Correlation analysis of community population problems and community sports service using data mining

Jiwei Yao & Xiushi Ding

Hunan University of Science and Technology
Xiangtan, Hunan, People's Republic of China

ABSTRACT: Education is a way to resolve population problems. Data mining technology used in this study is a computer-assisted analysis technology and an advanced learning tool. Data mining can not only analyse mass data, but can also find potential relationships among data, thus, obtaining valuable information and helping decision-making. Community sports services provide body-building exercises for community residents and is an important factor in determining the health of the community. In China there are a series of population problems that need to be solved urgently, such as health, population structure and life quality in the community. Reported in this article is correlation analysis on community population problems and community sports services on the basis of basic data collection and data investigation. The results show that population aging, population health condition, physical fitness and age distribution have certain correlations with community sports services.

INTRODUCTION

Technical and engineering education is one of the disciplines focusing on acquiring and using technical, scientific and mathematical knowledge to solve problems [1]. Initially, the primary value of engineering education was for functionality and profit. Most engineering education curricula reinforced these values [2]. Data mining (DM) is a computer-assisted technology involving knowledge discovery in databases and was put forward at the 11th International Joint Conference on Artificial Intelligence in 1989 [3].

Data mining is an application technology involving information intelligence. Generally speaking, it is the process of using computer and database technology to extract potentially significant information or knowledge from data, including incomplete data [4].

Data mining is the mining of information and knowledge from data, although there may be no clear conclusions. The information or knowledge obtained may be unexpected. That is to say, data mining can find information or knowledge that may not be intuitively obvious and might even go against intuition [5]. The more unexpected the information mined, the more valuable it might be. This does not mean that conclusions obtained by data mining violate common sense and the law. Instead, it is acceptable, understandable and applicable. Teaching strategies can be optimised by making scientific and reasonable decisions tailored to the demands of learners.

Population is the main problem that China faces at the beginning of the new century [6]. Important manifestations are the uneven distribution of population, problems of an aging population, poor levels of health, low life quality, and an imbalance of males and females, etc [7]. Population problems are manifested not only in the country as a whole, but also in small-scale communities. Community life has become the main lifestyle for Chinese urban residents [8].

Community services have become the object of attention by many researchers in recent years [9]. Researchers obtain basic information about communities by various means, such as questionnaire surveys, interviews and field observations. An important aspect of research on the community and its services is how to extract realistic, significant community data relevant to the people from the myriad data available and, then, identify meaningful and persistent phenomena, and seek to explain these phenomena.

Data mining is well suited to the demand to extract significant data from mass data. It is a reliable tool and a means to resolve the current problem of *rich data and poor knowledge*. Data mining has been successfully applied in marketing, insurance, finance, medical treatment and public health by numerous domestic and foreign scientific researchers. It has irreplaceable advantages in mining the relationships among noumena and rules. Meanwhile, data mining can analyse data from different perspectives, find rules and perspectives that cannot be found by other data analysis methods, and

suggest strategies and means for improvement. The community population problem and community sports services are two aspects of community life. Currently, no one has studied their specific correlation. Data mining can be used to study the correlation between them.

THE DATA MINING PROCESS AND ALGORITHM

In terms of functionality, data mining includes: statistical analysis including histograms, linear and non-linear regression, knowledge discovery, including association rules, neural networks and genetic algorithms, and others, such as text mining and Web mining [10]. Knowledge discovery is a technology for indirectly extracting information from data [11].

This information is implicit and unknown and has potential value. The most common knowledge discovered includes five types, such as general knowledge, association knowledge and biased knowledge [12]. Association knowledge reflects the dependency or association of an event on other events. If two or more attributes are correlated, the value of one attribute can be predicted based on the other attribute value.

Data mining process

Data mining is a process which mines potentially effective and relevant information from large databases. Such information may assist decision-making or enrich knowledge [13]. The mining object is a large amount of data typically in a database [14]. This requires data mining tools, possibly visual so humans can acquire knowledge. Figure 1 shows the data mining environment.

Figure 1: Data mining environment.

Table 1 shows the steps involved in the process.

Table 1: Data mining steps.

Name	Meaning
Business object	Data source for mining
Selection	Search all internal and external data for information related to the business object
Pre-treatment	Quality of research data
Switch	Switch data into an analysis model
Data mining	Mine all data obtained and switched
Result analysis	Explain and evaluate the results
Knowledge assimilation	Integrate the knowledge obtained by analysis into the organisational structure of the business information system

Correlation Analysis Algorithm

Algorithms for data mining include decision-making, clustering, regression analysis and association rules. Algorithms related to correlation analysis are generally basic algorithms of association rules. Among algorithms of association rules, the most commonly used one is the Apriori algorithm put forward by Agrawal et al [15]. The Apriori algorithm is a breadth-first algorithm. It considers two important properties of the data:

1. Assuming that K is a frequent item of set P, all subsets of K are frequent items of set P.
2. Assuming that K is a non-frequent item of set NP, all supersets of K are non-frequent items of set NP.

The Apriori algorithm specifies that, when an item of set K cannot meet the minimum support degree, this item set is a non-frequent item set. If an item set A is added to K, it can be judged that its frequency of occurrence in the whole transaction database cannot be higher than that of the original item set K. Therefore, it can be judged that KA is a non-frequent item set c.f. Formula (1):

$$\text{support}(K) < \text{minsup} \Rightarrow \text{support}(K \cup A) < \text{minsup} \quad (1)$$

The main steps of this algorithm are: first find all frequent item sets meeting the condition and, then, produce strong correlation rules through the frequent item set. Formula 2 shows the conditional expression for generating association rules:

$$\left(\begin{array}{c} Y \subset X \\ Y \neq \varnothing \\ con(Y \Rightarrow (X\text{-}Y)) \geq \min conf \end{array}\right) \rightarrow X \Rightarrow Y \quad (S\%, C\%) \tag{2}$$

In the Apriori algorithm, the Lk-1 frequent item set is used to generate the Lk item set. The main processes are connection and pruning.

1. Connection: in connection, assuming that Ck is the collection of the candidate set K, then, when Lk-1 and Lk-2 are the same element, Lk-1 has connectability, i.e.:

$$C_K = L_{k-1} \infty L_{k-1} \tag{3}$$

When the element satisfies Formula (4), Formula (5) is also valid:

$$\left(l_i[1] = l_j[1]\right) \cap \left(l_i[2] = l_j[2]\right) \cap \cdots \cap \left(l_i[k\text{-}2] = l_j[k\text{-}2]\right) \cap \left(l_i[k\text{-}1] \neq l_j[k\text{-}1]\right) \tag{4}$$

$$l_i \infty l_j = \left\{l_i[1], l_j[1], \cdots, l_i[k\text{-}1], l_j[k\text{-}1]\right\}, l_i \infty l_j \in C_k \tag{5}$$

2. Pruning: pruning is a deletion process, which uses Apriori to delete item sets whose subsets are non-frequent item sets in a candidate item set. These non-frequent item sets meet the condition of Formula (6).

$$c_k \notin L_{k-1} \Rightarrow c_k \notin L_k (c_k \in C_k) \tag{6}$$

DATA MINING CORRELATION ANALYSIS

This research was focused on analysing the relationship between community population problems and sports service using data mining. Figure 2 shows the process.

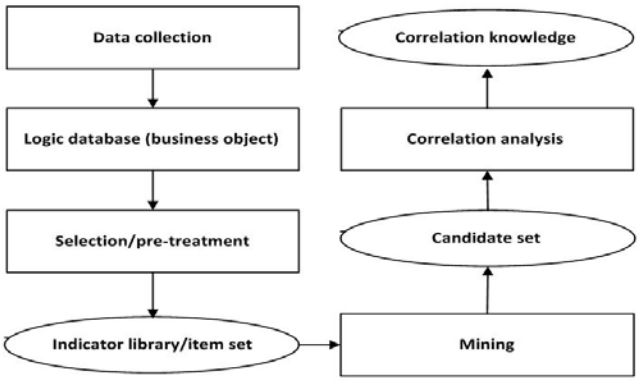

Figure 2: Data mining process.

1. Data collection: data of five (5) employee communities of state-owned enterprises, 10 teaching and administrative staff communities, eight (8) high-tech industrial communities, 14 commercial housing communities and four (4) affordable housing communities in the city of Hunan were obtained using methods, such as field surveys and data querying to establish a logical database business object data sheet. The data sheet was composed of a general community sheet, two secondary sheets about population and sports and various three-level sheets. Table 2 shows the structure of the general community sheet.

Table 2: General community sheet.

Community Number	Name	Unit	Property	Location	Number of community population sheet	Number of community sports sheet
1	###	State-owned enterprises	10	###	RK0001	TY0001
2	###	School	20	###	RK0002	TY0002
3	###	Real estate	30	###	RK0003	TY0003
……	……	……	……	……	……	……

2. The pre-treatment: data are organised based on fields to be used in the correlation analysis. Unnecessary fields are removed. The index fields required by the correlation analysis regarding community population problems and community sports services were obtained.

   Fields were divided into secondary and three-level indicators based on the grade of the sheet. As there were textual data in fields and the association rule generating algorithm can only process numerical data, corresponding indicators were replaced with number, as shown in Table 3.

Table 3: Field indicators.

Sheet type	Secondary indicators	Three-level indicators	Number
Population	Total population	Total number	1001
		Males	1002
		Females	1003
	Age structure	Under 20	1004
		Between 20 and 40	1005
		Between 40 and 60	1006
		Above 60	1007
	Educational level	Below junior college	1008
		Junior college or bachelor	1009
		Postgraduate	1010
	Health level	Healthy	1011
		Illness	1012
		Serious disease	1013
	Time for sports	More than 7 hours per week	1014
		2 to 7 hours per week	1015
		Less than 2 hours per week	1016
	……	……	N
Sports	Sports square	Sports square	2001
	Court	Number of basketball courts	2002
		Number of football fields	2003
		Number of volleyball courts	2004
	Other courts	Number of swimming pools	2005
		Billiards	2006
	Fitness center	Toll centre	2007
		Free centre	2008
	Sports service personnel	Full-time staff	2009
		Part-time staff	2010
	……	……	M

3. Data mining: conduct data mining for the data above and set the minimum support degree and minimum confidence level as 0.2 and 60% respectively.

Table 4 shows the candidate set obtained by using JAVA programming language to write the data mining algorithm and using an Access2003 database as the data support library.

Table 4: Result fragment of the candidate set.

ID	Items
1	10,1001,1004,1008,1010,1024,2002,2007,2014
2	10,1002,1005,1008,1012,1021,2001,2005,2019,2023
3	20,1003,1004,1018,1030,2006,2017
……	……

Confidence level is calculated according to Formula (7):

$$r=\frac{N\sum XY-\sum X\sum Y}{\sqrt{\left[N\sum X^2-(\sum X^2)\right]}\sqrt{\left[N\sum Y^2-(\sum Y)^2\right]}} \tag{7}$$

Table 5 shows the confidence coefficient correlation result obtained.

Table 5: Result fragment of correlation confidence coefficient.

Frequent items	Confidence coefficient
10, 1007, 2001, 1012, 2008	0.78
20, 1005, 1010, 2002, 2007	0.67
10, 1007, 1012, 2009, 2018	0.66
20, 1006, 1013, 2023, 2009	0.75
……	……

4. Correlation analysis: conduct pruning operation for the candidate set and conduct correlation analysis on the mining results for the minimum support degree and minimum confidence level meeting the requirements. Table 6 shows the analysis results.

Table 6: Correlation results fragment.

Frequent items	Confidence coefficient	Support degree
10, 1007, 2001, 1012, 2008	0.78	0.32
20, 1005, 1010, 2002, 2007	0.67	0.22
1007, 1012, 2009, 2018	0.69	0.31
1006, 1013, 2023, 2009	0.72	0.23
1002, 1007, 1014, 1017, 2023	0.60	0.21
10, 1006, 1008, 2003, 2006, 2024	0.8	0.33
30, 1004, 1009, 2004, 2005	0.76	0.35
……	……	……

Formula (8) was used for an efficiency analysis of the correlation result above:

$$h(x, y) = ((\text{order of } x) \times 10 + (\text{order of } y)) \bmod 7 \tag{8}$$

Association rules with an inadequate efficiency are pruned and deleted through the efficiency analysis, and Table 7 shows the correlation results obtained.

Table 7: Effective correlation result fragment.

Rules	Frequent item	Confidence coefficient	Support
1	10, 1007, 2001, 1012, 2008	0.78	0.32
2	20, 1005, 1010, 2002, 2007	0.67	0.22
3	1007, 1012, 2009, 2018	0.69	0.31
4	1006, 1013, 2023, 2009	0.72	0.23
……	……	……	……

5. Association rules: through the correlation analysis, community population and sports service have the following association rules:

Rule 1: community type 10; main population above 60, good health, sports time: over seven (7) hours per week, a few ball game fields.

Rule 2: community type 10; main population above 60, main sport field: square, no full-time sports service personnel, poor sports facilities, severe degradation of existing equipment.

Rule 3: community type 20; main population between 20 and 40, main sport field: basketball court or volleyball court, no full-time sports service personnel, good sports facilities, idle facilities and severe damage to facilities.

Rule 4: community type 20; main population between 20 and 40, educational level: postgraduate, health condition: easy to get ill, time for physical training: two (2) to seven (7) hours per week.

Rule 5: community type 30; main population between 20 and 40, toll construction centre, full-time sports coach, good athletic facility conditions, idle facilities and severe damage to facilities.

DISCUSSION

The following correlation analysis results can be obtained through the data mining described above:

1. Communities in this region, especially communities with the subjects supported by state-owned enterprises, have a main population above 60 and are an aging population. In communities with an aging population, the average time spent on sports by community residents is more than seven (7) hours and the main sports court is the community square.

   These community courts have low use and some of them are severely abraded. This indicates that these facilities have high potential demand but lack maintenance. This type of community does not have professional sports service personnel and residents tend to be engaged in group sports such as aerobics.

2. In communities at educational institutions and high-tech industrial institutions, residents generally have a good academic degree and are mainly between 20 and 40. They spend less time on sports. In this type of community, the population is poor physically. Such communities have many athletic fields and much fitness equipment. These fields have good usage, but the fitness equipment has low usage.

   There are no professional sports service personnel in such communities. In white collar communities at educational and high-tech industrial institutions, the main problem that was exposed is the physical fitness of community residents. Residents generally are in the state of sub-standard health. Due to the high pressure of urban life and work, community residents spend less time on sports. The main mode of exercise is the ball game. Although good outdoor sports equipment is supplied, it has low use and may be idle, thus, causing a great waste.

CONCLUSIONS

Aging populations and low physical fitness are the main population problems for communities in China. With regard to the population problems in these communities, effective sports services are important for improving the physical fitness of the old and ensuring their health and happiness. Meanwhile, the physical fitness of the young should be improved. This research used correlation analysis on community population problems and community sports services through data mining technology. The results showed that aging communities had strong participation in sports but that young people had inadequate sports participation.

Through analysis, this research leads to the recommendation that professional sports service personnel should be provided to guide the aged in fitness, and that outdoor physical fitness facilities should be enhanced to meet the demand of the aged for sports equipment in state-owned enterprise communities and other relevant communities with serious aging populations. In educational and high-tech communities low physical fitness is the main population problem. As a result, the construction of ball venues should be enhanced, while the construction of outdoor sports facilities that are little used should be reduced. Meanwhile, professional fitness centres should be set up to encourage and guide residents to get involved in healthy physical sports.

ACKNOWLEDGEMENTS

This work is supported from a project supported by National Social Science Fund [13CTY009], a project supported by Hunan Provincial Education Department(12C0134) and a project supported by Hunan Provincial Sport Bureau [KT12-030] carried out by Jiwei Yao.

REFERENCES

1. Rohana, H., Sarimah, I. and Kamarudzaman, M.I., Epistemology of knowledge for technical and engineering education. *Procedia - Social and Behavioral Sciences*, 56, 108-116 (2012).
2. Behrooz, P., Computer science and engineering education in a developing country: the case of Iran. *Educ. and Computing*, 2, **4**, 231-242 (1986).
3. Karmakar, M. and Bhattacharyya, D.K., Privacy preserving data mining using matrix algebraic approach. *Matrix World*, 23, **1**, 391-401 (2007).
4. Zhao, X., Liu, S. and Zhang, L., An integrative method of mining genes related to complex disease like Clear Cell Renal Cell Carcinoma with mRNA microarray data. *Inter. J. of Applied Mathematics and Statistics*, 40, **10**, 242-258 (2013).
5. Hamilton, H.J, Geng, L., Findlater, L. and Randall, D.J., Efficient spatio-temporal data mining with GenSpace graphs. *J. of Applied Logic*, 4, **2**, 192-214 (2006).
6. Liang, J., Study on urban community sport construction. *China Sport Science and Technol.*, 39, **9**, 23-25 (2003).
7. Qian Wenjun, Construction of community sport construction index system. *J. of Nanyang Normal University,* 9, **3**, 72-77 (2010).
8. Xiao, L., Concept and theoretical analysis of public sport service. *J. of Tianjin Institute of Physical Educ.*, 22, **2**, 97-101 (2007).

9. Zhang, H., On title, feature and function of community sport. *Sports and Science*, 22, **2**, 25-30 (2001).
10. Gibert, K., Spate, J., Sànchez-Marrè, M., Athanasiadis, I.N. and Comas, J., Data mining for environmental systems. *Developments in Integrated Environmental Assessment*, 3, **12**, 205-228 (2008).
11. Houtsma, M. and Swami, A., Set-oriented data mining in relational databases. *Data & Knowledge Engng.*, 17, **3**, 245-262 (1995).
12. Han, J., Nishio, S., Kawano, H. and Wang, W., Generalization-based data mining in object-oriented databases using an object cube model. *Data & Knowledge Engng.*, 25, **1**-**2**, 55-97 (1998).
13. Xu, S., Data mining using higher order neural network models with adaptive neuron activation functions. *Inter. J. of Advancements in Computing Technol.*, 2, **4**, 168-177 (2010).
14. Ghaderi, R. and Minaei-Bidgoli, B., Detecting data errors with employing negative association rules. *J. of Digital Content Technol. and its Applications*, 3, **3**, 91-95(2009).
15. Cai, Y., Qi, L. and Wang, C., A data mining model of complex system based on improved cluster analysis model and rough set theory. *Inter. J. of Applied Mathematics and Statistics*, 43, **13**, 45-51 (2013).