TIPid: Decision Support System on E-Shopping Products through Graphical Visualization and Data Mining

A Thesis Proposal

Presented to the Faculty of the

College of Computer and Information Sciences

Polytechnic University of the Philippines

In Partial Fulfilment

of the Requirements for the Degree

Bachelor of Science in Computer Science

Averos, Christian M.

Delicano, Jobea Ann F.

September 2017

TABLE OF CONTENTS

TIPid: Decision Support System on E-Shopping Products through Graphical Visualization and Data Mining i
TABLE OF CONTENTS ii
LIST OF FIGURES iii
LIST OF TABLES iv
LIST OF NOTATION v
Chapter 1 THE PROBLEM AND ITS BACKGROUND 1
Introduction 1
Background of the Study 2
Statement of the Problem 3
Theoretical/Conceptual Framework 4
Theoretical Framework 4
Conceptual Framework 4
Significance of the Study 5
Scope and Limitation of the Study 5
Definition of Terms 6
Chapter 2 REVIEW OF LITERATURE AND STUDIES 7
Related Literature and Studies 7
Synthesis of the Study 12
Chapter 3 METHODOLOGY 14
Research Design 14
Sources of Data 14
Instrumentation 15
Software/Hardware Tools 15
System Architecture 15
Development Details 16
Research Instrument 17
Data Generation 17
Data Analysis 18
REFERENCES 20
APPENDIX 22
APPENDIX A: SAMPLE EXPERIMENT PAPER 22

LIST OF FIGURES

Figure 1 System Theory Framework 4
Figure 2 System Architecture 15

LIST OF TABLES

Table 1 Conceptual Framework of the System 5

LIST OF NOTATION

DM	Data Mining
OSW	Online Shopping Website

Chapter 1

THE PROBLEM AND ITS BACKGROUND

This chapter introduces the problem and the background as well as the supporting information that will be used throughout the study.

Introduction

In the fast-paced modern world, where technology progresses rampantly, we try to keep up with today's trends as well as needs. With the rise of technology, clinging unto technology makes our lives easier, and some experts say that the internet can now be consider a need along with food and shelter (Pastore, 1998). Everything is being digitalized as for our communication and work-processes, and with the use of the internet and its rapid development, delivery systems are getting more and more popular alongside of online shopping websites (OSWs). With that being said, it has great effect on us, Filipinos, especially modern Filipina who buy products on the internet through OSWs.

Most people all over the Philippines have access to the internet and they are still growing in numbers but there's a problem, how can we diminish the hard work process of search through many OSWs just to get the best deals? As a Filipino, I'm used to seeing my mother go through many store just to get the optimal price that fits in our budget, sometimes, we even come to a point of bargaining or asking the store to lower the price if we're going to buy bulk of their products. Similarly, E-Shoppers would search for many online shopping websites like Amazon and Lazada to compare and analyze what is the best for their searched similar products, weighing in the reviews, quantity, and even quality just to make sure that they're making a good decision upon purchasing such product. Searching and rummaging through many OSWs is a tiresome job. The proposed solution helps the user to remove the users the problem of doing such a tiresome work by going through the process itself providing the user a single comparison interface page. Providing 3 ranked lists each sorted by highest

rating, by cheapest prices, and by the combination of the former two forming the best deals, respectively.

Alongside of many frameworks for building mobile and web applications, acquired data must be represented in a formal and understandable way for users to grasp in first glance the importance of this study. Visualizing dynamic data in a user-friendly interface can make users comprehend easily the content of the acquired unstructured data sets. Therefore, the researchers decided to develop a system to help with troubled online shoppers (OS).

Background of the Study

Most Filipino shoppers have a hard time jumping to another store into other store just to find the best deals of their wanted product, may it be a new pair of shoes, shining jeans, or maybe a newly released branded cap. As much as we can, we bargain, bargain, and bargain. Most modern Filipino women would go to mall after mall just to buy off on-sale product which is also true for old-fashion mothers who would prefer to bargain just to get deals from their market suppliers. In today's technological advantage, many shoppers prefer e-commerce as their mode of shopping (Kitonyi, 2017), primarily because internet is easy to access and buying is just a few clicks away.

Filipinos like to bargain, since the old times, Filipino mothers would go to the market with a fixed money and get much more worth of their money for. Just imagine the hard work of going through and bargaining to many stores just to lessen the pay for a similar product as well as avoiding similar low quality products. This is also true to online shoppers looking for looking for a specific product in an online store, they would search many online store just to get the best deal in their opinion. They would scan for reviews of the said product and sometimes to avoid getting scammed or getting less than what you expected. One of the most common problem in OSWs is that there are scammers always present (Bolido, 2014) even in the most trusted OSWs. As for saving, it is no wonder that getting the less payment

is the best. Some men would pay double just to get the product they want as they deemed it worth it. Ratings and reviews comes into play as helpful elements in deciding on what to buy on the store.

Visualization is the representation of data in which is any technique for creating images, animations, and etc. to convey a message. Truly, life would be boring if there weren't many colors or shapes around. Both became one of the major tools in life for differentiating and comparing. In today's application of visualization, one of the tools it is used for, when it comes to websites and computer applications, is helping the users' on their experiences. It help them in the form of 'navigation', or even the colour and theme of the website to give more feeling of appropriate based on the contents.

With the visualization as the main part of knowing, and criticizing, and differentiating can be implemented in a decision-support system in which colours and shape helps in choosing one in the items in question. The researchers plan to build a system that caters in the selection of similar products in which the user has to decide to buy or not. Getting the data only from the trusted OSW to minimize cyber market scamming and increasing the satisfaction rate of the users in selecting. Filtering only the 'best' deals in terms of reviews, quantity, and popularity to get the most-likely product to be picked by the user.

Statement of the Problem

This study aims to develop a system that gets the best deal from various online shopping websites based on user’s search terms and apply a visualization technique to display the best deals. Specifically, this will address the following sub-problems:

1. Using web scraper and spider, what is the accuracy of getting the relevant data from online shopping websites using percentage formula?
2. What is the accuracy of applying Interleaved Ranking using Recall formula in getting the ranked best deals based on?

a. The Ranking by Price?

b. The Ranking by Average Rating and Number of Customer Reviews?

3. What is the total accuracy of using the developed system using the mean formula?

Theoretical/Conceptual Framework

Theoretical Framework

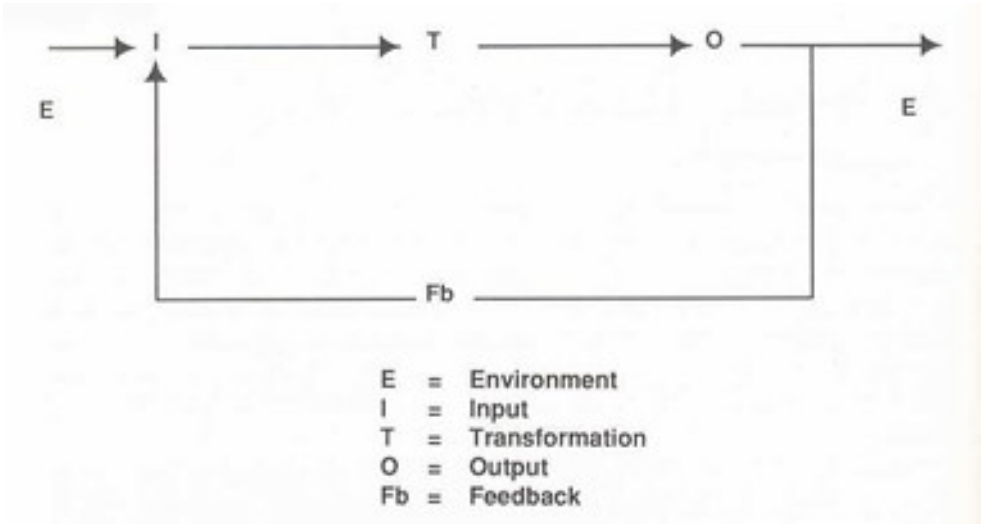

Figure 1 System Theory Framework

Figure 1 shows the theoretical framework of the system. System Theory is introduced by Ludwig von Bertalanffy. This theory points main four things that build up a system: objects, attributes, object-relationship and environment.

Conceptual Framework

Table 1 shows the conceptual framework of the system where the user's search terms are the input data. The processes are getting the data from OSW and filtering out redundant data. The output data are the collection of best deals.

INPUT	PROCESS	OUTPUT

User search term	Getting data from site Filtering out redundant data Analysis of the data set	Graphical Representation of the listed Best Deals

Table 1 Conceptual Framework of the System

Significance of the Study

The system will benefit the following people:

Online Shoppers. This study would benefit shoppers who wants to lessen their time on selecting the best deal.

Store Owners. This will benefit them by being able to search how they would price their own products.

Future Researchers. This study will be a help as a guiding reference in making a system related to e-commerce.

Scope and Limitation of the Study

This study will focus on the accuracy of the finding the best deals of electronic products and accessories selected by the user at the top online shopping websites. The system will mine data from popular retailer online shopping websites and will not include second hand online shopping websites, specifically:

1. Lazada (lazada.com.ph)
2. Shopee (shopee.ph)
3. Amazon (amazon.com)

The scope in assessing the best deals will base on the relevancy, availability, price excluding the shipping fee, and numerical reviews regarding the product. The product prices will be displayed in Philippine Peso.

The system will not be affiliated in selling products and will only act as a recommender system for finding the best deals.

Definition of Terms

Data Mining. It is the process of extracting information from a large sets of data.

Online shoppers. Product consumers that uses Online Shopping Websites as their medium for purchasing.

Online Shopping Websites. Form of e-commerce that allows users to purchase commodities over the internet with the help of a web browser.

Scammers. People who extorts or uses dirty tactics to gain information, power or money.

Chapter 2

REVIEW OF LITERATURE AND STUDIES

This chapter discusses papers related to the study including Data Mining: its techniques, tools and application; visualization techniques; and shopping: physical shopping and its problems, OSWs and its problem and application of data mining in OSWs.

Related Literature and Studies

Data Mining

In the ever expanding data collection and data storage, Data Mining (DM) has been a logical process used to search useful information through piles of data. (Ramageri, 2010) Since data sets has become more complex and larger than it was, DM was improved by different algorithms discovered and developed in the field of computer science. DM is an important task in knowledge discovery in databases (KDD). (Li & Beaubouef, 2010) KDD consists procedures including (Smita & Sharma, 2014):

selection of data from varying sources;

preprocessing or data reduction;

transformation of data;

identifying the desired result; and

interpretation and analysis to give relevant information.

There are several DM techniques developed and applied in KDD (Bhatnagar et al., 2012):

Association. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in reservation systems analysis to identify in which area customers frequently make reservations.

Classification. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics. Basically classification is used to categorize each item in a set of data into one of predefined set of classes or groups.

Clustering. Clustering is a data mining technique that makes meaningful or useful cluster of objects that have similar characteristic using automatic technique. Different from classification, clustering technique also defines the classes and put objects in them, while in classification objects are assigned into predefined classes.

Prediction. It is one of a data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables.

Sequential Patterns. Sequential patterns analysis is one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

Discrimination. Data discrimination produces what are called discriminant rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class.

DM is currently used in wide range of industries that holds a large amount of data and is commonly combining it in other tools that can enhance the power of DM in various fields. (Ramageri, 2010) Some of these fields include the field of biological science to analyze sequential pattern in genes and identify various diseases related to it. Finance industries use DM for price prediction, stock forecasting, identify frauds and money laundering. (Li & Beaubouef, 2010) Sales industry also use DM as in different ways (Al Essa & Bach, 2014):

Information broker for customer buying habits, from transaction histories to loyalty card usage. Because of this, supermarkets can predict customer behavior and act upon it for customer satisfaction and better sales;

Recommendation-based business that tracks customer's bought items and offer customer new items that they might also like;

Keeping good customers that boosts sales and avoiding fraudulent, bad customers;

Decision support system for both business holders and/or customers (Smita & Sharma (2014); and

Other applications that can build customer relationship.

User-interface and Visualization

Visualization provide users a comprehensive understanding about a data. It gives form to raw data making it coherent to the users. Using computer graphic effects, data sets can be visualized to display patterns, relationship and trends in a more advanced way. This gives users an ease to apprehend information by the use of visual reasoning rather than numerical reasoning. (Zhao, 2002) Presentation of data may involve the manipulation of graphical entities and attributes. A good data visualization must consider the effectiveness or ease of interpretation, accuracy or correct quantitative evaluation, efficiency or data redundancy removal, aesthetic or must be pleasant to the user's eye and adaptable or adjustable to serve multiple needs. According to Ward, common techniques on data visualization are following: charts, graphs, plots, maps, images, 3D surfaces and animation. Parsaye and Chignell listed the common steps followed in data visualization:

1) Numerical transformation of data by modifying the distribution

2) Data analysis to interpret data that will be used in graphical interpretation

3) Graphical interpretation by plotting the analyzed data onto graphs

4) User interaction by adding options for users to dynamically adjust mapping, zooming, panning, highlighting, et cetera.

Shopping

Shopping is the activity of trading goods and products for other goods and products. (Robertson, 2012) Simply put, exchanging a product for a more benefiting one, this can be; food, tools, money, ad etc. depending on your needs. In the early age of man, the art of exchanging goods took place in which leads to easier lives of the customers for they don't even have to hunt or gather food themselves in which fasten the growth of many communities. According to Oxford, markets or shopping malls have been the main place of trading, but they're often in urban areas or in the middle of the city which can be, depending where you live or just lazy, can be hard to go to. (Bund)

As technology develop and as our understanding of computers and the internet as well as the remaining difficulty of going to the market place, landlines, the internet and trading companies gave birth to ticket booking and food delivery services, and online shopping websites (OSWs). Pizzas, clothes, and pizza, can be delivered in front of your doorstep in a matter of minutes by the usage of the internet by visiting such OSWs. But be wary of scammers especially in the e-shopping category as they run rampant waiting for a user to be conned.

As the danger of scammers aren't enough, one of the most commonly problem encountered by many users is the tedious work of manually comparing and searching of a particular type of product in many OSWs. Users have to go to multiple OSW just to type in the same keywords and is faced with many similar product in that OSW plus many more with the other OSWs. Reviewing the 'reviews', as well as looking in to the quantity, and the quality of the searched product come hand in hand which makes it hard for the comparison process thus making a decisive and long decision.

Fortunately, 'Comparison of ecommerce products using web mining authors', have started a way for tackling this issue way easier with the use of Data mining. Studying the HTML code structure of international OSWs such as Lazada and Amazon, and using web-crawling spider to scrape out unstructured data and organizing them to a meaningful data. They let the user input search terms that will display the comparing pair deals, successfully lessening the user's time for looking in different browser tabs. Displaying it in a single-view page for users to view and pick what they desire. However, displaying and adding a table of comparison for such products still doesn't solve the problem of finding the best deals an OSW can offer. Finding the best deals have to account the reviews of other users, quality, quantity, and availability of the product. (Mo, Z., Li, Y.-F. and Fan, P., 2015)

Bayesian Estimation

In order to avoid from bad product or service, consumers rely on product reviews by other consumers. Consumers automatically prejudge products on a five-star rating. This method is powerful and easy to understand but it can be mishandled when ranked since five-star ratings depends not only on the average number of stars but the average number of reviews. Take for example Product A with an average rating of 5 and a single review, and Product B with an average rating of 4 but has a hundred reviews. Not all cons Bayesian Estimation is an empirical metric that considers both average and number of reviews. This is done by using a probability distribution for each reviews given by a consumer on a specific product. By making a threshold parameter and taking a minimum number of ratings as a confidence for the threshold, a product is weighed according to the reviews it has received while still taking into consideration the true average of the product. Bayesian estimation is used in review sites such as IMDb.

Interleaving

In product information retrieval where there are different factors to consider such as price and reviews, they, of course, are ranked differently depending how it is filtered.

In order to make data retrieval efficient, the approach is to compare both rankings and display a single ranked list to the user. Both ranking are interpreted in an unbiased way and presented to the user to be more interpretable. One way of interleaving is by teamdraft wherein in adapts the analogy of selecting teams in a sports match. This way, both ranks are considered equally appending each item on the interleaved list in an alternate manner.

Synthesis of the Study

There have been many studies that concerns in the field of data mining and visualization. One of the main researches that drives the researchers is the 'Comparison of E-commerce products in using web mining' research done in Savitribai Phule Pune University in where the users inputs two or more products to compare in a single page website. However, the main problem of Shah, et al.'s work is that the users are required to manually search and add the products for it to compare. Undisputedly, price is one of the considerations when people are trying to save money when buying a product. So is ratings and reviews as for considering the duration of use or the quality of the product. It's true that people looked up to reviews to see if it's worth it or not. In this study, the researchers attempt to gather data from the leading OSWs automatically responding the search terms or input of the users. The researchers limited to OSWs with branches in the Philippines to remove or reduce the cost of high delivery fee for products from foreign countries. Using the input of the user, the system then will collect products related to the input from the OSWs and finally, displayed according to their ranking. As a decision support system, the system focuses on visualizing the products in a way that help convey the difference among the products as well as magnify the pros and cons.

For the gathering data relevant to the desired product of the user, the researchers used web scraper for acquiring the frequent relevant item sets. The prune data in will be stored in the a database to remove repeating data or noise data then subjected to an algorithm called Interleaved ranking algorithm with Bayesian estimation to find the best

deals to be visualize in a single view page. For the visualization, the users can view graphs and statistics which can help in deciding on what to buy.

Chapter 3

METHODOLOGY

This chapter defines the research methods used to conduct the study. It involves the research method that will be used, paradigm, system architecture and the data gathering procedure.

Research Design

The research design of this study used the Experimental research design. The aim of the experiment is to determine if the system is a reliable and accurate approach in the delivery of gathered best deals in the selected OSWs. With the use of this design, the researchers have reached and come up with the answers to the questions stated in the statement of the problem.

In determining the overall performance of the system in gathering the best deals, the researchers lead an experiment. The respondents, described in the Research instrument in Instrumentation, compare the system's actual output and the expected result by the researchers to measure the system's performance in accuracy, reliability, and consistency in getting the best deals.

Sources of Data

The respondents are both the researchers and searching online shoppers. The respondents used search terms to train the system and is given a set of possible best deals. The respondents then evaluated the accuracy and reliability of the system.

Instrumentation

Software/Hardware Tools

System Architecture

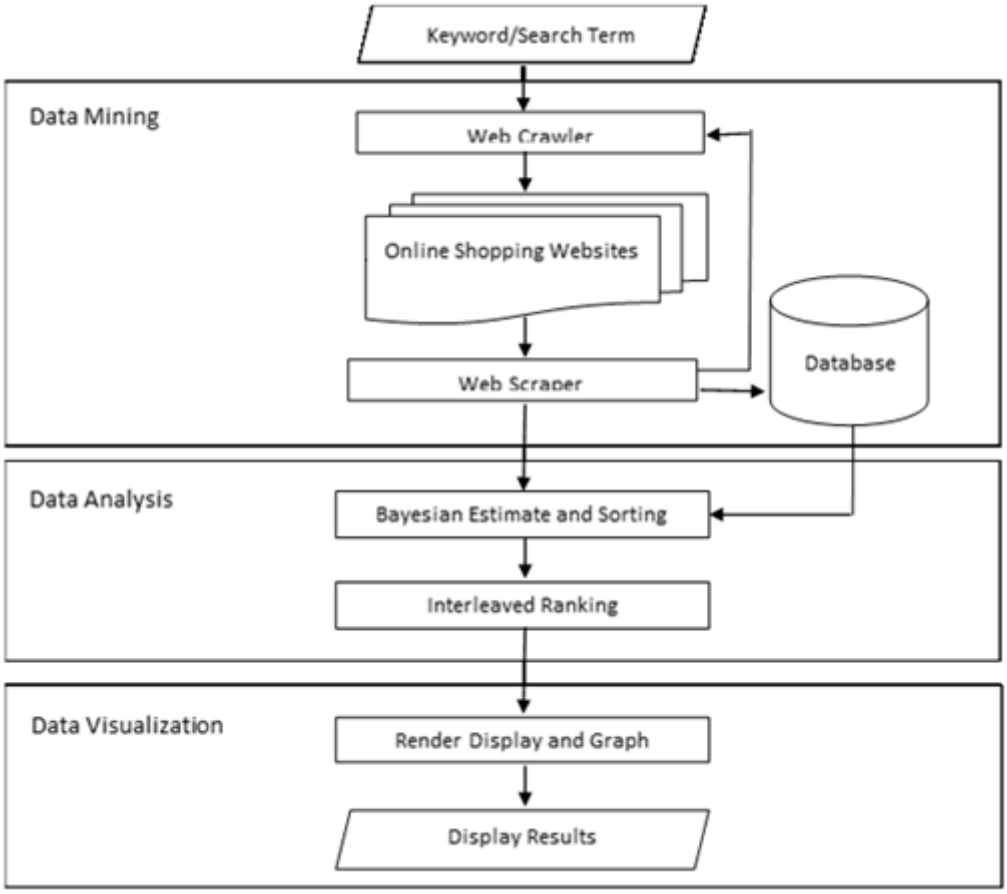

Figure 2 System Architecture

Data Mining. The product keyword will be used in order to get the specific items from OSWs. Using the Web Crawler to go to each OSWs, the Web Scraper will then mine relevant data and stored them in database.

Data Analysis. After the Web Scraper is done storing the relevant items in the database, the items will be duplicated in an exact duplicate set of two namely set A

and set B After computing the Bayesian estimate according to the ratings and reviews. Set A will be sorted starting with the lowest price and set B will be sorted according to the Bayesian estimate of the reviews. The two sets will undergo through interleaved ranking to get the ranked set.

Data Visualization. After acquiring the ranked set through interleaving, top 10 items will be displayed as the Best Deals item set. The top 10 in Set A and Set B, which were ranked by price and reviews will be displayed and called as the 'candidate' set. The scatter plot figure of Matplot library will be used to plot and display the item values onto a graph that will compare and show the difference between the ranked products.

Development Details

The researcher's will use the following tools in order to develop the system:

Python. A multi-paradigm programming language with high-level data structures and highly readable syntax that can be used on most platforms. This language will be used to create the backend for the system.

Django Framework. A server-side web framework written in Python which will be used to create the user-interface for the system.

BeautifulSoup. A web crawling framework written in Python. This will be used to crawl and extract data from various data sources.

Selenium Webdriver. Webdriver that will be used to run the spiders on browser.

Matplotlib. Matlab library in Python that will be used to render graphs.

Research Instrument

Experiment paper will be used to measure and record results while testing the system.

Data Generation

The following are the procedures in gathering data used by the researchers for the study:

Software Related Data

Maximum of (20) product data sets per OSW each for a maximum of sixty (60) products data sets were gathered will be stored in a database. The data acquired from the three (3) OSWs are then further reduced and extracted only the best out of all the data. The data sets of ten (10) 'Best deals' are sorted according to the respondent's preference and finally displayed.

Survey & Testing Data

A regression tests of twenty (20) unique search terms are used as testing data sets. The search terms will be used in order to extract item set from OSW. Accuracy of the relevant data in extracted item set will be evaluated by tallying the number of items related to the search term given. The extracted item set will be ranked by price and another similar extracted item set will be ranked by reviews. The top 10 in both ranked sets will be classified as the candidate items for the best deals. The two ranked item sets will undergo interleaved ranking and the top 10 will be the best deals. The accuracy will be acquired by counting the candidate items present in the best deals and using the recall formula. The total accuracy of the system will be computed by getting the average of the accuracy of the relevant items, accuracy of best deal by price and accuracy of by reviews.

Data Analysis

In order to solve for the first problem, accuracy of the scraped set by the Web Scraper is computed by the following formula:

$$x = \frac{n}{N} \times 100$$

where: x = accuracy

n = number of relevant item in set

N = total number of item in set

The formula of Bayesian Estimation in the second problem is used for evaluation of the ranking in the 'scraped' data sets:

$$x = \frac{C * m + R * v}{C + v}$$

where: x = Bayesian estimate

. R = mean over the ratings for the items.

C = confidence

v = number of reviews

m = threshold parameter

The formula of Recall will be used on solving second problem to get the accuracy:

$$x = \frac{cb}{cb + nb} \times 100$$

where: x = Accuracy

cb = no. of candidate (price/review) - best deal item

nb = no. of non-candidate (price/review) – best deal item

To solve the third problem, mean formula is used:

$$x = \frac{a.of\ relevant\ items + a.of\ best\ deal\ by\ price + a.of\ best\ deal\ by\ reviews}{3}$$

where: x = mean

REFERENCES

Al Essa, Bach, C. (2014, April). *Data Mining and Warehousing*.

Bhatnagar, Jadye, Nagar. (2012, November). *Data Mining Techniques & Distinct Applications: A Literature Review.*

Bolido, L (2014, April 23). Safety in Online Buying and Selling. Retrieved from http://lifestyle.inquirer.net/157606/safety-in-online-buying-and-selling/

Bund: friends of the earth, Germany. (Accessed 12 August 2017) Shopping by bike. Retrieved from http://www.einkaufen-mit-dem-rad.de/shopping_by_bike.shtml

Dr. Sudha, T. & Vasavi, G. (2017, March). Information Extraction from Online Shopping Sites using Web Content Mining Methods and Techniques.

Javadi, Dolatabadi, Nourbakhsh, Poursaeedi, & Asadollahi (2012, June). An Analysis of Factors Affecting on Online Shopping Behaviour of Consumers.

Kaidi, Z. (2002). Data visualization.

Kitonyi, N. (2017, March 07). E-commerce is Killing Traditional Retail. Retrieved from https://www.gurufocus.com/news/490164/ecommerce-is-killing-traditional-retail

Li, Y. & Beaubouef, T. (2010). Data Mining: Concepts, Background and Methods of Integrating Uncertainty in Data Mining.

Mo, Z., Li, Y.-F. and Fan, P. (2015) Effect of Online Reviews on Consumer Purchase Behavior. Journal of Service Science and Management, 8, 419-424. http://dx.doi.org/10.4236/jssm.2015.83043

Nazir, Tayyab, Sajid, Rashid, & Javed. (2012, May). How Online Shopping Is Affecting Consumers Buying Behavior in Pakistan?

Oxford: Oxford University Press. (Accessed 12 August 2017) Retrieved from https://en.oxforddictionaries.com/definition/market

Parsaye K. & Chignell M. Intelligent Database Tools & Applications.

Pastore, M. (1998, December 03). Internet Becoming Necessity to Users. Retrieved from https://www.clickz.com/internet-becoming-necessity-to-users/72138/

Ramageri, B. M. (2010, December). Data Mining Techniques and Applications.

Robertson, P. (2012, November). Robertson's Book of Firsts: Who did what for the first time?

Smita & Sharma P. (2014, June). Use of Data Mining in Various Field: A Survey Paper.

Ward, M. Overview of Data Visualization. Retrieved from www.cs.wpi.edu

APPENDIX

APPENDIX A: SAMPLE EXPERIMENT PAPER

#	Search Term/ Keyword	No. of relevant items in set	No. of Candidate – Best Deal items (C-BD)	No. of Non-Candidate – Best Deal items (NC-BD)	Total Accuracy
1	Drawing Tablet				
2	Bluetooth speaker				
3	Gaming Mouse				
4	LED keyboard				
5	Gaming Keyboard				
6	LED Mouse				
7	Silent Mouse				
8	Bluetooth Earphone				
9	Gaming Headset				
10	Digital Camera				
11	Bluetooth Controller				
12	PS3 Controller				
13	Silent Keyboard				
14	Wireless Keyboard				
15	Wireless Mouse				
16	Phone Charger				
17	Power Bank				
18	Wireless Earphone				

19	Bluetooth Mouse				
20	Midi Keyboard				
		Total:	**Total:**	**Total:**	**Total:**