ISSN: 2688-8645

Techniques of Online Reviews Mining

Shuang Zhang

School of economics and management, Xidian University, Xi'an 710000, China

zhangshuang_zhang@126.com

Abstract

With the development of web2.0, many online shopping websites have emerged, generating large number of online reviews. Data mining has proven very valuable in many fields, including online reviews mining. This paper introduced three kinds of techniques to do online reviews mining, including machine learning, rue mining, and natural language processing. Besides, the general mining process was introduced. By the way, some findings on the reviewed papers' data sources were presented.

Keywords

Data mining, online reviews.

1. Introduction

With the development of e-commerce, online consumption has become an important channel for people to consume. At the same time, there are more and more online shopping websites, and a large number of online reviews are produced. In recent years, the combination of social media and e-commerce has made online reviews more diversified. For example, live shopping is very similar to the previous TV shopping, but not the same. It depends on the host to recommend products to fans to promote sales, and the fans' message to the host also contains a lot of review information. Thus, A huge amount of online reviews have been generated and are being generated in various evaluation systems.

The evaluation system of online website usually includes reviews in the form of text, picture, video and star rating. In view of the variety of forms, complexity of data and valuable business information in online reviews, how to extract these business information from online reviews has become an important source of competitive advantage for enterprises ,which also attracts the research of academia. For example (Jamshidi,Soheil,et al., 2018) [1] examined the problem of detecting and characterizing incentivized reviews in two primary categories of Amazon products, because such incentivized (and often very positive) reviews can improve the rating of a product which in turn sways other users' opinions about the product.

On the other hand, researchers in different fields are applying data mining in their fields. Applications of data mining has covered human performance data, text data, geospatial data, bioinformatics, customer relationship management (CRM), computer and network security, image data, and manufacturing quality (Brzezinski, J, 2005) [2] . What's more, due to the development of text data mining technology and interests from both academia and industry, data mining has been applied to online reviews more deeply, which also promotes the development of opinion mining. Besides, the perspective of mining is also more diverse. Thus, a summary of the techniques is very necessary in the view of an increasing number of people rely on online reviews to make decisions.

This paper focuses on the techniques of online reviews mining, in which the first section introduced the background. The main section of this paper focuses on what techniques can be used to do these mining. At the same time, it introduced the general process of online reviews

ISSN: 2688-8645

mining. The following section was summary and findings.This paper covered 18 papers that focused on data mining of online reviews, all in English.

2. Background

The commercial value of online reviews should not be ignored . Many people can benefit from the application of online reviews mining.

Firstly, Mining online reviews can help managers and investors evaluate products or services .For example, hotel managers and related investors can clarify the heterogeneity of customer satisfaction in terms of latent dimensions. Take it to the next level, mangers can ascertain which dimensions influence customer satisfaction significantly and how consumer perceptions vary among different hotel classifications (Guo, et al. ,2017) [3].

Secondly, when users buy products or services, most of them will refer to word-of-mouth information, and their purchasing behavior and subsequent experience and evaluation of products or services will be directly or indirectly affected by word-of-mouth. So, online buyers can make better choices, enjoy better products and services by taking advantage of online reviews which are a form of electronic word-of-mouth.

What's more, online sellers can better understand online buyers' demands by online reviews mining , thus providing more competitive products and services, further benefiting online sellers. In addition, online reviews mining also provides some insights to online advertisers who aim to use customer-generated opinions to automatically devise an online advertising strategy for each product using the widely popular model of sponsored search advertising(Nikolay Archak, et al., 2011) [4].

Given that many roles of e-commerce can benefit, more attention should be payed to what technologies can be used to achieve them.

3. Techniques to do Online Reviews Mining

Online reviews can be posted on many websites conveniently and reviewers can post descriptive information, experiential information, and more. Based on the current research, what techniques can be used to do these mining?

There are different techniques that have been employed to online reviews mining. These techniques can be roughly divided into the following classes.

- Machine Learning
- Rule Mining
- Natural Language Processing(Varathan, KD ,et al., 2017)[5]

Machine learning (ML) is training computers to learn from data collected through past experience(Djenouri,Djamel, et al.,2019)[6]. Machine learning techniques used in online reviews mining mainly include SVM(Support vector machines), CRFs(conditional random fields) which directly models the conditional probability distribution of the output given the input, and can exploit the rich and global features of the inputs without representing the dependencies in the inputs, Clustering which is one of the most popular unsupervised learning technique. (Liu, Huang, et al.2013)[7] used SVM to identify comparative sentences of Chinese. (Li Chen, et al.,2012)[8] applied CRFs model to automatically label the words with the most suitable tags. (Xu, KQ,et al.,2011)[9]proposed a two-level CRF with unfixed inter-dependencies model which can model the dependencies between relations and entities as well as between relations and words, for capturing the features from entity and word levels. Then, they chose multi-class SVM as a benchmark to evaluate the proposed model. (Jae-Won Hong, et al.,2019)[10]conducted a text clustering analysis to explore the meaning of extracted core keywords.

ISSN: 2688-8645

Rule Mining is to mine relations from data. One of the most popular rule mining techniques is association rule mining which is mainly employed to find strong rules and patterns in online reviews mining.(Kim, EG, et al. ,2019)[11] employed text mining and association rule methods to examine consumer reviews of three different competitive automobile brands and analyzes the advantages and disadvantages of each vehicle. There are some other rule mining techniques, such as CSR(Class sequential rules),LSR(label sequential rules) but this paper didn't find their use in online reviews to the best of knowledge.

Natural Language Processing(NLP) is the interdisciplinary subject of computer science and linguistics that uses computers to process, understand and use human languages such as Chinese and English. NLP methods can be applied to analyze language at two different levels: syntactic analysis and semantic analysis. Dependency parsing is a type of syntactic analysis and aims to identify the syntactic relations between words. Semantic analysis tries to address the problem of language ambiguity by extracting the meaning of the sentence(Varathan, KD ,et al., 2017) [5]. (XU Xueke, et al., 2013) [12] used a NLP toolkit to segment reviews into sentences for hotel reviews. (SHAOZHONG ZHANG, et al.,2018) [13]used NLProcessor linguistic parser for entity word and high frequency word combined with public lexicon for sentiment word. (Shihao Zhou, et al.,2018)[14]develop a customer agility measure based on text analytics using natural language processing (NLP) and text similarity techniques.

4. Mining Process of Online Reviews

(Olson, David L, 2007) [15]introduced a data mining model which consists of six phases including business understanding, data understanding, data preparation, modeling, evaluation, deployment. This model can also be applied to online reviews.

• Business understanding: Data mining is very dependent on understanding the business, which determines the background and goal of data mining. Different businesses may also have different concerns about mining online reviews.

• Online reviews understanding: This include initial online reviews collection, online reviews description, online reviews exploration, verification of online reviews quality. The intent is to explore summary statistics, patterns, etc. of online reviews.

• Online reviews preparation: Once the online review data is initially understood, the data to be mined needs to be selected, cleaned and formatted. Online review data can be further explored deeper at this phase.

• Modeling: Some data mining software tools or methods such as clustering are suitable for preliminary analysis. According to the actual situation of online reviews, more detailed models can be applied. Generally, in order to model the data, dividing the data into training sets, test sets and even more refined sets is needed.

• Evaluation: Through the modeling and analysis of online reviews, the information and relationships contained in the data, such as user sentiment and trust, can be obtained. Looking back to the first step, evaluating these results in the context of business can lead to identifying valuable information such as user requirements not noticed before.

• Deployment: From the perspective of business, online reviews mining can be used in different business activities or stages for different purposes, and can also verify hypotheses that are previously held. At the same time, it should be noted that the models of online reviews mining need to be updated with the change of running conditions. These six stages are a useful framework and not a rigid. In fact, a certain step may be omitted or traced back to according to the actual situation and the level of analysts.

ISSN: 2688-8645

5. Data Sources

Figure1 shows the distribution of data source. As can be seen, most of the research is based on Amazon.com, which is one of the largest e-commerce companies in the United States and is one of the first to start doing e-commerce. Founded in 1995,Amazon has expanded into a wide range of other products, becoming the world's largest online retailer. Amazon has cracked down on manipulation of reviews and improved its system, while tweaking and upgrading its rating system to suit the situation. Now, Amazon calculates a product's star ratings using a machine learned model instead of a raw data average. The machine learned model takes into account factors including: the age of a review, helpfulness votes by customers and whether the reviews are from verified purchases. In general, the more real, the more natural the buyer likes, the easier it is to be accepted by the system, the easier it is to gain weight. Amazon datasets are widely used, probably because of Amazon's popularity and relatively mature rating system.

On the other hand, it's worth noting that China's Taobao and Jd are getting attention of scholars these years, which are the two largest e-commerce websites in the country. China's online retail sales reached 1.9520 billion yuan in the first half of 2019, accounting for 24.7 percent of the total retail sales of consumer goods, according to iiMedia Research (https://www.iimedia.cn/c1061/66692.html). iiMedia consulting analysts believe that since the rise of e-commerce industry in China, the importance of online shopping in people's lives continues to increase, online shopping has become an important channel for consumers to consume. In China, the user base is extensive and huge and massive online reviews are generated on kinds of e-commerce websites thus bringing great research potential.However, the problem is that some sellers pay people to pretend to be customers to fill in fake reviews, and some sellers would delete bad reviews. Researchers using this dataset need to spend more time on data preprocessing.

In addition, most of the datasets come from online shopping sites, the reason may be that shopping is an integral part of people's lives. Online shopping reviews are growing rapidly as a result of the growth of e commerce, so people can't ignore the commercial value of a large number of online reviews .The following is the hotel and tourism industry, then the civil aviation industry. These are all service industries, which pay more attention to customer experience and evaluation.

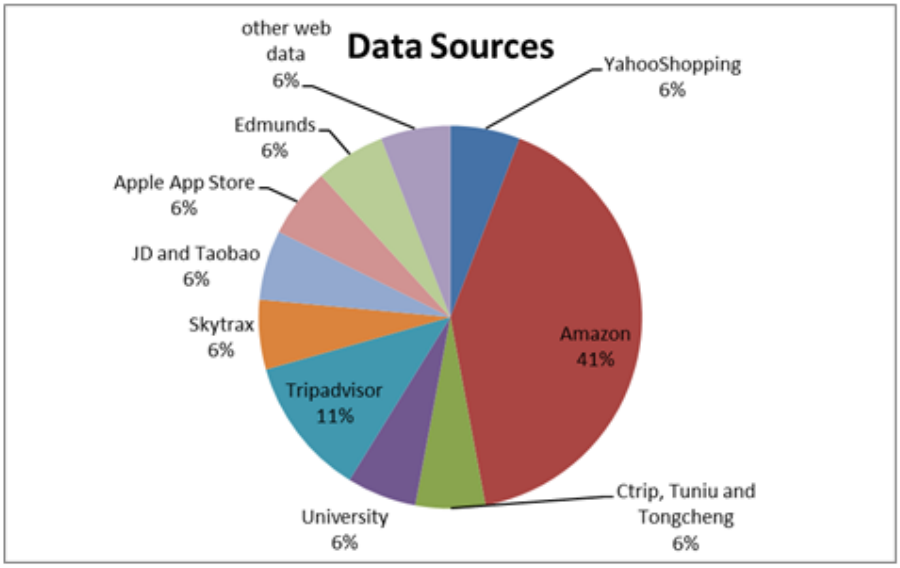

Figure 1. Data source distribution of reviewed papers.

ISSN: 2688-8645

6. Conclusion

Data mining has become a useful tool in many fields, while this paper focuses on the techniques of online reviews mining. This paper introduced three kinds of techniques, including machine learning, rue mining, and natural language processing. After that, this paper introduced the general process of online reviews mining. Finally, this paper summarizes the data sources in this field, and found that Amazon was the most commonly used website, while China's e-websites were getting more and more attention these years.

References

[1] Jamshidi, Soheil, et al., " Trojan Horses in Amazon's Castle: Understanding the Incentivized Online Reviews", "2018 IEEE/ ACM INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING (ASONAM)", 2018, pp.335-342.

[2] Brzezinski, J, "The handbook of data mining", "INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION", 2005, Vol.18, No.2, pp.233-234.

[3] Guo, Yue, et al.," Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation", "TOURISM MANAGEMENT", 2017, Vol.59, pp.467-483.

[4] Archak, Nikolay, et al., "Deriving the Pricing Power of Product Features by Mining Consumer Reviews", "MANAGEMENT SCIENCE", August 2011, Vol. 57, No. 8. pp.1485–1509.

[5] Varathan, Kasturi Dewi, et al., "Comparative Opinion Mining: A Review", "JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY", APR 2017, Vol.68, No.4, pp.811-829.

[6] Djenouri, Djamel, et al., "Machine Learning for Smart Building Applications: Review and Taxonomy", "ACM COMPUTING SURVEYS", 2019, Vol.52, No.2, pp.

[7] Liu, Q., et al., " Chinese comparative sentence identification based on the combination of rules and statistics. In Advanced Data Mining and Applications", "Berlin Heidelberg: Springer", 2013, pp.300–310.

[8] Chen, Li, et al., "Comparison of feature-level learning methods for mining online consumer reviews", "Expert Systems with Applications", 2012, Vol., No.39, pp. 9588–9601.

[9] Xu, Kaiquan, et al., "Mining comparative opinions from customer reviews for Competitive Intelligence", "DECISION SUPPORT SYSTEMS", MAR 2011, Vol.50, No.4, pp.743–754.