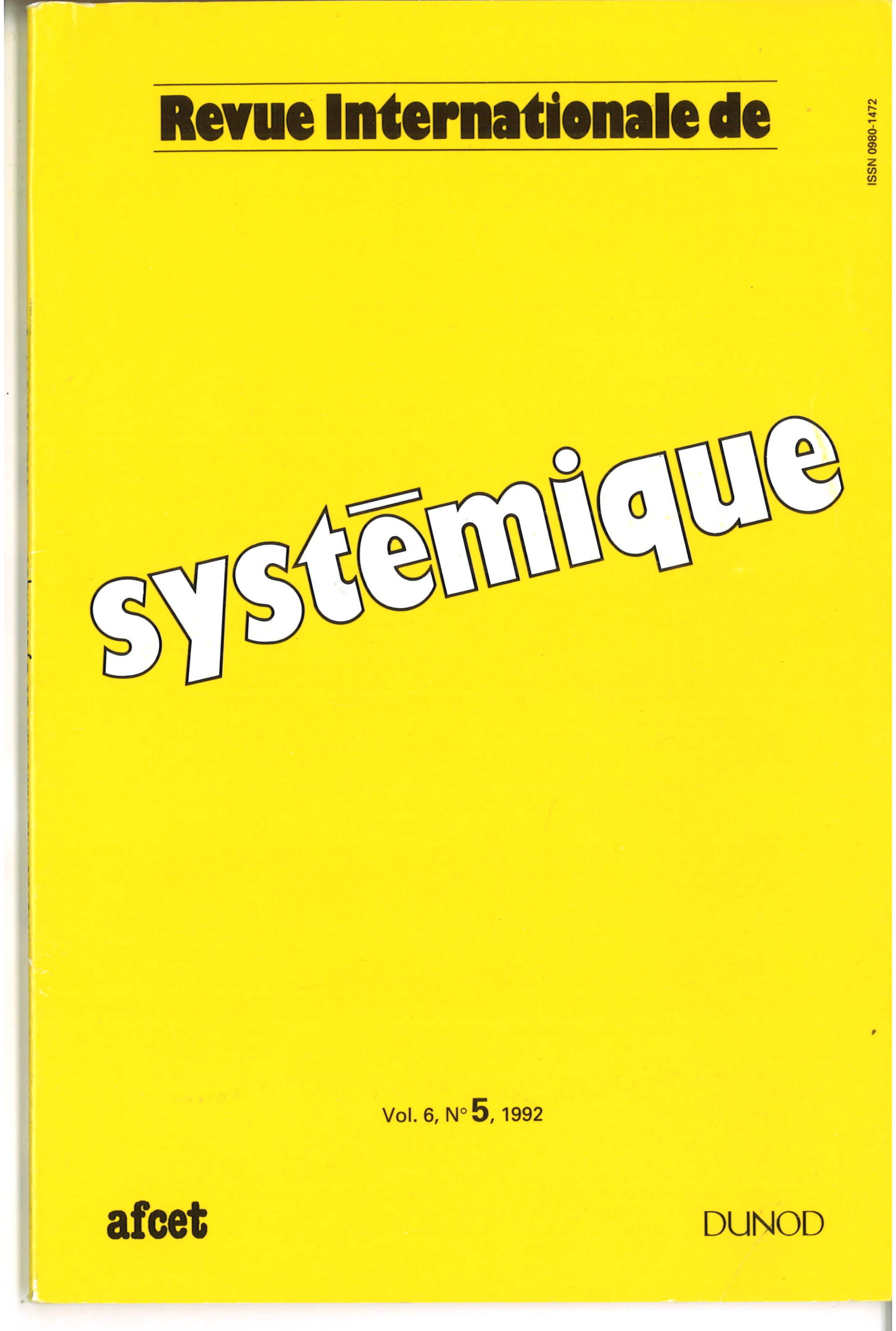

Revue Internationale de

systēmique

ISSN 0980-1472

Vol. 6, N° 5, 1992

afcet

DUNOD

AFSCET

Revue Internationale de systémique

Revue Internationale de Sytémique

volume 06, numéro 5, pages 465 - 481, 1992

An application of approximate reasoning to chemical knowledge

Matthais Otto, Ronald R. Yager

Revue Internationale de systémique

Rédacteur en chef : B. Paulré
Rédacteur en chef adjoint : E. Andreewsky

Revue Internationale de Systémique is published 5 times a year: March, May, July, September, December. Date of issue: January 1993.
Subscription price, per volume: Institutions US $ 185.
Second-class postage paid at Rahway, N.J. ISSN N° 0980-1472, USPS N° 007728.
U.S. Mailing Agent: Mercury Air-freight Intnl. Ltd., 2323 Randolph Ave., Avenel, NJ07001.
Published by Dunod, 15, rue Gossin, 92543 Montrouge Cedex France and Gauthier-Villars North America Inc., 875-81 Massachusetts Avenue, Cambridge, MA 02139, USA.
Postmaster. Please send all address corrections to: Dunod, c/o Mercury Air-freight Intnk. Ltd. 2323 Randolph Ave., Avenel, NJ 07001, USA.

REVUE INTERNATIONALE DE SYSTÉMIQUE
Vol. 6, N° 5, 1992, pp. 465 à 481

AN APPLICATION OF APPROXIMATE REASONING TO CHEMICAL KNOWLEDGE[1]

Matthais OTTO

Department of Chemistry Bergakademie Freiberg[1]

Ronald R. YAGER

Machine Intelligence Institute Iona College[2]

Abstract

We are concerned here with the problem of estimating the properties chemical compounds based upon knowledge of similar compounds which may exist in a chemical database. In addition we allow for inclusion of general knowledge of chemistry which may exist in the form of rules about typical situations. A characteristic of this problem is the fact that much of our knowledge is imprecise and unspecific. We suggest a methodology for addressing this estimation problem based upon the theory of approximate reasoning. This approach allows us to deal with the inherent imprecision, the problem of partial matching as well as being able function for numeric and nonnumeric properties. The approach essentially serves as the foundation for an intelligent database.

Résumé

On s'intéresse ici au problème de l'estimation de propriétés chimiques à partir de la connaissance de similarités entre des éléments qui peuvent exister dans une base de données de chimie. De plus, on prend en compte des connaissances générales de chimie qui peuvent être formulées sous forme de règles concernant des situations typiques. Une caractéristique de ce problème réside dans le fait que la majeure partie des connaissances

[1] 9200, Freiberg, Leipziger Str. 29, Germany.
[2] New Rochelle, NY 10801, USA.

Rev. intern. systémique. 0980-1472 Vol. 6/92/05/ 465 /17/$ 3.70/© Afcet Gauthier-Villars

est imprécise et non spécifique. On suggère une méthode fondée sur la théorie du raisonnement approximatif pour résoudre ce problème d'estimation. Cette approche nous permet de traiter l'imprécision inhérente ainsi que de prendre en compte des propriétés numériques aussi bien que non numériques. L'étude sert essentiellement à l'élaboration d'une base de données intelligente.

Introduction

In general, the representation and manipulation of chemical knowledge can be approached either in a numerical and algorithmic way or symbolically by means of logic and other artificial intelligence (AI) tools. In spite of the fact that quite powerful methods have been developed in recent years to estimate chemical reactivity, to model chemical strucutre-activity relationships, to design chemical synthesis or to predict properties of chemical compounds [1, 2] it is agreed among chemists that not all the chemistry can be converted into numbers, mathematical models and numerical dependencies. Therefore, the increasing use of AI-tools for storing and handling chemical knowledge can be envisaged as a natural step to help further integrate the computer in the chemical laboratory.

In the present work we address two objectives of working with chemical databases:

i) Predicting missing physical and chemical properties of compounds by taking into account available information in the database, and

ii) reasoning about possible properties by considering commonsense chemical knowledge. In this paper the theory of approximate reasoning is applied in a manner that enables crisp and/or uncertain and vague (fuzzy) knowledge to be manipulated. Examples are based upon a database of organic indicator dyes.

1. An introduction to the theory of approximate reasoning

The theory of approximate reasoning (AR), which was originally introduced by L. A. Zadeh [3, 4], provides a systematic methodology for reasoning with uncertain and imprecise information. In [5] Dubois & Prade provide a comprehensive survey of this area. AR makes intensive use of fuzzy subsets [4, 6]. In this section we shall provide a brief introduce to the theory of approximate reasoning.

Assume X is a set, fuzzy subset A of X is a subset that can have partial membership. This facility for partial membership allows us to capture concepts which have noncrisp boundaries. For example, if one tries to represent the concept of *tall* there exists some heights which are not completely tall but neither are they not tall. In [6] Zadeh first introduces the concept of a fuzzy subset and provides substantial justification for the necessity of such an object. Formally a fuzzy subset A has associated with it a membership function U_A such that for each $x \in X$, $U_A(x) \in [0, 1]$ indicates the degree to which x is a member of the set A. We shall take the liberty of using $A(x)$ instead of $U_A(x)$. A semantics that can be associated with $A(x)$ is that it is the degree to which the element x is compatible with the concept A. Implicit in this view is the idea of similarity, how similar x is to the ideal of the concept A. Ruspini [7, 8] has explored in considerable detail the connection between similarity and fuzzy set membership.

We say that a fuzzy subset A is normal if their exists at least one element x such that $A(x) = 1$, normality is closely related to nonullness in ordinary set theory. The crisp subset B of X consisting of all the elements for which $A(x) > 0$ is called the support of A.

In [6] Zadeh provides for the extension of the usual set operations to fuzzy subsets. Assume A and B are two fuzzy subsets of X. The intersection (conjunction) of these two subsets, denoted, $D = A \cap B$, is also a fuzzy subset of X which has as its membership function

$$D(x) = \mathrm{Min}[A(x), B(x)].$$

The set D can be seen as the elements which satisfy both A and B. It should be noted that the definition for conjunction as all the other definitions we shall provide are not unique, although they are the standard ones. In [9] Yager provides a comprehensive discussion of alternative fuzzy set operations.

The union(disjunction) of fuzzy subsets is a fuzzy subset F of X, denoted $F = A \cup B$ which has membership function

$$F(x) = \mathrm{Max}[A(x), B(x)].$$

The set F can be seen as the fuzzy subset of elements that satisfy either A or B both.

The negation of A, $\bar{A}$, is also a fuzzy subset of X defined such that

$$\bar{A}(x) = 1 - A(x).$$

If A and B are two fuzzy subsets of X, we say that A is contained in B, denoted $A \subset B$, if $A(x) \leq B(x)$ for all $x \in X$.

We recall that if X and Y are two sets their Cartesian product, $X \times Y$, is a set whose elements all are pairs (x, y) such $x \in X$ and $y \in Y$. If A and B are fuzzy subsets of X and Y respectively then their cartesian product, denoted $E = A \times B$, is a fuzzy subset of the space $X \times Y$ defined such that for each $x \in X$ and $y \in Y$, $E(x, y) = \mathrm{Min}[A(x), B(x)]$. More generally if $X_1, ..., X_n$ are sets and $A_1, ..., A_n$ are fuzzy subsets of $X_1, ..., X_n$ respectively then $X_1 \times X_2, ..., \times X_n$ is a set whose elements are all the n-tuples $(x_1, ... x_n)$ which $x_i \in X_i$. Furthermore, the cartesian product, $E = A_1 \times A_2, ... \times A_n$ is a fuzzy subset of $X_1 \times X_2, ..., \times X_n$ such that for each $(x_1, ..., x_n), x_i \in X_i$ it is the case $E(x_1, .., x_n) = \underset{i}{\mathrm{Min}}[A_i(x_i)]$.

Having introduced the basic machinery from fuzzy set theory we now turn to the theory of approximate reasoning. At the heart of the use of fuzzy sets in the theory of approximate reasoning is its ability to provide a semantics for the meaning of words used in natural language.

One very common way of defining the meaning of a concept is by extension. Essentially extension means that one illustrates the meaning of a concept by pointing to examples of the concept. Thus if I want to define what I mean by female I point to all the female objects in a particular class. If I want to describe the concept tall, in a given context, I list all the heights that I consider tall. A set provides a natural structure for defining the meaning of words or concepts by extension. Fuzzy subsets further enhance this ability by allowing for a more subtle definition of concepts where there is some grayness in the membership. For example, if I want to define red I can allow a very natural boundary from red to not red.

We are now in a position to provide a basic introduction to the theory of approximate reasoning(AR). Those interested in more detail can find it in the copious literature in the field [5, 10]. The basic elements used in AR are variables, which we denote as $V_1, ..., V_n$. A variable is generally associated with a property or attribute of some object. A variable could be a person's age, the temperature of a chemical process, or the truth of a proposition. Associated with each variable V_i is a set X_i called its base set or universe of discourse. The base set of a variable indicates the set of allowable values for that variable. If V is the variable age then its base set would be the set of integers between 0 and 120.

Information is conveyed in AR by propositions about the variables. Assume V_i is a variable whose base set is X_i and let A_i be a fuzzy subset of X_i an **atomic or canonical proposition** in AR is a statement of the form

$$V_i \text{ is } A_i.$$

The spirit of the above proposition is to indicate that the actual value of V_i lies in set A_i. More formally the above statement is seen to induce a possibility distribution with regards to the value of V_i where $A_i\,(x)$ is the possibility that V_i assumes the value x.

It should be noted that in the special case where A_i is a singleton $A_i = \{x^*\}$, then our proposition is reflecting the fact that $V_i = x^*$. In the other special case where $A_i = X_i$ our proposition is providing us with no information about V_i.

Assume $V_1, V_2, ..., V_n$ are a collection of variables with base sets $X_1, ..., X_n$ we indicate any subset collection of these variables as a joint variable, *ie.* (V_1, V_2, V_3), $(V_1, ..., V_n)$, (V_1, V_5, V_6).

Assume $(V_1, ..., V_n)$, is a joint variable then a proposition is a statement of the form

$$(V_1, ..., V_n) \text{ is } R$$

where R is a fuzzy subset of the cartesian product space, $X_1 \times X_2, ..., \times X_n$, the cartesian product of the base sets of the elements making up the joint variable V. The intention of such a joint proposition is to indicate that $R(x_1, ..., x_n)$ is the possibility of the joint solution $V_1 = x_1$ and $V_2 = x_2$ and, ..., $V_n = x_n$.

The reasoning process in AR is made up of a three step operation:

1. **Translation**
2. **Conjunction**
3. **Projection**

In the translation phase of the reasoning process we try to represent the information and knowledge we have available in terms of propositions of the kind previously described. While we shall just touch upon the various types of rules for translation from natural language to AR propositions it should be strongly emphasized that this is one of the most powerful features of AR, its ability to represent various pieces of knowledge in a unified format.

Conjunction consists of the process of fusing the individual pieces of information to obtain their combined effect. Essentially the idea here is that each proposition essentially restricts the allowable values of its constituent variables. Thus two or more propositions are combined by requiring that the constituent variables satisfy **all** the constraints imposed by the individual propositions.

The third step in the reasoning process is the projection process. This process allows us to obtain the value of any variable from a statement about a joint variable.

In describing the formal mechanism of AR we need first provide a description of the projection operation.

Definition: Assume V is a joint variable $V_1, \ldots, V_n$ and we have a proposition

$$(V_1, \ldots, V_n) \text{ is } R$$

where R is a fuzzy subset on the cartesian product of base sets of R, X. The projection of V on V_i is defined as the proposition

$$V_i \text{ is } A_i$$

where for each $x \in X$

$$A_i(x) = \underset{\substack{\text{over all} \\ (x_i,\, x_n) \\ \text{such that} \\ x_i = x}}{\mathrm{Max}} R(x_1, x_2, \ldots, x_n)$$

Using the definition of projection we define the first inference rule of AR, **the projection principle**: *From a proposition* ***(V_1, ...V_n) is R*** *we can always infer* ***V_i is A_i*** *where A_i is the projection on R onto X_i.*

Example: Assume (V_1, V_2) is R where

$$R = \left\{\frac{.7}{(a,\, 1)}, \frac{.9}{(a,\, 2)}, \frac{1}{(a,\, 3)}, \frac{.2}{(b,\, 1)}, \frac{.1}{b,\, 2}, \frac{.5}{(b,\, 3)}\right\}.$$

From this we can infer V_1 is $\left\{\frac{1}{a}, \frac{.5}{b}\right\}$ and V_2 is $\left\{\frac{.7}{1}, \frac{.9}{2}, \frac{1}{3}\right\}$.

It should be carefully noted that a proposition such as *V is A* indicates that the value of V must lie in the set A. From this it is easy to see that if we can infer that V also must lie in B where $A \subset B$. This observation forms the basis of the second important inference rule in AR the **entailment principle** [3]: *Given a proposition* ***V is A*** *we can infer the proposition* ***V is B*** *where $A \subset B$.* This principle is a very powerful inference mechanism in AR.

Two important concepts in AR are those of possibility and certainty. Assume *V is A* and *V is B* are two propositions where A and B are both fuzzy subsets of X. The **possibility** of the validity of the proposition *V is B* given the validity of the proposition *V is A*, is denoted, Poss[V is B|V is A] and is defined as

$$\text{Poss[V is B|V is A]} = \mathrm{Max}_x[D(x)]$$

where $D = A \cap B$ $(D(x) = A(x) \wedge B(x))$. The measure essentially measures the degree of intersection of the two propositions.

The **certainty** of the validity of the proposition *V is B* given the validity of the proposition *V is A*, is denoted, Cert[V is B|V is A] and is defined as

$$\text{Cert[V is B|V is A]} = 1 - \text{Poss[V is } \bar{B}\text{|V is A]}.$$

The measure of certainty essentially measures the degree to which A is contained in B and as such measures the degree to which the validity of A implies B.

One prototypical reasoning situation in AR consists of a case in which we have two propositions. The first proposition involves a relationship between two variables. The second proposition provides information about one of the variables. The object of this situation is to obtain information about the second variable. Let V and U be two variables with base sets X and Y. Assume the first proposition P_1 is *(U, V) is R*, where R is a fuzzy subset of X and Y. Thus P_1 provides a relationship between U and V where $R(x, y)$ indicates that possibility of the pair $U = x$ and $V = y$ occurring. The second proposition P_2 states *U is E*. The procedure used to obtain information about V from these two pieces of information is as follows:

1. Conjunct P_1 and P_2 giving us the proposition *(U, V) is H* where

$$H(x, y) = \mathrm{Min}[E(x), R(x, y)].$$

2. The second step is to project H onto V, giving us *V is E* where $E(y) = \mathrm{Max}[H(x, y)]$.

We have implicitly assumed that the form of the relationship R is already given. The process of obtaining these structures, the relationship between variables, is central to the representational step in the reasoning process. A considerable body of literature has been devoted to the issue of knowledge representation using approximate reasoning formalism.

One common type of relationship between U and V is a conditional relationship,

if U is A then V is B.

As suggested by Zadeh [3] one form for translating this statement is *(U, V) is R* where

$$R(x, y) = \bar{A}(x) \vee B(y).$$

If we use this form in the above we get

$$H(x, y) = (\bar{A}(x) \vee B(y) \wedge E(x) = (\bar{A}(x) \wedge E(x) \vee (B(y) \wedge E(x).$$

If we project H onto V we get *V is F* where

$$F(y) = \text{Poss}[\bar{A}|E] \vee B(y).$$

We see that if $\text{Poss}[\bar{A}|E] = 0$ then $F(y) = B(y)$ and the desired inference is made. In this case we note that $\text{Poss}[\bar{A}|E] = 0$ implies that $\text{Cert}[\bar{A}|E] = 1 - \text{Poss}[\bar{A}|E] = 1$. Thus in this case the proposition *V is E* assures us that A is true therefore allows us to make the appropriate conclusion, U is B.

Let us now consider the other extreme, $\mathrm{Poss}[\bar{A}|E] = 1$. In this case we get $F(y) = 1 \vee E(y) = 1$ and therefore *V is X*. Thus in this case we see that we infer that everything is possible and thus have been supplied with no information. We note that $\mathrm{Poss}[\bar{A}|E] = 1$ indicates that $\bar{A}$ is possible and therefore provides no assurance that V is A is satisfied.

In situations where $0 < \mathrm{Poss}[\bar{A}|B] < 1$, we get some deformed version of E

$$F(y) = \alpha \vee E(y)$$

where $\alpha = \mathrm{Poss}[\bar{A}|B]$.

The formalism just introduced easily extends to the case in which we have multiple antecedents. Consider

$$P_0\text{: if } U_1 \text{ is } A_1 \text{ and } U_2 \text{ is } A_2, \ldots, \text{ and } U_n \text{ is } A_n \text{ then } U \text{ is } B.$$

$$P_1\text{: } U_1 \text{ is } E_1$$

$$P_1\text{: } U_2 \text{ is } E_2$$

$$P_n\text{: } U_n \text{ is } E_n.$$

In this case it can be shown that we can infer *V is F* where

$$F(y) = \mathrm{Poss}[\bar{A}_1|E_1] \vee \mathrm{Poss}[\bar{A}_2|E_2] \vee \ldots \mathrm{Poss}[\bar{A}_n|E_n] \vee B.$$

Thus we see that if any of $\mathrm{Poss}[\bar{A}_i|E_i] = 1$ the rule doesn't fire.

2. Inferring chemical properties by approximate reasoning

Properties of chemical compounds are characterized by numerical attributes, such as "bonding energy" or "boiling point", by linguistic variables, such as "color" or "solubility" or by describing their reactivity by qualitative concepts, such as "the theory of hard and soft acids and bases" or "inductive" and "mesomeric" effects.

In the past most of effort has been put into attempts to quantify these properties by numerical values and to reason about them, in an algorithmic manner.

Using the methods of the theory of approximate reasoning there is no need to restrict chemical knowledge manipulation to numerical approaches since linguistic quantifiers and qualitative concepts can be handled either symbolically [11] or by representing these variables as fuzzy sets in a mathematically exact way.

Consider chemical compounds with attributes V_1, V_2, U taking their values respectively in the sets X_1, X_2, Y. If A, A_1, A_2, ..., A_n and B, B_1, B_2, ..., B_n and D_1, D_2, ... D_n are subsets in X_1, X_2, Y, respectively, then the chemical database is:

COMPOUND NO	V_1	V_2	U
0	A	B	?
1	A_1	B_1	D_1
2	A_2	B_2	D_2
n	A_n	B_n	D_n

(1)

In the above for compound number 0 only the attribute values for V_1 and V_2, A and B, are known while the value for attribute U is unknown and is to be estimated. We shall denote this unknown as D.

We can represent the information in the above database in terms of production rules as follows.

$$\begin{aligned} &\text{if } V_1 \text{ is } A_1 \text{ and } V_2 \text{ is } B_1 \text{ then U is } D_1\\ &\text{if } V_1 \text{ is } A_2 \text{ and } V_2 \text{ is } B_2 \text{ then U is } D_2\\ &\text{if } V_1 \text{ is } A_n \text{ and } V_2 \text{ is } B_n \text{ then U is } D_n \end{aligned} \tag{2}$$

In addition, we have the knowledge about compound 0 in the form of the following propositions

$$\begin{aligned} &V_1 \text{ is A}\\ &V_2 \text{ is B} \end{aligned} \tag{3}$$

Our objective is to use the above information to obtain the missing U value for compound 0 denoted by D. Based upon the theory each of the above production rules induces a relation H_i on Y such that

$$\begin{aligned} H_1(y) &= \mathrm{Poss}(\bar{A}_1/A) \vee \mathrm{Poss}(\bar{B}_1/B) \vee D_1(y)\\ H_2(y) &= \mathrm{Poss}(\bar{A}_2/A) \vee \mathrm{Poss}(\bar{B}_2/B) \vee D_2(y)\\ H_n(y) &= \mathrm{Poss}(\bar{A}_n/A) \vee \mathrm{Poss}(\bar{B}_n/B) \vee D_1(y) \end{aligned} \tag{4}$$

where $\bar{F}(x) = 1 - F(x)$, $\mathrm{Poss}(F/E) = \mathrm{Max}_x\,[F(x) \wedge E(x)]$ and $\wedge$= min, $\vee$= max.

The individual $H_i(y)$ are combined by *anding* them according to obtain the inferred value D for U.

$$D(y) = H_1(y) \wedge H_2(y) \wedge \ldots \wedge H_n(y) \tag{5}$$

To understand the inference pattern expressed in equation (4) we discuss it in a crisp sense in some more detail.

If a term $\mathrm{Poss}(\bar{A}_i/A) \vee \mathrm{Poss}(\bar{B}_i/B)$ evaluates to 1 then $H_i(y)$ also evaluates to 1 and hence provides no contribution to the determination of the desired value D. On the other hand if the term $\mathrm{Poss}(\bar{A}_i/A) \vee \mathrm{Poss}(\bar{B}_i/B)$ evaluates to zero then $H_i(y)$ evaluates to $D_i(y)$ and provides a contribution to the determination of D(y). The term Poss($\bar{A}_i$/A)$\vee$Poss($\bar{B}_i$/B) evaluates to one if either of the components evaluates to one. Let us look in more detail at a term of the form Poss($\bar{A}_i$/A). Poss ($\bar{A}_i$/A) = 1 if $\bar{A}_i \cap A \neq \Phi$. In this case our information, A, is such that we are not certain that the antecedent condition A_i is satisfied, therefore our rule doesn't fire. On the other hand Poss($\bar{A}_i$/A) = 0 if $\bar{A}_i \cap A = \Phi$. In this case our information, A, is such that $A \subset A_i$ and we are certain that our antecedent condition is satisfied thus **not** preventing the rule from firing. The same analysis holds for B and B_i.

As an example of the application of this technique we refer to the chemical data of Table 1.

In order to evaluate the performance of the method, colors of different indicators are to be estimated using known values of the maximum absorbance coefficient, ϵ_{max}, and the wavelength of maximum absorbance, λ_{max}. (The "solubility in water" and the negative logarithm of the dissociation constants, pK-values, are not considered within the frame of this task since it is well known that they do not influence the color of an indicator dye).

We note that the precision for measuring the ϵ_{max}-value is ± 20% relative to the actual value and that maximum deviations for the wavelength, λ_{max}, can be taken as ± 20 nm. In the light of this the values for ϵ_{max} and λ_{max} are taken as approximate values which we shall characterize at first, by the so called hard window approach, i.e. a membership value of 1 is assigned to values within the uncertainty range of the given values and a membership value of 0 is assigned elsewhere.

Table 1. *Part of a database of pH-indicators*

	Compound	**Solubility in Water**	**pK**	ϵ_{max}	λ_{max}	**Color**
1	Bromocresol green	more or less high	4.66	35040	616.5	blue
2	Bromothymol blue	very low	7.10	32400	616.5	blue
3	Thymol blue	very low	8.90	4224	597.5	blue
4	Phenol red	low	7.81	37740	558.7	red
5	Bromocresol purple	more or less low	6.12	63650	590	purple
6	Bromophenol blue	low	3.85	67840	590	purple
7	m-Cresol purple	low	8.3	9560	580	purple
8	Cresol red	low	8.25	24378	572	purple
9	Xylenol blue	low	8.8	16000	595	blue
10	Chlorophenol red	very low	5.6	23280	575	orange
11	Bromocresol green	more or less low	4.66	16370	438	yellow
12	Bromothymol blue	very low	7.10	16990	431	yellow
13	Thymol blue	very low	8.90	2007	433.7	yellow
14	Phenol red	low	7.81	16640	432	yellow

As a first example the color of bromophenol blue (no. 6, Table 1) is estimated by the reasoning scheme of equations (4) and (5). Table 2 reveals the results of applying (4) (ϵ_{max} is used as variable V_1 and λ_{max} for V_2).

Table 2. *Results for estimating the color (spectrum) of bromophenol blue (compound no. 6 in Table 1)*

Compound ♯	Poss($\bar{A}_i$/A)	Poss($\bar{B}_i$/B)	H_i
1	1 (1)[2]	1 (1)	1
2	1 (1)	1 (1)	1
3	1 (1)	0 (0.141)	1
4	1 (1)	1 (1)	1
5	0 (0.095)	0 (0) purple	(max [0.095, spectrum 5])
7	1 (1)	0 (0.25)	1
8	1 (1)	0 (0.81)	1
9	1 (1)	0 (0.063)	1
10	1 (1)	0 (0.563)	1
11	1 (1)	1 (1)	1
12	1 (1)	1 (1)	1
13	1 (1)	1 (1)	1
14	1 (1)	1 (1)	1

By applying equation. (5) we get as the final result D = {purple}.

In our database the color is not only given by a verbal expression but can be considered a linguistic variable characterized by a fuzzy set. This fuzzy set is based on the measurable visible spectrum as the membership function renormed to the interval [0, 1]. Thus the approximate reasoning scheme allows not only the retrieval of the most similar color name but can be also used to derive the most possible electronic spectrum.

Reasoning at H_i and D is then carried out at every position of the spectrum, Y, *i.e.* $F_i(y)$ and D(y).

If instead of using the crisp interval values to characterize our measurement uncertainty we describe the uncertainty in measuring the ϵ_{max}-values with a fuzzy set, *i.e.* "about ± 20%" relative to the given ϵ_{max}-value, the following membership function can be assigned

$$m(x_1) = [1 - c \mid x_1 - a \mid {}^2]^+ \qquad (6)$$

where x_1 stands for ϵ_{max}; the constant c renorms the membership function to the interval [0, 1] and is set to 1/(0.2 * 0.2); a represents the x_1 – value with

[2] Data in parenthesis refer to the fuzzy case

a membership value of 1 and the + sign denotes truncation of membership values to 0 at negative values. Impreciseness for the wavelength, $\lambda_{\max}$, is taken as "about ± 20 nm" expressed by the same type of membership function as in equation (6) with a constant $c = 1/(20 * 20)$.

The results of reasoning about the sought spectrum gives the bracketed values in Table 2. The inferred spectrum for D compares very well with the real spectrum of compound 6.

Searching for the spectrum of the alkaline form of bromocresol green (no. 1 in Table 1) also retrieves a quite similar spectrum compared to the real spectrum.

If one attempts to derive spectral information about the acidic forms of the indicators (compounds 11-14) two alternatives are found for phenol red by reasoning with the H_i's (cf. Table 3).

Table 3. *Estimation of the spectrum of phenol red in its acidic form (compound no. 14 in Table 1)*

Compound ♯	$\mathrm{Poss}(\bar{A}_i/A)$	$\mathrm{Poss}(\bar{B}_i)$	D_i
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
7	1	1	1
8	1	1	1
9	0.037	1	1
10	1	1	1
11	0.00	.09	max [0.09, spectrum 11]
12	0.01	0.0025	max [0.011, spectrum 12]
13	1	0.0072	1

Combining these alternatives by the minimum operator (equation [5]) reveals a spectrum that again is highly similar to the spectrum expected for phenol red.

The reasoning method works well in all cases where similar compound objects are present in the database. It may happen, however, that an attribute is to be estimated at a position in the data space where no neighboring compounds are available. As a consequence the inferred answer will be not too close the real values and therefore, it is recommended to apply in such cases an interpolation method based on fuzzy arithmetic as described elsewhere [12].

Sometimes the estimation of unknown properties may be imposible because no correlation (dependency) does exist between the attributes. This is the situation, e.g. for the pK-value in the small database of Table 1. The acid dissociation constants of the indicators can neither be correlated with optical properties of the compound nor with their solubility in water. Applying the reasoning mechanisms according to equation (4) and (5) would always infer 1 as the result for the sought attribute, i.e. nothing could be inferred. In such situations additional chemical knowledge should be available in the database to enable the attributes to be estimated by default reasoning.

3. Default reasoning about chemical properties

As an example we consider the estimation of the acid dissociation constants, i.e. the pK-values.

A chemists' reasoning would be as follows. All the compounds possess the same functional acidic group, i.e. the phenol group. Although the pK-value of pure phenol is exactly known to be 9.89 the dissociation constants of phenols in general may range between pK-values of 3.5 and 11.5 depending on the chemical environment, especially the number and kind of substituents and the degree of conjugation in the molecule. If additional conjugation of the phenolate anion occurs then the pK-value is expected to drop by about two pK-units. Substituents will further change this value mainly in dependence on their inductive effect and they either will increase the pK-value (higher degree of dissociation) if they push electron density to the phenolate ion (+ I effect) or they will decrease the pK-value if they withdraw electron density (– I effect).

Let us try to convert this knowledge into the approximate reasoning scheme. Let V be a variable indicating the type of substituent. This variable takes its values in the set of all acidic groups

$$X = \{-\text{phenol}, -\text{COOH}, -\text{OH}, -\text{SO}_3\text{H}, -\text{NR}^+{}_3\ldots\}.$$

The variable to be reasoned about, the pK value, will be denoted as U, takes its value in the set of all pK-values, $Y = [3.5, 11.5]$. Consider the rule

$$\text{If V is } A_1 \text{ then U is } D_1 \tag{7}$$

The antecedent condition V is the "phenol" group is represented by V is A_1, where $A_1 = \{\text{phenol}\}$, $A_2 = \{-\text{COOH}\}$ etc.

Given the data V is C the inferred value of U, U is F, is in analogy to equation (4)

$$F(y) = \mathrm{Poss}(\bar{A}_1/C) \vee D_1(y) \tag{8}$$

In the case where C = {phenol} the inferred value is $F(y) = D_1(y)$ for every y because the set $\bar{A}_1$ being "not phenol" causes $\mathrm{Poss}(\bar{A}_1/C) = 0$ and therefore $\max\,[0, D_1(y)] = D_1(y)$. On the other hand if $C \neq \{\text{phenol}\}$ we infer $F(y) = 1$, i.e. nothing can be inferred about U.

The inferred value for $D_1(y)$ lies in the range between 3.5 and 11.5 pK-values, Y, as already mentioned above. Therefore, the additional rules should be considered in order to sharpen the up to now very unspecific result.

Additional conjugation in the molecule:

Reasoning about additional conjugation of the phenol group can be performed by:

$$\text{If W is B when U is E} \qquad (9)$$

W is a variable indicating the appearance of additional conjugates that takes its values in the set $X' = \{1,0\}$ where 1 indicates an "additional conjugation" and 0 indicates "no additional conjugation in the above $B = \{1\}$. The possibility distribution for the set E is

$$E = \{\text{about 2 pK-units less than for the unconjugated phenols}\}.$$

With the data W is C the inferred value for U becomes U is G where

$$G(z) = \mathrm{Poss}(\bar{B}/C) \vee E(z) \qquad (10)$$

Considering the condition a default condition, in the sense that the effect of conjugates is only typical, the general reasoning scheme can be formulated as suggested [13]:

$$H_0 = \mathrm{Poss}(\bar{B}/C) \vee ((1 - \mathrm{Poss}(E/F) \vee E) \wedge F \qquad (11)$$

The inferred value will be F if no additional conjugation of the phenol group occurs since in that case $C \neq \{\text{additional conjugation}\}$, $\mathrm{Poss}(\bar{B}/C) = 1$, Red with $(1 - \mathrm{Poss}(E/F))$ gives 1 independent on what $(1 - \mathrm{Poss}(E/F))$ will be and *anded* with F results in F.

In the case where is C = {additional conjugation}, the $\mathrm{Poss}(\bar{B}/C) = 0$, $(1 - \mathrm{Poss}(E/F)) = 0$ since $\mathrm{Max}_y[E(y) \wedge F(y)] = 1$ (the sets E and F intersect) and as the consequence the inferred value for $H_0 = E(y) \wedge F(y)$. Thus, if additional conjugation can be observed in the molecule then it is taken into account by further specifying the possible range for the pK-value otherwise the original pK-range will be preserved.

Inductive Effects:

The influence of substituents on the acidity of the phenol group is formulated as a second default condition for inferring about the pK-value of the indicator. Qualitatively, the inductive effect of different substituents can be expressed, e.g. by Lucas' series [14].

(– I) NO_2 CN > SO_2 > COOH > C1 > Br > H > $NHCOCH_3$ > CH_3 > $OCOCH_3$. (+I)

In the present database example only Br, CH_3 and R_2HC are to be considered as substituents for H.

Let Z be the variable taking its value in the set of all possible substituents {NO_2, CN, SO_2...}. Then we define the rules

$$\begin{array}{l} \text{If Z is } B_1 \text{ then U is } E_1 \\ \text{If Z is } B_2 \text{ then U is } E_2 \\ \text{If Z is } B_n \text{ then U is } E_n \end{array} \qquad (12)$$

In order to express the consequence U is E_i for a certain substituent B_i we know at least for the substituents of interest that the substituent Br decreases the pK-value compared to the unsubstituted phenol by about 1 pK-unit (– I-effect), and the substituents CH_3 and R_2HC increase the pK-units (+ I-effect). The uncertainty about the exact value is described by a membership function of the form

$$m(y) = [1 - (1/(0.5 * 0.5))\ (y - b)]^+ \qquad (13)$$

with b being the most possible y-value with $m(y) = 1$.

Table 4. *Comparison of measured pK-values with inferred pK-values by default reasoning.*

Compound No.	Substituents	measured pK-value	inferred pK-value
1	4 Br, 2 CH_3	4.66	4.4
2	2 CH_3, 2 R_2HC, 2 Br	7.10	6.9
3	2 CH_3, 2 R_2HC	8.90	8.9
4	without (H)	7.81	7.9
5	2 Br, 2 CH_3	6.12	6.4
6	4 Br	3.85	3.9
7	2 CH_3	8.3	8.4
8	2 CH_3	8.25	8.4
9	4 CH_3	8.8	8.9

Thus reasoning about U is E_i can be undertaken in the same way as with the first default condition, i.e. one obtains with the data V is C for the inferred value H_i

$$H_i = \mathrm{Poss}(\bar{B}_i/C) \vee ((1 - \mathrm{Poss}\ (E_i/H_{i-1})) \vee E_i) \wedge H_{i-1} \qquad (14)$$

In this inference pattern the inferred value H_0 for considering additional conjugation in the system is taken into account as is the presence of several equal or different substituents.

Table 4 gives the results for inferring the pK-values by our reasoning scheme and compares them with the experimentally measured pK- values. The agreement between measured and inferred pK-values is satisfactory and these values could be used as good approximation for solving different chemical tasks.

We note that the reasoning system used is open to incorporate additional rules for further specifying the knowledge about this chemical property.

Conclusion

Approximate reasoning can be considered a very promising tool for manipulating chemical knowledge in knowledge-based systems. This is due to its well established capabilities for handling uncertain and imprecise observations, its facility for handling linguistic quantifiers of chemical properties and its ability to deal with partial matching of knowledge. More complicated inference patterns can easily be constructed and also aggregation of different rules could be carried out in a more sophisticated manner, e.g. by applying the ordered weighted averaging (OWA)-operator [15].

References

[1] E. ZIEGLER, *Computer in der Chemie*, Springer-Verlag: Heidelberg, 1985.

[2] D. L. MASSART, B. G. M. VANDEGINSTE, S. N. DEMING, Y. MICHOTTE and L. KAUFMAN, *Chemometrics: A Textbook, Data Handling in Science and Technology*, vol. **2**, Elsevier: Amsterdam, 1988.

[3] L. A. ZADEH, "A theory of approximate reasoning", in *Machine Intelligence*, vol. **9**, J. HAYES, D. MICHIE & L. I. MIKULICH, (eds.), New York: Halstead Press, pp. 149-194, 1979.

[4] R. R. YAGER, S. OVCHINNIKOV, R. TONG and H. NGUYEN, *Fuzzy Sets and Applications: Selected Papers* by L. A. ZADEH, John Wiley & Sons: New York, 1987.

[5] D. DUBOIS and H. PRADE, "Fuzzy sets in approximate reasoning Part I: Inference with possibility distributions, "*Fuzzt Sets and Systems*, 40, pp. 143-202, 1991.

[6] L. A. ZADEH, "Fuzzy sets", in *Information and Control*, vol. **8**, New York: Academic Press, pp. 338-353, 1965.

[7] E. H. RUSPINI, "On the semantics of fuzzy logic", Technical Note # 475, AI Center, *Stanford Research International (SRI)*, Menlo Park, CA, 1989.

[8] E. H. RUSPINI, "Similarity models for fuzzy logic", Proceedings of the *Third IPMU Conference*, Paris, pp. 56-58, 1990.

[9] R. R. YAGER, "Connectives and quantifiers in fuzzy sets", *Fuzzy Sets and Systems* 40, pp. 39-76, 1991.

[10] D. DUBOIS and H. PRADE, "Fuzzy sets in approximate reasoning Part 2: logical approaches", *Fuzzy Sets* 40, pp. 203-244, 1991.

[11] E. CHARNIAK and D. McDERMOTT, *Introduction to Artificial Intelligence*, Addison, Wesley: Reading, MA, 1985.

[12] M. OTTO and H. BANDEMER, "A fuzzy approach to predicting chemical data from incomplete, uncertain and verbal compounds", Proceedings of the *Beilstein-Workshop on Estimation of Physical Data for Organic Compounds*, Berlin: Springer-Verlag, (To Appear), *in*: C. JOCHUM, M. G. HICKS and J. SUNHEL, Eds., *Physical Property Prediction in Organic Chemistry*, Springer, Berlin, pp. 171-189, 1988.

[13] R. R. YAGER, "Using approximate reasoning to represent default knowledge", *Artificial Intelligence* 31, pp. 99-112, 1987.

[14] R. BRDICKA, *Grundlagen der Physikalischen chemie*, VEB Dt. Verlag Wiis.: Berlin, 1970.

[15] R. R. YAGER, "On ordered weighted averaging aggregation operators in multicriteria decision making", *IEEE Transactions on Systems, Man and Cybernetics* 18, pp. 183-190, 1988.