# Rare Item Detection in e-Commerce Site

Dan Shen
eBay Research Labs
No. 88 Keyuan Rd.
Shanghai, China
dashen@ebay.com

Xiaoyuan Wu
eBay Research Labs
No. 88 Keyuan Rd.
Shanghai, China
xiaowu@ebay.com

Alvaro Bolivar
eBay Research Labs
2145 Hamilton Avenue
San Jose, CA 95125
abolivar@ebay.com

## ABSTRACT

As the largest online marketplace in the world, eBay has a huge inventory where there are plenty of great rare items with potentially large, even rapturous buyers. These items are obscured in long tail of eBay item listing and hard to find through existing searching or browsing methods. It is observed that there are great rarity demands from users according to eBay query log. To keep up with the demands, the paper proposes a method to automatically detect rare items in eBay online listing. A large set of features relevant to the task are investigated to filter items and further measure item rareness. The experiments on the most rarity-demand-intensitive domains show that the method may effectively detect rare items ($> 90\%$ precision).

## Categories and Subject Descriptors

H.3.3 [**Information Systems** ]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

## General Terms

Algorithms, Measurement, Performance, Experimentation

## Keywords

long tail theory, rare item detection, rareness measure.

## 1. INTRODUCTION

In 2004, Chris Anderson [1] first coined a "long tail" concept to describe niche strategy of businesses. The long tail represents the evidence of statistical distribution that a high-amplitude population is followed by a low-amplitude population which gradually tails off asymptotically. Figure 1 shows a typical example of the long tail distribution, where the group in the long tail (about 80% in yellow color) comprises of a large number of low popularity items. The paper stated that significant profit is out of selling small volumes of hard-to-find items rather than only selling large volumes of popular items.

Above arguments are also true for eBay. As the largest online marketplace, the distribution of eBay items follows the long tail theory. Plenty of great rare items with potentially large, even rapturous buyers are obscured in the long tail of eBay item listing. A sucessful internet business will greatly benefit by additional efforts on promoting these items in some way, such as product recommendation and merchandising. The advantages of finding and promoting these items are spread widely.

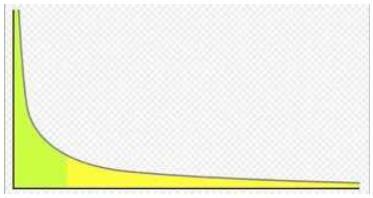

**Figure 1: An example of the long tail distribution**

- For e-commerce business, rare items are valuable, which exactly distinguish an e-commerce site from other common shopping sites and make the site much stickier to users. Promoting such items is a remarkable efficient form of marketing, making a site more interesting, allowing unique items to find suitable collectors.
- For consumers, it is hard to find rare items through existing searching methods although the items are equally or even more attractive. Buyers would never even learn about these items before they see them. Promoting such items will bring user more fun and better shopping experience.
- As the biggest auction site, eBay is a market with high freedom of deciding item prices. Price of an item lies on anxious extent of buyers. Once buyers miss a rare item, it will be really hard to find a repetition. Therefore, rare items always have the higher bidding prices. We believe finding and promoting rare items will bring more revenue to an auction site.

This paper proposes a method to automatically detect such valuable rare items from the large eBay item listing. As a following strategy, these items might be given more chance of exposure in item listing page, laid out in merchandising window or sent to other social network sites, such as facebook as widget. To our best knowledge, it is the first piece of work to explore rare items in an e-commerce site.

## 2. DEMAND OF RARITY

We develop a rare item detection algorithm to keep up with demand. To measure the demand of rare items, we


*WWW 2009,* April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

analyze user queries during 10-19 Oct. 2008 (10 days) on eBay site. A rarity query is defined as a query containing one of the rare words, such as *rare*, *infrequent*, *unusual*, *unique*, *only one*, *limited edition*, etc. The total rarity demand in a domain is measured as the proportion of the number of rarity queries to the total number of queries.

The following lists the domains containing the most demand for rarity, where the number in bracket is the permillage (‰) of rarity queries.

> Music (4.12), Toys & Hobbies (2.65), Coins & Paper Money (2.13), Entertainment Memorabilia (1.97), Stamps (1.70), DVDs & Movies (1.53), Collectibles (1.29), Art (1.21), Jewelry & Watches (1.18), Antiques (0.99)

Since eBay gets about 40 million queries everyday, the absolute rarity demand calculated with this proportion can not be ignored. Actually the potential market of rare items is even beyond the users explicitly asking for rarity since more users might haven't realized that they may search in this way or there are such rare items on eBay. Therefore, we propose a method to automatically detect valuable rare items and promote them directly to potential buyers.

# 3. METHOD

To find rare items among thousand millions of items on eBay site, we propose a method to rank items according to their rareness scores. Due to the huge size of item listing, it will be too costly to measure rareness for all of items. For efficiency, we firstly conduct a preprocessing step to quickly remove impossible items based on a set of heuristic rules (Section 3.1). Next, we rank the rest items by measuring their rareness (Section 3.2).

## 3.1 Item filtering

As a preprocessing step, the following features are used to filter items:

- **Listing quantity** The listing quantity of an item should be 1 if it is really rare.
- **Feedback score (FS)** Rare items are more likely to be selled by casual sellers ($100 \leq \text{FS} < 1000$) rather than power sellers and new sellers.
- **Seller's store** Since we prefer to cascual sellers, we further require the sellers without stores.
- **Positive feedback percent (PFP)** We choose items from the sellers with enough good selling reputation ($\text{PFP} > 95\%$).

## 3.2 Item rareness measure

According to the observation of the task, We design the following features to measure item rareness.

- **Title rareness** Given a certain item, we measure the simiarity between its title and the titles of all other items in the same category. Regarding a title as a term vector, we adapt cosine similarity measure to calculate the similarity between two titles $Sim(t_1, t_2)$, where each term is weighted with tf-idf. The title rareness $R(t)$ of the item $t$ in the category $C$ is calculated as: $R(t) = 1/\max_{t_i \in C \& t \neq t_i} Sim(t, t_i)$. The feature indicates item rareness at least on surface level.
- **Listing type** Since desired prices of rare items are often unpredicted, this kind of items is more likely to be selled as "*auction*" rather than "*Buy It Now*".
- **Description** There is always a story behind a rare item. Sellers don't merely list rare items but also share their experience about the items. Some features are extracted from descriptions, such as does a description use any templates? and, is it written in the first person?
- **Watch count** This feature awards valuable items and punishes naught items according to their popularities.
- **Bid count** People always are extremely anxious to win a rare item since it is really hard to find a repetition once they miss it. Therefore, the bid count may reflect the rarity of an item to some extent.

These features are further combined with a linear interpolation function, where the feature weights are estimated from a set of human-labeled rare items.

**Table 1: Performance**

| Domain | #items | P@10 | P@25 | P@50 |
|---|---|---|---|---|
| Music | 429,264 | 0.90 | 0.84 | 0.88 |
| Toys & Hobbies | 682,252 | 0.90 | 0.80 | 0.86 |
| Coins & Paper Money | 271,300 | 0.90 | 0.88 | 0.92 |
| Entertainment Memorabilia | 180,560 | 1.00 | 0.96 | 0.98 |
| Stamps | 196,484 | 1.00 | 1.00 | 0.98 |
| Overall | 1,759,860 | 0.94 | 0.90 | 0.92 |

# 4. RESULTS

The experiments are conducted on the most rarity-demand-intensive domains. The method applies to the items which are ended during Jan 2009. As a result, a set of top ranked items are returned for each domain. We measure precisions regarding to 10/25/50 items returned respectively. Recall can't be evaluated since it is unknown that how many real rare items exist in the huge eBay listing. Experts in each domain are asked to judge the items as "rare"/"not rare". As shown in Table 1, our method achieves above 90% precision on average for the listed domains. We believe eBay will supply users the better shopping experience by effectively detecting and recommending these rare items.

# 5. CONCLUSION

This paper takes the first step towards leveraging the "long tail" as a part of business for an e-commerce site. We propose a method to automatically detect rare items in eBay online listing by exploring a large set of relevant features. The method achieves above 90% precision on the most rarity-demand-intensitive domains. We believe that eBay will greatly benefit from additional efforts on promoting such valuable rare items in the long tail. Furthermore, we plan to explore the detection of query-related rare items in the future.

# 6. REFERENCE

[1] C. Anderson. The long tail. *Wired*, Oct. 2004.