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

Averos, Christian M.
Bachelor of Science in Computer Science
Polytechnic University of the Philippines Sta. Mesa, Manila
0995 771 7060
christianaveros@gmail.com

Delicano, Jobea Ann F.
Bachelor of Science in Computer Science
Polytechnic University of the Philippines Sta. Mesa, Manila
0955 855 3254
jobea.delicano@gmail.com

ABSTRACT — As new technologies arise, online food, apparel, and product delivery services came to spotlight. Along with the growth of online shopping websites comes with the growth of merchants offering different prices for the same product. In order to lessen the time consumed searching for the optimal product for buyers with a fixed budget, the study aimed to develop a recommender system that searches and scrapes data on online shopping websites to provide lists of top 10 ranking products based on a combination of price and reviews, based on price, and based on reviews in a single page layout based on the user's search term input. The system used classification technique in data mining to identify the products a user wants to buy and Interleaved Ranking Algorithm list out the best products in the online market. Through an experimental research method the system achieved 90.57% accuracy on acquiring relevant data, 72.80% on getting the best deals in terms of price ranking, and 58.40% in terms of reviews ranking. Therefore, having a 73.92% on the system's overall accuracy.

Keywords: Data Mining; E-Commerce; Interleaved Ranking Algorithm; Recommender Systems; Data Visualization

1. 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. [1] 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 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 searching through many OSWs just to get the best deals? The researchers are used to seeing their mothers go through many stores just to get the optimal price that fits in their budget, sometimes, they 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).

2. RELATED STUDY AND LITERATURE

2.1 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. [2] 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). [3] KDD consists procedures including [4]: selection of data from varying sources; pre-processing 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 [5]: Association, Classification, Clustering, Prediction, Sequential Patterns and Discrimination.

Before acquiring data to be displayed, websites should be visited first. Web crawlers are used to automate website visits from time to time and partnered with a scraper in order to get needed data. Data mining techniques like classification is used in order to crawl to needed webpages to avoid wasting resources and to lessen noise from the data that will be obtained. [6]

Web scraper functions as a tool for obtaining the specific data needed from the website. Different data mining techniques are applied to get the data in a systematic way by grouping the data to discover hidden aspects of the data. Classification technique is used to scrape data given that there is a predefined class given as a basis. [7]

2.1. DATA ANALYSIS

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.

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 Team Draft 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.

2.2. DATA 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. [8] 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. [9]

3. SYSTEM OVERVIEW

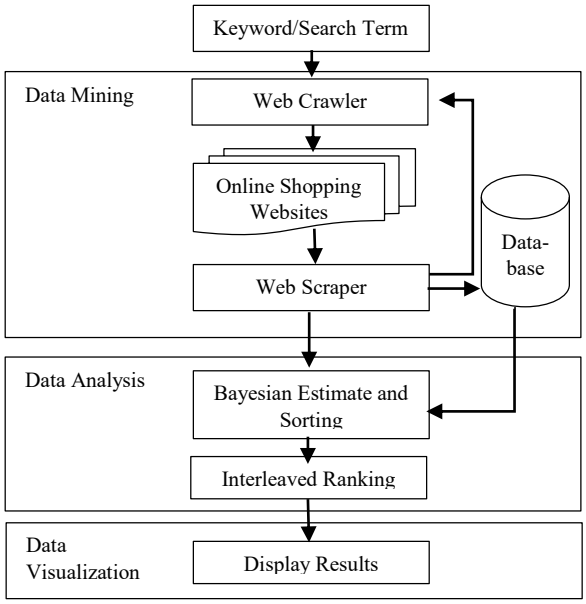


Figure 1. System Architecture

Figure 1 shows the system architecture of the system. The architecture is divided into three parts:

Data Mining. The search term will be used to search and extract products from three online shopping websites namely Amazon, Lazada and Shopee. The extracted products are stored in the 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.

4. TEST RESULTS

The test results contains data from fifty (50) unique search terms which were technology related. The raw results were recorded in an experiment paper then analysed and interpreted.

In order to get the performance of the system in getting the relevant items, the percentage formula was used:

$$\% = \frac{number\ of\ relevant\ items}{total\ number\ of\ items} \times 100$$

The number of relevant items is the count of the product items which is equal to the user's expectation. The total number of items is the count of the product items which was extracted from the online shopping websites. The performance of the system in getting relevant items using web crawler and web scraper is shown below:

Table 1
Performance of the developed system in using web crawler and spider

Total accuracy	90.57%

The total accuracy of the developed system in getting the relevant items using web crawler and spider is 90.57%. The researchers found out that the search term that got high accuracy displayed products which are aligned to the user's expectations. While search terms which got low accuracy displayed products which only match the name but not the user's expectations.

In order to get the performance of the system in getting the best deals in terms of price and customer review rankings, the formula for recall was used:

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

The *cb* is the number of candidate-best deals. It is the count of products in best deal rankings which is present in price rankings and in customer review rankings. The *nb* is the number of non-candidate-best deals. It is the count of products in best deal rankings which is not present in price rankings and in customer review rankings. The performance of the system in identifying the best deals in terms of price and customer review rankings.

Table 2
Performance of the developed system in identifying the best deals in terms of price and customer reviews

Accuracy of identifying the best deals in terms of:	
Price	**Customer Reviews**
72.8%	58.4%

The performance of the system in identifying the best deals in terms of price is 72.8% while 58.4% in terms of customer reviews. In terms of price, the search terms with high accuracy in identifying best deals listed majority of the products with low price and high reviews while search terms that got low accuracy listed majority of the products with low price but were poorly reviewed. In terms of customer reviews, search terms with high accuracy in identifying best deals displayed products with excellent reviews and low prices while those with low accuracy displayed excellent reviews yet high price.

In order to obtain the overall performance of the system, the mean formula was used:

$$\bar{x} = \frac{p1 + p2 + p3}{3}$$

The *p1* is the performance of the system in getting the relevant items. The *p2* and *p3* is the performance of the system in identifying the best deals in terms of price and customer reviews, respectively. Below is the overall performance of the developed system:

Table 2

Overall performance of the developed system

Accuracy of getting relevant items	Accuracy of getting the best deals in terms of:		Overall Accuracy
	Price	Customer Reviews	
90.57%	72.80%	58.40%	**73.92%**

The overall performance of the system is 73.92%. The factors that affect the performance of the system includes the ambiguity of the search term and the correlation of the price and customer reviews of products. The overall performance of the system is moderately high.

5. CONCLUSION

By testing the system and analysis of results, the following had been concluded by the researchers:

1. The researchers found that accuracy of getting the relevant data in online shopping websites heavily weighs on the ambiguity of the search term. One of the test results, “4GB ram” search term test got the lowest accuracy of getting the relevant data since the term can be used to specify a desktop computer hardware or a specifications of a smartphone unit. VR Headset got a 100% accuracy of getting the relevant data due to lack of ambiguity. Therefore, the ambiguity of the search term is a factor that affects the accuracy of getting the relevant data.

2. The total accuracy of getting the best deals in terms of price is higher than in terms of reviews due to the products in the price ranking having a good review rates and review rankings having listing highly reviewed yet expensive products. The highly reviewed expensive products will rank low in the price rankings and the products with good reviews which ranked high in the price rankings will be ranked on a middle rather than low. During the merging of two rankings, the products which topped the price ranking would get ranked high in the best deals rank. The researchers therefore concluded that the accuracy in getting the best deals in terms of price and customer reviews is dependent on the correlation between price and reviews of the products.

3. The overall accuracy of the System is moderately high. The researchers found that there are factors which affects the overall accuracy of the System. The factors that affects the accuracy of the system are the ambiguity of the search term, and the price and review of the products.

6. RECOMMENDATION

The researchers recommend the following to improve the effectiveness of the system:

1. Using filtering options in search function to increase the performance of the system in getting the relevant data
2. Using different merging method to combine price rankings and customer review rankings to increase the performance of system on getting the best deal ranking in terms of price and customer review rankings.

3. Adding other online shopping websites for wider selection of products.
4. Using different data mining approach to improve the speed of data extraction from different online shopping websites.
5. Improvement of user interface and user experience to be pleasing to the user's eyes.
6. Testing the system on different time intervals to check the differences and variations of results' effectiveness on the accuracy of time.

7. REFERENCES

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

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

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

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

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

[6] Mitra, P. & Tan, Q. (2010). Clustering-based Incremental Web Crawling

[7] Dai, K., Nespereira, C., Vilas, A., & Redondo, R. (2015). Scraping and Clustering Techniques for the Characterization of LinkedIn Profiles.

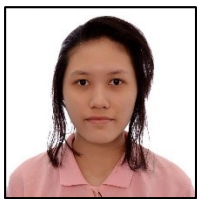
[8] Zhao, K. (2002). Data visualization.

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

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

8. CURRICULUM VITAE

Christian M. Averos. He was born on January 27, 1998 and currently living at Quezon City. He attended Manuel A. Roxas High School for his secondary education. He is currently a senior college student from Polytechnic University of the Philippines taking Bachelor of Science in Computer Science. He took his internship at Save 22, inc. under I.T/Engineering Department. His field of interest includes Game Development.



Jobea Ann F. Delicano. She was born on June 21, 1998 in Pasig City and currently residing at Antipolo City, Rizal. She attended Nativity of Our Lady Parochial School during her primary education and Roosevelt College Cainta during her secondary education. At the moment, she is a fourth year college student at Polytechnic University of the Philippines and currently taking Bachelor of Science in Computer Science. She took her internship at Save 22, inc. under I.T/Engineering Department. Her field of interest consists of Data Science and Web Development.