

***“Analyzing Survey Data Concerning the Construction of Central Taiwan Science Park using Association Rules and Geographical Information Systems”***

*Fang-Yie LEU*

**Fang-Yie LEU**

Dept. of Computer Science and Information Engineering  
Tunghai University  
Taiwan  
leufy@thu.edu.tw

**Abstract:** Recently, many data mining techniques have been developed for and deployed by scientific and industrial use to automatically mine, analyze and extract hidden knowledge from raw data given. Among them, association rule, one of the most commonly used ones, is often used to discover relationship between two set of items. Also, commercial Geographical Information Systems (GISs) and their functions have been quickly developed and significantly improved respectively in recent years. Researchers and policymakers can input environmental data to a GIS system to gain spatial analytical results which often show up how data is geographically dispersed. In this paper, we discuss how to apply association rules to analyze surveyed data collected from people living in the Situn district and Dayia village which are two areas surround Central Taiwan Science Park so that researchers can accordingly realize some facts that can not be superficially obtained from raw data concerning the construction of the science park (before and after). The results can be referred to by local and central governments as a reference when making public policies. Besides, if we can input the analytical results to GIS, the hidden meanings or rules embedded in the survey data can be then uncovered more deeply and precisely.

**Keywords:** GIS, Association rules, survey data, Central Taiwan Science Park, Shitun district, Dayia village

## 1. INTRODUCTION

Nowadays, Geographical Information Systems (GISs) are widely used, particularly in designing and showing a city's road networks, underground pipes, power lines, and et al. Users can search roads or landmarks on a electronic map or in internet if the map provides a web version, to realize the locations they are interested in.

Besides, machine learning [1-7] is also a well known intelligent technique/model. Most of the researchers or decision makers rely on computers to analyze their data in deep which are always stored in computer databases or files. However, databases or files are passive data. We can query or manipulate them only. They never actively tell us the knowledge deeply embedded or hidden in them. In the social or geographic domain, few applications deploy GIS and data mining at the same time.

In this paper, we will discuss how to apply association rules to analyze survey data collected from people living in the Situn district and Dayia village, which are two areas surrounding Central Taiwan Science Park (CTSP), so that researchers can accordingly realize facts that can not be superficially obtained from raw data concerning the construction of the science park (before and after). The results can be referred to by local and central governments as a reference when making public policies. Besides, if we can input the analytical results to GIS, the hidden meanings or rules embedded in the survey data can be then uncovered more deeply and precisely.

## 2. RELATED WORK

To date, many application domains have employed data mining or GIS techniques, but not both, to promote their business.

In health care domain, Mitchell [1] described several prototypical uses of data mining, including an expert system able to predict women at high risk of requiring an emergency C-section. Merck-Medco Managed Care, a pharmaceutical insurance and prescription mail-order unit of Merck, used data mining to help uncover less expensive but equally effective drug treatments for certain types of diseases or patients [2].

In finance domain, Bank of America deployed data mining to detect which customers were using which Bank of America products so they could offer the right mix of products and services to better meet customer needs [2].

In sports domain, Brian James, assistant coach of the Toronto Raptors professional basketball teams, used Advanced Scout, a data mining/warehousing tool developed by IBM especially for NBA, to create favorable player matchups and help call the best plays [3].

Besides, many commercial products of GIS have been released, such as ArcGIS [4], TomTom Navigator [5], Google Map [6], Yahoo Map [7]. Some of the products are for single client use, and others for web-based service. For analysis purpose, the ArcGIS is much more mature than others since it can perform almost every type of geographical analysis. or mobile or navigation purpose, Garmin and TomTom have released many products in this domain.

### 3. SYSTEM DESCRIPTION

Machine learning is a complex process. Computers are sometimes good at learning concepts. A concept is a set of objects, symbols, or events grouped together due to sharing certain characteristics. Concepts can be well designed and structured for future retrieval and management. Common concept structures include trees, rules, networks, and mathematical equations.

Data mining, a famous machine learning model, is the process of employing one or more computer learning techniques to automatically analyze and extract knowledge from data collected in a large database. Its purpose is to identify trends and patterns in data so that users can extract *hidden predictive information from the database*. It is a powerful technology with great potential to help researchers focus on the most important information in their raw data.

Many types of data mining techniques have been developed. Among them, association rule is one of the most commonly used ones. It is often used to discover relationship between two set of items.

#### 3.1. Association rules

Affinity analysis is the general process of determining which things go together. A typical application is market basket analysis, where the desire is to determine those items likely to be purchased by a consumer during a shopping experience. The output of the market basket analysis is a set of associations about consumer-purchase behavior.

To perform affinity analysis between two things A and B, confidence and support are two important parameters required to be considered. Confidence is the conditional probabilities of the occurrence of A given the occurrence of B, and the occurrence of B given the occurrence of A.

Support is simply the minimum percentage of transactions or instances in the concerned database that contain all items listed in a specific association rule. Confidence and support should be each over their given thresholds before we can start the affinity analysis by using association rules.

#### 3.2. Geographical Information System (GIS)

A GIS system (or GIS in short) is an application system for creating, storing, analyzing and managing spatial data and associated attributes [8-13]. In a more generic sense, a GIS is a software tool that enables users to create interactive queries, analyze spatial information, edit data and display geographically-referenced information.

GIS is often used for scientific investigations, resource management, asset management, environmental impact assessment, city development planning, cartography, and route planning, for example, to identify a polluted area that needs to be isolated from others.

#### 3.3. Data Creation

Modern GIS technologies rely on digital information, for which there are a number of collection methods. The most common and popular one is digitization, where a hardcopy

map or survey plan is transferred into a digital medium through the use of a digitization tool which is a computer-aided drafting (CAD) program with geo-referencing capabilities.

### **3.4. Data Formats**

GIS represents real world objects (roads, wetlands, buildings) with digital data. Raster and vector are two common methods used to store data in a GIS for discrete objects and continuous fields. Raster images consist of rows and columns of cells where a cell stores a single value. The value recorded for each cell may be a discrete value, a continuous value, or a null value (if no data is available).

Vector uses geometries such as points, lines (series of point coordinates), or polygons (shapes bounded by lines), to represent objects. Vector features can be made to respect spatial integrity constraints through the application of topology rules such as 'polygons must not overlap'. Vector data can also be used to represent continuously varying phenomena to show us the continuous change of objects, e.g., the annual development of last 20 years.

Additional non-spatial data can also be stored besides the spatial data, e.g., names and addresses, collected through questionnaires or interview. In vector data, attributes of object are required. For example, a city inventory polygon may also have an identifier value and information about its population. In raster data, the cell value can be attribute information, or an identifier relating to records in another table.

## **4. DATA ANALYSIS ON SURVEY DATA OF CTSP**

Central Taiwan Science Park (CTSP) has started its running since 2003. Its surrounding environment has hugely changed. For example, a huge shopping mall has been constructed. Many luxurious restaurants are opened during the past three years. Many big houses and apartments have been constructed. Currently, about 20,000 people are employed by CTSP corporations. The ultimate number of employees will be 50,000, which would increase population of CTSP residents to about 200,000 (including their families). The more people, the more business opportunities and the more shops and restaurants.

In the survey on people who live in Shitun district and Daya village, a total of 613 (Shitun 401 and Daya 212) residents are successfully interviewed. Among the questions in questionnaires designed, eight are GIS related.

**Question 34:** *About 4 to 5 years ago (before CTSP started running), where had you gone shopping?*

**Question 35:** *In the passed one year, where have you gone shopping?*

**Question 36:** *About 4 to 5 years ago (before CTSP started running), where had you eaten out?*

**Question 36a:** *In the passed one year, where have you eaten out?*

**Question 37:** *About 4 to 5 years ago (before CTSP started running), where had you gone to spend your leisure time?*

*Question 37a: In the passed one year, where have you gone to spend your leisure time?*

*Question 38: About 4 to 5 years ago (before CTSP started running), where was your office or business location?*

*Question 38a: In the passed one year, where was your office or business location?*

#### **4.1. Distance Analysis**

According to the survey data, about 65% of answers of question 34 are the same as those of question 35. Most people as their usual go to middle-scale supermarkets or nearby smaller-scale supermarkets to purchase their daily needs. It seems that they do not change their shopping behaviors. The reasons are that they are familiar with everything in the supermarket or supermarkets they have often gone, and know what the supermarkets have and the prices. People can also easily find out what they want and used to buy. Therefore, seldom of them change their shopping locations. Additionally, no large-scale supermarkets, that are currently available and can attract them to visit, have been constructed and opened, even several are now under planning and designing.

From questions 36 and 36a, we can realize that many residents have their meals at home. But those who often eat out gradually change their restaurants (from 23.1% to 31.8%) from far away ones to nearby since many new restaurants have been opened in Shitun district and Daya village. The reason is simple. More than 70,000 people (CTSP employees and their families) come to live near CTSP. Many restaurants are required. Further, most restaurants (old and new) are much more expensive than before since high-tech employees make much more money than other original residents. Of course, restaurants offer higher quality foods. What we can also learn from the phenomenon is that many schools, including elementary, junior and senior high schools, are required to fulfill future educational needs because most CTSP employees are young. Many more high quality houses and apartments are also required since high-tech people have much more money to buy luxurious ones.

From questions 37 and 37a, we can conclude that about 83.7% of people do not change locations where they like to spend their leisure time. There are three reasons. One is leisure facilities, e.g., new parks and department stores, are not newly constructed and equipped. Nothing can attract them to change their original leisure facilities. The second is original residents currently are not rich enough to change leisure facilities to new ones. The last one is they enjoy their current leisure ways. As we further study the reasons by interviewing concerned people again, the three reasons truly exist. We also conclude that people living in these two areas due to prosperous business activities have earned much more money than before.

Questions 38 and 38a show that about 89% of people do not change their jobs since most people living in Daya village are less educated. Seldom residents have opportunities to enter CTSP as workers. Most Shitun original residents, even having a little bit higher education, have their own jobs before 2003. Their professions, positions and ages are or are not acceptable by CTSP corporations. So only few original residents are employed by the corporations. But, the remaining 11% of residents is business people. They move their

businesses back to the area near CTSP. They do benefit from this science park since CTSP employees have higher consuming capabilities due to a higher high-tech salary and payment.

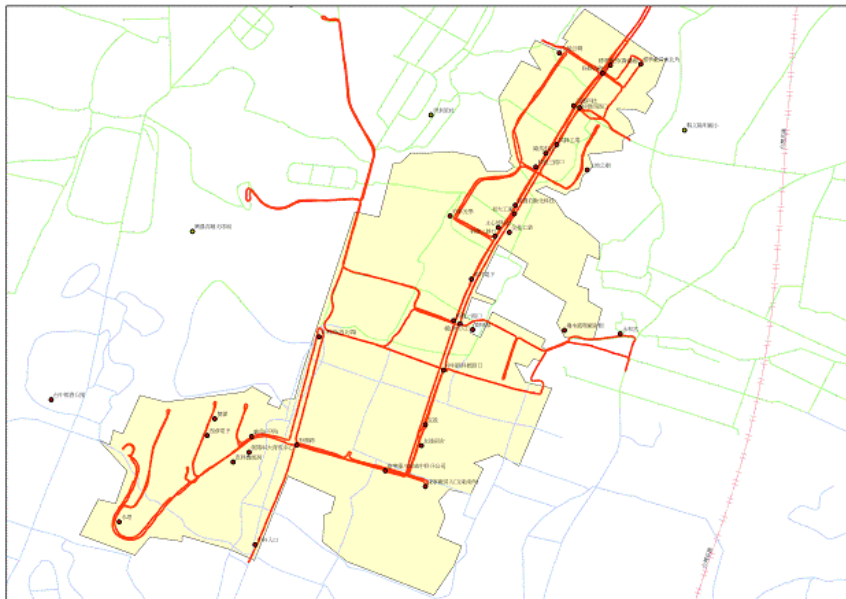
#### 4.2. GIS Applications

In this paper, several GIS applications have been developed, including drawing roads that are newly constructed but not drawn on any maps. The way to do that is carrying a global positioning system (GPS) to drive a car through the roads newly constructed with breath first approach [14] and retrieve longitude and latitude from GPS periodically, e.g., once a second, along the roads. After that, we plot the positioning signal on an existing map. As shown in Fig. 1, new roads are clearly illustrated on the map. The area that CTSP locates is colored by yellow.

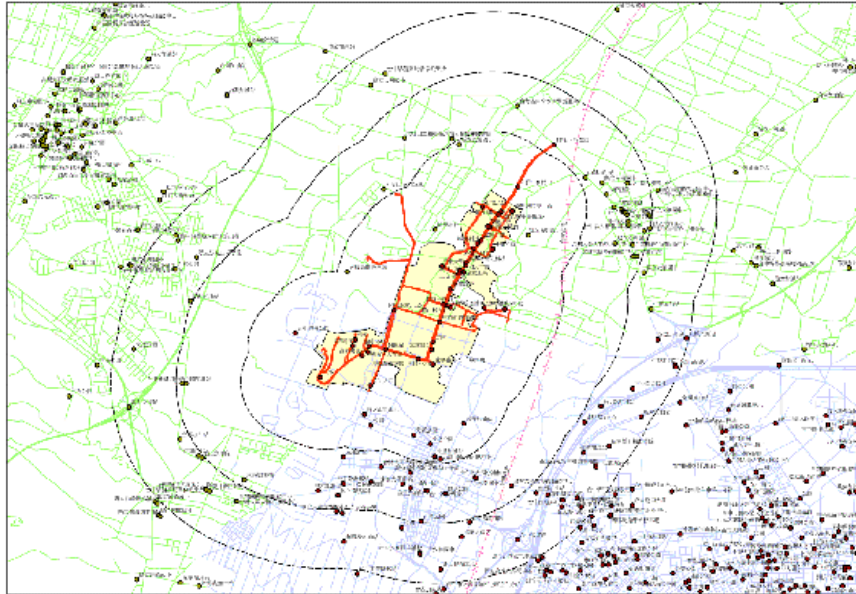
Another application is to plot the areas within a specific distance from CTSP. Fig. 2 shows an example. We can overlap data or objects plotted on other layers, e.g., population statistics and buildings, on the map. From which researchers can then realize the distribution of the objects he/she is interested in. For example, the distribution of population living in the areas within one, two and three kilo meters away from CTSP.

The third application is given students an easy way to point out locations an interviewed person answers if the question raised is to give a specific point he/she likes to go and the location is hard to be described by oral description. Of course, as shown in Fig. 3, a coordinate system should be given, e.g., x-axis is marked by A, B, C, ...and y-axis is marked by 1, 2, 3, .... Students can record the position given by recording the corresponding points in the lattices, e.g., 2-A and 6-D, instead of writing name of a location or an address. After interview, students can easily translate the records of these points into their longitudes and latitudes, which can be then shown on an electronic map.

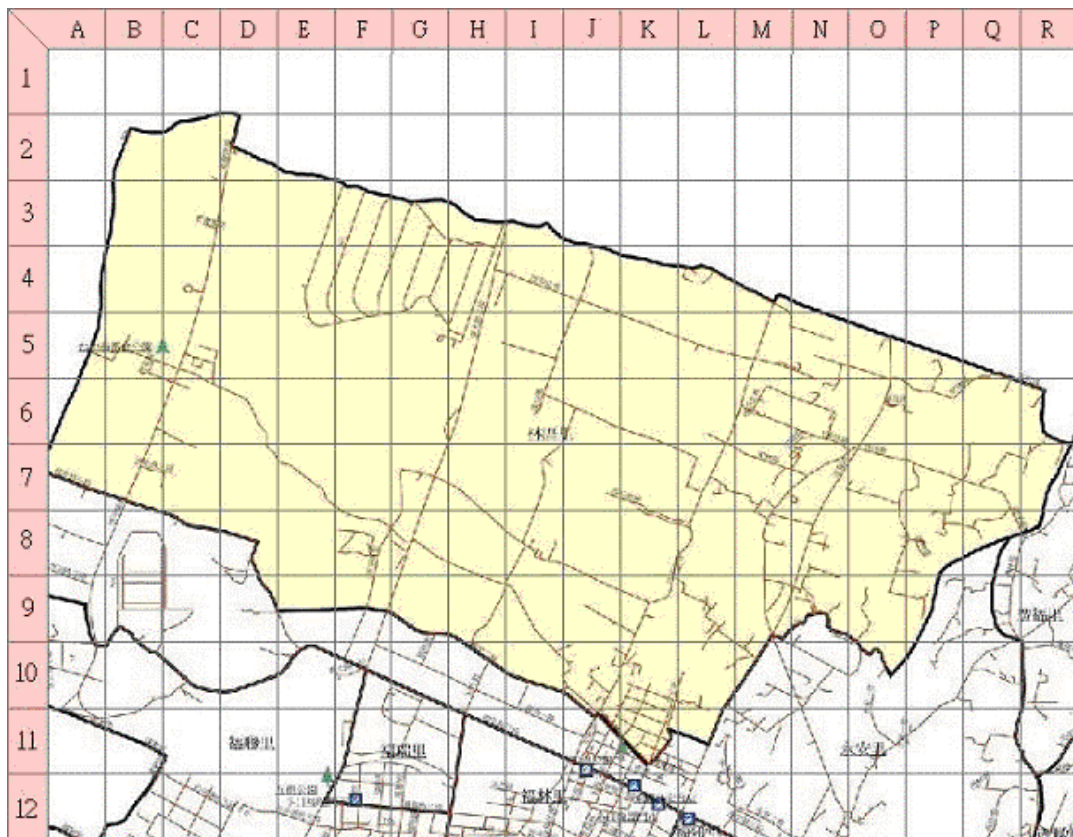
**Fig. 1: Roads newly constructed in CTSP and the area that CTSP locates.**



**Fig. 2: Buffer areas within 1km, 2km, and 3km away from CTSP.**



**Fig. 3 A map that provides x and y axis so that students can easily locate a location that an interviewed person figures out by oral description**



### 4.3. Data Mining – Two Examples

We use association rules to identify large-item sets between two or among  $n$  attributes,  $n > 2$ , and discover that a group of people consider that traffic does not change. But, they feel that public security is worse than before. However, the other group on the contrary thought that traffic is terrible and public security is fine. Opinions of the two groups are quite opposite. We further study the reason and realize that most people of the first group are housekeeping women. When their husbands and children go to offices, they stay at home, thus very often experiencing no rush hours during eight to night AM and five to seven PM since during these time periods they are cleaning houses and cook dinner for their families, respectively. However, many salespersons sell something at day time, particularly home by home, and in these areas burglars often break into houses also at day time, making these women feel that they are unsafe at home.

Their husbands and children are the second group. They suffer traffic jam during rush hours and when they stay at home, the whole family comes together, making them feel safe and happy.

The other example is that many people answered that they very much concern events that occur near or surrounding their homes. Assume that this is question 0, Q0 in short. But, their answers on the following questions are highly contradictory. Table 1 lists their confidences and supports.

**Q1:** Do you know there are several protests against poor environmental quality of the area near your home recently?

Many residents answered “I do not know”. The other choices are yes, no, and just can not remember.

**Q2:** Who is the key institute or organization that has the strongest power to determine environmental quality of the area near your home?

Many residents answered “residents themselves”. The other choices are government, community, and non-profit organization.

**Q3:** Have you ever raised your opinions or suggestions when local government makes a decision or defines policies that may affect environmental quality, particularly affecting that of the area near your home?

Many residents answered “never”. The other choices are always, frequently, and seldom.

	Q1	Q2	Q3	Q1&Q2	Q1&Q3	Q2&Q3	Q1&Q2&Q3
confidence	45.3	57.2	48.6	33.3	29.8	37.6	21.4
support	54.4	61.7	59.9	40.1	45.5	39.0	32.1

Our conclusion is that residents like to participate in activities that can improve environmental quality. They also know that they play the key role in those activities. However, some of them do not know any recent protests against poor environmental quality of the area near their homes and of course absent from those activities. The reasons are that they are disappointed in what have been done by government, even their suggestions and opinions are known to government (53.7%). The second is that they are

busy all day long, and have no time to participate in those activities (32.4%). The third is that some of them satisfy their current environmental quality (13.9%).

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a way to analyze survey data collected from interviewing residents living near CTSP through the use of association rules and GIS so that we can realize the residents' feeling and opinions on the environment surrounding CTSP before and after the park started in 2003. Association rules can give users the deep relationship between two or among several attributes. GIS shows relationship among spatial data. We only give several examples. Users can analyze the remaining attributes and develop many more GIS applications to fulfill their analytical needs.

We will continue analyzing the data by using other data mining techniques, like neural networks and time series, to predict the trends of some specific data, e.g., amounts of salaries of CTSP employees and incomes of CTSP corporations.

## REFERENCES

- Mitchell, T.M. (1997) Does Machine Learning Really Work? *AI Magazine*, vol.18, n<sup>o</sup>.3, pp.11-20.
- McCarthy, V. (1997) Strike It Rich. *Datamation*, vol.43, no.2, pp.44-50.
- Baltazar, H. (2000) NBA Coaches' Latest Weapon: Data Mining. *PC Week*, pp.69-69.
- ESRI - The GIS Software Leader, <http://www.esri.com/>.
- Systèmes de navigation routière GPS portables de TomTom, <http://www.tomtom.com/index.php>.
- Google Maps, <http://maps.google.com/>.
- Yahoo! Maps, Driving Directions, and Traffic, <http://maps.yahoo.com>.
- Garcia-Holina, H., Ullman, J.D. and Widoma, J. (2000) *Database System Implementation*. Prentice Hall.
- Roiger, R.J. and Geatz, M.W. (2003) *Data Mining: A Tutorial-Based Primer*, Addison Wesley.
- Adriaans, P. and Zantinge, D. (1996) *Data Mining*, Addison Wesley.
- Moustakis, V.S. Lehto, M. and Salvendy, G. (1996) Survey of expert opinion: which machine learning method may be used for which task? Special issue on machine learning of *International Journal of HCI*.
- Lavrac, M. and Wrobel, S.K. (1995) *Machine Learning: ECML-95*, New York: Springer Verlag.
- Wikipedia, the free encyclopedia, <http://en.wikipedia.org/wiki/>.
- Horowitz, E., Sahni, S. and Freed S.A., *Fundamentals of Data Structures in C*, Computer Science Press, 1993.