Amazon Cloths Recommender

by Gopika Vinod

ABSTRACT

Most recommender systems are still largely dependent on user ratings to find the most likely product that consumers want to buy. In this project, we present an advanced recommender system based on the visual and textual representation of the product. This approach helps us extract certain valuable features of the products that are not possible with a conventional recommender system. We have used the json dataset for over 180,000 apparel images available on Amazon and recommended similar clothing using content based search. In this paper, we reviewed previous work on visual and content based recommendations. We applied VGG16 Convolutional Neural Network on product images and Bag of Words and TF-IDF on product titles. We recommended top products to the users using these techniques and finally conducted an empirical evaluation of these methods.

KEYWORDS

Visual Recommendation, Textual Recommendation, VGG Convolutional Network, TF-IDF, Bag of Words, Amazon Recommender

1 INTRODUCTION

On the Internet, where the number of choices is overwhelming, there is need to filter, prioritize and efficiently deliver relevant information in order to alleviate the problem of information overload, which has created a potential problem to many Internet users. Recommendation algorithms solve this problem by searching through large volume of dynamically generated information to provide users with personalized content and services[1]. Recommendation algorithms are best known for their use on e-commerce Web sites, where they use input about a customerfis interests to generate a list of recommended items. 34.4% of Amazon's revenue is generated by its Recommendation Engine[2]. This shows us how important recommendation systems are in generating profits for any company.

In this project i have built an Amazon Cloths Recommender which uses the following models: VGG Convolutional Neural Network(CNN), TF-IDF and Bag of Words to give us the top recommendations of items most similar to the selected item. We are using TF-IDF and Bag of Words on product titles and CNN on images for recommendations. We have taken Women Apparel Data of Amazon from Kaggle.com which is in json format and consists of 183138 data points and 19 features. Out of the 19 features we take into consideration 7 features for recommending items to the user. We performed preprocessing on our data by removing values with NULL price and colour and thereafter removing duplicates. We also did text preprocessing to remove stop words, special characters,etc. We then do feature extraction and obtain the vector representation of the products using the 3 techniques mentioned above. Then we perform cosine similarity to obtain the similarity between the product selected by the external user and the remaining products in the database. The last step is recommending products to the user and checking the accuracy of our models.

To summarize,

1. Visual Recommendation using:
   - VGG Convolutional Neural Network
2. Textual Recommendation using:
   - TF-IDF
   - Bag of Words

2 RELATED WORK

The work done in this area includes [3] which is an apparel recommendation system which provides its recommendation using decision trees. [4] is a popular amazon product recommender research paper which talks about item to item collaborative filtering and the scalability of the recommender system.[5] explains the working of a visual recommender system based on styles and attributes. [6] gives a detailed explanation of different recommender systems based on neural networks mainly CNN. eliminating the limitations of the previous works by focussing upon visual and textual representation of the product rather than just focussing upon the ratings given by the user.

3 PROBLEM FORMALIZATION

Following is the mathematical significance of the research problem and approach:

3.1 VGG16 Covolution Neural Network Model

VGG-16 is a multi-layer neural networks model designed primarily to extract features from images. It always uses 3 x 3 filters with stride of 1 in convolution layer and uses SAME padding in pooling layers 2 x 2 with stride of 2.[7]

3.2 TF-IDF

3.2.1 *Term Frequency.*

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

It is the ratio of number of times the word appears in a document compared to the total number of words in that document.[8]

3.2.2 Inverse Document Frequency.

$$idf_i = \log \frac{|D|}{|d : t_i \in d|}$$

It is used to calculate the weight of rare words across all documents in the corpus. The words that occur rarely in the corpus have a high IDF score.[8]

3.3 Bag of Words

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves a vocabulary of known words and a measure of the presence of known words.[9]

3.4 Cosine similarity

Cosine similarity measures the similarity between two vectors by calculating the cosine of the angle between them.[10]

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \quad (1)$$

4 PROPOSED MODEL

The flowchart of our system is shown in Figure 1 Firstly, we took Women Apparel Data of Amazon from Kaggle.com. We then choose a subset of these features for our recommender system which are:

(1) asin (Amazon standard identification number)
(2) color (Color information of the clothes, it can contain many colors as a value ex: red and black stripes)
(3) brand (brand of the product)
(4) product_type_name (type of the apparel, ex: SHIRT/TSHIRT)
(5) medium_image_url (url of the image)
(6) title (title of the product)
(7) formatted_price (price of the product)

We then did general preprocessing which included removing products with NULL price and color and eleminating duplication by deleting products with almost similar titles. Text processing is then done before feature extraction to obtain more accurate results. After preprocessing we use the three models: VGG CNN (for feature extraction from images) and TF-IDF and Bag of Words (for feature extraction from product titles). After this step we find cosine similarity between the product selected by the external user and the rest of the products and obtain the top 10 recommendations for each of the 3 models described above. We then compare our three models using euclidean distance measure to compare their performance and accuracy. Finally, we conclude by stating the results from our built models.

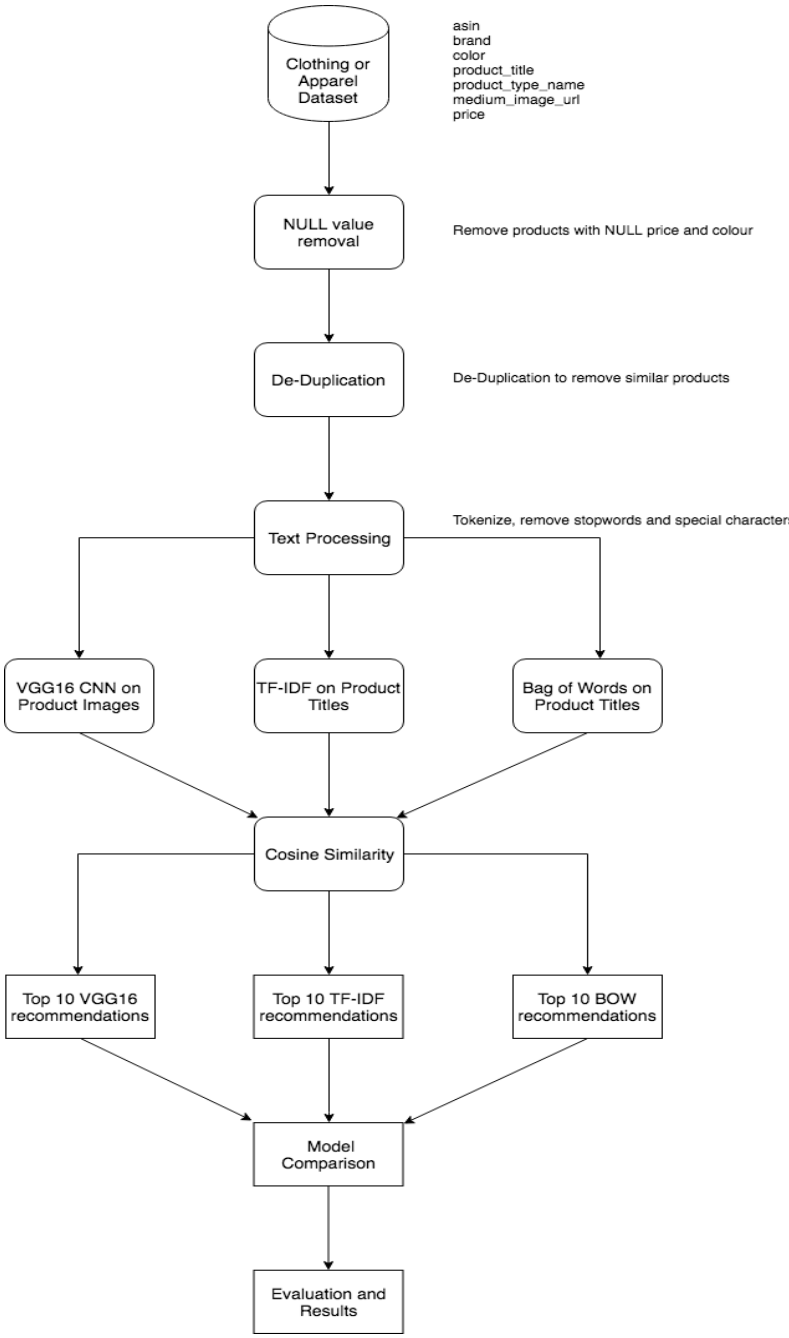

Figure 1: Flow Diagram of the System

5 EXPERIMENTS AND RESULTS

my Amazon Women Apparel Dataset is json format so we loaded the dataset using Pandas. performed the following analysis and modelling:

(1) Data consists of 183138 data points and 19 features.
Out of the 19 features we take into consideration 7 features for recommending items to the user.

(2) Data preprocessing was performed by removing the NULL values by eliminating the rows with NULL product price or color, also we eliminated the entries with a very short description in the product titles.
Number of data points with price and color = NULL is 7871

(3) We eliminated the products which have duplicate titles, so if the number of words in which both strings (product

titles) differ are less than 3 , we consider them as same and hence we ignore them.
For example:
Nike Ultra Strong Women's Shirt X-Large
Nike Ultra Strong Women's Shirt Small
We remove one product out of the above two product titles because they are almost equal and hence they add no value in our analysis.
The number of entries with duplicate title is 2325.

(4) Text processing for NLP is done on the semi structured data so that it could be fed into our recommendation algorithms. In this process, we tokenize followed by removing stop words and special characters.

(5) The main focus is on building a recommendation algorithm and we have dealt with it in two ways. First method is obtaining vector representation of product images by doing feature extraction. This is done by applying VGG16 convolutional neural network model on the product images. In the Second and Third method, vector representation of products is done by application of bag of words and TF-IDF on the product titles.

(6) For each of the three methods described above, then applied cosine similarity to get similarity between a given product id given by an external user and all the products in our dataset.

Figure 2: Inputted Clothing Item

(7) Then we sorted the products on the basis of their similarity score and the top 10 items are recommended to the user for each of the proposed methods.
A small snippet of how our recommendation system works can be seen in Figure 2 and Figure 3. Here Figure 2 displays the selected item and Figure 3 shows the items recommended to the user when using TF-IDF model.

(8) Comparison between the three different models is then done using Euclidean distance measure. We observe in Figure 4 which is the line chart of the euclidean distance

Figure 3: Recommended items

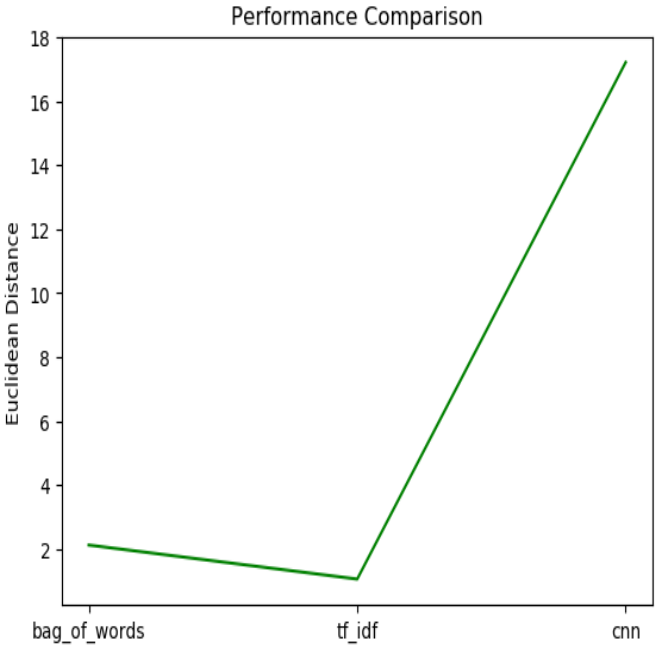

Figure 4: Line Chart of Euclidean Distance Measure

measure of the three models that the average euclidean and manhattan distance of the recommended products from the product selected by the user is highest in case of VGG16 CNN model and is the lowest for TF-IDF model. It means that TF-IDF is performing the best in terms of recommending products and the CNN model is not able to perform well for this use case. Figure 5 which is a bar chart of the same plot as in Figure 2 is plotted to give a clear idea of the euclidean distances of the three models.

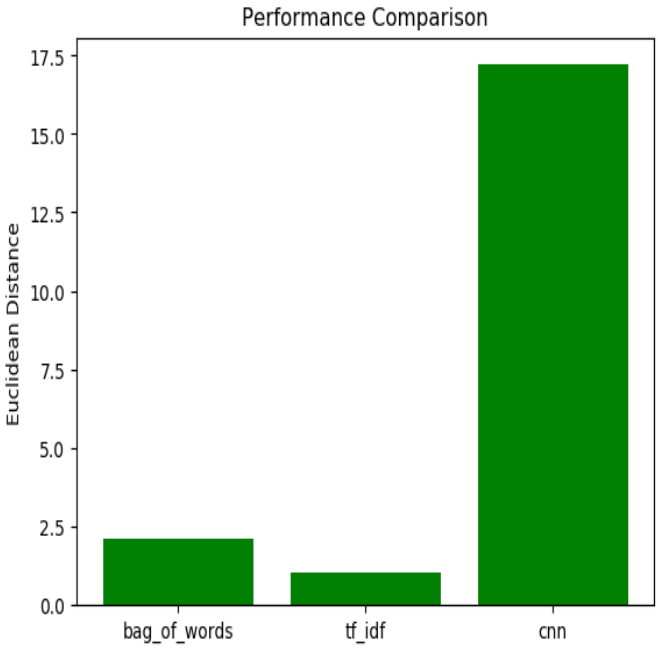

Figure 5: Bar Chart of Euclidean Distance Measure

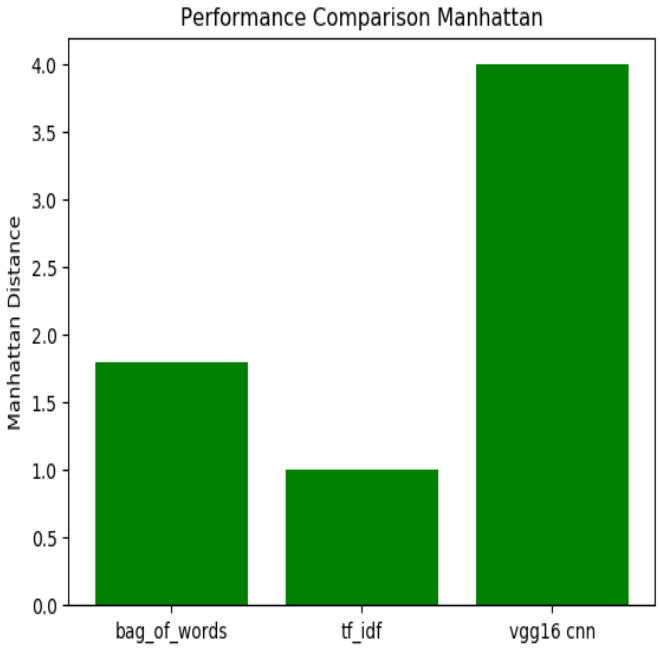

Figure 7: Bar Chart of Manhattan Distance Measure

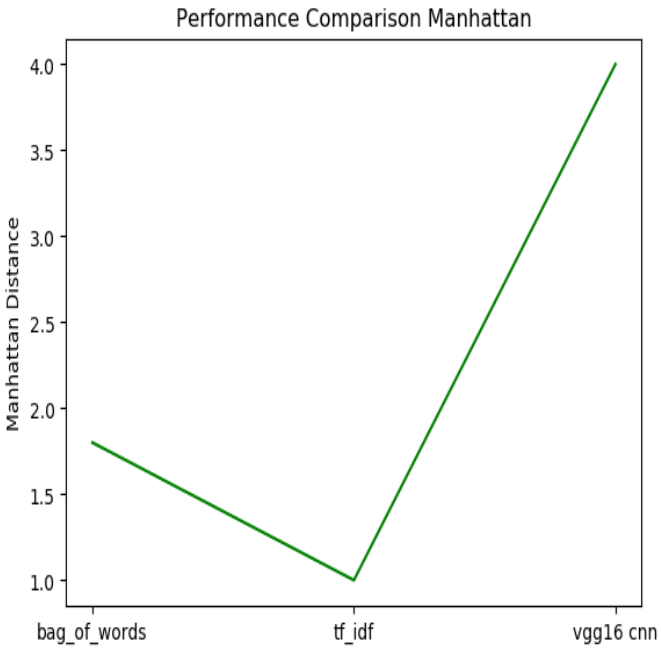

Figure 6: Line Chart of Manhattan Distance Measure

6 CONCLUSIONS AND FUTURE SCOPE

We have built a different kind of recommender system which doesn't focus upon the ratings given by the user but extracts the unknown features from the product's visual and textual properties. For this, here used VGG16 CNN on product images and TF-IDF and Bag of Words on product titles. On comparing the performances of these three models using Euclidean Distance Measure, found that TF-IDF gave the best performance followed by Bag of Words and then CNN.

In the future, we can introduce ResNet CNN for visual recommendation and Word2Vec for textual recommendations and compare their performances with our existing models.

We can scale up the system for all the categories of Amazon products by using the Apache Spark framework. We can also investigate the trend of time complexity by increasing the number of nodes in the cluster step by step.

ACKNOWLEDGEMENT

We would like to express our very great appreciation to professor Yongfeng Zhang and also Yunqi Li and Hanxiong Chen for their valuable and constructive guidelines during the planning and development of this project.

REFERENCES

[1] F.O.Isinkayea Y.O.Folajimib B.A.Ojokohc Recommendation systems: Principles, methods and evaluation 2015

[2] https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

[3] Xinjuan Zhu ; Junfang Huang ; Yang Qi An Apparel Recommender System Based on Data Mining

[4] Greg Linden, Brent Smith, and Jeremy York Amazon.com Recommendations Item-to-Item Collaborative Filtering 2003

[5] Julian McAuley Christopher Targett Qinfeng (fiJavenfi) Shi Anton van den Hengel Image-based Recommendations on Styles and Substitutes

[6] RUIHUI MU A Survey of Recommender Systems Based on Deep Learning

[7] https://medium.com/@sidereal/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5

[8] https://medium.freecodecamp.org/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3

[9] https://medium.freecodecamp.org/an-introduction-to-bag-of-words-and-how-to-code-it-in-python-for-nlp-282e87a9da04

[10] https://medium.com/@mark.rethana/building-a-song-recommendation-system-using-cosine-similarity-and-euclidian-distance-748fdfc832fd