Geobotanical Database Exploration

Ireneusz R. Moraczewski[1], Robert Zembowicz[2] and Jan M. Żytkow[3]

[1] Department of Plant Systematics and Geography, Warsaw University,
Al. Ujazdowskie 4, 00-478 Warsaw, Poland, (Fulbright scholar at Wichita State University)

[2] Computer Science Services Group, LLC, 9415 E. Harry, Suite 302, Wichita KS 67207
(also: Department of Computer. Science, Wichita State University)

[3] Department of Computer Science, Wichita State University, Wichita, KS 67260-0083
(also: Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland)

Scientific database exploration

As the scientific databases are growing in size and number, the automation of knowledge discovery becomes a necessity. Human researchers are unwilling to consider huge numbers of hypotheses in order to find regularities hidden in the data. But the effort is worthwhile. Our data mining experience shows that databases reveal large numbers of significant regularities that represent a great variety of knowledge, and the systematic approach of the machine captures abundance of regularities never considered by humans.

Effective automation of discovery must cope with many requirements. One of them is the breadth of search for knowledge. Ideally, we would like to capture all knowledge available in the data. Another concern is efficiency of search. Different forms of knowledge may require search in different hypotheses spaces, but the search should be avoided if data do not fit any hypothesis in a given space or if a simpler regularity can capture the same knowledge. Still another requirement is effective knowledge presentation to the user. Large numbers of regularities discovered in a database may make the user confused and frustrated rather than happy, unless they can be organized into concise theories that preserve the empirical contents of the discovered knowledge.

Each of these requirements can be satisfied only to a degree, but the unbound progress is possible on all of them, as well as on their combination. The design goal of our 49er system (Żytkow & Baker, 1991; Żytkow & Zembowicz 1993) has been the balanced progress along these and other dimensions of discovery. The initial strategy applied by 49er in absence of specific user's expectations is to search for statistically significant 2-dimensional contingency tables in all combinations of variables and in many subsets of data. 2-dimensional contingency tables, treated in many statistics textbooks (Jobson, 1992; Johnson & Bhattacharyya, 1992), offer a homogenous way of expressing knowledge implicit in databases. They can guide the further discovery process because different forms of regularities can be detected as various special instances of contingency tables (Troxel et al. 1994). 49er applies simple tests to contingency tables to distinguish many special forms of knowledge. Two processes follow. First, additional search leads to knowledge refinements. For instance, after a functionality test reveals a functional relationship between two variables in a contingency table, the search in the space of equations determines the form of functional dependency (Żytkow & Zembowicz, 1993). Second, regularities of one type are combined into specialized theories, such as multidimensional equations, causal graphs, taxonomies and inclusion graphs. These theories can capture large numbers of regularities.

In this paper we will focus on 49er's exploration of a geobotanical database. We describe the results of a general purpose search mechanism, followed by tests which detected large numbers of regularities in the form of equivalence and subset relation. Those special regularities are subsequently summarized into more concise theories. In particular, we show how a large number of subset relations can be summarized into an inclusion graph.

Equivalences and implications inferred from contingency tables

Some contingency tables express equivalence or implication. This can be clearly seen in 4-cell (2×2) contingency tables. Such tables are natural for two-valued attributes, like Boolean, but also for many-valued attributes, when values are binned in two groups. 2×2 contingency tables are useful summaries of data in all areas, but they may be the dominant form of regularity in some disciplines, e.g. in geobotany, when the Boolean attributes express presence or absence of particular species in various geographic areas.

Let us denote the cells in a (2×2) contingency table as follows:

	B	$\neg B$
A	a	c
$\neg A$	b	d

$a, b, c, d \geq 0$

	B	$\neg B$	$a, d > 0$
A	a	0	$A = B$
$\neg A$	0	d	or $A \equiv B$

Figure 1: 4-cell contingency table that expresses equality of two sets, or equivalence of two statements.

	B	$\neg B$	$a, b, d > 0$
A	a	0	$A \subset B$
$\neg A$	b	d	or $A \rightarrow B$

Figure 2: 4-cell contingency table that expresses inclusion between a pair of sets, or implication.

where a, b, c, d are numbers of records of each type. For instance, a is the number of records in which two Boolean attributes A and B are both true.

If one diagonal of a 4-cell contingency table consists of zeros then the regularity can be expressed as the equality of two sets or the equivalence of two statements (Fig. 1). A larger number of equivalences can be summarized by a taxonomy (Troxel et al. 1994). Since in geobotany we want to make claims about the ranges of species, rather than their complements, we distinguish between the same distribution pattern ($b = c = 0$) when the ranges are equal, and complementary distribution pattern ($a = d = 0$), when ranges are disjoint and exhaustive.

Contingency tables in which exactly one cell is empty express inclusion or implication (Fig. 2). Four cases of inclusion are possible. When $c = 0$, the contingency table expresses inclusion $A \subset B$ or implication $A \rightarrow B$. Analogously, if $b = 0$, then $B \subset A$ ($B \rightarrow A$). If A and B denote the ranges of two species, $B \subset A$ means that the range of species B is included into the range of A. If $a = 0$, the contingency table expresses inclusion of A in the complement of B: $A \subset \neg B$ (and inclusion of B in the complement of A: $B \subset \neg A$). When we want to make claims about the ranges of species, not their complements, when $a = 0$ we speak about the exclusion of two ranges, which is a symmetrical relation. If $d = 0$, $\neg A \subset B$, $\neg B \subset A$ the species overlap, and their union covers the whole space. This is also a symmetrical relation between ranges, which can be called exhaustive overlap.

The above distinctions between different subset and equivalence relations make sense for Boolean attributes when the two values mean existence and non-existence of an object. This is the case in the geobotanical data we discuss in this article. These distinctions, however, do not apply to other 2-valued attributes, when the values are just two different values that an object may have.

In practical applications we must admit some exceptions. We assume, for example, that A is included in B if $c/(a + c) < 0.1$. Analogous rules apply in the remaining cases.

Case study: flora of Warsaw

We will now describe one case study: the automated exploration of a geobotanical database, with the focus on equivalence and subset relations between species. The data are the result of 10-year investigations of Warsaw flora carried out by Sudnik-Wójcikowska (1987). The city was divided into 225 squares of 1.5 kilometer side. The inventory of vascular plants growing at each square has been made. 1181 higher plant taxa (species and subspecies) were found within the limits of 225 squares. The presence and absence of each taxon has been represented by a separate binary attribute, leading to the binary data matrix of the size 1181×225. Each value of 1 indicates the presence, while 0 indicates the absence of a particular taxon in a particular square. In addition, each square has been described by four abiotic factors: anthropopressure (reflecting the intensity of man's impact on the natural conditions), air pollution, building density (influencing the amount of space available to plants) and potential for vegetation (reflecting mainly the soil structure, fertility and moisture).

According to the research paradigms accepted in floristics, the ultimate goals of the analysis are (1) discovery of relationships between habitat requirements of different species and (2) explanation of plant distribution patterns in terms of environmental, geographical and historical factors. In this paper we will limit our attention to the regularities for 1181 variables that represent species, thus our results contribute to the first goal.

Exploration of `Warsaw_flora` database

We used 49er to explore the `Warsaw_flora` database. For the 1181 attributes that represent species, 49er considered all pairs of attributes. The total number of hypotheses is $0.5 \times 1181 \times 1180 \approx 7 \times 10^5$. There is always a chance that some randomly created patterns look like real relationships. Statistical measure of significance estimates the probability that a given patterns has been created randomly. To minimize the number of pseudo-regularities generated by random fluctuation we set the threshold of acceptance to 10^{-5}, because in random data this would lead to approximately 7 spurious regularities for the number of hypotheses considered during the search. On the flip side we risk ignoring regularities which are real but for which the probability of random fluctuation is greater than 10^{-5}.

A typical database exploration reveals many interesting regularities in data subsets. In the `Warsaw_flora`, however, even the simplest slicing method which generates two complementary subsets for each attribute leads to 2×1181 data subsets and to $0.5 \times 1180 \times 1179 \times 2 \times 1181 \approx 1.6 \times 10^9$ additional hypotheses. The task would occupy 49er for some 3 years.

After considering all pairs of 1181 attributes, 49er returned 16577 statistically significant contingency tables. 594 of them capture equivalence of attributes, 6364 different subset relations, while the remaining 9619 represent statistically significant overlaps of species.

From equivalence to taxonomy

Equivalence captures the same distribution pattern for two plant taxa. Equivalent species behave in the same manner in a given territory. 49er uses equivalences to build a taxonomy (Troxel et al. 1994). In the first step all species equivalent with each other have been grouped together. This led to interesting discoveries. A 29-element class of species has been discovered, all species cohabitating a single square and absent everywhere else. This class of species consists of ephemerophytes, that is alien species not established permanently in the flora. It turns out that they occur in the vicinity of a mill. Another 9-element class contains species being the fugitives from gardens (= ergasiophygophytes). The locations of both classes are depicted in Fig. 3, together with several 6-species classes. Since all equivalences led only to one- or two-element sets, no interesting taxonomy has been formed.

6364 contingency tables obtained by 49er represent approximate inclusions of all four types. This number has been reduced to 2995 after merging equivalent nodes. 2878 contingency tables express "positive" inclusion ($A \subset B$) while 117 can be treated as "exclusions of ranges" ($A \subset \neg B$) and none by "exhaustive overlap" ($\neg A \subset B$). The dramatic difference in the number of both types of regularities can be ascribed to low significance of many relations between small subsets, as the majority of species are distributed over small numbers of squares.

Inclusion graph

A large number of inclusions can be conveniently combined into an inclusion graph I, whose vertices represent ranges of attributes, and edges represent inclusions between them. Formally, each directed edge $\langle A, B\rangle$ between nodes A and B in graph I represents the inclusion $A \subset B$. Examining and interpreting inclusion graphs is the topic of this section.

Although each subset link has a clear interpretation, the I graph for `Warsaw_flora` is very large, difficult to visualize and comprehend. Further analysis of the graph attempts to decompose I into meaningful parts, each having clear, domain specific interpretation. The I graph for `Warsaw_flora` splits into four coherent subgraphs I_1, I_2, I_3 and I_4, consisting of 1, 1, 2 and 2874 edges, respectively. Each coherent subgraph may still contain some redundant edges resulting from transitivity of inclusion. For instance, in a graph $\{\langle A, B\rangle, \langle B, C\rangle, \langle A, C\rangle\}$, $\langle A, C\rangle$ is a redundant edge. Elimination of those edges simplifies the graph's analysis while preserving its empirical content.

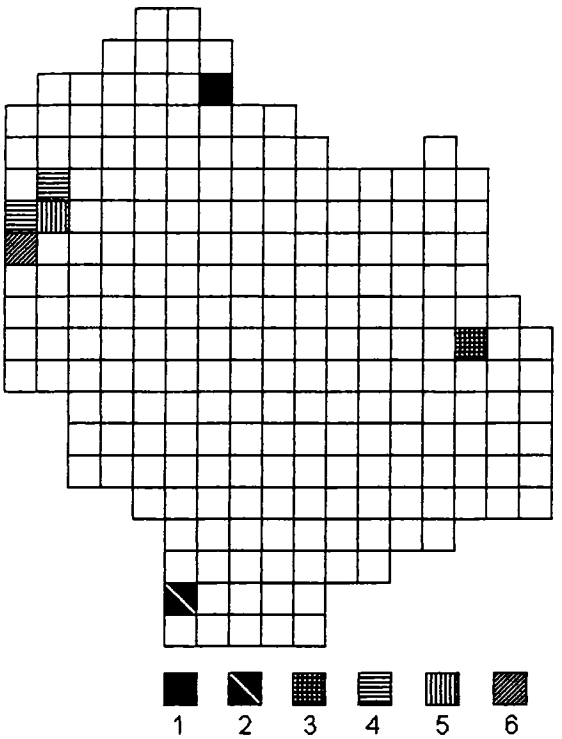

Figure 3: Distribution of the areas covered by the largest numbers of equivalence species: 1: Glaucium corniculatum class (29 species), 2: Callistephus chinensis class (9 species), 3: Thalictrum minus class (6), 4: Trollius europaeus class (6), 5: Chenopodium ficifolium class (6), 6: Borago officinalis class (6).

Let I_4^r stand for the graph I_4 after the redundant edges have been eliminated. I_4^r contains 2354 edges. Of particular interest are maximum elements, that is those species in I, which are not subsets of any other species. Each maximum element represents one of the maximally spread, that is the least selective species. There are 302 maximum elements in I_4^r. For each maximum element M, a simple algorithm can select a graph G_M of all subsets of M in I_4^r. For each M, G_M represents all species which have their habitat requirements more selective than M. The further away from M in I_4^r, the narrower is the ecological amplitude of the species.

Some species in G_M can be also subsets of other maximum elements. Therefore in the next step of the interpretation process, G_M is constrained into the graph G_M^* which contains all and only those edges in G_M that do not belong to any other maximum subgraph. In geobotanical terms, the species in G_M^* are characteristic to a specific environment determined by species M.

The following algorithm summarizes the above procedure:

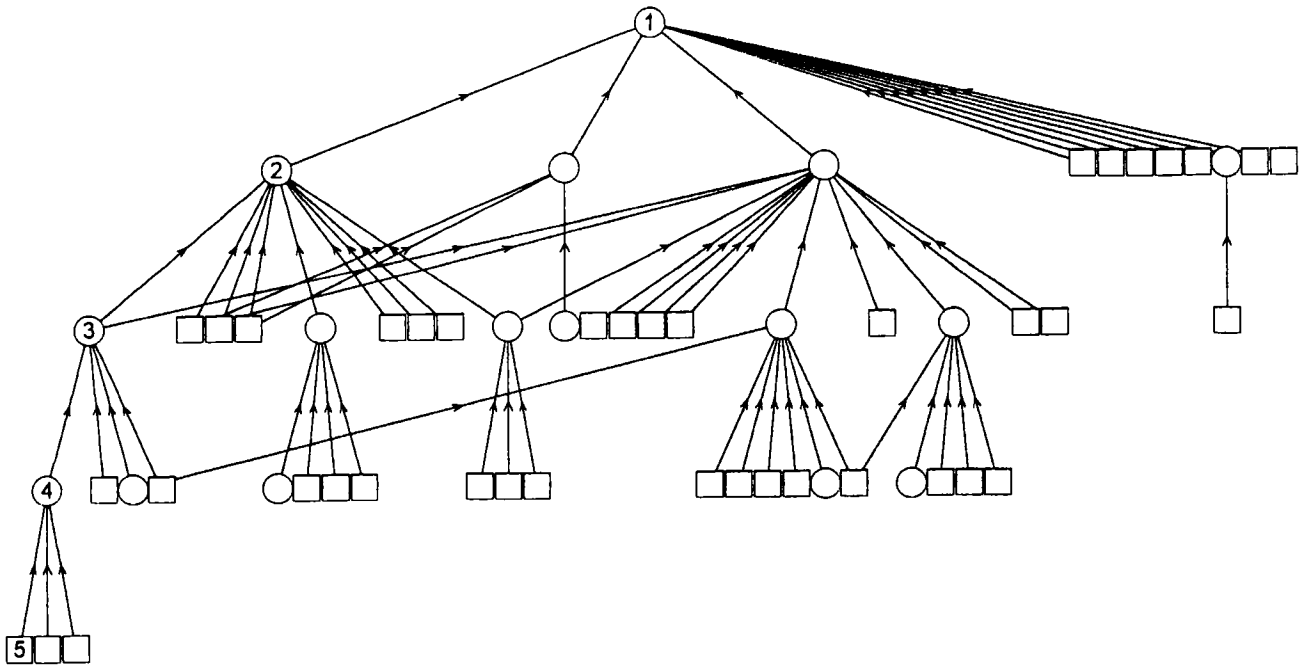

Figure 4: One of the largest G^*_M graphs found in the Warsaw_flora database. Nodes are species, directed edges correspond to inclusions between them. Circles denote attributes not connected with other G^*_M graphs, squares stand for nodes connected to at least one other graph. 1: *Ranunculus acris*, 2: *Myosotis palustris*, 3: *Linum catharticum*, 4: *Parnassia palustris*, 5: *Calla palustris*.

Algorithm: Build & analyze the inclusion graph
Given all significant contingency tables
Select proper inclusions and combine them in graph I
Remove transitive edges
Determine all coherent subgraphs of I
For each maximum node M in I
  Let G_M be a graph of all subsets of M in I
  Let G^*_M result from G_M by removing all edges
    that belong to any other graph $G_{M'}$

Fig. 4 portrays one of the largest G^*_M graphs found in the `Warsaw_flora` database. Intergraph links, which by the definition of G^*_M can only occur at the leaves, are indicated by square nodes. The graph can be easily interpreted in ecological terms. All vertices of this graph are species preferring moist areas. Directed edges lead from species of special habitat requirements, towards more commonly found taxa of broader ecological amplitude. Consider, for example, the path marked 1–5 in Fig. 4 which is also presented as a map in Fig. 5. The largest subset of squares corresponds to *Ranunculus acris* (# 1), a common meadow plant, growing on wet or fresh soils, that is missing only in the downtown and several dryer squares. *Myosotis palustris* (# 2) is also a quite common species connected with wet habitats and clearly avoiding the center of the city. The next three species have much narrower range and are homophobic. The last of the three is *Calla palustris* (# 5), growing only in peat-bogs, that is in very wet, natural habitats.

Comparison with previous work

`Warsaw_flora` database has been extensively studied. The relationships between species distributions, however, has not been addressed, mainly because of the huge computational cost of such analysis. The approach presented above is the first attempt at revealing the regularities between distribution patterns. Some of the regularities found have been known to the expert of the field, e.g. 29-element equivalency class (Sudnik-Wójcikowska, personal communication), but the vast majority is new and noteworthy.

The relations between species distribution and abiotic factors received plenty of attention. Sudnik-Wójcikowska (1986) has shown some interesting relationships between distribution patterns of selected vascular plants and anthropopressure zones. The monograph on flora of Warsaw (Sudnik-Wójcikowska 1987) provided, apart from detailed information about higher plant taxa distribution, a comprehensive analysis of flora synanthropization processes. As the estimation

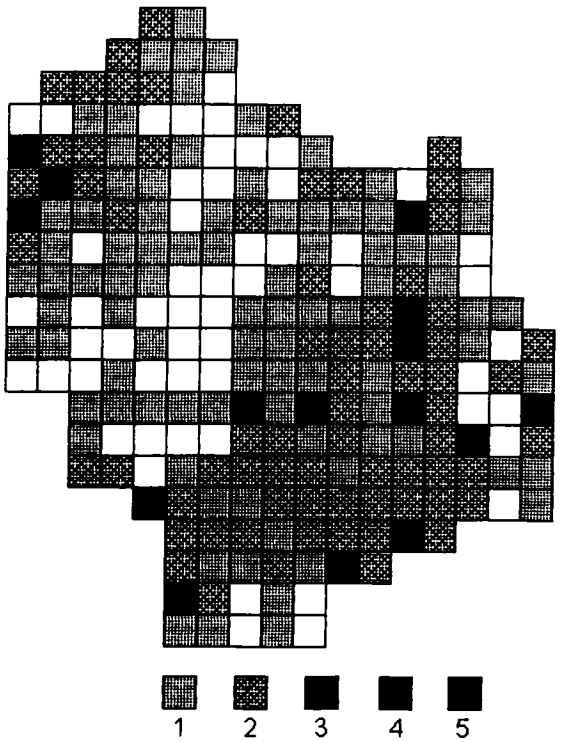

Figure 5: Path that is marked in Fig. 4 is here displayed as a map. 1: *Ranunculus acris*, 2: *Myosotis palustris*, 3: *Linum catharticum*, 4: *Parnassia palustris*, 5: *Calla palustris*.

of synanthropization advancement is of great practical importance, Sudnik-Wójcikowska (1988, 1991) introduced several floristic indices reflecting the intensity of anthropopressure. Sudnik-Wójcikowska and Moraczewski (1993) applied clustering techniques to compare the 1181-dimensional "species space" with four sets of floristic parameters with regard to their ability to reflect the human impact.

The above-mentioned studies were done in intuitive way and involved some extra expert-derived knowledge. 49er has also explored systematically the relations between species distribution and abiotic factors, discovering plenty of regularities (877 regularities at the significance level 10^{-3}). Although 49er has been confined to the original, raw data, yet it obtained many new and noteworthy results. The comparison of these results with those discovered by "manual" analysis will be presented in a separate publication.

Conclusions and future work

Large-scale automated search in a geobotany database revealed very large number of regularities. In this paper we concentrated on regularities that express subset relations. We have shown that a very large number of such relations can be represented as the inclusion graph, without losing the empirical contents of the theory. The inclusion graph can be decomposed into parts, which possess distinct geobotanical interpretation. Graphs G_M and G^*_M reveal important groups of species and their relations to other groups.

Many regularities between species detected by 49er cannot be captured neither by the taxonomy nor by the inclusion graph. Both 117 significant exclusions and 9619 significant overlaps can be interpreted in terms of distances between species in the space of environmental factors. The attempt to reconstruct such a space from data on the distribution of species will be a subject of another paper.

References

Jobson, J.D. 1992. *Applied Multivariate Data Analysis, Vol.II*, Springer-Verlag.

Johnson, R.A. & Bhattacharyya, G.K. 1992. *Statistics; Principles and Methods*, John Wiley & Sons.

Sudnik-Wójcikowska, B. 1986. Distribution of some vascular plants and anthropopressure zones in Warsaw. *Acta Societatis Botanicorum Poloniae, 55*, No 3: 481–496.

Sudnik-Wójcikowska, B. 1987. *The flora of Warsaw and its changes in 19th and 20th Centuries* (in Polish). Wyd. Uniw. Warszawskiego, Warszawa, 242+435 pp.

Sudnik-Wójcikowska, B. 1988. Flora synanthropization and anthropopressure zones in a large urban agglomeration (exemplified by Warsaw). *Flora 180:* 259–265.

Sudnik-Wójcikowska, B. 1991. Synanthropization indices of urban floras — an attempt at definition and assessment. *Acta Societatis Botanicorum Poloniae, 60*, No 1–2: 163–185.

Sudnik-Wójcikowska, B. and Moraczewski, I.R. 1993. Floristic evaluation of anthropopressure zones in Warsaw. *Feddes Repertorium 104:* 81–92.

Troxel, M., Swarm, K., Zembowicz, R. & Żytkow, J.M. 1994. Concept Hierarchies: a Restricted Form of Knowledge Derived From Regularities, in Z.W. Raś, M. Zemankova eds. *Methodologies for Intelligent Systems (Proc. of ISMIS'94 Symposium)*, Springer Verlag, 437–447.

Żytkow, J.M. & Baker, J. 1991. Interactive Mining of Regularities in Databases, in G. Piatetsky-Shapiro & W. Frawley eds. *Knowledge Discovery in Databses*, AAAI Press.

Żytkow, J.M. & Zembowicz, R. 1993. Database Exploration in Search of Regularities, *Journal of Intelligent Information Systems, 2*: 39–81.