A pollen identification expert system: An application of expert system techniques to biological identification.

A thesis presented in partial fulfilment
of the requirements for the degree of
Master of Science
in Computer Science.

Colin G. Eagle

Massey University

1990

Abstract

The application of expert systems techniques to biological identification has been investigated and a system developed which assists a user to identify and count air-borne pollen grains. The present system uses a modified taxonomic data matrix as the structure for the knowledge base. This allows domain experts to easily assess and modify the knowledge using a familiar data structure. The data structure can be easily converted to rules or a simple frame-based structure if required for other applications. A method of ranking the importance of characters for identifying each taxon has been developed which assists the system to quickly narrow an identification by rejecting or accepting candidate taxa. This method is very similar to that used by domain experts.

Acknowledgements

My sincere thanks go to my supervisor, Mr Ray Kemp, for his guidance and useful criticisms during the development and presentation of this research.

I would also like to thank the domain experts, Dr Clive Cornford, Mr Richard Burr and Dr David Fountain of the Botany and Zoology Department for their willingness to provide time and expertise.

Thanks are due to the staff of the School of Information Sciences, who have provided encouragement during the production of this report.

I would like to thank my parents, Gordon and Judy Eagle, for their encouragement and support.

Finally I wish to thank my wife, Susan, for her patience, support and understanding during this time.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Purpose

The purpose of the present study is to investigate the suitability of using expert systems technology in the field of biological identification, using pollen identification as an example. In addition, the present study examines the use of a taxonomic data matrix as the core of the knowledge base structure, and also develops a method of assigning importance values to characters.

Chapter 1 describes the objectives of the present study, and presents an investigation into expert systems technology and design, graphical user interfaces and pattern recognition and their relevance to expert systems. Chapter 2 investigates the techniques of biological identification and how expert system techniques are applied to this. Chapter 3 contains descriptions of prototype systems for pollen identification which were intended to determine the practicality of expert systems for pollen identification. Chapters 4 describes the user view, knowledge base organisation and inference engine of the present system. Chapter 5 contains a summary of results achieved and proposals for future developments of the present system.

1.2 Objectives of this project

The main objective of the present study is the development of an expert system designed to quickly and accurately identify and count New Zealand pollens based on morphological descriptions given by the user. The system is

designed to run on a computer beside a microscope, assisting the user to interactively identify and count the pollens seen in the microscope.

This study is designed to meet an identified need for a pollen identification system for use in allergen research. Pollen allergens have been identified by the World Health Organization as a research priority. In New Zealand current research (Cornford, Fountain, Burr & O'Leary, 1988) has aimed to build a reference bank of pollens and their extracts, measuring the occurrence of hazardous pollens in the atmosphere, and purifying pollen extracts for use in allergen analysis and treatment programs. This research requires the collection of pollens from throughout New Zealand. Identification of these is primarily carried out by trained but non-specialised staff. These staff would be assisted by an expert system designed to take into account a variety of interacting factors which are crucial to an accurate analysis of pollens. Experienced staff would also benefit from a system which enables them to identify unusual pollens.

In addition to allergen research, there are several other fields where pollen identification may be assisted by an expert system. For example, forensic scientists may need to investigate the approximate area and season in which a crime took place. Apiculturalists can benefit from a pollen identification system to ensure optimum placement of hives, and in palynology pollen identification can aid understanding of plant distribution and geology (Kemp, Greenwood, Tse & Eagle, 1988).

The present study was designed primarily to assist those involved in allergen research. It is intended that the completed system will be used to:

- assist the user to count different types of pollen;

- lead the non-specialised user through an identification process, asking for data which either confirm or negate the most likely candidate pollen;
- assist more experienced users in routine identification and in identifying unusual pollens, via an option which omits the questioning process and allows direct description of an unidentified pollen;
- report when it is not possible to clearly differentiate between two pollens;
- explain the process used to confirm or negate candidate pollens;
- be easily amended to provide identifications in other fields of biological identification;
- incorporate a graphical user interface so that the system is simple and intuitive to use;
- be easily extended to incorporate real-time pollen recognition.

1.3 Expert Systems

Expert (knowledge-based) systems are computer programs which can solve 'real world' problems, that is, problems for which a solution requires judgement and experience. The emphasis of expert systems is on the heuristic knowledge which reflects the experience of the expert and the structure of that knowledge, rather than on reasoning from first principles (Michaelsen, Michie & Boulanger, 1983; Wolfgram, Dear & Galbraith, 1987).

An important aspect of expert systems is a capability for explaining their knowledge of the domain and the reasoning processes used to produce results and recommendations. This assists users and system builders to understand the contents of the system's knowledge base and reasoning processes, and

facilitates the debugging of the system during development. It educates users about both the domain and the capabilities of the system, and gives information which assures users that the system's conclusions are correct. Explanation can also help a user to discover when the limits of the system's knowledge are being exceeded (Moore & Swartout, 1988).

In order to make use of judgemental knowledge, expert systems normally include a method for reasoning with uncertainty. This allows better modelling of expert behaviour, including the use of guesses and degrees of belief (Atkinson & Gammerman, 1987).

Other useful aspects of expert systems include the capacity to mimic human reasoning, making the logical progress toward a problem solution easily understood by users. It is also possible to build generalisable systems, that is, an expert system designed to identify one type of biological specimen can, by changing the knowledge base, be used to identify another type of specimen (Woolley & Stone, 1987).

Hayes-Roth, Waterman and Lenat (1983), Wolfgram et al (1987) and Poo and Lu (1989) have identified distinct categories of expert systems designed to solve particular types of problems.

Firstly, fixed instant diagnosis systems (i.e., those in which interpretation of a diagnosis at a point in time depends on the data available), may be used, for example, in medical, electronic, mechanical and software diagnosis (Poo et al, 1989). MYCIN is an example of a medical diagnosis system which attempts to diagnose infectious blood diseases from available knowledge or data supplied by a physician. Clancey (1984) has described various methods of designing fixed instant diagnosis systems.

Secondly, interpretation systems can be used in areas such as surveillance, speech understanding, image analysis and signal interpretation. They attempt

to explain observed data by assigning to them symbolic meanings describing the system state accounting for the data (Hayes-Roth et al, 1983). DENDRAL analyses experimental chemical data in order to infer the plausible structures of an unknown compound (Wolfgram et al, 1987).

Thirdly, prediction systems infer likely consequences from given or hypothetical situations (Wolfgram et al, 1987). This category includes weather forecasting, demographic predictions, traffic predictions and military forecasting.

Planning systems compose sequences of actions for achieving some prescribed effect. This category includes automatic programming, and robot, route, experiment and military planning problems (Hayes-Roth et al, 1983).

Configuration systems construct descriptions of objects in various relationships with one another, and verify that these configurations conform to stated constraints (Wolfgram et al, 1987). These systems include computer configuration (e.g., R1, the DEC VAX computer equipment configuration system), circuit layout, building design and budgeting.

Advice giving systems use recommendations and explanations in attempting to provide the user with a supportive environment for problem solving (Coombs & Alty, 1984; Jackson and Lefrere, 1984). This category includes plan formation and computer programming.

Finally, computer-aided instruction systems incorporate diagnosis and debugging subsystems that address the student as the system of interest. Typically, these systems construct a model of the students knowledge which interprets the students behaviour, diagnose weaknesses in the students knowledge, identify an appropriate remedy, and then plan a tutorial intended to convey the remedial knowledge to the student (Hayes-Roth et al, 1983; Farrell, Anderson & Reiser, 1984; Clancey & Bock, 1988).

The architecture of a typical expert system consists of a fact base, a knowledge base, an inference engine and an explanation facility (Hayes-Roth et al, 1983; Ramsey, Reggia, Nau & Ferrentino, 1986; Poo et al, 1989). (See Figure 1). A fact base may be defined as a store of unchanging knowledge about the domain of interest of the expert system. A knowledge base consists of extensive knowledge regarding the domain of interest, and is used to make inferences about unknown facts, based on information in the fact base. An inference engine is responsible for control of the problem solving process, that is, manipulating the knowledge base, updating the state of the world, and remembering the chain of reasoning being used. It makes use of knowledge in the knowledge base in order to reason about the problem using information in the fact base. In order to provide a more transparent and explainable design, Buchanan and Duda (1983) and Clancey et al (1988), have proposed that inference procedures be represented abstractly, as rule sets, separate from the domain knowledge they operate on. This has advantages for design and maintenance of the system, making it easier to debug and modify, as hypotheses and search strategies are not embedded in rules. The explanation facility of the inference engine consists of an identification of steps used in the reasoning process and justification of each step.

The knowledge bases of expert systems are commonly divided into two types of knowledge representation: rules and frames. Rule-based (production) systems consist of the knowledge and experience of a human expert encoded into a set of rules which consist of antecedents (conditional statements) that define a pattern or state; and consequents, that is, instructions to be carried out in the event that the current state matches the hypothetical pattern described in the antecedent (Woolley et al, 1987). The skill of a rule-based system increases at a rate proportional to the enlargement of its knowledge

base. Rule-based systems are modular, in that each rule defines a small, relatively independent piece of knowledge; this allows relatively simple addition of new rules and updating of old rules (Bratko, 1986). By adaptively selecting the best sequence of rules to execute, and by combining the results in appropriate ways, rule-based systems can solve a wide range of possibly complex problems. They can explain their conclusions by retracing lines of reasoning and translating the logic of each rule into natural language (Hayes-Roth, 1985).

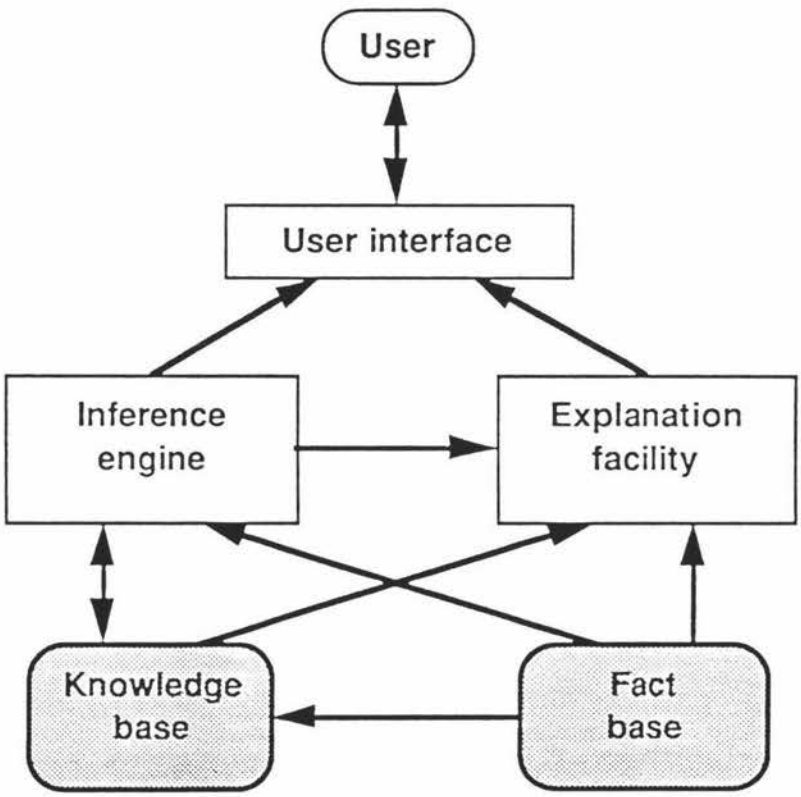

Figure 1: Structure of a typical expert system

(arrows show direction of information flow).

Frame-based expert systems are based on a structured representation of an object or a class of objects (a frame). Frames incorporate sets of attribute descriptions called slots, which are used to describe attributes of the object or class represented by the frame. Constructs are available which allow an

expert system designer to describe relationships between frames (Hayes, 1979; Brachman, 1983; Fikes & Kehler, 1985; Wolfgram et al, 1987). For example, birds can be described as animals in addition to a set of properties which distinguish birds from other classes of animals.

In addition to rule- and frame-based systems, Tschudi (1988) has proposed another type of knowledge representation based on a matrix similar to a taxonomic data matrix. The knowledge in the matrix can easily be encoded to produce rules or a decision tree.

1.4 Graphical User Interfaces

Apple Computer (1987) have defined a computer interface as:

> ". . . the sum of all communication between the computer and the user. It's what presents information to the user and accepts information from the user. It's what actually puts the computer's power into the user's hands." (p. xi).

Graphical user (direct-manipulation) interfaces are common on many types of computer system. They provide a human-computer interface which is easier to learn and simpler and more pleasant to use than the traditional command-line interface (Gould & Lewis, 1983; Foley & van Dam, 1984).

Direct manipulation interfaces have been defined by Shneiderman (1983) as a variety of graphical user interface in which the user sees a continuous representation of the world of action. The objects of interest and the permissible actions on those objects are represented on the screen in a visual format which takes into account the user's knowledge of the task domain. Physical actions replace typed commands and actions are rapid, incremental and reversible. These design principles lead to several important benefits. Users with knowledge of the domain find the system easy to learn, users need

learn only a small number of computer concepts, and can therefore concentrate on the task. In addition, designers can reduce the number of situations in errors can be made, users feel free to explore 'what-if' possibilities, and long-term retention is facilitated (Baroff, Simon, Gilman & Shneiderman, 1986).

In expert systems, effective use of direct manipulation interfaces can assist in containing complexity and make the system intuitive and credible to use. This can be done by exploiting the user's expectations regarding how ideas are organized and expressed within the system domain (Potter, 1988). Direct manipulation interfaces have been used in expert system development, allowing designers to display rules and heuristics in graphical format and to graphically display actual and possible interactions between rules (Poltrock, Steiner & Tarlton, 1986; Baroff et al, 1986).

In addition to expert system development, direct manipulation interfaces may be used in the user-computer interface. For example, 'The Student Advisor' (Baroff et al, 1986), assists students in planning course schedules and uses a windows and buttons in order to simplify the interface. The apple problem diagnosis system (Kemp & Boorman, 1987) attempts to determine the cause of inadequate quality or quantity of fruit using 'windows', 'icons', 'mice' and 'pull-down menus' ('wimps') for more effective user interaction and therefore allowing the user to adapt quickly to the system, even though he/she may not use it for extended periods.

1.5 Pattern Recognition

Pattern recognition refers to the act of recognising a given object from a complex input stream. For example, identifying a chair from the wider class of 'furniture' (Pao & Ernst, 1982). Three interrelated but distinct processes take

place during a typical pattern recognition process. Data acquisition is the process of converting incoming data from its physical source (pictures, speech, character string, etc.) into an acceptable form for further processing. Pattern analysis is concerned with organising the converted body of data into a form for further processing by determining the different pattern classes which might exist in the data. Finally, pattern classification refers to the process whereby pattern classes are matched with a known class (Chien, 1978). Pattern classification has used expert systems techniques since the early 1960's, particularly where there is imperfect correspondence between input data and a known class (Ballard, Brown & Feldman, 1977; Ogawa, Kurioka, Kitahashi & Tanaka, 1980; Brady, 1982; Magee & Nathan, 1985). For example, galaxy classification (Thonnat, Granger & Berthod, 1985), inspection of mechanical parts (Kanal, 1974), and the interpretation of medical images to provide diagnoses (Ellam & Maisey, 1986).

The application of computerised pattern recognition has been largely directed toward computer vision (e.g., object classification) and speech recognition. A summary of pattern recognition techniques has been provided by Rohlf and Ferson (1983).