PERGAMON

Expert Systems with Applications 25 (2003) 425–430

Expert Systems with Applications

www.elsevier.com/locate/eswa

An application of expert systems to botanical taxonomy

W. Fajardo Contreras[a], E. Gibaja Galindo[a,*], A. Bailón Morillas[b], P. Moral Lorenzo[a]

[a]*Universidad de Granada, E.T.S. de Ingeniería Informática, Departamento de Ciencias de la Computación e Inteligencia Artificial, C/Periodista Daniel Saucedo Aranda, 18071 Granada, Spain.*

[b]*Universidad de Almería, Escuela Politécnica Superior, Departamento de Lenguajes y Computación, Carretera Sacramento s/n, 04120 La Cañada de San Urbano, Almería, Spain.*

Abstract

The implementation of intelligent systems is not particularly widespread in the field of Botany and even less so on **Internet**. At present, we can currently only find hypertext documents or databases which store unprocessed information. The GREEN (*Gymnosperms Remote Expert Executed Over Networks*) System is presented as the application of Artificial Intelligence techniques to the problem of botanical identification. GREEN is an Expert System for the identification of Iberian Gymnosperms which allows online queries to be made. It can be consulted in: http://drimys.ugr.es/experto/index.html

Keywords: Gymnosperms; Identification keys; Expert Systems; Artificial Intelligence; World Wide Web; Iberian Peninsula

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. In fact, the systems which exist are basically databases which store files on the specimens. Artificial Intelligence can offer a more productive approach to these systems by processing the information they contain in order to obtain knowledge which has not been stored explicitly in the database.

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 GREEN (*Gymnosperms Remote Expert Executed Over Networks*), a pioneering system in the application of Artificial Intelligence techniques to the field of botany. GREEN is an online decision aid system, resulting in a much greater and faster diffusion of knowledge and a broader receptor spectrum.

2. Material and methods

We have divided this study on GREEN into 5 parts:

- A first part (Sections 2.1 and 2.2) 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 (Sections 2.3, 2.4, 2.5, 2.6) 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 2.7) is devoted to the reasoning process which the System uses.

* Corresponding author. Tel.: +34-958240468; fax.: + 34-958243317.
E-mail addresses: gibaja@decsai.ugr.es (E.G. Galindo).

0957-4174/03/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0957-4174(03)00066-6

- A fourth part (Section 2.8) in which we discuss other important features of the System.
- Finally, we finish (Section 3) with conclusions drawn directly from what has been presented in this article and from the bibliography used.

2.1. 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 classify 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* (Luger and Stubblefield, 1993) 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 will explain 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 Section 2.2 we outline the process for the design and implementation of the System, detailing the Artificial Intelligence techniques which have been applied.

2.2. 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 (Castroviejo et al., 1986) 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.

2.3. 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 (Blanca and Morales, 1991, Font Quer 1979, López González 1982, García Rollán, 1983, Krüssmann, 1972). 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 have been compiled (Durkin, 1994), which gather the identifying diagnostic characters for each taxon 'Table 1'. As it is not

Table 1
Decision table

	Arrangement of the 'berry' cones	Color of the 'berry' cone	*Pruinose* 'berry' cone	Size of the 'berry' cone	No. of seeds in the 'berry' cone
Juniperus communis subsp. *communis*	Axillary	Bluish-black	Yes	Between 0.6 and 1 cm	3
Juniperus communis subsp. *hemisphaerica*	Axillary	Bluish-black	Yes	Between 0.6 and 1 cm	3
Juniperus communis subsp. *alpina*	Axillary	Bluish-black	Yes	Between 0.6 and 1 cm	3
Juniperus oxycedrus subsp. *oxycedrus*	Axillary	Brown	No	Between 0.6 and 1 cm	1–3
Juniperus oxycedrus subsp. *badia*	Axillary	Brown	No	More than 1 cm	1–3
(···)	(···)	(···)	(···)	(···)	(···)

advisable for these tables to have many empty cells, they have been filled in since many were not necessary when the taxon were identified using the traditional method.

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.

2.4. Obtaining the Knowledge Base

A set of *rules* (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 (Ignizio, 1991), 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 GREEN perfectly and the user becomes familiar with the reasoning process followed by the human experts who have contributed their knowledge to the System (learning).

2.5. 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, GREEN incorporates a module to deal with uncertainty. Uncertainty is modeled using *certainty factors* (Shortlife & Buchanan, 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.

2.6. Consistency reinforcer

During the development of the Knowledge Base, inconsistencies may arise mainly due to errors during the knowledge acquisition and elicitation stage or during the design or implementation of the technique for automatically obtaining the rules.

Another important note is that GREEN is capable of accommodating uncertainty which is why inconsistencies about the certainty of results cause an additional impact. Consequently, this makes it necessary for GREEN to incorporate a *consistency reinforcer* which systematically analyzes each of the rules in the Knowledge Base in order to be able to *detect possible errors* (Gonzalez and Dankel

1993) which have been introduced during the design process thereby guaranteeing that the Knowledge Base has been correctly designed and implemented.

2.7. 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 Motor, Ignizio's BASELINE with forward chaining 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:

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.

2.8. Other characteristics

As we have already mentioned, GREEN 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

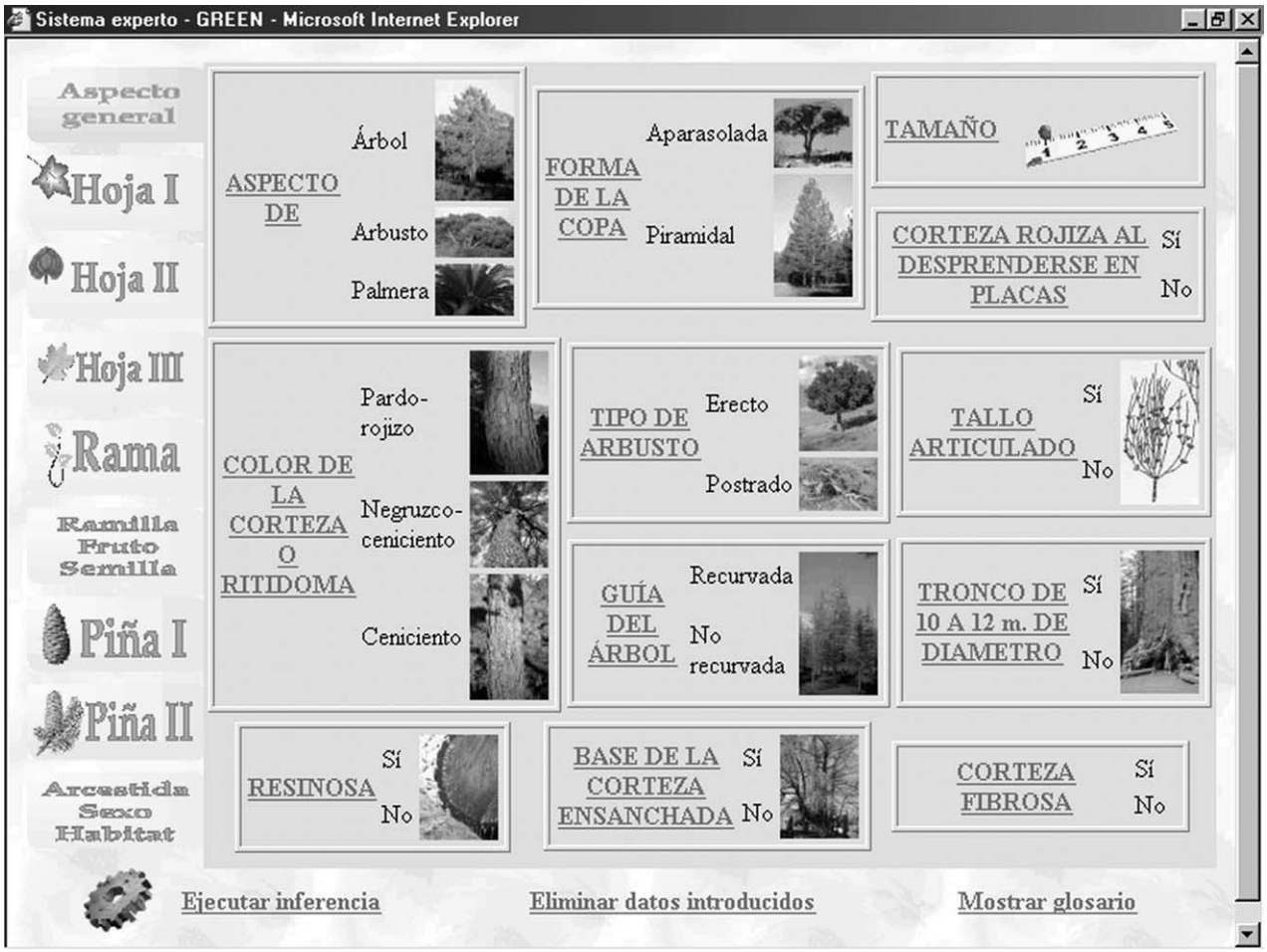

Fig. 1. A screen shot for the user interface for introduction of data. Author: Eva Lucrecia Gibaja Galindo.

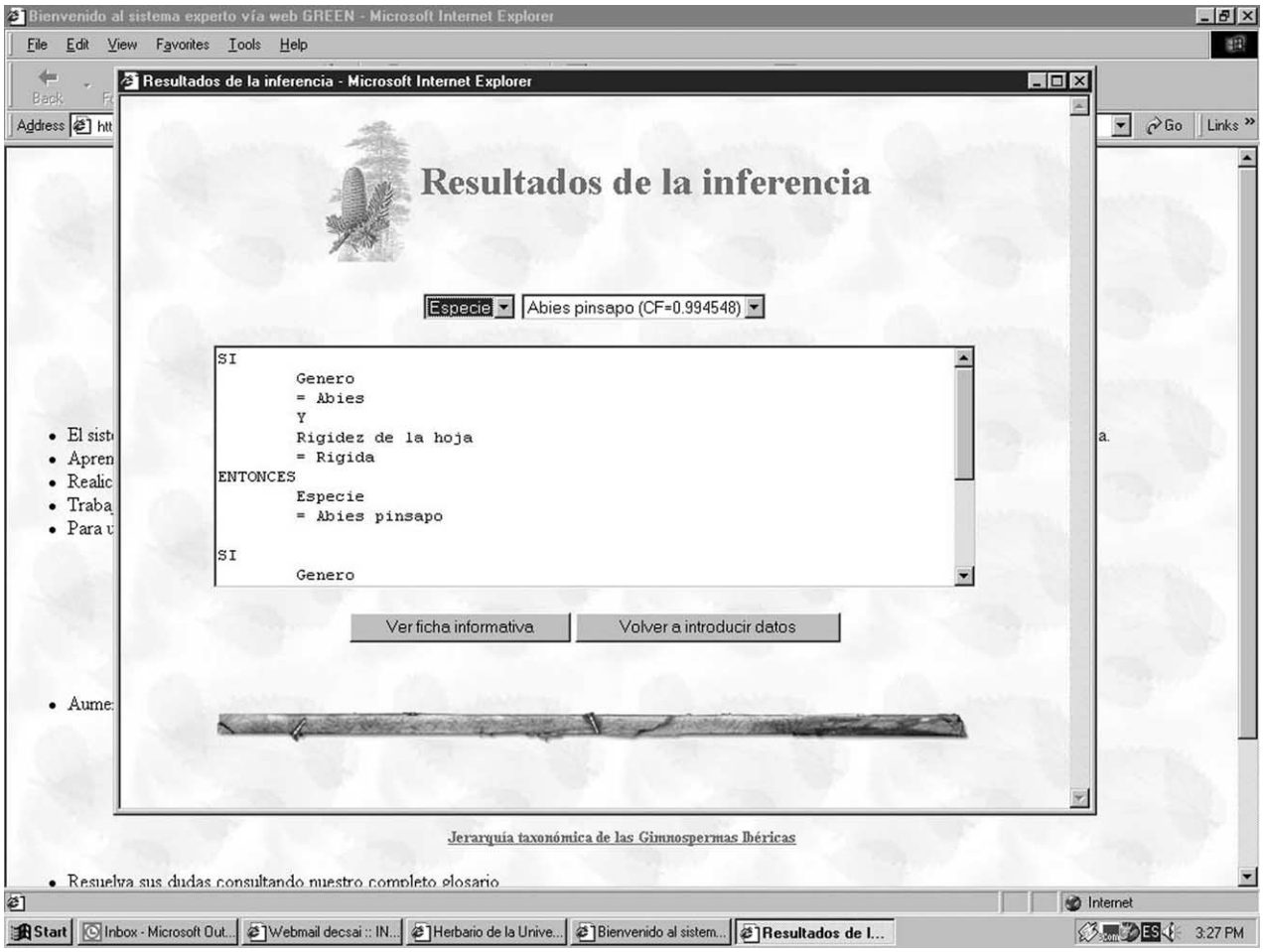

Fig. 2. A screen shot for the user interface for identification results. Author: Eva Lucrecia Gibaja Galindo.

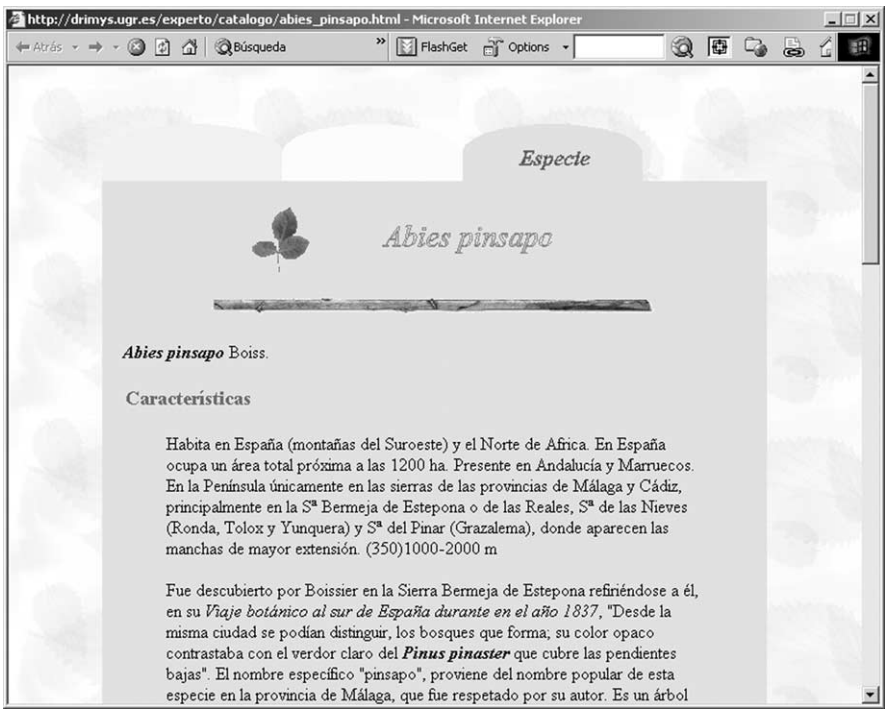

Fig. 3. A screen shot for the user interface for additional information. Author: Eva Lucrecia Gibaja Galindo.

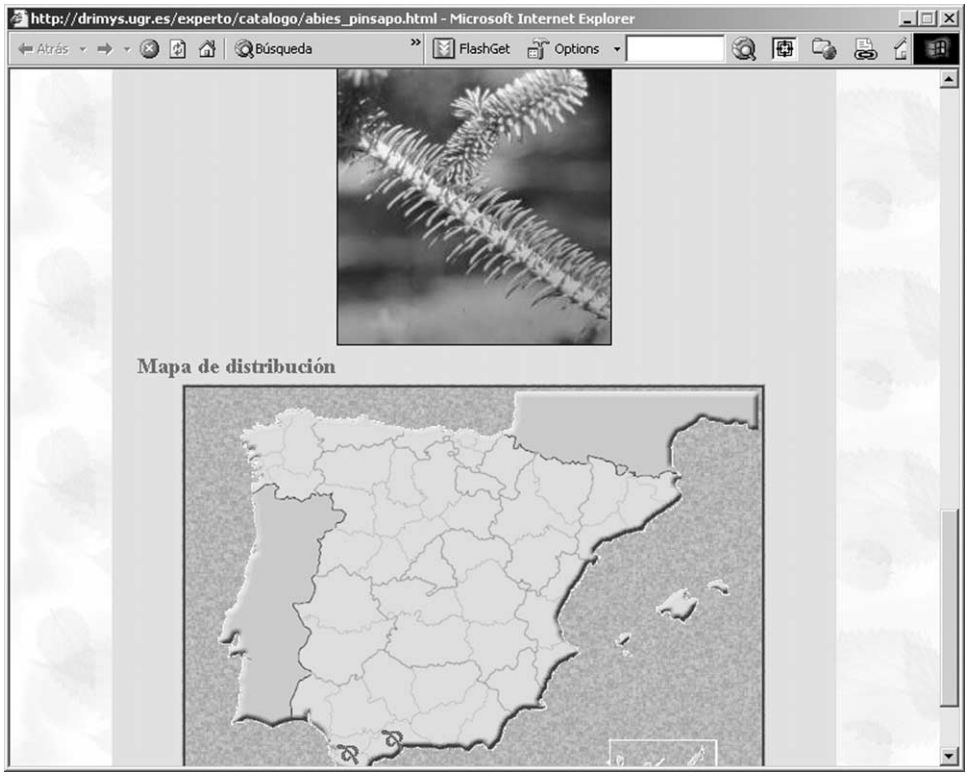

Fig. 4. Other screen shot for the user interface for additional information. Author: Eva Lucrecia Gibaja Galindo.

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 ordered according to how well they fit the query and an outline of the reasoning followed in order to reach these conclusions. If the user wishes, it is possible to increase the information about the specimen by accessing the multimedia database. GREEN 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.

GREEN 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. Figs. 1–4.

3. Conclusions

1. Computing offers 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.
3. The GREEN System is a practical tool which may be used online and which will enable different taxa comprising the Iberian Gymnosperm flora to be recognized.

References

Blanca, G., & Morales, C. (1991). *Flora del Parque Natural de la Sierra de Baza*. Granada: Servicio de Publicaciones de la Universidad de Granada.

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*, Madrid: Real Jardín Botánico.

Durkin, J. (1994). *Expert systems. Design and development*. London: Prentice Hall International.

Font Quer, P. (1979). *Diccionario de Botánica*. Barcelona: Labor.

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

Gonzalez, A. J., & Dankel, D. D. (1993). *The Engineering of knowledge-based systems. Theory and practice*. Englewood Cliffs, NJ: Prentice-Hall International.

Ignizio, J. P. (1991). *Introduction to expert systems. The development and implementation of rule-based expert systems*. New York: McGraw-Hill.

Krüssmann, G. (1972). *Manual of cultivated conifers*. Portland: Timber Press.

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

Luger, G. F., & Stubblefield, W. A. (1993). *Artificial intelligence. Structures and strategies for complex problem solving. The Benjamin/Cummings series in artificial intelligence*, Redwood City: Benjamin/Cummings.

Shortlife, E., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, *23*, 351–379.