December 2003

Automating Ontology Generation for Information Systems Research Using GHSOM

Mohammad Al-Ahmadi
Oklahoma State University

Ramesh Sharda
Oklahoma State University

AUTOMATING ONTOLOGY GENERATION FOR INFORMATION SYSTEMS RESEARCH USING GHSOM

Mohammad Al-Ahmadi
College of Business Administration
Oklahoma State University
mohamas@okstate.edu

Ramesh Sharda
College of Business Administration
Oklahoma State University
sharda@okstate.edu

Abstract

Building ontology for a specific field of research is a very tedious task; yet, very important. Ontologies can help in defining the boundaries of a discipline and identifying new emerging streams of research. Automating this process reduces, if not eliminates, the overhead associated with manual ontology building methods and gives a big jump to continue refining and improving the generated ontology. Growing Hierarchical Self-Organizing Map (GHSOM) is a promising unsupervised artificial neural networks architecture that can help in identifying hierarchical relations embedded into datasets. Our project-in-progress is exploring the use of GHSOM to generate ontology for the Information Systems (IS) published research.

Keywords: Information systems, IS research, research, ontology, taxonomy, neural networks, SOM, GHSOM, growing hierarchical self-organizing map, decision support systems, DSS

Introduction

Chandrasekaran et al. (1999) classify ontologies as content theories about objects (and their properties) in a specific domain of knowledge, and any relationship that may exist between these objects. Ontology, in one of the two meanings they provide, refers to "a body of knowledge describing some domain". Gruber (1993) defines ontology as "an explicit specification of a conceptualization". A conceptualization is a simple abstracted view of the world we would like to model. Welty and Guarino (2001) also state that ontology means conceptual models in the general industrial context. Taxonomies are located at the center of most conceptual models. We use ontology and taxonomy interchangeably.

The purpose of this paper is to explore the use of hierarchical clustering techniques to build taxonomies semi-automatically. Specifically, we adapt neural-network based algorithm Growing Hierarchical Self-Organizing Map GHSOM (Dittenbach, et al., 2002) for developing a knowledge map for all the published papers for a journal.

In this paper, section 2 identifies the benefits, in general, of generating ontology for any published research field (or part of it). Section 3 discusses the main tasks in the ontology field and how GHSOM can facilitate some of these tasks. Section 4 introduces GHSOM algorithm itself. In section 5, we present the systematic steps that should be followed to generate ontologies for information system research using GHSOM. Section 6 briefly describes our preliminarily experiment to use GHSOM to generate ontology for the published research in *Decision Support Systems* journal.

Benefits of Generating Ontology for a Research Field

Ontology for a research field facilities field visualization from different perspectives and through the different lenses of main players in the research field. In this section, we present a model that represents the accumulated knowledge in a published research field at different levels. Although the focus of our project is information systems research, the model (and the discussion in this section) is presented in a general format to demonstrate the universality of the model.

Scientific papers appearing in scientific peer-reviewed journals in a research field represent the basic building block for this model. Each journal can be characterized by the main field it belongs to, the subfields that it usually covers, and policies that its editorial board sets. Published research of a researcher presents another interesting dimension. Figure 1 shows an abstracted model, that we call the Accumulated Research Model (ARM) for Ontology Generation, for the published research in a domain. ARM is intended to simplify the discussion of the benefits we can get by building ontology for each level appearing in the model. We don't claim that the model is complete although it represents the most interesting levels, in our opinion, for ontology generation. For example, ARM doesn't consider a journal issue as a separate level by itself because it doesn't usually have enough information to be represented by ontology. Moreover, this is consistent with the emerging trend of publishing accepted papers online without waiting for them to appear in a journal issue.

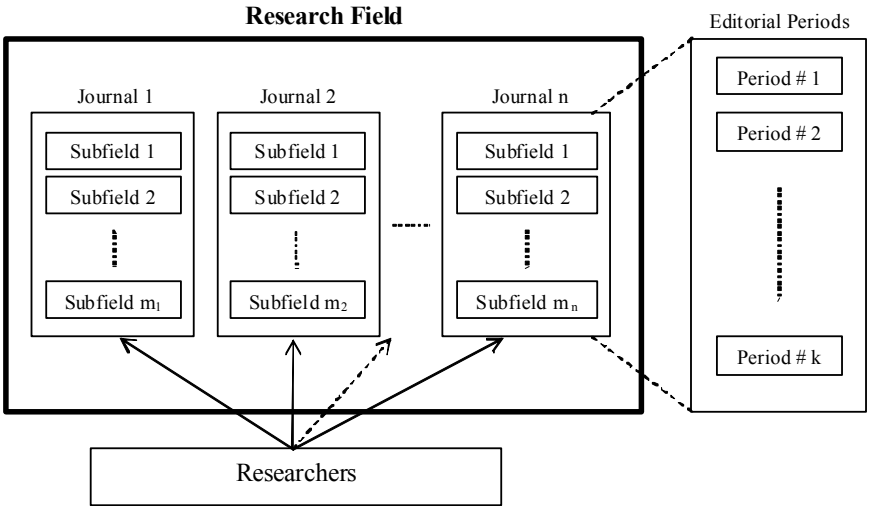

Figure 1. Accumulated Research Model (ARM) for Ontology Generation

Five levels appear in ARM for possible ontology generation: field, subfields, journals, editorial periods, and researchers. We can study the benefits of building these ontologies through the lens of the three main players in any research field: researchers, editors and reviewers. These roles are not usually played exclusively, resulting in an overlap in the gained benefits. Some examples of the gained benefits are presented in Table 1. On the other hand, there are other roles, such as research grants-providers, which are not included for brevity. Moreover, the presented examples are independent of the methods used to generate ontologies.

Applications of GHSOM in Ontology Field

Most ontologoies have been built manually (Ding and Foo, 2002b; Sugumaran and Storey, 2002). "Further research is encouraged to find appropriate and efficient ways to detect or identify the relations either semi-automatically or automatically",(Ding and Foo, 2002b). Newly emerging algorithms from neural networks foundations, such as Growing Hierarchical Self-Organizing Map (GHSOM) which will be introduced in the next section, offer capabilities to achieve this. The unsupervised fashion of GHSOM makes it ideal to fill part of the need for tools to automate ontology generation. GHSOM can be employed as a semi-automated ontology generation methodology that is able to model taxonomical "is-a" relationships. "Is-a" relationships are naturally represented in a hierarchy.

From another perspective, there are three main tasks related to ontology: *generation*, *mapping*, and *evolving* (Ding and Foo, 2002a; Ding and Foo, 2002b). GHSOM (and its hierarchical structure) has associations, at least partially, with each task. Ding and Foo (2002b) discuss five ontology generation projects in details. Two of them have used hierarchical structures to represent generated ontologies. GHSOM can be used to automate creating such hierarchies. Many ontology mapping approaches have been proposed. Visser and Tamma (1999) have developed ontology clustering approach, which clusters similar resources and organizes them hierarchically (Ding and Foo, 2002a). Moreover, recent research has been performed to automate or semi-automate ontology evolving (Ding and Foo, 2002a). However, such research has received limited success. The ability of GHSOM to resume its learning process with new specifications or content is well suited for automating ontology maintenance (evolving).

Table 1. Examples of Benefits of Generating Ontologies for a Research Field

		Gained Benefits through the Lens of		
		Researchers	**Editors**	**Reviewers**
Generating Ontology for	**Field**	Understanding how a field has evolved Discovering any emerging trends Identifying possible ways to combine research from different subfields	Understanding how a field has evolved	Understanding how a field has evolved
	Subfield	Helping writing literature reviews Helping building conceptual models Identifying gaps for further research	Encouraging research in emerging subfields by issuing special issues	Helping evaluating literature reviews Helping evaluating conceptual models
	Journal	Targeting appropriate journals for research publication Identifying relevant journals as main sources of a subject for research	Understanding the focus of a journal Enriching editors with ideas from different journals	Identifying which kind of research fits a journal in general
	Editorial Periods	Observing any change in editorial policies Finding the most suitable place to publish a type of research	Understanding past editorial policies Helping setting new editorial strategies and directions	Identifying which kind of research fits a journal according to its editors' views
	Researchers	Identifying their research streams and contributions Opening doors for more chances to cooperate	Identifying scholars that can be invited to write in a special issue Assigning reviewers according to their specialties and contributions	Identifying scholars that they or their research might be consulted to facilitate a reviewing process

Growing Hierarchical Self-Organizing Map (GHSOM)

Kohonen self-organizing features map (abbreviated as either SOM or SOFM) is an unsupervised artificial neural network architecture that has been widely adopted because of its ability to map high-dimensional datasets into 2-diminstions maps and to cluster similar documents in neighbor regions on the maps (Dittenbach, et al., 2002; Kohonen, 1982). Dittenbach et al. (2002) have proposed a Growing Hierarchical Self-Organizing Map (GHSOM) algorithm, which is a promising extension to SOM. GHSOM overcomes two limitations of SOM algorithm. First, SOM map has a fixed size of units on the map and there is always a need to determine the arrangement of these units before any training (learning process) can take place. Both tasks, finding the appropriate number of units and their arrangement, are difficult and based on prior knowledge about the data. Second, it is believed that the 2-dimension space representation in SOM oversimplifies any complex relationships that might exist in datasets. For example, SOM doesn't reflect hierarchical structures usually exist in many document archives. (Dittenbach, et al., 2002).
The hierarchical structure of GHSOM has multiple layers. Each layer has a number of independent SOMs. Each SOM, except the one in the first layer, is a natural expansion for a unit on one of SOMs in its parent layer. In addition, GHSOM has the flexibility to control the breadth and depth of the required hierarchical map by specifying two parameters (thresholds). GHSOM is a dynamic algorithm that can expand horizontally and vertically during the learning process to reflect the structure of the data. The two parameters work as stopping rules for the learning process. GHSOM usually produces unbalanced trees that are more representative for the actual structure embedded in the data than SOM. Another interesting feature, GHSOM learning process can be resumed if there is a change in the requirements such as the need for a deeper structure. A detailed description of GHSOM algorithm is beyond the scope of this paper; interested readers can refer to Dittenbach et al. (2002). We provide a diagrammatic view of GHSOM in Figure 2.

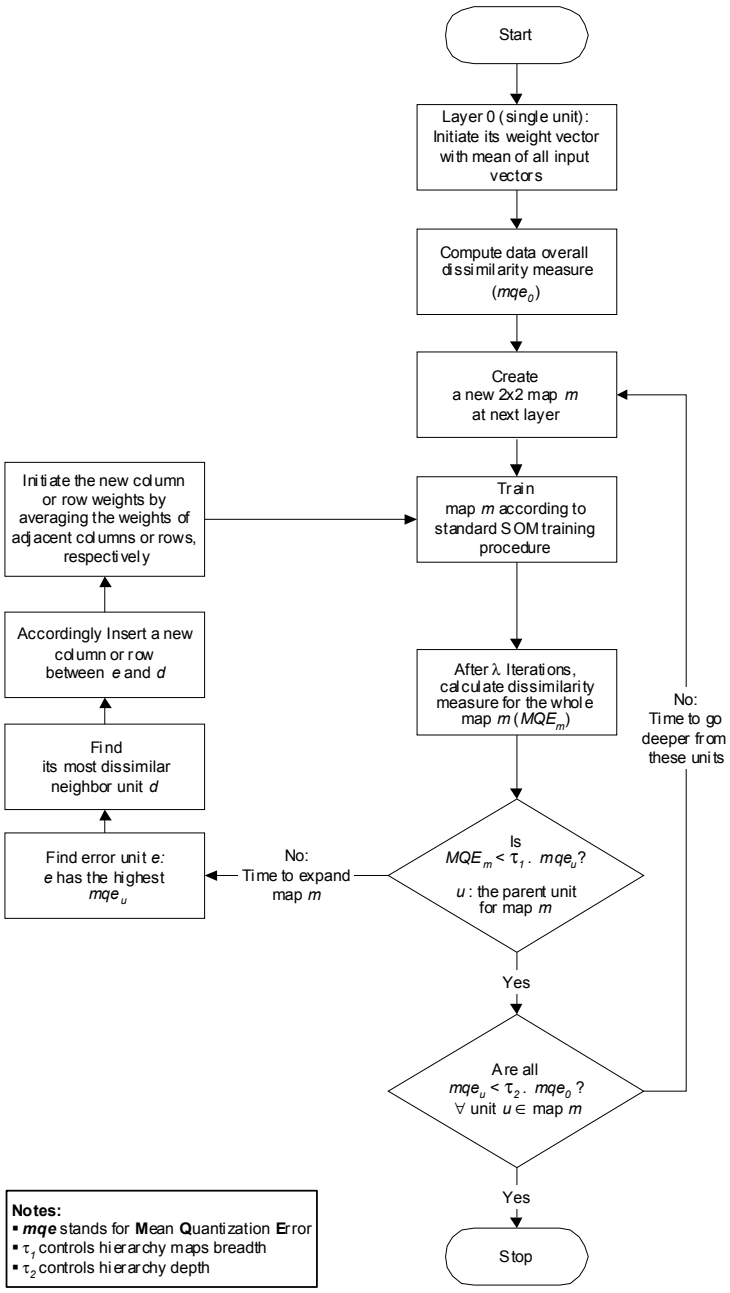

Figure 2. Diagrammatic View of GHSOM (Dittenbach et al., 2002)

Information System Research Ontology Automation Project

Generating ontology automatically using GHSOM consists of seven steps. Collecting data is the first step in the ontology generating process. In our case, we collect data, such as keywords, abstracts and titles, related to the published papers in the domain of interest (Information Systems field, subfields, journals, etc.). Second step is to extract the concepts (keywords and unique terms) after eliminating stopping words (e.g. and, or, etc.). Third, these concepts are used to design the Features Vector Structure (FVS). In addition, we ensure that different terms referring to the same concept (for example, Electronic Commerce and EC) are placed in the same place in the features vector structure. Fourth, concepts in each paper are represented using a features vector built by mapping the concepts to the FVS. Fifth, these papers' features vectors are presented to a GHSOM neural network for the purpose of training and generating ontology of the field of interest. Sixth, the generated ontology quality is evaluated and reviewed for possible enhancements. This may require an interference of a human expert. Possible enhancements can be due to the need of deeper hierarchy for better representation. The seventh step is to feed back any required improvements and repeat the process to enable the ontology to evolve. Figure 3 summarizes the whole process cycle.

Decision Supports Systems Journal Experiment

As a proof-of-concept, we have run an experiment to build ontology for the *Decision Support Systems (DSS)* journal. The nature of generated ontology in this experiment is semi-automated. We collected the data using *ISI Web of Science*. The collected data covers 768 articles appeared in *DSS* between January 1991 and March 2003.

For GHSOM training, we used a MATLAB implementation called GHSOM Toolbox (Chan and Pampalk, 2003). We developed a small MATLAB program to combine together the two parser output files (normalized individual vectors file and its template file). This is required by GHSOM Toolbox. Two other programs are developed to work in conjunction of the first program: one allows us to manipulate manually which term to include or exclude, and the others to automatically remove stop words.

Then we use GHSOM_TRAIN function to train the network. We start with a single map (the top one). Then, we try different breadth settings to grow the top map. At this stage, it is better to prevent any hierarchical growing in order to accelerate the training process. Different generated maps are evaluated to check if their sizes are enough to represent balanced number of aggregated subjects. Once a map is believed to be sufficient to represent top concepts, hierarchical growing is allowed for hierarchical ontology to evolve. Table 2 shows some training examples with different settings.

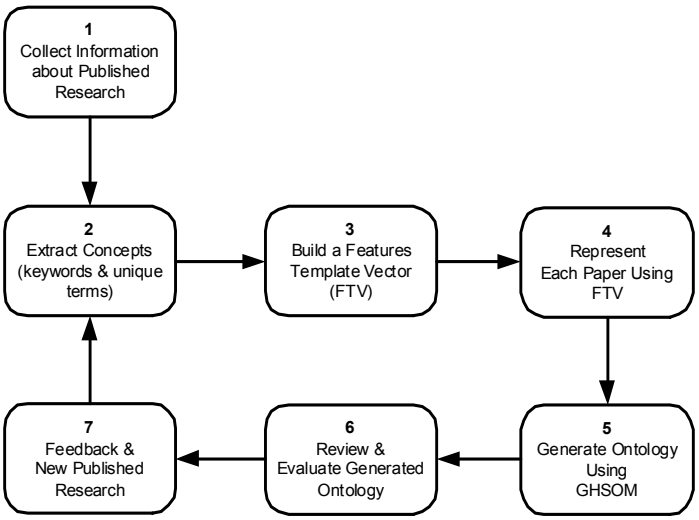

Figure 3. Ontology Generation Life Cycle Using GHSOM

Table 2. Training Examples

Training Settings		Training Outputs			
Breadth	**Depth**	**Map#1 Size (Row x Col)**	**Maps No**	**Layers No**	**(Maps No) / (Layer)**
0.90	0.10	8x9	1	1	1/1
0.90	0.01	8x9	48	4	1/1, 32/2, 14/3, 1/4
0.928	0.01	7x6	55	4	1/1, 29/2, 24/3, 1/4
0.928	0.02	7x6	27	3	1/1, 21/2, 5/3
0.95	0.01	7x4	63	5	1/1, 24/2, 30/3, 7/4, 1/5
0.95	0.02	7x4	31	4	1/1, 22/2, 7/3, 1/4
0.95	0.03	7x4	16	3	1/1, 14/2, 1/3
0.95	0.05	7x4	6	2	1/1, 5/2
0.95	1	7x4	1	1	1/1

Once the training has been completed, we need to label the units that appear on the hierarchy's different maps. We employ two labeling methods: GHSOM_DATALABELS and GHSOM_LABELSOM. The first method labels every unit (cluster) on a map based on the documents assigned to them. The LABELSOM method (Rauber, 1999) labels maps' units based on the main features that characterized them.

Once a hierarchy is labeled, it is ready to be visualized. We developed two visualization functions: GHSOM_WEBVIS (to be used with GHSOM_DATALABELS and to produces *documents view*) and GHSOM_WEBFVIS (to be used with GHSOM_LABELSOM and to produce *features (keywords) view*). Both represent the hierarchy as set of connected Web pages. Each Web page represents a single map. The first function can take a list of the documents description files names and use them to link every label representing a document number to its description file.

The features view of the hierarchy shows all of the important keywords associated with different maps' clusters. These keywords provide a good description of every cluster of papers. However, these labels are not good enough for presentation as a taxonomy. For our experiment, we studied the two hierarchy views (*documents* and *features* views) to try to come up with reasonable labels for the ontology. Each author tried to come with his own labels alone. Then we discussed and finalized the labels jointly. In some cases, labeling the maps units is easy and direct while in other cases hard and difficult. The former represents units that are very

homogenous and their dominant subject(s) are easy to identify (by investigating the features maps alone). The latter represents heterogeneous units where one hardly finds one or few subjects that can describe the group. This is usually found on the maps located near the top of ontology hierarchy.

The results of our experiment, using different settings for the training, can be found at **http://catt.okstate.edu/mohamas/DSS_ontology/**. Figure 4 shows an example for the generated hierarchical ontology. It is interesting to notice that similar topics are mapped to adjacent cells.

Conclusion

Most ontologies have to be built manually. This makes the whole process inefficient and error-prone. There is a major need for new tools to automate this process. One of the contributions of this paper is the introduction of the applicability of GHSOM in the context of ontology field. We believe that GHSOM is an important addition for any ontology engineer's tools-bag. It can be used as a stand-alone tool or in combination with other existing tools to semi-automate, if not automate, ontology generation. The other contribution of this paper is its plan to use this new tool to generate an automated ontology for *Decision Support Systems* journal. Figure 4 shows that it is possible to develop an overall view of a journal's research diversity and depth. Nonetheless, we are early in the process to judge its success. The semi-automated process of ontology generation needs to be replicated across journals and fields. In addition, the validity of generated ontology should be verified.

References

Chan, A., and Pampalk, E. "GHSOM Toolbox," 2003, **http://www.ai.univie.ac.at/~elias/ghsom/**.

Chandrasekaran, B., Josephson, J.R., and Benjamins, V.R. "What are ontologies, and why do we need them?," *IEEE Intelligent Systems & Their Applications* (14:1), 1999, pp. 20-26.

Ding, Y., and Foo, S. "Ontology research and development. Part 2 - a review of ontology mapping and evolving," *Journal of Information Science* (28:5), 2002a, pp. 375-388.

Ding, Y., and Foo, S. "Ontology research and development. Part I - a review of ontology generation," *Journal of Information Science* (28:2), 2002b, pp. 123-136.

Dittenbach, M., Rauber, A., and Merkl, D. "Uncovering hierarchical structure in data using the growing hierarchical self-organizing map," *Neurocomputing* (48), 2002, pp. 199-216.

Gruber, T.R. "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition* (5:2), 1993, pp. 199-220.

Kohonen, T. "Self-Organized Formation of Topologically Correct Feature Maps," *Biological Cybernetics* (43:1), 1982, pp. 59-69.

Rauber, A. "LabelSOM: On the Labeling of Self-Organizing Maps," *Proceedings of the International Joint Conference on Neural Networks (IJCNN'99)*, Washington, DC, 1999, pp. 3527-3532.

Sugumaran, V., and Storey, V.C. "Ontologies for conceptual modeling: their creation, use, and management," *Data & Knowledge Engineering* (42:3), 2002, pp. 251-271.

Welty, C., and Guarino, N. "Supporting ontological analysis of taxonomic relationships," *Data & Knowledge Engineering* (39:1), 2001, pp. 51-74.

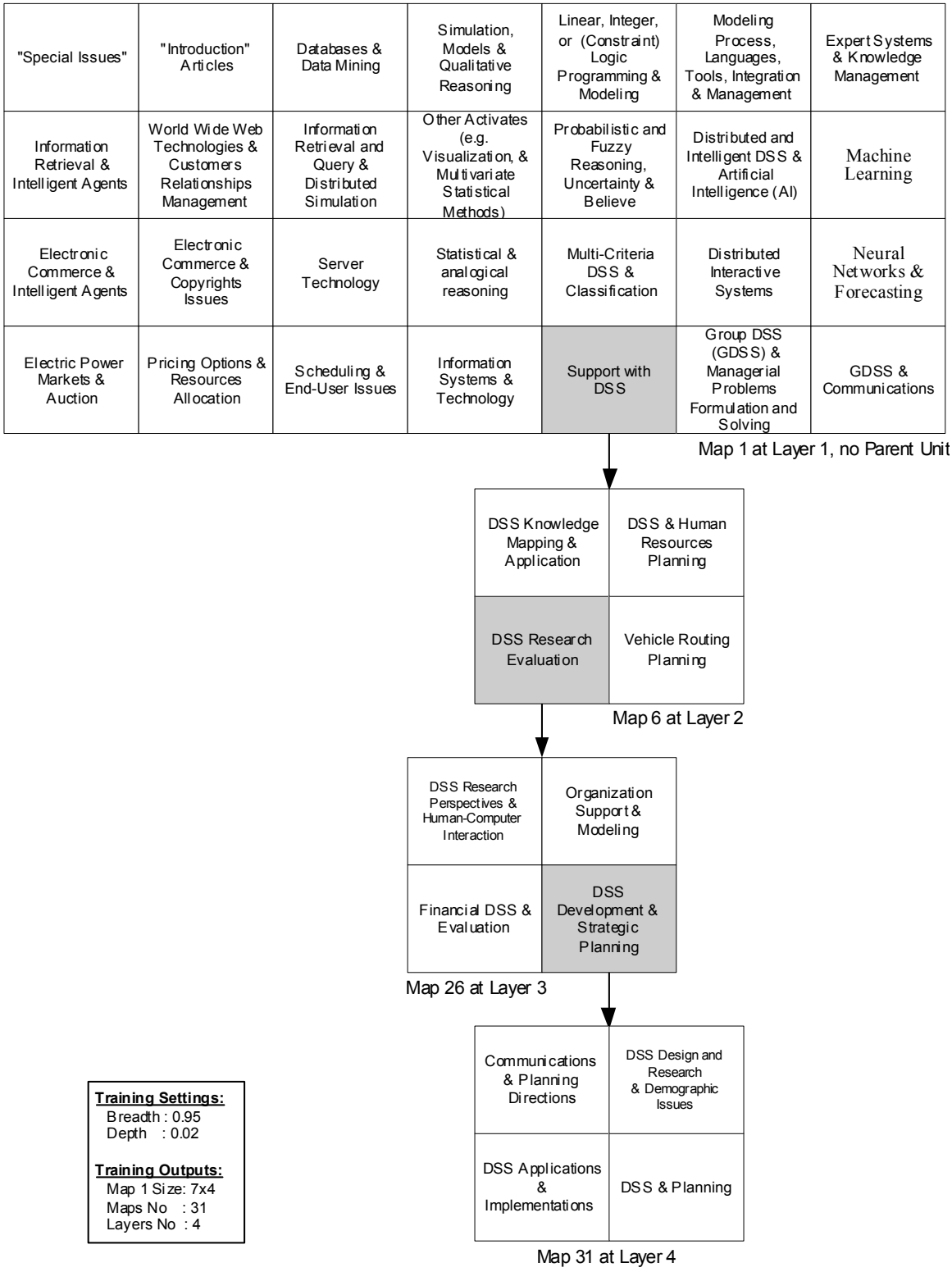

Figure 4. Example for DSS Journal Hierarchical Ontology