Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 3, Issue, 11, pp.100-103, October, 2011

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

DATA MINING APPROACH TO STUDY E-GOVERNMENT –A CASE STUDY

[1]Srimani, P.K. and [2]Udaya Rani, S

[1]Former Chairman, Department of Computer Science and Mathematics, Bangalore University, Director, R&D, B.U., Bangalore, India
[2]Mother Teresa University, Kodaikanal, Tamilnadu, India

ARTICLE INFO

Article History:
Received 22nd July, 2011
Received in revised form
16th August, 2011
Accepted 18th September, 2011
Published online 15th October, 2011

Key words:
E-government,
Data Mining,
SPSS, Correlation,
Multi-dimensional cross- table.

ABSTRACT

E-government is a modern way that government department provides services for the public. The level of e-government development is an important standard for a national informationization, e-government can improve government management efficiency, so it is very important that how to improve the public service by the public's need to e-government's development. This model applied data mining technique in e-government construction. Firstly investigates the populace opinion, then process the collected data by SPSS cross tab and correlation analysis, find the immanent rule of people real need, then can provide the better support for government decision, government department also provides the better services for public and achieves humanist truly.

INTRODUCTION

E-government is a modern way that government department provides services for the public. The level of e-government development is an important standard for a national informationization degree. Since e-government can improve government management efficiency, it is very important to improve the public service by the public's need to e-government's development. This model applied the data mining technique in the e-government construction. Firstly, the populace opinion was investigated and then the collected data was processed by SPSS crosstab and correlation analysis, to find the immanent rule of people's real need, and to provide the better support for government decision. By e-government, meant that government uses modern information and communication technique, to i) integrate management and service by network technique, ii) realize optimization recombination of government organization structure and iii) understand the workflow on the internet. At present, e-government's build and application has played an active role in our country. Each department such as person entity base library, natural resources and space geography based database, macroeconomic economy database etc. has produced mass space data and Nonspatial data (Sheng Yu, 2007). Data Mining, also known as knowledge discovery in the database, refers to extract implicit potential useful information and knowledge from a large quantity of incomplete, noise, and blurring, random data which are previously unknown (Jiawei Hart Michelin Kamber, 2001). Data mining method can usually be divided into two categories: The first category is statistical, and its technology used probability analysis, relevance, cluster analysis and discriminated analysis; the other is machine learning based on the artificial intelligence approach, through the training and learning a large number of samples that need to set the mode or parameters.

The related options

As a new way to provide services, one of the characteristics of the public service of e-government is guided by public demand. 1) The services provided by the Government is not to allow the public to adapt to the settings and functions of the departments of the need, but government services should be the maximum from the needs of the public, based on "the interests of the public as the center" design services, and improve service efficiency, reduce service costs, improve service quality, providing the public with the largest service efficiency (Dr. Srimani P.K., and Udaya Rani. S., 2008). Therefore, the Government's electronic public service is not just to change to the mode of service, but to provide more important to government services awareness and the concept of service innovation. 2) Government public services need to attach importance to the principle of demand-oriented and then carry out all system construction and services work. To

*Corresponding author: profsrimanipk@gmail.com, udayamurthy@yahoo.com

achieve this goal, government departments and the public must be interactive with information on the public information needs and study the real need of public (Sheng Yu, 2007).

This Process needs to establish the Public Opinion Collection Mechanism. To start with firstly, the network system of the collection of public opinions has to be established. Usage of the computer technology fully, enables the different opinion of each aspect to reflect to the production public product department prompt and accurately. Secondly, each main body of public should establish information system; allocate specialized personnel to work professionally, so that different views from the public can be reflected in different categories for policy-makers with timely and accurate decision-making signal. Lastly, the final collected public views should be finalized as a management system.

METHODOLOGY

The methodology uses the information architecture (IA) way to carry on the investigation and analysis

Architecture of information act is to organize the information and design information environment, information space and information architecture to meet the information needs of user's art and science. The current method has been successfully applied to various construction sites, as a blueprint for the building site, viz i) whether the site solution with the organization's business objectives, and ii) whether to meet the information needs of users, and other practical issues, such as in a series of indicators weighted analysis IA Government websites can be drawn after the customer satisfaction rankings (Xiao Beigeng, and Zhang Jianping, Jan. 2008, Chen Mingliang, 2003).

Data Mining Technique

Multi-dimensional Cross-table Analysis

Multi-dimensional cross-table analysis predicts that two or more variables join the frequency distribution table. It belongs to the scope of discrete multivariate analysis. To understand the different age levels and qualifications that are concerned about the relationship between the content of the Government, the process can be used to form a two-dimensional tables (Chen YuChen, and Gant Jon, April 2001, WFMC TC00-1011, 1994.). To show that different age groups and all education levels are concerned about the number of different frequency content distribution, correlation, and to choose the suitable way to carry out inspection, Multi-dimensional cross-table analysis of the selected output variables can be performed.

Correlation analysis method

The objective things are interrelated and mutual influence and mutual restraint reflect the interconnection between things to quantity, the correlation between the variables. For instance height and body weight, income and expense. The correlation analysis will find the latent rule that is valuable and the description variable relates the data. Through the co-relational dependence statistics, and the relationship between the variables, one can determine the variable's connection close to the degree and the linear correlation direction. Most commonly Pearson's correlation coefficient R is used. If variables X and Y carry on the observation, on a group of data: x_i, y_i (i = 1, 2,..., n) , then the correlation coefficient formula is

$$R_{xy} = \frac{\sum (x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 - \sum (y_i - \bar{y})^2.}}$$

Where $\bar{x}, \bar{y}$ respectively are the arithmetic average values.

Here $| R_{xy} | \leq 1$. $0 < | R_{xy} | < 1$, means X and Y are right relevant; if $-1 < R_{xy} < 0$, then there is negative correlation between X and Y; and $| R_{xy} |$ closer to 1, implies that their exists a remarkable linear relationship between X and Y. If R_{xy} is close to 0, then X and Y are not related. When $| R_{xy} | = 1$, then X and Y are completely related.

Example of data mining in e-government affair's service – a case study

A careful study pertaining to e-government affairs reveals that there is a considerable improvement in the efficiency and the transparency of the government related works. Although there exists a massive accumulation of data, no effective decision – making policies are available. For example, every government's website has similar "public opinion investigation" columns. It is a very good way of understanding the public's demand, but looked from the website's announcement, the conclusion also poses problems. Take the Karnataka Government net investigation as an example, which investigates tour sites within Karnataka residents (like being possible to elect), to mainly understand the populace demand. The surveyor's basic document includes age level, school record level and occupation and so on. The website has made the simple statistics to the questionnaire, for example the voting results, age level, the school record level and preferred tour sites as shown in Table 1.

It is not difficult to see that above analysis still remains at the surface of the problem, for it did not reveal the intrinsic link of all factors. For example, we can't get the conclusions of the relationship between certain age level and the corresponding enjoyment or whether different degree levels will affect their choice. Therefore, we studied the information collected in-depth with the method of data mining. In order to provide the intrinsic link among the variables such as: certain age conditions, analysis of different qualifications and the different tour sites by using multi-dimensional cross-table. The specific steps followed were:

Step1: Analysis of the data as shown in Table 1.
Step2: Define variables for multiple choice questions; define a variable for each topic. Variables defined are as shown in Table 2.
Step3: The data file after transformation (the partial data) is listed, as shown in Table 3.
Step 4: Multi-dimensional cross-table analysis.

Select 30 samples randomly from the data file. Use the analysis of SPSS cross-table, the results of the analysis are as shown in Tables 4, 5 and 6.

RESULT ANALYSIS

The website has provided the simple statistics to the related questionnaire. For example, the voting characteristics and results at different age groups, the school record level, preference for different tour sites etc., are presented in Table 1. A careful study of the table shows that it is absolutely necessary to predict in detail the intrinsic link of all the factors associated with the problem. For example, i) the relationship between the age levels and the enjoyments ii) the relationship between the degree levels and their preference and iii) the level of optimum preference. In order to predict the above mentioned relationships and choices, the Data mining technique is applied on the information collected. The table contains information about different age levels, levels of academic education and tour sites. The Multi-dimensional cross-table analysis is made by using SPSS cross table and correlation analysis (a sample of 30 is selected at random). The results are shown in Tables 1 to 6. From the cross-table, clearly the distribution among different age groups, degrees and favorite spots were observed. For example in the 18 to 30-year-old age group, a total of ten individuals, degrees for elementary school, junior high school, high school and universities, including high school group had two persons interested in Historic places, five persons interested in Amusement park, one person in Resort and two persons interested in Educational visit. In a word, less than 30 years old age group had greater interest in Amusement park, as they might have kids. The result fits on young people's interest. 30 to 50 years old and over the age of 50 were interested mainly in the Temples and Pilgrimages, which reflected that the adult are interested in historical places only.

Table 1. Karnataka Government Net Investigation

Sight No	Sight Name	No. of Vote
1	Amusement park	55480
2	Educational visit	54329
3	Historic places	53918
4	Resorts	53679
5	Temples & pilgrimages	52834

Table 2. Variables and Variable Value Transformation Comparative Table

Age	Degree	Tour Site
<18 – 1	<Primary – 0	Amusement park – 1
18 – 30 – 2	Primary – 1	Educational visit – 2
30 – 50 – 3	Junior – 2	Historic places – 3
> 50 – 4	High School – 3	Resorts – 4
	University – 4	Temples & pilgrimages – 5
	≥ Graduate – 5	

Table 3. Data Form after Transformation (Partly)

Sl.No.	Age	Degree	Care
1	2	1	1
2	2	2	1
3	2	3	1
4	2	1	1
5	4	2	1
6	2	2	2
7	1	2	3
8	1	1	3
9	1	1	3
10	1	3	3
11	2	3	3
12	3	2	3
13	3	3	3
14	3	5	3
15	4	2	4
16	4	3	4
17	3	1	4
18	3	2	4
19	3	3	4
20	1	2	5

Table 4 Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age * degree * care	30	100.0%	0	.0%	30	100.0%

Table 5. Care * degree * Age Cross tabulation

Count						
	Age		degree			Total
			1	2	3	5	
1	Care	2	0	0	1		1
		3	2	1	1		4
		5	0	1	0		1
	Total		2	2	2		6
2	Care	1	2	2	1		5
		2	1	1	0		2
		3	0	0	2		2
		4	0	0	1		1
	Total		3	3	4		10
3	Care	1	0	1	1	0	2
		2	0	0	1	0	1
		3	0	1	1	1	3
		4	1	2	1	0	4
	Total		1	4	4	1	10
4	Care	1		1	0		1
		2		1	0		1
		4		1	1		2
	Total			3	1		4

The most important results observed was: when the correlation degree between the degree and the corresponding favorite scenic spot is high. (Table 4; age level 4), the Pearson's correlation coefficient is 0.556. In other words, there is a strong positive correlation between the age levels and the preferences for scenic spots. (Irrespective of the degree level). Previous data preparation work is very important regarding the data mining success or failure. This paper demonstrated the application of data mining technique to the public electronic government affairs, and obtains an analysis result from a small sample data. Certainly very accurate results could be obtained when the current methodology is applied to practical situations with large data.

Conclusion

At present, the application of data mining in the e-government public services is relatively small. Through this paper, i) in Karnataka government e-government public services

Table 6. Symmetric Measure

	Age		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
1	Ordinal by Ordinal	Kendall's tau-b	-.289	.309	-.866	.386
		Kendall's tau-c	-.250	.289	-.866	.386
		Spearman Correlation	-.354	.348	-.756	.492[c]
	Interval by Interval	Pearson's R	-.227	.239	-.467	.665[c]
	Measure of Agreement	Kappa	.[d]			
	N of Valid Cases		6			
2	Ordinal by Ordinal	Kendall's tau-b	.485	.231	1.975	.048
		Kendall's tau-c	.480	.243	1.975	.048
		Spearman Correlation	.552	.253	1.871	.098[c]
	Interval by Interval	Pearson's R	.588	.197	2.057	.074[c]
	Measure of Agreement	Kappa	.[d]			
	N of Valid Cases		10			
3	Ordinal by Ordinal	Kendall's tau-b	-.294	.228	-1.291	.197
		Kendall's tau-c	-.267	.207	-1.291	.197
		Spearman Correlation	-.363	.265	-1.101	.303[c]
	Interval by Interval	Pearson's R	-.207	.212	-.599	.566[c]
	Measure of Agreement	Kappa	.[d]			
	N of Valid Cases		10			
4	Ordinal by Ordinal	Kendall's tau-b	.516	.246	1.414	.157
		Kendall's tau-c	.500	.354	1.414	.157
		Spearman Correlation	.544	.261	.918	.456[c]
	Interval by Interval	Pearson's R	.556	.275	.945	.444[c]
	Measure of Agreement	Kappa	.[d]			
	N of Valid Cases		4			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.
d. Kappa statistics cannot be computed. They require a symmetric 2-way table in which the values of the first variable match the values of the second variable.

application data mining technique is introduced, ii) the question of independence is linked together, iii) the nature and potential link of the problem is demonstrated, iv) better decision-making information and support to the Government for the actual work is provided and finally v) one can understand the actual needs of the public by these knowledge, so that the accuracy in the scientific decision-making in the departments, could be enhanced and better service to the people could be provided. The results are tested and validated and are high practical interest.

REFERENCES

Chen Mingliang, 2003. Sourcing Models of Electronic Governance and Government Process Reengineering in China,"Journal of Zhejiang University (Humanities and Social Sciences), pp.27-29.

Chen YuChen, and Gant Jon, 2001. Transforming E-government Services", the Use of Application Service Providers. Government Information Quarterly, 18: 343-355, April 2001.

Jiawei Hart Michelin Kamber, 2001.Concept and Technique of Data Mining," Beijing: Machine Industry Press, 2001 (in Chinese).

Sheng Yu, 2007. The Application of Data Mining in Government Electronic Public Service", Journal of Information, pp.88-89. (in Chinese).

Dr. Srimani P.K., and Udaya Rani. S, 2008. A study of DM techniques for CRM", proceedings of first national women's science congress, 1: 65 – 74.

Workflow Management Coalition, 1994. Workflow Management Coalition Terminology & Glossary", WFMC TC00-1011.

Xiao Beigeng, and Zhang Jianping, 2008. On the Exterior Stipulativeness of the Governmental Power Operation under the Electronic Governance," Present Law Science, 25: 101-104.

****