Home Depot Product Search Relevance

Kewen Zhang, Pengfei Wang, Xiaoci Xing, Ziyue Wu

***Abstract*—In this search relevance project, our goal was to build a model to predict the relevance of search items and product on homedepot.com, given the searching items, resulting product titles and product descriptions. Our teams solution relies heavily on feature extraction/selection and model ensembling.**

Our solution consists of three parts:

1. Text cleaning

Before generating features, we have realized that its reasonable to process the data. So we cleaned the data with spelling correction, synonym replacement, removing dots and stop words. Then we selected the optimal solution of N for each N-grams feature.

2. feature extraction

We had tried three types of feature:

a counting features(which is not used in the final result)

b distance features(1-8 grams)

c TF-IDF features (1-8 grams)

3. selection and model ensembling.

Model ensembling consisted of two main steps. Firstly, we trained model library using different models, different parameter settings, and different subsets of the features. Secondly, we generated ensemble submission from the possible ensemble selections. Performance was estimated using cross validation within the training set. We tried both classification and regression to compare our results.

a Neural Net

b General Linear Model

c Machine Learning Methods

I. Introduction

Shoppers rely on Home Depots product authority to find and buy the latest products and to get timely solutions to their home improvement needs. From installing a new ceiling fan to remodeling an entire kitchen, with the click of a mouse or tap of the screen, customers expect the correct results to their queries quickly. Speed, accuracy and delivering a frictionless customer experience are essential.

In this project, we were asked to help them improve their customers' shopping experience by developing a model that can accurately predict the relevance of search results.

Search relevancy is an implicit measure Home Depot uses to gauge how quickly they can get customers to the right products. Currently, human raters evaluate the impact of potential changes to their search algorithms, which is a slow and subjective process. By removing or minimizing human input in search relevance evaluation, Home Depot hopes to increase the number of iterations their team can perform on the current search algorithms.

The relevance is a number between 1 (not relevant) to 3 (highly relevant). For example, a search for "AA battery" would be considered highly relevant to a pack of size AA batteries (relevance = 3), mildly relevant to a cordless drill battery (relevance = 2), and not relevant to a snow shovel (relevance = 1).

II. Data Cleaning

id	product_uid	product_title	search_term	product_description	relevance
2	100001	Simpson Strong-Tie 12-Gauge Angle	angle bracket	Not only do angles make joints stronger, they also pro...	3.00
3	100001	Simpson Strong-Tie 12-Gauge Angle	l bracket	Not only do angles make joints stronger, they also pro...	2.50
9	100002	BEHR Premium Textured DeckOver 1-gal. #SC-141 Tu...	deck over	BEHR Premium Textured DECKOVER is an innovative s...	3.00
16	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in C...	rain shower head	Update your bathroom with the Delta Vero Single-Han...	2.33
17	100005	Delta Vero 1-Handle Shower Only Faucet Trim Kit in C...	shower only faucet	Update your bathroom with the Delta Vero Single-Han...	2.67
18	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Micro...	convection otr	Achieving delicious results is almost effortless with thi...	3.00
20	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Micro...	microwave over stove	Achieving delicious results is almost effortless with thi...	2.67
21	100006	Whirlpool 1.9 cu. ft. Over the Range Convection Micro...	microwaves	Achieving delicious results is almost effortless with thi...	3.00
23	100007	Lithonia Lighting Quantum 2-Light Black LED Emergen...	emergency light	The Quantum Adjustable 2-Light LED Black Emergenc...	2.67
27	100009	House of Fara 3/4 in. x 3 in. x 8 ft. MDF Fluted Casing	mdf 3/4	Get the House of Fara 3/4 in. x 3 in. x 8 ft. MDF Flute...	3.00
34	100010	Valley View Industries Metal Stakes (4-Pack)	steele stake	Valley View Industries Metal Stakes (4-Pack) are 9 in. g...	2.67
35	100011	Toro Personal Pace Recycler 22 in. Variable Speed Self...	briggs and stratton lawn mower	Recycler 22 in. Personal Pace Variable Speed Self-Prop...	3.00
37	100011	Toro Personal Pace Recycler 22 in. Variable Speed Self...	gas mowe	Recycler 22 in. Personal Pace Variable Speed Self-Prop...	3.00
38	100011	Toro Personal Pace Recycler 22 in. Variable Speed Self...	honda mower	Recycler 22 in. Personal Pace Variable Speed Self-Prop...	2.00
48	100012	Hampton Bay Caramel Simple Weave Bamboo Rollup S...	hampton bay chestnut pull up shade	The 96 in. wide Caramel Simple Weave Rollup Bamboo...	2.67
51	100013	InSinkErator SinkTop Switch Single Outlet for InSinkEra...	disposer	The InSinkErator SinkTop Switch Single Outlet for InSin...	2.67
65	100016	Sunjoy Calais 8 ft. x 5 ft. x 8 ft. Steel Tile Fabric Grill G...	grill gazebo	Make grilling great with this handsome and functional ...	3.00
69	100017	MD Building Products 36 in. x 36 in. Cloverleaf Alumin...	door guards	The MD Building Products 36 in. x 36 in. x 1/50 in. Al...	1.00

Fig. 1. 74067 observations

A. Word Replacement

By exploring the provided data, it seems important to perform some word replacements/alignments, e.g., spelling correction and synonym replacement, to align those words with the same or similar meaning.

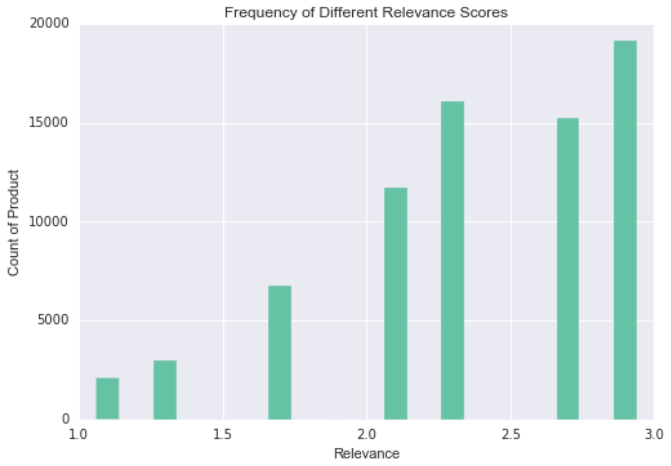


Fig. 2. Relevance

1) Spelling Correction: We just fixed the typo like "helloWorld" to "hello World", which seperated two individual word. What is more, we delete the stop words like "the", "and", for these words don't lend any support to our training.

2) Synonym Replacement: We replaced the synonym to reduce and simplify our data size.

3) Other Replacements: Including but not limited to removing insignificant punctuation like "# , .", translating plurality like "feet" to "foot".

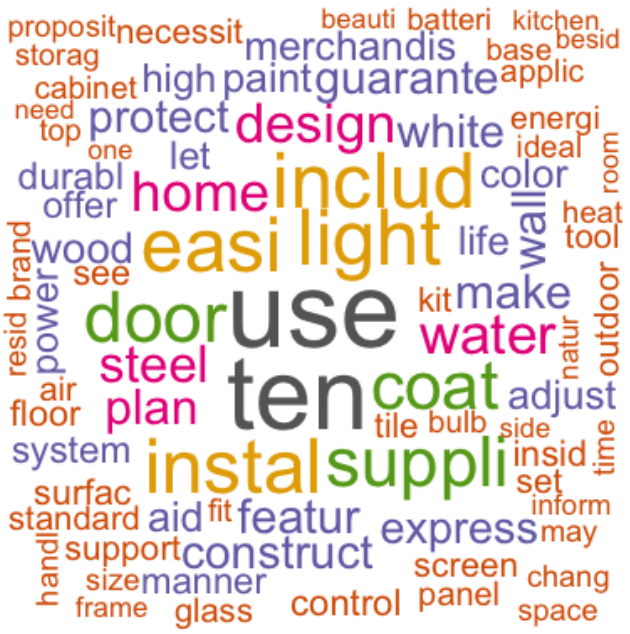


Fig. 3. Combined List

III. Feature Extraction/Selection

We use function ngram(s, n) to extract string/sentence ss n-gram (splitted by whitespace), where n = 1, 2, 3....

For example ngram(big red apple, 2) = [big red, red apple].

All the features are extracted for each run (i.e., repeated time) and fold(used in cross-validation and ensembling), and for the entire training and testing set (used in final model building and generating submission).

A. Counting Features

We generate counting features by calculating the total number of words in product title and search term.

For some search terms, it only contains one word, like some ambiguous word car or some meaningless stopword to, or. While there are still many meaningful one-word search term, multi-word search items are more likely to be specific and clear.

Another counting features are the number of digits appeared in the product title and search term.

	id	product_uid	product_title	search_term	relevance
1697	5185	100900	ShelterLogic 10 ft. x 15 ft. x 8 ft. Tan Cover...	car	1.67
3782	11752	102051	Teks #12 2-3/4 in. Phillips Flat-Head Self-Dri...	To	1.00

Fig. 4. Counting Features

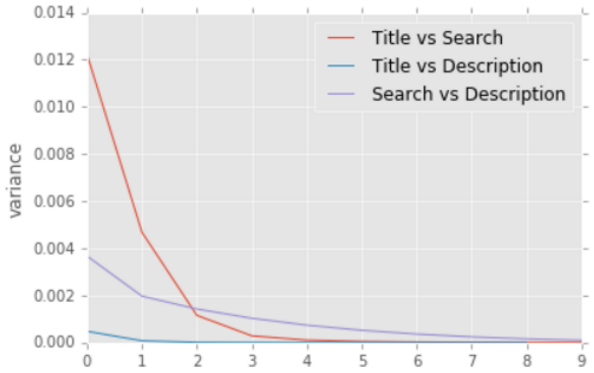
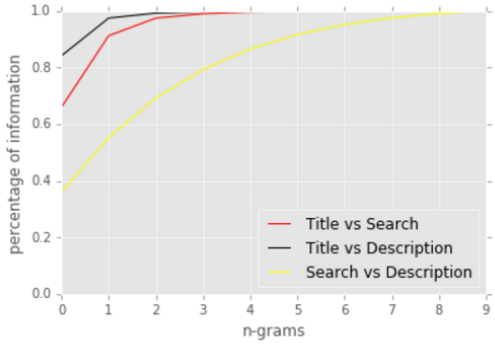


Fig. 5. Distance Feature

B. Jaccard Distance

JaccardCoef(A,B) = $|A \bigcap B|/|A \bigcup B|$ We calculated the distance between "Search Item", "Product Description" and "Product Title" respectively. We thought it was important to compute the distance between "Product Description" and "Product Title" because ... We tried many "n"s for our results, and found the optimal "n"s, n=3 for Product Title and Search Item, n=2 for Product Title and Product Description, n=8 for Search Item and Product Description.



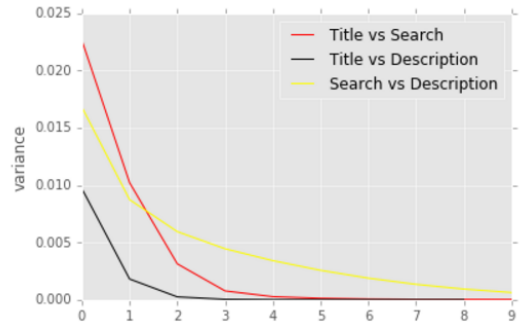
Thus, we choose n as follow:

- Title vs Search: n = 3,
- Title vs Description: n = 2
- Search vs Description: n = 8

Fig. 6. n-gram for Distance Feature

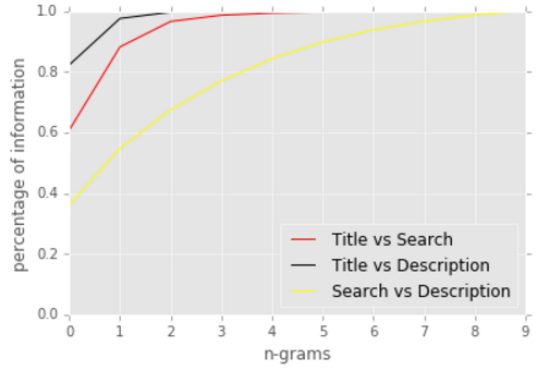
C. Cosine Similarity with tf-idf

In the case of the term frequency tf(t,d), the simplest choice is to use the raw frequency of a term in a document. The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. For n-gram selection of tf-idf, we have the similar result as Jaccard Coefficient distance.



Therefore, we will choose n = 4 for our training model

Fig. 7. TF-IDF Feature



Percentage of information: cumlative variance / sum of variance

Thus, we choose n as follow:

- Title vs Search: n = 3,
- Title vs Description: n = 2
- Search vs Description: n = 8

Fig. 8. n-gram for TF-IDF Feature

D. Customized Features

As most of items have their attributes in data, we choose the most three common attributes among products, brand, color and material. In these features, we used the method the same as Jaccard coefficient. Obviously, as many items only have one word in their color and material description, we would prefer 1-gram here.

E. Combined Features

Combined all the features above as our predictors.

IV. Regression Analysis

Because we found the relevance scores in training set are 1, 1.33, 1.5, 1.67, 2, 2.33, 2.5, 2.67. 3, we think regression should be better than classification.

	product_uid	name	value
279	100010	Color/Finish	Silver/Gray
282	100010	Material	Steel
283	100010	MFG Brand Name	Valley View Industries

Fig. 9. Combined Features

For each model selection method, we use 5-fold cross-validation to avoid overfitting problem. The training data is broken up into 5 sections called folds. At each iteration, we hide one fold and train the selected model on the remaining 4 folds. Grid search is applied to find several possible costs for most models. For each cost we quantify the models performance by taking the average test error over all folds, then pick the cost parameters with the lowest overall test error.

A. Simple Linear Regression

- The basic understanding of relationship
- Try as a starting point

B. Ridge Regression

- Alleciate multicollinearity among predictor variables
- Grid-search to find optimal penalty
- Search Range:
  Alpha: $10e^{-6}$ to $10e^{-2}$

C. Random Forest

Random forest is a substantial modification that builds a large collection of de-correlated trees and then averages them. It is a popular ensemble selection method as the performance of random forest is very similar to boosting and corresponding parameters are simpler to train and tune.

- Parameters range for grid-search:
  Number of trees in forest: 10-35
  Maximum depth of tree: 2-20

The final output shows that the optimal number of trees in forest is 33 and the maximum depth of tree is 10.

D. Extreme Gradient Boosting

XGBoost, short for Extreme Gradient Boosting, is also a ensemble model for supervised learning problem. It applies a more regularized model formalization to control over-fitting, which often gives a better performance.

- Search Range:
  Number of tree depth for base learners: 2-20
  Number 10-35

The final output shows that the optimal number of tree depth is 9 and number of boosted trees to fit is 31.

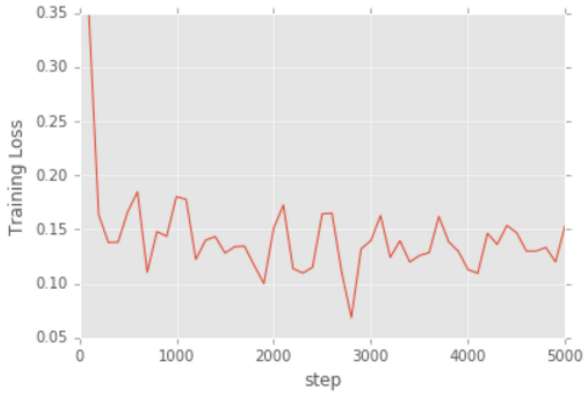


Fig. 10. Neural Network Loss

E. Neural Network Regressor

- Layers: 4
- Nodes: 50/layers
- Steps: 5000

V. Conclusion

We applied 5-folds cross validation for each regressors to generated the following table:

References

[1] https://github.com/tensorflow/skflow
[2] https://github.com/ChenglongChen/Kaggle_CrowdFlower
[3] https://github.com/dmlc/xgboost