A Review Based Web Mining for Online Store Selection

Sanjeev P. Kaulgud
Department of Computer Science and Engineering, Presidency University, Bengaluru
sanjeev.kaulgud@gmail.com

Abstract- Online trading even in the retail sector is gaining more and more popularity with the widespread use of Internet. The main advantage of online shopping is the convenience of finding the right kind of products with the required quality of service, on-time delivery, shipping charges, in addition to the right price, at the click of the mouse. However, the problem with such online shopping is the bewildering array of options, which are available on the Internet. This causes customers to go by the rating of the online stores made by the users. These ratings may not be the true indicators of the online stores.This leads to erroneous selection of the online stores.In this paper, a novel method is proposed which works in two stages: In the first stage, it identifies and extracts the values of the overall rating and the presence of those features in the full review. The second stage focuses on establishing the co-relation between overall rating of the online store and opinion given by the customer in the full review. The paper also reveals that, there is no similarity between overall rating and full reviews posted by the customers. The authors are of the strong opinion that the customer's decision of selecting the right online store has to be based on summary of the full review and overall rating of only those reviews which are correlated.

Keywords: Online Shopping, Reviews, Rating values Extractor, Co-Relation.

I. Introduction

Online trading is one of such sector that is gaining more and more popularity with the widespread use of Internet. Online shopping offers many advantages in terms of choice, access to goods and services. A considerable number of studies have been conducted on the effects of comments on the web. People's appraisal has high confidence to express their behaviors and aspects. Online shopping allows customers to openly buy things or services from a seller over the Internet using a web browser. Customers find a merchandise of interest by visiting website of the retailer directly or by searching amongstother vendors by means of a shopping search engine, which will display the same product's availability and price at different e-retailers. As of 2016, customers can shop online using a range of different computers and devices, including desktop computers, laptops, tablet computers and smartphones.

An Internet presentation management company held a survey [1] on over 1400 customers across 11 countries in Middle-East, North America, Asia and Europe and the results of survey are given below:

- Online retailers must increase the speed of website.
- Online retailers should ease customers' fear of security.

These worries majorly distress the decisions of nearly two thirds of the customers.

Fig 1: Year-wise Online shopping growth

Thus, the appraisals on the web are significant information for customer making their decision. Through the concept of web, people could share their opinion on the web freely Many companies develop their business around this concept. Surfer can obtain lots of useful information on the web with these reviews added by user rapidly [2]. We get product information only from the manufacturer or our friends who experienced it in the past time, but through the experienced people share their comment, we could know a product more clearly and do better decision. Thus, it is very essential to know which online store is the best one where the customer can buy a product [3]. In order to make this decision the customer has to go through the opinions on the different online stores.

There are two review formats by which the customer can give his review
Format(1) – Overall Rating:The reviewer is asked to give his opinion in the form of rating, in the scale of 1-5 on some important features of the online stores. The example of overall rating is shown in the sample web page of Fig.2.
Format (2) – Full review:The reviewer is asked to express his opinion in the form of sentences. The format of the full review is shown in the sample web page of Fig2.

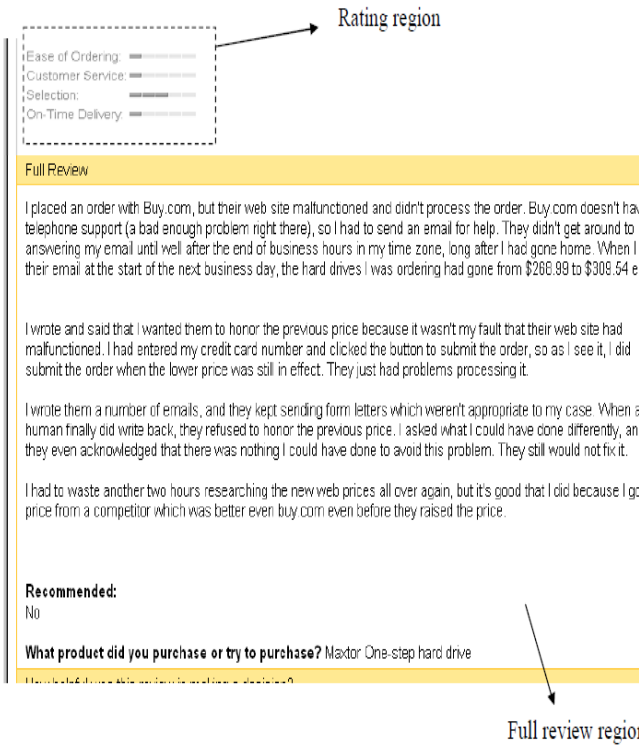

Fig. 1 Sample Rating Region and Full Review Region

Fig 2: Rating Region and Full Review Regions

To buy any particular product available on different online stores, the customers need to select a better online store which gives him all the services that he is looking for. The better way of selecting a online store is by reading the earlier customer opinions expressed in two formats. It is expected that, the opinion expressed by the customer in both the format should be similar [4]. In this paper, I propose a novel and an effective technique which reveals that, there is very less co-relation between the overall rating and full review.

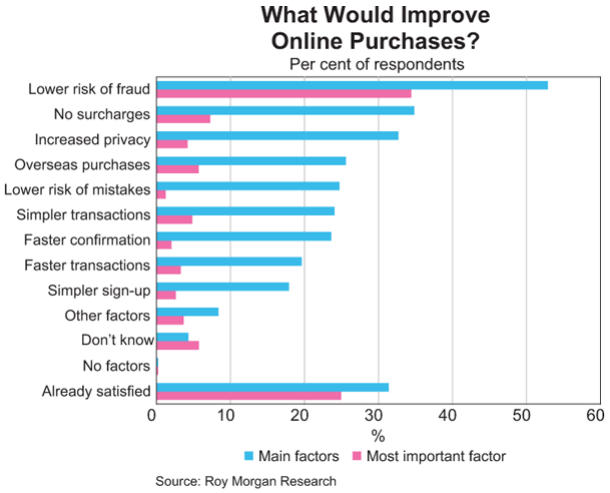

Fig 3: Factors to be considered for online purchases.

The input to our system is web review pages that contain the overall rating and the detailed reviews given by the user on online stores. Then the system works in three stages.

Stage 1: It identifies and extracts the Review Region containing only overall rating and full review region.

Stage 2: It extracts values of the overall rating and finds the existence of the features in the full review

Stage 3: Focuses on establishing the co-relation between overall rating and full reviews

II. Proposed technique

In this paper, I propose a novel method called "Web Mining For Online Store Selection: A Review Based Approach" which identifies and extracts the values of overall rating and the presence of those features in the full review and then focuses on establishing the correlation of overall rating and full review. When multiple online stores are available for shopping, the issue is which one is the better choice. To select a particular online store, generally the customer looks into the opinions of the earlier buyers. These opinions are expressed in two formats as mentioned in introduction section.The reviews posted by the customers in both the formats are supposed to be similar. This paper reveals the correlation between the overall ratings and the full review. The system model of the proposed system is shown in Fig.3, and it consists of the following components

- Review Region Extractor
- Rating values extractor and Feature existence
- Relation component

The output of each component is the input for the next component.

A. Region extractor

Given a URL of a web page containing customer reviews on the online stores, region extractor identifies and extracts the region that contains the rating region and the full review region which is the largest rectangle [5] as shown in fig 4.

In order to identify the largest rectangle, the HTML file is scanned for the tags [10]. For each tag, there exists an associated rectangular area on the screen. This rectangle is called the bounding rectangle for the particular tag. For each tag encountered, the co-ordinate of the top left corner, height and width of the bounding rectangle of that tag is determined.

Based on the height and width of the bounding rectangle, the area of the bounding rectangles of each of the children of the BODY tag is determined. Then the largest rectangle among these bounding rectangles is determined.

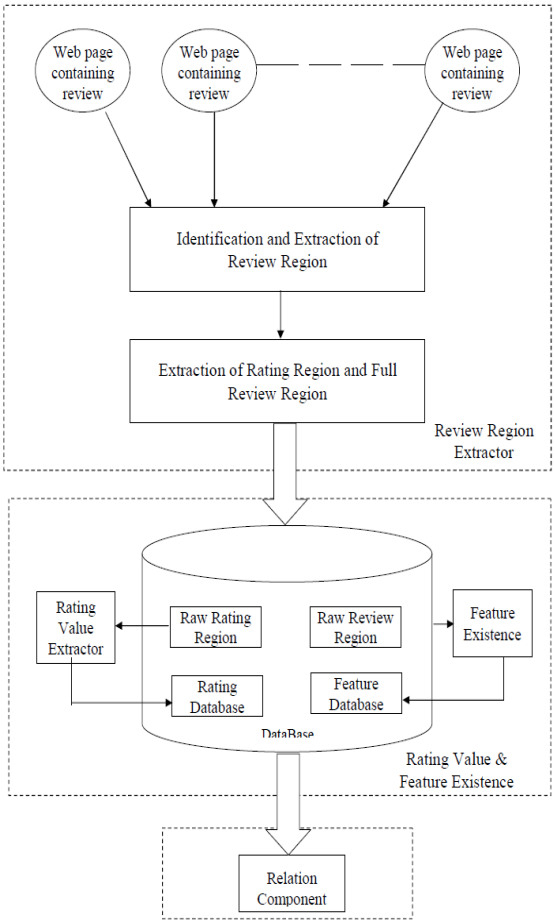

Fig.2 The System Model

Fig 4: System Model

This rectangle will be the review region containing rating region and the full review region as shown in the fig 2. The algorithm Find Region (BODY) finds the Region that contains rating region and full review region as in fig 5.

Algorithm *FindRegion(BODY)*

{

for each child of BODY tag

find the co-ordinates of the bounding rectangle for the child

if the area of the bounding rectangle > area of maxRect

maxRect = child

}

Fig.4 Algorithm to find the largest rectangle

Fig 5: Algorithm to find the largest rectangle

B. Extraction of Rating and Full Review Region

The input to this component is the review region after eliminating all the irrelevant regions, i.e. the output of previous step. In this step we extract and separate the overall rating region and full review region. It is observed that the area covered by bounding rectangle of full review region is always greater than the overall rating region on most websites. Hence I find the area under both the rectangles and compare them. The largest is considered to be full region and the other one is overall rating region [5]. The algorithm FindFullRevRec&RatingRect (maxRect) finds the rating region and the full review region.

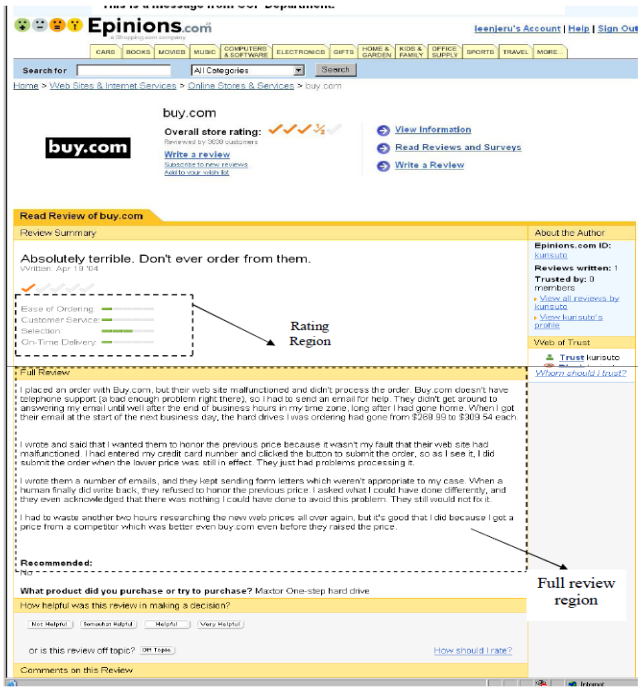

Fig.3 An example of a Review region containing Rating Region and Full Review Region

Fig 6: An example of a Review region containing Rating Region and Full Review Region

C. Rating values Extractor and Feature Existence

This section discusses about the extraction of rating values and the presence of same feature in the full review. They are discussed in detail in the following section.

Rating values Extractor

The input to this component will be the rating region. This component extracts the value of each feature and stores it in a database. The Algorithm RatingValExt identifies the feature and extracts its values and stores it in database. The HTML source file of the rating region is scanned and for every feature in the rating region the count is set to 0 and is incremented as soon as the fill color is green and is stored in the rating value database.

```
Algorithm FindFullRev&RatingRect(maxRect)
{
  for each tag associated with maxRect
     if area of bounding rectangle of tag>half the area of maxRec
          FullRevRegion = tag
          else
          RatingRegion = tag
}
```

Fig. 5 Algorithm to find the Rating Region and Full Review Region

Fig 7: Algorithm to find the Rating Region and Full Review Region.

This is repeated for all m reviews stored in the raw rating region and is stored in the database in the following format

Table.1 Sample Rating Values

Review No	Ease of ordering	Customer service	Selection	On time Delivery
R1	5	4	5	5
R2	4	4	5	4
R3	2	3	4	5
R4	5	5	5	4
R5	4	4	4	5
:				
R_m	-	-	-	-

Table 1: Sample Rating Values

Feature value existence

Here I present the most effective method of finding the existence of the feature from the full review [7]. The input to this component is the full review region and the feature set. The full review region is scanned for all the features present in the rating region and its presence is checked. If the feature exists in the full review region then its existence is marked as 1 and stored in the matrix otherwise 0 as shown in the table [8].

The sample feature existence matrix constructed for few reviews using the above algorithm is given in table2

Table.1 Sample Feature existence matrix that contains few reviews

Rev No	Ease of Ordering	Customer Service	Selection	On-time Delivery
	F1	f2	f3	f4
R1	0	0	0	0
R2	0	0	1	0
R3	0	0	0	0
R4	1	1	1	0
R5	0	0	0	0
R6	0	0	0	0
R7	0	0	1	0
R8	0	0	0	0
R9	0	1	0	0
:				
Rm	-	-	-	-

Table 2: Sample Feature existence matrix that contains few reviews.

D. Relation Component

In this section an attempt is made to find the co-relation between overall rating(A) and the full review(B). The data obtained in the previous section are given to the relation component which checks how strongly overall rating implies the full review. The Co-relation between A and B can be measured by

$$r_{A,B} = \frac{\sum (A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A \sigma_B}$$

Where , n - Number of tuples
$\overline{A}, \overline{B}$ - Mean values of A and B
$\sigma_A \sigma_B$ - Standard deviation of A and B

$$\sigma_A = \sqrt{\frac{\sum (A-\overline{A})^2}{n-1}}$$

If the value of r in the above equation is greater than zero, then overall rating and full review are positively correlated. If the resulting value is equal to zero, then they are independent and there is no correlation between them. If the resulting value is less than zero then they are negatively co-related.

III. Experimental Results

The proposed technique is applied and the correlated results are obtained from the overall rating and the full review. Experiments are conducted by taking reviews from the web pages and assessing them as how they are correlated to each other.

Here, 1000 customer reviewshave been randomly taken on one online store namely Amazon.com. The results obtained are tabulated in table 3

Table. 3 Summary of Corelation

No. of Reviews	Positevly Corelated	Not Corelated	Negatively Corelated
1000	160	720	120

Table 3: Summary of Correlation.

The experimental results in the table show that 72% of the reviews are not co-related. 16% of reviews are positively co-related and only 12% of the reviews are negatively co-related.

IV. Conclusion

This paper focuses on a new approach to correlate the overall rating and the full review given by the customer on the online stores. It is observed that there is no direct correlation between overall rating and full review. Hence, to select a particular online store for online trading it is necessary to consider the opinion expressed in both formats not only one.

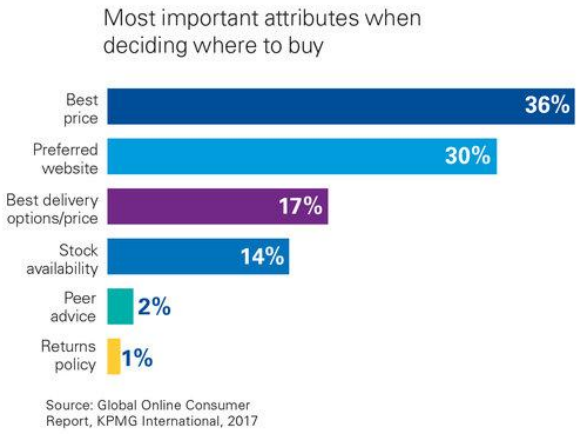

Fig 8: Attributes to be considered for online shopping

The summarization of full reviews and extracting the other important features and it synonyms gives the scope for future study.

References

[1] N. Xiang, C. w. Chung and S. Shang, "A Decision Making Method Based on TOPSIS and Considering the Social Relationship," 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, 2018, pp. 90-97.

[2]"Consumers and their online shopping expectations – Ecommerce News". 20 February 2015. Retrieved 29 August 2016.

[3] Shalini, G. R., &Malini, K. S. H. (2015). A study of online shopping website characteristics and its impact on consumer intention to purchase online in Chennai. International Research Journal of Engineering and Technology, 2(9).

[4] Dewi, M. A. A., Nurrohmah, I., Sahadi, N., Sensuse, D. I., &Noprison, H. (2017, October). Analysing the critical factors influencing consumers'e-impulse buying behavior. In Advanced Computer Science and Information Systems (ICACSIS), 2017 International Conference on (pp. 81-92). IEEE.M. A.

[5] Benchalli, S.S, P.S Hiremath, Siddu Algur, and Renuka Udapudi. "Mining Data Regions from Web Pages." COMAD2005b. 2005 DEC.

[6] Dejong, G. "An Overview of the FRUMP System." In Stratregies for Natural language Parsing, 149-176. 1982.

[7] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." KDD. 2004.

[8] Jacquemin, C, and D Bourigault. "Term Extraction and automatic indexing." In Handbook of Computational Linguistics. Oxford University Press, 2001.

[9] Kupiec, J, J Pedersen, and F Chen. "A Trainable Document Summarizer." SIGEIR. 1995.

[10]Liu, Bing, R Grossman, and Y Zhai. "Mining Data Records in Web pages." SIGKDD 2003. 2003.

[11]Liu, Jingjing, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. "Low-Quality Product Review Detection in Opinion Summarization." Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague, 2007. 334-342.

[12]Paice, C D. "Constructing Literature Abstracts by Computer: techniues and prospects." Information Processing and Management, 1990: 26.

[13]Tait, J. Automatic summarising of english text. PhD Dissertation, Cambridge: University of Cambridge, 1983.

[14]Wu, Y C, T K Fan, Y S Lee, and Show-Jane Yen. "Extracting Named Entities Using Support Vector Machines." KDLL. 91-103, 2006.

[15]Y, Zhai, and B Liu. "Web Data Extraction Based on Partial Tree Alignment." WWW-05. China,Japan, 2005.

[16] JifengLuo et al. (2012) , "The Effectiveness Of Online Shopping Characteristics And Well-Designed Websites On Satisfaction". MIS Quarterly Vol. 36 No. 4, pp. 1131-1144.

[17]Sannakki, S. S., & Kaulgud, S. P. (2012). Memory learning framework for retrieval of neural objects. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(6), pp-100.