RESEARCH ARTICLE OPEN ACCESS

Predict the relevance of search results from Ecommerce sites

Mr. Santosh Kumar Mishra*, Mukul Singhal**,Dikshant Awasthi***, Parshvi Verma****,Sumit Kumar Malik******
**(Assistant Professor, Department of Computer Science, RKGIT College Ghaziabad*
*** (Department of Computer Science, RKGIT College Ghaziabad*
**** (Department of Computer Science, RKGIT College Ghaziabad*
***** (Department of Computer Science, RKGIT College Ghaziabad*
****** (Department of Computer Science, RKGIT College Ghaziabad*

ABSTRACT
Search engines have become the dominant model of online search. Large and small ecommerce provide built-in search capability to their visitors to examine the products they have. While most large business are able to hire the necessary skills to build advanced search engines, small online business still lack the ability to evaluate the results of their search engines, which means losing the opportunity to compete with larger business .Due to today's transition from visiting physical stores to online shopping, predicting customer behaviour in the context of e- commerce is gaining importance. It can increase customer satisfaction and sales, resulting in higher conversion rates and a competitive advantage, by facilitating a more personalized shopping process. With the rapid growth of e-Commerce, online product search has emerged as a popular and effective paradigm for customers to find desired products and engage in online shopping. However, there is still a big gap between the products that customers really desire to purchase and relevance of products that are suggested in response to a query from the customer. In this synopsis, we propose a robust way of predicting relevance scores given a search query and a product, using techniques involving machine learning, natural language processing and information retrieval.

Keywords – CrowdFlower, E-commerce, Online Shopping, Products, Query-Based Search, Search Engines,

I. INTRODUCTION

Large e-commerce sites typically use query based search to help consumers to find information on their websites. They are able to use technology to provide user with a better experience. Because they understand the importance of search relevance, and that long and unsuccessful searches can turn their users away because users are accustomed to and expect instant, relevant searches results. This project aims at ranking those product items, which are not only relevant to the user's query but with higher probability to be purchased by the user, into higher position. To evaluate our new ranking methods, we collect a large scale dataset of search and transaction logs for a commercial product search engine. Experimental results demonstrate that our new ranking method is more effective for locating product items that users really buy to desire items at higher ranking positions without hurting the search relevance. High quality search is all about returning relevant results even when the data is changing or poorly structured, the queries are imprecise. Here is where we are going today: given only raw text as input, our goal is to predict the relevancy of the results at the e-commerce site. As a result, help them to improve their customer shopping experience.

In this work, we show the use of machine-learning algorithms, as well as the use and implementation of preprocessing methods, applied to the prediction of text search relevance. For this purpose, the paper describes first some basic concepts about data preprocessing and text retrieval. In this context, some properties of tfidf are described next. Then, we describe the dataset used for this study (from CrowdFlower) as well as some proposal and hypothesis for data preprocessing towards feature extraction. Afterwards, we describe the application of two machine-learning algorithms on the processed data. Finally, on the last section, we show benchmark results we obtained on running four combinations of feature extraction methods and machine-learning algorithms[1].

1.1 Proposed Work

In proposed work model, data is taken as input. After input part, for preprocessing we mainly performed HTML tags dropping, word replacement, and stemming part. For feature extraction part, counting features, distance features, TF-IDF features and Query Id are used. For supervised learning method, I used ensemble selection to generate ensemble from a model library. The model library built with models trained using various algorithms, various parameter settings, and various feature sets. Parameter tuning is used to choose parameter setting from a pre-defined parameter space for training different models. In feature extraction, counting features, distance features, TF-IDF features and query id are used. In Ensemble Selection, XGBoost Linear Booster, XGBoost Tree Booster, Gradient Boosting Regressor, Extra Trees Regressor, Random Forest Regressor, SVR, Ridge, Keras NN, RGF Regression. Python is used in this project. For model training part, XGBoost, Sklearn, keras and rgf[2].

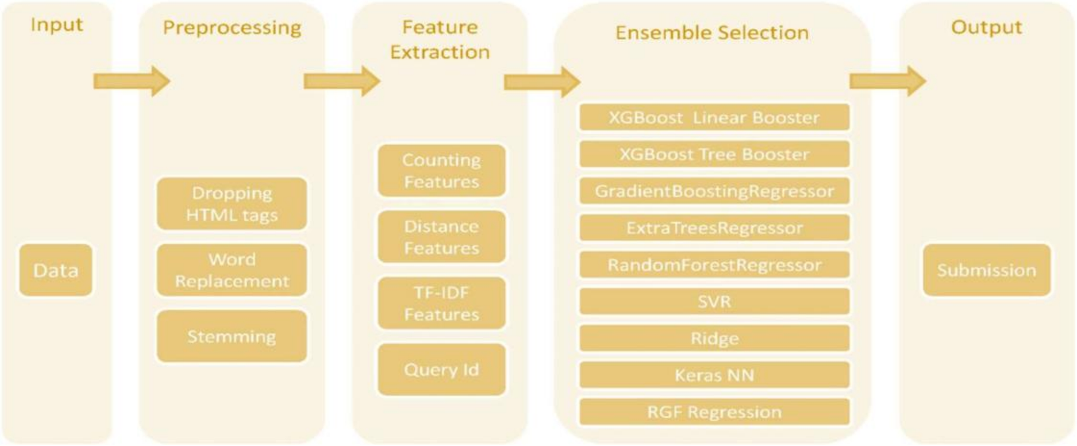

FIGURE 4.1 Flow Chart of the Work The measurement of the separate model and the ensemble model[2]

II. REVIEW SURVEY

So many of our favorite daily activities are mediated by proprietary search algorithms. Whether you're trying to find a stream of that reality TV show or shopping on eCommerce site for a new set of things, the relevance of search results is often responsible for your happiness. Currently, small online businesses have no good way of evaluating the performance of their search algorithms, making it difficult for them to provide an exceptional customer experience .Large online retailers typically use query-based search to help consumers find information/products on their websites. They are able to use technology to provide users with a better experience. Because they understand the importance of search relevance, and that long and/or unsuccessful searches can turn their users away because users are accustomed to and expect instant, relevant search results like they get from Google and Amazon. While search is critical to the success of any eCommerce business, it is not always as easy as it seems, in particular, for middle or small online retailers, because it often requires huge volumes of manually labelled data and machine learning techniques. The improvement to searching is not just limited to the exact match. Machine learning can display for a variety of related products, increasing chances For Additional sales[3].

Older, traditional site searches are called “recommender” or “product recommendation” searches. They have little imagination and deliver only results focused on the keyword. Some search apps can’t understand misspelled words returning no results at all. The customer must try again or go else where . In an SLI study, 73% of customers left a site after 2 minutes if they hadn’t yet found what the searches enhanced with machine learning returned a wider choice of results to each query. They map products and interconnect them in new ways. For example, a search for “cat food” returns cat food wet, dry, mat, bowls, container, dispenser, y were searching for. Mobile users are even less patient. Product lid and canned. Adding one of those extra keywords will yield more related choices. The program improves the search results based on the preferences clicked by customers[4].

III. SOLUTION INFERENCES FROM REVIEW SURVEY

In this paper, we were able to show the whole cycle and steps for performing a data analysis on real world data, from data preprocessing until feature prediction. Moreover, as described on the last section, such approaches can be effectively applied on large datasets conversely. Throughout this study, we could testify scikit-learn package, as well as other built-in Python packages, saves substantial development time. With our benchmark on C test set, we could attest that Random Forest is an efficient machine-learning algorithm. It shown better accuracy compared to a simple SVM implementation. This, in part, is due to the ensemble nature of Random Forest, which allows multiple learning algorithms to be run. Beside this fact, the nature of the dataset is also another reason for SVM disadvantage, since such algorithm is likely to provider poorer performances when the number of features is much greater than the number of samples. Finally, we could also testify that the preprocessing step is crucial for the whole data analysis. It also consumes most of the time and implementation efforts needed on the whole analysis. Moreover, our study also shown that preprocessing may be more critical for precision than the machine-learning algorithm itself . Earlier in the research papers the accuracy that was calculated was less but in our model we will try to increase the accuracy by at least 10-15 percent .

IV. CONCLUSION

With our benchmark on CrowdFlower test set, we could attest that Random Forest is an efficient machine-learning algorithm.Random Forest shown better accuracy compared to a simple SVM implementation. This, in part, is dueto:
->The ensemble nature of Random Forest, which allows multiple learning algorithms to be run.
-> Beside this fact, the nature of the dataset is also another reason for SVM disadvantage, since such algorithm is likely to provider poorer performances when the number of features is much greater than the number of samples.

Employing tf-idf for preprocessing features and, Random Forest is a powerful and effective approach for predicting, and measuring, the relevance of text search in e-commerce scenarios. Such approach suits small e- commerce businesses emerged in big data necessities. The preprocessing step is crucial for the whole data analysis because:

-> It consumes most of the time and implementation efforts needed on the whole analysis.

->Preprocessing may be more critical for precision than the machine-learning algorithm itself.[5]

REFERENCES

Proceeding Papers:

[1] Predicting the Relevance of Search Results for E-Commerce Systems Mohammed Zuhair Al-Taie1 , Siti Mariyam Shamsuddin1 , and Joel Pinho Lucas.

[2] Qiqi Wang E-commerce Sites Search Results Relevance Prediction Based on Ensemble Approach

Journal Papers:

[3] https://towardsdatascience.com/predict-search-relevance-using-machine-learning-for-online-retailers- 5d3e47acaa33

[4] https://www.eventige.com/blog/machine-learning-e-commerce-search

[5] https://www.slideshare.net/MohammedTaie/predicting-the-relevance-of-search-results-for-ecommerce- systems