Identifying Interestingness in Ebay products using Pinterest Data

Vaibhav Dwivedi (vdwivedi@buffalo.edu) 5013-2711
Pritika Mehta (pritikam@buffalo.edu) 5013-4965

ABSTRACT

Pinterest is a social media platform which lets users pin pages on their boards. Research shows that majority of these pages are something which the user “wants” or “needs”, i.e. something the user is interested in owning. This project aims to extract this interestingness from a set of pinterest items and apply it on a set of ebay items to find potentially interesting ebay items. The project uses word2vec to generate feature vectors for the set of pinterest and ebay items and uses libsvm to train a model to classify interesting and non-interesting items and then test it in a set of ebay items to identify potentially interesting items. We evaluate our project by taking a survey to find a more interesting set of items from a given set of items classified as interesting by our model and a set of pinterest items.

INTRODUCTION

There are three main type of modern recommendation system algorithms –

1. Collaborative Filtering
2. Content Based Filtering
3. Trust Aware Recommendation Systems

Collaborative filtering methods are based on collecting and analyzing a large amount of information on users’ behaviors, activities or preferences and predicting what users will like based on their similarity to other users.A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items such as movies without requiring an "understanding" of the item itself.

Content-based filtering methods are based on a description of the item and a profile of the user’s preference. In a content-based recommender system, keywords are used to describe the items; beside, a user profile is built to indicate the type of item this user likes. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past. In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended.

Trust Aware Recommendation Systems - Research shows users are more comfortable using products that are suggested by the people that they know rather than a stranger. This type of recommendation system information from user’s “trust-circle” like facebook to filter products based on the history and preferences of their friends.

The problem with these above mentioned algorithms is that they either require a large user base as in case of collaborative filtering or huge amounts of analyzable user history as with content based filtering. Trust aware recommendation fails to when the user does not provide his social information or is socially inactive.

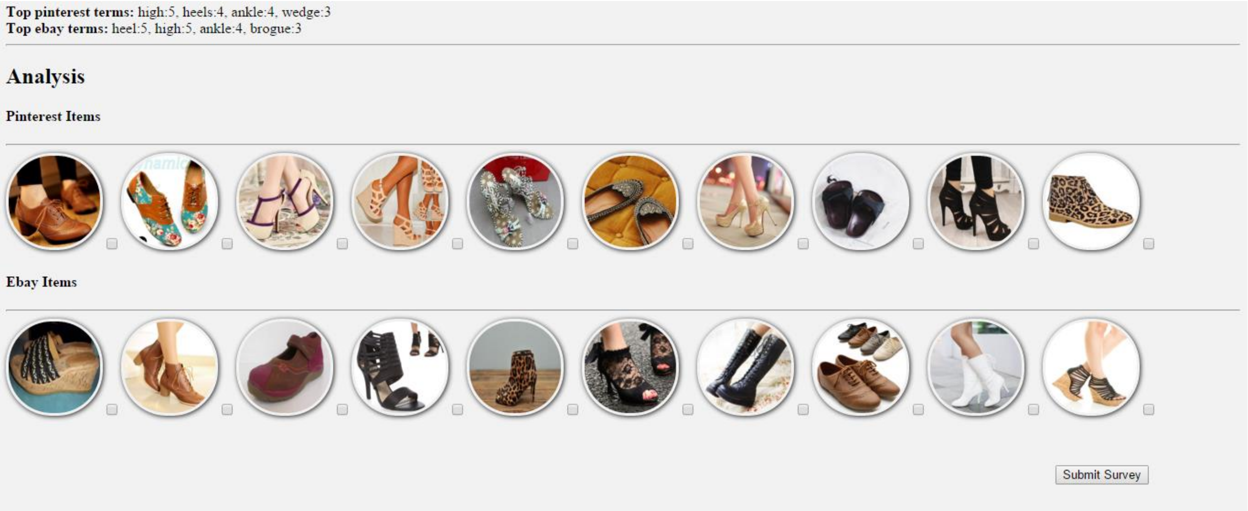
In this project, we aim to provide recommendations based on a product's interestingness to the user. To classify products as interesting and not interesting we use the pins on the pinterest as our ground truth data set. This is based on the research that the majority of pins on pinterest are something which the user "needs" or "wants". Thus in case of products these pins are something which the user is "interested" to buy.

Most of the previous work in extracting interestingness is based on the domain of computer vision. This though is highly impractical in a real world scenario, as a typical e-commerce site consists of millions of images and extracting meaningful information from these many images can be time consuming and also fairly inaccurate.

In this project we aim to extract interestingness based on the textual data of the product. Research and our preliminary analysis showed that users are interested in certain fashion vocabulary such as vintage, modern, classic, timeless, etc and thus this information can be mined to train a model to classify products as either interesting or not interesting.

PRELIMINARY ANALYSIS

In order to establish pinterest vocabulary as ground truth, we collected a randomly generated set of pinterest pins and ebay products related to women's shoes category with similar titles. We then took a survey where people where first given a set of 10 pinterest items and were asked to find the 5 items which they find to be most interesting or something they would consider to buy. They were then shown as set of 10 ebay products and were asked to do the same thing. The results of the survey taken by 100 different people were aggregated and their fashion vocabulary was analyzed. It was found that the top words in both the sets were "high heels", "laces" and "classic". This thus helped us establish pinterest pins as the ground truth for determining interestingness.



Top pinterest terms: high:5, heels:4, ankle:4, wedge:3
Top ebay terms: heel:5, high:5, ankle:4, brogue:3

Analysis

Pinterest Items

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Ebay Items

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

DATASET

Ebay –

Ebay was crawled using an API to aggregate a set of 10,000 ebay products tagged under “women shoes” category. The title and image of the products were extracted. The textual data was passed through a pipeline where –

- Duplicate items from multiple sellers were removed.
- Price was removed from title.
- EBay specific terms like new, used, NIB(new in box), etc were removed.
- Titles with less than 4 characters were removed.

Pinterest –

Pinterest was crawled using an API to aggregate a set of 10,000 pinterest products tagged under “women shoes” category. Pins with popularity score of zero were ignored. The title and image of the products were extracted. The textual data was passed through a pipeline where –

- Products that were concepts were removed based on information from its title.
- Pinterest specific terms like “pin”, “want”, “need”, etc are removed.
- Titles with less than 4 characters were removed.

FEATURE EXTRACTION

Before we could train our classifier to classify products and identify potentially interesting items, we needed to generate features for each of the titles of the products so that it could then later be used by the classifier to identify similarity between two items. We used word2vec library for this purpose.

Word2Vec –

Word2Vec uses skip-gram model to generate feature vectors for given training set. Word2vec model can be used to map each word to a vector of typically several hundred elements, which represent that word's relation to other words. This vector is the neural network's hidden layer. In our case, we trained our Word2Vec model with a ebay dataset of 10,000 products and a pinterest dataset of 10,000 products.

It took around 15 Minutes to train our model using these 20,000 words using 10 dimensions and minimum word count of 5. Since the generated feature vector set consists of features for individual words, the word vectors for every word in its title is taken and normalize by the total number of words. The final vector of 10 dimensions contributes 10 features.

Popularity Feature –

Though these features similarity of one title from another, there is one other factor which contributes towards interestingness of products and that is its popularity. Thus we added a 11th feature which accounts for the popularity of the product. For pinterest items, we determined this feature by taking a ratio of the no of likes for that item to the total number of likes in all pinterest items. For ebay the case is a little different as there is no way to determine popularity of product as the same product can be sold by multiple sellers at different prices thereby accumulating different ratings. Thus we gave an equal popularity score of 1/N to all ebay items where N is the total number of ebay items.

EXPERIMENTATION –

We use LIBSVM, an open source machine learning library, as our classifier. We then classified interestingness under two classes – positive and negative. Since we used pinterest pins as our ground truth for interestingness, all the pinterest pins are classified as positive. Since there is no conclusion about the interestingness of ebay items, we classify them all as negative.

Training –

To avoid unbalanced classes we used equal number of ebay and pinterest items. We train our classifier using a smaller dataset of 2000 ebay and 2000 pinterest items with ebay items marked as negative and pinterest items marked as negative.

Although we treat all eBay items to be negative in our classification task, some with positive class attributes fall on the Pinterest side of the hyperplane and are thus classified as positive and thus are potentially interesting items.

Thus ebay items with similar features which in our case similar fashion vocabulary are grouped together with the pinterst items, thereby classifying them as potentially interesting items.

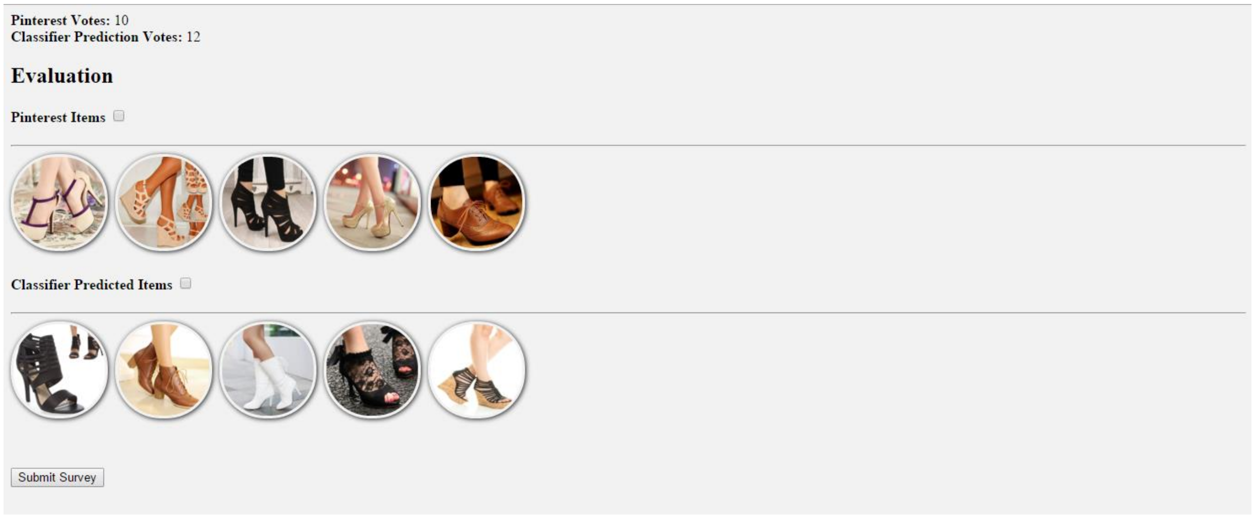
Testing –

A set consisting only of 8000 ebay items are then tested on the trained classifier and a prediction model is generated. The prediction model consists of the classifier's prediction for each of the ebay item. All the ebay items which are classified by the classifier under positive class are considered to be potentially interesting and are thus stored separately for evaluation. Our classifier predicted around 550 ebay items as possibly interesting.

EVALUATION –

The evaluation is performed with the help of human survey in a manner similar to our preliminary analysis. Each person is provided with two randomly generated sets. The first set consists of a set of images consisting of pins from pinterest. The second set consists of items that are classified by our classifier as potentially interesting.

Each person was then asked to select the set which they feel to be most interesting. The results were aggregated from each survey and it was found that 54% of people found the dataset generated by our classifier to be interesting. Since the other dataset was taken straight from pinterest which by our ground truth are interesting, more than half of people agreeing on the dataset generated by our classifier to be interesting proves the correctness of our system.



Pinterest Votes: 10
Classifier Prediction Votes: 12

Evaluation

Pinterest Items □

Classifier Predicted Items □

CONCLUSION –

By our evaluation, we conclude that identifying and extracting interestingness from pinterest items and then applying them to the ebay products helps us recommend products that are potentially interesting to the users thereby increasing the probability of sale of these products. We also conclude that this provides us with a good system to recommend users at the initial stages of a website when there are few users and less analyzable user history.

In this project, we identified interestingness by mining fashion vocabulary from the pinterest and ebay pins. We believe this system can be expanded into other domains such as electronics and advertisements using a similar concept of interestingness.

REFERENCES –

[1] Identifying Interestingness in Fashion E-commerce using Pinterest Data Paper
https://www.cs.utexas.edu/~nrajani/srs.pdf

[2] Word2Vec Library
https://code.google.com/p/word2vec/

[3] LIBSVM: Setup & Installation
http://vgl-ait.org/cvwiki/doku.php?id=libsvm:tutorial:install:ubuntu