GREEN An Expert System to identify Gymnosperms

W. Fajardo
University of Granada

Eva Gibaja
University of Cordoba (Spain)

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

Waldo Fajardo, Eva Gibaja, Pedro Moral
Department of Computer Science and Artificial Intelligence
University of Granada
Periodista Daniel Saucedo Aranda, 18071, Granada, Spain
E-mail: aragorn, gibaja@decsai.ugr.es
Laura Baena, Carmen Quesada
Herbario de la Universidad de Granada.
University of Granada
C/ Rector López Argüeta, 8. Colegio Mayor Isabel La Católica. 18071. Granada, Spain.
E-mail: laurab, cquesada@ugr.es

Summary

The application of **Artificial Intelligence** 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" [26]. In this paper we present the G.R.E.E.N. (*Gymnosperms Remote Expert Executed Over Networks*) System as an expert System for the identification of Iberian Gymnosperms which allows **on-line uncertainty** queries to be made. The System can be consulted in http://drimys.ugr.es/experto/index.html.

Keywords

Gymnosperms, Identification Keys, Expert Systems, Artificial Intelligence, World Wide Web, Iberian Peninsula, certainty factors.

1.Introduction

Plant Taxonomy is a complex, meticulous science which 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 are 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 [9], MEKA [30] and **rule-based expert systems** like IKBS [19] and RIH [20]. 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. We can also divide interactive identification systems in on-line systems like NAVIKEY [1], LUCid [28], POLLYCLAVE [14] or INTKEY or non-on-line systems like MEKA or XID [36]. Dallwitz have done a comparison of interactive identification programs [12] and it seems that INTKEY and LUCid are the most complete . Some characteristics described in his paper are not contained in G.R.E.E.N. (like guidance about the next character to use, and subsets). Nevertheless we introduce in our system a very desirable and not much studied characteristic in other identification systems: the manage of uncertainty and imprecise information. So Artificial Intelligence offers a productive approach to identification systems by a management uncertainty technique in order to obtain a better response when user's observations don't match exactly with the set of characters represented in the system. In this sense only a few identification systems have been related (Atkinson & Gammerman, 1987; Fermanian & Michalski 1989) but their main disadvantage is that the botanical expert must provide a probability distribution. In this paper we propose an alternative 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 using Artificial Intelligence techniques with a specific study of the group of Gymnosperms (*Gymnospermae*) in the Iberian peninsula. This group was chosen due to the presence 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*), a system that applies Artificial Intelligence techniques of Machine Learning and Manage of uncertainty to the field of botany. G.R.E.E.N. is an online decision aid System, resulting in a great and fast diffusion of knowledge and a broader receptor spectrum.

2.Material and methods

We have divided this study on G.R.E.E.N. into 5 parts:

- A first part (sections 1 and 2) in which we describe the structure of the system, and defines the main modules which comprise the system and the knowledge gathered.
- A second part (section 3 to 6) in which we develop the process for acquiring and validating the knowledge available on the problem domain until a knowledge base is finally obtained. In this part, the processing of imprecise information, common to this type of problem, is also discussed.
- A third part (section 7) is devoted to the reasoning process which the System uses.
- A fourth part (section 8) in which we discuss other important features of the System.
- Finally, we finish (section 9) with conclusions drawn directly from what has been presented in this article and from the bibliography used.

3. System structure

The System structure is directly derived from the way in which botanical experts work. Dichotomic keys of the type IF-THEN are used for the classification and recognition of plant species. That is to say, that each key leads to either another key or a plant species. In this way, 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 rules.
- A process of the **use of this knowledge** in order to solve the particular problem (keys are searched until a particular species is identified).

This description coincides perfectly 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 the two essential modules described in the previous paragraph and reflecting the ideal structure of a Knowledge-Based System, the System has:

- An **uncertainty processing module** fitting the nature and subjectivity of the observer.
- A **justifying module** which explains the results achieved to the System in a language close to the natural language.

We will also add user support modules.

- A **multimedia database** to reference known species.
- A **glossary of scientific terms** to make the System more accessible to users who are not botanical experts.

Additionally to design and implement a **server** which will deal with user (or client) requests and send the results by Internet is needed.

In next section we outline the process for the design and implementation of the System, detailing the Artificial Intelligence techniques which have been applied.

4. Knowledge gathered by the System

The first stage is to determine its application domain, that is, the type of knowledge the System will manage. As we have mentioned, the group of Gymnosperms has been chosen from which information is provided on 46 taxa present in the Iberian Peninsula, both autochthonous and cultivated.

In addition to the Knowledge Base, which has been optimized in order to obtain results in the queries, the System gathers information on the System domain in other formats and these are incorporated into a multimedia database which provides images and data about its distribution and ecology and a glossary of botanical terms which make the arduous task of species identification easier and more enjoyable.

5. Knowledge acquisition and elicitation

The first problem when developing the System is that the information available on the problem domain does not have a structure which may be directly translated to a Computer System. The information is dispersed, incomplete; it is imprecise and unstructured. In order to be able to represent the knowledge in an appropriate way, a process of knowledge acquisition and elicitation is needed, and on which the final functions of the System depend to a large extent. In order to begin the acquisition and elicitation process, we begin with different keys. We gather and summarize their information, thereby producing a list of diagnostic characters (descriptors or attributes) at **family, genus, species** and **subspecies** level. This hierarchical organization of the information offers the advantage of multilevel answers so that, even with little information, some objective may be reached in the higher levels of the hierarchy. This has a simple explanation:

Generally, in order to reach an objective in the higher levels of the hierarchy only a small amount of information is needed, which is also what is observed more easily. Heuristically, this leads us to suppose that the minimum amount of information which the user knows will be that which will allow inference in the highest levels. As information becomes known, the response will be refined until the lower and less general levels of the hierarchy are reached. The more information we have, the more we will know, nevertheless, results may generally be obtained with little information. All information has subsequently been compared by observing nature and consulting herbalist documents and experts.

The most important taxonomical characters in Gymnosperms have been 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, decision tables [15] have been compiled from traditional dichotomic keys, This tables gather the identifying diagnostic characters for each taxon “Table 1”. As it is not advisable for these tables to have many empty cells (the more information, the better), they have been filled in since many were not necessary when the taxon were identified using the traditional method.

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

	Aspecto General/ Appearance	Resinosa/ Resiniferous	Características de la hoja/ Leaf Characters	(........)
Ephedraceae	Arbusto/Shurb	No	Escamosa/ Scale-like	(........)
Cupressaceae	Arbol y Arbusto/Tree and Shurb	Si/Yes	Acicular y Escamosa/Acicular and Scale-like	(........)
Taxaceae	Arbol y Arbusto/Tree and Shurb	No		(........)
Pinaceae	Arbol/Tree	Si/Yes	Acicular/ Acicular	(........)
Araucariaceae	Arbol/Tree	Si/Yes		(........)
Taxodiaceae	Arbol/ Tree	Si/Yes	Acicular/ Acicular	(........)
Cephalotaxaceae	Arbol/ Tree	No		(........)
Cycadaceae	Palmera/ Palm	No		(........)
Ginkgoaceae	Arbol/ Tree	No	En forma de abanico /Fan-Shaped	(........)

Although initially filling in a table of this type supposes a greater effort than using dichotomic keys directly, this investment is easily compensated for since these will enable us to apply Artificial Intelligence techniques in order to obtain keys which are different from the standard ones.

Botany uses identification keys, whereas applied Artificial Intelligence techniques determine the minimum set of diagnostic characters in order to recognize the different taxa. Artificial Intelligence allows us to find determining characters, which exclude others, and this enables quicker identification than that provided by the traditional method.

6. Treatment of uncertainty

Information about the domain is based on what normally happens, but every rule has its exceptions. As it is usual for some data not to be known with absolute certainty and since expert knowledge is not always defined with complete certainty, errors of measurement may be committed. But this does not mean that the information that we have should be rejected as not only are experts able to work with uncertainty but good results can also be obtained regardless.

Given this large amount of sources of uncertainty, G.R.E.E.N. incorporates a module to deal with uncertainty. Uncertainty is modeled using **certainty factors** [35] 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. A certainty factor (**CF**) is a number which is associated with each component of a condition or a consequence of a rule and represents a **degree of belief**. This number is usually in the range –1(definitely false) to +1 (definitely true). A positive value represents a degree of belief while a negative value indicates a degree of disbelief. There are rules to propagating certainty factors while the inference runs and several rules are activated. So on one hand the user can tells the system how sure is he about his own observations and in 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.

7. Obtaining the Knowledge Base

A set of **rules** with a certainty factor associated (represented in the Knowledge Base) is obtained automatically from the tables. For this, we apply Artificial Intelligence learning techniques (*machine learning*), in particular we modify the ID3 algorithm proposed by Quinlan [23], so that it allows us to obtain more than one rule per objective. For this:

- We use Occam's razor criterion as a heuristic for ramification (“*simple explanations are preferable to more complex explanations*”) quantifying this criterion through the use of the concept of entropy. In this way, rules of minimum length are created which exclude irrelevant knowledge, since irrelevant descriptors will not be taken into account.

- We obtain a Knowledge Base, the content of which is more complete than that of the dichotomic keys, since it 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 G.R.E.E.N. perfectly and the user becomes familiar with the reasoning process followed by the human experts who have contributed their knowledge to the System (learning).

8. Consistency reinforcer

During the development of the Knowledge Base, inconsistencies may arise mainly due to errors during the knowledge acquisition and elicitation stage.

Another important note is that G.R.E.E.N. is capable of accommodating uncertainty which is why inconsistencies about the certainty of results cause an additional impact. Consequently, this makes it necessary for G.R.E.E.N. to incorporate a **consistency reinforcer** which systematically analyzes 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 and checking for **completeness** means checking for missing rules, unreferenced attribute values, illegal attribute and decision Values, unreachable conditions and unreachable goals [21].

9. System reasoning

The **Inference Engine** provides the control mechanism and knowledge inference (a process used in an expert System in order to derive new information from information known). 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 has been taken as a model [23].

The Inference Engine incorporated into the System is quite a different module from the Knowledge Base. This differentiation is important since:

1. Knowledge may be represented more naturally. The knowledge model together with the inference process reflects the problem-solving mechanism followed by a human being better than a model which incrusts knowledge within the inference process.
2. The System designers can focus on capturing and organizing the knowledge common to the problem domain independently of its implementation.
3. It enables the content of Knowledge Base to be changed without the need to change the control System so that a) the Knowledge Base may be updated without changing the Inference Engine b) a single Inference Engine may be used to solve different problems.

10. Other characteristics

G.R.E.E.N. is specifically designed to work on Internet which is why interaction with the user is carried out using forms which send the data and the queries to a remote server. The entire transfer of information online has been minimized so as not to overload the server and in order to obtain a satisfactory System response time for the user.

G.R.E.E.N. has been designed independently of the type of botanical database on which it is employed, so that it may be easily adapted in order to classify species other than Gymnosperms.

As we have already mentioned, G.R.E.E.N. is extremely easy to use (see Fig. 1). The specimen descriptors are grouped into general categories (general appearance, leaf, branch, cone, etc.) with names which are familiar to all users. Within each category, users select the descriptor they know and enter a value for the degree of belief.

The System has been provided with two methods for entering the query: basic and advanced. In the basic mode, the user has a set of options, so that the use of certainty factors is clear. In the advanced mode, the user must manually enter the certainty value of the observation. After entering the data, the inference process is executed and the System gives the user a set of results and an outline of the reasoning followed in order to reach these conclusions. The results are ordered according to how well they fit the query . In addition the user can increase the information about the specimen by accessing the multimedia database.

For example, supose a user has done an observation where "características de la hoja" (leaf characters) is "con seguridad escamosa" (for sure scale-like) (see Fig. 2) and "resinosa" is "seguro que si" (for sure resiniferous), with only this information the system could conclude that the item observed was a *Cupressaceae* whose CF value is 1. If the user introduce "consistencia de la fructificación" (fruit consistence) is "carnosa" (fleshy) then the system could reach the conclusion "genus is *Juniperus*". If he added to the information "número de semillas de la arcéstida es de dos a cuatro" (number of seed of the berry cone is two to four) and "tipo de arbusto" (type of shrub) is "probablemente postrado "(probably postrated) the system could conclude that the item is a "*Juniperus sabina*" with CF equal to 0.7 (see Fig. 3).

The system could have also reached the same objective with different input information because it's able to provide for each objective. For example, we could conclude the item was a *Cupressaceae* with "consistencia de la fructificación carnosa" (fruit consistence fleshy), "semilla con arilo = no" (seed with fleshy aril=no), "persistencia de la hoja = persistente" (leaf persistence= persistent), "hoja de tono parduzco = no" (brownish leaf=no), "características sexuales=dioica" (sexual character=dioecius) and "semillas numerosas=no" (numerous seeds = no).

Fig. 1: A screen shot for the user interface for introduction of data

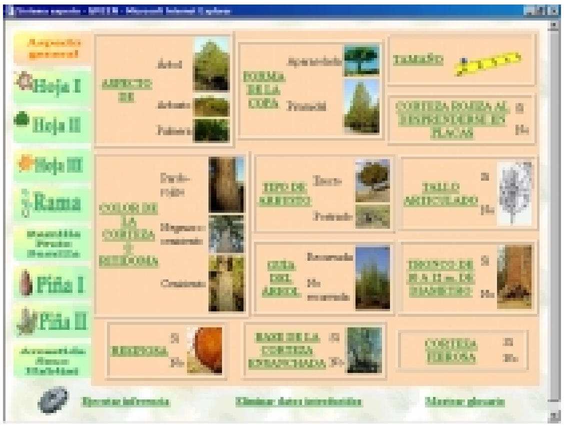


Fig.2. A screen hotspot for the character "characteristics of the leaf"

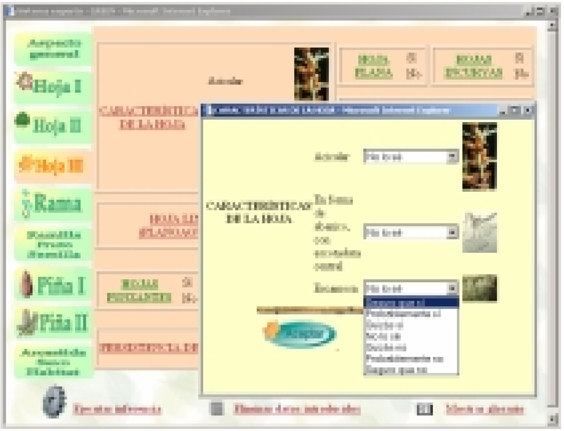
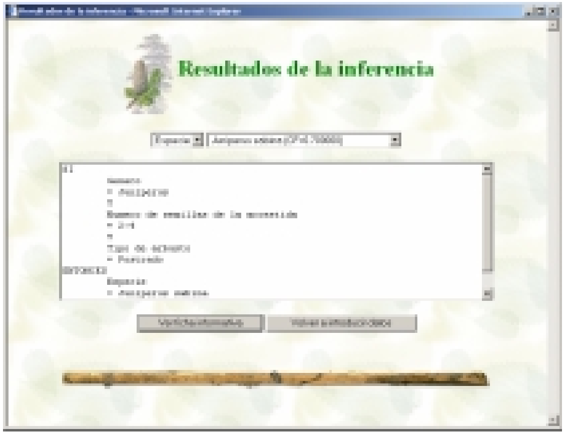


Fig.3: A screen shot for the user interface for identification results



```
SI
        Genero
        = [illegible]
        Y
        Numero de semillas de la [illegible]
        = 2-4
        Y
        Tipo de arbusto
        = Postrado
ENTONCES
        Especie
        = Juniperus sabina
```

11.Conclusions

1. Artificial Intelligence and Internet technology offer new advantages to the popularization of Botany, including the production of automatic keys or computer-generated keys, which will make it possible for non-experts to identify plants.

2. 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..

3. The G.R.E.E.N. System is a practical tool which may be used online and which will enable different taxa comprising the Iberian gymnosperm flora to be recognized.

Bibliography

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

[2] Blanca, G. & Morales, C. 1991. *Flora del Parque Natural de la Sierra de Baza*. Ed. Servicio de Publicaciones de la Universidad de Granada, Granada.

[3] Calvo-Flores, M. D. 1996. *La Inteligencia Artificial. Lección Inaugural. Apertura del Curso Académico*. Ed. Servicio de Publicaciones de la Universidad de Granada, Granada.

[4] 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 Lycopodiaceae-Papaveraceae, ed. Real Jardín Botánico, Madrid.

[5] Ceballos, L., L. Fernández de Córdoba & Ruiz de la Torre, J. 1971. *Árboles y arbustos de la España Peninsular.* Madrid.

[6] Dallwitz, M. J. 1974. A flexible computer program for generating identification keys. *Syst. Zool.* 23, 50–7.

[7] Dallwitz, M. J. 1980. A general system for coding taxonomic descriptions. *Taxon* 29, 41–6.

[8] Dallwitz, M. J. 1992. A comparison of matrix-based taxonomic identification systems with rule-based systems. *Proceedings of IFAC workshop on expert systems in agriculture.*

[9] 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/*

[10]Dallwitz, M. J., Paine, T. A. & Zurcher, E. J. 1995. User's guide to Intkey: a program for interactive identification and information retrieval. 1st edition. *http://biodiversity.uno.edu/delta/*

[11] Dallwitz, M. J. 2000a. Interactive identification using the Internet. *http://biodiversity .uno.edu/delta*

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

[13]Dallwitz, M. J., Paine, T. A. & Zurcher, E. J. 2000c. Principles of interactive keys. *http://biodiversity.uno.edu/delta/*

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

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

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

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

[18]Gonzalez, A. J. & Dankel, D.D. 1993. *The Engineering of Knowledge-Based Systems. Theory and Practice*. Ed. Prentice-Hall International, Englewood Cliffs, N. J.

[19]Grosser D. And 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).*

[20]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*

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

[22]Hopgood A.A. 2001. *Intelligent Systems for Engineers and Scientists*. Ed CRC press

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

[24]Jones, D.L. 1993. *Cycads of the world*. Ed. Red New Holland, Australia.

[25]Krüssmann, G. 1972. *Manual of cultivated conifers*. Ed. Timber Press, Portland.

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

[27]López González, G. (1982). *La Guía de Incafo de los árboles y arbustos de la Península Ibérica*. Ed. INCAFO, Madrid.

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

[29]Luger, G. F. & Stubblefield, W. A. 1993. *Artificial Intelligence. Structures and strategies for complex problem solving*. Ed. The Benjamin/Cummings Series in Artificial Intelligence, Redwood City.

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

[31]Molero Mesa, J., Pérez Raya, F. & Valle Tendero, F. 1992. *Parque Natural de Sierra Nevada*. Ed. Rueda, Madrid.

[32]Nilsson, N. J. 2001. *Inteligencia Artificial. Una Nueva Síntesis*. Ed. Mc Graw Hill, Madrid.

[33]Royal Botanic Gardens, Kew. 1999. *Introduction to Plant Identification.* Notes of International Diploma Course in Herbarium Techniques' 1999, pp.: 1-32. Ed. Royal Botanic Gardens, Kew.

[34]Russell, S. & Norvig, P. 1996. *Inteligencia Artificial, un Enfoque Moderno*. Ed. Prentice International, Mexico.

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

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