Design and implementation of an integrated knowledge system

d

aDepartment of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hunghom, Hong Kong, People’s Republic of China
bDepartment of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hunghom, Hong Kong, People’s Republic of China
cDepartment of Manufacturing Engineering and Engineering Management, The City University of Hong Kong, Hong Kong, People’s Republic of China
dDepartment of Mechanical Engineering, The University of the West Indies, St Augustine, Trinidad, Jamaica, WI

Received 29 January 2001; revised 22 November 2001; accepted 8 April 2002

Abstract

Case Base Reasoning (CBR), which is characterized by its capability to capture past experience and knowledge for case matching in various applications, is an emerging and well-accepted approach in the implementation Knowledge Management (KM) systems. The data format of CBR belongs to the “free” type and therefore is dissimilar to the traditional relational data model which emphasizes on specified data fields, field lengths and data types. However, there is a lack of research regarding the seamless integration of these heterogeneous data models for achieving effective data communication, which is essential to enhance business workflow of enterprises. This paper attempts to propose an integrated knowledge system to support the extrapolation of projected outcomes of events based on knowledge generated by the relational database model and CBR knowledge model, both of which supplement and complement each other by virtue of their distinct structural features.

q 2002 Published by Elsevier Science B.V.

Keywords: Data warehouse; Online analytical processing; Case base reasoning; Knowledge management

1. Introduction

Knowledge Management (KM) is meant to promote ongoing business success through a formal, structured initiative to improve the creation, distribution, or use of knowledge in an organization. Certain enterprises consider that their continued survival in the industry mainly depends on the successful implementation of the concept of KM [1]. For example, the KnowMore project used to support people who work on knowledge-intensive tasks by providing adequate information [10]. To implement KM in organizations effectively, computation intelligence tools are incorporated to achieve the self-growth of knowledge with past cases and examples, for the entire operation of relevant disciplines. It has been suggested that companies are becoming more knowledge intensive rather than capital intensive [2]. Due to the increasingly concern about shortage of knowledge workers in relevant KM activities, the capability to capture and retain knowledge within an organization without worrying about human turnover is of great interest to many executives. Moreover, the recent research in KM demonstrates the use of knowledge based system for cost control and analysis in order to support an automated alternative design analysis with online schematic drawing, material selection, crew selection, and productivity analysis for residential building project [11].

There is no doubt that the fundamental input of any intelligent and KM system is data. A successful company would treat data as the most valuable thing that it ever owns. Data is the raw material that is used to generate useful information and helpful knowledge, which is, in another extreme, a vital input for decision making and problem tackling. In today’s business world, the speed of product and service innovation has made KM crucial to any organization which is keen to stay on top in an information-oriented age [3]. The survival kit includes (i) locating knowledge from both within and outside an organization and capture the information, (ii) leveraging and storing the knowledge in an efficient manner, and (iii) using the knowledge derive profits for the company. In simple words, companies would need an effective and efficient framework and/or application to manage and to allow fluid access to an integrated knowledge repository [4].

* Corresponding author. Address: Department of Manufacturing Engineering, The Hong Kong Polytechnic University, Hunghom, Hong Kong, People’s Republic of China. Tel.: +852-2766-6628; fax: +852-2362-5267.
E-mail address: mfhenry@inet.polyu.edu.hk (H.C.W. Lau).
In order to develop a timely and flexible data access framework, and to support the corporate KM system, the combination of data warehouse, Online Analytical Processing (OLAP), and Case Base Reasoning (CBR) would be proposed. Whilst there are adequate publications about the data processing of CBR models, there is little literature related to the seamless integration of the data warehouse, OLAP, and CBR models to achieve effective bi-directional data communication. To achieve the implementation of an effective KM system, it is essential that relevant information is to be seamlessly unified to form an integrated knowledge system.

2. Review of related studies

Data warehouse is a large scale storage facility for data [3]. Data warehouse is a database designed to answer business queries by serving as a system for types of business data from many sources. OLAP, on the other hand, is application software that is used by users to explore the data in ways that are decision oriented. In general, data warehouse and OLAP systems include a data collector, data converter, data scrubber, data aggregator, data warehouse, OLAP server, and OLAP interface. However, some of these components may not exist depending on the organization needs and corporate constraints.

The functionality of the data collector of the data warehouse is to load data from several data sources. The data sources can be a mainframe located in another region or closely a hard drive of a personal computer located in another floor of the same building. Data collector would determine whether the information has been modified in the data sources and how it should be added or updated to the data warehouse. The data collector would also decide particularly which set of data should be downloaded to the warehouse while maintaining a schedule of the activities.

Data converter is responsible for translating data, since data from different data sources may have different formats [3]. It ensures only the data which has the standard formats that can be stored to the data warehouse. This is done to ensure data formats are standardized in the data warehouse, to prevent problems with data retrieval and data update.

To avoid problems occurring in the future, data scrubber cleans the data by identifying and remedies errors in data before storing them in the warehouse. This is done, for example, replacing a default value for missing values or modifying nonsense data before placing the data into data warehouse.

Data aggregator performs summarization of data by, for example, putting the annually purchasing information of a client into one single variable to allow timely access. The summarization criteria should be defined carefully based on the business strategic objectives. Since the detailed information regarding each purchase would be stored in the data warehouse, so generating aggregated information for the data warehouse does not mean there will be data loss. The purpose of performing summarization is to speed up some of the pre-defined and frequently used query.

Data warehouse would be responsible for storing all the data and performs a role as a database [5]. Once the data are in the warehouse, they are heavily indexed and combined for very fast access. Data warehouses often result in performance gain of several orders of magnitude even for standard database queries [2]. However, since the main purpose is for decision support and answer queries, it usually stores historical information in a more fluid-accessing manner.

OLAP systems is data manipulation software system. It is very flexible as comparing with the relational database management and expert system, where both only provide pre-defined queries for user. However, OLAP system allows end-user to select, mix and match variables, constraints, and measures without predefining the query while developing the system. OLAP functionality includes: (i) calculation and modelling applied across dimensions via hierarchies and/or across values, (ii) trend and seasonal analysis over time periods, (iii) slice and dice data in almost any manner, (iv) drill down into data to get deeper level of detail, (v) reach-through of underlying detail data, and (vi) rotation of new dimension of data viewing. There are numbers of approaches to OLAP representation, the most commonly used is a multi-dimensional approach to data storage [3]. To obtain a specific data by going to the appropriate dimension and select the correct input and output cell.

Both data warehouse and OLAP system allow flexible and timely data manipulation and data access to users. Answers to most of the queries have already loaded into the cells in the OLAP hypercube data representations. With the combination of extensively indexed and combined data in the data warehouse, OLAP tools allow user to dig through megabytes or gigabytes of data without having to wait hours for results.

Case contains a set of attributes which describe the problem and the solution for the problem, or describe a state of affairs. A collection of cases is called a case base. Within a case base, cases may share some common attributes, so there may be overlaps between cases. CBR system solves a problem by searching through the case base for cases whose attributes are similar.

CBR is popularly used to solve problem which has no clear-cut theories. It is based on human reasoning model, i.e. mentally search for similar situations happened in the past and reusing the experience gained in those situations [6]. CBR consists knowledge capturing and knowledge reuse features. In other words, CBR system will ‘grow’ as it ‘learns’ from experience and solving new problems. Its learning capabilities enhance problem solving skills and knowledge base.

CBR mimics human experts’ methodology to retrieve case rapidly. According to theory, experts employ effective indexing schemes [3]. When a case is recorded to their mind, the expert somehow indexes it using a set of attributes regarding the case. The more important the case, the larger number of indexes the expert is likely to use to record the case. The indexes serve as pointers to locate similar cases which are similar to the problem [7]. If the case is not heavily indexed, the expert may have to spend more time for retrieval. Since CBR simulates experts indexing schemes, good indexing is vital to create an efficient CBR system, as the case indexing takes a crucial rule for timely and accurate retrieval [8].

Review of contemporary publications indicates that whilst many research activities are done on data warehouse, OLAP system and CBR, the research in the area related to the seamless integration between them and the drill-down process have not received the attention it deserves. This issue is addressed in this paper with the introduction of an integrated system by integrating data warehouse, OLAP system, and CBR, providing a synergetic combination of various techniques and technologies related to information storage.

3. Existing practice

The unique feature, allowing user to ask “What if” to different scenarios, makes OLAP a great decision making tool for determining the best course of action for the company’s business and operation. Data warehouse and OLAP systems work closely and complement each other; the data warehouse stores and manages the data while OLAP manipulates the stored data into useful information.

Data warehouse and OLAP systems enable companies and manufacturing firms as a whole to combine, store, manipulate, and respond more quickly to market demands. This is possible because they provide the ability to solve real business and/or production problems, making better decisions for the conduct of the organization, and use people and material resources more efficiently [9]. Market responsiveness often yields improved revenue and profitability. An example from the real business world, Office Depot, on of the biggest office equipment suppliers in the US, significantly improved its sales by using OLAP tools [5].

On the other hand, CBR system solves a problem, rather than starting from scratch, it searches its case base for cases that have similar attributes [3]. It will then create a solution by synthesizing the similar cases and adjusting the final answer for differences between the current situation and the one described in the problem. Besides the attributes have to be heavily indexed, it also requires a database and retrieval software application in order to support an efficient access of attributes and match similar attributes in timely manner. Therefore, in this paper, a framework will be proposed to achieve timely access of useful information from CBR while taking the advantages of CBR self-learning features for support of problem solving and decision making.

4. An integrated knowledge system

The model, recommended in this paper, contains three main components: data warehouse, OLAP, and CBR systems. This combination would bring upon the advantages of each module while complementing each other’s weaknesses.

4.1. Data warehouse module

Case is usually in free format, which means it does not have a structured and standardized format for the cases no matter how well the system has designed or how simple is the business nature. Therefore, it is not very wise to store each case in record and table format, which includes query and solution, in a database. Since a standardized format cannot be developed for the cases, case-matching problem may occur. For example, imagine a new inquiry input into the CBR system in a structure that is different from all the cases in the system. The system will have difficulties to generate a match for the new inquiry since it cannot match with any cases due to structure differences, even though there is indeed an existing case in the case base that has 99.99% similarity to the new inquiry. Therefore, in order to solve the problem, cases should be stored and searched in their subset format—attributes. A set of attributes describes a stage of affairs of a case. In fact, since some cases consist common attributes, which indicate these cases have certain degree of similarities, it will save a lot of resources in the process of matching cases.

Data warehouse would take the role as a database where contains all the attributes. It is also responsible for collecting, converting, cleaning, aggregating, and indexing data, i.e. attributes, from different data sources. The attributes within data warehouse would be extensively indexed. Each attribute point to designated cases that contain the attribute. Furthermore, the aggregated data, i.e. aggregated attributes, are generated to indicate similar cases, which are cases that share common attributes, and answers that may solve queries in the future.

4.2. OLAP module

The purpose of OLAP systems in this framework are to be the bridge between CBR and data warehouse, the software to facilitate timely access and manipulation of the data, and the application to drill down into data to obtain further information. It will also locate the aggregated data into the data block, so that answer can be given in timely manner for frequently asked questions. The result of the aggregation would speed up attributes matching and cases
searching. However, this paper will not go into details regarding how the aggregated data should be done, since it mainly depends on the business nature.

The model will take advantages of OLAP systems of its flexible and timely manipulation of data and its special functionalities, such as drill, slice, and dice. OLAP is used for searching and matching of similar attributes for new inquiry, which is delivered from CBR, by manipulating the attributes in the data warehouse and retrieving the cases that are directed by the matched attributes. In the meantime, it will drill down further for the differences that are found between the inquiry and the similar cases in order to get the best solution. OLAP would use multi-dimensional approach to obtain a specific data by going into proper dimension based on the new inquiry and choose input and output cell.

4.3. CBR Module

The CBR module will take the role as the interface between the system and the user, and send updating attributes messages to the data warehouse if there are modifications to attributes.

Once a new inquiry has been input into the CBR system, it will be sent to OLAP which performs search and match to find the attributes that are similar to the inquiry’s attributes in the data warehouse. OLAP will retrieve one or more similar cases from the case base, and drill, slice, and dice the data warehouse in order to generate solution. If the database does not contain a matching case that is 100% match to the new inquiry, a proposed solution would be made and a newly generated case will be retained in the case base (Fig. 1).

5. Case example

A medical school has developed a case base learning tool for the medical students to enhance their practical skills in helping patients who, in the real world, have different symptoms. This learning tool grows as more cases are captured by the system. Each case has a collection of symptoms and information on the symptoms. The Microsoft SQL Server is selected to be OLAP application, while Case Advisor 4.12 is deployed to be the case base application. Since there was no existing database which can be used to build the application, three steps have to be taken to builds the model and case base. Firstly, a data model is necessary to be defined, while the data model contains descriptors that include all the characteristics the potential user is likely to require. For the medical learning program, the advantage of a case base would associate the symptoms with...
the appropriate treatments. Secondly, the descriptors of a domain will be defined in a hierarchical way—a domain contains several subclasses. For example, for the descriptor, Symptoms, the case base would contain a lot of different symptoms (keywords). The last step is to create the case base by building a questionnaire to capture inputs from the user.

Let’s assume a case base contains three cases in current status. Case #1 includes attributes A, B, C, and D. Case #2 has attributes C, D, E, F, and G. Lastly, Case #3 contains attributes X and Y. Each attribute represents a symptom and the detail information on the symptoms. For example, attribute A represents running nose and attribute B stands for 3 days (Fig. 2).

A hypercube and multi-dimensional data representation of OLAP makes it convenient to query data along any dimension. Both Case #1 and Case #2 includes attributes C and D, which stand for pain on lower back and itchy eyes, respectively. They are considered to be aggregated data stored in the database since they are common attributes of different cases. In OLAP system, another type of aggregated data based on pre-defined rules will also be computed from each case. For example, attribute M represent prescribed Aspirin is generated from Case #1, attribute P stands for refer to allergy specialist is generated from Case #2, and attribute Z represents refer to dermatologist is generated from Case #3. These attributes are generated for providing answer that may solve and answer future inquiries that are asked quite often. These attributes can be pointer that points to other cases or simply a number calculated based on a specific case. In other words, these aggregated data are pre-computed by rules based on business nature and user needs (Fig. 3).

Pseudocode of the predefined rules:

Function generate_aggregate_data
User input symptoms
Check similar symptoms
Rank and Suggest treatment
Evaluate treatment by user
Store case

Function check_similar_symptoms
Compare input_symptoms and stored_symptoms
Retrieve cases with similar symptoms

Function rank_suggest_treatment
Compare cases
Rank cases

Function store_case
Submit case with treatment solution to database

A user input an inquiry Case #4 with the attributes E, X and Y, which represent watery eyes, itchy skin, and swollen skin, respectively, to the system. This inquiry is sent to OLAP system which finds that both Case #2 and Case #3 contains part of Case #4—Case #2 contains attribute E while Case #3 contains attribute X. OLAP system drills down and finds aggregated attributes P and Z (cannot be shown in Fig. 4 since it is on the second level of the plane). The multi-dimensional structure of the OLAP system allows timely access of information related to the input. To answer the inquiry, system retrieves all the cases contains the attributes E, X, and Y, which is Case #2 (attributes C, D, E, F, G) and Case #3 (attributes X, Y) and the aggregated attributes P and Z, which has already generated and located in data bock in OLAP for timely access. A temporary solution will be given with the attributes C, D, F, G, P, and Z, which are pain on the lower back, itchy eyes, dry mouth, 2 days, refer to specialist of allergy, and refer to dermatologist, respectively (Fig. 4).

Refinement would be done by OLAP system to drill, slice and dice the data for the differences between Case #4 and Case #2 and Case #3. The system finds Case #2 only has one attribute out of five that matches one attribute in Case #4. There is little similarity between these two cases. Therefore, the aggregated data generated from Case #2 would not be very appropriated to become part of the solution. In other words, attribute P would not be included in the solution since it is generated by all five of the attributes in Case #2. For the aggregated attribute C and D, which exist in both Case #1 and Case #2 would be reserved as the user may find them necessary for refining the answer. On the other hand, Case #4 and Case #3 are closely matched between the attributes. So, the aggregated...
attribute that is generated from Case #3 would be included in the solution. Therefore, the answer to the new inquiry, Case #4, will be attribute Z, which stands for refer to dermatologist with attributes C, D, E, and G as related answer to allow user to drill down for more information. Case #4 will be retained in the case base and aggregated attributes will be generated based on pre-defined rules to answer future inquiry.

It should be noted that the answer to Case #4 cannot be 100% satisfied, since it can only provide an answer to attribute X and Y but not E, X, and Y. Users are required to send another inquiry to obtain further information and/or more details by drilling down the related answer provided along with the proposed solution. However, this framework allows user to drill down further if they want to know more about attribute E which represents watery eyes (Figs. 5 and 6).

6. Discussion

The proposed model of the integrated knowledge system serves the purpose of providing the closest answer to the user inquiry while user can dig through the data in timely
manner. This framework also takes the advantage of CBR self-learning abilities. There are aggregated data generated and stored in the system, which provides quicker search between related cases. Some of these aggregated data are generated by pre-defined rules to answer user frequently asked questions. Furthermore, the learning ability of CBR enlarges the database, which makes it able to answer wide range of query as the system grows (Fig. 7).

On the other hand, since data warehouse does not update in real time. It only performs update periodically during non-office hours to avoid network traffic. The solution provided to the inquiry may not be the most update information. Also, the workloads are focusing on the data warehouse and OLAP system as the data manipulations are mainly done in those two levels. Therefore, if data are loaded to the data warehouse in real time, part of the resources in the system will be allocated to load and process the data, as the result answer to user inquiry cannot be generated quickly with lower level of resources.

7. Conclusion

The development of the integrated knowledge system which is incorporated with data warehouse, OLAP system, and CBR knowledge model has been described in this paper. The unique feature of this knowledge system is concerned with the creation of solutions by synthesizing the similar cases and adjusting the final answer for differences between the new inquiry and the ones in the case base in timely manner. This system is a significant contribution in the KM in terms of information interchange and knowledge updating across heterogeneous platforms. However, in order to resolve the drawbacks of non-real-time data, further research is required.

References

[7] A. Aamodt, E. Plaza, Case-based reasoning: foundational issues, Methodological variations, and system approaches, AICom—


