International Journal of Electrical Electronics & Computer Science Engineering
Special Issue - NCSCT-2018 | E-ISSN : 2348-2273 | P-ISSN : 2454-1222
March, 2018 | Available Online at www.ijeecse.com

A Review Paper on Data Extraction from E-Commerce Website and Its Analysation

Amit Kumar[1], Akansh Pathak[2]
[1-2] UG Scholar, Department of CSE, Poornima College of Engineering, Jaipur, Rajasthan, India
[1]amitpceevn510@poornima.org, [2]akanshkumarpceevn505@poornima.org

Abstract: Data Crawling is basically the technique with which we crawl the data through a series of steps like data collection, data pre-processing, data analysis and data visualisation technique. Firstly, we collect the data which is the desired and the resultant data that the user actually wants and use it for the visualisation technique. Then , on that collected data ,data pre-processing is done in which various types of filtration are done like removal of extra words and case conversions. Secondly, analysis of data is done after pre-processing in which we are going to visualize it on the basis of analysed data on the dashboard using tables, pie-charts and histograms

Keywords: Crawl, Visualize, Histograms, Dashboard.

I. INTRODUCTION

Product classification for E-commerce sites is a backbone for successful marketing and sale of products listed on several online stores like Amazon, eBay, and craigslist etc. Since a large number of business users list their products and expect to find buyers for their products, it is crucial that the products are listed in accurate categories. The data on e-commerce websites is increasing at a very rapid rate. We create 2.5 quintillion bytes of data each day. That's the problem businesses face. Companies need a way to access the latest information in as little time as possible, and they need those data sets to be coherent. Data visualization tools offer a way to communicate large amounts of information clearly and efficiently. Data visualization works because 65 percent of people are visual learners. Visualization helps the eye quickly compare different pieces of data by using graphics such as tables, bubble charts, maps, and scatter plots. Our minds often respond better to pictures than to rows of numbers. Business managers who use visual data discovery tools are 28 percent more likely to find timely information than peers who only use managed reporting and dashboards.

II. DIFFERENCE BETWEEN WEB AND INTERNET

The Internet and World Wide Web are not one and the same. The internet is a collection of interconnected computer networks, linked by copper wires, fiber-optic cables, wireless connections, etc. In contrast, the Web is a collection of interconnected documents and other resources, linked by hyperlinks and URLs. The World Wide Web is one of the services accessible via the internet, along with various other along with e-mail, file sharing, online gaming and others described below. However, the Internet and the Web" are commonly used interchangeably.

A. The Internet:

The Internet is a massive network of networks, a networking infrastructure. It connects millions of computers together globally, forming a network in which any computer can communicate with any other computer as long as they are both connected to the Internet. Information that travels over the Internet does so via a variety of languages known as protocols.

B. The Web:

The World Wide Web, or simply Web, is a way of accessing information over the medium of the Internet. It is an information-sharing model that is built on top of the Internet. The Web uses the HTTP protocol, only one of the languages spoken over the Internet, to transmit data. Web services, which use HTTP to allow applications to communicate in order to exchange business logic, use the Web to share information. The Web also utilizes browsers, such as Internet Explorer or Firefox, to access Web documents called Web pages that are linked to each other via hyperlinks. Web documents also contain graphics, sounds, text and video.The Web is just one of the ways that information can be disseminated over the Internet. The Internet, not the Web, is also used for email, which relies on SMTP, Usenet news groups, instant messaging and FTP. So the Web is just a portion of the Internet, albeit a large portion, but the two terms are not synonymous and should not be confused.

III. IMPLEMENTATION MODEL

Implementation model is basically an architecture of designing the modules , which basically helps to tell the blue print of our work. It is basically the sequential model which helps to tell that what process should be done next to enhance performance of our goal.

1. Data Collection:

To evaluate and test our approach to test which classifier with the given feature set would perform best in product classification, information on 35,000 products for 45 categories were gathered by crawling amazon site and scraping the pages to extract the attributes (title and description) . The category tree can be very deep and possibly contain many levels of depth Category Selection: The graph in picture illustrates the distribution of the category classes. This reflects the typical distribution of long tail products e-commerce

sites carry on the categories.

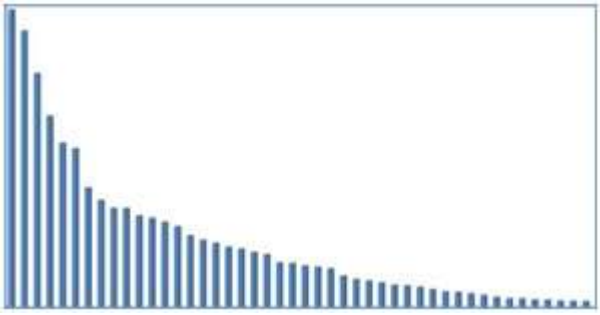


Table 1. Top 5 Categories with Highest Product

Automotive	3000
Watches	2500
Shoes	2000
Electronics	1654
Bikes	1500

2. Data Pre-Processing:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.
Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications (like neural networks).Data goes through a series of steps during preprocessing:

Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

Data Integration: Data with different representations are put together and conflicts within the data are resolved.

Data Transformation: Data is normalized, aggregated and generalized.

Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.

Data Discretization: Involves the reduction of a number of values of a continuous attributeby dividing the range of attribute intervals.

For pre-processing, Apache Lucene libraries were used for tokenization, normalization, stop word removal, and stemming. As a first step in preprocessing, all the text was converted to lower case, and applied tokenization based on delimiters. (Tokenization can be very complex for multilingual scenarios and the current focus was restricted to English).

The following table illustrates the various combinations feature list were prepared (the cells that are checked are the ones against which the experiments were run):

Table 2.

Model	Frequency Based	Info Gain	Chi-Square	LDA
Unigrams	✓	✓	✓	✓
Bigrams	✓			
Unigram + Bigram (50-50 Split)	✓	✓	✓	

Similarly for, Bigram modals, feature selection was conducted by picking top word pairs are that co-occurring in each of the categories and merging to get the final feature list. Once top X features are identified, then various experiments were carried out with varying number of features that included 100 features, 600 features, 1200 features, 4000 features and 10000 features. In addition to varying the number of features, separate experiments were done once by using only from title and other time using both title and description. The results of this analysis were detailed in

3. Data Analysis:

Data Analysis is basically the analysation on the resultant data. This data may be analyzed in the form of report manuals. It means analyzing the data on the basis of graphs for e.g.- pie -charts, histograms, bargraph. We can extract the data and can make their analysis with the help of various tools like Power BI. This tool has the capacity to analyses the data in the form of tables, histograms, pie charts, etc. Power BI offers basic data wrangling capabilities similar to Excel's Power Query. Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery for predictive rather than purely descriptive purposes. Analysis refers to breaking a whole into its separate components for individual examination. Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses or disprove theories.

The results of the experiment with NB, SVM and K-NN give the following results:

1. The unigrams as whole group outperform the bigrams for accuracy in classification.
2. Feature set size at 100 : When the feature set was small at 100 size, K-NN performed the best with all

feature selection models. (refer to Figure 1 in next page) maintain accuracy of at least 1 to 2 percentage points above the rest. SVM trailed right behind K-NN with 1-2 percentage points below. However Naïve Bayes accuracy rates were much lower with a difference of 9 % points. A unique result set to is that, when the feature set of 100 derived using Chi-square model was used, it significantly increased the accuracy rates by more than 10 percentage points for Naïve Bayes and for SVM and K-NN at least by 4 percentage points. of all the classifiers as shown in Figure 2 on next page. The reason Chi-Square performed better was, it was able to identify words like "pedomet, trampoline rower sled" which have high degree of accuracy in associating with the respective category.

3. Performance at feature set size of 600 showed that all models got close to 2 or more % point boost using the chi-square model compared to the rest with SVM faring the best and KNN just trailing behind(refer to Figure -2 on next page). At the feature set size of 1200, Naïve Bayes started displaying higher accuracy gain compared to smaller data sets previously reaching 80% while SVM and KNN had steadily improved their accuracy rate till reaching 4000 feature set size with all models.
4. Optimal Feature set size seems to be at 4000 set size where theperformance gap between regular unigram choice based on frequency and info-gain and Chi-square fared almost comparably with frequency based unigram model showing the best results for all classifiers.
5. LDA also showed decent levels of accuracy at 4000 data set butfrequency-based and Chi-square models surpassed LDA. Refer to Table -3 below.
6. Naïve Bayes remained almost flat at the same levels as 4000feature set size when the feature set size increased to 10,000 and surpassed all other models at 1000 features.
7. SVM steady improvement with increased feature set size till thepoint of 4000 and started a downward curve at 10,000 feature set size, the reason is that it started suffering from over fitting.
8. KNN was the worst performing as the feature set size increased.
9. Running time for `Naïve Bayes for fastest for all feature set sizesup to 10,000 with 1-2 minutes. SVM approximately max 15 minutes. K-NN ranged from 30min to 5 hours depending on the feature set size.

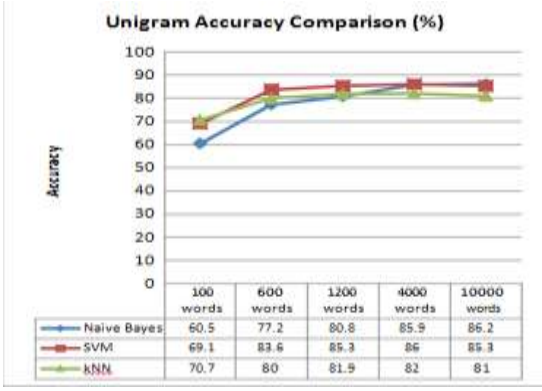


Fig. 1. Unigram Frequency Model

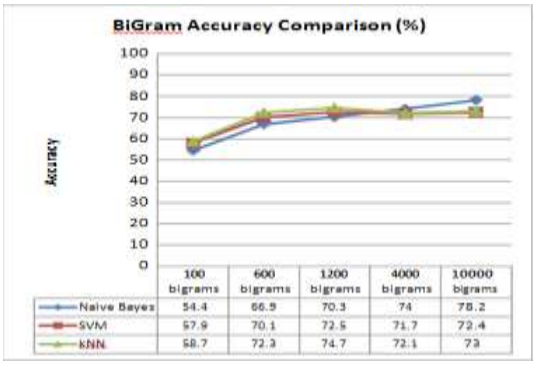


Fig. 2. Bigram Frequency Based Model

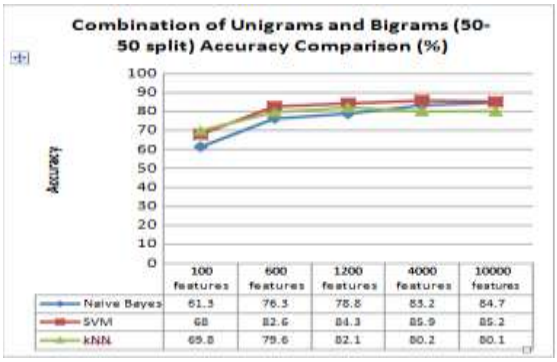
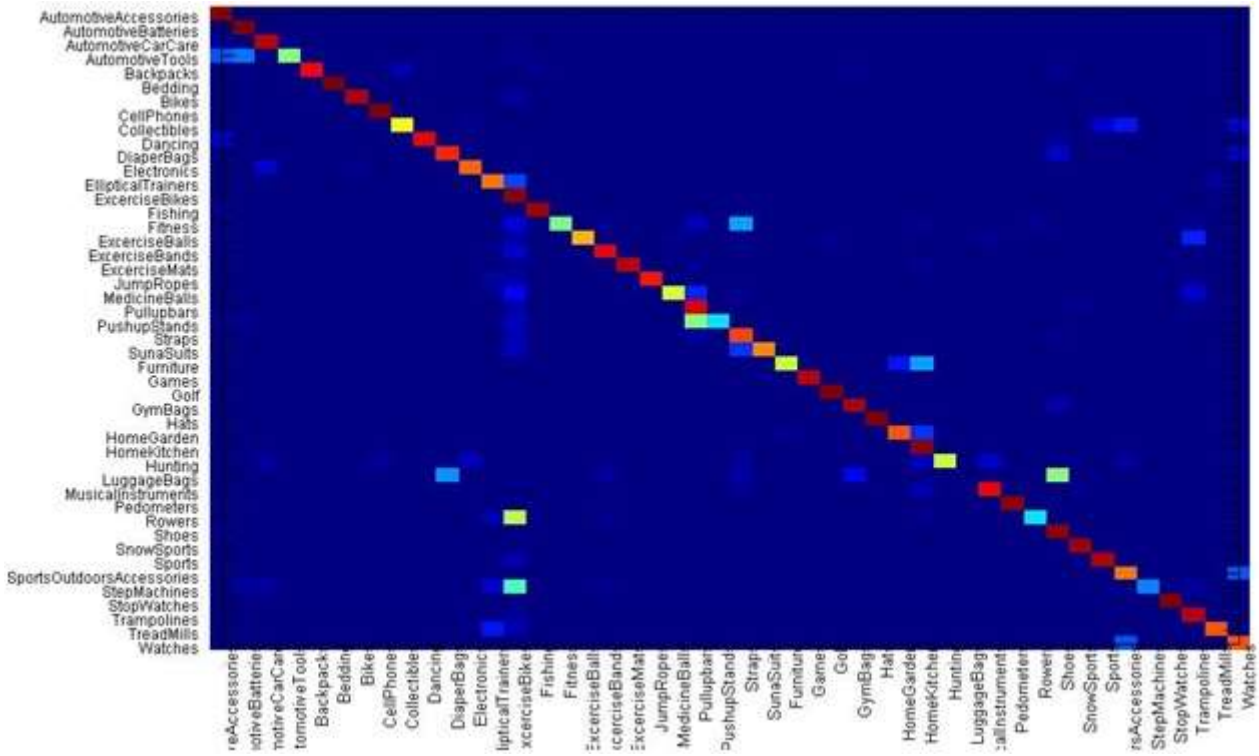


Fig. 3. Unigram and Bigram split Model

Below is the confusion matrix derived from the best performing Chi-Square feature selection unigram modal with 4000 features based on SVM classifier:



As seen in the above confusion matrix, classifier was getting confused while classifying the items in 'Sports Outdoor Accessories' with the items in' Watches' category. The reason is that, 'Sports Outdoor Accessories' have "sporty watch items" and as watches category is dominant in training, classifiers were trying to maximize towards watch category. We have found similar patterns with other categories as well, where the product classification could fall into more than one category. This prompts for further research into multiple class relations.

IV. CONCLUSION

The information has lead to an explosion of information available to user. As there is large amount of data in web browsers, so for selecting desired data from the web browsing sites we use a crawler for collection of data, and for its pre-processing and its analysis and visualization. The results of the experiments clearly indicate that with a small feature size set a Chi-square model of feature selection gives a significant boost to the classifiers. The main aim is to visualize the data on the dashboard on the basis of tables, pie-charts, histograms, and bar-graphs. We have basically collected the data in the form of tables which is stored in the form of XML sheet and just apply further modules of data pre-processing, analysis and visualization. In the process of data-preprocessing, the tables are going to be normalized as removing the redundancies.

V. REFERENCE

[1] https://en.wikipedia.org/wiki/Web_crawler

[2] searchsoa.techtarget.com

[3] https://www.promptcloud.com/data-scraping-vs-data-crawling/

[4] https://scrapinghub.com/

[5] bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining/

[6] https://www.woorank.com/en/blog/how-a-crawler-works-back-to-the-basics

[7] trackmaven.com/marketing-dictionary/web-crawling/

[8] https://scrapy.org/

[9] https://en.wikipedia.org/wiki/Data_mining

[10] www.thearling.com/text/dmwhite/dmwhite.htm

[11] searchsqlserver.techtarget.com › BI and Data Warehousing › Software applications

[12] www.investopedia.com/terms/d/datamining.asp