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

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.

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".

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

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.

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.

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.

A. Simple Linear Regression

B. Ridge Regression

C. Random Forest

- Trial and error Search Range:
- Grid-search to find optimal parameter
  Number of trees in forest: 5-20
  Maximum depth of tree: 2-20

D. XGBoosting

- Grid-search to find optimal parameter
- Search Range:
  Number of tree depth for base learners: 5-20
  Number 2-20

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] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

John Doe Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.