Zero to One Billion: The Journey to a Rich Product Graph

XIN LUNA DONG
AMAZON
6/2021
Knowledge Graph Example for 2 Songs

- **Entity**
  - mid345
  - name: "Shake it off"
  - genre: Country pop
  - artist: Taylor Alison Swift

- **Entity**
  - mid346
  - name: "Love Story"
  - genre: Dance-pop
  - artist: Taylor Swift

- **Relationship**
  - mid345 - artist - mid127
  - mid346 - song_writer - mid128
  - mid127 - name - "Taylor Alison Swift"
  - mid128 - name - "Taylor Swift"
  - mid129 - birth_date - 12/13/1989
  - mid129 - type - Country pop
Product Graph Example for 2 Products
Product Graph

- Mission: To answer any question about products and related knowledge in the world
Use Case I: Providing Information

Brand: Cetaphil
Ingredients: Water, Cetyl Alcohol, Propylene Glycol, Iodopropynyl Butylcarbamate, 2-Bromo-2-Nitropropane-1, 3-Diol, Sodium Lauryl Sulfate, Stearyl Alcohol, Methylparaben, Propylparaben, Sodium Citrate, Butylparaben, Allantoin, Zinc Gluconate.

Scent: Fragrance free
Additional Item Information: Non-Comedogenic, Fragrance-free, Natural

Skin Type: Sensitive

About this item:
- Gentle for everyday use; Cetaphil gentle skin cleansing cloths will leave your skin feeling clean, refreshed and balanced after every use.
- Removes makeup & dirt: Thoroughly remove makeup and dirt, leaving skin clean.
- Mild & non-irritating: Soap free formulation won’t strip skin of its natural protective oils and emollients.
Use Case II: Providing Choices
Use Case III: Improving Search
Use Case III: Improving Search
Use Case III: Improving Search
Use Case IV: Improving Recommendation
Product Graph vs. Knowledge Graph

(A) Generic KG

(B) Generic KG

(C) Generic KG

PG
Knowledge Graph vs. Product Graph

(A) Generic KG

(B) Generic KG

(C) Generic KG

Product Graph

(Hardline, softline, consumables, etc.)

Movie, Music, Book, etc.
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

<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>

- #Types ↑ 3X
- Defect rate ↓ up to 68 percent points

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
But, Is The Problem Harder?
Challenges in Building Product Graph I

Sparse and noisy structured data

Scott's Cakes Dark Chocolate Toffee Cream Filling Candies with Dark Blue Foils in a 1 Pound Red Roses Box
by Scott's Cakes
Be the first to review this item
Challenges in Building Product Graph II

- Extremely complex domains
- How to identify the millions of product types?
- How to organize types into a taxonomy tree?
Challenges in Building Product Graph III

- Big variety across product types
- Different attributes apply to different product types
- Different value vocabularies and different patterns
Challenges in Building Product Graph III

- Big variety across product types
- Different attributes apply to different product types
- Different value vocabularies and different patterns

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
Scale Up in 3 Dimensions

- Thousands of attributes
- Millions of categories
- Hundreds of languages

Big challenge: Limited training labels for large-scale, rich data
A 100-Year Project
Deliver the Data Business

1,000,000,000,000
Deliver the Data Business

High precision models
Deliver the Data Business

1,000

High precision models + E2E pipeline + AutoML to reduce modeling cost
Deliver the Data Business

1,000,000

High precision models
E2E pipeline + AutoML to reduce modeling cost

Scale-up to reduce #models

100s attributes
1000s categories
10s languages
Deliver the Data Business

High precision models

E2E pipeline + AutoML to reduce modeling cost

Scale-up to reduce #models

1000s categories
100s attributes
10s languages

Higher yield from multi-modal models
From 0 to 1: High-Quality Data
From Zero to One: A Core Algorithm

1

High precision models
## OpenTag Extraction from Product Profiles

<table>
<thead>
<tr>
<th>name</th>
<th>form</th>
<th>scent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tide Detergent with Febreze Freshness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain Apple Mango Tango Liquid Laundry Detergent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gain Joyful Expression Powder Detergent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tide PODS Original Scent HE Turbo Laundry Detergent Pacs 81-load Tub</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tide PODS Free &amp; Gentle HE Turbo Laundry Detergent Pacs 35-load Bag</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
OpenTag Extraction from Product Profiles

\[ x = \{w_1, w_2, ..., w_n\} \text{ input sequence} \]
\[ y = \{t_1, t_2, ..., t_n\} \text{ tagging decision} \]

{beef meal} \rightarrow \text{Flavor Extractions} \rightarrow \{ranch raise lamb\}
OpenTag Extraction from Product Profiles

Word Embedding

Bi-LSTM

Attention

CRF
OpenTag Extraction from Product Profiles

BiLSTM+CRF+Attention obtains best results

Extraction on new values is comparable to already known values

Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.
OpenTag in Practice

Understand domain and attributes, and generate LOTS OF training data

Train and fine-tune models

OpenTag

Postprocess extraction results to further improve data quality

Pre-publish evaluation as gatekeeper to guarantee high quality data
Still the Origin Point

- Millions of categories
- Thousands of attributes
- Hundreds of languages
From 1 to 1K: Reducing Modeling Cost
From 1 to 1K: E2E AutoML Pipeline

High precision models + E2E pipeline + AutoML to reduce modeling cost
An End-to-End Pipeline

Understand domain and attributes, and generate LOTS OF training data → Train and fine-tune models → Postprocess extraction results to further improve data quality → Pre-publish evaluation as gatekeeper to guarantee high quality data
An End-to-End Pipeline

- **Train and fine-tune models**
- **Postprocess extraction results to further improve data quality**
- **Pre-publish evaluation as gatekeeper to guarantee high quality data**

**Automatic Training Data Generation**
- Distant supervision
- Data programming

**OpenTag**

**Benchmarking**
An End-to-End Pipeline

- **Automatic Training Data Generation**
  - Distant supervision, Data programming

- **OpenTag**
  - Train and fine-tune models

- **Deep Learning Data Cleaning**
  - Postprocess extraction results to further improve data quality

- Pre-publish evaluation as gatekeeper to guarantee high quality data
An End-to-End Pipeline

Automatic Training Data Generation

Train and fine-tune models

OpenTag

Deep Learning Data Cleaning

Postprocess extraction results to further improve data quality

Benchmarking

Scale-up pre-publish evaluation w. lower labeling needs

Distant supervision, Data programming
An End-to-End Pipeline

AutoML

Automatic Training Data Generation
- Distant supervision,
- Data programming

OpenTag

Deep Learning Data Cleaning

Benchmarking

Scale-up pre-publish evaluation w. lower labeling needs
Transformer-Based Anomaly Detection

Love of Candy Bulk Candy - Pink Mint Chocolate Lentils - 6lb Bag
Brand: Love of Candy
Price: $84.99 ($0.96 / Ounce) + $16.92 shipping
Pay $14.17/month for 6 months, Interest-free with your Amazon Prime Rewards Visa Card
Flavor Name: Pink
Blue, Green, Orange, Pastel Assortment, Pink, Red
Size: 6 Pound

- Love of Candy’s huge selection of bulk candy now includes Premium Mint Chocolate Lentils in a variety of bold & striking colors. Available in small to large sizes ranging from 1 to 10 lb bags. These beautiful chocolate morsels feature gourmet, dairy-free dark chocolate coated in a crispy and crunchy mint candy shell. Similar to M&M’s, these mint chocolate candy lentils are fun, bite-sized snacks that can be enjoyed during any occasion.
- Sourced from the most esteemed candy makers from around the world, we’ve put together an extremely broad collection of wholesale candy to fulfill your every need. Whether you’re in need of candy for vending machines, pillboxes or candy buffets, you can trust that Love of Candy’s got you covered. Our consistent product quality and unmatched customer satisfaction have quickly made Love of Candy the market’s most trusted source of high-quality, wholesale bulk candy.
Transformer-Based 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
Transformer-Based Anomaly Detection
Transformer-Based Anomaly Detection

Category as input for model training
Transformer-Based 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>
From 1K to 1M: Scaling Up
From 1K to 1M: One Size Fits All

High precision models

E2E pipeline + AutoML to reduce modeling cost

Scale-up to reduce #models

1000s categories

10s languages

100s attributes
Scale up for Millions of Categories

Millions of categories

Thousands of attributes

Hundreds of languages
Scale-up I: Millions of 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
Scale-up I: Millions of Categories

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

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.
Scale-up I: Millions of Categories

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.
Scale-up I: Millions of Categories

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

Attention conditioned on category representation
Scale-up I: Millions of Categories

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

Attention conditioned on category representation

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
Scale-up I: Millions of Categories

- 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>
<tr>
<td>5 Baby car seat cover, Nursing covers Breastfeeding cover carseat canopy (Style5)</td>
<td>scent: style5</td>
<td>scent: -</td>
</tr>
</tbody>
</table>

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
Scale up for Thousands of Attributes

- Millions of categories
- Thousands of attributes
- Hundreds of languages
Scale-up II: Thousands of Attributes

First Aid Beauty Ultra Repair Cream: Vegan and Gluten-Free Intense Moisturizer for Dry Sensitive Skin. Perfect for Skin Conditions and Eczema. Pink Grapefruit (14 ounce)

About this item
- HEAD-TO-TOE: Head-to-toe moisturizer that provides instant relief and long-term hydration for dry, distressed skin, even eczema. The beautiful, whipped texture is instantly absorbed with no greasy after-feel. Grapefruit has a bright citrus fruit scent that is fresh, juicy and sparkling.
- CLINICALLY PROVEN: Formulated with Colloidal Oatmeal, Shea Butter, Ceramide 3 and the FAB Antioxidant Booster, it provides immediate relief and visible improvement for parched skin and it is clinically proven to increase hydration by 169% immediately upon application.

Product description
Banish dry skin with First Aid Beauty’s Ultra Repair Cream. Suitable for all skin types, especially dry, flaky skin, this hydration wonder leaves skin feeling smooth, hydrated and comfortable after just a single use.

Mentioned Attributes: Brand, SkinType, Scent, Quantity
Scale-up II: Thousands of Attributes

Scale-up II: Thousands of Attributes

Scale-up II: Thousands of Attributes

MoE (Mixture-of-Experts)
• Question: Attribute -> Layer Weight (Matrix)
  • Use a weighted combination of basis matrices to generate the weight.

Scale-up II: Thousands of Attributes

HyperNet (Hyper Network)

• Question: Attribute -> Layer Weight (Matrix)
  • Use a network to generate the weight of another network.
Scale-up II: Thousands of Attributes

- Train one model for 32 attributes, obtaining higher quality than single-attribute models


**Diagram:**
- Higher F1 than one model per attribute
- One model for multiple attributes
From 1M to 1B: Increasing the Yield
From 1M to 1B: Multi-Modal Extraction

High precision models

E2E pipeline + AutoML to reduce modeling cost

Scale-up to reduce #models

1000s categories

100s attributes

10s languages

Higher yield from multi-modal models

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
Multi-Modal Signals

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
Multi-Modal Extraction

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
Multi-Modal Extraction

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
Multi-Modal Extraction

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
Multi-Modal Extraction

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
Multi-Modal Extraction

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
## Multi-Modal Extraction

- Improve over OpenTag (text-only) by 11% on F-measure

<table>
<thead>
<tr>
<th>Models</th>
<th>P(%)</th>
<th>R(%)</th>
<th>F1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM w/o text</td>
<td>79.9</td>
<td>63.4</td>
<td>70.7</td>
</tr>
<tr>
<td>PAM w/o image</td>
<td>88.7</td>
<td>72.1</td>
<td>79.5</td>
</tr>
<tr>
<td>PAM w/o OCR</td>
<td>82.0</td>
<td>69.4</td>
<td>75.1</td>
</tr>
<tr>
<td>PAM</td>
<td>91.3</td>
<td>75.3</td>
<td>82.5</td>
</tr>
</tbody>
</table>

Text still plays the most important role

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
### Successful vs. Not-Yet-Successful Fields in Industry

<table>
<thead>
<tr>
<th>SUCCESS</th>
<th>NOT-YET SUCCESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Entity linkage</td>
<td>- Schema mapping</td>
</tr>
<tr>
<td>- Knowledge extraction (ClosedIE)</td>
<td>- Small scale: Manual</td>
</tr>
<tr>
<td>- Knowledge cleaning</td>
<td>- Large scale: “Replaced” by ClosedIE</td>
</tr>
<tr>
<td>- Knowledge-based QA</td>
<td>- OpenIE</td>
</tr>
<tr>
<td></td>
<td>- “Replaced” by Reading comprehension</td>
</tr>
<tr>
<td></td>
<td>- Knowledge fusion</td>
</tr>
<tr>
<td></td>
<td>- Not fully needed yet</td>
</tr>
<tr>
<td></td>
<td>- Knowledge inference</td>
</tr>
<tr>
<td></td>
<td>- Quality not high enough yet</td>
</tr>
</tbody>
</table>

1. One cannot live without it
2. The techniques are ready
Take Aways

- We are building an authoritative product knowledge graph for millions of categories, thousands of attributes, and hundreds of languages.

- High-accuracy modeling is the first step for building an authoritative knowledge graph.

- AutoML E2E pipelines, one-size-fits-all solutions, and multi-modal models are critical for enriching the knowledge.
Thank You!