RESEARCHERID
THOMSON REUTERS

[Agarwal * *et al.*, 6(10): October, 2017]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

EDUCATIONAL DATA SETS AND TECHNIQUES OF RECOMMENDER SYSTEMS: A SURVEY

Ajay Agarwal[*1] & Minakshi[2]
[*1&2]KIET Group of Institutions, Ghaziabad

DOI: 10.5281/zenodo.1036288

ABSTRACT

Due to growth of World Wide Web, enormous data are created. To get the information out of available data it is necessary to store these data in a particular format. These formatted data are called datasets. These datasets are important for extracting information in such a way so that decision can be taken to recommend the trend embedded in the datasets. In addition, they can be used to test and train many information processing applications. A general practice to use available datasets obtained from different application environments is to evaluate developed recommendation techniques. Such techniques, in turn, are used as benchmarks to develop new recommendation techniques and compare them with other techniques under same applications. In this paper, we explored available public datasets collected for educational applications. These data sets can be used to evaluate and compare the performance of different recommendation techniques for learning. From basic techniques to the state-of-the-art, this paper also attempts to explore recommendation techniques, which can be served as a roadmap for research and practice in this area.

KEYWORDS: Cold Start, Collaborative Filtering, Content-Based Recommendation, Recommendation System, Sparsity Problem.

I. INTRODUCTION

There is an exponential growth of internet data. If these data are stored in form of data-sets then they can be made useful and meaningful. Some of these datasets are already publicly available, whereas others are still under preparation and are not yet publicly accessible. In this paper, we have shown datasets collected as a result of comprehensive survey. This paper provides information on educational datasets that include usage related data such as ratings, tags, reads or downloads. Such sets may form a basis to demonstrate and evaluate recommendation techniques.

Usually, the users rely on search engine to obtain the information. On the internet one gets overwhelming number of choices. There is a need to filter and prioritize such choices. As a result, relevant information may be efficiently delivered to user to lessen information overload. Recommendation systems are a useful alternative to search techniques since they help users to discover items users might not have found by themselves. Recommendation systems are programs which attempt to predict items that a user may be interested in. These information filtering system seek to predict 'rating' or 'preference' that a user would give to an item (such as music, books or movies) or social element (e.g. people or group) they had not yet considered. Such filtering can be categorized as Editorial, Simple aggregates (Top 10, Most popular, Recent uploads) and tailored to individual users (Amazon.com: for Books, CDs, Netflix and Movielens: for Movies). Recommender systems are beneficial to both users and service providers [1]. Such systems reduce costs of finding and selecting items, e.g., in an online shopping environment [2]. They have also improved decision making process and quality [3]. In e-commerce area, recommender systems enhance revenues, for the fact that they are effective means of selling more products [3]. These systems allow users to move beyond catalog searches in scientific libraries. Therefore, the need to use efficient and accurate recommendation techniques that will provide relevant recommendations for users cannot be ruled out.

In Section 2, background information regarding dataset and recommender system are provided. Recommender system is defined more formally in Section 3. Section 4 discusses different approaches of Recommender System. Problems faced using recommender system are discussed in Section 5. Section 6 introduces different educational data-sets publicly available and their analysis on various parameters. Concluding remarks are given in Section 7.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

II. BACKGROUND

A dataset is a collection of data. It may be used to train and test a new system under development. As new systems work on data, it is necessary to validate and verify their behavior with sufficient datasets prior to their deployment. As more and more leaning applications are data-driven, high-quality datasets have become critical for training these applications. Most of the recommender systems use such datasets for giving ranking/preferences for items users may be interested in.

Many approaches like collaborative filtering, content based filtering, or hybrid and many others [4] have been used to provide recommendations. "Collaborative Filtering" was introduced by Goldberg et al (1992). It uses rating structure. In content based filtering (Basu et al,1998), items or services are recommended on basis of user's previous actions. Different techniques and approaches are there to provide recommendations that may either use rating information or content information. However, both (collaborative and content based) types of filtering faces certain limitations. To overcome these limitations, Pazzani has attempted by proposing hybrid approach that combines both rating as well as content information. Recommender system will remain an active research area. This includes disciplines like data mining, information retrieval, context awareness, personalization and group recommendations.

III. RECOMMENDER SYSTEM

Recommender System is an intelligent system. It makes suggestion about items that might interest to the users. Some of the practical applications of these systems include recommending books, CD on flipkart, movies by Movielens, music by last.fm. The formal definition of recommender system is:

C: The set of all users
S: The set of all possible items that can be recommended (such as books, movies, or restaurants).
U: A function that measures utility (usefulness) of a specific item s $\in$ S to user c $\in$ C, i.e., U: C×S $\rightarrow$R, where R is a set of positive integers or real numbers in a predefined range.

Note that the space *S* of possible items and C of possible users can be very large. In recommender system, for each user c $\in$ C, we select item s $\in$ S that *maximizes* the user's utility. In Table 1, each cell U(u, i) corresponds to the ratings of user u for item i. The task is to predict the missing rating U(a, i) for active user a. More formally, in recommender systems the utility of an item is usually represented by a *rating*, which indicates how a particular user liked a particular item. In general, the range of possible ratings a user can give to an item is 1 (minimum, *disliked*) to 5 (maximum, *higlhy liked*).

Table 1. User rating matrix. R={1(dislike),2,3,4,5(highly liked)}

	item1	item2	…	item i	…	item m
user 1	5	3		1	2	
user 2		2				4
user 3	3	4	5	2	1	
⋮						
user u					4	
⋮						
user n				2		
user a	3	5		?	1	

Usually, rating is not done on a complete dataset or space C×S and thus only rating on subset is available. Empty cells mean that user has not yet seen the associated item. The main aim of a recommender system is to predict ratings of the non-rated user/item combination and thus providing appropriate recommendations. Recommender system may either provide the highest estimated rating item or alternatively provide a list of top N items as recommendation to a user or set of users.

IV. RECOMMENDATION SYSTEM APPROACHES

On the basis of their approach to rating estimation, recommendation systems are usually classified:

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

- Content-based System (CB)
- Collaborative Filtering System(CF)
- Hybrid System

Content-Based Recommendation Systems

Content-based recommendation systems depend on similarities in order to make recommendations. Note that the similarity measurement is restricted to the history of the user's interests with which the item recommendation is to be made. Specifically, content-based recommender systems calculate the similarity between items unseen by the user with those that the user liked in the past based on their descriptions [5]. Those items with the best-matching are recommended to the user.

For example, user X may have given higher ratings to a number of books by author Y. The system will in time learn to recommend more books by author Y to user X. The author feature is in this case being used to measure the similarity between items seen by the user and those unseen, and recommend to the user unseen items that have a higher similarity score. Other book features include plot, genre, character, form and setting. Content-based filtering is generally employed in systems that offer text-based items such as books, news and documents [6]. These systems will normally store an item profile of each item which can be recommended to the user. Naturally, this item profile will consist of all (or most) of the features that characterize the items on offer. For example, a book item profile may have all the book features mentioned earlier. Information about items the user has liked in the past is stored in a user history. The user history will at least contain the rating given by a user on each item they have liked. Formally, a content-based recommender system may be represented as follows [6]:

let IProfile(s) be the item profile of item s, i.e., the set of features characterizing item s. Also, let UHistory(c) be the user history of user c, containing items that the user has liked in the past. The utility function u(c, s) can then be defined as in equation (1):

u(c, s) = sim (UHistory(c), IProfile(s)) (1)

The sim function in equation (1) can be a similarity function such as the cosine similarity measure defined in equation (2). Cosine similarity is a measure of similarity between two non-zero vectors $\vec{X}$ and $\vec{Y}$ of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1(implies exactly similar), and it is less than 1 for any other angle.

$$\cos(\vec{X},\vec{Y}) = \frac{\vec{X}\,.\vec{Y}}{\|\vec{X}\|\|\vec{Y}\|} = \frac{\sum(\mathrm{Xi}\mathrm{Yi})}{\sqrt{\{\sum(\mathrm{Xi})^2\}}\sqrt{\{(\mathrm{Yi})^2\}}} \qquad (2)$$

Similarity measures may be some other heuristics. Other than the traditional heuristic methods which depend on a formula to calculate the utility prediction, machine learning techniques and Bayesian classifiers such as clustering, neural networks and decision trees can also be used in content-based recommendation. These techniques employ a slightly different approach. They calculate the utility prediction using a model learned from the user and item data.

Research related to content based recommendations has been focused on recommending items with associated textual content. These are web pages, books etc. This problem, treated as *information retrieval* task, describes the user's preference and on basis of similarity with this query, unrated documents are scored.

As content in text based system is usually described with keywords, the "importance" of word ki in document dj is determined with some weighting measure wij that can be defined in many different ways. One of the best-known measures for specifying keyword weights in Information Retrieval is the term frequency/inverse document frequency (TF-IDF) measure [7]. Term frequency (TF) is calculated by simply counting the number of times the word is found in the target document and has been shown in equation (3). For example, assume we're calculating for the term "sun" in document having line "We can see the shining sun, the bright sun" then TF(sun)=2. Inverse document frequency (IDF) is a count of how many documents in the entire corpus contain the term. The calculation of IDF for a term t is shown in equation (4). Suppose we have a corpus of 100 documents with 20 of those documents containing the word "sun". The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document) (3)
IDF(t) = log_10(Total number of documents / (1+Number of documents with term t in it)) (4)
IDF(sun)=$\log_{10}(100/(1+20))=\log_{10}(4.7619)=0.6778$
TF*IDF weight= 2*0.6778=1.3556

Nearest neighbor technique is another popular technique to provide recommendations on the basis of textual information stored in memory (i.e. training data). To classify a new unlabelled item, the technique determines the nearest neighbor or k nearest neighbors using a similarity function (Euclidean or cosine similarity according to the type of textual information) and comparing it to all stored values. The class of the unseen item can then be determined from the class labels of nearest neighbors.

Collaborative Filtering
Collaborative filtering (CF), also referred to as social filtering systems collects user feedback in the form of ratings for items in a given domain. It is based on the idea that people who agreed in their ratings of certain items in the past are likely to agree again in the future. A person who wants to purchase a costly product, for example, might ask for recommendations from friends. The recommendations of some friends who have similar interests are trusted more than recommendations from others. This information is used in the decision on whether to purchase or not. Therefore, CF exploits similarities in rating behavior amongst several users in determining how to recommend an item. Based on item ratings given by other users, collaborative filtering predicts the utility value a user would assign to an item [6]. CF problem can more formally be formulated as follows:

Let C be the set of all users of similar taste and S the set of all possible items that can be recommended. Let n be the number of users in set C. We define u(c, s), a utility function that measure all the ratings assigned to item s by each user c_j where $\forall j$: $0 \leq j \leq n : c_j \in C$. The utility measure u(c, s) of item s by user c can then be predicted as follows:

Mean rating for user i can be calculated as in equation (5), which is further used for predicting rating of the active user c. Predicted rating of the active user c has been calculated in equation (6) where R(i,j) is the rating of user i on item j and sim (a,i) are similarity measures of users in set C and k is a normalizing factor and is usually calculated as $\sum_{j\in C}|\mathrm{sim}(a,i)|$. The most commonly used similarity measures are cosine and correlation coefficient similarity [6].
avg(Ri)= $(\sum_{j\in I(i)} R(i,j))/|I(i)|$ where I(i) is set of items ranked by user i. (5)
p(c,j) = avg(Ra) + $(\sum_{j\in C}(R(i,j) - \mathrm{avg}(Ri))*\mathrm{sim}(a,i)/k$ (6)

Hybrid Recommendation Systems
Several researchers have attempted to combine collaborative filtering and content based approaches in order to smoothen their disadvantages and gain better performance while recommendations. Depending on domain and data characteristics, several hybridization techniques are possible to combine CF and CB techniques which may generate different outputs. Some of the techniques are weighted; feature augmentation, feature combination, mixed, switching, cascade etc. Different ways of hybridization are[9,10,11]:

- Solving CF and CB separately and combine their predictions.
- Incorporating some content based characteristics into collaborative approach.
- Incorporating some collaborative characteristics into content based approach.
- Constructing a general unifying model that incorporates both content-based and collaborative

V. PROBLEMS OF RECOMMENDATION SYSTEMS

Various methods used in a recommender system have some of the hurdles:

Sparsity Problem
It is one of the key problems faced by recommender system and has great influence on the quality of recommendation. The reason being that as the numbers of users and items increases the the user-item matrix dimensions gets increased which implies sparsity of ratings in it. As Collaborative filtering is dependent over the rating matrix, it suffers mainly from this problem. Many researchers [12], [13], [14] have attempted to alleviate this problem; still this area demands further research.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Cold Start problem

This refers to the situation when a new user or item enters the system. Three types of cold start problems are: *new user problem*, *new item problem* and *new system problem*. In these cases, it becomes complex to provide recommendation as in case of new user, there is very fewer information available about user. For a new item, no ratings are generally available and thus CF cannot make useful recommendations in case of new item as well as new user. However, content based methods can provide recommendation in case of new item as they do not depend on any previous rating information of other users to recommend the item. In *new system*, the information about user as well as item is required.

Scalability

To handle growing amount of information in a graceful manner is called as scalability of a system. With tremendous growth in information over internet, it is evident that the recommender systems are having enormous data. Definitely it is a great challenge to handle continuously growing data. In CF, computations grow rapidly and become expensive which may lead to inaccurate results sometimes. Proposed techniques for handling this scalability problem and speeding up recommendation formulation are based on approximation mechanisms. Even if they improve performance, most of the time they result in reducing accuracy [15]. **Over Specialization**

Problem

A content-based filtering system will not select items if the previous user profile [16] does not provide evidence for this. It prevents user from discovering new items and other available options. Additional techniques have to be augmented to give the system the capability to make suggestion outside the scope of what the user has already shown interest in. The reason is diversity of recommendations is a required feature of all recommendation system.

VI. DATA SETS USED BY RECOMMENDER SYSTEM

To evaluate recommendation techniques, the practice is to use publicly available datasets obtained from different application environments (e.g. Book-Crossing, MovieLens, or Each Movie). In given settings [17], these datasets are used as benchmarks to (i) develop new recommendation techniques and to (ii) compare them with other techniques. In such datasets, a representation of selected item is stored using implicit or explicit feedback obtained from users. This feedback allows the recommender system to produce a recommendation. Depending on the filtering approach, this feedback can be in several forms. For example, in the case of collaborative filtering systems, it can be ratings or votes (i.e. if an item has been viewed or bookmarked). In the case of content-based recommenders, it can be product reviews or simple tags (keywords) that users provide for items. Additional information is also required, such as a unique way to identify who provides this feedback (user identifier) and upon which item (item identifier). The user rating matrix used in collaborative filtering is a well-known example [18]. Several educational datasets have been collected as a result of survey. Some of these datasets are already publicly available, whereas others are still under preparation and not yet publicly accessible. In this paper we have selected 29 educational datasets explained below from D1 to D29 [19].

D1. Anonymous Microsoft Web Data: The data was created by sampling and processing the www.microsoft.com logs. The data records the use of www.microsoft.com by 38000 anonymous, randomly-selected users. For each user, the data list all the areas of the web site (Vroots) that the user visited in a one week timeframe.
D2. Artificial Characters: Dataset artificially generated by using first order theory which describes structure of ten capital letters of English alphabet.
D3. Computer Hardware: Relative CPU Performance Data, described in terms of its cycle time, memory size, etc.
D4. Internet Advertisements: This dataset represents a set of possible advertisements on Internet pages.
D5. Meta-Data: It was used in order to give advice about which classification method is appropriate for a particular dataset (taken from results of Statlog project).
D6. Optical Recognition of Handwritten Digits: An extraction of normalized bitmaps of handwritten digits from a preprinted form. From a total of 43 people, 30 contributed to the training set and different 13 to the test set. 32x32 bitmaps are divided into non-overlapping blocks of 4x4 and the number of pixels is counted in each block. This generates an input matrix of 8x8 where each element is an integer in the range 0 to 16. This reduces dimensionality and gives invariance to small distortions.
D7. Pen-Based Recognition of Handwritten Digits: the digit database was created by collecting 250 samples from 44 writers. The samples written by 30 writers are used for training, cross-validation and writer dependent testing, and the digits written by the other 14 are used for writer independent testing. This database is also available in the UNIPEN format.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

D8. Teaching Assistant Evaluation: The data consist of evaluations of teaching performance; scores are "low", "medium", or "high"
D9. Tic-Tac-Toe Endgame: Binary classification task on possible configurations of tic-tac-toe game.
D10. University: Each observation concerns one university. In some cases, more information is provided about the attribute (e.g., units or domain). Some duplicates may exist and a single observation may have more than one value for a given attribute (esp. academic emphasis).
D11. CMU Face Images: This data consists of 640 black and white face images of people taken with varying pose (straight, left, right, up), expression (neutral, happy, sad, angry), eyes (wearing sunglasses or not), and size.
D12. Internet Usage Data: This data contains general demographic information on internet users in 1997.
D13. UJI Pen Characters: Data consists of written characters in a UNIPEN-like format.
D14. Bag of Words: This data set contains five text collections in the form of bags-of-words.
D15. Dexter: DEXTER is a text classification problem in a bag-of-word representation. This is a two-class classification problem with sparse continuous input variables. This dataset is one of five datasets of the NIPS 2003 feature selection challenge.
D16. Character Trajectories: Multiple, labeled samples of pen tip trajectories recorded whilst writing individual characters. All samples are from the same writer, for the purposes of primitive extraction. Only characters with a single pen-down segment were considered.
D17. UJI Pen Characters (Version 2): A pen-based database with more than 11k isolated handwritten characters.
D18. Semeion Handwritten Digit: 1593 handwritten digits from around 80 persons were scanned, stretched in a rectangular box 16x16 in a gray scale of 256 values.
D19. Amazon Commerce reviews set: The dataset is used for authorship identification in online Writeprint which is a new research field of pattern recognition.
D20. Amazon Access Samples Amazon's InfoSec is getting smarter about the way access data is leveraged. This is an anonymized sample of access provisioned within the company.
D21. First-order Theorem Proving: Given a theorem, predict which of five heuristics will give the fastest proof when used by a first-order prover. A sixth prediction declines to attempt a proof, should the theorem be too difficult.
D22. User Knowledge Modeling: It is the real dataset consisting of the students' knowledge status about the subject of Electrical DC Machines. The dataset had been obtained from Ph.D. Thesis.
D23. BlogFeedback: Instances in this dataset contain features extracted from blog posts. The task associated with the data is to predict how many comments the post will receive.
D24. REALDISP Activity Recognition Dataset: The REALDISP dataset is devised to evaluate techniques dealing with the effects of sensor displacement in wearable activity recognition as well as to benchmark general activity recognition techniques.
D25. Student Performance: Predict student performance in secondary education (high school).
D26. Educational Process Mining (EPM): A Learning Analytics Data Set: Educational Process Mining data set is built from the recordings of 115 subjects' activities through a logging application while learning with an educational simulator.
D27. Default of Credit Card Clients: This research aimed at the case of customers' default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods.
D28. Online Retail: This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.
D29. GPS Trajectories: The dataset has been fed by Android app called Go!Track. It is available at Goolge Play Store(https://play.google.com/store/apps/details?id=com.go.router).

These twenty nine educational data sets are further classified under five properties mentioned in Table 2. The abbreviations used in Table 3 are taken from Table 2. The classification of selected data sets based on different properties namely Default Task, Attribute Type, Data Type, Number of Attributes, Number of Instances has been depicted in Figure1, Figure2, Figure3, Figure4, and Figure 5 respectively.

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

Table 2. Acronyms used in classification of properties of educational data sets.

Default Task	Attribute Type	Data Type	# Attributes	# Instances
C-Classification Reg-Regression Cl-Clustering CD-Casual Discovery	Cat-Categorical I-Integer , R-Real	M-Multivariate Seq-Sequential TS-Time-Series T-Text DT-Domain-Theory Im-Image	Less than 10, 10 to 100, Greater than 100	Less than 100, 100 to 1000, Greater than 1000

Table 3. Summary of classifications

Data set	Data Types	Default Task	Attribute Types	# Instances	# Attributes
D1	N/A	RS	Cat	37711	294
D2	M	C	Cat, I, R	6000	7
D3	M	Reg	I	209	9
D4	M	C	Cat, I, R	3279	1558
D5	M	C	Cat, I, R	528	22
D6	M	C	I	5620	64
D7	M	C	I	10992	16
D8	M	C	Cat, I	151	5
D9	M	C	Cat	958	9
D10	M	C	Cat, I	285	17
D11	Im	C	I	640	N/A
D12	M		Cat, I	10104	72
D13	M, Seq	C	I	1364	N/A
D14	T	Cl	I	8000000	100000
D15	M	C	I	2600	20000
D16	TS	C, Cl	R	2858	3
D17	M,Seq	C	I	11640	N/A
D18	M	C	I	1593	256
D19	M, T, DT	C	R	1500	10000
D20	TS, DT	Reg, Cl, CD	N/A	30000	20000
D21	M	C	R	6118	51
D22	M	C, Cl	I	403	5
D23	M	Reg	I, R	60021	
D24	M, TS	C	R	1419	
D25	M	C, R	I	649	
D26	M, Seq, TS	C, Reg, Cl	I	230318	
D27	M	C	I, R	30000	
D28	M, Seq, TS	C, Cl	I, R	541909	8
D29	M	C, Reg	R	163	15

Column charts are drawn corresponding to those properties where one or more cells of the data sets have multiple categories. For example, D20 data set has Regression, Clustering, and Casual Discovery categories under its Default Task property. In all other cases pie charts are drawn. From Table 3, it can be concluded that more than

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

79% (shown as 23 out of 29) of selected educational data sets are used for classification under default task property. As a result, such data sets can be used by a researcher to validate proposed classification technique(s).

Majority of attribute types of these data sets are integers (shown as 21 out of 29) which provide simple arithmetic as compared to real numbers. Further, most of the data type is multivariate (24 out of 29) which facilitates researchers to calculate statistical results based on multiple variables like correlation etc. Number of attributes of these data sets is more than 10 (14 out of 21). This will imply dimensional aspect of data set. It is noted that number of instances of these data sets are more than 100 is 100%. These will help researcher to validate proposed technique.

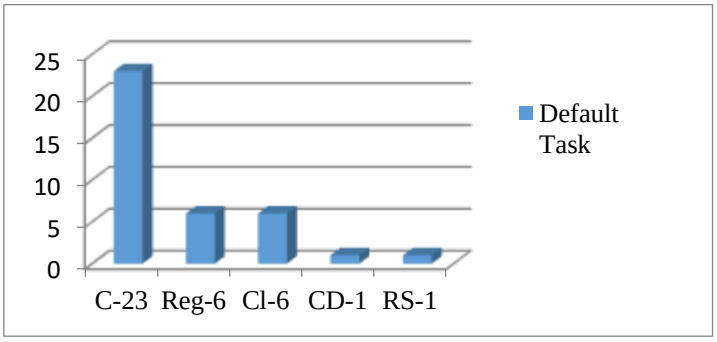

Figure1. Classification of data set based on Default Task

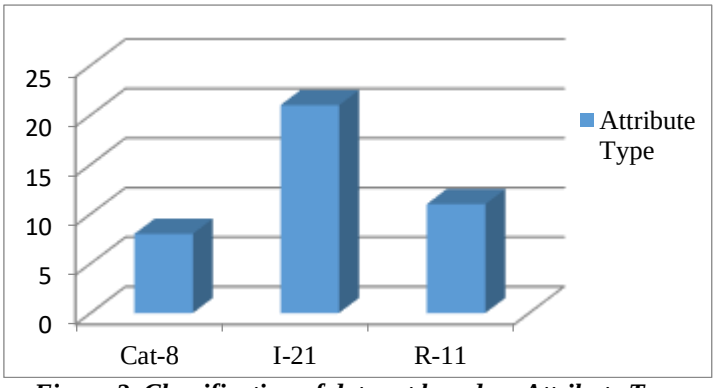

Figure 2. Classification of data set based on Attribute Type

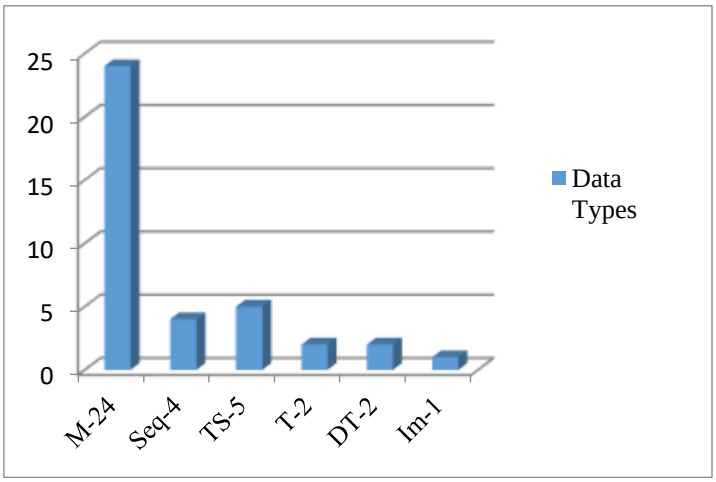

Figure 3. Classification of data set based on Data Type

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

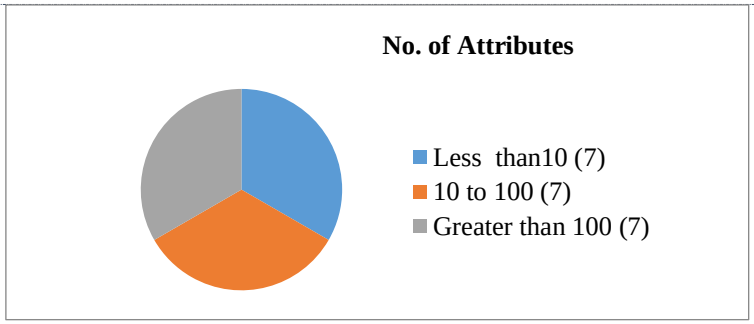

Figure 4. Classification of data set based on No. of Attributes

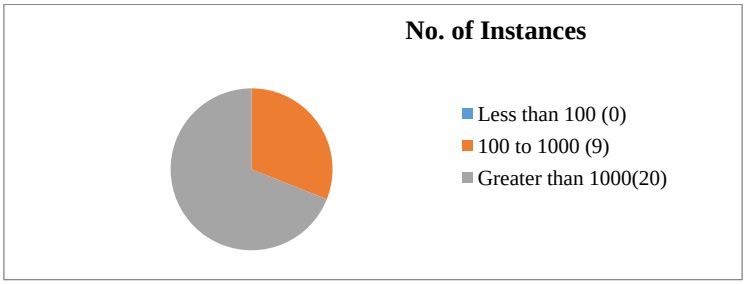

Figure 5. Classification of data set based on No. of Instances

VII. CONCLUSION

Due to tremendous growth of data over internet, data sets are being available in every field. In this paper educational datasets are mentioned and explained their purpose so that these may be analyzed by method(s) used in recommender systems. A number of recommendation systems have been proposed. These are based on content based filtering, collaborative filtering, and hybrid recommendation techniques. Most of them have been able to solve the problems by giving better recommendations. However, due to exponential growth of information, it is required to work on this research area to explore and provide new methods. So that such methods can provide recommendation in a wide range of applications while considering both the quality and privacy aspects. In a nutshell, there is a need to improve the current recommendation system for present and future requirements of better recommendation qualities.

VIII. REFERENCES

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms", In WWW '01: Proceedings of the 10th international conference on World Wide Web, New York, NY, USA, 2001.

[2] Pazzani, Michael, and Daniel Billsus. "Content-based recommendation systems." *The adaptive web* (2007): 325-341.

[3] Claypool, M., A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combining content-based and collaborative filters in an online newspaper". In ACM SIGIR'99, Workshop on Recommender Systems: Algorithms and Evaluation, August 1999.

[4] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans. On Knowl. and Data Eng., 17(6):734{749, June 2005.

[5] P. Kantor, L. Rokach, F. Ricci, and B. Shapira, *Recommender Systems Handbook*. Springer Science+Business Media, LLC 2011, 2011. [Online]Available: http://cds.cern.ch/record/1412605

[6] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*,vol. 17, no. 6, pp. 734–749, Jun. 2005. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1423975

[7] G. Jeh and J. Widom, "SimRank: a measure of structural-context similarity," *Proceedings of the eighth ACM SIGKDD internationalconference on Knowledge discovery and data mining*, pp. 1–11, 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=775126

ISSN: 2277-9655
Impact Factor: 4.116
CODEN: IJESS7

[8] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste, "A constant time collaborative filtering algorithm", Information Retrieval, 4(2):133–151, 2001.
[9] Sarwar, Badrul M., George Karypis, Joseph A. Konstan and John T. Riedl, "Application of Dimensionality Reduction in Recommender System – A Case Study", In ACM WebKDD Workshop, 2000.
[10] Boddu Raja Sarath Kumarmaddali and Surendra Prasad Babuan:, "Implementation of Content Boosted Collaborative Filtering Algorithm", IJEST.
[11] J. Wang, A. P. de Vries, and M. J. T. Reinders, "Unifying User-Based and Item-Based Collaborative Filtering Approaches by Similarity Fusion". In SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY,USA, 2006.
[12] C. Desrosiers and G. Karypis, "Solving the Sparsity Problem: Collaborative Filtering via Indirect Similarities", Technical Report, Dec 2008.
[13] Zhou, Jia, and Tiejian Luo. "A novel approach to solve the sparsity problem in collaborative filtering." *Networking, Sensing and Control (ICNSC), 2010 International Conference on*. IEEE, 2010.
[14] [14] Yibo Chen, chanle Wu, Ming Xie and Xiaojun Guo, "Solving the Sparsity Problem in Recommender Systems Using Association Retrieval", Journal of Computers, Vol. 6, No. 9, September 2011.
[15] [15] M. Morita and Y. Shinoda. Information f i ltering based on user behavior analysis and best match text retrieval. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '94, pages 272–281, New York, NY, USA, 1994. Springer-Verlag New York, Inc.
[16] N. Manouselis, R. Vuorikari, and F. V. Assche. Simulated Analysis of MAUT Collaborative Filtering for Learning Object Recommendation. In Workshop proceedings of the EC-TEL conference: SIRTEL07 (EC-TEL07), pages 27–35, 2007.
[17] H. Drachsler, T. Bogers, R. Vuorikari, K. Verbert, E. Duval, N. Manouselis, G. Beham, S. Lindstaedt, H. Stern, M. Friedrich, and M. Wolpers. Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. Procedia Computer Science, 1(2):2849-2858,2010.
[18] Sarwar, Badrul M., George Karypis, Joseph A. Konstan and John T. Riedl, "Application of Dimensionality Reduction in Recommender System – A Case Study", In ACM WebKDD Workshop, 2000.
[19] Publicly available data sets: https://archive.ics.uci.edu/ml/datasets.html
[20] T. Raghunadha Reddy, B. Vishnu Vardhanb, P. Vijayapal Reddy, " A Document Weighted Approach for Gender and Age Prediction Based on Term Weight Measure", ***International Journal of Engineering-Transactions B: Applications*** Vol. 30, No. 5, (May 2017) 643-651
[21] Hamidi, H. and Daraee, A., "Analysis of pre-processing and post-processing methods and using data mining to diagnose heart diseases", ***International Journal of Engineering-Transactions A: Basics***, Vol. 29, No. 7, (2016), 921-929.
[22] Darvishi, A. and Hassanpour, H., "A geometric view of similarity measures in data mining", ***International Journal of Engineering-Transactions C: Aspects***, Vol. 28, No. 12, (2015), 1728-1735

CITE AN ARTICLE

Agarwal, A., & M. (2017). EDUCATIONAL DATA SETS AND TECHNIQUES OF RECOMMENDER SYSTEMS: A SURVEY. *INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY*, *6*(10), 434-443.