www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 7 July, 2014 Page No. 6855-6858

Music Recommendation System Using Association Rule Mining and Clustering Technique To Address Coldstart Problem

V.Manvitha[1], M.Sunitha Reddy[2]

[1]M.Tech CSE, Vasavi College of Engineering
vmanvitha@gmail.com

[2]Assistant Professor, Vasavi College of Engineering
sashu2006@gmail.com

Abstract – ***Recommender system based on data mining is very useful, more accurate and provides worldwide services to the user. Recommender systems are becoming very popular in recent years. More and more people rely on online sites for purchasing songs, apparels, books, rented movies etc. The competition between the online sites forced the web site owners to provide personalized services to their customers. So the recommender systems came into existence. Recommender systems are active information filtering systems that attempt to present to the user, information items in which the user is interested in. The recommender systems also suffer from issues like cold start, sparsity. Cold start problem is that the recommenders cannot draw inferences for users or items for which it does not have sufficient information. This paper attempts to propose a solution to the cold start problem by combining association rules and clustering technique.***

Keywords: Recommender system, Cold-start problem, clustering, association rules, Apriori.

1.INTRODUCTION

The development of the Internet provided more ways for people to interact but also a place where they could find information about almost everything and anything. Recommender systems can be considered a way of combining these two aspects in order to help people find the information they need or somethingthey would be interested in.

Recommender systems are now an integral part of online sites. They are very useful in recommending items or products to user according to their interests. The benefits of online sites having a recommender system in place are cross-selling, personalization, keeping the customers informed and customer retention. Some of the websites that use recommenders are Amazon, MovieLens, eBay, CDNow, MovieFinder. In collaborative filtering approach, the system recommends new items to the user by analyzing items purchased by similar users (Amazon.com). In content based approach, they recommend items with similar contents to the items preferred by the target users (PandoraRadio). In hybrid approaches, both the content based and collaborative approaches are used to provide recommendations (Netflix). These approaches provide the customers with a number of recommendations.

Cold start problem (new user, new item) is one of the major issues in recommender system. In the case of a new user, the number of ratings will be very less. This implies that the user profiles (consist of ratings given to the items) will be very short. The new user will be given non personalized recommendations till an adequate number of ratings are collected for the user. For a new item updated in the system, initially there will be no ratings. The possibility that this item will be recommended to the user is minimal. These problems should be addressed because the initial recommendations given to a new user plays an important role in deciding the user satisfaction and retention. Only if good quality recommendation is given, the users will come back to the site.

Various methods exist for addressing the cold start problem. Some of these solutions are based on association rules, clustering, classification etc. Many hybrid recommenders also exist for solving this issue. In this paper, association rules and clustering technique is used for solving cold start. Association rules are used to create and expand the user profile so that it will contain more number of ratings/domains of interest so as to solve new user problem. The clustering technique is used to group items and make prediction for item to solve new item problem. Experiments have been done to show that combining two techniques such as association rule and clustering gives better recommendations than using a single technique such as association rule only.

2. RELATED WORKS

In this section review of some of the works related to the proposed approach is done. Much work has been done in the area of recommender systems. Qing Li and Byeong Man Kim explain how the clustering techniques can be applied to the item-based collaborative filtering framework to solve the cold start problem. The work done byGavin Shaw, YueXu and ShlomoGeva explains how to expand a user profile from a dataset with the help of association rules. The collaborative filtering, content based and hybrid approaches and the issues in recommender systems are clearly explained in the survey done by Adomavicius and G. Tuzhilin [1]. Schein and Popescul et al... , proposed a hybrid recommender system, the aspect model, to recommend items that is not yet recommended [2]. The relationship between ontology's and recommender system and how to exploit this synergy to solve cold start problem is given by Middleton, Alani et al… [3]. Ziegler, C.N. Lausen and G. Schmidt proposes a method in which the taxonomic background knowledge is used for computing personalized recommendations in particular domain [4]. In [5] Leung, et al… Chung implements a hybrid recommendation algorithm which makes use of Cross-Level Association Rules (CLARE) to integrate content information about domain items into collaborative filters. Pasquier, N. Taouil, et al… explains how to remove the redundant association rules without reducing the information in [6]. In [7], [8], [13] by Shaw, Xu and Geva they implemented a method which allows the removal of hierarchically redundant approximate basis rules from multi-level datasets through the use of the dataset's hierarchy or taxonomy. The new algorithmic elements that increase the accuracy of collaborative filtering is discussed by Herlocker, Konstan et al... in [9]. Q. Li and B. Kim described a new filtering approach that combines the content-based filter and collaborative filter to achieve a good performance [10]. The basics of recommenders and the e-commerce sites which use the recommenders are described in [11] by Schafer, Konstan, and Riedl. Al Mamunur Rashid et al… proposed an online simulation framework to address cold start problem in their paper[12] . Another interesting work done by Sunita B Aher, and Lobo proposed various combinations of algorithms in recommending the courses to students in E-learning [14]. A combination of clustering and association rules was used to improve recommendation for digital library in the study conducted by HuiLia and XinyueLiub [15]. In [16] Herlocker, Konstan et al… explain the key decisions in evaluating the collaborative filtering recommender system.

3. PROPOSED WORK

In this section an outline of the proposed approach for solving the cold-start problem in recommender systems is addressed. The approach is to combine two existing approaches in a sequential manner. Clustering technique is one of the most important techniques which have got a wide variety of applications. First clustering technique is applied to group the similar users into the clusters. Association rule technique is applied for recommendation.

3.1 Clustering Technique

Clustering technique is one of the most important techniques which have got a wide variety of applications. In this approach clustering algorithm is applied to group the users. . Here in this work, k-means algorithm is used to cluster the users which are similar. For the similarity calculation between two users, Cosine similarity is used..

3.1.1User Clustering

User clustering technique work by identifying groups of users who appear to have similar ratings. Some clustering techniques represent each user with partial participation in several clusters. Once the user clustering is complete then a transaction database is built for each and every cluster.

3.2 Association Rule Technique

Association rule mining is an important component of data mining. Association rules are an important class of methods of finding regularities/patterns in data. Association rules find patterns in the information available about user preferences. These patterns are then used to make predictions based on the information available for the selected user. Agrawal (1993) give a formal statement of association rule mining for transaction databases.

Item : It is a field of the transaction database.

Transaction : It is corresponding to a record of the database. Transaction usually is marked as small letter t to mark item i.ti={i1,i2,…,ip}. Each transaction has an only identifier called TID. The whole set of transaction ti constitutes a database D.

D={t1,t2,…,tn}

Support: The support of association rule X-> Y in transaction database is a ratio. The ratio is between the count of item set which contains X and Y, and the count of all of item set. That marks support(X->Y). That is the percent of the item set containing X and Y at the same time in the transaction database.

Confidence: It is the ratio between the count of transaction containing X and Y and the count of transaction containing X.That is marked as confidence(X ->Y). Confidence is the percent of the transaction sets containing X and Y at the same time in the transaction database.

Frequent Item set: The item set, whose support is not lower than the minimum support (Min Sup).

Strong rule and Weak rule: If support(X -> Y) >=MinSupport and Confidence (X->Y)>=MinConf, then mark association rule X-> Y as strong rule, otherwise mark it as a weak rule.

Let I = {i1, i2, . . . , im} be the universe of items. A set X of items is called an itemset. A transaction t = (tid ,X) is a tuple where tid is a unique transaction ID and X is an itemset. A transaction database D is a set of transactions. The count of an itemset X in D, denoted by count(X), is the number of transactions in D containing X. The support of an itemset X in D, denoted by supp(X), is the proportion of transactions in D that contain X. The rule X ->Y holds in the transaction set D with confidence c where c = conf(X->Y) and conf(X->Y) = supp(XY)/supp(X), where supp(XY) denotes the support of items X and Y occurring together. In association rule mining the objective is to retrieve all rules of the form X->Y where supp(XY) > s and conf(X->Y) >c, with s and c being user-supplied thresholds on minimum support and minimum confidence respectively. (Agrawal,1993).

4. DATASET COLLECTION AND PREPROCESSING

The dataset used contains information about songs, users and the user song count .It was taken from Million song dataset and it is named Taste profile subset.

Taste Profile Subset: Taste Profile subset, the official user dataset of the Million Song Dataset. The dataset contains real user – play-counts from private partners, all songs already matched to the MSD.

- 1,019,318 unique users
- 48,373,586 user - song - play count triplets

The data looks like this:

- User ID Song ID count

  b80344d063b5ccb3... SOYHEPA12A8C13097F 8

  b80344d063b5ccb3... SOYYWMD12A58A7BCC9 1

  85c1f87fea955d09... SOACWYB12AF729E581 2

The data set is split into 75%/25% disjoint splits into training and test data. This is used for fold cross validation.Data preprocessing is done by taking the constraint as the user has listened at least three songs and the songs are preprocessed using song listened by at least 30 users.

6. ALGORITHM

The algorithm used is Apriori Algorithm. Apriori algorithm according to the property of association rule is the sub sets of the frequent item set is also frequent item set, the supersets of non-frequent item set is also non- frequent item set. The algorithm each time makes use of k-frequent item set carrying on conjunction to get k+1 candidate item set. The key idea of Apriori algorithm is to make multiple passes over the database. It employs an iterative approach known as abreadth-first search (level-wise search) through the search space, where k-item sets are used to explore (k+1)-item sets.

The working of Apriori algorithm is fairly depends upon the Apriori property which states that" All nonempty subsets of a frequent item sets must be frequent" [5]. It also described the anti monotonic property which says if the system cannot pass the minimum support test, all its supersets will fail to pass the test.Therefore if the one set is infrequent then all its supersets are also frequent and vice versa. This property is used to prune the infrequent candidate elements. In the beginning, the set of frequent 1-itemsets is found. The set of that contains one item, which satisfy the support threshold, is denoted by . In each subsequent pass, we begin with a seed set of item sets found to be large in the previous pass. This seed set is used for generating new potentially large item sets, called candidate item sets, and count the actual support for these candidate item sets during the pass over the data. At the end of the pass, we determine which of the candidate item sets are actually large (frequent), and they become the seed for the next pass. Therefore, is used to find , the set of frequent 2-itemsets, which is used to find , and so on, until no more frequent k-item sets can be found. The feature first invented by in Apriori algorithm is used by the many algorithms for frequent pattern generation. The basic steps to mine the frequent elements are as follows :

Generate and test: In this first find the 1-itemset frequent elements by scanning the database and removing all those elements from which cannot satisfy the minimum support criteria.

Join step: To attain the next level elements join the previous frequent elements by self join i.e. known as Cartesian product of i.e. This step generates new candidate k-item sets based on joining with itself which is found in the previous iteration. Let denote candidate k-item set and be the frequent k-item set.

Prune step: is the superset of so members of may or may not be frequent but all frequent item sets are included in thus prunes the to find frequent item sets with the help of Apriori property. i.e. This step eliminates some of the candidate k-item sets using the Apriori property A scan of the database.

7. EXPERIMENTAL RESULTS

Using divide and conquer strategy, we converted taste profile dataset into a format that can be used by the association rule mining algorithm. Due to limitations in JVM we have considered a dataset of 50,000 records for which we got about 476 users and 2264 songs with the constraint mentioned above. Next splitting the database using holdout method into 75% as training data i.e 357 records & 25% as test data as 119 records. Generate the frequent itemsets and association rules for the splitted datasets with the thresholds support as 15% and confidence as 60%.

$$support(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

$$confidence(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

After the association rules are formed then the antecedent of rules in training set are compared with the antecedent part of rules in test set and precision is checked then recommendations are done finally to the users in the test set.

THRESHOLD VS CLUSTERS

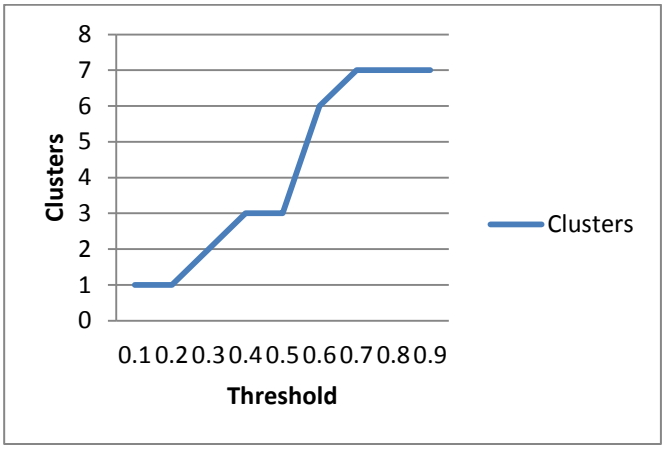
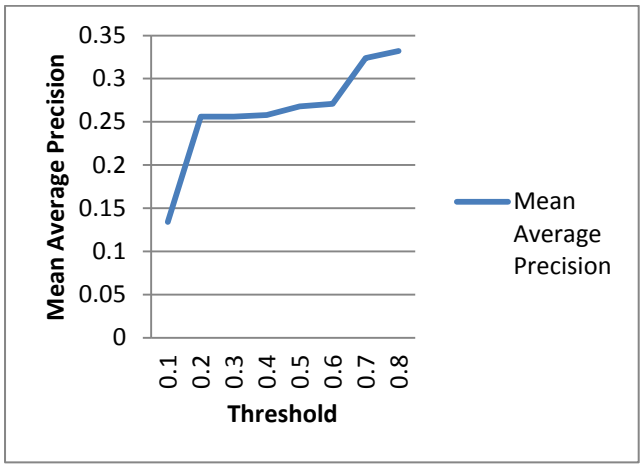


Fig 7.1 Threshold vs Clusters

Threshold x=[0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9]

Clusters y= [1 1 2 3 3 6 7 7 7];

THRESHOLD VS MEAN AVERAGE PRECISION



CONCLUSION

In this paper we proposed a method to solve cold start problem in recommender system. The system is implemented and experimentation is done with the data available. The dataset used is the Taste Profile dataset collected from the Million Song Dataset. The dataset size is 10M.For the experiment only a subset of the dataset is used because of the computational constraints. Using the results a comparative study of the quality of recommendations when association rule technique is applied and when combination of association rule and clustering technique is used rather than applying association rule only.

REFERENCES

[1] Adomavicius, G., Tuzhilin, A. 2005. Toward the Next Generation of Recommender Systems:A Survey of the State-of-the-Art and Possible Extensions. IEEE Transactionson Knowledge and Data Engineering 17, 734–749.

[2] Schein, A.I., Popescul, A., Ungar, L.H., Pennock, M. 2002. Methods and Metrics forCold-Start Recommendations. In: 25th Annual International ACM SIGIR Conferenceon Research and Development in Information Retrieval (SIGIR'02). Pp.253–260. Tampere, Finland.

[3] Middleton, S.E., Alani, H., Shadbolt, N.R., Roure, and D.C.D. 2002. Exploiting Synergybetween Ontologies and Recommender Systems. In: The Semantic Web Workshop,World Wide Web Conference (WWW'02). pp. 41–50. Hawaii, USA.

[4] Ziegler, C.N., Lausen, G., Schmidt-Thieme, L. 2004.Taxonomy-driven Computationof Product Recommendations. In: International Conference on Information andKnowledge Management (CIKM'04). pp. 406–415.Washington D.C., USA.

[5] Leung, C.W., Chan, S.C., Chung, F. 2007. Applying Cross-level Association Rule Miningto Cold-Start Recommendations. In: IEEE/WIC/ACM International Conferenceon Web Intelligence and Intelligent Agent Technology - Workshops. pp. 133–136. Silicon Valley, California, USA. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] Pasquier, N., Taouil, R., Bastide, Y., Stumme, G. 2005.Generating a Condensed Representationfor Association Rules. Journal of Intelligent Information Systems 24,29–60.

[7] Shaw, G., Xu, Y., Geva, S. 2008. Eliminating Association Rules in Multi-level Datasets.In: In 4th International Conference on Data Mining (DMIN'08). pp. 313–319. LasVegas, USA.

[8] Shaw, G., Xu, Y., Geva, S. 2008. Extracting Non-Redundant Approximate Rules fromMulti-Level Datasets. In: In 20th IEEE International Conference on Tools withArtificial Intelligence (ICTAI'08). pp. 333–340. Dayton, Ohio, USA.

[9] J. Herlocker, J. Konstan, B. A., and J. Riedl. 1999. An algorithmic framework for performing collaborative filtering. In Proc.ACM-SIGIR Conf., pages 230–237.

[10] Q. Li and B. Kim. 2003. An approach for combining content-based and collaborative filters. In Proc. of IRAL2003.

[11] J. B. Schafer, J. Konstan, and J. Riedl. 1999. Recommender systemsin e-commerce. In Proc. 1st ACM. Conf. on ElectronicCommerce (EC'99).

[12] http://labrosa.ee.columbia.edu/millionsong/tasteprofile

[13] Qing Li, Byeong Man Kim. 2003. Clustering Approach for Hybrid Recommender System, Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03)© 2003 IEEE