

Catalog Classification at the Long Tail using IR and ML

Neel Sundaresan (nsundaresan@ebay.com)

Team: Badrul Sarwar, Khash Rohanimanesh, JD
Ruvini, Karin Mauge, Dan Shen



Then There was One...



When asked if he understood that the laser pointer was broken, the buyer said "Of course, I'm a collector of broken laser pointers"

Divine Reward!



What we sell on a daily basis?



The Importance of Structured Information

- Search Experience
- Recommender Systems
- Fraud and Counterfeit Detection

Discovering Catalogs: Challenges

- Our goal is to build catalogs using
 - An unsupervised metadata extraction system
- Challenges
 - Huge volume of raw text
 - Highly unstructured
 - High level of noise
 - Lack of consistency/standardization of attribute name and value usage

Take Advantage of the Community

- Savvy sellers provide plenty of useful information
 - We need to combine techniques that can
 - Extract attribute names and values from this large collection
 - Remove noise and normalize attribute names and value usage

We have the data



Brand New GUCCI Sunglasses 2597/F/S 584 SHINY BLACK

Item condition: **New**

Ended: Apr 03, 2010 17:54:08 PDT

Bid history: **0 bids**

Starting bid: **US \$109.99**

or

Price: **US \$129.00**

Shipping: **\$29.00** USPS Express Mail International | [See all details](#)
Estimated delivery time varies for items shipped from an international location.

Returns: 7 day exchange, buyer pays return shipping | [Read details](#)







eBay Buyer Protection

eBay will cover your purchase price plus original shipping.
[Learn more](#)



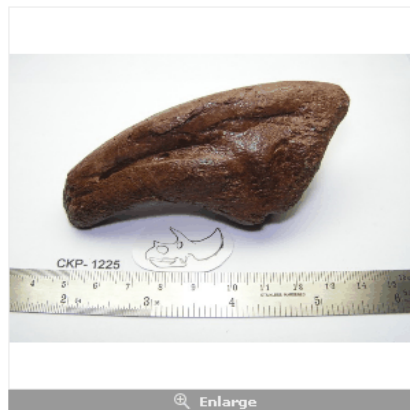
Description

- Guaranteed 100% Authentic GUCCI Sunglasses.
- Condition: Brand New.
- Made in Italy.
- Model No. 2597/F/S Color: 584 LF (BLACK FRAMES, GRAY GRADIENT LENSES)
- An Original Gucci Trademark Case And Cleaning Cloth Included.
- Size (mm):

Eye	Bridge	Vertical(B)	Temple
			
60	10	38	120



We have the data



Dinosaur Fossil Tyrannosaurus rex Foot Pes Claw ckp1225

Real Fossil from the Hell Creek Formation of Montana

Item condition: --

Time left: 4d 12h (Apr 08, 2010 06:03:30 PDT)

Bid history: 0 bids

Starting bid: **US \$6,500.00**

Your max bid: **US \$** **Place bid**

(Enter US \$6,500.00 or more)

or

Price: **US \$7,500.00** **Buy It Now**

[Watch this item](#)

Item: Tyrannosaurus rex Foot Claw

Size: 1mm between lines on ruler (25mm = 1 inch)

Collection Area: *Channel Deposit, sub-surface, Garfield County, Montana, Upper Hell Creek Formation, Late Cretaceous (65-68mya)*

Collection Date: 3/29/2010

Preparation Date: 3/29/2010

Sediment: *Coarse Sand (grains the size of table salt), reddish in color, high energy water flow, most areas contain shell fragments with the fossil. Layer below is a mudstone containing plant fossils.*

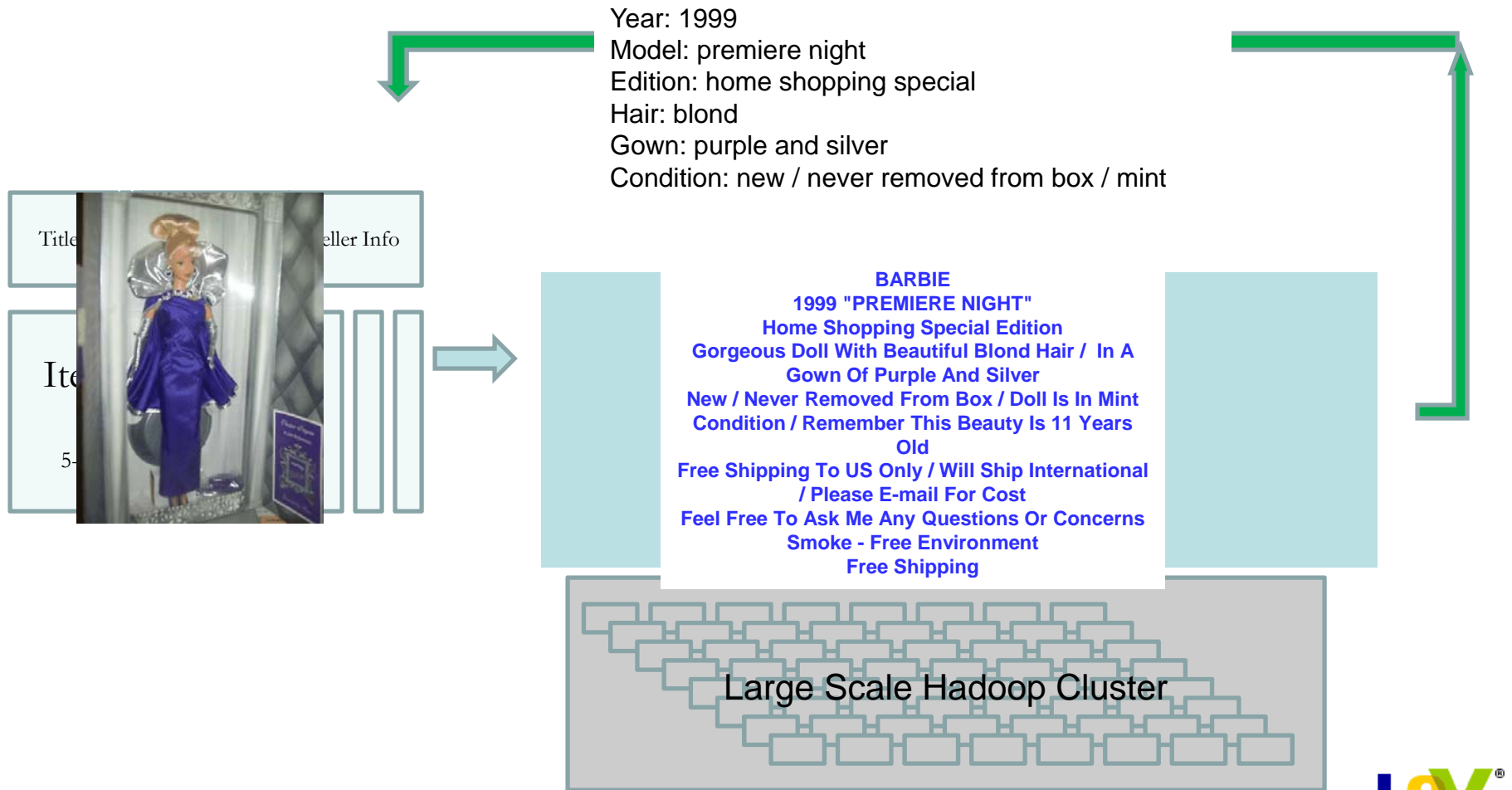
Description

We are back in the field!

Field Notes: We started digging again after another long and snowy winter. This claw was found on our second trip this year. It was discovered on 3/29/2010 by C.M. at the main channel site (NCD01). It was warm, near 70 degrees F, but there were 30-40 mph winds causing a huge sand storm in the quarry. You can see it in one of the field pictures. This fossil came out great and prepared well. -C.M.

This is a super unique find. This is a Tyrannosaurus rex foot claw. It is in good condition with excellent color and good surface detail. The claw is missing a small portion of the tip. This happened before the claw was fossilized. We are 100% sure this is a T-rex foot claw. It did tumble a bit before being fossilized, but all the features are present for a positive identification. This is not a huge T-rex foot claw. It is either the dew claw from an adult, or a toe claw from a juvenile-sub adult. It measures 4.5 inches along the outer curve and about 1.25 inches at its widest. There is slight crushing on the left side, but very minor causing only slight distortion which is very good for this size claw. This is the first T-rex foot claw to come out of this quarry(NCD01).

The BIG Picture



Our Approach

- To build an automatic product catalog we follow the following steps
 - Grouping items into categories
 - Category classification
 - Weeding out noise through accessory classification
 - Extraction of attribute names and values
 - Simple two pass approach
 - Cleaning and normalization
 - Capturing human expertise through machine learning

Catalog Discovery

- Improve value coverage for important names
 - Use machine learning to expand value coverage
- Product building
 - Organize items in a hierarchical collection
- Matching inventory to products
 - Adoption
- At each step we perform machine learning/text mining techniques

Item Categorization

- Near similar titles
 - “Apple IPOD Nano 4GB Black NEW! Great Deal!”
 - “Apple IPOD Nano 4GB Black NEW! Skin Great Deal!”
- Category Classification
 - Feature Selection
 - Smoothing
- Accessory classification (NBC)

Class Pruning

- Class Pruning – unique to eBay
 - we compute the posterior probability in NBC i.e, $P(C | \text{Title words})$ then for some title words the number of class they appear is huge (for instance, the word “harry potter” appears in thousands of categories and that puts a strain on the online posterior probability computation. To fix this, we use class pruning—for a given feature we only keep a few top classes in the computation.

Item Categorization on eBay



SELL YOUR ITEM

1. SELECT A CATEGORY 2. CREATE YOUR LISTING 3. REVIEW YOUR LISTING

Select a category

Find a matching category

Enter at least 3 keywords about your item to find a relevant category to list in.

For example: Transformers action figure

Larry Bird Boston Celtics Signed

Search

← Seller describes his item with a few keywords

Search categories

Browse categories

Recently used categories

Buyers will see your listing in the category that you select.

Sports Mem, Cards & Fan Shop

- Fan Apparel & Souvenirs > Basketball-NBA
- Cards > Basketball
- Autographs-Original > Basketball-NBA > Jerseys
- Fan Apparel & Souvenirs > Baseball-MLB
- Autographs-Original > Football-NFL > Jerseys
- Autographs-Original > Baseball-MLB > Jerseys
- Autographs-Original > Basketball-NBA > Photos
- Fan Apparel & Souvenirs > College-NCAA

← eBay recommends 15 categories for his/her consideration

Tip: Reach more buyers by selecting two categories. (Fees apply)

Continue

Start over

About eBay | Announcements | Buy Hub | Security Center | Resolution Center | Buyer Tools | Policies | Government Relations | Stores | Site Map | Help



Item Categorization on eBay



Larry Bird Boston Celtics Signed Adidas Classic Jersey

Price: US \$399.99

Buy It Now

■ Categorize into ?

- Sports Mem, Cards & Fan Shop > Manufacturer Authenticated > Basketball-NBA
- Sports Mem, Cards & Fan Shop > Fan Apparel & Souvenirs > Basketball-NBA
- Sports Mem, Cards & Fan Shop > Autographs-Original > Basketball-NBA > Jerseys
- Clothing, Shoes & Accessories > Men's Clothing > Athletic Apparel
- Clothing, Shoes & Accessories > Men's Clothing > Shirts > T-Shirts, Tank Tops
- Collectibles > Advertising > Clothing, Shoes & Accessories > Clothing



Challenge I

■ Large collection of categories

- 30K categories
- Hard to distinguish

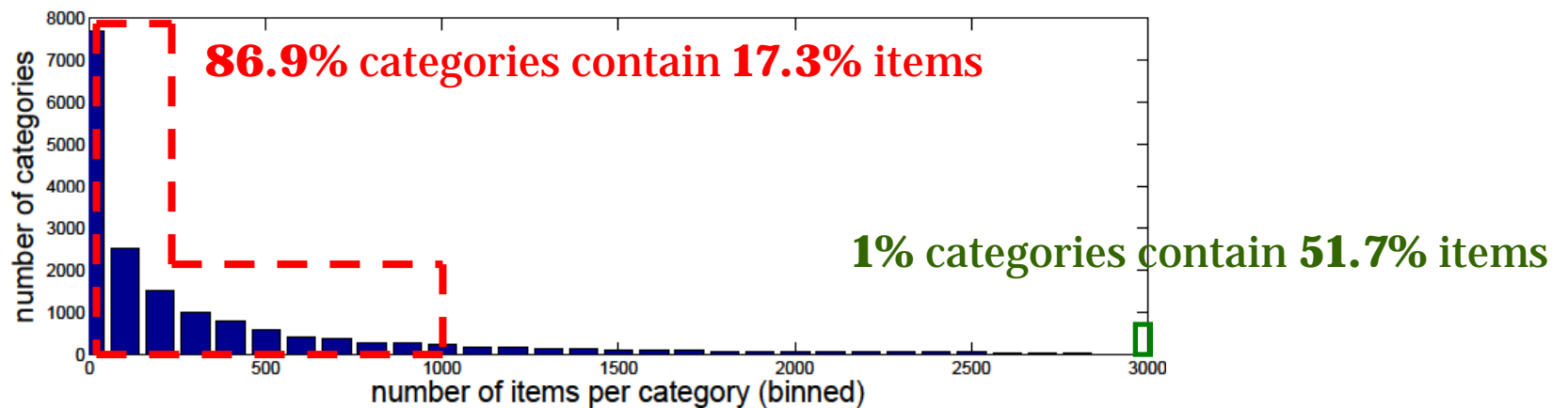
1. Clothing, Shoes & Accessories → Costumes & Reenactment Attire → Costumes → Women
2. Everything Else → Adult Only → Clothing, Shoes & Accessories → Costumes & Fantasy Wear → Women

■ Insufficient information of items

- Limited length of item title: 10 words
- Inaccurate or fraud title description

Challenge II

■ Highly skewed item distribution



■ Scalability and efficiency

- 4 million items daily
- Real-time response
- Good scalability and high efficiency

Applications

- Recommending category candidates for seller's listing
- Monitoring misclassification rate on current site
- Detecting outlier items

Method

- Multinomial Bayesian algorithm
- Smoothing
 - Scaling-up to cope with highly skewed item distribution
 - Data sparseness problem
 - Common or non-informative word problem

Bayesian Learning Framework

- We employ the Naive Bayes with Multinomial likelihood function which is to find the most likely class c with the maximum posterior probability of generating item t

Approach

- Exploit Data to the Maximum
- Apply simple algorithms at the same time

Smoothing Algorithms

- Laplace Smoothing
- Jelinek-Mercer Smoothing
- Dirichlet Prior
- Absolute Discounting
- Shrinkage Smoothing

Experiments

■ Train:

- Sold items on eBay site in about one month
- 18 million items
- 18K categories

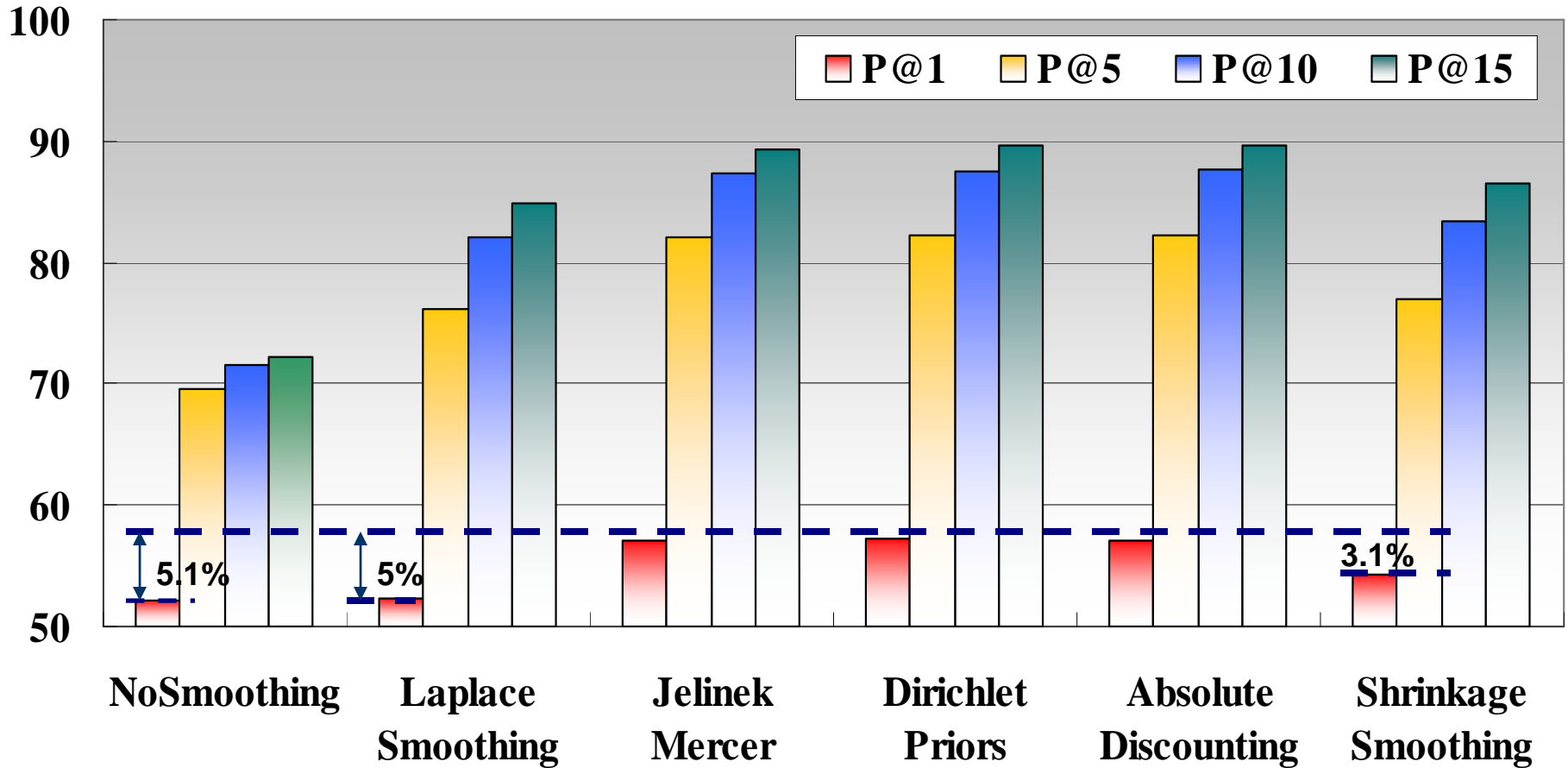
■ Test:

- Sold items in the day following the training period
- 278K items

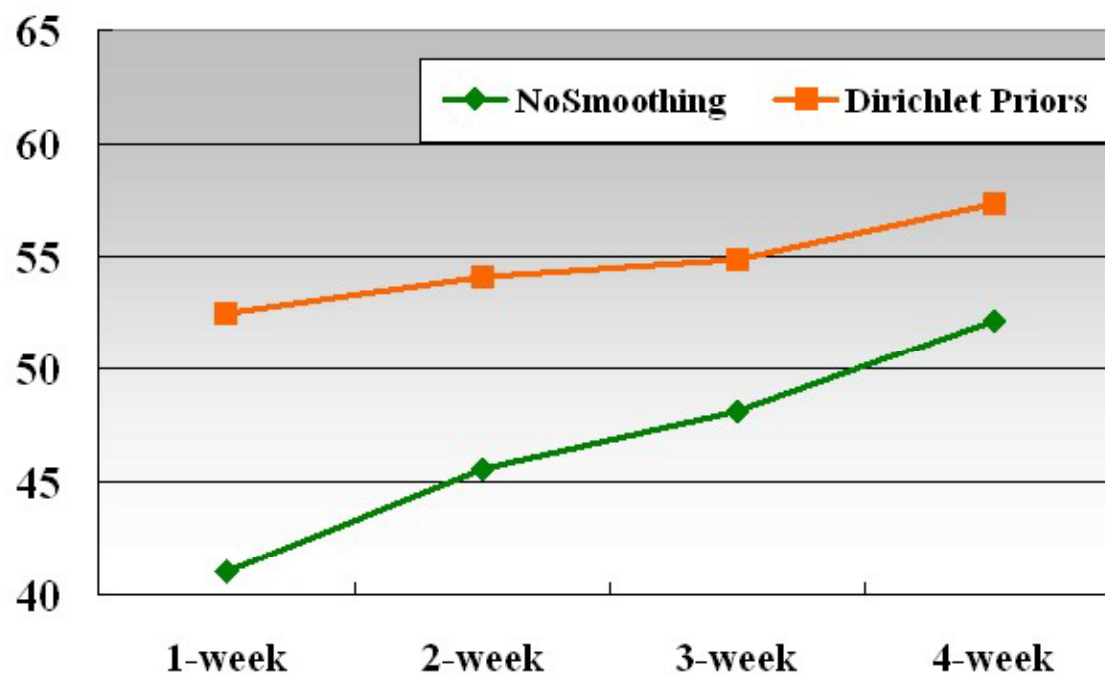
Research Questions

- How various smoothing methods perform on our task?
- How does smoothing interact with
 - size of training data set
 - size of category
 - vocabulary focusedness of category

Overall Precision of Smoothing Methods



Influence of size of training set on Smoothing



- Larger data set leads to better performance
- The rate of the increase is not fixed when multiplying data set blindly
- Quality of training data
- Increasing prior sample size μ improves performance

Influence of Category Size on Smoothing

- Smoothing for insufficient sample problem
 - Eliminate zero probability of unobserved words
- Two data sets:
 - LargeCat: the cat. containing >10K training instances
 - SmallCat: the cat. containing <1K training instances

		LargeCat	SmallCat
No Smoothing		69.4	35.9
Dirichlet Prior	$\mu = 100$	69.8	39.1
	$\mu = 500$	69.4	45.1
	$\mu = 1000$	70.1	43.7
	$\mu = 2000$	71.0	41.0
	$\mu = 5000$	72.8	35.1

- LargeCat significantly outperforms SmallCat by 27.7%
- Smoothing saves the system
 - +3.4% on LargeCat
 - +9.2% on SmallCat

Influence of word specificity on Smoothing

- Smoothing for common or non-informative words
 - Decrease the discrimination power of such words
- Two data sets:
 - SpecCat: the cat. containing words with high IDF values
 - NotSpecCat: the cat. containing words with low IDF values

		SpecCat	NotSpecCat
No Smoothing		71.4	42.8
Dirichlet Prior	$\mu = 100$	73.0	43.7
	$\mu = 500$	75.7	44.3
	$\mu = 1000$	75.1	44.7
	$\mu = 2000$	75.6	44.5
	$\mu = 5000$	76.4	45.0

- SpecCat significantly outperforms NotSpecCat by 31.4%
- Smoothing saves the system
 - +5.0% on SpecCat
 - +2.2% on NotSpecCat

Cataloging: Extraction of Attribute Names and Values

- We extract attribute names and values from millions of descriptions
- Harder than named entity recognition
 - Attribute names are not known beforehand
- We employ a two-pass process



■ Pass 1—Name identification

- Use a high-precision low-recall extraction based on pattern search
- Use seller count, items count and other statistics to find names



■ Pass 2—Improve recall

- Extract more names and values using the phase 1 results

■ Observation

- Being community supplied information these names and values are often
 - Noisy
 - Not normalized
 - Of low coverage

Cleaning and Normalization

- Sellers use different vocabulary
 - We need to normalize/merge names into more standardized form
- We use a semi-supervised learning method
 - We use a simple UI to view the names per category
 - Using this UI human subjects can choose which names (e.g., brand and manufacturer) are to be merged
 - From this human supervision we automatically draw training data to build a classifier that can predict given two or more names whether they should be merged

MaxEntropy Classifier

- Training a MaxEnt classifier
 - Automatically build training examples
 - Positive: cases coming from same clusters
 - Negative: cases coming from different clusters

Features...

- Features used

- The context of attribute names and values
- Jaccard distance between two names in term of their values
- Mutual information
- Seller adoption as confidence score
- Lexical equivalence (synonyms/related terms)
- ...

Improve value coverage for important names

- For a set of important names we seek to maximize the value coverage
 - Critical for automatic catalogs
 - Our extraction and merging/cleaning process give us a set of seed values
 - We apply a machine learning approach to discover more values from titles

Classification

- Algorithm: Convergent boundary classification (combine a high-precision and a high-recall classifier)
 - For each important name, we train a classifier
 - Training process is fully automatic, we go over the product titles
 - values from the seed dataset are taken as positive training examples.
 - all other tokens in the product title are added as negative training examples.
 - if a product title doesn't contain any value from the seed data, all tokens are added as test cases.

Features

- Feature generation
 - Positional features
 - Position of the token, statistical confidence for positions etc
 - Sequence modeling features
 - The names to the left and right of each token etc.
 - Syntactic features
 - Contains letters, digits, or mixed
 - Demand features
 - Is part of top queries etc.

Classifier

- We then train an SVM classifier that accurately classifies each token whether it is a value of a name
 - Example: if we see Canon, Sony, Samsung in the training examples for brand, then “mustek”, which is unknown but shares similar features as training cases will be classified as a value of brand. In several categories we achieve ~85% accuracy for brands and ~90-95% accuracy for model classification.

Summary

- Structured Data Discovery Problem –
Categorization and Catalogs
- Intelligent Use of High Volume even though
Low Quality Data
- Simple Algorithms with Good Features
(textual, author, adoption) can give good
results

Questions?

- We are hiring!
- Contact: nsundaresan@ebay.com