Research on data analysis of scientific leteracy online survey

Jue Li, Haifeng Zheng

QiuZhen school of HuZhou teachers college,HuZhou, China

QiuZhen school of HuZhou teachers college,HuZhou, China

e-mail: 851645457@qq.com, 360260134@qq. com

Keywords:rough set; scientific literacy survey; data fusion; data mining.

Abstract. Scientific literacy has been regarded as the science education especially the basic education stage of the important goal of education of science. In 1990, the internationally accepted research methods had been introduced into our country, and now it has formed a survey of ideas and index system suited to our national characteristics. The online survey of scientific literacy is an important source of survey data, it can make up for the shortcomings of the cost of household survey, but the credibility of the data is relatively high. In this paper, we use data mining to analyze the scientific literacy survey data, and propose a data fusion based on online survey data and household survey data analysis algorithms to compensate for both their own lack of.

Introduction

With the intensification of international integration, the world is more competitive. National governments and research institutions have realized the importance of the national scientific quality in all aspects of the country's construction. Especially after the Second World War, various national competitions are mainly in science and technology competition. The national scientific quality for the development of science and technology has very important significance. Science and technology research and development must be public support, there must be more and more young people engaged in scientific and technical work, make its traditional economy to knowledge economy rapid transition.

Western countries are now the focus of the study to the investigation of modern science and technology impact on human beings and public understanding of modern technology development trends. For example, information technology and biotechnology has had a major impact on human life and work. These perceptions of the public affected the degree of support of science of technology research, but also affected the development and progress of society. Keep abreast and investigation of the public view of advanced technology is an important part of the public scientific literacy survey. The United States is the world's earliest public scientific literacy research state; its representative is Miller J.D. The United States conducted the first survey in 1957. The EU Commission of Inquiry is a very important investigative agency in Europe. The EU Commission of Inquiry carried out the first survey of public attitudes and understanding of scientific and technological development status in 1992. The second survey, conducted during May, 2001 to June was the basic research for Science and Society Action Plan of EU that was officially launched on December 4, 2001.

Asia and other developing countries also pay much more attention to their own public scientific literacy research. Japan adopted the Miller's ideological system and index system from time to time to investigate and participate in international comparison. India's national characteristics and cultural characteristics literacy surveys are conducted from time to time according with their conditions. All of these surveys have important enlightening significance for us.

Since 1990, the international accepted research method has been introduced into China and has experienced of a study, digestion, try and extensive process. At present, China has already have investigate ideas and index system for their own characteristics. The survey thought, reference to international scientific quality concepts and definitions, drawing lessons from the international general index system, was designed for domestic situation combined with Chinese national condition.

Published by Atlantis Press, Paris, France.

Still, it can participate in international comparative index system. In past practice, we basically have a relatively mature index system, and accumulated a certain experience. After participating in several international public meeting of the scientific quality, the index system and survey methods and data analysis results obtained widespread international attention and concern. The survey data was also included in the research center database of international public scientific literacy, founded by Dr. Miller.

However, we should also recognize that China is still a developing country; the social basis of such large-scale investigations is still relatively weak. For example, telephone surveys, such as commonly used in developed countries will not be suitable to our country because of the low rate of telephone penetration. We can only use the manual survey. Artificial surveys need a reliable, high-quality network and investigation teams. This is the significance of building our website. Establishing a relatively stable, high-quality team of investigators is the basis of obtaining accurate data and ensuring survey quality.

The connotation of public scientific literacy online survey

Generally, scientific literacy can be summarized for three components in international, which is to achieve a basic level of understanding of scientific knowledge; to achieve a basic level of understanding of scientific research process and methods; and to achieve a basic level of understanding of the impact of science and technology on society and individuals. Now all countries adopt the standard to measure the national public scientific literacy, making the survey results internationally comparable.

The online survey, which is to put questionnaire on the internet, is completed by the internet users visiting and browsing the internet home page according to their personal interest in the initiative. This web-based survey depends on the visitor entered voluntarily, and this survey has its unique advantages due to the anonymity feature of the network that is it can be able to reflect the true thoughts and wished of the internet users. At present, web-based survey carried out in China, such as the China Internet Network Information Center (CNNIC) conducted eight times since 1978, belongs to this form of network survey.

Public Scientific Literacy Survey, carried out in accordance with internationally accepted sociology, anthropology and statistical methods, using multi-stage probability proportional to size (PPS) systematic sampling, and there are ten stages: research, demonstration, questionnaire design, random sampling, training of investigators, household surveys and questionnaires summary of data collection and processing, statistical analysis, and report writing etc. Online survey eliminates the need for investigator training and the household survey, greatly reduce the workload, saving time and money.

The actual value of scientific literacy survey

There are problems with traditional survey: samples difficult to collect, expensive research costs, lag of research aspects monitoring etc. With the further exposure of the problems of traditional survey, it is increasing unsuitable to today's society. The number of Internet users in China is constantly on the rise, efficient and convenient features of the online survey, as well as controllability of quality enhancement, a broad prospect for the online survey; online survey is bound to become the dominant trend of the future investigation.

The online survey of the scientific quality is to make the traditional investigation process entirely online and intelligence, making a deep analysis, and ultimately form a professional survey report. The survey is divided into eight modules: the establishment of the questionnaire, questionnaire testing, questionnaire sent to recover the data, statistical reports, project management and system permissions etc.

1) Sharing: Online survey is open to any Internet users, they can vote and view the results, after the initial automatic processing voting information through statistical analysis software, they can immediately see to the stage investigation results.

2) The convenience and low cost: Generally speacking, the implementation of online survey just need a copmuter, a modem and a phone, then researchers can communicate trhough the Internet with the repondent. Researchers need only electronic questionnaire posted on the Internet, counducted a questionnaire survey on Internet users, this is not only very convenient, but a very low cost. At the same time, web-based survey is not affected by weather, distance and time constraints, do not need to print the questionnaire; still the most onerous, crtical collected information and entry tasks in the course of the investigation is distributed to Internet usrs with the software, which will reduce the survey required manpower and material consumption. The exchange of information through the Internet can also avoid mailing questionnaires and interviewing which cost a lot of time.Therefore, compare to traditional survey the web-based survey need less time and cost to complete the same size and number of research.

3) Interaction The greatest advantage of the network is interactive. Respondents can put forward their own views and recommendations in a timely manner on the related issues of the online survey questionnaire, leading the findings of bias can be reduced due to the unreasonable design of the questionnaire.

4) Reliability and objectivity Respondents are completely voluntary principles involved in the investigation, and the investigation of the targeted, thus ensures the reliable collection information and objective findings.

5) No time and geographical restrictions Online market research is avlailable for 24 hours a day, which is different from traditional survey constrainted by region and time. In general, respondents often troubled and failed to get questionnaire return. In this situation, quesionnaire should be sent again to respondents to ensure an acceptable recovery rate, and the online survey can avoid this problem. Intenet make it easy to return questionnarie as long as the repondents are willing to, they just need to click the mouse after finishing quesionnaire, and then the survey is completed.

6) Efficiency As the information can be transformed fast on Internet, and the questionnaire may be sent to respondents in very short time, which guarnatees the investigation in a very short period of time, and get a lot of survey results. At the same time, it is very convinent to monitor the recovery of questionnaire. Compare to traditional survey, as soon as the online survey researchers put the questionnaire on Internet, the network management will start at the same time, researchers can keep informed of network data collection, processing and analysis situation, and timely detecte of problems, modify the questionnaire, which makes web-based survey is more efficient than traditional survey. This high efficiency of the network survey is unmatched by the tradifitional method of investigation.

7) Accuracy During the online survey, the data entry process required by the traditional survey is replaced by the statistical software, thus reduce the problem of omission and coding. A large number of standardized statistical analysis can be accurately completed in a short peroid with the statistical software, it also in large part to ensure the reliability of the network survey results, reducing the statistical error of traditrional research.

Scientific literacy online survey data analysis

Scientific literacy online survey data approximate reduction algorithm. Scientific literacy survey data is very large, it is necessary to carry out the reduction in order to facilitate further processing of data mining. After attribute reduction, the number of properties is reduced, but most of the information system, the number of properties is still a lot. And the differences between some of the properties is caused by the noise of the data, so these properties can be further reduction, but also remained the style of the face of the original information system.

We have established a pairwise difference on the set of attributes, using the collection to identity the relative importance of the attributes. After finding the sets with less attributes to the complete reduction attribute sets, there will be a small amount of pairwise left. And omit these attributes will not affect the original system to a certain extent.

Set n objects of the hypothetical information system overall for the U, it is preferable to the r possible values $\{u_1, u_2, \ldots\ldots, u_r\}$, the number of samples for each value is$\{ p_1, p_2, \ldots\ldots, p_r\}$;after

reduce several properties, the number of samples for each value is{ p_1',p_2',……,p_r'}. Now to verify: H_0: $p_i = p_i$' i=1,2, ……,r.

According to the law of large numbers of mathematical statistics, it can be concluded that when H_0 holds, the difference between p_i and $p_i^{'}$ should be less, otherwise, the difference is large, and naturally reason to believe H_0 does not hold. Accordingly, Pearson had constructed statistic to reflect the difference: $\chi^2 \sim \sum_{i=0}^{i} \frac{(p_i^{'} - p_i)^2}{p_i}$ .And he also demonstrated that when H_0 holds and n is sufficiently large, regardless of the overall subject to any distribution, the test statistics χ^2 always approximately obeys χ^2 distribution whose degree of freedom is r-1.

To make Arm be reduction matrix for the attribute information on table S, the element of Arm is $RM(a_k) = \{u_i, u_j\}$, it is the set of all the pairwise object of the attribute a_j in the information table. Create a property ArF, and its element is $RM(a_k)$, which corresponds to the number of the set of all pairwise distinct object attribute a_k. Property approximate reduction algorithm can be described as follows:

Set R= Ø, calculate ArM, and do initial classification to information table, r represents the initial number of categories, calculate the number of every sample p={ p_1, p_2,……, p_r}. Do loops:

8) Calculate ArF, and ArF(ak)=Max(aj);

9) R<--R∪{a_k};

10) Let ArM<--ArM(a)- ArM(ak);

11) Classification of objects according to ArM;

4.1) Scan ArM, merging object u_j to u_i for every $\{u_i, u_j\}$;

4.2) Recalculate for each category the number of samples p'={ p_1', p_2',……, p_r'};

12) Calculate the statistic of $\chi^2 \sim \sum_{i=0}^{i} \frac{(p^{'}_i - p_i)^2}{p_i}$

Until $\chi^2 < \chi_\alpha^2(r-1)$. Then the set R is a property approximate reduction of information table at the significant level α.

The online survey data mining algorithms.*Set (U, A) is a set of performance data system, U = {u_1, u_2, ..., u_n}, a and y is two attributes of the set.*

L1: i=1, j=1, s=1, v_1={u_1};

L2: if i<|U|, then i=i+1, j=1;

L3: if j=s then s=s+1, v_s={u_i}, go to L2;

L4: j=j+1;

L5: if a(u_i)=a(v_j) , ui∈v_j, go to L_2; else go toL3;

L6: i=1, j=1, t=1, w_1={u_1};

L7: if i<|U|, then i=i+1, j=1;

L8: if j=t then t=t+1, w_t={u_j}, go to L7;

L9: j=j+1;

L10: if y(u_i)=y(w_j) , ui∈w_j, go to L7, else go to L8;

L11: i=1;

L12: while i<t:

L12.1: j=1, L_i=∅ ;

L12.2: if v_j∈w_i, then L_i=L_i∪V_j;

L12.3: if j<s then j=j+1, go to L12.2;

L12.4: $spt_\alpha(W_i) = |L_i| / |W_i|$

L13: end .

$spt_\alpha(W_i)$ is the support of a(u)=a(Li) that implies y(u)=a(w_i). The time complexity of this algorithm is O($|U|^2$)

Online survey data and household survey fusion method. We propose the data repair and correction data fusion algorithm on the basis of data preprocessing methods for the online survey

data and household survey data. The method can be summarized as the following steps:

Step 1: First look for the observational data records with complete data field as the donor and defects in the data field observational data as the receptor, then calculate the Euclidean distance between all of the donor and acceptor, finally using the Set Pair Analysis (SPA) theory project team are working to process the same, different and contrary three types of data.

Step 2: Calculate the default distance threshold to select from the receptor close to the donor (based on the experience of a preset distance threshold), while avoiding a donation by excessive use to prevent the results obtained over a single.

Step 3: According to the additional conditions, such as the geometry of location, time and information to select donor among the donors selected in step 2, then get the donor's data used directly to repair in the data domain of the receptor defect.

Conclusion

Scientific literacy started relatively late in China, but now from the central and local governments using the method to conduct surveys. HuZhou city, Zhejiang province, has also conducted three large-scale surveys. The survey now commonly use the household survey, it need to put in a lot of manpower and material, the cost is huge. The online survey is an important source of survey data, it can compensate for the shortcomings of the household survey, but the credibility of the data is relatively high, so we use data mining to analyze scientific literacy survey data, and propose an online survey data and investigation data fusion analysis algorithm to compensate for their shortcomings. These two technologies have played a very large role in the recent scientific literacy data analysis of HuZhou.

References

[1] Qin Qin, Huiying Ren. Handbook of Sampling in Research[M]. Beijing: China's era of economic press.2004

[2] Jeffrey Richter, Balena, Francesco. Applied Microsoft .NET Framework Programming in Microsoft Visual Basic .NET[M]. Wuhan: Huazhong University of Science and Technology Press

[3] Dino Esposito, Andrea Saltarello. Microsoft .NET architecting applications for the enterprise[M]. Beijing: People's Posts & Telecom Press.2009

[4] Ramez Elmasri , Shamkant B. Navathe. Fundamentals of database systems[M]. Beijing: Tsinghua University Press,2011

[5] Bin Shao, Yunliang Jiang, Zhen Yang. The Research on Video supervision technology based on Mathematical Morphology[J]. 2009 International Conference on Industrial Mechatronics and Automation,ICIMA .2009.5:57-60

[6] Bin Shao, Yunliang Jiang, Qing Shen. The Multi-Agent Simulation and Philosophy on Economic Game[J]. Journal of Information & Computational Science.2009.12,vol(6) 6:2305:2309

[7] Bin Shao, Nan Xiang, Zhen Yang.Mathematical Morphology Structure Element Construction Based on Genetic Algorithm and Its Application in Traffic Vehicle Detection[J]. The Internation Conference on Multimedia technology（2010 Icmt) .2010.10:540-543

[8] Bin Shao, Lingli Wu, Yunliang Jiang. A Reduction Algorithm of Rough Sets Against Imperfect Information Systems[J]. Microelectronics & Computer.2007,4