Expert Systems with Applications 39 (2012) 11870–11879

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Discovering business intelligence from online product reviews: A rule-induction framework

Wingyan Chung [a,*], Tzu-Liang (Bill) Tseng [b]

[a] UNC Fayetteville State University, School of Business and Economics, 1200 Murchison Road, Fayetteville, NC 28301, USA
[b] Department of Industrial, Manufacturing and Systems Engineering, The University of Texas at El Paso, 500 West University Avenue, El Paso, TX 79968, USA

ARTICLE INFO

Keywords:
E-commerce
Online reviews
Data mining
Text mining
Association rule mining
Rough set theory
Business intelligence
Online reputation

ABSTRACT

Online product reviews are a major source of business intelligence (BI) that helps managers and marketers understand customers' concerns and interests. The large volume of review data makes it difficult to manually analyze customers' concerns. Automated tools have emerged to facilitate this analysis, however most lack the capability of extracting the relationships between the reviews' rich expressions and the customer ratings. Managers and marketers often resort to manually read through voluminous reviews to find the relationships. To address these challenges, we propose the development of a new class of BI systems based on rough set theory, inductive rule learning, and information retrieval methods. We developed a new framework for designing BI systems that extract the relationship between the customer ratings and their reviews. Using reviews of different products from Amazon.com, we conducted both qualitative and quantitative experiments to evaluate the performance of a BI system developed based on the framework. The results indicate that the system achieved high accuracy and coverage related to rule quality, and produced interesting and informative rules with high support and confidence values. The findings have important implications for market sentiment analysis and e-commerce reputation management.

1. Introduction

As e-commerce supports higher interactivity among users with Web 2.0 applications, user-generated content posted on these sites is growing significantly. Users not only consume Web content, but also produce massive data of their participation, often affecting other users' decisions. A study finds that more than three-quarters of the 2078 users reported that online product reviews had a significant influence on their purchase decisions (comScore, 2007). These online product reviews contain descriptions about user preferences, comments, and recommendations that serve as a major source of business intelligence (BI), helping managers and marketers to better understand customers. Management scholar Peter Drucker emphasizes that "what is value to the customer" may be the most important question to answer in order to realize a business's mission and purpose (Drucker, 2003). However, the large volume of online product review data creates significant information overload problems (Bowman, Danzig, Manber, & Schwartz, 1994), making it difficult to discover BI from the reviews and to analyze customer concerns.

Two major pieces of information available in each online review are its textual content and the numerical rating, which respectively indicate the aspects of customer concerns and the customer sentiment. However, neither of these two alone provides the full account of a product's real "value" (Drucker, 2003), which is the true explanation of the customer's satisfaction. An important task of a manager is therefore to correlate between the numerical ratings and the textual content of the reviews in order to understand what the customer values in a product. This task is typically done by manually reading and extracting key phrases or words that indicate customer concerns and by manually relating between the extracted phrases and the numerical ratings. Despite its usefulness, such analysis is time-consuming and does not scale up to the rapidly growing online reviews. Automated tools and techniques have been proposed to analyze online reviews. These works try to study the reviews' impact on sales (Zhu & Zhang, 2010), to recommend products (Aciar, Zhang, Simoff, & Debenham, 2007), to calculate the utility of the reviews (Ding & Liu, 2007), to identify key product features (Zhang, 2008), to detect false reviews (Jindal & Liu, 2007), and to summarize review content (Zhuang, Jing, & Zhu, 2006). However, research that supports the managerial task of correlating between the numerical ratings and textual content of the reviews is not widely found. The problem of how the reviews' textual content contributes to the numerical ratings is thus not widely

* Corresponding author.
E-mail address: wchung@uncfsu.edu (W. Chung).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.02.059

addressed. Understanding this correlation in large amounts of online review data could help e-commerce managers to make effective decisions on brand management, product promotion, and reputation management.

In this paper, we discuss existing works on analyzing online product reviews and critically review these existing approaches. Following a design science paradigm (Hevner, March, Park, & Ram, 2004), we develop a new framework for designing a new class of BI systems that correlate the textual content and the numerical ratings of online product reviews. In contrast to behavioral science, the design science paradigm was chosen because it emphasizes on building and evaluating innovative artifacts that address the analysis needs of e-commerce managers, marketers, and BI practitioners. In the process of building our artifacts, we drew upon the theoretical and computational foundations of data mining (Liu, 2007; Pawlak, 1982) and information retrieval (Salton, 1989; Salton, Wong, & Yang, 1975). Based on the rough set theory and inductive rule mining methods used in the framework, we developed as an instantiation a system for extracting the relationship between hundreds of customer ratings and their corresponding textual reviews posted on Amazon.com's Web site. To demonstrate the applicability of the system, two data mining methods were implemented to extract automatically decision rules to guide the understanding of the relationship. Using quantitative and qualitative experiments, we empirically tested the system that was configured under different methods and settings. The system's enhanced performance was demonstrated over different types of products' online reviews. The results have strong implications for brand management and online market sentiment analysis.

2. BI research and online product review analysis

The term "business intelligence" (BI) is defined as the acquisition, interpretation, collation, analysis, and exploitation of information in business (Chung, Chen, & Nunamaker, 2005; Davies, 2002). BI systems enable organizations to understand their internal and external environments. To support understanding of internal data, one class of BI systems (Carvalho & Ferreira, 2001) manipulates massive operational data to extract essential business information. Some examples of these systems are decision support systems, executive information systems, online-analytical processing (OLAP), data warehouses and data mining systems that are built upon database management systems to reveal hidden trends and patterns. Another class of BI systems tries to systematically collect and analyze information from the external business environment to assist in organizational decision making. They gather information from public sources such as the Internet and provide insights into various knowledge discovery processes. Examples include customer review analysis, Web search log mining, and opinion mining. Technologies to support the second class of BI systems are in general less matured than those for the first class mentioned above.

Research and industry developments in BI have been growing in recent years due to the growing amounts of business data and the widespread use of the Internet as a medium of communication. Many of these works develop Web-based business intelligence systems to assist in data analysis and decision-making (Chung & Chen, 2009; Lawton, 2006). The most recent trends in BI concern about user-generated data analysis. Opinion and sentiment are extracted from large amounts of textual data to facilitate managerial decision-making. There is much room for the design science research community to contribute to this area.

Design science is concerned with the creation and evaluation of new information technology (IT) artifacts with a goal of meeting business needs (Hevner, March, Park, & Ram, 2004). Hevner et al., (2004) provided seven guidelines for design-science research: design as an artifact, problem relevance, design evaluation, research contributions, research rigor, design as a search, and communication of research. While most BI systems are IT artifacts and the BI domain provides sufficient challenges to satisfy the guideline for "problem relevance," there is a need for design science research to address the remaining guidelines mentioned above. The design-science-based information systems research can contribute significantly to BI because the design and evaluation of new IT artifacts within organizational and managerial context can bring new insights about BI technologies, practices, and challenges (Chen, Chiang, & Storey, 2009). Information systems and technologies utilizing data/text/Web mining techniques have been developed to analyze BI from online product reviews.

2.1. Online product review analysis

Data mining and machine learning techniques identify patterns from large amounts of data using statistical and heuristics methods (Mitchell, 1997). These techniques have been applied to a large number of domains, such as business stakeholder classification (Chung, Chen, & Reid, 2009), crime analysis (Chen et al., 2004), and medical data prediction (Brown et al., 2000). Text mining applies data mining techniques to analyzing unstructured, text data (Trybula, 1999). Web mining further uses data and text mining techniques to extract the content, structure, and usage information from Web data (Kosala & Blockeel, 2000). These techniques are applied to a variety of online product review analyses. For example, Zhu and Zhang studied the reviews' impact on sales of online games and found that these reviews are more influential for less popular games and games whose players have greater Internet experience (Zhu & Zhang, 2010). Yan et al. developed a dictionary-based method to represent review textual features and used machine-learning techniques to classify the review sentiment (Dang, Zhang, & Chen, 2010). Zhang used lexical similarity, shallow syntactic features, and lexical subjectivity clues to distinguish useful from useless reviews (Zhang, 2008). To address ambiguity in review text, Ding and Liu used linguistic rules to determine the semantic orientations of words in customer reviews (Ding & Liu, 2007). To support spam detection, reviews were categorized into false opinion (overly positive or negative comments), brand reviews (based only on brand but not product), and non-reviews (advertisements without comment) (Jindal & Liu, 2007). Besides, product recommendation was done through mapping automatically each sentence of a review into a manually-created ontology (Aciar et al., 2007). Different actors were considered in (Zhuang et al., 2006) to summarize movie reviews using WordNet and statistical analysis.

While much of previous research tried to extract sentiment and opinion and to distinguish among different types of product reviews, identifying rules and patterns from online reviews is not widely studied. According to the Merriam-Webster Dictionary, a "pattern" is defined as "a discernible coherent system based on the intended interrelationship of component parts (pattern, 2010)." A rule is defined as "a usually valid generalization" and can be considered a specific type of patterns. Discovering rules from data is a major task in data mining (Liu, 2007), in which a rule is often specified as an association in the form "antecedents ⇒ consequents" such that the left-hand side (e.g., words used in online reviews) of the rule determines the right-hand side (e.g., product rating). These rules are a specific type of patterns that represent associations among any extracted entities. Such rules and patterns often represent valuable knowledge assets in organizations (e.g., tacit knowledge as discussed in p.112 of (Alavi & Leidner, 2001)), providing insights for managerial decision making.