Chapter 2

State of the Art

2.1 Recommender system overview

**2.1.1 Recommender system**

**What is a recommender system?**

Often termed Recommender systems, they are simply a kind of content filtering techniques and tools used to provide users with the most accurate and relevant product suggestions filtered from a huge database of information. Recommendation system works by discovering information patterns from the dataset, learning user choices and yield results that are co-related with their requirement and likings.

The most common use of recommender system can be seen in the commercial applications. From recommending movies on Netflix to predicting user ratings on amazon, the recommender system can be seen in a variety of areas. Recommender systems are also being used to explore research articles and experts [4], collaborators [5], and financial services [6]. Many social networks and online dating sites use social network graphs to recommend potential friendships.

One of the major areas where we see applications of recommendation systems is e-learning environments. i.e. within the context of Technology-enhanced learning [33] to improve students' self-directed learning capabilities.

**Information filtering techniques**

Recommender system strategies generally use two types of techniques for generating recommendations one is content-based filtering approach and the other being collaborative filtering approach. Some times knowledge-based systems are also used (also known as rule-based). A combination of the above three techniques is also used commonly known as the hybrid approach. According to [9] five types of RSs exists: content-based, knowledge-based, demographics, community-based, collaborative and hybrid.

Content-based filtering some times, also known as Item-based collaborative approach look at the individual items and their attributes to compute the similarity between the given item *i* and then selects k most similar items [7]. At the same time, their corresponding similarities are also computed.

On the other side, Collaborative filtering techniques look at the users and their attributes and compute the similarity between the users and provide a recommendation based on it. The basic idea behind CF-based algorithms is to produce recommendations to predictions about the item based on the opinions of other like-minded users. The data can be obtained either implicitly or explicitly asked from the user. Example of explicit data collection is the ratings and review a user gives to an item. Implicit data like the previous purchase history, browsing pattern of the user can also be useful in making a meaningful recommendation.

Knowledge-based techniques for recommendation takes into account the explicit knowledge about the product and the user and some predefined rules to generate a recommendation. This is also known as rule-based recommendations [8]. The rule-based technique applies the relationship rule discovery algorithm to find an association between similar purchased items and then generates a list of recommended items based on the strength of the association [14].

For the purpose of this study, the focus is on the most widely used and perhaps the most effective technique [9] that is collaborative filtering.

Collaborative filtering can be broadly divided into two categories 1. User based collaborative filtering and the 2. Item based collaborative filtering.

In user-based collaborative filtering, we incorporate the behaviour of other users in the system to give more weights to the items which users like me have purchased or expressed and opinion about. The intuition being the more similar a user is to our target user, the more likely it is that the target user will like the items that other similar users have liked.

**Challenges in user-based collaborative filtering algorithms**

Although user-based collaborative algorithms are very popular and have been successful in past amongst both small scale and large commercial-scale applications, their extensive use has shown some potential challenges such as follows:

**Scalability:** Computational requirements for nearest neighbour algorithms grows with both the number of items and the users.

**Sparsity:** In a large commercial use (like Amazon recommending books form large DB), even if a user has purchased <1% of the items(millions of books) the nearest neighbour algorithm will not be able to recommend any item for that user. Resulting in poor accuracy.

In conventional collaborative filtering recommender system, the quantity of work increases as the number of users in the system increases. For producing accurate results quickly on a large dataset we will explorer the Item-based collaborative filtering technique.

The item-based technique first examines user-item matrix to identify relationships between all the different items present and then use this information to implicitly calculate the recommendations for the user. **[23]** discussed various techniques in to compute the item-item similarity matrix (e.g. cosine similarity, Pearson correlation) and also different methods ( E.g., regression model vs weighted sum ) to obtain recommendations from them. Because the association amongst the item is relatively steady, item-based algorithms may be able to produce equivalent or sometimes superior results [23].

[13] studies collaborative filtering algorithms and their bottlenecks. Tapestry [10] shows the earliest applications of collaborative filtering-based recommender systems. This research relied on the direct inputs of people from the closed group, like an college working group. However, recommender systems for large groups like Scratch cannot depend on each person knowing every other one. In the course of time, many rating based automated recommenders were developed. The GroupLens research group at the University of Minnesota[17,18] gives pseudonymous collaborative filtering solutions for the movies and Usenet. Video Recommender [20] and Ringo [19] are web-based music recommendation systems.

**Recommender systems in e-learning environments**

E-learning is a broad term that outlines an environment in which a person can learn at any time and anywhere using a computer, generally connected to a computer network.It is well established that e-learning can be as large and useful as the traditional classroom experience or even more than it [34]. With e-learning, one can learn and master the skills and knowledge just like he would in a traditional learning counterpart.

As e-learning systems begin to expand and are ever-expanding, the users need to first process a large amount of information before it meets their needs and provides them with items that are relevant to them.E-learnings fast development has altered traditional learning conduct and presented both the students and teachers with the fresh scenario. students find it difficult to navigate through a huge number of exercises and courses, and teachers find it difficult to recommend students with materials. One of the solutions to this information overload problem in e-learning is a recommendation system [35]

A personalized recommendation system in e-learning and online learning environments provides learning suggestions to students. In [25] Lu discusses a framework for a personalised learning recommender system. [35] discusses a fully mature recommendation system architecture model for an e-learning arrangement.

In [42] we see the application of Hierarchical Clustering (HC ) on a TEL dataset built with Coursera data known as DAJEE. Recommendation from this system was done on the basis of three educational entities, Instructors, Courses and Lessons. The main teaching context considered were Teaching preference of instructors, the course information and the lesson information.

[43] Discusses a hybrid approach for recommendation MOOC platforms which uses social network analysis and association rule mining techniques. User contributions on the forums and their social interaction and peer review activities were taken in to account to extract information for social network analysis and use this information for collaborative filtering approach. This approach was again tested on the Coursera dataset which showed a well-performing system given the limited information available in the dataset.

**Recommendations in scratch**

In [8] (Recommending exercises in Scratch: an integrated approach for enhancing the learning of computer programming) the authors discuss recommender system in scratch with respect to a course framework for teaching Foundations of Computer Programming at Universidad Estatal de Milagro (State University of Milagro), Ecuador, with Scratch. consisting of various exercises that a student follows.

Benavides [10] proposes CARAMBA a Scratch extension that recommends exercises, based on taste and program complexity. CARAMBA can recommend personalized exercises for students. This again considers problem statements as exercises that would be recommended. It also conducts a study to evaluate the improvements in student activity as a result of the recommended exercises.

**Evaluation of a recommendation system**

Recommendation algorithms typically perform differently on different recommendation tasks and domains. Therefore it is critical from a practical and research perspective to select the proper algorithm that matches the domain and tasks that is of the interest. The standard way to make such a choice is by comparing different algorithms offline using some assessment metric. [36] discusses different matrics for evaluating recommending algorithms.

In [12] discusses various evaluation methods for a recommender system, particularly on the top N recommendations tasks. It suggests that while algorithms in the top-N recommendation task, to choose the test set carefully otherwise accuracy metrics are strongly biased.

The recommendation system research area has used various types of matrics to evaluate the quality of the recommendation system. Mainly these are classified into two categories [23].

1. **Statical accuracy metrics evaluation** it computes the accuracy of a recommendation system by equating the numerical recommendation score with the actual rating by the user in the test-dataset. MAE is a widely used method for measuring the statical accuracy. The Mean absolute error is calculated between the predictions and the user ratings. It measures the deviation of recommendation from their true user given value.
2. **Decision support accuracy metrics evaluation it** evaluates how useful and effective the predictions are at helping the user find the relevant item from the set of all items.These metrics consider the process of recommendation as a set of binary task, wither the recommendation is good or bad, i.e. the item is either predicted or not considered at all. With this in mind, the prediction can be anything like 1 or 2.5, it doesn't matter as the user only selects items with a rating higher than 4. Most common techniques in this category are the ROC sensitivity and the reversal rate [37].

**Other techniques in recommendation systems**

Additionally, other techniques have been applied to the recommender system area, like Bayesian networks, Horting and clustering. Bayesian networks work by creating models based on a training set associated with a decision tree. Each edge and node of the tree represents user information. The model can be trained and completed in the time as less as hours or a few days. This model tends to be very small and very fast and can be as accurate as the nearest neighbour methods [21]. Bayesian can be used practically in environments in which the information about the user preferences changes steadily in reference to the time which the model building requires. But Bayesian networks have been known to fail in areas where user preferences change very rapidly.

On the other hand, clustering algorithms work by classifying users with similar preference into different groups known as clusters. After the clusters have been finalized, predictions for a user are made by taking the average of the opinions of other users in the cluster. Some of the clustering techniques consider users in the form of participation across several clusters and the averaging is done by the weighted sum of the degree of the participation. Clustering is known to produce non-personal results in the recommendation, in fact, in some cases, it may give even worse accuracy than the nearest neighbour algorithms [21]. However once the clustering is done, there is a large increase in performance because the comparing needs to be done on very small clusters. Clustering techniques can also be used in conjunction with nearest neighbour algorithms to reduce the size of the dataset that must be analysed.

Horting is a graph-based approach in which the nodes are represented by the users and the connections between them are edges representing the degree of closeness between the users [22]. Recommendations are given by traversing the graph to the nearest node and averaging the opinions of nearby nodes/users. Horting and nearest neighbour differ the sense that in horting graph may be traversed through users who have not had an opinion about the item and therefore considering the changing relationships that nearest neighbour does not consider. In a study with artificial datasets, Horting has shown superior performance to the nearest neighbour algorithms [2].