SPECIFICATION AND IDENTIFICATION OF RELATIONSHIPS BETWEEN PRODUCTS IN THE FOOD SEGMENT

Sian Brumm, Rolf Krieger, Christoph Brosch

Umwelt-Campus Birkenfeld | HOCHSCHULE TRIER

Motivation

- Retail companies manage data for hundreds of thousands or millions of products.
- Product data is utilized in various functional areas, e.g. sales, puchasing, etc.
- Knowledge of product relationships is essential, but manually maintaining these relationships is prone to errors.

Usage of product relationships:

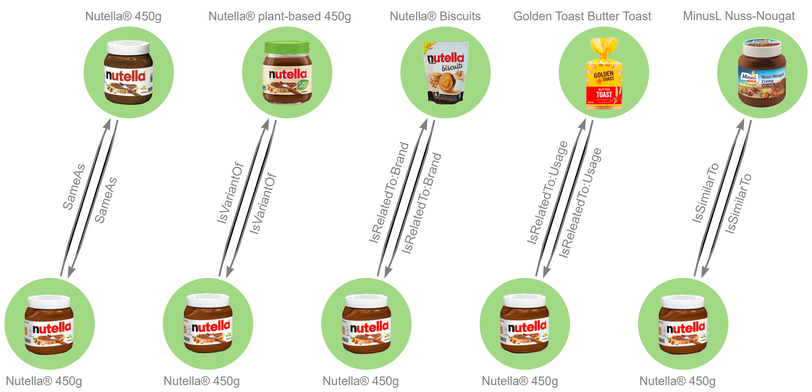
- Product recommendations (e.g., accessories, upselling opportunities).
- Conducting a competitive pricing analysis requires identifying product relationships.
- Support of process automation and error correction by suggesting attribute values from similar products in master data management.
- Acquisition of new knowledge for the construction of prodcut graphs

Related Work

- Christen, P.: Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer Berlin Heidelberg, 2012
- Primpeli, A., Peeters, R., Bizer, C.: The WDC Training Dataset and Gold Standard for Large-Scale Product Matching. In: Companion Proceedings of The 2019 World Wide Web Conference, San Francisco USA, ACM, 2019, pp. 381–386
- Peeters, R., Bizer, C.: Entity Matching using Large Language Models, 2024
- Schema.org: Homepage. https://schema.org/, retrieved on 07.09.2024

Specification of Product Relationships

Our specification of product relationships is based on the following properties: brand, type, content, packaging, refill bag, and use. DiffPackaging is introduced for specializing the product relationships SameAs, IsVariantOf and IsSimilarTo. DiffPackaging is used if products only differ in their packaging.



Data

Data about food products were collected by crawling German online shops.

Attributes	Product 1 (Shop A)	Product 2 (Shop B)
Image		
Name	Nutella® 450g	Nutella® plant-based 450g
Description	Das Produkt der Marke Nutella ist im 450g-Behälter verfügbar. Mit dem süßen Aufstrich starten Sie geschmackvoll in den Tag...	Nutella Plant-Based ist da. Das unverwechselbare Nutella-Erlebnis mit Zutaten pflanzlichen Ursprungs statt Milch...
Categories	Brot, Cerealien & Aufstriche / Süße Aufstriche / Nuss- & Schokoaufstriche	Startseite / Lebensmittel / Brot & Frühstück / Aufstriche
Brand	Nutella	Nutella

Three datasets were created to analyse the impact of the size of the training dataset on model performance:

- Each dataset contains procuct pairs labeled with SameAs, IsVariantOf, IsSimilarTo, NotSpecified
- The largest dataset contains 15,253 pairs and is based on 21,245 unique products from 67 GPC classes.
- Most frequent categories: alcoholic beverages (17%), sweets (11%), non-alcoholic ready-to-drink beverages (10%), herbs/spices/extracts (5%), and sauces/spreads/dips/seasoning sauces (5%).

Model Development

A procedure was developed for the automated determination and classification of the product relationships. It consists of three steps:

1. Data Preparation

- Removal of special characters and stop words
- Segmentation of product names and extraction of attributes, e.g. brand, packaging size, and packaging volume

2. Blocking

- Products are grouped by their **Type** and **Brand**.
- An Approximate Nearest Neighbor-algorithm calculates k-best candidates.

3. Multi-Class Classification of Product Relationships

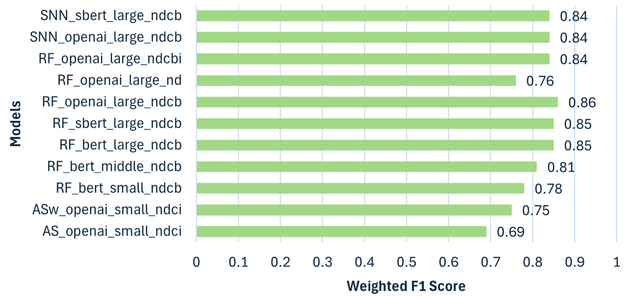
- **Approach 1: Classification based on weighted cosine similarity of attributes (AS):**
  - Input: vector representations of product attributes, e.g. name, description, categories, and image.
  - Thresholds are used for classifying the relationships.
- **Approach 2: ML-based Classification:**
  - Input: Concatenated vector representations of text and image data.
  - Models: Random Forest (RF) and Siamese Neural Network (SNN)
- **Classification of sub-relationships**:
  - **DiffPackaging** and **IsConsumableFor** are determined by rule-based methods.
  - **IsRelatedTo:Brand** is determined by analyzing products of the same brand but having different types.
  - **IsRelatedTo:Usage:** Determination of ingredients often used together in recipes.

Experiments

Analysis of the impact of different parameters on the performance of the multi-class classification model for product relationships:

- Approaches: AS, ASw (weighted attributes), SNN, RF
- Datasets with product pairs: small (6,800), middle (10,751), large (15,253)
- Embeddings: bert (bert-base-german-uncased), sbert (distiluse-base-multilingual-cased-v2), openai (text-embedding-3-small)
- Attributes: n=name, d=description, c=categories, b=brand, i=image

Performance of the classification model for product relationships



Results:

- Enlarging the size of the training data improves the classification performance by 0.07 with respect to the F1 score.
- The model using the Openai embeddings combined with the largest dataset and all text attributes achieved the best classification performance with an F1 score of 0.86.
- SNNs achieved slightly weaker results compared to the RF models.

Future Work

- Improvement of the accuracy and the generalizability of our models by enlarging the training dataset and incorporating additional attributes, such as ingredient lists.
- Analysis of the overall performance of the process, in particular by taking into account the reduction of the candidate set through the blocking procedure.
- Research for utilising large language models for more effective classification of defined product relationships.

Funded by: Bundesministerium für Bildung und Forschung as part of the funding program KI4KMU, FKZ 01|S23060
Project coordinator: retailsolutions GmbH
Project partner: Hochschule Trier, Umwelt-Campus Birkenfeld
Project duration: 01.10.2023 -30.09.2025

Contact:
Prof. Dr. Rolf Krieger, Hochschule Trier, Umwelt-Campus Birkenfeld, 55761 Birkenfeld, Email r.krieger@umwelt-campus.de

Andreas Mohr, retailsolutions GmbH, Otto-Kaiser-Straße 4, 66386 St. Ingbert, Email andreas.mohr@retailsolutions.de

Bundesministerium für Bildung und Forschung