NAIST-IS-DT9561042

Doctoral Dissertation

STUDIES ON QUALITATIVE INTERPRETATION OF INACCURATE DATA

Qi Zhao

December 1995

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

Dissertation submitted to
Graduate School of Information Science
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering

Qi Zhao

Thesis committee: Toyoaki Nishida, *Professor*
Yuji Matsumoto, *Professor*
Yoshiteru Ishida, *Associate Professor*

Studies on Qualitative Interpretation of Inaccurate Data*

Qi Zhao

Abstract

In this dissertation, I address the following four issues of artificial intelligence: (1) qualitative interpretation of inaccurate data, (2) possibility propagation and uncertain reasoning, (3) constraint satisfaction problems, and (4) knowledge-based systems. First, I present a novel method for interpreting inaccurate data by using qualitative correlations among related data as confirmatory or disconfirmatory evidence. Second, I present a novel method for reasoning under uncertainty by extracting and propagating qualitative correlations among hypotheses. Third, I present a practical method for solving constraint satisfaction problems, including an efficient pattern-driven algorithm for generating initial solutions and an overlap-reduce heuristic for repairing the initial solutions. Finally, I introduce a knowledge-based system for infrared spectrum interpretation. The above three methods are all successfully applied to the system, and the implementation of the system indicates that it is significantly better than many similar systems.

Keywords:

Qualitative Correlations, Support Coefficient Function, Inaccuracy Handling, Possibility Propagation, Uncertain Reasoning, Knowledge-Based Systems, Solution Generating and Repairing, Constraint Satisfaction Problems

---

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9561042, Dec. 1995.

Preface

Many problems of artificial intelligence (AI) are solved by analyzing input data. However, when input data are inaccurate, analyzing them becomes very difficult, and may not lead to any useful conclusion. Therefore, interpreting inaccurate data to determine what they mean or what they should be has long been regarded as a significant and difficult problem in AI.

Meanwhile, in practical problems, especially in data rich problems such as diagnosis and signal processing, input data are often inaccurate. One reason is that the measuring and entering methods are error-prone, and the other is that the real data are not noise-free. As a matter of fact, interpreting inaccurate data is an unavoidable problem in many applications of AI.

In this dissertation, I present qualitative methods for interpreting inaccurate data. The methods consider qualitative dependencies among data, called qualitative correlations among related data, as confirmatory or disconfirmatory evidence of interpreting inaccurate data.

First, I introduce and discuss the definitions of related data and qualitative correlations among related data. Then, I put forward a new concept called support coefficient function (SCF). SCF can be used to extract, represent, and calculate qualitative correlations among related data within a dataset.

I propose an approach to calculating shift intervals of inaccurate data which, based on SCF, dynamically determines how inaccurate an inaccurate data item is allowed to be. Then, I propose an approach to calculating possibility of identifying inaccurate data in the dynamic shift intervals.

On the basis of the above two approaches, I present a novel method for interpreting inaccurate data by considering qualitative correlations among related data as confirmatory or disconfirmatory evidence.

Then, I extend the method to a wider scope to propagate qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence for uncertain reasoning. I present a method for extracting, representing and propagating

qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence to update the possibilities of hypotheses. The function of the method is similar to the probability propagation on Bayesian networks. But compared with traditional methods for probability propagation, the method has the following advantages: (1) it can be applied to the problems where evidence is not explicitly given; (2) few numbers and assumptions need to be provided by domain experts in advance; and consequently, (3) the knowledge acquisition is simple, and the inconsistency in knowledge bases can be avoided.

I have applied the methods to infrared spectrum interpretation, and have thoroughly tested the methods against about three hundred real spectra. The experimental results show that the methods are significantly better than the conventional ones used in many similar systems. This dissertation also describes the implementation of the methods and the experiments.

In this dissertation, I also present a knowledge-based system for infrared spectrum interpretation. The system employs the above two methods to handle inaccuracy of infrared spectral data. Since handling inaccurate data is only one issue of infrared spectrum interpretation, I introduce the other issues of infrared spectrum interpretation, and give the overall picture of the system.

First, I discuss the principle and process of infrared spectrum interpretation, and propose a knowledge model for integrating qualitative reasoning and quantitative analysis. Then, I introduce the design and development of the knowledge-based system, and present the architecture and working process of the system.

In addition, I also present a new method for solving constraint satisfaction problems in this dissertation. I propose an efficient pattern-driven algorithm for generating initial solutions, and an overlap-reduce heuristic for repairing the initial solutions. The method is initially developed for solving the constraint satisfaction problems in infrared spectrum interpretation, but it is applicable to a class of similar problems.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Professor Toyoaki Nishida, for his invaluable guidance and encouragement during my doctoral study. He not only supervised my research, but also provided me with financial assistance and various support. I enjoyed my days working with him very much.

I would like to thank Professor Yuji Matsumoto and Associate Professor Yoshiteru Ishida at Nara Institute of Science and Technology for their guidance. They read the draft of this dissertation and gave me many excellent comments and suggestions. My thanks also go to Associate Professor Hideaki Takeda and other staff and students at Nishida's laboratory.

Part of the work in this dissertation was conducted when I was a researcher at ASTEM Research Institute of Kyoto, Japan, during 1991 to 1994. I would like to thank Dr. Yutaka Ohno, then President of ASTEM, for inviting me to work there. Thanks to the colleagues and managers at ASTEM for their various help and support when I worked there.

The research was once partially supported by Horiba Ltd. of Kyoto, Japan. I would like to thank the managers and colleagues there for their support and cooperation. I would also like to thank NEC Foundation for C&C Promotion for granting me the 1995-96 C&C Research Fellowship.

I owe a great debt of gratitude to Prof. Wuyi Yue at Konan University, Prof. Chiaki Sakama at Wakayama University, and Prof. Tatsuya Mikami at Ritsumeikan University for their help and friendship.

Special thanks are due to Mr. Hisao Kumagai and his family. They helped me so much during my stay in Japan.

Finally, I would like to thank my wife, Chunling. She encouraged me to return to university to pursue my long cherished doctoral degree after I had worked for several years, and supported me in the whole period of the stressful doctoral study. As a computer professional, at the same time, she continuously gave

me suggestions, criticisms and encouragement. Without her love and support, I could not have completed the dissertation in such a short time. I would also like to thank my parents in China for their love and understanding.

Contents

Preface v

Acknowledgments vii

1 Introduction 1
- 1.1 Significance of the Research 1
- 1.2 Motivations 3
- 1.3 Objectives 5
- 1.4 Contributions 6
- 1.5 Outline of the Dissertation 9

2 Background Problem 11
- 2.1 Problem Description 11
- 2.2 Example 12
- 2.3 Formal Representation 15

3 Preliminaries 19
- 3.1 Qualitative Dependency and Related Data 19
  - 3.1.1 Definition 19
  - 3.1.2 Example 20
- 3.2 Qualitative Correlations among Related Data 20
  - 3.2.1 Definition 21
  - 3.2.2 Example 21
- 3.3 Support Coefficient Function 22
  - 3.3.1 Definition 22
  - 3.3.2 Example 22
- 3.4 Dynamic Shift Interval 23
  - 3.4.1 Definition 23

3.4.2 Example 24
3.5 Evidence Based on Qualitative Correlations 27
3.5.1 Definition 27
3.5.2 Example 28
3.6 Summary 28

4 Qualitative Interpretation of Inaccurate Data **31**
4.1 Introduction 31
4.2 Predicate "$d_i@R_j$" 32
4.2.1 Defining Support Coefficient Function 33
4.2.2 Determining Dynamic Shift Interval 34
4.2.3 Calculating Value of Predicate "$d_i@R_j$" 36
4.2.4 Procedure $d_i@R_j$ 38
4.3 Predicate "$R_j@MD$" 39
4.3.1 Calculating Value of Predicate "$R_j@MD$" 39
4.3.2 Procedure $R_j@MD$ 39
4.4 Algorithm for Qualitatively Interpret Inaccurate Data 40
4.4.1 Algorithm 40
4.4.2 Analysis 41
4.5 Discussion 42
4.5.1 Intuition 42
4.5.2 Applicability 43
4.5.3 Comparison 44
4.6 Summary 45

5 Propagation of Qualitative Correlations **47**
5.1 Introduction 48
5.2 Probability Propagation on Bayesian Networks 49
5.3 Propagation of Qualitative Correlations 51
5.3.1 Two Extensions 52
5.3.2 Concepts and Definitions 52
5.4 Method for Qualitative Correlation Propagation 56
5.4.1 Algorithm for Obtaining Qualitative Correlations 56
5.4.2 Algorithm for Propagating Qualitative Correlations 58
5.5 An Example 59
5.6 Discussion 61
5.6.1 Properties 61
5.6.2 Comparison 62

5.7 Summary 63

6 Implementation and Experiments 65
6.1 Infrared Spectrum Interpretation 65
6.2 Applying the Proposed Methods to Infrared Spectrum Interpretation 66
6.3 System for Interpreting Infrared Spectra 69
6.4 Examples 70
6.4.1 Case I: Considering the First Kind of Related Data 71
6.4.2 Case II: Considering the Second Kind of Related Data . . 73
6.5 Analysis of Experimental Results 74
6.6 Comparison with Related Systems 79
6.6.1 Systems Based on Yes/No Classification 79
6.6.2 Systems Based on Fuzzy Logic 80
6.6.3 Systems Based on Pattern Recognition 81
6.6.4 Systems Based on Neural Networks 82
6.7 Summary 82

7 Knowledge-Based System for Infrared Spectrum Interpretation 83
7.1 Introduction 83
7.2 Design of the System 85
7.2.1 Qualitative Process 86
7.2.2 Quantitative Process 89
7.3 Architecture of the System 91
7.4 Summary 93

8 A Method for Solving Constraint Satisfaction Problems in Infrared Spectrum Interpretation 95
8.1 Introduction 95
8.2 Delay of Some Constraints 97
8.3 Pattern-Driven Algorithm 99
8.4 Overlap-Reduce Heuristic 102
8.5 Discussion 104
8.5.1 Defects of the Method 104
8.5.2 Effectiveness 104
8.6 Summary 105

9 Related Work and Discussion 107

10 Conclusions 111

List of Publications 113

Bibliography 117

Chapter 1

Introduction

> *I viewed uncertainty as a normal, unavoidable result of interactions between agents and their environments. ... Yet humans are rarely crippled by uncertainty and generally manage with meager computational resources.*
>
> — *Paul R. Cohen*

Interpreting inaccurate data can be regarded as a special problem of coping with uncertainty. The problem has long been a significant and difficult problem in artificial intelligence (AI). Since the beginning of AI, researchers have been trying various methods to deal with the problem.

1.1 Significance of the Research

In daily life, there are many uncertainties caused by inaccuracy. People may pronounce some words improperly in their speech, or make some spelling mistakes in their writing. If the mistakes are not very serious, they can usually be understood, and cause no confusion.

The ability of interpreting and understanding minor inaccurate data or phenomena is an important feature of human intelligence.

In science and engineering problems, the uncertainties caused by inaccurate data are more common and unavoidable since few problems can provide completely accurate data [Cohen, 1984 & 1987]. The reasons of having inaccurate data involved are various. One main reason is that the methods of measuring and

entering data are error-prone sometimes. For example, a patient's temperature may be inaccurately measured or entered, and a signal or an experimental data item may be inaccurately observed or recorded. The other reason is that the real data are not noise-free in some cases. For example, among the received satellite signals, there may be some noise mixed up, and what is worse, infrared spectral data (peaks) themselves may be noisy, i.e., some peaks may be affected by noise or other factors [Colthup, Daly, & Wiberley, 1990][Oppenheim & Nawab, 1992].

The ability of interpreting and understanding inaccurate data plays an important role in solving complex science and engineering problems.

AI researchers have long viewed the problem of interpreting inaccurate data as a significant and difficult problem, and have been attempting to integrate the human intelligence into computer systems [Kuipers, 1988][Zadeh, 1978].

In many problems of AI, inferences are drawn on the basis of interpretation or analysis of measured data [de Kleer & Williams, 1987][Sacks, 1991]. However, when measured data are inaccurate, interpreting or analyzing them is very difficult. In diagnosis or signal analysis, for example, the general reasoning method is to compare measured data with reference values [Reiter, 1987][Shortliffe & Buchanan, 1975][Voscovi & Robles, 1992]. When measured data are not accurate due to noise or other unforeseen reasons, the comparison between measured data and reference values may not lead to any useful conclusion. A rule like "*if there is a strong peak in 3000 cm^{-1}- 3100 cm^{-1} on the infrared spectrum of an unknown compound, then the unknown compound may contain at least one benzene-ring*" may work in ideal cases (here "*strong*" and "*3000 cm^{-1}- 3100 cm^{-1}*" are reference values, and "*infrared spectrum of unknown compound*" is a measured dataset). However, the rule can not work in general cases. For example, when the spectral data are inaccurate, e.g., the measured peak in 3000 cm^{-1}- 3100 cm^{-1} is not a strong peak but a medium one, or a measured strong peak is not exactly located in 3000 cm^{-1}- 3100 cm^{-1} but shifts slightly, the rule may not be applied.

In the past decades, many methods have been proposed to deal with the problem. Fuzzy logic provides a mathematical framework for representation and calculation of inaccurate data [Zadeh, 1978 & 1989]. By fuzzy logic, reference value x_0 is associated with a fuzzy interval $\triangle x$. If a measured data item falls into $[x_0 - \triangle x, x_0 + \triangle x]$, then it can be interpreted as the reference value with a corresponding membership degree. Probability theory and possibility theory are also widely used for handling inaccuracy and uncertainty [Dempster, 1968][Duda, Hart, & Nilsson, 1976] [Pearl, 1988][Shafer, 1976]. In many practical systems, when statistical samples are insufficient or absent, subjective statements are

used to take the place of statistics of inaccurate data or uncertain evidence, such as certainty factors in MYCIN [Shortliffe, 1976], and prior probabilities in PROSPECTOR [Duda, and et al, 1977]. The above methods are commonly used in AI systems. The way of applying them, however, depends on the nature of domain problems, and there is not yet a standard and generally accepted method thus far.

Due to the difficulty of interpreting inaccurate data, many AI systems suppose that all input data have been formalized, and are accurate, then make inference based on the precondition [Kawata, and et al, 1987][Moldoveanu & Rapson, 1987]. However, the precondition is not tenable to non-toy or non-experimental problems.

1.2 Motivations

I began my research on interpreting inaccurate data around 1984 when I developed an experimental natural language interface for a road inquiring system [Zhao, 1984]. The function of the system was to automatically provide information about the roads, traffic and main scenic spots of Beijing City. The interface was developed to enable users to input their inquiries by using limited words and sentences, such as:

How long will it take from **A** *to* **B***?* or

Where is Street **C** *located?*

Because users sometimes made mistakes when they entered inquiries, especially when they entered the addresses or building names, the interface needed to check its dictionary first to detect wrong entrances. Doing that was not difficult since the dictionary was very limited. However, frequently telling users that their inquiries were invalid and that their inquiries needed to be entered again made the use of the interface and the system very boring, especially when users only had one or two letters typed wrong. As an attempt, I tried to interpret unmatched strings based on how many letters in the string were matched and how many letters were not. For example, suppose the entered string was *Wangfujingsajie*. In the dictionary, there was no string as *Wangfujingsajie*, but there was a very close string as *Wangfujingdajie*. Because 14 out of 15 letters in the two strings were the same (i.e., the matching rate is about 93%), the

interface interpreted *Wangfujingsajie* as *Wangfujingdajie* by assuming that letter *s* came from letter *d* due to users' inaccurate (improper) typing. In these cases, the interface first reminded users of the wrong typing, then went ahead to answer user's inquiries. Only when the interface made an interpretation for an unmatched string but the matching rate was not high, or it could not make any interpretation at all, users were required to enter their inquiries again. Although the method for processing inaccurate (improper) data in the interface was very simple, it significantly improved the use of the interface.

Later in almost every system I developed, I met the similar uncertainties caused by inaccurate data, although the inaccurate data appeared in various ways.

For example, in developing the knowledge-based system for general cargo stowage, I met many inaccurate data and phenomena [Zhao, 1986a] [Zhao & Lin, 1987 & 1988]. The system can be briefly described as follow.

> *Given the capacity and shape of a ship hold, and given a batch of general cargoes with their weights, volumes, shapes, packing forms and stowage requirements, the system should make stowage plans to properly arrange all cargoes in the ship hold.*

One of the difficult tasks of the system was to deal with inaccurate data. Because cargoes were all general cargoes, the shapes of them were irregular. As a result, the volumes of cargoes were always estimated, and hence inaccurate. It was impossible to measure cargoes for accurate volumes before making stowage plans, so the system had to imitate domain experts to use other information, such as weights, specific gravities, shapes and packing forms to interpret inaccurate volumes, i.e., to infer the accurate volumes.

I realized two important facts from developing the system:

1. Interpreting inaccurate data is unavoidable in solving practical engineering problems;

2. Related data can remedy inaccurate data, therefore, they can be used as evidence of interpreting inaccurate data.

In 1991, I undertook a large project on infrared spectrum interpretation at ASTEM Research Institute [Zhao, 1991-1993]. The objective of the project was to develop programs to interpret infrared spectra of unknown compounds in order to identify what the unknown compounds contain, or what they are. The

input datasets of the project (i.e., infrared spectra of unknown compounds) were typical inaccurate datasets. Many noises were included in obtaining the datasets. And what is worse, even if the datasets were thresholded and filtered, they were still inaccurate in most cases because data themselves (i.e., peaks on infrared spectra) affected each other very often. Therefore, I did much work to interpret the inaccurate datasets to give correct solutions.

From then on till my doctoral study at Nara Institute of Science and Technology (NAIST), I have been studying on the general, effective and efficient methods for interpreting inaccurate data. The research described in this dissertation is about my work then at ASTEM, and later at NAIST.

1.3 Objectives

Inaccurate data mean data which shift from the correct values (or the reference values). For example, if we consider a word as a set of data, then a letter typed wrong in the word is an inaccurate data item. On the other hand, if we consider a sentence as a set of data, then a word spelled wrong in the sentence is an inaccurate data item.

Interpreting inaccurate data is to find the correct values or the most suitable values for inaccurate data [Riese, 1993]. In signal processing or infrared spectrum interpretation, for example, interpreting inaccurate data is to get the meanings of inaccurate signals or infrared spectral data [Oppenheim & Nawab, 1992].

Traditional methods for interpreting inaccurate data are primarily based on quantitative analysis. For example, in fuzzy logic, interpreting inaccurate data is based on quantitative calculation of the membership degree of an inaccurate data item in a fuzzy region [Bowen, Lai, & Bahler, 1992][Mukaidono, Shen, & Ding, 1989]. In probabilistic reasoning, prior probabilities or statistic values are needed beforehand for making interpretation of datasets [Laskey & Lehner, 1989][Ramer & Lander, 1991]. In continuous methods, a distortion function is needed to determine the possible values of an inaccurate data item [Bose & Rajamoney, 1993][Console, Friedrich, & Dupre, 1993][Raskutti & Zukerman, 1991].

However, the main problems of only using quantitative analysis are:

1. Quantitative analysis needs many numbers and mathematical models provided by domain experts in advance. But, unfortunately, in many problems, these numbers and models are not always available.

2. Quantitative analysis is usually very complex, and in some cases, it may become intractable.

The objectives of the research are using qualitative methods to effectively interpret inaccurate data, and applying the methods to practical problems.

1.4 Contributions

I present a method for interpreting inaccurate data on the basis of qualitative correlations among related data. The method is based on the essential consideration that some data items within a dataset are qualitatively dependent: a set of data may describe the same phenomenon, or refer to the same behavior. For example, a patient's temperature, blood pressure and other symptomatic data reflect the patient's disease, and a couple of peaks on an infrared spectrum indicate the presence of a partial component. The dependency among data within a dataset is called *qualitative correlation among related data*[1].

By considering qualitative correlations among related data, the confirmatory or disconfirmatory evidence can be obtained to interpret inaccurate data. In general, related data should be simultaneously present or absent, so if most of the related data have been completely identified, these data will enhance the identification of the rest. For example, a benzene-ring can create many other peaks besides the strong peak in 3000 cm^{-1}- 3100 cm^{-1}. All the peaks created by the benzene-ring are related data which have qualitative correlations. If all the peaks except that in 3000 cm^{-1}- 3100 cm^{-1} have been completely identified, the benzene-ring is quite likely to be contained by the unknown compound. Therefore, the inaccurate peak around 3000 cm^{-1}- 3100 cm^{-1} may still be identified. In fact, spectroscopists frequently use the following knowledge in addition to the rules given in Section 1.1:

> *If there is a strong peak around 3000 cm^{-1}- 3100 cm^{-1}, then the spectrum may be partially created by benzene-rings —- check peaks around 1650 cm^{-1}, 1550 cm^{-1} and 700 cm^{-1}- 900 cm^{-1} to make sure because a benzene-ring may have other peaks there at the same time.*

The central idea of the method is to find evidence for interpreting inaccurate data by considering qualitative correlations among related data. The idea is

[1]Detailed definitions will be given later.

very common in human thinking. When all the data except blood pressure of a patient show that the patient has a certain disease, we would naturally suspect that the blood pressure of the patient was inaccurately entered. Similarly, when all the peaks except one indicate that a partial component is present, we would naturally suspect that the unmatched peak was inaccurately measured or the peak was affected by noise or something else. If acceptable solutions can be made by assuming an inaccurate data item to be a reference value based on qualitative correlations between the data item and its related data, the inaccurate data item may be compensated and hence identified.

The key point is a new concept called support coefficient function (SCF) for extracting, representing, and calculating qualitative correlations among related data. When measured data are inaccurate, the qualitative correlations among related data can provide evidence for confirming or disconfirming the hypothesis that the measured data are the same as the reference values. An approach to determining dynamic shift intervals of inaccurate data, an approach to calculating possibility of interpreting inaccurate data, and an algorithm for using the above two approaches are proposed on the basis of SCF, respectively.

The method uses much dynamically obtained information, so it does not require many assumptions in advance, and is more robust. The method interprets inaccurate data by considering qualitative correlations among related data, so it is quite effective and efficient, especially in the case of problems where dependencies among data apparently exist. In general, qualitative correlations among data can always, more or less, be extracted. In the worst case where qualitative correlations are not known a priori, the method degenerates to a conventional fuzzy method[2].

Some extensions which allow the qualitative correlations to propagate among related data enable the method to interpret inaccurate symbolic data. Based on the extensions, I develop a new method for propagating qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence for uncertain reasoning. I present an algorithm for extracting and representing qualitative correlations among hypotheses, and an algorithm for propagating qualitative correlations as confirmatory or disconfirmatory evidence to update the possibilities of hypotheses, respectively. The function of the method is similar to the probability propagation on Bayesian networks. But compared with traditional methods for probability propagation, the method has the following advantages: (1) it can

[2]I refer to the fuzzy methods which use an empirical fuzzy interval for each inaccurate data item as conventional fuzzy methods.

be applied to the problems where evidence is not explicitly given; (2) few numbers and assumptions need to be provided by domain experts in advance; and consequently, (3) the knowledge acquisition is simple, and the inconsistency in knowledge bases can be avoided.

I have applied the above two methods to infrared spectrum interpretation, and have thoroughly tested the methods against several hundred real spectra. The experimental results show that the methods are significantly better than the traditional methods used in many similar systems.

I also present a knowledge-based system for interpreting infrared spectra. The primary task of the system is to identify unknown compounds by interpreting their infrared spectra. I propose a knowledge model for integrating qualitative reasoning into infrared spectrum interpretation. The implementation of the system indicates that both the efficiency and quality are improved by employing the knowledge model. The rate of correctness (RC) and the rate of identification (RI) of the system are near 74% and 90% respectively, and the former is the highest among known systems[3].

In addition, I present a new method for solving constraint satisfaction problems. The method is initially developed for solving the constraint satisfaction problems in infrared spectrum interpretation, but it is applicable to a class of similar problems. I propose an efficient pattern-driven algorithm for generating initial solutions, and an overlap-reduce heuristic for repairing the initial solutions.

Briefly, my contributions mainly include:

1. A qualitative method which interprets inaccurate data by using qualitative correlations among related data as confirmatory or disconfirmatory evidence, and a corresponding algorithm which crystallizes the method;

2. A qualitative method which propagates qualitative correlations among known hypotheses to update the possibilities of the hypotheses, and a corresponding algorithm which crystallizes the method;

3. Successful applications of the above two qualitative methods to a practical problem;

4. A knowledge-based system which integrates qualitative reasoning and quantitative analysis to interpret infrared spectra;

[3] RC and RI are two important standards for evaluating the solutions of infrared spectrum interpretation. I will give the detailed definitions of RC and RI, and show the experimental results of the system later.

5. A new method for solving constraint satisfaction problems including an efficient pattern-driven algorithm for generating initial solutions and an overlap-reduce heuristic for repairing the initial solutions.

1.5 Outline of the Dissertation

This dissertation consists of ten chapters. The rest of the dissertation is organized as follows.

In Chapter 2, I describe the problem of interpreting inaccurate data. First, I discuss the concepts of inaccurate data and inaccurate data interpretation, and give some examples. Then, I introduce the formal representation of interpreting inaccurate data.

In Chapter 3, I give the preliminaries of my research including some new definitions and notions. First, I define the concepts of related data and qualitative correlations among related data. Then, I put forward a novel concept called support coefficient function (SCF), and discuss the dynamic shift intervals of inaccurate data based on SCF. Finally, I introduce the concept of determining the possibilities of interpreting inaccurate data in the dynamic shift intervals by considering qualitative correlations among related data as confirmatory or disconfirmatory evidence.

In Chapter 4, I present a method for interpreting inaccurate data by considering qualitative correlations among related data. First, I discuss the method of extracting, representing and calculating qualitative correlations among related data, and the method of representing and calculating SCF. Then, I propose an approach to calculating shift intervals of inaccurate data which, based on SCF, dynamically determines how inaccurate an inaccurate data item is allowed to be, and an approach to calculating possibility of identifying inaccurate data in the dynamic shift intervals. Finally, based on the above two approaches, I propose an algorithm for interpreting inaccurate data by considering qualitative correlations among related data as confirmatory or disconfirmatory evidence, and briefly discuss the applicability and complexity of the algorithm.

In Chapter 5, I present a method for propagating qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence for uncertain reasoning. The method can automatically extract and propagate qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence to update the possibilities of hypotheses. First, I introduce the method. Then, I

give an example to demonstrate the use of the method, and discuss its properties and applicability.

In Chapter 6, I introduce the application of the methods to a practical system for infrared spectrum interpretation, and discuss my experiments and the implementation of the proposed methods. First, I briefly describe the architecture of the system, and the way of applying the proposed methods to the system. Then, I introduce the experimental results with the system. I have thoroughly tested the system against about three hundred real infrared spectra. The experimental results show that the proposed methods are significantly better than the conventional methods used in many similar systems.

In Chapter 7, I present a knowledge-based system for infrared spectrum interpretation. I propose a knowledge model for integrating qualitative reasoning into infrared spectrum interpretation, and introduce the design, architecture and working process of the system.

In Chapter 8, I present a new method for solving constraint satisfaction problems in infrared spectrum interpretation. I propose an efficient pattern-driven algorithm for generating initial solutions, and an overlap-reduce heuristic for repairing them.

In Chapter 9, I address the related work, and analyze the advantages and disadvantages of my own work. I claim that in general cases, qualitative correlations among related data are always available since there are always structures existing in complex datasets. Therefore, the methods proposed in this dissertation are generally effective and efficient. I also claim that in some cases when the qualitative correlations among related data are not known a priori, the methods degenerate to a traditional fuzzy method. I also address my future research work in Chapter 9.

Finally, I conclude the dissertation in Chapter 10.

Some chapters of this dissertation are based on our published papers. Roughly, Chapter 3 and 4 are based on [Zhao & Nishida, 1995a & 1995c]. Chapter 5 is based on [Zhao & Mikami, 1994]. Chapter 6 is based on [Zhao & Nishida, 1995b]. Chapter 7 is based on [Zhao & Nishida, 1994a, 1994b & 1995d]. Chapter 8 is based on [Zhao, 1994].

Chapter 2

Background Problem

In this chapter, I introduce the background problem of the research including the problem description of interpreting inaccurate data, an example of the problem in science and engineering, and the logic representation of the problem.

2.1 Problem Description

Analyzing datasets is a commonly used method for solving science and engineering problems [Fringuelli, and et al, 1991][Wang, 1994]. For example, the behaviors of a device form a dataset of the device. By analyzing the dataset, the state of the device can be known, and the troubles of the device can also be diagnosed [Biswas & Yu, 1993][Huberman & Struss, 1989][Iwasaki, and et al, 1993][Zhao, 1991]. In spoken or written language understanding, a string of letters form a dataset of a word, and a set of words form a dataset of a sentence. Understanding the word or the sentence requires analyzing the corresponding dataset. In signal processing and image understanding, the main task is to analyze datasets, i.e., received signals or images, to get the interpretation or structures behind them [Blaffert, 1986][Luinge, and et al, 1987].

In practice, however, datasets are often inaccurate due to various reasons most of which are unforeseen, or unknown at all. When a dataset contains inaccurate data, analyzing it becomes very difficult [Berry, 1992][Cullen, Hull, & Srihari, 1992].

There are primarily three kinds of inaccuracy in datasets including:

1. Inaccuracy caused by including noises or irrelevant data in datasets. For example, in signal processing, many noises may be received along with signal series;

2. Inaccuracy caused by inaccurately measuring or entering data. For example, a letter in a word may be typed wrong, and a word in a sentence may be spelled improperly;

3. Inaccuracy caused by data themselves affecting each other. For example, in infrared spectrum interpretation, due to co-existence of two different partial components in a compound, their peaks on infrared spectra may shift from their reference values.

The occurrences and causes of inaccurate data are unpredictable, or unknown at all in some cases, so interpreting inaccurate data in datasets is a very difficult task.

Many systems bypass the problem, and directly assume that all data in their datasets have been accurate. However, practical datasets can rarely be guaranteed to be completely accurate.

2.2 Example

Take the infrared spectrum interpretation for example to introduce the problem of interpreting inaccurate data [Colthup, Daly, & Wiberley, 1990][Savitzky, 1987].

Infrared spectrum interpretation is a typical problem having inaccurate data included in datasets. Some inaccurate data are caused by noises, but in most cases, they are caused by unforeseen effects among data themselves.

The task of infrared spectrum interpretation is to interpret infrared spectra of unknown compounds to identify the unknown compounds, or to identify the

compositions of the unknown compounds (i.e., to identify what partial components (PC) the unknown compounds contain).

The process of infrared spectrum interpretation for unknown compounds is shown in Figure 2.1.

The input datasets of infrared spectrum interpretation are infrared spectra of unknown compounds, and the solutions are sets of partial components which the unknown compounds may contain.

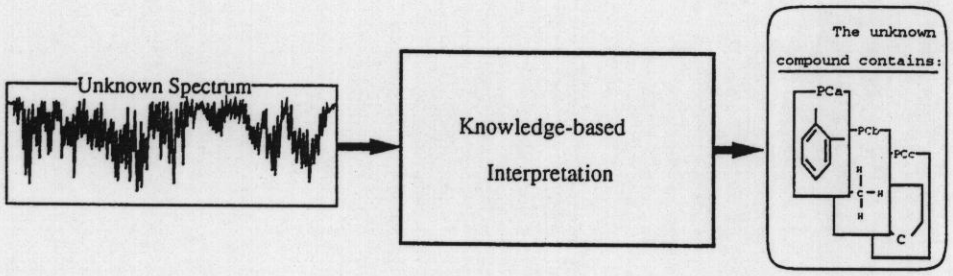

Figure 2.1: Process of Infrared Spectrum Interpretation

An infrared spectrum can be represented as a set of peaks:

$$Sp = \{p_1, p_2, ..., p_n\}$$

where peak p_i consists of its frequency position f_i, strength s_i and width w_i, that is,

$$p_i = (f_i, s_i, w_i) \qquad i = 1, 2, ..., n$$

The peak lists of partial components are known in advance, each of which is a list of peaks that the partial component can create. Table 2.1 shows the examples of partial components and their peak lists.

CH_3-	*Peak List*						
	Fre.	Str.	Wid.	Fre.	Str.	Wid.	...
	2960	1.3626	0.7105	2870	0.7920	0.3011	...

$[CH_3]_2-CH-$	*Peak List*						
	Fre.	Str.	Wid.	Fre.	Str.	Wid.	...
	1170	1.1034	0.6918	1145	0.6672	0.1082	...

$-[CH(C_6H_5)-CH_2]_n-$	*Peak List*						
	Fre.	Str.	Wid.	Fre.	Str.	Wid.	...
	1170	1.1034	0.6918	1145	0.6672	0.1082	...

Table 2.1: Examples of Partial Components and Their Peak Lists

Suppose the peak list of partial component PC_α is

$$PL(PC_\alpha)=\{p_{\alpha_1}, p_{\alpha_2}, ..., p_{\alpha_m}\}$$
$$=\{(f_{\alpha_i}, s_{\alpha_i}, w_{\alpha_i}) \mid i = 1, 2, ..., m\}$$

then if PC_α is contained by a compound, peaks in $PL(PC_\alpha)$ will appear on the infrared spectrum of the compound.

Ideally, if all spectral data are accurate, the process of infrared spectrum interpretation for unknown compounds can be simply described by the following steps:

1. Select a peak, p_i, from Sp. Retrieve all partial components whose peak lists have the same peak, and put the partial components in a candidate list: CL;

2. Select a partial component, PC_j, from CL. If $PL(PC_j) \subset Sp$, then put the partial components in a solution list: SL; Otherwise, delete the partial component from CL;

3. Goto 2 until all partial components in CL are checked;

4. Goto 1 until all peaks in Sp are identified;

5. Delete conflicts (overlaps) among partial components in SL, and output SL as the solution.

The above process can be briefly represented as the following predicate calculi.

$$\forall p_i \forall PC_j((p_i \in PL(PC_j)) \rightarrow (PC_j \in CL)), \quad and$$

$$\forall PC_j((PC_j \in CL) \wedge (PL(PC_j) \subset Sp) \rightarrow (PC_j \in SL))$$

In practice, however, spectral data are often inaccurate due to various reasons most of which are unforeseen, or unknown at all. The main reasons causing inaccuracy of spectral data include:

1. Infrared spectral data are very easy to be affected by noise;

2. Infrared spectral data vary along with different conditions and purity of compounds;

3. Finally and most importantly, infrared spectral data are often shifted by effects among co-existing partial components, or in other word, effects among data themselves in datasets.

Because the peak lists of partial components are accurate but Sp is inaccurate (i.e., $p_i \in PL(PC_j)$ and $PL(PC_j) \subset Sp$ are uncertain), it is not simple to determine whether $p_i \in PL(PC_j)$ and $PL(PC_j) \subset Sp$ are $true$ or $false$.

2.3 Formal Representation

In practical problems, measured data can be represented as a finite set:

$$MD = \{d_1, d_2, ..., d_n\},$$

and reference values can also be represented as a finite set:

$$RV = \{r_1, r_2, ..., r_N\}.$$

Interpreting or analyzing measured data is typically carried out on the basis of so-called "if-then" rules in which the premises are comparisons between MD

and RV like "*if* $d_i = r_j$ *then* ...", or "*if* $(r_i \in MD) \wedge (r_j \in MD)$ *then* ...". When MD is accurate, the main operation implied by these premises is usually to find a corresponding reference value from RV for each data item in MD. However, when MD is inaccurate, the operation becomes complicated. In this case, it is difficult to determine to which reference value an inaccurate data item corresponds, e.g., for some measured data no reference value may be simply identified, while for others more than one may be available.

For example, if received signals are known to be accurate, and an expected signal (reference value) can not be found from the signal series (measured data), then we can conclude that the expected signal does not appear. However, if received signals are inaccurate, and an expected signal can not be identified from the signal series, it is hard to determine whether the expected signal does not appear or appears but looks different due to the inaccuracy [Oppenheim & Nawab, 1992].

Most currently known approaches to dealing with inaccurate data such as fuzzy logic and probabilistic reasoning are mainly based on quantitative similarity or closeness between measured data and reference values [Kruse, 1984][Wang, 1983]. However, the identity of qualitative features is much more effective and reliable than quantitative similarity or closeness in many cases [Forbus, 1983 & 1987].

Consider signal analysis again. If an inaccurate signal has the same qualitative features as the expected one such as the interval of frequency, the signal may still be identified even though its quantitative features are slightly different from those of the expected one such as strength etc.; conversely, an inaccurate signal may not be identified if it is quantitatively similar to an expected signal but does not have the same qualitative features as the expected one.

I have discussed the following points before:

1. Some data items within a dataset are qualitatively dependent (i.e., they are related data);

2. There are qualitative correlations among related data;

3. Qualitative correlations among related data enable us to confirm or disconfirm the interpretation of qualitative features.

Therefore, RV and MD can be, explicitly or implicitly, divided into finite groups on the basis of qualitative dependencies among data, and the data in each

group are related to each other. For example, RV can be divided into R_1, R_2, ... and R_k:

$$RV = R_1 \cup R_2 \cup ... \cup R_k,$$

where

$$R_j = \{r_{j_l} \mid r_{j_l} \in RV, 1 \leq l \leq N\}.$$

The qualitative correlations among related data in R_j include:

1. Data in R_j should be simultaneously present or absent which means that all reference values in R_j should have corresponding data in MD;
2. The presence of r_{j_p} may enhance the presence of r_{j_q}, and the absence of r_{j_p} may depress the presence of r_{j_q}.

Consequently, considering the qualitative correlations among related data will lead to evidence for the interpretation of inaccurate data.

Suppose the corresponding reference values of measured data MD can be represented as $IN(MD)$, then $IN(MD)$ should be a subset of RV, that is,

$$IN(MD) \subset RV.$$

So the problem of interpreting/analyzing inaccurate data is to make qualitative hypotheses for MD, or in other words, to find an $IN(MD)$ from RV. The problem can be briefly represented as the following predicate calculi:

$$\forall d_i \forall R_j((d_i @ R_j) \rightarrow (R_j \in CL)), \quad \textit{and}$$

$$\forall R_j((R_j \in CL) \wedge (R_j @ MD) \rightarrow (R_j \in IN(MD)))$$

where "$d_i @ R_j$" and "$R_j @ MD$" are two essential qualitative predicates in my methods which represent that d_i possibly (qualitatively) belongs to R_j (i.e., ?

$d_i \in R_j$), and R_j possibly (qualitatively) belongs to MD (i.e., ? $R_j \subset MD$), respectively. Determining "$A@B$" is based on qualitative correlations among related data.

Since $R_j \in CL$ is a certain predicate, I simply use the following expression to represent the above predicate calculi:

$$\{d_i@R_j, R_j@MD\} \Rightarrow \{R_j \in IN(MD)\}.$$

Chapter 3

Preliminaries

In this chapter, I put forward and discuss several new concepts which are used, and play important roles, in the research. I first define these concepts, then give corresponding examples and explanations of them, respectively.

3.1 Qualitative Dependency and Related Data

Data in a dataset are rarely completely independent. There are qualitative dependencies or connections among some data in a dataset. For example, a set of data may describe the same behavior, structure, object or phenomenon.

3.1.1 Definition

Definition 3.1 *Related data: If data d_1, d_2, ..., and d_m describe a common phenomenon, or they refer to the same behavior simultaneously, then they can be treated as related data.*

For example, a patient's temperature, blood pressure and other symptomatic data are related data, and all the features for describing a criminal are also related data. If we consider a word, or correctly, a string, as a dataset, then all letters in the word can be viewed as related data. If we consider a sentence as a dataset, then all words in the sentence can be viewed as related data.

The phenomenon that some data within a dataset are related data is more apparent in engineering problems. In the following section, I give an engineering problem to show the phenomenon that some data items are related to each other.

3.1.2 Example

The datasets of infrared spectrum interpretation are infrared spectra. Or in other words, they are sets of peaks on infrared spectra. Figure 3.1 shows an infrared spectrum of a compound from which two kinds of related data can be noticed.

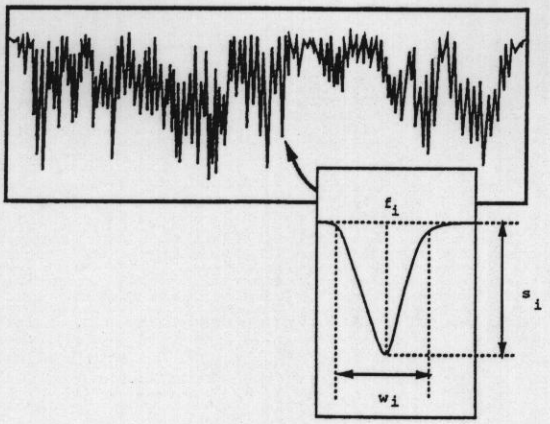

Figure 3.1: Example of Related Data in Spectrum Interpretation

First, as far as a single peak is concerned, the frequency (position) f_i, strength (height) s_i, and width (shape) w_i of the peak are related data. Second, a partial component may create numerous peaks at the same time. If we consider all the peaks that a partial component may create, all of these peaks are related data.

3.2 Qualitative Correlations among Related Data

Related data in a dataset are qualitatively dependent on each other. Suppose a dataset describes the behaviors of a device, the related data describing the same behavior of the device are qualitatively dependent. For example, if a behavior is identified, then all related data describing the behavior should be contained in the dataset. Or in other words, if and only if all related data describing the same behavior are found in the dataset, then the behavior can be identified.

3.2.1 Definition

Definition 3.2 *Qualitative Correlation among Related Data: If d_i and d_j are two related data items, then the presence of d_i qualitatively enhances the presence of d_j, and the absence of d_i qualitatively depresses the presence of d_j. This kind of effects among related data are called qualitative correlations among related data.*

For example, some symptomatic data such as temperature, blood pressure and pulse refer to a certain disease. In describing the disease, all of these symptomatic data are related data. Making a definite diagnosis of a certain disease for a patient, all related data referring to the disease should be definitely found from the dataset of the patient.

In practice, if some symptomatic data of a patient presage a certain disease, the other data related to these data which refers to the disease are usually required to be tested and diagnosed.

3.2.2 Example

Consider the example of infrared spectrum interpretation in Section 3.1.2 again. The frequency (position), strength (height) and width (shape) are related data items, so these three data items should be present or absent simultaneously. When spectral data are accurate, interpreting a single peak on an infrared spectrum requires that these three data items should be all identified. On the other hand, when spectral data are inaccurate (i.e., some measured peaks look like but are not exactly the same as reference peaks), the identification of a data item among the related data may prompt and enhance the identification of others since all related data should be present or absent at the same time. Similarly, the peaks that a partial component can create are also related data items, so these peaks should also be present or absent simultaneously. When spectral data are accurate, interpreting a partial component requires that the peaks created by the partial component should be all identified. When spectral data are inaccurate, on the other hand, the identification of some peaks may prompt and enhance the identification of other peaks since all these peaks should be present or absent at the same time.

3.3 Support Coefficient Function

There are qualitative correlations among related data. The presence of some data items can qualitatively enhance the presence of their related data items, and the absence of some data items can qualitatively depress the presence of their related data items. Therefore, qualitative correlations among related data may lead to either confirmatory or disconfirmatory evidence of interpreting inaccurate data, depending on the degree called support coefficient.

3.3.1 Definition

Definition 3.3 *Support coefficient function (SCF): If there are m-1 data related to d_i, then SCF of d_i is the function to calculate the total effects of the m-1 related data of d_i.*

Suppose $c_i(d_j)$ represents the qualitative correlation between d_i and its related data item d_j, then the support coefficient function of d_i can be defined as:

$$SCF_i = \beta(\sum_{j=1,j\neq i}^{m} c_i(d_j), m).$$

where β is a function which should reflect the ratio of how many and how much related data support d_i. When SCF_i is greater than a certain value given by domain experts, the related data tend to support d_i; Otherwise, the related data tend to depress d_i.

For example, if most symptomatic data referring to a certain disease can not be found, then other data referring to the disease will be depressed, and are not likely to be caused by the disease but by others, although they looks quite like. Conversely, if most symptomatic data referring to a certain disease are found, then other data which are related to the disease but can not be exactly found will be supported, and will usually be analyzed carefully.

3.3.2 Example

Consider the following four strings:

(a) *i-n-a-c-c-u-r-a-t-e*

(b) *i-m-a-c-c-u-r-a-t-e*

(c) *i-m-a-c-c-u-l-a-t-e*

(d) *i-m-a-c-u-l-a-t-e*

(a) is a correct word, but (b), (c) and (d) are all spelled wrong. There is one letter, "*m*", in (b) different from that in (a), so the SCF for interpreting "*m*" in (b) as "*n*"in (a) is quite great. As a result, (b) can be easily interpreted as (a). Further, there are two letters, "*m*" and "*l*", in (c) different from those in (a). Although (c) may be interpreted as (a), the SCFs for interpreting "*m*"and "*l*" in (c) as "*n*"and "*r*" in (a) would not be great. Finally, there are three letters, "*m*", "*c*" and "*l*", in (d) different from those in (a). As a result, (d) will hardly be interpreted as (a) since the SCFs for interpreting these three letters will be very small (even smaller than those for interpreting (d) as word "*immaculate*").

The definition of the threshold that the SCF of an inaccurate data item tends to support the inaccurate data item is domain-dependent. The principle for defining and calculating SCF is that SCF should directly depend on how many and how much related data support an inaccurate data item.

3.4 Dynamic Shift Interval

I use "$d_i@R_j$" to express that d_i can be qualitatively identified from R_j. Realizing "$d_i@R_j$" requires to define a shift interval $\triangle$ for R_j like:

$$R_j \pm \triangle = \{(r_{j_l} \pm \triangle) \mid l = 1, 2, ..., m\},$$

then to determine the possibility of "$d_i \in R_j \pm \triangle$".

The above formula is similar to that in fuzzy logic, but contains completely different meanings. The primary difference is that the shift intervals are dynamically determined by SCF_i, while in fuzzy logic, the fuzzy intervals are usually provided by domain experts in advance or calculated with quantitative criteria.

3.4.1 Definition

Definition 3.4 *Shift Interval: Shift interval is a dynamic region for inaccurate data. Given a standard fuzzy interval for inaccurate data, the shift interval of d_i varies around the standard fuzzy interval on the basis of SCF_i.*

When SCF_i shows that the related data support d_i, the shift interval of d_i becomes wider than the standard fuzzy interval. On the other hand, when SCF_i shows that the related data do not support d_i, the shift interval of d_i becomes narrower than the standard fuzzy interval.

Suppose $\triangle$ is a standard fuzzy region for inaccurate data, then the shift interval, $\triangle d_i$, of inaccurate data item d_i can be dynamically determined the following way:

$$\triangle d_i = \theta(\triangle, SCF_i)$$

where θ is a function which should make $\triangle d_i$ directly depend on SCF_i.

In traditional fuzzy methods, $\triangle d_i$ is simply $\triangle$. When qualitative correlations among related data are extracted and considered, $\triangle$ becomes a standard reference, and the shift interval will be determined by SCF_i.

3.4.2 Example

I first discuss an example of using fuzzy logic to interpreting inaccurate data.

Suppose many inaccurate data are provided to describe the characteristic of a criminal, such as height, color of skin, age, and etc. Because all these data are inaccurate, a fuzzy region $\triangle_x$, and a membership function F_x are needed to calculate the degree that a real data item Y is interpreted as the provided inaccurate data item X.

Briefly, by fuzzy logic, the degree of Y being interpreted as X varies along with the curve shown in Figure 3.2 [Bousson & Trave-Massuyes, 1993][Kruse, Gebhardt, & Klawonn, 1994].

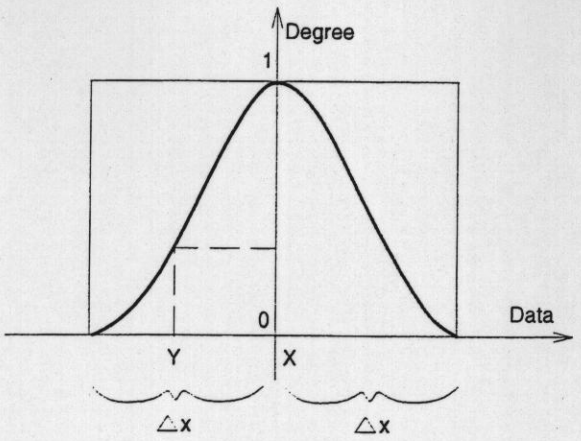

Figure 3.2: Membership Curve of Inaccurate Data

When a real data item Y is located outside the fuzzy region, the degree of Y being interpreted as X is *zero*; Otherwise, the degree varies from 0 to 1 along with the membership curve.

The problem of fuzzy logic is that $\triangle_x$ and F_x are usually fixed and provided by domain experts in advance, so dynamic information and correlations among data can not be properly used.

For example, if the description about the criminal's age is $30 - 40$, then $X_{age} = 35$, and $\triangle_{age} = 5$. As a result, ages outside [30, 40] will be viewed as unidentified.

With the definition of shift interval, the dynamic information and qualitative correlations among related data can be used to dynamically determine how wide a data item is allowed to shift, or in other words, how inaccurate an inaccurate data item is allowed to be.

Suppose all data except *age* have been identified already. These identified data will provide *age* with a great SCF. Therefore, the dynamic shift interval for *age* may become larger than $\triangle_{age}$. With the same membership function, the

degree of identifying *age* will vary along the new membership curve shown in Figure 3.3.

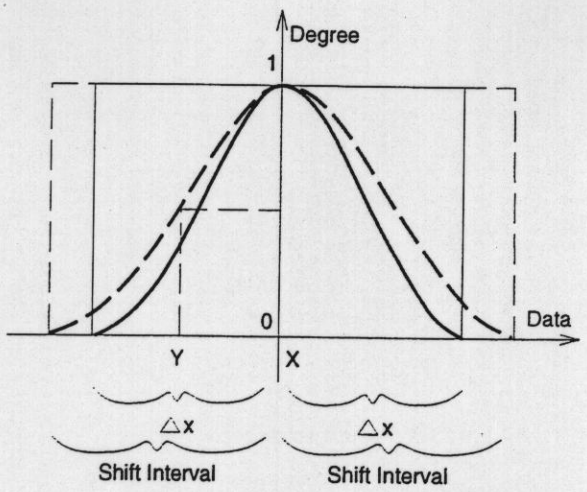

Figure 3.3: A Dynamic Shift Interval (1)

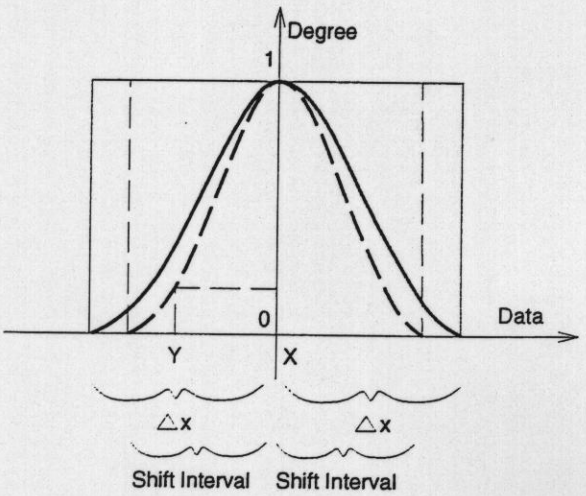

Figure 3.4: A Dynamic Shift Interval (2)

On the contrary, suppose most of other data can not be identified, then these unidentified data will provide *age* with a small SCF. Therefore, the dynamic shift interval for *age* may become narrower than $\triangle_{age}$. With the same membership function, the degree of identifying *age* will vary along the new membership curve shown in Figure 3.4.

3.5 Evidence Based on Qualitative Correlations

I have discussed that the values of SCF_i determines the shift interval of d_i, that is, SCF_i determines how widely d_i is allowed to shift. The wider the shift interval, the more easily d_i is identified. Therefore, SCF_i provides confirmatory or disconfirmatory evidence for identifying d_i.

3.5.1 Definition

Definition 3.5 *Confirmatory Evidence from Qualitative Correlations: If dynamic shift interval of an inaccurate data item becomes wider than the standard fuzzy region by considering the qualitative correlations among related data, then the qualitative correlations provide confirmatory evidence for interpreting the inaccurate data item.*

Definition 3.6 *Disconfirmatory Evidence from Qualitative Correlations: If dynamic shift interval of an inaccurate data item becomes narrower than the standard fuzzy region by considering the qualitative correlations among related data, then the qualitative correlations provide disconfirmatory evidence for interpreting the inaccurate data item.*

If the SCF_i of d_i is greater than a certain value, then the related data tend to support d_i. As a result, the dynamic shift interval of d_i becomes greater than the standard fuzzy region, and confirmatory evidence is provided for interpreting d_i; Otherwise, if the SCF_i of d_i is smaller than the certain value, then the related data tend not to support d_i. As a result, the dynamic shift interval of d_i becomes narrower, and disconfirmatory evidence is provided for interpreting d_i.

3.5.2 Example

The example shown in Section 3.4.2 actually indicates the confirmatory and disconfirmatory evidence got from qualitative correlations among related data. In this section, I discuss the infrared spectrum interpretation as another example.

Partial component CH_3 usually creates numerous peaks including a strong peak located at 2960 cm^{-1}. Because the peaks of CH_3 on real infrared spectra are always inaccurate, especially in the case of CH_3 being contained by a compound together with partial component OH, CO and *Benzene-ring*, spectroscopists usually consider the peaks of CH_3 with a region instead of the exact locations. For example, peaks in 2960 ± 20 cm^{-1} can all be viewed as the peaks of CH_3. Giving a region for an inaccurate data item is similar to the treatment of the values in fuzzy logic, but contains completely different meanings. Spectroscopists determine the region based on static information, but modify it dynamically.

Since CH_3 can create many peaks besides that at 2960 cm^{-1}, the qualitative correlations among the peaks created by CH_3 can be used as confirmatory evidence to enhance the interpretation of the peak at 2960 cm^{-1}, or as disconfirmatory evidence to depress the interpretation of the peak. For example, if the SCF tends to support the identification of the peak, then the dynamic shift interval of the peak will become greater than 20 cm^{-1}. As a result, peaks which can not be identified before may be identified. Similarly, if the SCF tends not to support the identification of the peak, then the dynamic shift interval of the peak will become smaller than 20 cm^{-1}. As a result, the peaks which can be identified before may be considered again.

3.6 Summary

I put forward several new concepts in this chapter, including *related data*, *qualitative correlations among related data*, *support coefficient function*, *dynamic shift interval*, *confirmatory evidence* and *disconfirmatory evidence*.

For each concept, I gave corresponding examples and detailed discussion in order for readers to understand them.

These concepts are used, and play important roles, in the methods to be presented in the following chapters.

With these concepts and examples, I tried to explain the intuition behind the research.

The first concept is *related data* which is based on the phenomenon that data in a dataset are rarely completely independent. For example, a set of data may

describe the same object, or refer to the same behavior. The second concept is *qualitative correlations among related data* which means that related data in a dataset may qualitatively enhance or depress each other. For example, related data should be simultaneously present or absent in general, so if most of the related data have been completely identified, they will enhance the identification of the rest. The third concept is *support coefficient function* which enable qualitative correlations among related data to be represented and calculated. The fourth concept is *dynamic shift interval* which represents the dynamic region that an inaccurate data item may shift. It is not a static region, but varies dynamically along with qualitative correlations among related data. The more an inaccurate data item is supported by its related data, the wider the dynamic shift interval should be. The last two concepts are *confirmatory evidence* and *disconfirmatory evidence*. The central idea of the research is on the basis of these two concepts, that is, to use qualitative correlations among related data as confirmatory or disconfirmatory evidence to interpret inaccurate data.

Chapter 4

Qualitative Interpretation of Inaccurate Data

In this chapter, I introduce a method for qualitatively interpreting inaccurate data. The fundamentals of the method are concepts introduced in Chapter 3, and the central idea is to extract, represent and use the qualitative correlations among related data as confirmatory or disconfirmatory evidence.

4.1 Introduction

I use predicate "$d_i@R_j$" to express that d_i possibly (qualitatively) belongs to R_j, where d_i is a measured data item, and R_j is a set of reference values. And I use predicate "$R_j@MD$" to express that R_j possibly (qualitatively) belongs to MD, where MD is a set of measured data.

Predicate "$d_i@R_j$" and "$R_j@MD$" are two essential qualitative predicates in the research. "$d_i@R_j$" provides a logic framework for interpreting a single inaccurate data item on the basis of qualitative correlations among related data, and "$R_j@MD$" provides a logic framework for interpreting a set of inaccurate data on the basis of qualitative correlations among related data.

The values of "$d_i@R_j$" and "$R_j@MD$" are determined by considering qualitative correlations among related data, which differs the method from fuzzy logic and other methods in the following two aspects:

1. Determining the value of predicate "$d_i@R_j$" (i.e., interpreting inaccurate data item d_i) is not only based on the calculation of d_i, but also based on the calculation of the related data of d_i, since qualitative correlations

among related data can provide more important evidence of interpreting inaccurate data in some cases. Similarly, determining the value of predicate "$R_j@MD$" (i.e., interpreting a set of inaccurate data R_j) is not only based on the calculation of R_j , either;

2. Dynamic information can be properly used in determining the values of predicate "$d_i@R_j$" and predicate "$R_j@MD$". In interpreting inaccurate data, qualitative correlations among related data are extracted and used, and the shift intervals for inaccurate data are dynamically calculated on the basis of qualitative correlations among related data.

In this chapter, I present the method with an emphasis on the realization of these two predicates. First, I discuss how to define and determine the value of predicate "$d_i@R_j$", and introduce a procedure for realizing the predicate. Then, I discuss how to define and determine the value of predicate "$R_j@MD$", and introduce a procedure for realizing the predicate. Finally, I present the method with the form of an algorithm. I also discuss and analyze the applicability and complexity of the algorithm.

4.2 Predicate "$d_i@R_j$"

When d_i is accurate, "$d_i@R_j$" is equal to "$d_i \in R_j$". If there is a reference value in R_j which corresponds to d_i (i.e., $r_{j_p} \in R_j$ and $r_{j_p} = d_i$), then $d_i@R_j = T$. If there is no reference value corresponding to d_i, then $d_i@R_j = F$.

When d_i is inaccurate, however, it is not sure whether r_{j_p} corresponds to d_i. In this case, "$d_i@R_j$" means that d_i possibly (qualitatively) belongs to R_j, or in other words, r_{j_p} possibly (qualitatively) corresponds to d_i. The value of "$d_i@R_j$" is not T or F, but the possibility of "$r_{j_p} = d_i$" or "$d_i \in R_j$".

I have discussed that the identity of qualitative features is much more robust and reliable than quantitative similarity or closeness in many cases. I have also discussed that qualitative correlations among related data can lead to evidence for the identity of qualitative features in diagnosis or interpretation. So if r_{j_p} ($r_{j_p} \in R_j$) is assumed to correspond to d_i, and there are m-1 reference values (i.e., $r_{j_1}, r_{j_2}, ..., r_{j_{p-1}}, r_{j_{p+1}}, ..., r_{j_m}$) related to r_{j_p}, then each of the m-1 reference values should correspond to a certain data item in MD, and the $m-1$ data

items in MD are also related to each other. Therefore, qualitative correlations between d_i and its m-1 related data items in MD should be considered.

The method first determines the possibility of "$r_{j_p} = d_i$" by calculating the similarity or closeness between r_{j_p} and d_i as the same as conventional fuzzy methods, then considers qualitative correlations among related data to obtain evidence for updating the possibility. When the qualitative correlations show that the related data support "$r_{j_p} = d_i$", the possibility of "$r_{j_p} = d_i$" will increase. When the qualitative correlations show that the related data do not support "$r_{j_p} = d_i$", the possibility will decrease.

4.2.1 Defining Support Coefficient Function

Suppose r_{j_q} corresponds to d_t, where $r_{j_q} \in R_j$, and $d_t \in MD$. Because $r_{j_p} \in R_j$, r_{j_q} is related to r_{j_p}, and d_t is related to d_i. As I have discussed, the qualitative correlation between d_i and d_t means that if d_t exists, then d_i is enhanced; otherwise, d_i is depressed.

First, I define the qualitative correlation between two related data items, d_i and d_t, as:

$$c_i(d_t) = \begin{cases} 1 & \text{if } d_t \text{ can be found from } MD \text{ which satisfies:} \\ & r_{j_q} - d_o \leq d_t \leq r_{j_q} + d_o \\ 0 & \text{if } d_t \text{ can not be found from } MD \text{ which satisfies:} \\ & r_{j_q} - d_o \leq d_t \leq r_{j_q} + d_o \end{cases}$$

where d_o is a standard fuzzy interval of inaccurate data, and $c_i(d_t)$ expresses the qualitative correlation between d_i and d_t. $c_i(d_t)$=1 means that d_i is enhanced by its related data item d_t since d_t can be found from the measured dataset, and $c_i(d_t)$=0 means that d_i is depressed by d_t since it can not be found from the measured dataset. The definition of $c_i(d_t)$ is simply based on the consideration that if a data item is identified, then the data item will support its related data items (i.e., the coexisting data items).

As there are m reference values in R_j, the support coefficient function SCF_i of d_i can be defined on the basis of $c_i(d_t)$ $(t = 1, 2, ..., m,\ t \neq i)$:

$$SCF_i = \frac{1 + \sum_{t=1, t \neq i}^{m} c_i(d_t)}{m}$$

where (1) SCF_i expresses the total qualitative correlations between d_i and all of its related data. In other words, SCF_i reflects the support coefficient of r_{j_p} corresponding to d_i; (2) $0 < SCF_i \leq 1$.

If $m = 1$, then $SCF_i = 1$. When $m > 1$, SCF_i is in the direct ratio to the number of the related data which may be identified from MD.

4.2.2 Determining Dynamic Shift Interval

Suppose d_o is a standard fuzzy interval of inaccurate data, the dynamic shift interval of d_i can be defined on the basis of SCF_i as:

$$\triangle d_i = \frac{(2m-1)d_o}{m} \times SCF_i$$

where (1) $\triangle d_i$ expresses how inaccurate d_i is allowed; (2) $0 < \triangle d_i < 2d_o$; and (3) $\triangle d_i$ is in the direct ratio to SCF_i.

If $m = 1$, then $SCF_i = 1$, and $\triangle d_i = d_o$. In other words, when qualitative correlations among data are not known a priori, $SCF_i = 1$ and $\triangle d_i = d_o$. In this case, the method degenerates to a conventional fuzzy method.

When m is fixed, the more the related data are identified, the greater the SCF_i, therefore the greater the $\triangle d_i$. When SCF_i is fixed, $\triangle d_i$ depends on the number of related data.

Table 4.1 shows the relation among $\triangle d_i$, m and SCF_i.

$\triangle d_i$		m					
		1	10	50	100	500	1000
SCF_i	1	d_o	$1.9000d_o$	$1.9800d_o$	$1.9900d_o$	$1.9980d_o$	$1.9990d_o$
	0.8	/	$1.5200d_o$	$1.5840d_o$	$1.5920d_o$	$1.5984d_o$	$1.5992d_o$
	0.5	/	$0.9500d_o$	$0.9900d_o$	$0.9950d_o$	$0.9990d_o$	$0.9995d_o$
	0.3	/	$0.5700d_o$	$0.5940d_o$	$0.5970d_o$	$0.5994d_o$	$0.5997d_o$
	0.1	/	$0.1900d_o$	$0.1980d_o$	$0.1990d_o$	$0.1998d_o$	$0.1999d_o$

Table 4.1: Relation among $\triangle d_i$, m and SCF_i

The following properties can be drawn from the above formulas.

Property 4.1: *With the same m, the more the related data are identified, the greater the SCF_i; otherwise, the smaller the SCF_i.*

Property 4.2: *With the same m, the greater the SCF_i, the greater the $\triangle d_i$. In other words, the more the related data support d_i, the more widely d_i is allowed to shift.*

Property 4.3: *With the same SCF_i, the greater the m, the less $\triangle d_i$ varies along with m. In other words, the greater the number of related data, the less a single related data item can affect d_i.*

Property 4.2 and Property 4.3 are illustrated in Figure 4.1.

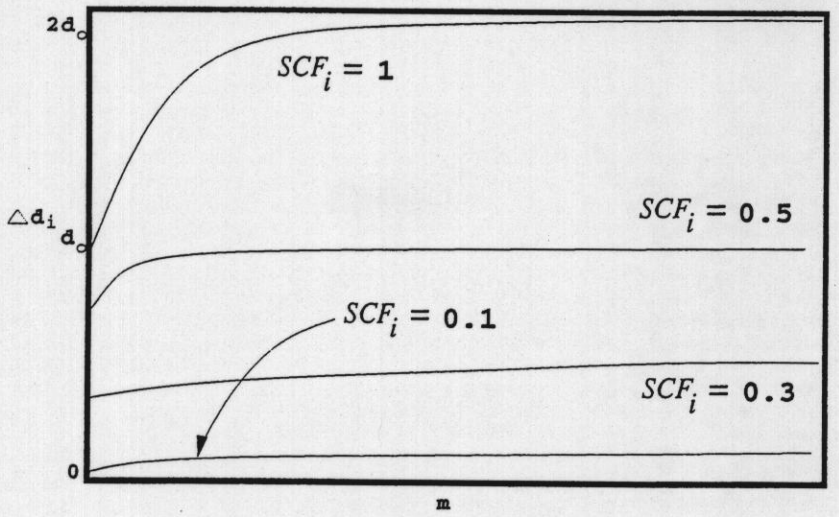

Figure 4.1: $\triangle d_i$ versus m with Different SCF_i

The following properties can also be drawn from the above formulas.

Property 4.4: *$\triangle d_i$ is directly proportional to SCF_i. The slope is equal to, or greater than 1.5, which means that $\triangle d_i$ heavily depends on SCF_i.*

Property 4.5: *Along with the increase of m, the slope increases very slightly. In other words, $\triangle d_i$ depends on the number of the related data which support d_i, rather than the total number of related data.*

Property 4.4 and Property 4.5 are illustrated in Figure 4.2.

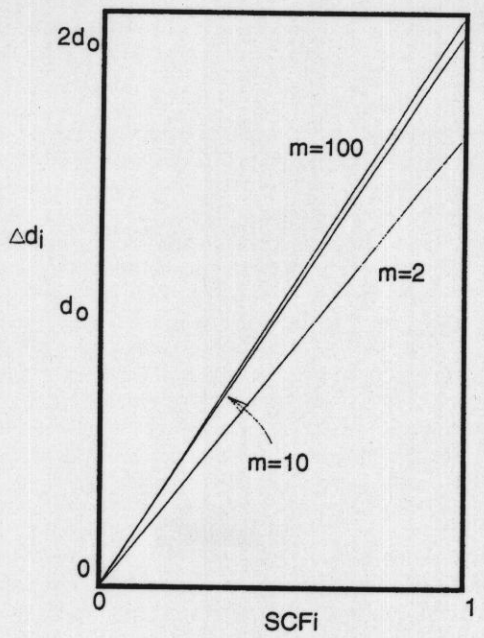

Figure 4.2: $\triangle d_i$ versus SCF_i with Different m

4.2.3 Calculating Value of Predicate "$d_i@R_j$"

The value of "$d_i@R_j$" is equal to the possibility of "$r_{j_p} = d_i$", where $d_i \in MD$, $r_{j_p} \in R_j$, and d_i corresponds to r_{j_p}. The value of "$d_i@R_j$" can be calculated by using the following formula:

$$\mu_i = 1 - \frac{\mid d_i - r_{j_p} \mid}{\triangle d_i}$$

where (1) $\mid d_i - r_{j_p} \mid$ means the real distance between d_i and r_{j_p}; (2) $\triangle d_i$ means the maximum distance between d_i and r_{j_p} which is dynamically determined by SCF_i; and (3) $\mu_i \leq 1$.

At a glance, the representation of μ_i looks like the membership degree of "$r_{j_p} - \triangle d_i \leq d_i \leq r_{j_p} + \triangle d_i$" in fuzzy logic. However, the meaning is completely different, since $\triangle d_i$ is neither provided by domain experts nor determined by quantitative similarity or closeness. Here $\triangle d_i$ is determined on the basis of qualitative correlations among related data. When qualitative correlations among related data are not considered, $\triangle d_i$ is d_o, and the possibility is $1 - \frac{|d_i - r_{j_p}|}{d_o}$. With the consideration of qualitative correlations, the possibility is updated.

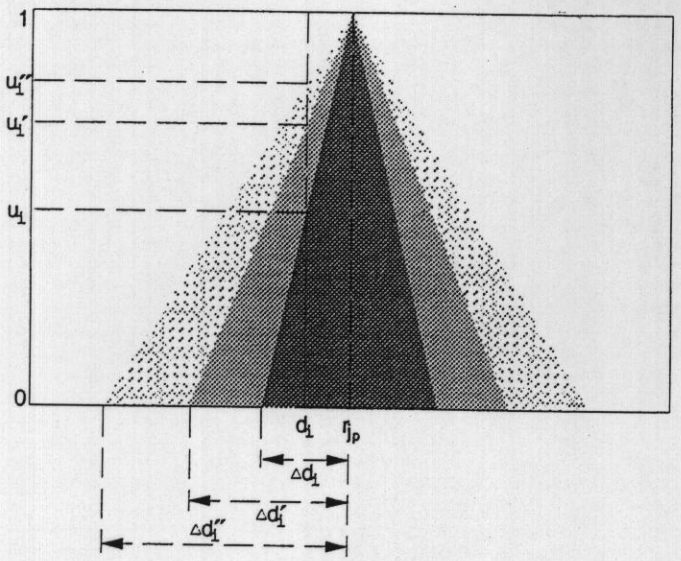

Figure 4.3: Value of "$d_i@R_j$" versus Various $\triangle d_i$

Two new properties can be drawn from the above formula for calculating μ_i.

Property 4.6: *With the same d_i, the greater the $\triangle d_i$, the greater the μ_i. In other words, the wider the dynamic shift interval, the greater the value of "$d_i@R_j$". Formally, if $\triangle d_i'' \geq \triangle d_i' \geq \triangle d_i$, then $\mu_i'' \geq \mu_i' \geq \mu_i$.*

Property 4.7: *SCF_i provides qualitative evidence for accepting or rejecting d_i as r_{j_p}, since μ_i is in the direct ratio to $\triangle d_i$, and $\triangle d_i$ is in the direct ratio to SCF_i.*

Property 4.6 and Property 4.7 are illustrated in Figure 4.3.

4.2.4 Procedure $d_i@R_j$

The above process of calculating the value of "$d_i@R_j$" in Section 4.2.1 to Section 4.2.3 can be expressed by the following procedure.

Procedure $d_i@R_j$

select r_{j_p} *from* R_j;
$SCF_i = 0$;
if $d_i = r_{j_p}$ {
    $SCF_i = 1$;
    $\mu_i = 1$;
}
else {
    for each $r_{j_l} \in R_j$ $(l = 1, ..., m, l \neq p)$ {
        calculate $c_i(d_t)$;
        $SCF_i = SCF_i + c_i(d_t)$;
    }
    $SCF_i = (1 + SCF_i)/m$;
    $\triangle d_i = d_o \times SCF_i \times (2m - 1)/m$;
    $\mu_i = 1 - \mid d_i - r_{j_p} \mid / \triangle d_i$;
}
if $\mu_i > 0$
    return μ_i;
else
    return NIL
end procedure

In the procedure, d_t stands for the data item in MD which corresponds to r_{j_l}. When d_i can be identified with a certain possibility (i.e., $\mu_i > 0$), the procedure returns the value of μ_i ("$\mu_i > 0$" means "T"); otherwise, the procedure returns F.

4.3 Predicate "R_j@MD"

When MD is accurate, "R_j@MD" is equal to "$R_j \subset MD$", where R_j is a set of reference values, and MD is a set of measured data. If all of the m reference values in R_j can be identified from MD, then $R_j@MD = T$; otherwise $R_j@MD = F$.

When MD is inaccurate, however, "R_j@MD" means that R_j is possibly (qualitatively) a subset of MD. The value of "R_j@MD" is not T or F, but the possibility that all the reference values in R_j can be identified from MD.

4.3.1 Calculating Value of Predicate "R_j@MD"

The value of "R_j@MD" is equal to the possibility of "$R_j \subset MD$". Suppose there are m reference values in R_j, the value of "R_j@MD" reflects the possibilities of all of the m reference values being identified.

μ_1, μ_2, ... and μ_m can be respectively calculated by using the algorithm presented in Section 4.2.4. If $\mu_l > 0$ ($l = 1, 2, ..., m$), then R_j can be regarded as a subset of MD with a certain possibility. Let s_1, s_2, ..., and s_m be the priorities of the reference values in R_j which are usually determined by domain experts, then the value of "R_j@MD" can be calculated based on μ_1, μ_2, ..., and μ_m by using the following formula:

$$R_j@MD = \frac{\sum_{l=1}^{m} s_l \times \mu_l}{\sum_{l=1}^{m} s_l}$$

where (1) $s_l > 0$; and (2) $\mu_l > 0$.

To most problems, s_1, s_2, ... and s_m are the same. To some problems, however, there may be a priority chain among related data. For example, in infrared spectrum interpretation, the peaks of CH_3 in frequency section 2800 cm^{-1} to 3000 cm^{-1} are more distinct than its peaks elsewhere. Therefore, the peaks in this section are prior to those located in other sections. In the case of problems where some data are prior to others, s_1, s_2, ... and s_m represent the priorities of the data.

4.3.2 Procedure $R_j@MD$

Suppose μ_i has been calculated by using **procedure** $d_i@R_j$, then the process of calculating the value of "R_j@MD" can be expressed by the following simple procedure.

Procedure $R_j@MD$

$P = s_i \times \mu_i;$
$S = s_i;$
$for\ l = 1\ to\ m\ (l \neq p)\{$
$\mu_l = d_t@R_j;$
$if\ \mu_l > 0\{$
$P = P + s_l \times \mu_l;$
$S = S + s_l;$
$\}$
$else\{$
$P = 0;$
$exit;$
$\}$
$\}$
$if\ P > 0$
$return\ P/S;$
$else$
$return\ NIL$

end procedure

In the procedure, d_t stands for the data item in MD which corresponds to r_{j_l}. When R_j can be identified as a subset of MD with a certain possibility (i.e., P/S), the procedure returns the value of P/S ("$P/S > 0$" means "T"); otherwise, the procedure returns F.

4.4 Algorithm for Qualitatively Interpret Inaccurate Data

In this section, I first present the method for interpreting inaccurate data based on qualitative correlations among related data with the form of algorithm, then I discuss and analyze the algorithm.

4.4.1 Algorithm

I give the following algorithm for interpreting/analyzing measured data based on **procedure** $d_i@R_j$ and **procedure** $R_j@MD$. When measured data are not accu-

rate, the algorithm can identify inaccurate data items by considering qualitative correlations among related data.

Algorithm *Making-Qualitative-Hypotheses*

$IN(MD) = \emptyset$;
for $i = 1$ to n {
  for $j = 1$ to k {
    $P(R_j) = 0$;
    if $d_i@R_j$ (*i.e.*, **Procedure** $d_i@R_j$)
      if $R_j@MD$ (*i.e.*, **Procedure** $R_j@MD$) {
        $R_j \rightarrow IN(MD)$;
        $P(R_j) = R_j@MD$;
      }
      end if
    end if
  }
  end for
}
end for

end algorithm

In the algorithm, $P(R_j)$ represents the value of "$R_j@MD$". The algorithm is actually the realization of the logic expression: $\{d_i@R_j, R_j@MD\} \Rightarrow \{R_j \in IN(MD)\}$ which has been discussed in Chapter 2.

4.4.2 Analysis

I discuss and analyze the algorithm in the following two aspects:

1. *Validity of the algorithm*: The algorithm uses qualitative correlations among related data as confirmatory or disconfirmatory evidence. In general, data in a dataset are rarely completely independent, so qualitative correlations among related data can always be extracted, represented and used with the algorithm. In the specific cases that qualitative correlations among related data are not known a priori, the algorithm degenerates to a traditional fuzzy method;

2. *Complexity of the algorithm*: For each measured data item in MD ($MD = \{d_1, d_2, ..., d_n\}$), the algorithm searches RV ($RV = \{R_1, R_2, ..., R_k\}$) once. For each R_j ($R_j = \{r_{j_1}, r_{j_2}, ..., r_{j_m}\}$), the algorithm checks other n-1 measured data items for m times, and other m-1 reference values for n times. Therefore, with blind search, the number of operations is about (at worst): $n \times k \times [m \times (n-1) + n \times (m-1)] = (2m-1)kn^2 - kmn$. Since k and m are two constants, the complexity of the algorithm is $O(n^2)$.

4.5 Discussion

In the above sections of this chapter, a method for interpreting inaccurate data by considering qualitative correlations among related data as confirmatory or disconfirmatory evidence was proposed. In this section, I discuss the method in more detail.

4.5.1 Intuition

The intuition behind the method can be summarized in the following three aspects.

1. The idea is very common in human thinking. When all data except blood pressure of a patient show that the patient has a certain disease, we would naturally suspect that the blood pressure of the patient was inaccurately entered. Similarly, when all peaks except one indicate that a partial component is present, we would naturally suspect that the unmatched peak was inaccurately measured or the peak was affected by noise or something else. If acceptable solutions can be made by assuming an inaccurate data item to be a reference value based on qualitative correlations between the data item and its related data, the inaccurate data item may be compensated and hence identified;

2. In practical problems, the idea is commonly used by domain experts when inaccuracy occurs. In infrared spectrum interpretation, for example, spectroscopists frequently use the qualitative analysis like

   If there is a strong peak around 3000-3100 cm^{-1}, then the unknown spectrum may be partially created by benzene-rings —— check peaks around 1650, 1550 and 700-900 cm^{-1} to make sure

since a benzene-ring may have other peaks there at the same time. Or

If there is a sharp peak in 2950-2960 cm^{-1}, and the peaks around 1500-1600 cm^{-1} look like the peaks of CO, then the peak in 2950-2960 cm^{-1} is likely to be the peak of CH_3 even if it is not strong.

So if spectral data are inaccurate (i.e., some measured peaks look like but are not exactly the same as reference peaks), considering qualitative correlations among related data may lead to qualitative evidence for the identification of inaccurate data. For example, suppose the strength of a peak is slightly different from the reference value, but both the frequency and shape of the peak are the same as the reference values, then the strength of the peak may still be identified since both of its related data, frequency and shape, support it. Similarly, if peaks at low frequency sections are inaccurate, considering related peaks at high frequency sections may help identify these peaks;

3. The qualitative correlations among related data are similar to evidence of inference, and identifying an inaccurate data item is similar to making hypotheses of inference. So to find qualitative correlations among related data is actually to search for evidence for hypotheses, and to propagate probabilities on inference networks. The similar idea has been widely accepted in evidence theory and probabilistic reasoning. However, the difference lies in that the proposed method dynamically calculates the values of "$d_i@R_j$" and "$R_j@MD$", so it does not need many assumptions in advance, and can avoid inconsistency in knowledge and data bases as well. Concerning this point, more descriptions are available in the following chapter.

4.5.2 Applicability

Data in a dataset are rarely completely independent in general, so qualitative correlations among related data can always be used as evidence of interpreting inaccurate data. Therefore, the method can be applied to many science and engineering problems.

For example, in infrared spectrum interpretation, data concerning a single peak (i.e., the frequency position, strength and width of the peak) are related to each other. If one of them can not be identified due to its inaccuracy, but all others can be completely identified, then the inaccurate data item may be supported

and compensated, since related data should be present or absent simultaneously. The more the related data support the inaccurate data item, the more it is compensated. Analogously, data concerning the pattern of a partial component (i.e., the peaks created by the partial component) are also related to each other. If one peak can not be identified due to its inaccuracy, but all other peaks can be completely identified, then the inaccurate peak may be compensated and hence identified.

The method can be easily employed to solve the problems such as infrared spectrum interpretation. As a matter of fact, I have done many experiments on applying the method to the problem of infrared spectrum interpretation. Later in Chapter 6, I will introduce the application of the method to the problem, and give corresponding examples.

4.5.3 Comparison

In determining a dynamic shift interval for an inaccurate data item, the method is similar to fuzzy method [Zadeh, 1978]. The differences between the method and conventional fuzzy methods include:

1. The method dynamically determines shift intervals for inaccurate data, while fuzzy intervals for inaccurate data in conventional fuzzy methods are usually provided and fixed by domain experts in advance. Consequently, much more information can be used in determining how inaccurate an inaccurate data item is allowed to be;

2. The method determines dynamic shift intervals on the basis of qualitative correlations among related data, while fuzzy intervals in conventional fuzzy methods are solely determined in general. Consequently, the interpretation of inaccurate data can be more reliable.

In the case of problems where qualitative correlations among related data exist, the method is better than conventional fuzzy methods. In the case of problems where qualitative correlations among related data are not known a priori, the method degenerates to conventional fuzzy methods.

Using qualitative correlations among related data to determine the possibility of interpreting inaccurate data is a little bit similar to using evidence to update the probabilities of hypotheses in Bayesian methods [Duda, and et al, 1977]. The advantage of the method over traditional Bayesian methods is that it needs few assumptions in advance, while traditional Bayesian methods usually need

many numbers provided by domain experts in advance, such as the degree that a piece of evidence is sufficient for a hypothesis and the degree that a piece of evidence is necessary for a hypothesis. I will further compare the use of qualitative correlations among related data and the probability propagation in Bayesian methods in Chapter 5.

4.6 Summary

In this chapter, I presented a method for interpreting inaccurate data by considering qualitative correlations among related data as confirmatory or disconfirmatory evidence.

I discussed the way of extracting, representing and calculating qualitative correlations among related data, and the way of representing and calculating SCF of related data. SCF reflects the total qualitative correlations between an inaccurate data item and its related data, and provides either confirmatory or disconfirmatory evidence of interpreting the inaccurate data item. Then, I proposed an approach to calculating dynamic shift intervals of inaccurate data based on SCF which dynamically determines how wide an inaccurate data item is allowed to shift, and an approach to calculating possibility of identifying inaccurate data in the dynamic shift intervals. Finally, based on the above two approaches, I presented an algorithm for interpreting inaccurate data by considering qualitative correlations among related data as confirmatory or disconfirmatory evidence. I also briefly analyzed the applicability and complexity of the algorithm, and gave the discussion about the intuition behind the method, the applicability of it, and the comparison of it with other similar methods.

Chapter 5

Propagation of Qualitative Correlations

In this chapter, I present a novel method for extracting, representing and propagating qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence of uncertain reasoning. Part of the method is based on some extensions of the method presented in Chapter 4. The extensions allow qualitative correlations among related data to propagate which enable the possibilities of inaccurate data being interpreted to be updated by considering qualitative correlations among related data. as confirmatory or disconfirmatory evidence. The function of the extended method is similar to the probability propagation on Bayesian networks [Duda, Hart, & Nilsson, 1976], but the advantages of the method are evident. First, it can be applied to the problems where evidence is not explicitly given, or is not available. Second, fewer numbers and assumptions need to be provided by domain experts in advance, since both the degree that a piece of evidence enhances a hypothesis (i.e., LS in Subjective Bayesian Methods) and the degree that a piece of evidence does not enhance a hypothesis (i.e., LN in Subjective Bayesian Methods) are dynamically calculated from qualitative correlations among related data, rather than are provided by domain experts. Third, the knowledge acquisition procedure is simpler, and the inconsistency in knowledge bases can be avoided, since most numbers are dynamically calculated in the interpreting process, rather than are obtained from domain experts beforehand. As an additional result, the method can be used to qualitatively interpret inaccurate symbolic data as well as inaccurate numerical data.

5.1 Introduction

Bayesian inference has been proved to be effective for reasoning under uncertainty, and has been used in many AI systems [Dempster, 1968][Kleiter, 1992]. However, the problems of using Bayesian inference are: (1) evidence must be explicitly provided, and the relation between evidence and hypothesis must be explicitly provided too. Unfortunately, in many cases, evidence and the relation between evidence and hypothesis are not always available. In addition, after evidence has been considered and hypotheses have been made, it is still possible to refine the hypotheses by using other knowledge and information; (2) Bayesian inference requires many statistical numbers in advance. Unfortunately, in many practical problems, it is impossible to have all numbers provided beforehand. Although subjective Bayesian methods provide a practical framework for using subjective statements or assumptions to take the place of statistical data when they are insufficient or absent, the problems still remain since subjective statements are not always available and the inconsistency in knowledge bases is hard to avoid [Duda, Hart, & Nilsson, 1976].

Chapter 4 provided a method for interpreting inaccurate data by considering qualitative correlations among related data, and discussed the effectiveness of using qualitative correlations among related data as confirmatory or disconfirmatory evidence in interpreting inaccurate data. The qualitative correlations among related data are always available within a dataset, so the method is applicable in general cases.

Since there are qualitative correlations among related data, if the possibilities of inaccurate data in a dataset have been calculated or provided, then the qualitative correlations among related data may be used as confirmatory or disconfirmatory evidence to update the known possibilities of inaccurate data.

In addition, in most cases, data in a dataset are numerical data, such as the position, strength and width of a peak, the length and frequency of a signal, and the voltage and current intensity at an electric circuit. Numerical data are objects that the method presented in Chapter 4 concerns. From a wider scope, however, data in a dataset may also be in other forms, such as symbols and letters. For example, when the strength of a peak is concerned, the strength is a numerical data item. On the other hand, when a peak is concerned, the peak is a symbolic data item. Similarly, a letter in a word and a word in a sentence are both symbolic data items.

No matter whether data are in numerical or symbolic forms, there are the following preliminaries concerning qualitative correlations:

1. Some data items in a dataset are usually qualitatively related to each other;
2. There are qualitative correlations among related data;
3. Qualitative correlations among related data can provide confirmatory or disconfirmatory evidence of interpreting inaccurate data.

Based on the preliminaries, the method presented in Chapter 4 can be extended to a new form which enables qualitative correlations to propagate among related data, and at the same time, can be used to interpret inaccurate symbolic data.

What is more important, on the basis of the extensions, a new method can be developed for uncertain reasoning.

In the following sections of this chapter, I first briefly introduce the probability propagation on Bayesian networks, then discuss some extensions of the method presented in Chapter 4. Based on the extensions, I present a novel method for uncertain reasoning. The method automatically extracts, represents and propagates qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence to update the possibilities of the hypotheses. The function of propagating qualitative correlations and updating possibilities of hypotheses in the proposed method is similar to the function of propagating and updating the probabilities of hypotheses in Bayesian inference. But unlike Bayesian inference, the proposed method automatically obtains and uses qualitative correlations among hypotheses as qualitative evidence, so the above problems of using Bayesian inference can be easily avoided. I put forward a new concept called qualitative correlations among hypotheses and a new concept called qualitative correlation propagation. Then I propose an algorithm for extracting and representing qualitative correlations among hypotheses and an algorithm for propagating qualitative correlations and updating possibilities of hypotheses respectively. After presenting the method, I give an example to demonstrate its applications, and compare it with other similar ones.

5.2 Probability Propagation on Bayesian Networks

Bayesian methods provide mathematical fundamentals for updating probabilities with conditional probabilities [Dempster, 1968][Kleiter, 1992]. Subjective Bayesian methods are proved to be useful in using subjective statements to take

the place of statistics of data or evidence when statistical samples are insufficient or absent [Duda, Hart, & Nilsson, 1976]. Many systems use subjective Bayesian methods to handle inaccuracy and uncertainty [Duda, and et al, 1977][Ramer & Lander, 1991].

Subjective Bayesian methods update probabilities by calculating the probability propagation for the inference rule with the following form:

$$IF \ \{E, P(E)\} \ then \ \{H, P(H)\} \ \{LS, LN\}$$

where

1. E is a piece of evidence;
2. $P(E)$ is the probability that E is true;
3. H is a hypothesis;
4. $P(H)$ is the probability that H is true;
5. LS represents the degrees that E enhances H, $LS = \frac{P(E|H)}{P(E|\neg H)}$;
6. LN represents the degrees that $\neg E$ enhances H, $LN = \frac{P(\neg E|H)}{P(\neg E|\neg H)}$.

$P(E)$, $P(H)$, LS and LN are all provided by domain experts in advance. When E is obtained, $P(H)$ is updated with two conditional probabilities:

$$P(H \mid E) = \frac{LS \times P(H)}{(LS-1)P(H)+1} \qquad and$$

$$P(H \mid \neg E) = \frac{LN \times P(H)}{(LN-1)P(H)+1}$$

The final conditional probability of H can be calculated with the following formula:

$$P(H \mid S) = P(H) + \frac{P(H \mid E) - P(H)}{1 - P(E)} \times [P(E \mid S) - P(E)]$$

where S stands for relevant observations.

The above process can be graphically represented in Figure 5.1.

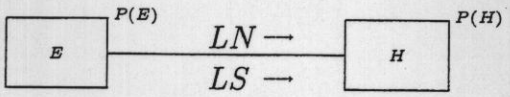

Figure 5.1: Probability Propagation

If an inaccurate data item is viewed as a hypothesis, and its related data are viewed as pieces of evidence, then the form of subjective Bayesian methods is similar to the method presented in Chapter 4. However, the problems of using subjective Bayesian methods are:

1. The degree that a piece of evidence supports or does not support a hypothesis is determined in advance, therefore, useful dynamic information can not be used properly;

2. Many assumptions and numbers, such as LS and LN, are needed, therefore, the knowledge acquisition procedure is very tedious in general;

3. The inconsistency in knowledge bases can hardly be avoided.

5.3 Propagation of Qualitative Correlations

I have discussed that there are qualitative correlations among related data, and that qualitative correlations among related data can be used as confirmatory or disconfirmatory evidence for interpreting inaccurate data. I presented a method to crystallize the use of qualitative correlations among related data as confirmatory or disconfirmatory evidence in interpreting numerical inaccurate data in Chapter 4. When possibilities of inaccurate data have been known, qualitative correlations among related data can be used to update the known possibilities as confirmatory or disconfirmatory evidence. Or in other words, when hypotheses have been made with corresponding possibilities, qualitative correlations among hypotheses can be used as evidence of uncertain reasoning.

5.3.1 Two Extensions

With the following two extensions, the method presented in Chapter 4 can propagate qualitative correlations among related data, and can use qualitative correlations among related data as confirmatory or disconfirmatory evidence to interpret inaccurate symbolic data. Based on the extensions, I propose a new method for uncertain reasoning:

1. In Chapter 4, I used d_o to express the standard fuzzy region of inaccurate numerical data. d_o means how much a numeral is allowed to shift, and can actually be viewed as the qualitative correlation between a data item and itself. To the problems where the possibilities of inaccurate data have been known and the problems where the datasets consist of inaccurate symbolic data, the qualitative correlation between a data item and itself can be viewed as 1. In these cases, $d_o = 1$;

2. In Chapter 4, I used d_i to express an inaccurate data item, and r_{j_p} to express a reference value corresponding to d_i. Then, the possibility of d_i being interpreted as r_{j_p} is:

$$\mu_i = 1 - \frac{\mid d_i - r_{j_p} \mid}{\triangle d_i}$$

When the possibility of an inaccurate data item has been known as μ_i^o, or d_i is an inaccurate symbolic data item, "$\mid d_i - r_{j_p} \mid$" is "$1 - \mu_i^o$".

5.3.2 Concepts and Definitions

In inference, hypotheses are rarely completely independent. A group of hypotheses may refer to the same object, i.e., different hypothesis refers to the different aspect of the object. For example, in medical diagnosis, several symptoms may refer to the same hypothesis, and several hypotheses may refer to the same disease. The hypotheses referring to the same disease may have some qualitative correlations.

Due to the qualitative correlations among hypotheses, some hypotheses may be related to each other. If the possibilities of hypotheses have been calculated or provided, then the related hypotheses may be used as confirmatory or disconfirmatory evidence to update the known possibilities of hypotheses.

The following concepts are given to define the related hypotheses, the qualitative correlations among related hypotheses, and the confirmatory and disconfirmatory evidence from the qualitative correlations among related hypotheses.

Definition 5.1 *Related hypotheses: If hypotheses h_1, h_2, ..., and h_m refer to the same object, then they can be treated as related hypotheses.*

I use rh_α to represent a group of related hypotheses, and use $h_i \& h_j$ to represent that h_i is related to h_j, then $rh_\alpha = \{h_k \mid \forall h_{k'}((h_{k'} \in rh_\alpha) \wedge (h_{k'} \neq h_k) \rightarrow (h_k \& h_{k'}))\}$.

For example, the hypotheses describing various features of a criminal are related hypotheses since they refer to the same criminal. If we consider a string as an object, then the hypotheses for all letters in the string can be viewed as related hypotheses. If we consider a sentence as an object, then the hypotheses for all words in the sentence can also be viewed as related hypotheses.

Definition 5.2 *Qualitative correlations among related hypotheses: If h_i and h_j are two related hypotheses (i.e., $h_i \& h_j$), then the great possibility of h_i qualitatively enhances h_j, and the small possibility of h_i qualitatively depresses h_j. This kind of effects among related hypotheses are called qualitative correlations among related hypotheses.*

I use qc_i^j to represent the qualitative correlations of h_j to h_i. The greater the possibility of h_i, the more greatly h_i qualitatively supports h_j, and the smaller the possibility of h_i, the more weakly it qualitatively supports h_j (or in other words, the more greatly it qualitatively depresses h_j).

The main effectiveness of qualitative correlations among hypotheses includes:

1. Great qualitative correlations imply strong support among hypotheses which are related to each other. The greater the qualitative correlations, the stronger the support;

2. Small qualitative correlations imply weak support among hypotheses which are related to each other (or in other words, small qualitative correlations imply strong depression among related hypotheses). The smaller the qualitative correlations, the weaker the support (or in other words, the smaller the qualitative correlations, the stronger the depression).

For example, some symptomatic hypotheses such as temperature, blood pressure and pulse refer to a certain disease. In describing the disease, all of these

hypotheses are related to each other. Making a definite diagnosis of a certain disease for a patient, all related hypotheses referring to the disease should be completely confirmed. So in practice, if some symptomatic hypotheses of a patient presage a certain disease, the other hypotheses related to these hypotheses which refers to the disease are usually required to be made and confirmed.

Consider the infrared spectrum interpretation. The hypotheses that some peaks are created by a certain partial component are related to each other. Because all peaks that a partial component can create should be present or absent simultaneously, identifying a partial component requires that all these related hypotheses be made and confirmed. As a result, making some hypotheses with great possibilities may prompt and enhance making other hypotheses that are related to these ones.

Definition 5.3 *Sum-degree of qualitative correlations among related hypotheses: If there are $m-1$ hypotheses related to h_i, then the Sum-degree of qualitative correlations among related hypotheses of h_i reflects the total qualitative correlations between h_i and all of its related hypotheses.*

I use SD_i to represent the Sum-degree of the qualitative correlations between h_i and its related hypotheses.

For example, consider the following four strings again:

(a) *i-n-a-c-c-u-r-a-t-e*

(b) *i-m-a-c-c-u-r-a-t-e*

(c) *i-m-a-c-c-u-l-a-t-e*

(d) *i-m-a-c-u-l-a-t-e*

(a) is a correct word, but (b), (c) and (d) are all spelled wrong. For a wrong spelled word, determining what a letter in the word should be is to make a hypothesis for the letter. Because all letters in a word refer to the same word, the hypotheses for these letters are related to each other, and have qualitative correlations each other. There is one letter, "*m*", in (b) different from that in (a), so the Sum-degree of qualitative correlations from other letters which support interpreting "*m*" in (b) as "*n*" is quite great. As a result, (b) can be easily interpreted as (a). Further, there are two letters, "*m*" and "*l*", in (c) different from those in (a). Although (c) may be interpreted as (a), the Sum-degree of qualitative correlations from other letters which support interpreting "*m*"and "*l*" in (c) as "*n*"and "*r*" would not be great. Finally, there are three letters,

"*m*", "*c*" and "*l*", in (d) different from those in (a). As a result, (d) will hardly be interpreted as (a) since the Sum-degree of qualitative correlations from other letters which support interpreting these three letters will be very small (even smaller than that for interpreting (d) as word "*immaculate*").

The principle for defining and calculating the Sum-degree of qualitative correlations among related hypotheses is that the qualitative correlations among related hypotheses should reflect the ratio of how many and how much related hypotheses qualitatively support each other.

My method bases on the above two definitions. The method extracts qualitative correlations among known hypotheses, and propagates them among related hypotheses as confirmatory or disconfirmatory evidence, and updates the possibilities of the hypotheses.

Definition 5.4 *Confirmatory evidence from qualitative correlations among related hypotheses: If the Sum-degree of qualitative correlations among related hypotheses of h_i is greater than a certain value given by domain experts, then the qualitative correlations among related hypotheses provide confirmatory evidence for h_i. As a result, the possibility of h_i may increase.*

Definition 5.5 *Disconfirmatory evidence from qualitative correlations among related hypotheses: If the Sum-degree of qualitative correlations among related hypotheses of h_i is smaller than a certain value given by domain experts, then the qualitative correlations among related hypotheses provide disconfirmatory evidence for h_i. As a result, the possibility of h_i may decrease.*

For example, partial component CH_3 usually creates numerous peaks each of which should have an exact location on infrared spectra. Because the peaks of CH_3 on real infrared spectra are always inaccurate, especially the peak located at 2900 cm^{-1}, real peaks on infrared spectra can not be directly identified as the peaks of CH_3. Instead, hypotheses need to be made to assume the similar peaks to be those of CH_3. Suppose a peak around 2900 cm^{-1} is assumed to be the peak of CH_3. Since CH_3 can create many peaks besides that at 2900 cm^{-1}, the qualitative correlations among the hypotheses for these peaks created by CH_3 can be used as confirmatory evidence to enhance the hypothesis for the peak around 2900 cm^{-1}, or as disconfirmatory evidence to depress the hypothesis. For instance, if other hypotheses all have very great possibilities, then these hypotheses tend to support the hypothesis for the peak around 2900 cm^{-1}, and the Sum-degree of qualitative correlations among related hypotheses of the peak around 2900 cm^{-1} will be very great. As a result, the possibility of identifying the

peak around 2900 cm^{-1} will increase, that is, the possibility of the hypothesis for the peak will be updated with a greater one. Similarly, if the related hypotheses all have quite small possibilities, then they tend to depress the hypothesis for the peak around 2900 cm^{-1}, and the Sum-degree of qualitative correlations among related hypotheses of the peak will be quite small. As a result, the possibility of identifying the peak around 2900 cm^{-1} will decrease, that is, the possibility of the hypothesis for the peak will be updated with a smaller one.

5.4 Method for Qualitative Correlation Propagation

I have discussed that there are qualitative correlations among related hypotheses, and that qualitative correlations among related hypotheses can be propagated as confirmatory or disconfirmatory evidence for reasoning under uncertainty. In this section, I present a method to realize the use of qualitative correlations among related hypotheses as confirmatory or disconfirmatory evidence. First, I present an algorithm for extracting and representing qualitative correlations among hypotheses. Then, I present another algorithm for propagating qualitative correlations and for updating the possibilities of hypotheses by propagating qualitative correlations among related hypotheses as confirmatory or disconfirmatory evidence.

5.4.1 Algorithm for Obtaining Qualitative Correlations

I describe the algorithm for obtaining qualitative correlations among hypotheses with the following steps.

Step 1: Grouping related hypotheses

Suppose the known hypotheses are h_1, h_2, ..., and h_n which form a hypothesis set H. If some hypotheses in H refer to the same object, they can be treated as related hypotheses. Therefore, H is divided into some subsets, that is,

$$H = \{h_1, h_2, ..., h_n\} = rh_1 \cup rh_2 \cup ... \cup rh_k$$

where (1) $rh_i = \{h_{i_p} \mid h_{i_p} \in H \wedge \forall h_{i_q}(h_{i_q} \in rh_i \wedge h_{i_q} \neq h_{i_p} \rightarrow h_{i_p} \& h_{i_q})\}$,

(2) $rh_i \neq \emptyset$, and

(3) $rh_i \cap rh_j = \emptyset$ or $rh_i \cap rh_j \neq \emptyset$, $i \neq j$.

Step 2: Extracting qualitative correlations among related hypotheses

For each subset of H, i.e.,

$$rh_i = \{h_{i_1}, h_{i_2}, ..., h_{i_m}\}, \ i = 1, 2, ..., k,$$

suppose the corresponding set of possibilities is

$$\mu^o_i = \{\mu^o_{i_1}, \mu^o_{i_2}, ..., \mu^o_{i_m}\}, \ i = 1, 2, ..., k.$$

The principle for defining the qualitative correlation between two related hypotheses is that if the possibility of a hypothesis is greater than a certain value, then it is qualified to qualitatively support its related hypotheses; otherwise, it is not qualified. For example, suppose 0.5 is given as the certain value, then

$$qc^{i_q}_{i_p} = \begin{cases} 1 & h_{i_p} \in rh_i \wedge h_{i_q} \in rh_i \wedge \mu^o_{i_q} \geq 0.5 \\ 0 & h_{i_p} \in rh_i \wedge h_{i_q} \in rh_i \wedge \mu^o_{i_q} < 0.5 \end{cases}$$

where $qc^{i_q}_{i_p} = 1$ means that h_{i_p} is qualitatively supported by h_{i_q}, and $qc^{i_q}_{i_p} = 0$ means that h_{i_p} is not qualitatively supported by h_{i_q}.

Step 3: Calculating Sum-degree of qualitative correlations

There are m hypotheses in rh_i related to each other, so the Sum-degree of qualitative correlations of h_{i_p} is calculated by considering $qc^{i_l}_{i_p}$, $l = 1, 2, ..., m$ and $l \neq p$, that is,

$$SD_{i_p} = \frac{1 + \sum_{l=1, l \neq p}^{m} qc^{i_l}_{i_p}}{m}$$

where $0 < SD_{i_p} \leq 1$.

SD_{i_p} expresses the total qualitative correlations between h_{i_p} and all of its related hypotheses. If $m = 1$, then $SD_{i_p} = 1$. When $m > 1$, SD_{i_p} is in the direct ratio to the number of the related hypotheses in rh_i which are qualified to qualitatively support their related hypotheses.

5.4.2 Algorithm for Propagating Qualitative Correlations

I describe the algorithm for propagating qualitative correlations among hypotheses to update the possibilities of hypotheses with the following steps.

Step 1: Calculating possibility propagation factor

I use $P^2_{i_p}$ to represent the possibility propagation factor of h_{i_p} from all of its related hypotheses, and calculate $P^2_{i_p}$ by considering the confirmatory or disconfirmatory evidence obtained from SD_{i_p}:

$$P^2_{i_p} = \frac{(2m-1) \times SD_{i_p}}{m}$$

where $SD_{i_p} < P^2_{i_p} < 2SD_{i_p}$.

$P^2_{i_p}$ is in the direct ratio to SD_{i_p}. If $m = 1$, then $SD_{i_p} = 1$, and $P^2_{i_p} = 1$. In other words, when qualitative correlations among hypotheses are not available, $SD_{i_p} = 1$ and $P^2_{i_p} = 1$. This is the only case to which the method is not applicable. However, in practical problems, qualitative correlations among hypotheses are always available, so the method can always be applied.

When m is fixed, the greater the number of related hypotheses which are made with great possibilities, the greater the SD_{i_p}, therefore the greater the $P^2_{i_p}$. When SD_{i_p} is fixed, $P^2_{i_p}$ depends on the number of related hypotheses.

Step 2: Propagating qualitative correlations as (dis)confirmatory evidence

The possibility of h_{i_p} has been known as $\mu^o_{i_p}$, and the possibility propagation factor of h_{i_p} has been calculated as $P^2_{i_p}$. So a new possibility of h_{i_p} can be calculated with the following formula after using qualitative correlations among related hypotheses as confirmatory or disconfirmatory evidence:

$$\mu_{i_p} = 1 - \frac{1 - \mu^o_{i_p}}{P^2_{i_p}}$$

where $0 \leq \mu_{i_p} \leq 1$.

The function of SD_{i_p} is similar to the function of LS and LN in Subjective Bayesian methods. However, both SD_{i_p} and $P^2_{i_p}$ are automatically calculated by considering qualitative correlations among related hypotheses, while in Subjective Bayesian methods both LS and LN are provided by domain experts in

advance. In addition, using $P^2_{i_p}$ to calculate μ_{i_p} is also similar to using evidence to update the probabilities of hypotheses. However, the method is applicable to any problem where hypotheses have been made with corresponding possibilities, while Subjective Bayesian methods are usually applicable to the problems where evidence and the relations between evidence and hypotheses are explicitly provided.

Step 3: Updating possibilities of hypotheses

For $h_{i_l} \in rh_i$ $(l = 1, 2, ..., m)$, if μ_{i_l} exists, then μ_{i_l} is used to replace $\mu^o_{i_l}$.

5.5 An Example

I discuss the application of the proposed method through the following example.

Figure 5.2 shows the peak of partial component CH_3 in 3000 cm^{-1} to 2900 cm^{-1}. The accurate peak of CH_3 in this region should be a strong peak located at 2960 cm^{-1}, but in this example, the real peak of CH_3 is only a medium peak located at 2918 cm^{-1}. By considering the peak itself, the possibility that the real peak at 2918 cm^{-1} is identified as the peak of CH_3 at 2960 cm^{-1} is 0.352 [Zhao & Nishida, 1995a], that is,

$$\mu^o_{p_{2960}} = 0.352.$$

CH_3 can mainly create 4 peaks, p_{2960}, p_{2870}, p_{1460} and p_{1380}. These peaks are related to each other.

If CH_3 is contained by the real infrared spectrum, then all peaks of CH_3 should be identified. Therefore, if other peaks are all identified with great possibilities, CH_3 is quite likely to be contained by the real infrared spectrum, and

the other peak will qualitatively support the identification of the inaccurate peak at 2918 cm^{-1} as the peak at 2960 cm^{-1}.

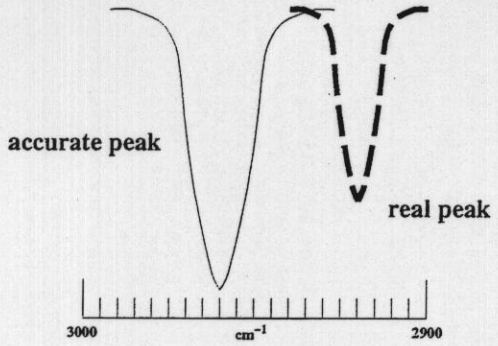

Figure 5.2: Peaks of CH_3

Suppose the possibilities of other peaks are obtained with the same method:

$$\mu^o_{p_{2870}} = 0.850, \quad \mu^o_{p_{1460}} = 0.921 \quad and \quad \mu^o_{p_{1380}} = 0.975.$$

According to the proposed method, the qualitative correlations between two related peaks are respectively calculated as:

$$qc^{p_{2870}}_{p_{2960}} = 1, \quad qc^{p_{1460}}_{p_{2960}} = 1 \quad and \quad qc^{p_{1380}}_{p_{2960}} = 1.$$

Then

$$SD_{p_{2960}} = 1, \quad and \quad P^2_{p_{2960}} = \frac{2 \times 4 - 1}{4} \times 1 = 1.75.$$

So

$$\mu_{p_{2960}} = 1 - \frac{1 - 0.352}{1.75} = 0.629.$$

Therefore, the possibility that the real peak at 2918 cm^{-1} is identified as peak p_{2960} of CH_3 increases from 0.352 to 0.629 due to the qualitative correlations among related hypotheses. CH_3 can not be identified before since one main peak of it can not be found from the real infrared spectrum (i.e., $\mu^o_{p_{2960}} < 0.5$). After

considering qualitative correlations among related hypotheses as confirmatory evidence, CH_3 can be identified with considerably great possibility (i.e., $\mu^o_{p2960} = 0.629$).

The above process is similar to the probability propagation in probabilistic reasoning. However, neither evidence nor relation between evidence and hypotheses is required beforehand.

5.6 Discussion

In Section 5.4, I presented a new method for uncertain reasoning. In this section, I discuss the properties of the method, and compare it with related work.

5.6.1 Properties

The following main properties of the method can be drawn.

Property 5.1: *With the same number of related hypotheses, m, the greater the number of related hypotheses whose possibilities are greater than a certain value provided by domain experts, the greater the SD_{i_p}; otherwise, the smaller the SD_{i_p}.*

Property 5.2: *With the same m, the greater the SD_{i_p}, the greater the $P^2_{i_p}$.*

Property 5.3: *With the same SD_{i_p}, the greater the m, the less $P^2_{i_p}$ varies along with m.*

Property 5.4: *With the same $\mu^o_{i_p}$, the greater the $P^2_{i_p}$, the greater the μ_{i_p}.*

Property 5.5: *SD_{i_p} provides qualitative confirmatory or disconfirmatory evidence for h_{i_p} since μ_{i_p} is in the direct ratio to $P^2_{i_p}$, and $P^2_{i_p}$ is in the direct ratio to SD_{i_p}.*

The method can be graphically represented in Figure 5.3.

Due to the propagation of qualitative correlations among related hypotheses h_{i_1}, h_{i_2}, ..., and h_{i_m}, the possibility of h_{i_p} changes from $\mu^o_{i_p}$ to μ_{i_p}.

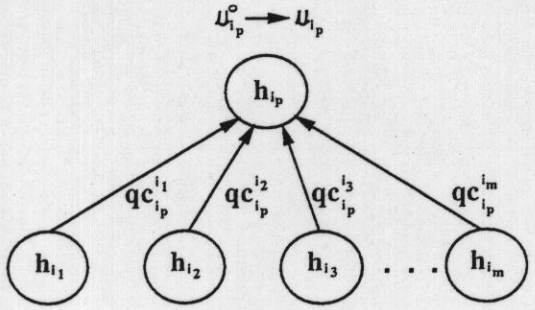

Figure 5.3: Propagation of Qualitative Correlations

5.6.2 Comparison

The propagation of qualitative correlations among hypotheses in the method is similar to the probability propagation in Bayesian methods – if we view the qualitative correlations among related hypotheses as pieces of evidence [Duda, and et al, 1977].

In solving the problems where qualitative correlations among related hypotheses can be extracted and used, the method is better in the following aspects:

1. In traditional Bayesian methods, evidence and its prior probability, hypotheses and their prior probabilities, and the relations between evidence and hypotheses (e.g., LS and LN) are all provided and fixed by domain experts in advance, so knowledge acquisition is a difficult task, and consequently, inconsistency can hardly be avoided. In the proposed method, on the other hand, only a few numbers are needed in advance. Instead, qualitative correlations among related hypotheses and much dynamic information are automatically obtained and propagated;

2. Traditional Bayesian methods require that the evidence which supports or depresses hypotheses be explicitly provided, and are only applicable to the

problems where the relations between evidence and hypotheses are known. The proposed method, however, is applicable to the problems where the relations between evidence and hypotheses are unknown as well as the problems where the relations between evidence and hypotheses are known.

3. In some cases, the proposed method does not need any assumption in advance. For example, in interpreting an inaccurate peak, the possibility of the peak, which corresponds to the prior probability of a hypothesis in Bayesian methods, can be automatically obtained by using the method presented in Chapter 4. Then, qualitative correlations between the inaccurate peak and its related peaks are directly used to update the possibility.

When qualitative correlations among related hypotheses are available, and assumptions necessary for Bayesian methods are hard to obtain, the proposed method is better than traditional Bayesian methods. However, when qualitative correlations among hypotheses are not known a priori, the method is not applicable. The method is especially effective to interpret inaccurate numerical and symbolic data by considering qualitative correlations among known hypotheses as confirmatory or disconfirmatory evidence.

5.7 Summary

In this chapter, I presented a novel method for propagating qualitative correlations among related hypotheses as confirmatory or disconfirmatory evidence of uncertain reasoning. The function of the method is similar to the probability propagation in Bayesian methods. However, compared with traditional Bayesian methods, the proposed method can be applied to the problems where evidence or the relation between evidence and hypotheses is not explicitly given, or is not available. In addition, the proposed method needs few numbers and assumptions in advance. Therefore, it is quite simple, and can effectively avoid inconsistency in knowledge bases. The method is especially effective to interpret inaccurate numerical and symbolic data by considering qualitative correlations among related data as confirmatory or disconfirmatory evidence.

Chapter 6

Implementation and Experiments

I have developed a knowledge-based system for interpreting infrared spectra by applying the proposed methods, and have thoroughly tested the system against several hundred real spectra. The experimental results show that the proposed methods are significantly better than the conventional methods used in many similar systems.

6.1 Infrared Spectrum Interpretation

The primary task of infrared spectrum interpretation is to identify unknown objects by interpreting their infrared spectra. In this chapter, I will focus on the problem to interpretation of infrared spectra of compounds to determine composition of unknown compounds without loss of generality.

Infrared spectrum interpretation is a very good test-bed of the research for the following reasons:

1. Interpreting infrared spectra is a very significant problem in both academic research and industrial application. For example, in chemical science and engineering, interpreting infrared spectra of compounds is the most effective way to identify unknown compounds, and to analyze the composition and purity of compounds [Colthup, Daly, & Wiberley, 1990].

2. Interpreting infrared spectra is a very difficult problem. First, spectral data are huge in quantity, and complex in representation. Second, both

symbolic reasoning and numerical analysis are needed to interpret infrared spectral data [Puskar, Levine, & Lowry, 1986][Sadtler, 1988].

3. Interpreting infrared spectra is a typical problem dealing with inaccurate data since spectral data are often inaccurate. They often shift from their theoretical values due to various reasons. For example, the following is an assertion for spectrum interpretation:

   The high frequency peak of partial component PC_α is located at F_i.

   In practice, however, the peak of PC_α may irregularly shift around F_i due to noise or other unforeseen reasons. When the above assertion is used to identify real spectra, uncertainty arises.

6.2 Applying the Proposed Methods to Infrared Spectrum Interpretation

Interpreting infrared spectra is a special problem of diagnosis. Suppose the infrared spectrum of an unknown compound can be thresholded and represented as a finite set of peaks (i.e., the measured dataset MD):

$$Sp = \{p_1, p_2, ..., p_n\},$$

where every peak consists of the frequency (position) f, strength (height) s, and width (shape) w, respectively:

$$p_i = (f_i, s_i, w_i) \qquad i = 1, 2, ..., n.$$

Because f_i, s_i and w_i refer to the same peak p_i, they are related data. This is one kind of related data in infrared spectrum interpretation.

Suppose there are finite partial components (i.e., reference values RV):

$$\begin{aligned} PC &= \{PC_1, PC_2, ..., PC_k\} \\ &= \{\{p_{j_1}, p_{j_2}, ..., p_{j_m}\} \mid j = 1, 2, ..., k\} \\ &= \{\{(f_{j_p}, s_{j_p}, w_{j_p}) \mid p = 1, 2, ..., m\} \mid j = 1, 2, ..., k\}. \end{aligned}$$

Because f_{j_p}, s_{j_p} and w_{j_p} also refer to the same reference peak p_{j_p}, they are related data as well.

The spectroscopic knowledge for interpreting infrared spectra is usually in the form like "*if p_i is equal to p_{j_p}, then p_i may be created by partial component PC_j*", where "p_i is equal to p_{j_p}" represents that f_i, s_i, and w_i are equal to f_{j_p}, s_{j_p}, and w_{j_p} respectively.

This kind of related data has the following qualitative correlations:

1. f_i, s_i and w_i should be identified simultaneously, that is,

   - if f_i corresponds to f_{j_p}, then s_i corresponds to s_{j_p}, and w_i corresponds to w_{j_p}, and

   - if s_i corresponds to s_{j_p}, then f_i corresponds to f_{j_p}, and w_i corresponds to w_{j_p}, and

   - if w_i corresponds to w_{j_p}, then f_i corresponds to f_{j_p}, and s_i is s_{j_p}.

2. related data support each other. For example, if both f_i and s_i have been identified, then they will enhance the identification of w_i. Conversely, if f_i and s_i have not been identified, then they will weaken the identification of w_i.

The methods for identifying f_i, s_i and w_i based on the qualitative correlations among them presented in Chapter 4 can be formalized as the following predicate calculi, respectively[1]:

$$\{f_i@p_{j_p}, p_{j_p}@p_i\} \Rightarrow \{p_i \textit{ is created by } PC_j\}, \text{ and}$$

$$\{s_i@p_{j_p}, p_{j_p}@p_i\} \Rightarrow \{p_i \textit{ is created by } PC_j\}, \text{ and}$$

$$\{w_i@p_{j_p}, p_{j_p}@p_i\} \Rightarrow \{p_i \textit{ is created by } PC_j\},$$

where "p_i *is created by* PC_j" means that f_i, s_i and w_i can be qualitatively identified to be f_{j_p}, s_{j_p} and w_{j_p}.

In general, each partial component may create finite peaks at the same time. So if p_i is created by PC_j, then Sp is partially created by PC_j; if Sp is partially created by PC_j, then all the peaks that PC_j may create should be contained by Sp simultaneously. Therefore, all the peaks created by a partial component are also related data. This is another kind of related data in infrared spectrum interpretation.

This kind of related data has the following qualitative correlations:

1. all the peaks of a partial component should be identified simultaneously, that is,

$$\text{if } p_i \text{ corresponds to } p_{j_p}, \text{ then } p_{j_l} \in Sp \quad (l = 1, 2, ..., m, l \neq p).$$

2. the peaks created by the same partial component support each other. For example, if most of the peaks of a partial component have been identified, these peaks will enhance the identification of the rest peaks. Conversely, if most of the peaks of a partial component can not be identified, then the identification of the rest peaks will be depressed.

[1]The process described in this chapter does not concern the constraint satisfaction problems in infrared spectrum interpretation which will be presented in Chapter 8.

The method for identifying related peaks based on the qualitative correlations can be formalized as the following logic expression:

$$\{p_i@PC_j, PC_j@Sp\} \Rightarrow \{PC_j \in IN(Sp)\}.$$

6.3 System for Interpreting Infrared Spectra

The system is implemented with C and MS-WINDOWS. Figure 6.1 shows the data flow diagram of the system[2].

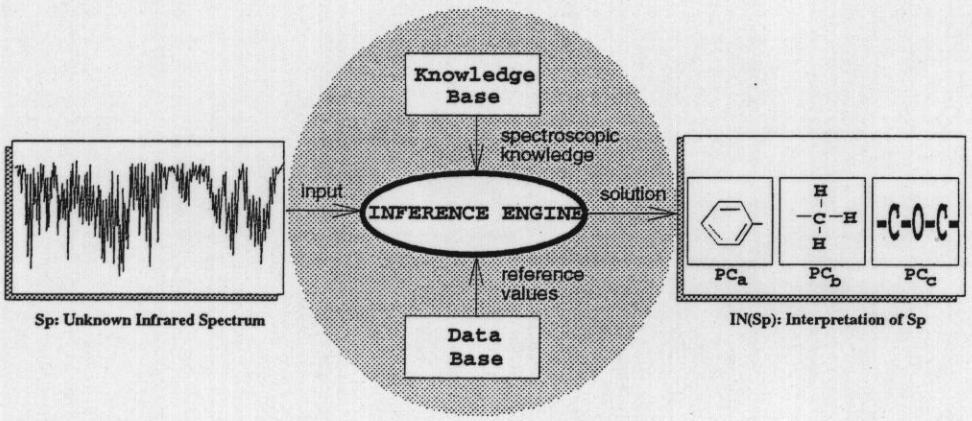

Figure 6.1: Data Flow Diagram of the System

The input data of the system are infrared spectra of unknown compounds, and the solutions are partial components that the input spectra may contain. Because inferences are based on qualitative features of spectral data and qualitative

[2]This section only briefly mentions the architecture of the system. The design and development of the system will be presented in Chapter 7.

correlations among related data, the system can gain high correct interpretation performance with noisy spectral data.

As I mentioned before, there are two kinds of related data in infrared spectrum interpretation: all the features of a single peak (i.e., f_i, s_i and w_i of p_i), and all the peaks of a single partial component (i.e., p_1, p_2, ... and p_m). The inference engine of the system employs the proposed methods to both kinds of the related data when inaccuracy arises.

6.4 Examples

I discuss the performance of the system through the following example. Figure 6.2 shows an infrared spectrum of an unknown compound. The spectrum is very hard to interpret since the peak with an arrow (named p_1) shifts substantially. The system correctly identifies that p_1 is created by partial component *benzene-ring*.

In contrast, many similar systems can not correctly identify the peak [Clerc, Pretsch, & Zurcher, 1986][Hasenoehrl, Perkins, & Griffiths, 1992][Wythoff, Buck, & Tomellini, 1989], since the peak of a *benzene-ring* at this frequency position (named p_{b_1}) should be a strong peak (i.e., $s_{b_1} > 1.000$) according to spectroscopic knowledge, not a medium one ($s_1 = 0.510$) as the case in this example. Systems based on conventional fuzzy methods usually assume a fuzzy interval for each inaccurate peak, then determine the membership degree that the inaccurate peak is in the fuzzy interval. Suppose the reference value for a strong peak is 1.000, and the fuzzy interval for a strong peak is 0.300 [Colthup, Daly, & Wiberley, 1990], then only peaks with strength of 1 ± 0.300 can be regarded as strong peaks. Obviously, by conventional fuzzy methods, the possibility of p_1 being a strong peak is zero, i.e., $\mu_{benzene-ring}(s_1) = 0$.

Inferring on the basis of qualitative correlations among related data, the system makes a correct interpretation of the spectrum. Through the following

two cases, I introduce the inference process of the system, and at the same time demonstrate the use of the methods presented before.

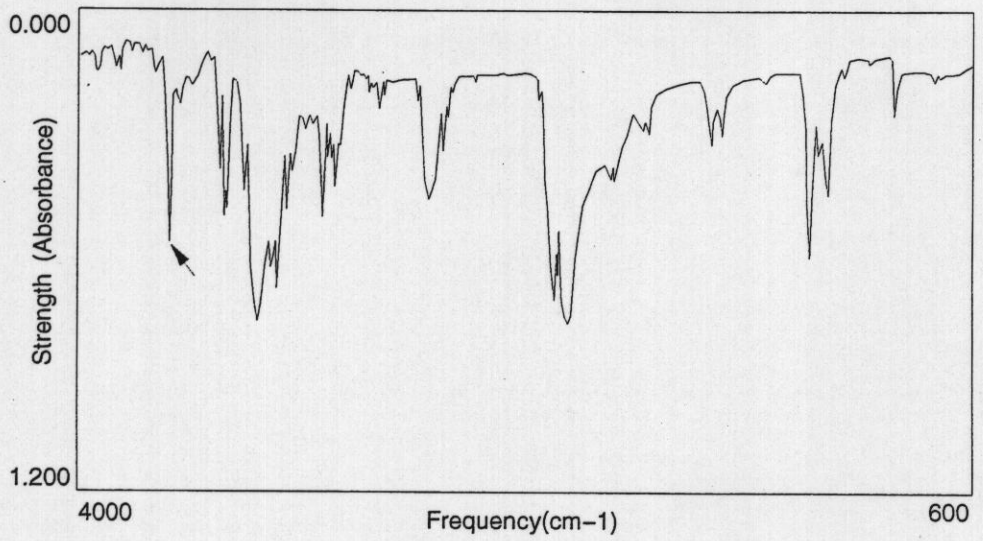

Figure 6.2: An Example of Infrared Spectrum

6.4.1 Case I: Considering the First Kind of Related Data

Because the frequency (position) and width (shape) of p_1 are both the same as those of *benzene-ring*, the possibility of f_1 being identified as f_{b_1} is 100% (i.e., $\mu_{benzene-ring}(f_1) = 1$), and the possibility of w_1 being identified as w_{b_1} is also 100% (i.e., $\mu_{benzene-ring}(w_1) = 1$[3].

[3] $\mu_*(d)$ means the possibility of d being identified by conventional fuzzy methods, i.e., SCF is not considered.

As I have discussed before, f_1, s_1 and w_1 are related data, so we can obtain confirm evidence for identifying s_1 by considering qualitative correlations among s_1, f_1 and w_1:

$\mu_{benzene-ring}(f_1) = 1$,

so, $c_{s_1}(f_1) = 1$ ($c_{s_1}(f_1)$ represents the qualitative correlation between s_1 and f_1),

$\mu_{benzene-ring}(w_1) = 1$,

so, $c_{s_1}(w_1) = 1$ ($c_{s_1}(w_1)$ represents the qualitative correlation between s_1 and w_1)

so, $SCF_{s_1} = \frac{1+2}{3} = 1$, and

$\triangle s_1 = \frac{(6-1)\times 0.300}{3} \times 1 = 0.500$, and

$s_1@p_{b_1} = 1 - \frac{1-0.510}{0.500} = 0.02$.

By considering SCF_{s_1}, the possibility of p_1 being regarded as a strong peak of *benzene-ring* increases from 0 to 0.02. Many near-misses may be handled by the negligible possibility. For example, in most systems based on fuzzy and other methods [Clerc, Pretsch, & Zurcher, 1986], it is impossible to identify p_1 to be "strong" (i.e., $\mu_{benzene-ring}(s_1) = 0$), but considering qualitative correlations among related data makes it possible, although the possibility is only 0.02 ("$\mu \leq 0$" means "impossible", but "$\mu > 0$ means "possible").

As f_1 and w_1 are both the same as the reference values, so $f_1@p_{b_1} = 1$, and $w_1@p_{b_1} = 1$.

Suppose the priorities of f_1, s_1 and w_1 are 2, 1 and 1 respectively, then the possibility of p_1 being identified as p_{b_1} is:

$$\mu_1 = p_{b_1}@p_1 = \frac{2 \times 1 + 0.02 + 1}{4} = 0.755.$$

while the possibility of p_1 being identified before is 0, since one data item can not be identified (i.e., $\mu_{benzene-ring}(s_1) = 0$).

6.4.2 Case II: Considering the Second Kind of Related Data

The process of considering the second kind of related data is quite similar.

We have got that the possibility of p_1 being created by a *benzene-ring* is μ_1 ($\mu_1 = 0.755$). Suppose the *benzene-ring* can create m peaks: $\{p_{b_1}, p_{b_2}, ..., p_{b_m}\}$, then the m peaks are related to each other. If p_1 is created by the *benzene-ring*, then Sp is partially created by the *benzene-ring* (i.e., the *benzene-ring* is contained by the unknown spectrum); if Sp is partially created by the *benzene-ring*, then the other m-1 peaks of the *benzene-ring* should also be identified.

By using the same procedure as obtaining μ_1, we can get μ_2, μ_3, ... and μ_m as well. According to the method presented in Chapter 5, the qualitative correlation between two related peaks, p_i and p_j, is defined as:

$$qc_i^j = \begin{cases} 1 & if \ \ \mu_j \geq 0.5 \\ 0 & if \ \ \mu_j < 0.5. \end{cases}$$

So

$$SD_i = \frac{1 + \sum_{j=1, j \neq i}^{m} qc_i^j}{m}, \quad 0 < SD_i \leq 1.$$

Then

$$P_i^2 = \frac{2m - 1}{m} \times SD_i, \quad 0 < P_i^2 < 2,$$

and

$$p_i@benzene - ring = 1 - \frac{1 - \mu_i}{P_i^2}, \quad p_i@benzene - ring \leq 1.$$

Roughly, when $SD_i > 0.5$, related peaks tend to support p_i. When related peaks support p_i, $P_i^2 > 1$. When $P_i^2 > 1$, $p_i@benzene - ring > \mu_i$.

Table 6.1 shows the relation among $p_i@benzene - ring$, μ_i and P_i^2.

$p_i@benzene - ring$		μ_i				
		1	0.8	0.5	0.3	0
P_i^2	1.3	1	0.846	0.615	0.462	0.231
	1.1	1	0.818	0.545	0.364	0.091
	1	1	0.8	0.5	0.3	0
	0.9	1	0.778	0.444	0.222	/
	0.7	1	0.714	0.286	0	/

Table 6.1: Relation among $p_i@benzene - ring$, μ_i and P_i^2

In the above example, $SD_1 = 0.850$, and $P_1^2 = 1.658$, so

$$p_1@benzene - ring = 1 - \frac{1 - 0.755}{1.658} = 0.852.$$

Therefore, the possibility of p_1 being identified as p_{b_1} increases from 0.755 to 0.852 due to qualitative correlations among related peaks. The process is similar to the probability propagation in probabilistic reasoning. Here identifying p_1 is a hypothesis, and qualitative correlations among related data of p_1 are pieces of evidence.

After all the peaks of the *benzene-ring* are identified, the possibility that the *benzene-ring* is contained by Sp can be finally calculated.

6.5 Analysis of Experimental Results

I compare two methods in the experiments. The first method (called "AF") is a conventional fuzzy method which is used by most similar systems [Clerc, Pretsch, & Zurcher, 1986][Wythoff, Buck, & Tomellini, 1989]. To use AF, each reference value must be associated with a fuzzy interval for dealing with inaccuracy. Both

reference values and fuzzy intervals are empirically determined [Colthup, Daly, & Wiberley, 1990].

Table 6.2 lists some reference values and their fuzzy regions used by AF.

CH_3	$2960 \pm 15cm^{-1}$	$strong \pm 0.3$	$sharp \pm 1$
	$2870 \pm 15cm^{-1}$	$strong \pm 0.3$	$sharp \pm 1$
	$1450 \pm 10cm^{-1}$	$medium \pm 0.3$	$sharp \pm 0.5$
	...		
$benzene - ring$	$3055 \pm 25cm^{-1}$	$strong \pm 0.3$	$sharp \pm 1.5$
	$1645 \pm 10cm^{-1}$	$medium \pm 0.3$	$sharp \pm 0.5$
	$1550 \pm 30cm^{-1}$	$medium \pm 0.3$	$sharp \pm 1$
	$1450 \pm 3cm^{-1}$	$medium \pm 0.3$	$sharp \pm 0$
	...		
$-CH_2 - OH$	$3635 \pm 5cm^{-1}$	$strong \pm 0.3$	$broad \pm 1$
	$3550 \pm 25cm^{-1}$	$strong \pm 0.3$	$sharp \pm 1$
	...		

Table 6.2: Some Reference Values and Their Fuzzy Regions

The membership function of AF is:

$$\mu_r(d) = max\left\{0, 1 - \frac{\mid d - r \mid}{\triangle d}\right\},$$

where d is a measured data item, r is a reference value, $\triangle d$ is the fuzzy interval of r, and $0 \leq \mu_r(d) \leq 1$.

The second method (called "AF^*") is the combination of the methods proposed in Chapter 4 and Chapter 5. AF^* uses the same reference values and fuzzy intervals as AF in identifying a single peak, but the fuzzy intervals in AF^* are only used as standard fuzzy intervals based on which dynamic shift intervals are determined by considering qualitative correlations among related data.

AF and AF^* use the same reference values and empirical fuzzy intervals. The formula for calculating membership degrees in AF (i.e., $\mu_r(d) = max\{0, 1 - \frac{|d-r|}{\triangle d}\}$) is also similar to the formula for calculating possibility in AF^* (i.e., $\mu_i = 1 - \frac{|d_i - r_{j_p}|}{\triangle d_i}$). However, in AF, $\triangle d$ is simply an empirical fuzzy interval, while

in AF^*, $\triangle d_i$ is a dynamic shift interval based on qualitative correlations among related data.

I have tested the system against about three hundred real infrared spectra of organic compounds. The experimental results show that AF^* is significantly better than AF.

There are two important standard metrics for evaluating solutions of infrared spectrum interpretation:

Definition 6.1 *Rate of correctness (RC): the rate that the identified partial component set is exactly the same as the partial component set in the correct solutions.*

Definition 6.2 *Rate of identification (RI): the rate that how many partial components in the correct solutions are identified.*

Table 6.3 shows the comparison between AF and AF^* with the two standard metrics.

	RC	(error-rate)	RI	(error-rate)
AF	0.455	(0.545)	0.812	(0.188)
AF^*	0.736	(0.264)	0.894	(0.106)

Table 6.3: Evaluation of AF & AF^* with RC and RI

Table 6.3 demonstrates that both the RC and RI increase by integrating SCF, but the RC increases more significantly. The reason is that although AF can identify most partial components of unknown compounds, it is hard to identify all partial components of unknown compounds because there are always some partial components whose measured peaks seriously shift from the reference values.

Table 6.4 and 6.5 list part of the experimental results in which the first column indicates the solutions obtained by AF; the second column indicates the solutions obtained by AF^*; and the third column shows the correct solutions.

AF (Without *SCF*)	*AF** (With *SCF*)	Correct Solutions
◎ –CH2– CH3– –[CH2]n–	◎ –CH2– CH3– –[CH2]n–	–CH2– CH3– –[CH2]n–
(2/3) –CH2– –C(–)–	◎ –CH2– CH3– –C(–)–	–CH2– CH3– –C(–)–
◎ –CH2– CH3– –CH(CH3)–	◎ –CH2– CH3– –CH(CH3)–	–CH2– CH3– –CH(CH3)–
◎ –CH2– CH3– –C(–)–	◎ –CH2– CH3– –C(–)–	–CH2– CH3– –C(–)–
(2/3) CH3– –CH(CH3)–	◎ CH3– –CH(CH3)– ⬡	CH3– –CH(CH3)– ⬡
(3/4) –CH2– CH3– >C=CH–	◎ –CH2– CH3– >C=CH– ⬡	–CH2– CH3– >C=CH– ⬡
◎ CH3– ⬡	(2/2) –CH2– CH3– ⬡	CH3– ⬡

◎ : identified PC set is the same as the PC set in the correct solution(in this case, RI=1)
(n) : identified PC set is not the same as the PC set in the correct solution(the number indicates the RI)

Table 6.4: Experimental Results with *AF* & *AF** (Part 1)

AF (Without *SCF*)	*AF** (With *SCF*)	Correct Solutions
(2/3) -CH2- CH3-	(2/3) -CH2- CH3-	-CH2- CH3-
(1/3) >C=CH-	(2/3) CH3- >C=CH-	CH3- >C=CH-
◎ -[CH2]n- -C≡CH	◎ -[CH2]n- -C≡CH	-[CH2]n- -C≡CH
(4/5) -CH2- CH3- >C=CH- -CH[CH3]2	◎ -CH2- CH3- >C=CH- -CH[CH3]2	-CH2- CH3- >C=CH- -CH[CH3]2
(2/3) -CH2-	◎ -CH2- CH3-	-CH2- CH3-
(1/2) C(Cl)(Cl)	◎ -C=C- C(Cl)(Cl)	-C=C- C(Cl)(Cl)
(3/4) CH3- NH2-	◎ CH3- NH2- -C(-)(-)-	CH3- NH2- -C(-)(-)-

◎ : identified PC set is the same as the PC set in the correct solution(in this case, RI=1)
(n) : identified PC set is not the same as the PC set in the correct solution(the number indicates the RI)

Table 6.5: Experimental Results with *AF* & *AF** (Part 2)

6.6 Comparison with Related Systems

Related systems mainly fall into the following four categories: (1) Systems based on Y/N classification, (2) Systems based on fuzzy logic, (3) Systems based on pattern recognition, and (4) Systems based on neural networks.

6.6.1 Systems Based on Yes/No Classification

The method commonly used by spectroscopists in practice is numerical analysis [Colthup, Daly, & Wiberley, 1990]. Numerical analysis is primarily based on comparison between spectral data and reference values. Reference values are usually some regions like $frequency : 3615 \pm 5cm^{-1}$ or $strength : 1.000 \pm 0.300$. If spectral data are in certain regions, the answer of classification is yes; otherwise, the answer is no.

Most systems for interpreting infrared spectra use this method [Hasenoehrl, Perkins, & Griffiths, 1992][Puskar, Levine, & Lowry, 1986][Wythoff, Buck, & Tomellini, 1989]. For example, in Wythoff's system, rules for comparing spectral data are in following forms as shown in Table 6.6.

ANY PEAK(S)	*FREQUENCY:1700-1707*	*STRENGTH:0.7-1.0*
	WIDTH:SHARP TO BROAD	
ANSWER -YES-		
*ACTION - ****		

Table 6.6: Rules for Comparing Spectral Data

The advantage of these systems is that they are very easy to develop because they can directly use spectroscopic knowledge, and do not need further computation. However, the problem is that each of these systems is only applicable to a specific class of compounds, or pure compounds because when spectral data are seriously inaccurate, the reference values (regions) can not reflect the inaccuracy. For example, Hasenoehrl's system is only for distinguishing compounds containing at least one carbonyl functionality from other compounds, although the *RI* of the system is about 98% (naturally, the *RC* is not available), and Puskar's system is only for identifying hazardous substances.

In fact, spectroscopists also use qualitative analysis in some specific cases in addition to the formal spectroscopic knowledge, such as "*if the peaks in 600 cm^{-1}- 900 cm^{-1} look like the peaks of benzene-rings, then the peaks in 3000 cm^{-1}- 3100 cm^{-1} are quite likely to be created by a benzene-ring.*" Unfortunately, the qualitative analysis was hardly applied to these systems since it can not be used directly. In contrast, my system can successfully use the qualitative analysis like spectroscopists. The way of using it is the methods proposed in this Chapter 4 and Chapter 5. As a result, the system is applicable to all compounds with very high correct performance.

6.6.2 Systems Based on Fuzzy Logic

Since spectral data are always inaccurate, and the representation of spectroscopic knowledge is quite like that in fuzzy logic, some systems naturally use fuzzy logic or some techniques similar to fuzzy logic [Clerc, Pretsch, & Zurcher, 1986]. In these systems, fuzzy intervals which are similar to the regions described in Section 6.6.1 are given for reference values, and memberships of inaccurate data are calculated on the basis of the degrees that the inaccurate data are in the fuzzy intervals. These systems are better than those described in Section 6.6.1 in some cases, but the degrees that inaccurate data are in fuzzy intervals do not necessarily reflect the possibility of the inaccurate data being the reference values. For example, in Figure 6.3, it is difficult to determine which peak is closer to the reference value only by considering the degrees that *peak a* and *peak b* are in the fuzzy interval.

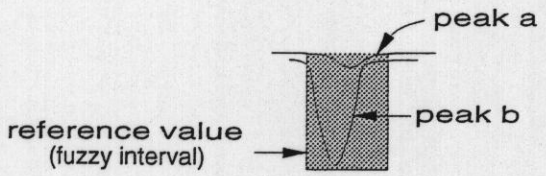

Figure 6.3: Two Peaks in a Fuzzy Interval

However, by applying the methods proposed in Chapter 4 and Chapter 5, the above problem can be easily solved. As I discussed before, in practice, spectroscopists also frequently use knowledge about correlations among peaks in addition to the formalizable spectroscopic knowledge. This kind of knowledge

is essential to my methods which enable qualitative correlations among related data to be used as evidence for the identification of inaccurate data.

I have compared the fuzzy method used by these systems with the proposed methods in section 6.5. So far as I know, the *RC* of my system is the highest among the similar systems, and the *RI* of my system is higher than that of many similar systems.

6.6.3 Systems Based on Pattern Recognition

Some systems use pattern recognition techniques to interpret infrared spectra [Jalsovszky & Holly, 1988][Sadtler, 1988], of which Sadtler is the most popular commercial system. The system compares known patterns with unknown ones, and determines the possibility of an unknown pattern being a known one by calculating the quantitative similarity or closeness between the two patterns.

Unlike fuzzy techniques, pattern recognition considers a group of data (i.e., a pattern) at the same time. However, pattern recognition is primarily based on quantitative analysis. I have discussed that in many cases especially when the inaccuracy of spectral data is not slight, qualitative features of spectral data are much more important than quantitative ones. For example, Figure 6.4 shows two simple cases. The difference between the two patterns in (a) is smaller than that in (b). From the viewpoint of Sadtler, the two patterns in (a) are closer than those in (b). However, the two patterns in (b) may be the same in some cases, while the two patterns in (a) may not be the same in any case. The reason is that the qualitative features (frequency positions of peaks) of the two patterns in (a) are different.

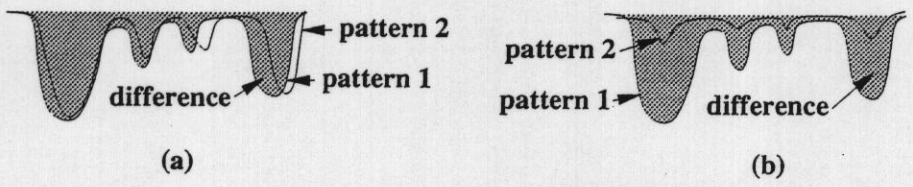

Figure 6.4: Quantitative Differences between Patterns

Because quantitative similarity and closeness are not always sound, most systems based on pattern recognition including Sadtler can not give concrete

solutions. In general, the solutions of these systems are only a series of candidates from which users have to finally decide the possible one by themselves.

It is difficult to compare these systems with mine because the solutions of these systems are quite loose, and neither the *RC* nor the *RI* is available. Sadtler, for example, usually gives the list of all known patterns associated with the values of quantitative differences between the unknown patterns and these known ones.

6.6.4 Systems Based on Neural Networks

Recently, neural networks have been applied to infrared spectrum interpreting systems [Anand, Mehrotra, Mohan, & Ranka, 1991][Becraft, Lee, & Newell, 1991][Robb & Munk, 1990]. In Anand's system, a neural network approach is used to analyze the presence of amino acids in protein molecules. To this specific classification, the *RI* of Anand's system is about 87%, and the *RC* is not available. In Robb's system, a linear neural network model is developed for interpreting infrared spectra. The system is for general purpose like mine. Without prior input of spectrum-structure correlations, the *RC* of Robb's system is 53.3%.

Although the *RC* and *RI* of my system are both higher than those of the two systems, using neural networks is still promising, especially when model training or system learning is a must.

6.7 Summary

In this chapter, I introduced the implementation of the proposed methods, and discussed the corresponding experiments. I applied the methods to a practical system for infrared spectrum interpretation, a typical problem dealing with inaccurate data. I fully tested the system against about three hundred real infrared spectra of organic compounds. The experiments show that the methods are significantly better than the conventional methods used in many similar systems.

I also gave two examples in this chapter to demonstrate how to use qualitative correlations among related data as evidence to identify inaccurate data in practical problems. In addition, I compared my system with similar systems, the *RC* and *RI* of my system are higher than those of others so far as I know. I also compared the methods with related methods, and discussed when and how they are better or different.

Chapter 7

Knowledge-Based System for Infrared Spectrum Interpretation

In this chapter, I present a knowledge-based system for infrared spectrum interpretation. In Chapter 6, I once discussed the application of my methods to the system for interpreting inaccurate spectral data. Since interpreting inaccurate spectral data is only one of the difficult issues of system, in order to give an overall picture of the system, in this chapter, I introduce other issues of the system. First, I briefly introduce the design and development of the system. Then, I demonstrate the working process of the system with examples.

7.1 Introduction

Traditional methods for infrared spectrum interpretation require comparing infrared spectra of unknown compounds with infrared spectra of known compounds to interpret what the unknown ones are. The principle behind traditional methods is that compounds exhibiting similar infrared spectra will also have similar chemical structures which can be expressed as the following formula.

$$\begin{cases} Sp = F(St) \\ St = F^{-1}(Sp), \end{cases}$$

where Sp means spectra of compounds, and St means their chemical structures.

Performing the task of comparing infrared spectra of unknown and known compounds traditionally relies on quantitative analysis. However, only using quantitative analysis has two critical problems:

1. Quantitative analysis is generally very complex, and in some cases it may even become intractable. For example, the number of known compounds is very large, and the comparison between unknown and known infrared spectra is very complex, so reducing the number of known compounds to be compared must be done before quantitative analysis starts;

2. Spectral data are always inaccurate due to noise and other unforeseen reasons. When spectral data are inaccurate, only using quantitative analysis is hard to give concrete solutions.

As we know, a compound usually consists of several different partial components, and different partial component has different spectral pattern. Therefore, the infrared spectrum of the compound will somewhat exhibit the patterns of the partial components, that is,

$$\begin{cases} Partial\_Sp_i = G_i(Partial\_Component_i) \\ Partial\_Component_i = G_i^{-1}(Partial\_Sp_i), \end{cases}$$

and

$$\begin{cases} Sp = \sum(Partial\_Sp) \\ St = \sum(Partial\_Component). \end{cases}$$

My system interprets infrared spectra with three separate phases. Instead of investigating all known compounds to directly determining what compounds the unknown compounds are, the system first qualitatively analyzes the unknown spectra to determine what partial components the unknown compounds contain. Then, at the second phase, it gives, based on the partial components, a very limited list of candidates of compound. Finally, at the third phase, it investigates the limited compound candidates to determine the correct one by using quantitative analysis. Because the number of partial components is much smaller than the number of compounds, and qualitative rather than quantitative information can be used in analyzing partial components, the first phase is quite simple. And because the number of compounds to be compared with unknown compounds can be significantly reduced by qualitative analysis (Usually the space can be narrowed down from over thousand known compounds to several candidate),

the quantitative analysis at the third phase becomes very convenient. In addition, qualitative analysis relies on qualitative features of infrared spectra more than quantitative features of infrared spectra, inaccuracy of spectral data can be handled well at the same time.

7.2 Design of the System

Three phases of the system described in Section 7.1, can be roughly viewed as two different processes, that is, a qualitative process for analyzing what partial components an unknown compound contains by qualitatively interpreting the infrared spectrum of the compound, and a quantitative process for analyzing partial components to produce a list of candidates of compound which contains the partial components identified by the qualitative process, and for analyzing the list of candidates of compound to determine what the unknown compound is.

The two processes are shown in Figure 7.1.

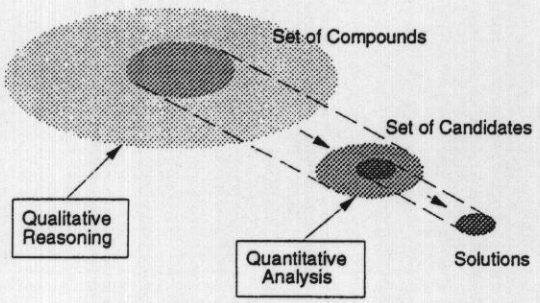

Figure 7.1: Qualitative and Quantitative Processes

By doing qualitative analysis first, the system only needs to apply complex quantitative analysis to a very limited list of candidates. The method for qualitatively interpreting inaccurate data presented in Chapter 4 and the method for uncertain reasoning presented in Chapter 5 are employed by the qualitative reasoning.

7.2.1 Qualitative Process

Before introducing the qualitative process, I first put forward the following three new concepts.

Definition 7.1 (*Peaks' Unique Pattern*): *Peaks' Unique Pattern (PUP) is a combination of some peaks of a partial component which distinguishes the partial component from others. The peaks in the PUP of a partial component are called key peaks of the partial component.*

For example, the peaks of *meta-sub benzene* are distributed from 4000 cm^{-1} to 600 cm^{-1}. In 3100 cm^{-1} to 3000 cm^{-1}, only *alkane CH*, *meta-sub benzene*, *ortho-sub benzene*, *mono-sub benzene*, and *para-sub benzene* have peaks. Therefore, peaks of these partial components in (3100, 3000) can be selected as one of their key peaks to distinguish them from others. Similarly, in 900 cm^{-1} to 800 cm^{-1}, only *benzenes* have peaks. Therefore, peaks of *benzenes* in (900, 800) can also be selected as their key peaks to distinguish them from others. Among *benzenes*, only *meta-sub benzene* has peaks in 800 cm^{-1} to 700 cm^{-1}, so the peak in (800, 700) can be selected as its key peaks to distinguish it from other *benzenes*. As a result, the peaks in (3100, 3000), (900, 800) and (800, 700) can form a *PUP* of *meta-sub benzene*.

The following theorem can be directly given from the definition of *PUP*.

Theorem 7.1: *If $PUP_j \not\subset Sp$, then $PC_j \notin SL$, where PC_j is a partial component, PUP_j represents the PUP of PC_j, Sp is the spectrum of unknown compound, and SL is the solution list.*

Proof: 1) $PUP_j \subseteq PL(PC_j)$, where $PL(PC_j)$ is the peak list of PC_j,

so if $PUP_j \not\subset Sp$, then $PL(PC_j) \not\subset Sp$;

2) if $PL(PC_j) \not\subset Sp$, then $PC_j \notin SL$. □

If the *PUP* of a partial component can be identified from the *Sp* of an unknown compound, the partial component is perhaps contained by the unknown compound; if the *PUP* can not be identified, the partial component is definitely not contained by the unknown compound.

With the theorem, we can easily eliminate obviously impossible partial components from our consideration by only checking the PUPs of partial components.

Definition 7.2 (*Splitted Spectral Section*): *Splitted Spectral Sections (SSS) are spectral regions on spectra in each of which only the key peaks of some specific partial components appear.*

For example, in the region of 3700-3100 cm^{-1}, partial components OH, NH and CH are active. The key peaks of these partial components always appear in this region. So this region can be selected as an SSS. Similarly, the region of 3100-3000 cm^{-1} can also be selected as an SSS in which *Substituted-benzenes* and *Alkane-CH* are always active.

If there are peaks in an SSS, a hypothesis, *the partial components which have key peaks in this SSS possibly exist*, is made. The further matter is to find evidence to prove the hypothesis, or to negate it. If other peaks of these partial components are also found, the hypothesis is enhanced. If all peaks of these partial components are found, the hypothesis gets proved.

The following theorem can be drawn from the definition of SSS.

Theorem 7.2: *If $p_i \notin SSS_j$, then $PC_j \notin SL$, where p_i is a key peak of PC_j, and SSS_j is an SSS of PC_j in which p_i should appear.*

Proof: 1) SSS_j is a section of Sp where p_i should be,
so if $p_i \notin SSS_j$, then $p_i \notin Sp$;
2) $p_i \in PUP_j$,
so if $p_i \notin Sp$, then $PUP_j \not\subset Sp$, $PL(PC_j) \not\subset Sp$;
3) if $PL(PC_j) \not\subset Sp$, then $PC_j \notin SL$. □

In a single SSS, it is impossible to identify what partial components exist, but it is possible to identify what partial components do not exist. Because if the key peak of a partial components should be but can not be found in an SSS, then the partial component is definitely not contained by the Sp.

If a key peak of a partial components can be found in an SSS, then the partial component is likely to exist. The further work is to find other peaks of the partial component from Sp.

Definition 7.3 (*Vague Frequency Position*): *Vague Frequency Positions (VFP) are frequency intervals of peaks around their standard frequency positions.*

Because spectral data are always inaccurate due to noise or other unforeseen reasons, the real peaks on Sp are always slightly different from their theoretical

positions. In qualitative analysis, the location of a peak is more important than its exact frequency position. In eliminating impossible partial components, the location of a peak in a specific SSS is used to represent the vague position of the peak which is called Vague Frequency Position (VFP). No matter what their frequency positions are, once the peaks fall into some SSSs, corresponding hypotheses will be made.

Using VFP can significantly reduce the complexity of checking partial components, and more importantly, can enable us to avoid handling inaccuracy of spectral data in the process of eliminating obviously impossible partial components, but leave inaccuracy handling for the process of deciding the correct partial components.

The following is a simple algorithm for the qualitative process of analyzing partial components.

Algorithm *From-Sp-to-PC*
  Procedure *From-Sp-to-Possible-PC*
    $Possible\_PC = \{PC_j \mid j = 1, 2, ..., N\}$; ;;; N PCs
    for $i = 1$ to L { ;;; L SSSs
      $Possible\_PC_i = \{PC$ whose PUP is in $SSS_i\}$;
      if *no_peak_in* SSS_i
        $Possible\_PC = Possible\_PC$ - $Possible\_PC_i$;}
    return$\{Possible_PC\}$;;; all possible PCs
  end procedure
  Procedure *From-Possible-PC-to-Solutions*
    $SL = \emptyset$;
    for $i = 1$ to M ;;; M PCs in Possible_PC
      if $PC_i \subset Sp$
        $SL = SL \cup \{PC_i\}$;
    return$\{SL\}$;;; final solution
  end procedure
end algorithm

The way of handling inaccurate spectral data in qualitatively analyzing partial components has been extensively discussed in Chapter 4, 5 and 6, so I made no mention of it in this section.

7.2.2 Quantitative Process

Before describing the quantitative process, I first introduce two relevant concepts.

Definition 7.4 ($DB1$): *$DB1$ is the data base of compounds which consists of all known compounds and their classifications.*

The classification of a compound includes the class and subclass that the compound belongs to.

In the system, $DB1$ is organized in the following form.

Class	Subclass	Compound
$Hydrocarbon$	$n\text{-}Paraffin$	$CH_3\text{-}[CH_2]_5\text{-}CH_3$
		$CH_3\text{-}[CH_2]_{10}\text{-}CH_3$
	$Isoparaffin$	$CH_3\text{-}[CH\text{-}CH_2]\text{-}CH_3$
	...	

Definition 7.5 ($DB2$): *$DB2$ is the data base of PCs which consists of all known partial components and their classifications.*

The classifications of PCs are determined by the classifications of compounds in which the PCs first appear. For example, CH_2, CH_3 and CH first appear in the class of $Hydrocarbon$, so they belong to the class of $Hydrocarbon$. CH_2 and CH_3 appear in both the subclasses of $n\text{-}Paraffin$ and $Isoparaffin$, but appear in $n\text{-}Paraffin$ first, so they belong to the subclass of $n\text{-}Paraffin$.

In the system, $DB2$ is organized in the following form.

PC	Class	Subclass
CH_2, CH_3, $[CH_2]_n$	$Hydrocarbon$	$n\text{-}Paraffin$
CH	$Hydrocarbon$	$Isoparaffin$
	...	

After partial components that an infrared spectrum contains have been determined by the qualitative process presented in Section 7.2.1, $DB1$ and $DB2$

are checked to produce a list of compounds as candidates each of which consists of these partial components.

The algorithm shown in Figure 7.2 performs the task of producing candidates from partial components. In the algorithm, Lpc is the list of partial components.

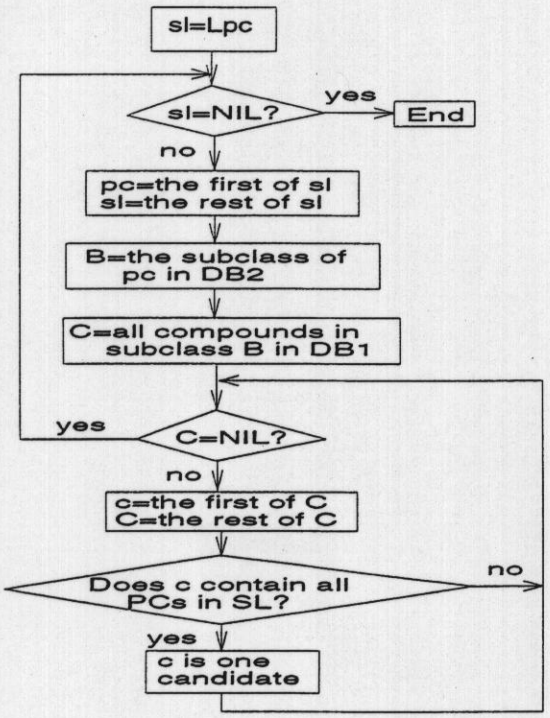

Figure 7.2: Algorithm for Generating Candidate List

Since compounds consisting of the same partial components are very limited, the number of compounds in the list of candidates is generally very small. Therefore, using complex quantitative analysis to determine the correct compound from the list of candidates becomes very simple.

The algorithm in Figure 7.2 can create a list of compound candidates. However, deciding which candidate the unknown compound is still requires quantitative analysis. For example, so far we may be sure that an unknown compound is either $CH_3 - [CH_2]_5 - CH_3$ or $CH_3 - [CH_2]_{10} - CH_3$. The further work de-

termining the number n in $CH_3 - [CH_2]_n - CH_3$ has to be done by quantitative analysis.

I adopt a commonly used method for quantitative analysis. The method calculates the quantitative closeness between two *Sp*s as the possibility of one being the other, and gives the known compound with the greatest possibility of being the unknown compound as its solution [Colthup, Daly, & Wiberley, 1990][Sadtler, 1988].

There are two critical requirements for quantitative analysis of *Sp*. First, the correct solution must be included in the compound candidates. Second, the number of compounds in the compound candidates should be as few as possible.

Because the qualitative analysis is based on what *PC*s the unknown compounds contain, the first requirement can be satisfied in general cases. Concerning the second requirement, the system is also quite satisfactory. The average number of compound candidates in my experiments is about 3.76 which is much better than other known methods. For example, in Sadtler's system, the average number is about twenty [Sadtler, 1988].

7.3 Architecture of the System

The system is developed by using C under MS-WINDOWS. Figure 7.3 shows the architecture of the system.

In the system, main parts include:

1. Inference Engine

   Inference engine is the kernel of the system. It analyzes the spectra of unknown compounds to identify what partial components the unknown compounds contain, and what compounds the unknown compounds are. It also performs the task of dealing with inaccurate data.

2. Knowledge Base

   Knowledge base provides spectroscopic knowledge for interpreting infrared spectra, identifying partial components and compounds, and dealing with related data and qualitative correlations among related data.

3. Data Bases

   Data bases consist of spectroscopic data, references of compounds, and patterns of partial components.

4. Explanation

Explanation is developed for displaying various solving paths and solving methods of the system to help users to understand why and how the inference engine gives the results.

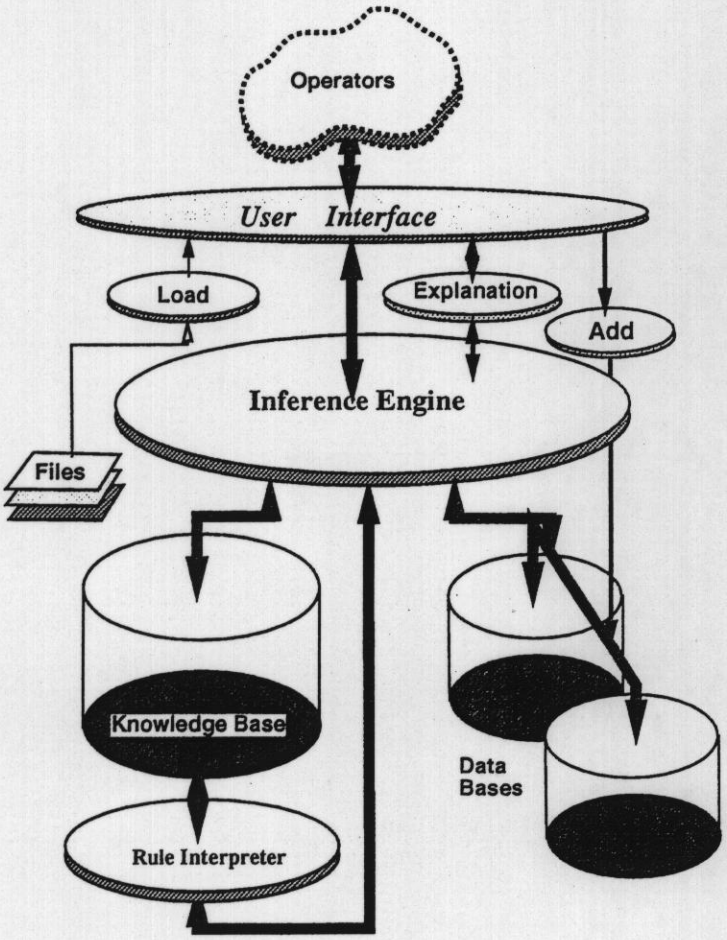

Figure 7.3: Architecture of IR Spectrum Interpreting System

5. Rule Interpreter

   The rules in knowledge base are written in semi-natural language for the convenience of operation. As a result, a rule interpreter is necessary to transfer various rules to the forms that the inference engine can understand.

6. User Interface

   Microsoft windows softwares are used to develop the user interface of the system. Users can use the system under the convenient working environment.

7. Load

   Users can search libraries of references, or load files in disks under the user interface by using *Load*.

8. Add

   Users can add new references to the libraries, or input new unknown compounds in the process of operation by using *Add*.

Figure 7.4 shows an example of the system. Figure 7.4 (a) is the *Sp* of an unknown compound. Figure 7.4 (b) shows the solution of the qualitative process, and Figure 7.4 (c) shows the solution of the quantitative process.

7.4 Summary

In this chapter, I introduced a knowledge-based system for infrared spectrum interpretation. I mentioned the system in Chapter 6 as the background problem of applying the methods for qualitatively interpreting inaccurate data. In this chapter, I gave the overall introduction to system. First, I described the design and architecture of the system. Then, I demonstrated the working process of the system with a real example.

Qualitative reasoning is widely believed to be able to guide and simplify quantitative analysis [Forbus, 1984][Nishida, 1991-1994][Yip, 1991]. The central idea of my research is to conduct qualitative reasoning to narrow down the space of objects at the first stage. Then, at the second stage, complex quantitative

analysis will only be applied to the candidates generated at the first stage. Both the efficiency and quality of the system are improved this way.

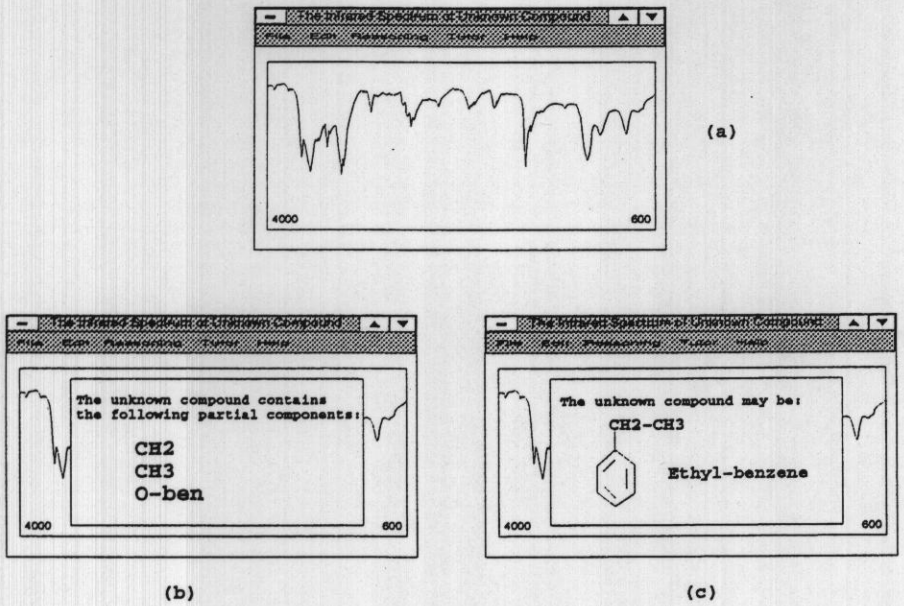

Figure 7.4: An Example of the System

Chapter 8

A Method for Solving Constraint Satisfaction Problems in Infrared Spectrum Interpretation

In this chapter, I introduce a simple method for constraint satisfaction solving. In infrared spectrum interpretation, getting an initial solution is much more difficult than refining the initial solution. Therefore, the most difficult constraints concerning that different variables cannot take on the same value (called constraint α) are considered after the other constraints (called constraint β) have been considered. An efficient *pattern-driven* algorithm is proposed to generate initial solutions which satisfy constraint β. Then an *overlap-reduce* heuristic is applied to minimize the redundancy in the initial solutions until all conflicts among the variables are eliminated.

8.1 Introduction

Constraint satisfaction problem (CSP) is an important problem in reasoning and problem solving [Dechter & Pearl, 1988]. The problem arises in the diagnosis, analysis, interpretation and other AI systems [Minton, and et al, 1990][Stefik, 1981].

Take the infrared spectrum interpretation for example. The problem variables are a set of peaks on an infrared spectrum of an unknown compound:

$$\mathbf{Sp} = \{p_1, p_2, ..., p_n\}$$

where p_i=(f_i, s_i, w_i) represents a peak on the unknown spectrum in which f_i, s_i and w_i are the frequency position, strength and width of the peak, respectively.

The associated domains of the variables are a set of known partial components:

$$\mathbf{PC} = \{PC_1, PC_2, ..., PC_m\}$$

where PC_j =$\{p_{j_1}, p_{j_2}, ..., p_{j_l}\}$ represents a known partial component which can create peaks p_{j_1}, p_{j_2}, ... and p_{j_l} if it is contained by the unknown compound.

The task of IR spectrum interpretation involves assigning a value for every variable:

$$p_i \in PC_j$$

where $p_i \in \mathbf{Sp}$ and $PC_j \in \mathbf{PC}$, that is, determining by which partial components in set **PC** the peaks on the unknown spectrum are created, and what partial components the unknown compound contains. So the solution of IR spectrum interpretation is a subset of **PC**:

$$\mathbf{SL} = \{PC_1, PC_2, ..., PC_k\}, \qquad \mathbf{SL} \subset \mathbf{PC}$$

where

$$\begin{cases} PC_1 = \{p_{1_1}, ..., p_{1_p} \mid p_{1_i} \in \mathbf{Sp} \wedge p \geq 1\} \\ PC_2 = \{p_{2_1}, ..., p_{2_q} \mid p_{2_i} \in \mathbf{Sp} \wedge q \geq 1\} \\ ... \\ PC_k = \{p_{k_1}, ..., p_{k_r} \mid p_{k_i} \in \mathbf{Sp} \wedge r \geq 1\} \end{cases}$$

and

$$\begin{cases} PC_1 \cup PC_2 \cup ... \cup PC_k = \mathbf{Sp} \\ PC_i \cap PC_j = \emptyset \quad (i, j = 1, 2, ..., k, i \neq j) \end{cases}$$

Assigning values for all variables in **Sp** is similar to searching a complete graph constructed by **Sp** to find the possible divisions which has been proved to be a NP Complete problem [Karp, 1975]. The constraints, especially those concerning that several variables cannot take on the same value, or that one variable cannot take on several values, make the problem harder to solve.

The primary methods of solving constraint satisfaction problems are heuristic backtracking and constraint propagation [Bitner & Reingold, 1975][Fox, Sadeh, & Baycan, 1989]. However the searching space and time consumption are two big problems [Brown & Purdom, 1981][Friedrich, Gottlob, & Nejdl, 1991]. What is more, in some practical problems such as infrared spectrum interpretation where the number of variables is huge and the assignment for the variables is complex, generating an initial solution itself is much more difficult than refining it. Traditional methods that consider all constraints to give an initial assignment for the variables and then apply local heuristics to repair the assignment is not efficient.

I propose a simple constraint-based reasoning method to solve this kind of constraint satisfaction problems. The key point of the method is that all constraints are classified into two groups. The difficult constraints like $PC_i \cap PC_j = \emptyset$ $(i, j = 1, 2, ..., k, i \neq j)$ which restrict that different variables cannot take on the same value are classified as constraint α. And the other constraints, like $PC_1 \cup PC_2 \cup ... \cup PC_k = \mathbf{Sp}$ are classified as constraint β. Firstly, only constraint β are considered to generate an initial solution. Because constraint α are not considered at this stage, conflicts among variables in the initial solution are permitted so that the problem of searching a complete graph constructed by **Sp** can be transferred into a process of checking whether the patterns of the possible partial components are contained by **Sp**, which is a *pattern-driven* process without backtracking. Secondly, constraint α are considered to eliminate the conflicts among variables. However, the effective range of these constraints has been narrowed down, and the heuristic for refining the initial solution has become more effective. An *overlap-reduce* heuristic is applied which minimizes the number of partial components in the initial solution so as to eliminate the conflicts among variables.

8.2 Delay of Some Constraints

Among all constraints about variables, some constraints like $PC_1 \cup PC_2 \cup ... \cup PC_k = \mathbf{Sp}$ indicate that every variable p_i in **Sp** must be assigned a value PC_j, that is, $p_i \in PC_j$.

These constraints actually require to search a complete graph constructed by **Sp** to find values for all nodes.

On the other hand, some constraints like $PC_i \cap PC_j = \emptyset$ (i,j=1, 2, ..., k, $i \neq j$) indicate that every variable p_i in **Sp** can only have one value PC_j.

These constraints actually express the restrictions on solutions to avoid multi-assignments for variables in **Sp**.

For example, suppose peak p_1 of **Sp** can be created by partial components PC_1, PC_2 and PC_3 respectively, and the three partial components whose patterns are known as followings will be checked.

$$\begin{cases} PC_1 = \{p_1, p_3, \ \ldots\} \\ PC_2 = \{p_1, p_4, \ \ldots\} \\ PC_3 = \{p_1, q_1, \ \ldots\} \end{cases}$$

Since $q_1 \notin$ **Sp**, partial component PC_3 will not be considered. If all peaks in the patterns of PC_1 and PC_2 can be identified from **Sp**, either PC_1 or PC_2 can be considered as a possible partial component to be included in the initial solution, that is, the following two kinds of assignment for variables p_1, p_3, p_4 and others in PC_1 and PC_2 can be given:

$$\begin{cases} p_1 \in PC_1, \ p_3 \in PC_1, \ \ldots \\ p_1 \in PC_2, \ p_4 \in PC_2, \ \ldots \end{cases}$$

But PC_1 and PC_2 cannot be included in the solution simultaneously because $PC_1 \cap PC_2 = p_1$ violates the constraint of $PC_i \cap PC_j = \emptyset$.

Suppose peak p_2 can be created by partial components PC_4 and PC_5 respectively, and the patterns of PC_4 and PC_5 are:

$$\begin{cases} PC_4 = \{p_2, p_3, \ \ldots\} \\ PC_5 = \{p_2, p_5, \ \ldots\} \end{cases}$$

Similarly, PC_1 and PC_4 cannot exist simultaneously due to $PC_1 \cap PC_4 \neq \emptyset$. The result is that $\{PC_1, PC_5\}$, $\{PC_2, PC_4\}$ or $\{PC_2, PC_5\}$ can be considered as possible partial components, that is, the following three kinds of assignment are all possible.

$$\begin{cases} p_1 \in PC_1, \ p_2 \in PC_5, \ p_3 \in PC_1, \ p_5 \in PC_5, \ \ldots \\ p_1 \in PC_2, \ p_2 \in PC_4, \ p_3 \in PC_4, \ p_4 \in PC_2, \ \ldots \\ p_1 \in PC_2, \ p_2 \in PC_5, \ p_4 \in PC_2, \ p_5 \in PC_5, \ \ldots \end{cases}$$

The above process continues until an assignment for all variables in **Sp** is generated, which has the exponential complexity in general.

If some constraints, like $PC_i \cap PC_j = \emptyset$ (i, j=1, 2, ..., k, $i \neq j$), which restrict the multi-assignments among variables (called constraint α), are considered after

others (called constraint β), an initial solution which satisfies constraint β can be generated much more quickly and easily. An efficient *pattern-driven* algorithm can be used to generate the initial solution. The other important point of the delay is that the effective range of constraint α can be significantly narrowed down, so later, an *overlap-reduce* heuristic can be used to refine the initial solution toward the optimal solution.

The following is the solving procedure:

1. From **PC**, the partial components by which one or more peaks in **Sp** may be created are picked out to form a possible partial component set, no matter whether these partial components satisfy all constraints or not:

$$\forall p_i \forall PC_j((p_i \in PC_j) \rightarrow (PC_j \in \mathbf{PC}'))$$

   where $\mathbf{PC}' \subseteq \mathbf{PC}$.

2. If all constraints except constraint α are considered, an initial solution can be generated from $\mathbf{PC}'$:

$$\forall PC_i((PC_i \in \mathbf{PC}') \wedge \forall p_j((p_j \in PC_i) \rightarrow (p_j \in \mathbf{Sp})) \rightarrow (PC_i \in \mathbf{PC}''))$$

   where $\mathbf{PC}'' \subseteq \mathbf{PC}'$.

3. Because constraint α is not considered in the process of generating $\mathbf{PC}''$, multi-assignments for variables may exist, so $\mathbf{PC}''$ may contain redundant partial components, although all partial components in the optimal solution, **SL**, have been contained by $\mathbf{PC}''$ (i.e., **SL** must be a subset of $\mathbf{PC}''$):

$$\mathbf{PC}'' \supseteq \mathbf{SL}.$$

   Then, considering constraint α can refine $\mathbf{PC}''$ towards **SL**.

8.3 Pattern-Driven Algorithm

Since the constraints such as $PC_i \cap PC_j = \emptyset$ are not considered at the first stage, overlaps among the partial components in $\mathbf{PC}''$ are permitted. If one peak of a partial component is found from **Sp**, then the partial component can be viewed as

a possible partial component, and if all peaks of the partial component are found from **Sp**, then the partial component will be included in the initial solution.

Therefore assigning values for all variables in **Sp** becomes checking whether the patterns of these possible partial components can be found from **Sp** or not, which can be represented as the following procedure.

Procedure I (*Forming a possible partial component set*)

$\mathbf{PC}' = \emptyset$;
for $i = 1$ *to* n { ;;; n is the number of peaks in **Sp**
  for $j = 1$ *to* m { ;;; m is the number of PCs in **PC**
    if $p_i \in PC_j$
      $\mathbf{PC}' = \mathbf{PC}' \cup PC_j$;
  }
}
return$\{\mathbf{PC}'\}$
end procedure I

Suppose there are M possible partial components in $\mathbf{PC}'$, and there are l_j peaks in the pattern of the possible partial component PC_j, then the second procedure can be given in brief:

Procedure II (*Forming an initial solution*)

$\mathbf{PC}'' = \emptyset$;
for $j = 1$ *to* M {
  $Symbol = 1$;
  for $i = 1$ *to* l_j {
    if $p_{j_i} \notin \mathbf{Sp}$ {
      $Symbol = 0$;
      exit;
    }
  }
  if $Symbol = 1$
    $\mathbf{PC}'' = \mathbf{PC}'' \cup PC_j$;
}

return{**PC**″};

end procedure II

Both **Procedure I** and **Procedure II** have linear complexities.

Figure 8.1 shows an unknown spectrum and the partial components in the initial solution generated by the above algorithm.

In the initial solution shown in Figure 8.1, both *partial component i* and *partial component j* recognize *Peak* p_o as their own peak, that is, variable p_o has been assigned twice. Therefore, an overlap between *partial component i* and *partial component j* exists, which means that *partial component i* or/and *partial component j* should be eliminated later from the initial solution.

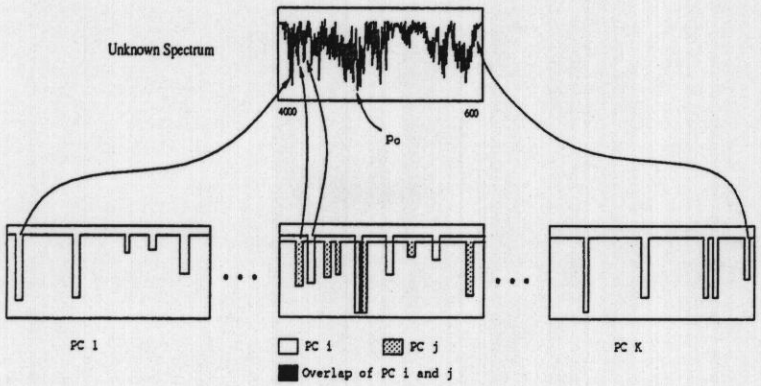

Figure 8.1: Unknown Spectrum and Its Interpretation

In the pattern of a partial component, some peaks are more distinctive than others, so the peaks can be sorted according to their distinctions. Therefore, in practice, partial components are not checked one by one, but a group by a group. When a peak on an unknown spectrum is concerned, the partial components whose most distinctive peaks are the same as this peak will be considered together. Only those whose second distinctive peaks can also be found from the spectrum are left to enter the next layer. The above process is shown in Figure 8.2.

8.4 Overlap-Reduce Heuristic

The *pattern-driven* algorithm can generate an initial solution with linear complexity. The initial solution is not a random assignment, but has the following characteristics:

1. It satisfies all constraints except constraint α such as $PC_i \cap PC_j = \emptyset$.
2. It may contain some redundant partial components, that is, conflicts among variables may exist. However, the partial components in the optimal solution are all included in the initial solution.
3. The difference between the initial solution and the optimal solution is that the former contains more partial components than the latter. Therefore, minimizing the number of partial components in the initial solution may make the initial solution close to the optimal one.

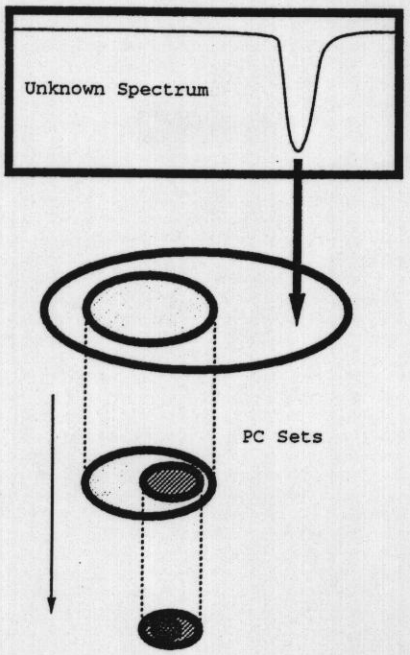

Figure 8.2: Hierarchical Procedure

A repair heuristic called *overlap-reduce* heuristic can be applied to minimize the number of partial components in the initial solution:

1. Pick out partial components having overlaps with others:

$$\forall PC_i \forall PC_j((PC_i \neq PC_j) \wedge (PC_i \cap PC_j \neq \emptyset) \rightarrow (PC_i \notin \mathbf{SL}) \vee (PC_j \notin \mathbf{SL}))$$

2. From the above result, eliminate a partial component, PC_i, whose peaks can be distributed to other partial components, and reassign values for those variables which have been assigned as to PC_i:

$$\forall p(p \in PC_i \rightarrow \exists PC_j((PC_i \neq PC_j) \wedge (p \in PC_j)))$$

Figure 8.3 shows the process of refining the initial solution, where $\circ$ represents the variable, and $\cup$ represents the partial component, and $\bullet/\uplus$ means being in conflict/overlap.

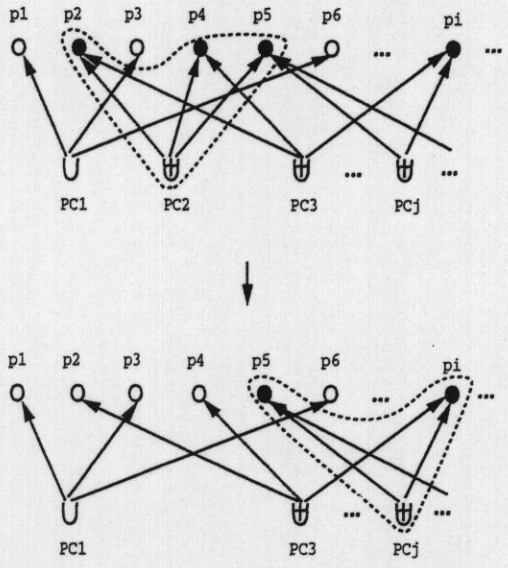

Figure 8.3: Conflict/Overlap Reducing

8.5 Discussion

The method presented in this chapter is a not a perfect method. I discuss the defects and effectiveness of the method in this section.

8.5.1 Defects of the Method

The method for solving constraint satisfaction problems has the following two main defects:

1. The *overlap-reduce* heuristic can be used to reduce the overlaps among variables, but it cannot guarantee that all overlaps in the initial solution can be eliminated. Or in other words, in the worst case, optimal solution can not be obtained with polynomial complexity;

2. The method is only effective to a class of constraint satisfaction problems, such as the constraint satisfaction problems in infrared spectrum interpretation, in which generating an initial solution is much more difficult and complex than refining it, and conflicts in the initial solution are not numerous, but the method may not be effective to other problems.

To some problems, the above defects of the method can not lead to any fatal mistake. I will discuss the effectiveness of the method to these problems in Section 8.5.2.

To some other problems, however, the method may not be effective at all. Because the method is initially developed for solving constraint satisfaction problems in infrared spectrum interpretation only, the research concerning its applications to other problems is left for my future work.

8.5.2 Effectiveness

The method is effective to a class of constraint satisfaction problems in which getting an initial solution is much more difficult and complex than refining it, and conflicts in the initial solution are not numerous.

I discuss the effectiveness of the method in the following two aspects:

1. Constraint satisfaction problem is a difficult problem in AI. Sometimes, even human experts can hardly eliminate all conflicts with a time limitation. The strategy of the proposed method is to use repair heuristic

to eliminate as many conflicts as possible with the limited computational resources;

2. To many problems, incomplete solutions are definitely unacceptable, but superfluous solutions are usually acceptable, because further means can be employed to analyze the superfluous solutions. For example, in infrared spectrum interpretation, if not all partial components are identified, many useful inferences can not be drawn. But if extra partial components are identified incorrectly, using the knowledge about the possible combination of partial components may detect extra partial components. The method guarantees that the optimal solutions are completely included in its solutions, although it does not guarantee that all of its solutions are optimal.

8.6 Summary

In this chapter, I introduced a simple method for constraint satisfaction solving. First, I mentioned that in infrared spectrum interpretation, getting an initial solution is much more difficult than refining it. Then, I discussed that the most difficult constraints concerning that different variables cannot take on the same value (called constraint α) can be considered after the other constraints (called constraint β) have been considered. Based on this idea, an efficient *pattern-driven* algorithm was proposed to generate initial solutions which satisfy constraint β, and an *overlap-reduce* heuristic was proposed to minimize the redundancy in the initial solutions until all conflicts among the variables are eliminated.

The method described in this chapter is not a optimal method, and is only effective to a certain class of constraint satisfaction problems. In this chapter, I also discussed the defects and effectiveness of the method.

Chapter 9

Related Work and Discussion

My work roughly falls into four areas of AI: (1) qualitative interpretation of inaccurate data, (2) possibility propagation and uncertain reasoning, (3) knowledge-based system for infrared spectrum interpretation, and (4) constraint satisfaction problems.

Interpreting inaccurate data has long been regarded as a significant and difficult problem in AI. Many methods and techniques have been proposed.

Fuzzy logic provides the mathematical fundamentals of representation and calculation of inaccurate data [Bowen, Lai, & Bahler, 1992][Negoita & Ralescu, 1987][Zadeh, 1978 & 1989]. My method for qualitatively interpreting inaccurate data is primarily based on fuzzy theory. But compared with conventional fuzzy techniques, the advantages of the method include: (1) shift intervals of inaccurate data are dynamically determined so that dynamic information can be used; (2) shift intervals are based on qualitative features of data and qualitative correlations among related data so that the solutions are more robust. The limitation of the method is that when qualitative correlations among related data are not known in advance, the method degenerates to a conventional fuzzy method. For instance, if SCF is unavailable, the two methods described in Section 6.5 become the same.

Pattern recognition provides the techniques for interpreting measured data in group [Jalsovszky & Holly, 1988][Raskutti & Zukerman, 1991]. By using pattern recognition methods, related data and connections among data can be considered. However, there are two preconditions which must be satisfied for complex data analysis by pattern recognition to be successful. The first precondition is that we have to obtain adequate data bases from which we can derive the patterns we need to recognize, and the second precondition is that we have to demonstrate

that there are suitable metrics of similarity between patterns. When patterns explicitly exist, and measured patterns are not seriously noisy, pattern recognition methods are effective. However, if patterns are not explicit, or patterns change irregularly which implies that there is not a stable metrics for determining the similarity between patterns (e.g., spectrum interpretation), the proposed method is more practical and robust.

Probabilistic reasoning provides a practical framework for reasoning under uncertainty [Dempster, 1968] [Duda, Hart, & Nilsson, 1976][Pearl, 1988][Shafer, 1976]. For example, by using Bayesian theory, uncertain evidence can be calculated and propagated on inference networks. In many systems, subjective statements are used to take the place of statistics of uncertain evidence when statistical samples are insufficient or absent, such as certainty factors in MYCIN [Shortliffe & Buchanan, 1975], and prior probabilities in PROSPECT [Duda, Hart, & Nilsson, 1976]. My method for propagating qualitative correlations as evidence of uncertain reasoning is similar to the method for probabilistic reasoning. However, the essential difference is that my method dynamically calculates qualitative correlations as evidence so it does not need many assumptions in advance, and can avoid inconsistency in knowledge and data bases.

When statistical samples are sufficient, or subjective statements can be consistently obtained like in MYCIN and PROSPECTOR, probabilistic reasoning methods can be applied. When statistical samples of inaccurate data are not enough and consistent subjective statements are not available, the proposed method is very effective.

Traditional methods and systems of infrared spectrum recognition are primarily based on quantitative analysis techniques which identify infrared spectra of unknown compounds by calculating the quantitative similarity or closeness between the infrared spectra of known and unknown compounds [Jalsovszky & Holly, 1992][Sadtler, 1988]. Due to the huge number of known compounds, these methods and systems usually require users to provide a range to which the unknown compounds belong in advance, then apply quantitative analysis to the known compounds in the range. In some cases, users may roughly provide the information by using the physical features of unknown compounds, but in more cases they can not. When users can not provide the information, or the information provided is not certain, quantitative analysis may be very complex. In contrast, the presented knowledge-based system analyzes what partial components the unknown compounds contain, and then analyzes the partial components to determine what the unknown compounds may be, so it can always give

a very limited list of candidates in which the unknown compounds are included. Therefore, the complexity of quantitative analysis can be significantly reduced.

Some recently developed systems also recognize infrared spectra by decomposing them [Hasenoehrl, Perkins, & Griffiths, 1992]. Compared with these systems, my system has two advantages. First, it uses the new concepts of PUP, SSS and VFP to generate the preliminary solutions, so its efficiency is very high. Second, it uses qualitative correlations among related peaks as evidence, so the inaccuracy of spectral data can be effectively handled. Both RC and RI of my system are higher than those of the systems using fuzzy or other techniques to deal with the inaccuracy [Anand, Mehrotra, Mohan, & Ranka, 1991].

My future research concerning the system is to consider the interaction among identified partial components. As I discussed before, spectroscopists frequently use the knowledge like: *"if C_6H_6 coexists with CH_3, then the peaks of CH_3 around 2900 cm^{-1} may shift"*, or *"if -C-O-C- has been identified, then the strength of the peaks of CH_3 may change"*. Therefore, it is possible to update the possibilities of identified partial components by considering the interaction among them. Analyzing the effects among identified partial components would not only help us identify inaccurate data, but also provide us with the reason why the data are inaccurate.

Constraint satisfaction problems have invited various research and applications [Dechter & Pearl, 1988]. To a certain class of constraint satisfaction problems where generating an initial solution is very difficult, the proposed method is efficient, since it can employ a *pattern-driven* algorithm to quickly generate an initial solution in which the optimal solution must be contained. However, the proposed method is not an optimal method, since it cannot guarantee that its final solution be optimal in any time. Developing the method further is another research of my future work.

Chapter 10

Conclusions

In this dissertation, I have presented a novel method for qualitatively interpreting inaccurate data by using qualitative correlations among related data as confirmatory or disconfirmatory evidence. First, I introduced a new concept called support coefficient function (SCF). Then, I proposed an approach to determining dynamic shift intervals of inaccurate data based on SCF, and an approach to calculating possibilities of interpreting inaccurate data, respectively. Based on these two approaches, I introduced a method for using qualitative correlations among related data as confirmatory or disconfirmatory evidence for the interpretation of inaccurate data.

I have presented a novel method for propagating qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence for uncertain reasoning. The method can extract, represent and propagate qualitative correlations among hypotheses as confirmatory or disconfirmatory evidence to update the possibilities of hypotheses. The function of the method is similar to the probability propagation on Bayesian networks. But compared with traditional methods for probability propagation, the method has the following advantages: (1) it can be applied to the problems where evidence is not explicitly given; (2) few numbers and assumptions need to be provided by domain experts in advance; and consequently, (3) the knowledge acquisition is simple, and the inconsistency in knowledge bases can be avoided.

I have applied the above two methods to infrared spectrum interpretation, and have fully tested the methods against several hundred real infrared spectra. The experimental results show that the methods are significantly better than the traditional methods used in many similar systems.

I have also presented a knowledge-based system for infrared spectrum interpretation. The primary task of the system is to identify unknown compounds by interpreting their infrared spectra. I proposed a knowledge model for integrating qualitative reasoning into infrared spectrum interpretation. The implementation of the system indicates that both the efficiency and quality are improved by employing the knowledge model.

Finally, I have presented a new method for solving constraint satisfaction problems. I proposed an efficient *pattern-driven* algorithm for generating initial solutions, and an *overlap-reduce* heuristic for repairing the initial solutions, respectively. I discussed the advantages, disadvantages and applicability of the method.

Briefly, my contributions mainly include:

1. A qualitative method which interprets inaccurate data by using qualitative correlations among related data as confirmatory or disconfirmatory evidence, and a corresponding algorithm which crystallizes the method;

2. A qualitative method which propagates qualitative correlations among hypotheses to update possibilities of them, and a corresponding algorithm which crystallizes the method;

3. Successful applications of the above two qualitative methods to a practical problem;

4. A knowledge-based system which integrates qualitative reasoning and quantitative analysis to interpret infrared spectra;

5. A simple method for solving constraint satisfaction problems including an efficient pattern-driven algorithm for generating initial solutions, and an overlap-reduce heuristic for repairing the initial solutions.

List of Publications

1. Zhao, Q. & Nishida, T. (1995a). Using Qualitative Hypotheses to Identify Inaccurate Data. *Journal of Artificial Intelligence Research (JAIR), Morgan Kaufmann Publishers, Vol. 3, pp. 119-145.*

2. Zhao, Q. & Nishida, T. (1995b). Qualitative Interpretation of Spectral Images: Reasoning with Uncertain Evidence. *Proc. of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95), Montreal, Canada, pp. 42-47.*

3. Zhao, Q. & Nishida, T. (1995c). Making and Refining Qualitative Interpretation for Spectral Images. *IEICE AI, Vol. 94, No. 57, pp. 9-16.*

4. Zhao, Q. & Nishida, T. (1995d). Qualitative Decomposition and Recognition of Spectral Images. *IEICE Trans. on Information and Systems* (to appear).

5. Zhao, Q. & Nishida, T. (1995e). Propagating Qualitative Correlations as Evidence for Uncertain Reasoning. submitted to *International Journal of Approximate Reasoning.*

6. Zhao, Q. & Nishida, T. (1994a). A Knowledge Model for Infrared Spectrum Processing. *Proc. of International Symposium on Information Theory and Its Applications (ISITA'94), Sydney, Australia, pp. 781-786.*

7. Zhao, Q. & Nishida, T. (1994b). Integrating Qualitative Reasoning into Complex Quantitative Analysis: Introduction to a Real System. in *Zhang, Debenham, & Lukose (eds.), Artificial Intelligence – Sowing the Seeds for the Future, pp. 299-306, World Scientific.*

8. Zhao, Q. (1994). An Efficient Method of Solving Constraint Satisfaction Problems in IR Spectrum Interpreting. *Proc. of the 2nd International Con-*

ference on Expert Systems for Development (ICED'94), Bangkok, Thailand, pp. 165-170.

9. Zhao, Q. & Mikami, T. (1994). A Model of Handling Uncertainty in Expert Systems. *Proc. of the 2nd World Congress on Expert Systems (WCES'94), Lisbon, Portugal, pp. 105-114.*

10. Zhao, Q. (1993a). Summary on KIRSAS - A Knowledge-Based IR Spectrum Analyzing System. *ASTEM Technical Report: TR-P-059-93, Kyoto, Japan.*

11. Zhao, Q. (1993b). New Progress on KIRSAS. *ASTEM Technical Report: TR-P-057-93, Kyoto, Japan.*

12. Zhao, Q. (1993c). Fuzzy Techniques for IR Spectra Interpreting. *ASTEM Technical Report: TR-P-056-93, Kyoto, Japan.*

13. Zhao, Q. (1993d). Computer-Based IR Spectra Analyzing. *ASTEM Technical Report: TM-P-030-93, Kyoto, Japan.*

14. Zhao, Q. (1993e). An Algorithm of Inferring Compounds from Atomic Groups. *ASTEM Technical Report: TM-P-029-93, Kyoto, Japan.*

15. Zhao, Q. (1993f). An Expert System for IR Spectra Interpreting, Inquiring and Tutoring. *ASTEM Technical Report: TR-P-051-92, Kyoto, Japan.*

16. Zhao, Q., Mikami, T., & Sugimoto, T. (1992). KIRSAS: An Knowledge Model for IR Spectra Interpretation. *ASTEM Technical Report: TR-P-045-92, Kyoto, Japan.*

17. Zhao, Q., Sugimoto, T., & Mikami, T. (1991). Knowledge-based IR Spectrum Analyzing System – Prototype, Related Concepts and Techniques. *ASTEM Technical Report: TR-P-036-91, Kyoto, Japan.*

18. Lin, Y., Zhao, Q., & Chen, Y. (1990). The Planning Stowage Expert System and Its Subdividing Method of General Cargo. in *Balagurusamy & Howe (eds.), Expert Systems for Management and Engineering, pp. 308-320, Ellis Horwood Limited.*

19. Lin, Y. & Zhao, Q. (1989). The Subdividing Method in General Cargo Stowage Planning Expert System. *Proc. of the 1st International Conference on AI in Industry and Government, Hyderabad, India.*

20. Zhao, Q. & Lin, Y. (1988). A Model of Space Planning and Its Application to Building an Expert System. *Proc. of the 8th International Workshop on Expert Systems and Their Applications, Avignon, France.*

21. Zhao, Q. & Lin, Y. (1987). DF-BT, A New Technique of Heuristic Searching. *Proc. of the 3rd National Conference on AI and PR, Taiyuan, China.*

22. Zhao, Q. (1986a). An Expert System for General Cargo Stowage Planning. *M.Sc. Thesis, Tsinghua University, Beijing, China.*

23. Zhao, Q. (1986b). A Forward Reasoning by Using Meta Rules. *Proc. of the 5th National Conference on Machine Intelligence, Xian, China.*

24. Zhao, Q. (1984). Pattern-Based Natural Language Understanding Techniques. *B.Sc. Thesis, Tsinghua University, Beijing, China.*

Bibliography

Anand, R., Mehrotra, K., Mohan, C. K., & Ranka, S. (1991). Analyzing Images Containing Multiple Sparse Patterns with Neural Networks. *Proc. of IJCAI'91, pp. 838-843.*

Becraft, W. R., Lee, P. L., & Newell, R. B. (1991). Integration of Neural Networks and Expert Systems for Process Fault Diagnosis. *Proc. of IJCAI'91, pp. 832-837.*

Berry, P. M. (1992). SCHEDULING: A Problem of Decision-Making Under Uncertainty. *Proc. of 10th European Conference on Artificial Intelligence, pp. 638-642.*

Biswas, G. & Yu, X. (1993). A Formal Modeling Scheme for Continuous System: Focus on Diagnosis. *Proc. of IJCAI'93, pp. 1474-1479.*

Bitner, J. R. & Reingold, E. M. (1975). Backtrack Program Techniques. *Comm. ACM, Vol. 18, pp. 651-655.*

Blaffert, T. (1986). An Expert System for Infrared Spectra Evaluation. *Chemica Acta, 191(86), pp. 161-168.*

Bose, P. & Rajamoney, S. A. (1993). Compositional Model-Based Design. *Proc. of IJCAI'93, pp. 1445-1450.*

Bousson, K. & Trave-Massuyes, L. (1993). Fuzzy Causal Simulation in Process Engineering. *Proc. of IJCAI'93, pp. 1536-1541.*

Bowen, J., Lai, R., & Bahler, D. (1992). Lexical Imprecision in Fuzzy Constraint Networks. *Proc. of AAAI'92, pp. 616-621.*

Brown, C. A. & Purdom, J. (1981). An Average Time Analysis of Backtracking. *SIAM J. Comput. Vol. 10.*

Clerc, J. T., Pretsch, E., & Zurcher, M. (1986). Performance Analysis of Infrared Library Search Systems. *Mikrochim. Acta[Wien], II, pp. 217-242.*

Cohen, P. R. (1984). Heuristic Reasoning about Uncertainty: An Artificial Intelligence Approach. *Pitman.*

Cohen, P. R. (1987). The Control of Reasoning under Uncertainty: A Discussion of Some Programs. *The Knowledge Engineering Review, 2(1), pp. 5-25.*

Colthup, L., Daly, H., & Wiberley, S. E. (1990). Introduction to Infrared and Raman Spectroscopy. *Academic Press INC.*

Console, L., Friedrich, G, & Dupre, D. T. (1993). Model-Based Diagnosis Meets Error Diagnosis in Logic Programs. *Proc. of IJCAI'93, pp. 1494-1499.*

Cullen, P. B., Hull, J. J., & Srihari, S. N. (1992). A Constraint Satisfaction Approach to the Resolution of Uncertainty in Image Interpretation. *Proc. of CAIA'92, pp. 127-133.*

Dechter, R. & Pearl, J. (1988). Network-Based Heuristics for Constraint-Satisfaction Problems. *Artificial Intelligence, Vol. 34, pp. 1-38.*

de Kleer, J. & Williams, B. (1987). Diagnosing Multiple Faults. *Artificial Intelligence, Vol. 32, pp. 97–130.*

Dempster, A. P. (1968). A Generalization of Bayesian Inference. *Journal of the Royal Statistical Society, Vol. B-30, pp. 205-247.*

Duda, R. O., Hart, P. E., & Nilsson, N. J. (1976). Subjective Bayesian Methods for Rule-Based Inference Systems. *Proc. of National Computer Conference, pp. 1075-1082.* or *Tech. Note 124, SRI Int., Menlo Park, Ca.*

Duda, R. O., and et al. (1977). Development of the Prospector Consultation System for Mineral Exploration. *Ann. Rep. SRI Proj. 5821 and 6415, AI Center, SRI Int., Menlo Park, Ca.*

Forbus, K. (1983). Measurement Interpretation in Qualitative Process Theory. *Proc. of IJCAI'83, pp. 315-320.*

Forbus, K. (1984). Qualitative Process Theory. *Artificial Intelligence, Vol. 24, pp. 85-168.*

Forbus, K. (1987). Interpreting Observations of Physical Systems. *IEEE Tran. on Systems, Man, and Cybernetics, Vol. SMC-17, No. 3, pp. 113-117.*

Fox, M. S., Sadeh, N., & Baycan, C. (1989). Constrained Heuristic Search. *Proc. of IJCAI'89, pp. 309-315.*

Friedrich, G., Gottlob, G., & Nejdl, W. (1991). Formalizing the Repair Process. *Proc. of 2nd Int'l Workshop on Principles of Diagnosis, pp. 11-22.*

Fringuelli, B., Marcugini, S., Milani, A., & Rivoira, S. (1991). A Reasoning Maintenance System Dealing with Vague Data. *Proc. of 7th Int'l Conference on Uncertainty in AI, pp. 111-117.*

Hasenoehrl, E. J., Perkins, J. H., & Griffiths, P. R. (1992). Expert System Based on Principal Components Analysis for the Identification of Molecular Structures from Vapor-Phase Infrared Spectra. *Journal of Anal. Chem., Vol.64, pp. 656-663.*

Huberman, B. & Struss, P. (1989). Chaos, Qualitative Reasoning and the Predictability Problem. *Proc. of 3rd Int'l Workshop on Qualitative Physics.*

Iwasaki, Y., Fikes, R., Vescovi, M, & Chandrasekaran, B. (1993). How Things Are Intended to Work: Capturing Functional Knowledge in Device Design. *Proc. of IJCAI'93, pp. 1516-1522.*

Jalsovszky, G. & Holly, G. (1988). Pattern Recognition Applied to Vapour-Phase Infrared Spectra: Characteristics of vOH Bands. *Journal of Molecular Structure, Vol. 175, pp. 263-270.*

Karp, R. M. (1975). On the Computational Complexity of Combinational Problems. *Networks, Vol. 5, pp. 45-68.*

Kawata, S., and et al. (1987). Spectral Searching by Fourier-phase Correlation. *Applied Spectroscopy, Vol. 41, pp. 1176-1182.*

Kleiter, G. D. (1992). Bayesian Diagnosis in Expert Systems. *Artificial Intelligence, Vol. 54, pp. 1-32.*

Kruse, R. (1984). Statistical Estimation with Linguistic Data. *Information Sciences, Vol. 33, pp. 197-207.*

Kruse, R., Gebhardt, J., & Klawonn, F. (1994). Foundations of Fuzzy Systems. *John Wiley & Sons.*

Kuipers, B. J., and et al. (1988). Using Incomplete Quantitative Knowledge in Qualitative reasoning. *Proc. of AAAI'88, pp. 324-329.*

Laskey, K. B. & Lehner, P. E. (1989). Assumptions, Beliefs and Probabilities. *Artificial Intelligence, Vol. 41, pp. 65-77.*

Luinge, H. J., and et al. (1987). Artificial Intelligence Used for the Interpretation of Combined Spectral Data(3): Automated Generation of Interpretation Rules for Infrared Spectral Data. *Chem. Inf. Comput. Sci., 27 (87), pp. 95-99.*

Minton, S., Johnston, M. D., Philips, A. B., & Laird, P. (1990). Solving Large-Scale Constraint Satisfaction and Scheduling Problems Using a Heuristic repair Method. *Proc. of AAAI'90, pp. 17-24.*

Moldoveanu, S. & Rapson, C. A. (1987). Spectral Interpretation for Organic Analysis Using an Expert System. *Anal. Chem., Vol. 59, pp. 1207-1212.*

Mukaidono, M., Shen, Z., & Ding, L. (1989). Fundamentals of Fuzzy Prolog. *Int'l Journal of Approximate Reasoning, Vol. 3, pp. 179-193.*

Negoita, C. V. & Ralescu, D. (1987). Simulation, Knowledge-Based Computing, and Fuzzy Statistics. *Van Nostrand Reinhold Company Inc.*

Nishida, T. & Doshita, S. (1991). A Geometric Approach to Total Envisioning. *Proc. of IJCAI'91, pp. 1150–1155.*

Nishida, T., and et al. (1991). Automated Phase Portrait Analysis by Integrating Qualitative and Quantitative Analysis. *Proc. of AAAI'91, pp. 811-816.*

Nishida, T. (1993). Generating Quasi-symbolic Representation of Three-dimensional Flow. *Proc. of AAAI'93, pp. 554–559.*

Nishida, T. (1994). Qualitative Reasoning for Automated Exploration for Chaos. *Proc. of AAAI'94, pp. 1211-1216.*

Oppenheim, A. V. & Nawas, S. H. (1992). Symbolic and Knowledge-Based Signal Processing. *Prentice Hall.*

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. *Morgan Kaufmann Publishers, San Mateo, California.*

Puskar, M. A., Levine, S. P., & Lowry, S. R. (1986). Computerized Infrared Spectral Identification of Compounds Frequently Found at Hazardous Waste Sites. *Journal of Anal. Chem., Vol.58, pp. 1156-1162.*

Ramer, A. & Lander, L. (1991). Formal Model of Uncertainty for Possibilistic Rules. *Proc. of the 7th Int'l Conference on Uncertainty in AI, pp. 295-299.*

Raskutti, B. & Zukerman, I. (1991). Handling Uncertainty during Plan Recognition in Task-Oriented Consultation Systems. *Proc. of the 7th Int'l Conference on Uncertainty in AI, pp. 308-315.*

Reiter, R. (1987). A Theory of Diagnosis From First Principles. *Artificial Intelligence, Vol. 32, pp. 57-95.*

Riese, M. (1993). Diagnosis of Communicating Systems: Dealing with Incompleteness and Uncertainty. *Proc. of IJCAI'93, pp. 1480-1485.*

Robb, E. W. & Munk, M. E. (1990). A Neural Network Approach to Infrared Spectrum Interpretation. *Mikrochim. Acta[Wien], I, pp. 131-155.*

Sacks, E. (1991). Automatic Analysis of One-parameter Planar Ordinary Differential Equations by Intelligent Numeric Simulation. *Artificial Intelligence Vol. 48, pp. 27–56.*

Sadtler Research Laboratories. (1988). Sadtler PC Spectral Search Libraries, Product Introduction & User's Manual. *Sadtler Research Lab.*

Savitzky, A. (1987). The Evolution of an Automated Infra-Red Spectra Interpretation System. *Computer-Enhanced Analytical Spectroscopy, Vol. 1, Plenum Press, pp. 183-199.*

Shafer, G. (1976). A Mathematical Theory of Evidence. *Princeton Uni. Press, Princeton.*

Shortliffe, E. H. (1976). Computer-Based Medical Consultations: MYCIN. *American Elsevier Publishing Inc.*

Shortliffe, E. H. & Buchanan, B. G. (1975). A Model of Inexact Reasoning in Medicine, *Mathematical Biosciences, Vol. 23, pp. 351-379.*

Stefik, M. (1981). Planning with Constraints. *Artificial Intelligence, Vol. 16, pp. 111-139.*

Vescovi, M. R. & Robles, J. (1992). Fuzzy Diagnosis of Continuous Processes. *Proc. of ECAI'92, pp. 749-753.*

Wang, J. C. (1994). Identifying Key Missing Data for Inference Under Uncertainty. *Int'l Journal of Approximate Reasoning, Vol. 10, No. 4, pp. 287-309.*

Wang, P. Z. (1983). From the Fuzzy Statistics to the Falling Random Subsets. *Wang, P. P. ed., Advances in Fuzzy Sets, Possibility and Applications, pp. 81-96, Plenum Press.*

Wythoff, B. J., Buck, C. F., & Tomellini, S. A. (1989). Descriptive Interactive Computer-Assisted Interpretation of Infrared Spectra. *Analytica Chimica Acta, Vol. 217, pp. 203-216.*

Yip, K. M. (1991). Understanding Complex Dynamics by Visual and Symbolic Reasoning. *Artificial Intelligence, Vol. 51(1-3), pp. 179–222.*

Zadeh, L. A. (1978). Fuzzy Set as a Basis for a Theory of Possibility. *Fuzzy Sets Syst., Vol. 1, pp. 3-28.*

Zadeh, L. A. (1989). Knowledge Representation in Fuzzy Logic. *IEEE Trans. on Knowledge and Data Engineering Vol. 1, No. 1, pp. 89-100.*

Zhao, F. (1991). Extracting and Representing Qualitative Behaviors of Complex Systems in Phase Spaces. *Proc. of IJCAI'91, pp. 1144–1149.*