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 50 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 immense growth in technology and overflooding data it has become really difficult for the users to find the information they prefer.They have a vast data with them to look at. Information filtering systems of which recommender system is a subpart helps in decision making by providing personalized recommendations by looking at the past history of the user.
Book recommender system will be used to analyze borrowing history of the users by using the techniques of data analysis and data mining Subject headings of borrowed books by the user will be used for generating recommendations which would appear in the user's feed.
In data mining, clustering is the most popular, powerful and commonly used unsupervised learning technique This basically divides the users into different clusters based on 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 works by looking at the items preferred by a user in the past and then making suggestions which could interest the user, hence making the recommendations more customized.

Each cluster would be of a particular preference. Customers within each of these cluster would then receive recommendations.
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 II deals with the research methodologies and the research questions whose solutions will be looked upon in section III. Section IV 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 papers 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. 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 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. Some common 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 algorithm is used to find a partition such that for some particular number of clusters it minimizes the square error. The sum of the Euclidean distances is calculated between each pattern and its cluster center .The most common example of this type of algorithm is k-means clustering.

It is a clustering methodology in which data objects are partitioned into k clusters and each data object belongs to the cluster which has the nearest mean to the object.

Time complexity :O(NkL)

N: no of data objects where

k: no of clusters

L: no of iterations

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 divided into two types . One is hard type and the other is fuzzy type. In hard type clustering all the patterns are divided in well defined clusters but it fails when the boundaries overlap. In such cases some pattern classes may be specified in different groups. To overcome this problem, fuzzy type clustering algorithm is used in real life.

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

It divides the dataset into smaller sets using techniques of hierarchy. It forms a tree of clusters by grouping data instances.It is of two types agglomerative and divisive.In agglomerative type clusters are formed in bottom up fashion opposed to divisive type where it is top down fashion.

The time and space complexity was found out to be:

Time complexity : $O(n^3)$

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 that there are *'n' number of* Gaussian and then the algorithm tries to fit the data into those 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

A lot of papers have made use of gaussian process model for recommender system also . It appears expensive computationally because it calculates inverse of a matrix which can have dimensions as large as 10000*1000.However, in

real-world recommender systems, we never need to compute or evaluate the inverse. All we need is to compute is Σ^{-1}_{00} efficiently since the number of books the user may have searched for will not be more than some 100's which means the size of $\Sigma_{00}$is at most some 100 squared. Calculating the inverse of a matrix of this size is computationally trivial. Hence it proves to be a really fast algorithm in real life cases.

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

A lot of research papers were analyzed for different algorithms and we gained a lot of insight regarding which algorithms are suitable where and what are the pros and cons of choosing a particular algorithm.

K-Means algorithm was mostly preferred in papers where they used large datasets, because it ends at local optima.Also the algorithm worked quite well on datasets which had numeric values.
The only cases where it failed were the cases with overlapping clusters. Also sometimes it may not give the best result because it follows gradient descent strategy and it might end up finding local minimum instead of the global minimum.

Fuzzy c-means algorithm is an unsupervised algorithm and works quite well in cases where some objects are lying on the boundary of clusters. It does not force those objects to fully belong to one class.
However it takes a lot of time to output the results , hence all the recommendation systems using this are not optimal. Also it gets affected by noise really fast, hence all the papers using this had to be extra cautious regarding noise.

Gaussian EM algorithm is mostly preferred in the cases when likelihood belongs to an exponent family. However a lot of papers refrained from using this algorithm because of its complex nature, and inability to converge to the global minimum.
Hierarchical clustering was mostly preferred by papers working on smaller datasets. Also the order of the data was taken into consideration. It was highly sensitive to outliers too and didn't allow the clusters to be moved around after assigning instances to the clusters. No recent paper preferred this algorithm.

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 Young Choi, 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

[21] K. Kameshwaran and K. Malarvizhi, “Survey on Clustering Techniques in Data Mining”, International Journal of Computer Science and Information Technologies (0975-9646), Vol. 5(2), 2014

[22] Anil K. Jain, “Data Clustering: 50 years beyond K-means”, Pattern Recognition Letters – ELSEVIER, 2009

[23] Murtagh, Fionn; Contreras, Pedro, “Methods of Hierarchical Clustering”, CSIR, Vol 1, pp, 1-21, May 3, 2011.

[24] Han, J. and Kamber, M. Data Mining-Concepts and Techniques, 3rd Edition, 2012, Morgan Kaufmann Publishers.

[25] R. Krishnapuram and J. Keller, “A possibilistic approach to clustering,IEEE Trans. Fuzzy Syst., vol. 1, pp. 98–110, 1993.

[26] W. Pedrycz, “Algorithms of fuzzy clustering with partial supervision,”Pattern Recognit. Lett., vol. 3, pp. 13–20, 1985

[27] W. Tracz,Software Reuse: Emerging Technology Los Alamitos, CA: IEEE Computer Soc. Press, 1990.

[28] X. L. Xie and G. A. Beni, “Validity measure for fuzzy clustering,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 3, pp. 841–846, 1991.

[29] Shi Zhong and Joydeep Ghosh, Scalable, balanced model-based clustering, SIAM Data Mining, 2003, pp. 71–82.

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

[31] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schroedl, Constrained k-means clustering with background knowledge, In ICML, Morgan Kaufmann, 2001, pp. 577–584.

[32] Kiri L. Wagstaff, When is constrained clustering beneficial, and why, in AAAI, 2006.

[33] Chandrima Sarkar, Sarah Cooley, and Jaideep Srivastava, Robust feature selection tech-nique using rank aggregation, Applied Artificial Intelligence 28 (2014), no. 3, 243–257.

[34] Manuel Laguia and Juan Luis Castro, Local distance-based classification, Know.-Based Syst. 21 (2008), 692–703.

[35] Katherine A. Heller and Zoubin Ghahramani, Bayesian hierarchical clustering, ICML,
2005, pp. 297–304.

[36] Dwivayani Sentosa, Budi Susetyo, Utami Dyah Syafitri, Sutoro, Applied Expectation Maximization (EM) Clustering for Local Variety Corn, International Journal of Scientific & Engineering Research, Volume 8, Issue 1, January -2017 ISSN 2229-5518.

[37] Epps J, Ambikairajah E. Visualisation of Reduced Dimension Microarray Data Using Gaussian Mixture Model. 2008.

[38] Fraley C, Raftery AE. 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association 97:611

[39] Sambandam, Rajan. 2003. Cluster Analysis Gets Complicated. Marketing Research, Vol 15 (1)

[40] Calinski T, Harabasz J. 1974. A Dendrite Method for Cluster Analysis. Communications in Statistics Theory and Method. 3(1): 1-27

[41] Ding, C., Jin, R., Li, T., Simon, H.D.: A learning framework using Green’s function and kernel regularization with application to recommender system. In: ACM SIGKDD, pp. 260–269 (2007)

[42] Gunawardana, A., Meek, C.: A unified approach to building hybrid recommender systems. In: ACM RecSys, pp. 117–124 (2009)

[43] Marlin, B.M., Zemel, R.S.: Collaborative prediction and ranking with non-random missing data. In: ACM RecSys, pp. 5–12 (2009)

[44] Umyarov, A., Tuzhilin, A.: Improving collaborative filtering recommendations using external data. In: IEEE ICDM, pp. 618–627 (2008)

[45] Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: NIPS, vol. 20, pp. 1257–1264 (2008)

[46] Albayrak, S; Wollny, S; Lommatzsch, A., Milosevic, D.: Agent Technology for Personalized Information Filtering: The PIA System.In Scalable Computing: Practice and Experience, vol. 8, Elsevier, 2007.

[47] Adomavicius, G.; Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Trans. on Knowledge and Data Engineering, vol.17, Piscataway, NJ, 2005.

[48] Polikar, R.: Ensemble based systems in decision making. In Circuits and Systems Magazine, vol. 6, IEEE, 2006.

[49] Herlocker, J.; Konstan, J.; Borchers, A.; Riedl,J.: An algorithmic framework for performing collaborative filtering. In Proc. Int. Conf. on Research and Development in Information Retrieval, New York, NY, 1999.

[50] Sahoo, N.; Krishnan, R.; Duncan, G.;Callan, J.:Collaborative filtering with multicomponent rating for recommender systems. In Proc. Workshop on Information Technologies and Systems, Milwaukee, WI, 2006.