International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Survey Paper on Clustering-based Collaborative Filtering Approach to Generate Recommendations

Rohit C. Joshi[1], Ratnamala S. Paswan[2]

[1]University of Pune, Department of Computer Engineering, Pune Institute of Computer Technology, Pune 411043, India

[2]Professor, University of Pune, Department of Computer Engineering, Pune Institute of Computer Technology, Pune 411043, India

Abstract: ***The rapid development of information technology takes our shopping into the orbit of information. With the network construction of resources, the amount of shopping resources increases rapidly. Collaborative filtering (CF) is a technique commonly used to build personalized recommendations on the Web. Some popular websites that make use of the collaborative filtering technology include Amazon, Netflix, iTunes, IMDB. The most important issue which influences the collaborative filtering recommendation accuracy is the so-called data sparseness. Data sparseness causes the system difficulty in determining the nearest neighbors of the target user accurately. Clustering can solve this problem to some extent. Grouping a set of physical or abstract objects into classes of similar objects, this process is called as clustering. This paper presents the methods to generate recommendations using clustering-based collaborative filtering approach.***

Keywords: Clustering, Collaborative Filtering, Data Sparseness, Personalized Recommendations, Nearest Neighbors

1. Introduction

The rapid development of information technology takes our shopping into the orbit of information. With the network construction of resources, the amount of shopping resources increases rapidly. For example, more and more customers come to buy the watch, but due to the large number of watches of different brands, it costs increasingly greater for the shoppers to find the needed information. Search engine, to a certain degree, eases the pressure of information retrieval, but the traditional search engine takes all watches as a group and doesn't take the differences of the brands, cost, material, etc, into account. As a result, it is difficult to meet the buyer's individual needs. Therefore, the personalized recommendation technology in the field of e-commerce becomes crucial. Among the recommendation techniques, the collaborative filtering recommendation technology is the main direction of the current studies.

At present, the most important issue which influences the collaborative filtering recommendation accuracy is the so-called data sparseness. Among the nearest neighbor's resources, the amount of resources which users have scored is negligible for the total number of system resources, which leads to the few evaluation data and the sparse data. Data sparseness causes the system difficulty in determining the nearest neighbors of the target user accurately, rendering it difficult to provide the target user with high-quality and efficient personalized recommendation.

Many commercial recommender systems use to evaluate large item sets (e.g., Amazon.com recommends books and CDnow.com recommends music albums). In these systems, even active users may have purchased well under 1% of the items (1% of 2 million books is 20,000 books). Accordingly, a recommender system based on nearest neighbor may unable to make any item recommendations for a particular user. As a result accuracy of recommendations may be poor.

In addition to data sparseness, there are still many factors influencing the accuracy of recommendation. One reason is that the collaborative filtering recommendation algorithm is based on the users' score for each resource, but the differences between users lead to a poor recommendation quality; the second reason is the uncertainty of the users and the limitations of the traditional collaborative filtering technology, namely, it processes all the resources which have been scored by the nearest neighbors of the user. The recommendation of resources only based on the similarity of users, to some extent, reduces the influence of the target user's own interests in the recommendation process, and reduces the accuracy of the recommendation system.

2. Literature Survey

Users having similar characteristics are assembled in the cluster analysis according to the web visiting message data. However, user's preference on web visiting may be irrelevant from preference on purchasing. Mittal, [7] projected to obtain the predictions for a user by first minimizing the size of item set, the user needed to explore. Movies are partitioned based on the genre requested by the user using *k*-means clustering algorithm. However, users have to give some extra information.

High-dimensional parameter free, divisive hierarchical clustering algorithm applied by, Simon, [8] needs only implicit feedback based on past user purchases to determine the relationships between the users. Products of high interest were endorsed to the users according to clustering results. However, implicit feedback does not always give correct information about the user's preference.

Data-Providing service in terms of vectors is described by, Zhou, [9] which considers the composite relation between input, output, and semantic relations between them. Refined fuzzy C-means algorithm is applied to cluster the vectors. The capability of service search engine was enhanced

considerably by merging similar services into a same cluster, especially in large Internet based service repositories. However, it is pretended that domain ontology exists for aiding semantic interoperability. Besides, this approach is inconvenient for some services which are lack of parameters.

Pham, [10] projected the neighborhoods of the users in social network is determined by applying network clustering technique, and then provide the traditional CF algorithms to produce the recommendations. This work is relying upon on social relationships between users.

Li, [11] projected to include multidimensional clustering into a collaborative filtering recommendation model. Background data in the form of user and item profiles was composed and clustered using the projected algorithm in the first stage. Then the poor clusters with analogous features were deleted while the appropriate clusters were further picked based on cluster pruning. At the third stage, an item prediction was made by operating a weighted average of deviations from the neighbor's mean. Such an approach was likely to trade-off on increasing the variety of recommendations while preserving the accuracy of recommendations.

Thomas, [12] proposed collaborative filtering based on weighted co-clustering algorithm. User and item neighborhoods are simultaneously produced via co-clustering and generate predictions based on the average ratings of the co-clusters. The entry of new users, items and ratings is handled by using an incremental co-clustering algorithm.

J. Kelleher, [13] proposed a collaborative recommender that uses a user-based model to predict user ratings for specified items. The model generates summary rating information derived from a hierarchical clustering of the users. Its accuracy is good and coverage is maximal. Proposed algorithm is very efficient; predictions can be made in time that grows independently of the number of ratings and items.

Rashid, [14] projected ClustKnn approach, a simple and intuitive algorithm that is well suited for large data sets. The projected method first compresses data tremendously by building a straightforward but efficient clustering model. Recommendations are then generated quickly by using a simple Nearest Neighbor-based approach. ClustKnn provides very good recommendation accuracy.

Sarwar, [15] proposed a new approach in improving the scalability of recommender systems by using clustering techniques. Experiments suggest that clustering based neighborhood provides comparable prediction quality as the basic CF approach. Author uses a variant of K- means clustering algorithm called the bisecting K-means clustering algorithm. This algorithm is fast and tends to produce clusters of relatively uniform size, which results in good cluster quality.

Zhirao, [16] proposed Community-based collaborative filtering algorithm, its idea is that the users who belong to the same community have the same interests, who do not belong to the same community do not have the same interests, which narrows the scope of the neighbors. To a certain extent, it solves the problem of data sparseness.

Tseng, [17] proposed Default voting scheme using the cloud model which represents the user's global preference that is computed from users' past ratings to ameliorate the sparsity problem. In order to describe the user's interests and preferences more accurately and reduce the data sparsity.

ZHANG, [18] Considers the user's level of consumption, using the association rule mining formalized the competitive relationship between goods; using the time-based Bayesian probability formalize the complementary relationship between commodities, and through these relationship between the two commodities matches the user's requiring preferences and price preferences into the item sets of user evaluation..

3. Overall Survey

3.1 Clustering

Grouping a set of physical or abstract objects into classes of similar objects, this process is called as clustering. A cluster is a collection of data objects that are similar to each other within the same cluster and dissimilar objects are in other clusters. Clustering partitions large data sets into groups according to their similarity, therefore, clustering is also called data segmentation in some applications.

3.1.1Partitioning Method

Partitioning organizes the objects of a set into several exclusive groups or clusters. It is a simplest and most fundamental version of cluster analysis. These are of two types.

a) **K-Means**

K points [1] are set in the space produced by the objects. These points serve as initial group centroids. Each object is accredited to the group that has the closest centroid. Recalculate the positions of the K centroids after all the objects have been assigned. This continues until the centroids no longer move. This produces a separation of the objects into clusters.

b) **K-Medoids**

Randomly pick k of the n data points as the medoids. A medoid is a data point in finite dataset, whose average dissimilarity to all the data points is minimal. It is the most centrally located point in the set. Associate each data point to the closest medoid, for each medoid 'd' and each data point 'p' associated to 'd'. Swap 'd' and 'p' to compute the total cost of the configuration, that is, the average dissimilarity of 'p' to all the data points associated to 'd'. Pick the medoid 'd' with the lowest cost of the configuration.

3.1.2Hierarchical Method

Hierarchical method organizes data objects into a tree of clusters. This method can be further classified into agglomerative and divisive.

a) **Agglomerative Hierarchical Clustering**

It is a bottom-up strategy. [1] Each object is placed in its own cluster and then merged these atomic clusters into larger and

larger clusters, until all objects are in a single cluster or until certain termination conditions are satisfied.

b) **Divisive Hierarchical Clustering**

It is a top-down strategy. It starts with all objects in one cluster. It subdivides the cluster into smaller and smaller pieces, until each object forms its own cluster or until satisfies certain termination conditions. This strategy is reverse of agglomerative hierarchical clustering.

3.1.3Density Based Method

Clustering based on density (local cluster criterion), such as density-connected points. Features are as follows:

- Determines clusters of arbitrary shape
- One scan
- Handle noise.

It needs density parameters as termination condition. There are mainly three types.

a) **DBSCAN**

(**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise) is a density based clustering algorithm. The algorithm determines sufficiently high density regions into clusters and clusters are of arbitrary shape with noise in spatial databases. It forms a cluster as a maximal set of density-connected points.

b) **OPTICS**

(**O**rdering **P**oints **T**o **I**dentify the **C**lustering **S**tructure) [2] It is similar to DBSCAN, but it overcomes the problem of detecting meaningful clusters in data of varying density, one major weakness of DBSCAN. In order to do so, the points of the database are (linearly) ordered. Points which are spatially closest become neighbors in the ordering. Additionally, a special distance is stored for each point that represents the density that forms a cluster in order to have both points belong to the same cluster. It can be represented by graphically or using visualization techniques.

c) **DENCLUE**

(**DEN**sity-based **CLU**st**E**ring) [3] It is a clustering method based on a set of density distribution functions. Overall density of the data space can be calculated as the sum of the influence function of all data points. An influence function describes the impact of a data point within its neighborhood. However, most of the data points, do not actually contribute to the overall density function. It uses a local density function. This algorithm considers only the data points which actually contribute to the overall density function.

3.1.4Grid Based Clustering

Define a set of grid-cells. Assign objects to the appropriate grid cell and compute the density of each cell. Cells whose density is below a certain threshold are eliminated. It uses a multi resolution grid data structure.

a) **STING**

It is [4] a grid based multi resolution clustering technique in which the spatial area is divided into rectangular cells. There are several levels of rectangular cells that correspond to different levels of resolution. These cells form a hierarchical structure. Each cell at a high level is partitioned into a number of smaller cells in the next lower level. Statistical information of each cell is calculated and stored and is used to answer queries.

b) **CLIQUE**

(**CL**ustering **I**n **QUE**st) was the first algorithm proposed that Automatically identifying sub spaces of a high dimensional data space that allow better clustering than original space. Partition the data space and find the number of points that lie inside each cell of the partition. It identify the sub spaces that contain clusters.

3.1.5Model Based Clustering

Method attempts to optimize the fit between the given data and some mathematical model. Such methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. The probability model for clustering will often be a mixture of multivariate normal distributions. Each component in the mixture is what we call a cluster. These are of three types- Expectation-Maximization, Conceptual clustering and a neural network approach to clustering.

a) **EM**

(**E**xpectation-**M**aximization) This algorithm can be used for finding the parameter estimates that follows iterative refinement. It can be seen as a variation of the k-means paradigm, which assigns an object to the cluster with which it is most similar, based on the cluster mean.

b) **Conceptual Clustering**

It is a form of clustering in machine learning. It produces a classification scheme over the objects from given set of unlabeled objects. The process generates a concept description for each generated class. Most conceptual clustering methods can generate hierarchical category structures.

c) **Neural Network**

It is [5] a useful technique for implementing competitive learning based clustering, which have simple architectures. A competitive learning-based neural networks used for clustering include adaptive resonance theory models, self-organizing map and learning vector quantization.

3.2 Collaborative Filtering

Collaborative filtering (CF) is a technique commonly used to build personalized recommendations on the Web. Some popular websites that make use of the collaborative filtering technology include Amazon, Netflix, iTunes, IMDB.

3.2.1Memory Based Collaborative Filtering

User rating is used to calculate similarity or weight between users and items. [6] Make predictions or recommendations according to those calculated similarity values. Similarity values are based on common items and therefore are unreliable when data are sparse. The common items are therefore few.

a) **User Based Collaborative Filtering**

It predicts user's interest in particular item based on rating information from similar user's profiles. Ratings by more similar users contribute to more predicting the item rating. Set of similar users can be identified by employing a threshold.

b) **Item Based Collaborative Filtering**

Apply the same idea as user based filtering but use similarity between items instead of users. Unknown rating of the item can be predicted by averaging the ratings of

other similar items. Where item similarity can be approximated by the cosine measure or pearson correlation.

3.2.2Model Based Collaborative Filtering

These algorithms make model of the user rating to provide item recommendation. [6] Probabilistic approach is used in this category. Different machine learning algorithms can be applied in model building process.

a) **Bayesian Network:** It drafts the probabilistic model for collaborative filtering problem.
b) **Clustering:** It considers the collaborative filtering problem as classification problem. It processes by keeping similar users in the same class.
c) **Rule Based Approaches:** It employs association rule discovery algorithms to discover association between co-purchased items and the item recommendations are produced on the basis of strength of association between items.

3.2.3Content Based Filtering

It recommends items based on the comparison between content of the items or user profiles which is also mentioned as cognitive filtering. Content of each item is expressed as set of descriptors or terms, typically words that appear in the document. Terms can either be assigned automatically or manually. When terms are assigned automatically, method has to be chosen that can extract these terms from items. These terms have to be represented such that both user profiles can be compared in meaningful way.

4. Conclusion

In this paper, we presented a survey on analysis of personalized recommendation techniques based on clustering and collaborative filtering. We also gave an overview of different approaches of clustering and collaborative filtering. It is a new area of research. It has provided a new way to generate recommendations in an effective manner with relatively good accuracy and low cost.

References

[1] A.K. Jain, M.N. Murty, P.J. Flynn, “Data Clustering: A Review”, 1996 IEEE Computer Society Press.
[2] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jorg Sander, “OPTICS: Ordering Points To Identify the Clustering Structure”, Institute for Computer Science, University of Munich Oettingenstr. 67, D-80538 Munich, Germany.
[3] Alexander Hinneburg, Daniel A. Keim, “An Eficient Approach to Clustering in Large Multimedia Databases with Noise”, Copyright (c) 1998, American Association for Arti_cial Intelligence (www.aaai.org).
[4] Gholamhosein Sheikholeslami, Surojit Chatterjee, Aidong Zhang, “WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases”, Proceedings of the 24th VLDB Conference New York, USA, 1998.
[5] Farhat Roohi, “Artificial Neural Network Approach to Clustering”, The International Journal Of Engineering And Science (Ijes) Volume2, issue3, Pages33-38, 2013.
[6] Xiaoyuan Su, Taghi M. Khoshgoftaar, “A Survey of Collaborative Filtering Techniques”, Hindawi Publishing Corporation Advances in Artificial Intelligence Volume 2009, Article ID 421425, 19 pages.
[7] N. Mittal, R. Nayak, M. C. Govil, and K. C. Jain, “Recommender system framework using clustering and collaborative filtering,” in Proc. 3rd Int. Conf. Emerging Trends Eng. Technol., Nov. 2010, pp. 555558.
[8] R. D. Simon, X. Tengke, and W. Shengrui, “Combining collaborative filtering and clustering for implicit recommender system, ” in Proc. IEEE 27th Int. Conf. Adv. Inf. Netw. Appl., Mar. 2013, pp. 748755.
[9] Z. Zhou, M. Sellami, W. Gaaloul, M. Barhamgi, and B. Defude, “Data providing services clustering and management for facilitating service discovery and replacement,” IEEE Trans. Autom. Sci. Eng., vol. 10, no. 4, pp. 116, Oct. 2013.
[10] M. C.Pham, Y. Cao, R. Klamma, and M. Jarke, “A clustering approach for collaborative filtering recommendation using social network analysis,” J. Univ. Comput. Sci., vol. 17, no. 4, pp. 583604, Apr. 2011.
[11] X. Li and T. Murata, “Using multidimensional clustering based collaborative filtering approach improving recommendation diversity,” in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol., Dec. 2012, pp. 169174.
[12] George Thomas, Srujana Merugu, “A scalable collaborative filtering framework based on co-clustering.”, In Proceedings of the IEEE ICDM Conference. 2005.
[13] Jerome Kelleher, Derek Bridge, “RecTree Centroid: An Accurate, Scalable Collaborative Recommender”, In Procs. of the Fourteenth Irish Conference on Artificial Intelligence and Cognitive Science, pages 89–94, 2003.
[14] Al Mamunur Rashid, Shyong K. Lam, George Karypis, John Riedl, “ClustKNN: A Highly Scalable Hybrid Model& MemoryBased CF Algorithm”, WEBKDD ’06, August 20, 2006, Philadelphia, Pennsylvania, USA, Copyright 2006 ACM.
[15] Badrul M. Sarwar, George Karypis, Joseph Konstan, John Riedl, “Recommender Systems for Large-scale E-Commerce: Scalable Neighborhood Formation Using Clustering”, Proceedings of the Fifth International Conference on Computer andInformation Technology, 2002.
[16] Jiang Zhirao, “Based on Java Technology System and Implement the Personalized Recommendations of the system”, Jilin: Jilin University, 2011.
[17] Kuo-Cheng Tseng, Chein-Shung Hwang, Yi-Ching Su, “Using Cloud Model for Default Voting in Collaborative Filtering”, Journal of Convergence Information Technology (JCIT) Volume6, Number12, December 2011.
[18] ZHANG Yao, FENG Yu-qiang, “Hybrid Recommendation method IN Sparse Datasets: Combining content analysis and collaborative filtering”, International Journal of Digital Content Technology and its Applications (JDCTA) Volume6,Number10,June 2012.