close matches to any particular users.

2.5.1 Collaborative Filtering Systems

Many research projects have exploited the potential of CF in recommender systems.

GroupLens system [3]: It is a classical example of CF based system. It applies Collaborative Filtering to the personalized selection of Netnews. It provides an open architecture wherein people can rate articles and their ratings are distributed through the net. GroupLens employs *Pearson correlation coefficients* to determine similarity value between users. That is, it uses similarity between user's ratings on the same rated items (co-rated items) to find similarity value between users.

Ringo[5] is a collaborative filtering system which makes personalized music recommendations. People describe their listening pleasures to the system by rating some music. These ratings constitute the person's profile. Ringo uses these profiles to generate recommendations to individual users. Ringo compares user profiles to determine which users have similar tastes (they like the same albums and dislike the same album). Once similar users have been identified, the system can predict how much the user may like an album/artist that has not yet been rated by computing a weighted average of all ratings given to such album by other users who have similar tastes.

Ringo is an online service accessed through electronic mail. Users may sign up with Ringo by sending e-mail to ringo@media.mit.edu with the word "join" in the body of mail. People interact with Ringo by sending commands and data to a central server via e-mail. Once an hour, the server processes all incoming messages and sends the replies as necessary.

2.5.2 The Advantages of Collaborative Filtering

CF systems do not use any information regarding the actual content of the documents, but use the judgments of human as whether the document is valuable. Accordingly, it becomes possible to discover new items of interest simply because other people liked them (CF systems provide serendipitous discovery). It is also easier to provide good recommendations even when the item attributes of user interest are unknown or hidden (independence of content). For example, many movie viewers may not want to see a particular actor or genre so much as "a movie that makes me feel good" or "a smart, funny movie" (the quality of items and taste on items).

Therefore, the CF technique can reduce three limitations of IR technique mentioned above, because of the following three reasons: (i) It uses user preference in recommending items for users; (ii) It does not require users to know the appropriate query; and (iii) Independence of content (not support only machine parsable items). Moreover, CF technique provides three keys advantages that are not provided by the Content-Based Filtering[27][28]: (i) Independence of content (not support only machine parsable items); (ii) It has the ability in providing serendipitous recommendations, and (iii) It has the ability in filtering items based on quality and taste.

2.5.3 The Limitations of Collaborative Filtering

While collaborative filtering has been a substantial success, there are several problems that researchers and commercial applications have identified.

First-rater problem: If a user is the first one in the system. He/she will rate items without receiving any recommendation. This problem is inherent to the CF technique, because the recommendations are items that similar users have rated, an item cannot be recommended until other users rate such item.

Sparsity rating problem: It occurs when a user is very likely to rate only a small percentage of total number of items. In online retailers such as Amazon.com, there are millions of books that a user could never possibly rate. The overlap between user's ratings (number of co-rated items) is small. Accordingly, it is difficult to find similar people for the active user accurately. In other words, the correlations between other users and active user based on tiny co-rated items frequently prove themselves to be low quality in producing recommendation results for the active user. Another problem from tiny percentage rating is the co-rated item may not occur. Therefore, the CF system could not calculate correlation between the active user and any user. Accordingly, the CF system could not produce any recommendation result for that active user.

Synonymy problem: Different item names may be used for the same objects. The CF technique, which uses co-rated items in finding correlation between users, cannot find this latent association and treats these items differently. For example, one customer purchases 10 different *recycled letter pad* products, while another customer purchases 10 different *recycled memo pad* products. The CF based systems would see no match between product sets in computing correlation and would not be able to discover the talent association that they like *recycled office products*.

Scalability problem: As the number of users and items grows, the process of finding recommendations becomes very time consuming. In fact the computation is approximately linear with the number of users. This is especially problematic for large, high volume websites that wants to do a lot of personalizations among millions items.

2.6 Demographic Filtering

Demographic information can be used for identifying the type of users that like a certain object such as age, gender, education, etc. Demographic Filtering technique makes recommendation based on group that the current user belongs.

There is a research paper that focuses on the Demographic information by using the Winnow algorithm [19].

Demographic Filtering Systems

Pazzani [19] had discussed approaches to learning the user profiles; each approach uses a different type of information and has a different representation of a user profile on the domain of restaurant.

Table 2.2: Comparison of recommender techniques

Recommendation Techniques	Precision
Content-Based Filtering	61.2%
Collaborative Filtering	67.9%
Demographic Filtering	57.7%
Collaborative via Content (Hybrid)	70.1%

As the results in table 2.2, it shows that Demographic Filtering gives the lowest result because this technique ignores both content and user' preference.

In contrast, the Collaboration via Content gives the most accurate result because it makes recommendations based upon the experiences of users that may not be reflected by the content of the description. In order to prove the recommendation results, the combination of Content-Based and Collaborative have more accurate result than the use of individual one.

As the results in table 2.2, the use of Demographic information in Recommender System gives lower quality of recommendation so there is a researcher tried to combine with other technique.

Fred had combined the demographic into Big Five Personality Test on the domain of study course [29]. The students were given a list of elective courses after completed a survey about their demographics and the Big Five personality test. The Big Five personality system is based on the five proven independent elements: Extroversion, Emotional Stability, Orderliness, Accommodation, and Intellect. These elements make up the primary colors of personality; the interaction of elements in each person yields their overall personality profile.

Table 2.3: Combination of Demographic techniques

Recommendation Techniques	Precision
Demographic Model	83.1%
Big Five Personality Test Model	85.8%
Big Five Personality vs Demographic Model	75.7%

As the results in table 2.3, it shows that Demographic Filtering gives the lower accurate result than the Big Five. To combine Big Five Personality vs Demographic Model, the accurate result is even worst because this technique ignores both content and user' preference.

2.7 Hybrid Filtering

One common thread in recommendation researches is the need to combine recommendation techniques to achieve peak performance. All of the known recommendation techniques have strengths and weakness, and many researchers have chosen to combine techniques in different ways. Hybrid recommender systems combine two or more recommendation techniques to gain better recommendation quality and performance. Since, the Demographic Filtering is shown in section 2.6; it would not be worthy to combine Demographic Filtering with other techniques. Hybrid methods usually combine collaborative filtering and content-based filtering. Such methods are utilized in order to realize the benefits from both approaches, while at the same time minimize their disadvantages.

2.7.1 Hybrid Filtering Systems

There are various hybrid systems which have combined content-based and collaborative filtering, which is called content/collaborative hybrid systems. Burke [30] divides combination methods into seven categories: weighted, switching, mixed, feature combination, cascade, feature augmentation, and meta-level. Following details various content/collaborative hybrid systems on each combination method.

2.7.2 The Advantages of Hybrid Filtering

The combination of two or more techniques has its advantages to take benefits from all combined techniques and reduce the drawbacks of the individual one.

No Sparsity rating: the combination of Content-Based and Collaborative Filtering techniques reduce the sparsity because the Hybrid system uses the user profile to find the neighbor instead of using co-rated items.

No uncovered preference feature: the Hybrid system concern about the Multiple Criteria so it could cover the various of user preference on the items.

2.7.3 The Limitations of Hybrid Filtering

Although the hybrid system based on each combination method can reduce the drawbacks of the individual one, but it also remains some problems. There is no combination method can reduce all problems of all combined techniques.

Without distinction of interest: the current hybrid systems use the rating value on the items for evaluating user's preference opinions. The rating value represents the overall preference of the user. A user might express his/her opinion based on some specific features of the item [31]. For more accurate recommendations, the users' interest in more detailed features should be considered. This problem is called without distinction of interest problem. For example, if user (A and B) rate the same score 3 for the movie “The Matrix", but A likes its actor and B likes its genre, so the systems conclude that they have the same tastes as shown in table 2.4. Therefore, the neighbor from these systems tends to be low quality of Recommender Systems. Fortunately, many Recommender Systems that solve this problem were proposed by using the Multiple Criteria.

Table 2.4: without distinction of interest

Movies	User A	User B	User C
Bad Boys		2	3
The Matrix	3	3	
101 Dalmatians	1		1

Without weight feature problem: the current hybrid systems based on Multiple Criteria do not concern the weight of features that affects to the user's preference [31]. For example, if two users (A and B) like the same movie feature, i.e. same genre (Action), same actor (Brad Pitt), current systems conclude that both of them will be the good neighbors for each other. However, this conclusion may not be true. If A usually selects movies based on genre but not select from the actor. The weight of genre has a higher priority than weight of actor in A's opinion. On the other hand, if B select movie based on actor then the weight of actor has a higher priority than genre in B's opinion. It can be concluded that, although each couple of users that like the same movie features, they may select the different movie.

Unsuitable weight feature problem: the current hybrid system tried to solve the without weight feature problem by weighting the biggest component of each feature which is losing other component of each feature [31]. Unfortunately, each component of the feature is important so all components in each feature should be kept and weighted also.

Losing a huge of rating data problem: the contextual information of user characteristics might affect to the user preference on selected movie including where he saw the movie, when the movie was seen and with whom. Gediminas[32] concentrated on Multidimensional by using Reduction-Based approach. The Reduction-Based approach uses the intersection of Multidimensional data. For example, if the third dimensional is a Day dimension, they use Day = “Weekday" to be intersected value for filtering the data in the database as Figure 2.3. It would eliminate the Day dimension by selecting only the “Weekday" rating from the set of all ratings. Then, Gediminas[32] used any of basis recommendation technique such as Collaborative Filtering to generate the recommendation. This called losing a huge of rating data problem.

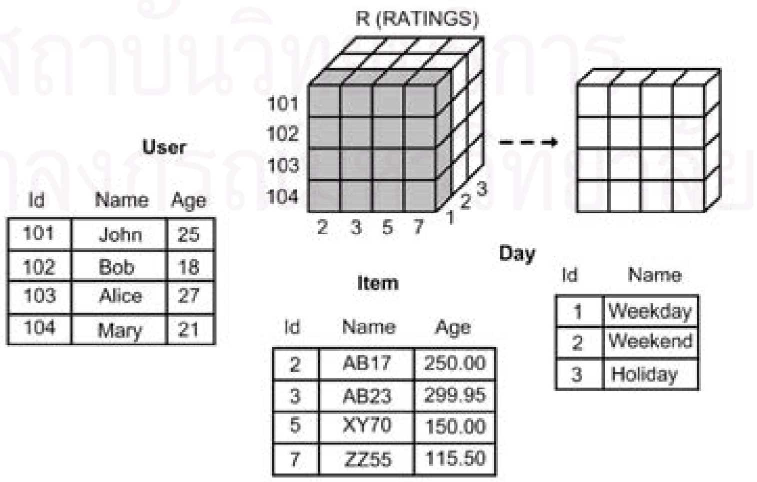


Figure 2.3: Multidimensional by using Reduction-Based Approach

2.8 Summary of Recommendation Strategies

In order to help reader understand each recommendation strategies described in this chapter more clearly. The advantage and disadvantage of each recommendation strategy are summarized in Table 2.5.

Table 2.5: Summary of recommendation strategies

Techniques	Advantage	Disadvantage
IR	- Narrowing scope of recommendations. - Users can directly identify interesting items.	- Require users to know appropriate query. - Capture only the specific query; no information about user preference.
CBF or IF	- No first-rater problem. - No sparsity rating problem. - No synonymy problem. - Not require user to know the appropriate query.	-Support only machine parsable items. - Lack of serendipitous discovery. - Lack of ability to filter items based on quality and taste.
CF	- Provide ability of serendipitous discovery. - Provide ability to filter items based on quality and taste. - Independency of content - Not require user to know the appropriate query.	- First-rater problem. - Sparsity rating problem. - Synonymy problem. - Scalability problem.
Hybrid	- Each combination has its advantages to take benefits from all combined techniques and reduce some problem of each combined technique.	- There is no combination method can reduce all problems. - Without distinction of interest. - Without weight feature problem. - Unsuitable weight feature. - Losing a huge of rating data.

CHAPTER 3

IMPROVING THE QUALITY OF RECOMMENDATIONS

This chapter will explain about the movement of hybrid recommender system and show to improve the quality of recommendation by using the Weighted Multiple Criteria and Multidimensional user profile.

3.1 Motivation

Hybrid System is well known in the recommendation research area which has reduced the problem of using either Contented-Based or Collaborative filtering. Although, there are still remaining problems such as sparsity rating, synonymy, uncovered preference features.

3.1.1 Multiple Criteria

The current system uses the rating value towards the items for evaluating user's preference opinions. The system does not have the capacity of recognizing the two distinct interests that represented in the same rating value. For more accurate recommendations, the interpretation of the user's interest must involve various features, related to the focus items; this is called *without distinction of interest problem*. The current system is also missing the weight of features that affect to the user's preference; this is called *without weight feature problem*.

Thai-Music system [31] is a hybrid recommender system. It shows that the Content-Based user profile does not cover necessary features of user's interest. To cover *without distinction of interest problem* and *without weight feature problem*, Thai-Music system based on weighted Multiple Criteria by weighting the biggest component of each feature. Unfortunately, Thai-Music system is still remaining problem on the weighted Multiple Criteria which is called *unsuitable weight feature problem*.

3.1.2 Multidimensional

There are only a few researchers that thinking of the contextual information of user. The contextual information of user characteristics might affect to the user preference on selected movie including where he saw the movie, when the movie was seen and with whom. Likewise, this contextual information becomes influential for recommending.

To cover the Multidimensional data, Gediminas[32] concentrated on Multidimensional by using Reduction-Based approach. The Reduction-Based approach uses the intersection of Multidimensional data. From intersection of each dimension, it might lose a lot of data that affect to the recommendation results. This problem is called *losing a huge of rating data problem*.

(i) Why based on Thai-Music System?

In Hybrid Recommender System, user's preference varies and always has Multiple Criteria [33]. For example, if user (A and B) rate the same score 3 for the movie “The Matrix", but A likes its actor and B likes its genre, so the systems conclude that they have the same tastes. Therefore, the neighbor from these systems tends to be low quality of Recommender Systems. This is called *without distinction of interest problem*.

Fortunately, many Recommender Systems that solve this problem were proposed by using the Multiple Criteria. Moreover, the current systems based on Multiple Criteria do not concern the weight of features that affects to the user's preference. For example, if two users (A and B) like the same movie feature, i.e. same genre (Action), same actor (Brad Pitt), current systems conclude that both of them will be the good neighbors for each other. However, this conclusion may not be true. If A usually selects movies based on genre but not select from the actor. The weight of genre has a higher priority than weight of actor in A's opinion. On the other hand, if B select movie based on actor then the weight of actor has a higher priority than genre in B's opinion. It can be concluded that, although each couple of users that like the same movie features, they may select the different movie. This called, *without weight feature problem*.

Thai-Music[31] Recommender System figures the *without weight feature problem* out by weighting only the biggest component of each feature but the way of this weight is not suitable because it will lose other components of each feature. This is called *unsuitable weight feature problem.* Until now, the Thai-Music system is focused on the Weighted Multiple Criteria in the domain of music.

(ii) The problem of Thai-Music System

Thai-Music[31] Recommender System figures the *without weight feature problem* out by weighting only the biggest component of each feature but the way of this weight is not suitable because other component of each feature might be lost. Moreover, people always select the movie from the movies' style such as most of the people select movie from genre so the frequency of selection might be considered. This problem is called *unsuitable weight feature problem*.

Additionally, the contextual information of user characteristics also affect to the user preference on each selected movie such as where he saw the movie, when the movie was seen and with whom. Gediminas[32] concentrated on Multidimensional by using Reduction-Based approach. The Reduction-Based approach uses the intersection of Multidimensional data. For example, if the third dimensional is a Day dimension, they use Day = “Weekday" to be intersected value for filtering the data in the database as Figure 3.1. It would eliminate the Day dimension by selecting only the “Weekday" rating from the set of all ratings. Then, Gediminas[32] used any of basis recommendation technique such as Collaborative Filtering to generate the recommendation. This is called *losing a huge of rating data problem*.

3.2 Proposed Method

This research focuses on providing the higher quality of recommendations. User Profile should be represented on various necessary features and the components of each feature should be weighted in more suitably way to avoid unsuitable weight feature problem. Instead of weight only the biggest component of feature, this proposed method weights all the component of each features. In real life, people always select movie from the style of movie, it implies that the selection by genre should have more weight than other features so the frequency of the selection movie features is taken into account. Moreover, to create user profile, the contextual information should be taken into consideration and use an appropriate approach to avoid losing a huge of rating data problem. This paper uses the Multiple Linear Regression analysis to perform the Multidimensional instead of using Reduction-Based approach.

3.3 The Basic Element of the Method

3.3.1. Characteristic of Movie

Movie Feature Vector (MFV)

Movie data in this case are stored in a database with characteristic data for each item. The movie characteristics are represented in the form of Movie Features Vector (MFV) which contains 24 elements (18 elements of movie genre feature, 3 elements of year feature and 3 elements of award feature) as shown in Table3.1.

Table 3.1: The Characteristic of Movie

Movie Features	Movie Component
Feature(1): Genre	Action(1), Adventure(2), Animation(3), Children(4), Comedy(5), Crime(6), Documentary(7), Drama(8), Fantasy(9), Film-Noir(10), Horror(11), Musical(12), Mystery(13), Romance(14), Sci-Fi(15), Thriller(16), War(17), Western(18)
Feature(2): Release Period	2009-2005(1), 2004-2002(2), 1999-before(3)
Feature (3): Movie Award	Oscar(1), Golden Globe(2), No award(3)

The MFV is constructed when a new item is introduced into the system. For a movie(i) MFV_i is shown as follow.

$$MFV = ((f_{11}, f_{12}, \dots, f_{1m_1}), (f_{21}, f_{22}, \dots, f_{2m_2}), \dots, (f_{N1}, f_{N2}, \dots, f_{Nm_N}))$$

Where fij is the value that represents movie characteristics component j of features i, m is the number of component in each feature and N is number of features. The value in the vector is presented in 0 or 1.

The elements in the vector MFV are MFV = ((Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western), (2009-2005, 2004-2002, 1999-before), (Oscar, Golden Globe, No award))

For example, MoviePlanet Web site notified that movie name “The Matrix" has component of each feature as

Genre = Action (1), Adventure (2), Sci – Fi (15) and Thriller (16)

ReleasePeriod = 1999 - 1995(3), and

Award = Oscar (1). It has characteristic

So, the MFV_{Matrix} is shown as follow

$$MFV_{Matrix} = ((1, 1,0, \dots ,1,1,0,0), (0,0,1), (1,0,0))$$

Thus, the value of the elements that represent the movie “The Matrix” are set to be 1, other elements are set to be 0.

3.3.2. Characteristic of User

(i) User Preference Vector (UPV)

This vector represents a user's opinion on feature or show how much each user feels towards what features affect the selection of each movie. The UPV will automatically create for each movie every time when each user gives opinion for that movie. To construct the UPV, the MFV is needed to transform by multiplying normalized rating value in range 0-1 toward each movie.

For example, if user A gives the rating value 2 (1 is dislike, 2 is neutral and 3 is like) for the movie “The Matrix", then the rating value is normalized to 0.67 (it is calculated by 2/3 = 0.67). After that, multiply the normalized rating value into the MFVMatrix to get the transformed MFVMatrix. The transformed vector is shown as follow.

$$MFV_{Matrix} = ((1, 1,0, \dots ,1,1,0,0), (0,0,1), (1,0,0))$$

$$\downarrow$$

$$Transformed\ MFV_{A,Matrix} = \big((1, 1,0, \dots ,1,1,0,0), (0,0,1), (1,0,0)\big) * Norm.Rating\ (0.67)$$

$$\downarrow$$

$$Transformed\ MFV_{A,Matrix} = \big((0.67,0.67,0, \ldots ,0.67,0.67,0,0), (0,0,0.67), (0.67,0,0)\big)$$

After transform the MFV vectors, the UPV (i) is the direct sum of the transformed MFV of rated movies and divided by the number of rated movies by user (i). Thus, the value of the elements that represent UPV of user(i) are set to be in the range of 0-1. As shown in Figure 3.1.

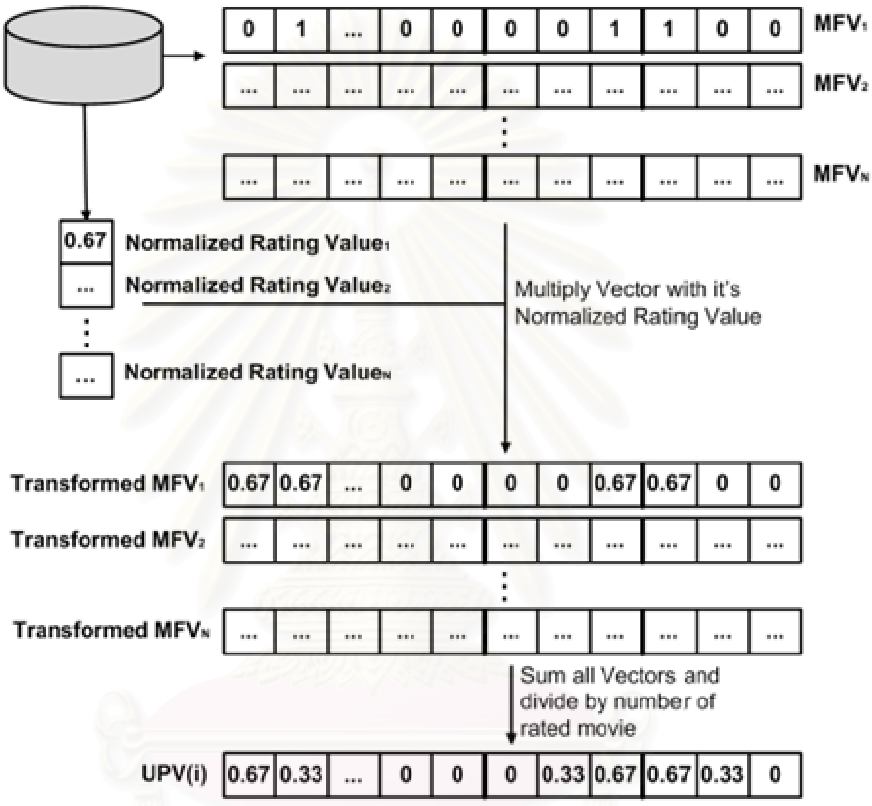


Figure 3.1: User Preference Vector

(ii) Selection on Movie Features Vector (SMV)

In the real life, people always select the movie by style of movie; it implies that genre should have more weight than other features. To increase weight of user's preference opinions, the frequency of feature selection is considered. This vector contains 7 elements of selection features which are title, genre, release period, actor, actress, director and award. Accordingly, this vector constructs automatically after the user give the opinion. Its characteristic is shown below

$$SMV = (S_1, S_2, \dots, S_N)$$

Where s_i is frequency of selection toward feature (i) and N is number of component.

The elements in the vector SMV are SMV = (title, genre, release period, actor, actress, director, and award)

For example, if the user (i) searches the first movie by genre and give the rate of that movie is 2 (normalized rating value = 0.67) then user search the second movie from genre, release period and give score 1 (normalized rating value = 0.33).

Therefore,

S_2 (genre) = 0.67+0.33/2 = 0.5

S_3 (release period) = 0.33/2 = 0.17

Accordingly,

SMV (i) = (0, 0.5, 0.17, 0, 0, 0, 0).

Thus, the value of the elements that represent SMV of user(i) are set to be in the range of 0-1. The process of this vector is also shown in figure 3.2.

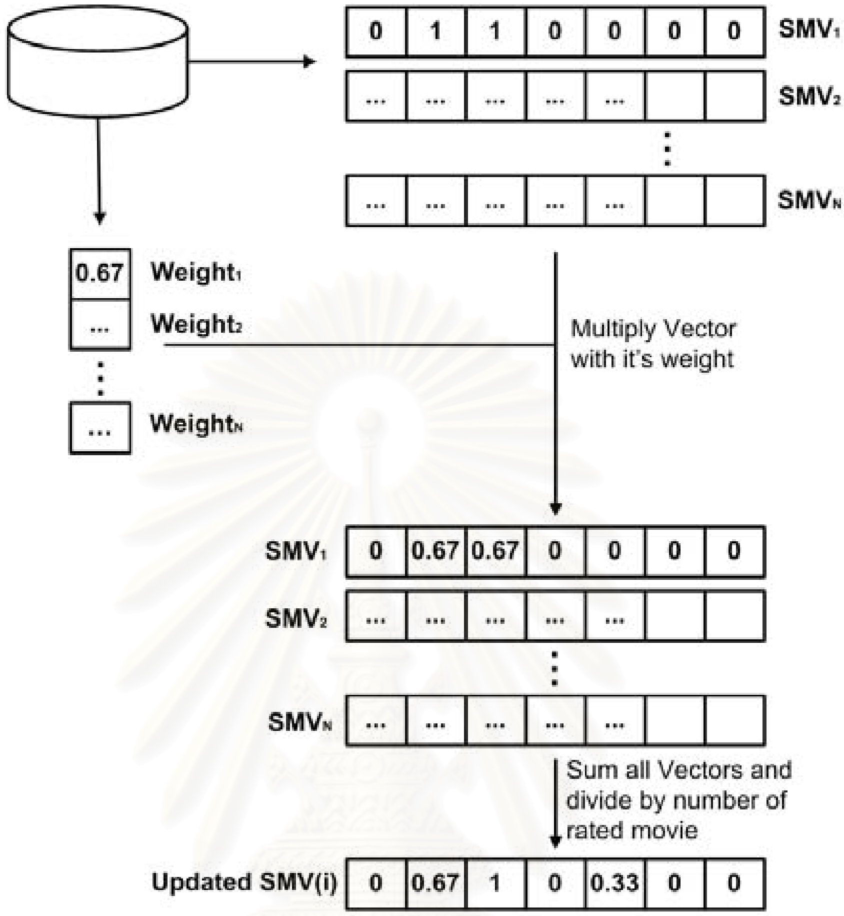


Figure 3.2: Selected on Movie Features Vector

(iii) Multidimensional Vector (MDV)

Normally, Recommender Systems ask users to give the rating value for the movie but now it's not sufficiency. To do the Multidimensional, the system needs to ask users to give more information about their contextual information which is place, day, time and companion. This paper uses the contextual information about user characteristics to create Multidimensional Vector by using Multiple Linear Regression. The form of Multiple Linear Regression equation is represented as follow.

$$y = \beta_0 + \beta_{1x_1} + \beta_{2x_2} + \ldots + \beta_{Nx_N}$$

Where y is rating value, x_j is dimension j in contextual information and x_j is the coefficient valued of each dimensions. This paper considers four dimensions which are place, day, time and companion. Its characteristic is shown bellow.

$$MDV(i) = (\beta_0, \beta_1, \dots, \beta_n)$$

Where β_i is the coefficient value from Multiple Linear Regression equation. Multiple Linear Regression will be explained in the next section.

Multiple Linear Regressions

Multiple linear regression[34] attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable *x* is associated with a value of the dependent variable *y*. The population regression line for *p* explanatory variables x_1, x_2, ... , x_p is defined to be $\mu_y = \beta_0 + \beta_{1x1} + \beta_{2x2} + \dots + \beta_{NxN}$. This line describes how the mean response μ_y changes with the explanatory variables. The observed values for *y* vary about their means μ_y and are assumed to have the same standard deviation δ . The fitted values b_0, b_1, ..., b_p estimate the parameters β_0, β_1, ..., β_n of the population regression line.

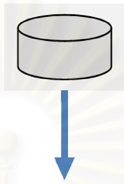
Since the observed values for *y* vary about their means μ_y, the multiple regression model includes a term for this variation. In words, the model is expressed as DATA = FIT + RESIDUAL, where the "FIT" term represents the expression $\beta_0 + \beta_{1x1} + \beta_{2x2} + \dots + \beta_{NxN}$. The "RESIDUAL" term represents the deviations of the observed values *y* from their means μ_y, which are normally distributed with mean 0 and variance δ. The notation for the model deviations is ε.

Formally, the model for multiple linear regression, given *n* observations, is

$$y_i = \beta_0 + \beta_{1xi_1} + \beta_{2xi_2} + \dots + \beta_{Nxi_N} + \varepsilon_i \qquad \text{for i = 1, 2, … n}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{21} & X_{31} & \cdots X_{p1} \\ 1 & X_{22} & X_{32} & \cdots X_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots X_{p3} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

For example,



Data of User(i)

$$\begin{pmatrix} 3_{\ like} \\ 2_{\ neutral} \\ \vdots \\ 1_{\ dislike} \end{pmatrix} = \begin{pmatrix} 1 & 1_{\ theater} & 1_{\ friend} & 3_{\ weekend} & 1_{\ daytime} \\ 1 & 3_{\ home} & 3_{\ family} & 1_{\ weekday} & 3_{\ nighttime} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1_{\ theater} & 1_{\ friend} & 3_{\ weekend} & 1_{\ daytime} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$

$$\downarrow$$

$$(X'X)\beta = X'Y$$

$$\downarrow$$

$$\beta = X'Y(X'X)$$

$$\downarrow$$

$$y(i) = 4.5 + 2x_1 + 1.35x_2 + 0.5x_3 + 3.1x_4$$

$$\downarrow$$

$$MDV(i) = (4.5, 2, 1.35, 0.5, 3.1)$$

3.4 Detail Steps Finding Neighbor Process

Neighbor of the target user is derived from three vectors; User Preference Vector (UPV), Selection on Movie Features Vector (SMV) and Multidimensional Vector (MDV). There 5 methods of finding neighbor processes are shown in below and the process is shown in figure 3.3.

Step 1: For the target user, the updated UPV is selected. In order to reduce the unsuitable weight problem, the biggest component of each feature in UPV is taken to calculate the weight value as show in equation (2).

$$w_i = \frac{f_i}{\sum_{i=0}^{N}(f_i)} \tag{2}$$

where w_i is the weight value for feature (i), f_i is the biggest component of feature(i) and N is number of component. Then, multiply the weight value to all component of UPV by using its own weight value of that feature. This step is also shown in Figure 3.3. Repeat this step to do the other users in the system. Step 2: In order to reduce the unsuitable weight problem, this research also use the frequency of the selection movie features to improve the recommendation by selecting the SMV vector of target user

Step 3: To do the Multidimensional and reduce the losing a huge of rating data problem, the Multidimensional Vector (MDV) should be used

Step 4: To construct the vector that represent the taste of each user in the system, the UPV, SMV and MDV are merged.

$$User\ Vector = UPV + SMV + MDV$$

Step 5: To find the association of each pair of user (the target user and another user in the system), this approach use the distance of vector. Distance of vector calculates by equation (3).

$$Distance = \sqrt{\sum_{i=0}^{N}(v_{1i} - v_{2i})^2} \tag{3}$$

Step 6: Neighbor is produced by selecting the user who has the smallest value of Distance Value.

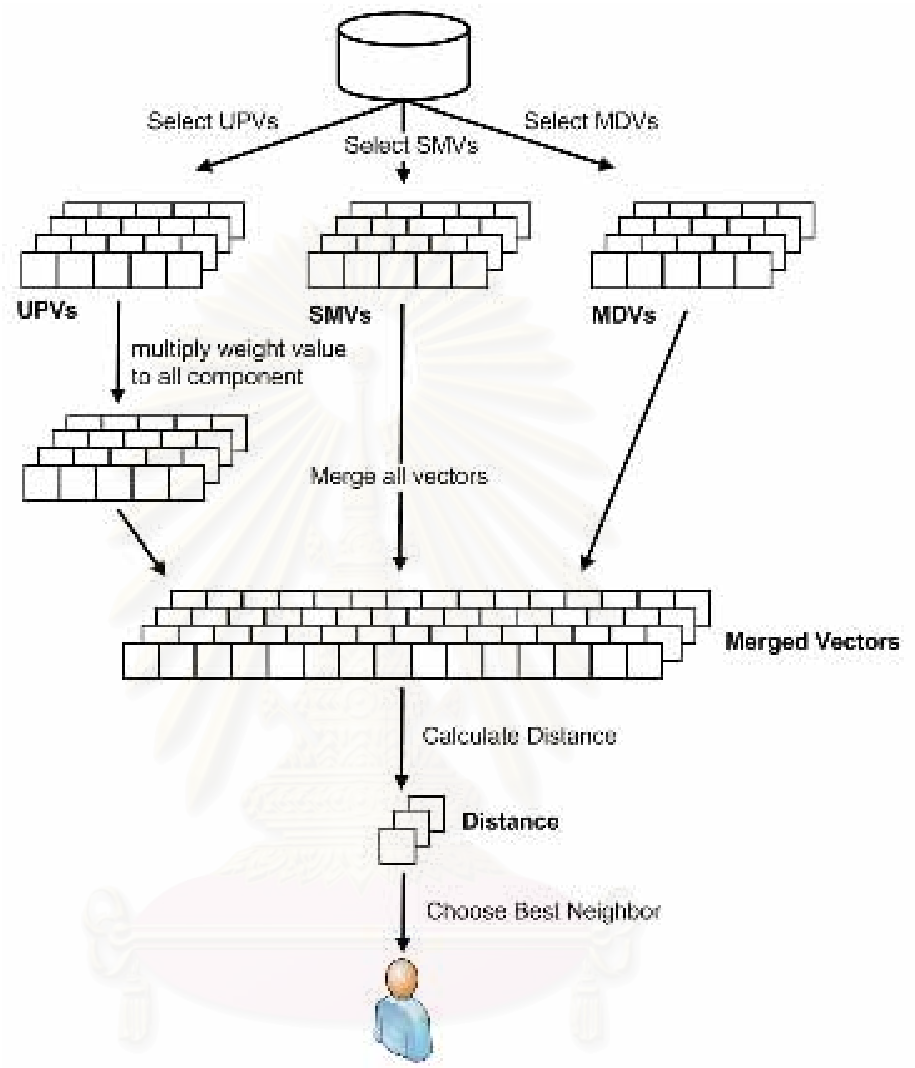


Figure 3.3: Finding Neighbor Process

CHAPTER 4

MOVIEPLANET SYSTEM

4.1 Prototype System

A prototype recommender system for movie, called MoviePlanet System, is created to implement and evaluate the proposed neighborhood formation method. This research has chosen to work within the domain of movie selection.

The MoviePlanet system is an online movie recommender system accessible through the Web. MoviePlanet is implemented by Microsoft Visual Studio .Net on Microsoft XP Professional and acts as WWW server. It uses Microsoft SQL Server to be data storage.

4.2 Database Structure

The MoviePlanet system has to main databases. One is Movie Database and another is User Database. The information of the available movies obtained from the Internet Movies' Database Site and the Thai-movies obtained from the www.pantip.com.

4.2.1 Movies Database

The Movie Database contains 1063 movie items and can be divided into two tables which are Metadata Table and Movie Feature Table.

(i) Moviedata Table

It contains movie metadata of each movie for all movies. This moivedata is movie detail presented to the user in order to help his/her make a decision whether he/she wants to see that movie as shown in the “Movie Detail Page”. Each record (row) of table refers to all metadata of each movie.

There are 9 fields (columns) in the Metadata table.

(a) Id: id number of each movie item.

(b) Title: title name of each movie item.

(c) Period of time: the released year of each movie.

(d) Genre: category or genre of each movie. For instance, the genre field of the movie "The Matrix" contains "Action", "Adventure", "Sci – Fi" and "Thriller"

(e) Actor: actor name of each movie item.

(f) Actress: actress name of each movie item.

(g) Director: director name of each movie item.

(h) Award: information about the award that each movie obtains, such as Oscar or Golden Globe.

(i) Trailer: The URL linked to the trailer (movie preview) of each movie.

(ii) MovieFeatures Table

It contains the movie feature vector (MFV) of each movie for all movies. Each record (row) of table refers to the MFV of each movie.

There are 25 fields (columns) in the MovieFeature table which are Id, Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western, 2009-2005, 2004-2002, 1999-before, Oscar, Golden Globe and No award

4.1.2 Users Database

The User Database contains the data of 100 users and can be divided into? Tables which are UserInfo Table, UserPreference Table, SelectionMovie Table, Multidimension Table and PredictionMovie Table.

(i) UsersInfo Table

It contains user's information data of each user for all users.

There are 8 fields in the UserInfo Table.

(a) UsedId: id number of each user.

(b) UserName: The user name for log in of each user.

(c) Password: the user password for log in of each user.

(d) FirstName : the first name of each user.

(e) LastName: the last name of each user.

(f) Age: the age of each user.

(g) Gender: the gender of each user.

(h) Occupation: the occupation of each user.

(ii) UserRating Table

It contains the information of the rating movies by the user for all user.

There are fields in the UserRating Table

(a) UserId: Id number of each user.

(b) MovieId: Movie Id of each movie that user was selected to give the rating.

(c) Rating: the number that present how much user like the selected movie (1-dislike, 2-neutral and 3-like).

(d) Place: the place where user watch the selected movie (home or theater).

(e) Companion: the person who user watch the selected movie with (boy/girl friend, friend, family and other).

(f) Day: the day that user watch the selected movie (Weekday or Weekend/holiday).

(g) Time: the time that user watch the selected movie (Daytime or Nighttime).

(iii) UserPreference Table

It contains the updated User Preference Vector (UPV) of each user for all users in the system. Each record (row) of table refers to the UPV of each user.

There are 25 fields in UserPreference Table

(a) UserId: Id number of each user.

(b) Action : the average value of selected action movies

(c) Adventure : the average value of selected adventure movies

(d) Animation : the average value of selected animation movies

(e) Children : the average value of selected children movies

(f) Comedy : the average value of selected comedy movies

(g) Crime : the average value of selected crime movies

(h) Documentary : the average value of selected documentary movies

(i) Drama : the average value of selected drama movies

(j) Fantasy : the average value of selected fantasy movies

(k) Film-Noir : the average value of selected film-noir movies

(l) Horror : the average value of selected horror movies

(m) Musical : the average value of selected musical movies

(n) Mystery : the average value of selected mystery movies

(o) Romance : the average value of selected romance movies

(p) Sci-Fi : the average value of selected sci-fi movies

(q) Thriller : the average value of selected thriller movies

(r) War : the average value of selected war movies

(s) Western : the average value of selected western movies

(t) 2009-2005 : the average value of selected movies this period of time

(u) 2004-2002 : the average value of selected movies this period of time

(v) 1999-before : the average value of selected movies this period of time

(w) Oscar : the average value of selected movies that have Oscar award

(x) Golden Globe : the average value of selected movies that have Golden Globe award

(y) No award : the average value of selected movies that have no award

(iv) SelectionMovie Table

It contains the updated Selection on Movie Features Vector (SMV) of each user for all users in the system. Each record (row) of table refers to the SMV of each user.

There are 8 fields in SelectionMovie Table

(a) UserId: Id number of each user.
(b) Title : the average value of searched movies by tiltle
(c) Genre : the average value of searched movies by genre
(d) Release Period : the average value of searched movies by release period
(e) Actor : the average value of searched movies by actor name
(f) Actress : the average value of searched movies by actress name
(g) Director : the average value of searched movies by director name
(h) Award : the average value of searched movies by award

(v) Multidimensional Table

It contains the updated Mutidimensional Vector (MDV) of each user for all users in the system. Each record (row) of table refers to the MDV of each user.

There are 8 fields in Multidimensional Table

(a) UserId: Id number of each user.
(b) 1^{st}coeffient: the β_0 from the Multiple Linear Regression equation as shown in section 3.3.2
(c) 2^{nd}coeffient: the β_1 of x_1 from the Multiple Linear Regression equation as shown in section 3.3.2
(d) 3^{th}coeffient: the β_2 of x_2 from the Multiple Linear Regression equation as shown in section 3.3.2
(e) 4^{th}coeffient: the β_3 of x_3 from the Multiple Linear Regression equation as shown in section 3.3.2

(f) 5^{th}coeffient: the β_4 of x_4 from the Multiple Linear Regression equation as shown in section 3.3.2

4.3 Rating Item

The MoviePlanet System offers the supplement feature, which make the system to increase the quality of recommendations. This feature is to provide the proper items for the user to give the opinions or rating.

In order to make the recommendation, the system needed acquire some information from the user. The direct way to do this task is to ask the user to rate the by selecting some movies from the presenting items of the system. If the system requires too much effort, user will give up which result will be poor recommendation. The system should provide the list that can minimize the user's effort and encourage him/her to use the system with fun until getting the good recommendations.

4.4 Process of MoviePlanet System

4.4.1 Entering user's opinion

Each user starts by registering to the MoviePlanet system. Then, the search page is shown up for entering an ykeywords about title, genre, year, award, director, actor and actress to queries the movie from the database as shown in Figure 4.1. After that, the result page displays the list of the movies search result. See Figure 4.2. On the result page, user clicks on the movie name. Then, user is asked to give the contextual information and the opinion about that movie. The contextual information is information about where he saw the movie (a movie was seen either in the theater or at home), when the movie was seen (he had seen either in the Day Time or Night Time and either on Weekday or Weekend) and with whom (Friend, Boyfriend/Girlfriend, Family or Other) as shown in Figure 4.3.

In the MoviePlanet System, there are three levels of the opinion, which are 1-3(1 is dislike, 2 is neutral and 3 is like). After that, the UPV, SMV and MDV are automatically updated.

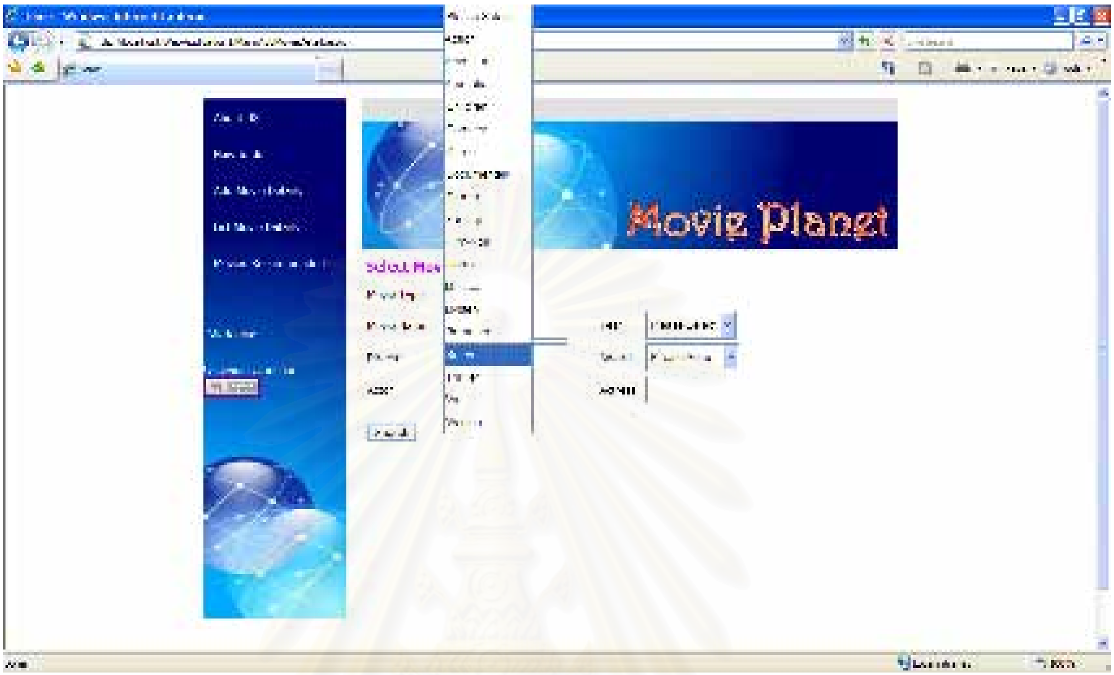


Figure 4.1 Search page of MoviePlanet system

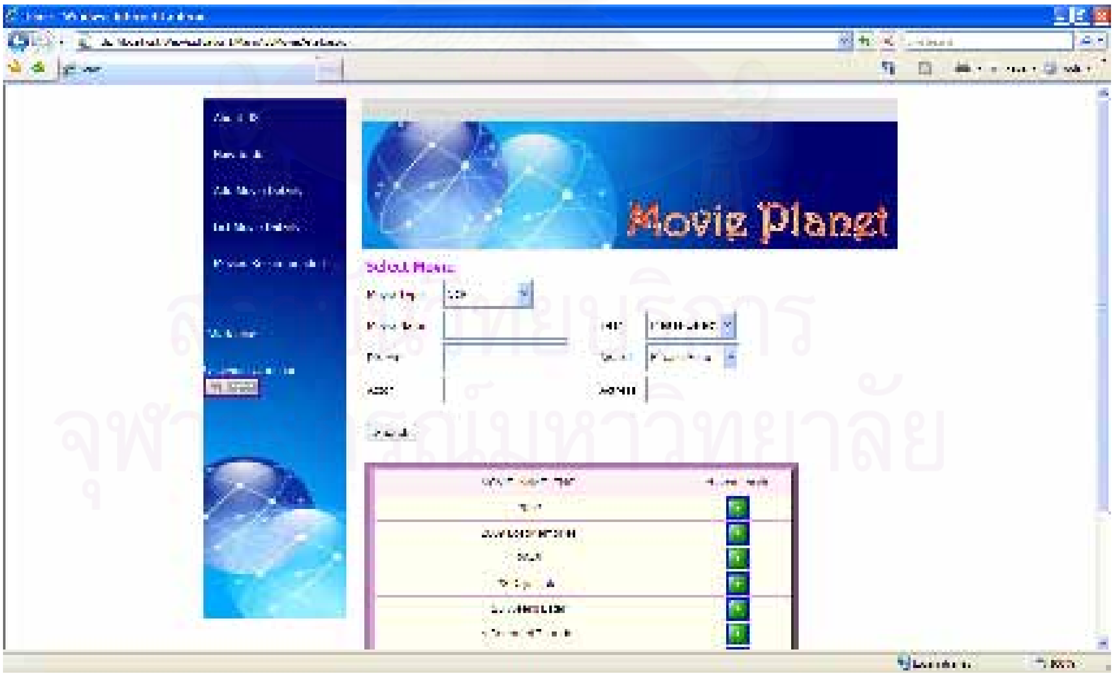


Figure 4.2 Result page (from Search page) of MoviePlanet system

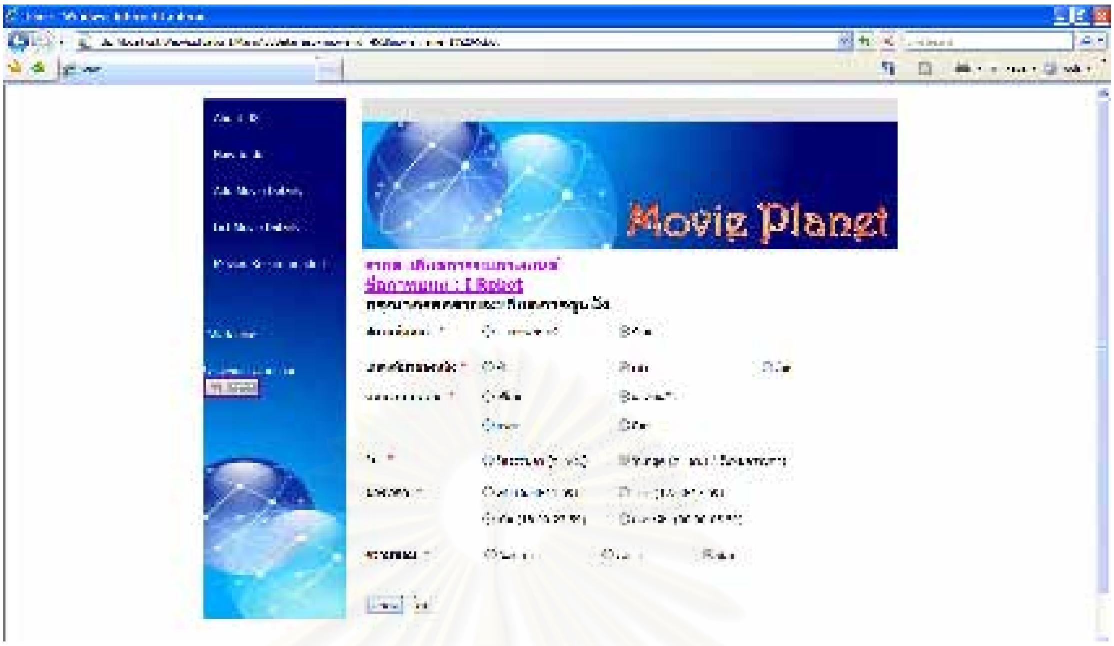


Figure 4.3 give the contextual information of user for MoviePlanet system

4.4.2 Finding Neighbor Process

This research is considered the higher quality of recommendation by using Multiple Criteria and Multidimensional user profile. In order to find neighbor, the merged vector (UPV + SMV + MDV) is constructed and then the distance of merged vector between the target user and another user (for all the rest of users in the system) are calculated. The selected neighbor is the person who has the smallest value of the Distance.

4.4.3 Generating the Recommendations

As the Recommender System usually show user's favorite or like most item, the MoviePlanet System presents the movie list of liked movies by the neighbor as the recommendations. See Figure 4.4.

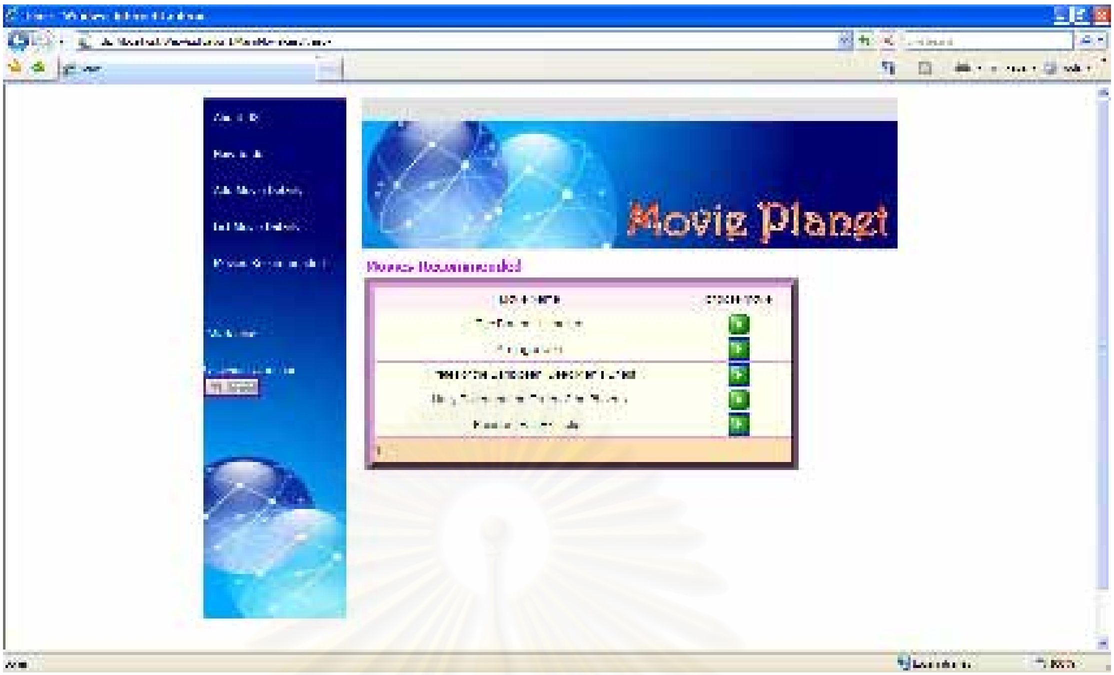


Figure 4.4: Recommended movies from MoviePlanet system

CHAPTER 5

EXPERIMENTS AND EVALUATION

As mentioned about the use of combining the Multiple Criteria and Multidimensional user profile, it is realized that the quality of recommendation results will be increased when the quality of neighbor increased. In order to get higher quality of neighbors, the user's opinion should be representing in various feature of interest.

5.1 Experiments

In order to prove the effectiveness of the proposed method, there are four kinds of experiment were taken.

For the experiment (1), this research tried to answer the following question. Could the new approach (Weight all the component of each feature) provides more accurate result than the current hybrid system (Thai-Music system - weight only the biggest component of each feature)?

For the experiment (2), this research tried to answer the following question. Could the Weight all the component of each feature and use the frequency of selection approach provides more accurate result than weight all the component of each feature approach?

For the experiment (3), this research tried to answer the following question. Could the new approach which based on Weighted Multiple Criteria and Multidimensional user profile by weighting all the component of each feature and concern about the contextual information, provides more accurate recommendation results than the current approach which based on Weighted Multiple Criteria by weighting all the component of each feature?

For the experiment (4), this research tried more to answer the following question. Could the proposed system (MoviePlanet) which based on Weighted Multiple Criteria and Multidimensional user profile by weighting all the component of each feature, using the frequency of selection and also concern about the contextual information, provides more accurate recommendation results than the Weight all component of each feature and use the Frequency of selection movies approach?

5.2 Data

In the experimental evaluation, the data of 1063 movies was inserted into the movie database and 100 users were willing to use the system. Each user was asked to rate at least 10 movies. The total of collected opinion from the experiments, are sum up to 1522 rating (1209 ratings are set to be a training set and 313 ratings are in the test set). The accuracy of this recommendations generated by the system will be revealed when the users say that they like the favorite recommend movies and dislike the undesirable recommend movies.

5.3 Evaluation Criteria

There are four criteria that used for determining the accuracy and quality of the recommendations.

5.3.1 MAE (Mean Absolute Error)

MAE (Mean Absolute Error) [5] is the average absolute deviation between the system's recommendation value and the user's actual preference value. It is a metric, which measures how close the recommender system's recommendation values (or predicted ratings) are to the actual user's preference values. The MAE is represented as equation (5.1). The lower the MAE is the more accurate the recommendation results.

$$|\bar{E}| = \frac{\sum_{i=1}^{T} |R_i - p_i|}{T} \tag{5.1}$$

Where, R_i is a recommendation value (or predicted value) for each item by the system.

p_i is the user's actual preference value for each item.

T is the number of items in the test set (see meaning of our test set in the evaluation process below).

To detail the four criteria below, the famous contingency table is introduced (Table 5.1). A is a set of relevant items by the user, and $\bar{A}$ is a set of non-relevant items by the user. B is a set of items accepted by the system, and $\bar{B}$ is a set of items rejected by the system.

Table 5.1: Contingency Table

	$Relevant\ by\ User(A)$	$Non-Relevant\ (\bar{A})$
$Accepted\ by\ system\ (B)$	$A \cap B$	$\bar{A} \cap B$
$Rejected\ by\ system\ (\bar{B})$	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$

5.3.2 Recall (or Sensitivity)

Recall (or Sensitivity) [36] is the probability that the relevant items will be accepted by the system. Recall is defined as the ratio of the user's relevant items, which are accepted by the system, to the total number of user's relevant items as shown in equation (5.2).

$$Recall = \frac{|A \cap B|}{|A|} \tag{5.2}$$

5.3.3 Precision (or Positive Predictive Value)

Precision (or Positive Predictive Value) [36] is the probability that the accepted items are relevant. Precision is defined as the ratio of the user's relevant items, which are accepted by the system, to the total number of accepted items as shown in equation (5.3).

$$Precision = \frac{|A \cap B|}{|B|} \tag{5.3}$$

5.3.4 F-measure

F-measure [32] is the weighted harmonic mean of precision and recall as shown in equation (5.4). The higher F-measure is more accurate the results.

$$F-measure = \frac{2(Recall)(Precision)}{Recall+Precision} \tag{5.4}$$

5.4 Experiment I

The objective of experiment I is to answer the following question. In the area of current hybrid system, Could the new approach (weight all the component of each feature) provides more accurate result than the current hybrid system (Thai-Music system - weight only the biggest component of each feature)?

Therefore, this research simulates both of two hybrid recommender systems on the same dataset and then compares these two systems.

5.4.1 Evaluation Process

Step 1: the Thai-Music system and Weight all component of each feature approach are created and simulated the methods on the same dataset.

Step 1.1 the Thai-Music system was created according to the process describe below.

1.1.1 Input Data: User Ratings (User's Opinion on various features)

1.1.2 Find Neighbor:

i. The UPV is constructed and then weight only the biggest component of each feature.
ii. Find the distance of Weighted UPV between target user and other users in the system.
iii. The person who has the smallest distance value is the neighbor.

Step 1.2 the Weight all component of each feature approach, was created according to the process describe below.

1.2.1 Input Data: User Ratings (User's Opinion on various features)

1.2.2 Find Neighbor:

i. The UPV is constructed and then weight all the component of each feature with its weight value.

ii. Find the distance of Weighted UPV between target user and other users in the system.

iii. The person who has the smallest distance value is the neighbor.

Step 2: Top 5 movies of neighbor are selected to be recommendations

Step 3: Calculate Mean absolute error (MAE) of each system. The error refers to the difference between a recommendation value (predicted value) of the movie in the test set of the target user and rating value of recommended movies that is given by the neighbor. For the predicted value, like most has score = 1, neutral has score = 2 and dislike has score = 3.

Step 4: Calculate the Recall, Precision and F-measure

Step 5: Compare MAE, Recall, Precision and F-measure between both systems.

Table 5.2: The result from Thai-Music and Weight all component of each feature approach.

Approaches	MAE	Recall	Precision	F-measure
Thai-Music	0.3341	0.6065	0.5606	0.5827
Weight all components	0.2986	0.6902	0.6414	0.6649

5.4.2 Evaluation Results

All the criteria are employed to compare Thai-Music system, which find the neighbor by weighting only on the biggest component of each feature with Weight all component of each feature approach. As presented in the table 5.2, the MAE of Weight all component system is lower than the Thai-Music system. It can be conclude that the Weight all component of each feature approach provides more accurate recommendation results than the Thai-Music system.

In addition, the results show that the capability of Weight all component of each feature approach in retrieving relevant movies is higher than the Thai-Music system because the Recall and Precision values from Weight all component of each feature approach are higher than Thai-Music system. The F-measure of Weight all component system is also higher than Thai-Music.

5.4.3 Discussion

Both Thai-Music method and Weight all component of each feature approach are concentrated on the Multiple Criteria which considered the user's interest in more detailed feature. Since the weight of all components of each feature approach give more accurate result than the Thai-Music which weight only the biggest component of each feature, it can conclude that not only the biggest component of each feature affect to the recommendations but other components of each feature should be considered also.

5.5 Experiment II

The objective of experiment II is to answer the following question. In the area of current hybrid system, Could the Weight all the component of each feature and use the frequency of selection approach provides more accurate result than weight all the component of each feature approach?

Therefore, this research simulates both of two hybrid systems on the same dataset and then compares these two systems.

5.5.1 Evaluation Process

Step 1: the Weight all component of each feature approach and the Weight all component of each feature and using the frequency of selection approach are created and simulated the methods on the same dataset.

Step 1.1 the Weight all component of each feature approach was created according to the process describe below.

1.1.1 Input Data : User Ratings (User's Opinion on various features)

1.1.2 Find Neighbor :

i. The UPV is constructed and then weight all the component of each feature with its weight value.

ii. Find the distance of Weighted UPV between target user and other users in the system.

iii. The person who has the smallest distance value is the neighbor.

Step 1.2 the Weight all component of each feature and using the frequency of selection approach was created according to the process describe below.

1.2.1 Input Data : User Ratings (User's Opinion on various features) and Frequency of selection movies

1.2.2 Find Neighbor :

i. The UPV is constructed and then weight all the component of each feature with its weight value.

ii. The SMV is constructed.

iii. The UPV and SMV are merged.

iv. Find the distance of merged vector between target user and other users in the system.

v. The person who has the smallest distance value is the neighbor.

Step 2: Top 5 movies of neighbor are selected to be recommendations

Step 3: Calculate Mean absolute error (MAE) of each system. The error refers to the difference between a recommendation value (predicted value) of the movie in the test set of the target user and rating value of recommended movies that is given by the neighbor. For the predicted value, like most has score = 1, neutral has score = 2 and dislike has score = 3.

Step 4: Calculate the Recall, Precision and F-measure

Step 5: Compare MAE, Recall, Precision and F-measure between both systems.

Table 5.3: the result from Weight all component of each feature approach and Weight all component of each feature and using Frequency of selection approach.

Approaches	MAE	Recall	Precision	F-measure
Weight all component	0.2986	0.6902	0.6414	0.6649
Weight all component + Frequency	0.2591	0.7384	0.6720	0.7036

5.5.2 Evaluation Results

All the criteria are employed to compare Weight all component of each feature approach with Weight all the component of each feature and using the frequency of selection movies. As presented in the table 5.3, the MAE of Weight all component of each feature and using Frequency of selection approach is lower than the Weight all component of each feature. It can be conclude that the Weight all component of each feature and using Frequency of selection approach system provides more accurate recommendation results than the Weight all component of each feature.

In addition, the results show that the capability of Weight all component of each feature and using Frequency of selection approach in retrieving relevant movies is higher than the Weight all component of each feature because the Recall and Precision values from Weight all component of each feature and using Frequency of selection approach are higher than Weight all component of each feature. The F-measure of Weight all component of each feature and using Frequency of selection approach is also higher than Weight all component of each feature.

5.5.3 Discussion

Both approaches are concentrated on the Multiple Criteria which considered the user's interest in more detailed feature. The first approach based on weight all components of each feature but the second approach based on weight all component of each feature and use the frequency of selection. As the result that shows in table 5.3, the second approach gives more accurate result than the first one, it can be conclude that the addition of weight by using the selection's style of people (people always select movie by styles of movie) affect to the recommendation results.

5.6 Experiment III

The objective of experiment III is to answer the following question. In the area of current hybrid system, could the new approach which based on Weighted Multiple Criteria and Multidimensional user profile by weighting all the component of each feature and concern about the contextual information, provides more accurate recommendation results than the current approach which based on Weighted Multiple Criteria by weighting all the component of each feature?

Therefore, this research simulates both of two hybrid systems on the same dataset and then compares these two systems.

5.6.1 Evaluation Process

Step 1: the Weight all component of each feature approach and the Weight all component of each feature and Multidimensional user profile approach are created and simulated the methods on the same dataset.

Step 1.1 the Weight all component of each feature approach was created according to the process describe below.

1.1.1 Input Data : User Ratings (User's Opinion on various features)

1.1.2 Find Neighbor :

i. The UPV is constructed and then weight all the component of each feature with its weight value.

ii. Find the distance of merged vector between target user and other users in the system.

iii. The person who has the smallest distance value is the neighbor.

Step 1.2 the Weight all component of each feature and Multidimensional user profile approach was created according to the process describe below.

1.2.1 Input Data: User Ratings (User's Opinion on various features) and contextual information of user when watch the movie.

1.2.2 Find Neighbor:

i. The UPV is constructed and then weight all the component of each feature with its weight value.

ii. The MDV is constructed.

iii. The UPV and MDV are merged.

iv. Find the distance of merged vector between target user and other users in the system.

v. The person who has the smallest distance value is the neighbor.

Step 2: Top 5 movies of neighbor are selected to be recommendations

Step 3: Calculate Mean absolute error (MAE) of each system. The error refers to the difference between a recommendation value (predicted value) of the movie in the test set of the target user and rating value of recommended movies that is given by the neighbor. For the predicted value, like most has score = 1, neutral has score = 2 and dislike has score = 3.

Step 4: Calculate the Recall, Precision and F-measure

Step 5: Compare MAE, Recall, Precision and F-measure between both systems.

Table 5.4: The result from Weight all component of each feature approach and Weight all component of each feature and Multidimensional user profile approach.

Approaches	MAE	Recall	Precision	F-measure
Weight all component	0. 2986	0. 6902	0. 6414	0. 6649
Weight all component + Multidimensional	0.2515	0.7477	0.7339	0.7407

5.6.2 Evaluation Results

All the criteria are employed to compare the Weight all component of each feature approach with Weight all component of each feature and Multidimensional user profile approach, which find the neighbor by weighting all the component of each feature and concern about the contextual information. As presented in the table 5.4, the MAE of Weight all component of each feature and Multidimensional user profile approach is lowers than the Weight all component of each feature approach. It can be concluding that the Weight all component of each feature and Multidimensional user profile approach provides more accurate recommendation results than the Weight all component of each feature approach.

In addition, the results show that the capability of Weight all component of each feature and Multidimensional user profile approach in retrieving relevant movies is higher than the Weight all component of each feature approach because the Recall and Precisions values from Weight all component of each feature and Multidimensional user profile approach are higher than the Weight all component of each feature approach. The F-measure of Weight all component of each feature and Multidimensional user profile approach is also higher than the Weight all component of each feature approach.

5.6.3 Discussion

Both approaches are concentrated on the Multiple Criteria which considered the user's interest in more detailed feature. The first approach based on weight all component of each feature. Then, the second approach based on weight all component of each feature and concerned about the Multidimensional user profile. As the result that shows in table 5.4, it can be conclude that other dimensions of user's characteristic such as where he saw the movie, when the movie was seen and what time affect to the recommendation results.

5.7 Experiment IV

The objective of experiment (4) is to answer the following question. In the area of hybrid system, could the proposed system (MoviePlanet) which based on Weighted Multiple Criteria and Multidimensional user profile by weighting all the component of each feature, using the frequency of selection and also concern about the contextual information, provides more accurate recommendation results than the Weight all component of each feature and use the Frequency of selection movies approach?

Therefore, this research simulates both of two hybrid systems on the same dataset and then compares these two systems.

5.7.1 Evaluation Process

Step 1: the Weight all component of each feature and use the Frequency of selection movies approach and the MoviePlanet system are created and simulated the methods on the same dataset.

Step 1.1 the Weight all component of each feature and use the Frequency of selection movies approach was created according to the process describe below.

1.1.1 Input Data: User Ratings (User’s Opinion on various features) and frequency of selection movies.

1.1.2 Find Neighbor:

i. The UPV is constructed and then weight all the component of each feature with its weight value.

ii. The SMV is constructed.

iii. The UPV and SMV are merged.

iv. Find the distance of merged vector between target user and other users in the system.

v. The person who has the smallest distance value is the neighbor.

Step 1.2 the MoviePlanet system was created according to the process describe below.

1.2.1 Input Data: User Ratings (User's Opinion on various features), Frequency of selection movies and contextual information of user when watch the movie.

1.2.2 Find Neighbor:

i. The UPV is constructed and then weight all the component of each feature with its weight value.

ii. The SMV is constructed.

iii. The MDV is constructed.

iv. The UPV, SMV and MDV are merged.

v. Find the distance of merged vector between target user and other users in the system.

vi. The person who has the smallest distance value is the neighbor.

Step 2: Top 5 movies of neighbor are selected to be recommendations

Step 3: Calculate Mean absolute error (MAE) of each system. The error refers to the difference between a recommendation value (predicted value) of the movie in the test set of the target user and rating value of recommended movies that is given by the neighbor. For the predicted value, like most has score = 1, neutral has score = 2 and dislike has score = 3.

Step 4: Calculate the Recall, Precision and F-measure

Step 5: Compare MAE, Recall, Precision and F-measure between both systems.

Table 5.5: the result from MoviePlanet system and Weight all component of each feature and use Frequency of selection approach.

Approaches	MAE	Recall	Precision	F-measure
Weight all component + Frequency	0. 2591	0. 7384	0. 6720	0. 7036
MoviePlanet	0.2018	0.7895	0.7965	0.7930

5.7.2 Evaluation Results

All the criteria are employed to compare Weight all component of each feature and use the Frequency of selection movies approach with MoviePlanet system, which find the neighbor by weighting all the component of each feature, using the frequency of selection movie and concern about the contextual information. As presented in the table 5.5, the MAE of MoviePlanet system is lower than the Weight all component of each feature and use the Frequency of selection movies approach. It can be conclude that the MoviePlanet system provides more accurate recommendation results than the Weight all component of each feature and use the Frequency of selection movies approach.

In addition, the results show that the capability of MoviePlanet system in retrieving relevant movies is higher than the Weight all component of each feature and use the Frequency of selection movies approach because the Recall and Presicion values from MoviePlanet system are higher than Weight all component of each feature and use the Frequency of selection movies approach. The F-measure of MoviePlanet system is also higher than Weight all component of each feature and use the Frequency of selection movies approach.

5.7.3 Discussion

Both approaches are concentrated on the Multiple Criteria which considered the user's interest in more detailed feature. The first approach based on Weight all component of each feature and use the Frequency of selection movies and the MoviePlanet system based on weight all components of each feature, frequency of selection movies and concern about the contextual information (Multidimensional user's profile). Since weight all component of each features and other dimensions affect to the recommendation results as show in experiment III so the MoviePlanet add the other dimensions to see how it work. As the results that show in table 5.5, the MoviePlanet provides the better result which can be conclude that the contextual information of user's characteristic such as where he saw the movie, when the movie was seen and what time affect to the recommendation results.

5.8 Discussion

There are four experiments which explain the results and the reason why the results become like that in section 5.4 to 5.7. Here, this section will summarize the four experiments above in more detailed.

Table 5.6: Results from all approaches

Approaches	MAE	Recall	Precision	F-Measure
Thai-Music	0.3341	0.6065	0.5606	0.5827
Weight all component	0.2986	0.6902	0.6414	0.6649
Weight all component + Frequency	0.2591	0.7384	0.6720	0.7036
Weight all component + Multidimensional	0.2515	0.7477	0.7339	0.7407
MoviePlanet	0.2018	0.7895	0.7965	0.7930

As present the experiment I to IV and the results shown in table 5.6, the result of MoviePlanet is the most accurate result. As the discussion above, it shows that combination of Weighted Multiple Criteria (by weight of all components and use the frequency of selection) and Multidimensional user profile (by using the contextual information of user's characteristics) give more accurate result than Thai-Music which weight only the biggest component of each feature.

5.9 Significant of Proposed Method

The proposed method combines three approaches which are the weight of all components of each feature, the use of frequency of selection and the use of contextual information as Multidimensional user profile. To see how the result of proposed method (MoviePlanet system) is significant by comparing with the Thai-Music system by using the Z-test.

Z-Test

Z-Test [36] is a test of any of a number of hypotheses in inferential statistics that has validity if sample size is sufficiently large and the underlying data are normally distributed by using equation 5.5 and 5.6. The methods of inference used to support or reject claims based on sample data are known as *tests of significance*.

$$z = \frac{\bar{x}-\bar{y}}{\sqrt{\left(\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}\right)}} \tag{5.5}$$

$$S_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{5.6}$$

Where, x_i is the different rating value between the target user and neighbor on recommended movies. $\bar{x}$ is a mean value of x_i. n is the number of the recommended movies that both of target user and neighbor have been rated.

The hypothesis is

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

In this case, the Z-test is calculated by using 95% of confidence so α = 0.05. The graph is shown in Figure 5.1.

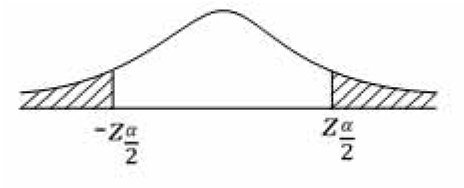


Figure 5.1: Z-Test graph.

Then, the Z value of Thai-Music and MoviePlanet is calculated by using equation 5.5

Which is $Z = 3.7020$

After that, find the $Z_{\alpha/2}$ and $-Z_{\alpha/2}$

$Z_{\alpha/2} = Z_{0.025}$ which is 1.95996, and

$-Z_{\alpha/2} = -Z_{0.025}$ which is -1.95996

Finally, Compare value of Z with $Z_{\alpha/2}$ and $-Z_{\alpha/2}$

$1.95996 < 3.7020$

We reject H_0 and accepted H_A which means that the recommendation results from Thai-Music and Proposed Method (Movie Planet) are significant.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this thesis, an advanced content/collaborative hybrid movie recommender system has been proposed. This proposed method based on Weighted Multiple Criteria and Multidimensional user profile.

6.1 Conclusion

Instead of weighting only biggest value of the features, the proposed method weights all components with its weight value and also use the frequency of the selection movie features to increase the weight of Multiple Criteria. For the varies of user's preference and always have Multiple criteria, this proposed system represent the users' opinion using both User Preference vector (UPV) which is a vector that express many features of User Preference and Selection on Movie Feature (SMV) which is a vector that express the frequency of selection movie feature that user always think about when he/she searches/selects for any movie. In other word, it can overcome unsuitable weight problem which is in the Thai-Music system. To concentrate on the Multidimensional user profile, the contextual information of user is needed such as where he saw the movie, when the movie was seen and with whom. Moreover, it can incorporate the contextual information without the losing a huge of data problem by using Multiple Linear Regression to perform the Multidimensional user profile instead of using the Reduction-based approach.

For evaluating the proposed method, a movie Recommender System called MoviePlanet has been created. As presented in the experimental evaluation, the combining of Multiple Criteria and Multidimensional user profile system provides more accurate recommendation results than the Hybrid System based on current Multiple Criteria Rating method.

6.2 Future Work

There is a few researcher that concern about contextual information of user's characteristic. As the experiments and evaluations shown in chapter 5, the Multidimensional user profile becomes attractively to use in recommender system. This thesis uses the Multiple Linear Regression to do the Multidimensional user profile part instead of Reduction-Based approach so there might be a better approach to work on Multidimensional user profile. Moreover, this thesis not yet concerns about the weight of contextual information of user's characteristic so this might be future work.

REFERENCES

[1] Shani, G., Heckerman, D. and Brafman, R. I. 2005. An MDP-Based Recommender System. The Journal of Machine Learning Research : 1265-1295.

[2] Adda, M. Valtchev, P., Missaoui, R. and Djeraba, C. (2007). Toward Recommendation Based on Ontology-Powered Web-Usage Mining. Internet Computing, IEEE : 45-52.

[3] Resnick, P., Iacovou, N., Suchak, M. and Bergstrom, P. and Riedl, J. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. Proceeding of 1994 Conference on Computer Supported Collaborative Work : 175-186.

[4] Good, N., Schafer, J. B., Konstan, J. A., Borchers, A., Sarwar, B., Herlocker, J. and Riedl, J., Combining Collaborative Filtering with Personal Agents for Better Recommendations. Proceedings of the 1999 Conference of the American Association of Artificial Intelligence (AAAI-99) : 439-446.

[5] Shardanand, U. and Maes, P. (1995). Social Information Filtering: Algorithms for Automating Word of Mouth. Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems : 210-217.

[6] Breese, J., Heckerman, D. and Kadie, C., Empirical Analysis of Predictive Algorithms for Collaborative Filtering. Proceedings of UAI : 43-52.

[7] Stolze, M. and Strobel, M., (2001) Utility-Based Decision Tree optimization: A Framework for Adaptive Interviewing. Proceedings of User Modeling 2001 : 105-116.

[8] Flemming, M. and Cohen, R. 1999, User Modeling in the Design of Interactive Interface Agents. Proceedings of User Modeling 1999 : 67-76.

[9] Balabanovic, M. and Shoham, Y., (1997). Fab: Content-Based, Collaborative Recommendation. Communications of the ACM : Vol. 40, No. 3, 66-72.

[10] Flemming, M. and Cohen, R. 1999, User Modeling in the Design of Interactive Interface Agents. Proceedings of User Modeling 1999 : 67-76.

[11] Recommendz: The Movie Recommender. Website[online]. (n.d.). Available from: http://recommendz.mcgill.edu:8080/protorype/.

[12] Kamba, T., bharat, K. and Albers, M. C. (1995). The Krakatao Chronicle: An Interactive Personalized Newspaper on the Web. International World Wide Web Conference.

[13] Balabanovic, M. and Shoham, Y., (1997). Fab: Content-Based, Collaborative Recommendation. Communications of the ACM : Vol. 40, No. 3, 66-72.

[14] Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. and Sartin, M. (2009). Combing Content-Based and Collaborative Filters on an Online Newspaper. Proc. Recommender System Workshop at ACM SIGIR 1999.

[15] Claypool, M., Le, P., Wased, M. and Brown, D. (2001). Implicit interest indicators. Proceedings of the International Conference on Intelligent User Interfaces : 33-40.

[16] Terveen, L., Hill, W., Amento, B., McDonal, D. and Creter, J. (1997). PHOAKS: A System for Sharing Recommendations. Communications of the ACM : Vol. 40, No. 3, 59-62.

[17] Salton, G. and McGill, M. J. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, Inc.

[18] Belkin, N. J. and Croft, B. W. (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin? CACM. : Vol. 35, No. 2.

[19] Pazzani, M. (1999). A framework for collaborative, content-based and demographic filtering. Artificial Intelligence Review : 393-408.

[20] Lang, K. Newsweeder: Learning to filter news. Proceedings of the 12th International Conference on Machine Learning : 331-339.

[21] Maes, P. (1995). Agents that Reduce Work and Information Overload. Readings in Human-Computer Interaction. Toward the Year 2000, Morgan Kauffman.

[22] Krulwich, B. and Burkey, C. (1996). Learning User Information Interests through Extraction of Semantically Significant Phrase. Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access : 110-113.

[23] Lieberman, H. (1997). Autonomous Interface Agents. Proceedings of ACM CHI 97 : 67-74.

[24] Cohen, W. (1996). Learning Rules that Classify E-mail. Proceeding of the AAAI Spring Symposium on Machine Learning in Information Access.

[25] Boone, G. (1998). Concept Features in Re: Agent, an Intelligent Email Agent. The Second International Conference on Autonomous Agents (ACM) : 141-148.

[26] Pandora Music Genome Project. Website[online]. (n.d.). Avalable from: http://www.pandora.com/corporate/mgp. [2007, July 8].

[27] Dias, R. and Junior, T. (2004). Combining Collaborative and Content-based Filtering to Recommend Research Papers. Master's thesis, University of Minnesota.

[28] Herlocker, J., Konstant, J., Borchers, A. and Rield, J. (1999). An algorithmic framework for performing collaborative filtering. Proceeding of the SIGIR ACM.

[29] Kitchens, F.L. and Rebman, C. (2007). Evaluating the Big Five Personality Test as a Foundation for an Online Student Course Recommendation System. Proceedings of the 2007 Decision Sciences Conference : 287-294.

[30] Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-adapted Interaction : Vol. 12, No. 4, 331-370.

[31] Saranya, M. (2007). High-quality Neighbor Formation for Music Recommender System. IMECS: 805-811.

[32] Gediminas, A., Ramesh, S., Shahana, S. and Alexander, T. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. ACM Transactions on Information Systems (TOIS): 103-145.

[33] Basu, C., Hirsh, H. and Cohen, W. (1998). Recommendation as Classification: Using Social and Content-Based Information in Recommendation. Proceedings of the American Association for Artificial Intelligence : 714-720.

[34] EFFROYMSON, M.A. (1960). Multiple regression analysis. Mathematical Methods for Digital Computers: 191-203.

[35] MedCalc Organizations. 1993. ROC curve analysis. Website[online]. (n.d.). Available from: http://www.medcalc.be/manual/mpage06-13a.php

[36] z-test. Website[online]. (n.d.). Available from:
http://www.geocities.com/chalong_sri/1_z.htm [2009, March 1].

VITAE

Keittima Chapphannarungsri was born in 1983. I received Bachelor Degree in Science (Majoring Computer Science) from Thammasart University in 2006 and Bachelor Degree in Business (Majoring Management) from Sukhothai Thammathirat Open University in 2007. I am working as Senior Application Developer for Wealth Management System Limited, Bangkok and I am also pursuing a Masters degree in Computer Science.