Unsupervised Construction of a Product Knowledge Graph

Omar Alonso, Vasileios Kandylas, Rukmini Iyer
{omalonso|vakandyl|rukmini}@microsoft.com

Introduction

- Constructing a commercial knowledge graph data asset that revolves around brands, products, and categories
- Query for a brand (Microsoft) and retrieve associated products (Surface 3, Windows 10, Xbox) and to query for a product (jeans) and retrieve the associated brands (Calvin Klein, Hudson, Armani)
- Challenges:
    - Very dynamic domain (brands and products appear/disappear)
    - Lack of major sources with clean brand/product data
    - Hard to define and detect products
    - Distinction between brand and product is sometimes blurred
    - Retailers don't always provide clean data
- A *brand* is a term or phrase that distinguishes an organization or product (e.g., Adidas, Microsoft). A *domain* is the most common URL associated with the brand (e.g., Microsoft's domain is `microsoft.com`). A *product* is an item that is manufactured for sale (e.g., Microsoft Surface 3, Gucci Guilty,) or a service provided (e.g., insurance, pet cleaning). An *alias* is a name that an item is otherwise called or known as (Apple Inc., AAPL). *Categories* group items into a given label or name (e.g., BMW -> vehicles).
- Unsupervised approach: 1) brand generation, 2) tag brands with domains and categories, 3) product generation
- New data sources can be added without affecting the existing process
- Bottom-up strategy with a focus on simplicity and data cleaning
- Implementation in Scope/Cosmos

Brands and Brand Domains

- Use as input n number of sources; each source assigns 1 vote to the presence of a term in their respective sources
- Compute voting above a threshold as final score
- Use a combination of Satori data and Bing search query logs to extract domain information

Term	Alias	Domain	score
Apple	Apple Inc.	`apple.com`	1
Bose	Bose audio	`bose.com`	0.7
Gap	The Gap Inc.	`gap.com`, `gapinc.com`	0.8

Categorization

- Approach using the following sequence: brand -> brand URL -> search queries to URLs -> query categories -> aggregation of categories -> categories
- Use Bing search query logs to get query-URL information
- Use categorizer to get query categories
- Probability of a query given URL $P(Q|U) = \frac{clicks(Q,U)}{clicks(U)}$
- Probability of category given brand $P(C|B) = \sum_U \left[\sum_Q P(C|Q) \cdot P(Q|U) \right] P(U|B)$
- Final score depends on number of queries and clicks available for brand

$$Score(Q|B) = max\left(0, P(Q|B) - \frac{\alpha_1}{\sqrt{clicks(B)}} - \frac{\alpha_2}{\sqrt{queries(B)}}\right)$$

- Brand popularity $P(B) = clicks(B) / \sum_b (clicks(b)$

Products

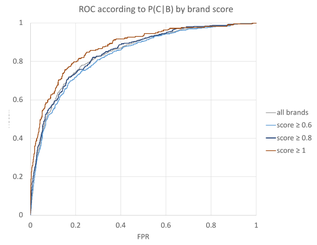
- Products from Retail Catalogs
    - Grouping of products within each brand
    - CDSSM trained on product ads to embed product names
    - K-means with heuristics to split clusters of larger size
    - Generate a set of M representative labels for each product cluster
- Products from Bidded Keywords
    - Bidded keywords are matched to brands by brand name and domain
    - Generate n-grams from the keywords selected
    - Score on brand-specific language model and ranked using KL divergence

Brand	Bidded keyword products	Retailer catalog-based products
microsoft	microsoft word	Microsoft Office Home & Student 2016 For Mac
microsoft	microsoft office	Microsoft Office 365 Personal Subscription 1 Year 1 User Pc/Mac
microsoft	microsoft office 365	Microsoft Office Professional Plus 2016 365 Pro Word Excel Key Windows
microsoft	microsoft outlook	Microsoft Office Home and Business 2016 Retail License
microsoft	microsoft surface	Microsoft 13.5" Surface Book 2-in-1 Notebook 256GB SSD 8GB RAM Core i5 2.4 GHz
microsoft	microsoft teams	12.3 Surface Pro 6 - 128GB / Intel Core m3 / 4GB RAM (Silver)
microsoft	microsoft project	Surface Pro 6 - 512GB / Intel Core i7 / 16GB RAM (Platinum)
microsoft	microsoft excel	Surface Pro 4 12.3" Bundle: Core i5, 4GB RAM, 128GB SSD, Surface Pen, Type Cover
microsoft	microsoft office 2016	Microsoft Project Professional 2016 - Digital Download
microsoft	microsoft powerpoint	Microsoft Project 2019 Professional w/ 1 Server CAL Open License
apple	apple watch	Apple Watch Series 3 44MM Rose Gold
apple	apple airpods	Apple Watch Series 4 44 mm Gold Stainless Steel Case with Gold Milanese Loop
apple	apple watch series 3	24K Gold Plated 42MM Apple Watch Series 3 Diamond Polished Modern Gold Link Band
apple	apple watch series 4	Apple Airpods MMEF2J/A Wireless Earphone For Iphone/Apple Watch/Ipad/Mac
apple	apple iphone	Apple Airpods Genuine Left-Only Airpod (Without Charging Case)
apple	apple watch 4	Apple Airpods - White MMEF2AM/A Genuine Airpod Retail Box
apple	apple ipad	Apple iPhone 8 Plus - 64GB - Space Gray (Unlocked) A1864 (CDMA + GSM)
apple	apple iphone xs max	Apple iPhone XR 256GB, Yellow - Sprint
apple	apple tv	Apple iPad Wi-Fi 128GB Silver
apple	apple earpods	Apple iPad 6th Gen. 32GB, Wi-Fi + Cellular (Unlocked), 9.7in - Silver #15718

Results and Evaluation

- Statistics: 130K brands, 1M products, and 22 top-level categories
- Human evaluation

Brand score	Precision
1	92.2%
0.83	87.8%
0.66	87.2%
0.5	85.6%



Conclusion

Presented an efficient and scalable brand-product graph generation data pipeline

bing

Microsoft