Semantic Labeling of Geo-Concept Clusters

S. Kontaxaki[1], M. Kokla[2], M. Kavouras[3]

National Technical University of Athens, 9, H. Polytechniou Str., 157 80 Zografos Campus, Athens, Greece
Emails: [1]skontax@mail.ntua.gr, [2]mkokla@survey.ntua.gr, [3]mkav@survey.ntua.gr

1. Introduction

In literature, clustering methods usually put emphasis on cluster creation; however, cluster labeling is equally important, since it accentuates the meaning of the clustering results. Furthermore, most of the existing cluster labeling approaches are based on the assessment of term frequencies within documents grouped into the same cluster (Merkl and Rauber, 1999; Popescul and Ungar, 2000; Treeratpituk and Callan, 2006) or on the discovery of candidate labels from concepts of an upper ontology (Stein and zuEissen, 2004) or the Web (Pantel and Ravichandran, 2004).

In the geospatial domain, semantic information sources are usually taxonomically structured, i.e., geographic concepts are described on the basis of terms, definitions and their relations with other concepts. Semantic-based applications such as the development of Spatial Data Infrastructures (SDIs) and the Semantic Web require the integration of several such taxonomies, classifications, ontologies, etc. In this context, clustering approaches maybe used to group similar concepts from different taxonomies.

The present paper does not deal with the clustering problem as such but with the problem of assigning labels to these new clusters. A method is introduced for generating semantic labels for clusters of geographic concepts described by natural language definitions. The aim is to create labels by taking advantage of the semantic information immanent in the definitions of geographic concepts instead of resorting to external information. The resulting labels are structured so as to epitomize the meaning of the clusters, in order to be readily understood by users in the context of semantic-based applications.

2. Semantic Information Extraction

Definitions constitute a prominent source of semantic information, which can be automatically extracted due to their special structure and content (Jensen et al., 1993). Usually the definition of a geographic concept consists of two parts: 1) the *genus* and (2) the *differentiae* (Kavouras and Kokla, 2008). The genus part comprises the hypernyms or superordinates of the defined concept, whereas the differentiae refer to the remainder elements of the definition, which help to distinguish the concepts with the same genus.

Semantic information extraction is a process consisting in: (a) the syntactic analysis of an input set of geographic concepts' definitions, and (b) the application of patterns in order to extract semantic information in terms of semantic elements. The semantic information extraction process used thereinafter is based on the methodology described in (Kokla 2008, Kavouras and Kokla 2008). The genera of the definitions are mapped to IS-A semantic elements while the differentiae to other semantic elements (e.g. SHAPE, SIZE, ADJACENCY, etc).

In order to generate semantic labels, the proposed method starts with the semantic information extraction from geographic concepts' definitions. For example, Figure 1

shows the semantic elements and their corresponding values extracted from the definitions of the geographic concepts *Fosse*, *Gorge* and *Trough*, found in the Glossary of Landform and Geologic Terms (USA National Resources Conservation Service[1]). In next, the paper will use this small running example to illustrate the steps followed by the method.

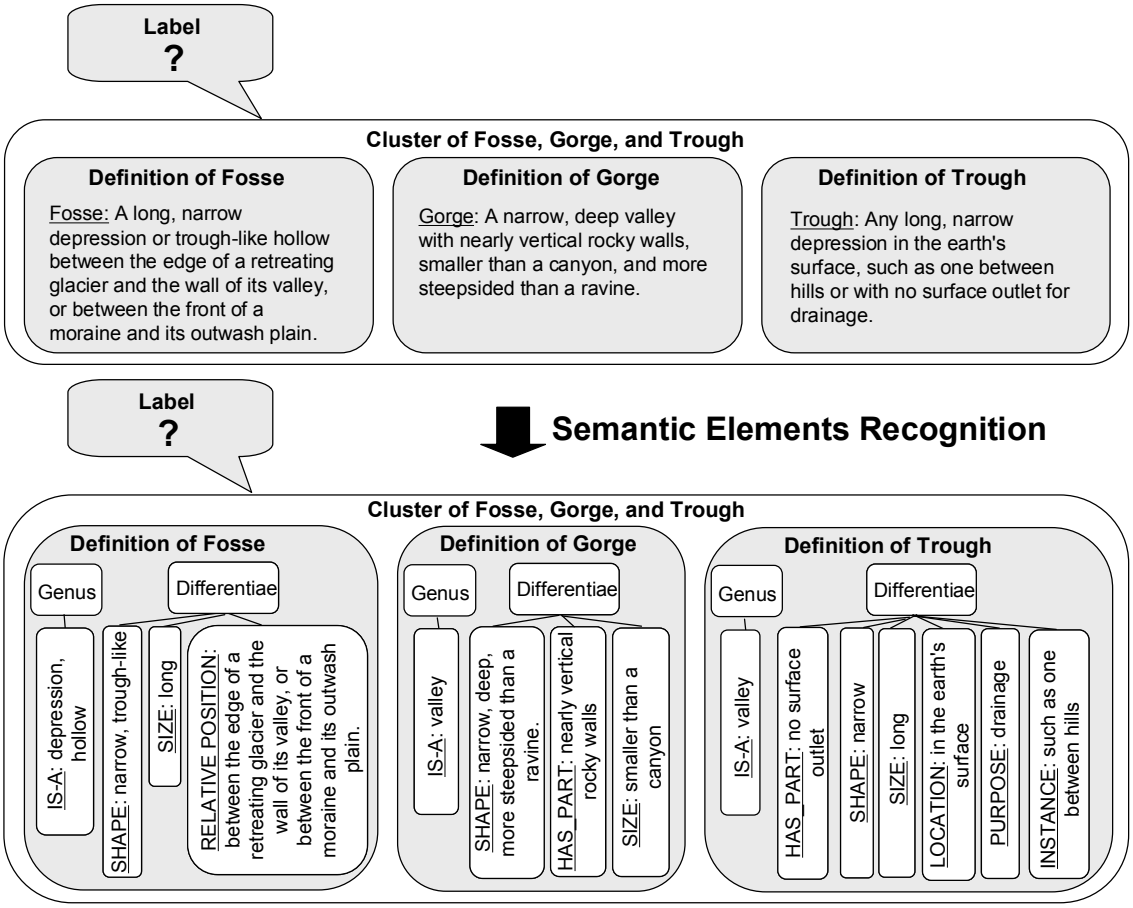

Figure 1. Example of semantic elements recognition from definitions.

3. Determination of a Label's Genus

The method proposes that labels are structured similarly to definitions, i.e., they are described by natural language phrases with genus and differentiae parts (Figure 2).

[1] Available at http://soils.usda.gov/technical/handbook/contents/part629.html

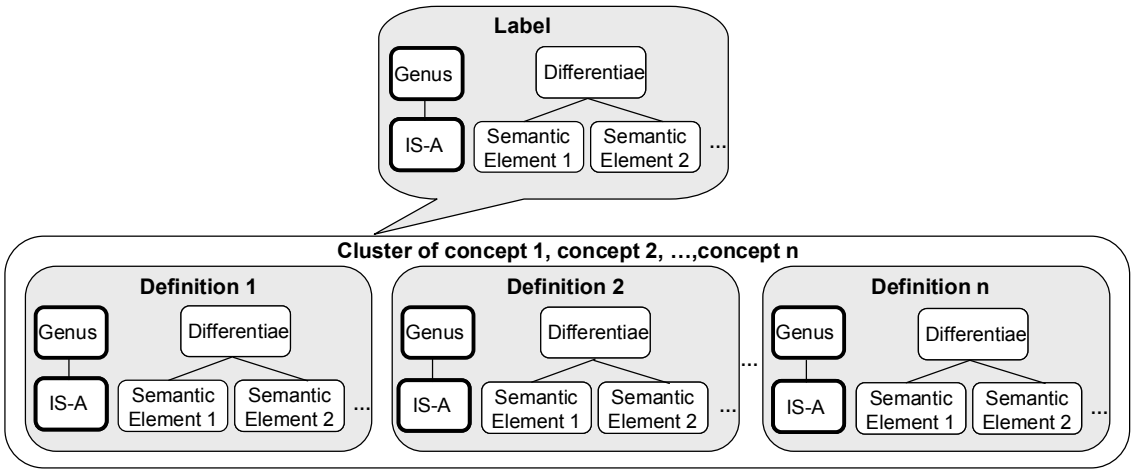

Figure 2. Definitions and labels structure.

The determination of a label's genus requires that both the terms of the concepts to be clustered and the genera of their definitions consisting of the recognized IS-A semantic elements, are firstly disambiguated and enriched with extra semantic information (synonyms and hypernyms). The disambiguation, which distinguishes between homonym terms, and the retrieval of the extra information can be done from knowledge bases, like *Wordnet*[2].

Once the disambiguation and enrichment are completed, the *common and most specific term* among the following is selected to represent the label's genus:

1) Concept terms and synonyms,
2) Genera and synonyms, concept terms' hypernyms, and
3) Genera hypernyms.

1) to 3) defines a sequence that starts from the most specific terms to the more abstract. Obviously, the more semantically different the concepts regrouped into a cluster are, the more abstract the label's genus will be.

Figure 3 shows how the label of the cluster consisting of the concepts *Fosse*, *Gorge* and *Trough*, is assigned the genus *Depression*.

[2] Available at http://wordnet.princeton.edu/

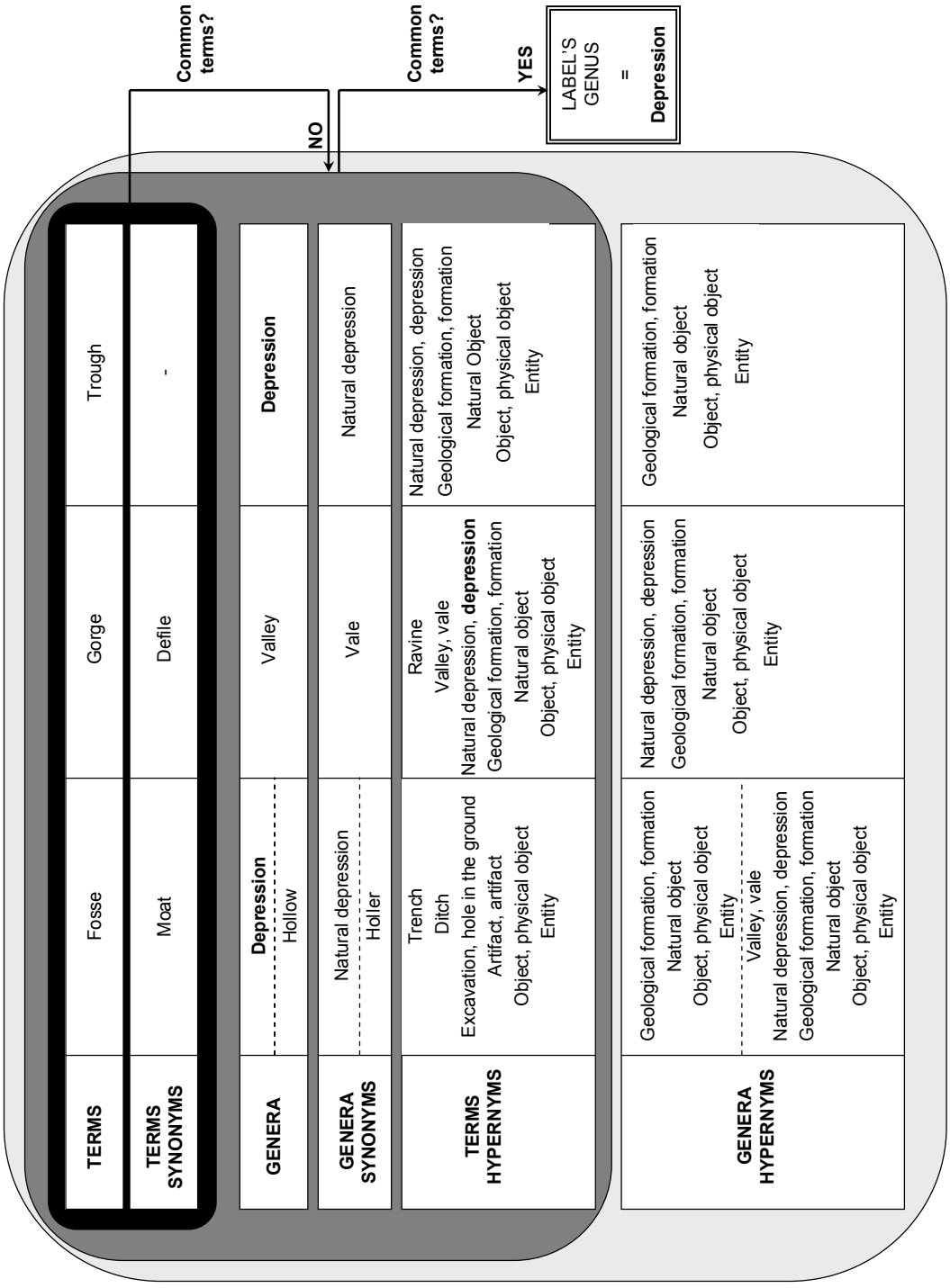

Figure 3. Label's genus determination example.

4. Determination of a Label's Differentiae

The determination of a label's differentiae requires finding the largest part of semantic information shared by the *common* semantic elements of the definitions. For this purpose, the method proposes that semantic elements are further decomposed into smaller and "comparable" segments of information, called *semantic particles* (Figure 4).

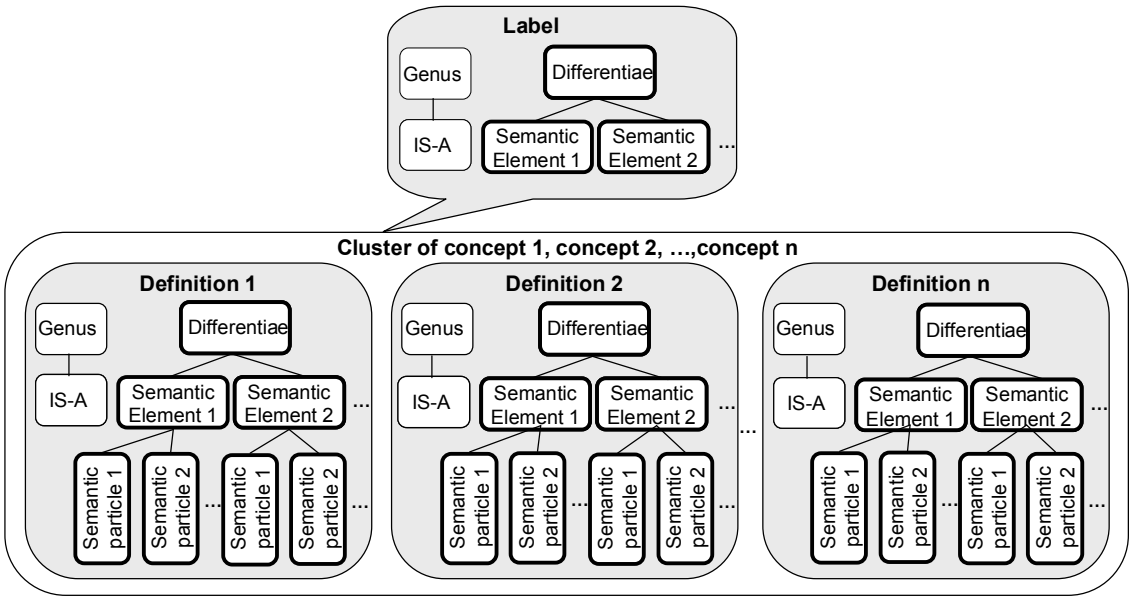

Figure 4. Semantic elements decomposition into semantic particles.

This further decomposition, made by parsing the semantic elements on the basis of predefined patterns, results to one parse tree *per* semantic element. Each parse tree starts from a token-root, from which hang the main subtrees, namely the semantic particles, and ends in leaves consisting of terminal tokens, which are either lexemes (e.g. *AND*, *NOT*, etc), or identifiers (e.g. *narrow*, *island*, etc) (Figure 5).

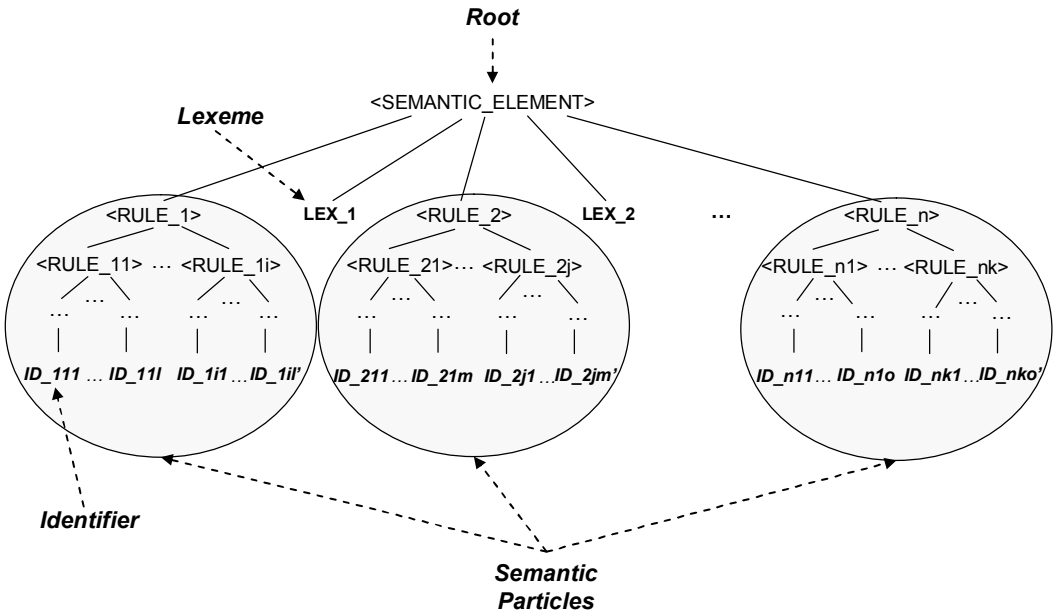

Figure 5. Parse tree and semantic particles.

As an illustrative example, the decomposition of the semantic element <SHAPE> of the concepts *Fosse*, *Gorge* and *Trough*, which results to semantic particles is shown in figure 6.

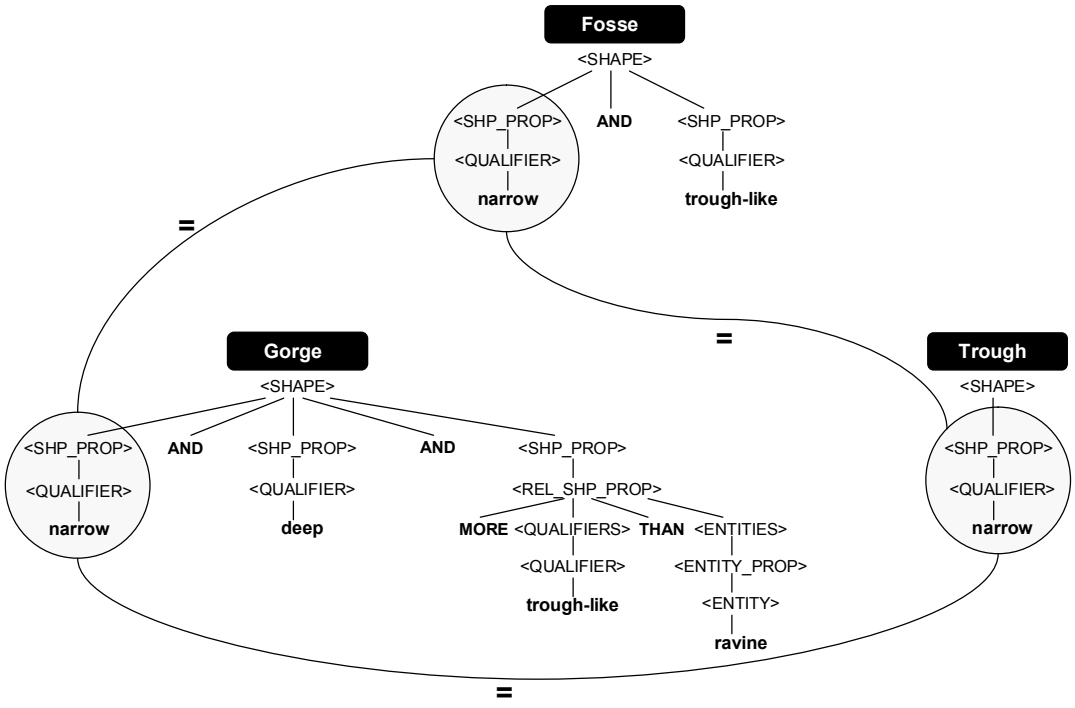

Figure 6. Example of common semantic information revealed by semantic particles.

In order to convey similar semantic information, the method assumes that semantic particles should have the same syntactic structure that ends to semantically similar identifiers. The semantic comparison of identifiers depends on the nonterminal tokens from which they are derived.

For example, for the semantic element <SHAPE>, identifiers are derived either from <QUALIFIER> or <ENTITY>. In the first case, two identifiers are semantically similar if they represent the same adjective or constitute synonyms or share common synonyms. If the lexemes *NOT* or *NO* precede, then identifiers should either constitute antonyms or share common antonyms. In the second case, two identifiers are semantically similar if they represent the same concept or constitute synonyms or share common synonyms.

The decomposition of the common semantic element <SHAPE> of the concepts *Fosse*, *Gorge* and *Trough* demonstrates that the largest part of semantic information shared by these concepts, is indicated by the semantic particles ending with the identifier *narrow* (Figure 6). Similarly, it could be demonstrated that these concepts do not share any information on the common semantic element <SIZE>.

Hence, in this small example, the label's differentia is set to *narrow* and the complete label becomes *Depression, narrow*.

5. Conclusion – Future Work

The present paper introduces a cluster labeling method based on the processing of a set of geo-concepts' definitions. In order to be semantically coherent and representative, the labels are structured as clear and concise definitions that both epitomize the meaning of clusters and differentiate one from the other.

The method can be used to label clusters of semantically similar concepts in several semantic-based applications, subjects of future work. Examples of such applications are the keyword-retrieval of large sets of concepts' definitions from geospatial data collections like glossaries or thesauri or the hierarchical organization of unstructured collections of geographic concepts for automatic ontology creation and integration.

Furthermore, the step-wise implementation of the method facilitates its future standardization and automation.

References

Jensen K., Heidorn G., and Richardson S. (Eds.), 1993, *Natural Language Processing: The PLNLP Approach*, Kluwer Academic Publishers, USA.

Kavouras M and Kokla M, 2008, *Theories of Geographic Concepts: Ontological Approaches to Semantic Integration*. CRC Press, Boca Raton, FL, USA.

Kokla M, 2008, GEONLP: A Tool for the Extraction of Semantic Information from Definitions. In: *Proceedings of the ISPRS 2008 Congress*, Beijing, Volume XXXVII, Commission II, 691–696.

Merkl D and Rauber A, 1999, Automatic Labeling of Self-Organizing Maps for Information Retrieval. *Lecture Notes in Artificial Intelligence*, Springer Verlag, 1574: 228–237.

Pantel P and Ravichandran D, 2004, Automatically Labeling Semantic Classes. In: *Proceedings of Human Language Technology / North American Association for Computational Linguistics (HLT/NAACL-04)*, Boston, MA, 321–328.

Popescul A and Ungar LH, 2000, *Automatic Labeling of Document Clusters*. Unpublished Manuscript, Department of Computer and Information Science, University of Pennsylvania, http://citeseer.nj.nec.com/popescu100automatic.html.

Stein B and zu Eissen SM, 2004, Topic Identification: Framework and Application. In: *Proceedings of the 4th International Conference on Knowledge Management*, Graz, Austria, 353–360.

Treeratpituk P and Callan J, 2006, Automatically Labeling Hierarchical Clusters. In: *Proceedings of the 6th National Conference on Digital Government Research*, San Diego, CA, 167–176.