Recognizing Emergent Nodes in Aligning Multiple Document Taxonomies

Tim Musgrove

TextDigger, Inc.
305 Vineyard Town Center #375
Morgan Hill, CA 95037 USA
tmusgrove@textdigger.com

Abstract. A document taxonomy alignment method, relying on document glosses and utilizing a soft ontology expansion, enables us to devise some all-new hierarchical leaf nodes for the purpose of better aligning a plurality of document taxonomies.

1. Introduction

In our past work of mapping different document taxonomies, we frequently were left with some "isolated nodes", i.e. categories of documents seeming to have no correlate in the other taxonomies. An example was in the Archery category on Yahoo, the sub-category of "Kyudo" (traditional Japanese archery). There was no equivalent to this category on DMOZ or About.com, the two taxonomies we were hoping to correlate. However, a soft ontology expansion we had devised to assist in the mapping meanwhile produced numerous candidate ontology nodes, such as "coaching/training" or "competitions/tournaments," and in this particular case, "traditional archery." While not a node in any of three reference taxonomies, "traditional archery" nonetheless applied to a great number of documents in all three, and especially in Yahoo's "Kyudo" category. Having used DMOZ as our "master taxonomy", we not only added "traditional archery" to it, but devised a method of automatically adding every other similar example, with the result of adding these new nodes: *Traditional Archery, Coaching & Training, Equipment & Gear, Stories & Discussion*.

The first of these, "traditional archery", included (as a child node) all the Kyudo documents, plus numerous documents from the other two indices, all of which pertain to traditional forms of archery. Since there are other traditional forms of archery (such as medieval European forms) besides Kyudo, it made sense that Kyudo be subsumed in the new node. We found that it was rather straightforward to devise a heuristic for automating this addition of nodes according to the following heuristic:

1. Find an expanded concept that is instantiated disproportionately in the document glosses of an unmapped node.
2. Test if that node is instantiated also in numerous documents not classified at a leaf node in a plurality of taxonomies.

3. If such a node is found, then create a new node with that concept and place the relevant documents under it.

In order to explain how this was accomplished, we will outline (1) our general approach to taxonomy node alignment by semantic resemblance; (2) our conception of a soft ontology expansion (3) the way in which results of the soft ontology expansion can be leveraged to create new nodes as described above.

2. Taxonomy alignment by semantic resemblance

One approach to taxonomy alignment is the intensional method, which examines the semantics of the names of the nodes, and the titles of documents, as well as the glosses applied to those documents by the taxonomy editors. We applied such a method to human-crafted document taxonomies bearing short glosses. These glosses are meant to summarize what the documents are about and what differentiates each one from others in the same topic, hence they are obviously valuable to our task.

We take the content words of the document titles and glosses, as well as bi-grams containing a topic word in any derived form (e.g., in the archery category we would take "field archery" and "archer's union", in addition to single words such as "arrows" and "bows"). We then check to see which of these may be closely related by semantic resemblance. For measuring semantic resemblance, we test for "semantic proximity" in WordNet, which we define as having a maximum distance of 2 in the WordNet hierarchy, with the additional limitations:

1. Only synonyms, hyponyms, hypernyms, and sister-terms are to be considered.
2. Sister-terms are considered proximate only if they share multiple content words in their glosses and/or example sentences in WordNet.
3. Hypernyms are included only if they are at least 4 levels down in the WordNet hierarchy from the root.

Note that this is similar to (Leacock 1998) in that it considers the depth of the taxonomy as counting toward semantic nearness, though our implementation is heuristic rather than statistical. (Since our application is to Web documents, we found it necessary to ignore certain words that are excessively frequent across all categories, and hence not useful, such as "photos", "contact details", "site map", etc.). Table 1 shows an outline of one of our case studies.

Table 1. Comparison of Archery in DMOZ, Yahoo and About.com

DMOZ	Yahoo	About.com
Chats & Forums	Bow Hunting	Shop for Archery & Bowhunting Gear
Clubs & Associations	Clubs & Organizations	Archery & Bowhunting Gear Manufacturers
Equipment Manufacturers	Competitions	Archery & Bowhunting Organizations
For Kids and Teens	Gear & Instruction	
Guides & Directories	Kyudo	
News & Media	Magazines	
Personal Pages	National Teams	
Tournaments & Events	Web Directories	

The result of our method is, for example, that "clubs" and "organizations" are treated as equivalent terms. This happens by means of a simple percentage match

scoring of the content words in node names. For example, the pair of "Equipment Manufacturers" and "Archery and Bowhunting Gear Manufacturers" receives a score of 0.80, owing to the following facts: First, "Archery" is omitted because it is the same as the overarching topic of "Archery" and hence implicit in all node names. Second, the stop word "and" is discarded. Third, "gear" is matched to "equipment" as a hypernym. That leaves five words total, with only one of them ("bowhunting") lacking a match: hence the score of 4/5 = 0.80. By trial and error we decided 0.66 was sufficient for alignment.

The virtue of this simple node name resemblance test is that it lets us align, for example, "Clubs and Organizations" with "Clubs and Associations" in two different taxonomies. However it leaves us with the different problem of the numerous documents not assigned a leaf node. In other words, in all three indices, many documents were simply classified in "Archery" without being assigned to a sub-category. In some cases, this seems correct, in that the documents in question were very general archery documents (or websites) not belonging to any particular sub-class. But in many other cases, it seemed that a node in a different taxonomy was a natural place for such documents. For example, a website of personal anecdotes and combined with feedback from others, was classified in one taxonomy simply as an "archery" document, but it would have found a perfect home in "Chats and Forums". This defeats taxonomy alignment, in that it is implied that none of the documents in the one taxonomy would belong in "Chats & Forums" of the other – and yet many of them did.

This type of predicament was later resolved, in some cases, by the results of a soft ontology expansion of all three taxonomies. In other words, after having enriched the ontological characterization of each specific leaf node, we could often align it with an appropriate subset of the documents lumped together in a more general topic of a different taxonomy.

3. Soft ontology expansion of document taxonomy leaf nodes using WordNet

For this exercise, we went back to our extracted words and bi-grams (e.g. "calendar" and "field archery", etc.), examined their WordNet glosses and example sentences and compared them with collocations and phrases in the document glosses, and found the following to hold true: if two words were frequently paired (collocated after skipping non-content words) in the taxonomy document glosses and also were found in each other's WordNet glosses, they were, without exception (in our case studies), genuinely related and of ontological import in the category. Our operational definition of "frequent" was: having at least one occurrence in all three taxonomies and having multiple occurrences (2 or more) in at least two of three taxonomies.

This technique has similarities to (Beneventano 2003) and (Martin 2004), in that it employs WordNet to develop one's taxonomy and/or ontology. The difference is that we are driving the process by reference to the glosses already created by editors of the various taxonomies. Our procedure derived the following soft concepts in Archery:

[calendar,schedule] having a relation to [event]

[tournament,competition] having relations to both [results] and [standings]
[outdoor] having a relation to [ranges]
[bow] having relations to [crossbow], [compound bow], and [long bow]

We call these "concepts" rather than merely "word occurrences" in view of the following: each is based on a small web of similar words, (e,g. "calendar"|"schedule"); each has an additional word relation ("events," etc.); all are contextualized to the local topic of Archery. The totality of all such extracted concepts we call a "soft ontology," in that it delineates raw materials of the local ontology, but obviously falls short of a formal representation of the relations between the concepts, such as those discussed in (Gaurino 1998).

Next, when checked the non-leaf-node documents' glosses for the presence of these concepts. If they matched, then we moved them to the newly created node. For example, several documents with glosses containing "discussion" and "stories" found their way into "Stories & Discussion." In the end, 37 of 189 documents were thus "migrated downward" to a leaf node, with the result that, on inspection, it seemed the alignment between taxonomies was more complete and intensionally unequivocal. This illustrates that taxonomy alignment cannot be divorced from issues of taxonomical scope and adequacy. If one taxonomy lacks the scope or granularity of another, then the only way to achieve proper alignment is to sort through some of the items in the less granular taxonomy so as to "multiply align" it to other nodes.

4. Emergent Nodes

Finally, we reached the result that certain of our soft ontology concepts embrace otherwise isolated nodes of one taxonomy, together with non-leaf-node documents of another. A clear example was the topic mentioned earlier, "Kyudo." Our soft ontology expansion had derived a "traditional archery" as a bi-gram. This was very dense in the Kyudo category (occurring in all but one of its items), while being found also in 16 non-leaf-node documents in DMOZ, including:

Donadoni **Archery**: Supplier of **traditional archery** equipment in Italy…
The **Archery** Centre: Specialists in field, **traditional**, and re-enactment **archery…**
Perris **Archery**: Recurve, compound and **traditional archery** equipment…

Our procedure was to use the concept string as a new node name (inserting "and" between words that had been found separately rather than directly collocated), and including as a child node the originally isolated node. So our master taxonomy now included "Archery/Traditional Archery/Kyudo" with several DMOZ documents placed in the new node "traditional archery." This satisfied us as being a far better alignment than we would have without the new node. Kyudo documents now had a closer parent than just being a direct child of "Archery." And the new interstitial node of "traditional archery" functions to explain where "Kyudo" belongs. We think the same of "Stories and Discussion" introduced as a parent of "Chat and Forums", and of "Coaching and Training" as a parent for "Instruction" documents that Yahoo had mixed with "Gear". Table 2 shows the overall alignment results.

Table 2. Results of alignment – New Nodes

New Nodes -Child node	DMOZ	Yahoo	About
Stories & Discussion - Chats & Forums	Chats & Forums	Glosses with "stories, ""discussion"	Glosses with "stories, ""discussion"
Equipment and Gear	Equipment Manufacturers	Glosses with "equipment" and "gear"	Archery & Bowhunting Gear Manufacturers, Shop for Archery & Bowhunting Gear
Bow Hunting	Glosses with "bow hunting"	Bow Hunting	Glosses with "bow hunting"
Coaching & Training	Glosses with "instruct", "coach", "train"	Glosses with "instruct", "coach", "train"	Glosses with "instruct", "coach", "train"
Traditional Archery - Kyudo	Glosses with "traditional"	Kyudo	Glosses with "traditional"

Regarding accuracy, the introduction of new nodes carried just one misclassified document, which had been misclassified already on of the third party indices. In general, the accuracy of this method should be as good as the accuracy of the classification of the participant taxonomies. However, we are concerned about the naming of the newly created nodes. In the Archery case above, all the names read nicely, but when we did Soccer, one node received the name "Instructing" when we would prefer to see "Instruction." Further work could be done on node naming

5. Conclusions

Editorially created document glosses are a boon to taxonomy alignment, in that they constitute a rich resource to guide semantic resemblance analysis, and have the added bonus, when soft ontology expansion is applied via WordNet, of enabling us to create new interstitial nodes for a more complete and unequivocal alignment of taxonomies.

References

Beneventano D., Bergamaschi S., Guerra, F., Vincini, M. 2003. Building an integrated Ontology within SEWASIE system. *Proceedings of the First International Workshop on Semantic Web and Data-bases (SWDB)*

Gaurino, N. 1998. Formal ontologies and information systems. *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS'98)*, Trento, Italy.

Leacock, Claudia, Martin Chodorow & George A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics,* 24(1): 147-165.

Martin, T. ., Ben Azvine, Behrad Assadian. 2004. Acquisition of Soft Taxonomies. *IPMU-04*, 1021-1032.