ProMap: Datasets for Product Mapping in E-Commerce
Project Type: Implementation

Omkar Metri, Saptarshi Mondal, Shubham Sanjay Darekar, Sparsh Navneet Prabhakar

metri@usc.edu, saptarsh@usc.edu, sdarekar@usc.edu, sprabhak@usc.edu

Introduction

Product mapping (PM), is a crucial process in e-commerce that involves aligning identical products across various online stores. Each product is characterized by diverse graphical and textual data, making it a valuable tool for general marketplace analysis and price comparison.

The challenge in product mapping arises from the absence of a universal product identification system across websites, necessitating the training of models to measure similarity based on textual and image data. Unfortunately, existing freely available datasets often fall short, lacking comprehensive product information and predominantly featuring distant non-matching pairs.

To address this gap, the paper introduced new freely available datasets for product matching, specifically ProMapCz for Czech and ProMapEn for English. Through web scraping, detailed product information was collected. The goal of product mapping is to create new dataset to efficiently predict, whether two listings from two different e-shops describe the same products.

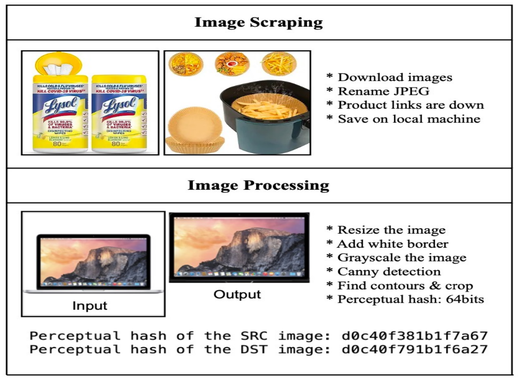
Aim

Replicate methods and technologies outlined in the paper to achieve end-to-end results for the ProMapEn dataset.

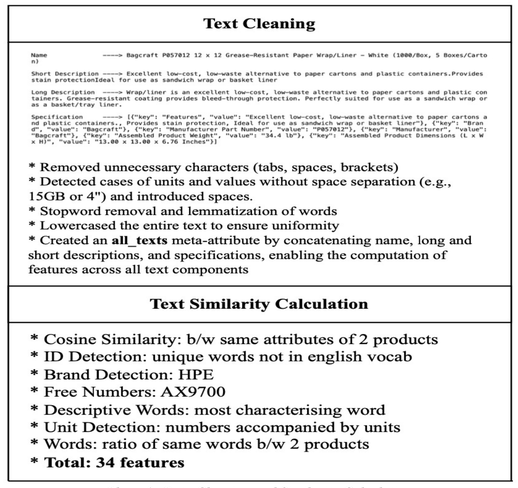
1. Image Scraping and Processing: Collect approximately 10-12k images from detailed information.
2. Data Cleaning and Feature Extraction: Perform data cleaning procedures and extract features by computing similarities among attributes of product pairs.
3. Model Building and Dataset Evaluation: Construct models for English product mapping. Evaluate the credibility and reliability of the ProMapEn dataset.
4. Model Training and Testing: Train the model using the ProMapEn feature set and test the trained model on Amazon-Google, Amazon-Walmart, and ProMapCZ datasets.
5. Comparative Analysis: Compare results across datasets to draw conclusions

Dataset: ProMapEn consists of 1,555 matching and non-matching pairs from different categories for training models for English product mapping from Walmart and Amazon.

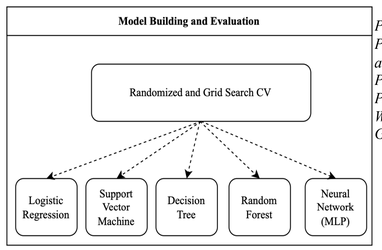
Method



Phase 1: Image Processing, Perceptual hash and hash similarity between images



Phase 2: Data Cleaning and Similarity Calculation



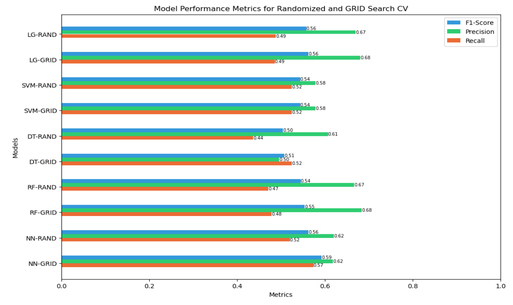
Phase 3: Trained on ProMapEn dataset and evaluated on ProMapEn, ProMapCz, Amazon-Walmart, Amazon-Google datasets

Results

- ProMapEn: 1,244 vectors for training and 311 for testing
- **Features:** Overall similarity of two products is characterized by vector of 34 features.
- **Labels:** Matching or non-matching products
- Utilized both grid search and random search techniques, with 20% validation set, to identify optimal hyperparameters
- Conducted hyperparameter tuning for linear regression, support vector machines, decision trees, random forests and neural network models.
- Model selection was based on maximizing the F1 score
- Best results were obtained by neural network-based models followed by logistic regression and random forests
- Neural networks have the most balanced precision and recall, while the random forests have larger differences

Model	Parameter	Possible Values
LogisticReg.	penalty	l1, l2, elasticnet, none
	solver	lbfgs, newton-cg, liblinear sag, saga
	max_iter	10, 20, 50, 100, 200, 500
SVM	kernel	linear, poly, rbf, sigmoid
	degree	2, 3, 4, 5
	max_iter	10, 20, 50, 100, 200, 500
DecisionTree	criterion	gini, entropy
	max_depth	5, 10, 15, 20
	min_samples_split	2, 5, 10, 15, 20
RandomForest	n_estimators	50, 100, 200, 500
	criterion	gini, entropy
	max_depth	5, 10, 20, 50
	min_samples_split	2, 5, 10, 20
NeuralNetwork	hidden_layer_sizes	(10, 10), (50, 50), (10, 50) (10, 10, 10), (50, 50, 50) (50, 10, 50), (10, 50, 10)
	activation	relu, logistic, tanh
	solver	adam, sgd, lbfgs
	learning_rate	constant, invscaling, adaptive
	learning_rate_init	0.01, 0.001, 0.0001
	max_iter	50, 100, 500

Parameter settings for the random and grid searches



Comparison of ML models trained and evaluated on ProMapEn dataset. Results are from models with best parameters from random and grid searches

	F1 Score	Precision	Recall
ProMapEn Test	0.659341	0.7407	0.5941
ProMapCz Test	0.583658	0.4717	0.7653
Am-Wm Test	0.535032	0.9767	0.3684
Am-Go Test	0.560440	1.0000	0.3893

Best model (MLP Classifier) trained on ProMapEn and evaluated on all datasets

- Transfer learning capabilities were tested by selecting the optimal model of ProMapEn and evaluating it on all other datasets.
- Missing attribute values from other test sets were filled with zeroes to address differences in attributes across datasets
- Overall, lower F1 metrics can be due to fewer IDs in the product descriptions of the English e-shops, different selection of products and categories or the need for further finetuning of preprocessing and hyperparameters.

Conclusion

This paper presents a comprehensive approach to generate English product mapping dataset through image and data scraping and evaluate the credibility by data cleaning, feature extraction, model building and dataset evaluation.

The study extends to comparative analysis across Amazon-Google, Amazon-Walmart and ProMapCZ datasets to draw insightful conclusions on the dataset's credibility and reliability. Overall results of models for Amazon-Walmart and Amazon-Google datasets are much higher than for ProMapEn datasets confirming that ProMapEn datasets are more challenging. As part of further research, focus will be on expanding English datasets for cross-e-shop product comparisons, emphasizing enhanced data preprocessing and advanced feature analysis.

Acknowledgements

- Thank you, Prof Dani Yogatama, TAs, CPs for guidance and feedback
- Kateřina Macková, Martin Pilát for wonderful paper
- Paper Link: https://arxiv.org/pdf/2309.06882.pdf