KROGER TECHNOLOGY

ENTERPRISE ITEM STRATEGY

Pages / … / Data Quality in Product and Item Domains

DQ Algorithm Iterations

Created by SAMIRAN BASAK, last modified on Jul 14, 2023

@SHILADITYA CHAKRABORTY

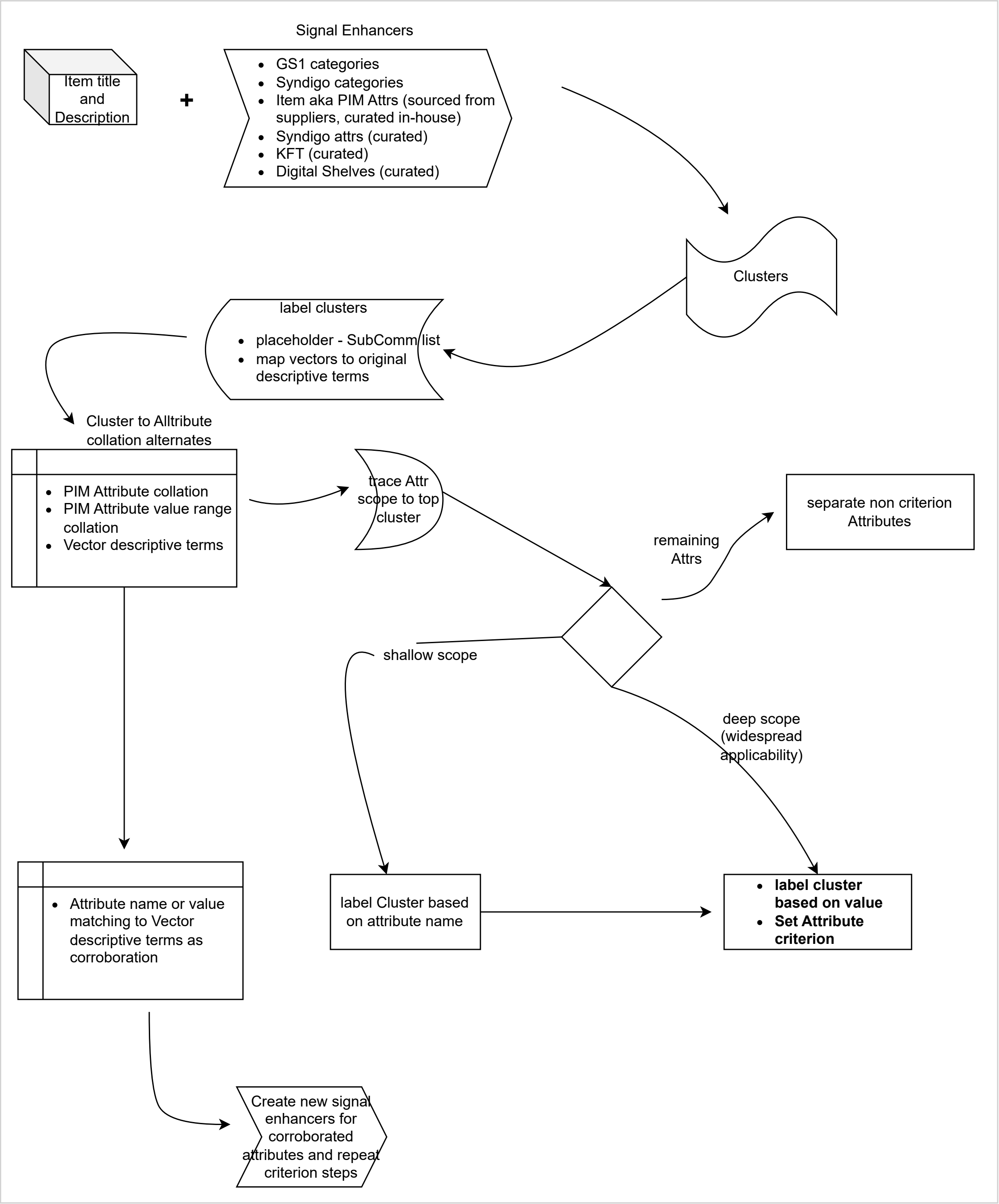
Objective	Detailed objective	Iterations & Rationale	Results	Follow ups
Classification into core taxonomy	**Algorithm selection**: Identify a selected small set of signals top iterate classification algorithms	Select 1) Product Title and description: (assumption) most product "customer facing" attributes are covered in description 2) KFT Sub comm: (assumption) this curated level should represent product type quite faithfully 3) leave out the other attributes - (assumptions) we may have doubts about attribute DQ and need selective sampling 4) Use one vertical Department eg Grocery - (assumption) lead to more homogeneity in product population, so that the algorithm itself could be better tuned than if the population was more widely assorted. Algorithms: consider clusters based on density rather than equal groupings based on mean distance 1. TBD algorithm with supervised taxonomy 2. BERTTopic - unsupervised clustering with hierarchical relationships between clusters (TBD binary splits only) a. no control on the size/population of clusters b. control the above	good convergence of clusters convergence becomes if groups under different parents are considered distinct based on their id (and not the same based on name) since "the misc transactions" are present in every department)	concerns: large size of population Action: remove all fees/taxes/donations/promotions Action: Run on a sample across the board and take stats on tree composition and convergence with KFT. (note leave out Deli, Pharmacy).
		Eliminate "Sub commodity" name as a signal (to eliminate the bias that might be leading to high convergence with KFT)	Done.	**Does the convince suffer in a big way?** **Action: Examine**
		Introduce more curated classification as signals (eg from GS1 and Syndigo)	@SHILADITYA CHAKRABORTY based on your recommendation today 6/30- other attributes are harder, but we can certainly use some curated txonomy	
		Introduce more attributes as signals (attributes, customer search)	Later	
		Tune Clustering with minimum Cluster size limitation (eg > 10) so that granular clusters (too much classification) is avoided		
		Explain- eg aberrations from KFT to determine positive or negative contribution of the algorithmic approach		**Action**: Examine some aberrations
		Verifications with Standards/ other licensed content providers		**Action**: with GS1, or Syndigo, top levels and lowest level convergence
	Measurements:			
	Filter all Misc Transactions which are not Products			

Objective	Detailed objective	Iterations & Rationale	Results	Follow ups
	Per Prime Dept clustering: Determine if vertical separation of products according to Depts leads to better classification per cohort	Introduce more departments (without Sub comms)	Done 6/27/23. @SHILADITYA CHAKRABORTY to share results	TBD Follow up small clusters, esp those with 1:1 with sub comm Follow up subcomms that are brand specific, or merchandizing specific (eg NFL) understand comms: 7505: SWIM/POOL CHEMICALS,PLUMBING REPAIR,WATER BOTTLES,GADGETS/TOOLS 297 SWIM/POOL CHEMICALS,SURFACE PROTECTION,BASIC OPENSTOCK,FLOTATION AND LIFE VESTS
	All Departments clustering: Representative sampling of all commodities to control processing resources. Note: not including Pharmacy, and Deli - which is a duplication/reflection of other Grocery)		Done 6/27/23. @SHILADITYA CHAKRABORTY to share results	TBD
	optimal tree config;			
	Optimally levelled (unbalanced) classification tree (depth relevant to variation, TBD splitting binary leads to unmanagemable depth)	tune n-way splits or recombine binary splits afterwards		Action: get the metrics first
	labelling the taxonomy			Action: prep the top keywords report for each cluster
	metrics	1. starting from the fourth level in core tax, three metrics for each group - for all gtin members: a. parent differences with KFT (can be iterated with Syndigo); next ancestor; next ancestor - which is the top level. i. Count number of distinct ancestors from KFT b. Plot three metrics above with their statistical metrics - min/max/avg/mode/median/standard deviation 2. Lowest level correspondence with subcommittee in KFT a. for each in lowest level core tax, compute i. number of distinct sub comms ii. number of non-exclusive subcomms (that extend across core tax clusters) b. for each in KFT Subcommittee, compute i. number of distinct lowest level core tax levels ii. number of non-exclusive lowest level core (that extend across sub comm		
Attribute criterion	Attribute relevance/completeness: identify attributes relevant to a cluster (hierarchy level) - split into those that have			Action: Collate all core attributes with each low level cluster, and analyse metrics: 1. populated attributes 2. same value attributes (indicates parent attributes

Objective	Detailed objective	Iterations & Rationale	Results	Follow ups
	common value in the group versus those that make a product unique (shirt size)			derived) 3. different values (distinguishing factor for product in cluster, might still include parent attributes) 4. range and distribution for different values
	Attribute derivation/generation from description/embeddings			
	Attribute value correctness: Range of values in attribute, and range that is relevant to products in group			
Product matching	other than correct values of unique 3P ids (including gtins) what attributes deternine the distinctness of a Product, or relate to other products			
	create new category in existing tree			
consumers	Search, Merchandizing			
Secondary Taxonomy production				
Higher order Product Association/ generation				
	Attribute generation: BI- or multi-lingual generation			

System Context Diagram

Algorithm flow


Item title and Description
+
Signal Enhancers
GS1 categories
Syndigo categories
Item aka PIM Attrs (sourced from suppliers, curated in-house)
Syndigo attrs (curated)
KFT (curated)
Digital Shelves (curated)
Clusters
label clusters
placeholder - SubComm list
map vectors to original descriptive terms
Cluster to Alltribute collation alternates
PIM Attribute collation
PIM Attribute value range collation
Vector descriptive terms
trace Attr scope to top cluster
remaining Attrs
separate non criterion Attributes
shallow scope
deep scope (widespread applicability)
Attribute name or value matching to Vector descriptive terms as corroboration
label Cluster based on attribute name
label cluster based on value
Set Attribute criterion
Create new signal enhancers for corroborated attributes and repeat criterion steps

Old notes:

Several Clustering iterations taken place:

unlimited clusters

set limit to 2000

set limit to 200

Suggested Metrics to eval each iteration:

1. Comparison of top level correspondence
   a. starting from the fourth level in core tax, three metrics for each group - for all gtin members:
      i. parent differences with KFT (can be iterated with Syndigo); next ancestor; next ancestor - which is the top level.
         1. Count number of distinct ancestors from KFT
      ii. Plot three metrics above with their statistical metrics - min/max/avg/mode/median/standard deviation
   b. Lowest level correspondence with subcommittee in KFT
      i. for each in lowest level core tax, compute
         1. number of distinct sub comms

        2. number of non-exclusive subcomms (that extend across core tax clusters)
    ii. for each in KFT Subcommittee, compute
        1. number of distinct lowest level core tax levels
        2. number of non-exclusive lowest level core (that extend across sub comms
c. Taxonomy structure
    i. Number of clusters (lowest level)
    ii. distribution (count, min, max, median) of path lengths, and count of members for each path, variation by top level cluster.
    iii. distrtbution ((count, min, max, median) of terms above a certain threshold for each cluster

No labels

2 Comments

DEEVRAJ NAIR

Thanks @SAMIRAN BASAK for the details. Have you also thought about governance of the core tax, if the lowest level has so many distinct clusters. Grocery alone had given us 5000+ lower level clusters when algorithm was executed without restriction. We might be able to bring it down, but considering the complete dataset we might have a huge volume of distinct clusters to deal with at the lowest level.. Please let me know your thoughts

SAMIRAN BASAK

Yes it's a challenge to manage! But do we have to manage, if the clusters are effective tools for attribute quality determination?