Self-Driving Product Understanding for Thousands of Categories

XIN LUNA DONG, AMAZON

8/2020
Product Graph

- Mission: To answer any question about products and related knowledge in the world
Knowledge Graph Example for 2 Songs

[Diagram showing a knowledge graph with nodes and relationships for two songs: "Shake it off" and "Love Story"]

- **Node mid345** with properties:
  - **name**: "Shake it off"
  - **genre**: "Country pop"

- **Node mid346** with properties:
  - **name**: "Love Story"

- **Node mid127** with properties:
  - **name**: "Taylor Alison Swift"

- **Node mid128** with properties:
  - **name**: "Taylor Swift"

- **Node mid129** with properties:
  - **name**: "Dance-pop"
  - **birth_date**: 12/13/1989

- **Node "Pop"** with properties:
  - **genre**: "Pop"

- **Node "Dance-pop"** with properties:
  - **genre**: "Dance-pop"

- **Node "Taylor Alison Swift"** with properties:
  - **name**: "Taylor Alison Swift"

- **Node "Taylor Swift"** with properties:
  - **name**: "Taylor Swift"

- **Node "Country pop"** with properties:
  - **genre**: "Country pop"
Product Graph Example for 2 Songs

“Shake it off”
- name: mid345
- genre

“Love Story”
- name: mid346
- song_writer

“Pop”
- name

“Dance-pop”
- name

“Taylor Alison Swift”
- name

“Taylor Swift”
- birth_date

12/13/1989

“Country pop”
- type

Genre
Product Graph Example for 2 Songs
Product Graph for Media Products

- Main challenge: **Heterogeneous** Data everywhere
- Key techniques

- Support ~20 Amazon Music applications
Product Graph for Retail Products
Product Graph for Retail Products

- Main challenge: **Sparse & Noisy** Data everywhere
Scott's Cakes Dark Chocolate Toffee Cream Filling Candies with Dark Blue Foils in a 1 Pound Red Roses Box

Price: $19.95 + $14.95 shipping
You can get 5% back on all Amazon.com purchases with the Amazon Prime Store Card. No annual fee.

Note: Not eligible for Amazon Prime.

In Stock. Ships from and sold by Scott's Cakes.

38 Flavors: 1 lb. Red Rose Box
- 1 lb. Christmas... $19.95
- 1 lb. Daisy Box $19.95
- 1 lb. Red Rose Box $19.95

19 Sizes: Dark Blue Foils
- Aqua Foils $19.95
- Burgandy Foils $19.95
- Dark Blue Foils $19.95

Get It Thu, Aug 22 - Tue, Aug 27
Get It Tue, Aug 20 - Fri, Aug 23 if you choose paid shipping at checkout.

Deliver to Yang - Seattle 98109
Qty: 1

Turn on 1-click ordering
Product Graph for Retail Products

- Main challenge: **Sparse & Noisy** Data everywhere
- Key techniques

- Live on Alexa Shopping, Detail Page, and Amazon Search
How to Scale Up?
Scale Up in 3 Dimensions

Focus of this talk

One vertical, A few sources

Hierarchy of 100K categories

 Millions of sources

Behavior info from tens of applications over years

Big challenge: Limited training labels for large-scale, rich data
Challenges for 100K Categories (I)

- Require 100K+ models
- Different value vocabularies
- Different patterns to express the same attribute
Challenges for 100K Categories (II)

- Different attributes apply to different product categories
- A lot of work to manually define ontology
- Hard to catch the trend of new product categories and properties
Challenges for 100K Categories (III)

- It is hard to build a product taxonomy
- What are product types in customers’ language?
- How to organize types into a taxonomy tree?

Sellers’ view

- Kitchen & Dinning
  - Cutlery & Knife Accessories
    - Chef's Knives
    - Bread knife
    - ...
  - Sports & Fitness
    - ...
    - Hunting
      - Knife Sets
      - Hunting Knives

Buyers’ view

- Knife
  - Kitchen Knife
    - Chef's Knife
    - Bread knife
    - ...
  - Outdoor Knife
    - Hunting Knife
    - Survival Knife
A 100-Year Project
Our Solution: **Self-Driving**
Product Understanding
AutoKnow: Self-Driving Product Knowledge Collection

<table>
<thead>
<tr>
<th>Product</th>
<th>Type</th>
<th>Flavor</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product 1</td>
<td>Snacks</td>
<td>Cherry</td>
<td></td>
</tr>
<tr>
<td>Product 2</td>
<td>Candy</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Product 3</td>
<td>Candy</td>
<td>Choc.</td>
<td>Gold</td>
</tr>
</tbody>
</table>

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

Input Data
- PT Taxonomy
- Catalog
- Behavioral Signals (e.g., search logs, reviews, Q&A)

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

Input Data
- PT Taxonomy
- Catalog
- Behavioral Signals (e.g., search logs, reviews, Q&A)

Ontology Suite
- Taxonomy Enrichment
- Relation Discovery

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

Input Data
- PT Taxonomy
- Catalog
- Behavioral Signals (e.g., search logs, reviews, Q&A)

Ontology Suite
- Taxonomy Enrichment
- Relation Discovery

Data Suite
- Data Imputation
- Data Cleaning
- Synonym Discovery

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
Self-Driving to Navigate 100K Categories

- **Automatic:** Fully ML-based
- **Annotation free:** Weak learning based on existing Catalog data and user behavior
- **One-size-fits-all:** Few taxonomy-aware models
- **Self guidance:** Identify important attributes and categories to focus efforts

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
I. GNN-Based Taxonomy Enrichment

- Step 1. Type extraction from product titles and queries by OpenTag tagging system

Figure 2: OpenTag Architecture: BiLSTM-CRF with Attention.
### I. GNN-Based Taxonomy Enrichment

- Consumables taxonomy **size increases by 2.9X**

<table>
<thead>
<tr>
<th>Source</th>
<th>Text</th>
<th>Product Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product</td>
<td>4 Country Pasta Homemade Style Egg Pasta - 16-oz bag</td>
<td>Egg Pasta</td>
</tr>
<tr>
<td>Product</td>
<td>Hamburger Helper Lasagna Pasta, Four Cheese, 10.3 Ounce (Pack of 6)</td>
<td>Lasagna Pasta</td>
</tr>
<tr>
<td>Product</td>
<td>COFFEE MATE The Original Powder Coffee Creamer 35.3 Oz. Canister Non-dairy, Lactose Free, Gluten Free Creamer</td>
<td>Coffee Creamer</td>
</tr>
<tr>
<td>Query</td>
<td>mccormick paprika 8.5 ounce</td>
<td>paprika</td>
</tr>
<tr>
<td>Query</td>
<td>flax seeds raw</td>
<td>flax seeds</td>
</tr>
</tbody>
</table>

Mao et al., OCTET: Online catalog taxonomy enrichment with self-supervision, SigKDD, 2020.
I. GNN-Based Taxonomy Enrichment

- Step 2. Taxonomy expansion with hypernym discovery by Graph Neural Network

Mao et al., OCTET: Online catalog taxonomy enrichment with self-supervision, SigKDD, 2020.
I. GNN-Based Taxonomy Enrichment

- Step 2. Taxonomy expansion with hypernym discovery by Graph Neural Network

Mao et al., OCTET: Online catalog taxonomy enrichment with self-supervision, SigKDD, 2020.
I. GNN-Based Taxonomy Enrichment

- Improve hypernym identification F1 by 18% over state-of-the-art

<table>
<thead>
<tr>
<th>Child Type</th>
<th>Parent Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coconut flour</td>
<td>Baking flours &amp; meals</td>
</tr>
<tr>
<td>Tilapia</td>
<td>Fresh fish</td>
</tr>
<tr>
<td>Fresh cut carnations</td>
<td>Fresh cut flowers</td>
</tr>
<tr>
<td>Bock beers</td>
<td>Lager &amp; pilsner beers</td>
</tr>
<tr>
<td>Pinto beans</td>
<td>Dried beans</td>
</tr>
</tbody>
</table>

Mao et al., OCTET: Online catalog taxonomy enrichment with self-supervision, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

- **Input Data**
  - PT Taxonomy
  - Catalog
  - Behavioral Signals (e.g., search logs, reviews, Q&A)

- **Ontology Suite**
  - Taxonomy Enrichment
  - Relation Discovery

- **Data Suite**
  - Data Imputation
  - Data Cleaning
  - Synonym Discovery

- **Broad Graph**
  - Ontology
  - {product, attribute, value}
  - {value, synonym, value}

---

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
### II. Behavior-Based Attribute Importance

<table>
<thead>
<tr>
<th>Child safety car seats</th>
<th>Almond</th>
</tr>
</thead>
<tbody>
<tr>
<td>brand</td>
<td>flavor</td>
</tr>
<tr>
<td>style</td>
<td>brand</td>
</tr>
<tr>
<td>size or capacity</td>
<td>ingredients</td>
</tr>
<tr>
<td>maximum weight recommendation</td>
<td>size or capacity</td>
</tr>
<tr>
<td>color</td>
<td>number of items</td>
</tr>
<tr>
<td>model</td>
<td>energy content per serving</td>
</tr>
<tr>
<td>number of items</td>
<td>bar code</td>
</tr>
<tr>
<td>usage</td>
<td>color</td>
</tr>
<tr>
<td>bar code</td>
<td>scent</td>
</tr>
<tr>
<td>flavor</td>
<td>usage</td>
</tr>
<tr>
<td>ingredients</td>
<td>style</td>
</tr>
<tr>
<td>scent</td>
<td>model</td>
</tr>
<tr>
<td>energy content per serving</td>
<td>maximum weight recommendation</td>
</tr>
</tbody>
</table>

**Important (ranked by score)**

**Unimportant**

**Unapplicable**

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
II. Behavior-Based Attribute Importance

Instances: \{ITK, attribute\} pairs
Labels: importance scores

Extract Signals
- coverage
- refinement
- Text Mining
  - title
  - queries
  - keywords
  - ......

Aggregate each signal at 3 levels
- Category
- Sub-category
- ITK

Predict importance using regression
- Random forest

Collect ground truth labels
- Mech. Turk

ITK = item type keyword

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
II. Behavior-Based Attribute Importance

Leveraging various behaviors

ITK = item type keyword

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
II. Behavior-Based Attribute Importance

- Predict for **7.5K categories** the applicability of **3K attributes** with $P=0.94/R=0.84$, and importance with 0.74 correlation (Spearman)

**Table 15: Attributes identified as most important for two example types.**

<table>
<thead>
<tr>
<th>Cereals</th>
<th>Shampoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>brand</td>
<td>brand</td>
</tr>
<tr>
<td>ingredients</td>
<td>hair type</td>
</tr>
<tr>
<td>flavor</td>
<td>number of items</td>
</tr>
<tr>
<td>number of items</td>
<td>ingredients</td>
</tr>
<tr>
<td>energy content</td>
<td>liquid volume</td>
</tr>
</tbody>
</table>
AutoKnow: Self-Driving Product Knowledge Collection

Input Data
- PT Taxonomy
- Catalog
- Behavioral Signals (e.g., search logs, reviews, Q&A)

Ontology Suite
- Taxonomy Enrichment
- Relation Discovery

Data Suite
- Data Imputation
- Data Cleaning
- Synonym Discovery

Broad Graph
- Ontology
- {product, attribute, value}
- {value, synonym, value}
III. Value Extraction—Traditional

**OpenTag**

- **CRF**
  Captures correlations between BIOE tags
- **Attention**
  Identifies important terms leading to attribute values
- **Bi-LSTM**
  Captures sequence info
- **Word Embedding**
  Captures semantics of each token

Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.
III. Value Extraction—100K Categories

Option 1. Train a single model? \textit{Train/Test Distribution shift -> Invalid predictions}

- Samsung UN58RU7100FXZA Flat 58-Inch 4K UHD 7 Series Ultra HD Smart TV with HDR and Alexa Compatibility (2019 Model)
- Taylors of Harrogate Classic Tea Variety Box, 48 Count (Pack of 1)
- Caribou Coffee Caribou Blend, Medium Roast Ground Coffee, 20 Ounce Bag, Rainforest Alliance Certified
III. Value Extraction—100K Categories

Option 1. Train a single model?  
*Train/Test Distribution shift -> Invalid predictions*

- Samsung UN58RU7100FXZA Flat 58-Inch 4K UHD 7 Series Ultra HD Smart TV with HDR and Alexa Compatibility (2019 Model)
- Taylors of Harrogate Classic Tea Variety Box, 48 Count (Pack of 1)
- Caribou Coffee Caribou Blend, Medium Roast Ground Coffee, 20 Ounce Bag, Rainforest Alliance Certified

Option 2. Train a model for each category?

- Store/orchestrate 100K+ OpenTag models
- Most categories are very sparse

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
III. Value Extraction—Taxonomy-Aware

Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
III. Value Extraction—Taxonomy-Aware

Figure 2: Our TXtract architecture for hierarchical multi-task learning.
III. Value Extraction—Taxonomy-Aware

Figure 2: Our TXtract architecture for hierarchical multi-task learning.
III. Value Extraction—Taxonomy-Aware

Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Attention conditioned on category representation
III. Value Extraction—Taxonomy-Aware

- Train one model on 4K categories, and improve state-of-the-art by 10.4% in F1, and by 11.7% in coverage

<table>
<thead>
<tr>
<th>Title</th>
<th>OpenTag</th>
<th>TXtract</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Controlled Labs Purple Wraath 90 Servings - Purple Lemonade</td>
<td>flavor: -</td>
<td>flavor: purple lemonade</td>
</tr>
<tr>
<td>2 Click - Espresso Protein Drink Vanilla Latte - 16 oz.</td>
<td>flavor: espresso</td>
<td>flavor: vanilla latte</td>
</tr>
<tr>
<td>3 Mason Vitamins Melatonin 500 mcg Fast Meltz Tablets, Fruit, 60 Count</td>
<td>flavor: -</td>
<td>flavor: fruit</td>
</tr>
<tr>
<td>4 Fashion Glitter Matte Eye Shadow Powder Palette Single Shimmer Eyeshadow</td>
<td>scent: palette</td>
<td>scent: -</td>
</tr>
</tbody>
</table>
| 5 Baby car seat cover, Nursing covers Breastfeeding cover carseat canopy (Style5) | scent: style5        | scent: -             

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
AutoKnow: Self-Driving Product Knowledge Collection

Input Data
- PT Taxonomy
- Catalog
- Behavioral Signals (e.g., search logs, reviews, Q&A)

Ontology Suite
- Taxonomy Enrichment
- Relation Discovery

Data Suite
- Data Imputation
- Data Cleaning
- Synonym Discovery

Broad Graph
- Ontology
- {product, attribute, value}
- {value, synonym, value}
IV. Taxonomy-Aware Anomaly Detection

Wang et al., Automatic validation of textual attribute values in eCommerce Catalog by learning with limited labeled data, KDD’20
IV. Taxonomy-Aware Anomaly Detection

Is the flavor “Pink”?

Wang et al., Automatic validation of textual attribute values in eCommerce Catalog by learning with limited labeled data, KDD’20
IV. Taxonomy-Aware Anomaly Detection

Wang et al., Automatic validation of textual attribute values in eCommerce Catalog by learning with limited labeled data, KDD’20
IV. Taxonomy-Aware Anomaly Detection

Category as input for model training
IV. Taxonomy-Aware Anomaly Detection

- Identify **1.77MM** incorrect values for Flavor and Scent for Consumables with **90% precision**

<table>
<thead>
<tr>
<th>Product</th>
<th>Attr</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love of Candy Bulk Candy - Pink Mint Chocolate Lentils - 6lb Bag</td>
<td>Flavor</td>
<td>Pink</td>
</tr>
<tr>
<td>Scott's Cakes Dark Chocolate Fruit &amp; Nut Cream Filling Candies with Burgandy Foils in a 1 Pound Snowflake Box</td>
<td>Flavor</td>
<td>1 lb. snowflake box</td>
</tr>
<tr>
<td>Lucky Baby - Baby Blanket Envelope Swaddle Winter Wrap Coral Fleece Newborn Blanket Sleeper Infant Stroller Wrap Toddlers Baby Sleeping Bag (color 1)</td>
<td>Flavor</td>
<td>color 1</td>
</tr>
<tr>
<td>ASUTRA Himalayan Sea Salt Body Scrub Exfoliator + Body Brush (Vitamin C), 12 oz</td>
<td>Scent</td>
<td>vitamin c body scrub - 12oz &amp; body brush</td>
</tr>
<tr>
<td>Folgers Simply Smooth Ground Coffee, 2 Count (Medium Roast), 31.1 Ounce</td>
<td>Scent</td>
<td>2Packages (Breakfast Blend, 31.1 oz)</td>
</tr>
</tbody>
</table>
AutoKnow: Self-Driving Product Knowledge Collection
V. Synonym Discovery

- **Identify 7.6K synonym pairs for flavor and scent with 92.6% precision**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type synonyms</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Phrase synonyms</strong></td>
<td>cr2032 battery : lithium coin / toys 6 months : engage baby / minnie dress : short cocktail / mtm 50 : 50 caliber / coq10 200mg : energizing antioxidant</td>
</tr>
</tbody>
</table>
V. Synonym Discovery

- Identify **7.6K** synonym pairs for flavor and scent with 92.6% precision

<table>
<thead>
<tr>
<th>Type synonyms</th>
<th>Examples</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Attribute synonyms</th>
<th>Examples</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Phrase synonyms</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>cr2032 battery : lithium coin / toys 6 months : engage baby / minnie dress : short cocktail / mtm 50 : 50 caliber / coq10 200mg : energizing antioxidant</td>
<td></td>
</tr>
</tbody>
</table>
### Key Techniques in AutoKnow

<table>
<thead>
<tr>
<th>Techniques</th>
<th>AK-Taxonomy</th>
<th>AK-Relations</th>
<th>AK-Imputation</th>
<th>AK-Cleaning</th>
<th>AK-Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graph structure</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taxonomy signal</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Distant supervision</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Behavior information</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Take Aways

- We need to scale up for 100K categories for building an authoritative product knowledge graph.
- We are investigating techniques for self-driving end-to-end product understanding.
- Graph Neural Network and Learning w. limited labels play an important role in various aspects.
Thank You!