Knowledge Acquisition For Reconstruction And Identification Of Dimensionless Groups By Means Of Case Based Reasoning

Fragiskos Batzias and Odysseas Kopsidas

Laboratory of Simulation of Industrial Processes
Department of Industrial Management and Technology
University of Piraeus
Karaoli & Dimitriou 80, 18534 Piraeus, Greece
fbatzi@unipi.gr

Abstract. When searching with Data Mining Techniques to identify or find out dimensionless groups (DGs) in technical literature, it is possible to meet errors/faults/omissions concerning both, the form and the content of such groups. In the present study, a methodological framework has been developed in terms of a logical flow chart, including 11 activity stages and 7 decision nodes, to acquire/process/store/retrieve knowledge for reconstruction and identification of these groups. Case Based Reasoning (CBR), especially modified to meet the needs of this work, has been used for tracing causality paths by similarity and making correction suggestions. Two case examples are presented to prove the functionality of the proposed methodology.

Keywords: dimensional analysis, knowledge acquisition, dimensionless groups, case based reasoning, data mining, Blake number, modified Reynolds number, adsorption, sedimentation, flocculation, filtration, scale up.

INTRODUCTION

Dimensional analysis was initially developed early in the last century [1-6] by scientists and engineers for expressing the behavior of a physical system in term of the minimum number of independent variables and in a form that is unaffected by changes in the magnitude of the units of measurement. The first attempt to a similar approach in Economics was made by Allais [7], who presented a systematic treatment of the theory of dimensions and its foundations. Later, some authors contributed to the diffusion of this scientific topic (e.g., see [8]) or used relevant techniques either for simply stressing the distinction between stock and flow variables [9,10] or for techno-economic analysis [11,12] or for model checking [13]. The physical and/or economic magnitudes are usually arranged in dimensionless groups (DGs), the number of which is $m=n-r$, where n is the number of dimensional variables/parameters/constants (VPCs) and r is the number of primary quantities used, like mass M, length L, time T, and temperature G. This rule is a general relation based on Linear Algebra and Group Theory and is mostly quoted as Buckingham's pi theorem.

DGs may have physical meaning and in such a case we can search for similar groups through Case Based Reasoning (CBR) in order to properly describe quantitatively and solve similar problems, possibly occurring due to errors either in dimensions or in form or in content. CBR has been formalized for purposes of computer-aided processing as a four-step algorithmic procedure, which might be named 4R from the initial letter of the four verbs used as computer program commands in the corresponding steps: (I) Retrieve the cases considered to be relevant to solving a given target problem. (II) Re-use the solution (if it was successful) from the previous case to the target problem by making the necessary modifications to adapt the solution as needed to fit the new situation. (III) Revise by testing the solution of the new problem directly in the real world or indirectly under suitable conditions or reliable simulation. (IV) Retain by storing the solution as a new case in the computer memory, provided that the adaptation was a successful one.

Computer-aided solving by CBR the problems quoted above is a difficult task, especially when this approach uses unpublished evidence as its main Knowledge Base (KB); since the available information from the past may be not relevant/reliable to the required level for backing a successful 'implementation-by-analogy', there is no guarantee that the necessary generalization (that ensures applicability) is correct. Nevertheless, inductive reasoning in case of scarce data/information for statistical relevance/reliability is usually based on incomplete evidence in practice. Moreover, there is always the possibility of developing CBR within a statistical framework and formalizing case-based inference as a specific type of probabilistic/possibilistic inference; therefore, it becomes feasible to make case-based estimates with a preset level of confidence.

METHODOLOGY

The methodological framework we have developed, in terms of a logical flow chart with 11 activity stages and 7 decision nodes (interconnected as shown in Fig. 1), to acquire/process/store/retrieve knowledge for reconstruction and identification of dimensionless groups by means of CBR, is presented below:

1. Analysis of the DG in error or the dimensionally inconsistent VPC.
2. Checking of the definitions of the VPCs registered under widely known names in relevant KBs and, subsequently, in the corresponding original sources of technical literature.
3. Translation of these verbal or symbolic definitions into dimensional definitions in accordance with the dimensional system adopted from the start as most appropriate.
4. Replacement of the VPCs suspected to be in error with the new/checked ones and testing of the DG under consideration for dimensional homogeneity.
5. Checking for operability within a proper/successful paradigm extracted from a KB by means of data mining and knowledge acquisition techniques to be used in accordance with CBR.
6. Combinations of DGs to find out those which are more close to the DG under consideration.
7. Checking these combinations for operability *ab initio* (i.e., without having necessarily a past successful paradigm).
8. Discrimination of relevant DGs and ranking of them in descending order of similarity degree as regards the DG to be corrected.
9. DA by replacing *a priori* the DG under correction with the most similar (but unexamined so far) DG within the set of DGs resulted as the solution of the dimensional matrix corresponding to the system of equations obtained according to Rayleigh's Method of Indices.
10. Multicriteria ranking of combinations performed in stage 6, according to the order of decreasing degree of closeness to the DG under consideration.
11. Checking of the first combination (among the ones not examined so far) for operability.
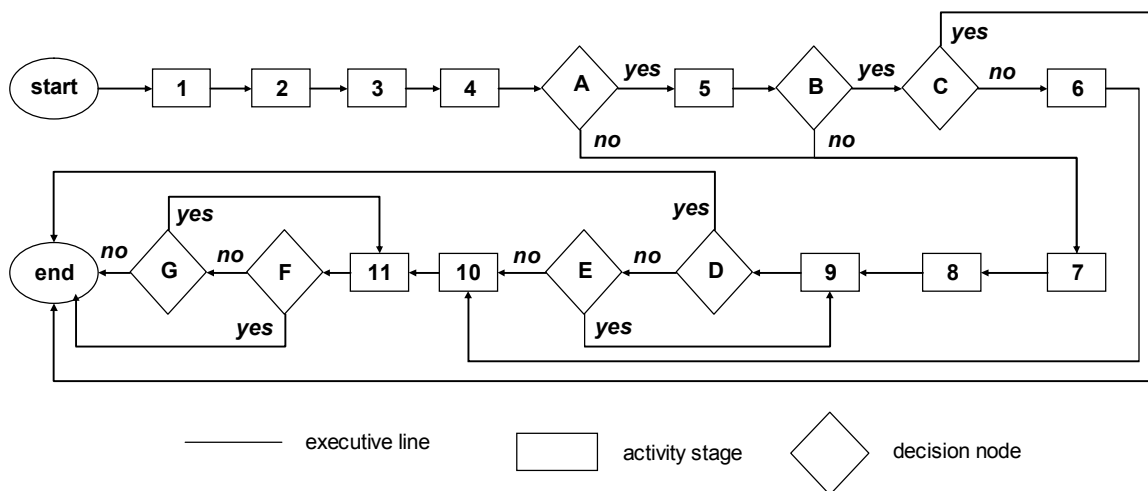

FIGURE 1. The logical flow chart developed for acquiring/processing/storing/retrieving knowledge for identification and correction of either dimensionally inconsistent VPCs or/and DGs in error.

A. Has dimensional homogeneity been restored?
B. Is operability satisfactory?
C. Is the new DG identified as the correction of the old one?
D. Is this replacement successful?
E. Is there another ranked DG with lower similarity degree?
F. Is its operability satisfactory?
G. Is there another ranked combination not examined?

IMPLEMENTATION BY CASE EXAMPLES

We have implemented successfully the methodology described above in a plethora of cases. Subsequently, we present two of them. The first refers to modified Reynolds or Blake Number, defined in Perry's Chemical Engineering Handbook [14] as $B = V\rho / [\mu(1-\varepsilon)S]$, where V: characteristic or average fluid velocity (m/s); ρ: fluid density (kg/m^3); μ: fluid viscosity (Pa s); ε: void fraction (m^3); S: particle area/volume (1/m); all magnitudes measured in SI units. By replacing the units with the corresponding dimensions, we obtain $B \in [LT^{-1}]$ $[ML^{-3}]$ $[ML^{-1}T^{-1}]^{-1}$ $[L^3]^{-1}[L^2L^{-3}]^{-1}$ or $B \in [L^{-3}] \neq [1]$. Evidently, this is in error, since the result should be $B \in [1]$, i.e., dimensionless. Searching in external KBs by means of data mining and knowledge acquisition techniques incorporated within an Intelligent Agent (IA), according to [15], we find the most recent reference source [16] with all common DGs, where $B = V\rho / [\mu(1-\varepsilon)d]$, with $\varepsilon \in [1]$, while d is the characteristic particle diameter $d \in [L]$ as recognized by the computer aided DA procedure adopted herein and thoroughly described in [17]. Working out as above, we obtain $B \in [LT^{-1}][ML^{-3}][ML^{-1}T^{-1}]^{-1}[L]^{-1}[L^2L^{-3}]^{-1}$ or $B \in [L^{-1}] \neq [1]$. By continuing the automatic searching, as shown in the flow chart of Fig. 1, the computerized procedure identified finally the correct dimensions to be $\varepsilon \in [1]$ and $S \equiv d \in [L^{-1}]$, giving B $\in [1]$.

The second case example used for implementation is the Freundlich isotherm, which we have used recently [18] to study by experimental simulation the dye adsorption in batch and fixed-bed systems:

$Q = KC^{1/n}$

where q = the amount of mass adsorbed per unit mass of the adsorbent [1]; C = the equilibrium concentration of the adsorbate, $[ML^{-3}]$; K = parameter related to adsorption capacity; n = parameter related to adsorption intensity, [1]. Most authors do not specify the dimensions of K, while several of the rest of them assign either $[L^3M^{-1}]$ or [1] (e.g., see [19] or [20,21], respectively). Evidently, these dimensional assignments are in error, as the dimensional homogeneity of this isotherm requires $K \in [M^{1/n} L^{3/n}]$, although this assignment is meaningless, when its physical content is considered thoroughly. By referring mass M to dye in solution, dye adsorbent, adsorbent itself, we can split this primary dimension to M_s, M_d, M_t, respectively. Consequently, we can write the isotherm model in dimensional form as follows: $q = K_s K_d K_t C^{1/n}$, where $q \in [M_d M_t^{-1}]$, $C \in [M_s L^{-3}]$, $K_s \in [M_s^{-1} L^3]^{1/n}$, $K_d \in [M_d]$, $K_t \in [M_t^{-1}]$, which is meaningful as representing the real situation (facilitating also the scale up/down of the corresponding simulation procedures).

DISCUSSION AND CONCLUDING REMARKS

Even in physical sciences and engineering/technology, this identification attempt becomes cumbersome and the computer program may fail to suggest a realistic solution without substantial human intervention. The probability to fail increases when the DG under investigation is incorporated into relations that change the original VPCs with others for sake of better fitting to system's specifications. For example, when the system is 'fluid flow through porous media', B can be combined with the 'resistance coefficient' DG expressed by the ratio $\Psi = A^2\varepsilon^3 P / [Q^2\rho(1-\varepsilon)lS]$ to give a relation of the general form $\Psi = K/B^n$ in the procedure known as the Blake-Carman correlation, where A is the cross-sectional area of packing perpendicular to flow direction, P is the pressure difference, Q is the volumetric flow rate ($Q=VA$), l is the length of packing in the flow direction, K and n are parameters. In the low B-range (approximately, $B<1$), $n=1$, and by re-arranging the VPCs, we obtain the Kozeny-Carman equation $Q = A\varepsilon^3 P / [K\mu(1-\varepsilon)^2 S^2 l]$, which is dimensionally homogenous, since $Q \in [L^3T^{-1}]$ and $A\varepsilon^3 P / [K\mu(1-\varepsilon)^2 S^2 l] \in [L^2][1][ML^{-1}T^{-2}][1][ML^{-1}T^{-1}][L^{-1}]^{-2}[L]^{-1}=[L^3T^{-1}]$.

In such a case, we cannot investigate any dimensional inhomogeneity in *B* since this DG does not appear *per se* in the final model while the investigation of any error propagation should be based on CBR where the similar cases to be retrieved should belong to a broader family of Kozeny-Carman equations. If, according to [16], *S* is replaced by *d* in *B*, then the Kozeny-Carman equation becomes $Q = A\varepsilon^3 P / \left[K\mu(1-\varepsilon)^2 Sdl\right]$. Consequently, the dimensional vector of the right-hand side of this equation becomes $[LT^{-1}]$ which is inconsistent with the left hand-side dimensional vector $[L^3T^{-1}]$ denoting the correct dimensions of volumetric flow rate *Q*. The IA cannot identify the source of error unless provided with further information corresponding to ancestor relations. On the other hand, the inference engine searching within the KB may identify *Q* with the 'superficial velocity' u_0 (i.e., the average linear velocity the fluid would have in the column if no packing were present) quoted as the dependent variable of Blake-Kozeny equation in the classic work [22].

As a matter of fact, searching for the possible routes of error propagation and the corresponding implications is a 'direct problem' while searching for the possible source of error through an identification procedure is the 'inverse problem'. In the case examined herein, searching for implications of error in Blake Number *B* is the direct problem while searching for the possible source of error in Kozeny-Carman equation is the inverse problem. It is also worthwhile noting that the same family of equations is frequently met under all pairwise combinations (i.e., Kozeny-Carman, Blake-Carman, Blake-Kozeny) leading to additional difficulties faced by the IA searching in external KBs using the keywords interface suggested in [15].

In conclusion, we have indicated the functionality of the methodological framework presented herein by analyzing two simple cases, one for DG and another for VPC. If the DGs or VPCs in eror have been incorporated/combined into/with other expressions, traceability decreases and human intervention, at least for changing the searching pattern through CBR, increases the effectiveness of computer aided performance while facilitates the KB's enrichment/restructuring.

REFERENCES

1. L. Rayleigh, *Proc Roy. Soc. London* **66**, 68 (1899-1900).
2. L. Rayleigh, *Nature* **95**, 66 (1915).
3. E. Buckingham, *Phys. Rev.,* **4**, 345-376 (1914).
4. H.E. Langahaar, *Dimensional Analysis*, London, 1951.
5. H.E. Huntley, *Dimensional Analysis,* London, MacDonald, 1952
6. J. Palacios, (translated by P. Lee with L.Roth), *Dimensional Analysis*, London, Mac Millan, 1964 (originally published in Spanish, 1956).
7. M. Allais, *Traite d' economie pure*, Vol. I, *Les Donnees generals de l' economie pure* , Paris and vol. IV, *Annexes* (2nd Ed. of a textbook published in 1943 under the title *A la recherché d' une discipline economique*).
8. F.J. de Jong, *Dimensional Analysis for Economists*, Amsterdam, North Holland Publ. Co, 1967.
9. K.E. Boulding, *Economic Analysis*, 4th Ed., Vol. II, New York, 1966.
10. G. Auckley, *Macroeconomic Theory*, New York, 1961 .
11. T. Nussbaumer and P. Neuenschwander. *Biomass Bioenerg.* **18**, 181-188 (2000).
12. Le Corre, O., GuillaumB. and Mohand, T., *Appl. Thermal Eng.* **84**, 177-185 (2002).
13. F.A. Batzias, D.K. Sidiras, *J. Hazard. Mater.* **114**, 167-174 (2004).
14. R.H. Perry, D.W. Green, *Perry's Chemical Engineers' handbook,* 7th Ed., McGraw-Hill 1997, pp. 648-650.
15. F.A. Batzias, E.C. Marcoulaki, *Comp. Aided Chem. Eng.* **10**, 829-834 (2002).
16. http://www.uni-magdeburg.de/ivt/mvt/englisch/framemvt.html and R.H. Perry, D.W. Green, J.O. Maloney, Perry's Chemical Engineers' Handbook (CD version), pp. 6-49, McGraw-Hill, New York, 1999.
17. F.A. Batzias, A.S. Kakos, N.P. Nikolaou, *J. Computat. Meth. Sci. Eng.* **5**, 47-60 (2005).
18. F.A. Batzias, D.K. Sidiras, *J Hazard. Mater.*, **114**, 167-174 (2004).
19. G.M. Walker, L.R. Weatherley, *J. Chem. Eng.* **83**, 201-206 (2001).
20. C. Costa, M.C. Marquez, *Wat. Res.* **32**, 107-114 (1998).
21. C. Ng, J.N. Losso, W.E. Marshall, R.M. Rao, *Bioresour. Technol.* **84**, 177–185 (2002).
22. R.B. Bird, W.E. Stewart, E.N. Lightfoot, *Transport Phenomena*. Wiley International, New York, 1960, pp. 196-200.