PHYSICOCHEMICAL AND THERMOPHYSICAL DATABASE DAFIT IN THE PROSPECT OF JAVA AND DATA WAREHOUSE

ANDRZEJ BYLICKI AND MARCIN GORAWSKI

Institute of Coal Chemistry,
Sowinskiego 5, 44–100 Gliwice, Poland
andrzej@bird.karboch.gliwice.pl

(Received 21 October 1999; revised manuscript received 28 December 1999)

Abstract: In this paper, we are discussing the meaning of DAFIT database for knowledge on physicochemical properties, methods of database access and graphical presentation. The paper presents the state of development of DAFIT database and concentrates on the application of the Java and DW technologies.

Keywords: physicochemical and thermophysical properties, database, data warehouse, OLAP, data aggregation

1. Introduction

The knowledge of the physicochemical properties of chemical substances and their mixtures constitutes an important part of physical chemistry and has an essential meaning for other branches of chemistry and related disciplines such as pharmacology, biochemistry, technology and chemical engineering. This knowledge based so far almost exclusively on the experimental investigation describes the physicochemical properties of substances only in a very fragmentary manner.

In comparison to the very large extent of this problem, the experimental data presented scientific literature in the world give only a very fragmentary description of the physicochemical properties of substances. For the multicomponent systems they are practically not available, especially for organic substances.

In the Physicochemical and Thermophysical Database DAFIT built in the Institute of Coal Chemistry of Polish Academy of Sciences (PAS) in co–operation with the Institute of Physical Chemistry PAS and the Thermodynamic Research Centre, Texas University, USA, the data collection comprises nearly 100 of property data for about 100.000 of substances.

The physicochemical properties of complex systems, such as function of state parameters, temperature, pressure and composition which constitute a multidimensional space practically can't be expressed by the set of experimental data.

Therefore in order to find a solution to this problem and increase the knowledge of the physicochemical properties of multicomponent mixtures of chemical substances the computation methods of correlation and prediction based on theoretical models were used. Selected critically evaluated experimental data were applied as reference data for testing the range of application of prediction methods and for determination of the value of parameters of correlation and prediction equations.

To illustrate of the magnitude of the problem of physicochemical properties, data of chemical substances and the necessity of processing and management of very large data sets we may present the number of binary systems formed by the most important organic substances, which is about 1010. For determination of 40 physicochemical properties of these substances as a function of state parameters, 1014 data are required.

The number of ternary systems is about 1015 and increases in a similar manner for each additional component of the mixture, which justifies the application of Data Warehouse technology. In order to make available the data collection in the DAFIT database in the INTERNET, the Oracle platform: Oracle Web Server and Developer/2000 were used. An integral part of the WWW service of DAFIT data base is a module realised with the application of Java language.

The paper presents the state of development of DAFIT database and concentrates on the application of the Java and DW technologies.

2. Database DAFIT platform

Applications including data editing programmes (insert, modify, delete), data browse programmes and data searching programmes were created using the application generator Oracle Developer/2000. Taking into consideration the physicochemical data search specificity a special group of data searching applications was built. There are three main programme classes distinguished:

- Simple programmes operating on single database tables
- Complex data presentation programmes operating on multiple tables and logical tables connections
- Dedicated searching programmes — allow a user to specify complex searching conditions

3. DAFIT database in prospect of Java

Oracle WebServer — HTTP server integrated with Oracle 7 server, allowing static and dynamic html documents creation using data from database– was used in order to make DAFIT database available on the Internet.

The graphic module is an integrated part of DAFIT WWW service. It was created using Java language. It is activated from a dynamic www page. Its main job is pure substance experiment results presentation — measurement points, their correlation — using mathematical equation models.

Equation models describing pure substance properties are presented as graphs with possibility of:

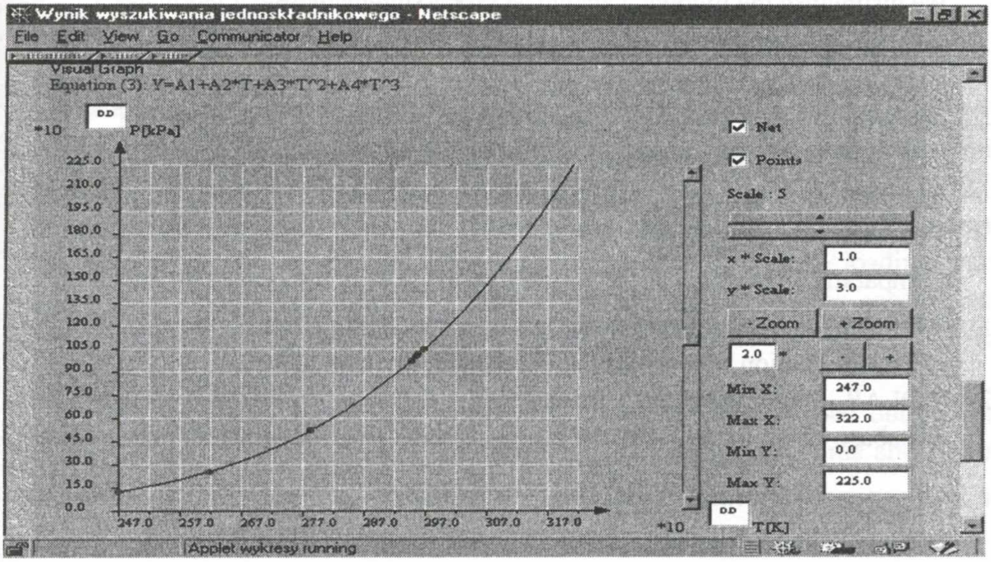
— reading co–ordinates of selected points;
— magnifying and scaling marked area;
— changing axis scale;
— showing grid and measurement points.

Java graphic presentation module shows all experiment data in the form of an applet. Such a solution was chosen because Java's features expand www pages interactivity in a significant way (in comparison with CGI) and allow to enrich their multimedial aspect. Also, applet that works directly on a user machine eliminates loss of time caused by data transmission (needed in the case of programmes working on the server) — applet is not limited by the net capacity in any way.

Access to the results of an experiment from the applet is realised from a dynamic www page that takes experiment's data directly from a database server. The information needed for graphical presentation (equation model, parameter values and measurements) is passed from a www page to the applet as its parameters.

The only condition that makes it possible to use the graphic module of the DAFIT database web site is to have a Java compatible web browser. Such browsers include standard net and graphic libraries.

The combination of www dynamic pages possibilities and Java merits allows to create WWW web sites that not only publish data but also make data interpretation and user interaction possible in a brand new way not demanding time–consuming data transmissions.



Wynik wyszukiwania jednoskładnikowego - Netscape

Visual Graph
Equation (3): Y=A1+A2*T+A3*T^2+A4*T^3

☑ Net
☑ Points
Scale : 5
x * Scale: 1.0
y * Scale: 3.0
2.0 *
Min X: 247.0
Max X: 322.0
Min Y: 0.0
Max Y: 225.0

Applet wykresy running

Figure 1. Graphical experiment results and their correlation presentation

4. DAFIT database in prospect of Data Warehouse

Taking into consideration DAFIT database specificity — a dominant operation is data searching. Data set supplementation and expansion follows the cycle of appearing of new publications describing substances and experiments and as a result of including new physicochemical properties. Modification and removing operation are very rare.

The goal of a user searching the database is, as a rule, to reach information about experiments carried out on a specific substance or group of substances that meet some conditions. Search criteria, depending on user needs, can concern also information source, publicating authors and edition years, changes of the ranges of experiment parameters [1]. Range of specified results conditions is dynamical and can undergo changes in work progress (decreasing or increasing the search area). Most frequent operations among search operations are searching according to substances' names synonyms, chemical formula, chemical abstract, substance properties, kind of property, authors' names, publication titles, edition years, parameters range. For the user it is most important to be able to receive processed data (using critical assessment and correlation) with high credibility degree as recommended data. The fast growth in data volume in multicomponent systems justifies the necessity to use Data Warehouse technology for DAFIT database. This technology is analysed in relation to on–line analytical processing for DAFIT database.

4.1 On–Line Analytical Processing

It is becoming necessary to use **On–Line Analytical Processing** and **Data Mining** technology, for more effective access to DAFIT database.

We distinguish the following analytical processing methods:

— **Multidimensional On–Line Analytical Processing (MOLAP)** — based on multidimensional database server;

— **Relational On–Line Analytical Processing (ROLAP)** — based on relational database server with star or snowflake data schemas;

— **Hybrid On–Line Analytical Processing (HOLAP)** — based on hybrid database, relational–matrix database or on several different database types.

Comparative analysis leads to the conclusion that, in the case of architecture:

1. MOLAP — aggregated data are pre–calculated and stored in database. Database construction allows data to be viewed multidimensionally;
2. ROLAP — aggregated data are stored in a database or calculated by request. Database construction facilitates searching and calculating data for presentation.

4.2 Comparative analysis: MOLAP vs ROLAP

Data in database have three features: depth, breadth and atomicity.

Data depth is related to the level of data aggregation. Data warehouse should allow OLAP analysis from the lowest data level (atomic) to the highest–aggregated data–level.

Data breadth applies to data access with reference to the number of dimensions and attributes which can be analysed by the user.

Data atomicity refers to data granulate (number of data appearances)

From the presented graph (Figure 2) it follows, that for databases with:

— limited number of dimensions (data breadth);

— limited data aggregation (data depth);

MOLAP is a sufficient solution. As data depth and breadth grow it is necessary to migrate towards ROLAP model.

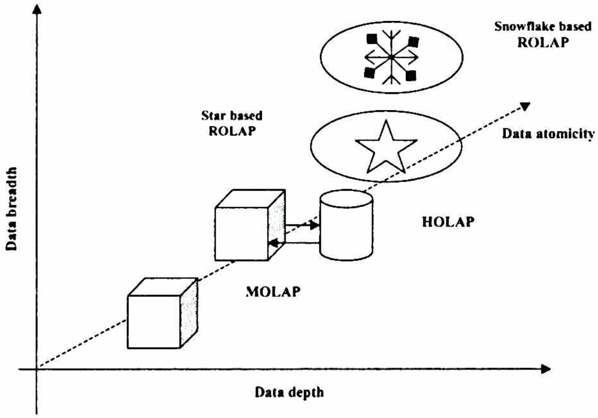


Figure 2. MOLAP and ROLAP architecture scalability

Data extraction and presentation process can be divided into several stages:

— transactional database data extraction;

— pre–calculation, data aggregation;

— search index creation;

— data presentation.

Depending on the architecture most of these operations are made initially in batch process or by request during data searching.

If data is initially processed, calculated and prepared for presentation, the system has better searching efficiency and response time. Periodical data loading batch processing is time–consuming and demands full computer resources engagement. Such systems are called high degree of aggregation (85–100%) systems — for example typical MOLAP databases.

In low degree of aggregation systems (0–15%), the system load during data loading is low, but data searching and presentation demand more time and system resources.

As degree of aggregation grows it results in:

— batch time processing and processor requirements growth during data loading;

— decreasing query response time and processor requirements during data presentation.

Processor requirements graph presented below is typical for about 10 dimensions (Figure 3).

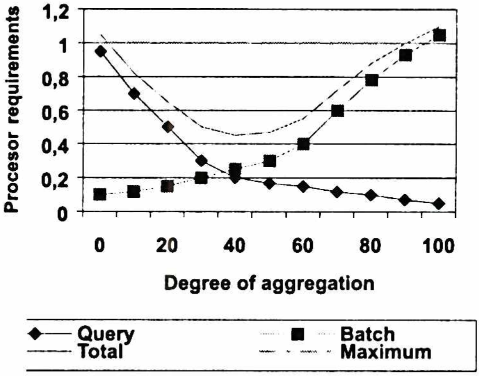


Figure 3. Processor requirements vs degree of aggregation

Processor requirements curves analysis for different degrees of aggregation, different data warehouses (number of dimensions, amount of atomic data, aggregated data volatility etc.) leads to the following conclusions:

— ROLAP systems can be implemented at any degree of data aggregation;

— MOLAP systems require high degree of data aggregation;

— ROLAP systems are better equipped to deal with fluctuating query–demand profiles;

— ROLAP systems are better equipped to handle large numbers of dimensions and can scale to large volumes of atomic data.

5. Conclusions

Creation of a database of physicochemical properties of complex organic substances is the first intuitive one in the world, based on calculation methods application with full use of experimental data and up–to–date technologies. DAFIT is a complex multimodule system. Its graphic Java module is based on methods and models having source in modern thermodynamics achievements. It allows for critical data interpretation. Fast growth of data amount, a complex system and transactional database limitations lead to one conclusion — there is a strong necessity for application of data warehouse with high degree of aggregation.

Recommended OLAP technology for DAFIT database is ROLAP with snowflake or star schema.

References

[1] Gorawski M., Konopacki A. and Koziatek A., *Physicochemical and Thermophysical Database (DAFIT) at Internet — informative aspects*, Proceedings INFOBAZY'97, Gdansk, 23–25 June 1997 (in Polish)

[2] Gorawski M. and Frączek J., *Comparative Analysis of realisation of DAFIT Database in Oracle/Magic and Oracle/Developer2000 environments*, Zeszyty Naukowe Politechniki Slaskiej, Seria Informatyka, Zeszyt 30, 1997 (in Polish)