A Novel Approach to the Analysis of Volcanic-Domain Data using Self-Organizing Maps: A Preliminary Study on the Volcano of Colima

JRG[1]Pulido, EMR[2]Michel, MA[3]Aréchiga, and G[4]Reyes

[1] Faculty of Telematics, University of Colima, México, jrgp@ucol.mx
[2] Faculty of Telematics, University of Colima, México, ramem@ucol.mx
[3] Faculty of Telematics, University of Colima, México, mandrad@ucol.mx
[4] Volcanic observatory (RESCO), University of Colima, México, gard@ucol.mx

Abstract. This paper describes an approach for helping in the enduring task of analysing volcanic-domain data. This proposal allows domain experts to have a view of the knowledge contained in and that can be extracted from the digital archive. Specific-domain ontology components with further processing, and by embedding that knowledge into the digital archive itself, can be shared with and manipulated by software agents. In particular, we deal with the issue of applying an artificial learning algorithm, Self-Organizing Maps, to volcano-tectonic signals originated by the activity of the Volcano of Colima, Mexico. By applying this algorithm we have generated clusters of volcanic activity and can readily identify situations of risk for predicting important events.

1 Introduction

Every day the activity of the different volcanoes in the world attract the attention of the goverment and scientists. This activity varies in intensity. Volcanic seismology is, in most cases, one of the most deadly natural disasters in the world. In the worst case, whole areas are devastated by erupting volcanoes, including communities living near by. A number of computacional architectures and resources have been set up all around the world to monitor, forecast, and alert people regarding volcano activity. This paper is a first approach to the problem of volcanic seismology from the computational perspective, in particular applying Self-Organizing Maps.

The next generation of volcano domain computational tools require that the huge amount of information generated by volcanoes and contained into digital archives is structured [3]. In the last few years a number of proposals on how to represent knowledge via ontology languages have paraded [8, 14, 11, 22]. Now that OWL has become an standard [20], the real challenge, in the context of the semantic web, has started. In this paper in particular, the volcanic-domain problem is addressed. Eventually, the kowledge contained into volcano activity digital archives will become semantic knowledge, ie software agents will be able to understand, manipulate, and even carry out inferencing and reasoning tasks

Advances in Artificial Intelligence: Algorithms and Applications
Research in Computing Science 40, 2008, pp. 49-59

for us. Converting such as digital archives into semantic ones is to take much longer if no semi-automatic approaches are taken into account to carry out this enterprise. This is what our paper is all about.

The remainder of this paper is organized as follows. In section 2 some key concepts on ontologies are introduced. Some related work is presented in section 3. Our approach is described in section 4. The paper concludes in section 5 with some thoughts on the approach we have applied to analyse volcanic-domain data.

2 The Purpose of Ontology

Scientists among disciplines require a framework in order to be able to interact with each other. Ontology is a framework that makes it possible for people to communicate in a consistent, complete, and distributed way. Even more, we are able to encode for a particular domain:

- entities, objects, processes, and concepts.
- relationships of entities, objects, processes, and concepts.
- relationships across discipline areas.
- domain-dependant axioms.
- multilingual knowledge of the domain.
- assumptions, parameter settings, experimental conditions as well.

These are useful forms of knowledge representation which may be used to support the design and development of intelligent software applications and expert systems. One of the most common uses of ontologies is to support the development of agent-based systems for web searching, for example those described in [23]. For this interaction to be possible, agents must share a common ontology, or at least a common wrapper to existing information structures.

In Table 1, an excerpt of the *volcano* ontology in OWL defined by the Semantic Web for Earth and Environmental[5]Terminology is presented. Some superclasses are shown. From this, a taxonomy can then be derived or vicev-ersa in a semi-automatic way by means of appropriate ontology software tools. Representing knowledge about a domain as an ontology is a challenging process which is difficult to do in a consistent and rigorous way. It is easy to lose consistency and to introduce ambiguity and confusion [2]. Ontologies can be expressed with varying degrees of formality according to the level of formalisms they can be written, however the following four categories are the most common ways to express them [38, 7]:

1. **Highly informal** written using unstructured natural language, usually as a list of terms, no axioms, no glosses at all, stored in a raw file.
2. **Semi-informal** restricted and structured using natural language, no axioms, glosses appear usually as a data-dictionary, may use more complex data structures to be stored.

[5] http://sweet.jpl.nasa.gov

Table 1. A excerpt of a volcano ontology.

```
<owl:Class rdf:about="#Volcano">
    <rdfs:subClassOf rdf:resource="#TopographicalRegion"/>
    <rdfs:subClassOf rdf:resource="#VolcanicSystem"/>
.
.
.
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#primarySubstance"/>
        <owl:someValuesFrom rdf:resource="&substance.owl;#Magma"/>
      </owl:Restriction>
    </rdfs:subClassOf>
  </owl:Class>
```

3. **Semi-formal** using a formally defined language, a collection of concepts with a partial order induced by inclusion, basic axioms, some basic searching tasks may be carried out, stored in centralized databases.
4. **Rigorously formal** including axioms, theorems and proofs, inference and reasoning tasks can be carried out, stored in distributed repositories.

The last two are the most appropriate for software agents to use in the context of the semantic web, in particular the last one as it provides mechanisms to carry out inference and reasoning.

3 Related Work

One of the most important aspects of monitoring volcano activity is forecasting, on one hand. An important number of research papers on this area are found in the literature. On the other hand, in the context of the *semantic web*, perhaps the most important aspect is related to mapping unstructured data into software agent enable knowledge [3]. In the next subsections we have a brief look at some work done on the computational aspects of volcanology. We move then onto the ontology construction and taxonomy systems aspects.

3.1 Volcanology

A vast source of research is [41]. In this book, the properties of volcano-tectonic earthquakes are described. A methodology and some applications for predicting eruptions are discussed. A classification of volcanic earthquakes is also presented. A study of volcanic explosions carried out onto four volcanos is described in [39]. This study focuses on applying several basic statistical techniques to small-scale events in trying to find clustering properties. An important software tool for volcanic-domain data is visualization. In [16] a study that explores these

techniques is presented. Researchers in the geoscience areas consider increasingly important using visualization and clustering software tools as an useful device to analyse data. The Volcano of Colima, Mxico, is one of the most active volcanoes in the world and the Telemetric Seismic Network (RESCO) monitors it. In Table 2, an excerpt of volcano signal sampling is presented.

Table 2. Seismic signal samples. Date and time omitted

#####	TipEvent	EZV4	EZV5	Lat.	Long.	Mag.	Prof.	VelAp.	#E	Archivo
00046	ve	408	416	19.519	-103.629	3.8	0.8	17.97	6	02030131.rss
00047	lp	38	46	19.528	-103.612	1.0	2.8	14.68	6	02030202.rss
00048	ve	380	385	19.525	-103.607	3.7	2.9	11.57	6	02030240.rss
00049	lp	---	25	19.831	-103.526	0.6	15.0	10.43	3	02030255.rss
00050	lp	26	26	19.815	-103.489	0.7	15.0	12.09	3	02030257.rss
00051	lp	34	24	19.826	-103.512	0.8	15.0	10.31	3	02030258.rss
00052	rf	75	---	------	-------	---	---	-----	1	02030640.rss
00053	lp	12	12	19.827	-103.516	-0.1	15.0	11.00	3	02030813.rss
00055	ve	401	410	19.525	-103.628	3.7	1.7	17.00	6	02031045.rss

3.2 Constructing Ontologies

For the volcano domain ontology construction process it is important to identify knowledge components and not to start from scratch. A good ontology assures scientists that software agents can reason properly about the domain knowledge and, for instance, forecast important events. Web ontologies can take rather different forms [36]. In [7] an early approach, the so-called *Simple HTML Ontology Extension* (SHOE) in a real world internet application is described. This approach allows authors to add semantic content to web pages, relating the context to common ontologies that provide contextual information about the domain. Most web pages with *SHOE* annotations tend to have tags that categorize concepts, therefore there is no need for complex inference rules to perform automatic classification.

Two ubiquitous and inter-related concepts in meta-level descriptions of information are *hierarchy* and *proximity*. Data samples, in a *dataset*, can be described as being *close* to one another if they are similar in some sense (eq.1). Two samples might be close in one respect, say writing style, but distant in another respect, for example content. We are more interested in the latter. On one hand, a distance measure applied to a set of samples results in a partial order relation which can form the basis for an ontology [31], for instance by using the Euclidean distance:

$$c_{ij} = \frac{\sum_k x_k y_k}{\sqrt{\sum_k x_k^2 \sum_k y_k^2}} \tag{1}$$

It is desirable to have an objective measure of the *quality* of a given ontology in order that a decision can be made as to whether one representation is better or worse than another. At this point, it is very important to state that a domain expert must be always part of the team for validating the ontology.

3.3 Taxonomy systems

Creating a volcano domain taxonomy scheme may help improve predicting software systems. Again, ontology is a useful framework to construct such a schemes. Support for browsing using classification hierarchies is an important tool for users of digital archives, eg *Yahoo*[6]categories. Users would like the data to be structured in a way that makes sense from their point of view. The purpose of browsing an environment is to present the data in a structured way such that this facilitates the discovery of information for a given purpose. We are able to do so by using ontologies as taxonomy systems as well. In [32] a distributed architecture for the extraction of meta-data from WWW documents is proposed which is particularly suited for repositories of historical publications. This information extraction system is based on semi-structured data analysis. The system output is a meta-data object containing a concise representation of the corresponding publication and its components. These meta-data objects can be classified and organized and then interchanged with other web agents. In [17] an intelligent agent for libraries is described. This inhabits a rich virtual

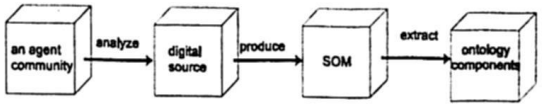

Fig. 1. Basic Approach. The ontology components from the domain are clustered together by a SOM. Further processing would allows us to embed this knowledge by means of OWL, for instance, into semantic digital archives for the current web to be transformed into software agent enable knowledge.

environment enhanced with various information tools to support searching.

Another interesting project is presented in [19], where the results of applying the *WEBSOM2*, a document organization, searching and browsing system, to a set of about 7 million electronic patent abstracts is described. In this case, a document map is presented as a series of HTML pages facilitating exploration. A specified number of best-matching points are marked with a symbol that can be used as starting points for browsing. Documents are grouped using Self-Organizing Maps (SOM), and then a graphical real-world metaphor is used to present the documents to users. That system was used as a front-end to a search engine. SOMLib and libViewer [28]. In SOMLib, maps can be integrated

[6] http://www.yahoo.com

to form a high-level library. This allows users to choose sections of a library to create personal libraries. Hierarchical feature maps consist of a number of individual self-organizing maps, and are able to represent the contents of a document archive in form of a taxonomy [24].

4 Our Approach

Our system (fig.1) can be regarded as a set of software tools that helps in the semi-automatic construction of domain-specific ontologies, in particular by clustering together a number of elements of the following sets [31]:

1. **Set of objects** (entities, concepts).
2. **Set of functions** (for example *is-a*).
3. **Set of relations** (*has* for instance).

Domain experts are always needed in order to validate the ontology components that have been identified. It can be inferred that an ontology should be produced in a *bespoke* manner to suit its purpose. This of course raises the crucial question of how such a purpose may be identified and specified. Linguistic resources such as *Wordnet* may help the domain expert in the validation of the ontology. Links to a set of *hyponyms*, incluiding instances, in *WordNet*[7] as explained in [25] can be introduced. Orology, speleology, and geophysics are hyponyms of geology. Asama, Pinatubo, and Colima are instances of Volcano for example.

4.1 Preparing the dataset

The obvious source of information for constructing a volcano-domain ontology is the data contained in the digital archives themselves. Datasets can be regarded as high dimensional vector spaces and can be represented either in a tabular form as shown in the following table:

D	v_1	$\cdots$	v_m
s_1	a_{11}	$\cdots$	a_{1m}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_n	a_{1n}	$\cdots$	a_{nm}

or in a mathematical way as follows:

$$d_j = \sum_k a_{jk} e_k \tag{2}$$

where $\{v_1, \cdots, v_n\}$ are n-dimensional *variables*, and $\{s_1, \cdots, s_n\}$ are m-dimensional *samples*, e_k is the unit vector and a_{jk} is the frequency of occurrence of v_j in s_k.

[7] http://wordnet.princeton.edu/perl/webwn

Our system consists of two applications: Spade and Grubber [5]. The former pre-processes data and creates a dataspace suitable for training purposes. The latter is fed with the dataspace and produces knowledge[8]maps that allow us visualize ontology components contained in the digital archive. As we have mendtioned, in a semantic context, they may later be organized as a set of *Entities*, *Relations*, and *Functions*. Problem solvers use this triad for inferring new data from [9, 10, 26] and carrying out reasoning.

4.2 Visualzing ontology components

By using Self-Organizing Maps we are able to cluster together volcano-domain ontology components. SOM can be viewed as a model of unsupervised learning and an adaptive knowledge representation scheme. Adaptive means that at each iteration a unique sample is taken into account to update the weight vector of a neighbourhood of neurons [18]. Adaptation of the model vectors take place according to the following equation:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)] \tag{3}$$

where t $\in \mathcal{N}$ is the discrete time coordinate, $m_i \in \Re^n$ is a node, and $h_{ci}(t)$ is a neighbourhood function. The latter has a central role as it acts as a smoothing kernel defined over the lattice points and defines the stiffness of the surface to be fitted to the data points. This function may be constant for all the cells in the neighbourhood and zero elsewhere. A common neighbourhood kernel that describes a natural mapping and that is used for this purpose can be written in terms of the Gaussian function:

$$h_{ci}(t) = \alpha(t) \exp(-\frac{||r_c - r_i||^2}{2\sigma^2(t)}) \tag{4}$$

where $r_c, r_i \in \Re^2$ are the locations of the winner and a neighbouring node on the grid, $\alpha(t)$ is the learning rate ($0 \leq \alpha(t) \leq 1$), and $\sigma(t)$ is the width of the kernel. Both $\alpha(t)$ and $\sigma(t)$ decrease monotonically. The major steps of our approach are as follows:

1. Produce a *dataspace*. A dataset is created with the individual vector spaces from the domain by *spade*. In some cases, when the dataset already exists, spade carries out a pre-processing validation task.
2. Construct the SOM. A second software tool, *grubber*, is fed and trained with the dataset and ontology maps are then created .

Ontology components can be visualized clustered together from the knowledge maps created. In most cases, raw datasets have to be pre-processed. Once the dataset is a valid one, *grubber* can be fed into with them. We start with a randomly initialized map and after a training process, clusters of ontology components can be readily identified from the map. The regions on the maps are

[8] Ontology maps and knowledge maps are here used indistinctly.

formed by merging nodes that have the same most representative samples. After the maps are trained through repeated presentations of all the samples in the collection, a labelling phase is carried out. Neighbouring nodes that contain the same winning elements merge to form concept regions. The resulting maps represent areas where neighbouring elements are similar to each other. The software interface created allows us to relate information from the samples in such a way that each node has a feature that relates to its corresponding subfeatures. This can be seen as we browse the maps and that help us understand the clusters that have been formed. Some classic approaches to the problem of clustering, on one hand, include partitional methods [29], hierarchical agglomerative clustering [34], and unsupervised bayesian clustering [27]. A widely used partitional procedure is the k-means algorithm [15]. A problem with this procedure is the selection of k a priori. PCA, on the other hand, is an excellent tool for reducing the size of the dataset. It allows the distance between samples to be measured in a well-defined and consistent manner [6]. An alternative to these methods is SOM which does not make any assumptions about the number of clusters a priori, the probability distributions of the variables, or the independence between variables. A comparative of these methods is not presented here. Preliminary results were surprisingly close to our intuitive expectations. After this, some other ontology tools such as editors can be used to organize this knowledge. Then, it can be embedded into the digital archive where it was extracted from by means of any of the ontology languages that exist (fig.2). Some results of applying our approach in other domains have been reported [35], and we are now further researching on the volcano domain in order to validate our results. At this stage we have already considered the use of hybrid systems that in combination of our approach will help in the semi-automatic construction of specific-domain ontologies [21].

5 Conclusions

The vast amount of data generated by volcanoes has eventually to be transformed into semantic data. In the context of the semantic web, by using semantic knowledge, software agents are able to carry out inference and reasoning tasks for us. In the volcano-domain, software agents may be of help in forecasting important events. An ontology is a form of knowledge representation that provides a common vocabulary of concepts and relationships which may be used to inform a viewer, a search engine or to inform other software entities. Agents have to interact with other agents using these dissimilar concepts. Therefore mechanisms and forms to exchange information and knowledge among different disciplines are needed. As we have already seen, ontologies can be used to give a sense of order to unstructured digital sources such as volcano-domain data. The acquisition and representation of knowledge needs to take into account the complexity that is often present in domains as well as the needs of the agents carrying out the search, a volcano-domain expert is always needed in order to assure the quality of the ontology created.

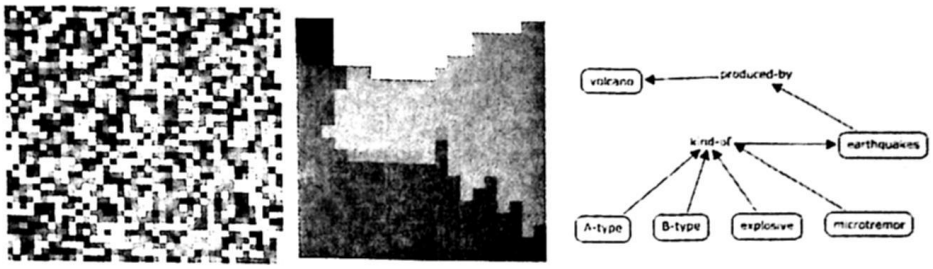

Fig. 2. After a training process, a randomly initialized SOM (left) becomes a categorized one (middle). Then an ontology can be derived (right).

In this paper we have presented a novel approach that generates clusters of volcanic activity and can readily help us identify situations of risk for predicting important events. However, some more research is required in order to fine tune the semi-automatic specific-domain ontology creation process.

References

1. K Bontcheva. Generating tailored textual summaries from ontologies. In A Pérez and J Euzenat, editors, *The Semantic Web Research and Applications*, volume 3532 of *LNCS*, pages 531–545–. Springer, 2005.
2. R Brachman. What is-a and isn't: An analysis of taxonomic links in semantic networks. *IEEE Computer*, 16(10):10–36, 1983.
3. L Crow and N Shadbolt. Extracting focused knowledge from the Semantic Web. *Int.J.Human-Computer Studies*, 54:155–184, 2001.
4. A Duineveld et al. Wondertools? a comparative study of ontological engineering tools. *Int.J.Human-Computer Studies*, 52:1111–1133, 2000.
5. D Elliman and *JRG Pulido*. Visualizing ontology components through self-organizing maps. In D Williams, editor, *6th International Conference on Information Visualization (IV02), London, UK*, pages 434–438. IEEE Computer Soc.Press, Los Alamitos, 2002.
6. G Foody. Applications of the self-organising feature map neural network in community data analysis. *Ecological Modelling*, 120:97–107, 1999.
7. A Gangemi et al. Ontology integration: Experiences with medical terminologies. In N Guarino, editor, *Formal Ontology in Info.Systems*, volume 46, pages 163–178. IOS Press, Amsterdam, 1998.
8. A Gómez and Oscar Corcho. Ontology languages for the Semantic Web. *IEEE Intelligent Systems*, 2002.
9. A Gómez et al. Knowledge maps: An essential technique for conceptualisation. *Data & Knowledge Engineering*, 33:169–190, 2000.
10. J Gordon. Creating knowledge maps by exploiting dependent relationships. *Knowledge-Based Systems*, pages 71–79, 2000.
11. J Heflin et al. Applying ontology to the web: A case study. *Engineering Applications of Bio-Inspired Artificial Neural Networks*, 1607, 1999.

12. J Hendler and E Feigenbaum. Knowledge is power: The Semantic Web vision. In N Zhonh et al., editors, *Web intelligence: Research and development*, volume 2198 of *LNAI*, pages 18–29. Springer-Verlag, Berlin, 2001.
13. V Hodge and J Austin. Hierarchical word clustering – automatic thesaurus generation. *Neurocomputing*, 48:819–846, 2002.
14. I Horrocks et al. From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of web semantics*, 1(1):7–26, 2003.
15. R Johnson and D Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, New Jersey, 4th edition, 1998.
16. B Kadlec et al. Visualization and analysis of multi-terabyte geophysical datasets in an interactive setting with remote webcam capabilities. In X Yin et al., editors, *Computational earthquake physics: simulations, analysis and infrastructure PART II.*, pages 2455–2465. Birkhauser-Verlag, Basel, 2006.
17. D Kirsh. Designing virtual libraries to help users find what they want. In C Landauer and K Bellman, editors, *The Virtual Worlds and Simulation Conf.-VWSIM'98*, pages 221–224. Soc.Computer Simulation Int, San Diego, 1998.
18. T Kohonen. *Self-Organizing Maps.* Information Sciences Series. Springer-Verlag, Berlin, 3rd edition, 2001.
19. T Kohonen et al. Self organization of a massive text document collection. In E Oja and S Kaski, editors, *Kohonen Maps*, pages 171–182. Elsevier Sci, Amsterdam, 1999.
20. L Lacy. *OWL: Representing Information Using the Web Ontology Language.* Trafford Publishing, USA, 2005.
21. S Legrand and *JRG Pulido.* A hybrid approach to word sense disambiguation: Neural clustering with class labeling. In P Buitelaar et al., editors, *Workshop on knowledge discovery and ontologies, 15th European Conference on Machine Learning (ECML), Pisa, Italy*, pages 127–132, September 2004.
22. P Martin and P Eklund. Embedding knowledge in web documents. *Computer Networks*, 31:1403–1419, 1999.
23. J McCormack and B Wohlschlaeger. Harnessing agent technologies for data mining and knowledge discovery. In *Data Mining and Knowledge Discovery: Theory, Tools and Technology II*, volume 4057, pages 393–400, 2000.
24. D Merkl. Document classification with self-organizing map. In E Oja and S Kaski, editors, *Kohonen Maps*, pages 183–192. Elsevier Sci, Amsterdam, 1999.
25. G Miller et al. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1991.
26. E Motta et al. Ontology-driven document enrichment: principles, tools and applications. *Int.J.Human-Computer Studies*, 52:1071–1109, 2000.
27. J Principe. *Neural and Adaptive Systems, Fundamentals through Simulations*, chapter 7. Wiley, USA, 2000.
28. A Rauber and D Merkl. The SOMLib digital library system. *LNCS*, 1696:323–342, 1999.
29. B Ripley. *Pattern Recognition and Neural Networks*, chapter 1,9. University Press, Cambridge, 1996.
30. H Ritter and T Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254, 1989.
31. G Salton. *Automatic Information Organization and Retrieval.* McGraw-Hill, New York, 1968.
32. I Sanz et al. Gathering metadata from web-based repositories of historical publications. In A Tjoa and R Wagner, editors, *9th Int. Workshop on Database and*

Expert Systems Apps, pages 473–478. IEEE Computer Soc.Press, Los Alamitos, 1998.

33. F Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.
34. H Tanaka et al. An efficient document clustering algorithm and its application to a document browser. *Information Processing and Management*, 35:541–557, 1999.
35. *JRG Pulido* et al. Identifying ontology components from digital archives for the semantic web. In *IASTED Advances in Computer Science and Technology (ACST)*, pages 1–6, 2006. CD edition.
36. *JRG Pulido* et al. Ontology languages for the semantic web: A never completely updated review. *Knowledge-Based Systems*, Elsevier volume 19, issue 7:489–497, 2006.
37. *JRG Pulido* et al. Artificial learning approaches for the next generation web: part I. *Ingeniería Investigación y Tecnología, UNAM (CONACyT), México*, 9(1):67–76, 2008.
38. M Uschold and M Gruninger. Ontologies: Principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155, 1996.
39. N Varley et al. Applying statistical analysis to understand the dynamics of volcano explosions. In H Mader et al., editors, *Statistics in volcanology.*, pages 57–76. Geological society for IAVCEI, London, 2006.
40. Y Yang et al. A study of approaches to hypertext categorization. *J.Intelligent Information Systems*, 18(2/3):219–241, 2002.
41. V Zobin. *Introduction to volcanic seismology.* Elsevier, Amsterdam, 2003.