An Efficient Knowledge Discovery Technique to Identify the Product of Interest for Web-buyers

1S. Sandhiya and 2Dr. A. Kannan

1Lecturer, Department of Information Technology, University College of Engineering Villupuram, Tamil Nadu, India
2Professor, Department of Information Science and Technology, College of Engineering Guindy, Chennai, Tamil Nadu, India

ABSTRACT

Web shopping is one of the most important industries for growing and fast environment all across the world. Information technology plays a vital role in promoting this type of industries. Understanding and adapting to changes of buyer’s behaviour is an important aspect for a web shopping industry to endure in an incessantly changing environment. The foremost challenge for this type of service provider is to identify the buyer’s product of interest with the continuously changing environment. The aim of this study is to expand the methodology which sense buyer’s product of interest using knowledge discovery techniques. We have utilized web crawler to obtain online buyer’s Account, orders, wish list, reviews and rating information’s. This study espouse a new perspective in analyzing buyer’s profiles using both quantitative and qualitative analysis methods were performed on the data collected from one of the India’s largest sales website. The final results from both data analysis method combined together which discovers the buyer’s product of interest based on their profile information such as gender, age group, and their occupations etc. This result can also useful for the organization to select profiles for posting ad in social websites like Facebook, Twitter, etc. Also recommend various offers for new customers by matching with analyzed profile data.

INTRODUCTION

Shopping is one of the essential part of our daily life. We’re using different types of shops to buy different kind of things every day. Online shopping has grown in popularity over the years, mainly because people find it convenient. One of the most attractive factor about online shopping, particularly during a holiday season, it alleviates the need to wait in long lines or search from store to purchase for a particular item.

Among various kinds of online information sources Online shopping has been growing very fast in India. Online shopping offers fast, easy, money saving and interesting shopping experience, it has many advantages like 24 hours shopping, shopping with coupon to get discount, shopping from Home, rich product availability and specifications, free shipping offers etc. Ebay and Amazon are the market leaders in online shopping in the world. However local Shopping deal site are more popular in India. Based on voting the top five popular websites offering online shopping in India are amezon, flipkart, snapdeal, paytm, and ebay as on October 2014. Flipkart is one of the most popular shopping site in india, founded by sachin bansal and binny bansal in 2007. It has google page rank #6, it receive 2.2 million page views per day and generates 46,574 every day in advertising revenue. In addition, it has a comprehensive storage of customer’s data, which could be used in our research.

This paper focuses on the role of data mining in the online shopping industry. In essence, we analyzed buyer behaviour patterns so as to discover useful and previously hidden knowledge in order to recommend appropriate product for buyer. Moreover, (Atae et al., 2014) quantitative and qualitative analyses have been undertaken simultaneously in this paper.

In the world of research, there are two general approaches to gathering and reporting information: qualitative and quantitative approaches. The qualitative approach to research is focused on understanding a phenomenon from a closer perspective. The quantitative approach tends to approximate phenomena from a larger number of individuals using survey methods. The aim of qualitative analysis is complete, detailed
description. No attempt is made to assign frequencies to the linguistic features which are identified in the data, and a rare phenomenon receives the same amount of attention as more frequent phenomena. The qualitative approach to gathering information focuses on describing a phenomenon in a deep comprehensive manner. This is generally done in interviews, open-ended questions, or focus groups. In most cases, a small number of participants participate in this type of research, because to carry out such a research endeavor requires many resources and much time.

The qualitative and quantitative approaches to research allow a different perspective of situations or phenomena. These two main approaches to research are highly informative, especially if used in combination. Each approach has its benefits and detriments, and being aware of the methods used to gather information can help practitioners and policy-makers understand the extent to which research findings can be applied. In this paper both quantitative and qualitative methods are used for knowledge discovery process. Quantitative analysis k-means and Apriori algorithm was applied using Weka machine learning software and qualitative analysis was performed using HyperRESEARCH software, based on buyer’s data. This study can support e-shop agencies to promote attractive products for new customers.

Related work:

In 2014, Atae Rezaei and Aghdam Mostafa proposed knowledge discovery technique to analyze tourists’ profiles where both quantitative and qualitative analysis methods. Results from both data analysis methods that supported each other were used to build a more definite model. This result shows tourist’s places of interest which is useful for tourism organizations in Malaysia. In 2013, Wei Sun et al, proposed a knowledge discovery system for detecting and visualizing knowledge evolution patterns of a research field. It is mainly focused on co-word technology and core-based algorithm of tracking knowledge evolution. They proposed six kinds of knowledge evolution patterns also they elaborated key technologies involved in the system construction, knowledge structure building, knowledge evolution pattern detection and visualization.

In 2013, Emma Ben et al, introduced the temporal data visualization techniques which are used in the KDD stages to increase the user participation as well as its confidence in the result in order to improve the decision support quality and their applicative context is the fight against nosocomial infections in the intensive care unit. In 2014, Tsu-Kuang et al introduced Knowledge Discovery technique to Intangible Asset and Real Estate Value. Their study improves product quality, enhance added-values and upgrade customer satisfaction for the accomplishment of the goal for sustainable development. It is hoped to study if applying knowledge to intangible asset and real estate values may help enhance the competitiveness of real estate industry. Through related analysis and regression model as well as application of balanced scoreboards, it was aimed to study the application of knowledge discovery to real estate values as well as its impacts on developers, consumers and real estate agents. Should the enterprises change the existing patterns and move forward to strategic alliances, it shall help create a win-win situation.

In 2013, Lina Hou et al, proposed knowledge discovery method for selection rules of acupuncture points in respiratory diseases therapy based on the theory of Structural Partial-Ordered Attribute Diagram and association rule mining. Their proposed technique works well in discovering new knowledge from medical treatises and clinical cases of acupuncture treatment. This method provides a scientific and advanced technological means for the heritage of Traditional Chinese Medicine. In 2014, Dennis Thom and Harald Bosch introduced two strategies for Social Media Location Discovery that are suitable to assign probable locations of origin to social media messages of unknown locations. They are based on aggregated knowledge about the author and/or the textual content of the message. Their result shows that they can locate up to 74% of all messages that were written in specific cities and about 20% of messages written in specific districts.

Methodology:

Our methodology is classified into five phases such as (a) Data collection, (b) Data Representation, (c) Data pre-processing, (d) Knowledge Discovery, and (e) Knowledge. Figure 1 represents the phases of proposed study with input and output in each phase.

The first phase of this methodology is data collection in this phase the required data is collected from the studied site (which is one of the India’s largest online shopping site) this data includes Gender, Age, Orders, Wish list, Reviews, Ratings etc. There are different types of products available in this site, but we have taken buyers interest in Mobile phones for analyzing. Data collected during particular time period shows 439 buyer’s profiles that have been taken into account for analyzing buyer’s interest on mobile phones. The required data has been collected automatically using specific web crawler.

In the second phase of our methodology the retrieved unstructured data sorted into structured data which provides buyer’s profile information in CSV format. Data pre-processing is an important step in knowledge discovery process. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization,
transformation, feature extraction and selection. In the data pre-processing phase, buyer’s profile information in CSV file has been preprocessed and converted into ARFF (Attribute-Relation File Format) file to load it into Weka machine learning software and HyperResearch software.

In the next phase the given ARFF file has been analyzed using two techniques such as quantitative analysis (K-means and Apriori) using Weka and qualitative analysis using HyperResearch. In the final phase the result of both analysis methods are combined and gives us results on buyer’s interest according to the attributes such as Gender, Age, Order, Wish list and Rating given as input.

![Fig. 1: Methodology Architecture.](image)

**Implementation and result analysis:**

For implementation we have taken 439 buyer’s profile data as input. The input taken was in raw CSV file format which has been pre-processed and converted into ARFF file format. The sample training set which has been pre-processed in ARFF file format is shown below.

```plaintext
@relation buyer.profile
@attribute name {vijay, suriya, saranya}
@attribute age {17, 28, 45}
@attribute gender {male, female}
@attribute wishlist {tablet, slimfit, smallsize}
@attribute likes {yes, no}
@data
vijay,17,male,tablet,yes
vijay,17,male,slimfit,no
```
ARFF file has been given as input to Weka explorer for quantitative analysis. For clustering we have used K-means algorithm which has processed the training set data into group of clusters based on relevant attributes. Sample output from clustering has been shown below with three buyer’s profile.

For association we have used Apriori algorithm for data analysis. Below shows the output processed using Apriori algorithm for the above clusters.

--- Run information ---
Relation: buyers.profile
Best rules found:
1. age=17 3 ==> name=vijay 3 conf:(1)
2. name=vijay 3 ==> age=17 3 conf:(1)
3. gender=male 3 ==> name=vijay 3 conf:(1)
4. name=vijay 3 ==> gender=male 3 conf:(1)
5. age=28 3 ==> name=suriya 3 conf:(1)
6. name=suriya 3 ==> age=28 3 conf:(1)
7. name=suriya 3 ==> gender=female 3 conf:(1)
8. age=45 3 ==> name=saranya 3 conf:(1)
9. name=saranya 3 ==> age=45 3 conf:(1)
10. name=saranya 3 ==> gender=female 3 conf:(1)

By combining these two algorithms quantitative analysis on given training set has been examined. For qualitative analysis buyer’s review, ratings and comments on product are given as input. The raw training set data has been processed using hyperresearch software and the result shows buyer’s preference in choosing the product and key factors of the product. By combining both quantitative and qualitative methodologies knowledge discovery has been achieved on training set data. Finally buyer’s product of interest has been analyzed from the 439 buyer’s profile training set data.

Recommendations for buyers:
According to the results provided in the previous sections, we can recommend a list of appropriate product and services for specific groups of customers based on result analysis. For product of interest, we can suggest the model below for buyer. This model demonstrates suitable product for each group of buyer’s for instance, it can be seen that android mobile is one of the best options for both the male and female buyers.

<table>
<thead>
<tr>
<th>Gender/Age</th>
<th>Male</th>
<th>Battery Life</th>
<th>0-19</th>
<th>20-39</th>
<th>40 above</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td></td>
<td>Battery Life</td>
<td>NA</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mobile size</td>
<td>Big</td>
<td>Slim</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model</td>
<td>Tablet</td>
<td>Smart phones</td>
<td>classic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type</td>
<td>Touch</td>
<td>Touch</td>
<td>keypad</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td>Battery Life</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mobile size</td>
<td>Big</td>
<td>Slim</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Model</td>
<td>Tablet</td>
<td>Smart phones</td>
<td>classic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Type</td>
<td>Touch</td>
<td>Touch</td>
<td>keypad</td>
</tr>
</tbody>
</table>

Table 1: A recommendation model for product of interest in mobile phones based on both quantitative and qualitative analysis
Conclusion:
This study exhibited the use of information technology in online shopping industry in order to discover hidden knowledge from online buyer’s profiles. Quantitative and qualitative analysis techniques were applied together to extract knowledge from buyer’s profile. The consequences of this research are beneficial for the online shopping organizations to identify buyer’s product of interest and buyer’s behavior in order to improve facilities and services in shopping sites.

This study only applied for 439 buyer’s profiles from the online shopping site and data collection was limited to particular product (mobile phones) Future study can be expanded to the various products and the use of data from several online shopping sites.

REFERENCES

Atae Rezaei Aghdam, Mostafa Kamalpour, Dong Chen and Alex Tze Hiang Sim, 2014. Identifying Places of Interest for Tourists using Knowledge Discovery Techniques. IAICT 2014, 978-1-4799-0, IEEE.
Wei Sun., Xuefu Zhang and Huai Wang, 2013. A Knowledge Discovery System for Detecting and Visualizing Knowledge Evolution of a Research Field. 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) 978-1-4673-5253-6 IEEE.