Remediating Intentionally Corrupted Ontology-Based Data Sets

Tim MUSGROVE [a,1] and Robin WALSH [a]
[a] *Callisto Media Lab*

Abstract. Many organizations have taken up the task of using an ontology to help align a folksonomy with a well-formed taxonomy. In one such approach, folksonomies (tagged by users) are taken as having detectable ontological implications, however poorly formed. The folksonomies are then used as a sort of reference corpora, to test and validate a formal ontology. When highly attested tag co-occurrences in the folksonomy are found to be unrelated in the formal ontology, it implies that the formal ontology should be examined for possible revision or expansion. For example, if my formal ontology fails to relate "mayor" to "politician", but the folksonomy commonly applies both those tags to the same entities, then it points to a place where my formal ontology needs improvement. This has been a useful approach for various projects in the information sciences. However, this approach breaks down when bad actors intentionally infect the folksonomy with inappropriate tags on a wide basis. Unfortunately, since many folksonomies are on the Web, where there is either a profit motive or a political motive to manipulate the semantic network of tags, we have found it necessary to develop special handling of "bad actors" intentionally corrupting the implicit ontology that is emergent within the folksonomy. We explain how a systematic detection of these bad actors, together with a programmatic pruning of the ontological clues gleaned from the folksonomy, can not only be achieved at scale, but also can be subjected to an ontological "integrity test", so that the improvement in the associated ontology can be demonstrated and validated.

Keywords. data corruption, bad actors, ontological distance, knowledge discovery, ontology mapping, semantic similarity, book publishing, ontology revision, cosine similarity, middle-level ontology

1. Introduction

Callisto Media is the fastest-growing publisher of non-fiction books in North America and is unique in being an entirely data-driven publisher. We use formal ontologies to filter and organize many hundreds of millions of data points from around the Web, in order to compile outlines of subject matter that would constitute a book meeting a consumer informational need. An integral part of that process is the use of ontologies. We have ontologies for the general domain of making and publishing books, as well as for the audiences of books, and for numerous non-fiction topic domains. We have

[1] Timothy Musgrove, PhD., Research and Development, Callisto Media Publishing, 6005 Shellmound Street, Suite 175, Emeryville, CA, USA. E-mail: tmusgrove@callistomedia.com.

mapped our internal working ontologies to various "folksonomies" on the World Wide Web, including that of Amazon.

Amazon has a complex taxonomy of many thousands of nodes, with free cross-referencing of nodes, and further allows millions of product owners to self-declare their own nodes, including that they can multiply classify any product into as many nodes as they want. Further, these same vendors can freely add a plurality of metatag keywords to each item in each node. The result is a massive folksonomy, created by countless Amazon vendors, in concert with Amazon's internal taxonomy designers.

Unfortunately, today some vendors are intentionally seeding the system with false information. For example, on Amazon, product vendors are motivated to apply irrelevant metatags to their product to engage in "black-hat" SEO (search engine optimization), as well as classifying their product in an irrelevant taxonomical category on Amazon to wrongfully win a "bestseller" flag that their product's sales rank would never merit in its proper taxonomy node. We focus on the latter phenomenon, and how to detect the intentional misclassifications, pruning away the misleading cross-references that would erroneously be created otherwise.

The initial endeavor that prompted this study was to determine which topics and phrases were most strongly associated with the most successful books on a given subject matter, as indicated by the titles, subtitles, and front- and back-matter descriptions of those books, together yielding a plurality of self-declared PERTAINS-TO relations for an ontology. This could be informative as to what related topics might appear to help books perform well, but also, which among those related topics are very "saturated" among the top-performing books in the given subject matter (meaning they no longer provide a strong differentiator of one book among others). Product specification is often the search for a medium amount of match between the new product and successful competing products. In the case of a book, one needs to know both how one will be similar, but also a bit different, from the way other books approach the same subject.

Toward this end, Callisto has fashioned the Callisto Content Ontology, a proprietary multi-level ontology built partly upon relations in WordNet such as the hyponym/hypernym relation, similar to the OntoWordNet project[1], but in this case with significant modifications and additions made by members of Callisto's R&D team. For example, "ketogenic" does not appear in WordNet, and thus was added manually to our ontology, along the following lines (rendered here in Cognitum's Controlled English[2], for readability):

> Every diet is a subject-matter.
> Every specialized-diet is a diet.
> Every ketogenic-diet is a specialized-diet.
> Every health-outcome is a subject-matter.
> Weight-Loss is a health-outcome.
> Every diet has-purpose health-outcome.
> Every syndrome-management is a health-outcome.
> Epilepsy-Management is a syndrome-management.
> Every ketogenic-diet has-purpose Weight-Loss.
> Every ketogenic-diet has-purpose epilepsy-management.
> Atkins-Diet is a ketogenic-diet.

This ontology is intended to be supplemented by the folksonomy of Amazon, in that content words with strong statistical characteristics among the cohort of books for a given

Amazon node act as a "bootstrap" of our ontology expansion process. But instead, we are now experimenting with using the ontology itself to detect maleficent infection of the folksonomy, as is explained below. Our present view is that only after purging the folksonomy of corruption, will it be safe to use it programmatically to grow our working ontology.

For an example of how the folksonomy (when all goes well) helps accelerate ontological discovery, consider the concept of something being "vegeterranean" (belonging at once to vegetarian and Mediterranean cuisine). Our internal ontologists were, at one point, not aware of this concept's existence, but became aware of it because of how our systems interact with the Amazon folksonomy. When one of the very first "vegeterranean" cookbooks had a page put up on Amazon for pre-sale (before the book actually existed), the page was classified folksonomically as both vegetarian and Mediterranean. The portmanteau "vegeterranean" promptly showed up on one of our internal reports, came under review, and the concept was added to our ontology forthwith.

The salient part of the example is that the value of this process for our organization comes from ***early*** detection of a new concept. We have an ontologist's version of the "time-to-market" challenge; our goal is to be among the earliest to enrich and expand the ontology of non-fiction book subject matter, so that our editors always have an accurate view of the complete range of non-fiction book opportunities in the market. This means that a sample size of one is sufficient for our system to flag a new ontological concept for human attention.

Naturally then, we were very alarmed to notice vendors' bad habit of misclassifying their books, watching it become more and more common on Amazon. The practice threatens to break down entirely the benefit we had aimed at originally, of building a bridge between the Amazon folksonomy and our own formal ontology construction process. Therefore, eliminating irrelevant book classifications has become a critical need.

2. Construction of the Sample Cohort of Books

We focused on non-fiction instructional books, in English, available in the US market, in many dozens of categories, including cooking, psychology, language learning, business, medical conditions, crafts and hobbies, sports training, and many more. The book cohort for our project was assembled by taking 100 applicable book titles that sell well (sales rank on Amazon better than 25,000), including 50 such books that we deemed bad actors and 50 legitimate books as annotated by our SMEs (subject matter experts). Bad actors are those that appeared to our SMEs to have very low production quality but nonetheless a disproportionately high number of positive reviews.

3. Understanding the Nature of Intentionally False Category Labels

Once it had become clear to us that the purpose of mapping our ontology to the Amazon folksonomy would be thwarted by a significant number of false category labels given to products on Amazon, we naturally looked to understand the manifest causes of the phenomenon. For example, one Italian cookbook was flying the "bestseller" flag because its publisher had classified it as a "personal travel guide" despite that it had nothing to with travel. From our sample corpus, approximately 64% of the bad actors had the appearance of misapplication of an Amazon category, as judged by our internal human

annotation, while only about 6% of the legitimate books did. To make clear how this can happen and what the obvious motive is, consider the following table, showing the Amazon sales rank of the top twelve books, as of this writing, in two category nodes:

Table 1. The sales rank (relative to all 17+ million English books on Amazon) of the twelve bestselling books in "Indian Cookbooks" and "Birdwatching"

Sales Ranks of Amazon's Top Selling Books	Indian Cookbooks	Birdwatching
1	31	8,477
2	2,563	57,710
3	2,584	63,782
4	4,023	63,987
5	5,825	73,669
6	7,445	76,216
7	9,166	77,566
8	12,116	80,407
9	13,209	87,909
10	13,519	137,783
11	15,855	141,393
12	20,833	142,420

Now suppose you are the publisher of the 5th bestselling Indian Cookbook. With a sales rank of 5,825, your book is nowhere near the top-selling Indian cookbook that has an impossible-to-beat sales rank of 31. Clearly your Indian cookbook will never get the coveted "bestseller" flag on its Amazon product page. However, notice that if you added a classification of your cookbook as "Birdwatching", where the bestselling book ranks at only 8,477, you would instantly win a bestseller tag with your book's sales rank of 5,825.

This is just a specific example of how some highly coveted tags can tempt people to "game" the system. The bestseller flag on Amazon is powerful. It shows up at the top of the product page, and shows immediately in all search results where that product appears. By contrast, the additional superfluous classification – revealing that the Indian cookbook is a #1 seller only within "birdwatching" – is very hard for users to notice. It is "below the fold," i.e., low down on the product page. And it does not show up at all on search results pages. Therefore, if you add the "birdwatching" classification to your 5th-ranking Indian cookbook, most consumers will perceive only that Amazon is flagging the book as a "bestseller" and they will not notice how the book actually "earned" this flag.

Many "bad actors" seem to believe this is an effective approach to boosting the sales of a book on Amazon. It has become a common practice, and it now populates the Amazon ontology with myriad intentional misclassifications. However, this examination supplies the clues we need to build a programmatic detection (and removal) of such mischief from our mapping.

Although there is previous work on measuring the quality of tags in a folksonomy, these center mostly around the usefulness of tags introduced by users[3], rather than getting at the intentional misapplication of tags already established. Thus, we deemed it necessary to devise our own methods of detecting intentionally false tag applications.

4. Two Approaches to Detecting Bad Category Classifications

We decided to try two approaches to detecting if a classification seemed out-of-place: an extensional approach that depends on most of the competitive books to be in similar categories, and an intensional approach that expected the title and subtitle of a book to have significant semantic similarity to the node name in which it is placed.

4.1. An Extensional Approach to Detecting False Category Labels

By an extensional approach we mean, as in linguistics, to validate the meaning of a class label by an examination of the members of that class, including their shared membership in multiple other classes. Our extensional approach could be called "graph-based", in that we measured the ontological distance between the most common category to books found on the Amazon SERP (search engine results page) for the primary keyphrase associated with a title, and any self-claimed category in which that title was evidently vying for a bestseller tag. When the distance was deemed too great, we flagged the claimed category as dubious.

More specifically, to determine the primary keyphrase associated with a title, we took the keyphrase that the title was most strongly converting from (producing sales from people searching for that keyphrase). To determine if a title was materially aiming at a bestseller tag, we checked if it ranked in the top 10 on the bestseller list for the associated category on Amazon. These two checks yielded two categories for us to compare distance between.

There are a few points of this measure that are less than obvious. Most books are multiply classified, without it being wrong or irrelevant to do so. For example, a weightlifting book is classified not only in Weight Training but also in Sports Training. Intuitively, these two nodes should be close. But in Amazon's folksonomy, they are not all that close. In fact, their step-wise ontological distance is five, as illustrated here:

- Books
    - Health, Fitness & Dieting
        - Exercise & Fitness
            - Weight Training
    - Sports & Outdoors
        - Sports Training

But generally, a distance of five is a signal of irrelevance. Consider that Islam and Orthodontics also have a step-wise distance of five, within Amazon's folksonomy:

- Books
    - Medical Books
        - Dentistry

        - Orthodontics
    - Religion & Spirituality
        - Islam

Because of this, we do not use the Amazon folksonomy directly for our measure. Instead we use the Callisto Books Taxonomy, which maps multiple Amazon nodes to our own category schema. The Callisto Taxonomy places weight training and sports training at a distance of only two, while placing Orthodontics a distance of six from Islam. Without the mapping between the folksonomy and our more well-formed ontology, we could not perform a worthwhile ontological distance calculation.

Even then, it is not so easy to determine the "maximum acceptable ontological distance" for two nodes to be deemed relevant to one another. To solve this, we had to employ the concept of a middle-level ontology preference, familiar in cognitive linguistics.[4] Our ontologists marked, throughout our hierarchical subject matter organization, nodes counting as the "middle level" – often the second or third level (where the total hierarchy is typically four or five levels deep). If the shortest path between two nodes requires a crossing of this "middle threshold", then the two nodes are deemed disparate, and each one an ontological alien to the other one.

For example, "Food & Beverage" is deemed a middle node in the Callisto Taxonomy. The only way to get from a Cooking book to a Parenting book is to pass up above Food & Beverage and come back down the tree to Parenting. Thus, cookbooks and parenting books are deemed unrelated.

Still there are problems, in that authors find ways to combine almost any two topics into a book. A book on "teaching your children to cook" could well be classified in both Parenting and Cooking. Thus, the extensional method, even with all our enhancements, is not foolproof by itself.

4.2. An Intensional Approach to Detecting False Category Labels

By an intensional approach we mean, as in linguistics, an explication of the semantics of a word or phrase in relation to other words or phrases that connect with it in a semantic network or matrix. Specifically, we combined two intensional methods, taking the best (highest) score of the two as our final intensional semantic relatedness score:

1. In intensional score 1, we measured the proportion of content words in a suspect category name that appear in titles/sub-titles on any books in the competitive cohort (including itself). This check used lemmatization and breaking up of compound words (e.g. to match "cooking" to "cookbook").
2. In intensional score 2, we determined the shortest WordNet (hypernym-tree or coordinate term) distance between any term in the claimed category name and any term in the primary keyphrase associated with that title (as determined by the same method explained above in our extensional measure). This is similar to other WordNet-based semantic distance measures in the literature[5,6].

Similar to our extensional approach, our intensional method was mostly correct, but not foolproof. This is largely because of the habit in publishing of giving books allusive titles that are not at all descriptive of their true subject matter. For example, one of the most popular books in "Motivational Self Help", as of this writing, has the title (and subtitle), "Unf*ck Yourself: How to Get Out of Your Head and Into Your Life". After discarding from this all of the non-content words and words not found in the dictionary,

the only words left are “head” and “life”, neither of which have a strong cosine similarity in our implementation to either of “motivational” or “self-help”. So, the intensional method wrongly deems this an irrelevant classification of the book. (This could possibly be corrected by including the full product description, or even reviews of the book, but such steps have been deferred to a later phase of this project.)

4.3. Comparison of Extensional and Intensional Approaches to Detecting False Category Labels

The system is sometimes wrong on one or the other approach, but so far, is seldom wrong when both approaches indicate a classification as dubious. A comparison of precision and recall of both methods individually and conjoined:

Table 1. Recall and precision of false category detection, depending on whether the extensional and intensional methods are employed singly, disjointly, or conjointly

Method	Recall	Precision	F-Measure
Extensional method	98%	74%	84%
Intensional method	67%	94%	78%
Conjunctive combination	67%	96%	79%

5. Conclusion and Best Practices Recommendations

Since the precision is very high when both intensional and extensional measures indicate a dubious classification, we now deem it safe to automatically prune such cases. For those detected as a dubious classification only by one or the other measure, we queue the case for human review. This combination of machine determination and human judgement enables a manageable workflow for the ongoing detection and elimination of false product categorizations.

6. Areas for Further Study

There are likely to be many other signals of intentional false categorization. For example, the average sales rank of a book in a falsely categorized node tends to be much better than its sales rank for its proper category (the one truly relevant to the book in question). Also, anecdotally, we see that independent publishers and others not on our internal “whitelist” of high-quality publishers tend to be more often guilty of the phenomenon of irrelevant categorization. These and similar signals will be programmatically included in our next upgrade to the system.

References

[1] Aldo Gangemi, Roberto Navigli, Paola Velardi. “The OntoWordNet Project: extension and axiomatization of conceptual relations in WordNet”, 2001. www.w3.org/2001/sw/BestPractices/WNET/ODBASE-OWN.pdf

[2] Cognitum Tehcnology Blog, 2014-05. http://techblog.cognitum.eu/2014/05/how-to-build-ontology-model-for.html

[3] Van Damme, C., Hepp, M., & Coenen, T. (2008). Quality metrics for tags of broad folksonomies. In Proceedings of International Conference on Semantic Systems (I-SEMANTICS) (pp. 118-125) https://www.academia.edu/376040/Quality_Metrics_for_Tags_of_Broad_Folksonomies

[4] Friedrich Ungerer, Hans-Jorg Scmid. *Introduction to Cognitive Linguistics* (Routledge, 2006), p.67.

[5] Budanitsky, A., & Hirst, G. (2001, June). Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and other lexical resources* (Vol. 2, pp. 2-2).

[6] Maki, W. S., McKinley, L. N., & Thompson, A. G. (2004). Semantic distance norms computed from an electronic dictionary (WordNet). *Behavior Research Methods, Instruments, & Computers*, *36*(3), 421-431.