Journal of Engineering and Applied Sciences 12 (Special Issue 12): 9394-9399, 2017
ISSN: 1816-949X

Data Mining Model to Classify Product Search Relevance

Kahkashan Firdous Shaikh and Manoj B. Chandak
Department of Computer Science and Engineering,
Ramdeo Baba College of Engineering and Technology, Nagpur, India

Abstract: Using an efficient search algorithm is valuable for any business providing a variety of products. Customers expect the search results to be accurate and fast. Home Depot an organization has hosted the competition on Kaggle website with an objective to automate evaluation of existing and new search algorithms. Currently, Home Depot has a team of human raters evaluate any changes in the search algorithm. The goal of this project is to accurately identify if a search phrase is relevant given a product title and description.

Key words: NLP, search relevance, crowdsourcing, data mining, website, organization

INTRODUCTION

The past couple of decades has seen an exponential increase in the use of e-Commerce websites (Demirkan, 2015). This increase has given rise to new complications. Kim *et al.* stated that "The rapid growth of e-Commerce has caused product overload where the customer is no longer able to effectively choose the products he/she is exposed to". Customers are willing to spend money to buy things online while they sit home and explore the buyer's websites. This has increased the expectation of the customers and the level of competition among the organizations. All organizations want their customers to come back to their websites and make sure customers have a good shopping experienceand customers expect to find what they need without any hardship.

Relevance labels and crowdsourcing: Many organizations use search relevancy as a measure to determine how quickly they can get customers to the right product. It is important that the word relevance being defined in terms of search engines and information retrieval. Merriam Webster defined relevance as "the ability (as of an information retrieval system) to retrieve material that satisfies the needs of the user". Traditional search engines use to return results based on the keywords matches used in user's query (Chowdhury, 2003). Such relevance was a replacement of recall concept (http://www.seobythesea.com/2012/07/relevance-search-engines/). For the current study, the assignment of relevance labels not only depend on the matches of keywords but also the important a search term holds for a product that a customer might use to search. Crowdsourcing has emerged as a viable alternative of gathering relevance labels in comparison to the traditional methods for the assessment of search engines. Kazai stated that this alternative offers a "solution to the scalability probleChom that hinders traditional approaches". For instance, a traditional approach known as Cranfield paradigm as cited by Carvalho *et al.* (2011) that depends on human judges to assess documents for topical relevance used in the evaluation of IR systems which tends to be slow, tedious and expensive. Carvalho *et al.* (2011) stated that "Crowdsourcing represents a promising new avenue for reducing effort, time and cost involved in evaluating search systems". However, this new approach poses new challenges and have its own limitations. In their study Carvalho *et al.* (2011) mentioned some challenges that include interaction with researchers, ensuring result's quality and aggregate crowd producing better annotation in less time and cost. Song *et al.* (2011) in their research to generate true relevance using "click through" data mentioned that the collection of superior quality labels for commercial search engines is expensive, time-consuming and labor-intensive.

Motivation to study : The prime motivation for such model development is the problem that some organizations for example Home Depot face where the impact of changes to the search algorithm are evaluated by human raters which is time-consuming (Song *et al.*, 2011; Clough *et al.*, 2013). Hence, this study is an effort to make this process efficient and increase the iteration count on the search algorithm. Some other challenges faced by the organizations are the dynamic, global and unpredictable business environment in which they operate. Customers prefer to shop online because of the flexibility and

Coressponding Author: Kahkashan Firdous Shaikh, Department of Computer Science and Engineering,
Ramdeo Baba College of Engineeringand Technology, Nagpur, India

convenience that these e-Commerce websites offer. It is essential for any e-Commerce website to offer their customers with results that should satisfy their request in term of effort, time and accuracy. For example, if a customer searches for a term "X" and the return result contain products organized based on the relevance of "X" for the given product will aid sellers to provide a customer with products quickly, accurately and efficiently which improves customer's shopping experience for relevance labels is widely accepted, it has received mixed opinions from different researchers. The two major problem with this approach is the quality of results and time-consumption. After the gathering of relevance labels, these labels are provided to human raters for rating. And based on these rating appropriate adjustment is made to the search algorithm. But this rating by the crowd is time-consuming, at least for organizations like Home Depot who wants to automate their rating system with some models that can predict efficiently and with better or equal quality in comparison to human raters.

MATERIALS AND METHODS

Traditional approaches and crowdsourcing: People today takes advantage of the systems that aid them in gathering or retrieving desired information. One such technology is the Information Retrieval (IR) that has made it possible to find required information with efficiency. Carvalho *et al.* (2011) stated that "development of these technologies has historically depended on slow, tedious and expensive data annotation". These systems depend on the relevance evaluation for information retrieval and relevance evaluation is a notoriously laborious and expensive job/task (Alonso *et al.*, 2008). During its early year, relevance evaluation involves the determination of the relevance of a test query which is done by some graduate student volunteers reading through every document in a corpus. This process was difficult and only a small set of test collection used to be created (Alonso *et al.*, 2008).

TREC (Text retrieval Conference) in 1992, represented a modern demonstration of Canfield methodology, that made the test collection with millions of full-text documents available to the researchers. However, a different approach, named polling was developed where instead of reading and evaluating every document, only top N documents retrieved by the participating system were investigated. Alonso *et al.* (2008) mentioned that TREC collections has been crucial in furthering IR research for a long time but these collections possess some limitations. They continued mentioning the obvious limitation that these collections have a limited study for different kind of IR tasks (Chowdhury, 2003). Furthermore, a different approach was required for the evaluation of general web search which possesses unique challenges. These reasons lead the use of editorial resources to develop relevance assessment specific to the needs of the system. Large editorial staffs are involved in judging the relevance of a web page which are hired by web search engines. Even in academic research, researchers depend on student participant for this relevance assessment. Some researchers suggested to harness user behavior as an evaluation signal. Rose pointed out that determination of the quality of web search result snippet cannot be done using user click behavior, since, it is not certain that the lack of a click represents a perfect snippet or a poor one that satisfy or fails the user's information need (Baccianella *et al.*, 2009).

However, this new approach poses new challenges and have its own limitations. In their study, Carvalho *et al.* (2011) mentioned some challenges that include interaction with researchers, ensuring result's quality and aggregate crowd producing better annotation in less time and cost (Chowdhury, 2003). In their research study Le *et al.* shows how dynamic learning environment affect the researchers results in examining search relevance task using Amazon Mechanical Turk. Song *et al.* (2011) in their research to generate true relevance using click-through data mentioned that the collection of superior quality labels for commercial search engines is expensive, time-consuming and labor-intensive.

Introduction of TF-IDF: As we had discussed lot more about the new approaches there has been an algorithm which had used to overcome the drawback of crowdsourcing. From the syllabus of information retrieval an TF-IDF algorithm has been used by us to rank and determine the relevance of the search term. Term frequency-inverse document frequency is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The TD-IDF algorithm has been used for its accuracy and it used with the vector space model which creates term-document matrix that describes the frequency of terms that occur in a collection of documents (Brill, 1995). In the vector space model documents and queries are represented as vectors:

$$D_j = \left(w_{1,j}, w_{2,j}, \cdots, w_{t,j}\right)$$

$$Q = \left(w_{1,q}, w_{2,q}, \cdots, w_{n,q}\right)$$

where, D and Q represent document and queries, respectively. Relevance rankings of documents in a keyword search can be calculated of document similarities

theory by comparing the deviation of angles between each document vector and the original query vector where the query is represented as the same kind of vector as the documents. The term similarity can be find by performing three major steps:

Finding TF-IDF weight (n is given):

TF(t) = (Number of times term t appears in a document)/ (Total number of terms in the document)
IDF$_j$ = Log [n/df$_j$]

Length normalization:

$$\left\|\vec{x}\right\|_2 = \sqrt{\sum_i x_i^2}$$

Dot product of the search terms:

$$\cos(\vec{q},\vec{d}) = \vec{q} \bullet \vec{d} = \sum_{i=1}^{|v|} q_i d_i$$

By applying these steps and vector space algorithm, we finally get the appropriate values of the search term.

Purpose of study: For this research the problem is to build a predictive model for the evaluation of search labels. This challenge is faced by the Home Depot , a firm that deals with selling home requirement products. In their current system, numerous human raters are provided with the product/search labels to evaluate relevance scores. Kaggle is an online platform which allows companies to crowd-source their effort to find the best data mining models for any problems they might be facing. Home Depot has hosted one such competition on Kaggle for ranking search results. Home Depot is an American retail company specializing in home improvement and construction supplies and has stores in all US states, Canada, Mexico and China. Home Depot crowdsource the product/search labels which contain products detail and real customer search labels/terms for its website. Furthermore, each pair is evaluated by at least three raters and the relevance score is the average of these ratings. Home Depot wants to provide its customers with a better shopping experience and minimize the human rater's involvement by automating the relevance score generation in their current system. Data for this research will be collected from Home Depot to study their current rating scheme and variables (Cheung *et al*., 2003).

The objective of the project is to find a model that can best predicts the relevance of each search term and product pair, i.e., predict the relevance of a product given a search keyword. The dataset provided by the Home Depot includes both training and testing sets in CSV format. In addition, product descriptions and product attribute descriptions are provided separately. We will be using only one of these additional files:

Product descriptions: Use of product description for the model is optional. However, product descriptions contain more information that the name of the product which are shorter and may contain a trade name. Hence including the product descriptions would enhance the accuracy of our model.

Product attributes: Product attribute descriptions are only available for some products in unstructured text format, to simplify the model, we have decided to exclude product attribute descriptions.

The data description table provides a description of how we used an attribute in our project. Since, most of the data is provided in unstructured text format, appropriate preprocessing is required before they can be analyzed. Preprocessing unstructured text data will most likely yield a sparse dataset with missing values. For instance, for a specific product description, punctuation marks in its value can be considered as noise and should be removed for the efficient working of the tokenization process, searching algorithm. Secondly, the current rating available from the organization contains numerical values between 1-3. But the requirement of the organization is to classify the label's relevance as "relevant", "not relevant" and "most relevant". Therefore, it is required to transform these numeric values to nominal values which are necessary in order to work on this classification problem (Manning *et al*., 2009). The goal of this project is to accurately identify if a search phrase is relevant given a product title and description and to compare the result of crowd source score and by the program score, i.e. whether the score matches accurately or not.

Implementation

Plan: Planning for this project involves consideration of each activity that accustomed the data mining task. The following list presents the different activities that this project will undertake.

Data collection (Brill, 1995): Data will be downloaded from the Home Depot's (https://www.kaggle.com/c/home-depot-product-search-relevance/data) data source. The organization has provided the files, containing current ratings and attributes on the Kaggle's website which is available to people who are interested in participating in the competition to solve their problem.

Data exploration (Cheung *et al.*, 2003): Data can be of any type and it is very crucial to comprehend the data and its properties at the early stage of any project to be able to utilize the data to its full extent. For example, a given field can be of a numeric type or nominal type that distinct it from each other on what values it can deal with and the operations that can be performed on that field. The files available at the Home Depot source are in CSV format and the fields mostly contain textual data.

Data preprocessing: This is the most crucial stage of this project and involves major task of data cleaning and data transformation. As mentioned before the files available at the website hold fields with textual data, it is important to clean this data, so that, every field have textual tokens for comparison which is the requirement for coding and model development. For example, a file contains fields such as "product_title" and "search_term" which might have punctuation marks in their values. These can be considered as noise in the actual values and should be removed for the efficient working of tokenization process, searching algorithm.

As defined earlier that the organization has provided the dataset which is in unstructured format, so first, we had done a task of bringing the data into the structured format. Next is the step where we had applied a weighting algorithm to rank and match the similarity between the terms, i.e., the TF-IDF algorithm. TF-IDF calculation was used for calculating number of times a word appeared in document. We calculate cosine similarity for finding similar matches. Study with phenomena, we will find a result with respect to the two terms. Further processing on the whole collection, we used vector space model where documents as well as queries are represented as vectors. For this, we had made two separate files one consisting of product-ID, product name and product description and another consisting of product-ID, product name and search term. We had compare these two files together with respect to the algorithm where product-id is used to match between two files. Creating a dataset which contain all the product-ID and search term which help in searching a right product. Searching for one term we get the list of all the product name in which the search term exists. So, we had used a java swing to create a small user interface which result like (Table 1).

After clicking on any of the product name a whole description of the product with respect to the relevance score generated by the algorithm for each search terms. For processing, a vocabulary is created which stores all the non-repeated term and assign weights according to their occurrences in the collection of the documents. The vocabulary here plays a major role in calculating or assigning weight to the term. Single occurrence of a term specifies one point to a term and that term is to be inserted in the vocabulary list (Chowdhury, 2003). This process continues till the whole terms are processed. If in case the same term occurs, then only the point or weight of the term increases and term will not be added in the vocabulary again as it is already available in the vocabulary (Fig. 1 and 2).

Table 1: File description

File name	Attribute	Description	Data type	Remarks
Train.csv (N = 74000)	ID	Identifies each row in the CSV file	Numeric	
	Product _UID	Uniquely identifies a product	Numeric	Used for joining the two CSV files
	Product title	Name of the product	Unstructured text	
	Search term	Search query used by the users	Unstructured text	Extracted for preprocessing
Product _description.csv	Product UID	Uniquely identifies a product	Numeric	Used for joining the two files
	Product description	Provides detailed description of the product	Unstructured text	

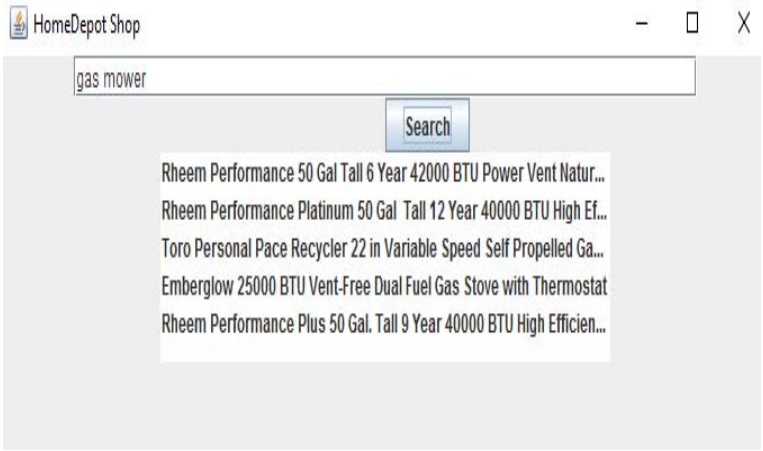

Fig. 1: Displaying list of product

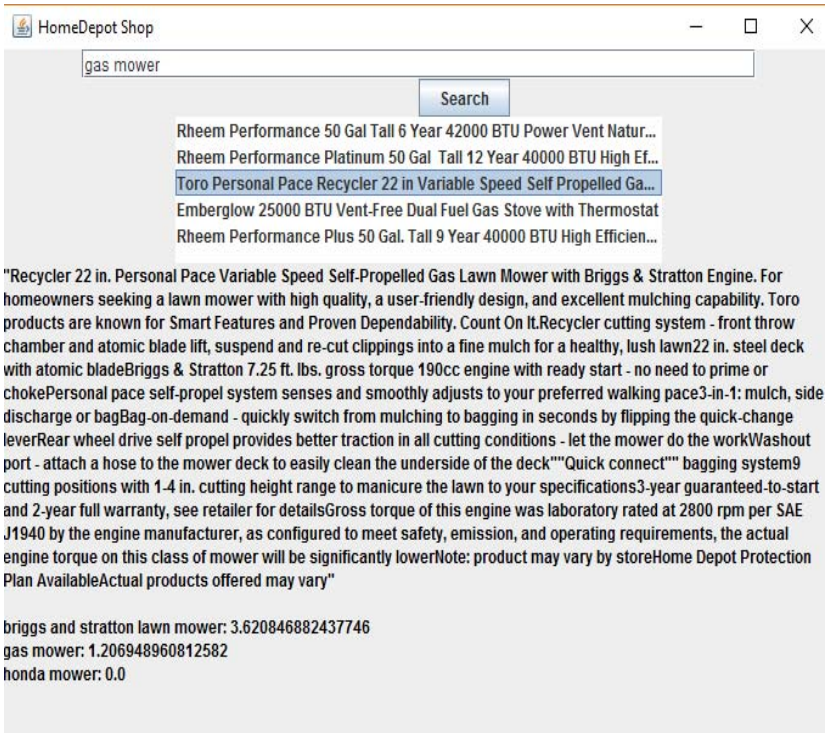

Fig. 2: Description of particular product with their search terms score

Experimental setup: The system is built using Java framework (Version jdk 7) on windows platform. The NetBeans (Version 8.1) is used as a development tool. The system doesn't require any specific hardware to run, any standard machine is capable of running the application.

RESULTS AND DISCUSSION

In this study an analysis had been performed on the outputted result of the search terms. The analysis had done on the basis of how many products correspond to a single search term. This has been shown by a bar chart representation with the help of Tableau Software where on the y-axis there has been defined a numbers which indicates the number of products and on the x-axis there has been written a search term name where each search term specifies that it belongs to how many number of products (Fig. 3).

Furthermore, it is important that the model should not only display good accuracy with training data but also with the testing data where relevance score for instances is missing. It is also important that more than one classification algorithms be applied to gain a better understanding of which algorithm provides the best accuracy. Specifically, the objective of this project is to find answers to the following questions.

- Is the model capable of predicting relevance with the same accuracy on test data?
- Which data mining algorithm provides the best accuracy for the given data set?

Potential benefits: Customer's procurement experience can be improved if the model can successfully predict the relevance of the search result. Another potential benefit is the customer satisfaction, both in term of effort and efficiency. As the model will allow quick retrieval of preferred products based on search relevance prediction, the level of a customer satisfaction increase. Finally, as the model will be used to predict the search relevance, human rater's involvement which is subject to quality of the search label and is time-consuming, can be minimized.

Applications: This approach to analyze reviews with help of data mining technique is an attempt to improve the usability of crowdsourced data. From business and marketing point of view these analytics is very crucial. Such kind of analysis is becoming very important with the growth in social media. Preprocessing and clustering these reviews remove the unwanted and redundant data respectively. This technique is useful for generating reports and will help in decision making for the analyst.

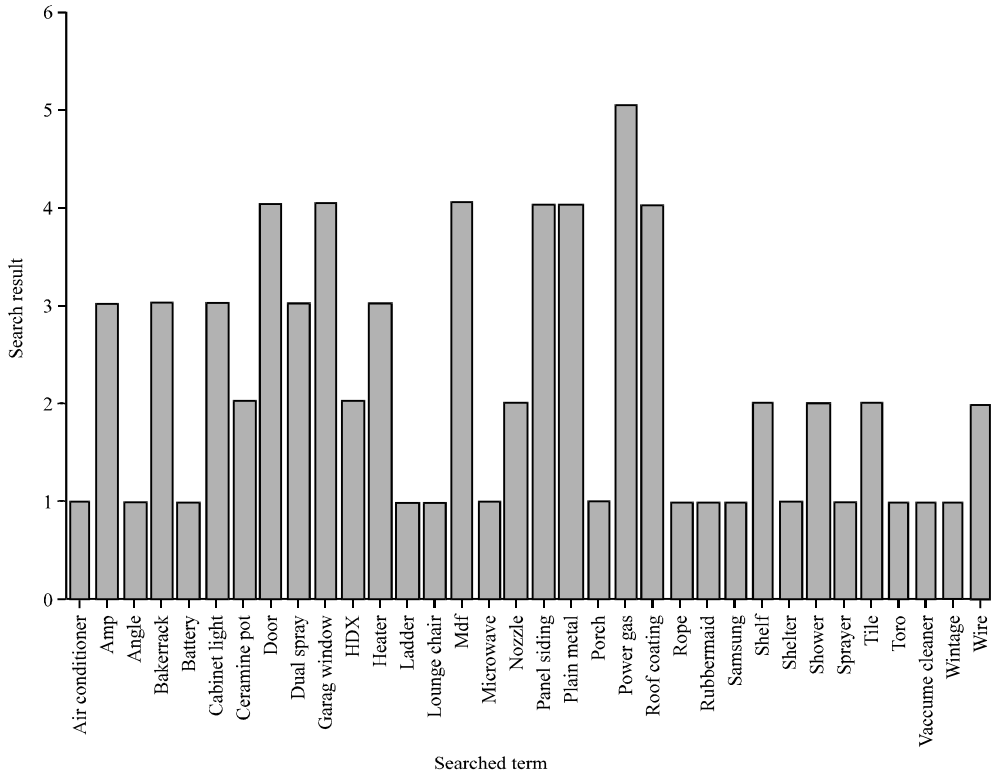

Fig. 3: Result of analysis of search term

CONCLUSION

In this research, our main scope was to find out the relevance of a search term in comparison of the crowdsource people. Further research on the dataset by applying TF-IDF algorithm and found result to be more relevant than the crowdsource people.

RECOMMENDATION

Further research on classifying the relevance labels has been under process and also a better approach will be used to accurately match the result.

REFERENCES

Alonso, O., D.E. Rose and B. Stewart, 2008. Crowdsourcing for relevance evaluation. ACM. SIGIR. Forum, 42: 9-15.

Baccianella, S., A. Esuli and F. Sebastiani, 2009. Multi-Facet Rating of Product Reviews. In: Advances in Information Retrieval, Boughanem, M., C. Berrut, J. Mothe and C.S. Dupuy (Eds.). Springer, Berlin, Germany, pp: 461.

Brill, E., 1995. Transformation based error driven learning and natural language processing: A case study in part-of-speech tagging. Comput. Ling., 21: 543-565.

Carvalho, V.R., M. Lease and E. Yilmaz, 2011. Crowdsourcing for search evaluation. ACM. SIGIR. Forum, 44: 17-22.

Cheung, K.W., J.T. Kwok, M.H. Law and K.C. Tsui, 2003. Mining customer product ratings for personalized marketing. Decis. Support Syst., 35: 231-243.

Chowdhury, G.G., 2003. Natural language processing. Ann. Rev. Inf. Sci. Technol., 37: 51-89.

Clough, P., M. Sanderson, J. Tang, T. Gollins and A. Warner, 2013. Examining the limits of crowdsourcing for relevance assessment. IEEE. Internet Comput., 17: 32-38.

Demirkan, H., 2015. Special section: Enhancing e-commerce outcomes with IT service innovations. Intl. J. Electron. Commerce, 19: 2-6.

Manning, C.D., P. Raghavan and H. Schutze, 2009. Introduction to Information Retrieval. University of Cambridge, Cambridge, England, UK.,.

Song, H., C. Miao and Z. Shen, 2011. Generating true relevance labels in chinese search engine using clickthrough data. Proceedings of the 25th AAAI Conference on Artificial Intelligence, August 7-11, 2011, AAAI Press, San Francisco, California, pp: 1230-1236.