Stages of Knowledge Discovery in E-Commerce Sites

Christophe Giraud-Carrier
Brigham Young University, USA

Matthew Smith
Brigham Young University, USA

INTRODUCTION

With the growth and wide availability of the Internet, most retailers have successfully added the Web to their other, more traditional distribution channels (e.g., stores, mailings). For many companies, the Web channel starts off as little more than an online catalog tied to a secure electronic point of sale. Although valuable in its own right, such use of the Web falls short of some of the unique possibilities it offers for intelligent marketing. Consider the following intrinsic differences between physical, brick-and-mortar stores, and online, Web-based stores. Physical stores are rather static and mostly customer-blind. In particular, 1) the store's layout and content are the same for all customers, 2) changes to layout and/or content are generally costly, and 3) visits are not traceable except for limited sale's data, such as what was bought, when it was bought and by what method of payment. Online stores or commercial Web sites, on the other hand, are naturally dynamic and customer-aware. Indeed, 1) layout and content can be modified easily and cheaply, 2) layout and content can be tailored to individual visitors, and 3) every visit automatically generates a rich trail of information about the customer's experience (e.g., visit duration, pages viewed, items bought if any, etc.), and possibly about the customer's persona (e.g., demographics gathered through an online questionnaire at registration time).

With such flexibility and nearly everything traceable and measurable, the Web is a marketer's dream come true. Although data-independent initiatives, such as offering social interactions (e.g., user forums) or providing virtual versions of physical stores (e.g., displays, lighting, music) (Oberbeck, 2004), can clearly enhance the user experience, the full benefit of the emerging and growing Web channel belongs to those who both gather and adequately leverage the rich information it provides.

BACKGROUND

Web mining emerged in the late 1990's as the branch of data mining concerned with the mining of data and structure on the Web. As pointed out early in the game, the transfer from traditional data mining to mining on the Web is not without significant challenges (Monticino, 1998; Spiliopoulou, 1999). Yet, many business analysts were just as quick to argue that the potential benefits far outweigh the costs (Greening, 2000; Edelstein, 2001), and researchers began developing specialized techniques and tools (Spiliopoulou & Pohle, 2001).

There have traditionally been three sub-areas of Web mining, based upon the type of data being mined: Web usage mining focuses on extracting information from users' interactions with a Web site; Web content mining focuses on extracting knowledge from the textual (and more recently, multimedia) content of Web pages; and Web structure mining focuses on discovering patterns of connection among Web pages (Kosala & Blockeel, 2000). A number of survey papers and texts have been dedicated to descriptions of various Web mining techniques and applications, as well as relevant research issues (Han & Chang, 2002; Kolari & Joshi, 2004; Scime, 2005; Liu, 2007).

In this chapter, we depart a little from this research-oriented approach. Indeed, we do not discuss the different types of data that may be mined on the Web, but rather highlight stages in the analysis of Web data, from simplest to most elaborate, so that business users may appreciate the potential of such analysis and have an implementation roadmap. At each stage, different

types of data (usage, content and structure) may be, and indeed are, used.

MAIN FOCUS

In the context of e-commerce, Web mining lends itself naturally to a staged approach, where moving from one stage to the next requires increasing sophistication, but also produces increasing return-on-investment. The first stage is limited to the analysis of the direct interaction of the user with the site; the second stage introduces behavioral information; and the third stage enables personalization of the user's experience. We examine each in turn.

Stage 1: Clickstream Analysis

The amount of data found in Web server logs is enormous and clearly evades direct human interpretation. Yet, it is rich in potential, making Web server logs the readiest source of data to analyze on a Web site (Srivastava et al., 2000; Fu et al., 2000; Moe & Fader, 2004). Web log analysis tools, such as AWStats, The Webalizer, SiteCatalyst, ClickTracks, Google Analytics, and NetTracker have been designed specifically to summarize and interpret Web server log data, allowing marketers to gain basic knowledge about e-commerce customers' activities, including unique number of visitors and hits, visit duration, visitors' paths through the site (i.e., clickstream), visitors' host and domain, search engine referrals, robot or crawler visits, visitors' browser and operating system, and search keywords used. These clickstream analysis reports may provide insight into business questions such as:

- Where do most visitors come from?
- What proportion of visitors come from a direct link or bookmark, a search engine, or a partner Web site (if any)?
- Which search engines (e.g., Google, Yahoo, MSN, etc.) and search terms are most frequently used?
- How long do visitors stay on the Web site?
- How many pages do visitors see on average?
- Which pages are most popular?
- How does Web site activity evolve throughout the day, week or month?
- From which pages do visitors commonly exit the Web site?

This information in turn helps e-retailers better understand the dynamics of customers' interactions with online offerings, and aids in such decisions as: which referrer to invest in, which pages to remove or replace, which pages to improve, etc. For instance, if clickstream analysis shows that a substantial number of visitors are accessing content several clicks deep into the Web site, then it might be valuable to make that content more accessible to visitors (e.g., maintaining and linking to "Top Sellers," "Most Wished for Items," or "Most Popular Items" pages on the home page).

Stage 2: Behavior Analysis

In general, transactional data and order information are not stored directly in standard Web server logs. Yet, both are essential in discovering patterns of buying and non-buying customers. The next stage of knowledge discovery requires linking clickstream data to order information. With adequate design, a host of new business-relevant questions may be answered when the front-end clickstream data is linked to, and mined together with, the back-end transactional data, allowing marketers to take further control of their e-commerce activity (Mobasher et al., 1996; Gomory et al., 1999; Rusmevichientong et al., 2004). The following are a few classical examples.

- What is the conversion rate (i.e., how many Web site visitors become buying customers)?
- How many would-be customers begin shopping (i.e., partially filling up their shopping cart) but drop out before proceeding to or completing check-out?
- How well did special offer *X* do (i.e., how much revenue vs. interest did it generate)?
- Who buys product *P*?
- Who are the most profitable customers?
- What is being bought by whom?

Answers to these questions help e-retailers better understand their customers' buying behavior as well as the value of their offerings. This information, in turn, leads to considerations such as what products to focus on, how to turn browsers into customers, how