Situation Awareness in Web-based Environmental Monitoring Systems

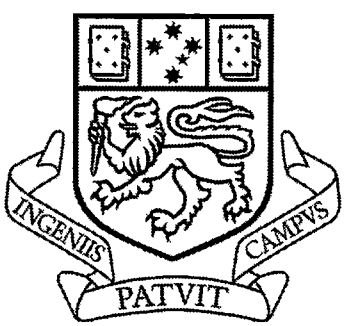
By

Meng Zhang, MComp

A dissertation submitted to the
School of Computing and Information System
In partial fulfilment of the requirements for the degree of

Master of Computing

October, 2010



University of Tasmania

Declaration

I, Meng, Zhang, declare that this thesis contains no material which has been accepted for the award of any other degree or diploma in any tertiary institution. To my knowledge and belief, this thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

Meng, Zhang

Abstract

The Tasmanian ICT of CSIRO developed a Sensor Web test-bed system for the Australian water domain in south-east of Tasmania. This system provides an open platform to access and integrate near real time water information from distributed sensor networks.

Traditional hydrological models can be adopted to analyse the data on the Sensor Web system. However, the requirements on high data quality and high level domain knowledge may greatly limit the application of these models. To overcome some these limitations, this project proposes a data mining approach to analyse patterns and relationships among different hydrological events. This approach provides a flexible way to make use of data on the Hydrological Sensor Web.

To make data mining rules easy to understand, this project developed the user friendly interface demo to visualize the data mining rules and describe the application of sorts of rules.

Acknowledgement

I would like to thank everyone that has supported and helped on completed this thesis, especially my supervisors, Professor Byeong Ho-Kang and Dr Quan Bai (Tasmania ICT Centre, CSIRO). I am deeply indebted to your valuable, welcome, timely and plentiful feedback. Without you I would still be getting big headache to figure out my research work. Great thanks to you for sharing your life experience and brighten up my future.

I also would like to thank Mr Andrew Pratt (Tasmania ICT Centre, CSIRO) for providing the resource about the Sensor Web and supporting about the basic system about the operation of the South-Esk Sensor web.

Also, I would like to thank to Mr. Chris Macgeorge and Mr. Simon McCulloch from the Bureau of Metrology. They give me perfect feedback about the data mining rules and explain the domain knowledge in water phenomenon patiently.

Most importantly, I am forever indebted to my family members for their physical and mental support and encouragement throughout this year.

Table of Content

List of Figures and Tables

1. Introduction

The Sensor Web is a distributed information system that connects sensors, observation archives, processing systems and simulation models using open web service interfaces and standard information models (Open Geospatial Consortium, 2007). As shown in Figure 1-1, the Sensor Web is targeting at providing an open platform for users in different domains and organizations to share their data, information and knowledge. It can enable the integration of web-enabled sensors and sensor systems for a global scale collaborative sensing.

The Tasmanian ICT Centre, as part of CSIRO Water for Healthy Country Flagship project, has developed a Sensor Web system called the South Esk Hydrological Sensor Web (SEHSW) (Liu et al. 2010). It collects sensor data from five water agencies, and real-timely monitors the South Esk catchment, which covers an area of approximately 3350 square kilometres in north-east of Tasmania. SEHSW makes the South Esk catchment a sensor-rich environment with an opportunity to study the hydrology phenomenon in a catchment and test-bed to study the development of Sensor Web systems. Figure 1-2 shows sensor distribution on the SEHSW.

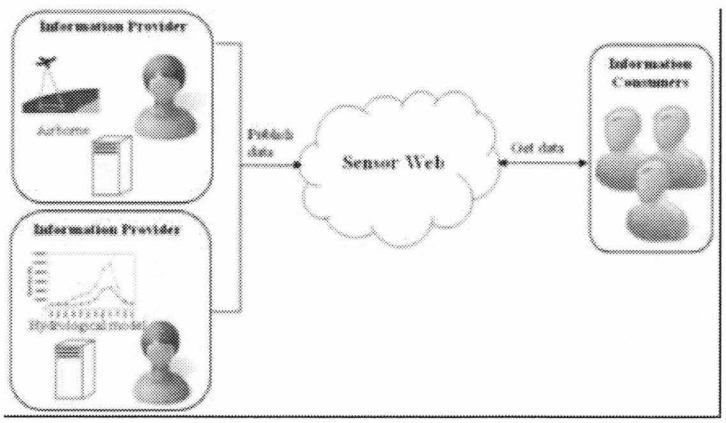


Figure 1-1: Conceptual Architecture of Sensor Web

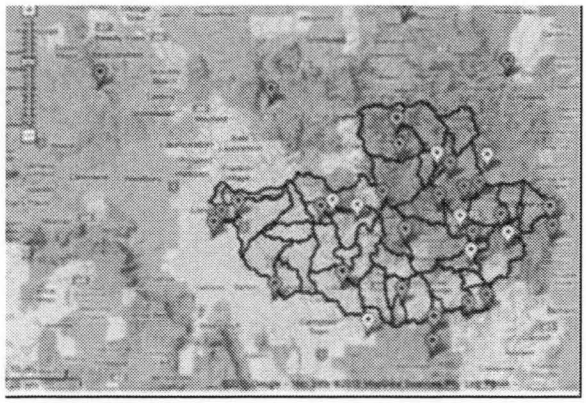


Figure 1-2: Sensor distribution on the South Esk Hydrological Sensor Web

With the deployment of the SEHSW, huge amount of real time data are collected from the physical environment and shared on an open platform. Meanwhile the value of data needs to be exploited by some data analysis and knowledge discovery approaches. In the water domain, hydrological data are normally manipulated and interpreted by traditional hydrological models. These models are very powerful in stream forecasting, water quality monitoring, etc. However, most hydrological models have some limitations which can block their applications:

- Most of hydrological models ask for high level domain knowledge to generate the structure and analyse data.
- Most of hydrological models require high quality data. Due to the particularity of water related data, the real water data ask for the integrity and veracity. The data also need to manipulate accuracy during the process.
- Most of hydrological models have the specific region.

Compared with hydrological models, data mining provides a more flexible way to achieve knowledge discovery (Liang et al. 2001). It can analyse hidden patterns and relationships among data from a data-driven perspective. Namely, it can be operated based on available data. In this paper, we propose the use of association rule based approach to analyse hydrological data. The proposed approach can interpret data on the SEHSW to useful information without high domain expertise.

2. Literature Review

T his chapter will describe some issues review which focused on the understanding of situation awareness and sensor network data. In addition, data mining methods and data visualization related information also take into the consideration. It discussed the related knowledge and selected suitable methods to become the points of this research.

2.1 Situation awareness

According to Endsley's definition (1995a), Situation Awareness (SA) is "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future". For environmental monitoring systems, SA means that to find ways to make improvements for the explanation of the environmental systems. SA can be described as a theory, an activity, a product or an understanding process (Durso & Sethumadhavan 2008). Some people have different views about the definition of the SA. For example, according to Garsoffky, Schwan and Hesse (2002), the video of soccer goals can give the participants different impression to recognize the memory of the details in that radio. Hence, researchers in different domains may hold different opinions and interpretations about SA. In this project, SA is the specific awareness related to events or phenomena in the nature environment.

2.1.1 SA in hydrological sensor system

Different people may hold different opinions about situations in the water domain. For example, bush walking amateur may interest on extreme weathers such as snow and storm. Water managers may interest on related phenomena related with catchment statuses, e.g., humidity, soil moisture, etc. In this research, we focus on the interpretation of sensor data, so that more useful information can be provided to users

to identify situations based on their requirements.

2.2 Sensor network data

With the development of sensor techniques, various kinds of sensors involving video cameras, thermometers, obstacle sensors have been used in different domains (Ikuhisa et al. 2009). Sensors can be installed by different companies or organizations. The data collected by sensors can not only used for the initial purposes of the sensor owners exclusively, but also many other people if the sensor data can be shared and published to the public. For an instance, observations of a weather station can not only provide useful meteorological information like wind speed and air temperature, but also road safety related information such as the melting speed of road surface ice (Anthony & Vinny 2004). Take another example, a calculagraph can be used to identify and respond the changes in spatial context in mobile computing devices (Anthony & Vinny 2004). Moreover, data collected by sensors can be processed, interpreted in different ways, and as a result, different interpretations may fit for purposes of different users (Peter et al. 2005).

2.2.1 The features of Sensor data

Observations from sensors can be generated and updated very frequently. As a result, it is a big task to navigate and manipulate sensor data (Kiyoharu et al. 2004). Especially for large scale sensor networks, a wide range of sensors, which can deliver numerical, discredited and digitized data streams, are used to monitor and manage the nature systems (Klein & Lehner 2009). In a sensor network, a sensor node may only possess very limited sensing and computing capabilities. Sensor nodes in a sensor network can organize and communicate themselves with others to transmit power control and antenna in order to adjust the transmission. Namely, sensor nodes can work individually for simple tasks like collecting audio, seismic, and other types of data, and then higher level tasks can be achieved through collaborations within a sensor network (Stephanie 2001) .

Sensor networks can be used to observe complex phenomena and collect high level information. Environmental monitoring and event detection are two basic applications of sensor networks. Environmental monitoring can collect and utilize the data in scientific or other applications. On the other hand, event detections can only focus on detecting specific events. The data in sensor network is expected to be correlated in both spatial and temporal domain (Kevin et al. 2009) . Sensor networks normally can collect huge amount of data which makes data gathering an important issue. In a sensor network, the proposed compressive data gathering can reduce the communication cost of the global scale without including intensive computation or complex transmission control (Chong et al. 2009).

2.2.2 Relevance of the research

Sensor networks can provide huge amount of sensor data, but all of them are raw data. The purpose of sensor network is to provide useful information for users from collected data which means the ways to get information from data are important. In addition, it is also necessary to find suitable ways to disposal raw data and find the relationship between every sensor from the data.

2.3 Data Mining Methods

There are a number of definitions about data mining. In this research, we adopt the following data mining definition: data mining is a practical topic and concludes learning in a practical to find and describe structural patterns of data like tools to help explain the relationships between data and make some predictions from it (Ian 2005). Data mining applications are widely applied for detecting fraud, assessing risk, and product retailing. Data mining allows discovering previously unknown, valid patterns and relationships in large dataset by using of data analysis tools (Su et al. 2004).

For many domains, data mining methods play important roles especially in the aspect of data classification, quality assurance and pattern discovery. Data mining

approaches can be included in data warehouse environments (Han 1998). These methods provide a generic way for data analysis (Su et al. 2004).

2.3.1 Clustering algorithms

Clustering techniques (see Figure 2-1) can be used without predefined classes (Ian 2005). According to JAIN, MURTY & FLYNN (1999), clustering is an unsupervised classification of information or data items based on their features. The outputs of clustering can be expressed different ways. Namely, one instance can belong to a unique cluster or a number of different clusters (Ian 2005).

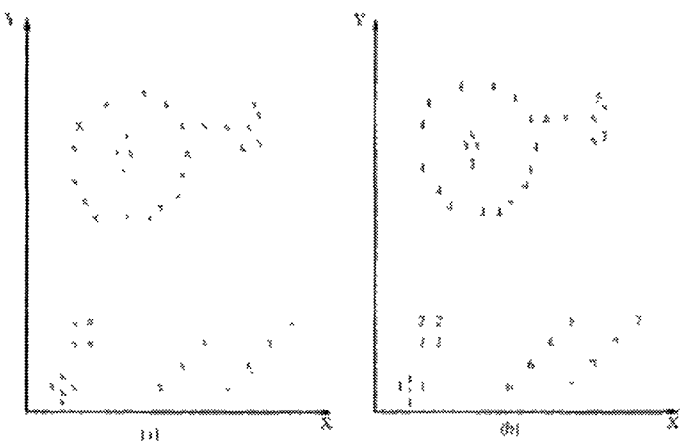


Figure 2-1: Data Clustering

Data clustering normally can be achieved in three steps, i.e., feature selection, inter-pattern similarity and grouping (see Figure 2-2). Feature selection is to identify the effectiveness when clustering data and feature extraction to generate the salient features. Pattern presentations are meant of the quality and quantity of the patterns. Grouping is to divide the data into suitable patterns (Jain 1999).

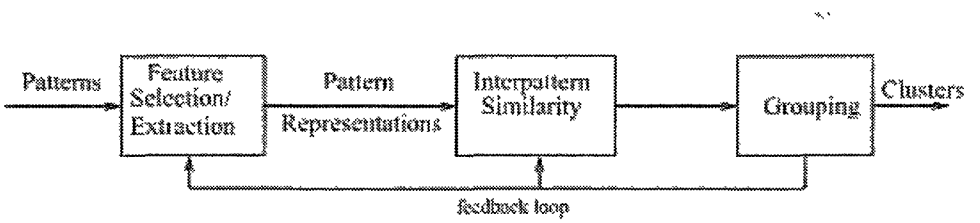


Figure 2-2: Stages in Clustering

In general, the clustering methods can be used to normalize data and transfer continuous data into nominal data type. It is obvious that the clustering can improve the quality of data and divide the data into reasonable groups.

Related technologies for clustering

To cluster the data, it needs to set some mechanism to manipulate the data. One of the classic clustering methods is called k-means clustering. k-means clustering is normally operated in the following steps (Ian 2005):

1. Define a parameter k which indicates the number of clusters;
2. Based on the ordinary Euclidean distance metric, all distance can be assigned to their closet group centre;
3. The centroid of instances is calculated which called the process of "means".
4. New cluster centres can be generated.

However, the problem of k-means is exist which is about the number of clusters and the location problem. One of them is to minimize of the sum of distance of the nearest centre called Euclidean k-medians, another one is to minimize the maximum distance from every point to its closets centre called k-centre problem. To solve these problems, the k-mean algorithm is generated which is based on a simple iterative scheme for finding a locally minimal solution (Tapas 2002).

To compare with k-mean algorithm, another effective method called EM (expectation-maximization) algorithm can have a better performance in statistical estimation problems which involve incomplete data or the problems like mixture estimation (Borman 2004). In that algorithm, "expectation" means to calculation the cluster probabilities which can be the "expected" class value and the "maximization" is meant that to calculate of the distribution parameters which is the data given by the likelihood (Ian 2005). So EM algorithm is the good way to calculate the Maximum likelihood estimate (Borman 2004).

There are a number of applications of the EM algorithm. In some applications, the EM algorithm is used to estimate data sequences transmitted instead by continuous phase modulation in a multi-path channel which concludes a random phase, amplitude and the time delay. In this situation, the simplified EM algorithm makes the maximization of desired likelihood function about the adequate amount of training data by some general formulation to do explicit calculation to find the estimation and provide the likelihood (Linda & Hisashi 2002).

There are several variants can be happened when use the k-means algorithm. Firstly, the resulting clusters can be splitting and merging. Secondly, the clusters can be split in a pre-specified threshold and the two clusters may be merged when the distance between these two clusters is lower than another pre-specified threshold (Su et al. 2004).

Both k-mean and EM algorithms has specific requirement for data. On the other hand, there is another specific algorithm is to build a probabilistic hierarchical tree by the CU (category utility) (Mi 2004). Based on the computation of the probabilities, when the category utility is maximal, the every instance can be read per iteration from the dataset and the algorithm can incorporates instance in to the trees by descending the tree with four possible operations at each node (Mi 2004).

- Put the instance into a exist cluster
- Create a new cluster
- Merge the best two clusters with respect to the values of CU.

Split one clusters into several clusters by lifting its children one level in the tree to replace it.

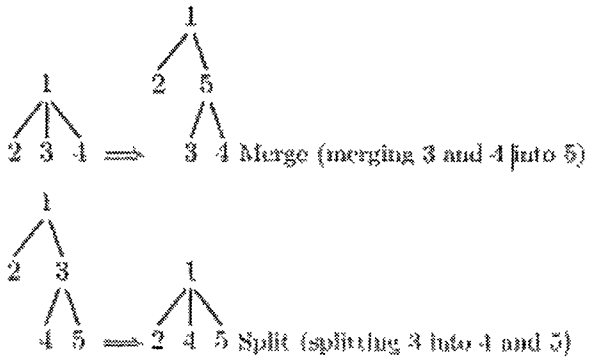


Figure 2-3: Merge and Split Operation

The node split and merges for the higher CU situation is displayed by Figure 2-3(Nachiketa et al. 2006).

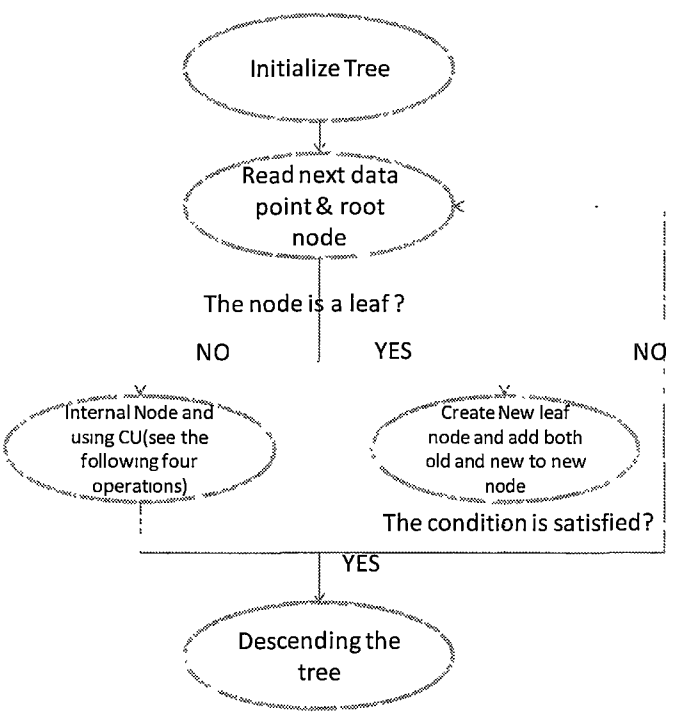


Figure 2-4: The Definition of the Cobweb Algorithm

Figure 2-4 shows the common steps of the process of the Cobweb algorithm. The tree in this figure can be constructed above those steps for every instance. So this algorithm may output a massive hierarchy for large datasets (Mi 2004).

Aims for hydrological sensor web by using clustering methods

Normally, the clustering methods are used for initial implantation during the whole process. Clustering methods can normalize or discrete continuous data. Considering that most of complex mining methods have data requirement with nominal type, clustering methods become the basis technique to disposal data as for the first step. In this research, we are going to take the first step in the clustering data which is collected by the sensors. After collecting the raw data from sensors, we prefer to use some clustering method to help us normalize the data and build up the suitable dataset as for the input and get ready to use more complex process or methods to manipulate the data and find more potential relationships or patterns.

Based on the introduction of some clustering algorithms, at the first step of using data mining methods, it is important to select suitable clustering methods to normalize data. However, this research does not focus on comparing algorithms or creating new algorithms so that the target is to get the nominal type of data rather than spending time on selecting algorithms or improving specific algorithms.

2.3.2 Association rules

An association rule is to find the potential relationship between data and make the rules to predict the class or instances of the dataset (Ian 2005). The classification rules is similar with association rules which can be used to make prediction for class of the dataset. However, the biggest difference between association rule and classification rule is that the association rules can make prediction for the attributes of the dataset. What is more, the amount of the association rules may be massive and some of them are related so that the rules need to reduce as well (Ian 2005).

Examples of association rules

A classic example is the market-basket problem (Arawal & Srikant 1994). In

transactional data-sets of supermarkets, rules like {bear}→{diapers} means that most of customers buy bears, meanwhile, they may buy diapers as well. Such rules will then suggest the supermarket managers to put bear and diapers together which may improve their profits. From this example, it can be seen that association rules provide a useful mechanism for discovering relationships among the underlying data. Also according to the Toivonen (1996), this kind of example can help manager to manage the situation of related goods and help customers to find their interesting products easily. In addition, those rules or relationships can improve the profits for the supermarket.

From this example in supermarket, it is obvious that the association rules can help firms to find the potential relationship in related goods and managers can make good planning and decision to distribute goods to improve profitability. Also for the sensor web, association rule mining can generate related relationships to help domain experts or normal user find more information from sensor data. The input of association rule mining is the data which requires lower quality than hydrological model and the output is rules. Those rules describe the relationship between data and hydrological events.

Related technology

In this research, we define a rule is defined as a form A = {L=> R}, where A is an association rule, L and R are two disjointed item sets of event E ($L \cap R = \phi$). L is called the antecedent of A; and R is called the consequent of A. There are two important constraints for selecting useful rules, i.e., the minimum thresholds on support and confidence (Jiawei & Micheline 2006). Support is the ratio of the number of transactions that include all items the antecedent (S(L, R) = P(L)). Confidence is the probability of finding the consequence in transactions under the condition that these transactions also contain the antecedent (C(L, R) = P(L | R)). There is another threshold called frequent which is calculated by support and confidence. This is the

third significant metric for association rules and named lift in association rule mining. Lift value can be calculate based on confidence and support: lift (X=>Y) = confidence (X=>Y) / Support (Y). Only rules with lift value greater than 1 can be considered into good rules. There are a number of learning algorithms which can find association rules based predefined support and confidence thresholds.

The support and confidence is the key standard to justify the quality of rules. The rules with both high support and confidence can be called useful roles. For example, in the supermarket, the rules: bread => rose has confidence value over 90%, however, with less than 0.01% of support. This means that about one of 10000 customers may buy both bread and rose together. It is obvious that this rule is useless for its frequency. If managers put bread and rose together, to compare with separating them in different locations, there will be nothing happen. Also the rules with high support and low confidence cannot be used into applications. So the key point to select useful rules is to set up the support and confidence value at initial time before utilizing the association rule mining and select the high level rules with both high confidence and support values.

Relevant of the research

In this research, we do not focus on the algorithms comparison and discovering new algorithms. Based on previous description of situation awareness, we focus on building models to make sensor web and sensor data meaningful, thus the association rules can become the aim to help us to find the relationship or patterns in sensor data. Because association rules mining is focus on finding potential relationships in data which is the target for the sensor web. So we set this method as a component of the model. If the result or relationships are not useful for sensor web, we can find other methods instead of association rules which can be the future work for this research.

2.4 Data Presentation and Visualization

In order to facilitate users to understand discovered knowledge, data presentation becomes to an important issue for the success of the approach. Data visualization tools or technologies can allow people easy to identify the interesting information or patterns by creating dimensional pictures or views. It is the keys to help people to make decisions about the quantities data (Soukup & Davidson 2002). According to Usama, Georges and Andres (2001), visualization can be used to explore different kinds of data, allow users to become a viewer and make complex data reification. Also, data visualization technology can prove to explore data analysis in large database especially when little is unknown about the data and related meaning. (Keim 2002)

It is a suitable way to combine data mining method and data visualization technology together to make application in hydrological sensor web in this research. Data visualization provides a flexible way to present the analysis and the result of data mining methods. Visualized data also can make communication between users, hydrological events and domain experts.

Data type in visualization

According to Keim (2002), different types of data have different methods to visualize. Figure 2.5 describes the general techniques for data visualization in different data types. It is easy to find that the data visualization methods can fulfil almost amounts of data so that it can not only present the sensor raw data in sensor web, but also visualize the relationship between data and hydrological events.

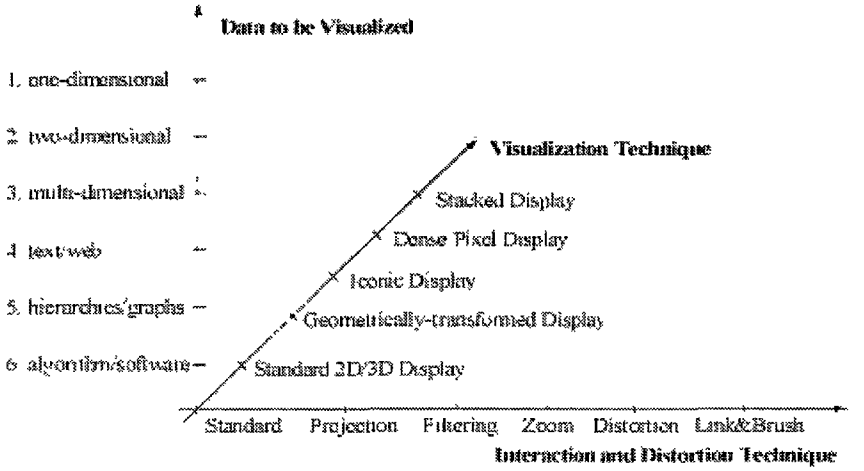


Figure 2-5: Information Visualization Techniques

The target of data visualization is to simplify the complex or huge data and transfer into friendly interface. In addition, the visualization achievement is justified the user preference model. User preference model is always required some AI technique for input in general (Linden & Hanks 1997), but it is not good for complex domains. So a reasonable approach is to build communications to allow user to support the decision and filter the information (Thomas & Fischer 1996). To build up the user preference model can also help both domain experts and normal users to understand complex process or information easily.

For this research, to build up the visualization interface and user preference model is the advanced awareness in the hydrological sensor web which can help people to make fully understand for hydrological events and the relationships. It is also the improvement for the data mining methods and allows data mining results to be visualized.

3. Methodology

This chapter describes the methodology of the research and related approach to implement this research. The methods of the methodology and the visualization of the result are based on the related literature as shown in last chapter (Chapter 2). The main highlight of this research is to use mathematic mechanism to find the potential relationships or patterns in hydrological related sensor data without the domain knowledge.

3.1 Introduction

The main idea of this project is to find the potential patterns in hydrological related data which is collected by the sensors. In order to calculate and analyse the data, data mining approaches provide the flexible ways and can be used. Data mining focus on the data and it does not require the data with high quality which means that the data can be not continuous which cannot make any influences to the result.

The approach of the data mining for this research was to use some of the methods to analyse data and give the result about the potential relationships or patterns of the data which can help people to understand the hydrological phenomenon used to facilitate water resource management, event predictions and catchment monitoring.

The objective of this research is to develop an effective and generic approach about the hydrological system and the approach can be used in any monitoring systems and generated by non-domain experts or people with a little domain knowledge.

Based on the miming methods and sensor data, the process flow of this research is illustrated in Figure 3-1. In the approach, hydrological data are transformed into a proper format and sent to the association rule analysis component. Then a number of

association rules will be generated from the rule analysis component. These rules will be pruned and stored in a rule base. Finally, the stored rules can be presented based on users' queries.

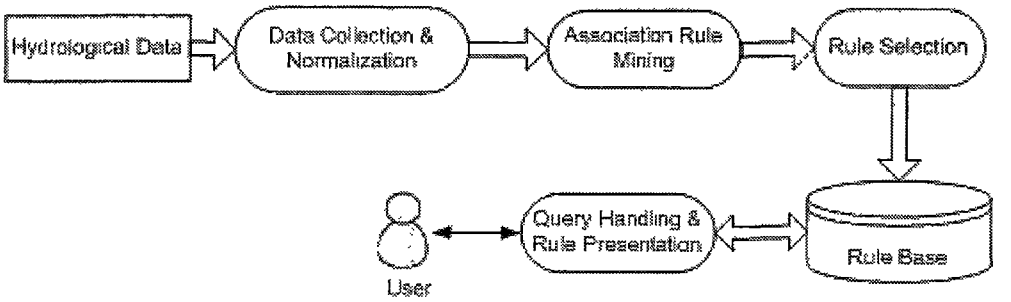


Figure 3-1: The Process Flow of Mining Data

The scope of using data mining methods and finding useful results was constrained in the following ways:

- Focus on the collecting and normalizing sensor data from the sensor web and generating the suitable format of the data before using mining methods.
- Focus on the result of the mining data methods and analyzing the result rather than comparing different algorithms
- Focus on the data visualization of describing the result of mining methods and which way the result can be used.

3.2 Building the data mining model

To compare with the other methods, data mining methods provide the flexible way to disposal hydrological data. Figure 3-2 shows the process flow of the data mining application. In order to build up the initial dataset, we need to set up the purpose to select different types of data from sensor web. The initial dataset concludes the specific spatio-temporal hydrological events or phenomenon. Then, because the association rule mining requires input data with nominal type, we need to discrete data from initial dataset and transfer the continuous data into nominal type and build

up the nominal dataset as for the input. After that, we can use association rule to generate related rules. Next we use the standard value of support and confidence to justify rules. Finally we filter the useless rules and put the useful rules into the rule base.

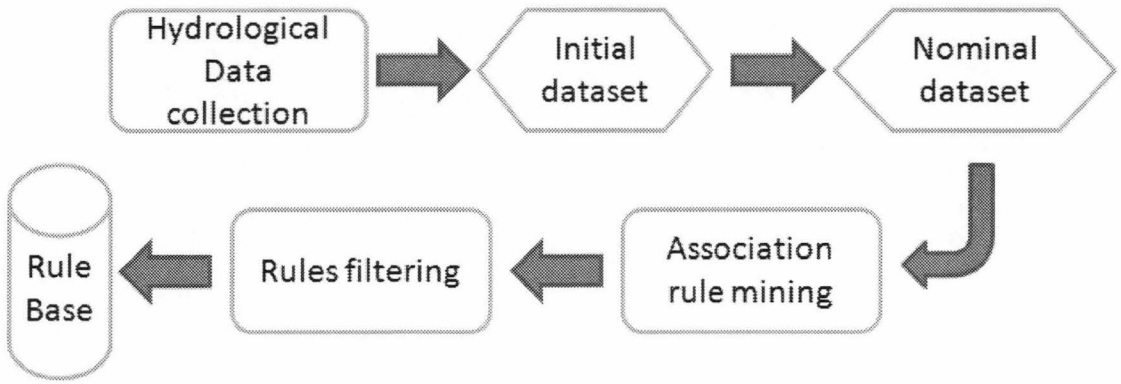


Figure 3-2: Data Mining Process Flow

3.3 Implementation

3.3.1 Data collection and normalization

Based on the features about the sensor data (Ikuhisa et al. 2009), the sensor data are huge and update frequent so that it needs to find ways to collect data and build up the suitable dataset before doing the data analysis. Therefore the data collection and normalization are important component during the process of the mining data.

3.3.1.1 Data Collection

We are expecting to discovery spatio-temporal patterns in different hydrological events. In this research, we are particularly focus on analysing the relationships between rainfall events and other phenomena, e.g., humidity, air-temperature, etc.

To achieve the purpose, firstly, we record the maximum or minimum values of rainfall events from the sensor web, and extract them into a new database. To record the rainfall event, there are two different sensor records as shown on Figure 3-3. The left

one sensor records the rainfall as a big bucket and the value will be recorded and improved when rainfall happens so that the value of this sensor is always increased until it reset. To catch the maximum value of rainfall in this sensor, we just calculate the time gap between each changed value and the lower the time gap is the higher the rainfall value is. Based on that phenomenon, we can catch the maximum rainfall value time.

The right status is another type of sensor to record the rainfall value. At this moment, this sensor records the value like a small test-tube which has small capacity, when this test-tube catch the rainfall inside, it will update every settled time and record the rainfall value at that time. In order to catch the maximum value in this sensor, we can record the maximum record value and the maximum update frequency.

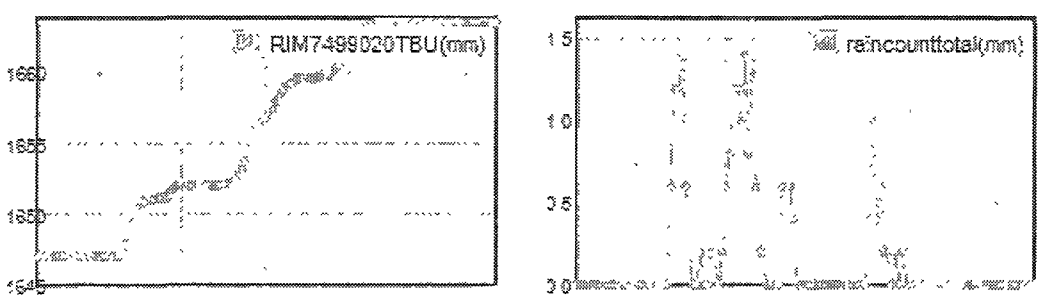


Figure 3-3: The Two Different Types of Sensor Record About Rainfall Event

After recording the maximum time of the rainfall, in addition, to simplify the implementation, we also assign an index to each location and phenomenon type. Table 3-1 describes index values for different location and phenomenon types. In this table, location 1 represents the locations that a corresponding phenomenon may happen (e.g. Humidity). Location 2 represents the locations of rainfall events.

Location 1	Index (J)	Phenomenon	Index (K)	Location 2	Index (L)
Ben Lomond	1	Humidity	1	English Town Road	1
Story Creek	2	Air-Temperature	2	Valley Road	2

Ben Ridge Road	3	Evaporation	3	Hogans Road	3
Avoca	4	Transpiration	4	Mathinna Plains	4
Tower Hill	5	Wind-Run	5	Tower Hill Road	5

Table 3-1: The Description of Location and Phenomenon

3.3.1.2 Data normalization

From Table 3-1, we can calculate the time gap between the maximum value or minimum value of a rainfall event and another event within the same day by using Equation below.

$$Max\_Gap(Min\_Gap)_K = Max(Min)_K - Max_L$$

Figure 3-4: Formula of Calculating Value Gap

Then, we can get two sets of time gaps which indicate the time differences between the rainfall event reaches the maximum value and another event reaches the maximum and the minimum values. The time gaps are described in a continuous data type. However, association rules can only take nominal or ordinary data types as inputs. To satisfy this requirement, the continuous values need to be transferred to a nominal style. Here, we use a simple clustering technique to achieve the conversion as described in chapter 2. Figure 3-5 shows the clustering method we used. The method transfers the continuous values into nominal items by generating different clusters. Each cluster contains a range of continuous time gaps. For instance, the cluster (Max_Gap(0_4)) covers the time gaps between 0 to 4 hours.

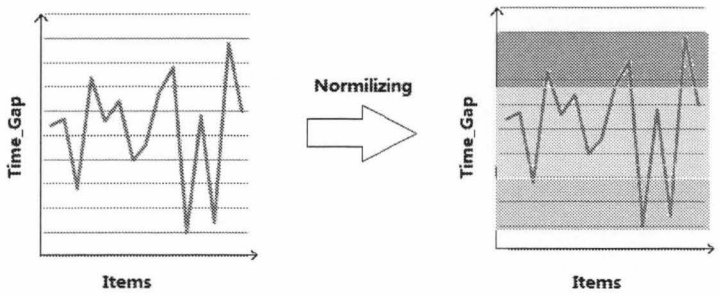


Figure 3-5: Transfer Data into Nominal Style

After that, we finish collecting and normalizing hydrological data and build a new database to store the nominal data. This primary database concludes the gaps between rainfall and other phenomenon, the locations and the phenomenon. The next step is to select a suitable platform or workbench to analyse data by using association rule method and find the rules about the relationships or patterns in hydrological event.

3.3.2 WEKA workbench based analysis tools

We described the processes for data pre-processing in the previous subsections. After that, we need to operate association rules mining on a platform or workbench. There are a number of data mining tools which can conduct association rule mining. In this project, we select WEKA workbench to analyse hydrological data.

WEKA(stand for Waikato Environment for Knowledge Analysis), was developed at the University of Waikato in New Zealand, provides a uniform interface to lots of algorithms for pre- and post processing and for evaluating the result of the learning schemes on any dataset(Ian 2005). This work bench can not only provide algorithms to analyze data, but also allow the researcher to advance the existing algorithms or implement the new algorithms without considering any support infrastructure(Mark et al. 2009).

To compare other platform, WEKA provides an open interface which allows users to manipulate it easily and flexible support algorithm enhancement. We can achieve our

target via WEKA workbench by using existing functional interface. In addition, WEKA provides opportunities to make enhancement of existing algorithms which can improve the continuing work of this project by using data mining methods in hydrological domain.

3.3.2.1 Data preparation for WEKA

WEKA requires a specific input file format named ARFF which is the best format. It also allows other formats of data file like .csv or C4.5 (Hall 2009). To analyze the hydrological data, we select to convert the normal data from database that we created in previous section into ARFF format file. There are three major components in ARFF files. The label "@relation" is used to make a simple description about the dataset; "@attribute" refers to the titles of each item; "@data" means that after this label all data are values of items in transactions.

In order to using WEKA workbench, we need to transfer the normal data format into ARFF file. As shown in Figure 3-6, the data in brackets (after each attribute) indicate the possible values of that attribute. The data after the label "@data" are values of items in transactions. At this moment, we set the initial dataset with five attribute, the location, item, compare_location refer to Location 1, phenomenon, Location2 that shown in table 3.1. The max_gap and min_gap refer to the time gap which is calculated by previous formula. Then each instance describes the values of each attribute and every time gap value of instance is based on same day. For example, the first instance means in Ben Lomond, the maximum value time of wind speed is 6 and half hours earlier than the maximum value time of rainfall in Rabbit Marsh. Also, the minimum value time for wind speed is 4 and half hours earlier than the maximum value time of rainfall in Rabbit Marsh. In addition, both gap happened in the same day.

test_data.arff ×

```
1 %this dataset is to record the time gap between some items and rainfall.
2 %all data is based on 24 hours series
3 @relation 'water'
4 @attribute location {BenLomond,StorysCreek,BenRidgeRoad,Avoca,TowerHill,Cox
5 @attribute item {Windspeed,Humidity,airtemperature,evaporation,transpiratio
6 @attribute compare_location {BenLomond,StorysCreek,BenRidgeRoad,Avoca,Tower
7 @attribute max_gap real
8 @attribute min_gap real
9 @data
10 BenLomond,Windspeed,RabbitMarsh,-6.5,-4.5
11 BenLomond,Humidity,RabbitMarsh,6.5,10.5
12 BenLomond,airtemperature,RabbitMarsh,2.5,1
13 BenLomond,Humidity,RabbitMarsh,11.5,-3
14 StorysCreek,airtemperature,RabbitMarsh,2.5,11.5
15 StorysCreek,evaporation,RabbitMarsh,0.5,0
16 StorysCreek,transpiration,RabbitMarsh,-3.5,4.5
17 BenRidgeRoad,Humidity,RabbitMarsh,4.5,0
18 BenRidgeRoad,airtemperature,RabbitMarsh,0,1
19 BenRidgeRoad,evaporation,RabbitMarsh,-0.5,0
20 BenRidgeRoad,transpiration,RabbitMarsh,4.5,-12.5
21 Avoca,Humidity,RabbitMarsh,-5.5,3
```

Figure 3-6: The Initial ARFF File in Hydrological Data

3.3.2.2 Using Explore to input dataset

There are four components in the initial interface of WEKA, based on the description of chapter 2, for this research, we prefer to use Explorer to generate the association rules. Figure 3-7 displays the initial interface of setting input value in WEKA explorer. The WEKA Explorer provides a friendly interface to realize different kinds of data mining methods (e.g. Classify, Association rule. etc). To manipulate data mining methods in Explorer, we need to select suitable dataset and select target methods and set up the target options of method. Then we can get the result of analysis on that interface directly.

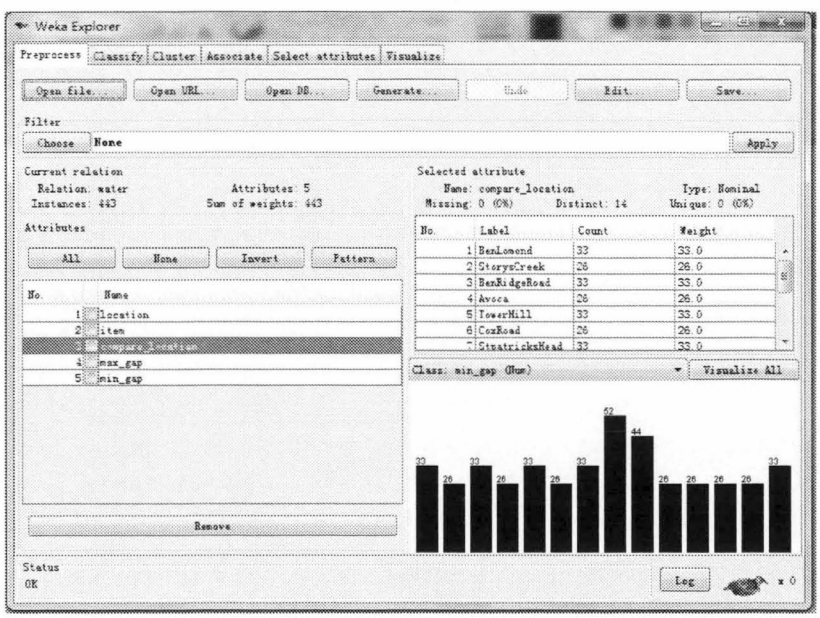


Figure 3-7: Interface of WEKA Explorer

We had made the ARFF format dataset file so that in this step, we just open the ARFF file we had made and select the target method to make analysis. For this research, we select the association rule method to analyse data and find the relationship or patterns in hydrological data. WEKA Explorer provides several algorithms to support and analyse the association rules. Based on the discussion of Chapter 2, we select the classic algorithms called Apriori algorithms. Because we are not focusing on the algorithms analysis which can be another big issue to discuss and analyse in this project, we just select one of the normal methods which are suitable for hydrological data analysis. Before we generate rules, we need to set up related options by using Apriori algorithms. For this experiment, we set the support threshold as 10% to 100%, and the confidence threshold as 0.5. Due to the features of hydrological data, the confidence threshold was not set to very high.

3.3.3 Generate Association Rules

For this implementation, in this research, we build up the dataset that concludes 443 instances to generate related rules. During this process, we input the 400 instances to generate rules and set up 43 instances (10% of total) to evaluate rules.

3.3.3.1 Rules filtering

As we mentioned in chapter 2, there are two important attributes, which named support and confidence, used to justify the quality of rules. After generating the hydrological data, we had got lots of rules to describe the relationship of the hydrological events. Different rules have different support and confidence value. Figure 3.8 shows the distribution of confidence and support for 17 rules. The X axis is the confidence value and the point is located into the related support value. The total number of instances is 443 and Y axis shows the support value. For instance, the

support value is 125 means that there are 125 instances satisfying the rules. We filtered the rules which have big gap between support and confidence value.

We had known that the rules will be useless if the gap between support and confidence value is large. For example, the rule: {(item = rainfall, location = StorysCreek) =>(compardlocation = Avoca, max_gap = [3-7])} have 100% confidence value but with 1 of support value which means in 400 instances, only 1 of them support this rules. It means that this phenomenon is the rare event and the related rule is useless.

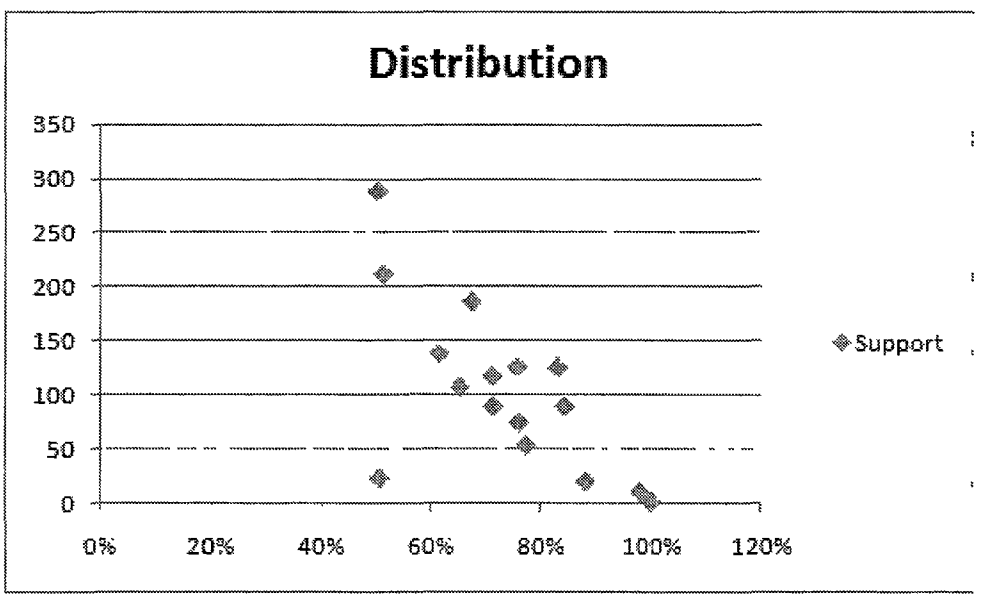


Figure 3-8: Distribution of Support and Confidence

After the filtering all of rules, we have got 10 rules as for the result which have more than 1/3 of total instances value support. Then we use another 44 instance to justify the rules. Table 3.2 shows the result and evaluation of association rules generation.

	Rules	Accuracy (40 instance evaluate)
Total number	10	Average 80%
With confidence (> 80%)	4	85%
With confidence (<80%)	6	75%

Table 3-2: The Result of Association Rules

3.3.3.2 Rules analysis

After the generation, we get some rules for hydrological data. The rules are basically related with phenomenon types and time gaps. Based on identified rules, we can find the relationships between some specific phenomena (e.g. rainfall and humidity), and furthermore, make some predictions on hydrological events. Here is the following are some typical rules examples:

Rule = {(max_gap:[3-7]) => (item: humidity)} conf 0.83

This rule means that regardless the location, the humidity should get the maximum value 3 to 7 hours later than rainfall get the maximum value. For instance, if a rainfall event gets the maximum value at 9:00 am, the humidity will get the maximum value between 12:00pm and 16:00pm. The support of this rule is the 20% of max_gap and 40% of humidity. This rule has the high confidence with 83%. So based on this rule, we can find that the phenomenon types and time gaps between humidity and rainfall. In addition, we can make prediction about the maximum time of humidity via the maximum time of humidity.

Rule = {(min_gap:[(-5)-(-2)]) => (item: airtemperature) } conf 0.675

This rule shows the relationship between minimum value time of air temperature and the maximum value time of rainfall. It means that regardless the location, the air temperature can get the minimum 2 to 5 hours earlier than rainfall get the maximum value. The related confidence is 67.5% which is lower than previous rule.

Compare with those two rules, they have similar content and different confidence value. Both of them are regardless the locations and focus on specific phenomenon with related values (max or min). Then, the different confidence means the different probability. It is obvious that the relationship between humidity and rainfall is

stronger than relationship between air temperature and rainfall. Besides, there are some other related rules that conclude relationship with locations.

Rule: {(Item = airtemperature), (location = BenRidgeRoad) => (max_gap[1.5-9]) conf 0.62

This rule has built up the relationship in most of attributes: location, item and value type. The detail information of this rule is that in Ben Ridge Road, the gap between air temperature of maximum value and rainfall of maximum value is 1.5 to 9 hours. It is easy to find that the range of time gap is large than previous rules because the related items in this rule has expanded. Also, the confidence value becomes lower with only 62%. In addition, this time range (1.5-9) almost covers 1/3 daytime which means the if we want to get more accurate value from this rule such as make the time range smaller, the confidence will decrease sharply even lower than 40%. So this kind of rule is not as good as previous even it concludes more items.

In conclusion, based on that dataset, most of rules can build the relationship between phenomenon and value type with different confidence. However, if the rules contain more items, the confidence will become lower and the time range will become larger. It is obvious that those rules with more items and high confidence are more useful and meaningful. To improve the quality of result, there are several ways: improving the algorithms or change the input dataset. In this research, we are not focus on the comparing related algorithms, thus we prefer to improve the input data set.

3.3.4 Rebuild the dataset and find more useful rules

In order to generate more useful rules, we prefer to rebuild the input dataset. There are five attributes in the initial dataset: two parts of locations, one item and two value type (max and min). It is hard to build the stronger relationships between locations and items. Then we reconsidered the component of the initial dataset and we found that

the locations are distributed separately in this dataset. For example, the same phenomenon in the same terrain location can get the similar maximum or minimum value time. Besides, the different location in different topography shows the different status in same phenomenon.

Figure 1-2 describes the distribution of locations in hydrological sensor web. In that Google map, different colours mean different topography. It is easy to find that locations are distributed in different topography. So we can combine the same topography locations into same group and generate several groups of locations to build up new dataset. Then we can generate new rules to find whether the new dataset can show stronger relationships with locations.

3.3.4.1 Build up new dataset

According to the distribution of locations, we can build up new dataset with additional attributes. In this project, we divide those locations into four different groups due to the topography. Based on table 3.2, we generate two new titles named location1area and location2 area to represent the meaning with groups of location1 and groups with location 2. The four level group of location are generated by the altitude of the topography. The highest locations are named as area 1 and lowest locations are named as area4. Figure 3.9 shows the new version dataset which concludes 7 attributes.

test_data_advanced.arff

```
1 %this dataset is to record the time gap between some items and rainfall.
2 %all data is based on 24 hours series
3 @relation 'water'
4 @attribute location {BenLomond,StorysCreek,BenRidgeRoad,Avoca,TowerHill,Coxi
5 @attribute item {Windspeed,Humidity,airtemperature,evaporation,transpiratio
6 @attribute compare_location {BenLomond,StorysCreek,BenRidgeRoad,Avoca,Tower
7 @attribute max_gap real
8 @attribute min_gap real
9 @attribute location1area {1,2,3,4}
10 @attribute location2area {1,2,3,4}
11 @data
12 BenLomond,Windspeed,RabbitMarsh,-6.5,-4.5,3,3
13 BenLomond,Humidity,RabbitMarsh,6.5,10.5,3,3
14 BenLomond,airtemperature,RabbitMarsh,2.5,1,3,3
15 BenLomond,Humidity,RabbitMarsh,11.5,-3,3,3
16 StorysCreek,airtemperature,RabbitMarsh,2.5,11.5,3,3
17 StorysCreek,evaporation,RabbitMarsh,0.5,0,3,3
18 StorysCreek,transpiration,RabbitMarsh,-3.5,4.5,3,3
19 BenRidgeRoad,Humidity,RabbitMarsh,4.5,0,1,3
20 BenRidgeRoad,airtemperature,RabbitMarsh,0,1,1,3
21 BenRidgeRoad,evaporation,RabbitMarsh,-0.5,0,1,3
22 BenRidgeRoad,transpiration,RabbitMarsh,4.5,-12.5,1,3
23 Avoca,Humidity,RabbitMarsh,-5.5,3,4,3
24 Avoca,airtemperature,RabbitMarsh,2,-11.5,4,3
```

Figure 3-9: The Advanced Version of the Input Dataset

As shown in Figure 3.9, location area1 and location area 2 are nominal data type so that they do not need to normalize before using association rule mining method. After building the new dataset, we can use the WEKA to implement new rules.

3.3.4.2 New rules analysis

At this moment, we still set up the support range from 1% to 100% and the confident threshold as 0.5. The amount of the instances is still 443 and 43 (10%) of them are used to evaluate the result.

After the generation, we got several rules. Some of them are same with previous because the initial attributes and instances are not changed. Besides, we got some new rules to describe the relationships between location areas and other items. Here are some successful rules:

Rule: {(location2area = 4), (max_gap = [3-6]) => (location1area = 3), (item = airtemperature)} conf 0.71

In this rule, it shows the stronger relationships between locations and other items than previous rules. In addition, the time range is smaller than previous as well. This rules means if one phenomenon in one location gets the maximum value 3 to 6 hours later than maximum value of rainfall in location area 4, the location will belong to location area 3 and the phenomenon will be the air-temperature. The confidence is 71% which shows that the probability is higher than previous rules.

Compare with two rules in different datasets, we can find that the second version rules concludes the first version rules, also the second version rules are better than first. The new version rules can make the connection in location, item and value type with high confidence. Then we also summarize ten reasonable rules to evaluate (we pruned the same rules in previous result). Table 3.3 shows the result of evaluation.

	Rules	Accuracy (40 instance evaluate)
Total number	10	Average 70%
With confidence (> 80%)	3	73%
With confidence (<80%)	7	67%

Table 3-3: The New Result of Association Rules

It is easy to find that the accuracy of this rules are lower than previous. However, those rules can show the stronger relationships in hydrological data and they are more meaningful than previous rules. It proves that if we want to get detailed rules, we cannot get the rules with high confidence.

The association rules mining are related the input dataset and it has some limitations about the size and content of the dataset. Also the accuracy of the rules are related the quality of datasets (Carlos 2006). So the dataset can make the influence for the quality of the rules.

4. Discussion

This chapter discusses the applications of patterns which are generated by the association rule mining approach introduced in Chapter 3. In this chapter, we will also introduce the feedbacks of our approach from some domain experts in the Bureau of Meteorology. Finally, we will introduce how the discovered patterns are visualized in our approach.

4.1 Introduction

After generating the rules by using the association rule mining, we invited some domain experts to provide some feedbacks. Facilitated by the CSIRO Tasmanian ICT Centre, we contacted and made an appointment with some experts in the Bureau of Meteorology to discuss about our approach and the rules we discovered. These discussions were motivated from three aspects:

- To find the objective ways about the association rules in hydrological events by discussing with domain experts.
- To find the objective ways about the association rules by comparing the difference between traditional hydrological model and association rules mining method.
- To find a suitable ways about visualizing rules in order to make rules easy to understand.

All the discussion of the result must be understood in the related of sample way, thus it also is indicative rather than conclusive. The conducted evaluations of this project have several limitations. If we want to justify the usefulness of the rules, we will find some potential users to test and verify the rules and give the feedback. However it is too hard to form a user group for this project due to time and resource limitations. To overcome this limitation, we conduct several discussions with some

experts in the water domain, and ask for their opinion about the approach we developed, and their preferred way to present the discovered patterns.

4.2 Discussion with domain experts

Based on the findings in Chapter 3, we know that the rules can be used to predict some hydrological events based on the occurrences of rainfall events. We need to evaluate these rules and find whether they can be used into the real world. Therefore, we ask for feedbacks from two domain experts in the field of hydrology, i.e., Mr Chris Macgeorge and Mr Simon McCulloch from the Bureau of Meteorology.

4.2.1 Meeting with Chris Macgeorge

Mr. Chris Macgeorge is an expert in the Bureau of Meteorology who is in charge of flood predictions and warnings. During the meeting with him, firstly, he introduced some traditional models to approach forecasting. Generally if he focuses on the forecast of a particular event, large amount of data that contains related data of that event in a long term (more than 10 years) are required to be collected to build and calibrate the forecast model. The model is normally based on complex mathematic formulas.

Figure 4.1 describes the processes of his forecasting approach. In his approach, he normally has a specific purpose at from the very beginning. Then, toward the purpose, he starts to design and develop a model. After that, large amount of related data are required to be collected from some particular locations. Obviously, the development of such models is complex and requires long term periods of data with high quality. In addition, such models are hard to be maintained and operated as they are normally based on complex mathematic formulas or methods. These models can generate very accurate forecasting result.

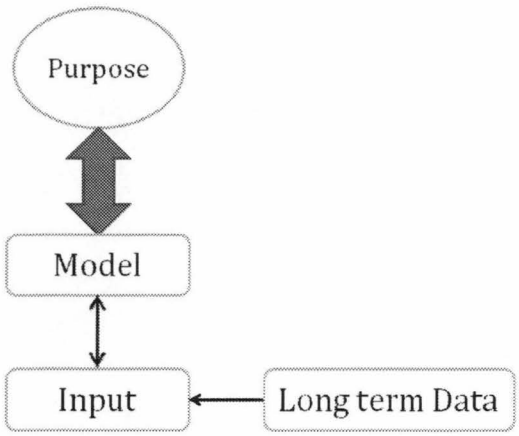


Figure 4-1: Simple Process of Traditional Data Model

In the meeting, we also introduced the rules that were generated by association rule mining. Comparing with the domain model, our rules are not powerful enough to generate forecasting in high accurate levels. However, Chris mentioned that data mining methods could be another suitable way to analyse data because this model is only focus on data itself. It means that if we can provide a wide range of data as input, data mining method can generate stronger and accurate result.

Finally, he suggested us to have discussion with another expert, i.e., Simon McCulloch who is in charge of modelling application and data analysis. He is also familiar with mathematical modelling, thus he can give us more feedbacks about related rules.

4.2.2 Meeting with Simon McCulloch

Mr. Simon McCulloch works in the Bureau of Meteorology and he takes charge in data analysis and observation. He was interested in our findings and gave some useful feedbacks and suggestions.

His work is normally to trace meteorological events and analyse data by using some domain methods. The system can trace and monitor meteorological events via images transferred from radars. During the discussion, he gave some suggestions about the

extension work for using data mining model. He listed some potential relationship between events and phenomenon. For example, weather situations are related with rainfall and temperature. They may indicate us to improve the input data quality for mining in future work. Besides, he mentioned that there are two important components during the analysis process: analysis result and the visualization. It is obvious that the results will mean nothing if they cannot be understood by domain experts or users.

He also suggested us to describe rules in a user friendly way rather than mathematic statistics. Then he showed us some models and applications in the radar monitoring system and the data collected from radar. In their system, users can easily find the details for every specific phenomenon. And it provided an open platform to allow users manipulating specific data and analysing data.

For our findings of that rules, he said it could be a good way to build a suitable interface to show how it works, then users can find the value of that rules easily. These comments indicate us that we should focuses on combining data mining models and data visualization together in the approach.

4.3 Implication for comparing different models

After the discussions, we have got some useful feedbacks from domain experts. Those feedbacks can be summarized as follows:

- Our findings cannot reach to the high accuracy by comparing the prediction or forecasting in hydrological model.
- Data mining is a good way to analyze hydrological data with its low data quality requirements. If we can improve the quality of dataset as for input, the mining model can generate more accurate results.
- It is important to visualize rules to improve the level of understanding which can prove the usefulness of rules

4.3.1 Traditional hydrological model

In the water domain, hydrological data are normally manipulated and interpreted by traditional hydrological models. These models are very powerful in stream forecasting, water quality monitoring, etc. However, most hydrological models have some limitations which can block their applications.

Firstly, Most of hydrological models ask for high level domain knowledge to generate the structure and analyse data. It means that only the domain experts can use this method to analyse data. However, from now on, more users prefer to use hydrological prediction service without domain knowledge. According to National Research Council (2006), there are many users groups requiring the hydrological prediction events such as farmers, schools, governments. In addition, most of users have less domain knowledge. If there is no domain support, they cannot analyse the data and get the result.

Then most of hydrological models require high quality data. Due to the particularity of water related data, the real water data ask for the integrity and veracity. The data also need to manipulate accuracy during the process.

Finally, Most of hydrological models have the specific region like different locations and topography. Also different climatic zone can influence the hydrological events.

4.3.2 Data mining approaches vs. hydrological model

In general, traditional hydrological model can generate more accurate results but with very high costs. For example, the hydrological model can give predictions about a specific location and phenomenon, e.g., water level in location A will increase by 10cms in the next 10 hours. However, these results are generated based on high quality input data. In addition, these models may require over 10 years' data for

calibration and training purposes. Such requirements will lead to very high cost on data preparation, and greatly limit the applications of the models. On the other hand, data mining methods have much lower requirements on data quality. Most data mining methods are purely data driven. Namely, the discovered information/knowledge is based on the data availability. Hence, even without long period and high quality data, we still can find some useful patterns from the data. Meanwhile, these patterns are easier to understand for general users without too much domain knowledge.

Also our findings in data mining model are shown by mining software which is the simple characters and statistics, thus this research can indicate combining data mining results and visualization together to describe rules and allow user to understand rules easily.

4.4 Rule Presentation

In order to facilitate users to understand discovered knowledge, data presentation becomes to an important issue for the success of the approach. Data presentation is a suitable way to transfer the obscure data information into perceptible information (Pittelkow & Wilson 2009). A good presentation can let data easy to understand and provide clear information from data to potential audience. Data presentation also can prune the redundant data and provide the friendly interface.

In this research, we focused on the presentation of discovered association rules and try to embed the presentation with the current interface of the South Esk Hydrological Sensor Web. Also that can indicate the combination model about the data mining methods and data visualization function.

4.4.1 Interface design

According to Chapter 2, data visualization can help users to understand information.

In this research, in order to describe the rules clearly, we built up a tree view structure to describe the relationships between locations, sensors and phenomenon types. As shown in Figure 4.2, in this structure, locations, sensor types and phenomenon types are shown clearly. So it can also help the users to understand the relationships between observed phenomenon types and locations.

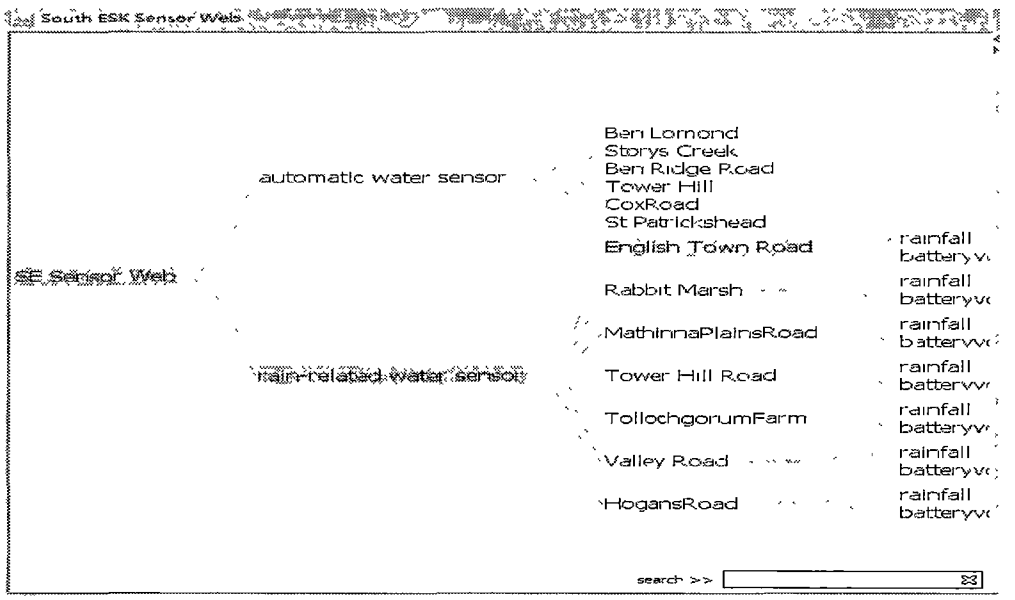


Figure 4-2: Tree View Structure of the Sensor Web

Comparing with the initial sensor web (See Figure 1-2), this structure is more clearly to show the relationship between locations and phenomenon. Also people can see details information by clicking every node of this tree view interface. To connect with the initial sensor web, we created the channel to show the details of locations and status in related phenomenon as shown in Figure 4.3. The red-signed location is signed by clicking related location and the status belongs to selected phenomenon (e.g. humidity). Also that can help people to understand the situation information in specific phenomenon and location.

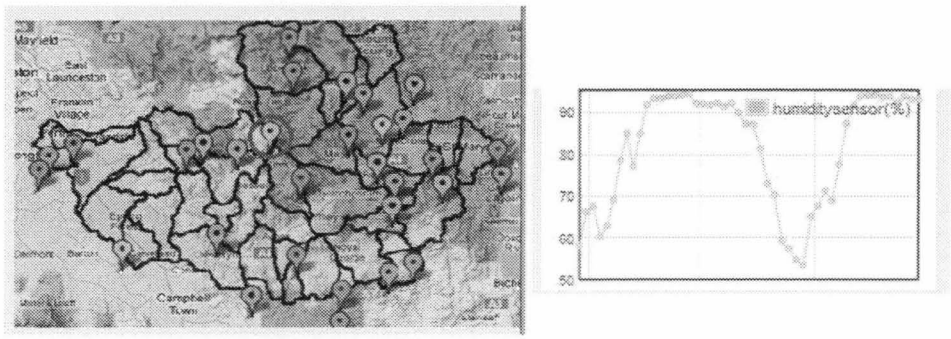


Figure 4-3: Related Location and Status in Specific Phenomenon

To present the rules, we provide a set of selection boxes to allow users to choose their preferred rules followed by Figure 4.4. Based on the "maximum value time" of a rainfall event, there are four selection boxes to allow users to select locations, phenomenon types, value types and time gaps. When a user sets three attributes from any three boxes, the value of the other box can be generated based on the association rules in the rule base. For example, if the rainfall gets the maximum value at 2:00 am, and we select the location as BenLomond, phenomenon type as Humidity and the value type as Maximum Value, then the rules will be invoked and it can be found that the maximum value time will at 21:00 pm in previous day. In addition, the application will give the confidence values of the association rule as well. This application of data presentation is the extension of the rules from data mining approaches which can visualize the rules and make rules easy to understand. In addition, it also provides a friendly interface to let user can manipulate rules and know how rules work in sensor data.

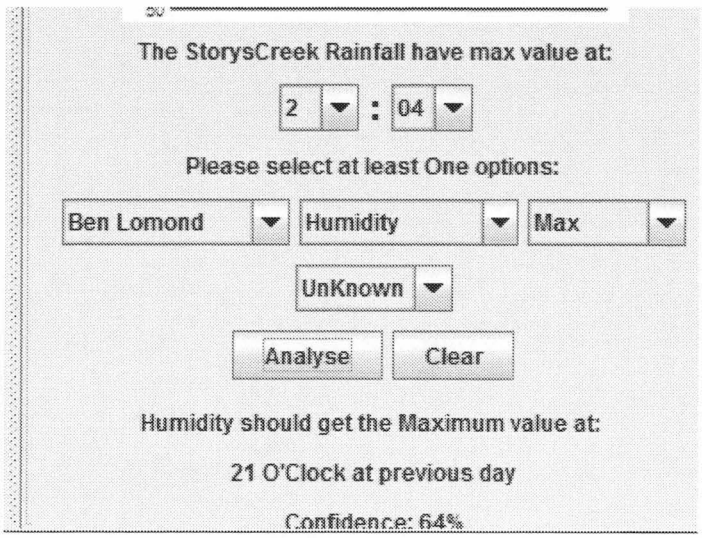


Figure 4-4: Application of Rules from Data Mining

Figure 4.5 is the integrated interface for visualize data mining rules and the conception of the hydrological sensor web. It has functions that indicate people to see and search related locations and phenomenon in SE sensor web. Also that allows users to manipulate the interface to understand rules.

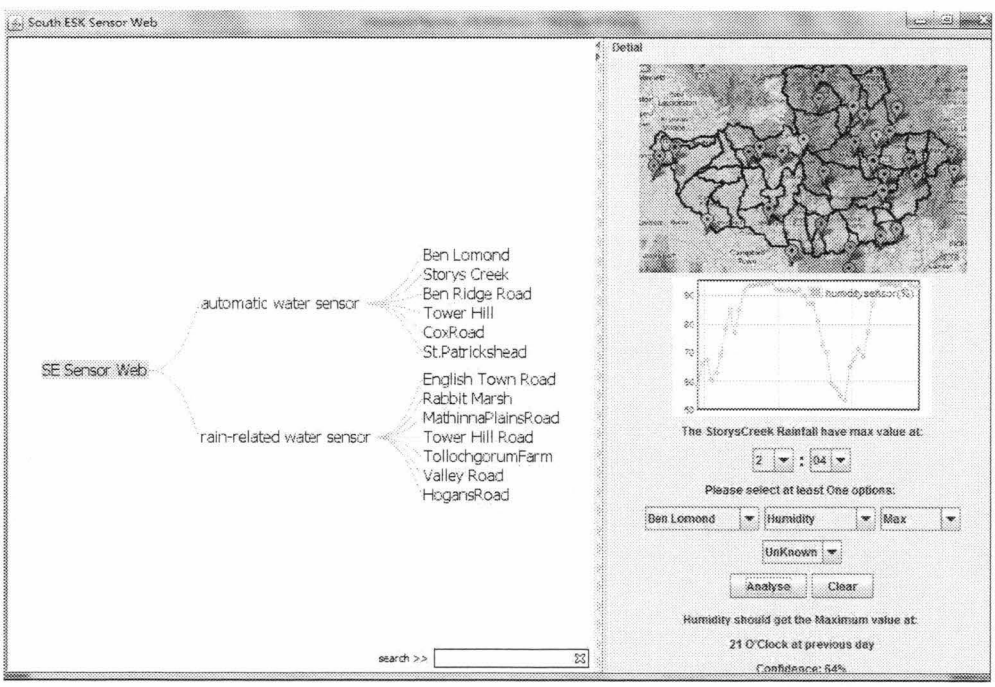


Figure 4-5: The Integrated Interface in Sensor Web

4.4.2 Implementation

This visualization interface was developed using the Java language and XML formatted file as for related database. Java is used to develop the interface and parse XML formatted file to get the data from XML-based database. XML file is used to store the sensor information and rules details.

To show the tree view of the sensor web, we prefer using the tree visualization toolkit for the frame and invoke data from XML file to build the structure. Then we built the connection between location map and status phenomenon with the tree structure. After that, we parse related rules which stored in XML file and show the descriptions in that interface.

To select and generate rules, we designed a simple algorithm to invoke rules based on users' selection in the interface. Figure 4.6 shows the rules selection. Users can select their interested target via selecting different items in that interface (e.g. select location, item and leave value type and time unknown). Then server part can receive the action for users and match related rules in database. If the rules match perfectly, server will obtain related rules and describe rules in that interface by prediction function and give the confidence to show the probability.

To match related rules, we design a simple algorithm to search and invoke. We set up a rules filter to divide different rules in different groups which conclude location-base, item-base, valuetype-base and valuetime-base. For example, rule {(min_gap:[(-5)-(-2)]) => (item:airtemperature) } belongs to item-base, valuetype-base and valuetime-base. If the system receive the related information which concludes one of them from users, this rule will be invoked and displayed in the related interface.

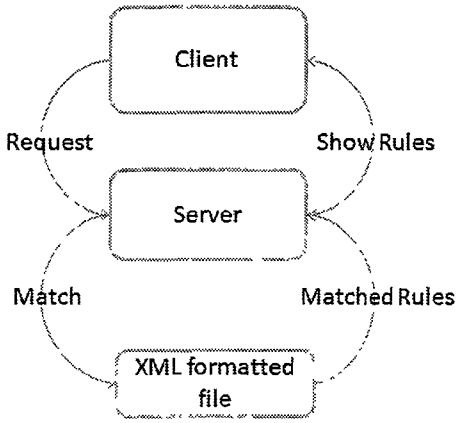


Figure 4-6: Rules Selection Process Flow

Unfortunately, we currently do not have a user group to evaluate the interface in the approach. However, we had built a combination model to show the result of analysis, and this interface provides a suitable way to describe our findings. In addition, comparing with initial hydrological web, this interface allows users to manipulate related data and events and they can get the good feedback via select related items. Such interactions can filter out useless rules based on users' preferences.

5. Conclusion & Further Work

This chapter shows the goals of this research and the review of the contributions in related work. It displays the future work for enhancement of the research as well.

5.1 Conclusion

Event prediction plays an important role in the water domain. In this research, we focused on using data mining approaches to achieve knowledge discovery. We select the association rules to analyse the relationships and patterns among different sensor observations on the Hydrological Sensor Web. Compare with traditional hydrological models, data mining methods have lower requirements on data quality and domain expertise. In this paper, we described how to collect and prepare data from the Sensor Web, and use a specific data mining workbench to achieve knowledge discovery by using association rules. In addition, we also introduced the presentation of discovered rules.

As mentioned in chapter 2 and chapter 3, this research focuses on the modeling work and finding relationships between hydrological data and events. The following data mining model can provides the flexible way to find the potential relationship in South-Eek sensor web. The generated rules can make simple prediction in specific hydrological phenomenon. Although those sorts of rules are not as stronger as traditional hydrological model, data mining methods require lower data requirement and less domain knowledge. In addition, it is easier to improve data mining rules by considering the dataset or related algorithms.

Then in chapter 4, we discussed the application of the rules and provided the interface to visualize rules. We combine the data visualization and data mining model together to present rules and help users to understand related rules easily.

In conclusion, to compare the traditional hydrological model, based on the features of South-Esk sensor web, we provide the specific model to disposal hydrological data. To improve the application of sensor web and indicate the usefulness of hydrological data.

5.2 Contributions

The greatest contribution of this research work is to build the specific model to disposal sensor data and find relationships between hydrological data and events for South-Esk sensor web. This research use data mining methods to find relationship and data visualization technology to present the rules and enhance the usefulness in the sensor web. In addition, this research had built up the user friendly interface to visualize rules and South-Esk sensor web which can allow users to manipulate demo to understand rules and phenomenon easily.

5.3 Future Work

Based on the model we had built in this research, the future work of this research will be the investigation of the application other data mining in environment monitoring. In addition, we will improve the integration of data mining model and data visualization technology to present data more clearly. Also we will combine the monitory system and web technology to present sensor data more comfortable.

For the improvement of the data mining methods, we will pay more attention to selecting more data mining methods and build different dataset. The potential relationship purpose could between wind and water related events. We will take time to compare existing algorithms and select or improve algorithms to achieve more useful relationship or patterns.

For the data visualization, we will focus on developing user friendly interface on web

environment and allow users manipulate online. We will embed functional parts such as WEKA Explore or other data mining platforms to allow users using data mining methods to realize their interesting information.

We will improve the modeling work for sensor web data and set this kind of model into widely use such as in meteorological and hydrological domain.

6. Reference

A.K. JAIN, MNM, P.J. FLYNN 1999, 'Data Clustering: A Review', *ACM Computing Surveys*, vol. 31, no. 3, pp. 265-323.

Anthony, H & Vinny, C 2004, 'Route profiling: putting context to work', *Proceedings of the 2004 ACM symposium on Applied computing*, ACM, Nicosia, Cyprus.

Arawal, R. and Srikant, R., (1994), *"Fast algorithms for mining association rules in large databases"*, Proceedings of the 20th International Conference on Very Large Data Bases, SanFrancisco, pp 487-499.

Borman, S 2004, 'The Expectation Maximization Algorithm A short tutorial'.

Carlos, O 2006, 'Comparing association rules and decision trees for disease prediction', Proceedings of the international workshop on Healthcare information and knowledge management, ACM, Arlington, Virginia, USA.

Chong, L, Feng, W, Jun, S & Chang Wen, C 2009, 'Compressive data gathering for large-scale wireless sensor networks', *Proceedings of the 15th annual international conference on Mobile computing and networking*, ACM, Beijing, China.

Committee to Assess the National Weather Service Advanced Hydrologic Prediction Service Initiative, National Research Council, Toward a new advanced hydrological prediction service, National Academy of Science, 2006

Durso,F & Sethumadhavan,A., 2008, *Situation Awareness: Understanding Dynamic Environments,* Human Factors: The Journal of the Human Factors and Ergonomics Society, pp. 443-448.

Endsley, M.R., 1995a, Towards a theory of situation awareness in dynamic systems, Human Factors, Vol. 37, pp. 32-64.

Garsoffky, B., Schwan S., & Hesse, F. W. (2002). *Viewpoint dependency in the recognition of dynamic scenes*. Journal of Experimental Psy-chology: Learning, Memory and Cognition, 28, 1035–1050.

Hall M., 2009, The WEKA Data Mining Software: An update, 'SIGKDD Explorations', Vol 11, Issue 1, pp10-18

Han, J. (1998). "*Toward on-line analytical mining in large databases*," ACM Sigmod Record, 27(1) 97-107

Ian H, EF 2005, *Data Mining Practical Machine Learning Tools and Techniques*, Second Edition edn, MORGAN KAUFMANN.

Ikuhisa, M, Michihiko, M, Tsuneo, A & Noboru, B 2009, 'Sensing web: to globally share sensory data avoiding privacy invasion', *Proceedings of the 3rd International Universal Communication Symposium*, ACM, Tokyo, Japan.

Jiawei, H. and Micheline, K., 2006, Data Mining: Concepts and Techniques, *MORGAN KAUFMANN PUBLISHER*.

Kevin, N, Nithya, R, Mohamed Nabil Hajj, C, Laura, B, Sheela, N, Sadaf, Z, Eddie, K, Greg, P, Mark, H & Mani, S 2009, 'Sensor network data fault types', *ACM Trans. Sen. Netw.*, vol. 5, no. 3, pp. 1-29.

Keim,D., 2002, "Information Visualization and Visual Data Mining", IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 7, NO. 1, JANUARY-MARCH 2002, pp.100-107.

Kiyoharu, A, Datchakorn, T, Shinya, K & Toshihiko, Y 2004, 'Efficient retrieval of life log based on

context and content', *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, ACM, New York, New York, USA.

Klein, A & Lehner, W 2009, 'Representing Data Quality in Sensor Data Streaming Environments', *J. Data and Information Quality*, vol. 1, no. 2, pp. 1-28.

Liang, X. and Liang, Y., 2001, "Applications of data mining in hydrology". *Proceedings of the IEEE international Conference on Data Mining*, pp 617-620.

Linda, MZ & Hisashi, K 2002, 'A simplified EM algorithm for detection of CPM signals in a fading multipath channel', *Wirel. Netw.*, vol. 8, no. 6, pp. 649-658.

Linden,G & Hanks,S., 1997, "Interactive Assessment of User Preference Models: The Automated Travel Assistant", Proceeding of the Sixth International Conference, pp. 67-79.

Liu, Q., Bai, Q. and Terhorst, A., 2010, "Provenance-Aware Hydrological Sensor Web", *the Proceedings of Hydroinformatics Conference*, pp. 1307-1315, Tianjin, China.

Mark, H, Eibe, F, Geoffrey, H, Bernhard, P, Peter, R & Ian, HW 2009, 'The WEKA data mining software: an update', *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10-18.

Mi Li, GH, Bernhard Pfahringer 2004, 'Clustering Large Datasets Using Cobweb and K-Means in Tandem', *Lecture Notes in Computer Science*, vol. 3339, pp. 368-379.

Nachiketa, S, Jamie, C, Ramayya, K, George, D & Rema, P 2006, 'Incremental hierarchical clustering of text documents', *Proceedings of the 15th ACM international conference on Information and knowledge management*, ACM, Arlington, Virginia, USA.

Open Geospatial Consortium, 2007, "OGC Sensor Web Enablement: Overview and High Level Architecture". *Technical Report OGC 07-165*.

Peter, D, Deepak, G & Prashant, S 2005, 'TSAR: a two tier sensor storage architecture using interval

skip graphs', *Proceedings of the 3rd international conference on Embedded networked sensor systems*, ACM, San Diego, California, USA.

Pittelkow, Y E. and Wilson, S R., 2009, "Visualization of Gene Expression Data", *The Berkeley Electronic Press*.

Jeffrey, SW, 2004, *Data Mining: An Overview, Congress Research Service,*.

Stephanie Lindsey, CR, Krishna Sivalingam 2001, *Data Gathering in Sensor Networks using the Energy Delay Metric,* Los Angeles.

Soukup,T & Davidson,I., 2002, *"Visual Data Mining: Techniques and Tools for Data Visualization and Mining* ", John Wiley & Sons, Inc, first edition.

Su, F., C. Zhou, V. Lyne, Y. Du and W. Shi. (2004). "*A data-mining approach to determine the spatio-temporal relationship between environmental factors and fish distribution.*" Ecological Modeling 174.4 421-431.

Thomas, C and Fischer, G., 1996. *"Using agents to improve the usability and usefulness of the World-Wide Web"*, In Proceedings of the Fifth International Conference on User Modeling, 5–12.

Tapas Kanungo, DMM, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, Angela Y. Wu 2002, 'An Efficient k-Means Clustering Algorithm: Analysis and Implementation', *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 24, no. 7, pp. 881-892.

7. Appendix

A. Related materials of data mining model

The initial input dataset arff file named "test_data.arff" is in the "Data Mining Materials" folder of the included CD-ROM.

The advanced input dataset arff file named "test_data_advacned.arff" is in the "Data Mining Materials" folder of the included CD-ROM.

B. Data visualization materials

The demo of the visualization is named "demo.wmv" in the "Data Visualization Materials/Demo" folder of the included CD-ROM.

The source code and the XML data base of the demo in the "Data Visualization Materials/Source" folder of the included CD-ROM.