Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System

Poonam B. Thorat
Computer Engineering
MIT Academy of Engineering
Pune India

R. M. Goudar
Computer Engineering
MIT Academy of Engineering
Pune India

Sunita Barve
Computer Engineering
MIT Academy of Engineering
Pune India

ABSTRACT
Recommender systems or recommendation systems are a subset of information filtering system that used to anticipate the 'evaluation' or 'preference' that user would feed to an item. In recent years E-commerce applications are widely using Recommender system. Generally the most popular E-commerce sites are probably music, news, books, research articles, and products. Recommender systems are also available for business experts, jokes, restaurants, financial services, life insurance and twitter followers. Recommender systems have formulated in parallel with the web. Initially Recommender systems were based on demographic, content-based filtering and collaborative filtering. Currently, these systems are incorporating social information for enhancing a quality of recommendation process. For betterment of recommendation process in the future, Recommender systems will use personal, implicit and local information from the Internet. This paper provides an overview of recommender systems that include collaborative filtering, content-based filtering and hybrid approach of recommender system.

1. INTRODUCTION
Recommender systems have become very popular in recent years and are used in various web applications. Recommender Systems (RSs) are software tools that are used to provide suggestions to user according to their requirement. The suggestions associate with various decision-making processes, such as which items to buy, what music to listen to. "Item" is the general term used to denote what the system recommends to users. A RS normally focuses on a specific type of item, its design, its graphical user interface and the core recommendation technique used to generate the recommendations are all customized to provide useful and effective suggestions for that specific type of item. Due to the increasing importance of recommendation, it has become an autonomous research field since the mid 1990s [1]. Broadly speaking, a RS suggests to a user those items that might be of users interest. Former work [10] distinguishes recommendation techniques into following four classes.

- **collaborative-filtering**
  A significant role is play by a Collaborative Filtering (CF) methods in the recommendation process and because of that Collaborative filtering is most extensively used approach to design recommender system [1, 2]**.** In this approach recommendation for each active user is received by comparing with the preferences of other users who have rated the product in similar way to the active user [27].

- **Content-Based filtering**
  In content-Based filtering recommendations depends on users former choices. Item description and a profile of the user's orientation play an important role in Content-based filtering. Content-based filtering algorithms try to recommend items based on similarity count [27].

- **Demographic Filtering**
  In demographic filtering recommendations is established on a demographic profile of the user. Here recommendation is based on the information provided by the user is considered to be similar according to demographic parameter such as nationality, age, gender etc [27].

- **Hybrid filtering**
  The hybrid filtering is a combination of more than one filtering approach [27]. The hybrid filtering approach is introduced to overcome some common problem that are associated with above filtering approaches such as cold start problem, overspecialization problem and sparsity problem. Another motive behind the implementation of hybrid filtering is to improve the accuracy and efficiency of recommendation process.

Table 1 shows the some popular sites which are currently using recommendation system for different purpose [26].

Table 1: Popular sites using recommender systems

Site	What is recommended
Amazon	Books/other products
Facebook	Friends
Netflix	DVDs
CDNOW	CDs/DVDs
CareerBuilder	Jobs

1.1 Major challenges in recommender system

- **Data sparsity**
  As we know that usage of recommender system increases very rapidly. So that many commercial recommender system uses large datasets. Therefore , the user-item matrix used for filtering could be very large and sparse and because of that performance of recommendation process may get degrade. The cold start problem is caused by the data sparsity . In collaborative filtering method recommendation of item is based on past preferences of users, so that new users will need to rate enough count of items to

allow the system to catch their preferences accurately and thus allows for authentic recommendations [26].

- **Scalability**
  Traditional CF algorithms will suffers from scalability problems as the numbers of users and items increases. For example, consider a ten millions of customers O(M) and millions of items O(N), with that the complexity of algorithm is 'n' which is already too large. As recommender system play an important role in E-commerce application where system must respond to the user requirement immediately and irrespective of users ratings history and purchases system must make recommendations, which requires a higher scalability. Twitter is large web company to scale the recommendations of their millions of users it uses clusters of machines [26].

- **Diversity**
  Recommender system are anticipated to increase diversity because they help us to discover new products. Some algorithms, may accidentally do the opposite. Here recommender system recommend popular and highly rated items which are appreciated by particular user. This lead to lower accuracy in recommendation process. To overcome this problem there is need to develop new hybrid approaches which will enhance the efficiency of recommendation process [26].

- **Vulnerability to attacks**
  Security is one of major issue in any system which are deployed on web. Recommender system play an important role in e-commerce applications and because of that recommender systems are probably targets of harmful attacks trying to promote or inhibit some items. This is one of major challenge faced by the developer of recommender system [26].

2. COLLABORATIVE FILTERING

Collaborative filtering is most extensively used approach to design recommender system**.** Collaborative Filtering (CF) methods play an significant role in the recommendation process, although Collaborative filtering is often used along with other filtering techniques like content-based, knowledge-based [19]. Basically Collaborative filtering methods are established on gathering and examining a large amount of information which based on users demeanor, activities or preferences and anticipating taste of that particular user by using their similarity with other users [5,8]. It does not depend on machine decomposable message and thus it is correctly recommending composite items and because of that it is a key benefit of the collaborative filtering approach. In collaborative filtering recommendation system recommended objects are selected on the basis of past evaluations of a large group of users.

Table2: Recommendation process in a nutshell

Persons / Movies	Joe	Bob	Carol
Titanic	5	1	5
The reader	1	5	2
Harry potter	4	2	?

Table 2 shows the recommendation process in nutshell where first we have to estimate the potential favorable opinion of Carol about Harry potter, one can use the similarity of her with those of Joe. Alternatively, one can note that ratings of Titanic and Harry potter follow a same pattern, which show that people who liked the former might also like the latter [26]. An example given in Table 2 will give brief idea about collaborative filtering.

2.1 Techniques related to memory-based collaborative filtering

To generate prediction Memory-based CF algorithms uses the total or a some part of database of the user-item. Here every user with similar interests is part of a similar group of people. By identifying the neighbors of a new user or currently active user, it can produces a anticipation of preferences on new items for him or her.

2.1.1 k nearest neighbors

The most extensively used algorithm for collaborative filtering is the k Nearest Neighbors (kNN) [1,5,4]. In the GroupLens Usenet article recommender it was first introduced. There are two types of k Nearest Neighbor algorithms:

1. User based Nearest Neighbor

2. Item based Nearest Neighbor.

1. User based Nearest Neighbor

In the user to user version, kNN executes the following three tasks to generate recommendations for an active user:

(a) Using the selected similarity measure, we produce the set of k neighbors for the active user a. The k neighbors for a are the nearest k (similar) users to u.

(b) Once the set of k users (neighbors) similar to active user a has been computed, in order to receive the prediction of item i on user a, one of the following aggregation approaches is often used: the average, the weighted sum and the adjusted weighted aggregation (deviation-from-mean).

(c) To obtain the top-n recommendations, we choose the n items, which provide most satisfaction to the active user according to our predictions.

User to user based kNN suffers from scalability problem.

2. Item based Nearest Neighbor

As the number of users increases User to user based kNN suffers from scalability problem. To overcome this drawback new method called item to item kNN is introduced by Sarwar et al. [22] and Karypis. The item-based approach investigates the set of items rated by target user and calculates their similarity with the target item *i* and then chooses *k* most similar items$\{i_1, i_2, \ldots i_k\}$. Their representing similarities$\{s_{i1}, s_{i2}, \ldots s_{ik}\}$ are also computed at the same time. Formerly the most similar items are discovered, after that by taking a weighted mean of the target user's ratings on these similar items the prediction is calculated. Similarity computation and the prediction generation are two important factors which make item-based recommendation more powerful. For similarity computation basically different types of similarity measures are used and weighted sum and regression used for prediction computation**.**

2.1.2 Dimensionality reduction techniques

To reduce the problems from high levels of sparsity in RS databases, certain studies have used dimensionality reduction techniques [6]. The reduction methods are based on Matrix Factorization [7,9,8]. Matrix factorization is especially

adequate for processing large RS databases and providing scalable approaches [10]. The model-based technique Latent Semantic Index (LSI) and the reduction method Singular Value Decomposition (SVD) are typically combined to achieve high performance [11,15,17]. SVD methods provide good anticipation results but are computationally it is very expensive. Its distribution relies on static off-line settings where it does not alter with time the known preference of information.

2.2 Techniques related to model-based collaborative filtering

The basic idea behind model-based recommendation systems is to build a "model" with the help of dataset ratings. In other words we can say that it is process of extraction of some information from the dataset and use that information as a "model" to make recommendations without having the use of complete dataset every time. This approach is beneficial in terms of both speed and scalability. Model based approach also improves prediction accuracy of algorithm.

2.2.1 CF algorithms based on MDP

Instead of viewing the recommendation process as a anticipation problem views it as a consecutive optimization problem and use a Markov decision processes (MDPs) model for recommender systems. An MDP is a model for sequential stochastic determination problems, which is often used in applications where an agent is influencing its surrounding environment through actions. More appropriate model is provided by Markov decision processes (MDPs) for implementation of recommender systems. The key advantage of MDP is they consider the long-term effects of each recommendation and the arithmetic mean of each recommendation. The MDP-based recommender system get succeed in practice because it employ a strong initial model which is solvable quickly .The MDP has memory less property and due to that it does not consume too much memory.

An MDP can be defined as a four-tuple: $< S, A, R, Pr >$, where S is a set of states, A is a set of actions, R is a reward function for each state/action pair, and Pr is the transition probability between every pair of states given each action [20].

2.3 Advantages of collaborative filtering

1. Memory-Based Collaborative filtering techniques makes implementation of recommendation system easier.
2. Using Memory-Based Collaborative filtering techniques one can add new data easily and in incremental manner.
3. Model-Based Collaborative filtering techniques improves prediction performance.

2.4 Limitation of collaborative filtering

1. Cold Start: CF systems often require a huge amount of existing data on which user can make exact recommendations [21].
2. Scalability: CF makes recommendations for various environments where billions of users and products exist. Therefore, a huge amount of computation power is often essential to compute recommendations.
3. Sparsity: On major e-commerce site the number of items sold are enormously large. Because of that only a small subset of the entire database is rated by most active users. Hence very few ratings are given to the most popular items[3].

3. CONTENT BASED-FILTERING

Content-based filtering (CBF) tries to recommend items to the active user based on similarity count which is rated by that user positively in the past [16,14,5]. For example, if a user likes a web page with the words ''mobile'', ''pen drive'' and ''RAM'', the CBF will recommend pages related to the electronics world. Item description and a profile of the user's orientation play an important role in Content-based filtering. Content-based filtering algorithms try to recommend items based on similarity count. The best-matching items are recommended by comparing various candidate items with items previously rated by the user.

The tf–idf representation is most extensively used algorithm (also called vector space representation). For creation of user profile mostly system concentrates on two types of information: 1. A user's preference model. 2. User's interaction log with the recommender system. Basically, item profile is used by these methods for (i.e. a set of distinct dimensions and characteristics) qualifying the item within the system. Creation of a content-based profile of users is done with help of weighted vector of item features. Importance of each feature to the user is denoted by the weights. It can be calculated from individually rated content vectors using a various proficiencies. Figure 1 shows the CBF mechanism, which includes the following steps:

1. Educe the attributes of items for recommendation.

2. Compare the attributes of items with the preferences of the active user.

3. Recommend items according to features that fulfill the user's interests.

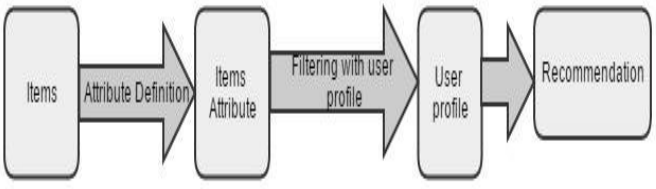

Figure:1 Content –based filtering

When the attributes of the items and the user profiles are known, the key role for CBF is to determine whether a user will like a specific item. This task is traditionally answered by using heuristic methods [18] or classification algorithms, such us: rule induction, nearest neighbors methods, Rocchio's algorithm, linear classifiers and probabilistic methods [13] .

3.2 Advantages of content-based filtering

1. Content-based recommender system provide user independence through exclusive ratings which are used by the active user to build their own profile.
2. Content-based recommender system provide Transparency to their active user by giving explanation how recommender system works.
3. Content-based recommenders system are adequate to recommend items not yet placed by any user. This will be advantageous for new user.

3.3 Limitation of content-based filtering

1. It is a difficult task to generate the attributes for items in certain areas.
2. CBF advocate the same types of items because of that it suffers from an overspecialization problem.
3. It is harder to acquire feedback from users in CBF because users do not typically rank the items (as in CF) and therefore, it is not possible to determine whether the recommendation is correct.

4. HYBRID RECOMMENDER SYSTEM

Recent research has proved that a hybrid approach could be more effective in some cases. Basically Collaborative filtering and Content-based filtering approaches most extensively used in information filtering application. As we know that every coin has two side similarly each approach has its own reward and weaknesses. Basically the main motive of hybrid approach is to aggregate collaborative filtering and content-based filtering to improve recommendation accuracy. Hybrid approaches can be implemented in various ways:

1. Implement collaborative and content-based methods individually and aggregate their predictions.
2. Integrate some content-based characteristics into a collaborative approach,
3. Comprise some collaborative characteristics into a content-based approach, and
4. Construct a general consolidative model that integrate both content-based and collaborative characteristics.

Cold start and the sparsity are common problems in recommender systems which are resolved by using these methods. Good example of hybrid recommender systems is Netflix. They make recommendations by comparing the looking out and exploring habits of similar exploiters (collaborative filtering) as well as by providing movies that share features with films that a exploiters has rated highly (content-based filtering).

The online DVD rental company Netflix released a data set containing approximately 100 million anonymous movie ratings in October 2006 and challenged investigators and practitioners to beat the accuracy of the company's recommendation system, Cinematch [23]. Although the released data set represented only a small fraction of the company's rating data, thanks to its size and quality it fast became a standard in the data mining and machine learning community. The data set contained ratings in the integer scale from 1 to 5 which were accompanied by dates. The year of release were provided for every movie and title. No information about users was given. Submitted predictions were evaluated by their root mean squared error (RMSE) on a qualifying data set containing over 2817,131 unknown ratings. Total 20,000 are registered teams out of that 2000 teams submitted at least one answer set. The grand prize of $1000,000 was awarded to a team on 21 September 2009 that performed better over the Cinematch's and also increases accuracy by 10%. In this competition we learned several lessons [24]. Firstly, the company acquired a superior recommendation system that improve users satisfaction and also company gained lot of publicity. Secondly, ensemble methods play an important role for improving the accuracy of predictions. Thirdly, we discovered that when RMSE drops below a certain level that time accuracy improvements are increasingly demanding. Finally, despite the company's effort, namelessness of its users was not sufficiently assured [25].

CBF and CF can be aggregated in different ways [1].following figures shows the different choices for aggregating CB and CBF. Figure 2 shows the methods that estimate CBF and CF recommendations individually and subsequently combine them to yield better recommendations across the board.

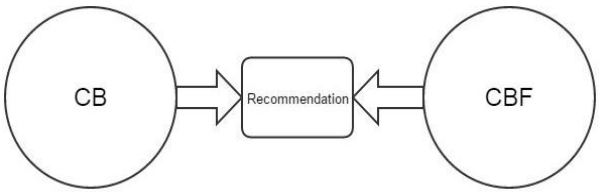

Figure: 2

Figure 3 shows the methods that integrate CBF characteristics into the CF approach. So that it will overcome the cold start problem in collaborative filtering and overspecialization problem of content-based filtering.

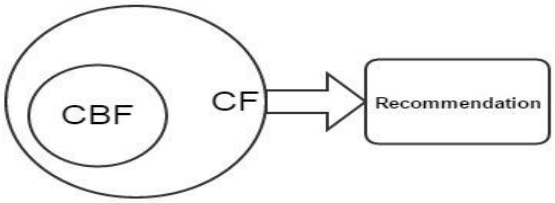

Figure: 3

Figure 4 illustrates the methods for construction of a unified utility system with both CBF and CF characteristics. In this method by combining some features of CBF and CF one unified model is constructed that can improve effectiveness of recommendation process.

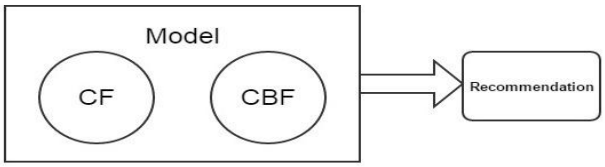

Figure: 4

Figure 5 shows the methods that incorporate CF characteristics into a CBF approach.

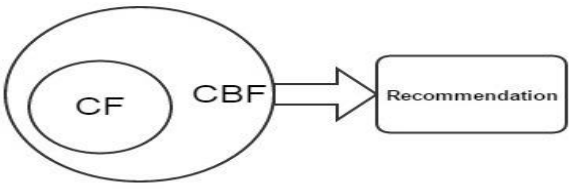

Figure: 5

Content-based filtering systems can allow recommendations for "cold-start" items for which no training information is available, but it suffers by lower accuracy than collaborative filtering systems. Conversely, collaborative filtering approach frequently provide accurate recommendations, but go wrong for cold start items. Hybrid schemes try to aggregate these different kinds of information to get efficient recommendation result.

4.1 Classification of Hybrid Recommendation Systems:

Burke [12] presented taxonomy for the hybrid recommendation systems he classifies Hybrid Recommendation Systems into following seven classes:

- **Weighted**:
  Different recommendation components scores are combined statistically. This class aggregates scores from each factor using additive formula.
- **Switching:**
  From available recommendation components system chooses particular component and applies the picked out one.
- **Mixed:**
  Different recommender provides their recommendation that will be introduced together. This class is based on merging and presentation of multiple rated list into single rated list.
- **Feature Combination:**
  Contributing and actual recommender are two different recommendation components are exist for this class. The working of actual recommender is depends on the data modified by the contributing one. The contributing one throws features of one source on to the other components source.
- **Feature Augmentation:**
  This class is similar to the feature combination hybrids but only difference is that the contributor gives novel characteristic. It is more elastic than feature combination method.
- **Cascade:**
  This class play an role of tie breaker. Here for every recommender assign some priority and according to that assign priority, lower priority recommenders play an tie breakers role over higher priority.
- **Meta-level:**
  Their exist contributing and actual recommenders but the early one completely substitutes the data for the latter one.

5. CONCLUSION

Recommender systems are turning out to be a useful tool that will provide suggestion to user according to their requirement. Filtering is used to improved recommendation accuracy in the first recommender systems. To achieve this accuracy most memory-based methods and algorithms were formulated and optimized under some circumstance (e.g., kNN metrics, singular value decomposition, etc.). At this stage, to improved the quality of the recommendations some hybrid approaches are used (primarily collaborative filtering and content filtering). In the second stage, algorithms that admitted social information with former hybrid approaches were accommodated and developed (e.g., trust-aware algorithms, social adaptive approaches, social networks analysis, etc.). Currently, the hybrid algorithms are used to integrate location information into existing recommendation algorithms. To improve the quality of recommender systems anticipations future research will concentrate on progressing the existing methods and algorithms. Novel lines of research will be formulated for following fields, such as on: (1) The existing recommendation methods that uses different types of available information will be combine in good order, (2) For recommender systems processes enable security and privacy, (3) Flexible frameworks are design for machine-controlled analysis of heterogeneous data.

6. REFERENCES

[1] Adomavicius, G.; Tuzhilin, A., "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* , vol.17, no.6, pp.734,749, June 2005.

[2] F. Ricci, L. Rokach, B. Shapira, P.B. (Eds.), "Kantor Recommender Systems Handbook", first ed., 2011, XXX, 842 p. 20 illus

[3] J. Bobadilla, F. Serradilla, "The effect of sparsity on collaborative filtering metrics", *in: Australian Database Conference*, 2009, pp. 9–17.

[4] J. Bobadilla, A. Hernando, F. Ortega, J. Bernal, "A framework for collaborative filtering recommender systems", *Expert Systems with Applications* 38 (12) (2011) 14609–14623.

[5] J. B. Schafer, D. Frankowski, J. Herlocker, S. Sen,"Collaborative filtering recommender systems", in: P. Brusilovsky, A. Kobsa, W. Nejdl (Eds.), The Adaptive Web, 2007, pp. 291–324 .

[6] B. Sarwar, G. Karypis, J. Konstan, J. Riedl, "Application of dimensionality reduction in recommender system – a case study", in: ACM WebKDD Workshop, 2000b, pp. 264–272.

[7] Y. Koren, R. Bell, CH. Volinsky, "Matrix factorization techniques dor recommender systems", IEEE *Computer 42 (8)* (2009) 42–49.

[8] X. Luo, Y. Xia, Q. Zhu, "Applying the learning rate adaptation to the matrix factorization based collaborative filtering", *Knowledge Based Systems* 37 (2013) 154–164.

[9] X. Luo, Y. Xia, Q. Zhu, "Incremental collaborative filtering recommender based on regularized matrix factorization", *Knowledge-Based Systems* 27 (2012) 271–280.

[10] G. Takacs, I. Pilaszy, B. Nemeth, D. Tikk, "Scalable collaborative filtering approaches for large recommender systems", *Journal of Machine Learning Research* 10 (2009) 623–656.

[11] M.G. Vozalis, K.G. Margaritis, "Using SVD and demographic data for the enhancement of generalized collaborative filtering", *Information Sciences 177* (2007) 3017–3037.

[12] R. Burke, "Hybrid recommender systems: survey and experiments", User Modeling and User-Adapted Interaction 12 (4) (2002) 331–370.

[13] M. Gemmis, P. Lops, G. Semeraro, P. Basile, "Integrating tags in a semantic content-based recommender", *in: Proceedings of the 2008 ACM conference on Recommender Systems*, 2008, pp. 163–170.

[14] M. Pazzani, "A framework for collaborative, content-based, and demographic filtering", Artificial Intelligence Review-Special Issue on Data Mining on the Internet 13 (5-6) (1999) 393–408.

[15] S. Zhang, W. Wang, J. Ford, F. Makedon, "Using singular value decomposition approximation for collaborative

filtering", in: IEEE International Conference on *E-Commerce Technology*, 2005, pp. 1–8.

[16] M. Balabanovic, Y. Shoham, "Content-based, collaborative recommendation", *Communications of the ACM* 40 (3) (1997) 66–72.

[17] F. Cacheda, V. Carneiro, D. Fernandez, V. Formoso, "Comparison of collaborative filtering algorithms: limitations of current techniques and proposals for scalable, high-performance recommender Systems", *ACM Transactions on the Web 5* (1) (2011). Article 2.

[18] C. Basu, H. Hirsh, W. Cohen, "Recommendation as classification: using social and content-based information in recommendation", in: Proceedings of the Fifteenth National Conference on Artificial Intelligence, 1998, pp. 714–720.

[19] Lops, Pasquale, Marco De Gemmis, and Giovanni Semeraro. "Content-based recommender systems: State of the art and trends." Recommender systems handbook. Springer US, 2011. 73-105.

[20] Sutton, R.S., Barto, A.G., 1998. "Reinforcement Learning: An Introduction". MIT Press, Cambridge, MA.

[21] K. Heung-Nam, E.S. Abdulmotaleb, J. Geun-Sik, "Collaborative error-reflected models for cold-start recommender systems", Decision Support Systems 51 (3) (2011) 519–531.

[22] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in ACM www '01, pp. 285–295, ACM, 2001.

[23] J. Bennett, S. Lanning, "The Netflix prize, in: Proceedings of KDD Cup and Workshop", 2007, pp. 3–6.

[24] R.M. Bell, Y. Koren, "Lessons from the Netflix prize challenge", ACM *SIGKDD Explorations Newsletter 9 (2007) 75–79.*

[25] A. Narayanan, V. Shmatikov, "Robust de-anonymization of large sparse datasets", in: IEEE *Symposium on Security and Privacy, 2008, pp. 111–125.*

[26] Linyuan Lua, Matus Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang,Tao Zhou, et al. "Recommender systems" *Physics Reports* 519.1 (2012): 1-49.

[27] Sánchez Sánchez, José Luis. "Improving Collaborative Filtering Based Recommender Systems Using Pareto Dominance*"*. Diss. E_Informatica, 2013.