# Application of Clustering Algorithms in Library Management System

Palak Singhal
16CO129
Computer Science
NITK Surathkal, India
smarty1palak@gmail.com

Sharanya Kamath
16CO140
Computer Science
NITK Surathkal, India
sherrykamath@gmail.com

**Abstract**
Recommender systems are powerful new technology mainly used for providing automatic personalized suggestion based on user's past history. It recommends products to the end users that are most appropriate. This paper describes the use of clustering algorithms for implementing book recommendation feature in library management systems based on features of content based filtering (CBF). Few of the major algorithms presently used in different kinds of recommendation systems are analyzed, and their properties and drawbacks are listed. A total of 42 papers were analyzed as part of the literature review. The final objective would be to find out the algorithm which works well for the library database to be used later and reduce the running time of the algorithm and to improve upon the recommendations.

**Keywords:**
Book recommendation, clustering, algorithms, content based filtering, python clusters implementation, Systematic Literature Review, Library management systems.

## Ⅰ. INTRODUCTION

With the advent of technology and overflooding data it has become extremely difficult for the users to look for the information they prefer. Information filtering systems of which recommender system is a subpart helps in decision making by providing personalized suggestions to the users based on their past preferences and history.

The paper focuses on analyzing the readers' borrowing records by using the following techniques: data analysis and data mining.Subject headings of borrowed books by the user will be used as the basis for generating pertinent recommendations which would appear in the user's feed.
In data mining, Clustering is the most popular, powerful and commonly used unsupervised learning technique used for data mining.It clusters the users in dataset according to their preferences and then a recommender algorithm is applied to each of these clusters that will cater to the interests of the users. A content based algorithm is used where system recommends items that a user has preferred in the past by matching them with the characteristics of the item and generate new items of potential interest to this particular user, hence making the recommendations more customized.
Each cluster would be assigned to typical preferences, based on preferences of customers who belong to the cluster. Customers within each cluster would receive recommendations computed at the cluster level.
This paper presents a comprehensive review of literature related to application of clustering algorithms for recommendation systems in academic journal between 2001 till present.
The rest of the paper is organized as follows: section two deals with the research methodologies and the research questions whose solution will be looked upon in section three. Section four deals with the conclusion and

the literature review ends with all the references used for the review.

## Ⅱ. RESEARCH METHODOLOGIES

The literature review will be carried out in
The following three phases:
1. Planning the review
2. Conducting the review
3. Reporting the review

### *2.1 PLANNING THE REVIEW*

In the planning phase we decided the following research questions that need to be answered in the review.

**RQ1. What clustering algorithms are being used in recommender systems specifically book recommender systems?**

**RQ2. What are the advantages and disadvantages of choosing a particular clustering algorithm in a recommender system?**

The review protocol is described in the next sections**.**

### *2.2 CONDUCTING THE REVIEW*

In the conducting phase a search protocol was created where research paper were searched based on the keywords identified in the research questions addressed in the planning phase. Some extra keywords were added to make the search more comprehensive and to ensure that proper research papers were captured.
The search presented more than 7000 journals, articles, research papers and blogs. The reference list for the papers. To reduce the number of searches so that proper analysis could be done proper keywords were added to the already used keywords and bigger sentences pertaining to the research questions were searched instead.

#### *2.2.1 PRIMARY STUDY SELECTION CRITERIA*

All the texts obtained from the above search which were in languages other than english were removed.
Studies related to recommender system with collaborative filtering, or with the use of any algorithm other than content based filtering were removed.

#### *2.2.2 QUALITY ASSESSMENT FOR QUANTITATIVE STUDIES*

Quality of the papers used was tested. Only papers from valid publications like IEEE, ACM, Springer were read as others might present a threat to the validity of data.

#### *2.2.2 DATA EXTRACTION*

The properties of the research questions were identified and tabulated

### *2.3 REPORTING THE REVIEW*

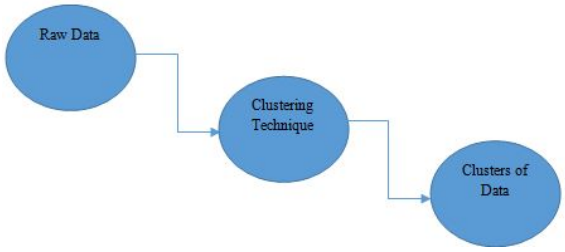
Some unwanted keywords like collaborative filtering, voting algorithm etc popped up while searching which might have some effect on the research questions we started with. Also there is no proper abstraction service which could help in search.

## Ⅲ. ANALYSIS OF RESULTS

**RQ1 .What clustering algorithms are being used in recommender systems specifically book recommender systems?**

*Clustering methodologies:*

Clusters divide data into groups so that we can gather beneficial and meaningful information. To obtain meaningful information clusters should capture the natural structure of data. Hence it is very critical how data is clustered.




After having decided on the clustering variables we need to decide on the clustering procedure to form our groups of objects. These approaches are: square error based, k-means, fuzzy c-means, hierarchical methods, gaussian probabilistic etc.

A lot of research papers, journals, blogs etc were read for different types of recommender systems used for various purposes like movie recommendations, book recommendations etc and specifically four major clustering algorithms could be identified which are described as follows:

***Square error based clustering***

This clustering is used to obtain a partition Such that for some fixed number of clusters it minimizes the square error. Square error can be defined as the sum of the Euclidean distances between each pattern and its cluster center. The most common example of this type of algorithm is k-means clustering.

***K-means clustering:***

It is the most popular clustering tool that is used in scientific and industrial applications. It is clustering methodology which involves partitioning of 'n' data objects into k clusters in which each data object belongs to the cluster with the nearest mean.

**Process:**

1. The initial seeds with the chosen number of clusters, *K*, are selected and an initial partition is built by using the seeds as the centroids of the initial clusters.
2. Each record is assigned to the centroid that is nearest, thus forming a cluster.
3. Keeping the same number of clusters, the new centroid of each cluster is calculated.
4. Iterate Step (2) and (3) until the clusters stop changing or stop conditions are satisfied.

Time complexity is O(nkl) where 'n' is the number of data objects, 'k' represents the number of clusters, and 'l' are the number of iterations.
Space complexity is O(k + n).

Some papers preferred k-means clustering (or variations of k-means) because of the algorithm's ability to work on large datasets efficiently. It is not a very complex algorithm which can be used in a straightforward manner. So papers on recommendation systems which did not have too many features, and most of the clustering was based on numeric values, used k-means clustering algorithms.

***Fuzzy c-means clustering:***

Clustering can be either hard or fuzzy type. In the first category, the patterns are distinguished in a well defined cluster boundary region. But due to the overlapping nature of the cluster boundaries, some class of patterns may be specified in a single cluster group or dissimilar group. This property limits the use of hard clustering in real life applications. To reduce such limitations fuzzy type clustering came into the picture.

**Equations:**

Let us consider a set of n vectors *(X=(x1,x2,…,xn)2≤c≤n )* for clustering into c groups.
Each vector $x_i \in R^s$ is described by s real valued measurements which represent the features of the object $x_i$.
A membership matrix known as Fuzzy partition matrix is used to describe the fuzzy membership matrix. The set of fuzzy partition matrices *(c×n)* is denoted by $M_{fc}$ and is defined in as

$$M_{fc} = \{W \in R^{cn} | w_{ik} \in [0,1], \forall i,k; \sum_{i=1}^{c} w_{ik} = 1, \forall k; 0 < \sum_{k=1}^{n} w_{ik} < n, \forall i\}$$

where $1 \le i \le c$ , $1 \le k \le n$

.

The objective function of the fuzzy c-means algorithm is computed by using membership value and Euclidean distance

$$J_m(W,P) = \sum_{\substack{1 \le k \le n \\ 0 \le i \le c}} (w_{ik})^m (d_{ik})^2$$

where $(d_{ik}) = || x_k - p_i ||$

where $m \in (1, +\infty)$ is the parameter which defines the fuzziness of the resulting clusters and $d_{ik}$ is the Euclidian distance from object $x_k$ to the cluster center $p_i$.

The minimization of the objective function $J_m$ through FCM algorithm is being performed by the iterative updation of the partition matrix using

$$p_i = \sum_{k=1}^{n} (w_{ik})^m x_k / \sum_{k=1}^{n} (w_{ik})^m$$

$$w_{ik}^{(b)} = \sum_{j=1}^{c} 1/[(d_{ik}^{(b)}/d_{jk}^{(b)})^{2/m-1}]$$

The FCM membership function is calculated as:

$$\mu_{i,j} = \left[ \sum_{t=1}^{c} \left( \frac{\|x_j - v_i\|_A}{\|x_j - v_t\|_A} \right)^{\frac{2}{m-1}} \right]^{-1}$$

$\mu_{i,j}$ is the membership value of *j*th sample and *i*th cluster. The number of clusters is represented by c, $x_j$ is the *j*th sample and $v_i$ cluster center of the *i*th cluster. $//\ //_A$ represents the norm function.

**Process:**

1. Initialize the number of clusters c.
2. Select an inner product metric Euclidean norm and the weighting metric (fuzziness).
3. Initialize the cluster prototype $P^{(0)}$, iterative counter b = 0.
4. Then calculate the partition matrix $W^{(b)}$.
5. Update the fuzzy cluster centers $P^{(b+1)}$.
6. If $||P^{(b)} - P^{(b+1)}|| < \varepsilon$ then stop, otherwise repeat step (2) through (4).

According to most of the papers the fuzzy c means recommender system provides better similarity metrics and quality than the ones provided by the existing clustering algorithms in recommender system but the computation time taken by the proposed recommender system is more. To solve this problem clustered data points can be taken as input datasets. For computing the quality measures of the existing systems it has been first implemented and then the results obtained were compared with the results of the proposed methodology.
Using this algorithm showed a lot of improvement in precision and accuracy, This proposed recommender system can be used in all content based filtering systems.

***Hierarchical clustering***

This algorithm works by dividing the data set into various smaller subsets in a hierarchical manner. It groups the data instances into a tree of clusters. There are two major methods that are available under this category:

**1. Agglomerative** : It forms clusters in bottom up fashion until all data instances belong to the same cluster i.e starts with N clusters and finally merges them into one cluster. Here the closest pair of clusters are merged.

**2. Divisive** : It splits up the data set into smaller clusters in a top down manner until each of cluster contains only one instance i.e starts with 1 cluster and divides repeatedly until singleton clusters are formed. If N data objects are present then $2^{N-1} - 1$ two subset divisions are possible which is computationally expensive.

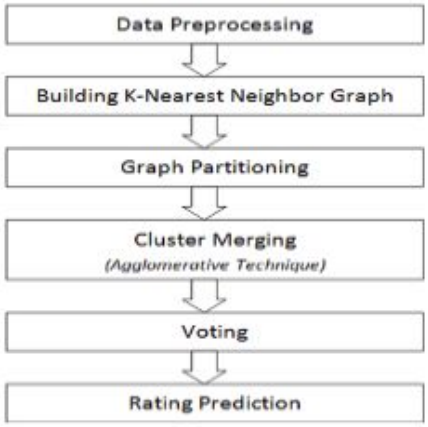
**Algorithm:**

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

**Complexity:**

Time complexity : $O(n^2)$

Space complexity: $O(n^2)$




A lot of papers were found to prefer hierarchical over k-means because Hierarchical clustering algorithm continuously produced lesserror as compared to K-Means.Although the error produced by it is lesser than k-means, the running time is quite high. So it has scope to reduce the running time which has not been done yet with the help of some parallel technologies like map reduce.

***Gaussian ( EM ) clustering***

Estimation via Mixture (EM) algorithm assumes apriori (after theoretical deduction) that there are *'n'* Gaussian and then the algorithm tries to fit the data into the *'n'* Gaussian by expecting the classes of all data point and then maximizing the maximum likelihood of Gaussian centers.
Data points are generated from density functions like multivariate gaussian etc.
Suppose the prior probability(mixing probability) is $P(C_i)$ for cluster $C_i$, i= 1, . . .k
(k is assumed to be known here) and
$p ( X | C_i \Theta_i )$ = Conditional probability density are already known. Then mixture probability density for the whole set is

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \sum_{i=1}^{K} p(\mathbf{x} \mid C_i, \boldsymbol{\theta}_i) P(C_i)$$

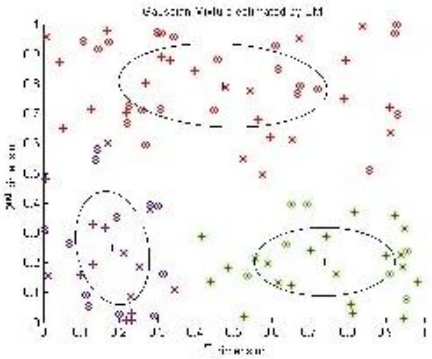
The EM algorithm algorithm tries to generate a series of estimates for the parameter theta and tries to reach the convergence criteria (t). It follows the following algorithm to do that:

**Algorithm**:

1. Initialize $\Theta^0$ and set t=0.
2. Compute the expectation of complete data log likelihood output.

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = E[\log p(\mathbf{x}^g, \mathbf{x}^m \mid \boldsymbol{\theta}) \mid \mathbf{x}^g, \boldsymbol{\theta}^t];$$

3. Select a new parameter estimate that maximizes the Q function.
4. Increment t by 1 and repeat steps 2-3 until the condition for convergence is satisfied.



*Result of gaussian algorithm for dataset of size N=60*

A lot of papers have made use of gaussian process model for recommender system also . It might appear computationally expensive because it calculates inverse of Σ where Σ is a p-by-p matrix, where p could be tens of thousands.However, in real-world recommender systems, we never need to compute or evaluate $\Sigma^{-1}$. All we need is to compute is $\Sigma^{-1}_{00}$ efficiently Since the number of movie items that each user can rate may only vary from several dozens to 200, which means the size of $\Sigma_{00}$is at most 200-by-200. Inverting a matrix of this size is computationally trivial, thus the prediction can be done very quickly and efficiently.

**RQ2. What are the advantages and disadvantages of choosing a particular clustering algorithm in a recommender system?**

***1.K-means:***

**Advantages:**

- By using this clustering algorithm, large data sets can be processed efficiently because it often terminates at a local optimum.
- This algorithm works only on numeric values. The shape of clusters formed are convex.

**Drawbacks:**

- It does not deal well with overlapping clusters and the clusters can be pulled out of center by outliers.
- The clustering result may depend on the initial seeds, but there is no mechanism to optimize the initial seeds.
- It may converge to a local minimum under certain conditions because it works as a hill-climbing strategy.

***2.Fuzzy c-means:***

**Advantages:**

- It is an unsupervised algorithm.
- The clusters converge.
- Fuzzy clustering is more natural than hard clustering because objects on the boundaries between several classes are not forced to fully belong to one of the classes.

**Drawbacks:**

- It does not work on high dimensions properly.
- Long computational time.
- Sensitivity to the initial guess (speed, local minima).
- Sensitivity to noise and one expects low (or even no) membership degree for outliers (noisy points).

***1.Gaussian EM***

**Advantages:**

- EM iteration increases the observed data (i.e., marginal) likelihood function.
- EM is useful when the likelihood is an exponential family: the E step becomes the sum of expectations of sufficient statistics, and the M step involves maximizing a linear function.

**Drawbacks:**

- EM typically converges to a local optimum, not necessarily the global optimum, and there is no bound on the convergence rate also.
- It can be arbitrarily poor in high dimensions and there can be an exponential number of local optima.
- The algorithm is highly complex in nature,

***Hierarchical clustering***

**Advantages:**

- It is easier to decide on the number of clusters by looking at the dendrogram.
- Easy to implement.

**Drawbacks:**

- It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around.
- Time complexity: not suitable for large datasets
- Initial seeds have a strong impact on the final results.
- The order of the data has an impact on the final results.
- Very sensitive to outliers.

## Ⅳ. CONCLUSION

With the dawn of Big data, clustering assists in grouping objects to analyze information, recognize patterns, and simplify the data. In this paper we have shown how clustering is performed, the different classifications, advantages and disadvantages of clustering types and several authors views on clustering.

We have also shared experimental results on the different clustering methods. Analysis of some clustering algorithms for use in recommendation systems was also included in this paper. The choice of which clustering algorithm will perform optimally in implementation of book recommendation in library management can be decided on the basis of trial and error on these shortlisted algorithms. This is because performance of an algorithm depends hugely on the dataset which is being used.

**REFERENCES:**


[1] P. Symeonidis, A. Nanopoulos and Y. Manolopoulos, “Providing justifications in recommender systems”, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, Vol. 38, Nov. 2008, pp. 1262-1272.

[2] Nidhi Singh,.Divakar Singh,”Performance Evaluation of K-Means and Hierarchical Clustering in Terms of Accuracy and Running Time’, International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012,4119-4121

[3] Namrata S Gupta, Bijendra S Agrawal, Rajkumar M. Chauhan, “Survey On Clustering Technique of Data Mining”, American International Journal of Research in Science, Technology, Engineering & Mathematics, ISSN:2328-3491

[4] Praveena Mathew, Bincy Kuriakose, Vinayak Hegde, “Book Recommendation System through Content Based and Collaborative Filtering Method”, 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)

[5] O.A. Abbas, “Comparisons Between Data Clustering Algorithms”, International Arab Journal of Information Technology (IAJIT), Vol.5, Issue.3 pp.321-325, 2008.

[6] M. Halkidi and I. Koutsopoulos, “Online clustering of distributed streaming data using belief propagation techniques”, In Mobile Data Management (MDM), 2011 12th IEEE International Conference on (Vol. 1, pp. 216-225), 2011

[7] Soumi Ghosh, Sanjay Kumar Dubey, “Comparative Analysis of K-Means and Fuzzy C-Means Algorithms”, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013

[8] Ran Vijay Singh, M.P.S Bhatia, “Data Clustering with Modified K-means Algorithm”, Recent Trends in Information Technology, 2011 IEEE International Conference on 3-5 June 2011 (pp. 717-721)

[9] E. Vozalis and K. G. Margaritis, “Analysis of recommender Systems’ algorithms”, Proc. 6th Hellenic European Conference on Computer Mathematics & its Applications (HERCMA 03), Athens, Greece, 2003.

[10] M. Dave and H. Gianey, “Different clustering algorithms for Big Data analytics: A review”, In System Modeling & Advancement in Research Trends (SMART), IEEE International Conference (pp. 328-333). 2016

[11] Chandrima Sarkar, Atanu Roy, “Using Gaussian Measures for Efficient Constraint Based Clustering”, Department of Computer Science and Engineering, University of Minnesota, Twin Cities.

[12] Witold Pedrycz and James Waletzky, Fuzzy clustering with partial supervision, IEEE Transactions on Systems, Man, and Cybernetics, Part B 27 (1997), no. 5, 787–795.

[13] Shunzhi Zhu, Dingding Wang, and Tao Li, Data clustering with size constraints, Know.-Based Syst.23 (2010), 883–889.

[14] P. Sharma, “Comparative Analysis of Various Clustering Algorithms Using WEKA”, International Research Journal of Engineering and Technology (IRJET), Vol.2, Issue.04, 2015.

[15] Qi Liu, Enhong Chen, Biao Xiang, Chris H. Q. Ding, Liang He, “Gaussian Process for Recommender Systems”, International

Conference on Knowledge Science, Engineering and Management 2011.
[16] Utkarsh Gupta, Dr Nagamma Patil, “Recommender System Based on Hierarchical Clustering Algorithm Chameleon”, Dept. of Information Technology, National Institute of Technology Karnataka,Surathkal, Mangalore.

[17] Deuk Hee Park, Hyea Kyeong Kim, Il YoungChoi, Jae Kyeong, Kim, “A literature review and classification of recommender systems research”

[18] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Analysis of recommendation algorithms for e-commerce,” in Proceedings of the 2nd ACM
conference on Electronic commerce. ACM, 2000, pp. 158–167.

[19] . M. Dakhel and M. Mahdavi, “A new collaborative filtering algorithm using k-means clustering and neighbors’ voting,” in Hybrid Intelligent Systems (HIS), 2011 11th International Conference on. IEEE, 2011, pp. 179–184.

[20] Amandeep Kaur Mann, Navneet Kaur Mann, “Review Paper On Clustering Techniques”, Global Journal Of Computer Science And Technology Software & Data Engineering, VOL. 13 ,2013