Chapter 5
Spidering Scripts for Opinion Monitoring

Antonella Capriello
University of Eastern Piedmont, Italy

Piercarlo Rossi
University of Eastern Piedmont, Italy

ABSTRACT

With the advent of Web 2.0 technologies, online forms of communication are rich sources of data to study socio-economic growth patterns and consumer behaviours. In this research field, the more robust development of data mining and opinion monitoring depends on fully automating data collection to monitor the evolution of customer opinions and preferences in real time. Although web crawlers or spiders can assist researchers in an innovative and effective way, this data collection approach could give rise to ethical concerns on the cost of web crawling processes and on data protection and privacy. With a focus on opinion monitoring, the chapter aims to discuss the ethical and legal issues of data mining in relation to spidering scripts. This contribution proposes a detailed analysis of the ethical and legal aspects of online data collection by comparing existing legislations. For illustrative purposes, a spidering software is presented to discuss its potential and explore ethical solutions in the data-mining sphere.

INTRODUCTION

The development of Web 2.0 technologies has provided innovative platforms for consumers; through social media applications, they can share their opinions on products, exchange information, learn through knowledge transfer, socialize and participate in co-creation and innovation processes. Aside from the developments in technology that have led to the emergence of the described phenomenon, online information is a rich data source that can facilitate measuring actual customer and enterprise behaviours. Online reviews, chat rooms, blogs and virtual brand communities have become important venues for sharing product and service opinions since consumers assign a high value to digital word-of-mouth messages.

Interest in opinion monitoring has significantly increased among marketing scholars: Senecal and Nantel (2004), Sen and Lerman (2007) and

DOI: 10.4018/978-1-4666-4078-8.ch005

Smith, Menom, and Sivakumar (2006) analyse the usefulness of online reviews in customer decision-making processes, while Duan, Gu, and Whinston (2008), Chen, Wang, and Xie (2011) and Liu (2006) study their impact on sales and firm marketing strategies. Capriello (2012) emphasizes the key role of opinion monitoring for strategy formulation and implementation in tourism services. More specifically, she underlines the instrumental function of online travel reviews in both online marketing communication activities and product design planning.

In addition to traditional field research methods, marketing researchers have recognized the need to analyse customer-generated content, but no single approach prevails as a best method to analyse such online qualitative data. With a focus on online consumer groups, Kozinets (2002) proposes 'netnography' as a methodology that adapts ethnography research techniques to the study of cultures and communities emerging through computer-mediated communications using publicly available information from online forums. This technique allows marketing researchers to gain insights into consumer experiences in a less costly, time-consuming or intrusive way than focus groups and personal interviews (Kozinets, 2002). In analysing blogs, Woodside, Sood, and Miller (2008) propose the storytelling method, as this approach allows creating narrative interpretation maps to gain insights on consumer experiences. Kwortnik and Ross (2007) combine the analysis of online forums with consumer ethnographic and introspective vacation planning tasks. Pan et al. (2007) rely on word frequencies and semantic network analysis to investigate consumer opinions in relation to a tourist destination, while Mason and Davis (2007) and Crotts et al. (2009) purport stance-shift analysis as an alternative approach to quantitative content analysis focusing on key language pattern identification.

Despite the divergence in qualitative data analysis methods and methodologies, Capriello et al. (forthcoming) highlight that the future development of data mining and opinion monitoring depends on fully automating the data collection and analysis processes to monitor the evolution of customer opinions and preferences in real time.

In this perspective, web crawlers and spiders are effective tools for qualitative data collection. These computer programs methodically and iteratively browse the web and webpages. Furthermore, spidering software can aid web data collection and extract useful information for researchers and practitioners.

However, the adoption of web crawling may raise some ethical and legal issues concerning key aspects such as privacy and copyright since spidering software can be used without the awareness of those being scrutinized. These tools for gathering information could consequently pose a threat to the protection of data relating to the personal information of consumers who generate the content as well as the protection of copyright in terms of unauthorized access to the databases and websites that host, categorize and reproduce these contents. This chapter aims to discuss these issues as well as the potential solution offered by the spidering software developed by the authors with the support of COMDATA[1].

BACKGROUND

Web Crawling and Ethical Issues

The evolution of internet communications, mainly social media websites where consumers generate their own content, has created a great deal of publicly available information. In terms of market research, these websites enable consumers to express their beliefs, preferences and ratings on products and services, which is vital information for industry and policy-makers in improving both the product offer and market efficiency. Consumer generated content is vast and is continuously increasing with millions of daily new insertions on social media websites.