H.I.D.R.A.: A Hierarchical, Interactive and Dynamic Recognition Architecture for Product Categorization

Renato Cordeiro[1,2], **Ígor Bonadio**[1,2],
Vinicius Resende[1], **Beatriz Marouelli**[1], **Junior Koch**[1], **Henrique Oliveira**[1],
Leandro Fadelli[1], **Helton Alponti**[1], **André Formento**[1], **Rodrigo Vedovato**[1]

[1] Elo7 Research Lab

[2]Instituto de Matemática e Estatística – Universidade de São Paulo (USP)

{renato.cordeiro,igor.bonadio}@elo7.com

{vinicius.resende,beatriz.marouelli}@elo7.com

{junior.koch,henrique.oliveira}@elo7.com

{leandro.fadelli,helton.alponti}@elo7.com

{andre.formento,rodrigo.vedovato}@elo7.com

{renatocf,igorbonadio}@alumni.usp.br

Abstract. *The Hierarchical, Interactive and Dynamic Recognition Architecture (H.I.D.R.A.) for Product Categorization is a new intelligent system architecture developed by Elo7 to easily evolve its category tree and automatically classify millions of products, thus improving the page ranking of our marketplace.*

1. Introduction

Brazil has a strong community of artisans, and handmade products that generate income for more than 8.5 million people. Elo7 is the largest online handcrafts marketplace in Brazil, with more than 7 million announced products divided in circa 800 categories and produced by more than 100,000 active sellers living in 4,000 cities across the country.

Due to the COVID-19 pandemic, fabric masks became a nation-wide top-seller product. To meet this demand, Elo7 created a new category for these products. A simple model based on user-defined rules was applied to find out if a new product is a mask and put it in a new category. Thanks to the improved classification, this category reached a high page ranking in search providers, thus increasing the number of buyers.

Unfortunately, Elo7's current categorization system has many limitations. Firstly, models based on user-defined rules require a lot of effort from curators. Creating rules is a manual process and more rules are required to describe complex categories. In Elo7, this is prohibitive because the curators expect to create a large number of categories (up to 10,000). One approach to solve this problem is to use machine learning algorithms to automate model creation. Secondly, the process of creating categories is complex since it involves changing Elo7's legacy monolith. To integrate this intelligent component in the system, it is necessary to strangle this functionality and create a new set of APIs that will allow curators to evolve the category tree.

To create a more robust solution, this paper presents the Hierarchical, Interactive and Dynamic Recognition Architecture (H.I.D.R.A.) for product categorization. For each category, our architecture proposes using a weak supervised model [Ratner et al. 2017] to generate a labeled training set and then to train a binary supervised model to classify products. To apply these models, our architecture uses reactive principles [Bonér 2016] and event streaming [Wampler 2019] to process new or updated products in near real time. Finally, the architecture includes three specialized microservices supporting different workloads to serve the front-end components that show data related to the categories, The goal is to allow Elo7 to easily evolve its category tree and automatically classify products so that the website gets a high page ranking.

2. Architecture

Figure 1 represents the Hierarchical, Interactive and Dynamic Recognition Architecture or H.I.D.R.A. The figure is divided into five sections to make it easier to understand each one of its components. Figure 1.1 shows the overall architecture with outputs that describe contracts between the components.

Figure 1.2 describes the CONTENT CURATION component. A team of curators uses an ontology manager to manipulate Elo7's category tree and to generate a RDF (Resource Description Framework) file describing it. For each category, they use an experimentation tool to create two artifacts: an query to select products that belong to the category based on data provided by sellers (title, tags, etc.); and a set of regular expressions used as heuristics to find whether a product should belong to the category (also using the product title). The initial dataset can be further refined via a curation tool that allows curators to visualize and remove products from the dataset.

Figure 1.3 describes the MODEL ORCHESTRATOR component. Each category has a corresponding categorizer model that predicts the probability of a product belonging to the category. The orchestrator is composed by a configurator that manages a distributed computing framework to run each categorizer. The categorizer receives a new or updated product via a event stream, and sends its results to models that classify in its subcategories. After the execution of every available categorizer, for each product the categorizer produces a list of all its categories and chooses the main category of the product.

Figure 1.4 details the MODEL TRAINER component. Each categorizer is a binary classification model. To find a suitable dataset to train it, an initial noisy set of positive examples is selected from a query defined by the curators. This dataset is initially refined using a curation tool. To minimize the number of products that the curators need to evaluate, a weak supervised learning technique is applied. This approach uses a label matrix, where each row represents a product and each column has a classification made by a labeling function. For Elo7's product categorization, labeling functions are created from from three sources: crowdsourcing votes, obtained via a crowdsourcing voting tool; heuristics, produced by curators as described above; and third party models, which are other simple models trained specifically for each category (such as image models). These labeling functions are applied on an augmented dataset that contains the products defined by curators and the remaining products from the website. In the end, a neural network is used to ensemble these classifications and generate two types of metrics.

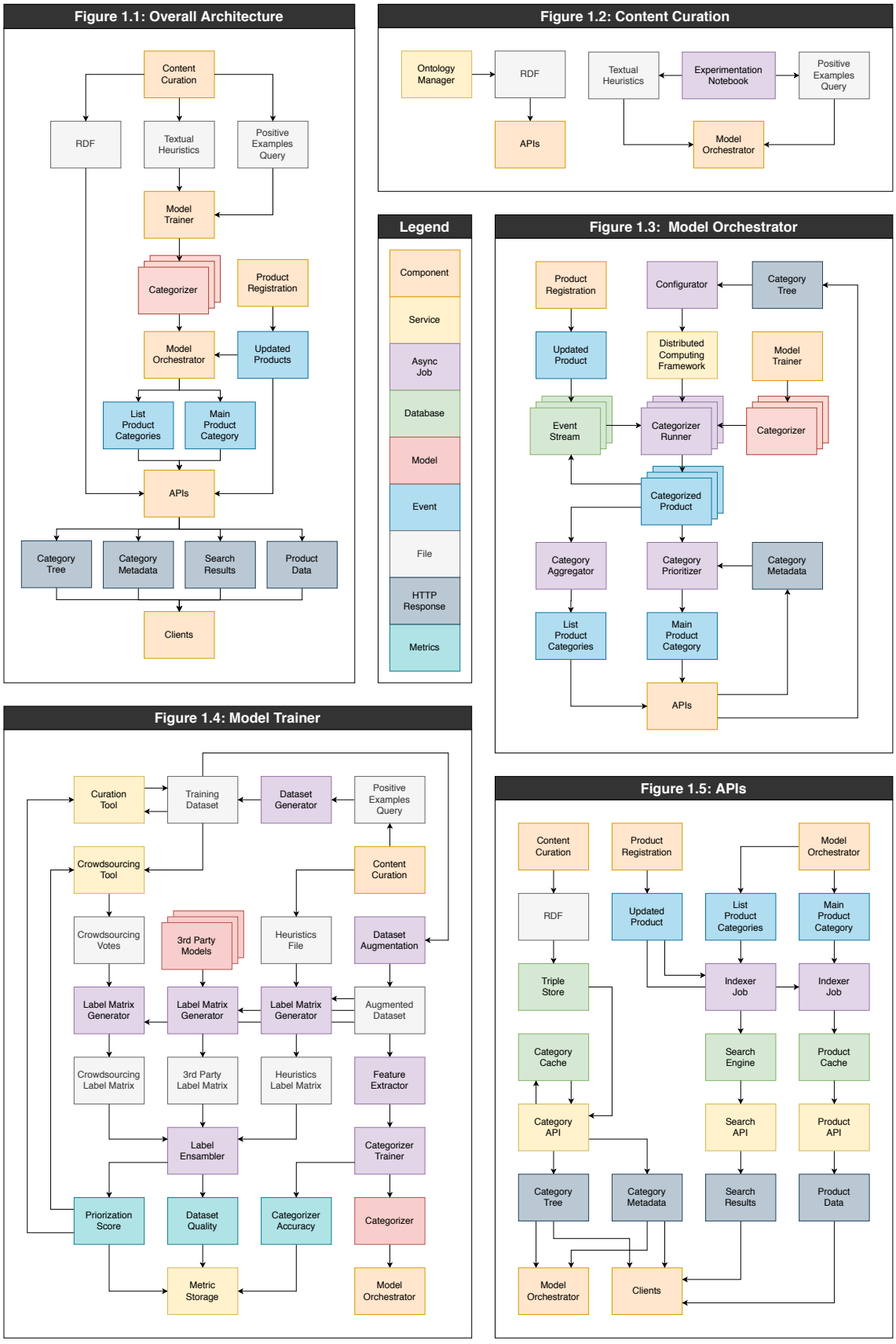

Figure 1. HIDRA: Hierarchical, Interactive and Dynamic Recognition Architecture

Firstly, it produces a priorization score, which is used as a feedback to the curation and crowdsourcing tools so that they can choose the best products to show to voters, i.e., products whose classification will help the most to increase the quality (decrease the uncertainty) of the label ensembler. Secondly, it produces a quality metric that helps to decide when to stop the process. Finally, when the final augmented training dataset is created, the categorizer is trained using a supervised learning technique. Its accuracy metrics determine wether the dataset needs to be further improved (because the model was not able to generalize) or if it can be used by the MODEL ORCHESTRATOR.

Figure 1.5 shows the CATEGORY SERVING component, which encompasses three services consumed by Elo7's client applications. The Category API serves the category tree exhibited in the sidebar of the website and other metadata used in the pages. It consumes the RDF File via a triple store (a type of graph database) and caches the queries results to speed up response times. The Search API lists products according to queries made in the search bar or clicks in a category from the sidebar. It uses a search engine that is asynchronously loaded with product data. Finally, the Product API return all data related to products, including its main category (used to create a breadcrumb) and metadata that should appear in the product visualization page. It uses a cache that joins data from the product registration with the main product category.

3. Development

Currently, the system is in its initial development stage. The CONTENT CURATION component is fully implemented and the team of curators is in the process of defining a new tree for a subset of high impact categories. The curators also defined the necessary data for two categories that are guiding the development of the other components. The CATEGORY LABELER component is fully developed and is in the final stage of testing. Its quality is being measured based on the sample categories chosen by the curators and the results are promising. The MODEL ORCHESTRATOR component still under discussion as its infrastructure and implementation details need further definitions. Finally, the CATEGORY SERVER component is in construction: the Category API was initiated whereas the Search and Product APIs will be worked on afterwards.

4. Conclusion

This paper presented the Hierarchical, Interactive and Dynamic Recognition Architecture (H.I.D.R.A.) for Product Categorization. The components described previously are currently being implemented to allow Elo7 to easily evolve its category tree and automatically classify products so that the website gets a high page ranking.

The system is expected to simplify the creation of new categories. The bottleneck will be the curation process while the classification of millions of products will be automatic. This is the most complex intelligent system ever implemented by Elo7. It is a cross-team effort involving many different skill sets, such as data science, machine learning engineering, software architecture, back-end and front-end software engineering.

In the upcoming months, the focus will be ending the development process, measuring its quality, and applying its capabilities to classify products in the high-impact categories defined by the curators. The project also allows other possible applications of the architecture described, including the definition of product attributes and the creation of seasonal marketing categories without manual product selection.

References

Bonér, J. (2016). Reactive Microservices Architecture. Technical report, Lightbend, Inc.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Wampler, D. (2019). Fast Data Architectures for Streaming Applications. Technical report, Lightbend, Inc.