*J. Basic. Appl. Sci. Res.*, **3(7)258-273, 2013**
**© 2013, TextRoad Publication**

ISSN 2090-4304
**Journal of Basic and Applied Scientific Research**
*www.textroad.com*

# New Recommendation System Model Based on Semantic Similarity in Movie Domain

**Amir Hossein Nabizadeh Rafsanjani[1], Naomie Salim, Nastaran Mohammadhossein, Karamollah Bagheri Fard**

Faculty of Computing, Universiti Teknologi Malaysia (UTM), Skudai, 81310 Johor, Malaysia

## ABSTRACT

Recommender systems (RS) automatically select the most appropriate items to each user, thus shortening his product searching time and adapting the selection as his particular preferences evolve over time. It has been found that different recommendation methods use different techniques to recommend objects to consumers. There is possibility that the majority of the previous researches have problems. Currently, the most important problem in recommendation system is cold start that during this study, we consider to this issue by proposing RECOMOVIER model. Four techniques have been developed for designing new model including: Collaborative Filtering (CF), Content Base Filtering (CBF), and Hybrid and Cascade method. The main goal of this research is omitting the cold start problem for new user that helps to increase the accuracy of results of recommendation system. Analytical review of existing recommendation system model is utilized as research methodology which guides us to understand more about pros and cons of current recommender methods and systems. In order to implement the proposed model the C# programming language and two sets of data that were downloaded and modified from Movielens website were used. This application is tested by users as every user filled the form and then got feedback from this application. The results of experiments show that, using user profile has positive influence on increasing the accuracy of results of recommendation system by omitting the cold start problem from RS systems. The proposed model in this paper would be valuable and beneficial for future researchers and practitioners interested in developing recommendation systems.

**KEYWORDS**: Recommendation Systems (RS), Collaborative Filtering (CF), Content Base Filtering (CBF), Hybrid

## 1. INTRODUCTION

Discovering products that meet the consumers' requirement are crucial in such competitive environments as online shopping. Recommender systems assist in advertising tasks by automatically selecting the most appropriate items for each user as per his/ her personal interests and preferences (Adomavicius & Tuzhilin, 2005). Research in recommender systems started back in the early 1990s, but the greatest advances have been due to the irruption of recent technologies like those of the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001). It has been proved that semantics-based recommender systems can outperform previous approaches by exploiting two main elements:

- A knowledge base, typically an ontology, that represents semantic features or attributes of the available items.
- Filtering strategies based on semantic reasoning techniques that discover relevant relationships between the users' preferences and the items to be recommended (Blanco-Fernandez, Lopez-Nores, Pazos-Arias, Gil-Solla, & Ramos-Cabrer, 2010; Blanco-Fernández, López-Nores, Pazos-Arias, Gil-Solla, & Ramos-Cabrer, 2008; Middleton, Shadbolt, & De Roure, 2004; Pazos-Arias et al., 2008).

Obviously, keeping the users' satisfaction high requires means to adapt the selection of items as their interests evolve over time. For many years, in most of the existing filtering strategies, data collection about the users' interests was regarded as a static process, weighing equally the ratings given by the users at different times. Later, some researchers proposed time-aware approaches that made the last observations more significant than the older ones, which means assuming that a user's interest in a product always decreases from the moment of the last purchase (Ding & Li, 2005; Lee, Park, & Park, 2009; Liu & Shih, 2005). This may be true in certain areas of application, such as personalized programming guides that recommend TV programs to the users. Notwithstanding, the interest in (or the need for) commercial products in general may actually increase or vary in diverse forms over time. For example, if a user has just bought a dishwasher, it is foreseeable that he/she will not need another one until the average lifetime of such appliances has passed; therefore, the interest estimations should follow an increasing function, and any recommender system should prioritize other products for some time. Likewise, the interest for seasonal clothes may vary along the year, while the interest in books and music may remain constant and school

***Corresponding Author:** Amir Hossein Nabizadeh Rafsanjani, Faculty of Computing , Universiti Teknologi Malaysia (UTM), Skudai, 81310 Johor, Malaysia. Cell phone : (0060)1112149015 Email: Hnzramir2@live.utm.my

equipment may have a peak at the beginning of the academic year(Blanco-Fernández, López-Nores, Pazos-Arias, & García-Duque, 2011).

The main goal of this research is omitting the new user cold start problem that helps to enhance the accuracy of results of recommendation system. "Cold start" means when recommendation systems cannot predict user or customer precedence in item choice because of lack of adequate information. At the first time that a customer chose and visit a site, a recommendation system cannot rate any of the items. Therefore, the recommendation systems cannot identify customers' things that they are fond or not fond.

The main scientific contributions of this paper are

- ✓ Omitting the cold start for new user.
- ✓ Increasing the accuracy of results of RS in movie domain.

The paper is organized as follows: Section 2 includes a review of recommender systems literature to highlight the differences among them and in our new filtering method. Next, Section 3 illustrates the research methodology which is utilized for this study. Section 4 details the main parts of our personalization model, while in Section 5 we focus on the results and discussion. Section 6 presents the conclusion of the current research. Finally, Section 7 and 8 provides information about limitation of current study and Acknowledgment, respectively.

## 2. LITERATURE REVIEW

Research in recommender systems is hectic nowadays, in an attempt to address the many new questions raised by the growing number of practical applications. Next, we provide an overview of types of recommenders, and thereafter focus on the issues of proposing a new model for recommender to decrease the issues and increase the accuracy.

Recommender systems have become an important research area since the emergence of the first research paper on collaborative filtering in the mid-1990s (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994; Shardanand & Maes, 1995). In general, recommender systems directly help users to find content, products, or services (such as books, digital products, movies, music, TV programs, and web sites) by aggregating and analyzing suggestions from other users, which mean reviews from various authorities, and users (Frias-Martinez, Chen, & Liu, 2009; Frias-Martinez, Magoulas, Chen, & Macredie, 2006; Kim, Ji, Ha, & Jo, 2010). These systems use analytic technology to compute the probability that a user will purchase one of the products at each place, so that users will receive recommendations for the right products to purchase.

Recommender systems are generally classified into Collaborative Filtering (CF), Content-Based filtering (CBF) and Hybrid method. In general, CF uses an information filtering technique based on the user's previous evaluation of items or history of previous purchases.

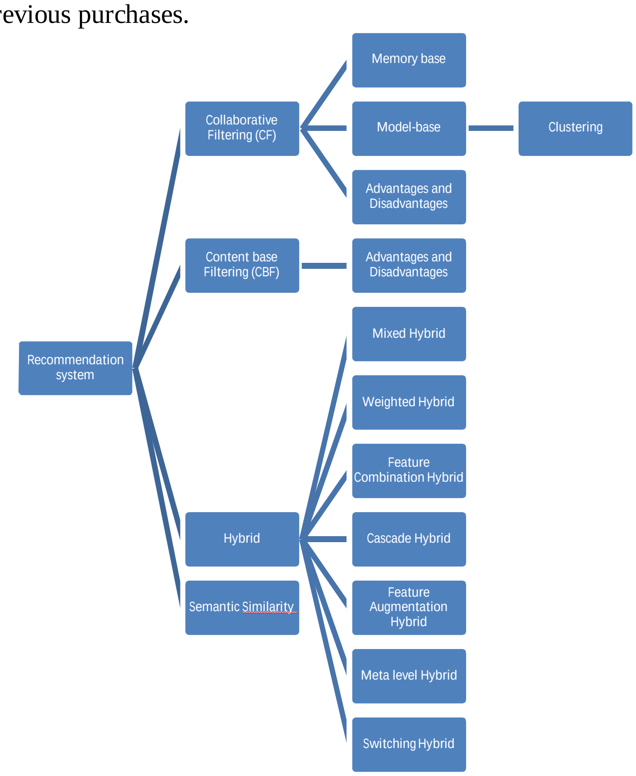


**Fig 1: Literature review**

### a. Collaborative Filtering (CF)

Collaborative Filtering technique is the procedure of evaluating or filtering items utilizing the opinions of other people. Despite the term that collaborative filtering method (CF) has only been around for a little more than a decade, Collaborative Filtering takes its basics from something humans have been doing for many years sharing ideas with others (Schafer, Frankowski, Herlocker, & Sen, 2007).

#### i. Memory-based Collaborative Filtering Techniques

Memory-based CF algorithms use the entire or a sample of the user-item database to generate a prediction. Every user is part of a group of people with similar interests. By identifying the so-called neighbors of a new user (or active user), a prediction of preferences on new items for him or her can be produced(Su & Khoshgoftaar, 2009). The neighborhood-based CF algorithm, a current memory-based CF algorithm, uses the following steps: calculate the similarity or weight, wi,j , which reflects distance, weight, between two users or two items, I and j; produce a prediction for the active user by taking the weighted average of all the ratings of the user or item on a certain item or user, or using a simple weighted average (Sarwar, Karypis, Konstan, & Riedl, 2001). When the task is to generate a top-N recommendation, we need to find k most similar users or items (nearest neighbors) after computing the similarities, then aggregate the neighbors to get the top-N most frequent items as the recommendation (Su & Khoshgoftaar, 2009).

#### ii. Model-Based Collaborative Filtering Techniques

The designing and development of models (such as machine learning, data mining algorithms) can allow the system to learn to recognize complex patterns based on the training data, and then make intelligent predictions for the collaborative filtering tasks for test data or real-world data, based on the learned models. Model-based CF algorithms, such as Bayesian models, clustering models, have been investigated to solve the shortcomings of memory-based CF algorithms (Basu, Hirsh, & Cohen, 1998; Breese, Heckerman, & Kadie, 1998; Su & Khoshgoftaar, 2009).

##### 1. Clustering Model

Clustering algorithms have been used to quickly locate a user's neighbors (Linden, Smith, & York, 2003). In these schemes, a user is compared to groups of users, rather than individual users. Clusters of users similar to the target are quickly discovered, and nearest neighbors can be selected from the most similar clusters. Both k-means clustering (MacQueen, 1967), and hierarchical divisive (Johnson, 1967) and agglomerative clustering (Lam & Riedl, 2004) can segment users into clusters.

One challenge in using clustering is that clustering schemes use distance functions, such as Pearson correlation, to both form the clusters and measure distance from a cluster. However, due to missing data, distance functions generally do not obey the triangle equality and are not true mathematical metrics. This can lead to unintuitive and unstable clustering (Schafer, et al., 2007).

#### iii. Advantages and Disadvantages of Collaborative Filtering Method (CF)

Collaborative Filtering (CF) algorithms has several pros, like capability for taking an object/item quality or defect into an account when suggesting objects/items, particularly in explicit customer rankings. For example, a local music band could fall into the same genre of music a rock band that is famous in all over the world, but this item does not assurance which they have same level of quality. This subject demonstrates that objects/items identification quality is obvious pros of Collaborative Filtering (CF). Collaborative Filtering (CF) can hinder deficient suggestions and recommendation by taking the precedence of customers which are actual into an account. Second pros is which the Collaborative Filtering (CF) algorithms are particularly applicable and useful in domains where the analysis of content is very expensive or difficult, like music and film suggestion, without demanding any domain of knowledge (Burke, 2002).

Although the Collaborative Filtering (CF) algorithms has several pros and the quality level of Collaborative Filtering (CF) algorithms improve during the time, but the most important problem is the phase of startup in recommendation system, as there are many objects and items are provided in the system while there are few customers and few or no rankings. This problem named "**cold start**" and means that recommendation system cannot produce any suggestion or recommendations (Schein, Popescul, Ungar, & Pennock, 2002). Remedies for solving this problem involve seeding the system by utilization other data sets, and using algorithms of recommendation system that are different in startup phase which do not suffer from "cold start" problem. Even after obtaining more ranking from customers, scantiness of the customer-object matrix can still be a problem for Collaborative Filtering (CF) (Schein, et al., 2002).

Second problem named "**gray sheep**" with regarding to Claypool et al., that is a description about the hardship of recommendation system for people who are not belong to the part of an obvious group (Claypool et al., 1999). Collaborative Filtering (CF) is useful and works very well for customers and users who are fit into a particular group with a lot of neighbors that are similar (Burke, 1999).

**Scalability** is the next challenge of CF. When the number of objects and customer increases, the traditional form of Collaborative Filtering (CF) suffers critical from scalability problem. For instance, with an enormous population of customers and also a big number of objects and items, then the intricacy of Collaborative Filtering (CF) will increased. At this time, we need many systems to response urgently for online demands that we require a higher level of scalability of a Collaborative Filtering (CF) (Burke, 2002).

Another challenge that Collaborative Filtering (CF) is faced is **synonymy**. This problem related to the inclination of numerous of very similar objects to have distinctive names. Recommendation systems usually are not capable to find this problem then faced with these objects differently. For instance, "adult automobile" and "adult car" are different statements but both of them allude to the similar object. In fact, the performance of Collaborative Filtering (CF) will decrease by propagation of synonyms (Burke, 2002).

**Shilling Attacks** can be another challenge for recommendation systems. It means when every item or object can be ranked by every customer, in comparison with other objects that belonging to other people, customers maybe give higher rank to own objects and items or even give negative rate to competitors' products. That's why in many cases, Collaborative Filtering (CF) systems must establish safety measure to dissuade customers and users from Shilling attacks (Burke, 2002).

### b. Content Base Filtering (CBF)

Content base recommendation system recommends an item to a user based upon a description of the item and a profile of the user's interests (Semeraro, 2010).

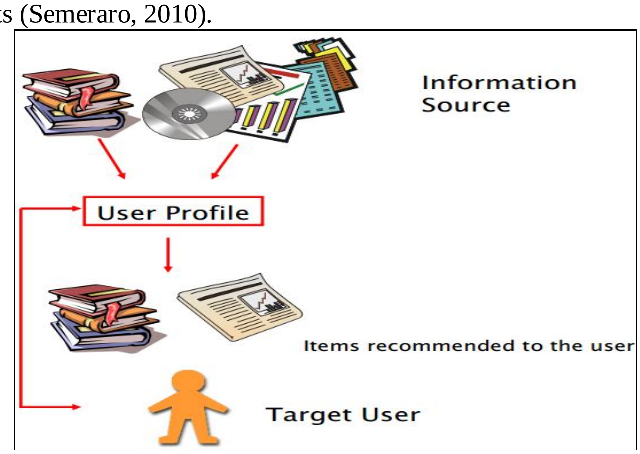


**Fig 2:Content-base filtering (Semeraro, 2010)**

#### i. Advantages and Disadvantages of Content Base Filtering (CBF)

One of the most obvious advantages of content-based filtering algorithms is these algorithms do not need to domain of knowledge. It is adequate to gather feedback from customers about their precedence (Rashid et al., 2002).

Next advantage of content-based filtering algorithms that we can consider to it is, these algorithms are better than Collaborative Filtering (CF) at finding locally similar objects. Because the explicit focus of content-based filtering algorithms is on similarity of text. However, this item can be a defect in domains where analysis of content in large number is impractical, impossible or difficult, like music and movies. The tendency of algorithms of content-based filtering is get stuck in a "well of similarity" (Rashid, et al., 2002) , where they suggest objects only from a restrict theme scope. Then the recommendations that are serendipitous can be very difficult to achieve.

### c. Hybrid Method

Hybrid recommendation systems are adjusted to join Content-based and Collaborative Filtering (CF) that control by one framework, to increase the benefits and to decrease the weaknesses of both techniques.

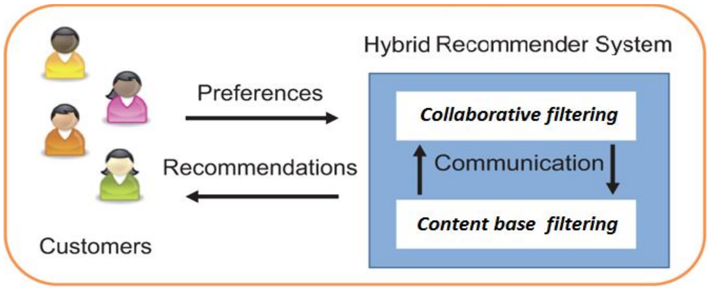


**Fig 3 : Hybrid recommendation system (Rodríguez, Espinilla, Sánchez, & Martínez-López, 2010).**

Following table provides a brief explanation of all hybridization methods that are:

| Hybridization method | Description |
|---|---|
| **Mixed Hybrid Recommender** | This method point to the suggestions and recommendations which are recommended from a set of various recommendation systems, are presented simultaneously. |
| **Weighted Hybrid Recommender** | Production a single recommendation by utilization of the votes and rates that are produced by some recommendation approaches. |
| **Feature Combination Hybrid Recommender** | The characteristics which are relate to various recommendation data resources are get together into a single recommendation system algorithm. |
| **Cascade Hybrid Recommender** | One of the recommendation systems purify the suggestions and recommendations that are presented by another recommendation system. |
| **Feature Augmentation Hybrid Recommender** | The results from one approach are utilized as input data and characteristics for another recommendation method. |
| **Meta level Hybrid Recommender** | The approach that is learned by one recommendation system is utilized as a input for another approach. |
| **Switching Hybrid Recommender** | In this method, recommendation system switches among recommendation approaches according to the current situation. |

**Table 1 : Summarization of Hybridization method**

### d. Semantic Recommender System

Recommender systems that use Semantic similarity are described by the inclusion of semantic knowledge in their processes to improve suggestion's quality. The majority of those utilize a concept based method to enhance the user profile representation (user modeling stage), and to use standard vocabularies and ontology languages (OWL) in standardized form (Victor CODINAa, 2010).

Regardless of the kind of system, we have understood that a usual feature in most semantic recommendation systems is utilization of profiles to depict the users´ information requirements and precedence. Hence, customer profiles have become a main part of effective filtering in recommendation systems, since an insufficient profile may cause cheap quality and impertinent user suggestions (Peis, del Castillo, & Delgado-López, 2008).

We will regard semantic recommendation systems as any system which bases is knowledge base, usually explained through conceptual maps (like a taxonomy or thesaurus) or an ontology, and that utilize technologies from the Semantic (Peis, et al., 2008).

Utilizing semantic similarity in recommendation systems limits and decreases specific problems, involving the following:

- To guarantee the inter operability of system resources and the similarity of the representation of information (Peis, et al., 2008).
- To allow for the dynamic contextualization of user preferences in specific domains.
- To facilitate performance in collaborative filtering (CF).
- To improve communication processes between recommendation systems and between recommendation systems and users.
- To omit or decrease the "cold start" problem by completing the incomplete information through inferences (Peis, et al., 2008).

## 3. RESEARCH METHODOLOGY

This research would direct based on the progress as illustrated in Figure 4. The operational framework is separated into several phases. The following subsections explain these steps.

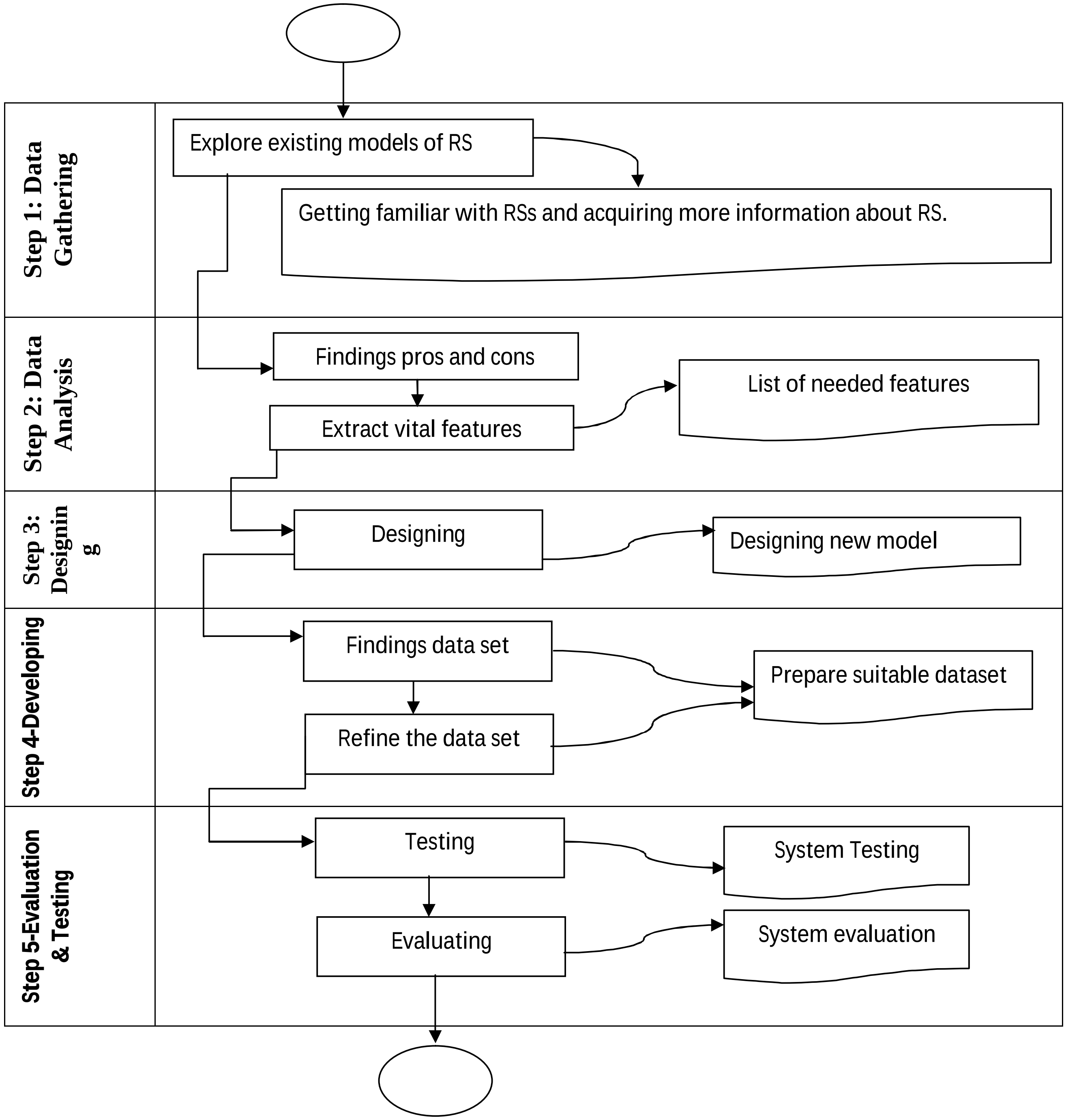


**Fig 4 : Operational Framework**

### a. Explore existing models

During Explore relevant existing models phase, investigating the existing recommendation system is the guideline to introduce the new model. Exploring current methods and solutions for recommendation systems, regardless of being successful or not, will help to understand what are the recommendation systems and most demanding qualities of an ideal recommendation system. Hence, reviewing the previous models selected as one of the research method in this phase. In this phase, a collection of all recommendation methods will select and gather. This collection generates a clear picture of current recommendation systems including their advantages and disadvantages. In this phase, more than 100 different articles and papers which worked on recommendation systems are explored and investigated.

**b. Findings weaknesses and strengths of previous study**

The number of researches and models that we focused on those in previous phase read with a critic point of view to understand their models accurately and find their weaknesses and strengths of each model. By critical viewpoint, we can find out what the common problems and weaknesses in recommendation system models are and try to find solutions for those by proposed model which can remove or decrease the severity.

**c. Features extraction**

Providing a set of all essential features, the most demanding qualities and properties in addition of deletion and omitting the problematic issues will create a clear picture of proposed recommendation system model. However, technical issues and considerations must be involved in the design of proposed recommendation model. Therefore, during in this phase we try to concentrate on important features that we need to decrease the weaknesses and increase the precision of recommendation system. Features and specifications needed for new systems are deliverable results of this phase.

**d. Designing new model**

This phase designs new model. During this phase, selected features from previous phase (feature extraction) are utilized to propose and design new recommendation system. During designing phase, we must consider the problems of recommendation system that we want to decrease or remove. During this phase, primary model of recommendation system will be designed. The result is a model that clearly described expectations of recommendation system. The result of this phase proposing new recommendation system model based on semantic similarity. Implemented model will explain how recommendation system works and how it supports the properties of ideal recommender system. Model supposed to response to system requirement and support selected properties of recommendation system.

**e. Findings & Refining data sets**

During this phase we want to determine our data set that we need to implement our model. MovieLens is a movie recommender project, which is a typical collaborative filtering (CF) system that collects movie preferences from customers and then groups users with same tastes. Two data sets are available at the MovieLens web site (http://movielens.umn.edu). The first one consists of 100,000 ratings for 1682 movies by 943 users. The second one includes of approximately 1 million ratings for 3883 movies by 6040 users.

The large data set consists in 3 text files, with tabular format, describing 1000209 anonymous ratings of 3883 movies made by 6040 MovieLens users who joined MovieLens in 2000. In the following, we describe the contents of each text file.

i. **Rating Dataset**

This Dataset includes data nearby 1000209 ratings in the format: UserID::MovieID::Rating::Timestamp where:

- UserID is an integer, ranging from 1 to 6040 which recognizes a user. Each user has rated at least 20 movies.
- MovieID is an integer, ranging from 1 to 3952 which recognizes a movie.
- Rating is an integer, ranging from 1 to 5, made on a 5-star scale.
- Timestamp is represented in seconds since 1/1/1970 UTC.

The structure of rating dataset is described in figure 5.

```
1::1193::5::978300760
1::661::3::978302109
1::914::3::978301968
1::3408::4::978300275
1::2355::5::978824291
```

**Fig 5: Overview of rating dataset**

Based on model and requirements, the rating data set must change and then after modification by omitting the Timestamp field, it changed to the new one

```
1::1193::5
1::661::3
1::914::3
1::3408::4
1::2355::5
```

**Fig 6 : Overview of modified Rating dataset**

### ii. Movie Dataset

This file contains data about 3883 movies (1 movie in each line) in the format: MovieID::Title::Genres ::Actor::Actress::Director::Company::Language where:

- MovieID is an integer, ranging from 1 to 3952, that identifies a movie.
- Title is a String that concatenates movie title and year of release (between brackets).
- Genres are a pipe-separated list of genres. Provided genres are: Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western.
- Actor who is the first actor in this movie.
- Actress who is the first actress in this movie.
- Director reveals the name of movie’s director.
- Company that shows the name of company that produce and distribute this movie.
- Language which shows the language of the movie.

The structure of movie dataset is described in figure 7.

```
1::Salt::2010::Action,Crime,Mystry::Live Schreiber::Angelina jolie::Phillip noyce::Columbia pictures::En
2:: Mr and Mrs Smith::2005::Romantic,Comedy,Action::Brad pitt:: Angelina jolie::doug liman::20th century fox::En
3::Seven pounds::2008::Drama:: Will smith::Rosario dawson::Gabriele muccino:: Columbia pictures::En
4::A separation::2011::Drama::Shahab hosseini::Leila hatami::Asghar farhadi::FilmIran::Farsi
```

**Figure 7 : Overview of movie dataset**

### f. Model Developing

During this phase has tried to develop the new model which is designed in previous phase based on features that were extracted. As it is mentioned before, the most important goal of this model is, decreasing the problem of previous recommendation system model like cold start, synonymy and etc while the precision of proposed model is better than models which have presented, before.

To develop this model, we use C# programming language. This model uses two data sets (movie data set, rating data set) and also gets some information from user at the beginning.

### g. Model testing

One of the most important phases in this study is testing. During testing, we use two sets of data which we find and refined them. The first data set is movie, and the second one is rating that every user gives to every movie. Movie data set has 3 fields: name of the movie, production year and also the genre of the movie. Second data set that is related to the rate of movie by users which has 3 fields: User ID, Movie ID and Rate of movie (1 to 5). Another information that is used in this model is users’ information that user enter to the system at the beginning. The workable model (application) is the main goal of this phase.

### h. System Evaluation

Evaluating the application is very important phase. Because of importance of limitations and properties of new model, and looking at practical aspects of development of model, the process of evaluation is designed in a practical running. To evaluate the application, this application runs by around 78 users. 78 users are selected since the results must be stable for analyzing. Users enter their information to the application and see the recommendations which this system gives them and they can check the results. The result of evaluation collected and analyzed to measure the level of success of proposed application. During evaluation the system, we increase the number of respondents time to time until our result become stable.

The following table shows the steps, actions and deliverables of each step to perform this research.

| Steps | Action | Deliverables |
|---|---|---|
| **1-Data Gathering (Literature review)** | • To explore existing models of recommendation systems<br>• To find weaknesses and strengths of different methods | • List of recommendation system methods and techniques<br>• Function of recommendation system<br>• List of common weaknesses and strengths of recommendation systems |
| **2-Data analysis (Requirement analysis)** | • Investigation to find the methods and systems in same area<br>• Finding out the weaknesses of this area | • List of features and specifications needed for new systems. |
| **3-Designing** | • To design the new model for recommendation system | • New model for recommendation system based on semantic similarity |
| **4-Developing** | • To develop the proposed model | • new recommendation system application by C# |
| **5-Evaluation and Testing** | • Implementation of model by two sets of data (movie, rating)<br>• Evaluations the model to measure the level f success. | • Movie datasets and rating data sets are used to implement the model<br>• Evaluate the model by around one hundred users |

**Table 2: The Applicable Method and Deliverables**

## 4. Model

The main goal of recommendation system is generating significant suggestions and recommendations information, products or objects for users' society that users could interest them. For instance, book recommendation on Amazon site, Netflix that recommend movies that use recommendation systems to identify users' tendencies and subsequently, attract users more and more (Melville & Sindhwani, 2010).

Based on the previous studies that are performed on same area, the proposed model is produced that it is shown in figure 8.

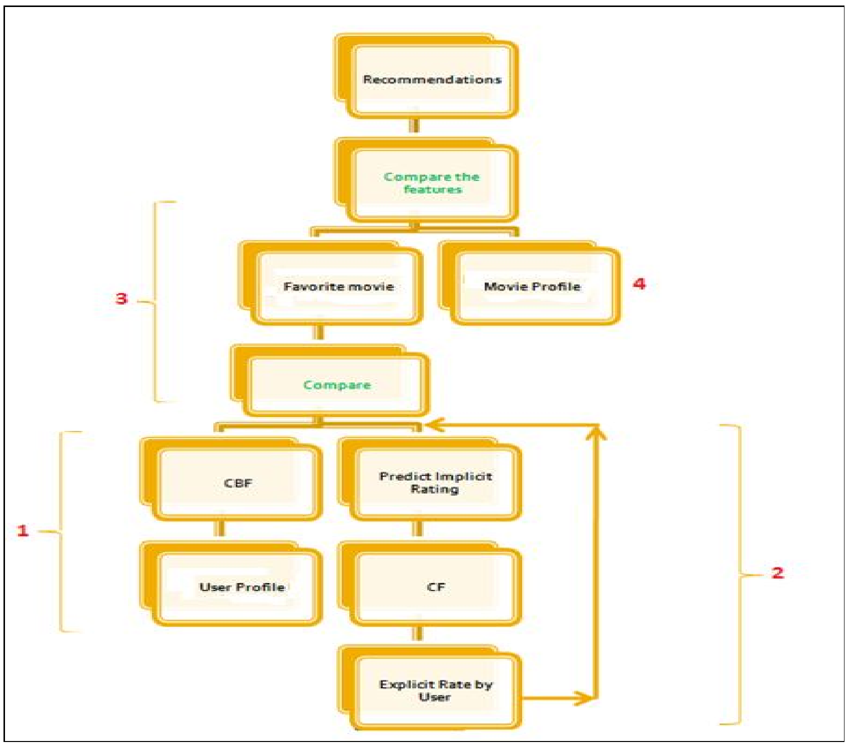


**Fig 8: Proposed model**

According to figure 3 this model divided into four sections. The first section uses semantic similarity for modeling user profile by semantic similarity and Content base filtering (CBF) which are illustrated in section 2.4 and 2.2 respectively during chapter 2 (Literature review). User profile is modeled in terms of our objectives and also our requirements. For this study that our data set is movie, we must design a profile that could meet users' needs. Therefore based on essential requirements, the user profile is designed which has presented in figure 9.

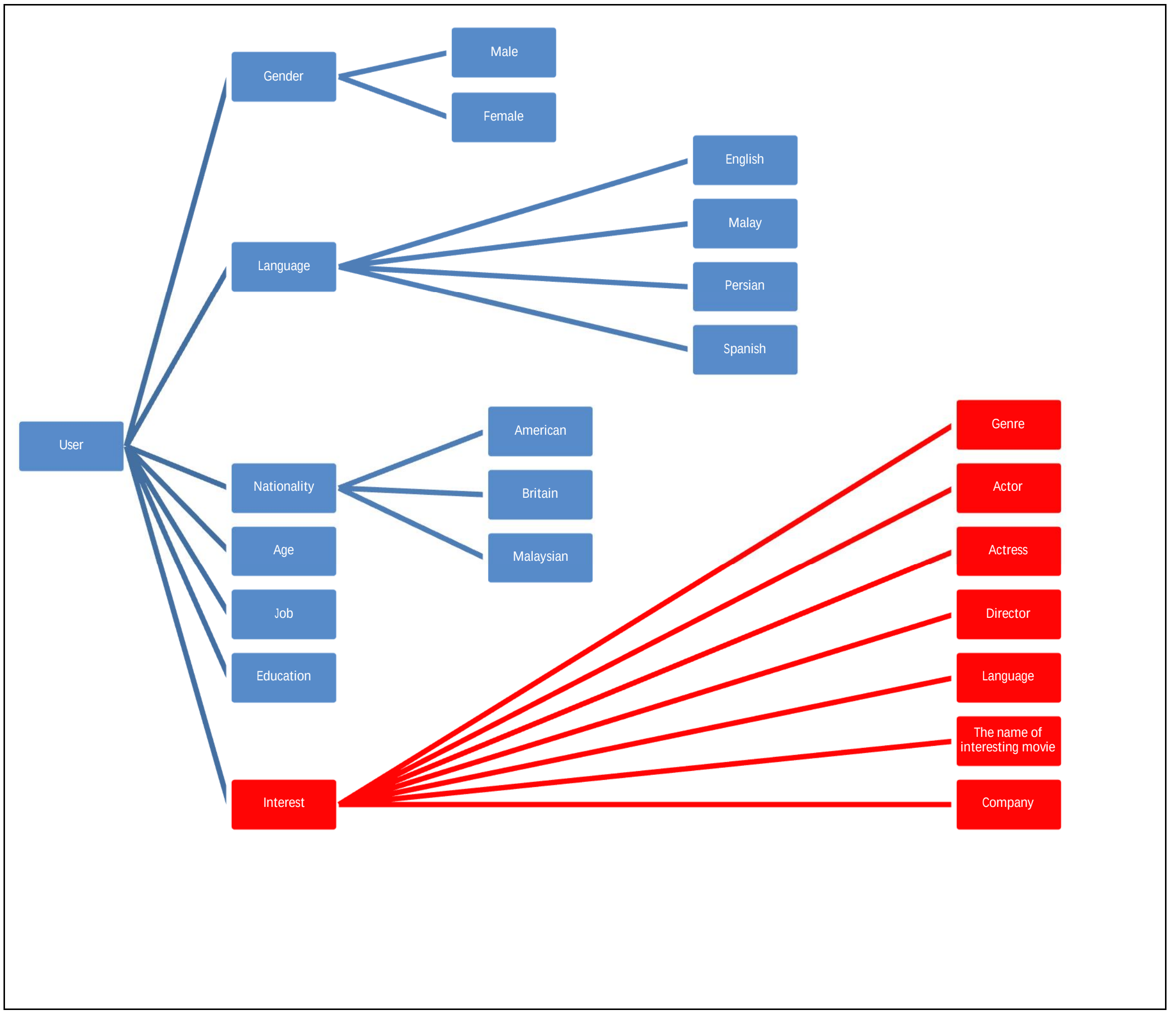


**Fig 9: User profile**

The most important section in user profile is user interest. In this part, the user enter his interests item that are related to the movies like Genre, Actor, Actress, Director, Language, Company of movie and etc.

Now, we run Content-Base Filtering (CBF) on user profile. Content-Base Filtering (CBF) suggests and recommends objects and information which are comparable in content to objects that the users have interested previously, or compared and matched to the users' characteristics. Another name of Content-Base Filtering (CBF) is search-base method means this method only searching for items which are similar to user preferences in terms of content. Therefore, the output of this step is a list of movies that are similar to user feature preferences that we model all these movies based on semantic similarity.

Second part use Implicit and explicit rating and also collaborative filtering (CF) that are explained during section 2.5 and 2.1 in chapter two. As it is shown in figure 8, the second step begins with explicit rate. Explicit means clearly expressed or readily observable by ranging the customer satisfaction by distribution and fulfillment the questionnaires or filling the form of precedence by customers. By running Collaborative Filtering (CF) on explicit rate we can extract the implicit rate of movies that user did not rate them, before.

At the end of this step we combine the result of implicit and explicit rate of user to movies and we have a list of movies that are rated by user. Now, we must list movies which have high rank among movies and suggest them as result of this step.

The result of section 1 and 2 are combined to use as input for third part then we use hybrid method (combination of CF and CBF methods). All the hybridization methods are described in section 2.3 during second chapter. For third stage based on figure 8, we must compare the result of first and second steps. The result of third stage in this model can be one or a list of movies with the highest ranks that user like them. At this time, we model all these movies based on semantic similarity. One of movie profile which is modeled based on semantic similarity has shown in figure 10.

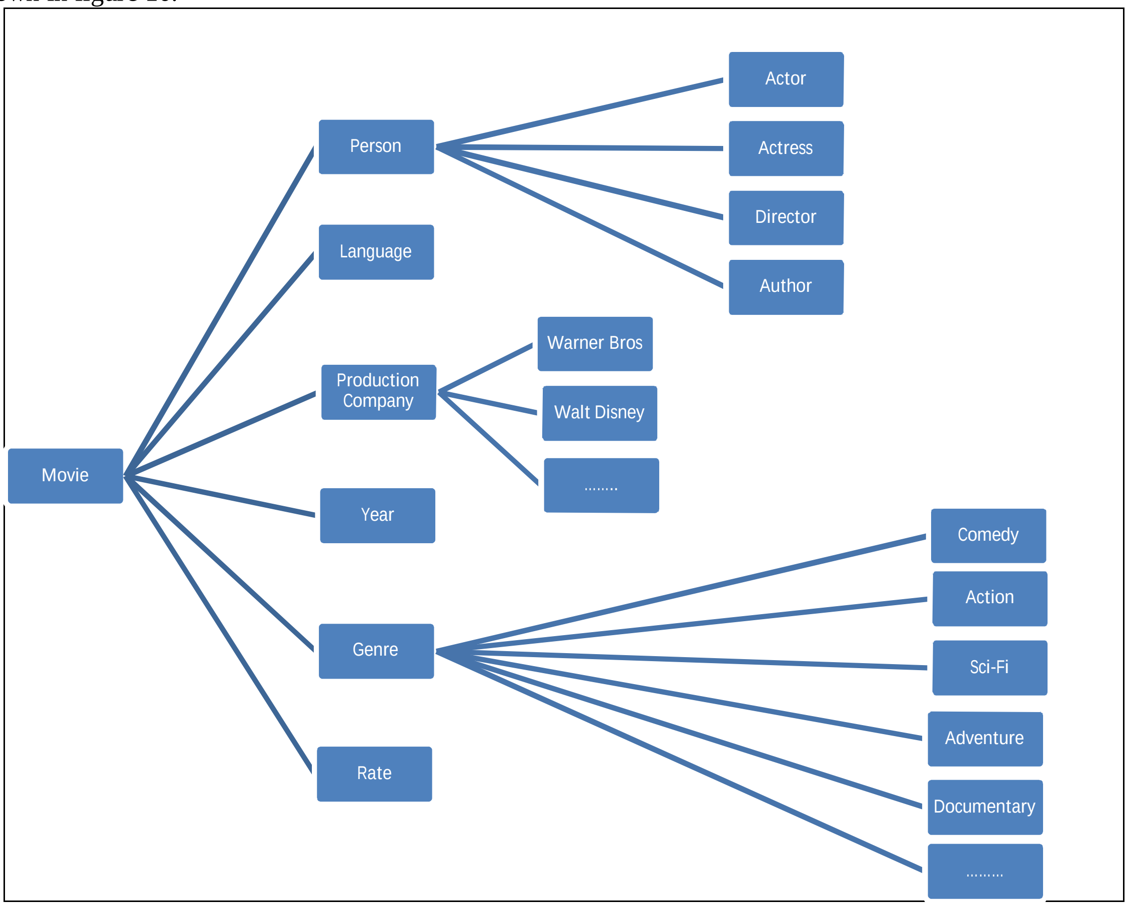


**Fig 10: Movie Profile**

In forth step, we use semantic similarity for modeling movie profile like first step. The output of first and second part is as input for third and forth part. It shows the results of first and second part are purified (figure 8) by third and forth part therefore this section utilizes Cascade hybrid method which is described in section 2.3 in detail during chapter two.

In fourth step of our model, we have to create a model from the movies that are located on our data set. This model is the same as movie profile that is presented in figure 10.

Therefore, in final operation in the proposed model of this study which is presented in figure 8, we must compare the result of third and forth stages in model. The last operation compares the features set of movies with the highest rank that are users' preferences with movies in data set that are modeled by semantic similarity. This operation creates a list of movies (five or less) that user like them. The manner of RECOMOVIER system to match and compare data is explained as follow.

As it has shown in table 3, new user inserts his information to the system. Filling the first five fields is optional while filling the others is compulsory. The cross symbol shows that this field did not fill by the user and tick symbol shows which this field is filled up by new user. At this time, content base filtering algorithm executes to extract the features that user insert to the system and match them with the similar features of movies in the movie data set.

| | Gender | Lan | Nationality | Age | Occupation | Genre | Actor | Actress | Director | Language | Co. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **User1** | Male | × | × | × | Student | Action | Pitt | Jolie | Spilberg | English | Warner Bros |

**Table 3 : Data entry into the system**

Table 4 shows how content base filtering works and compares the new user's inputs with similar fields from movie dataset. At this time, the content base algorithm (CB) executes on the information of user profile and tries to extract, compare and match the compulsory fields that user inserts to the system with the fields of data in movie data. During this process, CB algorithm checks the fields with the same name from movie data set and user profile and compare them with each other and if these fields have the similar context, CB extract the movie ID of this movie from movie dataset. During this phase, those ID of movies can extract that at least 3 fields of them are comparable with the fields of data that user inserts to the system (user profile).

| **User Information** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Gender** | **Lan** | **Nationality** | **Age** | **Occupation** | **Genre** | **Actor** | **Actress** | **Director** | **Language** | **Co.** |
| **User1** | **Male** | × | × | × | **Student** | **Action** | **Pitt** | **Jolie** | **Spilberg** | **English** | **Warner Bros** |

| **Data in movie Data set** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Movie ID** | Name of movie | Year | Genre | Actor | Actress | Director | Company | Language |
| **1** | Salt | 2010 | Action,Crime,Mistery | Liev Schreiber | Angelina Jolie | Phillip Noyce | Columbia Pictures | English |
| **2** | Mr and Mrs Smith | 2005 | Romantic,Comedy,Action | Brad pitt | Angelina Jolie | Doug Liman | 20th Century Fox | English |
| **3** | Seven pounds | 2008 | Drama | Will Smith | Rosario Dawson | Gabriele Muccino | Columbia Pictures | English |
| **4** | Troy | 2004 | Adventure,Drama,History | Brad pitt | Diane Kruger | Wolfgang Petersen | Warner Bros. Pictures | English |
| **5** | 3 Idiots | 2009 | Comedy,Drama, Romance | Aamir Khan | Kareena Kapoor | Rajkumar Hirani | Vinod Chopra Productions | Hindi |
| **6** | A separation | 2011 | Drama | Shahab Hosseini | Leila Hatami | Asghar Farhadi | FilmIran | Farsi |
| **7** | Inception | 2010 | Action,Adventure,Mystery | Leonardo DiCaprio | Ellen Page | Christopher Nolan | Warner Bros. Pictures | English |
| **8** | Source code | 2011 | Mystery,Sci-Fi,Thriller | Jake Gyllenhaal | Michelle Monaghan | Duncan Jones | Summit Entertainment | English |

**Table 4 : Matching user data with data in movie data set**

According to the table 4, among these 8 movies, the ID from four movies has extracted which are Salt , Mr and Mrs Smith ,Troy and Inception because in these four movies there are at least three fields which their context is similar to the new users' fields which are inserted into system.

Concurrently, there is a rating data set that includes thousands number of movies with their ID, users' rank (in scale of 5) and also the ID of users who rate to these movies. At this time, we must search for those users who have rated to the similar movies which are extracted from the previous step (movies that new user is interested to them) through the rating dataset. These kind of users must rate to at least one of extracted movies, then these users can be as neighbors for new user (a user who inserted his information into the system). Therefore, we take out the rate of each movie in terms of neighbors' rating and then this system calculates the average of rating based on neighbors' rating for each movie and match every ratings to the its movie which is extracted from previous phase. Table 5 provides more information about this process. In table 5, columns show movie ID and rows indicate user ID. For example, this table shows user 3 gave rank 4 to the movie 1 and user 501 gave rank 5 to the movie which its ID is 4 and etc. There are four movies which are took out from the previous step which are Salt (Movie ID:1), Mr and Mrs Smith (Movie ID:2),Troy (Movie ID:4) and Inception (Movie ID:7) then this system finds users who rate these movies (new user's neighbors) through rating dataset. Then, this system finds rank of each selected movie by averaging the rate of neighbors' rate. For instance, the rate of movie which its ID is 4 (Troy) is ((5+2+3+5)/4) = 3.75,thus the rate of troy is 3.75.

| | **Movie ID** | | | | | |
|---|---|---|---|---|---|---|
| **User ID`** | **1** | **7** | **6** | **4** | **2** | **3** |
| 1 | 2 | 4 | 5 | 5 | 4 | 3 |
| 1002 | | 1 | 3 | | | |
| 3 | 4 | | 4 | 2 | 3 | 3 |
| 98 | | 5 | 4 | 3 | | 4 |
| 501 | 2 | 2 | 2 | 5 | 2 | 2 |
| 6 | | 4 | 3 | | | |

**Table 5 : Rate of movies**

| **The movies which are extracted from movie data set that user is interested to them** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Movie ID** | **Name of movie** | **Year** | **Genre** | **Actor** | **Actress** | **Director** | **Company** | **Language** |
| **1** | Salt | 2010 | Action,Crime,Mistery | Liev Schreiber | Angelina Jolie | Phillip Noyce | Columbia Pictures | English |
| **2** | Mr and Mrs Smith | 2005 | Romantic,Comedy,Action | Brad pitt | Angelina Jolie | Doug Liman | 20th Century Fox | English |
| **4** | Troy | 2004 | Adventure,Drama,History | Brad pitt | Diane Kruger | Wolfgang Petersen | Warner Bros. Pictures | English |
| **7** | Inception | 2010 | Action,Adventure,Mystery | Leonardo DiCaprio | Ellen Page | Christopher Nolan | Warner Bros. Pictures | English |
| **Rating data set** | | | | | | | | |
| **Movie ID** | Rate | | | | | | | |
| **4** | **4** | | | | | | | |
| **7** | **5** | | | | | | | |
| **2** | **3** | | | | | | | |
| **1** | **3** | | | | | | | |

**Table 6 : Matching the movie ID with rate of movie**

As it is shown in table 6, we can find the rate of interesting movies based on movie ID. During this stage, probably the system finds many movies but this system must suggest the best one to the user, hence, we need to check movie IDs and their ranks to find a movie that its rank is the highest. For example in table 6, there are four movies in the movie list that user is interested to them but Inception has the highest rank among these movies. Thus, the system extracts the ID of this movie (movie with highest rank) and search again through movie dataset to find its details that are include: Movie ID, Name, Year, Genre, Actor, Actress, Director, and Language to suggest such information to the user.

Finally, new user after giving recommendation from application and checking them, assess this application by giving a rate between 1 to 5 which one shows the least and five shows the most user's satisfaction of system.

## 5. RESULT AND DISCUSSION

The result of system testing and evaluation is discussed during this section. C# programming language is utilized to develop and implement this system. As it mentioned before, firstly, user must insert his information to the system and after that he can get the feedback from this system in shape of several movies in details include movie id, name of movie, year, and genre. When users get feedback from the system, they must evaluate result by choosing a number between 1 and 5 which 1 shows the weakest acceptance of system by users and 5 indicates the most confirmation of the system that its interface has displayed in figure 11.

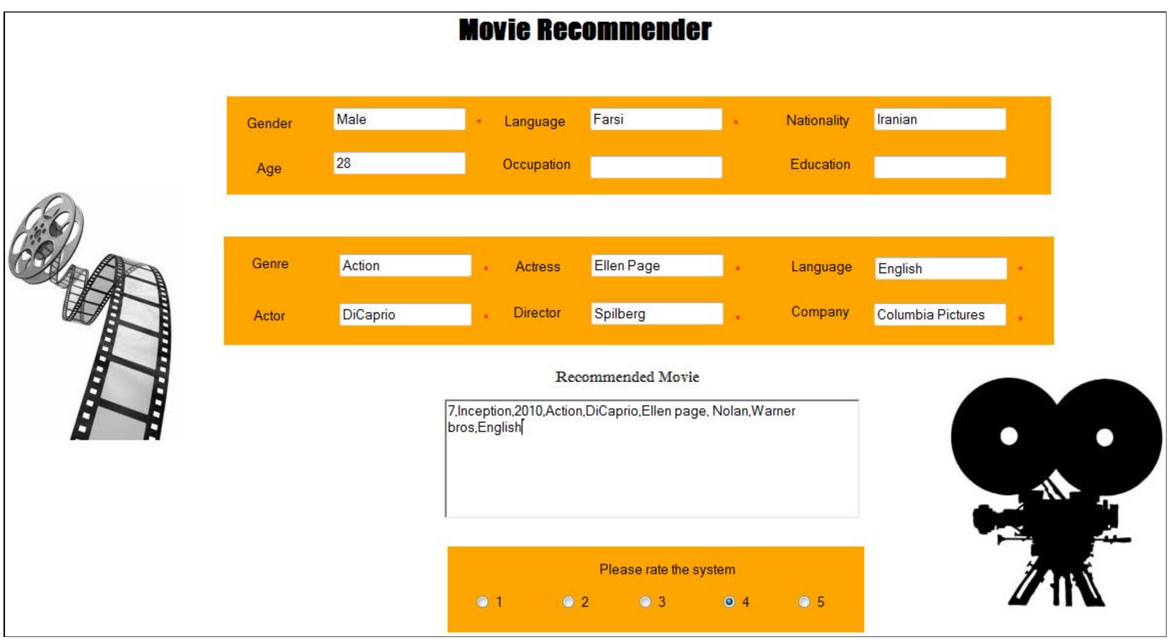


**Fig 11: System interface**

During system evaluation, 78 people are chosen randomly with distinct nationality such as Iranian, Malaysian, Chinese, Indonesian and etc to evaluate the system. Every user inserts his data into the system and gets the result from the system. As it shown in figure 12, our respondents are chosen from more than 10 nationalities to evaluate the system. For example, 15 users are chosen randomly from the Nigeria that each of them rated system based on scale of numbers between 1 and 5. During evaluation the system, we increase the number of respondents time to time until our result become stable.

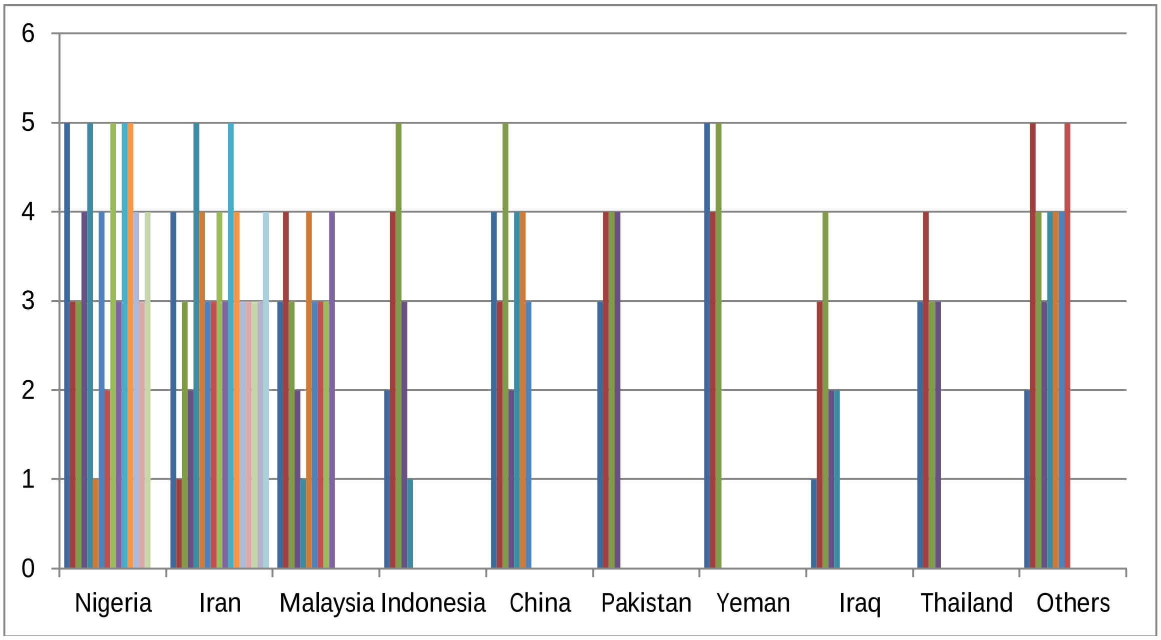


**Fig 12 : Respondents demographic**

Table 7 provides information about the analysis the result of the evaluation. This table shows how many people from each nationality participate in evaluation, and also the average of rank that every nationality gives to the system that finally we calculated the average of all users' ranks in the bottom of this column which is 3.5. The last column in this table shows the percentage of population which shows the percentage of each nationality that contributes during the system evaluation. According to the average of rank (3.5) that respondents gave to the system, the performance of the system is acceptable.

| No | Nationality | Count | Average of Rank(1-5) | Percentage of population |
|---|---|---|---|---|
| **1** | Nigeria | 15 | 4 | 19.2% |
| **2** | Iran | 17 | 3 | 21.9% |
| **3** | Malaysia | 10 | 3 | 12.8% |
| **4** | Indonesia | 5 | 3 | 6.4% |
| **5** | China | 7 | 4 | 9% |
| **6** | Pakistan | 4 | 4 | 5.1% |
| **7** | Yemen | 3 | 5 | 4% |
| **8** | Iraq | 5 | 2 | 6.4% |
| **9** | Thailand | 4 | 3 | 5.1% |
| **10** | Others | 8 | 4 | 10% |
| | | **Total=78** | **Total Ave: 3.5** | **Total =100%** |

**Table 7 : Evaluation analysis**

## 6. Conclusion

The main aim of recommendation system is creating significant suggestions and recommendations information, products or objects for users' society that users could interest in those. During this study, we proposed a model that has several phases which during these phases, user inserts several information into the system and receive number of recommendations from this system. The main goal of this research is omitting the cold start problem for new user which has direct effect on the preciseness of the result. Analytical review of existing recommendation system model is utilized as research methodology which guides us to understand more about pros and cons of current recommender methods and systems. According to the results of system evaluation, we found that the "cold start" problem was omitted from this system. Therefore, omitting "cold start" problem for new user from RS has direct positive effect on the accuracy of the result of RS.

Furthermore, this system Improves communication processes between recommendation systems and users. Previously, the majority of recommendation systems were generating and suggesting some recommendations to user automatically but during utilization of proposed model, user must insert some data to get some recommendations from the systems, therefore this model can increase such communications.

### 7. Limitations and Future Works

One of the most obvious limitations of this study is movie dataset. The movie dataset which proposed model needs must have 6 fields includes, person (cast), country, language, production country, award, year, genre and etc while our datasets does not have several required fields. Although, proposed model has acceptable operation by this data set but if we can find better dataset that has all required fields can be useful.

Because of shortages of required fields in our dataset, the results of model do not have such kind of accuracy that we had in our minds. However, the precious of results based on our dataset is convincing.

During this study semantic similarity use as basis for model. Now as future works, the others can apply other techniques like ontology to enhance the productivity of this model. Besides, they can apply a better movie data set to get more accurate recommendations from this model.

### 8. Acknowledgement

I would like to acknowledge my friends and participants who their guidance and participation helped to proceed the study.

## REFERENCES

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on, 17*(6), 734-749.

Basu, C., Hirsh, H., & Cohen, W. (1998). *Recommendation as classification: Using social and content-based information in recommendation.* Paper presented at the Proceedings of the national conference on artificial intelligence.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American, 284*(5), 28-37.

Blanco-Fernández, Y., López-Nores, M., Pazos-Arias, J. J., & García-Duque, J. (2011). An improvement for semantics-based recommender systems grounded on attaching temporal information to ontologies and user profiles. *Engineering Applications of Artificial Intelligence, 24*(8), 1385-1397.

Blanco-Fernandez, Y., Lopez-Nores, M., Pazos-Arias, J. J., Gil-Solla, A., & Ramos-Cabrer, M. (2010). Exploiting digital TV users' preferences in a tourism recommender system based on semantic reasoning. *Consumer Electronics, IEEE Transactions on, 56*(2), 904-912.

Blanco-Fernández, Y., López-Nores, M., Pazos-Arias, J. J., Gil-Solla, A., & Ramos-Cabrer, M. (2008). *Personalizing e-commerce by semantics-enhanced strategies and time-aware recommendations.* Paper presented at the Semantic Media Adaptation and Personalization, 2008. SMAP'08. Third International Workshop on.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). *Empirical analysis of predictive algorithms for collaborative filtering.* Paper presented at the Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence.

Burke, R. (1999). *Integrating knowledge-based and collaborative-filtering recommender systems.* Paper presented at the Proceedings of the Workshop on AI and Electronic Commerce.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction, 12*(4), 331-370.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999). *Combining content-based and collaborative filters in an online newspaper.* Paper presented at the Proceedings of ACM SIGIR Workshop on Recommender Systems.

Ding, Y., & Li, X. (2005). *Time weight collaborative filtering.* Paper presented at the Proceedings of the 14th ACM international conference on Information and knowledge management.

Frias-Martinez, E., Chen, S. Y., & Liu, X. (2009). Evaluation of a personalized digital library based on cognitive styles: Adaptivity vs. adaptability. *International Journal of Information Management, 29*(1), 48-56.

Frias-Martinez, E., Magoulas, G., Chen, S., & Macredie, R. (2006). Automated user modeling for personalized digital libraries. *International Journal of Information Management, 26*(3), 234-248.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241-254.

Kim, H. N., Ji, A. T., Ha, I., & Jo, G. S. (2010). Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. *Electronic Commerce Research and Applications, 9*(1), 73-83.

Lam, S. K., & Riedl, J. (2004). *Shilling recommender systems for fun and profit.* Paper presented at the Proceedings of the 13th international conference on World Wide Web.

Lee, T. Q., Park, Y., & Park, Y.-T. (2009). An empirical study on effectiveness of temporal information as implicit ratings. *Expert Systems with Applications, 36*(2), 1315-1321.

Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE, 7*(1), 76-80.

Liu, D.-R., & Shih, Y.-Y. (2005). Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. *Journal of Systems and Software, 77*(2), 181-191.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations.* Paper presented at the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability.

Melville, P., & Sindhwani, V. (2010). Recommender systems. *Encyclopedia of machine learning,* 829-838.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS), 22*(1), 54-88.

Pazos-Arias, J. J., López-Nores, M., García-Duque, J., Díaz-Redondo, R. P., Blanco-Fernández, Y., Ramos-Cabrer, M., et al. (2008). Provision of distance learning services over Interactive Digital TV with MHP. *Computers & Education, 50*(3), 927-949.

Peis, E., del Castillo, J. M. M., & Delgado-López, J. (2008). Semantic recommender systems. analysis of the state of the topic. *Hipertext. net, 6*, 1-5.

Rashid, A. M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., et al. (2002). *Getting to know you: learning new user preferences in recommender systems.* Paper presented at the Proceedings of the 7th international conference on Intelligent user interfaces.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). *GroupLens: an open architecture for collaborative filtering of netnews.* Paper presented at the Proceedings of the 1994 ACM conference on Computer supported cooperative work.

Rodríguez, R. M., Espinilla, M., Sánchez, P. J., & Martínez-López, L. (2010). Using linguistic incomplete preference relations to cold start recommendations. *Internet Research, 20*(3), 296-315.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). *Item-based collaborative filtering recommendation algorithms.* Paper presented at the Proceedings of the 10th international conference on World Wide Web.

Schafer, J., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. *The adaptive web*, 291-324.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). *Methods and metrics for cold-start recommendations.* Paper presented at the Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval.

Semeraro, G. (2010). Content-based Recommender Systems problems, challenges and research directions. *INTELLIGENT TECHNIQUES FOR WEB PERSONALIZATION & RECOMMENDER SYSTEMS (ITWP 2010)*, 89.

Shardanand, U., & Maes, P. (1995). *Social information filtering: algorithms for automating "word of mouth".* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence, 2009*, 4.

Victor CODINAa, L. C. (2010). *Taking Advantage of Semantics in Recommendation Systems.* Paper presented at the Artificial Intelligence Research and Development: Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence.