G.R.E.E.N.
An Expert System to identify Gymnosperms

Antonio B. Bailón, Eva Gibaja, Ramón Pérez
Department of Computer Science and Artificial Intelligence
University of Granada

Carmen Quesada
Herbario de la Universidad de Granada

Keywords: Gymnosperms, Identification Keys, Expert Systems, Artificial Intelligence, World Wide Web, Iberian Peninsula, Certainty Factors

Abstract: The application of Artificial Intelligence (AI) techniques to the problem of botanical identification is not particularly widespread even less so on Internet. There are several interactive identification systems but they usually deal with raw knowledge so it appears that "research and development of web-based expert systems are still in their early stage" (Li et al., 2002). In this paper we present the G.R.E.E.N. (*Gymnosperms Remote Expert Executed over Network*) system as an expert system for the identification of Iberian Gymnosperms which allows on-line uncertainty queries to be made. The system is operative and it can be consulted in http://drimys.ugr.es/experto/index.html.

1 INTRODUCTION

Plant Taxonomy is a complex, meticulous science that allows taxa to be identified by retrieving information contained on them in a classification system. There are various ways which this identification may be carried out, although the one most commonly used employs dichotomic keys (a process which requires knowledge of botanical terminology and organography). As a result of the complexity of this process, botany-related activities are not particularly automated.

A number of interactive identification systems have been reported in the literature. Taking into account the data structure chosen to represent the knowledge we can distinguish basically two kinds of systems: **matrix-based identification systems** like INTKEY (Dallwitz et al., 1993), MEKA (Meacham, 1996) and **rule-based expert systems** like IKBS (Grosser et al., 1999) and RIH (Grove et al., 1999). We can also divide interactive identification systems in **on-line** systems like NAVIKEY (Bartley, 1999), LUCid (CPITT, 1999), POLLYCLAVE (Dickinson, 1999) or INTKEY and **non-on-line** systems like MEKA or XID (XID, 1999). Dallwitz have done a comparison of interactive identification programs (Dallwitz, 2000) and it seems that INTKEY and LUCid are the most complete. Some of the characteristics described in his paper are not contained in our system (like guidance about the next character to use, and subsets), nevertheless, we introduce a very desirable and not too much studied characteristic in other identification systems: the management of uncertainty and imprecise information.

AI offers a productive approach to identification by managing uncertainty in order to obtain a better response when user's observations don't match exactly with the set of characters represented in the system. Successful rule-based expert systems and a strong mathematical theory have been developed since the first expert system (DENDRAL in 1965) to the present time. In spite of this, intelligent identification systems that deal with uncertainty are not particularly widespread. In this sense only a few systems have been related (Atkinson et al., 1987; Fermanian et al., 1989). Their main disadvantage is that the botanical expert must provide a probability distribution. In this paper we propose an alternative

B. Bailón A., Gibaja E., Pérez R. and Quesada C. (2004).
G.R.E.E.N. - An Expert System to identify Gymnosperms.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 216-221
DOI: 10.5220/0002642702160221

to avoid these disadvantages. The alternative is the use of expert systems whose uncertainty management technique is the certainty factors theory.

Within the wealth and variety offered by the plant kingdom, the subject of scientific disclosure has been dealt with a specific study of the group of *Gymnospermae* (46 autochthonous and cultivated taxa present in the Iberian Peninsula). This group was chosen due to the presence in this area of important forest species which it contains. In addition, many of these offer resources or are cultivated as ornamental, which makes their identification useful for non-botanical expert users.

This has all given rise to G.R.E.E.N. (*Gymnosperms Remote Expert Executed Over Networks*), an on-line decision aid system that applies AI techniques of machine learning and uncertainty management to the field of Botany.

2 SYSTEM STRUCTURE

The system structure is derived from the way in which botanical experts work. Particularly, dichotomic keys of the type IF-THEN are used for the recognition of plant species. That is to say, that each key leads to either another key or a plant species. When a botanist wants to identify a particular species, it is possible to distinguish:

- A source of knowledge comprising all the available information on each plant species in the form of dichotomic keys.
- A process of the use of this knowledge. Keys are searched until a particular species is identified.

This description coincides with that of a knowledge-based system and more specifically with that of a rule-based expert system with: a **knowledge base** which stores knowledge about the domain of the problem in the form of rules and an **inference engine** which extracts information from the knowledge base. In addition to these two modules, the system has:

- An **uncertainty-processing module** fitting the nature and subjectivity of the observer.
- A **justifying module** which explains the results achieved by the system in a language close to the natural language.
- A **multimedia database** to reference known species. This module provides images and data about species, ecology and distribution.
- A **glossary of scientific terms** to make the system more accessible to users who are not botanical experts.
- A **server** which will deal with user (client) requests and sends back the results by Internet. We can see the system's architecture in Fig. 1.

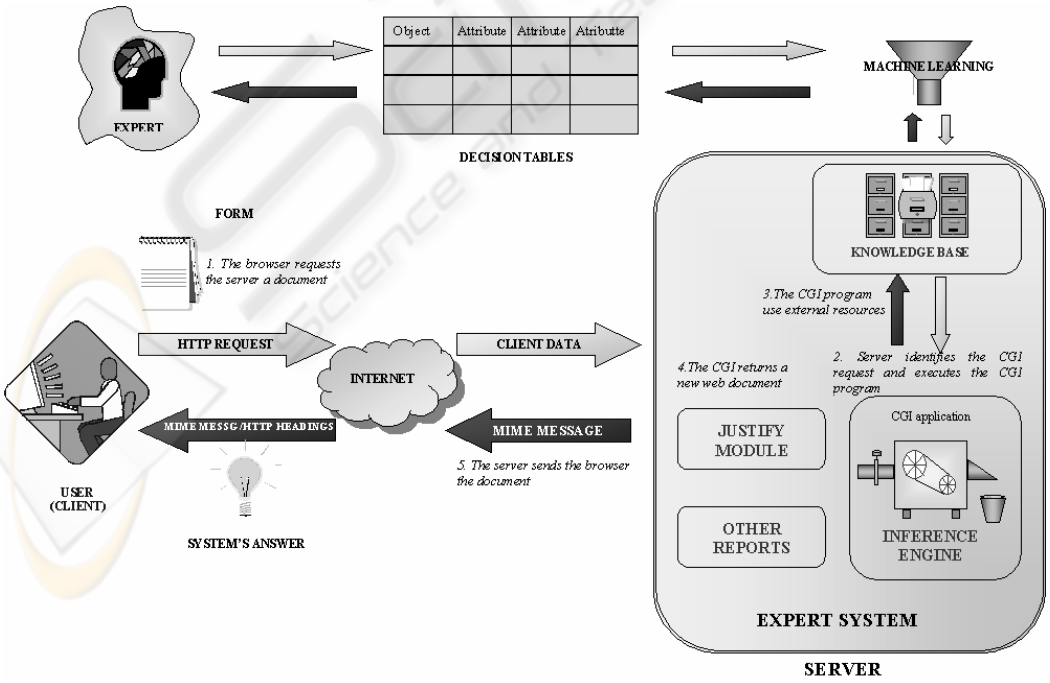

Figure 1: System's architecture

3 KNOWLEDGE ACQUISITION AND ELICITATION

The information available on the problem domain is dispersed, incomplete, imprecise and unstructured. Particularly, this complexity is manifested in:

- Character dependence. The presence of some character may depend on the presence/absence of other ones.
- Differentiating attributes. These attributes difference only one taxon so when they are present we can directly do an identification while this information is not relevant to identify other taxa.
- Continuous and uncertain values.

In order to be able to represent the knowledge in an appropriate way, a process of knowledge acquisition and elicitation was needed. The acquisition and elicitation process began with different dichotomous keys whose information was gathered and summarized, thereby producing a list of diagnostic characters (descriptors or attributes) and values at **family, genus, species** and **subspecies** level.

This hierarchical organization of the information offers the advantage of multilevel answers. Generally, only a small amount of information (which is also what is observed more easily) is needed in order to reach an objective in the higher levels of the hierarchy. Obviously, the more information we have, the more we will know. Due to the inherent complexity of taxonomical information all information was subsequently compared several times by observing nature and consulting herbalist documents and experts.

The most important taxonomical characters in Gymnosperms were divided into different groups: general aspect of the taxon, characteristics of the leaf, of the branches, of the shoots, monoecious or dioecious, characteristics of the fructification (cone and "berry" cone), of the seeds, and ecology of the taxon. With these characters and with additional expert information, decision tables (Durkin, 1994) were compiled, which gathered the identifying diagnostic characters for each taxon (see Table 1).

Table 1: A fragment of the decision table for Iberian Gymnosperms Families

Family	**General Appearance**	**Resiniferous**	**Leaf Characters**	**(.....)**
Ephedraceae	Shurb	No	Scale-like	(.....)
Cupressaceae	Tree and Shurb	Yes	Acicular and Scale-like	(.....)
Taxaceae	Tree and Shurb	No		(.....)
Pinaceae	Tree	Yes	Acicular	(.....)
Araucariaceae	Tree	Yes		(.....)
Taxodiaceae	Tree	Yes	Acicular	(.....)
Cephalotaxaceae	Tree	No		(.....)
Cycadaceae	Palm	No		(.....)
Ginkgoaceae	Tree	No	Fan-Shaped	(.....)

4 TREATMENT OF UNCERTAINTY

Information about the domain is based on what normally happens, but every rule has its exceptions. “We must take into account diversity and incompleteness, and exception is the only valid rule” (Grosser et al., 1999).

As it is usual for some data not to be known with absolute certainty errors of measurement may be committed. Given this large amount of sources of uncertainty, the system incorporates a module to deal with uncertainty.

Uncertainty is modelled using **certainty factors** (Shortlife, 1975) since it is a simple computational model which allows experts to estimate confidence in each hypothesis and in the conclusion, facilitating the expression of subjective certainty estimations. This model also enables knowledge to be represented easily in the form of rules and has successfully been used in many other systems. So on one hand the user can tells the system how sure is he about his own observations and on the other hand the system is able to give a response with a certainty degree associated based in the certainty of rules and user’s data.

Other advantage of the use of certainty factors is that they can be automatically estimated when we obtain the rule base from tables so the expert doesn’t have to give the system any probability distribution.

5 OBTAINING THE KNOWLEDGE BASE

A set of **rules** with a certainty factor associated (the knowledge base) was obtained automatically from the decision tables. For this, a modification of the ID3 algorithm proposed by Quinlan (Ignizio, 1991) was used in order to obtain more than one rule per objective. Rules of minimum length (entropy determines the minimum set of diagnostic characters in order to recognize a taxa) were created which

excluded irrelevant knowledge, since irrelevant descriptors were not taken into account. We obtained keys which were different from the standard ones. Particularly whose content is more complete than that of the dichotomic keys, since the knowledge base contains all the consistent rules which may be obtained according to the selected descriptors in order to determine the objectives.

The rules provide a structuring of the knowledge which the user can understand and which is similar to the dichotomic keys used by expert botanists. When the system presents its conclusions in the form of rules, the user understands the reasoning followed by the system perfectly and becomes familiar with the reasoning process followed by the human experts who have contributed their knowledge to the system (learning).

Additional advantages of computerized systems are discussed in (Dallwitz, 2000).

6 CONSISTENCY REINFORCER

During the development of the knowledge base, inconsistencies may arise mainly due to errors during the knowledge acquisition and elicitation stage. The system is capable of accommodating uncertainty which is why inconsistencies about the certainty of results cause an additional impact.

Consequently, this makes it necessary for the system to incorporate a **consistency reinforcer** which systematically analyses each of the rules in the knowledge base in order to guarantee its consistency and completeness.

Checking for **consistency** includes detecting redundant rules, conflicting rules, subsumed rules, rules with unnecessary conditions and circular rules while checking for **completeness** means checking for missing rules, unreferenced attribute values, illegal attribute and decision values, unreachable conditions and unreachable goals. The algorithm designed is based in (Grzymala-Busse, 1991) and it is shown in Fig. 2.

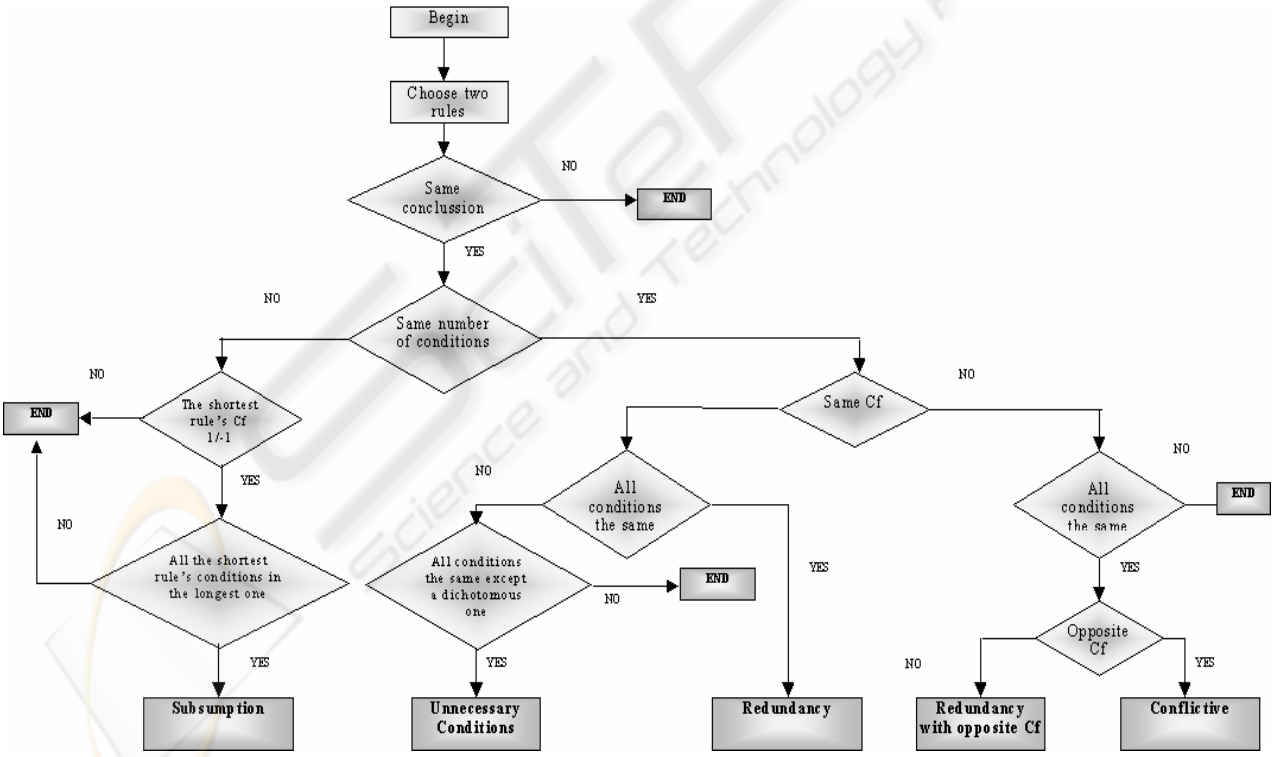

Figure 2: Algorithm to check the consistency

7 SYSTEM REASONING

The **inference engine** provides the control mechanism and knowledge inference (the process used in order to derive new information from known information). It combines the input facts with the knowledge gathered in the knowledge base thereby responding to user queries. In order to design the inference engine, Ignizio's BASELINE with forward chaining and a modification to deal with certainty

factors has been taken as a model (Ignizio, 1991). The inference engine incorporated into the system is quite a different module from the knowledge base. This differentiation is important since:

- The system designers can capture and organize the knowledge common to the problem domain independently of its implementation.
- It enables the content of knowledge base to be changed without the need to change the control system
- A single inference engine may be used to solve different problems.

8 OTHER CHARACTERISTICS

The system may be easily adapted in order to classify species other than Gymnosperms.

It is also easy to use. The specimen descriptors are grouped into general categories (general appearance, leaf, branch, cone, etc.) with familiar to all users names (see Fig. 3.a). Within each category, users select the descriptor they know and enter a value for the degree of belief.

The system provides two methods for entering the query: **basic** (the user has a set of options, so that the use of certainty factors is clearer) and **advanced** (the user manually enters the certainty value of the observation).

After entering the data, the inference process is executed and the system presents the user with a set of results (ordered according to how well they fit the query) and an outline of the reasoning followed in order to reach these conclusions (see Fig. 3.c).

If the user wishes, it is possible to increase the information about the specimen by accessing the multimedia database.

As the system has been specifically designed to work on Internet, the entire on-line transfer of information has been minimized so as not to overload the server and in order to obtain a satisfactory system response time for the user. For example, suppose a user has done an observation where “leaf characters” is “for sure scale-like” (see Fig. 3.b) and “for sure is resiniferous”, with only this information the system concludes that the item observed was a *Cupressaceae* whose CF value is 1.

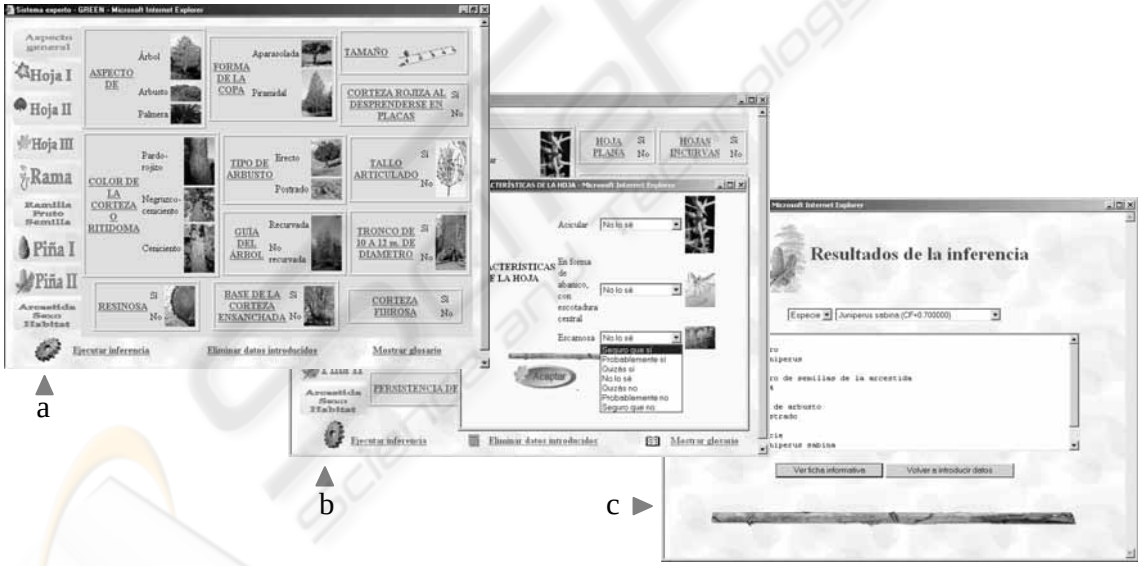

Figure 3: a) The user interface for introduction of data. b) A view of the character “characteristics of the leaf”. c) The user interface for identification results

If the user introduces “fruit consistence” is “fleshy” the system reaches the conclusion “genus is *Juniperus*”. By adding “number of seed of the berry cone is two to four” and “type of shrub” is “probably postrated” the system concludes “*Juniperus sabina*” with CF equal to 0.7 (see Fig. 3.c). The system also could have reached the same objective with a different input. For example, it could conclude the item was a *Cupressaceae* with “fruit consistence fleshy”, “seed with fleshy aril=no”, “leaf persistence=persistent”, “brownish leaf=no”, “sexual character=dioecius” and “numerous seeds = no”.

10 CONCLUSIONS

By way of conclusion and to sum up:

- In this paper, an expert system is presented which will offer the user a new interactive species identification method whose main contribution is the use of intelligent techniques to deal with uncertainty
- It solves problems in which incomplete data is handled. This is an important feature, since in taxonomical classification processes information and observations are not complete.
- IA and Internet technology offer new advantages to the popularisation of Botany.
- The user can easily learn the features to observe through interaction with the query interface. It also is able to spread expert knowledge by justifying the solution, so that the user can learn the reasoning followed by the system.
- The model to represent the knowledge is also useful in order to produce automatic keys or computer-generated keys.
- The system is a practical operative tool which may be used on-line and which will enable different taxa comprising the Iberian Gymnosperm flora to be recognized.
- The model can be extended to other branches of Biology as well; it is a question of generating a suitable knowledge base and a user interface, while the inference system does not change.

REFERENCES

Atkinson, W. D. & Gammerman, A. 1987. An application of expert systems technology to biological identification. Taxon 36, 705–714.

Bartley. M. 1999. http://www.herbaria.harvard.edu/computerlab/web_keys/navikey/. Consulted 17/2/03

Castroviejo, S., Laínz, M., López González, G., Montserrat, P., Muñoz Garmendia, F., Paiva, J. & Villar, L. 1986. *Flora Ibérica. Plantas vasculares de la Península Ibérica e Islas Baleares.* Vol. I ed. Real Jardín Botánico, Madrid.

Centre for Pest Information Technology and Transfer (CPITT) at the University of Queensland .1999. http://www.lucidcentral.com/, Consulted 17/2/03

Dallwitz, M. J., Paine, T. A. & Zurcher, E. J. 1993. User's guide to the DELTA System: a general system for processing taxonomic descriptions. 4th edition. *http://biodiversity.uno.edu/delta/*

Dallwitz, M. J. 2000. A comparison of interactive identification programs. *http://biodiversity .uno.edu/delta*.

Dickinson T. 1999. http://prod.library.utoronto.ca/polyclave/. Royal Ontario Museum Canada. Consulted 17/2/03

Durkin, J. 1994. *Expert Systems. Design and development*. Ed. Prentice Hall International, London .

Fermanian T.; Michalski R. S. 1989. Weeder: An Advisory System for the Identification of Grasses in Turf. *Agronomy Journal* ,81(2), pp 313-316

Font Quer, P. 1979. *Diccionario de Botánica*. Ed. Labor, Barcelona.

García Rollán, M. 1983. *Claves de la flora de España. Península y Baleares.* Vol. I., ed. Ed. Mundi-Prensa, Madrid.

Grosser D. & Conruyt N. 1999. Tree-based classification approach for dealing with complex knowledge in natural sciences. *Proceedings of ACAI'99 - Machine Learning and Applications. Chania (Greece).*

Grove R. F. & Hulse A. C. 1999. An Internet-Based Expert System for Reptile Identification. *The First International Conference on the Practical Application of Java, London*

Grzymala-Busse, J.W. 1991. *Managing Uncertainty in Expert Systems*. Ed. Kluwer Academic Publishers.

Ignizio, J. P. 1991. *Introduction to Expert Systems. The Development and Implementation of Rule-Based Expert Systems*. Ed. McGraw-Hill, New York.

Li D., Fu Z. & Duan Y. 2002. Fish - Expert: a web-based expert system for fish disease diagnosis. *Expert Systems with Applications* 23, 311–320.

Meacham C. 1996, The Meka for Windows FAQ page. http://ucjeps.berkeley.edu/meacham/meka/ Jepson Herbarium U.C. Berkeley, Consulted 17/2/03

Shortlife, E., & Buchanan, B. G. 1975. A Model of Inexact Reasoning in Medicine. *Mathematical Biosciences* 23: 351-379.

XID Services, Inc. 1999. http://www.xidservices.com/. Consulted 17/2/03