Knowledge discovery in scientific data using hierarchical modeling in dimensional analysis

Steffen Brückner and Stephan Rudolph

Institute for Statics and Dynamics of Aerospace Structures
Universität Stuttgart, Pfaffenwaldring 27
Stuttgart, Germany

ABSTRACT

In the automotive and the aerospace industry large amounts of expensively gathered experimental data are stored in huge databases. The real worth of these databases lies not only in easy data access, but also in the additional possibility of extracting the engineering knowledge implicitly contained in these data. As analytical modeling techniques in engineering are usually limited in model complexity, data driven techniques gain more and more importance in this kind of modeling. Using additional engineering knowledge such as dimensional information, the data driven modeling process has a great potential for saving modeling as well as experimental effort and may therefore help to generate financial benefit.

In a technical context, knowledge is often represented as numerical attribute-value pairs with corresponding measurement units. The database fields form the so-called relevance list which is the only information needed to find the set of dimensionless parameters for the problem. The Pi-Theorem of Buckingham guarantees that for each complete relevance list a set of dimensionless groups exists. The number of these dimensionless parameters is less than the number of dimensional parameters in the dimensional formulation, thus a dimensionality reduction can easily be accomplished. Additionally, dimensional analysis allows a hierarchical modeling technique, first creating models of subsystems and then aggregating them consecutively into the overall model using coupling numbers.

This paper gives a brief introduction into dimensional analysis and then shows the procedure of hierarchical modeling, its implications, as well as its application to knowledge discovery in scientific data. The proposed method is illustrated in a simplified example from the aerospace industry.

Keywords: Knowledge Discovery, Scientific Data, Dimensional Analysis, Hierarchical Modeling

1. INTRODUCTION

In engineering analysis top-down and bottom-up modeling procedures are used as complementary approaches. The top-down modeling allows to start with a "black-box" model which is continously refined until the desired level of modeling accuracy is reached. Bottom-up modeling on the other hand starts with detailed partial models which are then assembled together to form more complex models until the model has the desired model complexity. Both methods are supported by hierarchical modeling using dimensional analysis. For knowledge discovery systems the process of hierarchical modeling using dimensional analysis can be automated, since the procedure inherently processes the occuring physical couplings as will be shown later.

In the following, dimensional analysis is presented first. Then hierarchical modeling using dimensional analysis is introduced and the application to knowledge discovery in scientific data is discussed. The procedure of hierarchical modeling using dimensional analysis is illustrated using the example of an aerospace gas turbine.

Further author information: http://www.pigroup.de/, [steffen.brueckner|stephan.rudolph]@pigroup.de

2. DIMENSIONAL ANALYSIS

In the engineering domain and the natural sciences dimensional homogeneity is required for all valid relationships[1], since all dimensionally inhomogeneous relationships are agreed on to be wrong by definition. This principle provides a restriction on the search space of admissible functional relations for a given problem. Conventional knowledge discovery systems often ignore or even violate this requirement and have therefore to search a much larger search space containing also inadmissible functional relationships.

Based on the principle of dimensional homogeneity, the method of dimensional analysis yields m dimensionless groups $\pi = \{\pi_1, \ldots, \pi_m\}$ which are scaling-invariants of the problem and which form a dimensionless and therefore dimensional homogeneous relationship $F(\pi)$ corresponding to the dimensional relationship $f(x)$.

$$f(x) \xrightarrow{\Pi} F(\pi) \tag{1}$$

Due to the property of scaling-invariance, these dimensionless groups are often used for scale-up and scale-down (e.g. the Reynolds and Prandtl numbers in fluid dynamics are such dimensionless groups) and are called *similarity numbers*. Dimensional analysis provides an automated procedure to generate the dimensionless groups from the knowledge of the relevance list $x = \{x_1, \ldots, x_n\}$. This list consists of the physically relevant parameters and the dimensional representations thereof. The relevance list corresponds to the list of database fields along with the dimensional representation of the contents.

One possible scheme to generate the similarity numbers automatically from the knowledge of the relevance list is the matrix method[2] which is explained below.

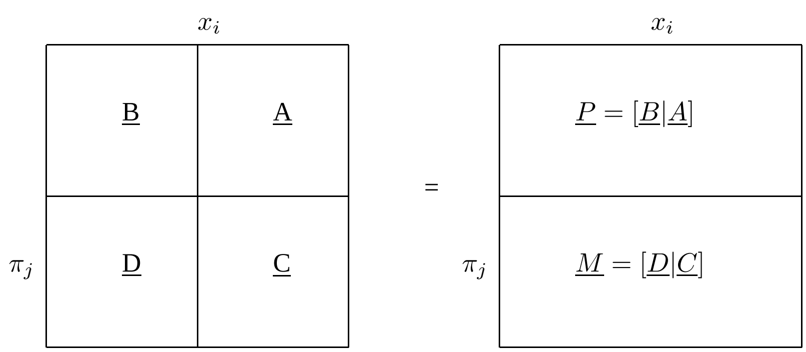

Figure 1. The matrix method of dimensional analysis, partitioning of the dimensional set[2]

The dimensional matrix

$$P = [B|A] \tag{2}$$

as the upper part of the dimensional set

$$\left[\begin{array}{c} P \\ \hline M \end{array}\right] = \left[\begin{array}{c|c} B & A \\ \hline D & C \end{array}\right] \tag{3}$$

also shown in figure 1, has columns corresponding to the physical variables x_i of the relevance list and rows corresponding to the dimensional exponents of these physical variables. Using the SI-system of units, the number of rows of the dimensional matrix P cannot exceed the seven base dimensions of this system (i.e. the length in meter, mass in kilogram, time in seconds, thermodynamic temperature in kelvin, amount of substance in mol, luminous intensity in candela, and electric current in ampere). The size of the matrix P is $r_P \times c_P$ where r_P is the number of base dimensions in the relevance list (which is often identical to the rank r of the matrix P) and c_P is identical to the number n of physical variables. If the rank of the dimensional matrix P is less than the number of base dimensions r_p, one or more linearly dependant rows have to be eliminated from the dimensional matrix. This procedure is explained in detail in Szirtes[2].

According to equation (2) the dimensional matrix is partitioned in the rightmost determinant, which is called the submatrix A of size $r_A \times c_A = r \times r$ and the remaining columns, called the submatrix B of size $r_B \times c_B = (n - r) \times r$.

The submatrix D of the dimensional matrix has to be a regular matrix[2] of size $r_B \times c_B = (n - r) \times (n - r)$ but besides this restriction D can be freely chosen by the engineer to accomodate for already known dimensionless groups. Usually, at least for a first try, the matrix D is chosen to be a identity matrix.

The dimensional analysis is then performed calculating the submatrix C using the equation

$$C = -D \cdot \left(A^{-1} \cdot B\right)^T \tag{4}$$

The dimensionless groups can now be determined from the combined submatrices $M = [D|C]$. The matrix elements m_{ji} are the exponents of the corresponding physical variables x_i in the dimensionless group π_j.

$$\pi_j = \prod_{i=1}^{n} x_i^{m_{ji}} \tag{5}$$

In knowledge discovery applications the list of relevant parameters (i.e. the relevance list) is given by the field names of the data base / data warehouse. For scientific data, also the dimensional representation (e.g. $[N] = [kg]\ \ [m] / [s]^2$ for the force measured in Newton) has to be stored. A set of dimensionless groups corresponding to the relevance list (i.e. the problem) can then be determined automatically using the procedure explained above.

Since the dimensionless groups form a minimal set of parameters for a given problem, it is the most compact formulation possible, which is also advantageous for most data mining algorithms. Additionally, using the similarity groups of a given problem, completely simlar cases can be determined and partially similar cases can be identified, which provides a similarity measure for a subsequent identification procedure of the automated modeling technique.

3. HIERARCHICAL MODELING

In general, as a rule of thumb, the complexity of the system identification procedure depends on the number of degrees of freedom (i.e. parameters). Simpler systems are much easier to identify than complex systems. Data for subsystems can often be taken from different experimental setups, even from different systems which contain the same subsystem. Hierarchical modeling provides a procedure to model simple subsystems first and store this knowledge for (re-)use. More complex systems are then formed by aggregating the known subsystems[3]. This procedure allows the effective reuse of the same subsystem knowledge in the modeling of different complex systems.

To model a complex system, the subsystems are aggregated using dimensional analysis. Additionally to the similarity parameters of the subsystems, so-called coupling numbers are generated which account for the couplings between the subsystems. Using this methodology the subsystem can be refined using new data and can be reused for all complex systems which do contain this subsystem.

Combining subsystems with n_i physical and m_i dimensionless variables and the ranks r_i of the dimensional matrices of the subsystems respectively, a new dimensional matrix P of size $r_P \times c_P = r \times \sum_i n_i$ with $r = rank(P)$ is created. The matrix D is then of the size $(\sum_i n_i - r) \times (\sum_i n_i - r)$.

The number of dimensionless groups for the aggregated system is then given by the relation

$$\begin{aligned} m &= \sum_i n_i - r \\ &= \sum_i (n_i - r_i) - r + \sum_i r_i \\ &= \sum_i m_i + \left(\sum_i r_i - r \right) \end{aligned} \tag{6}$$

This is the sum of the number of dimensionless groups of the respective subsystems and additional

$$m_c = \sum_i r_i - r = r \cdot \sum_i \frac{r_i}{r} - 1$$

dimensionless groups. These m_c additional dimensionless groups cannot be formed from variables from one of the subsystemes alone, but only by using variables from more than one subsystem. These additional numbers account for the couplings between the different subsystems and are therefore called *coupling numbers*[3].

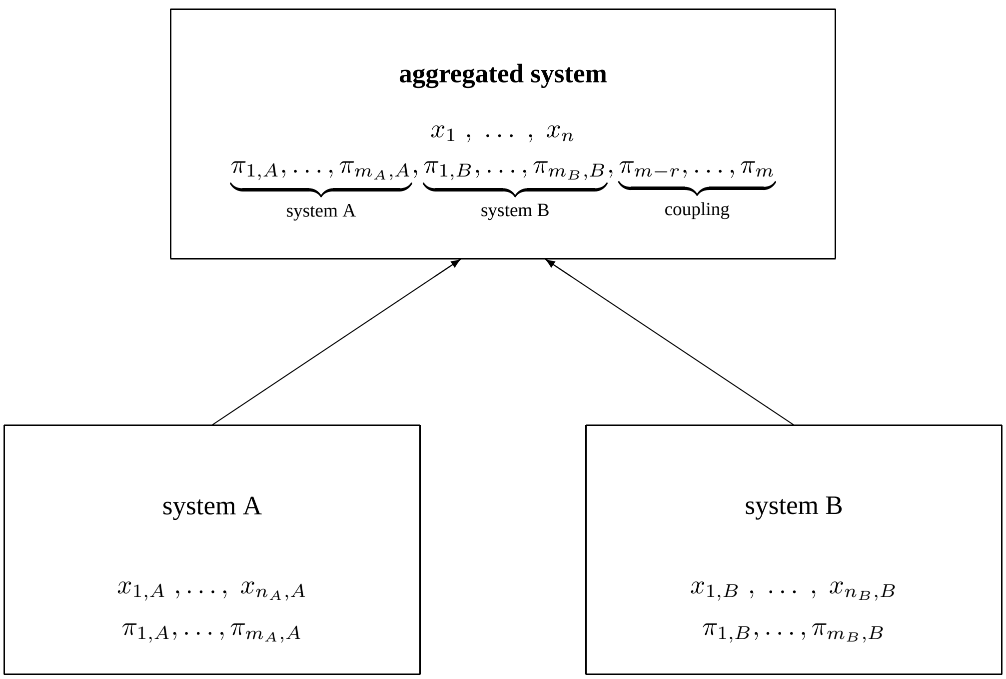

Figure 2. Aggregation of two subsystems with m_A and m_B dimensionless parameters respectively to an aggregated system with $n = n_A + n_B$ dimensional and m dimensionless parameters, including $r = m - (m_A + m_B)$ coupling numbers

Since for a given problem there exists an infinite number of equivalent sets of dimenensionless groups, sets with more than m_c dimensionless numbers consisting of variables of more than one subsystem can be found. However there cannot exist a set of dimensionless groups with less than m_c coupling numbers. Only when m_c coupling numbers are in the set of dimensionless groups, all the dimensionless groups from the subsystems are preserved in the aggregated system. This property is advantageous in data mining and knowledge discovery applications.

4. APPLICATION TO KNOWLEDGE DISCOVERY

If for engineering data the dimensional information is stored in the database, dimensional analysis can be performed automatically. Modeling simple systems (i.e. subsystems) using similarity parameters determined by dimensional analysis allows to build a library of such simple systems which can be modeled using data from different experiments. This information can then be used anytime a new complex system is built aggregating these subsystems.

5. EXAMPLE

To illustrate the hierarchical modeling using dimensional analysis the example of an aerospace gas turbine consisting of the three subsystems compressor "c", burner (combustion chamber) "b", and turbine stage "t". These three subsystems can be modeled individually and can then be combined to form the aggregated system of the gas turbine.

5.1. compressor

The first stage of an aerospace gas turbine is the compressor section, usually consisting of an axial compressor. The relevance list for axial compressors (independent of the specific use in an aerospace gas turbine) is given in table 1.

$\dot{m}_C$	mass flow
$c_{p,C}$	specific heat ratio
$T_{1,C}$	compressor inlet temperature
$T_{2,C}$	compressor outlet temperature
$\dot{W}_C$	compressor power

Table 1. Relevance list of the compressor

The dimensional set ist easily established. Here the rank of the submatrix A is 3 although the number of rows is 4. According to Szirtes[2] one row (the one corresponding to length $[L]$ has been chosen here) can therefore be eliminated. The submatrix D ist chosen to be the unity matrix of size 2×2 and the submatrix C is then calculated using equation (4).

	$T_{2,C}$	$\dot{W}_C$	$T_{1,C}$	$c_{p,C}$	$\dot{m}_C$
mass	0	1	0	0	1
time	0	-3	0	-2	-1
temp.	1	0	1	-1	0
$\pi_{1,C}$	1	0	-1	0	0
$\pi_{2,C}$	0	1	-1	-1	-1

Table 2. Dimensional set of the compressor

The number of parameters in the relevance list (see table 1) is equal to 5 and the rank of the dimensional matrix (see table 2) is equal to 3. Therefore $5 - 3 = 2$ dimensionless groups $\pi_{1,C}$ and $\pi_{2,C}$ for the compressor can be determined from the above dimensional set.

$$\pi_{1,C} = \frac{T_{2,C}}{T_{1,C}} \tag{7}$$

$$\pi_{2,C} = \frac{\dot{W}_C}{\dot{m}_C \; c_{p,C} \; T_{1,C}} \tag{8}$$

The functional relationship between these two dimensionless parameters $\pi_{1,C}$ and $\pi_{2,C}$ can easily be found using data mining algorithms. Instead of modeling a functional relationship between 5 dimensional variables, only the functional relationship between the 2 dimensionless parameters has to be identified.

In the case of reversible adiabatic compression of a fluid, the equation $\dot{W}_C = \dot{m}_C \cdot c_{p,C} \cdot (T_{2,C} - T_{1,C})$ holds and a data mining algorithm must find the relationship

$$\pi_{1,C} = \pi_{2,C} + 1 \tag{9}$$

which can then be transformed back into physical variables applying the inverse pi-transform to equations (7) and (8) and yields the well-know thermodynamic relationship between 5 variables

$$\begin{aligned} \frac{T_{2,C}}{T_{1,C}} &= \frac{\dot{W}_C}{\dot{m}_C\ c_{p,C}\ T_{1,C}} + 1 \\ \dot{W}_C &= \dot{m}_C \cdot c_{p,C} \cdot (T_{2,C} - T_{1,C}) \end{aligned} \tag{10}$$

5.2. combustion chamber

Burning fuel in the combustion chamber converts chemical energy into thermal energy in the fluid.

$\dot{m}_B$	mass flow
$c_{p,B}$	specific heat ratio
$T_{2,B}$	chamber inlet temperature
$T_{3,B}$	chamber outlet temperature
$\dot{Q}_F$	thermal power

Table 3. Relevance list of the combustion chamber

	$T_{3,B}$	$\dot{Q}_F$	$T_{2,B}$	$c_{p,B}$	$\dot{m}_B$
mass	0	1	0	0	1
time	0	-3	0	-2	-1
temp.	1	0	1	-1	0
$\pi_{1,B}$	1	0	-1	0	0
$\pi_{2,B}$	0	1	-1	-1	-1

Table 4. Dimensional set of the combustion chamber

The dimensionless parameters for the combustion chamber subsystem are then determined from the dimensional set shown in table 4. The two dimensionless groups $\pi_{1,B}$ and $\pi_{2,B}$ are

$$\pi_{1,B} = \frac{T_{3,B}}{T_{2,B}} \tag{11}$$

$$\pi_{2,B} = \frac{\dot{Q}_F}{\dot{m}_B\ c_{p,B}\ T_{2,B}} \tag{12}$$

For an isobaric heating the energy equation is given by $\dot{Q}_F = \dot{m}_B\ c_{p,B}\ \ (T_{3,B} - T_{2,B})$, and a data mining system must find the dimensionless relation

$$\pi_{1,B} = \pi_{2,B} + 1 \tag{13}$$

Transforming this equation back into physical variables, the thermodynamic equation is found to be

$$\dot{Q}_F = \dot{m}_B \cdot c_{p,B} \cdot (T_{3,B} - T_{2,B}) \tag{14}$$

5.3. turbine stage

The hot gas is expanded in the turbine stage and mechanical shaft energy is gained to drive the compressor and auxiliary units.

$\dot{m}_T$	mass flow
$c_{p,T}$	specific heat ratio
$T_{3,T}$	turbine inlet temperature
$T_{4,T}$	turbine outlet temperature
$\dot{W}_T$	turbine power

Table 5. Relevance list of the turbine subsystem

	$T_{4,T}$	$\dot{W}_T$	$T_{3,T}$	$c_{p,T}$	$\dot{m}_T$
mass	0	1	0	0	1
time	0	-3	0	-2	-1
temp.	1	0	1	-1	0
$\pi_{1,T}$	1	0	-1	0	0
$\pi_{2,T}$	0	1	-1	-1	-1

Table 6. Dimensional set of the turbine subsystem

According to the relevance list (table 5) and the dimensional set (table 6) for the turbine stage, dimensional analysis yields two dimensionless groups $\pi_{1,T}$ and $\pi_{2,T}$.

$$\pi_{1,T} = \frac{T_{4,T}}{T_{3,T}} \tag{15}$$

$$\pi_{2,T} = \frac{\dot{W}_T}{\dot{m}_T\ c_{p,T}\ T_{3,T}} \tag{16}$$

Assuming adiabatic-reversible expansion in the turbine section, the underlying functional relationship that has to be found by data mining algorithms is

$$\pi_{1,T} = \pi_{2,T} + 1 \tag{17}$$

Applying the inverse pi-transform to equations (15) and (16), the expression in physical variables is

$$\dot{W}_T = \dot{m}_T \cdot c_{p,T} \cdot (T_{4,T} - T_{3,T}) \tag{18}$$

5.4. Aggregation of the subsystems

The three subsystems compressor, combustion chamber, and turbine can be aggregated into the overall system of the aerospace gas turbine. The three subsystems have $m_C = 2$, $m_B = 2$, and $m_T = 2$ dimensionless parameters respectively. The aggregated system has $n = n_C + n_B + n_T = 5 + 5 + 5 = 15$ dimensional variables and the dimensional matrix has the rank $r = 3$. Therefore, the aerospace gas turbine has $m = n - r = 15 - 3 = 12$ dimensionless groups. Only $\sum m_i = 6$ dimensionless groups can be built using only variables from the same subsystem, the other $m - \sum m_i = 12 - 6 = 6$ dimensionless groups must be built using dimensional variables from different subsystems. Therefore the minimal number of coupling numbers is 6.

Combining the relevance lists of the three subsystems (table 1, 3, and 5) yields the relevance list of the aerospace gas turbine as shown in table 7.

$\dot{m}_C$	compressor mass flow
$c_{p,C}$	compressor specific heat ratio
$T_{1,C}$	compressor inlet temperature
$T_{2,C}$	compressor outlet temperature
$\dot{W}_C$	compressor power
$\dot{m}_B$	chamber mass flow
$c_{p,B}$	chamber specific heat ratio
$T_{2,B}$	chamber inlet temperature
$T_{3,B}$	chamber outlet temperature
$\dot{Q}_F$	thermal power
$\dot{m}_T$	turbine mass flow
$c_{p,T}$	turbine specific heat ratio
$T_{3,T}$	turbine inlet temperature
$T_{4,T}$	turbine outlet temperature
$\dot{W}_T$	turbine power

Table 7. Relevance list of the gas turbine

Again the rank of the dimensional matrix $P = [B|A]$ is 3 and therefore the length row has been omitted. The submatrix D has been chosen in a special form to accomodate the already known dimensionless groups from the subsystems in the set of dimensionless groups for the aggregated system as shown in table 8.

	$T_{4,T}$	$\dot{W}_T$	$T_{3,T}$	$c_{p,T}$	$\dot{m}_T$	$T_{3,B}$	$\dot{Q}_F$	$T_{2,B}$	$c_{p,B}$	$\dot{m}_B$	$T_{2,C}$	$\dot{W}_C$	$T_{1,C}$	$c_{p,C}$	$\dot{m}_C$
mass	0	1	0	0	1	0	1	0	0	1	0	1	0	0	1
time	0	-3	0	-2	-1	0	-3	0	-2	-1	0	-3	0	-2	-1
temp.	1	0	1	-1	0	1	0	1	-1	0	1	0	1	-1	0
π_1	1	0	-1	0	0	0	0	0	0	0	0	0	0	0	0
π_2	0	1	-1	-1	-1	0	0	0	0	0	0	0	0	0	0
π_3	0	0	1	0	0	0	0	0	0	0	0	0	-1	0	0
π_4	0	0	0	1	0	0	0	0	0	0	0	0	0	-1	0
π_5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	-1
π_6	0	0	0	0	0	1	0	-1	0	0	0	0	0	0	0
π_7	0	0	0	0	0	0	1	-1	-1	-1	0	0	0	0	0
π_8	0	0	0	0	0	0	0	1	0	0	0	0	-1	0	0
π_9	0	0	0	0	0	0	0	0	1	0	0	0	0	-1	0
π_{10}	0	0	0	0	0	0	0	0	0	1	0	0	0	0	-1
π_{11}	0	0	0	0	0	0	0	0	0	0	1	0	-1	0	0
π_{12}	0	0	0	0	0	0	0	0	0	0	0	1	-1	-1	-1

Table 8. Dimensional matrix of the gas turbine, submatrix D is regular (rank(D)=12)

For the given relevance list (see table 7) the above dimensional set (see table 8) yields 12 dimensionless groups for the aerospace gas turbine

$$\begin{aligned} \pi_1 &= \frac{T_{4,T}}{T_{3,T}} = \pi_{T,1} & \pi_2 &= \frac{\dot{W}_T}{\dot{m}_T\ c_{p,T}\ T_{3,T}} = \pi_{T,2} \\ \pi_3 &= \frac{T_{3,T}}{T_{1,C}} & \pi_4 &= \frac{c_{p,T}}{c_{p,C}} \\ \pi_5 &= \frac{\dot{m}_T}{\dot{m}_C} & \pi_6 &= \frac{T_{3,B}}{T_{2,B}} = \pi_{B,1} \\ \pi_7 &= \frac{\dot{Q}_F}{\dot{m}_B\ c_{p,B}\ T_{2,B}} = \pi_{B,2} & \pi_8 &= \frac{T_{2,B}}{T_{1,C}} \\ \pi_9 &= \frac{c_{p,B}}{c_{p,C}} & \pi_{10} &= \frac{\dot{m}_B}{\dot{m}_C} \\ \pi_{11} &= \frac{T_{2,C}}{T_{1,C}} = \pi_{C,1} & \pi_{12} &= \frac{\dot{W}_C}{\dot{m}_C\ c_{p,C}\ T_{1,C}} = \pi_{C,2} \end{aligned} \tag{19)-(30}$$

It is supposed in the following that the thermodynamic process underlying figure 3 is ideally adiabatic, reversible compression ($1 \rightarrow 2$), isobaric heating ($2 \rightarrow 3$), adiabatic-reversible expansion ($3 \rightarrow 4$), and isobaric cooling ($4 \rightarrow 1$). The connections between the subsystems are assumed lossless, i.e. mass flow and the specific heat ratio are equal for all subsystems, and the temperatures of the connected outlets and inlets are equal ($T_{2,C} = T_{2,B}, T_{3,B} = T_{3,T}$). Using data for such a system, data mining / knowledge discovery tools will identify constant parameters:

$$\begin{aligned} \pi_3 &= 1 & \pi_4 &= 1 \\ \pi_5 &= 1 & \pi_8 &= 1 \\ \pi_9 &= 1 & \pi_{10} &= 1 \end{aligned} \tag{31)-(36}$$

The set of dimensionless groups can be reduced by these 6 groups which are constant and do therefore not contribute to the modeling of the gas turbine system under the above stated assumptions.

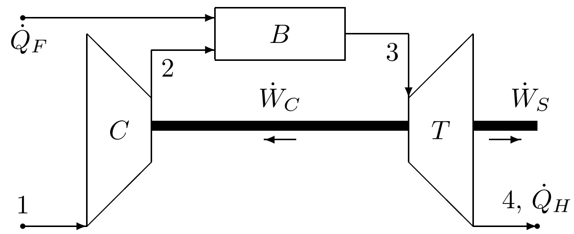

Figure 3. Component scheme of gas turbine[4]

The same result can be found by applying the restriction of lossless connections in the relevance list of the gas turbine and omitting the surplus physical variables, as is common in physical modeling in dimensional variables. The procedure of hierarchical modeling using dimensional analysis however is more general, since no modifications in the subsystems are necessary.

Should the connections between the subsystems not be ideal, then the dimensionless groups in eq. ((31)-(36)) are no longer equal to one, or not even constant any more, but account now for the losses on the connections between the subsystems.

Hierarchical modeling using dimensional analysis allows a formal aggregation of the subsystems without restrictions on the connections between the subsystems. Engineering knowledge about the connections can then be integrated in the assessment of the dimensionless groups of the aggregated system. The coupling numbers found by hierachical modeling account for non-ideal connections between the different subsystems.

6. SUMMARY

It has been shown how hierarchical modeling using dimensional analysis can be used in knowledge discovery in scientific data. In domains where the princple of dimensional homogeneity holds, dimensionless groups (*similarity numbers*) can be generated automatically from the knowledge of the dimensional representation of the database fields. The number of dimensionless groups is less than the number of physical parameters (database fields) and the dimensionless groups form a minimal set for a given problem. Simple (sub-)systems can then be identified more easily using data mining algorithms. Hierarchical modeling using dimensional analysis then allows to aggregate the known subsystems to form more complex systems, where only the couplings between the subsystems have to be newly identified using data mining algorithms. The procedure of hierarchical modeling using dimensional analysis allows to built a database of identified systems, update these system models when new data is available and use these simple system models to form complex systems.

REFERENCES

1. E. Buckingham, “The principle of similitude,” *Nature* **96**, pp. 396–397, 1915.
2. T. Szirtes, *Dimensional Analysis*, McGraw-Hill, 1998.
3. S. Brückner and S. Rudolph, “Hierarchical modeling using dimensional analysis,” in *3rd International Workshop on Similarity Methods 2000*, B. Kröplin, S. Rudolph, and S. Brückner, eds., pp. 93–110, Institute for Statics and Dynamics of Aerospace Structures, Universität Stuttgart, (Stuttgart), November 2000.
4. A. Frohn, *Einführung in die Technische Thermodynamik*, AULA-Verlag, Wiesbaden, 1989.