Application of Data Mining in Library-Based Personalized Learning

https://doi.org/10.3991/ijet.v12.i12.7967

Lin Luo
Chongqing Radio & TV University, Chongqing, China
ddtsg@qq.com

Abstract—this paper expounds to mine up data with the DBSCAN algorithm in order to help teachers and students find which books they expect in the sea of library. In the first place, the model that DBSCAN algorithm applies in library data miner is proposed, followed by the DBSCAN algorithm improved on demands. In the end, an experiment is cited herein to validate this algorithm. The results show that the book price and the inventory level in the library produce a less impact on the resultant aggregation than the classification of books and the frequency of book borrowings. Library procurers should therefore purchase and subscribe data based on the results from cluster analysis thereby to improve hierarchies and structure distribution of library resources, forging on the library resources to be more scientific and reasonable, while it is also conducive to arousing readers' borrowing interest.

Keywords—Digital Library, Clustering Algorithm, DBSCAN algorithm

1 Introduction

Library, as a first-hand source of intelligence database for teachers and students in the universities, treasures up abundant books involving a wide range of disciplines. There is a steady stream of new books purchased every year which makes the collection of books in the libraries stacks up to the peak. For this reason, it is a rather difficult task for the universities' teachers and students to find what they expect in the thousands of great books. In the actual process of borrowing, books borrowed by teachers and students do not always represent the user's interests and hobbies. Sometimes borrower checks out books on behalf of other students which leads to a fact that the results of the recommender may not be what teachers and students themselves really want. It requires another kind of thinking for them to choose. Consequently, it is of great importance for teachers and students' learning and research if an accurate and efficient optimization is achieved in the structure of the stock books, "as stated in [1]".

This paper applies the cluster algorithm to assist the librarians in acquiring book classification data about the borrowing frequency and the type of favorites of all kinds of books for fanciers, "as stated in [2]", and then recommend readers the appropriate resources according to their professional backgrounds, interests and hobbies, and other information.

2 Optimization design for DBSCAN algorithm

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) proposed by Martin Ester in 1996 is a typical density-based clustering algorithm. Unlike the partitional and hierarchical clustering, it defines the clusters as the maximum set of density-connected points, and will be able to partition the area with sufficiently high density into clusters. It can also find clusters of arbitrary shapes in a noisy spatial database.

The basic event complexity of the DBSCAN algorithm is O (N *, the time spent in finding the midpoint of the Eps field). In the worst case the time complexity is O (N * N); in the better case, the time complexity is O (N * logN). The most prominent advantage of this algorithm is the density-based classification. In relation to other algorithms, it features better noise countermeasures and processing clusters with arbitrary size and shape, however it will have a poor timeliness when the density of a cluster changes greatly. If it is a multidimensional, a great challenge we face is how to define the density to make their inherent relevance have a better reflex.

2.1 Application model of DBSCAN in the library

A suitable book recommended to the borrowing user, for example, requires to be dug up. For this purpose, the application model is introduced here. Other data mining analysis can be carried out using similar models.

1. Make word segmentation on the user information to obtain a set of user vector labels; repeat this step for the books in the library (information such as a book publisher, a book author, a book blurb, etc.) to capture a set of book vector labels; these word segmentations are mainly used for partition and classification of the categories for more detailed analysis.
2. Aggregate the set of book vector labels based on DBSCAN algorithm, the experiment gives individual book clusters and the cluster center for each cluster.
3. Extract the book-related vector in all the borrow records of individual subscriber as the user set.
4. Use the book cluster center as the initial cluster center of the user set, and then the book clustering is performed with the DBSCAN algorithm to capture the user book clusters and the user cluster center for each cluster.
5. Use the user cluster center as the cluster center applied in the set of book vector labels to form the aggregation clusters, the book vector in the aggregation cluster is used as candidate object (user) of cluster center;
6. Carry out a correlation analysis on the candidate books obtained for recommendations, further access to the user's professional background, interests and hobbies and other information.

2.2 Optimization design

This paper proposes an improved strategy in allusion to a great impact of EPS on the efficiency and the precision of DBSCAN algorithm, that is, two EPSs are used to perform the algorithm, called EPS1 and EPS2 respectively, where Eps1 <Eps2. This idea comes here as follows. The algorithm flow is visualized as figure 1.

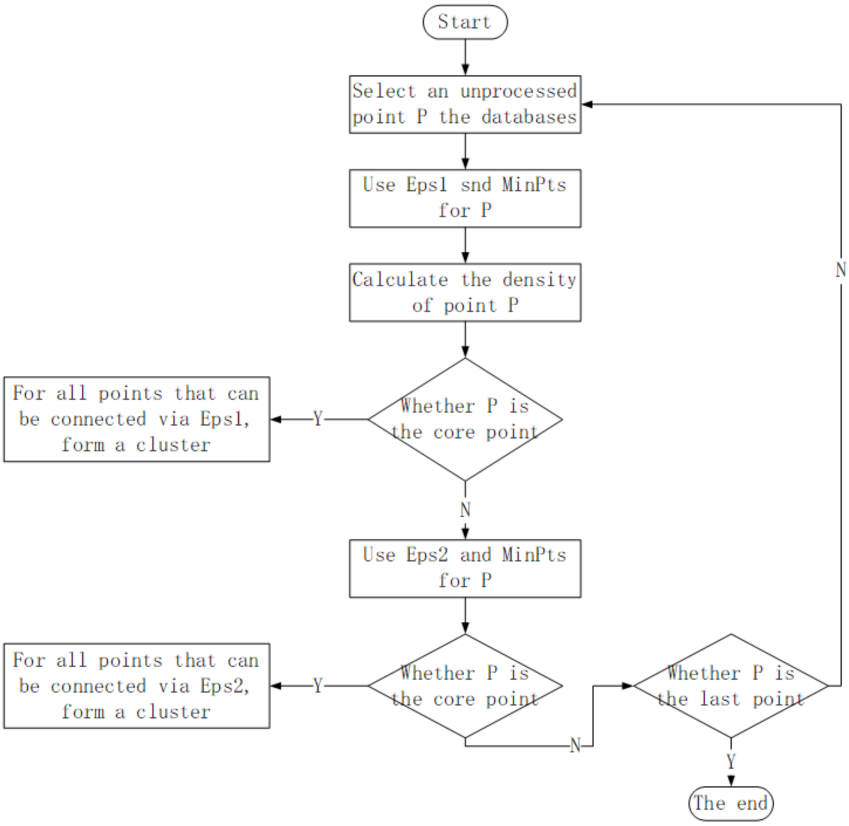
Start
Select an unprocessed point P the databases
Use Eps1 snd MinPts for P
Calculate the density of point P
Whether P is the core point
Y
For all points that can be connected via Eps1, form a cluster
N
Use Eps2 and MinPts for P
Whether P is the core point
Y
For all points that can be connected via Eps2, form a cluster
N
Whether P is the last point
N
Y
The end

Fig. 1. Improved DBSCAN algorithm flow

1. Build a cluster for unprocessed point object P using Eps1;
2. If the point object P can be named a core point in Eps1, then the other unprocessed point objects are processed;
3. If point object P is unable to be named a core point after using Eps1, Eps2 (Eps1 <Eps2) is used for the point object P;
4. If the point object P can be transferred into the core point by Eps2, then the other unprocessed point objects are proceeded;
5. If the point object has not yet been a core point after using Eps2, the point P is marked as a boundary point;
6. Go to the other point objects, repeat the Steps 1-5.

3 Experiment on DBSCAN algorithm

3.1 Empirical data

This experiment adopts bulk of desensitization data from 2012 to 2016 in a Polytechnic University as study sample, among which, four kinds of books are selected, i.e. TP312 programming, TP393 computer network, I247 contemporary works, H319 English language instruction. As shown in Fig. 2-5, the X-axis represents the year; the Y-axis represents the frequency of book borrowings. The dots in the figure reflect the frequencies of book borrowings each year between 2012 and 2016.

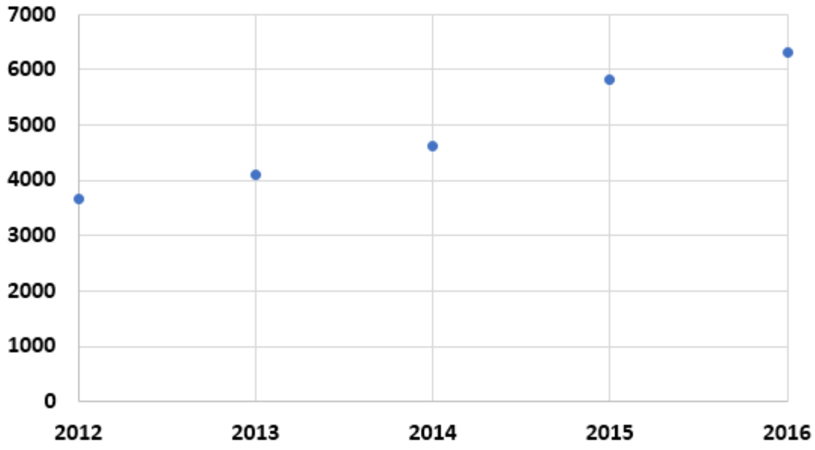

Fig. 2. TP312 Programming borrowing situation

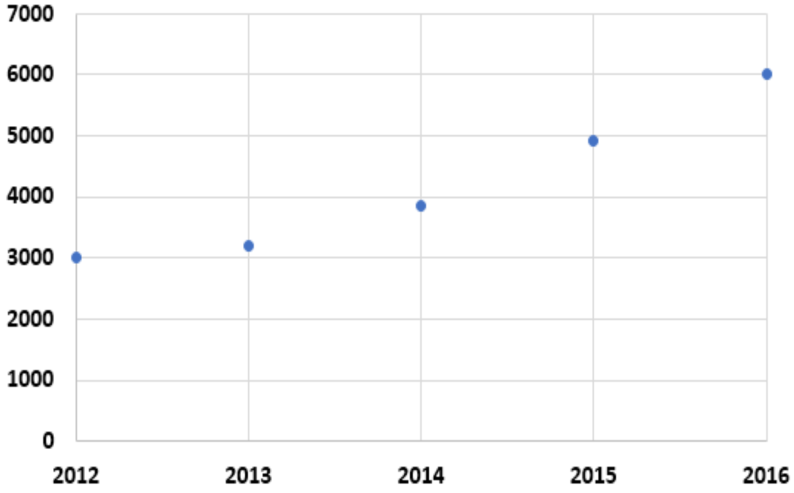

Fig. 3. TP393 Computer network borrowing situation

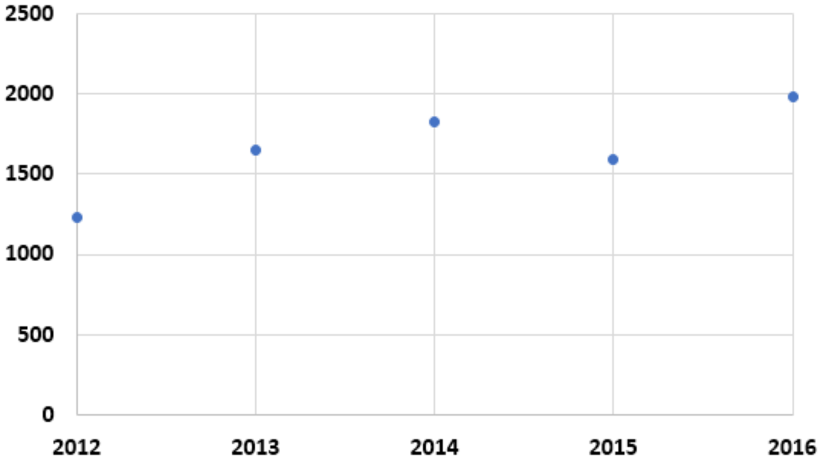

Fig. 4. I247 Contemporary works borrowing situation

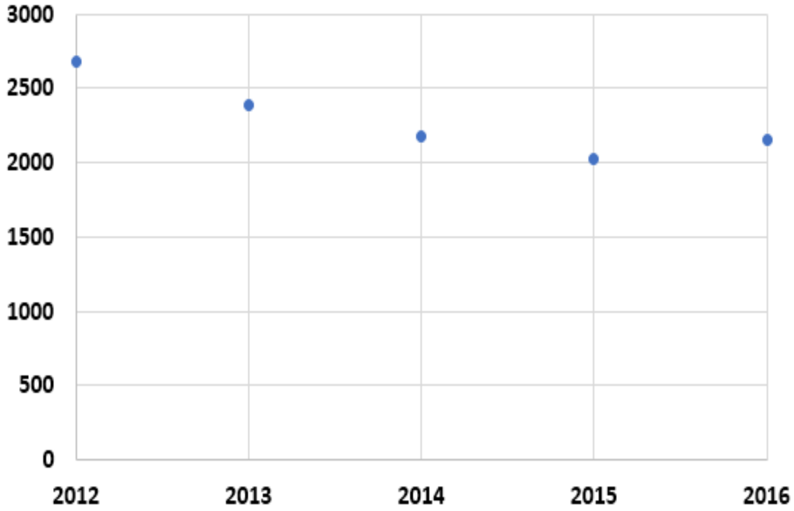

Fig. 5. H319 English language teaching borrowing situation

3.2 Evaluation criteria

As learned from Fig. 2-5, the TP312 programming ranks the first in the frequency of borrowings, and has been on the rise since 2012. The TP393 computer network also shows an upward trend, especially from 2015 to 2016, even has surpassed TP312 programming for some time. For the H319 English language instruction which is biased toward literal arts, the frequency of borrowings decreases year on year, and basically maintains the annual borrowings of 2,000 or so. I247 contemporary works also presents a relatively low level, but generally has a certain growth from 2012 to 2016.

As shown in the above figure, it turns out that the types of engineering books all have a higher borrowing rate than the liberal arts books in the university, which also depends on the selected Polytechnic University to a certain extent because most sciences and engineering majors all concentrate on there with relatively less liberal arts departments, so that the library books mostly involve these fields. While in the categories chosen for analysis, the computer books are usually borrowed at the highest frequency, which is also due to the rapid development of the current Internet. The Internet has spread across various fields with a certain impact on traditional industries, resulting in relative popularization of books on computers.

3.3 Summary on DBSCAN algorithm

When cluster analysis is carried out for selected data with the density-based DBSCAN algorithm mentioned above, the popularity is judged by 4 criterions, i.e. very popular, popular, unpopular and very unpopular. The experimental results derive the cluster center points as shown in Table 1:

Table 1. Experimental results

Clustering center	Average annual savings
Very popular	4637
Popular	2872
Unpopular	1621
Very unpopular	353

Based on the experimental results, librarians can purchase books popular among library users to increase the library's borrowing rate. The results reveal that the prices and the stocks of the books in the library have a little impact on the clustering, however, the classification of books and the frequency of borrowing books have a greater impact on it. Library procurers can purchase and subscribe data depending on the cluster analysis results, thereby to improve the hierarchical and structural distribution of library resources, forging on the library resources to be more scientific and reasonable. It is also conducive to improving readers' borrowing interests.

4 Conclusion

Mass and multifarious data exist in the database of library management system. How to make full use of these data for analysis in the studies, and how to exploit existing information to find out its association from other data, these has become urgent problems for digital library. This paper adopts the DBSCAN algorithm in the cluster analysis for data mining, which enables an improved genetic algorithm to acquire the samples from massive data stored in the libraries. Then the set of data samples is analyzed to extract the user's interests and hobbies, professional backgrounds and other profiles, so as to recommend them the targeted books. This paper uses cluster analysis, as well as other analysis methods such as association analysis and classification analysis for data mining. It is obvious that data mining can do far more than recommendation of library books described in this paper, there is still lots of spatial applications required to be mined.

5 References

[1] Mario, P. Jaume, N. (2015). Visual articulation of navigation and search systems for digital libraries. International Journal of Information Management, 35(5): 572-579. https://doi.org/10.1016/j.ijinfomgt.2015.06.005

[2] Maristella, A. Nicola, F. Gianmaria, S. (2016). Digital library interoperability at high level of abstraction. Future Generation Computer Systems, 55: 129-146. https://doi.org/10.1016/j.future.2015.09.020

[3] Agneta, E. L. Erwin-Christian, L. Corina, M. G. (2014). Digital Library of Mechanisms. Procedia-Social and Behavioral Sciences, 163: 85-91. https://doi.org/10.1016/j.sbspro.2014.12.290

[4] Yanan, Q. Qinghua, Z. Tetsuya, S. Junting, Y. (2015). Dynamic author name disambiguation for growing digital libraries. Information Retrieval Journal, 18(5): 379-412. https://doi.org/10.1007/s10791-015-9261-3

[5] Yi, N. Hui, L. (2017). Progress Report on the Role of Digital Resource Preservation and Utilization for Libraries in China. Publishing Research Quarterly, 1-10.

[6] Klaus, K. Jennifer, A. W. (2017). Applications of RISM data in digital libraries and digital musicology. International Journal on Digital Libraries, 1-10. https://doi.org/10.1007/s00799-016-0205-3

[7] Carlo, M. Nicolas, S. Tsuyoshi, S. (2014). A Model for Digital Libraries and its Translation to RDF. Journal on Data Semantics, 3(2), 107-139. https://doi.org/10.1007/s13740-013-0029-x

[8] Gobinda, C. (2014). Sustainability of digital libraries: a conceptual model and a research framework. International Journal on Digital Libraries, 14(3): 181-195. https://doi.org/10.1007/s00799-014-0116-0

[9] Alexander, R. L. (2015). Mark Y. Herring: Are Libraries Obsolete? An Argument for Relevance in the Digital Age. Publishing Research Quarterly, 31(3): 230-231. https://doi.org/10.1007/s12109-015-9418-3

[10] Mahesh, K. K. Rama, M. A. A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method. Pattern Recognition, 58: 39-48. https://doi.org/10.1016/j.patcog.2016.03.008

6 Author

Lin Luo is with Chongqing Radio & TV University, Chongqing, China.

Article submitted 08 November 2017. Published as resubmitted by the author 12 December 2017.