A B.Tech Project Report

on

E-Commerce Recommender System

Submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF TECHNOLOGY

(Computer Science and Engineering)

BY

1.VAMSI KRISHNA(2016KUCP1028)

2.GUMMADI MANOJ KUMAR(2016KUCP1016)

3.VUCHURU PURUSHOTHAM(2016KUCP1024)

UNDER THE GUIDANCE OF:-

Dr.Amit Kumar

INDIAN INSTITUTE
OF INFORMATION
TECHNOLOGY, KOTA

DECLARATION

We hereby declare that the work reported in the Bachelor of Technology report entitled **E-commerce Recommender System** submitted at Indian Institute of Information Technology, Kota is an authentic record of our work carried under the supervision of Dr.Amit Kumar. We have not submitted this work anywhere else for any other degree.

GUMMADI MANOJ KUMAR

(2016KUCP1016)

VUCHURU PURUSHOTHAM

(2016KUCP1024)

VAMSI KRISHNA

(2016KUCP1028)

CERTIFICATE

This is to certify that the report entitled, **"E-commerce Recommender System"** which is being submitted by Vuchuru Purushotham,Vamsi krishna,Gummadi Manoj kumar in fulfilment for the award of degree of B.Tech in Computer Science and Engineering by the Indian Institute of Information Technology, Kota is the record of candidates own work carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree.

Dr.Amit Kumar
Dept of CSE
IIIT KOTA

ACKNOWLEDGEMENT

We are profoundly grateful to Amit Kumar for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement. His timely and efficient contribution helped us shape our work into its final form and we express our sincerest gratitude for his assistance in any way that we may have asked. We appreciate his guidance in our project that has improved our project many folds, thanks for the comments and advise. We would also like to thank all other faculty members of IIIT Kota for their direct or indirect supports and advise without which this project would not have been possible.

Contents

1 ABSTRACT

Internet is speeding up and modifying the manner in which daily tasks such as online shopping, paying utility bills, watching new movies, communicating, etc., are accomplished.The shift to online shopping has made it incumbent on producers and retailers to customize for customers' needs while providing more options than were possible before.E-commerce is an online trading system that eases transactions for both sellers and consumers without having to meet in person. The prevalence of e-commerce has increased competition amongst sellers, hence the users of e-commerce has to increase their performance, one of them by using recommendation system.This research develops a hybrid recommendation system for e-commerce that implements Content-based Filtering and Collaborative Filtering methods, which will compute the similarities of product description (Based on Title,Brand and color). In experiment results, we found the recommendations for the similar products.In this scenario, we discuss about common recommender systems techniques that have been employed and their associated trade-offs.

2 INTRODUCTION

2.1 What is a Recommender system?

With the popularity of Internet and rapid development of e-commerce, many well-known e-commerce sites such as Amazon,Flipkart,E-bay etc., developed a recommendation system to provide personalized information recommendation services for customers. E-commerce recommendation system is used by e-commerce sites to provide goods information and advice to customers, and simulate the shop sales workers to help customers successfully completing the purchase process.

E-commerce recommendation system has been greatly developed both in theory and practice, especially the research of the recommended method is the core of which, so to adopt which recommended method is essential for the effectiveness and efficiency of the recommendation system.

2.2 Why use Recommender system ?

With the continuous growth of the intenet and the progress of electronic commerce the issue of product recommendation become increasingly important.

Revenue

With years of research, experiments and execution primarily driven by Amazon, not only is there less of a learning curve for online customers today. Many different algorithms have also been explored, executed, and proven to drive high conversion rate vs. non-personalized product recommendations.

Customer Satisfaction

Many a time, customers tend to look at their product recommendation from their last browsing. Mainly because they think they will find better opportunities for good products. When they leave the site and come back later; it would help if their browsing data from the previous session was available. This could further help and guide their e-Commerce activities, similar to experienced assistants at Brick and Mortar stores. This type of customer satisfaction leads to customer retention.

Personalization

We often take recommendations from friends and family because we trust their opinion. They know what we like better than anyone else. This is the sole reason they are good at recommending things and is what recommendation systems try to model. You can use the data accumulated indirectly to improve your website's overall services and ensure that they are suitable according to a user's preferences. In return, the user will be placed in a better mood to purchase your products or services.

Discovery

For example, the "Genius Recommendations" feature of iTunes, "Frequently Bought Together" of Amazon.com makes surprising recommendations which are similar to what we already like. People generally like to be recommended things which they would like, and when they use a site which can relate to his/her choices extremely perfectly then he/she is bound to visit that site again.

Provide Reports

Is an integral part of a personalization system. Giving the client accurate and up to the minute, reporting allows him to make solid decisions about his site and the direction of a campaign. Based on these reports clients can generate offers for slow moving products in order to create a drive in sales.

A few months ago,Netflix estimated that its recommendation engine is worth a yearly 1 billion dollars.

2.3 Various methods of Recommender system.

Recommended methods include knowledge engineering, content-based recommendation methods, collaborative filtering recommendation method, hybrid recommendation methods and data mining. Content-based filtering methods are based on a description of the item and a profile of the user's preferences.[41][42] These methods are best suited to situations where there is known data on an item (name, location, description, etc.), but not on the user. Content-based recommenders treat recommendation as a user-specific classification problem and learn a classifier for the user's likes and dislikes based on product features.At present, the collaborative filtering method is the most successful recommendation approach . With the growmg size of e-commerce system, collaborative filtering recommendation method is also facing many challenges, such as recommended quality, scalability, data sparseness, cold start problems, and so on.Collaborative filtering is based on the assumption that people who agreed in the past will agree in the future, and that they will like similar kinds of items as they liked in the past. The system generates recommendations using only information about rating profiles for different users or items. By locating peer users/items with a rating history similar to the current user or item, they generate recommendations using this neighborhood.There is no reason why several different techniques of the same type could not be hybridized. Hybrid approaches can be implemented in several ways: by making content-based and collaborative-based predictions separately and Evaluation is important in assessing the effectiveness of recommendation algorithms. To measure the effectiveness of recommender systems, and compare different approaches, three types of evaluations are available: user studies, online evaluations (A/B tests), and offline evaluations.The commonly used metrics are the mean squared error and root mean squared error, the latter having been used in the Netflix Prize.

3 LITERATURE REVIEW

In the literature review,we have brought forward the surveys,theories and methodologies given by other people in their research that has helped us form the objective of our project.

3.1 Recommendation System

“.(6)How to build a high quality Recommender system?”, S.Khusro et al.

The world wide web has brought numerous changes in a way that every individual is depending on the technologies like web search engines that search and retrieve relevant information on almost any aspect of life.To provide this relevant and reliable information we need to develop the high quality recommender system to cope with the issues in providing relevant suggestions to the users.

This paper discussed how to develop a good recommender system by identifying some of the prominent issues and challenges in the design and development of a fine-tuned Recommender system.The author also discussed about the solutions,techniques and research guidelines that might help in coping withsome of this issues and in designing a fine -tuned recommender system.

Some issues and challenges sited by the author in this paper

- Synonymy
- Privacy
- Limited content analysis and over specialization
- Grey sheep
- Latency

The author also explained about different filtering techniques that are used for recommender system where in each of them is suggested a suitable approach to develop them.

The author concluded the following results:

- The more we focus on the issues and the challenges the less we suffer in future.
- Recommender systems have been used among the many evaluable solutions in order to medicate information and cognitive overload problem by suggesting related and relevant items to the users.

“.(2)A Hybrid Recommender system for E-commerce based on product description”, Tessy badriyah,Erri try Wijayanto,Iwan Syarif ,Prima Kristalina.

To meet the expectation of the users on e-commerce platform the developers of the software had to come up with a futuristic model.For this the developer use the hybrid Recommender model which is a combination of content based and collaborative filtering techniques.

The author worked on two purposes mainly in developing a prototype of hybrid recommendation system.

- Text mining method to generate tags automatically based on product description.
- The system created combines generate tag results automatically with the user profile to get relavant recommendations.

To implement these purposes the author used TF-IDF method as a content based filtering and cosine similarity ,correlation similarity as a collaborative filtering technique.

The author concluded by showing the increament of precision value by this model and pointing out the advantage of creation of tag and their dynamic nature.He stated the size of precision and recall precision also depends heavily on what is really meant by relavant products and how to ensure whether a document is relavant or not.

3.2 Content Based Approach

Content-based filtering methods are based on a description of the item and a profile of the user's preferences. These methods are best suited to situations where there is known data on an item (title,brand,color, etc.), but not on the user. Content-based recommenders treat recommendation as a user-specific classification problem and learn a classifier for the user's likes and dislikes based on product features.

.(7)"TF-IDF Approach", Ari Aulia Hakin,Alva Erwin,Kho I Eng,Maulahikmah Galinium,Wahyu Muliady

To create a good word and weight dictionary the author suggested some steps cited below:

- Tokenization
- Bigram creation
- Duplicate Removal
- Stopword Removal
- term frequency fltering
- supervised word removal
- tf-idf

TF-IDF is a content based filtering technique where TF is Term frequency and IDF is Inverse Document Frequency.This method is used to eliminate the most common terms and extracting most relevant terms from the corpus.

The authors used TF-IDF as a weighting factor in information retrieval and text mining.The TF-IDF value increases proportionally to the number of times a word appears in the document, but its offset by frequency of the word in corpus. They used 12000 articles and 53 persons for this research purpose.

The authors concluded there by showing the efficiency of the algorithm.They also suggested to use better semantic relativity concepts which might be domain specific but provide the better results.They also pointed out weakness points of the algorithm.

“.(8)A Novel Recommender System for E-Commerce”, Pang-Ming Chu , Shie-Jue Lee

Word2vec is adopted to analyze the semantics of the text words and each word is represented by a unique vector. Then an item vector is developed for each item and dimensionality reduction is applied to project the acquired item vectors into a lower dimension space. This novel recommender system reads the user reviews and recommends them other books based on the reviews. Reviews are considered as items .

The author of the paper used Word2vec to convert the items into vectors. The original data downloaded from Amazon.Word2Vec considers all the words near the word to be trained as a feature. For this reason, the author inserted **asin** in several places of the comments to ensure it to be taken as a feature of items.

Then the author use Word2Vec to get the item vectors of the text words. we apply PCA to reduce the dimensionality of the item vectors.

The main contributions of this paper are:

- Proposing a way of detecting semantic meanings from user comments on items bought.
- Using dimensionality reduction to increase the accuracy and speed of clustering while dealing with high dimensional data.
- Increasing the efficiency of recommendation due to dimensionality reduction and clustering.

“.(9)Very Deep Convolutional Networks for Large -Scale Image Recognition”, Karen Simonyan, Andrew Zisserman

The term deep learning refers to Artificial Neural Networks has been considered to be one of the most powerful tools and become very popular in handling large data.One of the most popular deep neural network is the Convolutional Neural Network(CNN).CNN have multiple layers including

Convolutional layer,Non linearity layer,pooling layer and fully connected layer.

The author in this paper made an evaluation of networks by increasing depth using an architecture with very small(3x3) convolutional filters which shows that a significant improvement on prior-art configurations can be achieved by pushing depth to 16-19 weight layers.

The author defined an architecture named VGG which comprises of multiple CNN layers of different sizes followed by max pooling and softmax classifier.The testing phase error resulted 6.8 percent which tops all the other models like GoogLeNet,MSRA,Zeiler Fergus.

The author concluded by demonstrating that the representation depth is benificial for the classification accuracy and the state-of-the-art performance can be achieved using a conventional ConvNet architecture is substantially increased depth there by certifying the importance of depth in visual representations.

4 MOTIVATION

Recommender systems are used by E-commerce sites to suggest products to their customers. The products can be recommended based on the top overall sellers on a site, based on the demographics of the customer, or based on an analysis of the past buying behavior of the customer as a prediction for future buying behavior.

E-commerce became a largely revenue business now a days.everybody goes shopping online with a mind of buying something and end up buying someother other than what they expected to buy,thanks to recommendation system.recommendation system study user's behaviour towards buying items and effect of one item on another item.Using machine learning algorithms, recommendation systems build user profile on the way he is buying things.

E-commerce recommendations currently mainly focuses on collaborative filtering,while we are concerned with both content and collaborative filtering because it produces quite good results.In our content based, the word2vec and tf-idf are integrated, so that the results will much satisfying.Recommendation system virtually act as a guide for us in online shopping redirecting us to several unknown best results.Satisfying customer is the only way to get success in e-commerce which amazon does very precisely.

5 AIMS OBJECTIVES AND PROBLEM STATEMENT

5.1 Aim

The aim of this dissertation is to develop a E-commerce recommendation system that would recommend products for different people based on the similarities between the characteristics(title,brand,color,images) and Ratings.

5.2 Objectives

- To identify and extract relevant information about the product using its title and attributes.
- To get a compressed data which is as efficient as original data in terms of yielding results.
- To compute the similarities between the titles of the products and produce a good recommendation for the searched product.
- To compute similarities between the products based on the image of the products.
- Obtaining similarities between products using the user ratings on the products.
- Integrating all the results that are obtained from all the filterings and recommending products based on hybrid recommendation.

5.3 Problem Statement

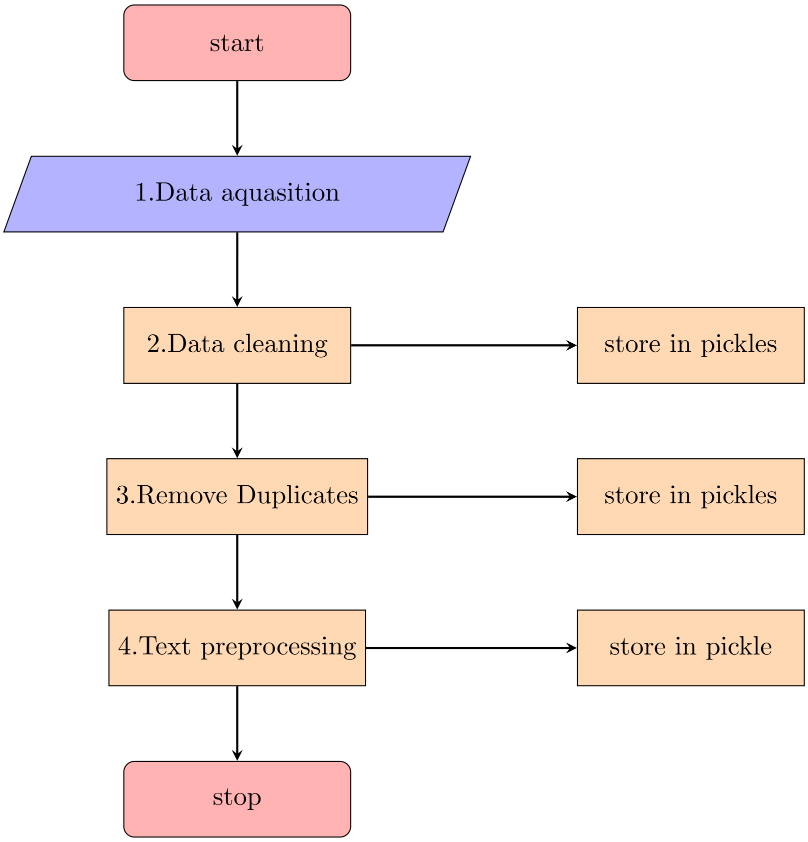
To develop a system which recommends the similar apparel items/ products in E-commerce.

6 METHODOLOGY

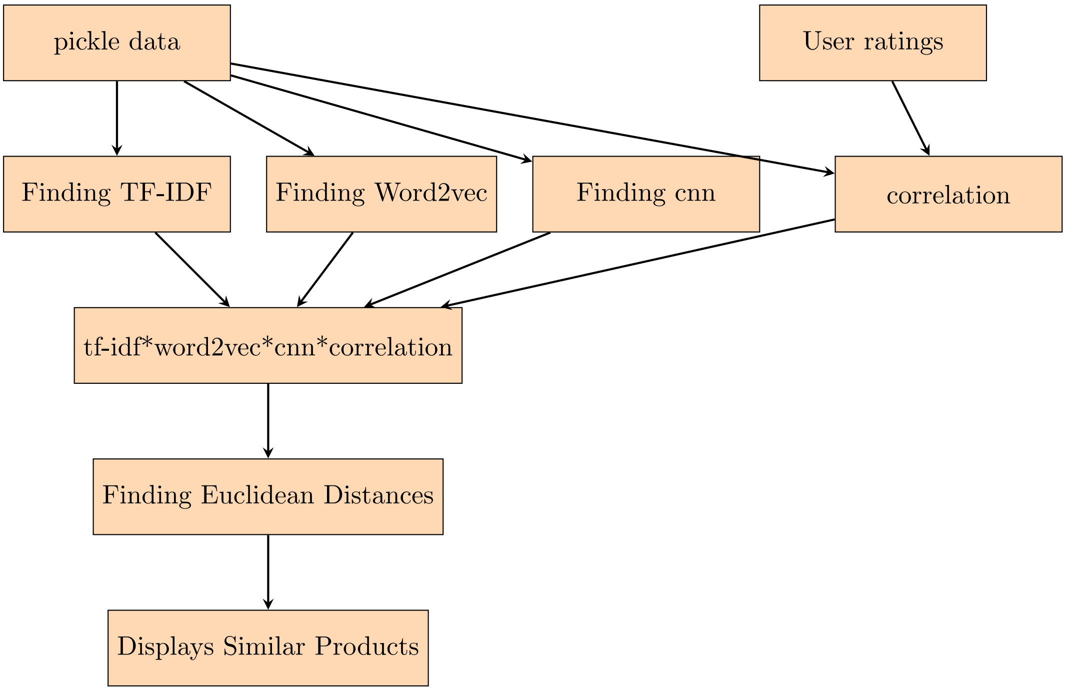
In the methodology we discuss about the problem of e-commerce like content based filtering,tf-idf,word2vec,bagofwords.

In the process of building this whole system we mainly go through 2 phases:

- Collection and cleaning of data
- Finding the similarity between the products



Flowchart for Data processing


pickle data
User ratings
Finding TF-IDF
Finding Word2vec
Finding cnn
correlation
tf-idf*word2vec*cnn*correlation
Finding Euclidean Distances
Displays Similar Products

Process of Building Recommender Systems

6.1 Collection and cleaning of data

Which data should we consider?

As we are talking about e-commerce products such as apparels has much information to process and produce results.So,we consider the most informative part of all the details that is the title of the product.The title of the product conveys large amount of useful information than the other attributes of the product. Lets take an example of womens shirt titled , **Nakoda cotton Self print straight kurti for women.**

From this we can extract :

- Brand-NAKODA
- Cloth type-COTTON
- Apparel type-STRAIGHT KURTI

- for women

Almost every word in the title gives us information which is useful in analysing the product.

6.1.1 Data Acquisation

Actually, here we try to work on real world data.we obtained data from amazon.com itself in a policy complained manner. we are using amazon's own Product Advertising API . we had obtained women's tops and shirts total of 183k products.
The json file we acquired through amazon API is given as input here.once we read no of data points, as we know there will be 183k and columns(features of product) are 19. If we take all the 19 features, it does involve a lot of computation and time consuming, so instead we only take 6 out of 19 features to make it easy for us. The features will be

- ASIN(Amazon Standard identification number)
- brand (brand to which product belongs to)
- color(color information of apparel,it can contain many colors as a value ex:red and black stripes)
- product type name (type of apparel,eg: SHIRT/TSHIRT)
- medium image url (url of the image)
- Title (title of the product)
- formatted price (price of the product)

6.1.2 Data Cleaning

This is extremely important stage in building the whole system.in the data we acquired, there may be many null values in many data points that is some product may have color null and some may have brand value null. Now when we came to formatted price having null values, we found that there are roughly 29k data points without null values in formatted price.so we will eliminate datapoints having null values at both formatted price and also color having null values. We remain with 29k datapoints i.e products.Now here comes one of the important step making pickels.we don't do the above process everything as the data is static.so we store this 29k in a file called pickles.From now on, we will load data from this if we want to do processing instead of that 183k data.

6.1.3 Removing Duplicates

Read data from pickle file.Find number of products which have duplicate titles.This is done by pandas data frama.duplicates . This function return boolean vector and gives true when title is duplicate.We found there are 2325 products with duplicated names out of 29k.

We said titles are short and much informative . So inorder to check title , title should not be too short and not too long.Title with too short are not informative. So we remove titles with less than 4 words in titles.Now sort all of our products based on titles in alphabetic order.Here we observed titles which are very similar except one or two words at the end of title.

For example,

T1-tokidoki The queen of diamonds women's shirt X-large

T2-tokidoki The queen of diamonds women's shirt Small

T3-tokidoki The queen of diamonds women's shirt Large

This kind of titles will be come one after the other when we sort all the data based on title.Now remove those kind of titles with set difference less than or equal to 2 for the titles. which means titles should atleast vary 3 different words at the end.After removing those kind of titles, we are left with 17k.Now pickle this data.

Now search those kind of titles for which set difference of words will be less than 2 but those distinct words lies in the somewhere in the middle of the title.

For example,

T1-EVALY Women's cool University of UTAH 3/4 sleeve raglan Tee

T2-EVALY Women's unique University of UTAH 3/4 sleeve raglan Tee

T3-EVALY Women's New University of UTAH 3/4 sleeve raglan Tee

Now apply brute force algorithm to remove those kind of titles.This takes a lot of time and also require minimum of 12GB ram to execute this code.

We used inverted indexes data structure to computing differing words among titles.These advanced data structure shrink this number of comparisions that happens in brute force algorithm.

We are left with 16k products .pickle this data for further use.

6.1.4 Text Preprocessing

This mainly includes of removing stop words.Stop words are those which are of no use in the title that means which are not informative.is, of, the are some of those.

As we are using TF-IDF,Word2vec stop word removal helps a lot in getting efficient results.Inorder to do this, we need to download NLTK toolkit for text processing.After downloading, load this toolkit into code. First remove all the special characters from title and then convert all letters into lower case inorder to distinguish easily.Then identify stop words in title and remove them.Pickle the data that is pre-processed.

6.2 Content Based Filtering

6.2.1 Text Based product similarity

Until now we made out data very compatible to use.Now its time for process.Every title is converted into N dimensional vector. This process is done by Tf-Idf. Each title is considered as a document and all the document collectively called as corpus.calculate TF(term frequency) and IDF for each word in document.

TF-IDF=(c(w,d)/T(w,d))*log(N/dw)

C(w,d)-Count of particular word in that document

T(w,d)-Total count of words in that document

N-No of documents

dw-no of documents that the particular word present

Now make n dimensional vector in which product of tf and idf is stored in the indexes pf that word in vector.tf tells us frequency of a word in document.idf tells us rarity of a word across all the documents.

We created a TF-IDF vector by importing the tf-idfvectorizer from scikit learn and applying it on all documents in the corpus.The products are recommended based on the pairwise distances between the input vector and remaining product vectors in the corpus.

6.2.2 Text semantics based product similarity

This tell us semantic similarities of a word in a corpus by creating 300 dimensional vector. Word2vec is one those algorithms which tells semantic relation between word and document. To implement this we are taking google's word2vec since it is made up of large corpus.This word2vec actually tells us how a word occurs in presence of other words that are present in a document. Now every word has a separate word2vec and tf-idf value. This word2vec vector is multiplied by its tf-idf value.It is done with every word and average all the vectors.This is the final vector vector which is very useful to calculate euclidean distance between two products.This vector has two specifics, due to tf-idf we can know word importance in document and by word2vec ,we know semantic similarity of words.

As we got vectors we can now imagine products are plotted in 300 dimensional space and each product is denoted by a point.We take brand and color of the product from the data.We will find vector for them using word2vec and concatenate to title vector i.e weighted tf-idf word2vec of title.By doing this,title,brand,color does major roles in getting similar products.We find euclidean distances between all the points and less distance between two vectors means that products corresponding to that vectors are much similar.we displayed top 20 similar products.

6.2.3 Image based product similarity

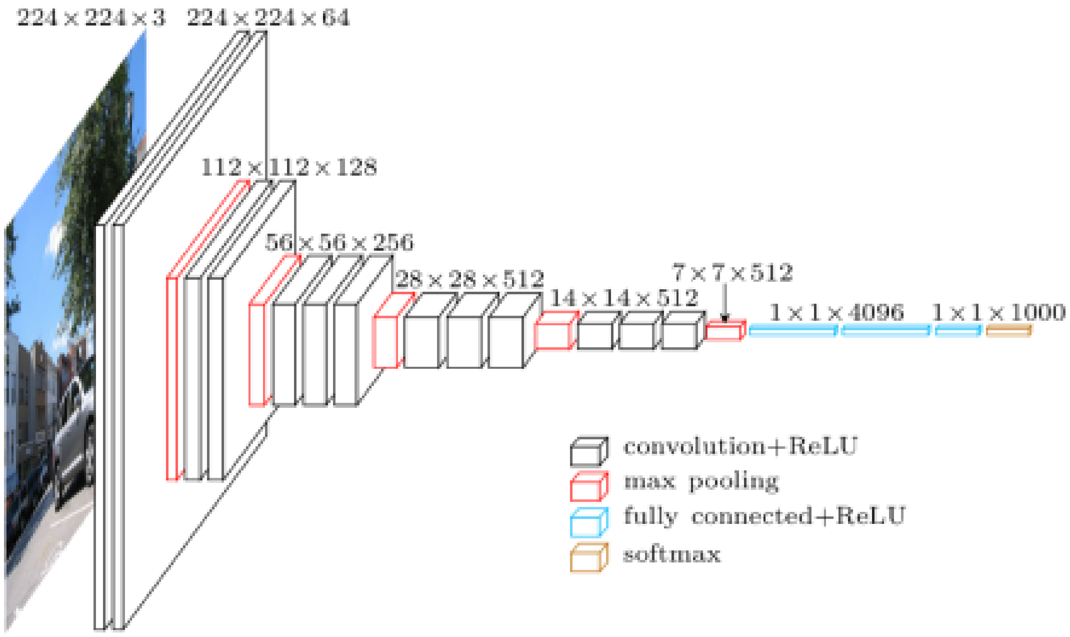
The main idea here is to recommend products based on the image of the main product.It takes every aspect of image like the boundary,designs or any certain patterns that can be visible on the input image.The algorithm to implement Convolutional Neural Networks is VGG-16.VGG-16 has a total of 16 layers comprising Convolution,Max pooling ,Fully connected and softmax layers.The preprocessing before applying VGG-16 here is,Every input image is converted to 224 x 224 x 3 dimension.

Procedure of VGG-16

1.The image pixels are scaled by a factor of 1/255.

2.Convolutional filter of size 3x3 is applied and iterated throughout the image.

3.Maxpooling(2x2) is done on the obtained image and thereby helps in reducing the size of the image.



4.A vector of size 25088 is obtained at the end of the fully connected layer.Every image gets its own same sized vector and all this vectors are stored in a file.Here the softmax classifier is neglected and the vectors that are saved are used to calculate the pairwise distances between the input image and the other images and recommendations are done accordingly.

6.3 Collaborative filtering

6.3.1 Using Correlation

The traditional collaborative filtering focuses on the user-user relation.It recommends products for a user based on his/her past orders.Our collaborative filtering recommends products based on the item-user relation.

We have a dataset of users,products and the user assigned ratings for that products.We need to find the correlation for a product with respect to all the products.But,here arises a problem while computing the correlation for all the products because the dataset of products is large i.e,16k.So,here we need to reduce the size of the data for which we are finding the correlation.In order to get this we are computing the utility matrix.This utility matrix consists of few products with all the user who gave ratings to that filtered items. To reduce utility matrix size, we are using SVD so that it gives us most optimistic matrix to calculate Correlation. So correlation matrix is calculated for this matrix that obtained after transforming the utility matrix with SVD.

Pearson Correlation is termed as

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \tag{1}$$

where

x_i- denotes the rating of user u on item i.

$\overline{x}$- is the average rating of the ith item.

y_i- denotes the rating of user u on item i.

$\overline{y}$ - is the average rating of the ith item.

6.3.2 Using Cosine Similarity

We are using the same utility matrix for this also and the result is obtained. **Cosine similarity is termed as**

$$cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||} \tag{2}$$

7 Experiment Results

7.1 Text based Results

Query image and title:

sleeveless lost printed vest tank ruffles swing women size xl

Results using TF-IDF Word2Vec:

1.2014 sleeveless heart breaker printed vest tank women size xl

2.binmertm women big size sleeveless zipper decoration vest tshirt tank top us 14

3.sleeveless america flag printed vintage vest tank women size xl

4.yabina women sleeveless loose racerback vest tank top us6 white

5.yabina womens sexy irregular vest sleeveless tank tops us6 white

6.bestpriceam sandistore womens summer lace chiffon vest tank tops xl

7.new cotton womens dlowers printed sleeveless vest tank tops red l

8.pink queen hot women skeleton printed sleeveless shirt vest tank tops galaxy1

Results using Word2Vec:

1.2014 sleeveless heart breaker printed vest tank women size xl

2.binmertm women big size sleeveless zipper decoration vest tshirt tank top us 14

3.sleeveless america flag printed vintage vest tank women size xl

4.yabina women sleeveless loose racerback vest tank top us6 white

5.yabina womens sexy irregular vest sleeveless tank tops us6 white

6.bestpriceam sandistore womens summer lace chiffon vest tank tops xl

7.pink queen hot women skeleton printed sleeveless shirt vest tank tops galaxy1

8.women plus size camisole strappy swing top cami vest wite polka dot xxxl

Results using TF-IDF:

1.girls chiffon tshirt blouse ruffles sleeveless tank tops

2.sleeveless america flag printed vintage vest tank women size xl

3.soprano plus size sleeveless swing tank top size 2x

4.2014 sleeveless heart breaker printed vest tank women size xl

5.women chiffon dot floral printed sleeveless vest dress women ladies

6.summer chiffon sleeveless shirt vest tank tops pink xl

7.km fashion printed vest slim sleeveless vest tshirt

8.sumlulu women summer vest sleeveless blouse casual tank tops tshirt xl

Results using IDF:

1.girls chiffon tshirt blouse ruffles sleeveless tank tops

2.white top blouse tank shirt sleeveless

3.womens tank top white

4.soprano plus size sleeveless swing tank top size 2x

5.summer chiffon sleeveless shirt vest tank tops pink xl

6.long sleeve top blouse tshirt

7.pink rose juniors sleeveless tank top size

8.studio printed long sleeve top size l

Results using CNN:

1.womens sullen needle pusher raglan tshirt black 2xl

2.yepme womens green cotton tees size 1

3.summer floral lace loose tshirts plus size

4.dear john denim womens dear john trista sweater xs navy

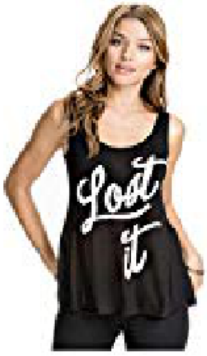
5.bella luxx womens bella luxx knit large heather gray large

6.dknyc tank top printed overlay 161871 periwinkle

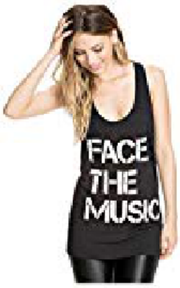
7.issac mizrahi short slv peplum knit top a265193 green olive xxs

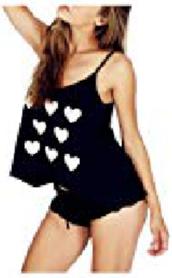
7.2 Results using combination of TF-IDF,Word2Vec,CNN and Correlation

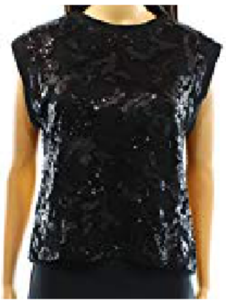
Input image

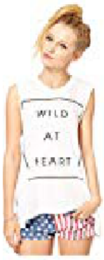


Final Recommended Products based on hybrid filtering are:









8 FUTURE SCOPE

So far we only seen results based on text mining,image and the user ratings that are given to products from the content that is available.

In our future work, we are going to take user reviews and do semantic analysis using word2vec on the user reviews and classify products based on the results and then appending the result with the previous content based techniques to obtain much precise data related to search query.

The next step of our future work is to adopt a complete hybrid algorithm to see how the combination of collaborative and content-based filtering techniques can gives us a better. Future collaborative filtering includes user-user relation and recommends according to user past history. recommendation compared to the adopted technique in this.

This addition of CNN and collaborative filtering makes it the best algorithm to produce recommended products

9 CONCLUSION

In this project we used different Data Cleaning techniques to make sure that our systems are able to perform required operations necessary for the project.
Then we used combined Content Based techniques to obtain the required results.According to the experimental results, the method of feature extraction based on TF-IDF and word2vec fusion can be used for mining user models. Aiming at the shortcomings of the current text vector representation method, we adopted the advantages of Word2Vec by combining Word2Vec and TF-IDF, and proposed a new feature extraction algorithm based on the Word2Vec vector.Finding distance between this combined tf-idf word2vec vector gives us better results than when are used alone. We took image of the products and applied CNN on them to get patterns and shapes of images. This CNN allots each image of product with 25k vector and finds distance between the vectors to get similarity between products. The collaborative filtering is done using user rating on products. We find similarity between products using correlation between products based on ratings given to them. We are checking every aspect of a product to recommend the best nearer products to it.

References

[1] Hu Jimning. “Application and Research of Collaborative Filtering in E-commerce Recommendation System.”

[2] Tessy Badriyah, Erry Tri Wijayanto, Iwan Syarif, Prima Kristalina(2017).“A Hybrid Recommendation System for E-Commerce based on Product Description and User Profile.”

[3] Prafulla Bafna, Dhanya Pramod, Anagha Vaidya(2016). “Document Clustering: TF-IDF approach”

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. “Distributed representations of words and phrases and their compositionality.” In Proceedings of International Conference on Neural Information Processing Systems, pp. 3111-3119, 2013.

[5] Jianqiao Hu1,2, Feng Jin1,2, Guigang Zhang2, Jian Wang2, Yi Yang2 (2017) “A User Profile Modeling Method Based on Word2Vec.”

[6] “How to build a high quality Recommender system?”,S.Khusro et al.

[7] “TF-IDF Approach”,Ari Aulia Hakin,Alva Erwin,Kho I Eng,Maulahikmah Galinium,Wahyu Muliady

[8] Pang-Ming Chu, Shie-Jue Lee (2017).“A Novel Recommender System for E-Commerce.”

[9] “Very Deep Convolutional Networks for Large -Scale Image Recognition”, Karen Si-monyan, Andrew Zisserman