2.2

DATA WAREHOUSING AND ITS POTENTIAL USING IN WEATHER FORECAST

Xiaoguang Tan*
Institute of Urban Meteorology, CMA, Beijing, China

1. INTRODUCTION

The main function of most current forecaster's workbench is data access. They are usually an integrative graphics system to show the data and derived data, and give some analysis tools, and graphics manipulation tools.

In China we have three problems with such systems. 1. Although the system can provide many data (over 2GB, several thousand fields one day), forecasters only use little of them (less than 1%) in operational forecast. 2. Massive storage system is developing in China, we will get many history data in operational forecast, how to make our system to use these data? 3. Many kinds of special observation system are developing, such as AWS, profiler, GPS/Met, storm lightning locator, etc. They are very useful for public service, but we must have a flexible data manager to use them in operational forecast.

The most serious problem is the first one above. Data warehouse is a new generation of Decision Support System (DSS). It may provide a good solution for the problems.

*

2. SOLUTION FROM DATA WAREHOUSING

Data warehouse is suggested by W. H. Inmon in the book "Building the Data Warehouse"(1993). He gave the first definition of data warehouse -- "A warehouse is a **subject-oriented, integrated, time-variant** and **non-volatile** collection of data in support of management's decision making process". Initially data warehouse is used in commercial business to help manager's decision. In these years, data warehouse is progressively used in wider fields, Include many scientific fields. But why should we concern a tool that is originally developed for business? Why should we use data warehouse technology in weather forecast field?

In operational process, we get data through collection, decoding, quality control, deriving process and then by inserting, modifying, deleting data process we settle them into the databases. Such data are operational and the most detailed data. Usually it is slow to be analysed. Imagine a short-range forecaster that face 2GB (several thousand fields) real-time data, plus massive historical data, plus heterogeneous data from special observing, and he has only short time (several hours) to make forecast decision before forecast products are sent out, how will he analyse such data in short time? Usually he only uses little data that he is used to use. So we should improve our system to help forecaster use more data.

How does the system make forecasters use more data? The ordinary way is providing some deeply processed data to concentrate data information or building some automatic forecast systems to save time, such as physical parameter diagnosis, traditional statistics, NWP models, MOS, MOD, etc. But the current development of above tools is still not qualified enough to consumedly reduce forecaster's work.

So whatever objective tools are provided, we still need an interactive system to help forecasters to make the final decision.

* *Corresponding author address:* Xiaoguang Tan, Institute of Urban Meteorology, 55 Beiwaxili, Haidian, Beijing, China; e-mail: xgtan@ium.cn

In data warehouse, forecasters will get more helps by fully using forecast expert's knowledge in the system. If the system build forecaster's knowledge into database, help forecasters accumulate new knowledge, provide a proper interactive graphics system to show and use the knowledge, forecasters will analyse data more quickly and use more data in operational forecast.

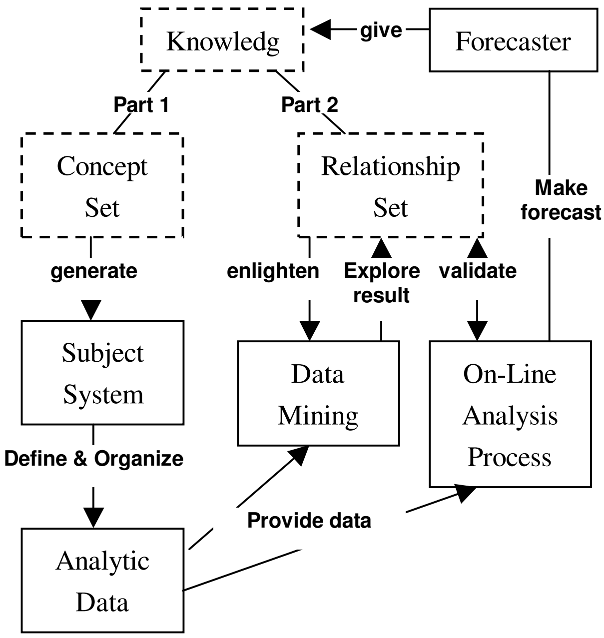

Figure 1
Solution from Data Warehouse

Figure 1 shows the data warehouse solution. The dash line box represents the module in forecaster's brain. Usually the forecaster's knowledge may be divided into a set of concepts and a set of relationship among the concepts, such as association rules, clusters, classes, etc.

Defining "**Subject**" is the object of analysis in forecast, e.g. concept in forecaster's knowledge.

Defining "**Analytic data**" is a real value corresponding to a subject. It is transformed from original operational data according to the definition of the subject.

Creating a subject system from the concept set of forecaster's knowledge, defining the data transformation to change operational data into analytic data for each subject in subject system, in operational process every day, we get real-time analytic data and store them in database. By this way, concept → subject → analytic data, the concept set of forecaster's knowledge is built into the database. Forecaster will analyses the analytic data that match the concepts in his brain.

Remember Inmon's definition, in data warehouse only analytic data are stored and they are subject-oriented.

The meaning of other terms used in Inmon's definition include: "integrated", means data warehouse should extract data from all data sources which contains useful data to the DSS mission; "time-variant", means data warehouse should contain many historical data that enables analysis of weather evolution and rules over time; "non-volatile", means the data in data warehouse are only for analysis, no more low level process (insert, modify, delete, etc.). This is the left half of figure 1.

In the right half, there are two important analysis tools.

One is Data Mining (**DM**). It is an exploring tool. It automatically explores the relationships among the subjects (i.e. concepts of forecasters) from the analytic data set. The exploring process is usually defined by forecasters, started may be automatically or by forecasters, and under the enlightenment of forecaster's knowledge. The result relationship from DM will be reinforced into forecaster's knowledge to help forecasters quickly accumulate their knowledge.

Another is On-Line Analysis Process (**OLAP**). It is a interactive validating tool. Forecasters use it viewing data, validating relationships (include forecaster's guess and results from DM) and then make forecast decisions. Its kernel technology is **multi-dimensional analysis**. It will be the main workbench of forecasters.

By these two tools, the relationship set in forecaster's knowledge is used and reinforced continuously.

In current time, a complete data warehouse solution not only include data warehouse defined by W.H.Inmon but also include DM and OLAP. Now we redefine data warehouse: Analytic data storage plus analysis tools DM and OLAP is data warehouse. To avoid confusion, we will call the data warehouse defined by Inmon "Data Storage" in following sections.

In following sections, a design of Data Storage, DM and OLAP will be discussed.

3. DATA STORAGE

Data Storage is the fundament of data warehouse. It changes operational data into analytic data, managing these data and provides data access tools or other support functions to applications especially DM and OLAP. Figure 2 is the flow chat. The main key technologies are as follows.

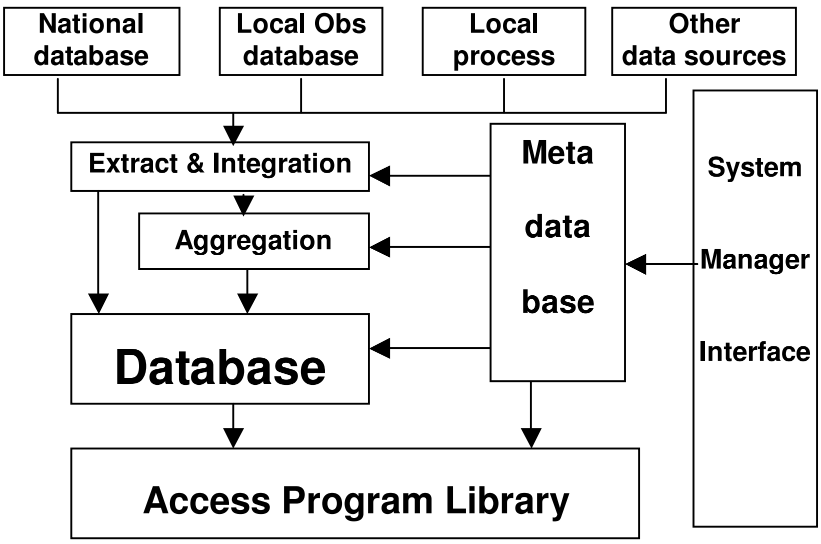

Figure 2
Flow chat of data store system

3.1 Subject System

The starting point of designing for data warehouse is collect forecaster's knowledge to build the subject system.

Subject system is the most important in data warehouse. In data storage, subject system is the basis of changing operational data to analytic data, basis of logic data model to store data. In OLAP, it is the basis of dimension set of multi-dimensional analysis. In DM, it is the basis of item set for association rules exploring.

According to forecaster's knowledge, in current there are 6 types subject in our subject system:

1. The **target element** of forecast must be included in subject system, e.g. rainfall, temperature, humidity, wind at Beijing.

2. **Index data**. Some Index data are useful, usually they are the flags or indicators of weather in future, e.g. surface wind direction of Wutaishan, 500hPa height change in west Siberia, etc.

3. **Statistical data**. To analyse statistic characters in a region or a time period, subjects like average of 5 stations, maximum temperature gradient in some region, minimum relative humidity in past 5 days, etc., should be included in the subject system.

4. **Transformed data**. Coefficients from orthogonal transformations, such as wavelet transformation, experience orthogonal function (EOF) transformation. And filtered data generated by lowpass, highpass, etc.

5. **Weather system**. Synoptic knowledge is essential in forecaster's knowledge, even if in the analysis of model output. So the weather system should be included in subject system. The example of weather system include: high, low, big gradient area, saddle area for scale element and convergence center, divergence

center, share line, jet stream, jet center for vector element. They should be easier to be identified by computer.

6. **Conceptual models**. They are compositive subjects. They are composed of above 5 types of subjects to form conceptual models.

Above subject types are only from forecast knowledge. It can be extended to include more area. For example, forecast service area, system management area, etc.

In our design, the subject system is a tree-like structure. Leaf node is corresponding to the most detailed concept. Starting from leaf nodes, subjects will be generalized to wider and wider. The root is corresponding to the widest concept. Correspondingly the name of subject at different tree level may use comparative words such as "more", "most", "less", "least", "better", "best", "bigger", "biggest", etc. Which tree level the subject is, may decided by forecasters when forecaster input the subject. The higher level subjects may also be produced from lower level subjects automatically, as in the following section. In following sections, we will see such structure is necessary for DM and OLAP.

3.2 Changing The Data into Analytic Data

In the Figure 2, data storage extracts data from data sources, then implements a integration process to check data quality, delete duplicate data, make data uniform, i.e. same data have same meaning, unit, measurement, accuracy, etc.

Then it will change data into analytic data by **transforming** and **aggregating**. The first step is transforming original data into the analytic data corresponding to one of the 6 type subjects. The second step is making aggregating on the transformed data.

Aggregating is the common method in ordinary data warehouse. It calculates count, sum, average, and find the maximum and minimum values of the transformed data in a spatial region or a time period. In our design, we define the networks in space with grid interval 1, 2, 5, 10, 20 degrees in longitude and latitude, calculate aggregation data in each network box. Divide time series into periods of every 1, 5, 10, 30 days, calculate aggregation data in each time period. Usually subject tree is formed from smaller (leaf, called **detailed granularity**) to bigger (root, called **rough granularity**) grid interval or time period.

Except the method above, we still define additional methods special for weather forecast to form subject tree for some subject types. They will weave with results of aggregating to get more granularity in subject tree.

For the 6 types of subject, data will be transformed as following method.

1. **Forecast target element**. Calculate the analytic data from original data, as the definition of the forecast target data, using fuzzy membership function and define several membership thresholds to get the subject tree, highest is leaf and lowest is root.

2. **Index data**. The calculation is same as above, but according to the definition of the index data.

3. **Statistical data**. The analytic data is the results of statistics. Subject tree is formed same as aggregation, no additional method.

4. **Transformed data**. Make wavelet transformation and EOF transformation to every data fields and take the first 10 coefficients as a vector to be the analytic data. The subject tree is formed from keeping all of 10 coefficients (leaf) to only keeping first coefficient (root). For filtered data, giving several levels of filter parameters, subject tree is formed from the sharpest (leaf) to the smoothest (root) field.

5. **Weather system**. Computer identify the weather system according to their definition, and take the attributes such as location, area, aspect, direction and length of long axis, intensity, longitude and latitude coordinate of vertexes of

characteristic line, etc. to form a vector as the analytic data. Weather system data just like the **Spatial Entities** of Geophysical Information, so many results from **Special Data Warehouse** can be used in their aggregating, storing and using. The subject tree is formed according to the match level from some match function that matches the vector components between real data and idea pattern. Define several thresholds, highest is leaf and lowest is root.

6. **Conceptual model**. They are composed of former 5 types. Giving an idea pattern of the conceptual model according to the forecaster's knowledge, the analytic data is the vector made up of attributes or values of its components. Note that, some complex weather system is also this type. For example, the cold front is composed of the weather system: wind shear line, temperature high gradient area, positive pressure change behind the shear line and negative pressure change before the shear line. It is the conceptual model of cold front. The subject tree is formed according to the match level, same as above.

3.3 System Management with Metadata

In Figure 2, the metadata is the kernel of storage management.

Metadata is data of data. In metadata, all data in data warehouse are completely described include data source, characters, quality, accuracy, transformation experience, location and access parameters, etc. It is a dictionary of data warehouse.

In data warehouse, there are still many system structure information included in the metadata base, include server IP or URL, access username, password, directories, definitions of data filename, descriptions of data format, descriptions of the algorithms, functions and parameters for data transformation, etc.

With this metadata base and providing a corresponding data access service function, data store system will be created. Later, metadata supports the maintenance and usage of data warehouse.

Subject system is also registered in this metadata base. The structure of part our metadata base shows in figure 3.

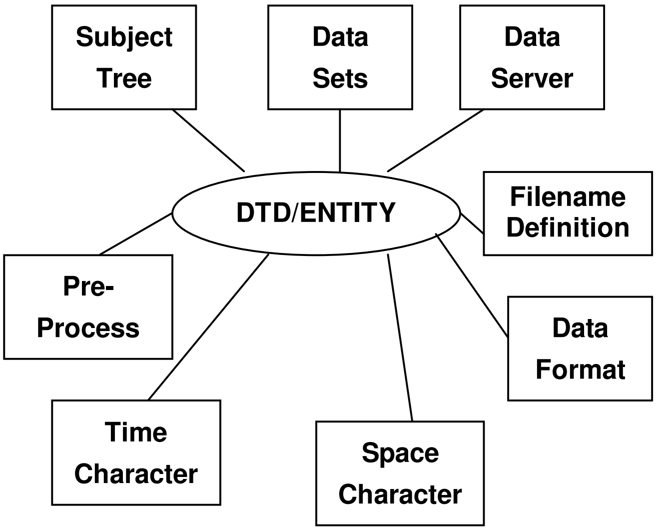

Figure 3
Structure of Metadata Base

Metadata are all in **XML**. The kernel is DTD. ENTITY is the member of DTD. The function of DTD/ENTITY is just like a pointer. Each ENTITY points each component of metadata base. Applications can access data starting from any component. For example, An application takes a data, it access **subject tree** first, there it find corresponding **data set**, then it can find the **server** and location of the data set, then it will find data by **filename definition**, then it read data from the file according to the **data format**, then it can use data according to the data **space character** and **time character**.

When system management applications run data ingest, it scans **data moving principle** in **data set** first, then it find the source **server** and original data location of the data set, then it will find data by **filename definition**, then it copy data file or read data according to the **data format** to put data to destination location, then it change original data into analytic data according to **pre-process** definition in subject tree.

In future, schema will be used instead of DTD.

By metadata, data management and maintenance become easier and flexible, historical data and heterogeneous data such as special observation data, even Internet data, will easier to be used by applications. If defining a standard schema, several data warehouses distributed in various provinces may be linked as one data warehouse through Internet.

3.4 Storing of Data

In data warehouse, there are two modes of data store: relational database and multi-dimensional database. In relational database, for convenience of OLAP's multi-dimensional analysis, should build dimension tables with data tables. There are star mode and snowflake mode.

Multi-dimensional database or **cube file** is a file that contains a multi-dimensional array. When it is used in a application, full of it will be read in memory, then build a multi-dimension array in memory. Then cube manipulations will be implemented in OLAP.

Usually, the data being used in OLAP are stored in cube files, the historical data are stored in relational database. In our design, all data are stored in cube files.

4. DATA MINING

Data Mining automatically explores the trends, models and association relationships that hide in database. It is from "Knowledge Discovery in Databases" (KDD) technology, but it only use mature methods in KDD. For weather forecast, it explores the relationships among the forecaster's concepts and speed up the accumulation of forecaster's knowledge on unused data.

Many methods used in data mining include multi-analysis, Bayes statistics, decision tree, classifying and clustering, neuro-network, time series analysis, etc. are very familiar by forecasters. In fact, forecasters have explored many years with them in traditional statistic forecast and model output interpretation.

Some unacquainted methods such as rough set, inheritance algorithms, especially the exploration of **association rules**, will open a new world to forecast researchers.

Association rule is a logic implication relationship. Let a item set I={ $i_1, i_2, \cdots\cdots i_m$ }, the item i_k (k=1,m) may be any weather event, such as relative humidity>70 at Beijing, rainfall>10 mm in Beijing urban area, a low pressure is at Mongolia, etc. Let D={ t_1, t_2, $\cdots\cdots$ t_n } is a historical sample set, any sample t_j (j=1,n) is a subset of I. If set A and B are all subset of I, and the intersection of A and B is empty. Then the implication relationship A→B, or statement "If A then B" is an association rule.

In statistical forecast, forecasters usually search correlation relationship. It is rigorous for most meteorological data. Let A and B are standardized random data, in correlation relationship we have B=r*A, r is correlation coefficient. We take a threshold at a point along the regression line to make A and B a (0,1) variant. If r is higher enough, we will get: if A=0 then B=0 and if A=1 then B=1. But in implication relationship, the relationship only need if A=1 then B=1, if A=0 B may be 0 or 1 any value. It is less rigorous than correlation relationship. It is more realistic for meteorological data. For example, let A is "There is a vortex near Beijing" and B is "It will be rainy in Beijing", correlation relationship requires if a vortex is near Beijing then Beijing will be rainy, if no vortex near Beijing then no rainy in Beijing. Obviously, because many weather system may cause rainy in Beijing not only vortex, the second half above is wrong. We will not get a good correlation. But implication relationship only requires the first half.

We will get a good implication. Most meteorological data are of this character. So **searching implication relationship is more rational than correlation relationship for meteorological data**.

In association rule exploring, giving proper item set is the key problem. In our design, item set is from subject system. Tree-like subject structure is necessary for DM, It makes us can explore association rules not only in same tree level but also across tree levels. By this way, if we do not find good rules in lower level (detailed data), we still try to explore rules across levels or in higher level. It is easier to find good rules.

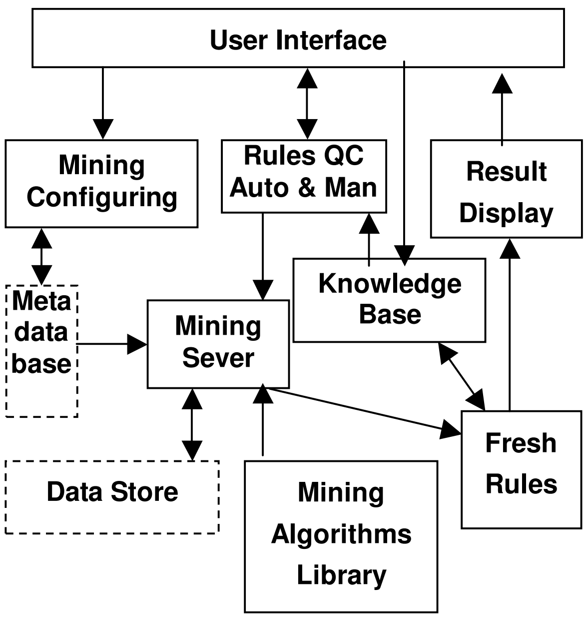

Figure 4
Flow chat of DM

Figure 4 is the flow chat of data mining system. In the flow chat, one important character is that there is a knowledge base in DM. It contains the relationship set from forecaster's knowledge and the quality controlled data mining results. With the subject system and this knowledge base, the forecaster's knowledge is completely built in the data warehouse.

Another important character is the knowledge base quality control system. There are three ways will be used in QC, forecasters check the rules with physical considerations, computer checks the rules with logical consistency, such as "no conflict", and computer check the Support and Confidence of association rules with historical data.

The interface in Figure 4 is a part of the interface of entire data warehouse. By this interface forecaster configure and enlighten the mining, control the quality of knowledge base, inputs his knowledge into knowledge base and get the results from DM.

5. OLAP

In data warehouse, Data Storage and Data Mining regularly run in background. The interactive workbench of forecasters is OLAP.

OLAP makes forecasters view data from multi aspects and analysis them. The kernel of OLAP is **Multi-Dimensional Analysis (MDA)**. The term "**Dimension**" in OLAP may be any aspect to view data. In our design **dimensions may be spatial, time or any subjects in subject tree.** OLAP is different from traditional meteorological graphics system. It view data in much more aspects, not only in a fixed frame of spatial and time dimensions. Because **the main aim of OLAP is view data to find the relationships or patterns in data, rather than just view data value or data pattern**. So it provides much wider scope than quondam graphics systems.

In Business field, usually managers view data with tables. But in weather forecast field, forecasters usually view data with graphics. Multi-dimensional data in OLAP form a data **cube**. Usually the number of dimension is more than three, we can not directly see the cube data. So there are some manipulations: **Slice**, **Dice** and **Pivoting** to reduce dimensions so as to see the data with graphics.

Slice is taking a 2 dimensions subset from the cube to display.

Dice is taking a 3 dimensions subset from the cube to display.

Pivoting is changing the dimensions in Slice and Dice.

The analytic data corresponding to lower level of subject tree are more detailed, corresponding to higher level of subject tree are more roughness. Two manipulations in OLAP make forecasters view data roughly or detailedly. “**Drill down**” is viewing data from higher level to lower level. “**Roll up**” is from lower level to higher level.

For weather forecast, only above manipulations is not enough. For finding more relationships, there are other 3 analysis tools based on multi-dimensional analysis: **Compare Analysis (CA), Multi-Analysis (MA) and Analog Analysis (AA)** in our design. The main functions of these four analysis tools are as follows.

5.1 Multi-dimensional Analysis

In multi-dimensional analysis, “**Dimension selection**” is the most important function. The manipulation “Slice”, “Dice” and “Pivoting” are all realized by this function.

The dimensions to be selected include natural spatial and time dimensions, and all subjects in subject tree.

We divided dimensions into 3 kinds. One is **frame dimension**, include spatial, time, the coefficient of subject type 4. It makes forecaster view data in spatial-time frame or orthogonal base frame from orthogonal transformation.

The second is **subject dimension**, include all subjects in subject tree. They are the content displayed in above frame.

The third is **element dimension**, include all of original data in database, such as height, temperature, wind, etc. They are also the content displayed in above frame. It is used for showing traditional graphics compatible with traditional meteorological graphics system.

OLAP shows selection boxes for these three dimensions.

The order of selected dimensions is important in following graphics display principles.

1. If forecaster only select two dimensions, then the first dimension must be a frame dimension and it is X axis and the second is Y axis, and show the curve of Y by X. Y must not be a frame dimension.

2. If forecaster select three dimensions, then the first dimension is X axis and the second is Y axis, and show the contour of the third dimension. The first two must be and third dimension must not be a frame dimension.

3. If forecaster select dimension number is more than three and the dimensions over three are all frame dimensions, then show a matrix of contour maps that are produced from the first three dimensions. Then arrange contour maps by the order of 4^{th}, 5^{th}, 6^{th}…. Also the first two must be and third dimension must not be a frame dimension. If there are some dimensions over three are subject or element dimensions, they will be overlapped in maps.

4. Forecaster can select some frame dimensions to show its scanning ruler. In scanning ruler, forecasters can move ruler mark by mouse. The dimension value will be changed easily. The ruler of space and time are always displayed in system interface.

With element dimensions, almost all of the traditional graphics may be displayed in OLAP by “Dimension selection”.

For example, if forecaster selects dimensions are in the order of longitude, latitude and height, it will show the contour of height and it is easy to change level and time by level ruler and time ruler; If selected dimensions are in the order of latitude, level, temperature, it will show the cross section along a longitude and it is easy to change longitude and time by the rulers.

Many unused graphics may also be displayed by “Dimension selection”. For example, if selected dimensions are in the order of time, latitude and height, it will show the latitude change by time for each height (e.g. 588).

5.2 Multi-Analysis

If all of selected dimensions, include the graphics frame dimensions are all subject dimensions or element dimensions, then enter the Multi-Analysis mode. In this case, the spatial dimensions for each subject or element are all taking fixed value.

We call the selected dimension "**factor**". Different factor may have different fixed value of spatial dimensions. Some of subjects already have the fixed value of spatial dimensions in their definition, for example rainfall in Beijing. If a factor has not fixed value of spatial dimensions, forecaster will appoint them by scanning ruler.

System shows the scattered map that the first factor is X axis, the second one is Y axis, within the map fill the value of the third factor (or its space location or time, specified by forecasters) at every point on time. Other factor will be used as above principle 3 and 4.

Forecasters can change the time period of MA.

5.3 Compare Analysis

Compare analysis is a multi-dimensional analysis with multi-data. The frame dimensions should be same for all of compared data. We can select compared data in subject tree or original data. If we select several data in subject tree, they will be overlapped in curve or contour maps. If we select several points in time, the graphics of same subjects at various time points will be overlapped in curve or contour map.

Up to now, Compare Analysis is just like MDA, and it can use all of functions of MDA. But it has some additional functions as follows.

All of compared data enter a **compare queue**. Forecaster can add or delete data in compare queue.

One data in the queue is **base data**. Other data in the queue will compare with it. Forecaster can appoint the base data.

When display compared data, forecasters can select:

1. To calculate the average and variance of the data in queue and display them in highlight.
2. To calculate the difference between each data in queue and the base data and display them in highlight.
3. To transform the data in the queue, such as normalization, filtering, orthogonal transformations, then compare the results.
4. To calculate count, frequency, average and variance of the value befell in specified interval of data in queue.

5.4 Analog Analysis

Analog analysis is a special compare analysis. Its compare queue is created by selecting the historical data that analog index with current data is more than a threshold.

Forecasters can choose the analog index, threshold, time period and spatial range.

After the compare queue is created, system will enter Compare analysis automatically and the current data will be the base data.

The aim of OLAP is similar to Data Mining. Data Mining finds the relationships among the subjects automatically, and OLAP do same things by human. So the same interface is used in Data Mining and OLAP. In OLAP, forecaster view data and get new knowledge, then put them in the knowledge base of DM. The results from DM will be displayed in the interface to give hints to forecaster. In these years, a new technology On-Line Analysis Mining (OLAM) has developed. It combines OLAP and DM.

6. DISCUSSION OF USAGE

In above sections, we discussed the main components of data warehouse separately. They should be integrated in data warehouse as in Figure 1. But how do forecasters use it?

One of running prospect is:

In background, data is extracted from various data sources (extract and integrating data) driven by time or event. Then system changes original data into analytic data (transform and aggregation) according to the definitions of subjects. Then analytic data are stored in Data Storage system. DM explores relationships in database of Data Storage driven by time or event according to the configurations from forecasters. The knowledge base of DM enlightens the explore process and saves the results.

In forecaster side, initially knowledge is entered the system by building subject system and inputting relationships into knowledge base. In regular work, he may add new concepts into subject system and new relationships into knowledge base at anytime when them occur in his brain or from summary after his case studies or other researches.

With interactive interface, forecaster checks or changes configurations of system components, controls the quality of relationships in knowledge base, validates the relationships with OLAP tools, do some case study or other research with OLAP and DM functions.

In operational forecast, forecaster views data in traditional maps or compare, analog analysis graphics of OLAP. He will get hints from DM automatically. He may designate some subjects and a proper data range, and send a temporal explore task to DM, then view the results in OLAP to validate some relationships needed in current forecast mission.

From above running prospect, we may look data warehouse as a new kind of Artificial Intelligence (AI) system that combines database and meteorological graphics technology. It helps forecasters accumulate, manage and use their knowledge in operational forecast. It is a new generation of DSS.

It is different from traditional knowledge base system. The bottle neck of traditional knowledge base system is knowledge acquirement. In data warehouse, forecaster may put their concept set into subject system firstly. It is easier than to put whole knowledge. Furthermore, he can input his new concepts at anytime later he remembers them. Then he can put some relationships that he remembered into knowledge base at anytime he remembers them and wait and get relationships from DM automatically. By this way it moderates the bottle neck problem.

Obviously data warehouse will not become whole system of forecaster's workbench, because operational forecast mission is very complex. But it is a system to help forecasters accumulate, manage and use their knowledge. It should be the kernel technology in next generation of forecaster's workbench.

REFERENCE

Aberer, K, Hemm, K, 1996, A Methodology for Building a Data Warehouse in a Scientific Environment, First IFCIS International Conference on Cooperative Information System, Brussels, Belgium

Ayyer, K, Data Warehousing and Knowledge Management, http://www.ischool.utexas.edu/~krishnan/i385q/DW%20and%20KM.htm#

Gorawski, M, Malczok, R, 2004, Distributed Spatial Data Warehouse Indexed with Virtual Memory Aggregation Tree, Proceedings of the Second Workshop on Spatio-Temporal Database Management (STDBM'04), Toronto, Canada, M.Gorawski@polsl.pl, R.Malczok@polsl.pl

Inmon, W.H.,1993, Building the Data Warwhouse , John Wiley & Sons.Inc

Nestorov, S, etc. Ad-Hoc Association-Rule Mining within the Data Warehouse, evtimov@cs.uchicago.edu