Nearly-Automated Metadata Hierarchy Creation

Emilia Stoica and **Marti A. Hearst**
School of Information Management & Systems
University of California, Berkeley
102 South Hall, Berkeley CA 94720
{*estoica,hearst*}*@sims.berkeley.edu*

Abstract

Currently, information architects create metadata category hierarchies manually. We present a *nearly*-automated approach for deriving such hierarchies, by converting the lexical hierarchy WordNet into a format that reflects the contents of a target information collection. We use the term "nearly-automated" because an information architect should have to make only small adjustments to produce an acceptable metadata structure. We contrast the results with an algorithm that uses lexical co-occurrence statistics.

1 Introduction

Human-readable hierarchies of category metadata are needed for a wide range of information-centric applications, including information architectures for web sites (Rosenfeld and Morville, 2002) and metadata for browsing image and document collections (Yee et al., 2003).

In the information architecture community, methods for creation of content-oriented metadata tend to be almost entirely manual (Rosenfeld and Morville, 2002). The standard procedure is to gather lists of terms from existing resources, and organize them by selecting, merging and augmenting the term lists to produce a set of hierarchical category labels. Usually the metadata categories are used as labels which are assigned manually to the items in the collection.

We advocate instead a *nearly*-automated approach to building hierarchical subject category metadata, where suggestions for metadata terms are automatically generated and grouped into hierarchies and then presented to information architects for limited pruning and editing. To be truly useful, these suggested groupings should be close to the final product; if the results are too scattered, a simple list of the most well-distributed terms is probably more useful (a similar phenomenon is seen in machine-aided translation systems (Church and Hovy, 1993)).

More specifically, we aim to develop algorithms for generating category sets that (a) are intuitive to the target audience who will be browsing a web site or collection, (b) reflect the contents of the collection, and (c) allow for (nearly) automated assignment of the categories to the items in the collection.

For a category system to be intuitive, modern information science practice finds that it should consist of a set of IS-A (hypernym) hierarchies[1], from which multiple labels can be selected and assigned to an item, following the tenants of faceted classification (Rosenfeld and Morville, 2002; Yee et al., 2003). For example, a medical journal article will often simultaneously have terms assigned to it from anatomy, disease, and drug category hierarchies. Furthermore, usability studies suggest that the hierarchies should not be overly deep nor overly wide, and preferably should have concave structure (meaning broader at the root and leaves, narrower in the middle) (Bernard, 2002).

Previous work on automated methods has primarily focused on using clustering techniques, which have the advantage of being automated and data-driven. However, a major problem with clustering is that the groupings show terms that are *associated* with one another, rather than hierarchical parent-child relations. Studies indicate that users prefer organized categories over associational clusters (Chen et al., 1998; Pratt et al., 1999).

We have tested several approaches, including K-means clustering, subsumption (Sanderson and Croft, 1999), computing lexical co-occurrences (Schutze, 1993) and building on the WordNet lexical hierarchy (Fellbaum, 1998). We have found that the latter produces by far the most intuitive groupings that would be useful for creation of a re-usable, human-readable category structure. Although the idea of using a resource like WordNet for this type of application seems rather obvious, to our knowledge it has not been used to create subject-oriented metadata for browsing. This may be in part because it is very

[1] Part-of (meronymy) relations are also intuitive, but are not considered here.

In the Proceedings of HLT-NAACL'04

large and the word senses are assumed to be too fine-grained (Mihalcea and Moldovan, 2001), or its structure is assumed to be inappropriate.

However, we have found that, for some collections, starting with the assumption that there will be a small amount of hand-editing done after the automated processing, combined with a bottom-up approach that extracts out those parts of the hypernym hierarchy that are relevant to the collection, and a compression algorithm that simplifies the hierarchical structure, we can produce a structure that is close to the target goals.

Below we describe related work, the method for converting WordNet into a more usable form, and the results of using the algorithm on a test collection.

2 Related Work

There has been surprisingly little work on precisely the problem that we tackle in this paper. The literature on automated text categorization is enormous, but assumes that a set of categories has already been created, whereas the problem here is to determine the categories of interest. There has also been extensive work on finding synonymous terms and word associations, as well as automatic acquisition of IS-A (or genus-head) relations from dictionary definitions and glosses (Klavans and Whitman, 2001) and from free text (Hearst, 1992; Caraballo, 1999).

Sanderson and Croft (1999) propose a method called *subsumption* for building a hierarchy for a set of documents retrieved for a query. For two terms *x* and *y*, *x* is said to subsume *y* if the following conditions hold: $P(x|y) \geq 0.8, P(y|x) < 1$. The evaluation consisted of asking people to define the relation that holds between the pairs of words shown; only 23% of the pairs were found to hold a parent-child relation; 49% were found to fall into a more general related-to category. For a set of medical texts, the top level consisted of the terms: *disease, post polio, serious disease, dengue, infection control, immunology*, etc. This kind of listing is not systematic enough to appear on a navigation page for a website.

Lawrie et al. (2001) use language models to produce summaries of text collections. The results are also associational; for example, the top level for a query on "Abuses of Email" are *abuses, human, States Act*, and *Nursing Home Abuses*, and the second level under *abuses* is *e-mail, send, Money, Fax, account, address, Internet*, etc. These again are too scattered to be appropriate for a human-readable index into a document collection.

Hofmann (1999) uses probabilistic document clustering to impose topic hierarchies. For a collection of articles from the journal *Machine Learning*, the top level cluster is labeled *learn, paper, base, model, new, train* and the second level clusters are labeled *process, experi, knowledge, develop, inform, design* and *algorithm, function, present, result, problem, model*. We would prefer something more like the ACM classification hierarchy.

The Word Space algorithm (Schutze, 1993) uses linear regression on term co-occurrence statistics to create groups of semantically related words. For every word, a context vector is computed for every position at which it occurs in text. A vector is defined as the sum of all fourgrams in a window of 1001 fourgrams centered around the word. Cosine distance is used to compute the similarity between word vectors.

Probably the closest work to that described here is the SONIA system (Sahami et al., 1998) which used a combination of unsupervised and supervised methods to organize a set of documents. The unsupervised method (document clustering) imposes an initial organization on a personal information collection which the user can then modify. The resulting organization is then used to train a supervised text categorization algorithm which automatically classifies new documents.

3 Method

WordNet is a manually built lexical system where words are organized into synonym sets (synsets) linked by different relations (Fellbaum, 1998). It can be viewed as a huge graph, where the synsets are the nodes and the relations are the links. Our algorithm for converting it to create metadata categories for information organization and browsing consists of the following steps:

1. Select representative words from the collection.
2. Get the WordNet hypernym paths for one sense of each selected word.
3. Build a tree from the hypernym paths.
4. Compress the tree.

3.1 Select Representative Words

To make the hierarchy size manageable, we select only a subset of the words that are intended to best reflect the topics covered in the documents (although in principle the method can be used on all of the words in the collection).

The criteria for choosing the target words is information gain (Mitchell, 1997). Define the set W to be all the unique words in the the document set D. Let the *distribution* of a word w be the number of documents in *D* that the word occurs in. Initially, the words in W are ordered according to their distribution in the entire collection D. At each iteration, the highest-scoring word w is added to an initially-empty set S and removed from W, and the documents covered by w are removed from D. The process repeats until no more documents are left in D.

3.2 Get Hypernym Paths

For every word in S, we obtain the hypernym path of the word from WordNet. In the current implementation, we take the hypernym for the first sense of the word only,

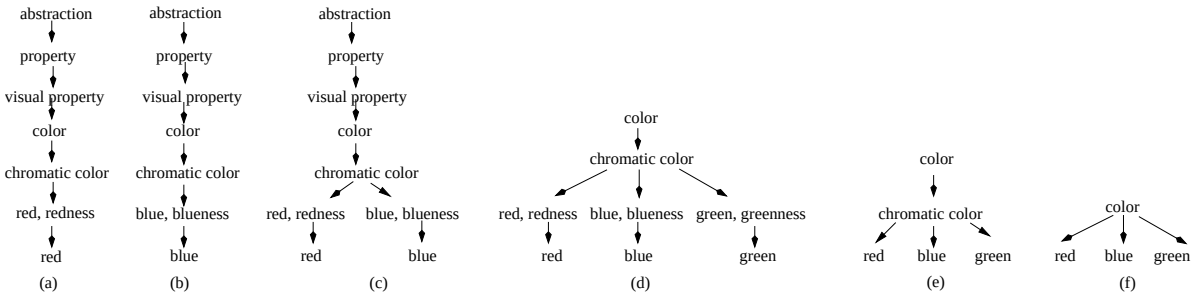

Figure 1: Building a hierarchy from WordNet. (a) The hypernym path for word *red*, and (b) *blue*. (c) Combining the paths of words *red* and *blue*, (d) The uncompacted tree for words *red*, *blue* and *green*, (e) The path after eliminating parents with less than two children, and (f) after eliminating children with name included in parent's name.

which is usually the most general. (In the future, we plan to explore how to disambiguate between senses based on the context in which the word appears in the document; see Discussion.) Figures 1(a) and 1(b) show the hypernym paths for words *red* and *blue*.

3.3 Build the Tree

Next we take the union of the hypernym paths of all words in set *S*, obtaining a tree, as shown in Figure 1(c).

3.4 Compress the Tree

The hypernym path length varies widely in WordNet, so we compress the tree using three rules:

1. Eliminate selected top-level (very general) categories, like *abstraction, entity*.
2. Starting from the leaves, eliminate a parent that has fewer than *n* children, unless the parent is the root.
3. Eliminate a child whose name appears within the parent's.

For example, consider the tree in Figure 1(d) and assume that $n = 2$ (eliminate parents that have fewer than two children). Starting from the leaves, by applying Rule 2, nodes *red, redness*, *blue, blueness*, and *green, greenness*, are eliminated since they have only one child. Figure 1(e) shows the resulting tree. Next, by applying Rule 3, node *chromatic color* is eliminated, since it contains the word *color* which also appears in the name of its parent. The final tree presented in Figure 1(f) produces a structure that is likely to be a good level of description for an information architecture.

Mihalcea and Moldovan (2001) describe a sophisticated method for simplifying WordNet, focusing on combining synsets with very similar meanings or dropping rarely used synsets. Their rules include what we define above as Rule 3. However, they focus on simplifying WordNet in general, rather than tailoring it to a specific collection, and focus on NLP applications that are likely to make use of every sense of a WordNet word. Nevertheless, it may be useful to explore using their simplified version of WordNet in future.

4 Results

We experimented with a collection of descriptions of approximately $35,000$ art documents containing about $23,000$ unique words.[2] Some sample documents are:

> *A French soldier clings to tree branches as a wolf stands beneath the tree.*
>
> *A Greek trellis with Ionic columns, meander crossing diagonally; few vines; trees background; trellis is in a circle.*

The descriptions are preprocessed by eliminating frequent words from a stop list. Information gain is used to select target words, in this case resulting in 849 words. Figure 2 shows partial results obtained using the WordNet algorithm (where compression reduced the number of nodes by 90%) and Word Space (Schutze, 1993).

Note that the WordNet-based organization is intuitive, but if not exactly what the designer wants, should be easy to adjust. For example, a designer may prefer to have a "nature" category that combines the subcategories of "geological formation," "body of water," and "vascular plant". Some terminology may also need renaming, but note that WordNet also provides thesaurus terms that can be used in an underlying search engine. Word Space, by contrast, produces associationally related terms.

5 Discussion and Future Work

We advocate the use of an existing rich lexical resource for the nearly-automated creation of hierarchical subject-oriented metadata for information browsing and navigation. We have created examples that show that a modified version of WordNet can produce a useful starting point for information organization projects. These have the added advantage of producing automated assignments of multiple labels to documents. We plan to augment the processing with more intelligent selection of hypernym

[2] This collection is also used in (Yee et al., 2003).

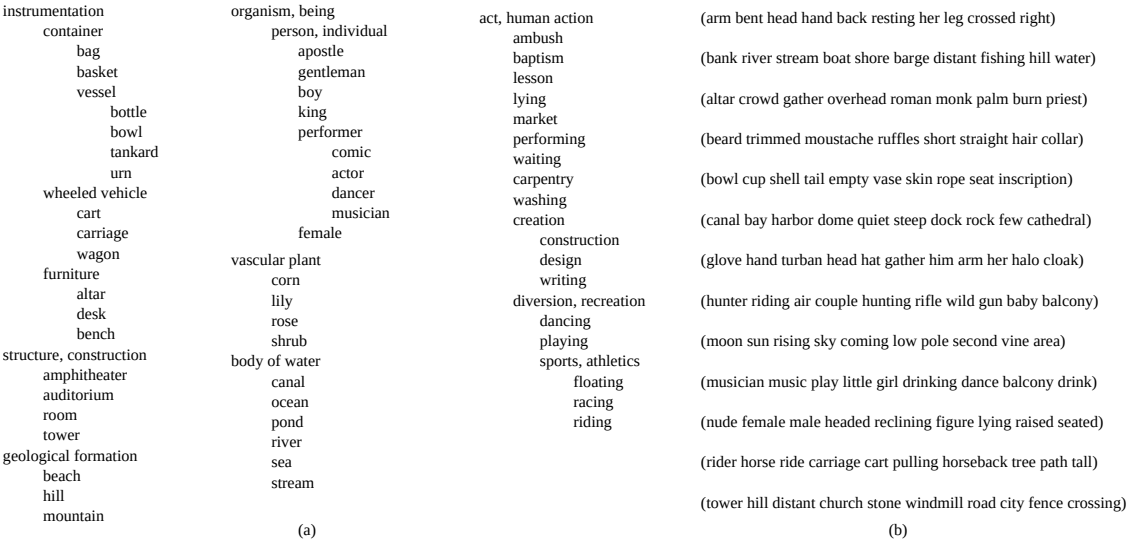

Figure 2: Comparison of partial results using (a) WordNet and (b) Word Space.

senses, as well as processing the descriptions to extract noun compounds and differentiate nouns from verbs. The method also worked well on a set of biomedical journal titles; we are in the process of determining how generally applicable the approach is. In addition, we are currently designing usability studies in which we will present different categorization suggestions to information architects to organize. Their subjective reactions, the amount of time it takes them to create the organizations, and the resulting quality and coverage of the organizations, as measured by users performing navigation tasks using the hierarchies, will be compared to other techniques.

Acknowledgements

This research was supported by NSF grants DBI-0317510 and IIS-9984741.

References

Michael L. Bernard. 2002. Examining the effects of hypertext shape on user performance. *Usability News*, 4(2).

Sharon A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of ACL '99*, College Park, MD.

Hsinchen Chen, Andrea L. Houston, Robin R. Sewell, and Bruce R. Schatz. 1998. Internet browsing and searching: User evaluations of category map and concept space techniques. *JASIS*, 49(7).

Ken Church and Eduard Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING '92*, Nantes, France.

Thomas Hofmann. 1999. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *Proceedings of IJCAI'99*, Stolckholm.

Judith Klavans and Brian Whitman. 2001. Extracting taxonomic relationships from on-line definitional sources using lexing. In *Proceedings of ACM/IEEE DL '01*, Roanoke, VA.

Dawn Lawrie, Bruce Croft, and Arnold L. Rosenberg. 2001. Finding topic words for hierarchical summarization. In *Proceedings of SIGIR '01*, New Orleans, LA.

Rada Mihalcea and Dan I. Moldovan. 2001. Ez.wordnet: Principles for automatic generation of a coarse grained wordnet. In *Proceedings of FLAIRS Conference 2001*.

Tom Mitchell. 1997. *Machine Learning*. McGraw Hill.

Wanda Pratt, Marti Hearst, and Larry Fagan. 1999. A knowledge-based approach to organizing retrieved documents. In *Proceedings of AAAI 99*, Orlando, FL.

Louis Rosenfeld and Peter Morville. 2002. *Information Architecture for the World Wide Web: Designing Large-scale Web Sites*. O'Reilly & Associates, Inc.

Mehran Sahami, S. Yusufali, and M. Q. W. Baldonaldo. 1998. SONIA: A service for organizing networked information autonomously. In *Proceedings of DL' 98*, New York.

Mark Sanderson and Bruce Croft. 1999. Deriving concept hierarchies from text. In *Proceedings of SIGIR '99*.

Hinrich Schutze. 1993. Word space. *Advances in Neural Information Processing Systems*, 5:895–902.

Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. Faceted metadata for image search and browsing. In *Proceedings of the CHI 2003*, Fort Lauderdale, FL.